



---

# Aerospace Vehicle Control And Perception Systems

*Version 3.0*

Jordan D. Larson

Copyright ©Jordan D. Larson 2026

All rights reserved. No parts of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the author.

Web page: <https://github.com/drjdlarson/aerospace-vehicle-control-and-perception-systems>

Digital Object Identifier (DOI): 10.5281/zenodo.10836294

MATLAB ©is a registered trademark of The MathWorks, Inc.

# Contents

<b>Contents</b>	<b>1</b>
<b>0 Preface and Acknowledgments</b>	<b>6</b>
0.1 Preface: Model-Based Design . . . . .	6
0.2 Acknowledgments . . . . .	11
<b>I Dynamical Systems and Control Theory</b>	<b>12</b>
<b>1 Dynamical Systems Theory</b>	<b>13</b>
1.1 Introduction to Dynamical Systems . . . . .	13
1.2 Linearization of Continuous-Time Time-Invariant Systems . . . . .	22
1.3 Modal Analysis of Continuous-Time LTI Systems . . . . .	26
1.4 Continuous-Time SISO LTI System Responses . . . . .	31
1.5 Continuous-Time MIMO LTI System Responses . . . . .	54
1.6 Discrete-Time Linear Systems . . . . .	64
<b>2 Continuous-Time LTI Feedback Control Theory</b>	<b>72</b>
2.1 Introduction to Continuous-Time Control Systems . . . . .	72
2.2 Classical Feedback Control System Analysis . . . . .	78
2.3 Classical Feedback Control System Design . . . . .	87
2.4 Classical Feedback Control System Performance Constraints . . . . .	110
2.5 MIMO LTI Feedback Control System Analysis . . . . .	117
2.6 MIMO LTI Command-Tracking Control System Design . . . . .	131
<b>3 Continuous-Time LTI Optimal Control Theory</b>	<b>136</b>
3.1 Introduction to Optimal Control . . . . .	136

<i>CONTENTS</i>	2
-----------------	---

3.2	Introductory Energy-, Time-, and Fuel-Optimal Control . . . . .	143
3.3	Convex Optimization in LTI Control . . . . .	150
3.4	$\mathcal{H}_2$ Optimal Control . . . . .	154
3.5	$\mathcal{H}_{\infty}$ Optimal Control . . . . .	160
<b>4</b>	<b>Discrete-Time Linear Control Theory</b>	<b>170</b>
4.1	Discrete-Time Linear Feedback Control Systems . . . . .	170
4.2	Discrete-Time Linear-Quadratic Regulator . . . . .	176
4.3	Iterative and Extended Linear-Quadratic Regulator . . . . .	180
4.4	Receding-Horizon Linear-Quadratic Regulator . . . . .	187
<b>5</b>	<b>Non-LTI Systems and Adaptive Control Theory</b>	<b>190</b>
5.1	Time-Varying Systems Theory . . . . .	190
5.2	Gain-Scheduled Adaptive Control . . . . .	195
5.3	Linear-in-Control Systems and Dynamic Inversion . . . . .	201
5.4	Model Reference Adaptive Control . . . . .	206
<b>II</b>	<b>Aerospace Vehicle Dynamics and Control Systems</b>	<b>211</b>
<b>6</b>	<b>Body Dynamics</b>	<b>212</b>
6.1	Reference Frames and Transformations . . . . .	212
6.2	Point-Mass Dynamics . . . . .	222
6.3	Rigid-Body Dynamics . . . . .	227
6.4	Elastic-Body Dynamics . . . . .	230
<b>7</b>	<b>General Aerospace Vehicle Dynamics</b>	<b>246</b>
7.1	Aerospace Vehicle Reference Frames and Rotations . . . . .	246
7.2	Point-Mass Aerospace Vehicle Dynamics . . . . .	260
7.3	Rigid-Body Aerospace Vehicle Dynamics . . . . .	271
7.4	Elastic-Body Aerospace Vehicle Dynamics . . . . .	281
7.5	Atmospheric Effects on Aerospace Vehicle Dynamics . . . . .	296
<b>8</b>	<b>General Aerospace Vehicle Guidance and Control Systems</b>	<b>309</b>
8.1	Aerospace Vehicle Actuation Systems . . . . .	309
8.2	Aerospace Vehicle Guidance and Control Systems . . . . .	317
8.3	Intercept and Rendezvous Guidance Systems . . . . .	320
8.4	Obstacle Avoidance Guidance Systems . . . . .	327
<b>9</b>	<b>Airplane Dynamics and Control Systems</b>	<b>328</b>
9.1	Introduction to Fixed-Wing Vehicles . . . . .	328
9.2	Point-Mass Dynamics for Airplanes . . . . .	335
9.3	Airplane Hold and Landing Guidance Systems . . . . .	339
9.4	Airplane Trim Analysis . . . . .	347

9.5 Rigid Airplane Dynamics and Stability . . . . .	360
9.6 Elastic Airplane Dynamics . . . . .	374
9.7 Airplane Attitude Control Systems . . . . .	382
<b>10 Helicopter Dynamics and Control Systems</b>	<b>396</b>
10.1 Introduction to Helicopters . . . . .	396
10.2 Point-Mass Dynamics for Helicopters . . . . .	399
10.3 Rigid Helicopter Dynamics and Stability . . . . .	399
10.4 Helicopter Attitude Control Systems . . . . .	400
<b>11 Orbital Vehicle Dynamics and Control Systems</b>	<b>401</b>
11.1 Introduction to Orbital Vehicles . . . . .	401
11.2 Point-Mass Dynamics for Orbital Vehicles . . . . .	403
11.3 Orbital Maneuvers . . . . .	413
11.4 Rigid Orbital Vehicle Dynamics and Stability . . . . .	413
11.5 Orbital Vehicle Attitude Control Systems . . . . .	425
<b>12 Ballistic Vehicle Dynamics and Control Systems</b>	<b>438</b>
12.1 Introduction to Orbital and Ballistic Vehicles . . . . .	438
12.2 Point-Mass Dynamics for Ballistic Vehicles . . . . .	440
12.3 Ascent and Descent Guidance Systems . . . . .	441
12.4 Rigid Ballistic Vehicle Dynamics and Stability . . . . .	442
12.5 Ballistic Vehicle Attitude Control Systems . . . . .	443
<b>III Probability and Perception Theory</b>	<b>444</b>
<b>13 Probability Theory</b>	<b>445</b>
13.1 Introduction to Probability Theory . . . . .	445
13.2 Random Variables . . . . .	449
13.3 Random Vectors . . . . .	459
13.4 Random Processes and Dynamical Systems . . . . .	471
13.5 Random Finite Sets . . . . .	484
<b>14 Optimal Parameter Estimation and Detection Theory</b>	<b>491</b>
14.1 Introduction to Optimal Parameter Estimation . . . . .	491
14.2 Batch Least-Squares Parameter Estimation . . . . .	498
14.3 Bayesian Optimal Parameter Estimation . . . . .	510
14.4 Hypothesis Testing and Optimal Detection Theory . . . . .	525
<b>15 Optimal Linear State Estimation Theory</b>	<b>535</b>
15.1 Introduction to Optimal State Estimation . . . . .	535
15.2 Discrete-Time Kalman Filtering and Smoothing . . . . .	543
15.3 Continuous-Time Kalman Filtering . . . . .	563

15.4 Multi-Modal and Heavy-Tailed Kalman Filtering . . . . .	567
15.5 Multiple Model Filtering . . . . .	568
15.6 Stochastic Linear-Quadratic Optimal Control . . . . .	574
<b>16 Nonlinear Bayesian State Estimation Theory</b>	<b>579</b>
16.1 Introduction to Nonlinear Bayesian State Estimation . . . . .	579
16.2 Extended Kalman Filtering and Smoothing . . . . .	581
16.3 Statistical Linearization and State Estimation . . . . .	595
16.4 Sigma-Point Kalman Filtering and Smoothing . . . . .	602
16.5 Error-State Kalman Filtering . . . . .	610
16.6 Equivariant Kalman Filtering . . . . .	613
16.7 Variational Bayes Kalman Filtering . . . . .	613
16.8 Particle Filtering . . . . .	613
<b>17 Stochastic Motion Models</b>	<b>628</b>
17.1 Uncoupled-Coordinate Stochastic Motion Models . . . . .	628
17.2 Turning Stochastic Motion Models . . . . .	633
17.3 Orbital and Ballistic Stochastic Motion Models . . . . .	639
<b>IV Aerospace Vehicle Sensors and Perception Systems</b>	<b>650</b>
<b>18 Sensor and Data Systems</b>	<b>651</b>
18.1 Inertial Sensors . . . . .	651
18.2 Electromagnetic Sensors . . . . .	658
18.3 Radionavigation Systems . . . . .	664
18.4 Air Data Systems . . . . .	673
18.5 Clocks and Timing Systems . . . . .	675
<b>19 Positioning Systems</b>	<b>681</b>
19.1 Introduction to Positioning Systems . . . . .	681
19.2 Pseudorange Positioning Systems . . . . .	692
19.3 Phase-Based Positioning Systems . . . . .	700
19.4 Satellite-Based Positioning Systems . . . . .	706
19.5 Map-Based Positioning Systems . . . . .	713
<b>20 Attitude Determination Systems</b>	<b>716</b>
20.1 Introduction to Attitude Determination Systems . . . . .	716
20.2 Attitude and Heading Reference Systems . . . . .	720
20.3 Multi-Source Direct Attitude Determination Systems . . . . .	729
<b>21 Navigation Systems</b>	<b>733</b>
21.1 Introduction to Navigation Systems . . . . .	733
21.2 Inertial Navigation Systems . . . . .	735

<i>CONTENTS</i>	5
21.3 Aided Inertial Navigation Systems . . . . .	741
21.4 Simultaneous Localization and Mapping Systems . . . . .	747
<b>22 Object Tracking Systems</b>	<b>748</b>
22.1 Introduction to Object Tracking Systems . . . . .	748
22.2 Probabilistic Data Association Filtering . . . . .	761
22.3 Multi-Hypothesis Tracking . . . . .	767
22.4 Probability Hypothesis Density Filtering . . . . .	771
22.5 Bernoulli Filtering . . . . .	782
22.6 Extended Object Tracking . . . . .	784
<b>23 Monitoring Systems</b>	<b>785</b>
23.1 Fuel and Battery Monitoring Systems . . . . .	785
23.2 Fault Monitoring Systems . . . . .	788
23.3 Integrity Monitoring Systems . . . . .	793
<b>V Appendices</b>	<b>796</b>
<b>A Fundamental Mathematical Concepts</b>	<b>797</b>
A.1 Set Theory . . . . .	797
A.2 Linear Algebra . . . . .	799
A.3 Numerical Methods . . . . .	810
A.4 Complex Analysis and Stability Criterion . . . . .	812
<b>B Airplane Stability and Control Derivative Models</b>	<b>818</b>
B.1 Finite-Wing Theory and Component Build-Up Model . . . . .	818
B.2 Longitudinal Stability and Control Derivatives . . . . .	825
B.3 Lateral-Directional Stability and Control Derivatives . . . . .	834
<b>Index</b>	<b>841</b>

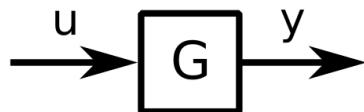
---

# Preface and Acknowledgments

## 0.1 Preface: Model-Based Design

**Model-based design** provides a common design environment facilitating communication, data analysis, design verification and validation (V&V) between various development groups for the entire system. Because modern computers allow the use of complex structures and extensive software code, engineers typically use model-based design with advanced simulation tools to provide rapid prototyping of new systems, including aerospace systems. In this way, engineers can locate and correct errors early in system design, when the impact of design modifications are minimized. Furthermore, design reuse for upgraded or expanded systems is made easier. However, as model-based design uses a coverall approach to standard embedded systems development, the time it takes to port between modeling software and embedded systems can outweigh the temporal value for alternative lab-based design.

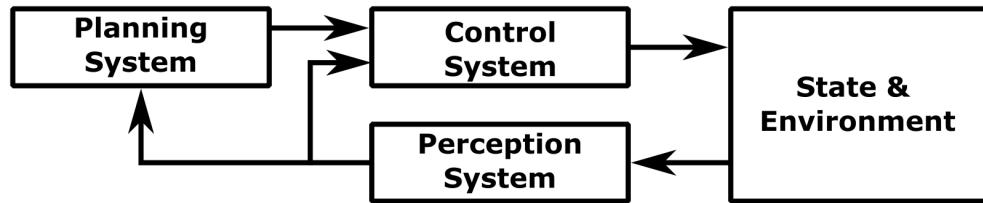
Model-based design software typically uses graphical modeling tools to provide a very generic and unified graphical modeling environment by reducing the complexity of model designs by breaking them into hierarchies of individual design blocks. This helps design engineers to conceptualize the entire system, typically with a **graphical user interface (GUI)**. The connection between signals and systems in these GUIs are typically depicted using **block diagrams**, for which each “block” represents a system and each “arrow” represents a signal. An example of a basic block diagram can be drawn as



which consists of a system  $G$  with a single input signal  $u$  and single output signal  $y$ .

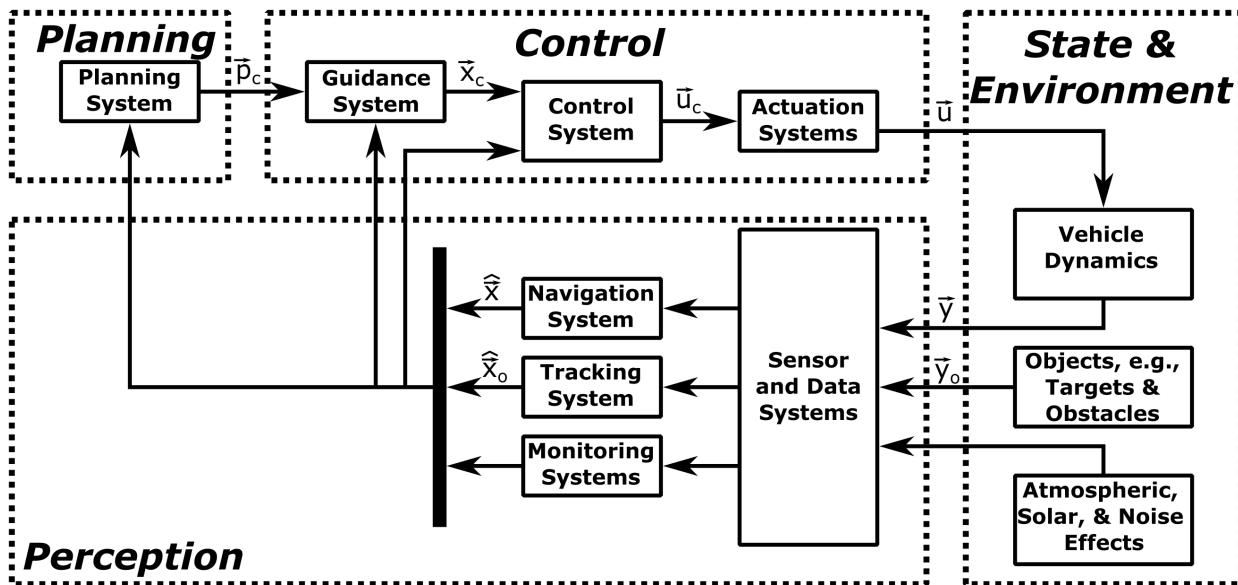
## Aerospace Planning, Control, and Perception Systems

In modern aerospace software, the design of the **planning, control, and perception** systems is fundamental to the operations of aerospace vehicles. These three software systems can be understood as connected together through the following



where the planning system determines what trajectory/path the aerospace vehicle should travel (or fly) to reach its current target, the control system executes the trajectory/path for the aerospace vehicle by actuating the control inputs while the perception system(s) senses and estimates the state of the aerospace vehicle, the state of its target(s), and the state of the environment. Thus, one can see that for aerospace vehicles, perception systems provide updates to the planning system for the trajectory/path as well as a (guidance and) control system to follow the planned trajectory.

Furthermore, one can represent the operation of an aerospace vehicle as an aerospace **planning, control, and perception** system in the following block diagram.



where the control block contains both the guidance, control, and actuation systems while the perception block contains the navigation, tracking, monitoring, and sensor and data systems. Of these systems, the majority of the content in this textbook is focused on the algorithm design of the guidance, control, navigation, tracking,

and monitoring systems for aerospace vehicles with relevant content for modeling aerospace vehicle actuation and sensor and data systems is provided to understand the design of the full control and perception system from a model-based perspective.

## Model-Based Design of Control and Perception Systems

During the 1920s, electrical control and perception systems were first implemented in electric powerplant industries. These large process facilities used process controllers for regulating continuous variables, e.g. temperature, pressure, and flow rate, with electrical relays. Throughout the 1930s and 1940s, control and perception systems began to see use in the automotive and aviation industries. In the 1950s and 1960s, the space race generated interest in embedded control and perception systems. Here, engineers constructed control and perception systems that could be part of the end product, e.g., engine control units and flight simulators. In the 1970s, computer-based control and perception systems were introduced and became standard bringing about a drastic shift in control and perception system design and has led to the modern use of model-based software design for control and perception systems.

Model-based design of control and perception systems can be described as four steps. Step one is **plant modeling** which consists of identifying the system model to be controlled, i.e., the **plant**, which comes from the “powerplant” in early days of control systems. Plant modeling can be based on first principles models or machine learning. First principles modeling implements a physics-derived mathematical model for the plant dynamics, e.g. Newton-Euler EOMs. Machine learning processes raw data from a real-world system and uses learning algorithms to identify the data-driven model for the plant dynamics. In this textbook, the plant is the aerospace vehicle.

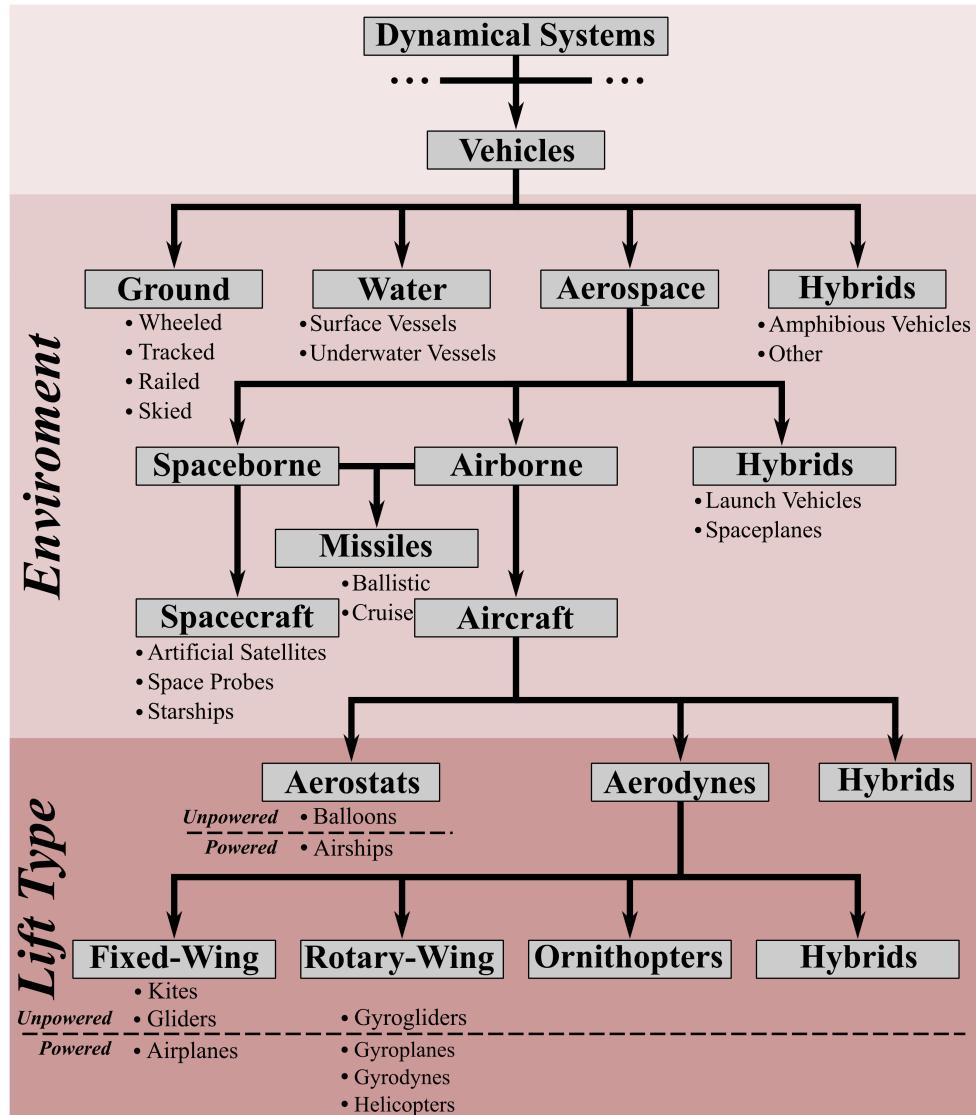
Step two is **software system design** which consists of **software system analysis** and **software system synthesis**. First, the mathematical aerospace vehicle model developed in the first step is used to identify suitable **design requirements** for the controlled aerospace vehicle. Based on these identified design requirements, a control strategy, including actuation, is chosen and the corresponding perception system requirements, including sensing, is defined and a perception strategy is chosen. Then, the software system composed of the control and perception systems are synthesized producing an explicit **controller**, i.e., a mathematical algorithm for computing the aerospace vehicle inputs automatically based on external commands, vehicle information, and environmental information, and/or an explicit **estimator**, i.e., a mathematical algorithm for computing the vehicle and environment information automatically from sensor data.

Step three is **system simulation** which may be performed as multiple simulations with lower/higher fidelity aerospace vehicle and sensor models as well as offline/real-time simulations. These simulations allow specification, requirements, and modeling errors to be found immediately, rather than later in the overall system design effort. Real-time simulation can be done by testing the control system, the perception system, sensors, and aerospace vehicle on a real-time modeling computer, also known as **software-in-the-loop (SIL)** simulation. Sometimes the software system may also be implemented on the embedded system hardware and simulated with real-time aerospace vehicle and sensor models, also known as **hardware-in-the-loop (HIL)** simulation. Step four is **validation and verification (V&V)** of the control and perception system. Here, the software systems are implemented on the embedded system hardware and tested with the actual aerospace vehicle and sensors. As the software systems are highly unlikely to work exactly the same on actual aerospace vehicle and sensors as in simulation, an iterative debugging process is carried out by analyzing the test results and updating the control system and perception system designs until the design

requirements are met in actuality.

This textbook approaches the design of components of control and perception systems for aerospace vehicles from a model-based theoretic based on the plant and sensor models. The control and perception system design primarily relies on linear, time-invariant systems theory for deterministic, uncertain, and stochastic systems with extensions provided for linear, parameter-varying and nonlinear systems. Thus, it is vital to develop the physics-based dynamics models inherent to aerospace vehicle control and perception relying on the field of **dynamical systems theory**, i.e., the mathematical discipline studying the behavior of time-varying systems, and **classical mechanics**, i.e., the study of the motion of physical objects. When modeling the motion of physical **objects**, also known as **bodies**, using classical mechanics, one can choose between three different models. First, a **point-mass model**, also known as **point particle model**, defines an object only by its mass at a point. Thus, the point-mass model assumes that a physical object's motion is completely characterized by the motion of its center of mass. Second, a **rigid-body model** defines an object by a static distribution of mass in space. Thus, the rigid-body model assumes that an object's motion is completely characterized by the motion of its center of mass and the rotation of its structure. Third, a **deformable-body model** defines an object by a dynamic distribution of mass in space that experiences deformations caused by stresses. A sub-type of the deformable body model is the **elastic-body model** which assumes the deformations caused by stress are completely recoverable. Thus, the elastic-body model assumes that an object's motion is completely characterized by the motion of its nominal center of mass, the rotation of nominal structure, and the vibration of the body about its nominal structure. These object models applied to aerospace vehicle dynamics are all utilized in the design of different control and perception systems for aerospace vehicles.

## Aerospace Vehicle Dynamics



A **vehicle** is defined as a physical object that transports a payload, e.g. cargo, people, munitions, sensors. Vehicles are typically classified by the environment in which they operate, i.e. ground, water, flying, or amphibious. **Aerospace vehicles** are a category of vehicle encompassing aircraft, spacecraft, airborne and spaceborne missiles, and air/space hybrids. These are also known as **flight vehicles**. There are many different types of aerospace vehicles. These are separated into aerostats, e.g. balloons, airships; aerodynes which includes fixed-wing aircraft, e.g., gliders and airplanes, and rotary-wing aircraft, e.g. gyrogliders, helicopters, autodynes, gyrodyne; spacecraft, e.g., orbital “satellites” and sub-orbital; and hybrid aerospace

vehicle that operate in air and space, e.g., launch vehicles, spaceplanes, and space capsules. This textbook provides a focus on airplanes, helicopters, satellites, launch vehicles, and ballistic missiles. However, it is important to note that the dynamics and control concepts in this textbook can be applied to *any* dynamical system to varying degrees as one diverges further away from aerospace vehicles.

**Vehicle dynamics and control** is the study of the changes in a vehicle's motion due to the forces and moments applied to and by that vehicle in its environment. The motion of a vehicle is typically described by its position, orientation/attitude, and velocity. When applied in particular to aerospace vehicle, this study is called **flight dynamics and control (FDC)** and is the subject of this part of the textbook. The important differentiating factor in FDC from other vehicles is that aerospace vehicle are affected by forces and moments from only *four* external sources: gravity, propulsion, aerodynamics, and radiation pressure, whereas ground and water vehicles experience ground forces and hydrodynamics, respectively. Modeling the forces and moments from these four external sources are the differentiating factors between aerospace vehicle dynamics and other physical object dynamics.

Aerospace vehicles are designed so that the **operator** of the aerospace vehicle, also known as the **pilot**, can affect the aerodynamic and propulsive forces and moments imparted on the aerospace vehicle in order to control or affect the vehicle's motion, i.e., its dynamics. The control of a aerospace vehicle can be performed completely by a human, also known as **manual control**), completely by a computer, also known as **automatic control**, which constitutes an **autopilot**, or partially from both, also known as **semi-automatic control**. This textbook primarily discusses the design of automatic control systems for aerospace vehicles.

## 0.2 Acknowledgments

The author acknowledges his former students who assisted in proofreading this textbook.

- Aabhash Bhandari
- Adam Hallmark
- Bo Landess
- Eric Becker
- Holden McNerney
- Isaiah Newell
- Ryan Thomas
- Tuan Luong
- Vaughn Weirens

## **Part I**

# **Dynamical Systems and Control Theory**

---

# Dynamical Systems Theory

## 1.1 Introduction to Dynamical Systems

The study of **flight dynamics and control (FDC)** most generally falls under the theory of signals and systems. A **signal** is a mathematical description of how a parameter varies with time. Signals can be continuous or discrete in time, as well as continuous or discrete in the values the parameter may take. **Analog signals** are continuous-time and continuous-valued. **Digital signals** are discrete-time and discrete-valued. **Discrete-time signals** are discrete in time and continuous-valued. Signals that are continuous-time and discrete-valued rarely occur and do not have a particular name. A **system** is any process that produces output signals in response to input signals. The output signal from a system is also known as the **system response**. A system may be characterized by its different properties.

A fundamental characterization is by how the output signal or system response depends on time. A **static system** produces an output signal that is not time-dependent while a **dynamical system** produces an output signal as a time-dependent quantity that evolves according to a fixed mathematical rule, also known as the **dynamics equation**. Here the term “dynamical” is used to correspond to the mathematical abstraction which is modeling some real-world dynamic process. For many physical dynamical systems, e.g. flight vehicles, the dynamics equation is also known as the **equation of motion (EOM)**. The study of dynamical systems encompasses three broad topics: simulation, system identification (SID), and control. Each of these concepts will be introduced in this introductory part of the textbook for FDC.

A dynamical system may be characterized by its signal types. **Analog systems** have only analog signals while **digital systems** have only digital signals. Importantly, any electronic dynamical system contains both analog and digital signals at some level which are traditionally analyzed separately along with the analog-to-digital and/or digital-to-analog converters.

A dynamical system may be characterized by the form of its dynamics equation. A **linear system** satisfies the **principle of superposition** which states that when one adds multiple input signals together as well as multiplies them by arbitrary scalars, then the overall system output signal will be equivalent to the added and scaled output signals of the individual input signals. A **nonlinear system** does not satisfy the principle of superposition. A **time-invariant system** has a dynamics equation that does not depend *explicitly* on time. A

**time-varying system** has a dynamics equation that does depend *explicitly* on time. Lastly, a **deterministic system** will always produce the same output signal for a given input signal. If this is not true, it is a **stochastic system**.

Lastly, a dynamical system may be characterized by its number of input and output signals. If a system has no inputs, it is an **autonomous system**. Otherwise, it may be a **single input, single output (SISO) system**; a **single input, multiple outputs (SIMO) system**; a **multiple inputs, single output (MISO) system**; or a **multiple inputs, multiple outputs (MIMO) system**.

This chapter of the textbook introduces the mathematical theory for continuous-time, **linear time-invariant (LTI)**, deterministic dynamical systems. Unfortunately, no dynamical system is perfectly linear, time-invariant, or deterministic. First, linearity implies that the operation of a system can be scaled to arbitrarily large magnitudes which is not physically possible in the extreme. Second, time-invariance is violated by aging effects that change the outputs of systems over time. Lastly, thermal noise and other random phenomena ensure that the operation of any system will have some degree of random or stochastic behavior. However, despite this limitation, these assumptions greatly simplify the mathematical theory of dynamical systems while still providing valuable insight and intuition into FDC.

## Continuous-Time System Models

The simplest mathematical model for representing continuous-time dynamical systems explicitly uses time derivatives in a system's dynamics equation. Continuous-time SISO systems can be modeled using univariate **ordinary differential equations (ODEs)** which have the general form

$$\frac{d^n y}{dt^n} = f \left( t, y, \frac{dy}{dt}, \dots, \frac{d^{n-1} y}{dt^{n-1}}, u, \frac{du}{dt}, \dots, \frac{d^m u}{dt^m} \right) \quad (1.1)$$

where  $y(t)$  is the single **output signal**,  $u(t)$  is the single **input signal**, and  $n$  is the **order of the ODE** with  $p \leq n$  denoting a **proper ordinary differential equation**. Furthermore, the ODE is said to be **unforced** if  $f()$  is not a function of an input  $u$ , i.e.

$$\frac{d^n y}{dt^n} = f \left( t, y, \frac{dy}{dt}, \dots, \frac{d^{n-1} y}{dt^{n-1}} \right) \quad (1.2)$$

The solution of the ODE, i.e. the specific output  $y(t)$  due to some specific input  $u(t)$ , requires the use of a **boundary condition** for the system output at some time  $t$  where a boundary condition given at  $t = 0$  is also known as an **initial condition**. It should be noted that from here on, this textbook will use Newton's dot notation to represent time derivatives up to the third order and a bracketed exponent for higher orders, i.e.

$$y^{[n]}(t) = f \left( t, y, \dot{y}, \dots, y^{[n-1]}, u, \dot{u}, \dots, u^{[m]} \right) \quad (1.3)$$

A multivariate extension of ODEs for continuous-time MIMO systems is the **continuous-time state-space model** which can be defined as the general form of two equations

$$\begin{aligned} \dot{\vec{x}}(t) &= f(t, \vec{x}, \vec{u}) \\ \vec{y}(t) &= h(t, \vec{x}, \vec{u}) \end{aligned} \quad (1.4)$$

where  $f : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_x}$ ,  $h : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \rightarrow \mathbb{R}^{n_u}$ ,  $\vec{u}(t) \in \mathbb{R}^{n_u}$  is the **input vector** of  $n_u$  input signals,  $\vec{y}(t) \in \mathbb{R}^{n_y}$  is the **output vector** of  $n_y$  output signals, and  $\vec{x} \in \mathbb{R}^{n_x}$  is the **state vector** of the  $n_x^{\text{th}}$ -order dynamical system. The first vector-valued differential equation is the **dynamics equation**, also known as the **state equation**, for continuous-time state-space systems. The second vector-valued algebraic equation is the **output equation** and relates the output vector to the state and input vectors. In this representation, the state of the system  $\vec{x}$  is dynamically *controlled* by the input  $\vec{u}$  and is *observed* through the output  $\vec{y}$  with the function.

It should be noted that this vector-valued state-space system can be rewritten as  $n_x$  first order ODEs for the state equation and  $n_y$  algebraic equations for the output in the general form

$$\begin{aligned}\dot{x}_1 &= f_1(t, x_1, \dots, x_{n_x}, u_1, \dots, u_{n_u}) \\ &\vdots = \vdots \\ \dot{x}_{n_x} &= f_{n_x}(t, x_1, \dots, x_{n_x}, u_1, \dots, u_{n_u}) \\ y_1 &= h_1(t, x_1, \dots, x_{n_x}, u_1, \dots, u_{n_u}) \\ &\vdots = \vdots \\ y_{n_y} &= h_{n_y}(t, x_1, \dots, x_{n_x}, u_1, \dots, u_{n_u})\end{aligned}\tag{1.5}$$

where  $x_i$  denotes the  $i^{\text{th}}$  element of  $\vec{x}$ ,  $u_j$  denotes the  $j^{\text{th}}$  element of  $\vec{u}$ , and  $y_k$  denotes the  $k^{\text{th}}$  element of  $\vec{y}$ .

The study of general continuous-time systems is an advanced topic and is discussed in subsequent parts of this textbook. This introductory chapter on dynamical systems focuses on the analysis of linear, time-invariant (LTI) systems. As stated previously, linear systems satisfy the **principle of superposition** described by two properties: scaling and additivity. Nonlinear systems do not, in general, satisfy these properties. To demonstrate these properties, let the output  $y_1(t)$  be a solution to a linear ODE with input  $u_1(t)$  and zero initial conditions and let  $y_2(t)$  be the solution with  $u_2(t)$  and zero initial conditions. The scaling property states that for any real number  $c$ , the solution of the linear ODE with input  $u_s(t) = cu_1(t)$  and zero initial conditions is given by  $y_s(t) = cy_1(t)$ . The additivity property states that the solution of the linear ODE with input  $u_a(t) = u_1(t) + u_2(t)$  and zero initial conditions is  $y_a(t) = y_1(t) + y_2(t)$ .

With these properties in mind, one can show that **linear ODEs** have the general form

$$y^{[n]}(t) + a_{n-1}(t)y^{[n-1]}(t) + \dots + a_1(t)\dot{y}(t) + a_0(t)y(t) = b_m(t)u^{[m]}(t) + \dots + b_1(t)\dot{u}(t) + b_0(t)u(t)\tag{1.6}$$

and **continuous-time linear state-space representations** have the general form

$$\begin{aligned}\dot{\vec{x}}(t) &= A(t)\vec{x}(t) + B(t)\vec{u}(t) \\ \vec{y}(t) &= C(t)\vec{x}(t) + D(t)\vec{u}(t)\end{aligned}\tag{1.7}$$

where  $A(t) \in \mathbb{R}^{n_x \times n_x}$  matrix is the **state matrix**. The  $B(t) \in \mathbb{R}^{n_x \times n_u}$  matrix is the **input matrix**. The  $C(t) \in \mathbb{R}^{n_y \times n_x}$  matrix is the **output matrix** and  $D(t) \in \mathbb{R}^{n_y \times n_u}$  is the **feedthrough matrix**.

As stated previously, time-invariant systems have dynamics equations that do not depend explicitly on time. Thus, **time-invariant ODEs** have the general form

$$y^{[n]}(t) = f(y, \dot{y}, \dots, y^{[n-1]}, u, \dot{u}, \dots, u^{[m]})\tag{1.8}$$

and **continuous-time time-invariant state-space representations** have the general form

$$\begin{aligned}\dot{\vec{x}}(t) &= f(\vec{x}, \vec{u}) \\ \vec{y}(t) &= h(\vec{x}, \vec{u})\end{aligned}\tag{1.9}$$

### Continuous-Time Linear Time-Invariant System Models

Thus, **linear time-invariant ordinary differential equations (LTI ODEs)** have the general form

$$y^{[n]}(t) + a_{n-1}y^{[n-1]}(t) + \cdots + a_1\dot{y}(t) + a_0y(t) = b_mu^{[m]}(t) + \cdots + b_1\dot{u}(t) + b_0u(t)\tag{1.10}$$

and **continuous-time linear time-invariant state-space models** have the general form

$$\begin{aligned}\dot{\vec{x}}(t) &= A\vec{x}(t) + B\vec{u}(t) \\ \vec{y}(t) &= C\vec{x}(t) + D\vec{u}(t)\end{aligned}\tag{1.11}$$

Thus, note that LTI systems theory borrows heavily from mathematical methods in differential equations and linear algebra which will be utilized throughout this textbook.

With this general form in mind, note that a particular LTI state-space model can be denoted by the quadruple  $(A, B, C, D)$ . It should also be noted that there are different choices for  $(A, B, C, D)$  for modeling the same dynamical system in terms of the input-to-output relationship,  $\vec{u}(t)$ -to- $\vec{y}(t)$ , although the internal state,  $\vec{x}(t)$ , will be different for each model. However, if one wishes to set  $\vec{y} = \vec{x}$ , then one may set the output matrix to the **identity matrix**, i.e.

$$C = I = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \vdots & 1 & 0 \\ 0 & 0 & \vdots & 0 & 1 \end{bmatrix}\tag{1.12}$$

and the feedthrough matrix to the **zero matrix**, i.e.

$$D = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}\tag{1.13}$$

For proper  $n^{\text{th}}$  order LTI ODE, i.e.  $m \leq n$ , one can rewrite this as

$$y^{[n]}(t) + a_{n-1}y^{[n-1]}(t) + \cdots + a_1\dot{y}(t) + a_0y(t) = b_nu^{[n]}(t) + \cdots + b_1\dot{u}(t) + b_0u(t)\tag{1.14}$$

which has an infinite number of equivalent state-space models. One common state-space model conversion is the **Controllable Canonical Form (CCF)** which is performed as follows. Let

$$\begin{aligned}x_1 &= y \\ x_2 &= \dot{y} \\ &\vdots = \vdots \\ x_{n_x-1} &= y^{[n-1]} \\ x_{n_x} &= y^{[n]}\end{aligned}\tag{1.15}$$

Then, the matrices of the CCF state-space system can be defined as

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix} \quad (1.16)$$

$$B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad (1.17)$$

$$C = [(b_0 - a_0 b_n) \quad (b_1 - a_1 b_n) \quad \cdots \quad (b_{n-2} - a_{n-2} b_n) \quad (b_{n-1} - a_{n-1} b_n)] \quad (1.18)$$

$$D = b_n \quad (1.19)$$

Note that if  $b_n = 0$  then the formula is much simpler. With this conversion in mind, this textbook will primarily use the state-space model instead of the ODE model in its presentation of dynamical aerospace systems which are intrinsically multivariate and when they are not, one can always easily convert from an ODE to state-space.

Another important, albeit abstracted, SISO LTI system model is the **transfer function** which uses a change of variables from the real variable  $t$  to a complex variable  $s$  under zero initial conditions by the **Laplace transform**

$$f(s) = \mathcal{L}\{f(t)\} = \int_0^\infty f(t)e^{-st} dt \quad (1.20)$$

and vice versa by the **inverse Laplace transform**

$$f(t) = \mathcal{L}^{-1}\{f(s)\} = \frac{1}{2\pi j} \lim_{T \rightarrow \infty} \int_{\gamma-jT}^{\gamma+jT} f(s)e^{st} ds \quad (1.21)$$

where  $j$  here represents  $\sqrt{-1}$ , a dynamical systems notation resulting from electrical systems already using lowercase  $i$  for current. From this transform, one can define the conversion of derivatives as

$$x^{[i]}(t) = s^i x(s) - \sum_{k=1}^i s^{i-k} x^{[k-1]}(0) \quad (1.22)$$

and linear expressions as

$$ax_1(t) + bx_2 = aX_1(s) + bX_2(s) \quad (1.23)$$

For a LTI ODE

$$y^{[n]}(t) + a_{n-1}y^{[n-1]}(t) + \cdots + a_1\dot{y}(t) + a_0y(t) = b_m u^{[m]}(t) + \cdots + b_1\dot{u}(t) + b_0u(t) \quad (1.24)$$

to find the transfer function,  $G(s)$ , which “transfers” the transformed input  $u(s)$  to the transformed output  $y(s)$  through simple multiplication, i.e.

$$y(s) = G(s)u(s) \quad (1.25)$$

one can convert to the equivalent transfer function model using the Laplace domain with zero initial conditions as

$$s^n y(s) + a_{n-1} s^{n-1} y(s) + \cdots + a_1 s y(s) + a_0 y(s) = b_m s^m u(s) + \cdots + b_1 s u(s) + b_0 u(s) \quad (1.26)$$

Rearranging, one has

$$\left( s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0 \right) y(s) = (b_m s^m + \cdots + b_1 s + b_0) u(s) \quad (1.27)$$

or

$$y(s) = \frac{b_m s^m + \cdots + b_1 s + b_0}{s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0} u(s) \quad (1.28)$$

Thus, by definition of the transfer function, the **standard transfer function form** can be defined as

$$G(s) = \frac{b_m s^m + \cdots + b_1 s + b_0}{s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0} \quad (1.29)$$

Notably, the numerator and denominator of a transfer function is a polynomial, thus, each has real and/or complex roots. A **zero**,  $z$ , of a transfer function is a root of the numerator while a **pole**,  $p$ , of a transfer function is a root of the denominator. These terms derive from the behavior of the transfer function’s magnitude,  $|G(s)|$ , in the complex plane, namely,  $|G(s)| \rightarrow 0$  as  $s \rightarrow z$  and  $|G(s)| \rightarrow \infty$  as  $s \rightarrow p$ . Plotting  $|G(s)|$  over the complex plane, this infinite value will look like a tent pole in a three dimensional plot. With this in mind, it should be noted that one can always factor a transfer function into

$$G(s) = \frac{K(s - z_1) \cdots (s - z_m)}{(s - p_1) \cdots (s - p_n)} \quad (1.30)$$

A transfer function has a **pole-zero cancellation** if any poles and zeros have the same value. If there are no pole-zero cancellations, then a transfer function is said to be a **minimal realization**. One can also factor a transfer function into a **partial-fraction decomposition**, i.e.

$$G(s) = \frac{r_1}{(s - p_1)} + \cdots + \frac{r_n}{(s - p_n)} \quad (1.31)$$

where  $r_i$  is the  $i^{\text{th}}$  constant factor known as a **residue** which will be zero if a pole-zero cancellation exists for  $p_i$ . An important result that one can use with the transfer function model for LTI systems is the **Final Value Theorem (FVT)** which states if every pole of a transfer function  $F(s)$  is not purely imaginary *except*  $F(s)$  has, at most, a single pole at the origin. Then,

$$\lim_{t \rightarrow \infty} f(t) = \lim_{s \rightarrow 0} s F(s) \quad (1.32)$$

where  $s \rightarrow 0$  denotes  $s$  approaching through the positive numbers.

For reference, some additional conversions for functions in the  $t$  and  $s$  domains are provided in the following table which are often used for converting between  $u(t)$  and  $u(s)$  and  $y(t)$  and  $y(s)$ , respectively.

Function	$t$ Domain ( $\forall t \geq 0$ )	$s$ Domain
Unit Step	1	$\frac{1}{s}$
Exponential	$e^{-at}$	$\frac{1}{s+a}$
Time power $n \in \mathbb{N}$ + Exponential	$\frac{t^{n-1}}{(n-1)!} e^{-at}$	$\frac{1}{(s+a)^n}$
Sine	$\sin \omega t$	$\frac{\omega}{s^2 + \omega^2}$
Cosine	$\cos \omega t$	$\frac{s}{s^2 + \omega^2}$
Exponentially Decaying Sine	$e^{-at} \sin \omega t$	$\frac{\omega}{(s+a)^2 + \omega^2}$
Exponentially Decaying Cosine	$e^{-at} \cos \omega t$	$\frac{s+a}{(s+a)^2 + \omega^2}$

For a continuous-time MIMO LTI systems as a state-space model, i.e.

$$\begin{aligned}\dot{\vec{x}}(t) &= A\vec{x}(t) + B\vec{u}(t) \\ \vec{y}(t) &= C\vec{x}(t) + D\vec{u}(t)\end{aligned}\tag{1.33}$$

one transforms to the Laplace domain with zero initial conditions as

$$\begin{aligned}s\vec{x}(s) &= A\vec{x}(s) + B\vec{u}(s) \\ \vec{y}(s) &= C\vec{x}(s) + D\vec{u}(s)\end{aligned}\tag{1.34}$$

To find the **transfer function matrix**,  $[G(s)]$ , which maps each input to each output through a matrix of transfer functions, i.e.

$$\vec{y}(s) = [G(s)]\vec{u}(s)\tag{1.35}$$

one requires finding  $\vec{x}(s)$  in terms of  $\vec{u}(s)$  and substituting into the output equation. Thus,

$$s\vec{x}(s) - A\vec{x}(s) = B\vec{u}(s)\tag{1.36}$$

$$(sI - A)\vec{x}(s) = B\vec{u}(s)\tag{1.37}$$

$$\vec{x}(s) = (sI - A)^{-1}B\vec{u}(s)\tag{1.38}$$

where  $(sI - A)^{-1}$  is known as the **resolvent** of  $A$ .

Then, by substitution

$$\vec{y}(s) = C(sI - A)^{-1}B\vec{u}(s) + D\vec{u}(s)\tag{1.39}$$

$$\vec{y}(s) = \left(C(sI - A)^{-1}B + D\right)\vec{u}(s)\tag{1.40}$$

Thus, by definition of the transfer function matrix, **standard transfer function matrix** can be defined as

$$[G(s)] = C(sI - A)^{-1}B + D\tag{1.41}$$

which is a  $n \times m$  matrix of transfer function elements,  $G_{ij}(s)$ , i.e.

$$[G(s)] = \begin{bmatrix} G_{11}(s) & \cdots & G_{1m}(s) \\ \vdots & \ddots & \vdots \\ G_{n1} & \cdots & G_{nm}(s) \end{bmatrix}\tag{1.42}$$

Lastly, one sometimes may use the **polynomial-matrix model** defined as

$$P(s)\vec{y}(s) = Q(s)\vec{u}(s) \quad (1.43)$$

where the system transfer function matrix is given by

$$[G(s)] = P^{-1}(s)Q(s) \quad (1.44)$$

### Equilibrium and Trim of Dynamical Systems

For ODEs, an output-input pair  $(\bar{y}, \bar{u})$  is an **equilibrium point** if all derivatives of  $y$  and  $u$  are zero for all  $t > 0$ , i.e. if

$$f(t, \bar{y}, 0, \dots, 0, \bar{u}, 0, \dots, 0) = 0 \quad (1.45)$$

is a valid solution for all  $t \geq 0$ . Thus, if one initializes  $y(t) = \bar{y}$  at  $t = 0$  and sets  $u(t) = \bar{u}$  for  $t \geq 0$ , then  $y(t) = \bar{y}$  for all  $t \geq 0$ , i.e.  $y$  is “steady.” Note that one may choose two variables  $\bar{y}$  and  $\bar{u}$  to solve for univariate equilibrium equation which may lead to multiple, even infinite, solutions. Thus, equilibrium points for systems with inputs are also known as a **trim point**, as one is said to be “trimming” the system input  $\bar{u}$  to accomplish equilibrium at a particular trim output,  $\bar{y}$ . By extension, for state-space models, a state-input pair  $(\vec{x}, \vec{u})$  is an **equilibrium point** if  $\vec{x}$  is zero for all  $t > 0$ , i.e. if

$$f(t, \vec{x}, \vec{u}) = 0 \quad (1.46)$$

is a valid solution for all  $t \geq 0$ . Similarly, if one initializes  $\vec{x}(t) = \vec{x}$  at  $t = 0$  and sets  $\vec{u}(t) = \vec{u}$  for  $t \geq 0$ , then  $\vec{x}(t) = \vec{x}$  and  $\vec{y} = h(\vec{x}, \vec{u})$  for all  $t \geq 0$ , i.e.  $\vec{x}$  is “steady.” Here, there also may be multiple, even infinite, solutions for a particular **trim state**,  $\vec{x}$ , since there are fewer equations than free variables for the vector-valued state equation, i.e.  $n_x$  elements and  $n_x + n_u$  free variables.

An important characterization of system equilibrium points is their stability which was first studied by Lyapunov in his dissertation *The General Problem of Stability of Motion*. With the condition  $u(t) = \bar{u}$ , an equilibrium point  $\bar{y}$  of an ODE is defined as **stable in the sense of Lyapunov (SISL)** if for every  $|\epsilon| > 0$  with  $|y(t) - \bar{y}| < \epsilon$ , there exists some  $\delta > 0$  for which  $|y(0) - \bar{y}| < \delta$  as  $t \rightarrow \infty$ , i.e., the system will “remain close” to the equilibrium point as long as one initializes the dynamical system “near enough” to the equilibrium. Moreover, if  $\delta$  exists for all  $\epsilon \in \mathbb{R}$ , then the equilibrium point is defined as **globally stable**. If  $y \rightarrow \bar{y}$  as  $t \rightarrow \infty$ , then the equilibrium point is defined as **asymptotically stable**. If  $y \rightarrow \bar{y}$  as  $t \rightarrow \infty$ , then the equilibrium point is defined as **asymptotically stable**. Lastly, if  $y \rightarrow \bar{y}$  for all  $\epsilon \in \mathbb{R}$ , then the equilibrium point is defined as **globally asymptotically stable (GAS)**.

To qualitatively assess the stability of equilibrium points, Lyapunov developed two methods. The first method of Lyapunov states that if the *linearized dynamical system* about the equilibrium point is strictly stable, then the original nonlinear dynamical system is also stable for some “stability neighborhood” about the equilibrium point. The size of this “stability neighborhood” is directly related to the effects of the neglected higher-order terms (HOT) in the linearization. Thus, for highly nonlinear ODEs, one typically analyzes the system’s stability using Lyapunov’s second method as the linearized stability neighborhood is too small to be practically useful. However, this section of the textbook focuses on LTI FDC and will only consider Lyapunov’s first method for time-invariant systems and will provide methods for assessing a flight vehicle’s stability and designing control systems. A later section of this textbook will discuss Lyapunov’s second method.

## General Solution to Continuous-Time LTI Systems

First, consider a proper first-order LTI ODE, i.e.

$$\dot{y}(t) + a_0 y(t) = b_0 u(t) \quad (1.47)$$

with boundary conditions,  $y(t_0)$ . From calculus, the general solution can be shown to be

$$y(t) = y_H(t) + y_P(t) \quad (1.48)$$

where  $y_H(t)$  is the **ODE homogeneous solution**, i.e. the solution for  $u(t) = 0$ , and  $y_P(t)$  is the **ODE particular solution** due to  $u(t) \neq 0$ .

Consider the **exponential function** of  $a(t - t_0)$ , i.e.

$$e^{a(t-t_0)} = \sum_{i=0}^{\infty} \frac{1}{i!} [a(t-t_0)]^i \quad (1.49)$$

with derivative

$$\frac{d}{dt} e^{a(t-t_0)} = \sum_{i=0}^{\infty} \frac{1}{i!} ai + 1(t-t_0)^i = ae^{a(t-t_0)} \quad (1.50)$$

Thus, the homogeneous solution to the first-order ODE is simply

$$y(t) = e^{-a_0(t-t_0)} y(t_0) \quad (1.51)$$

and the particular solution can be computed using a integrating factor  $e^{a_0 t}$

$$e^{a_0 t} \dot{y}_P(t) + e^{a_0 t} a_0 y_P(t) = e^{a_0 t} b_0 u(t) \quad (1.52)$$

By recognizing the product rule of the left side, one has

$$\frac{d}{dt} (e^{a_0 t} y_P(t)) = e^{a_0 t} b_0 u(t) \quad (1.53)$$

Next, integrating over the input signal from  $t_0$  to  $t$ , one has

$$e^{a_0 t} y_P(t) = \int_{t_0}^t e^{a_0 \tau} b_0 u(\tau) d\tau \quad (1.54)$$

Finally, multiplying both sides by  $e^{-a_0 t}$  yields

$$y_P(t) = \int_{t_0}^t e^{-a_0(t-\tau)} b_0 u(\tau) d\tau \quad (1.55)$$

Thus, the general solution for the first-order system is

$$y(t) = e^{-a_0(t-t_0)} y(t_0) + \int_{t_0}^t e^{-a_0(t-\tau)} b_0 u(\tau) d\tau \quad (1.56)$$

It is important to note that the fundamental characteristic of this first-order system depends on the value of  $a_0$ . Namely, if  $a_0 > 0$ , then the exponential terms in the general solution will decay to 0 as  $t \rightarrow \infty$ , i.e. the system output will reach some bounded output due to the integrated input signal. Thus, the LTI system is stable.

Next, consider the general continuous-time LTI state-space form

$$\begin{aligned}\dot{\vec{x}}(t) &= A\vec{x}(t) + B\vec{u}(t) \\ \vec{y}(t) &= C\vec{x}(t) + D\vec{u}(t)\end{aligned}\quad (1.57)$$

with boundary conditions  $\vec{x}(t_0)$ . By analogy, the general solution for the vector-valued first-order state equation can be shown to be

$$\begin{aligned}\vec{x}(t) &= e^{A(t-t_0)}\vec{x}(t_0) + \int_{t_0}^t e^{A(t-\tau)}B\vec{u}(\tau)d\tau \\ \vec{y}(t) &= C\left[e^{A(t-t_0)}\vec{x}(t_0) + \int_{t_0}^t e^{A(t-\tau)}B\vec{u}(\tau)d\tau\right] + D\vec{u}(t)\end{aligned}\quad (1.58)$$

where

$$e^{A(t-t_0)} = \sum_{k=0}^{\infty} \frac{1}{k!} [A(t-t_0)]^k = \lim_{k \rightarrow \infty} (I + \frac{1}{k} A(t-t_0))^k \quad (1.59)$$

is known as the **matrix exponential function** of  $A(t-t_0)$ , also known as the **LTI state-transition matrix**,  $\Phi(t, t_0)$ . Also by analogy to  $a_0$ , analysis of the values of the state matrix,  $A$ , provides the stability analysis of the general LTI system.

## References

For more information, please refer to the following

- Schmidt, D. K., “10.1 Linear System Analysis - A Just-In-Time Tutorial\*,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 548-562
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “3.2 State-Space Models,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 144-154
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “3.3 Transfer Function Models,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 155-170

## 1.2 Linearization of Continuous-Time Time-Invariant Systems

### Linearization of Time-Invariant ODEs

The linearization of time-invariant ODEs requires the **Taylor series** of a univariate function  $f(y)$  about  $\bar{y}$  which is defined as

$$f(y) = f(\bar{y}) + \left[\frac{df}{dy}\right]_{\bar{y}} (y - \bar{y}) + \frac{1}{2} \left[\frac{d^2f}{dy^2}\right]_{\bar{y}} (y - \bar{y})^2 + \dots \quad (1.60)$$

which is an infinite series of increasing order. The **linearization of  $f(y)$  about  $\bar{y}$**  approximates this series by the inclusion of the zeroth and first order terms, i.e. the constant and proportional terms, which produces a linear function as

$$f(y) \approx f(\bar{y}) + \left[ \frac{df}{dy} \right]_{\bar{y}} (y - \bar{y}) \quad (1.61)$$

which can be said to approximate the true function. Thus, the linearization is also known as a **first-order approximation**. The difference between the exact solution and the linearization is known as the **linearization error** and is a result of the removal of **higher-order terms (HOT)** from the Taylor series.

In addition, it should also be noted that often one uses the substitution

$$y = \bar{y} + \Delta y \quad (1.62)$$

where  $\Delta y$  is a **perturbation** about the point  $\bar{y}$ . This type of substitution is known as **perturbation form**. Then, the linearization can be rewritten as

$$f(\bar{y} + \Delta y) \approx f(\bar{y}) + \left[ \frac{df}{dy} \right]_{\bar{y}} \Delta y \quad (1.63)$$

which is known as the fundamental **linearization theorem** in **small-perturbation theory**, also known as **small-disturbance theory**.

The general form for a time-invariant ODE is given by

$$y^{[n]} = f \left( y, \dot{y}, \dots, y^{[n-1]}, u, \dot{u}, \dots, u^{[m-1]} \right) \quad (1.64)$$

Then, modeling the output signal in perturbation form as

$$y(t) = \bar{y} + \Delta y(t) \quad (1.65)$$

and the input signal in perturbation form as

$$u(t) = \bar{u} + \Delta u(t) \quad (1.66)$$

the linearization of a time-invariant ODE about a trim point,  $(\bar{u}, \bar{y})$ , can be written as

$$\Delta y^{[n]} + a_{n-1} \Delta y^{[n-1]} + \dots + a_1 \Delta \dot{y} + a_0 \Delta y = b_m \Delta u^{[m]} + \dots + b_1 \Delta \dot{u} + b_0 \Delta u \quad (1.67)$$

where the coefficients of this LTI ODE are

$$a_0 = - \frac{\partial f}{\partial y} (\bar{y}, 0, \dots, 0, \bar{u}, 0, \dots, 0) \quad (1.68)$$

$$a_i = - \frac{\partial f}{\partial y^{[i]}} (\bar{y}, 0, \dots, 0, \bar{u}, 0, \dots, 0) \quad (1.69)$$

for  $i = 1, \dots, n-1$ , and

$$b_0 = \frac{\partial f}{\partial u} (\bar{y}, 0, \dots, 0, \bar{u}, 0, \dots, 0) \quad (1.70)$$

$$b_j = \frac{\partial f}{\partial u^{[j]}} (\bar{y}, 0, \dots, 0, \bar{u}, 0, \dots, 0) \quad (1.71)$$

for  $j = 1, \dots, m$ . Note that the negative signs appear for the  $a$  coefficients due to the standard form for LTI ODEs where the left side of the equation contains all  $y$  terms and the right side contains all  $u$  terms. Lastly, note that the derivatives of  $y(t)$  and  $u(t)$  are equivalent to those of  $\Delta y(t)$  and  $\Delta u(t)$  since  $\bar{y}$  and  $\bar{u}$  are constants.

### Linearization of Time-Invariant Continuous-Time State-Space Models

The linearization of time-invariant state-space models requires the **Taylor Series** for multivariate function  $f(\vec{x}, \vec{u})$  about the vector pair  $(\bar{\vec{x}}, \bar{\vec{u}})$  is

$$\dot{\vec{x}}(t) = f(\vec{x}, \vec{u}) = f(\bar{\vec{x}}, \bar{\vec{u}}) + \left[ \frac{\partial f}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] (\vec{x} - \bar{\vec{x}}) + \left[ \frac{\partial f}{\partial \vec{u}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] (\vec{u} - \bar{\vec{u}}) + \text{HOT} \quad (1.72)$$

or

$$\vec{y}(t) = h(\vec{x}, \vec{u}) = h(\bar{\vec{x}}, \bar{\vec{u}}) + \left[ \frac{\partial h}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] (\vec{x} - \bar{\vec{x}}) + \left[ \frac{\partial h}{\partial \vec{u}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] (\vec{u} - \bar{\vec{u}}) + \text{HOT} \quad (1.73)$$

where  $\left[ \frac{\partial f}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right]$  is the **Jacobian** of  $f()$ . Thus, multivariate linearization is also known as **Jacobian linearization**.

Defining the state, input, and output **perturbation vectors** about constants  $\bar{\vec{x}}$ ,  $\bar{\vec{u}}$ , and  $\bar{\vec{y}}$  as

$$\Delta \vec{x}(t) = \vec{x}(t) - \bar{\vec{x}} \quad (1.74)$$

$$\Delta \vec{u}(t) = \vec{u}(t) - \bar{\vec{u}} \quad (1.75)$$

$$\Delta \vec{y}(t) = \vec{y}(t) - \bar{\vec{y}} \quad (1.76)$$

and recognizing for trim,  $f(\bar{\vec{x}}, \bar{\vec{u}}) = 0$  and  $h(\bar{\vec{x}}, \bar{\vec{u}}) = \bar{\vec{y}}$ , one has

$$\Delta \dot{\vec{x}}(t) = \dot{\vec{x}}(t) = f(\vec{x}, \vec{u}) = f(\bar{\vec{x}}, \bar{\vec{u}}) + \left[ \frac{\partial f}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] \Delta \vec{x}(t) + \left[ \frac{\partial f}{\partial \vec{u}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] \Delta \vec{u}(t) + \text{HOT} \quad (1.77)$$

$$\Delta \vec{y}(t) = \vec{y}(t) - \bar{\vec{y}} = h(\vec{x}, \vec{u}) - h(\bar{\vec{x}}, \bar{\vec{u}}) = \left[ \frac{\partial h}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] \Delta \vec{x}(t) + \left[ \frac{\partial h}{\partial \vec{u}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] \Delta \vec{u}(t) + \text{HOT} \quad (1.78)$$

Recall the general form for a time-invariant state-space model

$$\begin{aligned} \dot{\vec{x}} &= f(\vec{x}, \vec{u}) \\ \vec{y} &= h(\vec{x}, \vec{u}) \end{aligned} \quad (1.79)$$

Next, setting

$$A = \left[ \frac{\partial f}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial f_1}{\partial x_{n_x}}(\bar{\vec{x}}, \bar{\vec{u}}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{n_x}}{\partial x_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial f_{n_x}}{\partial x_{n_x}}(\bar{\vec{x}}, \bar{\vec{u}}) \end{bmatrix} \quad (1.80)$$

$$B = \left[ \frac{\partial f}{\partial \vec{u}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] = \begin{bmatrix} \frac{\partial f_1}{\partial u_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial f_1}{\partial u_{n_u}}(\bar{\vec{x}}, \bar{\vec{u}}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{n_x}}{\partial u_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial f_{n_x}}{\partial u_{n_u}}(\bar{\vec{x}}, \bar{\vec{u}}) \end{bmatrix} \quad (1.81)$$

$$C = \left[ \frac{\partial h}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] = \begin{bmatrix} \frac{\partial h_1}{\partial x_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial h_1}{\partial x_{n_x}}(\bar{\vec{x}}, \bar{\vec{u}}) \\ \vdots & \ddots & \vdots \\ \frac{\partial h_{n_y}}{\partial x_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial h_{n_y}}{\partial x_{n_x}}(\bar{\vec{x}}, \bar{\vec{u}}) \end{bmatrix} \quad (1.82)$$

$$D = \begin{bmatrix} \frac{\partial h_1}{\partial u_1}(\bar{x}, \bar{u}) & \cdots & \frac{\partial h_1}{\partial u_{n_u}}(\bar{x}, \bar{u}) \\ \vdots & \ddots & \vdots \\ \frac{\partial h_{n_y}}{\partial u_1}(\bar{x}, \bar{u}) & \cdots & \frac{\partial h_{n_y}}{\partial u_{n_u}}(\bar{x}, \bar{u}) \end{bmatrix} \quad (1.83)$$

yields an approximate LTI state-space model about  $\bar{x}$  and  $\bar{u}$  as

$$\begin{aligned} \Delta \dot{\vec{x}}(t) &\approx A \Delta \vec{x}(t) + B \Delta \vec{u}(t) \\ \Delta \vec{y}(t) &\approx C \Delta \vec{x}(t) + D \Delta \vec{u}(t) \end{aligned} \quad (1.84)$$

### Linearizations of Trigonometric Functions

As an aside, an important linearization is for the trigonometric sine and cosine functions about any angle  $\bar{\theta}$ . It can be shown that the Taylor series for sine and cosine are

$$\sin(\bar{\theta}) = \sin(\bar{\theta}) + \cos(\bar{\theta})(\theta - \bar{\theta}) - \frac{1}{2!} \sin(\bar{\theta})(\theta - \bar{\theta})^2 - \frac{1}{3!} \cos(\bar{\theta})(\theta - \bar{\theta})^3 + \dots \quad (1.85)$$

and

$$\cos(\bar{\theta}) = \cos(\bar{\theta}) - \sin(\bar{\theta})(\theta - \bar{\theta}) - \frac{1}{2!} \cos(\bar{\theta})(\theta - \bar{\theta})^2 + \frac{1}{3!} \sin(\bar{\theta})(\theta - \bar{\theta})^3 + \dots \quad (1.86)$$

where  $\theta$  must be in radians. Thus, the linearizations for sine and cosine about  $\bar{\theta}$  are given by

$$\sin(\bar{\theta}) \approx \sin \bar{\theta} + \cos(\bar{\theta})(\theta - \bar{\theta}) \quad (1.87)$$

and

$$\cos(\bar{\theta}) \approx \cos \bar{\theta} - \sin(\bar{\theta})(\theta - \bar{\theta}) \quad (1.88)$$

Of particular note, for  $\bar{\theta} = 0$  rad, the linearization simplifies to

$$\sin \theta \approx \theta \quad (1.89)$$

and

$$\cos \theta \approx 1 \quad (1.90)$$

In this case, one can also approximate

$$\tan \theta \approx \theta \quad (1.91)$$

These three approximations are commonly called the **small angle approximations** and will be used throughout this textbook. These are decent approximations when  $\theta < 15^\circ = 0.2618$  rad, producing a linearization error of 0.0028, 0.034, 0.0061 and for sine, cosine, and tangent, respectively.

### References

For more information, please refer to the following

- Schmidt, D. K., “1.1 Small Perturbation Theory for Nonlinear Systems,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 1-2

- Schmidt, D. K., “1.4 Vector Differentiation,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 10-14
- Schmidt, D. K., “1.6 Small Perturbation Analysis Revisited,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 18-21
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “3.7 Numerical Linearization,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 199-205

### 1.3 Modal Analysis of Continuous-Time LTI Systems

#### Characteristic Polynomial of Continuous-Time LTI Systems

Consider a  $n^{\text{th}}$ -order LTI system represented by the standard transfer function model, i.e.

$$y(s) = G(s)u(s) \quad (1.92)$$

where, using the standard ODE conversion, one has

$$G(s) = \frac{b_ms^m + \cdots + b_1s + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1s + a_0} \quad (1.93)$$

and the **ODE characteristic equation** can be defined as

$$\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 = 0 \quad (1.94)$$

where the left side of this equation is called the **ODE characteristic polynomial**. The solutions to this equation are the  $n$  roots of the ODE characteristic polynomial and may be repeated and will either be real numbers or complex-conjugate pairs.

This can be generalized for all LTI systems by the standard transfer function matrix model, i.e.

$$\vec{y}(s) = [G(s)]\vec{u}(s) \quad (1.95)$$

where, using the standard state-space conversion, one has

$$[G(s)] = C(sI - A)^{-1}B + D \quad (1.96)$$

Furthermore, by use of the **adjugate**, also known as the **classical adjoint** of  $A$  instead of the matrix inverse, i.e.

$$[G(s)] = C \left( \frac{\text{adj}(sI - A)}{\det[sI - A]} \right) B + D \quad (1.97)$$

one can see that each transfer function element,  $G_{ij}(s)$  of  $[G(s)]$ , will have the same denominator polynomial given by  $\det[sI - A]$ , and is known as the **state-space characteristic polynomial**.

Thus, the **state-space characteristic equation** for general LTI systems can be defined as

$$\det[\lambda I - A] = 0 \quad (1.98)$$

The solutions to this equation are the  $n_x$  roots of the state-space characteristic polynomial and may be repeated and will either be real numbers or complex-conjugate pairs. Lastly, the roots of the characteristic polynomials are also known as the **system poles**. Notably, if there are no pole-zero cancellations, then the poles of any  $G_{ij}(s)$  are also the system poles. Furthermore, each distinct real pole and each distinct complex-conjugate pair is called a **system mode** which are important in analyzing the characteristics of LTI systems.

### Modal Representation of Continuous-Time LTI Systems

To analyze the modes of an LTI system, first, note that the LTI system poles are equivalent to the eigenvalues of  $A$ ,  $\lambda_i$ . These eigenvalues solve the **eigenvalue equations** for right eigenvectors,  $\vec{v}_i$ , i.e.

$$A \vec{v}_i = \lambda_i \vec{v}_i \quad (1.99)$$

and for left eigenvectors,  $\vec{\mu}_i^T$ , i.e.

$$\vec{\mu}_i^T A = \lambda_i \vec{\mu}_i^T \quad (1.100)$$

If there are  $n_x$  distinct eigenvalues, then  $A$  is **diagonalizable** and one can perform an **eigenvalue decomposition** of  $A$  as

$$A = V \Lambda V^{-1} \quad (1.101)$$

where  $\Lambda$  is the diagonal matrix of  $n_x$  eigenvalues, i.e.

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{n_x} \end{bmatrix} \quad (1.102)$$

and  $V$  is the matrix of  $n_x$  linearly independent right eigenvectors, i.e.

$$V = [\vec{v}_1 \ \cdots \ \vec{v}_{n_x}] \quad (1.103)$$

and  $V^{-1}$  is the matrix of  $n_x$  linearly independent left eigenvectors, i.e.

$$V^{-1} = \begin{bmatrix} \vec{\mu}_1^T \\ \vdots \\ \vec{\mu}_{n_x}^T \end{bmatrix} \quad (1.104)$$

Next, define the **modal state vector**,  $\vec{z}$ , such that

$$\vec{x} = V \vec{z} = \vec{v}_1 z_1(t) + \cdots + \vec{v}_{n_x} z_{n_x}(t) \quad (1.105)$$

and

$$\vec{z} = V^{-1} \vec{x} = \begin{bmatrix} \vec{\mu}_1^T \vec{x} \\ \vdots \\ \vec{\mu}_{n_x}^T \vec{x} \end{bmatrix} \quad (1.106)$$

where notably,  $z_{n_x}(t)$ , is dimensionless. Then, substituting into the state equation, one has

$$V \dot{\vec{z}}(t) = AV \vec{z}(t) + B \vec{u}(t) \quad (1.107)$$

or

$$\dot{\vec{z}}(t) = V^{-1}AV \vec{z}(t) + V^{-1}B \vec{u}(t) \quad (1.108)$$

Thus, one can form the **Jordan canonical form (JCF)** of the state-space system as

$$\begin{aligned} \dot{\vec{z}}(t) &= \Lambda \vec{z}(t) + V^{-1}B \vec{u}(t) \\ \vec{y}(t) &= CV \vec{z}(t) + D \vec{u}(t) \end{aligned} \quad (1.109)$$

where the state-transition matrix for diagonalizable  $A$  is now

$$e^{\Lambda(t-t_0)} = \begin{bmatrix} e^{\lambda_1(t-t_0)} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & e^{\lambda_{n_x}(t-t_0)} \end{bmatrix} \quad (1.110)$$

and this model explicitly **decouples** the state equation into  $n_x$  LTI ODEs, i.e.

$$\dot{z}_i(t) = \lambda_i z_i(t) + \vec{\mu}_i^T B \vec{u}(t), \quad i = 1, \dots, n_x \quad (1.111)$$

Lastly, recalling **Euler's formula**, i.e.

$$e^{j\omega t} = \cos(\omega t) + j\sin(\omega t) \quad (1.112)$$

note that the exponential of a complex-conjugate pair  $\sigma \pm j\omega$  for an eigenvalue can be rewritten as

$$e^{(\sigma \pm j\omega)(t-t_0)} = e^{\sigma(t-t_0)} (\cos(\omega(t-t_0)) \pm j\sin(\omega(t-t_0))) \quad (1.113)$$

### Modal Analysis of Continuous-Time LTI Systems

Consider the zero-input response where the  $n_x$  modal states are given by

$$z_i(t) = e^{\lambda_i(t-t_0)} z_i(t_0) = e^{\lambda_i(t-t_0)} \vec{\mu}_i^T \vec{x}(t_0) \quad (1.114)$$

and the output response is given by

$$\vec{y}(t) = C (\vec{v}_1 z_1(t) + \cdots + \vec{v}_{n_x} z_{n_x}(t)) \quad (1.115)$$

Thus, one can see the modal states are characterized by the eigenvalues and left eigenvectors while the output response is additionally characterized by the right eigenvectors and the modal states. As these eigenvalues will occur as real numbers or in complex-conjugate pairs, the modal states may be complex-valued which describes the oscillatory nature of the modal responses. Regardless, the real part of each individual mode allows one to characterize the stability of each mode as well as the LTI system as a whole. In particular, a mode is **stable** if the real part of  $\lambda_i < 0$ , **marginally stable** if the real part of  $\lambda_i = 0$ , and **unstable** if the

real part of  $\lambda_i > 0$ . Furthermore, the LTI system is **stable** if *all* the modes are strictly stable, the LTI system is **marginally stable** if *all* the modes are stable or marginally stable, and if *any* mode is unstable, the LTI system is **unstable**. As these eigenvalues may be complex-valued, one can also state that all the system poles located in the left half of the complex plane, i.e. the **left half plane (LHP)**, correspond to stable modes.

Furthermore, each mode corresponds to a unique **modal time constant** which describes its exponential decay/growth rate, i.e.

$$\tau = -\frac{1}{\text{Real}(\lambda_i)} \quad (1.116)$$

where  $\text{Real}(\lambda)$  represents the real part of  $\lambda$ . Thus, if  $\tau > 0$ , the mode is stable. This also demonstrates modes have faster decay rates on the output for modes with higher eigenvalues. For complex-conjugate pairs,  $\lambda = -\frac{1}{\tau} \pm j\omega_d$ , one also defines the **modal damped frequency**,  $\omega_d$ , which are related to the **modal undamped natural frequency**,  $\omega_n$ , and **modal damping ratio**,  $0 < \zeta_i < 1$ , by

$$\omega_d = \omega_n \sqrt{1 - \zeta^2} \quad (1.117)$$

and

$$\tau = \frac{1}{\zeta \omega_n} \quad (1.118)$$

and one can compare

$$(\lambda + \frac{1}{\tau} + j\omega_d)(\lambda + \frac{1}{\tau} - j\omega_d) \quad (1.119)$$

to the standard second-order form

$$\lambda^2 + 2\zeta\omega_n\lambda + \omega_n^2 \quad (1.120)$$

and note  $\lambda = -\zeta\omega_n \pm j\omega_n\sqrt{1 - \zeta^2}$ .

Consider the transfer function matrix model using the eigenvalue decomposition with  $D = 0$ , i.e.

$$[G(s)] = C(sI - M\Lambda M^{-1})^{-1}B \quad (1.121)$$

which one can show results in the matrix

$$[G(s)] = CM \begin{bmatrix} \frac{1}{s-\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{s-\lambda_{n_x}} \end{bmatrix} M^{-1}B \quad (1.122)$$

Then by substitution for  $M$  and  $M^{-1}$ , one has

$$[G(s)] = C \left( \sum_{i=1}^{n_x} \frac{\vec{v}_i \vec{\mu}_i^T}{s - \lambda_i} \right) B \quad (1.123)$$

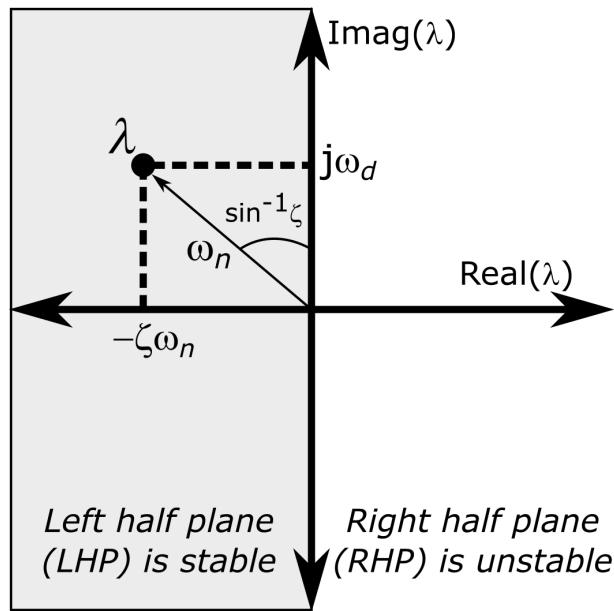
Furthermore, defining  $\vec{c}_j^T$  as the  $j^{\text{th}}$  row of  $C$  and  $\vec{b}_k$  as the  $k^{\text{th}}$  column of  $B$ , then for each input  $j$  and output  $k$ , one has the transfer function

$$G_{jk} = \sum_{i=1}^{n_x} \frac{\vec{c}_j^T \vec{v}_i \vec{\mu}_i^T \vec{b}_k}{(s - \lambda_i)} \quad (1.124)$$

which defines a partial-fraction decomposition.

The residues of these decompositions demonstrate that the  $i^{\text{th}}$  right eigenvector characterizes how the  $i^{\text{th}}$  mode affects the  $j^{\text{th}}$  output, while the  $i^{\text{th}}$  left eigenvalue characterize how the  $k^{\text{th}}$  input affects the *same*  $i^{\text{th}}$  modal state. Consequently, if any residue is zero, then either  $\vec{\mu}_i^T \vec{b}_k = 0$  which indicates the  $k^{\text{th}}$  input does not affect the  $i^{\text{th}}$  mode, i.e.  $i$  is an **uncontrollable** mode, and/or  $\vec{c}_j^T \vec{v}_i = 0$  which indicates the  $i^{\text{th}}$  mode does not affect the  $j^{\text{th}}$  output, i.e.  $i$  is an **unobservable** mode. Furthermore, if  $G_{jk}$  has an *approximate* pole-zero cancellation at  $s = \lambda_i$ , then the residue will be small and the contribution of the  $i^{\text{th}}$  mode will be relatively small on the output. It should be noted that if  $A$  is not diagonalizable, i.e. there are repeated eigenvalues, then, one must use Jordan matrices to form  $\Lambda$ , and a similar modal analysis can still be made with some caveats. These details are discussed in later parts of the textbook.

Lastly, it should be noted that often one desires to use a **reduced-order model (ROM)** for the purposes of simulation, control, and/or SID of dynamical systems. For LTI systems, one may use modal analysis to evaluate the most significant system effects in terms of the modes with significantly longer time constants, i.e. significantly smaller eigenvalues, as well as the higher gains, i.e. highest residues. A visual tool for quickly assessing LTI systems is the **pole-zero plot**, i.e. a plot of a transfer function's poles and zeros in the complex plane



which provides a visualization of the system stability, the slowest poles, and any approximate pole-zero cancellations.

## References

For more information, please refer to the following

- Schmidt, D. K., “10.1 Linear System Analysis - A Just-In-Time Tutorial\*,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 548-562
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “3.2 State-Space Models,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 144-154
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “3.3 Transfer Function Models,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 155-170

## 1.4 Continuous-Time SISO LTI System Responses

Consider a SISO LTI system represented by the standard LTI ODE

$$y^{[n]}(t) + a_{n-1}y^{[n-1]}(t) + \cdots + a_1\dot{y}(t) + a_0y(t) = b_mu^{[m]}(t) + \cdots + b_1\dot{u}(t) + b_0u(t) \quad (1.125)$$

with equivalent standard transfer function

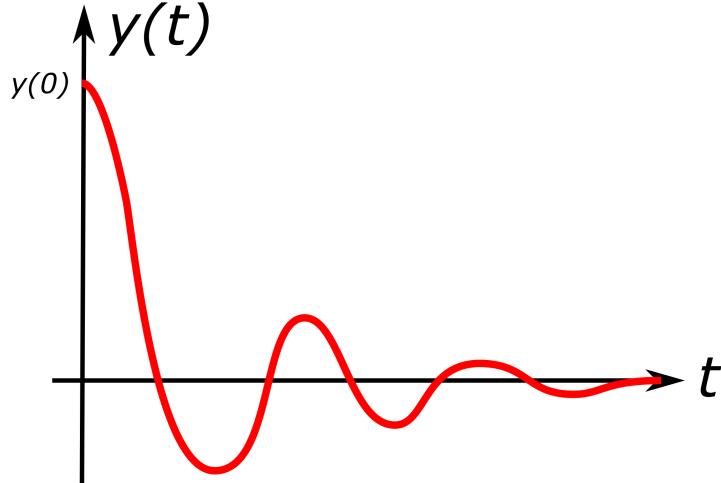
$$G(s) = \frac{b_ms^m + \cdots + b_1s + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1s + a_0} \quad (1.126)$$

Of particular importance are two system response types due to certain types of inputs and initial conditions for stable LTI systems, i.e., the zero-input and zero-state responses.

The **zero-input response** is given by  $u(t)$  as

$$u(t) = 0 \quad \forall t \quad (1.127)$$

and  $y(t)$  for boundary conditions given as initial conditions at  $t = 0$ , i.e.,  $(y(0), \dot{y}(0), \dots, y^{[n]}(0))$ . Thus, this is also known as the **free response** or the **initial condition response**. It also forms the **initial value problem (IVP)** for the ODE. In general, the initial condition response for stable SISO LTI systems has the form



This plot primarily displays whether a system is stable or not, i.e. decays to zero or grows exponentially, while its other characteristics generally also depend on the specific initial conditions and the system poles.

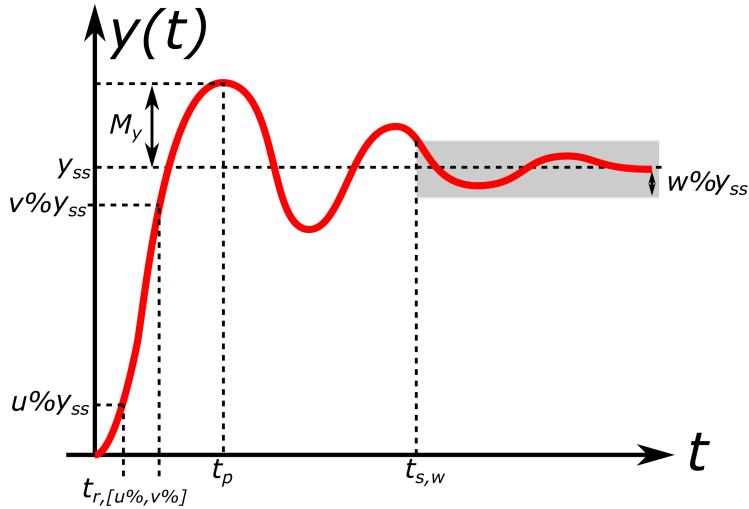
The **zero-state response** is given by

$$\begin{bmatrix} y(0) \\ \dot{y}(0) \\ \vdots \\ y^{[n]}(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (1.128)$$

For the zero-state response, there are three important choices for  $u(t)$ , namely, the step, impulse, and sinusoidal responses. This section will discuss the step and impulse responses. The **step response** given by  $y(0) = \dot{y}(0) = \dots = y^{[n]}(0) = 0$  and  $u(t)$  is the **step input**, i.e.

$$u(t) = \begin{cases} 0 & \forall t < 0 \\ k & \forall t \geq 0 \end{cases} \quad (1.129)$$

where  $k$  is the **step amplitude**. If  $k = 1$ , this is called the **unit step**. In general, the step response for stable SISO LTI systems has the form



First, as  $t \rightarrow \infty$ , the stable LTI system will reach some **steady-state condition**, i.e. the output will become steady in value as all derivatives of the input and output are zero. This is summarized by the equation

$$y_{ss} = \frac{b_0}{a_0} u_{ss} \quad (1.130)$$

where  $u_{ss} = k$  is the **steady-state input** signal,  $y_{ss}$  is the **steady-state output**, also known as the **final value**, and the ratio  $\frac{b_0}{a_0}$  is the **steady-state gain**, which for a unit step is equivalent to the final value. Furthermore, as the initial conditions are given specifically as zero, the step response has other specific characteristics. The **rise time**,  $t_{r,[u\%v\%]}$ , is defined as the time for the response to rise from  $u\%$  of  $y_{ss}$  to  $v\%$  of  $y_{ss}$ . Furthermore,

the **peak time**,  $t_p$ , is the time for the response to reach its peak value and the corresponding **maximum overshoot**, is given by

$$M_y = y(t_p) - y_{ss} \quad (1.131)$$

or expressed as a **percent maximum overshoot** as

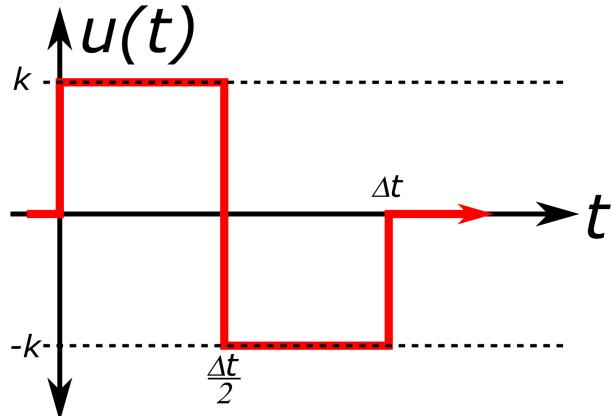
$$M_p = \frac{y(t_p) - y_{ss}}{y_{ss}} \quad (1.132)$$

Lastly, the **settling time**,  $t_{s,w}$ , is defined as the time for the response to reach and stay within  $w\%$  of  $y_{ss}$ . The next section will derive equations for the unit step response characteristics for unimodal SISO LTI systems, i.e. first-order and underdamped second-order. Then, the qualitative effects of additional poles and zeros on the dominant mode unit step response will be discussed qualitatively.

As an aside, a related input signal to the step is the **doublet input**, i.e.

$$u(t) = \begin{cases} 0 & \forall t < 0 \\ k & \forall 0 \leq t < \frac{\Delta t}{2} \\ -k & \forall \frac{\Delta t}{2} \leq t < \Delta t \\ 0 & \forall t \geq \Delta t \end{cases} \quad (1.133)$$

where  $k$  is the **doublet amplitude** and  $\Delta t$  is the **doublet length**.



The primary benefit of using the doublet over the step response is that for stable systems, the output and the *integrated* output will return to zero.

### Impulse Response

Similar to the initial condition response, consider the SISO LTI system response for an **impulse input**,

$$u(t) = \delta(t) \quad (1.134)$$

where  $\delta(t)$  is the **Dirac delta** which can be “loosely” considered a function on the real line which is zero everywhere except at the origin, where it is infinite, but is also constrained to satisfy the identity, i.e.,

$$\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases} \quad \int_{-\infty}^{\infty} \delta(x) dx = 1 \quad (1.135)$$

However, this is merely a heuristic characterization as the Dirac delta is not a function in the traditional sense as no function defined on the real numbers has these properties. The Dirac delta function can be rigorously defined either as a distribution or as a measure, but an intuitive understanding is as the special continuous-time **impulse function**.

Then, by the Laplace transform one has

$$u(s) = \int_0^{\infty} \delta(t) e^{-st} dt \quad (1.136)$$

or by the definition of  $\delta$  as zero everywhere except at  $t = 0$  and its integral constraint as well as  $e^{-s(0)} = 1$ , one has

$$u(s) = 1 \quad (1.137)$$

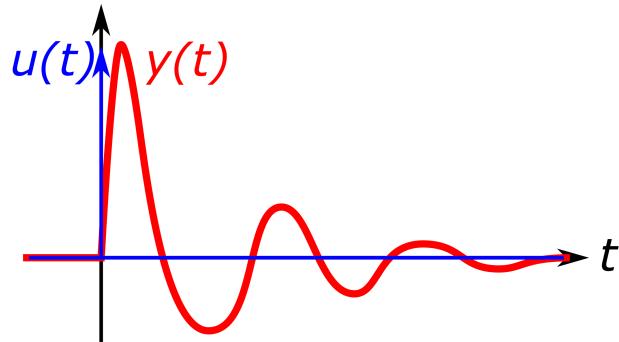
Thus,

$$y(s) = G(s) \quad (1.138)$$

or in the time domain by the inverse Laplace transform

$$y(t) = g(t) = \mathcal{L}^{-1}(G(s)) \quad (1.139)$$

where  $g(t)$  is called the **impulse response**. Thus, by simulating an LTI in the time domain using an impulse input, one can completely determine the transfer function transformed to the time domain. An example of an impulse response plot is



which is somewhat similar to an initial condition response, except the excitation of the system occurs via the input and not the initial conditions which are assumed to be zero for the impulse.

Furthermore, consider the output response in the Laplace domain

$$y(s) = G(s)u(s) \quad (1.140)$$

which can be transformed to the time domain as the **convolution integral** between the impulse response and *any* input as

$$y(t) = \int_0^t g(\tau)u(t - \tau)d\tau \quad (1.141)$$

Thus, a type of SID method is known as **deconvolution** which uses measured output and input signals in the time domain to estimate the impulse response.

### Unit Step Responses for Unimodal SISO LTI Systems

Consider a first-order LTI system with no zeros, i.e.

$$G(s) = \frac{b_0}{s + a_0} \quad (1.142)$$

Recalling the Laplace transform for a unit step, i.e.  $u(s) = \frac{1}{s}$ , the unit step response for a first-order LTI system in the Laplace domain is

$$y(s) = \frac{b_0}{s(s + a_0)} \quad (1.143)$$

which can be written in partial-fraction decomposition form as

$$y(s) = \frac{b_0}{a_0} \left( \frac{1}{s} + \frac{-1}{s + a_0} \right) \quad (1.144)$$

which, by the inverse Laplace transform, one has

$$y(t) = \frac{b_0}{a_0} \left( 1 - e^{-a_0 t} \right) \quad (1.145)$$

which is an exponential decay from 0 to the steady-state output

$$y_{ss} = \frac{b_0}{a_0} \quad (1.146)$$

and does not have a peak or maximum overshoot.

To compute the rise and settling times, note that one can write

$$\frac{a_0 y(t)}{b_0} = 1 - e^{-a_0 t} \quad (1.147)$$

$$1 - \frac{a_0 y(t)}{b_0} = e^{-a_0 t} \quad (1.148)$$

and, by the natural logarithm, one has

$$t = -\frac{\ln \left( 1 - \frac{a_0}{b_0} y(t) \right)}{a_0} \quad (1.149)$$

Thus, this system has a rise time from  $u$  to  $v$  given by

$$t_{r,[u,v]} = -\frac{\ln(1-u)}{a_0} + \frac{\ln(1-v)}{a_0} = \frac{v-u}{a_0} \quad (1.150)$$

and a  $w\%$  settling time

$$t_{s,w} = -\frac{\ln w}{a_0} \quad (1.151)$$

where notably

$$t_{s,2\%} = \frac{3.9}{a_0} \quad (1.152)$$

and

$$t_{s,5\%} = \frac{3}{a_0} \quad (1.153)$$

Next, consider a stable second-order LTI system with no zeros, i.e.

$$G(s) = \frac{b_0}{s^2 + a_1 s + a_0} \quad (1.154)$$

which may be written in the alternate quadratic form as

$$G(s) = \frac{b_0}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad (1.155)$$

where undamped natural frequency is

$$\omega_n = \sqrt{a_0} > 0 \quad (1.156)$$

and the damping ratio

$$\zeta = \frac{a_1}{2\omega_n} > 0 \quad (1.157)$$

has three different cases

1. **underdamped:**  $0 < \zeta < 1$ , i.e. one oscillatory decaying mode
2. **overdamped:**  $\zeta > 1$ , i.e. two exponentially decaying modes
3. **critically damped:**  $\zeta = 1$ , i.e. one exponentially decaying mode

thus, a unimodal system occurs for an underdamped second-order system.

Recalling the Laplace transform for a unit step, i.e.  $u(s) = \frac{1}{s}$ , the unit step response for an underdamped second-order LTI system in the Laplace domain is

$$y(s) = \frac{b_0}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)} \quad (1.158)$$

which can be written in partial-fraction decomposition form as

$$y(s) = \frac{b_0}{\omega_n^2} \left( \frac{1}{s} + \frac{-s - \zeta\omega_n}{s^2 + 2\zeta\omega_n s + \omega_n^2} \right) \quad (1.159)$$

or, completing the square, one has

$$y(s) = \frac{b_0}{\omega_n^2} \left( \frac{1}{s} - \frac{s + \zeta \omega_n}{(s + \zeta \omega_n^2 + \omega_n^2(1 - \zeta^2))} \right) \quad (1.160)$$

which, by the inverse Laplace transform, one has

$$y(t) = \frac{b_0}{\omega_n^2} \left[ 1 - e^{-\zeta \omega_n t} \cos(\omega_n \sqrt{1 - \zeta^2} t) \right] \quad (1.161)$$

or using the damped natural frequency, one has

$$y(t) = \frac{b_0}{\omega_n^2} \left[ 1 - e^{-\zeta \omega_n t} \cos(\omega_d t) \right] \quad (1.162)$$

is an oscillating exponential decay from 0 to the steady-state output

$$y_{ss} = \frac{b_0}{\omega_n^2} \quad (1.163)$$

It can be shown that the rise time from 0% to 100% is given by

$$t_{r,[0,1]} = \frac{\pi - \tan^{-1} \left( \frac{\sqrt{1-\zeta^2}}{\zeta} \right)}{\omega_d} \quad (1.164)$$

the peak time is given by

$$t_p = \frac{\pi}{\omega_d} \quad (1.165)$$

the maximum overshoot is given by

$$M_p = e^{-\frac{\zeta \pi}{\sqrt{1-\zeta^2}}} \quad (1.166)$$

and the w% settling time

$$t_{s,w} = -\frac{\ln(w \sqrt{1 - \zeta^2})}{\zeta \omega_n} \quad (1.167)$$

## Unit Step Responses for General SISO LTI Systems

For higher-order systems and system zeros, one can quantify the effects as follows. As an example of these effects, consider the nominal second-order LTI system,  $G_1(s)$ , as

$$G_1(s) = \frac{4}{s^2 + 2s + 4} \quad (1.168)$$

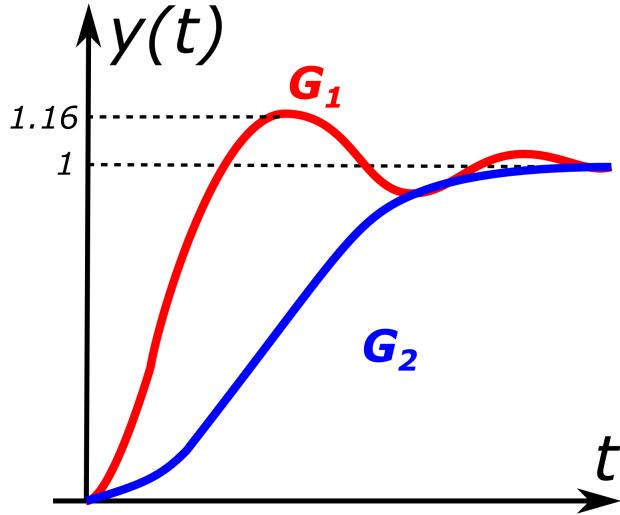
Notably,  $G_1$  is stable with poles at  $-1 \pm j1.73$  and has steady-state output of  $y_{ss} = \frac{b_0}{a_0} = \frac{4}{4} = 1$ . It is underdamped with  $\zeta = 0.5$  and  $\omega_n = 2$  and has a maximum overshoot,  $M_p \approx 0.16$ .

Due to the principle of superposition, additional poles for a system simply add additional exponential terms to the general analytical solution, additional stable poles will generally increase the peak time and

decrease the maximum overshoot from the nominal system. However, these effects are only significant if the additional poles are close in magnitude to the dominant poles, e.g. a factor of 4 is a good rule-of-thumb. As an example, consider an additional pole at  $s = -1$  for the nominal system, then one has the third-order LTI system

$$G_2(s) = \frac{4}{s^2 + 2s + 4} \left( \frac{1}{s + 1} \right) \quad (1.169)$$

Plotting the unit step response for both  $G_1(s)$  and  $G_2(s)$ , one has



One can explain the effects of a zero as follows. Let the nominal system be represented by

$$y_0(s) = G(s)u(s) \quad (1.170)$$

and let the new system,  $y_1$ , have an additional zero,  $s = z$ , i.e.

$$y_1(s) = G(s)u(s) \left( \frac{-1}{z}s + 1 \right) \quad (1.171)$$

Then,

$$y_1(s) = \left( \frac{-1}{z}s + 1 \right) y_0(s) \quad (1.172)$$

and in the time domain

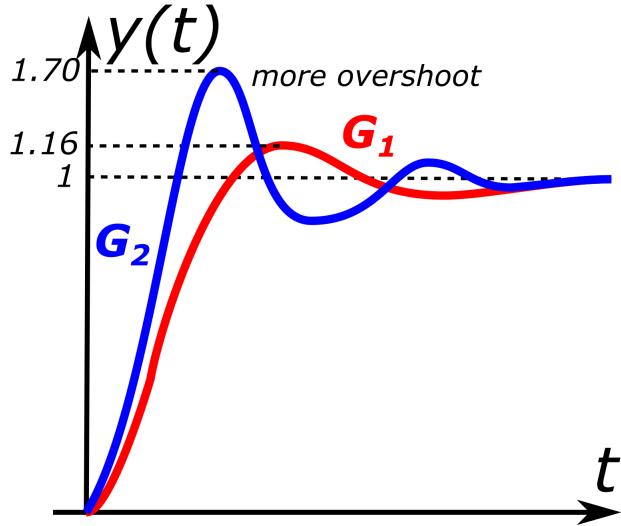
$$y_1(t) = \frac{-1}{z} \dot{y}_0(t) + y_0(t) \quad (1.173)$$

Thus, the step response consists of the nominal system response plus the “slope” of the nominal system response. Since the slope is initially positive for the nominal system response, one can infer that if  $\text{Real}(z) < 0$ , then there will be more maximum overshoot and if  $\text{Real}(z) > 0$ , then there will be also be an initial undershoot. However, these effects are only significant if the additional zeros are close in magnitude to the dominant poles, e.g. a factor of 4 is a good rule-of-thumb. If  $\text{Real}(z) > 0$ , then  $z$  is a **right half plane (RHP) zero**, otherwise if  $\text{Real}(z) < 0$ ,  $z$  is a **left half plane (LHP) zero**.

As an example, consider a LHP zero at  $s = -1$  for the nominal system, then one has the second-order LTI system

$$G_2(s) = \frac{4(s+1)}{s^2 + 2s + 4} \quad (1.174)$$

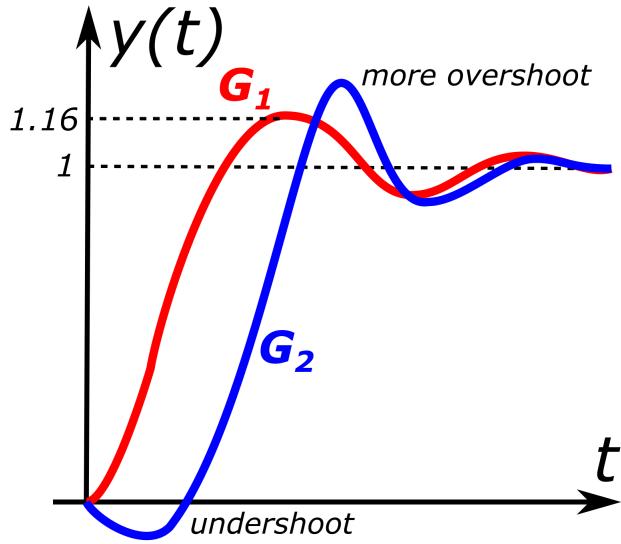
Plotting the unit step response for both  $G_1(s)$  and  $G_2(s)$ , one has



Lastly, as a second example, consider a RHP zero at  $s = 1$  for the nominal system, then one has the second-order LTI system

$$G_2(s) = \frac{4(-s+1)}{s^2 + 2s + 4} \quad (1.175)$$

Plotting the unit step response for both  $G_1(s)$  and  $G_2(s)$ , one has



## Sinusoidal Response

Dynamical systems use signals as the time-dependent input and output variables. An important type of signal is a **periodic signal** which is one that repeats its sequence of values exactly after a fixed length of time, known as the **period**. As an example, consider the simplest periodic input signal, i.e., a sinusoid at frequency  $\omega$

$$u(t) = \cos(\omega t) \quad (1.176)$$

Then, using the convolution integral for the output, one has

$$y(t) = \int_0^t g(\tau) (\cos(\omega(t - \tau))) d\tau \quad (1.177)$$

or

$$y(t) = \int_0^\infty g(\tau) (\cos(\omega(t - \tau))) d\tau - \int_t^\infty g(\tau) (\cos(\omega(t - \tau))) d\tau \quad (1.178)$$

where, if the system is stable, the first term is the steady-state sinusoidal response and the second term is the **transient response** which decays with  $t$ .

Thus, any stable system output will reach the steady-state sinusoidal response given by

$$y_{sss}(t) = \int_0^\infty g(\tau) (\cos(\omega(t - \tau))) d\tau \quad (1.179)$$

which using Euler's formula for  $\cos(\omega t)$  as

$$u(t) = \frac{1}{2} (e^{j\omega t} + e^{-j\omega t}) \quad (1.180)$$

one has

$$y_{sss}(t) = \frac{1}{2} \int_0^\infty g(\tau) e^{j\omega(t-\tau)} d\tau + \frac{1}{2} \int_0^\infty g(\tau) e^{-j\omega(t-\tau)} d\tau \quad (1.181)$$

or

$$y_{sss}(t) = \frac{1}{2} e^{j\omega t} \int_0^\infty g(\tau) e^{-j\omega\tau} d\tau + \frac{1}{2} e^{-j\omega t} \int_0^\infty g(\tau) e^{j\omega\tau} d\tau \quad (1.182)$$

and using the definition of the transfer function using the Laplace transform

$$y_{sss}(t) = \frac{1}{2} e^{j\omega t} G(j\omega) + \frac{1}{2} e^{-j\omega t} G(-j\omega) \quad (1.183)$$

or finally, the **steady-state sinusoidal response**

$$y_{sss}(t) = \text{Real}\{G(j\omega)\} \cos(\omega t) - \text{Imag}\{G(j\omega)\} \sin(\omega t) \quad (1.184)$$

which can alternatively be written in **polar form** as

$$y_{sss}(t) = |G(j\omega)| \cos(\omega t + \angle G(j\omega)) \quad (1.185)$$

where the **magnitude**, also known as the **gain**, is defined as

$$|G(j\omega)| = \sqrt{\text{Real}\{G(j\omega)\}^2 + \text{Imag}\{G(j\omega)\}^2} \quad (1.186)$$

and the **phase** is defined as

$$\angle G(j\omega) = \tan^{-1} \left( \frac{\text{Imag}\{G(j\omega)\}}{\text{Real}\{G(j\omega)\}} \right) \quad (1.187)$$

This derivation demonstrates that the steady-state sinusoidal response is a sinusoid at the same frequency, but with a different magnitude, and phase dependent on the transfer function evaluated at  $s = j\omega$ , i.e.  $G(j\omega)$ .

## Frequency Response

The importance of this substitution can be formally stated by the definition of the **Fourier transform** defined as

$$f(\omega) = \mathcal{F}\{f(t)\} = \int_{-\infty}^{\infty} f(t)e^{-j\omega t} dt \quad (1.188)$$

where  $\omega$  is the **angular frequency**, typically given in radians/second. Sometimes the **periodic frequency**,  $\xi$ , is used where  $\omega = 2\pi\xi$  and typically given in Hertz (Hz) or cycles/second. One intuitive way to understand the Fourier transform is through the **Fourier series** which states that a *periodic* function can be rewritten as the infinite summation of harmonically related sinusoids, similar to the Taylor series of arbitrary functions as infinite summations of polynomial terms. Then, as the “period” of any periodic function is allowed to approach infinity, i.e. it does not repeat in finite time, the Fourier series becomes the Fourier transform.

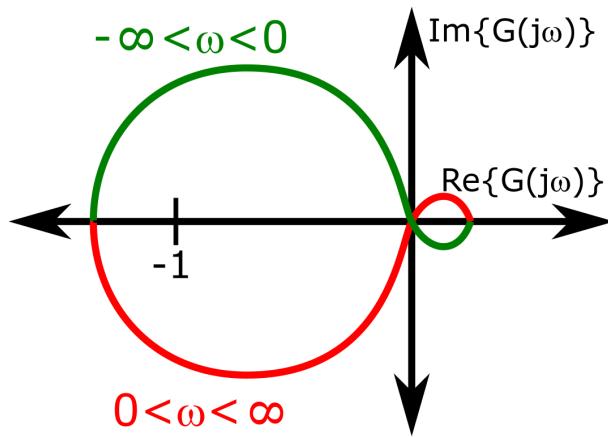
This transformation is useful through the **Fourier inversion theorem** which states that for “well-behaved” functions, e.g. LTI ODEs, it is possible to recover the function entirely from its Fourier transform using the **inverse Fourier transform** defined as

$$f(t) = \mathcal{F}^{-1}\{f(\omega)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\omega)e^{j\omega t} d\omega \quad (1.189)$$

Notably, the Fourier and inverse Fourier transforms are exactly equivalent to the Laplace transforms with the substitution  $s = j\omega$ .

Thus, the Fourier transform of a SISO LTI system,  $G(j\omega)$ , considered as a continuous function of all  $\omega \in [0, \infty)$  contains all responses of the SISO LTI system as a function of frequency, i.e. it models the **frequency response** of the SISO LTI system. With the Fourier inversion theorem in mind, the frequency response provides the SISO LTI system response to *all* harmonic sinusoids within any “well-behaved” input signal and from which one can exactly replicate the output signal. Notably, the class of well-behaved signals also includes the step signal, in fact, the *steady-state* output of the step response is equivalent to the frequency response at  $\omega = 0$ . Thus, the frequency response also provides the magnitude and phase values for all steady-state step and sinusoidal responses. Lastly, the frequency response can also be regarded as the Fourier transform of the impulse response, i.e.  $g(t) \rightarrow G(j\omega)$  for all  $\omega \in [0, \infty)$ . Two common plots used to analyze the frequency response of SISO LTI systems are the Nyquist and Bode plots. For MIMO LTI systems, one must employ multivariate frequency domain analysis which is discussed later in this textbook.

Similar to the Bode plot, the **Nyquist plot** can be used to analyze the frequency response of a transfer function, i.e.  $G(s)$  with  $s = j\omega$ . However, opposed to the Bode plot, the Nyquist plot visualizes the real and imaginary parts of  $G(j\omega)$  as a single curve with the real part on the horizontal axis and the imaginary on the vertical axis. It should also be noted the convention for the Nyquist plot is to plot over  $-\infty < \omega < \infty$  as opposed to  $\omega \geq 0$  for the Bode plot. This convention results in a reflected curve about the real axis for  $\omega < 0$  with respect to  $\omega > 0$  as a transfer function value  $G(j\omega) = \alpha + j\beta$  simply provides the complex conjugate for  $\omega < 0$  with respect to  $\omega > 0$ . An example of a Nyquist plot is the following:



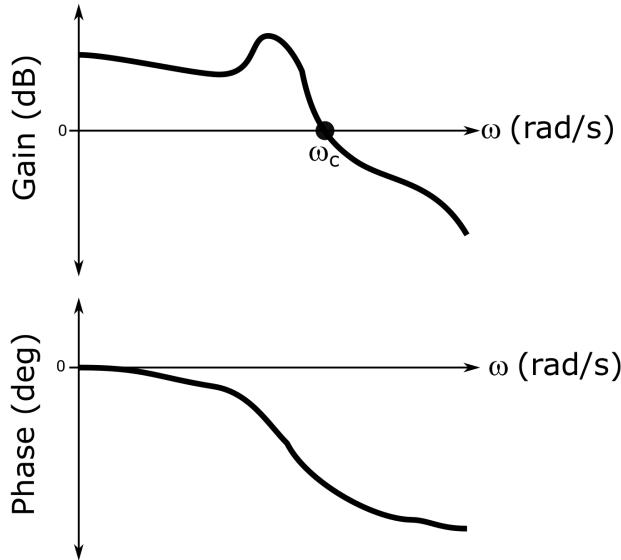
However, this textbook will focus on the Bode plot for analysis and design.

## Bode Plot

The **Bode plot** consists of two subplots:

1. the magnitude, also known as the gain,  $|G(j\omega)|$  (vertical) versus  $\omega$  (horizontal)
2. the phase  $\angle G(j\omega)$  (vertical) versus  $\omega$  (horizontal)

For the Bode plot, it is common to plot the frequency,  $\omega$ , on a  $\log_{10}$  scale in radians/sec, the phase,  $\angle G(j\omega)$ , in degrees ( $^\circ$ ), and the magnitude,  $|G(j\omega)|$ , in units of decibels (dB), i.e.  $20 \log_{10} |G(j\omega)|$ . In this manner, **crossover frequencies**,  $\omega_c$ , occur for  $|G(j\omega_c)|_{dB} = 0$  dB An example of a Bode plot is the following:



Some useful conversions for decibels to gain are as follows.

Decibels (dB)	Gain
40 dB	100
20 dB	10
6 dB	2
3 dB	$\sqrt{2}$
0 dB	1
-3 dB	$\frac{1}{\sqrt{2}}$
-6 dB	$\frac{1}{2}$
-20 dB	$\frac{1}{10}$
-40 dB	$\frac{1}{100}$

An important characteristic of the Bode plot, is that for multiplying complex numbers in polar form, e.g.  $n_1 = A_1 e^{j\phi_1}$  and  $n_2 = A_2 e^{j\phi_2}$ , one has

$$n_1 n_2 = A_1 A_2 e^{j(\phi_1 + \phi_2)} \quad (1.190)$$

which is a complex number with multiplicative magnitude, i.e.  $|A_1 A_2|$ , and additive phase, i.e.  $\phi_1 + \phi_2$ . However, if the magnitude is in dB, then one has

$$20 \log_{10} |A_1 A_2| = 20 \log_{10} |A_1| + 20 \log_{10} |A_2| \quad (1.191)$$

which is additive in dB, i.e.

$$|A_1 A_2|_{\text{dB}} = |A_1|_{\text{dB}} + |A_2|_{\text{dB}} \quad (1.192)$$

Recalling that SISO LTI systems can be rewritten into its poles and zeros, i.e.

$$G(s) = K \frac{(s - z_1) \cdots (s - z_m)}{(s - p_1) \cdots (s - p_n)} \quad (1.193)$$

the Bode plot for any SISO LTI system can be constructed by *adding* the Bode plots for each real and complex-conjugate pole and zero assuming the magnitude is in dB. As such, this section will discuss the Bode plots for differentiators, integrators, first-order zeros, first-order poles, underdamped second-order zeros, and underdamped second-order poles. This section will also discuss the **straight-line approximations** which provide a quick method to sketch Bode plots. From these plots and the previous additive property, one can construct Bode plots and sketch straight-line approximations for any SISO LTI system.

As an aside, this additive property also infers that if given the Bode plot of the frequency response, one could estimate the poles and zeros of an unknown SISO LTI system, i.e. the entire transfer function. This idea qualitatively forms the basis for frequency domain-based SID.

The **DC gain** of the Bode plot is the magnitude at  $\omega = 0$ , i.e.

$$|G(0)| = \frac{k z_1 \cdots z_m}{p_1 \cdots p_n} \quad (1.194)$$

This terminology comes from the origin of the Bode plot for frequency domain analysis of electrical systems using direct current (DC) and alternating current (AC) components. Note that the DC gain is also the steady-state gain for the step response of stable LTI systems.

The **asymptotic slope** of the Bode plot can be calculated by analyzing the limit of the gain asymptotically, i.e.

$$\lim_{\omega \rightarrow \infty} |G(j\omega)|_{\text{dB}} = 20 \log_{10} \omega^{m-n} \quad (1.195)$$

Next, consider increasing  $\omega$  by a factor of 10, i.e. a **decade** in frequency, and note

$$20 \log_{10}(10\omega)^x = 20 \log_{10} \omega^x + 20 \log_{10} 10^x = 20 \log_{10} \omega^x + 20x \quad (1.196)$$

Thus, as one moves up a decade in frequency, the asymptotic slope is

$$\left( \frac{d|G(j\omega)|_{\text{dB}}}{d(10\omega)} \right)_{\omega \rightarrow \infty} = 20(m - n) \quad (1.197)$$

Lastly, it should be noted that one must also account for the gain  $k \neq 1$ . If  $k < 0$ , then the phase will have  $\pm 180^\circ$  added. If  $|k| > 1$ , the magnitude will move up  $k_{\text{dB}}$ , while if  $|k| < 1$ , the magnitude will move down  $-|k|_{\text{dB}}$ .

## Frequency Response of Pure Differentiator

Consider a SISO LTI system with a transfer function of a **pure differentiator**

$$G(s) = s \quad (1.198)$$

which has a zero at  $s = 0$  and a DC gain of  $-\infty$  dB. The frequency response  $G(j\omega)$  can be computed as

$$G(j\omega) = j\omega \quad (1.199)$$

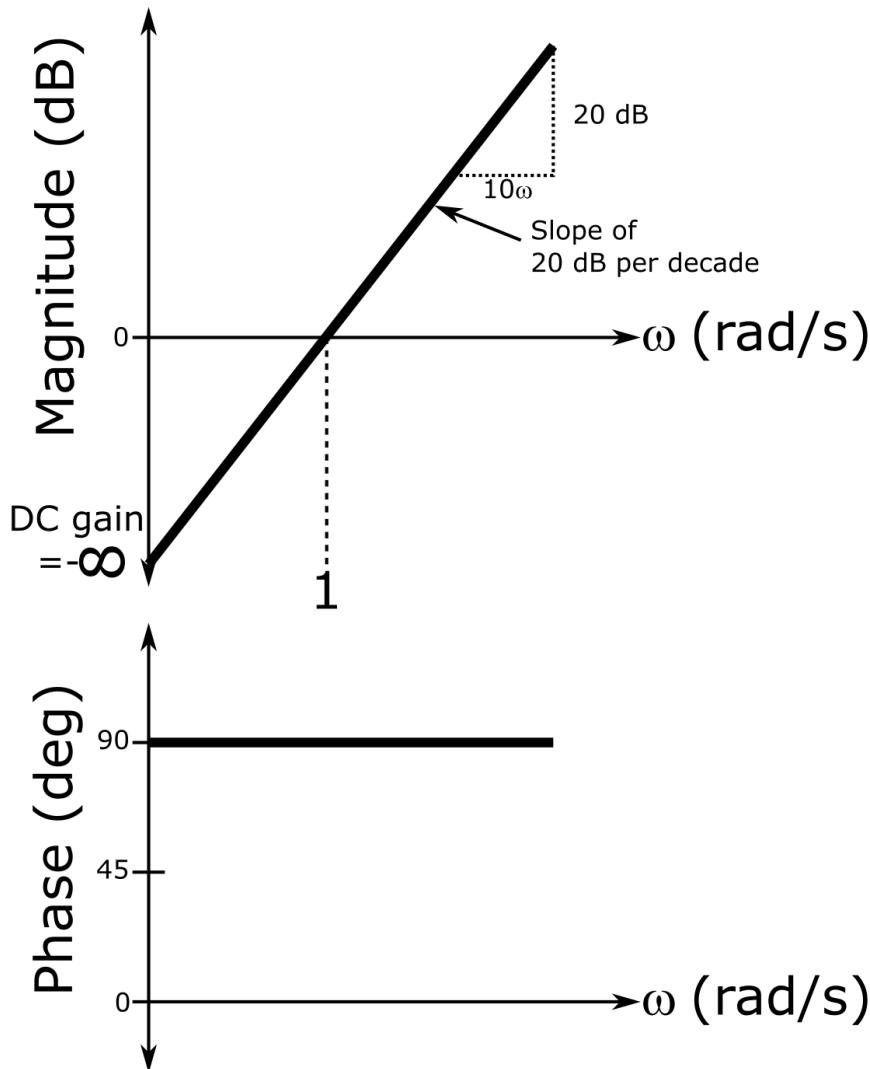
The magnitude can be computed as

$$|G(j\omega)|_{\text{dB}} = 20 \log_{10} |G(j\omega)| = 20 \log_{10} \omega \quad (1.200)$$

The phase can be computed as

$$\angle G(j\omega) = 90^\circ, \quad \forall \omega \quad (1.201)$$

Plotting, one has



### Frequency Response of Pure Integrator

Consider a SISO LTI system with a transfer function of a **pure integrator**

$$G(s) = \frac{1}{s} \quad (1.202)$$

which has a pole at  $s = 0$  and a DC gain of  $+\infty$  dB. The frequency response  $G(j\omega)$  can be computed as

$$G(j\omega) = \frac{1}{j\omega} = -j\frac{1}{\omega} \quad (1.203)$$

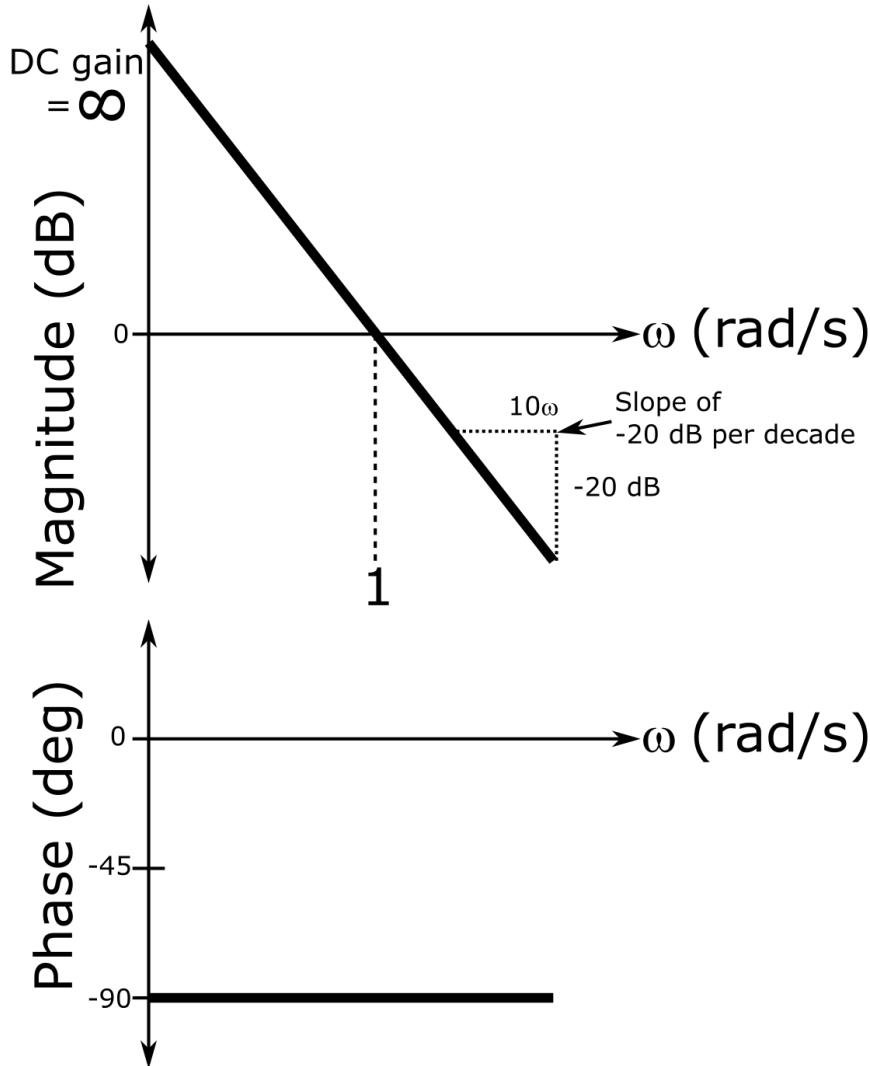
The magnitude can be computed as

$$|G(j\omega)|_{\text{dB}} = 20 \log_{10} |G(j\omega)| = 20 \log_{10} \frac{1}{\omega} = -20 \log_{10} \omega \quad (1.204)$$

The phase can be computed as

$$\angle G(j\omega) = -90^\circ, \quad \forall \omega \quad (1.205)$$

Plotting, one has



### Frequency Response of First-Order Zero

Consider a SISO LTI system with a transfer function of a first-order zero, i.e.

$$G(s) = s - z \quad (1.206)$$

which has a zero at  $s = z$ , a DC gain of  $20 \log_{10} |z|$ , and an asymptotic slope of 20 dB per decade. The frequency response can be computed as

$$G(j\omega) = -z + j\omega \quad (1.207)$$

The magnitude can be computed as

$$|G(j\omega)|_{\text{dB}} = 20 \log_{10} |G(j\omega)| = 20 \log_{10} \left( \sqrt{z^2 + \omega^2} \right) \quad (1.208)$$

where notably, at  $\omega = |z|$ , one has

$$|G(j|z|)|_{\text{dB}} = 20 \log_{10} |z| \sqrt{2} = 20 \log_{10} |z| + 3 \text{ dB} \quad (1.209)$$

or the DC gain plus 3 dB.

The phase can be computed as

$$\angle G(j\omega) = \tan^{-1} \left[ \frac{\omega}{-z} \right] \quad (1.210)$$

where notably

$$\angle G(j\omega) \rightarrow 90^\circ \text{ as } \omega \rightarrow \infty \quad (1.211)$$

for  $z > 0$

$$\angle G(j0) = \pm 180^\circ \quad (1.212)$$

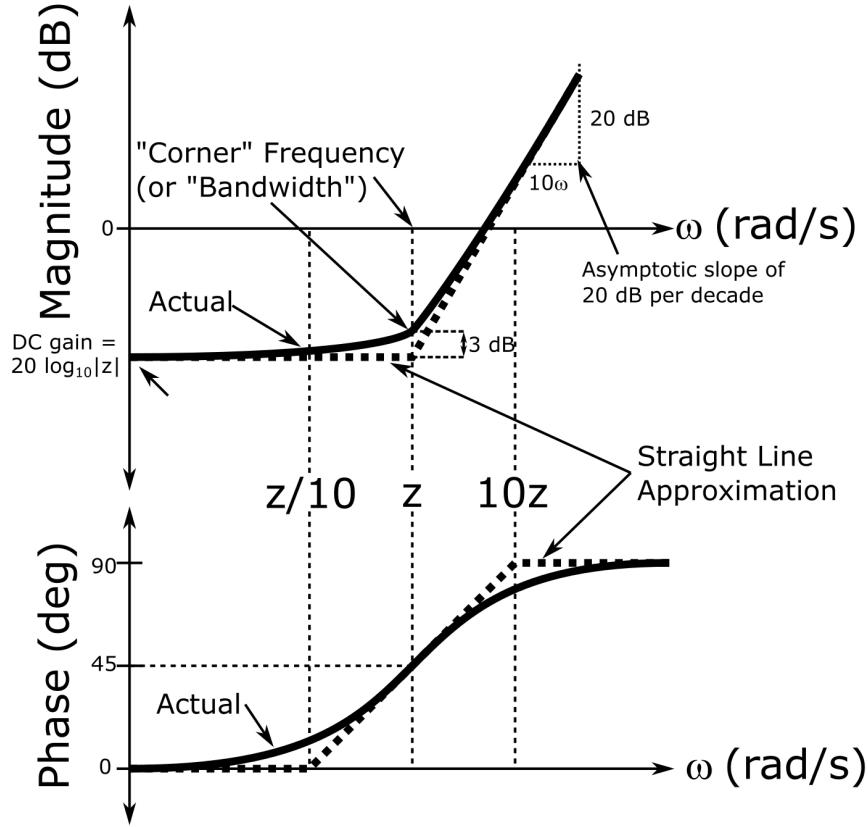
$$\angle G(jz) = 135^\circ \quad (1.213)$$

and for  $z < 0$

$$\angle G(j0) = 0^\circ \quad (1.214)$$

$$\angle G(-jz) = 45^\circ \quad (1.215)$$

Plotting for  $-1 < z < 0$ , one has



### Frequency Response of First-Order Pole

Consider a SISO LTI system with a transfer function of a first-order pole, i.e.

$$G(s) = \frac{1}{s - p} \quad (1.216)$$

which has a pole at  $s = p$ , a DC gain of  $-20 \log_{10} |p|$ , and an asymptotic slope of -20 dB per decade.

The frequency response  $G(j\omega)$  can be computed as

$$G(j\omega) = \frac{1}{j\omega - p} \quad (1.217)$$

$$G(j\omega) = \frac{1}{j\omega - p} \frac{-p - j\omega}{-p - j\omega} \quad (1.218)$$

$$G(j\omega) = \frac{-p}{p^2 + \omega^2} - j \frac{\omega}{p^2 + \omega^2} \quad (1.219)$$

The magnitude can be computed as

$$|G(j\omega)|_{\text{dB}} = 20 \log_{10} |G(j\omega)| = 20 \log_{10} \frac{1}{\sqrt{p^2 + \omega^2}} \quad (1.220)$$

where notably

$$|G(jz)|_{\text{dB}} = 20 \log_{10} |z| \sqrt{2} = -20 \log_{10} |p| - 3 \text{ dB} \quad (1.221)$$

or the DC gain minus 3 dB.

The phase can be computed as

$$\angle G(j\omega) = \tan^{-1} \left[ \frac{\frac{-\omega}{p^2 + \omega^2}}{\frac{p}{p^2 + \omega^2}} \right] \quad (1.222)$$

$$\angle G(j\omega) = \tan^{-1} \left[ \frac{-\omega}{p} \right] \quad (1.223)$$

where notably

$$\angle G(j\omega) \rightarrow -90^\circ \text{ as } \omega \rightarrow \infty \quad (1.224)$$

for  $p > 0$

$$\angle G(j0) = \pm 180^\circ \quad (1.225)$$

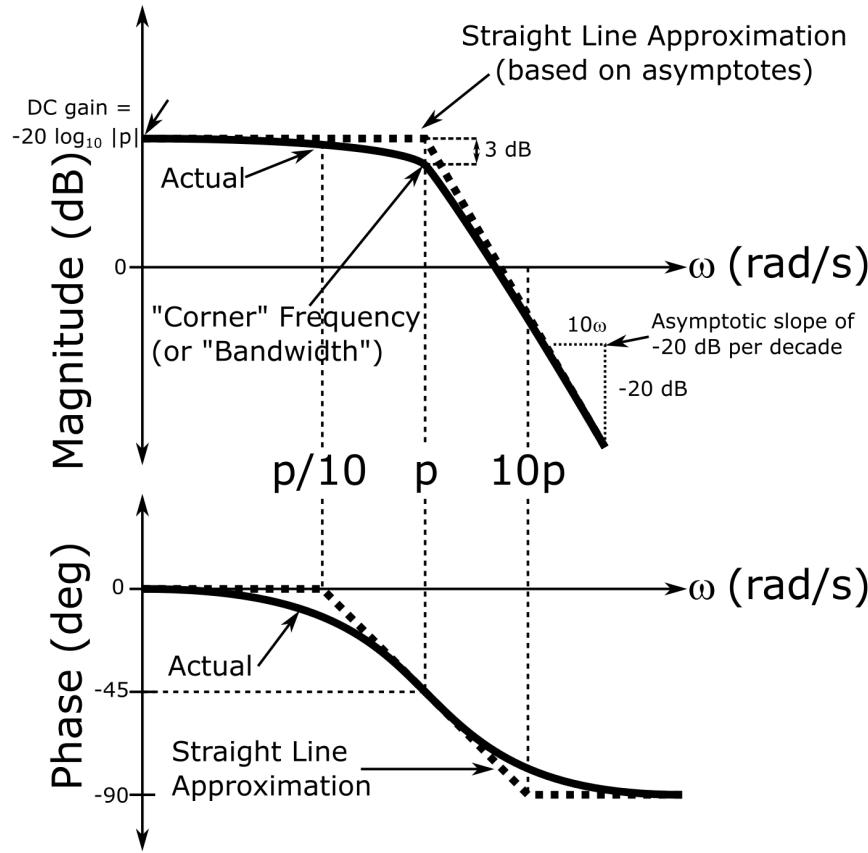
$$\angle G(jp) = -135^\circ \quad (1.226)$$

and for  $p < 0$

$$\angle G(j0) = 0^\circ \quad (1.227)$$

$$\angle G(-jp) = -45^\circ \quad (1.228)$$

Plotting for  $-1 < p < 0$ , i.e. a stable transfer function, one has



### Frequency Response for Underdamped Second-Order Zero

Consider a SISO LTI system with a transfer function of an underdamped second-order zero, i.e.

$$G(s) = s^2 + 2\zeta\omega_n s + \omega_n^2 \quad (1.229)$$

which has zeros at  $s = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}$ , a DC gain of  $20 \log_{10} \omega_n^2$ , and an asymptotic slope of 40 dB per decade.

The frequency response can be computed as

$$G(j\omega) = -\omega^2 + 2j\zeta\omega_n\omega + \omega_n^2 \quad (1.230)$$

$$G(j\omega) = (\omega_n^2 - \omega^2) + 2j\zeta\omega_n\omega \quad (1.231)$$

The magnitude can be computed as

$$|G(j\omega)|_{\text{dB}} = 20 \log_{10} |G(j\omega)| = 20 \log_{10} \left( \sqrt{(\omega_n^2 - \omega^2)^2 + 4\zeta^2\omega_n^2\omega^2} \right) \quad (1.232)$$

where notably

$$|G(j|\omega_n|)|_{\text{dB}} = 20 \log_{10} \left( \sqrt{4\zeta^2\omega_n^4} \right) \quad (1.233)$$

$$|G(j\omega_n)|_{\text{dB}} = 20 \log_{10} \omega_n^2 + 20 \log_{10} 2\zeta \quad (1.234)$$

or the DC gain plus some amount dependent on  $\zeta$ . For underdamped zeros, i.e.  $0 < \zeta < 1$ , one has

$$|G(j\omega_n)|_{\text{dB}} < 20 \log_{10} \omega_n^2 + 20 \log_{10} 2(1) \approx 20 \log_{10} \omega_n^2 + 6\text{dB} \quad (1.235)$$

and for  $\zeta < 0.5$ ,  $|G(j\omega_n)|_{\text{dB}} < 20 \log_{10} \omega_n^2$ . Furthermore, if  $\zeta < \frac{1}{\sqrt{2}} = 0.707$ , then it can be proven using calculus that at  $\omega = \omega_n \sqrt{1 - 2\zeta^2}$  there will be a **resonant valley**. This valley becomes more pronounced for smaller damping ratios and is significant for  $\zeta < 0.5$ .

The phase can be computed as

$$\angle G(j\omega) = \tan^{-1} \left[ \frac{2\zeta\omega_n\omega}{\omega_n^2 - \omega^2} \right] \quad (1.236)$$

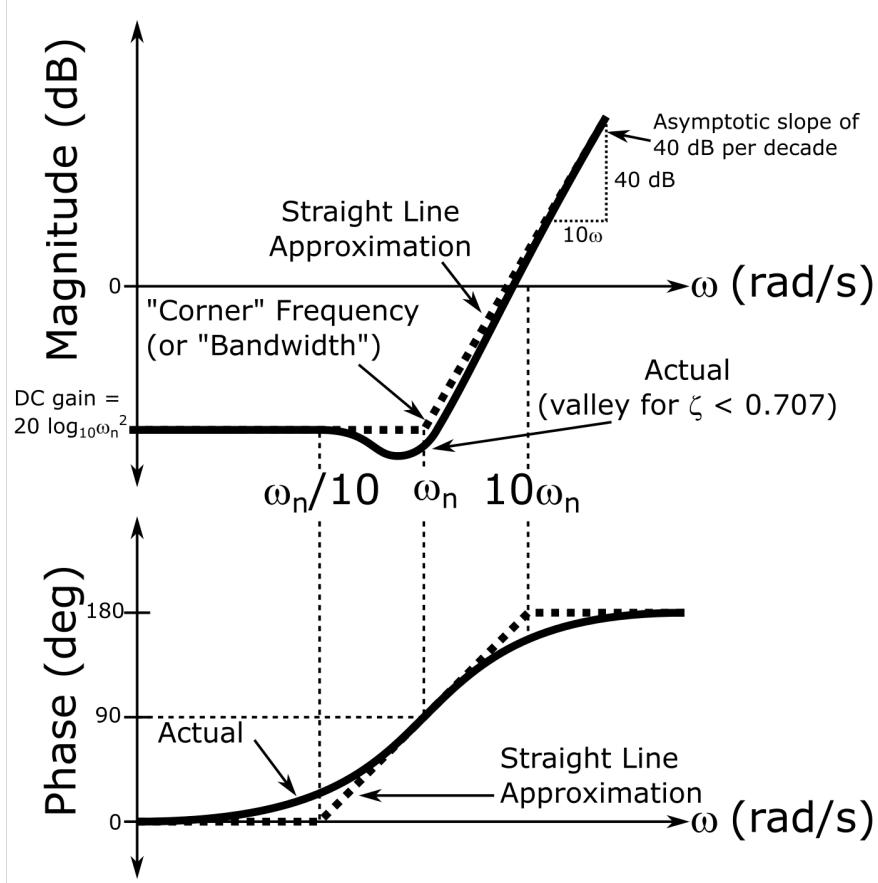
where notably

$$\angle G(j\omega_n) = 90^\circ \quad (1.237)$$

and

$$\angle G(j\omega) \rightarrow \pm 180^\circ \text{ as } \omega \rightarrow \infty \quad (1.238)$$

Plotting for  $0 < \omega_n^2 < 1$  and  $0 < \zeta < 1$ , one has



## Frequency Response for Underdamped Second-Order Pole

Consider a SISO LTI system with a transfer function of an underdamped second-order pole, i.e.

$$G(s) = \frac{1}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad (1.239)$$

which has poles at  $s = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}$ , a DC gain of  $-20 \log_{10} \omega_n^2$ , and an asymptotic slope of -40 dB per decade.

The frequency response can be computed as

$$G(j\omega) = \frac{1}{-\omega^2 + 2j\zeta\omega_n\omega + \omega_n^2} \quad (1.240)$$

$$G(j\omega) = \frac{1}{\omega_n^2 - \omega^2 + 2j\zeta\omega_n\omega} \frac{\omega_n^2 - \omega^2 - 2j\zeta\omega_n\omega}{\omega_n^2 - \omega^2 - 2j\zeta\omega_n\omega} \quad (1.241)$$

$$G(j\omega) = \frac{\omega_n^2 - \omega^2}{(\omega_n^2 - \omega^2)^2 + 4\zeta^2\omega_n^2\omega^2} - j \frac{2\zeta\omega_n\omega}{(\omega_n^2 - \omega^2)^2 + 4\zeta^2\omega_n^2\omega^2} \quad (1.242)$$

The magnitude can be computed as

$$|G(j\omega)|_{\text{dB}} = 20 \log_{10} |G(j\omega)| = 20 \log_{10} \frac{1}{\sqrt{(\omega_n^2 - \omega^2)^2 + 4\zeta^2\omega_n^2\omega^2}} \quad (1.243)$$

where notably

$$|G(j|\omega_n|)|_{\text{dB}} = 20 \log_{10} \frac{1}{\sqrt{4\zeta^2\omega_n^4}} \quad (1.244)$$

$$|G(j|\omega_n|)|_{\text{dB}} = -20 \log_{10} \omega_n^2 - 20 \log_{10} 2\zeta \quad (1.245)$$

or the DC gain minus some amount dependent on  $\zeta > 0$ . For underdamped zeros, i.e.  $0 < \zeta < 1$ , one has

$$|G(j|\omega_n|)|_{\text{dB}} > 20 \log_{10} \omega_n^2 - 20 \log_{10} 4 \approx 20 \log_{10} \omega_n^2 - 6 \text{dB} \quad (1.246)$$

and for  $\zeta < 0.5$ ,  $|G(j|\omega_n|)|_{\text{dB}} > 20 \log_{10} \omega_n^2$ . Furthermore, if  $\zeta < \frac{1}{\sqrt{2}} = 0.707$ , then it can be proven using calculus that at  $\omega = \omega_n\sqrt{1 - 2\zeta^2}$  there will be a **resonant peak** of magnitude where this peak becomes more pronounced for small damping ratios and is significant for  $\zeta < 0.5$ .

The phase can be computed as

$$\angle G(j\omega) = \tan^{-1} \left[ \frac{-2\zeta\omega_n\omega}{\omega_n^2 - \omega^2} \right] \quad (1.247)$$

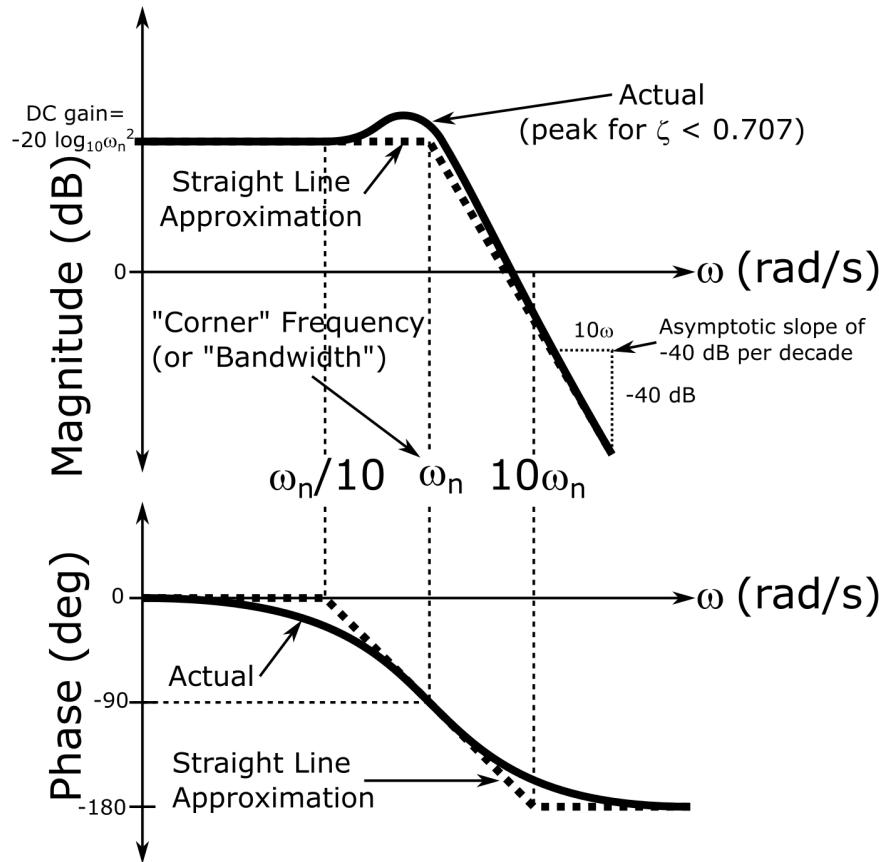
where notably

$$\angle G(j\omega_n) = -90^\circ \quad (1.248)$$

and

$$\angle G(j\omega) \rightarrow \pm 180^\circ \text{ as } \omega \rightarrow \infty \quad (1.249)$$

Plotting for  $0 < \omega_n^2 < 1$  and  $0 < \zeta < 1$ , one has



## References

For more information, please refer to the following

- Nelson, R. C., “4.2 Second-Order Differential Equations,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 133-138
- Nelson, R. C., “Appendix D: Review of Control System Analysis Techniques,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 439-434
- Schmidt, D. K., “10.1 Linear System Analysis - A Just-In-Time Tutorial\*,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 548-562
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “3.3 Transfer Function Models,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 168-170

## 1.5 Continuous-Time MIMO LTI System Responses

### Initial Condition Response for MIMO LTI Systems

For an initial condition at  $\vec{x}(0)$ , the initial condition response, i.e. the zero-input or free response, of a continuous-time state-space system representation can be reduced to the following

$$\dot{\vec{x}}(t) = A \vec{x}(t) , \quad (1.250)$$

By using the JCF, i.e.

$$\vec{x}(t) = V \vec{z}(t) \quad (1.251)$$

where  $V$  is the matrix of  $n$  eigenvectors of  $A$ , then, the state-space representation can be transformed to using a new state,  $\vec{z}(t)$  as

$$\dot{\vec{z}}(t) = \Lambda \vec{z}(t) \quad (1.252)$$

where  $\Lambda$  is in Jordan form, i.e. diagonal or nearly diagonal.

Assuming  $A$  is diagonalizable, then  $\Lambda$  is diagonal and the state-space is

$$\begin{bmatrix} \dot{z}_1(t) \\ \vdots \\ \dot{z}_n(t) \end{bmatrix} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} z_1(t) \\ \vdots \\ z_n(t) \end{bmatrix} \quad (1.253)$$

$$\begin{bmatrix} \dot{z}_1(t) \\ \vdots \\ \dot{z}_n(t) \end{bmatrix} = \begin{bmatrix} \lambda_1 z_1(t) \\ \vdots \\ \lambda_n z_n(t) \end{bmatrix} \quad (1.254)$$

Then, since each component is independent, the free response is given by the homogeneous solution to a first order ODE, i.e.

$$\begin{bmatrix} z_1(t) \\ \vdots \\ z_n(t) \end{bmatrix} = \begin{bmatrix} e^{\lambda_1 t} z_1(0) \\ \vdots \\ e^{\lambda_n t} z_n(0) \end{bmatrix} \quad (1.255)$$

$$\begin{bmatrix} z_1(t) \\ \vdots \\ z_n(t) \end{bmatrix} = \begin{bmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{bmatrix} \begin{bmatrix} z_1(0) \\ \vdots \\ z_n(0) \end{bmatrix} \quad (1.256)$$

$$\vec{z}(t) = \begin{bmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{bmatrix} \vec{z}(0) \quad (1.257)$$

and using the inverse transformation,  $\vec{x}(t) = V^{-1} \vec{z}(t)$ , one can solve for the free response solution in the original state as

$$V^{-1} \vec{x}(t) = \begin{bmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{bmatrix} V^{-1} \vec{x}(0) \quad (1.258)$$

$$\vec{x}(t) = V \begin{bmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{bmatrix} V^{-1} \vec{x}(0) \quad (1.259)$$

$$\vec{x}(t) = [\vec{v}_1 \ \cdots \ \vec{v}_n] \begin{bmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{bmatrix} V^{-1} \vec{x}(0) \quad (1.260)$$

$$\vec{x}(t) = [e^{\lambda_1 t} \vec{v}_1 \ \cdots \ e^{\lambda_n t} \vec{v}_n] V^{-1} \vec{x}(0) \quad (1.261)$$

and representing

$$V^{-1} \vec{x}(0) = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \quad (1.262)$$

where  $c_1, \dots, c_n$  are scalar constants, one has

$$\vec{x}(t) = [e^{\lambda_1 t} \vec{v}_1 \ \cdots \ e^{\lambda_n t} \vec{v}_n] \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \quad (1.263)$$

$$\vec{x}(t) = c_1 e^{\lambda_1 t} \vec{v}_1 + \dots + c_n e^{\lambda_n t} \vec{v}_n \quad (1.264)$$

which is a similar form to the SISO LTI free response except here the eigenvectors have an additional affect. By analyzing the relative strength of each element in an eigenvector, it is simpler to observe which eigenvalues affect which states more than others. If some eigenvalues/eigenvectors are complex conjugate pairs, then this expression can be rewritten using sin and cos functions to get only the real part of the solution for  $x(t)$ . This type of analysis is often called **modal analysis** where each real eigenvalue or complex conjugate eigenvalue pair denotes one **mode** of the solution.

It is also important to note that if  $A$  is not diagonalizable, the general solution can be found using **Jordan Chains** which are related to the generalized eigenvectors for repeated eigenvalues. This will include  $t, t^2, \dots, t^{n-1}$  terms and is quite similar to the SISO LTI free response for repeated roots of the characteristic equation. This result is left to other linear algebra resources.

In particular, continuous-time LTI systems are globally stable if and only if the eigenvalues of  $A$  are in the closed left half of the complex plane (LHP), i.e. the real part of the eigenvalues  $\leq 0$ . As a partial proof, consider the free response

$$\vec{x}(t) = c_1 e^{\lambda_1 t} \vec{v}_1 + \dots + c_n e^{\lambda_n t} \vec{v}_n \quad (1.265)$$

as  $t \rightarrow \infty$ , each  $i^{\text{th}}$  component with:

- $\text{Real}(\lambda_i) < 0$  will  $\rightarrow 0$
- $\text{Real}(\lambda_i) = 0$  will remain constant if  $\lambda_i$  is purely real
- $\text{Real}(\lambda_i) = 0$  will oscillate with a constant magnitude if  $\lambda_i$  is purely imaginary

all of which are bounded (i.e. there exists some  $\epsilon$ ). A full proof must include non-diagonalizable  $A$  and can be done in a similar fashion.

Each of these stabilities can be expanded to include two different types. Continuous-time MIMO LTI systems are **marginally stable** if  $\text{real}(\lambda_i) \leq 0 \forall i$  with  $\text{real}(\lambda_j) = 0$  for at least one index  $j$  and are **asymptotically stable** if all  $\text{real}(\lambda_i) < 0 \forall i$ .

### Frequency Response of MIMO LTI Systems

Recall the frequency response of a SISO LTI system is the Fourier transform of the system, i.e.  $G(j\omega)$ , which, by the Fourier inversion theorem, provides the system response to *all* harmonic sinusoids within any “well-behaved” input signal and from which one can exactly replicate the output signal. Thus, the frequency response also provides the magnitude and phase values for all steady-state step and sinusoidal responses. Furthermore, the Nyquist and Bode plots are used to analyze the frequency response of SISO LTI systems.

In the same way, the frequency response of MIMO LTI systems is the Fourier transform of the system, i.e.  $[G(j\omega)]$ , which provides the same information, albeit with potential coupling between all inputs and outputs via the transfer function matrix. However, recalling the definition of the singular value decomposition (SVD), one can model the transfer function matrix at a particular frequency as

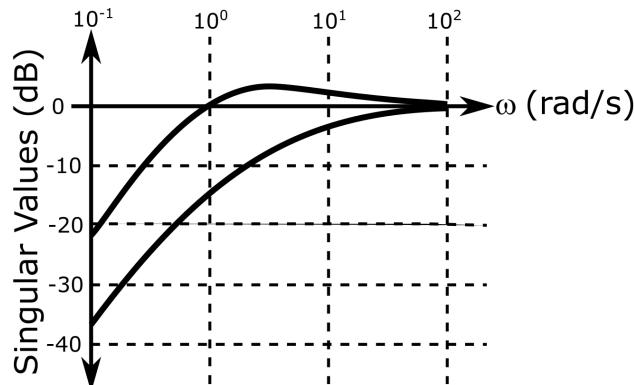
$$[G(j\omega)] = U(j\omega)\Sigma(j\omega)V^{-1}(j\omega) \quad (1.266)$$

where each of these matrices depends on frequency. Though the values of the unitary matrices,  $U$  and  $V$ , may change. The gain of the “directions” from input to output encoded in these unitary matrices is captured by the singular values within  $\Sigma$ . Thus, frequency response of MIMO LTI systems is typically done by plotting the singular values as a function of  $\omega$  called the **singular value plot** also known as the **sigma plot** or  **$\sigma$ -plot**.

As an example, consider the system

$$[G(s)] = \begin{bmatrix} \frac{s}{s+3} & \frac{-6s}{s^2+6s+9} \\ 0 & \frac{s}{s+3} \end{bmatrix} \quad (1.267)$$

which provides the following  $\sigma$ -plot



## Signal and System Norms

Consider a piecewise continuous signal vector,  $\vec{u}(t) \in \mathbb{R}^{n_u}$ . Then, one can define the **signal  $p$ -norm**, also known as the **signal  $\mathcal{L}_p$ -norm** as

$$\|\vec{u}(t)\|_p = \left( \int_0^\infty \sum_{i=1}^{n_u} |u_i(t)|^p dt \right)^{\frac{1}{p}} \quad (1.268)$$

where  $u_i(t)$  is the  $i^{\text{th}}$  element of  $\vec{u}(t)$ , which may or may not be finite. Of particular interest is the **signal  $\mathcal{L}_1$ -norm**, i.e.

$$\|\vec{u}(t)\|_1 = \int_0^\infty \sum_{i=1}^{n_u} |u_i(t)| dt \quad (1.269)$$

and the **signal  $\mathcal{L}_2$ -norm**, i.e.

$$\|\vec{u}(t)\|_2 = \left( \int_0^\infty \sum_{i=1}^{n_u} u_i(t)^2 dt \right)^{\frac{1}{2}} \quad (1.270)$$

which can be redefined as

$$\|\vec{u}(t)\|_2 = \left( \int_0^\infty \text{Tr} [\vec{u}(t)^T \vec{u}(t)] dt \right)^{\frac{1}{2}} \quad (1.271)$$

or by Parseval's theorem

$$\|\vec{u}(t)\|_2 = \left( \frac{1}{2\pi} \int_{-\infty}^\infty \text{Tr} [\vec{u}(j\omega)^* \vec{u}(j\omega)] d\omega \right)^{\frac{1}{2}} \quad (1.272)$$

Next, consider the LTI system  $\vec{y}(s) = G(s) \vec{u}(s)$ . Then, the  $p \leftarrow q$ -induced system norms, also known as the  $\mathcal{L}_{p,q}$  induced system norms, can be defined as

$$\|G\|_{p \leftarrow q} = \max_{0 \neq \|\vec{u}\|_q \leq \infty, \vec{x}(0)=0} \frac{\|\vec{y}\|_p}{\|\vec{u}\|_q} \quad (1.273)$$

which can also be stated as the smallest constant,  $c$ , such that  $\|\vec{y}\|_p \leq c \|\vec{u}\|_q$ . Note that if  $G$  is unstable, then  $\|\vec{y}\|_p \rightarrow \infty$  and  $\|G\|_{p \leftarrow q} = \infty$ . Of particular interest is the  $\|G\|_{2 \leftarrow 2}$  norms which is also known as the  $\mathcal{H}_\infty$ -norm of  $G$  for stable  $G$ , and is denoted, in this case by

$$\|G\|_\infty = \|G\|_{2 \leftarrow 2} \quad (1.274)$$

which is finite if and only if  $G$  is strictly proper, e.g.  $D = 0$  in its LTI state-space model, with no poles on the  $j\omega$  axis. By definition of the LTI system as  $\vec{y}(j\omega) = G(j\omega) \vec{u}(j\omega)$ , and the definition of the singular values of the transfer function matrix,  $G(j\omega)$ , one can write

$$\bar{\sigma}(G(j\omega)) = \max_{0 \neq \|\vec{u}(j\omega)\|_2} \frac{\|G(j\omega) \vec{u}(j\omega)\|_2}{\|\vec{u}(j\omega)\|_2} \quad (1.275)$$

where  $\bar{\sigma}(G(j\omega))$  is the maximum singular value of the transfer function matrix  $G(j\omega)$  evaluated at  $\omega$ . Thus, over the entire frequency spectrum, one has

$$\|G\|_\infty = \max_{\omega} \bar{\sigma}(G(j\omega)) \quad (1.276)$$

which for SISO systems is simply the maximum magnitude of the Bode plot. For MIMO systems, the  $\mathcal{H}_\infty$ -norm is simply the maximum singular value across all frequencies, i.e. the peak on the  $\sigma$ -plot.

Lastly, the  $\mathcal{H}_2$ -norm can be defined for a stable LTI system,  $G$ , as

$$\|G\|_2 = \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \|G(j\omega)\|_F^2 d\omega \right)^{\frac{1}{2}} \quad (1.277)$$

where  $\|G(j\omega)\|_F$  is the Frobenius or entry-wise  $L_{2,2}$  norm of the transfer function matrix evaluated at  $j\omega$ . This can be alternatively written as

$$\|G\|_2 = \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{Tr}[G(j\omega)G(j\omega)^*] d\omega \right)^{\frac{1}{2}} \quad (1.278)$$

which is finite if and only if  $G$  is strictly proper, e.g.  $D = 0$  in its state-space representation, with no poles on the  $j\omega$  axis. It should also be noted that for SISO systems, the  $\mathcal{H}_2$ -norm becomes

$$\|G\|_2 = \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} |G(j\omega)|^2 d\omega \right)^{\frac{1}{2}} \quad (1.279)$$

which is a measure of the area under the Bode plot of the system.

Using the SVD of  $G(j\omega) = U(j\omega)\Sigma(j\omega)V(j\omega)^*$ , one has

$$\|G\|_2 = \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{Tr}[V(j\omega)\Sigma(j\omega)^*U(j\omega)^*U(j\omega)\Sigma(j\omega)V(j\omega)^*] d\omega \right)^{\frac{1}{2}} \quad (1.280)$$

$$\|G\|_2 = \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{Tr}[\Sigma(j\omega)^*\Sigma(j\omega)] d\omega \right)^{\frac{1}{2}} \quad (1.281)$$

$$\|G\|_2 = \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \sum_{i=1}^n \sigma_i(j\omega)^2 d\omega \right)^{\frac{1}{2}} \quad (1.282)$$

Thus, the  $\mathcal{H}_2$ -norm can be considered as an average of the singular values also averaged across all frequencies.

An important interpretation of the  $\mathcal{H}_2$ -norm occurs by considering  $\vec{u}(t)$  as a white noise random process with unit variance, i.e.

$$\mathbb{E}[\vec{u}(t_1)\vec{u}(t_2)^T] = I_{n_u}\delta(t_1 - t_2) \quad (1.283)$$

Then, the  $\mathcal{H}_2$ -norm squared is the expected power in the output signal, i.e.

$$\|G\|_2^2 = \mathbb{E} \left[ \lim_{t_f \rightarrow \infty} \frac{1}{t_f} \int_0^{t_f} \text{Tr}[\vec{y}(t)^T \vec{y}(t)] dt \right] \quad (1.284)$$

which can be shown using the convolution integral for  $\vec{y}(t)$ . This provides a useful motivation for using  $\mathcal{H}_2$  OCPs when *unobservable* random processes are significant inputs to the plant.

To calculate the  $\mathcal{H}_2$ -norm, note that by Parseval's theorem one can write the transfer function matrix as the solution to the impulse response of the MIMO system with  $D = 0$ , i.e.

$$\|G\|_2 = \left( \int_0^{\infty} \text{Tr}[Ce^{At}BB^Te^{A^Tt}C^T] dt \right)^{\frac{1}{2}} \quad (1.285)$$

$$\|G\|_2^2 = \text{Tr} \left[ C \int_0^\infty e^{At} BB^T e^{A^T t} dt C^T \right] \quad (1.286)$$

and defining the **controllability gramian**,  $W_C = W_C^T$ , as

$$W_C = \int_0^\infty e^{At} BB^T e^{A^T t} dt \quad (1.287)$$

one has

$$\|G\|_2^2 = \text{Tr} [CW_C C^T] \quad (1.288)$$

where  $W_C$  is found by noting

$$\frac{d}{dt} e^{At} BB^T e^{A^T t} = Ae^{At} BB^T e^{A^T t} + e^{At} BB^T e^{A^T t} A^T \quad (1.289)$$

and integrating from 0 to  $\infty$ , one has

$$e^{At} BB^T e^{A^T t} \Big|_0^\infty = \int_0^\infty Ae^{At} BB^T e^{A^T t} dt + \int_0^\infty e^{At} BB^T e^{A^T t} A^T dt \quad (1.290)$$

or

$$-BB^T = AW_C + W_C A^T \quad (1.291)$$

Thus,  $W_C$  is the unique solution to

$$AW_C + W_C A^T + BB^T = 0 \quad (1.292)$$

which exists if and only if  $A$  is stable.

Alternatively, one can also write

$$\|G\|_2 = \left( \int_0^\infty \text{Tr} [B^T e^{A^T t} C^T C e^{At} B] dt \right)^{\frac{1}{2}} \quad (1.293)$$

from which  $W_O = W_O^T$  is the **observability gramian**

$$W_O = \int_0^\infty e^{A^T t} C^T C e^{At} dt \quad (1.294)$$

which can form the  $\mathcal{H}_2$ -norm as

$$\|G\|_2^2 = \text{Tr} [B^T W_O B] \quad (1.295)$$

and is the unique solution to

$$A^T W_O + W_O A + C^T C = 0 \quad (1.296)$$

which exists if and only if  $A$  is stable.

Both Equations 1.292 and 1.296 are types of matrix Lyapunov equations and can be solved using numerical algorithms. A **matrix Lyapunov equation** is a type of LME of the form

$$A^T X + X A + Q = 0 \quad (1.297)$$

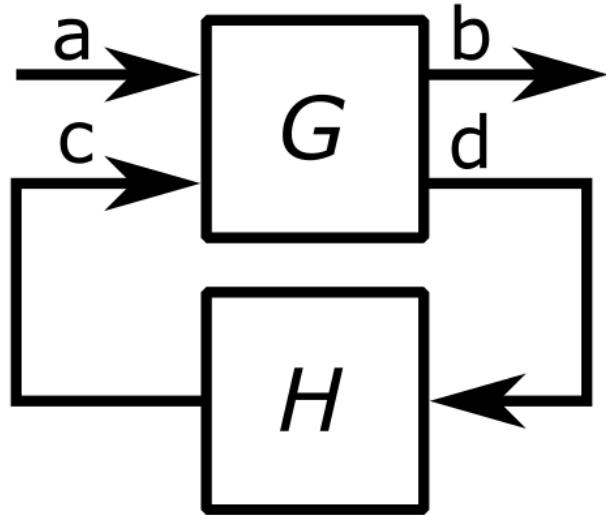
for which there is a unique solution

$$X = \int_0^\infty e^{A^T t} Q e^{At} dt \quad (1.298)$$

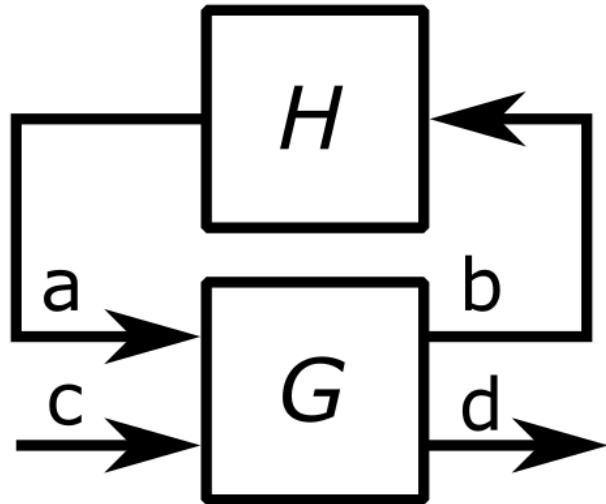
which is well-defined since  $A$  is assumed stable.  $X$  is typically obtained numerically using certain eigenvalue decompositions, not directly from this integral.

### Linear Fractional Transformations

A **linear fractional transformations (LFT)** models a feedback connection of two matrices which can be used to represent MIMO LTI systems via the transfer function matrix. These can be either an **lower LFT**, i.e.



which is denoted as  $F_L(G, H)$  or a **upper LFT**, i.e.



which is denoted as  $F_U(G, H)$ .

A lower LFT represents the following algebraic equations

$$\begin{aligned} \begin{bmatrix} \vec{b} \\ \vec{d} \end{bmatrix} &= \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} \vec{a} \\ \vec{c} \end{bmatrix} \\ \vec{c} &= H \vec{d} \end{aligned} \quad (1.299)$$

Substituting the third equation into the second equation provides

$$\vec{d} = G_{21} \vec{a} + G_{22} H \vec{d} \quad (1.300)$$

Thus, if  $I_{n_d} - G_{22}H$  is non-singular, i.e. invertible, then the lower LFT is well-posed and this equation can be solved as

$$\vec{d} = (I_{n_d} - G_{22}H)^{-1} G_{21} \vec{a} \quad (1.301)$$

which can be substituted into the first equation with  $\vec{c} = H \vec{d}$  to obtain

$$\begin{aligned} \vec{b} &= \left( G_{11} + G_{12}H (I_{n_d} - G_{22}H)^{-1} G_{21} \right) \vec{a} \\ \vec{b} &= F_L(G, H) \vec{a} \end{aligned} \quad (1.302)$$

An upper LFT represents the following algebraic equations

$$\begin{aligned} \begin{bmatrix} \vec{b} \\ \vec{d} \end{bmatrix} &= \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} \vec{a} \\ \vec{c} \end{bmatrix} \\ \vec{a} &= H \vec{b} \end{aligned} \quad (1.303)$$

Substituting the third equation into the first equation provides

$$\vec{b} = G_{11} H \vec{b} + G_{12} \vec{c} \quad (1.304)$$

Thus, if  $I_{n_b} - G_{11}H$  is non-singular, i.e. invertible, then the upper LFT is well-posed and this equation can be solved as

$$\vec{b} = (I_{n_b} - G_{11}H)^{-1} G_{12} \vec{c} \quad (1.305)$$

which can be substituted into the second equation with  $\vec{a} = H \vec{b}$  to obtain

$$\begin{aligned} \vec{b} &= \left( G_{22} + G_{21}H (I_{n_b} - G_{11}H)^{-1} G_{12} \right) \vec{c} \\ \vec{d} &= F_U(G, H) \vec{c} \end{aligned} \quad (1.306)$$

An important property of LFTs is that one can represent block diagram algebra for serial, parallel, feedback, and inverse interconnections of LFTs as LFTs.

As an example of a serial and parallel interconnection, consider the two upper LFTs

$$\begin{aligned} \begin{bmatrix} \vec{b}_1 \\ \vec{d}_1 \end{bmatrix} &= \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} \vec{a}_1 \\ \vec{c}_1 \end{bmatrix} \\ \vec{a}_1 &= H_1 \vec{b}_1 \end{aligned} \quad (1.307)$$

and

$$\begin{bmatrix} \vec{b}_2 \\ \vec{d}_2 \end{bmatrix} = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} \begin{bmatrix} \vec{a}_2 \\ \vec{c}_2 \end{bmatrix}$$

$$\vec{a}_2 = H_2 \vec{b}_2$$
(1.308)

connected in series from  $F_U(L, H_1)$  to  $F_U(N, H_2)$ , i.e.

$$\vec{c}_2 = \vec{d}_1$$
(1.309)

Combining, one has the upper LFT

$$\begin{bmatrix} \vec{b}_1 \\ \vec{b}_2 \\ \vec{d}_2 \end{bmatrix} = \begin{bmatrix} L_{11} & 0 & L_{12} \\ N_{12}L_{21} & N_{11} & N_{12}L_{22} \\ N_{22}L_{21} & N_{21} & N_{22}L_{22} \end{bmatrix} \begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vec{c}_1 \end{bmatrix}$$

$$\begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \end{bmatrix} = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix} \begin{bmatrix} \vec{b}_1 \\ \vec{b}_2 \end{bmatrix}$$
(1.310)

or connected in parallel, i.e.

$$\vec{c} = \vec{c}_1 = \vec{c}_2 \quad \& \quad \vec{d} = \vec{d}_1 + \vec{d}_2$$
(1.311)

Combining, one has the upper LFT

$$\begin{bmatrix} \vec{b}_1 \\ \vec{b}_2 \\ \vec{d} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 & L_{12} \\ 0 & N_{11} & N_{12} \\ L_{21} & N_{21} & L_{22} + N_{22} \end{bmatrix} \begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vec{c} \end{bmatrix}$$

$$\begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \end{bmatrix} = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix} \begin{bmatrix} \vec{b}_1 \\ \vec{b}_2 \end{bmatrix}$$
(1.312)

where for both LFTs, the  $H_1$  and  $H_2$  matrices form a “structured” upper LFT, i.e. zero matrices exist on the off-diagonal block elements.

As an example of a feedback interconnection, consider the upper LFT

$$\begin{bmatrix} \vec{b} \\ \vec{d} \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} \vec{a} \\ \vec{c} \end{bmatrix}$$

$$\vec{a} = H \vec{b}$$
(1.313)

with the negative feedback interconnection for  $\vec{c}$  as

$$\vec{c} = \vec{r} - \vec{d}$$
(1.314)

where  $\vec{r}$  is some reference command input to the feedback loop. From the second equation,  $\vec{d} = G_{21} \vec{a} + G_{22} \vec{c}$ , assuming  $I_{n_d} + G_{22}$  is invertible, one has by substituting for  $\vec{d}$

$$\vec{c} = (I_{n_d} + G_{22})^{-1} (\vec{r} - G_{21} \vec{a})$$
(1.315)

which can be substituted for  $\vec{c}$  to obtain the upper LFT

$$\begin{bmatrix} \vec{b} \\ \vec{d} \end{bmatrix} = \begin{bmatrix} G_{11} - G_{12}(I_{n_d} + G_{22})^{-1}G_{21} & G_{12}(I_{n_d} + G_{22})^{-1} \\ (I_{n_d} + G_{22})^{-1}G_{21} & G_{22}(I_{n_d} + G_{22})^{-1} \end{bmatrix} \begin{bmatrix} \vec{a} \\ \vec{r} \end{bmatrix}$$

$$\vec{a} = H\vec{b}$$
(1.316)

which is well-posed if  $I_{n_b} - (G_{11} - G_{12}(I_{n_d} + G_{22})^{-1}G_{21})H$  is invertible.

Consider the lower LFT

$$\begin{bmatrix} \vec{b} \\ \vec{d} \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} \vec{a} \\ \vec{c} \end{bmatrix}$$

$$\vec{c} = H\vec{d}$$
(1.317)

The **LFT inverse** can be obtained if  $G_{11}$  is invertible via the rewritten first equation

$$\vec{a} = G_{11}^{-1}\vec{b} - G_{11}^{-1}G_{12}\vec{c}$$
(1.318)

substituted into the second equation to obtain

$$\begin{bmatrix} \vec{a} \\ \vec{d} \end{bmatrix} = \begin{bmatrix} G_{11}^{-1} & -G_{11}^{-1}G_{12} \\ G_{21}G_{11}^{-1} & G_{22} - G_{21}G_{11}^{-1}G_{12} \end{bmatrix} \begin{bmatrix} \vec{b} \\ \vec{c} \end{bmatrix}$$

$$\vec{c} = H\vec{d}$$
(1.319)

which is well-posed if  $I_{n_d} - (G_{22} - G_{21}G_{11}^{-1}G_{12})H$  is invertible.

Likewise, consider the upper LFT

$$\begin{bmatrix} \vec{b} \\ \vec{d} \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} \vec{a} \\ \vec{c} \end{bmatrix}$$

$$\vec{a} = H\vec{b}$$
(1.320)

The **LFT inverse** can be obtained if  $G_{22}$  is invertible via the rewritten second equation

$$\vec{c} = -G_{22}^{-1}G_{21}\vec{a} + G_{22}^{-1}\vec{d}$$
(1.321)

substituted into the first equation to obtain

$$\begin{bmatrix} \vec{b} \\ \vec{c} \end{bmatrix} = \begin{bmatrix} G_{11} - G_{12}G_{22}^{-1}G_{21} & G_{12}G_{22}^{-1} \\ -G_{22}^{-1}G_{21} & G_{22}^{-1} \end{bmatrix} \begin{bmatrix} \vec{a} \\ \vec{d} \end{bmatrix}$$

$$\vec{a} = H\vec{b}$$
(1.322)

which is well-posed if  $I_{n_b} - (G_{11} - G_{12}G_{22}^{-1}G_{21})H$  is invertible.

## References

For more information, please refer to the following

- Skogestad, S., and Postlethwaite, I., “5 Limitations on Performance in SISO Systems,” *Multivariable Feedback Control: Analysis and Design*, 1st ed., Vol. 1, John Wiley & Sons, Chichester, England, 1996, pp. 159-212

## 1.6 Discrete-Time Linear Systems

A multivariate extension of difference equations or recursive relations for discrete-time MIMO systems is the **discrete-time state-space model** which can be defined as the general form of two equations

$$\begin{aligned}\vec{x}[k+1] &= f(k, \vec{x}[k], \vec{u}[k]) \\ \vec{y} &= h(k, \vec{x}, \vec{u})\end{aligned}\tag{1.323}$$

where  $k \in \mathbb{Z}$  is the integer-valued **time step**,  $f : \mathbb{Z} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_x}$ ,  $h : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \rightarrow \mathbb{R}^{n_y}$ ,  $\vec{u} \in \mathbb{R}^{n_u}$  is the **input vector** of  $n_u$  input signals,  $\vec{y} \in \mathbb{R}^{n_y}$  is the **output vector** of  $n_y$  output signals, and  $\vec{x} \in \mathbb{R}^{n_x}$  is the **state vector** of the  $n_x^{\text{th}}$ -order dynamical system. Just as for continuous-time systems, the first vector-valued difference equation is the **discrete-time dynamics equation**, also known as the **discrete-time state equation**, and the second vector-valued algebraic equation is the **output equation**. Contrary to continuous-time models, the recursive nature of discrete-time allows one to easily solve for  $\vec{x}[k]$  given any  $\vec{x}[0]$ .

The **discrete-time time-invariant (TI) state-space models** have the general form

$$\begin{aligned}\vec{x}[k+1] &= f(\vec{x}[k], \vec{u}[k]) \\ \vec{y}[k] &= h(\vec{x}[k], \vec{u}[k])\end{aligned}\tag{1.324}$$

The **discrete-time linear state-space models** have the general form

$$\begin{aligned}\vec{x}[k+1] &= F[k]\vec{x}[k] + G[k]\vec{u}[k] \\ \vec{y}[k] &= H[k]\vec{x}[k] + D[k]\vec{u}[k]\end{aligned}\tag{1.325}$$

where  $A \in \mathbb{R}^{n_x \times n_x}$  matrix is the **discrete-time state matrix**. The  $G \in \mathbb{R}^{n_x \times n_u}$  matrix is the **discrete-time input matrix**. The  $H \in \mathbb{R}^{n_y \times n_x}$  matrix is the **discrete-time output matrix** and  $D \in \mathbb{R}^{n_y \times n_u}$  is the **discrete-time feedthrough matrix**.

The **discrete-time linear time-invariant (LTI) state-space models** have the general form

$$\begin{aligned}\vec{x}[k+1] &= F\vec{x}[k] + G\vec{u}[k] \\ \vec{y}[k] &= H\vec{x}[k] + D\vec{u}[k]\end{aligned}\tag{1.326}$$

### Discrete-Time Trim and Linearization

For discrete-time state-space time-invariant systems similar to continuous-time time-invariant systems, a state-input pair  $(\vec{x}, \vec{u})$  is an **equilibrium point** if  $\vec{x}_k = \vec{x}$  is zero for all  $k \geq 0$ , i.e. if

$$f(k, \vec{x}, \vec{u}) = \vec{x}\tag{1.327}$$

is a valid solution for all  $k \geq 0$ . Similarly, if one initializes  $\vec{x}[k] = \bar{\vec{x}}$  at  $k = 0$  and sets  $\vec{u}[k] = \bar{\vec{u}}$  for  $k \geq 0$ , then  $\vec{x}[k] = \bar{\vec{x}}$  and  $\vec{y} = h(\bar{\vec{x}}, \bar{\vec{u}})$  for all  $k \geq 0$ , i.e.,  $\vec{x}$  is “steady.”

Defining the state, input, and output **perturbation vectors** about constants  $\bar{\vec{x}}$ ,  $\bar{\vec{u}}$ , and  $\bar{\vec{y}}$  as

$$\Delta\vec{x}[k] = \vec{x}[k] - \bar{\vec{x}}\tag{1.328}$$

$$\Delta \vec{u}[k] = \vec{u}[k] - \bar{\vec{u}} \quad (1.329)$$

$$\Delta \vec{y}[k] = \vec{y}[k] - \bar{\vec{y}} \quad (1.330)$$

and recognizing for trim,  $f(\bar{\vec{x}}, \bar{\vec{u}}) = \bar{\vec{x}}$  and  $h(\bar{\vec{x}}, \bar{\vec{u}}) = \bar{\vec{y}}$ , one has

$$\Delta \vec{x}[k+1] = \vec{x}[k+1] - \bar{\vec{x}} = f(\vec{x}, \vec{u}) - f(\bar{\vec{x}}, \bar{\vec{u}}) = \left[ \frac{\partial f}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] \Delta \vec{x}[k] + \left[ \frac{\partial f}{\partial \vec{u}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] \Delta \vec{u}[k] + \text{HOT} \quad (1.331)$$

$$\Delta \vec{y}[k] = \vec{y}[k] - \bar{\vec{y}} = h(\vec{x}, \vec{u}) - h(\bar{\vec{x}}, \bar{\vec{u}}) = \left[ \frac{\partial h}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] \Delta \vec{x}[k] + \left[ \frac{\partial h}{\partial \vec{u}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] \Delta \vec{u}[k] + \text{HOT} \quad (1.332)$$

by the Taylor series expansion.

Next, setting

$$F = \left[ \frac{\partial f}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial f_1}{\partial x_{n_x}}(\bar{\vec{x}}, \bar{\vec{u}}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{n_x}}{\partial x_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial f_{n_x}}{\partial x_{n_x}}(\bar{\vec{x}}, \bar{\vec{u}}) \end{bmatrix} \quad (1.333)$$

$$G = \left[ \frac{\partial f}{\partial \vec{u}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] = \begin{bmatrix} \frac{\partial f_1}{\partial u_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial f_1}{\partial u_{n_u}}(\bar{\vec{x}}, \bar{\vec{u}}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{n_u}}{\partial u_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial f_{n_u}}{\partial u_{n_u}}(\bar{\vec{x}}, \bar{\vec{u}}) \end{bmatrix} \quad (1.334)$$

$$H = \left[ \frac{\partial h}{\partial \vec{x}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] = \begin{bmatrix} \frac{\partial h_1}{\partial x_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial h_1}{\partial x_{n_x}}(\bar{\vec{x}}, \bar{\vec{u}}) \\ \vdots & \ddots & \vdots \\ \frac{\partial h_{n_y}}{\partial x_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial h_{n_y}}{\partial x_{n_x}}(\bar{\vec{x}}, \bar{\vec{u}}) \end{bmatrix} \quad (1.335)$$

$$D = \left[ \frac{\partial h}{\partial \vec{u}}(\bar{\vec{x}}, \bar{\vec{u}}) \right] = \begin{bmatrix} \frac{\partial h_1}{\partial u_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial h_1}{\partial u_{n_u}}(\bar{\vec{x}}, \bar{\vec{u}}) \\ \vdots & \ddots & \vdots \\ \frac{\partial h_{n_y}}{\partial u_1}(\bar{\vec{x}}, \bar{\vec{u}}) & \cdots & \frac{\partial h_{n_y}}{\partial u_{n_u}}(\bar{\vec{x}}, \bar{\vec{u}}) \end{bmatrix} \quad (1.336)$$

yields an approximate **linearized discrete-time LTI state-space model** about  $\bar{\vec{x}}$  and  $\bar{\vec{u}}$  as

$$\begin{aligned} \Delta \vec{x}[k+1] &\approx F \Delta \vec{x}[k] + G \Delta \vec{u}[k] \\ \Delta \vec{y}[k] &\approx H \Delta \vec{x}[k] + D \Delta \vec{u}[k] \end{aligned} \quad (1.337)$$

## Difference Equations and $z$ -Transform

Analogous to continuous-time differential equations for LTI SISO systems, discrete-time LTI SISO systems can use **difference equations** whose standard LTI form is

$$\begin{aligned} y[k] + a_{m-1}y[k-1] + \cdots + a_1y[k-m+1] + a_0y[k-m] \\ = b_p u[k] + b_{p-1}u[k-1] + \cdots + b_1u[k-p+1] + b_0u[k-p] \end{aligned} \quad (1.338)$$

Similarly, the Laplace transform for representing ODEs as continuous-time transfer functions is analogous to the **z-transform** for representing difference equations as a transfer function. For a discrete-time signal  $x[k]$ , the z-transform is given by

$$x(z) = \mathcal{Z}\{x[k]\} = \sum_{k=-\infty}^{\infty} x[k]z^{-k} \quad (1.339)$$

where  $z$  is a complex number. This also has an inverse z-transform

$$x[k] = \mathcal{Z}^{-1}\{X(z)\} = \frac{1}{2\pi j} \oint_C x(z)z^{k-1} dz \quad (1.340)$$

where  $C$  is a counterclockwise closed path encircling the origin and entirely in the **region of convergence (ROC)**, i.e. the set of points in the complex plane for which the z-transform summation converges. Thus, for a  $m^{\text{th}}$  order LTI difference equation, the **discrete-time transfer functions** is

$$H(z) = \frac{y(z)}{u(z)} = \frac{b_p z^p + b_{p-1} z^{p-1} + \cdots + b_1 z + b_0}{z^m + a_{m-1} z^{m-1} + \cdots + a_1 z + a_0} \quad (1.341)$$

For reference, some functions in the  $k$  and  $z$  domains are provided in the following table.

Function	$k$ Domain ( $\forall k \geq 0$ )	$z$ Domain
Unit Step	1	$\frac{z}{z-1}$
Power	$a^k$	$\frac{z}{z-a}$
Sine	$\sin(\omega k)$	$\frac{z \sin \omega}{z^2 - 2z \cos \omega + 1}$
Cosine	$\cos(\omega k)$	$\frac{z^2 - z \cos \omega}{z^2 - 2z \cos \omega + 1}$
Power Sine	$a^k \sin(\omega k)$	$\frac{z^2 - z \cos \omega}{z^2 - 2az \cos \omega + a^2}$
Power Cosine	$a^k \cos(\omega k)$	$\frac{az \sin \omega}{z^2 - 2az \cos \omega + a^2}$

One way to describe the similarity between the  $s$  and  $z$  transforms is to see that as the  $s$  variable operates in multiplication and division as a differentiator and integrator in continuous-time, respectively, the  $z$  variable operates in multiplication and division as a backward and forward **time shift operator** in discrete-time, respectively.

For the frequency response of a discrete-time system, one substitutes  $z = e^{j\omega}$ . Thus, the **discrete-time Fourier transform (DTFT)** with periodicity of  $2\pi$  can be written as

$$H(\omega) = \sum_{k=-\infty}^{\infty} h[k]e^{-j\omega k} \quad (1.342)$$

and the inverse discrete-time Fourier transform

$$H[k] = \frac{1}{2\pi j} \int_{-\pi}^{\pi} h(e^{j\omega})e^{j\omega k} d\omega \quad (1.343)$$

where the substitution  $z = e^{j\omega}$  has set the closed path  $C$  to the unit circle. It should be noted that the **discrete Fourier transform (DFT)** discretely samples the continuous DTFT, i.e.

$$H[n] = \frac{1}{N} \sum_{k=0}^{N-1} h[k] e^{-j \frac{2\pi n}{N} k} \quad (1.344)$$

where  $N$  is the number of evenly-spaced samples at some **sampling interval** is  $\Delta t$ . The **total sampling time** is  $T_s = N\Delta t$ , the **sampling frequency** in Hz is  $f_s = \frac{1}{\Delta t} = \frac{N}{T_s}$ , and the **frequency resolution** is  $\Delta f = \frac{f_s}{N}$ . Lastly, the **Nyquist frequency**,  $f_N$ , defined as

$$f_N = \frac{f_s}{2} = \frac{1}{2\Delta t} = \frac{N}{2T_s} \quad (1.345)$$

is the highest frequency component that can be resolved. This property of the Nyquist frequency leads us to **Shannons sampling theorem** which states that a continuous signal must be discretely sampled at least twice the frequency of the highest frequency in the signal. In this way, one can use the z-transform and apply SISO loop-shaping methods based on discrete-time frequency response requirements. One of the most important algorithms of the twentieth century is the **fast Fourier transform (FFT)** which efficiently computes the DFT.

### General Solution of Discrete-Time LTI Systems

Consider the state equation for the discrete-time linear state-space model

$$\vec{x}[k+1] = F_k \vec{x}[k] + G_k \vec{u}[k] \quad (1.346)$$

with a given initial state,  $\vec{x}[0] = 0$ . Then, iterating through this equation

$$\begin{aligned} \vec{x}[1] &= F_0 \vec{x}[0] + G_0 \vec{u}[0] = G_0 \vec{u}[0] + F_0 \vec{x}[0] \\ \vec{x}[2] &= F_1 \vec{x}[1] + G_1 \vec{u}[1] = G_1 \vec{u}[1] + F_1 G_0 \vec{u}[0] + F_1 F_0 \vec{x}[0] \\ \vec{x}[3] &= F_2 \vec{x}[2] + G_2 \vec{u}[2] = G_2 \vec{u}[2] + F_2 G_1 \vec{u}[1] + F_2 F_1 G_0 \vec{u}[0] + F_2 F_1 F_0 \vec{x}[0] \\ &\vdots \\ \vec{x}[k] &= G_{k-1} \vec{u}[k-1] + F_{k-1} G_{k-2} \vec{u}[k-2] + \dots + F_{k-1} \cdots F_1 G_0 \vec{u}[0] + F_{k-1} \cdots F_0 \vec{x}[0] \end{aligned} \quad (1.347)$$

This general solution can be rewritten as product of two matrices:

$$\vec{x}[k] = F_{k-1} \cdots F_0 \vec{x}[0] + \begin{bmatrix} G_{k-1} & F_{k-1} G_{k-2} & \cdots & F_{k-1} \cdots F_1 G_0 \end{bmatrix} \begin{bmatrix} \vec{u}[k-1] \\ \vec{u}[k-2] \\ \vdots \\ \vec{u}[0] \end{bmatrix} \quad (1.348)$$

which has the form

$$\vec{x}[k] = \Phi(0, k) \vec{x}[0] + C \vec{u} \quad (1.349)$$

where  $\Phi(0, k)$  is the **discrete-time state-transition matrix** and  $C$  is the discrete-time **state controllability matrix** and has the same structure as continuous-time. Note that for a discrete-time LTI system, one has

$$\Phi(0, k) = F^k \quad (1.350)$$

and

$$C = [G \quad FG \quad \cdots \quad F^k G] \quad (1.351)$$

Similar to the Laplace transform, one can use the Z-transform to obtain the discrete frequency solution from the LTI state equation as

$$z\vec{x}(z) - z\vec{x}[0] = F\vec{x}(z) + G\vec{u}(z) \quad (1.352)$$

or

$$(zI - F)\vec{x}(z) = z\vec{x}[0] + G\vec{u}(z) \quad (1.353)$$

which yields

$$\vec{x}(z) = z(zI - F)^{-1}\vec{x}[0] + (zI - F)^{-1}G\vec{u}(z) \quad (1.354)$$

With the output equation

$$\vec{y}(z) = H\vec{x}(z) + D\vec{u}(z) \quad (1.355)$$

or

$$\vec{y}(z) = Hz(zI - F)^{-1}\vec{x}[0] + \left(H(zI - F)^{-1}G + D\right)\vec{u}(z) \quad (1.356)$$

### Discrete-Time State-Space System Stability

For an initial condition at  $\vec{x}[0]$  and free response dynamics (i.e. no control), the discrete-time state-space can be reduced to

$$\vec{x}[k] = F\vec{x}[k-1] \quad (1.357)$$

which has the simple solution for the state at time step  $k$  as

$$\vec{x}[k] = F^k\vec{x}[0] \quad (1.358)$$

Next, define the **modal state vector**,  $\vec{z}$ , such that

$$\vec{x} = V\vec{z} = \vec{v}_1 z_1(t) + \cdots + \vec{v}_{n_x} z_{n_x}(t) \quad (1.359)$$

where  $V$  is the matrix of  $n$  eigenvectors of  $F$  and

$$\vec{z} = V^{-1}\vec{x} = \begin{bmatrix} \vec{\mu}_1^T \vec{x} \\ \vdots \\ \vec{\mu}_{n_x}^T \vec{x} \end{bmatrix} \quad (1.360)$$

where notably,  $z_{n_x}(t)$ , is dimensionless. Then, substituting into the state equation, one has

$$V\vec{z}[k+1] = FV\vec{z}[k] + G\vec{u}(t) \quad (1.361)$$

or

$$\vec{z}[k+1] = V^{-1}FV\vec{z}[k] + V^{-1}G\vec{u}[k] \quad (1.362)$$

Thus, one can form the discrete-time **Jordan canonical form (JCF)** of the state-space system as

$$\begin{aligned}\vec{z}[k+1] &= \Lambda\vec{z}[k] + V^{-1}G\vec{u}[k] \\ \vec{y}[k] &= GV\vec{z}[k] + D\vec{u}[k]\end{aligned} \quad (1.363)$$

where the state-transition matrix for diagonalizable  $F$  is now

$$\Lambda^k = \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_{n_x}^k \end{bmatrix} \quad (1.364)$$

and this model explicitly **decouples** the state equation into  $n_x$  LTI ODEs, i.e.

$$z_i[k] = \lambda_i^k z_i(t) + \vec{\mu}_i^T G \vec{u}(t), \quad i = 1, \dots, n_x \quad (1.365)$$

Consider the zero-input response where the  $n_x$  modal states are given by

$$z_i[k] = \lambda_i^k z_i[0] = \lambda_i^k \vec{\mu}_i^T \vec{x}[0] \quad (1.366)$$

and the output response is given by

$$\vec{y}[k] = H(\vec{v}_1 z_1[k] + \cdots + \vec{v}_{n_x} z_{n_x}[k]) \quad (1.367)$$

Thus, one can see the modal states are characterized by the eigenvalues and left eigenvectors while the output response is additionally characterized by the right eigenvectors and the modal states. As these eigenvalues will occur as real numbers or in complex-conjugate pairs, the modal states may be complex-valued which describes the oscillatory nature of the modal responses. Regardless, the magnitude of each individual mode allows one to characterize the stability of each mode as well as the discrete-time LTI system as a whole. In particular, a mode is **stable** if  $|\lambda_i| < 0$ , **marginally stable** if  $|\lambda_i| = 0$ , and **unstable** if  $|\lambda_i| > 0$ . Furthermore, the LTI system is **stable** if *all* the modes are strictly stable, the LTI system is **marginally stable** if *all* the modes are stable or marginally stable, and if *any* mode is unstable, the LTI system is **unstable**. As these eigenvalues may be complex-valued, one can also state that all the system poles located in the left half of the complex plane, i.e. the **left half plane (LHP)**, correspond to stable modes.

## Discretization Techniques

Often one wishes to discretize a continuous-time LTI system, i.e. convert it to a discrete-time LTI system. To discretize general MIMO systems, recall the continuous-time state equation

$$\dot{\vec{x}}(t) = A\vec{x}(t) + B\vec{u}(t) \quad (1.368)$$

which has the general solution

$$\vec{x}(t) = e^{At} \vec{x}(0) + \int_0^t e^{A(t-\tau)} B \vec{u}(\tau) d\tau \quad (1.369)$$

Thus, for every time step from  $k - 1$  to  $k$  and  $t = \Delta t$ , one can write

$$\vec{x}[k] = e^{A\Delta t} \vec{x}[k-1] + \int_0^{\Delta t} e^{A(\Delta t-\tau)} B \vec{u}(\tau) d\tau \quad (1.370)$$

which infers that

$$F = e^{A\Delta t} \quad (1.371)$$

Note that this explains the connection between the LTI stability conditions on  $F$  and the corresponding LTI stability condition on  $A$ , i.e. magnitude of eigenvalues of  $F$  are less than one and the real part of the eigenvalues of  $A$  are negative.

However, for the input matrix, one must assume that  $\vec{u}(\tau)$  follows some type of piecewise function between times steps. To only use the previous time step to approximate the integral, one may use a **zero-order hold (ZOH)** which assumes  $\vec{u}(\tau)$  is a piecewise constant input  $\vec{u}[k-1]$  from  $k-1$  to  $k$ , then

$$\int_0^{\Delta t} e^{A(\Delta t-\tau)} B \vec{u}(\tau) d\tau = \left( \int_0^{\Delta t} e^{A(\Delta t-\tau)} B d\tau \right) \vec{u}[k-1] \quad (1.372)$$

and one has for the discrete-time input matrix

$$G = \int_0^{\Delta t} e^{A(\Delta t-\tau)} B(\tau) d\tau \quad (1.373)$$

Other alternatives include higher-order holds, including a **first-order hold (FOH)** which assumes  $\vec{u}(\tau)$  is a piecewise linear input from  $k-1$  to  $k$ . However, other holds require knowledge of the inputs at additional time steps, e.g. FOH requires  $u[k]$  and  $u[k-1]$ . Lastly, note that  $H = C$  in this discretization.

Analogous to the ZOH and FOH, one can also approximate the transfer functions of SISO systems, using the **Euler Forward approximation**, i.e.

$$s \approx \frac{z - 1}{\Delta t} \quad (1.374)$$

or the **Tustin approximation**, i.e.

$$s \approx \frac{2(z - 1)}{\Delta t(z + 1)} \quad (1.375)$$

and another common method is the **Euler Backward approximation**, i.e.

$$s \approx \frac{z - 1}{z\Delta t} \quad (1.376)$$

which is numerically more stable than the forward version due to the backward requiring smaller sampling time  $\Delta t$ . Even so, using these latter two approximation methods to convert controllers typically requires that the sampling time satisfy

$$\Delta t < \frac{\pi}{5\omega_c} \quad (1.377)$$

in order for the controller to *likely* produce satisfactory closed-loop behavior where  $\omega_c$  is the crossover frequency of the open-loop transfer function.

## References

For more information, please refer to the following

- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “7 Digital Control,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 584-620

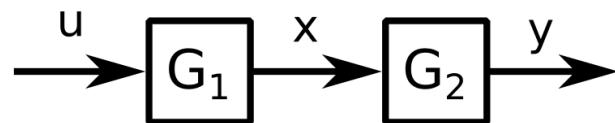
---

# Continuous-Time LTI Feedback Control Theory

## 2.1 Introduction to Continuous-Time Control Systems

### Interconnections of SISO LTI Systems

For model-based design of SISO LTI control systems, one must understand some of the block diagram algebra for interconnected SISO LTI systems. For building an overall system transfer function, it is useful to consider the input and output to various interconnected systems represented by transfer functions, in particular the serial/cascade, parallel, and feedback interconnections. A **serial interconnection**, also known as a **cascade interconnection** can be represented as



and in the Laplace domain

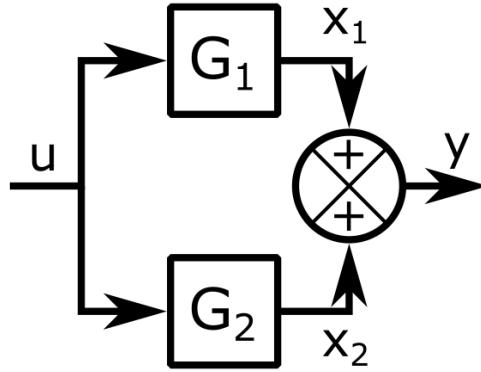
$$y(s) = G_2(s)x(s) \quad (2.1)$$

$$x(s) = G_1(s)u(s) \quad (2.2)$$

or for an overall system transfer function

$$y(s) = G(s)u(s) = [G_2(s)G_1(s)]u(s) \quad (2.3)$$

A **parallel interconnection** can be represented as



and in the Laplace domain

$$x_1(s) = G_1(s)u(s) \quad (2.4)$$

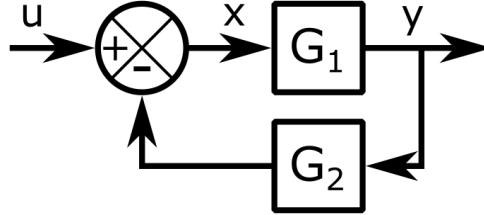
$$x_2(s) = G_2(s)u(s) \quad (2.5)$$

$$y(s) = x_1(s) + x_2(s) \quad (2.6)$$

or for an overall system transfer function

$$y(s) = G(s)u(s) = [G_1(s) + G_2(s)]u(s) \quad (2.7)$$

A (negative) **feedback interconnection** can be represented as



and in the Laplace domain

$$y(s) = G_1(s)x(s) \quad (2.8)$$

$$x(s) = u(s) - G_2(s)y(s) \quad (2.9)$$

$$y(s) = G_1(s)(u(s) - G_2(s)y(s)) \quad (2.10)$$

$$(1 + G_1(s)G_2(s))y(s) = G_1(s)u(s) \quad (2.11)$$

or for an overall system transfer function

$$y(s) = G(s)u(s) = \left[ \frac{G_1(s)}{1 + G_2(s)G_1(s)} \right] u(s) \quad (2.12)$$

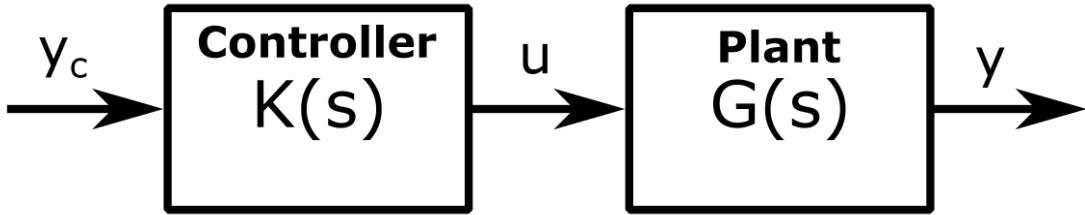
Lastly, when forming transfer functions for interconnected systems, it is important to take note of any pole-zero cancellations as this may result in an unstable feedback system due to the dynamics of interior signals of the system.

### Open- and Closed-Loop Control Laws

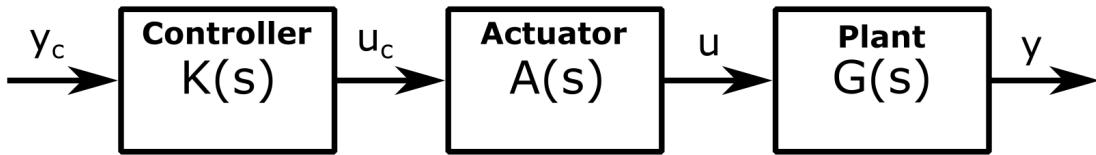
The simplest types of control laws are **open-loop control** strategies, also known as **feedforward control**. In general, this logically flows as the following

1. the user specifies a commanded output  $y_c$
2. the control law sets the input  $u$  as a function of that commanded output
3. this dynamically changes the output  $y(t)$

For the open-loop control of SISO LTI systems, one would have a simple system block diagram as



For real systems,  $u$  will undergo some dynamics itself, i.e. the plant input won't instantaneously reach the electric controller's computed  $u$ , but is actuated using some electro-mechanical interface which one can represent by an actuation system or **actuator**. This additional consideration would result in the following block diagram for SISO LTI systems.



Using dynamical systems theory, one can typically model actuation systems for vehicles,  $A(s)$ , as first- or second-order LTI systems, e.g.

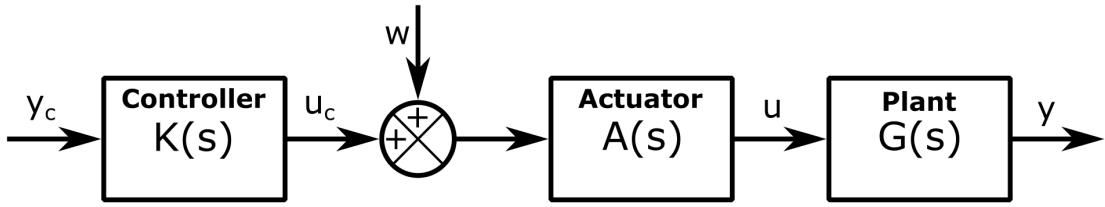
$$A(s) = \frac{\omega_a}{s + \omega_a} \quad (2.13)$$

or

$$A(s) = \frac{\omega_a^2}{s^2 + 2\zeta_a\omega_a s + \omega_a^2} \quad (2.14)$$

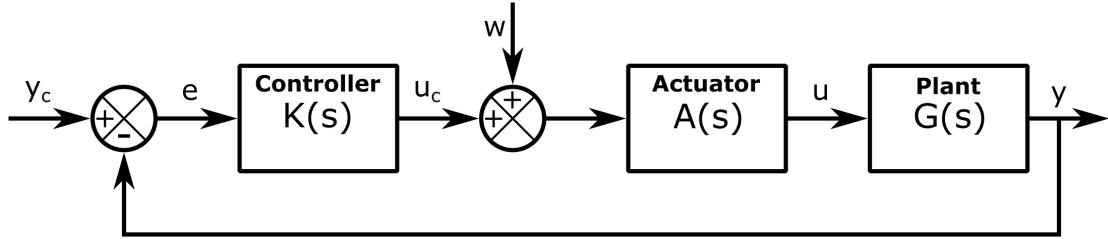
where  $\omega_a$  is the bandwidth of the actuator in rad/s and  $\zeta_a$  is the damping. Actuators also typically have hard limits on the minimum and maximum output as well as hard rate limits.

However, one consideration for real systems is the fact that one will have additional disturbances to the plant model. These effects can be modeled as an additive disturbance signal,  $w$ , to the control input,  $u$ , which results in a block diagram for SISO LTI systems as



Thus, if one does not have explicit knowledge of these disturbances, then one may not be able to guarantee the control system satisfies the design requirements using only open-loop control as this strategy does not gather any information about the disturbances as they occur. To reject these disturbances and account for system uncertainties, one must use alternative control strategies.

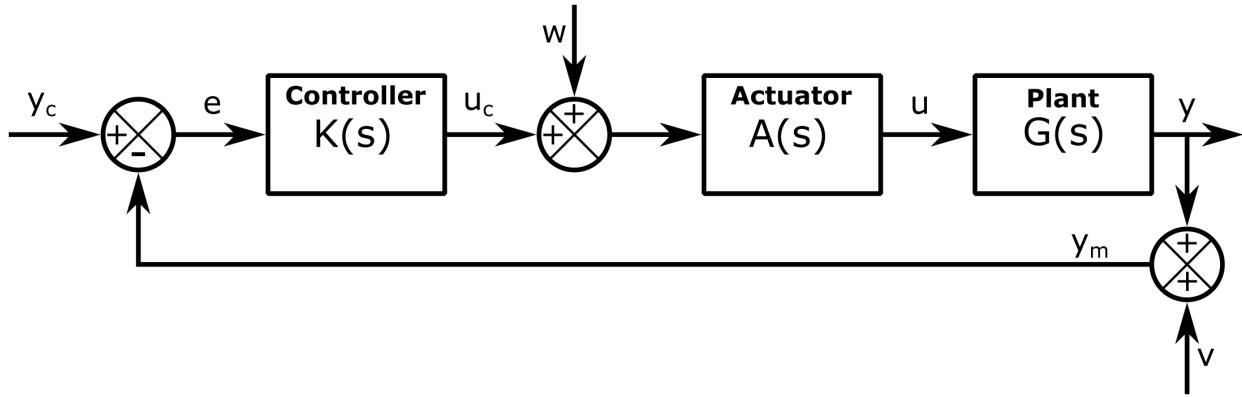
In particular, **closed-loop control** strategies, also known as **feedback control**, feed the plant output signal back to the control law, which results in a block diagram for SISO LTI systems as



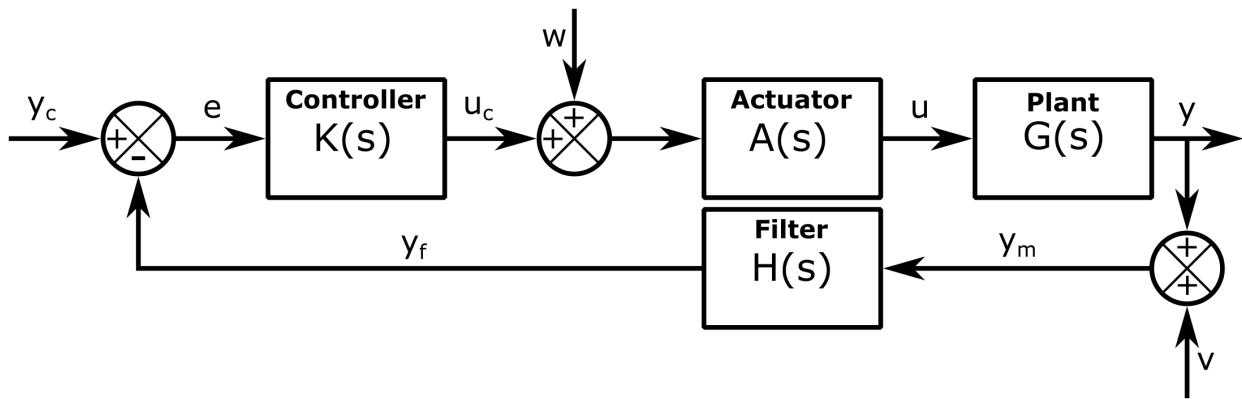
In this way, one can use an output of the plant,  $y(t)$ , to compute the tracking error,  $e(t)$ , which directly affects the control input to the plant,  $u(t)$ , which, in turn, affects the plant output and so on, as time continues. Recalling the negative feedback interconnection, a feedback controller produces an overall system transfer function

$$y(s) = \left[ \frac{G(s)A(s)K(s)}{1 + G(s)A(s)K(s)} \right] y_c(s) \quad (2.15)$$

For real systems, the output signal must be measured by a sensor system, or **sensor**, which measures the output of the plant using physical phenomena that can be converted into a measured signal,  $y_m$ , similar to the input signal being actuated in real systems. Likewise, as no sensor provides perfect measurement, one can consider an additive noise signal,  $v$ , to the output,  $y$ , which results in a block diagram for SISO LTI systems as



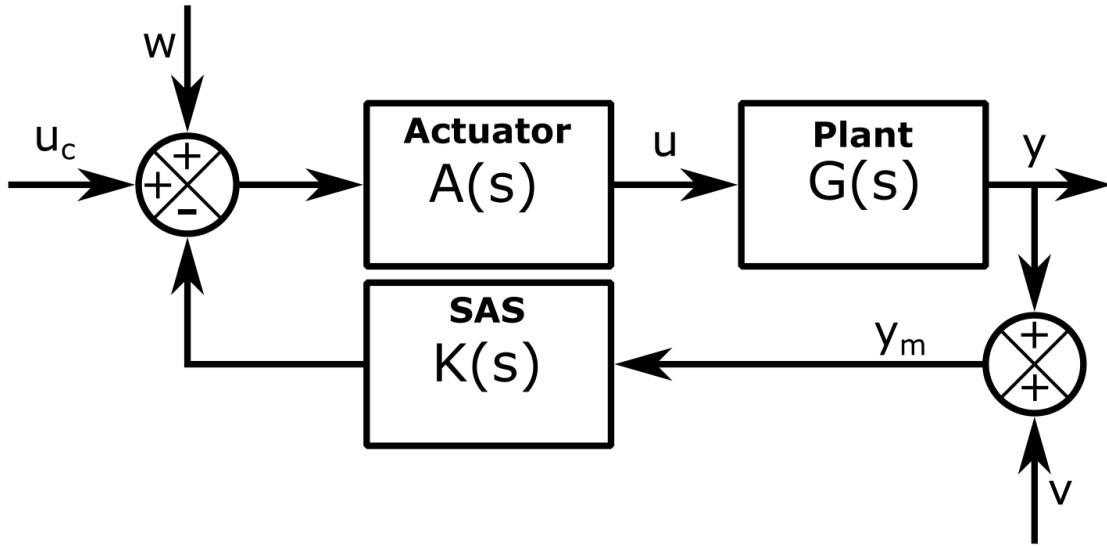
With this in mind, some control systems use a separate **filter** sub-system in the negative feedback loop to counteract the measurement noise,  $v$ , i.e.



Recalling the negative feedback interconnection, the inclusion of a separate filter into the feedback control system produces an overall system transfer function

$$y(s) = \left[ \frac{G(s)A(s)K(s)}{1 + H(s)G(s)A(s)K(s)} \right] y_c(s) \quad (2.16)$$

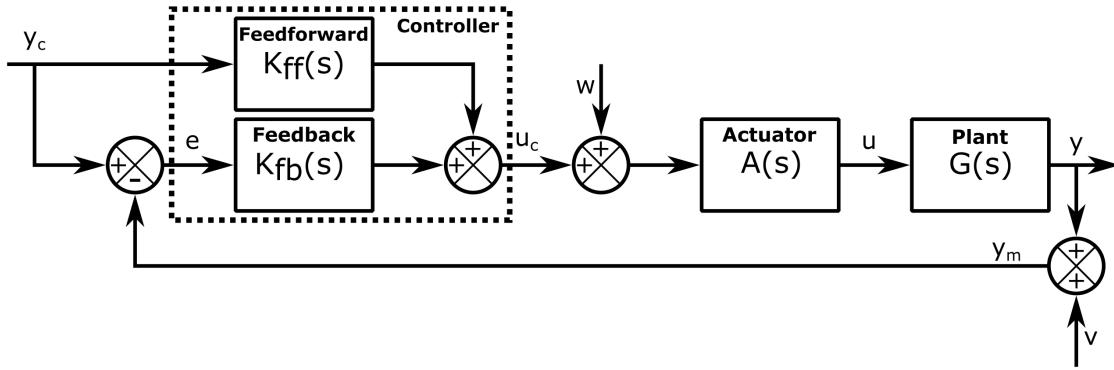
Similar to this is a special type of feedback control system called a **stability augmentation system (SAS)**. In this case, the plant is to be controlled by an external operator, e.g. pilot, but the inherent response characteristics, e.g. modal stability or damping, are not within the design specifications. Thus, a SAS is used to augment the plant dynamics to achieve certain dynamic responses, typically stability or excessive damping. An SAS for SISO LTI systems can be modeled without an actuator and sensor as the following block diagram.



Recalling the negative feedback interconnection, a SAS produces an overall system transfer function

$$y(s) = \left[ \frac{G(s)A(s)}{1 + G(s)A(s)K(s)} \right] u_c(s) \quad (2.17)$$

Lastly, it should be noted that feedforward and feedback control can also be combined which results in a block diagram for SISO LTI systems as



Recalling the negative feedback and parallel interconnections, a feedback and feedforward control system produces the transfer function

$$u_c(s) = \left[ \frac{K_{fb}(s)}{1 + G(s)A(s)K_{fb}(s)} + K_{ff}(s) \right] y_c(s) \quad (2.18)$$

and noting

$$y(s) = G(s)A(s)u_c(s) \quad (2.19)$$

one has an overall system transfer function

$$y(s) = \left[ \frac{G(s)A(s)K_{fb}(s)}{1 + G(s)A(s)K_{fb}(s)} + G(s)A(s)K_{ff}(s) \right] y_c(s) \quad (2.20)$$

## References

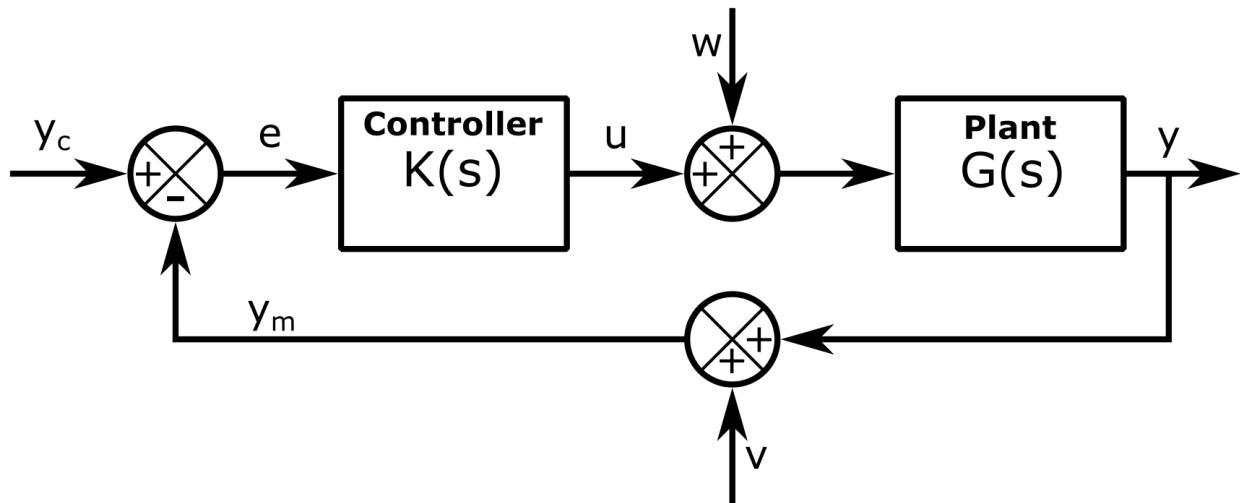
For more information, please refer to the following

- Schmidt, D. K., “12.1 Feedback Control-Law Synthesis Via Loop Shaping - A Just-In-Time Tutorial\*,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 548-562
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “3.9 Feedback Control,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 213-240

## 2.2 Classical Feedback Control System Analysis

### Classical Feedback Control System

The remainder of this chapter will focus on SISO LTI control systems, the study of which is known as **classical control theory**. When designing SISO LTI feedback control systems, one usually considers the following **classical feedback control system**



where  $G(s)$  is the plant (combined with potential actuator),  $K(s)$  is the controller (combined with potential filter),  $y_c$  is the commanded output signal,  $e$  is the tracking error signal,  $u$  is the control input signal,  $w$  is a noise signal to the dynamic system input, e.g. disturbance,  $y$  is the system output signal,  $v$  is a noise signal on the dynamic system output, e.g. sensor noise,  $y_m$  is the measured output signal. For this system, the **open-loop transfer function** is defined as

$$L(s) = G(s)K(s) \quad (2.21)$$

and plays a large role in the analysis of the SISO LTI control system. Note that  $L(s)$  would be the control system transfer function if one used an open-loop control design.

Note that this feedback control system has 3 external input signals:  $y_c, w, v$ , and 4 internal output signals:  $e, u, y, y_m$ . Then, using the previously described rules for building transfer functions, each possible input-output pair has an associated transfer function for a total of twelve and can be represented as a matrix

$$\begin{bmatrix} y(s) \\ y_m(s) \\ e(s) \\ u(s) \end{bmatrix} = \begin{bmatrix} \frac{GK}{1+GK} & \frac{G}{1+GK} & -\frac{GK}{1+GK} \\ \frac{GK}{1+GK} & \frac{G}{1+GK} & \frac{1}{1+GK} \\ \frac{1}{1+GK} & -\frac{G}{1+GK} & -\frac{1}{1+GK} \\ \frac{K}{1+GK} & -\frac{GK}{1+GK} & -\frac{K}{1+GK} \end{bmatrix} \begin{bmatrix} y_c(s) \\ w(s) \\ v(s) \end{bmatrix} \quad (2.22)$$

from which, ignoring the sign, one can see there are four **fundamental transfer functions**, namely

$$\frac{1}{1 + G(s)K(s)}, \quad \frac{G(s)K(s)}{1 + G(s)K(s)}, \quad \frac{G(s)}{1 + G(s)K(s)}, \quad \frac{K(s)}{1 + G(s)K(s)} \quad (2.23)$$

Thus, selecting  $K(s)$  affects these four fundamental transfer functions and the study of these transfer functions is key to classical control systems theory. Two of the four fundamental transfer functions have particular names. The first is the **error transfer function** defined as

$$S(s) = \frac{e(s)}{y_c(s)} = \frac{1}{1 + G(s)K(s)} = \frac{1}{1 + L(s)} \quad (2.24)$$

and is the transfer function from the closed-loop commanded output,  $y_c(s)$ , to the tracking error  $e(s)$ . The second is the **closed-loop transfer function** defined as

$$T(s) = \frac{y(s)}{y_c(s)} = \frac{G(s)K(s)}{1 + G(s)K(s)} = \frac{L(s)}{1 + L(s)} \quad (2.25)$$

and is the transfer function from the closed-loop commanded output,  $y_c(s)$ , to the actual output  $y(s)$ .

$S(s)$  is also known as the **sensitivity transfer function** as it characterizes how sensitive the system is to small changes in  $G(s)$ , an important consideration in real systems due to neglected dynamics, e.g. nonlinearities, or changing characteristics, e.g. component aging. To see this, consider the transfer function from  $y_c \rightarrow y$ , i.e.

$$\frac{Y(s)}{Y_c(s)} = \frac{G(s)K(s)}{1 + G(s)K(s)} = T(s) \quad (2.26)$$

Now consider the ratio of a small change in  $T$  to a small change in  $G$ , i.e.

$$\frac{\frac{\Delta T}{T}}{\frac{\Delta G}{G}} = \left[ \frac{\Delta T}{\Delta G} \right] \left[ \frac{G}{T} \right] \quad (2.27)$$

using the derivative for small  $\Delta$ 's

$$\frac{\frac{\Delta T}{T}}{\frac{\Delta G}{G}} = \left[ \frac{dT}{dG} \right] \left[ \frac{G}{T} \right] \quad (2.28)$$

$$\frac{\frac{\Delta T}{T}}{\frac{\Delta G}{G}} = \left[ \frac{d}{dG} \left( \frac{GK}{1+GK} \right) \right] \left[ \frac{G}{\frac{GK}{1+GK}} \right] \quad (2.29)$$

$$\frac{\frac{\Delta T}{T}}{\frac{\Delta G}{G}} = \left[ \left( \frac{(1+GK)K - GK^2}{(1+GK)^2} \right) \right] \left[ \frac{1+GK}{K} \right] \quad (2.30)$$

$$\frac{\frac{\Delta T}{T}}{\frac{\Delta G}{G}} = \left[ \left( \frac{K}{(1+GK)^2} \right) \right] \left[ \frac{1+GK}{K} \right] \quad (2.31)$$

$$\frac{\frac{\Delta T}{T}}{\frac{\Delta G}{G}} = \frac{1}{(1+GK)} \quad (2.32)$$

$$\frac{\frac{\Delta T}{T}}{\frac{\Delta G}{G}} = S \quad (2.33)$$

Lastly,  $T(s)$  is also known as the **complementary sensitivity transfer function** as one can write that

$$S(s) + T(s) = \frac{1}{1+G(s)K(s)} + \frac{G(s)K(s)}{1+G(s)K(s)} \quad (2.34)$$

or

$$S(s) + T(s) = 1 \quad \forall s \in \mathbb{C} \quad (2.35)$$

### Classical Feedback Control System Stability

When designing a control system, at a minimum, one requires that the system is stable. For the classical feedback control system model, this will only occur if and only if *all* the system transfer functions are stable. However, one can simplify this criteria by considering the numerator/denominator polynomials of  $G(s)$  and  $K(s)$ , i.e.

$$G(s) = \frac{n_G(s)}{d_G(s)} \quad (2.36)$$

$$K(s) = \frac{n_K(s)}{d_K(s)} \quad (2.37)$$

Then, the four fundamental transfer functions can be written

$$\begin{aligned} \frac{1}{1+GK} &= \frac{1}{1 + \frac{n_G n_K}{d_G d_K}} = \frac{d_G d_K}{d_G d_K + n_G n_K} \\ \frac{GK}{1+GK} &= \frac{\frac{n_G n_K}{d_G d_K}}{1 + \frac{n_G n_K}{d_G d_K}} = \frac{n_G n_K}{d_G d_K + n_G n_K} \\ \frac{G}{1+GK} &= \frac{\frac{n_G}{d_G}}{1 + \frac{n_G n_K}{d_G d_K}} = \frac{n_G d_K}{d_G d_K + n_G n_K} \\ \frac{K}{1+GK} &= \frac{\frac{n_K}{d_K}}{1 + \frac{n_G n_K}{d_G d_K}} = \frac{d_G n_K}{d_G d_K + n_G n_K} \end{aligned} \quad (2.38)$$

where notably  $d_G d_K + n_G n_K$  is the denominator for each one, i.e. the **SISO feedback control system characteristic polynomial**. Thus, the **SISO feedback control system characteristic equation** is

$$d_G(s)d_K(s) + n_G(s)n_K(s) = 0 \quad (2.39)$$

and if the roots/poles of this equation are in the left half of the complex plane, the feedback system is stable. This is true even in the case of any pole-zero cancellations which may occur between  $G(s)$  and  $K(s)$ . As a proof, let  $p_0$  be a pole such that a pole/zero cancellation occurs between  $G(s)$  and  $K(s)$ , i.e.

$$d_G(p_0) = n_k(p_0) = 0 \quad (2.40)$$

then recalling the characteristic equation

$$d_G(s)d_K(s) + n_G(s)n_K(s) = 0 \quad (2.41)$$

and substituting  $s = p_0$

$$d_G(p_0)d_K(p_0) + n_G(p_0)n_K(p_0) = 0 \quad (2.42)$$

one has

$$(0)d_K(p_0) + n_G(p_0)(0) = 0 \quad (2.43)$$

which equates to

$$0 = 0 \quad (2.44)$$

In addition, note that  $1 + L = 1 + GK$  appears in the denominator of each fundamental transfer function which can be rewritten as

$$1 + L(s) = 1 + G(s)K(s) = 1 + \frac{n_G n_K}{d_G d_K} = \frac{d_G(s)d_K(s) + n_G(s)n_K(s)}{d_G(s)d_K(s)} \quad (2.45)$$

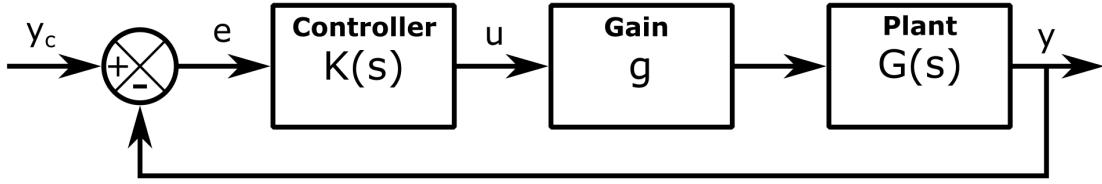
Thus, one can see that the numerator of  $1+L(s)$  is the SISO feedback control system characteristic polynomial. However, as  $1 + L(s)$  can still be affected by pole-zero cancellations, one can equivalently declare that one has a **stable SISO feedback control system** if and only if

1. there are no RHP pole-zero cancellations when forming  $L(s)$ ; and
2.  $1 + L(s)$  has no zeros in the RHP.

The analysis of the zeros of  $L(s)$  as regards stability and robustness will be discussed in the following lectures.

## Gain Margin

The **gain margin** is the amount of gain on the nominal plant model that can be added before closed-loop instability. The gain margin can be derived by considering a scaling perturbation to the plant dynamics model,  $G(s)$ , which can be modeled in block diagram form as



where  $g > 0$  is some scalar constant. Then, defining the perturbed open-loop transfer function as  $gL(s)$ , the gain margin is determined by assessing for what values of  $g$  the feedback control system will go unstable.

Recalling that the feedback control system is stable if (1) there are no pole-zero cancellations and (2) the zeros of  $1 + L(s) = 0$  are only in the LHP. Then, a critical gains,  $g_0$ , occurs when the zeros of  $1 + g_0L(j\omega_0)$  are on the imaginary axis for some critical frequency  $\omega_0$ , i.e.

$$1 + g_0L(j\omega_0) = 0 \quad (2.46)$$

Rewriting this expression, one has

$$L(j\omega_0) = -\frac{1}{g_0} \quad (2.47)$$

or in polar form (i.e. gain and phase), one has

$$|L(j\omega_0)|e^{j\angle L(j\omega_0)} = \frac{1}{g_0}e^{\pm j\pi} \quad (2.48)$$

or the equivalent magnitude/gain and phase requirements become

$$g_0 = \frac{1}{|L(j\omega_0)|} \quad (2.49)$$

and

$$\angle L(j\omega_0) = \pm 180^\circ \quad (2.50)$$

Thus, one can identify the gain margins by:

1. identifying all critical frequencies  $\omega_{0,i}$  where  $\angle L(j\omega_{0,i}) = \pm 180^\circ$ ;
2. calculating all associated “candidate” gain margins,  $g_{0,i}$ , using the inverse of the gain at  $\omega_{0,i}$ , i.e.

$$g_{0,i} = \frac{1}{|L(j\omega_{0,i})|} \quad (2.51)$$

3. setting the **upper gain margin**,  $\bar{g}$ , as  $\min_i g_{0,i} > 1$ ; and
4. setting the **lower gain margin**,  $\underline{g}$ , as  $\max_i g_{0,i} < 1$ .

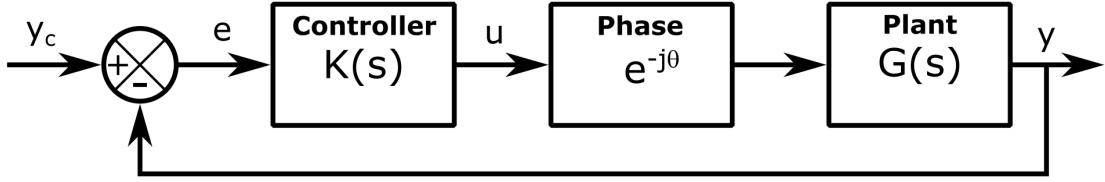
which can be done simply from the Bode plot of  $L(j\omega)$ . Then, one may state that the feedback control system is stable for

$$\underline{g} \leq g \leq \bar{g} \quad (2.52)$$

where it should be noted that not all systems will have an upper *and* lower gain margin. For FDC, a feedback control system is generally considered sufficiently robust if  $\underline{g} \leq 0.5$  and  $\bar{g} \geq 2$ , i.e.  $\underline{g} \leq -6$  dB and  $\bar{g} \geq 6$  dB, for the gain margin.

## Phase and Delay Margin

The **phase margin** is the amount of phase lead/lag that can be added to the input before closed-loop instability. The phase margin can be derived by considering a phase perturbation to the plant dynamics model which can be modeled in block diagram form as



Then, defining the perturbed open-loop transfer function as  $e^{-j\theta}L(s)$  the phase margin is determined by assessing for what values of  $\theta$  the feedback control system will go unstable.

Recalling that the feedback control system is stable if (1) there are no pole-zero cancellations and (2) the zeros of  $1 + L(s) = 0$  are only in the LHP. Then, a critical phase,  $\theta_0$ , occurs when the zeros of  $1 + e^{-j\theta}L(j\omega_0)$  are on the imaginary axis for some critical frequency  $\omega_0$ , i.e.

$$1 + e^{-j\theta_0}L(j\omega_0) = 0 \quad (2.53)$$

Rewriting this expression, one has

$$L(j\omega_0) = -e^{j\theta_0} \quad (2.54)$$

or in polar form (i.e. gain and phase), one has

$$|L(\omega_0)|e^{j\angle L(\omega_0)} = e^{\pm j\pi}e^{j\theta_0} = e^{j(\pm\pi+\theta_0)} \quad (2.55)$$

or the equivalent magnitude/gain and phase requirements become

$$|L(j\omega_0)| = 1 \quad (2.56)$$

and

$$\theta_0 = \pm 180^\circ + \angle L(j\omega_0) \quad (2.57)$$

where  $\theta_0$  is typically represented as a positive number as the phase margin can be referenced to either  $\pm 180^\circ$  and is symmetric for positive and negative  $\theta$ .

Thus, one can identify the phase margin by:

1. identifying critical frequencies  $\omega_{0,i}$  where  $|L(j\omega_{0,i})| = 1 = 0 \text{ dB}$ ;
2. calculating associated “candidate” phase margins,  $\theta_{0,i}$ , using distance from  $180^\circ$ , i.e.

$$\theta_{0,i} = \pm 180^\circ + \angle L(j\omega_0) \quad (2.58)$$

3. setting the phase margin,  $\bar{\theta}$ , as  $\min_i |\theta_{0,i}|$

which can be done simply from the Bode plot of  $L(j\omega)$ . Then, one may state that the feedback control system is stable for

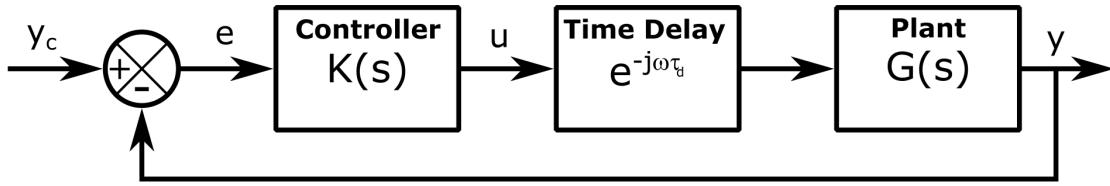
$$-\bar{\theta} \leq \theta \leq \bar{\theta} \quad (2.59)$$

For FDC, a feedback control system is generally considered sufficiently robust if  $\bar{\theta} \geq 45^\circ$  for the phase margin.

Alternatively, the phase margin can be rewritten in terms of a (time) delay margin. The **(time) delay margin** is the amount of time delay that can be added to the input before closed-loop instability. To which uses the fact that the critical time delay,  $\tau_d$ , is related to  $\theta_0$  by

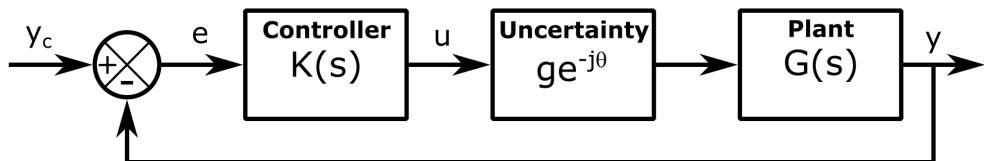
$$\tau_d = \frac{\theta_0}{\omega_0} \quad (2.60)$$

which can also be modeled in block diagram form as

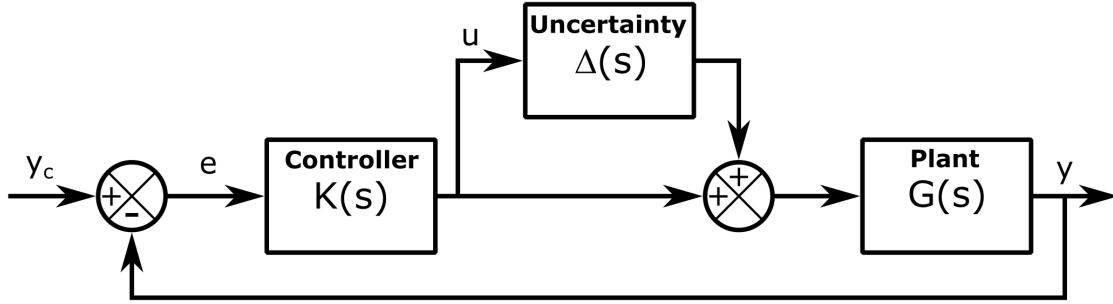


### Disk Margin

Consider a simultaneous gain and phase uncertainty model, i.e.



which encodes an arbitrary frequency response uncertainty to the nominal feedback control system. Alternatively, one could reform this form of uncertainty as  $\Delta(s)$  which can be modeled in block diagram form as



where  $\Delta(s) = 0$  refers to the nominal model.

Recalling that the feedback control system is stable if (1) there are no pole-zero cancellations and (2) the zeros of  $1 + L(s) = 0$  are only in the LHP. Then, a critical uncertainty,  $\Delta(j\omega_0)$ , occurs when the zeros of  $1 + (1 + \Delta(j\omega_0))L(j\omega_0)$  are on the imaginary axis for some critical frequency  $\omega_0$ , i.e.

$$1 + (1 + \Delta(j\omega_0))L(j\omega_0) = 0 \quad (2.61)$$

Rewriting this expression, one has

$$1 + L(j\omega_0) = \Delta(j\omega_0) - \Delta(j\omega_0)L(j\omega_0) \quad (2.62)$$

or in polar form (i.e. gain and phase), one has

$$\|1 + L(j\omega_0)\|e^{-j\angle(1+L(j\omega_0))} = \|\Delta(j\omega_0)L(j\omega_0)\|e^{j(\pm\pi-\angle(\Delta(j\omega_0)L(j\omega_0)))} \quad (2.63)$$

or the equivalent magnitude/gain and phase requirements become

$$\|\Delta(j\omega_0)L(j\omega_0)\| = \|1 + L(j\omega_0)\| \quad (2.64)$$

and

$$\angle(\Delta(j\omega_0)L(j\omega_0)) = \pm 180^\circ + \angle(1 + L(j\omega_0)) \quad (2.65)$$

Then, one can define the **disk margin**,  $d_{min}$ , as the minimum amount  $\Delta(j\omega_0)L(j\omega_0)$  adds to  $L(j\omega_0)$  at some  $\omega_0$  to create instability, i.e. the minimum perturbation “distance” allowable before closed-loop instability

$$d_{min} = \min_{0 \leq \omega < \infty} \|\Delta(j\omega_0)L(j\omega_0)\| \quad (2.66)$$

which, by the equivalence above, can also be considered as the minimum distance between  $L(j\omega)$  and the  $-1 \pm 0j$  point in the complex plane

$$d_{min} = \min_{0 \leq \omega < \infty} |1 + L(j\omega)| \quad (2.67)$$

which is minimized the closer  $L(j\omega)$  is to  $-1 \pm 0j$ . This also can be written in terms of the sensitivity transfer function as

$$d_{min} = \max_{0 \leq \omega < \infty} |S(j\omega)| \quad (2.68)$$

as  $|S(j\omega)| = |1 + L(j\omega)|^{-1}$ .

Furthermore, the gain and phase margins are two different “directions” in the complex domain that one is assessing the distance from  $-1 \pm 0j$  point. In particular, the gain margin criterion in polar form (i.e. gain and phase) is

$$\angle L(j\omega_0) = \pm 180^\circ \quad (2.69)$$

which restricts the minimization of the distance of  $L(j\omega)$  from -1 along the negative real axis for the gain margin, i.e.

$$d_{min,gm} = \min_{0 \leq \omega < \infty, \angle L(j\omega_0) = \pm 180^\circ} |1 + L(j\omega)| \quad (2.70)$$

or, as the constraint,  $\angle L(j\omega_0) = \pm 180^\circ$  means  $L(j\omega)$  is a negative real number, one has

$$d_{min,gm} = \begin{cases} \min_{0 \leq \omega < \infty, \angle L(j\omega_0) = \pm 180^\circ} 1 - |L(j\omega)| & L(j\omega) < -1 \\ \min_{0 \leq \omega < \infty, \angle L(j\omega_0) = \pm 180^\circ} |L(j\omega)| - 1 & L(j\omega) > -1 \end{cases} \quad (2.71)$$

and substituting  $|L(j\omega_0)| = \frac{1}{g_0}$ , one has

$$d_{min,gm} = \begin{cases} 1 - \frac{1}{g} = \min_{0 \leq \omega < \infty, \angle L(j\omega_0) = \pm 180^\circ} 1 - |L(j\omega)| & -1 < L(j\omega) < 0 \\ \frac{1}{g} - 1 = \min_{0 \leq \omega < \infty, \angle L(j\omega_0) = \pm 180^\circ} |L(j\omega)| - 1 & L(j\omega) < -1 \end{cases} \quad (2.72)$$

Furthermore, the phase margin criterion in polar form (i.e. gain and phase) is

$$|L(j\omega_0)| = 1 \quad (2.73)$$

which restricts the minimization to the distance of  $L(j\omega)$  from -1 along the unit circle for the phase margin, i.e.

$$d_{min,pm} = \min_{0 \leq \omega < \infty, |L(j\omega_0)| = 1} |1 + L(j\omega)| \quad (2.74)$$

or

$$d_{min,pm} = \min_{0 \leq \omega < \infty, |L(j\omega_0)| = 1} |1 + e^{j\angle L(j\omega_0)}| \quad (2.75)$$

and substituting  $\angle L(j\omega_0) = \pm\pi + \theta_0$ , one has

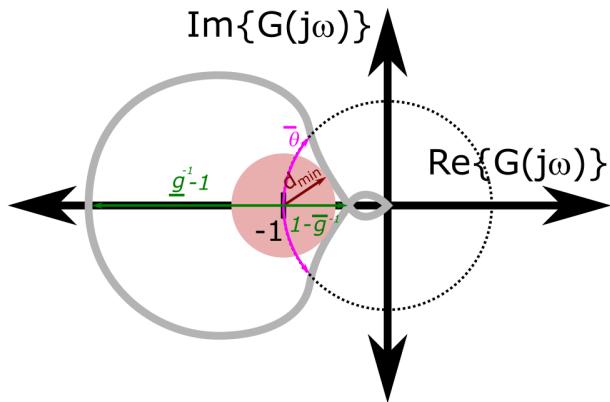
$$d_{min,pm} = \begin{cases} |1 + e^{j(\bar{\theta} + \pi)}| = \min_{0 \leq \omega < \infty, |L(j\omega_0)| = 1} |1 + e^{j\angle L(j\omega_0)}| & -\pi < \angle L(j\omega_0) < 0 \\ |1 + e^{j(\bar{\theta} - \pi)}| = \min_{0 \leq \omega < \infty, |L(j\omega_0)| = 1} |1 + e^{j\angle L(j\omega_0)}| & 0 < \angle L(j\omega_0) < \pi \end{cases} \quad (2.76)$$

Furthermore, the phase margin assesses the arc distance from  $-1 \pm 0j$  along the unit circle, i.e.

$$\theta_0 = \pm\pi + \angle \Delta(j\omega_0)L(j\omega_0) \quad (2.77)$$

With these definitions, a disk margin of 0.4 corresponds to a gain margins of  $\underline{g} = 0.71$  and  $\bar{g} = 1.67$  and a phase margin of  $\bar{\theta} = 23^\circ$ . However, the disk margin can also account for simultaneous gain and phase perturbations to the nominal feedback control system.

These relationships can be represented graphically in the Nyquist plot as



where the term “disk” derives from the disk that one forms this minimum distance that just intersects the frequency response of  $L(j\omega)$ .

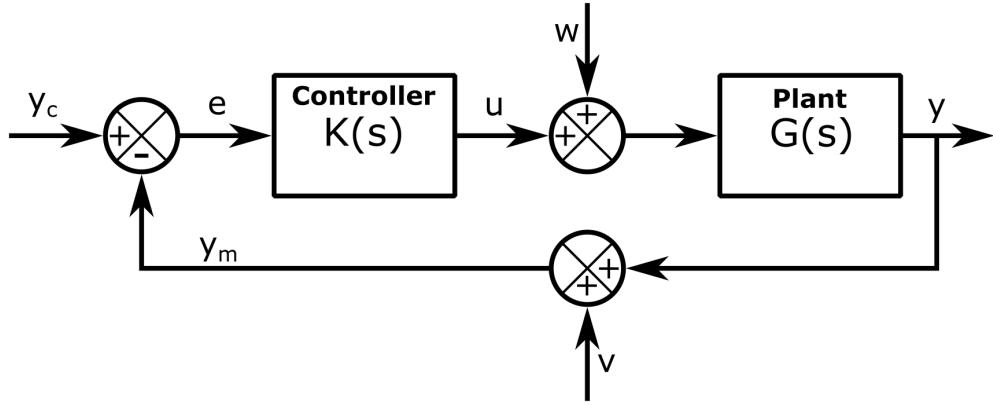
## References

For more information, please refer to the following

- Schmidt, D. K., “12.1 Feedback Control-Law Synthesis Via Loop Shaping - A Just-In-Time Tutorial\*,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 548-562
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “3.9 Feedback Control,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 213-240

## 2.3 Classical Feedback Control System Design

Recall the general feedback control system



with the objective to synthesize a controller  $K(s)$  such that it generally satisfies the following qualitative design requirements:

- 1 stable with good stability margins;
- 2 good tracking, i.e., the gain from  $y_c \rightarrow e$  is small;
- 3 disturbance rejection, i.e., the gain from  $w \rightarrow y$  is small;
- 4 sensor noise filtering, i.e., the gain from  $v \rightarrow e$  is small; and
- 5 the control effort is realistic, i.e.,  $|u|$  is not too large.

where design requirement 1 is typically referred to as the **stability requirements** and design requirements 2-5 are referred to as the **performance requirements**.

To satisfy all stability and performance requirements for SISO systems, one can use **loop-shaping** control design methods for synthesizing  $K(s)$  through “shaping” the SISO open-loop transfer function,  $L(s) = G(s)K(s)$ , to have certain frequency response characteristics. This section will establish the connections between these qualitative design requirements 1-5 and stability and performance requirements on  $L(j\omega)$ .

The primary aspect of loop-shaping is that  $K(s)$  has an additive effect on the magnitude subplot of the Bode plot when considered in decibels, i.e.

$$20 \log_{10} |L(j\omega)| = 20 \log_{10} |G(j\omega)K(j\omega)| = 20 \log_{10} |G(j\omega)| + 20 \log_{10} |K(j\omega)| \quad (2.78)$$

Thus, to form a suitable  $L(s)$ , one can use multiple **control stages** for different regions of the Bode plot using this additive property for shaping  $L(s)$ . Then, multiplying each stage together, i.e. using each stage in *series*, will provide the full controller transfer function  $K(s)$ , i.e. for  $N$  stages,

$$20 \log_{10} |K_1(j\omega) \cdots K_N(j\omega)| = 20 \log_{10} |K_1(j\omega)| + \cdots + 20 \log_{10} |K_N(j\omega)| \quad (2.79)$$

## Loop-Shaping Stability Requirements

Design requirement 1 specifies the feedback control system must be stable, i.e. the roots of the characteristic polynomial must be in the LHP, typically checked through the zeros of  $L(s)$  as there should also not be any pole-zero cancellations. In addition, one typically also requires some stability margin for the system due to uncertainty in the plant model  $G(s)$ . To provide some guidance on connecting stability margins to the loop-shaping control design method consider the following derivation. The **Bode gain-phase formula** states that if  $L(0) > 0$  and has all its poles and zeros in LHP, then the phase of  $L(j\omega)$  (in radians) at  $\omega$  is

$$\angle L(j\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{d|L|_{\text{dB}}(\zeta)}{d\nu} \log_{10} \coth \left| \frac{\nu}{2} \right| d\nu \quad (2.80)$$

where  $\nu = \log_{10} \frac{\zeta}{\omega}$ . Next, noting that

$$\log_{10} \coth \left| \frac{\nu}{2} \right| \approx \frac{1}{2} \pi^2 \delta(\nu) \quad (2.81)$$

one has the approximation

$$\angle L(j\omega) \approx \frac{\pi}{2} \left( \frac{d|L|_{\text{dB}}(\zeta)}{d\nu} \right)_{\zeta=\omega} \quad (2.82)$$

or in degrees

$$\angle L(j\omega) \approx \frac{90^\circ}{20 \text{ dB/decade}} (|L(j\omega)|_{\text{dB}}/\text{decade}) \quad (2.83)$$

Thus, if  $|L(j\omega)|_{\text{dB}}/\text{decade} = -30 \text{ dB/decade}$  over roughly 1 decade of  $\omega$ , then the phase approximately satisfies

$$\angle L(j\omega) \approx \frac{90^\circ}{20 \text{ dB/decade}} (-30 \text{ dB/decade}) \quad (2.84)$$

or

$$\angle L(j\omega) \approx -135^\circ \quad (2.85)$$

Then, recall that one typically requires a phase margin requirement of at least  $\pm 45^\circ$  and a gain margin requirement of at least  $\pm 6 \text{ dB}$ , which occur at  $\angle L(j\omega) = \pm 180^\circ$ , the open-loop transfer function,  $L(s)$ , should roughly satisfy the following:

1. no poles or zeros in the RHP;
2.  $|L(0)| > 0$ ;
3. a single gain crossover frequency  $\omega_c$ ;
4.  $|L(j\omega)|_{\text{dB}}/\text{decade} \geq -30 \text{ dB/decade}$  for  $\frac{\omega_c}{\sqrt{10}} < \omega < \sqrt{10}\omega_c$ ;
5.  $|L|_{\text{dB}} \geq 6 \text{ dB}$  for  $\omega \leq \frac{\omega_c}{\sqrt{10}}$ ; and
6.  $|L|_{\text{dB}} \leq -6 \text{ dB}$  for  $\omega \geq \sqrt{10}\omega_c$ .

Then, one can confidently claim that the feedback control system achieves the stability requirements.

Though these assumptions do not hold for all LTI systems, this approximation is generally decent and provides at least a good starting point for how much to limit the gain slope about  $\omega_c$ . It should be noted that if  $L(s)$  has zeros or poles in the RHP, then there may not exist a suitable  $K(s)$ , but at the very least there will be some additional phase introduced at  $\omega_c$  which will require a larger slope than  $-30 \text{ dB/decade}$ . This

additional phase is why a system with no poles or zeros in the RHP is called a **minimum phase system**. For non-minimum phase systems, additional control design must be performed case-by-case. Lastly, it should be noted that the *single* crossover frequency  $\omega_c$  for  $|L(j\omega_c)| = 0$  dB is also known as the **loop bandwidth** for the feedback control system, and is typically specified for some required frequency level in the design requirements.

## Loop-Shaping Performance Requirements

Design requirement 2 specifies the feedback control system to have good tracking, i.e., keep  $e$  small. Recalling the previously derived transfer function for  $y_c \rightarrow e$ , i.e.

$$\frac{e(s)}{y_c(s)} = \frac{1}{1 + GK} = \frac{1}{1 + L} = S \quad (2.86)$$

one requires that  $|\frac{1}{1+L}| \ll 1$  which occurs if  $|L| \gg 1$ . Furthermore, note that

$$|S(s)| = \left| \frac{1}{1 + L(s)} \right| \quad (2.87)$$

$$|1 + L(s)| = \left| \frac{1}{S(s)} \right| \quad (2.88)$$

and the triangle inequality, i.e.

$$|a + b| \leq |a| + |b| \quad (2.89)$$

Then, one can state

$$|1 - 1 + L(s)| \leq |-1| + |1 + L(s)| \quad (2.90)$$

$$|L(s)| \leq 1 + |1 + L(s)| \quad (2.91)$$

$$|L(s)| - 1 \leq |1 + L(s)| \quad (2.92)$$

Thus, for a requirement on  $|S(j\omega)| \leq e_{\omega,req}$ , one has that

$$|e| \leq e_{\omega,req} \rightarrow \left| \frac{1}{S(j\omega)} \right| \geq \frac{1}{e_{\omega,req}} \quad (2.93)$$

$$|e| \leq e_{\omega,req} \rightarrow |1 + L(j\omega)| \geq \frac{1}{e_{\omega,req}} \quad (2.94)$$

or, using the inequality above, one can define the approximate relationship

$$|e| \leq e_{\omega,req} \rightarrow |L(j\omega)| \geq 1 + \frac{1}{e_{\omega,req}} \quad (2.95)$$

Design requirement 3 specifies the feedback control system to reject disturbances, i.e.  $w$  should have little effect on  $y$ . Recalling the previously derived transfer function for  $w \rightarrow y$ , i.e.

$$\frac{y(s)}{w(s)} = \frac{G}{1 + GK} = \frac{G}{1 + L} \quad (2.96)$$

one requires that  $|\frac{G}{1+L}| \ll 1$ . However, noting that the effect of the disturbance with feedback control should be at least smaller than with no control, i.e.  $|\frac{G}{1+L}| \ll |G|$  or  $|\frac{1}{1+L}| \ll 1$ , which occurs if that  $|L| \gg 1$ .

Design requirement 4 specifies the feedback control system to filter out the sensor noise, i.e.  $v$  should have little effect on  $e$ . Recalling the previously derived transfer function for  $v \rightarrow e$ , i.e.

$$\frac{e(s)}{v(s)} = \frac{-GK}{1 + GK} = \frac{-L}{1 + L} \quad (2.97)$$

one requires that  $|\frac{L}{1+L}| \ll 1$  which occurs if  $|L| \ll 1$ . For example, one typically requires filtering out noise at high frequencies, i.e.

$$|T(j\omega)| \leq g \quad \forall \omega > \omega_{high} \quad (2.98)$$

But by definition

$$T(j\omega) = \frac{L(j\omega)}{1 + L(j\omega)} \quad (2.99)$$

Thus, one requires that

$$\left| \frac{1 + L(j\omega)}{L(j\omega)} \right| \leq \frac{1}{g} \quad (2.100)$$

or if

$$\left| \frac{1}{L(j\omega)} \right| \geq \frac{1}{g} + 1 \quad (2.101)$$

then, by the triangle inequality,

$$\left| \frac{1 + L(j\omega)}{L(j\omega)} \right| \geq \left| \frac{1}{L(j\omega)} \right| - 1 \geq \frac{1}{g} \quad (2.102)$$

Then, these type of requirements can be converted to

$$\left| \frac{1}{L(j\omega)} \right| \leq \frac{1 + g}{g} \quad (2.103)$$

or

$$|L(j\omega)| \leq \frac{g}{1 + g} \quad \forall \omega > \omega_{high} \quad (2.104)$$

Design requirement 5 specifies the feedback control system to keep  $u$  small enough. Recalling the previously derived transfer function for  $y_c \rightarrow u$ , i.e.

$$\frac{u(s)}{y_c(s)} = \frac{K}{1 + GK} \quad (2.105)$$

one requires that  $|\frac{K}{1+L}| \ll 1$ . Because  $G$  is fixed, one has 2 cases:

1. If  $|G| \gg 1$ , then all  $|K|$  will provide  $|\frac{K}{1+GK}| \ll 1$
2. If  $|G| \ll 1$ , then  $|K| \ll 1$  will provide  $|\frac{K}{1+GK}| \ll 1$

Summarizing these rough guidelines results in the following table.

General Requirement	Closed-Loop TF Requirement	Loop TF Requirement
1. Good Tracking	$ S  \ll 1$	$ L  \gg 1$
2. Disturbance Rejection	$ S  \ll 1$	$ L  \gg 1$
3. Noise Filtering	$ T  \ll 1$	$ L  \ll 1$
4. Small control effort	$ KS  \ll 1$	If $ G  \ll 1$ , then $ K  \ll 1$

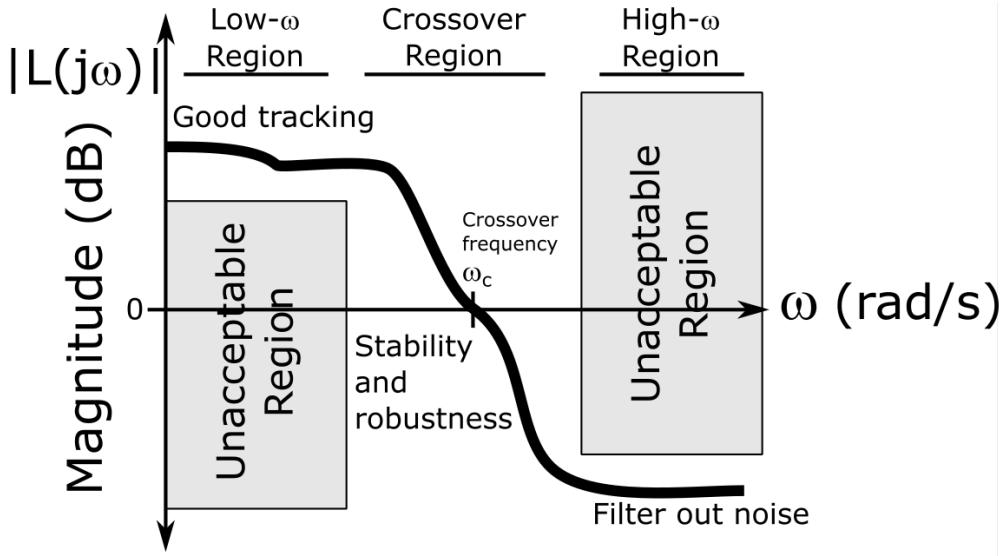
Thus, the stipulation that  $S + T = 1$  demonstrates that one cannot simultaneously satisfy all control design requirements. Thus, the *key idea* in frequency domain control design is that output commands are designed to be low frequency signals relative to any high frequency noise signals on the actual output. Thus, one requires the feedback control system at least satisfies

1.  $|S(j\omega)| \ll 1$  at low  $\omega$
2.  $|T(j\omega)| \ll 1$  at high  $\omega$

Then, translating these to the open-loop transfer function,  $L(j\omega)$ , provides the equivalent loop-shaping performance requirements

1.  $|L(j\omega)| \gg 1$  at low  $\omega$
2.  $|L(j\omega)|_{\text{dB}/\text{decade}} \geq -30 \text{ dB}/\text{decade}$  around  $\omega_c$
3.  $|L(j\omega)| \ll 1$  at high  $\omega$

which can be analyzed visually using the magnitude subplot of the Bode plot of  $L(j\omega)$ , i.e.



For the control effort requirement, one must analyze the Bode plot of  $K(j\omega)S(j\omega)$  in parallel with the loop-shaping of  $L(s)$  in these three regions. As a general rule, one typically should design  $K(j\omega)$  to not be too large where  $G(j\omega)$  is small. This is similar to the high-frequency requirement for sensor noise filtering and suppression of high-frequency model uncertainty.

### Frequency and Step Response Characteristics Relationships

Alternatively, in some cases, the design requirements for a control law may be specified as certain unit step response characteristics, i.e.

$$6 \quad t_s \leq t_{s,req}$$

$$7 \quad M_p < M_{p,req}$$

$$8 \quad |e_{ss}| \leq e_{ss,req}$$

where  $t_{s,req}$ ,  $e_{ss,req}$ , and  $u_{req}$  are some values where  $e_{ss,req}$  and  $u_{req}$  may alternatively be percentages. This subsection will discuss these alternative step response requirements and their loose relationships to the loop-shaping design requirements.

For design requirement 6,  $t_s \leq t_{s,req}$ , as a rough approximation, the loop bandwidth is related to the settling time roughly by

$$t_s \approx \frac{3}{\omega_c} \quad (2.106)$$

Thus, one can define the approximate relationship that

$$t_s \leq t_{s,req} \rightarrow \omega_c \text{ of } L(j\omega) = 0 \text{ dB} \geq \frac{3}{t_{s,req}} \quad (2.107)$$

However, for some systems this approximation can be significantly off, but, in general, increasing  $\omega_c$  will decrease  $t_s$  and is done through design iteration.

For design requirement 7,  $M_p < M_{p,req}$ , as a rough approximation, the phase margin can be related to the damping ratio roughly by

$$\bar{\theta} \approx 100\zeta \quad (2.108)$$

and using the second-order expression for the maximum overshoot in terms of  $\zeta$ , i.e.

$$M_p \approx e^{\frac{\zeta\pi}{\sqrt{1-\zeta^2}}} \quad (2.109)$$

One can define the approximate relationship that

$$M_p < M_{p,req} \rightarrow \bar{\theta} > 100 \sqrt{\frac{\ln M_{p,req}^2}{\pi^2 + \ln M_{p,req}^2}} \quad (2.110)$$

However, for some systems this approximation can be significantly off, but, in general, increasing  $\bar{\theta}$  will decrease  $M_p$  and is done through design iteration.

For design requirement 8,  $|e_{ss}| \leq e_{ss,req}$  can be regarded as a requirement on  $|S(0)| \leq e_{ss,req}$ , as  $\omega = 0$  describes the step input. Thus, one can define the exact relationship that

$$|e_{ss}| \leq e_{ss,req} \rightarrow |L(0)| \geq 1 + \frac{1}{e_{ss,req}} \quad (2.111)$$

which is simply another form of the low frequency requirement on  $L(j\omega)$ .

## SISO Loop-Shaping Control Stages

Loop-shaping uses multiple ***N* control stages** in order to shape different frequency regions of  $L(s)$ . Thus, one designs controllers in *series form* as

$$K(s) = K_1(s) \cdots K_N(s) \quad (2.112)$$

which is notably additive in the Bode plot which serves as the primary analysis tool. As each stage is in series, multiplying each stage together provides the controller transfer function  $K(s)$ . The four most common control stages are:

1. Proportional
2. Integral
3. Derivative
4. Low-Pass Filter

while four others are also common in some applications

1. High-Pass Filter
2. Band-Pass Filter
3. Band-Stop and Notch Filters
4. Low-Frequency Boost
5. Lag

## 6. Lead

It should be noted that all control stages are not necessary for a sufficient controller and typically the least number of stages, i.e., less complexity, is preferred by control engineers.

With this in mind, for the general SISO loop-shaping control design, one can consider the following iterative procedure for **loop-shaping design** of  $K(s)$  until all design requirements are satisfied.

1. Use a proportional control stage to set the loop bandwidth,  $\omega_c$ 
  - May need to obtain requirements on  $\omega_c$  from  $t_s$
2. Use integral control stage(s) to increase  $|L(j\omega)|$  at low  $\omega$ , i.e. good tracking.
  - May need to obtain requirements on  $L(j\omega)$  from  $S(j\omega)$  or  $e_{ss}$
3. Use derivative control stage(s) to reduce  $|L(j\omega)|$  slope about  $\omega_c$  to  $> -30 \text{ dB}/10\omega$  for stability and robustness.
  - Will need to lower proportional gain based on new loop bandwidth
4. Use low-pass filter control stage(s) to decrease  $|L(j\omega)|$  at high  $\omega$ , i.e. good noise filtering.
  - May need to obtain requirements on  $L(j\omega)$  from  $T(j\omega)$

It should be noted that iteration must typically be used as the control stages may slightly affect other frequency regions.

## Proportional Control Stage

A **proportional control stage** is used to shift  $L(s)$  up or down, i.e. to set the crossover frequency,  $\omega_c$ , a.k.a. the bandwidth, in the crossover-frequency region to have the system respond as fast as necessary. The transfer function for this control stage is defined as

$$K(s) = K_p \quad (2.113)$$

where  $K_p$  is the gain. Note that  $|K_p| > 1$  shifts  $L(s)$  up while  $|K_p| < 1$  shifts  $L(s)$  down. Note also that  $K_p > 0$  or  $K_p < 0$  should be chosen so that the feedback control system is stabilized while also not being too large as to violate any control effort constraints.

As an example, consider the following system

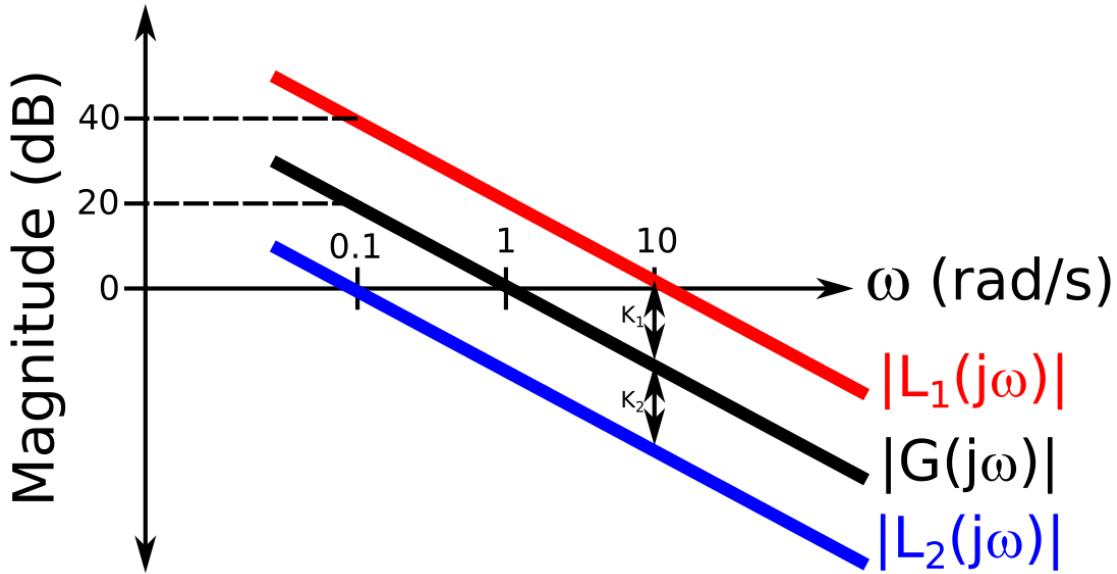
$$G(s) = \frac{1}{s} \quad (2.114)$$

$$K_1(s) = 10 \quad (2.115)$$

$$K_2(s) = 0.1 \quad (2.116)$$

$$L_1(s) = \frac{10}{s} \quad (2.117)$$

$$L_2(s) = \frac{0.1}{s} \quad (2.118)$$



### Integral Control Stage

An **integral control stage** is used to increase the gain of  $L(s)$  for the low-frequency region  $\omega < \omega_i$ , i.e. to have good tracking at low frequencies. The transfer function for this control stage is defined as

$$K(s) = \frac{s + \omega_i}{s} \quad (2.119)$$

where  $\omega_i$  is the chosen corner frequency/bandwidth to start the gain increase. This stage has the following notable properties:

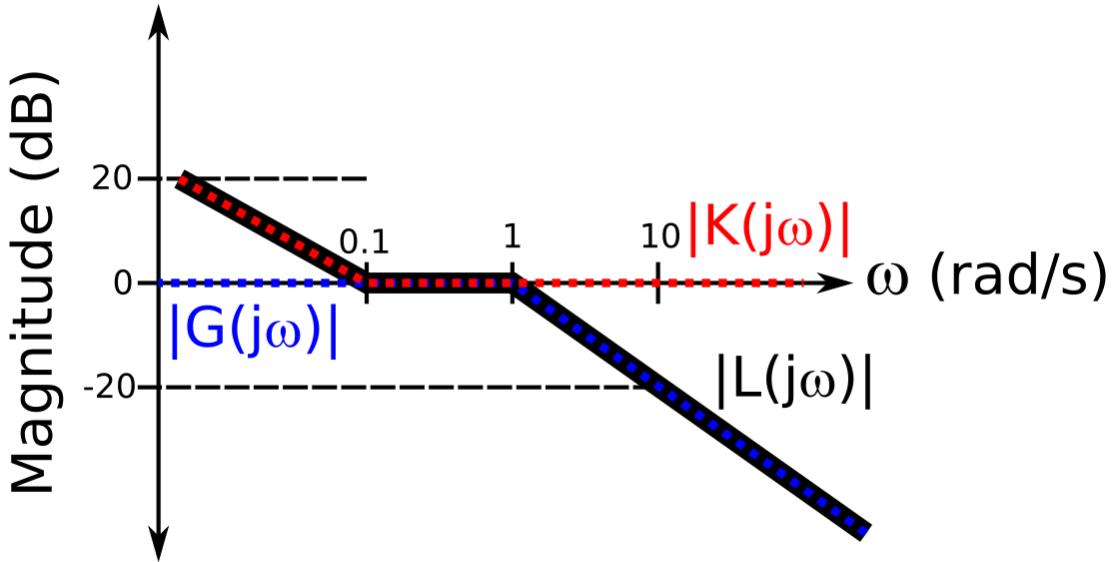
1.  $K(s)$  has a pole at  $s = 0$  and a zero at  $-\omega_i$ ;
2.  $K(j\omega) \rightarrow +\infty$  as  $\omega \rightarrow 0$ , thus  $e_{ss} = 0$ ;
3. at low  $\omega$ ,  $K(j\omega) \approx \frac{\omega_i}{j\omega}$ , i.e. for every  $10\omega \rightarrow K(j\omega)/10$  or a -20 dB/10ω slope; and
4. at high  $\omega$ ,  $K(j\omega) \approx 1$ , i.e. no effect (0 dB).

As an example, consider the following system

$$G(s) = \frac{1}{s + 1} \quad (2.120)$$

$$K(s) = \frac{s + 0.1}{s} \quad (\omega_i = 0.1) \quad (2.121)$$

$$L(s) = \frac{s + 0.1}{s^2 + s} \quad (2.122)$$



### Derivative Control Stage

A **derivative control stage** is used to increase the slope of  $L(j\omega)$  for the crossover-frequency region  $\omega > \omega_d$ . The transfer function for this control stage is defined as

$$K(s) = \frac{1}{\omega_d} s + 1 \quad (2.123)$$

where  $\omega_d$  is the rollup corner frequency/bandwidth. This stage has the following notable properties:

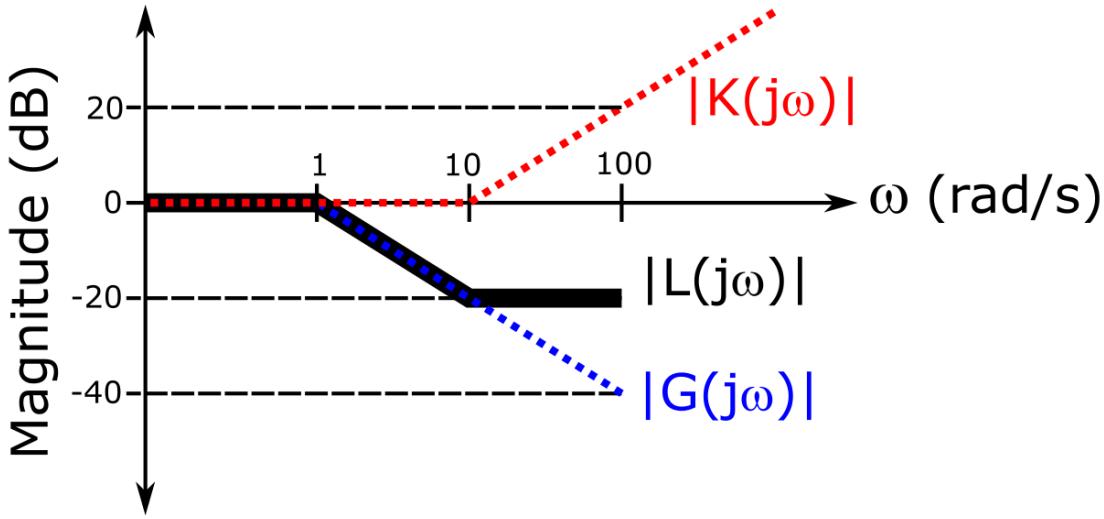
1.  $K(s)$  has a zero at  $s = -\omega_d$ .
2. At low  $\omega$ ,  $K(j\omega) \approx 1$  or 0 dB.
3. At high  $\omega$ ,  $K(j\omega) \approx \frac{j\omega}{\omega_d}$ , i.e. for every  $10\omega \rightarrow 10K(j\omega)$  or a 20 dB/10ω slope.

As an example, consider the following

$$G(s) = \frac{1}{s+1} \quad (2.124)$$

$$K(s) = 0.1s + 1 \quad (\omega_d = 10) \quad (2.125)$$

$$L(s) = 0.1 \frac{s+10}{s+1} \quad (2.126)$$



### Low-Pass Filter Control Stage

A **low-pass filter control stage** is used to decrease the gain of  $L(j\omega)$  for the high-frequency region  $\omega > \omega_{l-p}$ , i.e. to have noise filtering at high frequencies. This stage is “low-pass” as it passes frequencies below  $\omega_{l-p}$  with approximately unity gain, but significantly reduces the gain for higher frequencies. The types of low-pass filters used in practice can vary and include

- Butterworth filter
- Chebyshev filter
- Legendre-Papoulis filter
- Cauer filter
- Bessel filter

which balance the “unity gain” approximation and the attenuation or “rolloff” at the corner frequency. The order of these low-pass filter is typically a parameter as well where lower orders are typically preferred.

For SISO loop-shaping control design, one typically desires to keep a linear phase response, so low-pass Butterworth filters are often preferred. The  $n^{\text{th}}$ -order low-pass Butterworth filter control stage can be defined as

$$K_n(s) = \frac{1}{B_n \left( \frac{s}{\omega_{l-p}} \right)} \quad (2.127)$$

where  $\omega_{l-p}$  is the filter corner frequency/bandwidth and  $B_n$  is the **normalized Butterworth polynomial** defined as

$$B_n(s) = \begin{cases} \prod_{k=0}^{\frac{n}{2}-1} \left( s^2 - 2 \cos \left( 2\pi \frac{2k+n+1}{4n} \right) s + 1 \right) & n \text{ even} \\ (s+1) \prod_{k=0}^{\frac{n-1}{2}-1} \left( s^2 - 2 \cos \left( 2\pi \frac{2k+n+1}{4n} \right) s + 1 \right) & n \text{ odd} \end{cases} \quad (2.128)$$

It can be shown that the magnitude of an  $n^{\text{th}}$ -order low-pass Butterworth filter is

$$\|K_n(j\omega)\| = \frac{1}{\sqrt{1 + \left(\frac{\omega}{\omega_{l-p}}\right)^{2n}}} \quad (2.129)$$

The transfer function for a first-order Butterworth low-pass filter would be

$$K(s) = \frac{\omega_{l-p}}{s + \omega_{l-p}} = \frac{1}{\frac{1}{\omega_{l-p}}s + 1} \quad (2.130)$$

where  $\omega_{l-p}$  is the filter corner frequency/bandwidth. This stage has the following notable properties:

1.  $K(s)$  has a pole at  $s = -\omega_{l-p}$ .
2. At low  $\omega$ ,  $K(j\omega) \approx 1$  or 0 dB.
3. At high  $\omega$ ,  $K(j\omega) \approx \frac{\omega_{l-p}}{j\omega}$ , i.e. for every  $10\omega \rightarrow K(j\omega)/10$  or a -20 dB/10 $\omega$  slope.

The transfer function for a second-order Butterworth low-pass filter would be

$$K(s) = \frac{\omega_{l-p}^2}{s^2 + \sqrt{2}\omega_{l-p}s + \omega_{l-p}^2} \quad (2.131)$$

where  $\omega_{l-p}$  is the filter corner frequency/bandwidth and the damping ratio of the second-order system is  $\sqrt{2}/2$ . This stage has the following notable properties:

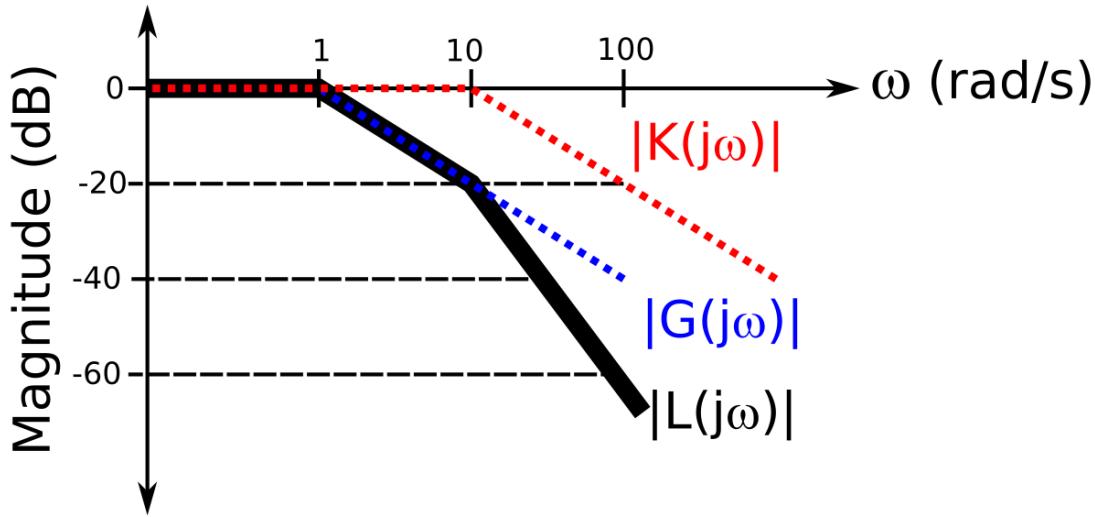
1.  $K(s)$  has poles at  $s = -\frac{\sqrt{2}}{2}\omega_{l-p} \pm j\frac{\sqrt{2}}{2}\omega_{l-p}$ .
2. At low  $\omega$ ,  $K(j\omega) \approx 1$  or 0 dB.
3. At high  $\omega$ ,  $K(j\omega) \approx \frac{\omega_{l-p}}{\omega^2}$ , i.e. for every  $10\omega \rightarrow K(j\omega)/100$  or a -40 dB/10 $\omega$  slope.

As an example, consider the following first-order low-pass Butterworth filter

$$G(s) = \frac{1}{s + 1} \quad (2.132)$$

$$K(s) = \frac{1}{B_n\left(\frac{s}{10}\right)} = \frac{10}{s + 10} \quad (2.133)$$

$$L(s) = \frac{10}{(s^2 + 11s + 10)} \quad (2.134)$$



### High-Pass Filter Control Stage

A **high-pass filter control stage** is used to decrease the gain of  $L(j\omega)$  for the low-frequency region  $\omega < \omega_{h-p}$ , i.e. to have noise filtering at high frequencies. This stage is “high-pass” as it passes frequencies above  $\omega_{h-p}$  with approximately unity gain, but significantly reduces the gain for lower frequencies.

For SISO loop-shaping control design, one typically desires to keep a linear phase response, so high-pass Butterworth filters are often preferred. The  $n^{\text{th}}$ -order high-pass Butterworth filter control stage can be defined as

$$K_n(s) = \frac{s^n}{B_n \left( \frac{s}{\omega_{h-p}} \right)} \quad (2.135)$$

It can be shown that the magnitude of an  $n^{\text{th}}$ -order high-pass Butterworth filter is

$$\|K_n(j\omega)\| = \frac{1}{\sqrt{1 + \left( \frac{\omega_{h-p}}{\omega} \right)^{2n}}} \quad (2.136)$$

The transfer function for a first-order Butterworth high-pass filter would be

$$K(s) = \frac{\omega_{h-p}s}{s + \omega_{h-p}} = \frac{s}{\frac{1}{\omega_{h-p}}s + 1} \quad (2.137)$$

where  $\omega_{h-p}$  is the filter corner frequency/bandwidth. This stage has the following notable properties:

1.  $K(s)$  has a pole at  $s = -\omega_{h-p}$ .
2. At high  $\omega$ ,  $K(j\omega) \approx 1$  or 0 dB.
3. At low  $\omega$ ,  $K(j\omega) \approx j\omega$ , i.e. for every  $10\omega \rightarrow 10K(j\omega)$  or a 20 dB/10ω slope.

The transfer function for a second-order Butterworth high-pass filter would be

$$K(s) = \frac{\omega_{h-p}^2 s^2}{s^2 + \sqrt{2}\omega_{h-p} s + \omega_{h-p}^2} \quad (2.138)$$

where  $\omega_{h-p}$  is the filter corner frequency/bandwidth and the damping ratio of the second-order system is  $\sqrt{2}/2$ . This stage has the following notable properties:

1.  $K(s)$  has poles at  $s = -\frac{\sqrt{2}}{2}\omega_{h-p} \pm j\frac{\sqrt{2}}{2}\omega_{h-p}$ .
2. At high  $\omega$ ,  $K(j\omega) \approx 1$  or 0 dB.
3. At low  $\omega$ ,  $K(j\omega) \approx -\omega^2$ , i.e. for every  $10\omega \rightarrow 100K(j\omega)$  or a 40 dB/10ω slope.

### Band-Pass and Band-Stop Filter Control Stages

A **band-pass filter control stage** is used to increase the gain of  $L(j\omega)$  over a frequency region  $\omega_{h-p} < \omega < \omega_{l-p}$ . This stage is “band-pass” as it passes frequencies between  $\omega_{h-p}$  and  $\omega_{l-p}$  with approximately unity gain, but significantly reduces the gain for outside these frequencies. A **band-stop filter control stage** is used to decrease the gain of  $L(j\omega)$  over a frequency region  $\omega_{l-p} < \omega < \omega_{h-p}$ . This stage is “band-stop” as it passes frequencies below  $\omega_{l-p}$  and above  $\omega_{h-p}$  with approximately unity gain, but significantly reduces the gain for between these frequencies.

Thus, a simple approach to these filters is to use a combination of low-pass and high-pass filters if the two stage don’t interact too much. However, a narrow band-stop filter can be designed as a **notch filter control stage**. A simple notch filter transfer function is defined as

$$K(s) = \frac{s^2 + \omega_z^2}{s^2 + 2\zeta_n\omega_p s + \omega_p^2} \quad (2.139)$$

where  $\zeta_n$  is the notch filter damping ratio, and  $\omega_z$  and  $\omega_p$  are chosen based on the type of notch filter, i.e.,  $\omega_z = \omega_p$  for standard notch,  $\omega_z > \omega_p$  for low-pass notch, and  $\omega_z < \omega_p$  for high-pass notch. It should be noted that sometimes the **Q-factor**,  $Q = \frac{1}{2\zeta_n}$  is used instead of the filter damping ratio. For the standard notch filter,  $\omega_z = \omega_p = \omega_c$  will be the center rejected frequency and  $\omega_w = 2\zeta_n\omega_c$  will provide the width of the rejected band, i.e.,

$$K(s) = \frac{s^2 + \omega_c^2}{s^2 + \omega_w s + \omega_c^2} \quad (2.140)$$

### Low-Frequency Boost Control Stage

A **low-frequency boost control stage** is used to alternatively increase the gain of  $L(s)$  for the low-frequency region  $\omega < \omega_i$ , i.e. to have good tracking at low frequencies, *without* an infinite gain at  $\omega = 0$  which may not be practical in hardware implementation. The transfer function for this control stage is defined as

$$K(s) = \frac{s + \omega_i}{s + \frac{\omega_i}{\beta}} \quad (2.141)$$

where  $\omega_i$  is the corner frequency/bandwidth and  $\beta > 1$  is the low frequency gain. This stage has the following notable properties:

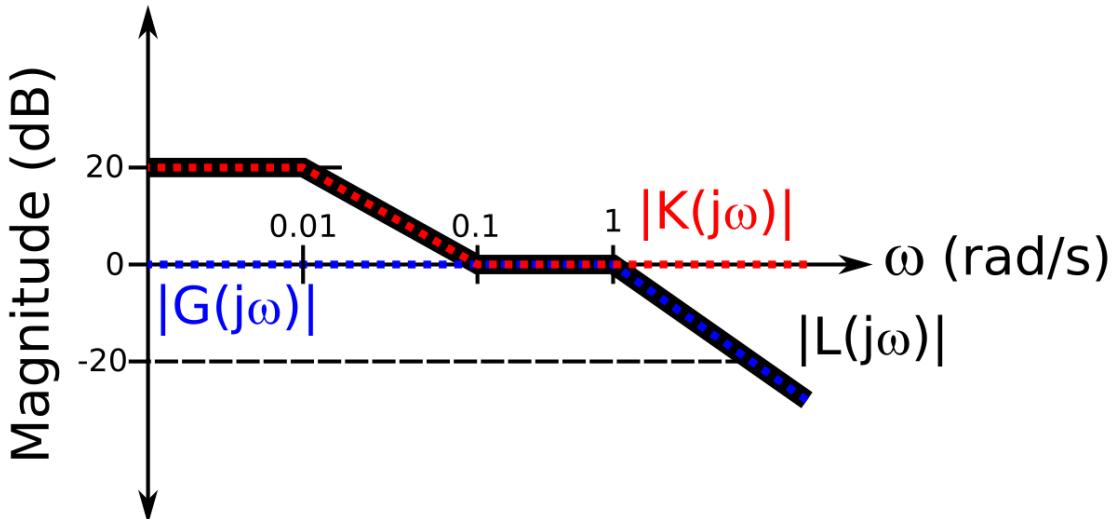
1.  $K(s)$  has a pole at  $s = \frac{-\omega_i}{\beta}$  and a zero at  $-\omega_i$ ;
2. at low  $\omega$ ,  $K(j\omega) \approx \beta$  or  $20 \log_{10} \beta$  dB;
3. at high  $\omega$ ,  $K(j\omega) \approx 1$  or 0 dB; and
4.  $K(s)$  is similar to integral stage with  $\omega_i$ , but levels off at the low frequency gain  $\beta$ , i.e. as  $\beta \rightarrow \infty$ , this stage approaches an integral stage.

As an example, consider the following

$$G(s) = \frac{1}{s + 1} \quad (2.142)$$

$$K(s) = \frac{s + 0.1}{s + 0.01} \quad (\omega_i = 0.1, \beta = 10) \quad (2.143)$$

$$L(s) = \frac{s + 0.1}{(s^2 + 1.01s + 0.01)} \quad (2.144)$$



### Lag-Lead Control Stage

A **lag control stage** is used to increase the low-frequency gain, i.e., better reference tracking, without any resulting instability, and to increase the phase margin of the system to yield the desired transient response. The transfer function for this control stage is defined as

$$K(s) = \frac{s + \omega_{lag}}{s + \frac{\omega_{lag}}{\alpha}} \quad (2.145)$$

where  $\alpha > 1$ .

A **lead control stage** is used to increase the bandwidth while providing a larger phase margin and a higher phase margin frequency. The transfer function for this control stage is defined as

$$K(s) = \frac{s + \omega_{lead}}{\beta s + \omega_{lead}} \quad (2.146)$$

where  $\beta < 1$ .

A **lag-lead control stage** may be more economical to use than separate lag and lead control stages. The transfer function for this control stage is defined as

$$K(s) = \frac{(s + \omega_{lead})(s + \omega_{lag})}{(s + \gamma\omega_{lead})\left(s + \frac{\omega_{lag}}{\gamma}\right)} \quad (2.147)$$

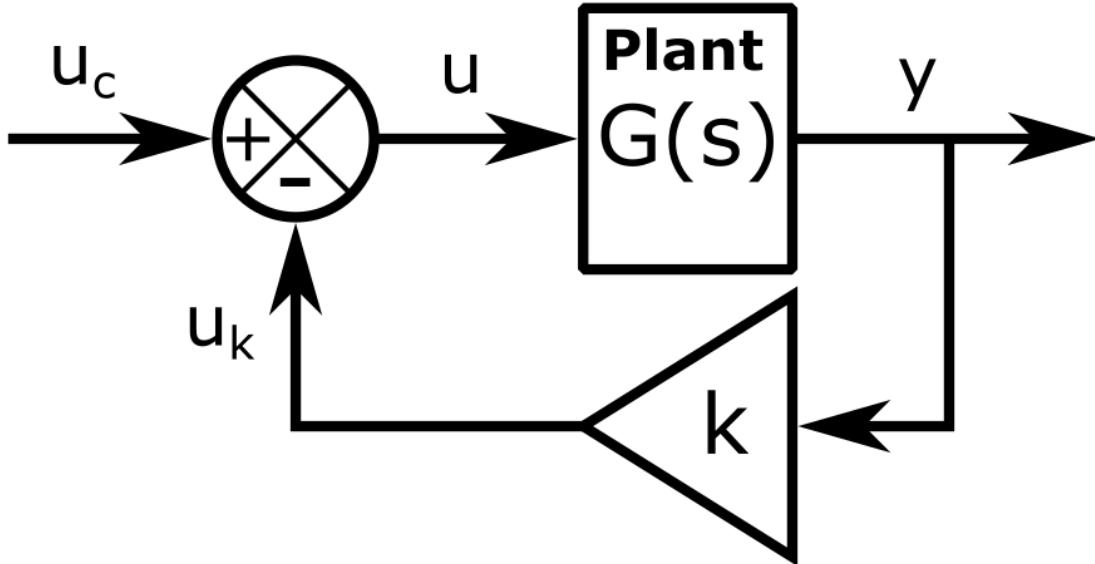
where  $\gamma > 1$  and  $\omega_{lag} < \omega_{lead}$ . Note that this form uses the design that  $\alpha = \beta^{-1}$ .

### Stability Augmentation System Design

The classical stability augmentation system design method is the root locus method defined as a **proportional SAS (P-SAS)** as

$$K_{P-SAS}(s) = \frac{u_k(s)}{y(s)} = k \quad (2.148)$$

where  $k > 0$  is the negative feedback gain as shown in the following block diagram



Thus, a P-SAS produces an overall system transfer function

$$y(s) = \left[ \frac{G(s)}{1 + kG(s)} \right] u_c(s) \quad (2.149)$$

with closed-loop system poles given by

$$d_G(s) + kn_G(s) = 0 \quad (2.150)$$

where  $n_G(s)$  is the numerator polynomial of  $G(s)$  and  $d_G(s)$  is the denominator polynomial of the minimum-phase  $G(s)$ . Thus, by sweeping through values of  $k$  from 0 to  $\infty$ , one can alter the system poles from  $d_G(s)$  to  $n_G(s)$  and/or  $\pm\infty$ , depending on the number of zeros relative to the number of poles. This behavior can easily be plotted on the **root locus** plot of the system poles and zeros in the complex plane as  $k$  varies from 0 to  $\infty$ , e.g. if one has a second-order plant given by

$$G(s) = \frac{s+2}{s^2 + 2s + 2} \quad (2.151)$$

which has poles at  $s = -1 \pm j$  and a zero at  $s = -2$ .

Then, for a P-SAS system, the closed-loop characteristic equation is

$$s^2 + 2s + 2 + k(s+2) = 0 \quad (2.152)$$

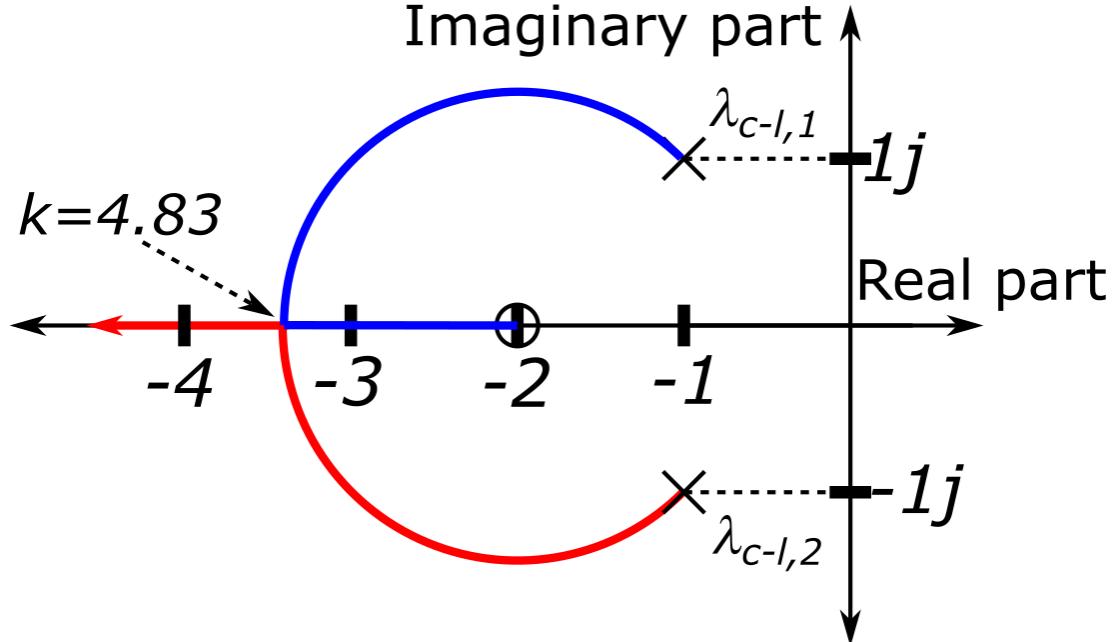
or

$$s^2 + (k+2)s + (2k+2) = 0 \quad (2.153)$$

for which the two closed-loop poles,  $\lambda_{c-l,1}$  and  $\lambda_{c-l,2}$ , are given by

$$\lambda_{c-l,1}, \lambda_{c-l,2} = \frac{-(k+2) \pm \sqrt{(k+2)^2 - 4(2k+2)}}{2} \quad (2.154)$$

Then, one has a root locus given by



where one can place  $\lambda_{c-l,1}$  and  $\lambda_{c-l,2}$  anywhere along the blue or red lines, respectively, depending on the choice of  $k$ .

Notably, the second-order closed-loop system becomes critically damped,  $\zeta = 1$ , at

$$k + 2 = 2\sqrt{2k + 2} \quad (2.155)$$

$$k^2 + 4k + 4 = 8k + 8 \quad (2.156)$$

$$k^2 - 4k - 4 = 0 \quad (2.157)$$

$$k = \frac{4 \pm \sqrt{(-4)^2 - 4(-4)}}{2} = 2 \pm 2\sqrt{2} \quad (2.158)$$

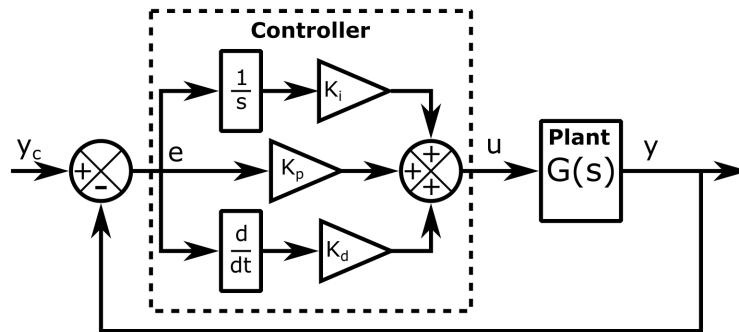
$$k = -0.83 \text{ or } 4.83 \quad (2.159)$$

### Proportional-Integral-Derivative Control Design

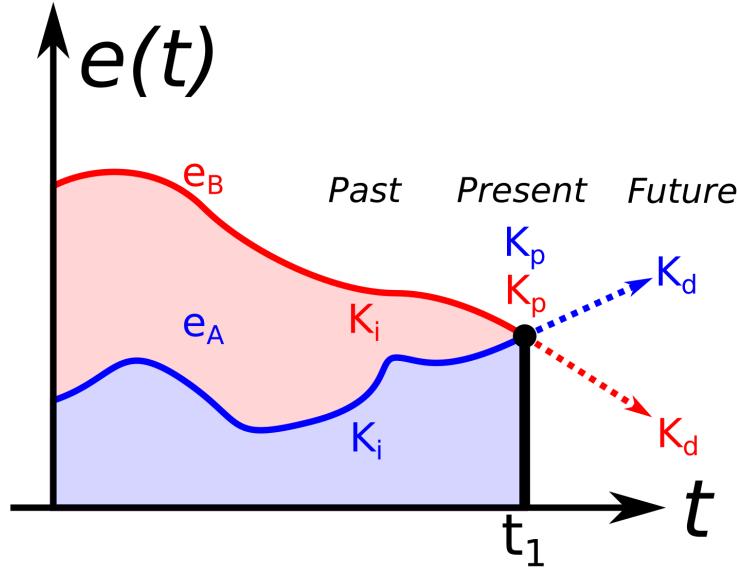
The most common classical control design method is the **proportional-integral-derivative (PID) control law** defined in **parallel form** as

$$K_{PID}(s) = \frac{u(s)}{e(s)} = K_p + \frac{K_i}{s} + K_d s \quad (2.160)$$

where  $K_p$  is the **proportional gain**,  $K_i$  is the **integral gain**, and  $K_d$  is the **derivative gain**.



It should be noted that if  $K_d = 0$ , one has a proportional-integral (PI) controller, if  $K_i = 0$ , one has a proportional-derivative (PD) controller and if  $K_d = K_i = 0$ , one has a simple proportional (P) controller. Here the control designer must choose the set of gains,  $(K_p, K_i, K_d)$ , to satisfy the design requirements. An intuitive understanding of PID control is to look at the following diagram of the tracking error history for two different systems as shown below.



Here the proportional term would notably have the same correction effect here at the present time,  $t_1$ , while the integral term has a correction effect based on the accumulated *past* error and continues to grow until the controller reaches  $u_{ss}$  for  $e_{ss} = 0$ . Lastly, the derivative term has a correction effect based on the *future* trend of the tracking error, i.e. its rate of change, and thus can improve the speed of response and reduce the damping.

A common alternative form for PID control design is the **standard form**

$$K_{PID}(s) = \frac{u(s)}{e(s)} = K_p \left( 1 + \frac{1}{\tau_i s} + \tau_d s \right) \quad (2.161)$$

where  $\tau_i = \frac{K_p}{K_i}$  is the **integral time** and  $\tau_d = \frac{K_d}{K_p}$  is the **derivative time**. However, for PID control design via loop-shaping, one typically considers the **series form**, also known as the **interacting form**, of the PID controller, i.e.

$$K_{PID}(s) = k \left( \frac{s + \omega_i}{s} \right) \left( \frac{1}{\omega_d} s + 1 \right) \quad (2.162)$$

where  $k$  is the overall gain,  $\omega_i$  is the integral frequency, and  $\omega_d$  is the derivative frequency. This form demonstrates PID control as a particular loop-shaping controller employing three control stages: a proportional gain of  $k$ , an integral boost at  $\omega_i$ , and a derivative stage at  $\omega_d$ . These three stages can be used to shape  $L(s)$  by choosing  $k$  for the loop bandwidth,  $\omega_i$  for the low frequency performance requirements, and  $\omega_d$  for the stability requirements. It should be noted that  $k$  is typically limited by the control effort requirements.

Furthermore, the PID series form design parameters can be related to the PID standard form design parameters by

$$K_p = k \left( 1 + \frac{\omega_i}{\omega_d} \right) \quad (2.163)$$

$$\tau_i = \frac{1}{\omega_i} + \frac{1}{\omega_d} \quad (2.164)$$

and

$$\tau_d = \frac{1}{\omega_i + \omega_d} \quad (2.165)$$

Noting the derivative boost stage, one can see that the derivative term increases the asymptotic slope of the open-loop transfer function,  $L(s)$ , which will amplify any high frequency sensor noise in  $y_m$  in the classical feedback control system. To overcome this unfavorable outcome, often the derivative term in PID control is alternatively implemented using a **filtered derivative**. This proportional-integral-derivative-filter (PIDF) controller can be defined in **PIDF parallel form** as

$$K_{PIDF}(s) = \frac{u(s)}{e(s)} = K_p + \frac{K_i}{s} + K_d \frac{s}{\frac{1}{\omega_\infty} s + 1} \quad (2.166)$$

or, alternatively, in **PIDF standard form**

$$K_{PIDF}(s) = K_p \left( 1 + \frac{1}{\tau_i s} + \tau_d \frac{s}{\frac{1}{\omega_\infty} s + 1} \right) \quad (2.167)$$

where  $\omega_\infty$  is the bandwidth of the “filtered” derivative, i.e. as  $\omega_\infty \rightarrow \infty$ , the transfer function  $\frac{s}{\frac{1}{\omega_\infty} s + 1}$  approaches a pure differentiator. Thus, one can rewrite this type of PID controller in **PIDF series form** as

$$K_{PIDF}(s) = k \left( \frac{s + \omega_i}{s} \right) \left( \frac{\beta^2}{\omega_\infty} s + 1 \right) \left( \frac{\omega_\infty}{s + \omega_\infty} \right) \quad (2.168)$$

where the series form design parameters can be related to the standard form design parameters by

$$K_p = k \frac{(\beta^2 - 1)\omega_i}{\omega_\infty} + 1 \quad (2.169)$$

$$\tau_i = \frac{\beta^2 - 1}{\omega_\infty} + \frac{1}{\omega_i} \quad (2.170)$$

and

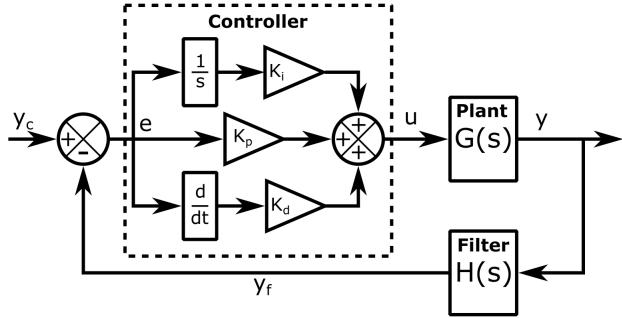
$$\tau_d = \frac{\beta^2}{((\beta^2 - 1)\omega_i + \omega_\infty)} - \frac{1}{\omega_\infty} \quad (2.171)$$

Equivalently, by the substitution  $\omega_d = \omega_\infty/\beta^2$ , one has

$$K_{PIDF}(s) = k \left( \frac{s + \omega_i}{s} \right) \left( \frac{1}{\omega_d} s + 1 \right) \left( \frac{\omega_\infty}{s + \omega_\infty} \right) \quad (2.172)$$

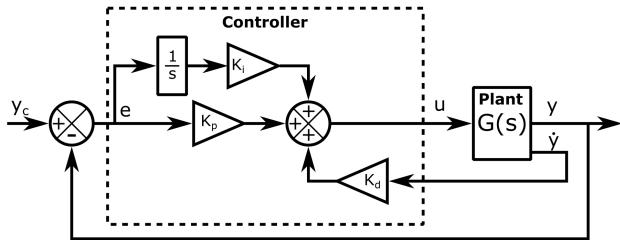
This form demonstrates filtered PID control has a particular loop-shaping controller employing four control stages: a proportional with  $k$ , an integral at  $\omega_i$ , a derivative stage at  $\omega_d$ , and a first-order filter at  $\omega_\infty$ . Thus, these stages allow one to use the loop-shaping design procedure to iterate on the values of these control stages to affect the low-frequency, crossover, and high-frequency regions of the  $L(j\omega)$ , while also implementing them as classical parallel PID control gains.

It should be noted that while PID controllers use zero or one low-pass filter control stages, one may be required to use additional filter stages in addition to the PID controller, though often these are implemented as part of the filter subsystem for  $y_f$  from  $y_m$  as shown in the following block diagram,



Note that one can also limit the update rate, i.e. frequency, of  $y_c$ , due to control signal sampling limitations or previous filtering.

As another alternative, one often can measure the output derivative directly instead of computing the derivative of the tracking error signal, so often the derivative term in PID control is alternatively implemented using **rate feedback control** where the derivative of the output, i.e. “rate,” is fed back. Thus, **PI control with rate feedback** would implement a system



This may be understood as a SIMO feedback control system as two separate outputs are fed back to the controller, the error and the output rate. It may also be considered as a **stability augmentation system (SAS)** using the rate and a second closed-loop PI controller using the output error.

### Ziegler-Nichols PID Tuning Method

The **Ziegler-Nichols (Z-N) PID tuning method** was developed by John Ziegler and Nathaniel Nichols as a simple heuristic method of tuning the gains of a PID-esque controllers which can be applied for any SISO system to maximize the disturbance rejection capabilities of PID. The Z-N PID tuning method works well when you have an analog controller for a stable LTI system whose response is dominated by a single stable mode. Though actual plants are unlikely to have only a first-order pole, this approximation is reasonable to describe the frequency response rolloff in many instances. Higher-order poles will introduce an extra phase shift, especially if the dominate mode is a second-order underdamped mode. Thus, the loop-shaping approach is typically the most flexible for LTI system design. These have also been adjusted by different control system designers into a few different rule sets. These methods are performed by first setting the  $K_i$  and  $K_d$  gains to zero, then increasing  $K_p$  from zero to the *ultimate gain*,  $K_p = K_u$ , which is defined as the

gain for which the output of the feedback control system has stable and consistent oscillations of period,  $T_u$ , or as the inverse of the gain margin for an LTI plant.

Then, one can set the P, PI, PID gains according to the following table for either the classic or standard forms.

Controller Type	$K_p$	$\tau_i$	$\tau_d$	$K_i$	$K_d$
P	$0.5K_u$	-	-	-	-
PI	$0.45K_u$	$0.8\bar{3}T_u$	-	$0.54K_u/T_u$	-
PD	$0.8K_u$	-	$0.125T_u$	-	$0.06K_u/T_u$
Classic PID	$0.6K_u$	$0.5T_u$	$0.125T_u$	$1.2K_u/T_u$	$0.06K_u/T_u$
Pessen Integral Rule PID	$0.7K_u$	$0.4T_u$	$0.15T_u$	$1.75K_u/T_u$	$0.06K_u/T_u$
Some Overshoot PID	$0.3\bar{K}_u$	$0.5T_u$	$0.3T_u$	$0.6K_u/T_u$	$0.1\bar{K}_u/T_u$
No Overshoot PID	$0.2K_u$	$0.5T_u$	$0.3T_u$	$0.4K_u/T_u$	$0.06K_u/T_u$

## References

For more information, please refer to the following

- Schmidt, D. K., “12.1 Feedback Control-Law Synthesis Via Loop Shaping - A Just-In-Time Tutorial\*,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 548-562
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “3.9 Feedback Control,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 213-240

## References

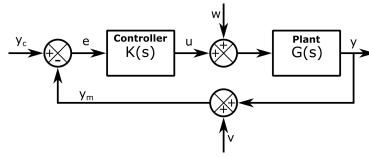
For more information, please refer to the following

- McCormack, A. S., and Godfrey, K. R., “Rule-Based Autotuning Based on Frequency Domain Identification,” *IEEE Transactions on Control Systems Technology*, vol, 6 no. 1, January 1998
- Nelson, R. C., “7.5 Time-Domain and Frequency-Domain Specifications,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 251-262
- Nelson, R. C., “7.7 Control System Design,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997
- Nelson, R. C., “7.8 PID Controller,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997
- Nise, N. S., “Chapter 10: Frequency Response Techniques,” in *Control Systems Engineering*, 8th ed., John Wiley & Sons, 2019
- Nise, N. S., “Chapter 11: Design via Frequency Response,” in *Control Systems Engineering*, 8th ed., John Wiley & Sons, 2019

- Schmidt, D. K., “12.1 Feedback Control-Law Synthesis Via Loop Shaping - A Just-In-Time Tutorial\*,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 548-562
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “3.9 Feedback Control,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 213-240

## 2.4 Classical Feedback Control System Performance Constraints

LTI feedback control systems must satisfy robustness requirements for stability *as well as* performance requirements. Thus, it is important to be aware of fundamental limits on the performance of the control system. To this end, recall the block diagram for a SISO LTI feedback control system.



Also, recall the error/sensitivity transfer function can be defined as

$$e(s) = S(s)y_c(s) \quad (2.173)$$

and the closed-loop/complementary sensitivity transfer function is defined as

$$y(s) = T(s)y_c(s) \quad (2.174)$$

where

$$S(s) + T(s) = 1 \quad \forall s \quad (2.175)$$

### Performance Limits from Sensitivity Tradeoff

One can deduce performance limits on the closed-loop response in the frequency domain by analyzing the Bode sensitivity integral theorem whose proof ultimately relies on Cauchy’s integral theorem which is a result from complex analysis. This theorem states that if  $1 + L(s)$  has no zeros in the RHP and  $L(s)$  has relative degree  $\geq 2$ , i.e. the denominator is at least two orders higher than the numerator, then

$$\int_0^\infty \log_{10} |S(j\omega)| d\omega = \pi \sum_{i=1}^{N_p} \operatorname{Re}\{p_i\} \quad (2.176)$$

where  $p_i$  for  $i = 1, \dots, N_p$  are the  $N_p$  open RHP poles of  $L(s)$ . Furthermore, if  $L(s)$  has no poles in the RHP, then one has

$$\int_0^\infty \log_{10} |S(j\omega)| d\omega = 0 \quad (2.177)$$

which illustrates the **sensitivity tradeoff** that any decrease in  $S(j\omega)$  must be balanced by an increase in  $S(j\omega)$ , i.e. the area under the curve of the Bode plot of  $S(j\omega)$  must be conserved.

Note that if there exist poles in the RHP, the area of sensitivity increase will exceed that of the sensitivity decrease. Furthermore, though this result does not depend on RHP zeros of  $G(s)$ , there are related results based on Poisson integrals which demonstrate that RHP zeros will further degrade this sensitivity tradeoff. In addition, note that the restriction for  $L(s) = G(s)K(s)$  to have relative degree  $\geq 2$  is essentially always satisfied in practice.  $G(s)$  will be strictly proper, i.e. its denominator will have higher order than its numerator, as all real plants eventually rolloff at high frequencies. Similarly,  $K(s)$  will also be strictly proper since it will be implemented on hardware whose dynamics rolloff at high frequencies.

An important part of the control design process is to set  $|S(j\omega)| \ll 1$  for low frequencies. This can be more formally stated as some limit on  $\log_{10} |S(j\omega)| \leq \alpha \ll 0$  for all  $\omega \leq \omega_L$  for some  $\omega_L$ . However, to push down  $\log_{10} |S(j\omega)|$ , by the Bode sensitivity integral, one has to increase the area above 0 dB. Thus, one may think the best approach is to shape the entire area of  $|S(j\omega)|$  to have some small amount over all frequencies,  $\omega > \omega_L$ . However, in practice, this is not possible as one does not have an accurate model of  $G(j\omega)$  at high frequencies and must be robust to this consideration. More formally stated, the control design process typically has as some upper limit on the **available bandwidth**,  $\omega_U$ , where  $|L(j\omega)|$  rolls off rapidly for  $\omega \geq \omega_U$ , i.e.

$$|L(j\omega)| \leq \epsilon \frac{\omega_U^2}{\omega} \quad \text{for all } \omega \geq \omega_U \quad (2.178)$$

for some  $\epsilon < 0.5$ . This practical consideration of  $\omega_U$  allows one to split the Bode sensitivity integral for no RHP poles into three pieces

$$\int_0^{\omega_L} \log_{10} |S(j\omega)| d\omega + \int_{\omega_L}^{\omega_U} \log_{10} |S(j\omega)| d\omega + \int_{\omega_U}^{\infty} \log_{10} |S(j\omega)| d\omega = 0 \quad (2.179)$$

or

$$-\left( \int_0^{\omega_L} \log_{10} |S(j\omega)| d\omega + \int_{\omega_U}^{\infty} \log_{10} |S(j\omega)| d\omega \right) = \int_{\omega_L}^{\omega_U} \log_{10} |S(j\omega)| d\omega \quad (2.180)$$

Then, assuming  $L(s)$  is stable and has relative degree  $\geq 2$  where each can be bounded by

$$-\int_0^{\omega_L} \log_{10} |S(j\omega)| d\omega \geq \omega_L \log_{10} \frac{1}{\alpha} \quad (2.181)$$

$$-\int_{\omega_U}^{\infty} \log_{10} |S(j\omega)| \geq -\frac{3\epsilon}{2} \omega_c \quad (2.182)$$

$$\int_{\omega_L}^{\omega_U} \log_{10} |S(j\omega)| d\omega \leq (\omega_U - \omega_L) \left( \max_{\omega \in [\omega_L, \omega_U]} \log_{10} |S(j\omega)| \right) \quad (2.183)$$

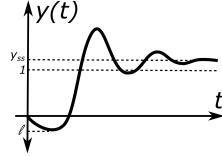
leads to the following

$$\max_{\omega \in [\omega_L, \omega_U]} \log_{10} |S(j\omega)| \geq \frac{\omega_L}{\omega_U - \omega_L} \log_{10} \frac{1}{\alpha} - \left( \frac{3\epsilon}{2} \right) \frac{\omega_c}{\omega_U - \omega_L} \quad (2.184)$$

which is also known as the **waterbed effect** which derives from the tradeoff between “pushing down” on  $S(s)$  causing it to “rise up” somewhere between  $\omega_L \leq \omega \leq \omega_U$  with the peak value,  $\max |S(s)| \rightarrow \infty$  as  $\alpha \rightarrow 0$ . Note that this formulation did not include RHP poles and zeros which will make the waterbed effect even worse.

### Performance Limit Due to Real RHP Zero

Assume that  $G(s)$  has a real RHP zero at  $s = z$ . This will cause the feedback control system response,  $y(t)$ , to initially undershoot by some amount,  $\ell$ , which can be visualized as



and defined as

$$\ell = \min_{t \geq 0} y(t) \quad (2.185)$$

If there is a zero,  $z$ , in the closed RHP such that  $G(z) = 0$ , then one has  $G(z)K(z) = 0$  as there can be no pole/zero cancellations, thus the closed-loop transfer function is then

$$T(z) = \frac{G(z)K(z)}{1 + G(z)K(z)} = 0 \quad (2.186)$$

Thus, the unit step response has a transfer function at  $s = z$  that satisfies

$$y(z) = T(z)y_c(z) = 0 \quad (2.187)$$

From the definition of the Laplace transform, one has

$$\int_0^\infty e^{-zt} y(t) dt = 0 \quad (2.188)$$

which can be split into two pieces

$$\int_0^{t_{s,w}} e^{-zt} y(t) dt + \int_{t_{s,w}}^\infty e^{-zt} y(t) dt = 0 \quad (2.189)$$

where the settling time,  $t_{s,w}$ , to some percentage  $w$  for the unit step response is defined as

$$t_{s,w} = \min t \quad \text{such that} \quad |y(t) - y_{ss}| \leq 0.05y_{ss} \quad t \geq t_{s,w} \quad (2.190)$$

where  $y_{ss}$  is the steady-state output, i.e.  $y(t)$  as  $t \rightarrow \infty$ .

Thus,

$$\int_{t_{s,w}}^\infty e^{-zt} y(t) dt = - \int_0^{t_{s,w}} e^{-zt} y(t) dt \quad (2.191)$$

By the definition of the minimum of  $y(t)$ , one has

$$-\int_0^{t_{s,w}} e^{-zt} y(t) dt \geq -\ell \int_0^{t_{s,w}} e^{-zt} dt \quad (2.192)$$

or

$$\ell \leq \frac{-\int_{t_{s,w}}^{\infty} e^{-zt} y(t) dt}{\int_0^{t_{s,w}} e^{-zt} dt} \quad (2.193)$$

For  $t > t_{s,w}$ , one has  $y(t) \geq (1-w)y_{ss}$  or  $-y(t) \leq -(1-w)y_{ss}$  and one can bound the undershoot by

$$\ell \leq \frac{-(1-w) \int_{t_{s,w}}^{\infty} e^{-zt} dt}{\int_0^{t_{s,w}} e^{-zt} dt} \quad (2.194)$$

which by integration for a real-valued  $z$ , one has

$$\ell \leq \frac{-(1-w)[\frac{1}{z}(e^{-zt_{s,w}})]}{[\frac{1}{z}(1 - e^{-zt_{s,w}})]} \quad (2.195)$$

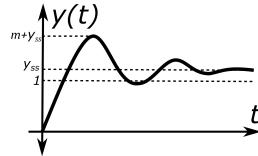
or

$$\ell \leq \frac{-(1-w)}{e^{zt_{s,w}} - 1} \quad (2.196)$$

Thus, for a fixed zero,  $z$ , for the plant,  $G(s)$ , as one changes  $K(s)$  to decrease the settling time, one must get a larger amount of undershoot, i.e. as  $t_{s,w} \rightarrow 0$ ,  $\ell \rightarrow -\infty$ , or conversely, to decrease the undershoot, one must increase the settling time,  $t_{s,w}$ . In general, a RHP zero, even if complex, places an upper bound on the bandwidth of the classical feedback control system.

### Performance Limit Due to Real RHP Pole

Assume that  $G(s)$  has a real RHP pole at  $s = p$ . This will cause the feedback control system response,  $y(t)$ , to have some overshoot by some amount,  $m$ , which can be visualized as



and defined as

$$m = \max_{t \geq 0} y(t) - y_{ss} \quad (2.197)$$

where  $y_{ss} = 1$  for simplicity.

If there is a pole,  $p$ , in the closed RHP such that  $G(p) = \infty$ , then  $G(p)K(p) = \infty$  as there can be no pole/zero cancellations, thus

$$S(p) = \frac{1}{1 + G(p)K(p)} = 0 \quad (2.198)$$

Thus, the tracking error has a transfer function at  $s = p$  that satisfies

$$e(p) = S(p)y_c(p) = 0 \quad (2.199)$$

From the definition of the Laplace transform, one has

$$\int_0^\infty e^{-pt} e(t) dt = 0 \quad (2.200)$$

which can be split into two pieces

$$\int_0^{t_{r,v}} e^{-pt} e(t) dt + \int_{t_{r,v}}^\infty e^{-pt} e(t) dt = 0 \quad (2.201)$$

where the rise time,  $t_{r,v}$ , for the unit step response is defined as

$$t_{r,v} = \min t \text{ such that } y(t) \leq vy_{ss} \text{ for all } t \leq t_{r,v} \quad (2.202)$$

Thus, as  $e(t) = 1 - y(t)$  for the unit step, one has

$$\int_{t_{r,v}}^\infty e^{-pt} e(t) dt = - \int_0^{t_{r,v}} e^{-pt} e(t) dt \quad (2.203)$$

By the definition of the overshoot for  $e(t)$ , one has

$$\int_{t_{r,v}}^\infty e^{-pt} e(t) dt \geq (1 - m - y_{ss}) \int_{t_{r,v}}^\infty e^{-pt} dt = -m \left[ \frac{1}{p} e^{-pt_{r,v}} \right] \quad (2.204)$$

and by the definition of the rise time for  $e(t)$

$$- \int_0^{t_{r,v}} e^{-pt} e(t) dt \leq -(1 - v) \int_0^{t_{r,v}} e^{-pt} dt = -(1 - v) \left[ \frac{1}{p} (1 - e^{-pt_{r,v}}) \right] \quad (2.205)$$

Combining these two bounds, one has

$$-m \left[ \frac{1}{p} e^{-pt_{r,v}} \right] \leq -(1 - v) \left[ \frac{1}{p} (1 - e^{-pt_{r,v}}) \right] \quad (2.206)$$

or

$$m \geq (1 - v)(e^{-pt_{r,v}} - 1) \quad (2.207)$$

Thus, for a fixed pole,  $p$ , for the plant,  $G(s)$ , as one changes  $K(s)$  to increase the rise time, one must get a larger amount of overshoot, i.e. as  $t_{r,v} \rightarrow \infty$ ,  $m \rightarrow \infty$ , or conversely, to decrease the overshoot, one must decrease the rise time,  $t_{r,v}$ . In general, a RHP pole places a lower bound on the bandwidth of the classical feedback control system. However, it should be noted that this limit assumes that the controller only receives the error signal  $e(t)$ . However, this can be avoided for MIMO inputs to the controller.

### Performance Limits Due to Real RHP Zero and Pole

Assume that  $G(s)$  has both a real RHP zero at  $s = z$  and a real RHP pole at  $s = p$ . This will cause the feedback control system response,  $y(t)$ , to have some undershoot by some amount  $\ell$ , and some overshoot by some amount,  $m$ . In this case, the unit step response has a transfer function at  $s = z$  that satisfies

$$y(z) = T(z)y_c(z) = 0 \quad (2.208)$$

and the tracking error has a transfer function at  $s = p$  that satisfies

$$e(p) = S(p)y_c(p) = 0 \quad (2.209)$$

From the definition of the Laplace transform, one has

$$\int_0^\infty e^{-zt}y(t)dt = 0 \quad (2.210)$$

and

$$\int_0^\infty e^{-pt}e(t)dt = 0 \quad (2.211)$$

Furthermore, as  $e(t) = 1 - y(t)$ , one has

$$\int_0^\infty e^{-zt}(1 - e(t))dt = 0 \quad (2.212)$$

and

$$\int_0^\infty e^{-pt}(1 - y(t))dt = 0 \quad (2.213)$$

Next, by rearranging, one has

$$\int_0^\infty e^{-zt}dt = \int_0^\infty e^{-zt}e(t)dt \quad (2.214)$$

and

$$\int_0^\infty e^{-pt}dt = \int_0^\infty e^{-pt}y(t)dt \quad (2.215)$$

and computing the integrals on the left sides, one has

$$\frac{1}{z} = \int_0^\infty e^{-zt}e(t)dt \quad (2.216)$$

and

$$\frac{1}{p} = \int_0^\infty e^{-pt}y(t)dt \quad (2.217)$$

Next, subtracting these from the original  $e(p)$  and  $y(z)$  integrals, one has

$$\int_0^\infty e^{-zt}y(t)dt - \int_0^\infty e^{-pt}y(t)dt = 0 - \frac{1}{p} \quad (2.218)$$

and

$$\int_0^\infty e^{-pt}e(t)dt - \int_0^\infty e^{-zt}e(t)dt = 0 - \frac{1}{z} \quad (2.219)$$

or

$$\int_0^\infty (e^{-zt} - e^{-pt})y(t)dt = -\frac{1}{p} \quad (2.220)$$

and

$$\int_0^\infty (e^{-pt} - e^{-zt})e(t)dt = -\frac{1}{z} \quad (2.221)$$

By the definition of the minimum of  $y(t)$ , one has

$$\ell \int_0^\infty e^{-zt} - e^{-pt} dt \leq \int_0^\infty (e^{-zt} - e^{-pt}) y(t) dt = -\frac{1}{p} \quad (2.222)$$

and the definition of the overshoot for  $e(t)$ , one has

$$-m \int_0^\infty e^{-zt} - e^{-pt} dt \leq \int_0^\infty (e^{-zt} - e^{-pt}) e(t) dt = -\frac{1}{z} \quad (2.223)$$

Calculating the integrals, one has

$$\ell \frac{p-z}{zp} \leq -\frac{1}{p} \quad (2.224)$$

and

$$-m \frac{z-p}{zp} \leq -\frac{1}{z} \quad (2.225)$$

Thus, if  $p > z$ , then dividing by  $\frac{p-z}{zp}$ , one has

$$\ell \leq -\frac{z}{p-z} \quad (2.226)$$

and if  $z > p$ , then dividing by  $-\frac{z-p}{zp}$ , one has

$$m \geq \frac{p}{z-p} \quad (2.227)$$

In either case, it should be noted that if  $p$  is near to  $z$ , then the denominator is close to zero and the bound has large magnitude. This implies that plants with “close” RHP poles/zeros are difficult to control.

This analysis can be extended to the sensitivity transfer function. First, one must know the **maximum modulus theorem** which states for a stable transfer function,  $G(s)$ , the peak magnitude over the imaginary axis is equal to the peak magnitude over the entire RHP, i.e.

$$\max_{\omega \in \mathbb{R}} \|G(j\omega)\| = \max_{s \in \mathbb{C}, \operatorname{Re}(s) \geq 0} \|G(s)\| \quad (2.228)$$

Suppose the classical feedback system is stable and  $G(s)$  has a real RHP zero at  $s = z$  and a real RHP pole at  $s = p$ , then one has  $S(p) = 0$  and  $S(z) = 1$ . Furthermore, consider factoring  $S(s)$  as

$$S(s) = \frac{s-p}{s+z} \tilde{S}(s) \quad (2.229)$$

where  $\tilde{S}(s)$  is stable and one can write

$$\max_{\omega} |S(j\omega)| = \max_{\omega} |j\omega - p j\omega + p \tilde{S}(j\omega)| = \max_{\omega} |\tilde{S}(j\omega)| \quad (2.230)$$

as

$$|j\omega - p j\omega + p| = 1 \quad \forall \omega \quad (2.231)$$

Then, by the maximum modulus theorem

$$\max_{\omega} |\tilde{S}(j\omega)| \geq |\tilde{S}(z)| \quad (2.232)$$

Finally, note that

$$1 = S(z) = \frac{z - p}{z + p} \tilde{S}(z) \quad (2.233)$$

or

$$\tilde{S}(z) = \frac{z + p}{z - p} \quad (2.234)$$

Thus, one can state

$$\max_{\omega} |S(j\omega)| \geq \left| \frac{z+p}{z-p} \right| \quad (2.235)$$

Then, recall that  $\|S(j\omega)\|$  is inversely proportional to the distance between  $L(s)$  and the critical  $s = -1$  point for any  $K(s)$ , i.e. the disk margin. Recalling

$$d_{min} = \frac{1}{\max_{\omega} |S(j\omega)|} \quad (2.236)$$

one can also state

$$d_{min} \leq \left| \frac{z-p}{z+p} \right| \quad (2.237)$$

Thus, this result also demonstrates that plants with “close” RHP poles/zeros are difficult to achieve good robustness.

## References

For more information, please refer to the following

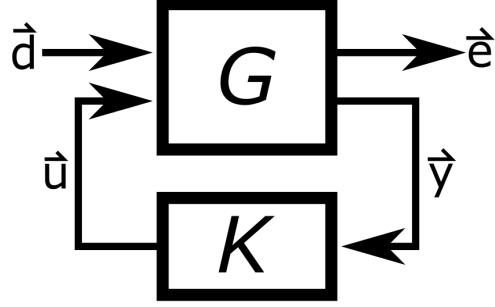
- Skogestad, S., and Postlethwaite, I., “5 Limitations on Performance in SISO Systems,” *Multivariable Feedback Control: Analysis and Design*, 1st ed., Vol. 1, John Wiley & Sons, Chichester, England, 1996, pp. 159-212

## 2.5 MIMO LTI Feedback Control System Analysis

When one can use MIMO LTI control systems, one may use a combination of feedforward terms, i.e. control gains dependent on  $\vec{r}$ , tracking error terms, i.e. control gains dependent on  $\vec{e}$ , output terms, i.e.  $\vec{y}$ , other terms, e.g. integrated signals and derivative signals as well as coupling between all of these. Thus, for studying MIMO LTI systems, it is useful to define a generalized framework to assess system properties including stability, controllability, observability, and robustness.

## Generalized LTI Feedback Control System

Consider the generalized LTI feedback control system architecture



which is denoted by  $F_L(G, K)$  and is a linear fractional transformation (LFT). The vector input,  $\vec{d} \in \mathbb{R}^{n_d}$ , to the generalized plant is known as the **generalized disturbance** which typically includes reference commands,  $\vec{r}$ , process noise,  $\vec{w}$ , and measurement noise,  $\vec{v}$ . The vector output,  $\vec{e} \in \mathbb{R}^{n_e}$ , from the generalized plant is known as the **generalized error** which typically includes at least the weighted tracking error and the weighted control effort. The other vectors are the **generalized control input** vector  $\vec{u} \in \mathbb{R}^{n_u}$  and the **generalized output**,  $\vec{y} \in \mathbb{R}^{n_y}$ .

In this context, the **generalized plant**,  $G$ , contains the original plant dynamics, actuators, weighting filters, and signal routing and operations on the disturbance, error, state, inputs, and output signals. Thus, the generalized output may be the tracking error instead of the “true” plant output or it may contain the reference command and the plant output. Using LFTs, the LTI state-space model for the generalized plant,  $G$ , can be written as

$$\begin{aligned}\dot{\vec{x}}(t) &= A\vec{x}(t) + [B_1 \quad B_2] \begin{bmatrix} \vec{d}(t) \\ \vec{u}(t) \end{bmatrix} \\ \begin{bmatrix} \vec{e}(t) \\ \vec{y}(t) \end{bmatrix} &= \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \vec{x}(t) + \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} \vec{d}(t) \\ \vec{u}(t) \end{bmatrix}\end{aligned}\tag{2.238}$$

with the **generalized LTI feedback controller**,  $K$ , designed as

$$\begin{aligned}\dot{\vec{x}}_K &= A_K \vec{x}_K + B_K \vec{y} \\ \vec{u} &= C_K \vec{x}_K + D_K \vec{y}\end{aligned}\tag{2.239}$$

Combining, one has for the input and output vectors

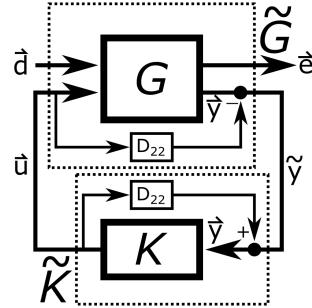
$$\begin{bmatrix} I & -D_K \\ -D_{22} & I \end{bmatrix} \begin{bmatrix} \vec{u}(t) \\ \vec{y}(t) \end{bmatrix} = \begin{bmatrix} 0 & C_K \\ C_2 & 0 \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \vec{x}_K(t) \end{bmatrix} + \begin{bmatrix} 0 \\ D_{21} \end{bmatrix} [\vec{d}(t)]\tag{2.240}$$

thus, the interconnection of  $G$  and  $K$  is well-posed if and only if  $(I - D_{22}D_K)^{-1}$  exists, i.e.  $\vec{u}, \vec{y}$  can be substituted into the state equation

$$\begin{bmatrix} \dot{\vec{x}}(t) \\ \dot{\vec{x}}_K(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & A_K \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \vec{x}_K(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \vec{d}(t) + \begin{bmatrix} B_2 & 0 \\ 0 & B_K \end{bmatrix} \begin{bmatrix} \vec{u}(t) \\ \vec{y}(t) \end{bmatrix}\tag{2.241}$$

and a unique solution can be found.

Note that in control synthesis,  $D_{22}$  is set to 0 without loss of generality as one can alternatively use **loop-shifting** which can be visualized as the following alteration to  $F_L(G, K)$  as



to form  $F_L(\tilde{G}, \tilde{K})$  from  $\tilde{G}$  and  $\tilde{K}$  and then undo the loop-shifting to get the original  $K$ . Thus, one typically assumes the generalized LTI plant,  $G$ , can be specified as

$$\begin{aligned} \dot{\vec{x}}(t) &= A\vec{x}(t) + [B_1 \quad B_2] \begin{bmatrix} \vec{d}(t) \\ \vec{u}(t) \end{bmatrix} \\ \begin{bmatrix} \vec{e}(t) \\ \vec{y}(t) \end{bmatrix} &= \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \vec{x}(t) + \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & 0 \end{bmatrix} \begin{bmatrix} \vec{d}(t) \\ \vec{u}(t) \end{bmatrix} \end{aligned} \quad (2.242)$$

In this case, the state equation for the closed-loop system can be rewritten as

$$\begin{bmatrix} \dot{\vec{x}}_K(t) \\ \vec{x}_K(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & A_K \end{bmatrix} \begin{bmatrix} \vec{x}_K(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \vec{d}(t) + \begin{bmatrix} B_2 & 0 \\ 0 & B_K \end{bmatrix} \begin{bmatrix} \vec{u}(t) \\ \vec{y}(t) \end{bmatrix} \quad (2.243)$$

with the following relationship for  $\vec{u}(t)$  and  $\vec{y}(t)$

$$\begin{bmatrix} I & -D_K \\ 0 & I \end{bmatrix} \begin{bmatrix} \vec{u}(t) \\ \vec{y}(t) \end{bmatrix} = \begin{bmatrix} 0 & C_K \\ C_2 & 0 \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \vec{x}_K(t) \end{bmatrix} + \begin{bmatrix} 0 \\ D_{21} \end{bmatrix} \begin{bmatrix} \vec{d}(t) \end{bmatrix} \quad (2.244)$$

$$\begin{bmatrix} \vec{u}(t) \\ \vec{y}(t) \end{bmatrix} = \begin{bmatrix} I & D_K \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 & C_K \\ C_2 & 0 \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \vec{x}_K(t) \end{bmatrix} + \begin{bmatrix} I & D_K \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 \\ D_{21} \end{bmatrix} \begin{bmatrix} \vec{d}(t) \end{bmatrix} \quad (2.245)$$

$$\begin{bmatrix} \vec{u}(t) \\ \vec{y}(t) \end{bmatrix} = \begin{bmatrix} D_K C_2 & C_K \\ C_2 & 0 \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \vec{x}_K(t) \end{bmatrix} + \begin{bmatrix} D_K D_{21} \\ D_{21} \end{bmatrix} \begin{bmatrix} \vec{d}(t) \end{bmatrix} \quad (2.246)$$

Then, by substitution, one can obtain the closed-loop LTI state-space system model as

$$\begin{aligned} \begin{bmatrix} \dot{\vec{x}}(t) \\ \dot{\vec{x}}_K(t) \end{bmatrix} &= \begin{bmatrix} A + B_2 D_K C_2 & B_2 C_K \\ B_K C_2 & A_K \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \vec{x}_K(t) \end{bmatrix} + \begin{bmatrix} B_1 + B_2 D_K D_{21} \\ B_K D_{21} \end{bmatrix} \vec{d}(t) \\ \vec{e}(t) &= [C_1 + D_{12} D_K C_2 \quad D_{12} C_K] \begin{bmatrix} \vec{x}(t) \\ \vec{x}_K(t) \end{bmatrix} + (D_{11} + D_{12} D_K D_{21}) \vec{d}(t) \end{aligned} \quad (2.247)$$

with the definitions for the closed-loop state matrix,  $A_L$ , as

$$A_L = \begin{bmatrix} A + B_2 D_K C_2 & B_2 C_K \\ B_K C_2 & A_K \end{bmatrix} \quad (2.248)$$

the closed-loop input matrix,  $B_L$ , as

$$B_L = \begin{bmatrix} B_1 + B_2 D_K D_{21} \\ B_K D_{21} \end{bmatrix} \quad (2.249)$$

the closed-loop output matrix,  $C_L$ , as

$$C_L = [C_1 + D_{12} D_K C_2 \quad D_{12} C_K] \quad (2.250)$$

and the closed-loop feedthrough matrix,  $D_L$ , as

$$D_L = D_{11} + D_{12} D_K D_{21} \quad (2.251)$$

which for stability of the generalized feedback control system, one requires that  $A_L$  is stable.

### State and Output Feedback Control

If  $A_K = C_K = B_K = 0$ , one has **static-controller feedback control**, a type of **fixed-gain controller**. The general case is also known as a.k.a. **output feedback control**, where one has an output feedback controller as

$$\vec{u}(t) = D_K \vec{y}(t) \quad (2.252)$$

which results in a simplified closed-loop LTI state-space model

$$\begin{aligned} \dot{\vec{x}}(t) &= A_L \vec{x}(t) + B_L \vec{d}(t) \\ \vec{e}(t) &= C_L \vec{x}(t) + D_L \vec{d}(t) \end{aligned} \quad (2.253)$$

which results in a closed-loop state matrix,  $A_L$ , as

$$A_L = A + B_2 D_K C_{12} \quad (2.254)$$

a closed-loop input matrix,  $B_L$ , as

$$B_L = B_1 + B_2 D_K D_{21} \quad (2.255)$$

a closed-loop output matrix,  $C_L$ , as

$$C_L = C_1 + D_{12} D_K C_2 \quad (2.256)$$

and a closed-loop feedthrough matrix,  $D_L$ , as

$$D_L = D_{11} + D_{12} D_K D_{21} \quad (2.257)$$

Thus, an output feedback control system is stable if and only if  $A + B_2 D_K C_{12}$  is stable.

A special case of fixed-gain feedback control is **state feedback control**, i.e.  $C_2 = I$ ,  $D_{21} = 0$ , and state feedback controller

$$\vec{u}(t) = D_K \vec{x}(t) \quad (2.258)$$

with closed-loop state matrix,  $A_L$ , is

$$A_L = A + B_2 D_K \quad (2.259)$$

the closed-loop input matrix,  $B_L$ , is

$$B_L = B_1 \quad (2.260)$$

the closed-loop output matrix,  $C_L$ , as

$$C_L = C_1 + D_{12} D_K \quad (2.261)$$

and the closed-loop feedthrough matrix,  $D_L$ , as

$$D_L = D_{11} \quad (2.262)$$

Thus, a state feedback control system is stable if and only if  $A + B_2 D_K$  is stable.

### Observer Feedback Control

The general case of dynamic-controller feedback control is also known as **observer feedback control**, one defines the controller state as the **state estimate**,  $\hat{x}$ , i.e.

$$\vec{x}_K = \hat{x} \quad (2.263)$$

where an observer is designed to form  $\hat{x}$  and uses this to form the control input as

$$\vec{u}(t) = -K \hat{x}(t) \quad (2.264)$$

Here, an **open-loop observer** could be formed based on the linear state-space model for continuous-time, assuming the disturbances  $\vec{d}$  are unknown, as

$$\dot{\hat{x}} = A \hat{x} + B_2 \vec{u} \quad (2.265)$$

However, for feedback control, one receives the output signal from the output equation, and forms a **closed-loop observer** for continuous-time as

$$\begin{aligned} \dot{\hat{x}}(t) &= A \hat{x}(t) + B_2 \vec{u}(t) + L \left( \vec{y}(t) - \hat{y}(t) \right) \\ \hat{y}(t) &= C_2 \hat{x}(t) \end{aligned} \quad (2.266)$$

where  $\hat{y}$  is the output estimate based on the output equation model and  $L$  is the **Luenberger observer matrix**. Thus, with  $\vec{u}(t) = -K \hat{x}(t)$ , one can form the continuous-time **observer feedback control system** as the generalized LTI feedback controller

$$\begin{aligned} \dot{\hat{x}}(t) &= (A - B_2 K - LC_2) \hat{x}(t) + L \vec{y}(t) \\ \vec{u}(t) &= -K \hat{x}(t) \end{aligned} \quad (2.267)$$

where

$$A_K = A - B_2 K - LC_2 \quad (2.268)$$

$$B_K = L \quad (2.269)$$

$$C_K = -K \quad (2.270)$$

$$D_K = 0 \quad (2.271)$$

Thus, one has the closed-loop dynamics

$$\begin{aligned} \begin{bmatrix} \dot{\vec{x}}(t) \\ \dot{\hat{x}}(t) \end{bmatrix} &= \begin{bmatrix} A & -B_2 K \\ LC_2 & A - B_2 K - LC_2 \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \hat{x}(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ LD_{21} \end{bmatrix} \vec{d}(t) \\ \vec{e}(t) &= [C_1 \ -D_{12} K] \begin{bmatrix} \vec{x}(t) \\ \hat{x}(t) \end{bmatrix} + D_{11} \vec{d}(t) \end{aligned} \quad (2.272)$$

However, to better assess the stability of the closed-loop system, consider the **state error**,  $\vec{e}_x(t)$ , defined as

$$\vec{e}_x(t) = \vec{x}(t) - \hat{x}(t) \quad (2.273)$$

Then, one has

$$\begin{aligned} \begin{bmatrix} \dot{\vec{x}}(t) \\ \dot{\vec{x}}(t) - \dot{\vec{e}}_x(t) \end{bmatrix} &= \begin{bmatrix} A & -B_2 K \\ LC_2 & A - B_2 K - LC_2 \end{bmatrix} \begin{bmatrix} \vec{x}(t) - \vec{e}_x(t) \\ \vec{x}(t) - \vec{e}_x(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ LD_{21} \end{bmatrix} \vec{d}(t) \\ \vec{e}(t) &= [C_1 \ -D_{12} K] \begin{bmatrix} \vec{x}(t) \\ \vec{x}(t) - \vec{e}_x(t) \end{bmatrix} + D_{11} \vec{d}(t) \end{aligned} \quad (2.274)$$

which can be rearranged as

$$\begin{aligned} \begin{bmatrix} \dot{\vec{x}}(t) \\ \dot{\vec{x}}(t) - \dot{\vec{e}}_x(t) \end{bmatrix} &= \begin{bmatrix} (A - B_2 K) & B_2 K \\ (A - B_2 K) & -A + B_2 K + LC_2 \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \vec{e}_x(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ LD_{21} \end{bmatrix} \vec{d}(t) \\ \vec{e}(t) &= [(C_1 - D_{12} K) \ D_{12} K] \begin{bmatrix} \vec{x}(t) \\ \vec{e}_x(t) \end{bmatrix} + D_{11} \vec{d}(t) \end{aligned} \quad (2.275)$$

And subtracting the second row of the state equation from the first row, one has

$$\begin{aligned} \begin{bmatrix} \dot{\vec{x}}(t) \\ \dot{\vec{e}}_x(t) \end{bmatrix} &= \begin{bmatrix} (A - B_2 K) & B_2 K \\ 0 & A - LC_2 \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \vec{e}_x(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ B_1 - LD_{21} \end{bmatrix} \vec{d}(t) \\ \vec{e}(t) &= [(C_1 - D_{12} K) \ D_{12} K] \begin{bmatrix} \vec{x}(t) \\ \vec{e}_x(t) \end{bmatrix} + D_{11} \vec{d}(t) \end{aligned} \quad (2.276)$$

Since the eigenvalues of an upper triangular matrix are dependent only on the block diagonal terms, an observer feedback control system is stable if and only if  $A - B_2 K$  and  $A - LC_2$  are stable. This fact of linear observer feedback control systems having independent criteria for the design of  $L$  and  $K$  is known as the **separation principle**, a foundational result in control and perception theory which justifies the separate use of control and perception algorithms in dynamical systems applications. although it only technically applies to linear systems. Thus, this part of the textbook focuses on LTI state feedback control design. The subsequent parts of the textbook consider optimal state estimation and its application to perception of a flight vehicle's "state."

With this analysis in mind, a straightforward linear controls design method is **eigenvalue placement**, also known as **pole placement** where one “places” the eigenvalues of the linear feedback control system in the complex plane by choosing the feedback gain matrix appropriately, thereby setting the modal characteristics for the system response. This method can be considered as a generalization of the root locus technique for selecting a single gain parameter in a SISO LTI control system to selecting two gain matrices in a MIMO LTI control systems and its effect on the system’s root/poles/eigenvalues for control and state estimation. It should be noted that there are computer algorithms to solve the eigenvalue placement problem, e.g. `place` in MATLAB, but derivations for these generic algorithms are beyond the scope of this textbook. However, for MIMO LTI systems, one may not be able to set the eigenvalues arbitrarily as the eigenvalue equations also depend on the generalized plant state, input, and potentially the output matrices. The ability to arbitrarily set the eigenvalues leads to the concepts of controllability and observability for LTI systems. To place the poles of state feedback, the system must be state controllable. To place the poles of output feedback, the system must be output controllable. To place the poles of observer feedback, the system must be state controllable to place  $A - B_2 K$  and observable to place  $A - LC_2$ .

### Controllability and Observability of Continuous-Time LTI Systems

**State controllability** is defined as the ability to control a system to *any* state using some finite input. Related to this is the concept of **state reachability** and can be stated as follows. For continuous-time dynamical systems, a state  $\vec{x}'$  is reachable if for every finite  $T > 0$ , there exists an input function  $u(t)$  with  $0 < t \leq T$  such that the state goes from  $x(0) = 0 \rightarrow x(T) = \vec{x}'$ . Thus, reachability is generally a slightly weaker notion than controllability. However, for linear state-space systems, the sequence for reaching any state can be inverted to return to zero from any initial conditions, thus state reachability is equivalent to state controllability. Using linear algebra, one can look at the conditions for state reachability which will imply state controllability for continuous-time LTI state-space systems.

Consider the state equation for a continuous-time LTI state-space system

$$\dot{\vec{x}}(t) = A\vec{x}(t) + B\vec{u}(t) \quad (2.277)$$

Given initial state  $\vec{x}(0) = 0$ , the general state-space solution is

$$\vec{x}(t) = \int_0^t e^{A(t-\tau)} B\vec{u}(\tau) d\tau \quad (2.278)$$

using a change of variables of  $\tau_2 = t - \tau$

$$\vec{x}(t) = \int_0^t -e^{A(\tau_2)} B\vec{u}(t - \tau_2) d\tau_2 \quad (2.279)$$

and using the **Cayley-Hamilton definition** of the matrix exponential, i.e.

$$e^{At} = \sum_{i=0}^{n-1} A^i \alpha_i(t) \quad (2.280)$$

where  $\alpha_0, \dots, \alpha_{n-1}$  are coefficients which depend on  $A$ , one has

$$\vec{x}(t) = \int_0^t \sum_{i=0}^{n-1} -A^i \alpha_i(\tau_2) B\vec{u}(t - \tau_2) d\tau_2 \quad (2.281)$$

which can be rearranged using the properties of integrals and sums

$$\vec{x}(t) = \sum_{i=0}^{n-1} A^i B \int_0^t -\alpha_i(\tau_2) \vec{u}(t-\tau) d\tau_2 \quad (2.282)$$

Then, by letting  $\beta_i(t) = \int_0^t -\alpha_i(\tau_2) \vec{u}(t-\tau) d\tau_2$ , one has

$$\vec{x}(t) = \sum_{i=0}^{n-1} A^i B \beta_i(t) \quad (2.283)$$

and the sum can be written out explicitly as

$$\vec{x}(t) = B\beta_0(t) + AB\beta_1(t) + \dots + A^{n-1}B\beta_{n-1}(t) \quad (2.284)$$

which can also be separated into the product of two matrices

$$\vec{x}(t) = [B \ AB \ \dots \ A^{n-1}B] \begin{bmatrix} \beta_0(t) \\ \beta_1(t) \\ \vdots \\ \beta_{n-1}(t) \end{bmatrix} \quad (2.285)$$

Finally, since any control input  $u(t)$  is allowed, any  $\begin{bmatrix} \beta_0(t) \\ \beta_1(t) \\ \vdots \\ \beta_{n-1}(t) \end{bmatrix}$  can be constructed, thus any  $\vec{x}(t)$  can be reached and the pair  $(A, B)$  is controllable if and only if **controllability matrix**,  $[B \ AB \ \dots \ A^{n_x-1}B]$ , is invertible, i.e.

$$\text{rank}([B \ AB \ \dots \ A^{n_x-1}B]) = n_x \quad (2.286)$$

A second test is the **Popov-Belevitch-Hautus (PBH) controllability test** which states  $(A, B)$  is controllable if and only if for all eigenvalues  $\lambda \in \mathbb{C}$  of  $A$ , one has

$$\text{rank}([\lambda I - A \ B]) = n_x \quad (2.287)$$

A weaker notion than controllability is **stabilizability**. A system is said to be stabilizable if all *unstable* state variables can be controlled. Thus,  $(A, B)$  is stabilizable if and only if for all eigenvalues  $\lambda \in \mathbb{C}$  of  $A$  with  $\text{Real}(\lambda) \geq 0$ , one has

$$\text{rank}([\lambda I - A \ B]) = n_x \quad (2.288)$$

Lastly, a third test uses the **controllability Gramian**, which for continuous-time can be written as

$$W_C(t) = \int_0^t e^{A\tau} BB^T e^{A\tau} d\tau \quad (2.289)$$

It can be shown that if and only if  $W_C(t)$  is positive definite for *any*  $t > 0$ , then the system is controllable. The equation above can be reduced to solving the equation for continuous-time as

$$AW_C + W_C A^T = -BB^T \quad (2.290)$$

then, checking if  $W_C > 0$ . Thus, the eigenvalues of the  $n \times n$  Gramian,  $W_C$ , characterize the relative degree of controllability.

In addition to state controllability, there are two other common notions of controllability. The first is **output controllability** which is the ability of an input to move the *output* from any initial condition to any final condition in finite time. This type of analysis naturally involves the output matrix in addition to the input matrix, e.g. the continuous-time **output controllability matrix** can be shown to be

$$[C \quad CAB \quad CA^2B \quad \cdots \quad CA^{n_x-1}B \quad D] \quad (2.291)$$

It is also important to point out that state and output controllability are not equivalent nor does one imply the other. The second form is **controllability under constraints** which may be imposed upon practical systems modeled as an LTI system. Such constraints may be inherent to the system, e.g. saturating actuator, or imposed by the control designer, e.g. due to safety-related concerns. The effect of constraints to a system is a vast larger topic in control and is mentioned later in this textbook.

**Observability** is defined as the ability to observe system's *past* initial state,  $\vec{x}(0)$ , i.e. if, for some finite time interval,  $[0, t_f]$ , inputs  $\vec{u}(t)$ , and outputs  $\vec{y}(t)$ , the initial state  $\vec{x}(0)$  can be determined. Consider the  $n_x - 1$  continuous output derivatives which necessitate measurements of  $y(t)$  over the time interval,  $[0, t_f]$

$$\begin{aligned} \vec{y}(0) &= C\vec{x}_0 \\ \dot{\vec{y}}(0) &= C\dot{\vec{x}}(0) = C(A\vec{x}(0) + B\vec{u}(0)) \\ \ddot{\vec{y}}(0) &= C\ddot{\vec{x}}(0) = C\frac{d}{dt}(A\vec{x}(0) + B\vec{u}(0)) \\ \ddot{\vec{y}}(0) &= CA(A\vec{x}(0) + B\vec{u}(0)) + CB\dot{\vec{u}}(0) \\ &\vdots = \vdots \\ \vec{y}^{[n_x-1]}(0) &= CA^{n_x-1}\vec{x}(0) + CA^{n_x-2}B\vec{u}^{[n_x-2]}(0) + \cdots + CB\dot{\vec{u}}(0) \end{aligned} \quad (2.292)$$

Thus, as  $\vec{y}(0)$  and  $\vec{u}(0)$  and all their derivatives are known, any  $\vec{x}(0)$  can be estimated and the pair  $(A, C)$

is observable if and only if the **observability matrix**,  $\begin{bmatrix} C \\ CA \\ \dots \\ CA^{n_x-1} \end{bmatrix}$ , is invertible, i.e.

$$\text{rank} \left( \begin{bmatrix} C \\ CA \\ \dots \\ CA^{n_x-1} \end{bmatrix} \right) = n_x \quad (2.293)$$

A second test is the **Popov-Belevitch-Hautus (PBH) observability test** which states  $(A, C)$  is observable if and only if for all eigenvalues  $\lambda \in \mathbb{C}$  of  $A$ , one has

$$\text{rank} ([\lambda I - A \quad C]) = n_x \quad (2.294)$$

A weaker notion than observability is **detectability**. A system is said to be detectable if all *unstable* state variables can be observed. Thus,  $(A, C)$  is detectable if and only if for all eigenvalues  $\lambda \in \mathbb{C}$  of  $A$  with  $\text{Real}(\lambda) \geq 0$ , one has

$$\text{rank} ([\lambda I - A \quad C]) = n_x \quad (2.295)$$

Lastly, a third method uses the **observability Gramian** to determine the observability for continuous-time LTI state-space systems as

$$W_O(t) = \int_0^t e^{A\tau} CC^T e^{A\tau} d\tau \quad (2.296)$$

It can be shown that if and only if  $W_O(t)$  is positive definite for *any*  $t > 0$ , then the system is observable. The Gramian can be shown to be the solution in continuous-time for the equation

$$AW_O + W_O A^T = -CC^T \quad (2.297)$$

then, checking if  $W > 0$ . Thus, the eigenvalues of the  $n_x \times n_x$  Gramian,  $W_O$  characterize the relative degree of observability.

Stability robustness to model uncertainties is a crucial design criterion for the feedback control systems of flight vehicles. Historically, the most widely used measure of stability robustness in flight vehicles has been single-loop gain and phase stability margins for SISO LTI systems, typically designated as at least  $\pm 6$  dB gain margin and  $30^\circ$  phase margin. However, more advanced methods exist for stability robustness analyses of LTI systems due to various types of uncertainties which try to determine bounds on how large the uncertainties can be before the LTI system becomes unstable. This section discusses general types of uncertainties, generalized framework for studying uncertain LTI systems, and primary robust stability results using this framework.

As mentioned previously, the plant model in a feedback control system is typically only an approximation due to model uncertainty, e.g linearization error, model parameter variability, simplified dynamics modeling, etc. This is especially true for FDC. To account for this, control engineers use safety factors when designing feedback control systems. These safety factors are known as **stability margins** and quantify the **feedback control system robustness**, i.e. a measure of the feedback control system will also perform adequately well under a different set of assumptions for the system model. This section will consider four primary SISO stability margins: gain margin, phase margin, delay margin, and disk margin.

## Types of Uncertainties in Dynamical Systems

Fundamentally, dynamical systems are models of real world processes which are typically not completely understood and thus have some uncertainty. Thus, to account for model uncertainty formally, one can define different types of model uncertainty, either as parametric uncertainty or dynamic uncertainty.

**Parametric uncertainty** is specified as parameters which are unknown, but are members of some set of values called an **uncertainty set**. For dynamical systems, these parameters can be **complex parametric uncertainty** or **real parametric uncertainty** where either of these types can be scalar, i.e. some  $\alpha \in \mathbb{C}$  or some  $\beta \in \mathbb{R}$ , or matrix-defined sets. A *probabilistic analysis approach* to modeling parametric uncertainty would be to specify the parameters as random variables distributed according to a distribution function where this approach also allows one to model dependencies between parameters. An alternative *worst-case analysis approach* is to use equally likely values within some finite uncertainty set. As probability theory is beyond the scope of this part of the textbook, robustness analysis of LTI systems under parametric uncertainty will use the worst-case analysis approach.

An important type of uncertainty set for complex parameters is the uncertainty disk defined as

$$\alpha \in Disk \left( \frac{1-m}{1+m}, \frac{1+m}{1-m} \right) = \left\{ \alpha = \frac{1-m\delta}{1+m\delta} \in \mathbb{C} : \delta \in \mathbb{C}, |\delta| \leq 1 \right\} \quad (2.298)$$

where  $m \in [0, 1)$ , the radius of the disk is

$$\frac{1}{2} \left( \frac{1+m}{1-m} - \frac{1-m}{1+m} \right) = \frac{1}{2} \left( \frac{1+2m+m^2 - 1+2m-m^2}{(1+m)(1-m)} \right) = \frac{2m}{(1+m)(1-m)} \quad (2.299)$$

and the center of the disk is

$$\frac{1-m}{1+m} + \frac{2m}{(1+m)(1-m)} = \frac{1-2m+m^2+2m}{(1+m)(1-m)} = m/2 \quad (2.300)$$

which can also be used to define *Disk* instead of  $m$ . Notably, as  $m \rightarrow 1$ , this disk converges to the right half of the complex plane (RHP). An important type of uncertainty set for real parameters is the closed interval defined as

$$\beta \in [\underline{\beta}, \bar{\beta}] = \left\{ \beta = \frac{\bar{\beta}+\beta}{2}\delta_\beta^2 + \frac{\bar{\beta}-\beta}{2}\delta_\beta + \beta_0 : \delta_\beta \in [-1, 1] \right\} \quad (2.301)$$

where  $\delta$  is the normalized uncertainty. Notably,  $\delta_\beta = 0$  corresponds to the nominal  $\beta_0$ ,  $\delta_\beta = -1$  corresponds to  $\underline{\beta}$  and  $\delta_\beta = 1$  corresponds to  $\bar{\beta}$ .

**Dynamic uncertainty** which can be **LTI dynamic uncertainty** or **nonlinear dynamic uncertainty**. Notably, one can relate LTI dynamic uncertainty to complex parametric uncertainty. This textbook will focus on LTI dynamic uncertainty for robustness analysis of LTI systems as robustness analysis for nonlinear and/or time-varying dynamic uncertainty requires more general **integral quadratic constraints** which will not be covered in this textbook.

An important uncertainty set for LTI dynamic uncertainty is defined as

$$G(s) \in \mathcal{M}_W = \{G(s) = G_0(s)(I + W(s)\Delta(s)) : \Delta(s) \text{ LTI}, \|\Delta\|_\infty \leq 1\} \quad (2.302)$$

where  $G_0(s)$  is the nominal LTI model,  $W(s)$  is an LTI weight, and  $\Delta(s)$  is the normalized uncertainty. Noticeably any  $G(s) \in \mathcal{M}_W$  satisfies

$$\left| \frac{G(j\omega) - G_0(j\omega)}{G_0(j\omega)} \right| \leq |W(j\omega)| \quad \forall \omega \quad (2.303)$$

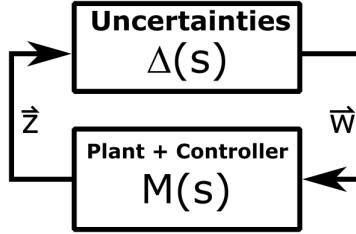
Thus,  $W(j\omega)$  is a bound on the relative error at all frequencies from  $G_0(j\omega)$ . Furthermore, as the relative error bound is unaffected by the phase of  $W(s)$ ,  $W(s)$  can be chosen without loss of generality to be stable and minimum phase and modeled as a transfer function or state-space model. Lastly, if  $\Delta$  is assumed to be stable, then one is restricting that  $G(s)$  can only have RHP poles at the same locations as  $G_0(s)$  which is often used in robustness analysis.

### Multiplicative Uncertain LTI System Model

### Additive Uncertain LTI System Model

### Generalized Uncertain LTI System Model

Consider the generalized LTI uncertainty model architecture



which is denoted by  $F_U(M, \Delta)$  and is a linear fractional transformation (LFT), i.e.

$$\begin{bmatrix} \vec{z} \\ \vec{d} \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \vec{w} \\ \vec{e} \end{bmatrix}$$

$$\vec{w} = \Delta \vec{z}$$
(2.304)

Here the generalized uncertainty set,  $\Delta(s)$  with  $\|\Delta\|_\infty \leq 1$ , contains all parametric and stable LTI dynamic uncertainties in the nominal stable LTI system,  $M(s)$ , e.g. the nominal generalized LTI plant and controller. Thus, the nominal stable LTI system dynamics correspond to  $\Delta = 0$  and this LFT is well-posed if  $I_{n_z} - M_{11}(\infty)\Delta(\infty)$  is invertible. Recalling the nature of LFTs,  $\Delta$  may have a “structure” depending on the types of uncertainties, e.g. multiple scalar uncertainties and/or real parametric uncertainties. However, if  $\Delta$  is a single complex uncertainty, i.e.  $\Delta \in \mathbb{C}^{n_w \times n_z}$  or a single LTI dynamic uncertainty, then  $\Delta$  is an **unstructured uncertainty set**.

Using LFTs, the LTI state-space model for the nominal model,  $M$ , can be written as

$$\begin{aligned} \dot{\vec{x}}(t) &= A \vec{x}(t) + [B_1 \quad B_2] \begin{bmatrix} \vec{w}(t) \\ \vec{d}(t) \end{bmatrix} \\ \begin{bmatrix} \vec{z}(t) \\ \vec{e}(t) \end{bmatrix} &= \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \vec{x}(t) + \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} \vec{w}(t) \\ \vec{d}(t) \end{bmatrix} \end{aligned}$$
(2.305)

with the generalized uncertainty set,  $\Delta$ , written as

$$\begin{aligned} \dot{\vec{x}}_\Delta &= A_\Delta \vec{x}_\Delta + B_\Delta \vec{z} \\ \vec{w} &= C_\Delta \vec{x}_\Delta + D_\Delta \vec{z} \end{aligned}$$
(2.306)

where if  $\Delta$  contains only real or complex uncertainty, then  $A_\Delta = B_\Delta = C_\Delta = 0$  with no state vector,  $\vec{x}_\Delta$ .

Combining, one has for the uncertainty input and output vectors

$$\begin{bmatrix} I & -D_\Delta \\ -D_{11} & I \end{bmatrix} \begin{bmatrix} \vec{w}(t) \\ \vec{z}(t) \end{bmatrix} = \begin{bmatrix} 0 & C_\Delta \\ C_1 & 0 \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \vec{x}_\Delta(t) \end{bmatrix} + \begin{bmatrix} 0 \\ D_{12} \end{bmatrix} \begin{bmatrix} \vec{d}(t) \end{bmatrix}$$
(2.307)

thus, the interconnection of  $M$  and  $\Delta$  is well-posed if and only if  $I_{n_z} - D_{11}D_\Delta$  is invertible. Thus, similar to the generalized LFT feedback control system using loop-shifting, one can form the closed-loop LTI state-space system model as

$$\begin{aligned} \begin{bmatrix} \dot{\vec{x}}(t) \\ \dot{\vec{x}}_\Delta(t) \end{bmatrix} &= \begin{bmatrix} A + B_1 D_\Delta C_1 & B_1 C_\Delta \\ B_\Delta C_1 & A_\Delta \end{bmatrix} \begin{bmatrix} \vec{x}(t) \\ \vec{x}_\Delta(t) \end{bmatrix} + \begin{bmatrix} B_2 + B_1 D_\Delta D_{21} \\ B_\Delta D_{12} \end{bmatrix} \vec{d}(t) \\ \vec{e}(t) &= [C_2 + D_{21} D_\Delta C_1 \quad D_{21} C_\Delta] \begin{bmatrix} \vec{x}(t) \\ \vec{x}_\Delta(t) \end{bmatrix} + (D_{22} + D_{21} D_\Delta D_{12}) \vec{d}(t) \end{aligned}$$
(2.308)

with the definitions for the closed-loop state matrix,  $A_L$ , as

$$A_L = \begin{bmatrix} A + B_1 D_\Delta C_1 & B_1 C_\Delta \\ B_\Delta C_2 & A_\Delta \end{bmatrix} \quad (2.309)$$

the closed-loop input matrix,  $B_L$ , as

$$B_L = \begin{bmatrix} B_2 + B_1 D_\Delta D_{12} \\ B_\Delta D_{12} \end{bmatrix} \quad (2.310)$$

the closed-loop output matrix,  $C_L$ , as

$$C_L = [C_2 + D_{21} D_\Delta C_1 \quad D_{21} C_\Delta] \quad (2.311)$$

and the closed-loop feedthrough matrix,  $D_L$ , as

$$D_L = D_{22} + D_{21} D_\Delta D_{12} \quad (2.312)$$

which for stability of the generalized LTI uncertainty system, one requires that  $A_L$  is stable.

However, for stability robustness analysis, if one assumes  $M$  and  $\Delta$  are stable and  $F_U(M, \Delta)$  is well-posed, then  $F_U(M, \Delta)$  has a pole at  $s_p$  if and only if  $\det(s_p I - A_L) = 0$  which can be shown through determinant formulas and  $M$  and  $\Delta$  stability. This only occurs if and only if  $\det(I - M_{11}(s_p)\Delta(s_p)) = 0$ . This implies that  $F_U(M, \Delta)$  is unstable if and only if  $\det(I - M_{11}(s_p)\Delta(s_p)) \neq 0$  for some  $s_p$  in the closed RHP. This can also be seen from the LFT from  $d$  to  $e$  as

$$F_U(M, \Delta) = M_{22} + M_{21}\Delta(I - M_{11}\Delta)^{-1}M_{12} \quad (2.313)$$

where  $\Delta$ ,  $M_{11}$ ,  $M_{12}$ ,  $M_{21}$ , and  $M_{22}$  are all stable by assumption. Hence,  $F_U(M, \Delta)$  is stable if and only if  $(I - M_{11}\Delta)^{-1}$  is unstable.

This allows the definition that the generalized LTI uncertainty system,  $F_U(M, \Delta)$ , is **robustly stable** if and only if  $I - M_{11}\Delta$  is invertible for all  $\Delta \in \Delta$  where  $\Delta$  is the modeled uncertainty set, structured or unstructured. A second consequence of this fact is that the robust stability margin can be computed by finding  $\Delta \in \Delta$  with the smallest  $\|\Delta\|_\infty$  such that  $\det(I - M_{11}(s_p)\Delta(s_p)) = 0$  for some  $s_p$  in the closed RHP. Thus, the **robust stability margin** of generalized LTI uncertainty system,  $F_U(M, \Delta)$ , is the largest value  $m$  such that  $F_U(M, \Delta)$  is well-posed and stable for all  $\Delta \in \Delta$  with  $\|\Delta\|_\infty < m$ . Note that if  $m > 1$ , then  $F_U(M, \Delta)$  is robustly stable. Furthermore, one can define the worst-case uncertainty,  $\Delta_{w-c} \in \Delta$ , which destabilizes  $F_U(M, \Delta)$  and has  $\|\Delta\|_\infty = m$ . Notably, this worst-case uncertainty causes  $F_U(M, \Delta)$  to have system poles on the imaginary axis at some  $\pm j\omega_{w-c}$ .

## Robust Stability for Unstructured Uncertainty

The **small-gain theorem (SGT)** provides a method for unstructured complex uncertainty. Namely, if  $\Delta(s) \in \mathbb{C}^{n_w \times n_z}$  and  $M_{11} \in \mathbb{C}^{n_z \times n_w}$ , then  $\det(I - M_{11}\Delta) \neq 0$  for all  $\Delta \in \mathbb{C}$  with  $\bar{\sigma}(\Delta) < m$  if and only if  $\bar{\sigma}(M) \leq \frac{1}{m}$ . This can be proven by recalling the maximum singular value is equal to the matrix 2,2 induced norm. Thus, for  $\bar{\sigma}(M) \leq \frac{1}{m}$ , then

$$\bar{\sigma}(M_{11}\Delta) \leq \bar{\sigma}(M)\bar{\sigma}(\Delta) < 1 \quad \forall \Delta \text{ with } \bar{\sigma}(\Delta) < m \quad (2.314)$$

which implies  $\|M\Delta\vec{v}\| < \|\vec{v}\|$  for any non-zero  $\vec{v} \in \mathbb{C}^{n_v}$  or  $(I - M\Delta)\vec{v} \neq 0$ , i.e.  $\det(I - M\Delta) \neq 0$  for all  $\bar{\sigma}(\Delta) < m$  which proves the sufficient condition. Furthermore, if  $\bar{\sigma}(M) > \frac{1}{m}$ , then the SVD of  $M$  provides

$$\bar{\sigma}(M)\vec{v} = M\vec{z} \text{ and } \|\vec{v}\|_2 = \|\vec{z}\|_2 = 1 \quad (2.315)$$

and by selecting  $\Delta_0 = \frac{1}{\bar{\sigma}(M)}\vec{z}\vec{v}^* \in \mathbb{C}^{n_z \times n_w}$  which provides  $\bar{\sigma}(\Delta_0) = \frac{1}{\bar{\sigma}(M)} < m$ , one has  $(I - M\Delta_0)\vec{v} = 0$  and  $\det(I - M\Delta_0) = 0$  which proves the necessary condition.

The SGT can be extended to unstructured LTI dynamic uncertainty through the fact that complex uncertainty can be interpolated at any finite, non-zero frequency using a stable LTI system and the LTI dynamic uncertainty has norm no larger than the chosen complex uncertainty. Formally, for a frequency,  $0 < \omega_0 < \infty$  and a rank-one matrix  $\Delta_0 = \vec{z}_0\vec{v}_0^* \in \mathbb{C}^{n_z \times n_v}$  there exists an  $n_z \times n_v$  stable, LTI system,  $\Delta(s)$  such that  $\Delta(j\omega_0) = \Delta_0$  and  $\|\Delta\|_\infty \leq \bar{\sigma}(\Delta_0)$ . Notably, the equivalence between complex and dynamic uncertainty breaks down at  $\omega_0 = 0$  and  $\infty$  as an LTI system with real-valued matrices has a real frequency response at  $\omega_0 = 0$  and  $\infty$ , i.e.  $\Delta(0), \Delta(\infty) \in \mathbb{R}^{n_z \times n_w}$ .

In summary, consider a generalized LTI uncertainty system,  $F_U(M, \Delta)$  where  $\Delta$  is an unstructured, stable LTI system and  $M$  is stable.  $F_U(M, \Delta)$  is well-posed and stable for all  $\|\Delta\|_\infty < m$  if and only if  $\|M_{11}\|_\infty \leq \frac{1}{m}$ , i.e.

$$\max_{\omega \in \mathbb{R} \cup \{\infty\}} \bar{\sigma}(M_{11}(j\omega)) \leq \frac{1}{m} \quad (2.316)$$

which can be proven using the above results. Notably, if  $M_{11}$  achieves its peak gain at  $\omega_0 = 0$  or  $\infty$ , then the interpolation is performed at some arbitrarily small or large finite frequency.

## Robust Stability for Structured Uncertainty

Notably, a consequence of the SGT is that  $\det(I - M_{11}\Delta) \neq 0$  for any  $\Delta \in \mathbb{C}^{n_w \times n_z}$  with  $\bar{\sigma}(\Delta) < \frac{1}{\bar{\sigma}(M_{11})}$ . Thus, one can interpret this fact as an optimization problem, i.e.

$$\begin{aligned} \frac{1}{\bar{\sigma}(M_{11})} &= \min_{\Delta \in \mathbb{C}^{n_w \times n_z}} \bar{\sigma}(\Delta) \\ \text{subject to: } &\det(I - M_{11}\Delta) = 0 \end{aligned} \quad (2.317)$$

Thus, for some structured  $\Delta \in \Delta$ , one can define the **structured singular value (SSV)** for some  $M_{11} \in \mathbb{C}^{n_w \times n_z}$  as

$$\mu_\Delta(M_{11}) = \begin{cases} \left[ \min_{\Delta \in \Delta} (\bar{\sigma}(\Delta) \text{ s.t. } \det(I - \Delta M_{11}) = 0) \right]^{-1} \\ 0, \text{ if no } \Delta \in \Delta \text{ causes } \det(I - \Delta M_{11}) = 0 \end{cases} \quad (2.318)$$

where notably the second case *may* occur if  $\Delta$  consists of only real parametric uncertainties.

With the SSV defined one can state a generalization of the SGT which follows almost directly from this definition. Namely, if  $\Delta(s) \in \Delta \subset \mathbb{C}^{n_w \times n_z}$  and  $M_{11} \in \mathbb{C}^{n_z \times n_w}$ , then  $\det(I - M_{11}\Delta) \neq 0$  for all  $\Delta \in \Delta$  with  $\bar{\sigma}(\Delta) < m$  if and only if  $\mu_\Delta(M_{11}) \leq \frac{1}{m}$ . Notably, if  $\Delta \in \mathbb{C}^{n_w \times n_z}$  is a full complex matrix, then the SGT-based maximum singular value optimization provides the SSV as

$$\mu_\Delta(M_{11}) = \bar{\sigma}(M_{11}) \quad (2.319)$$

However, computing  $\mu_\Delta(M_{11})$  for general uncertainty structures is a computationally difficult problem, thus, most implementations of SSV-based robustness bounds use upper and lower bounds on  $\mu_\Delta(M_{11})$ . Thus, as generally  $\Delta \subset \mathbb{C}^{n_w \times n_z}$ , a simple upper bound is

$$\mu_\Delta(M_{11}) \leq \bar{\sigma}(M_{11}) \quad (2.320)$$

which can be extended via tighter bounds through  $D$ -scalings which account for the non-uniqueness in the LFT representation of  $M$ .

A simple lower bound for  $\mu_\Delta(M_{11})$  can be computed if  $\Delta(s)$  that is a diagonal matrix whose diagonal is a single complex scalar  $\delta \in \mathbb{C}$ , i.e.

$$\Delta = \{\delta I : \delta \in \mathbb{C}\} \quad (2.321)$$

Substituting this into  $I - \Delta M_{11}$ , one has

$$I - \Delta M_{11} = I - \delta IM_{11} = \delta \left( \frac{1}{\delta} I - M \right) \quad (2.322)$$

which defines an eigenvalue problem for square  $M_{11}$ . Thus, for the simplest structure, one has

$$\mu_\Delta(M_{11}) = \bar{\rho}(M_{11}) = \max_i |\lambda(M_{11})| \quad (2.323)$$

where  $\bar{\rho}(M_{11})$  is the **spectral radius** of the matrix,  $M_{11}$ , i.e. the largest absolute value of the eigenvalues of  $M_{11}$ .

Thus, the SSV can be at least bounded above and below by

$$\bar{\rho}(M_{11}) \leq \mu_\Delta(M_{11}) \leq \bar{\sigma}(M_{11}) \quad (2.324)$$

and in general, numerical methods use versions of power iteration and  $D$ -scalings on general  $M_{11}$  to numerically compute the SSV bounds as

$$\max_Q \lambda(QM_{11}) \leq \mu_\Delta(M_{11}) \leq \inf_D \bar{\sigma}(DM_{11}D^{-1}) \quad (2.325)$$

Finally, assuming one can use interpolation to replace LTI dynamic uncertainty blocks with corresponding complex parametric uncertainty of the same size, one can state an exact, necessary, and sufficient condition for the robust stability margin of the generalized LTI uncertainty system with structured real and complex parametric uncertainty, and stable, LTI dynamic uncertainty,  $\Delta \in \Delta$  as follows. Consider a generalized LTI uncertainty system,  $F_U(M, \Delta)$  where  $\Delta$  is a stable LTI system in the structured set,  $\Delta$  and  $M$  is stable.  $F_U(M, \Delta)$  is well-posed and stable for all  $\Delta \in \Delta$  with  $\|\Delta\|_\infty < m$  if and only if

$$\max_{\omega \in \mathbb{R} \cup \{\infty\}} \mu_\Delta(M_{11}(j\omega)) \leq \frac{1}{m} \quad (2.326)$$

which can be proven using the above results.

## 2.6 MIMO LTI Command-Tracking Control System Design

This chapter has provided methods for MIMO LTI feedback control systems to be analyzed in terms of stability, controllability, observability, and robust stability using generalized LTI models. However, the purpose of a control system is often to track a certain command. This section will introduce two additional control system architectures to track reference commands, namely, the model-following control and the servomechanism control.

## Servomechanism Control for Command Tracking

In control system design, one often desires the control system to regulate the tracking error of a reference command,  $r(t)$ , to zero in the presence of unknown disturbances to the system,  $w(t)$ . From classical control theory for SISO systems, it is known that integral control action is necessary to achieve this zero steady-state tracking error. By the **internal model principle**, the number of integrators in the open-loop transfer function, i.e. the **system type**, must be greater than or equal to the reference and disturbance signal orders. Thus, if  $r(t) = w(t) = 0$ , then one needs 0 integrators and one has a type 0 control system, also known as a **regulator**. If  $\dot{r}(t) = \dot{w}(t) = 0$ , then one needs 1 integrator and one has a type 1 control system. If  $\ddot{r}(t) = \ddot{w}(t) = 0$ , then one needs 2 integrators and one has a type 2 control system.

For MIMO systems, state feedback controllers act as a type 0 control system, regulate system state to zero, but for any non-zero reference command or disturbance, there will be a steady-state error. Thus, the addition of integral action for state feedback is desirable for tracking constant non-zero reference signals with any potential constant non-zero disturbances. For other types of reference commands and disturbances, one requires additional feedback control considerations for zero-error tracking of an array of reference commands in the presence of an array of disturbances. A standard design approach for achieving this is the robust servomechanism control system which consists of two components, a servomechanism and state feedback. The “robust” terminology comes from its ability to reach zero steady-state tracking error in the presence of specified classes of *disturbances*, a type of uncertainty.

Consider the following continuous-time LTI state-space system

$$\begin{aligned}\dot{\vec{x}} &= A\vec{x} + B\vec{u} + M\vec{w} \\ \vec{y} &= C\vec{x} + D\vec{u}\end{aligned}\tag{2.327}$$

with an additive unknown bounded disturbance,  $\vec{w} \in \mathbb{R}^{n_w}$ . In addition, assume that the system is controllable and observable.

Next, assume the reference command is prescribed as some commanded output,  $\vec{r}(t) \in \mathbb{R}^{n_y}$ , which has the following  $p^{\text{th}}$  order ODE of the form

$$\vec{r}^{[p]} = \sum_{i=1}^p a_i \vec{r}^{[p-i]}\tag{2.328}$$

where the scalar coefficients,  $a_i$ , are known and the superscript  $[j]$  denotes the  $j^{\text{th}}$  derivative. Note that a constant command has the ODE form of  $\vec{r} = 0$  with  $p = 1$  and  $a_1 = 0$ , a ramp command has the ODE form of  $\ddot{\vec{r}} = 0$  with  $p = 2$  and  $a_2 = a_1 = 0$ , and a sinusoidal command at frequency  $\omega_0$  has the ODE form of  $\ddot{\vec{r}} = -\omega_0^2 \vec{r}$  with  $p = 2$ ,  $a_2 = -\omega_0^2$ , and  $a_1 = 0$ . Likewise, assume that the disturbance inputs have the same  $p^{\text{th}}$  order ODE form

$$\vec{w}^{[p]} = \sum_{i=1}^p a_i \vec{w}^{[p-i]}\tag{2.329}$$

with unknown  $w(0) = w_0$ .

Also, define the tracking error signal as

$$\vec{e} = \vec{y} - \vec{r}\tag{2.330}$$

which can be differentiated  $p$  times to obtain the error ODE

$$\vec{e}^{[p]} - \sum_{i=1}^p a_i \vec{e}^{[p-i]} = \left( \vec{y}^{[p]} - \sum_{i=1}^p a_i \vec{y}^{[p-i]} \right) - \left( \vec{r}^{[p]} - \sum_{i=1}^p a_i \vec{r}^{[p-i]} \right) \quad (2.331)$$

where, by definition, the second term on the right side will be zero and the first term can be written using the output equation and its derivatives as

$$\vec{e}^{[p]} - \sum_{i=1}^p a_i \vec{e}^{[p-i]} = C \left( \vec{x}^{[p]} - \sum_{i=1}^p a_i \vec{x}^{[p-i]} \right) + D \left( \vec{u}^{[p]} - \sum_{i=1}^p a_i \vec{u}^{[p-i]} \right) \quad (2.332)$$

which represents a set of coupled ODEs for the error differential equation.

Then, defining

$$\vec{\eta} = \vec{x}^{[p]} - \sum_{i=1}^p a_i \vec{x}^{[p-i]} \quad (2.333)$$

and

$$\vec{\mu} = \vec{u}^{[p]} - \sum_{i=1}^p a_i \vec{u}^{[p-i]} \quad (2.334)$$

the error differential equation becomes

$$\vec{e}^{[p]} - \sum_{i=1}^p a_i \vec{e}^{[p-i]} = C \vec{\eta} + D \vec{\mu} \quad (2.335)$$

Differentiating, one has

$$\dot{\vec{\eta}} = \vec{x}^{[p+1]} - \sum_{i=1}^p a_i \vec{x}^{[p-i+1]} \quad (2.336)$$

and by substitution for  $\dot{\vec{x}}$ , one has

$$\dot{\vec{\eta}} = A \left( \vec{x}^{[p]} - \sum_{i=1}^p a_i \vec{x}^{[p-i]} \right) + B \left( \vec{u}^{[p]} - \sum_{i=1}^p a_i \vec{u}^{[p-i]} \right) + M \left( \vec{w}^{[p]} - \sum_{i=1}^p a_i \vec{w}^{[p-i]} \right) \quad (2.337)$$

or

$$\dot{\vec{\eta}} = A \vec{\eta} + B \vec{\mu} \quad (2.338)$$

which is the original system without the additive disturbance term.

Finally, the **servomechanism state-space model**

$$\dot{\vec{z}} = \tilde{A} \vec{z} + \tilde{B} \vec{\mu} \quad (2.339)$$

where  $\vec{z}$  is the augmented  $n_x + p \times n_y$  state vector

$$\vec{z} = \begin{bmatrix} \vec{e} \\ \dot{\vec{e}} \\ \vdots \\ \vec{e}^{[p-1]} \\ \vec{\eta} \end{bmatrix} \quad (2.340)$$

with augmented state matrix

$$\tilde{A} = \begin{bmatrix} 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \\ a_p I & a_{p-1} I & \cdots & a_1 I & C \\ 0 & 0 & \cdots & 0 & A \end{bmatrix} \quad (2.341)$$

and augmented input matrix

$$\tilde{B} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ D \\ B \end{bmatrix} \quad (2.342)$$

In steady-state, this allows the state vector,  $\vec{x}$ , to be nonzero to achieve  $\vec{e} = 0$ , in which case

$$\vec{r} = \vec{y} = C\vec{x} + D\vec{u} \quad (2.343)$$

Lastly, note that to check for controllability of the servomechanism state-space model, one requires that  $(A, B)$  be controllable,  $n_u \geq n_y$ , and that  $(A, B, C, D)$  must not have any zeros in common with the ODE for  $\vec{y}$ .

For this system, one can design a servomechanism state feedback control law as

$$\vec{\mu} = K_z \vec{z} \quad (2.344)$$

where  $K_z$  is the feedback gain matrix which can be partitioned in the same manner as  $\vec{z}$ , i.e.

$$K_z = [K_p \ K_{p-1} \ \cdots K_1 \ K_x] \quad (2.345)$$

Next, recalling the equation for  $\vec{\mu}$ , one has

$$\vec{\mu} = \vec{u}^{[p]} - \sum_{i=1}^p a_i \vec{u}^{[p-i]} = [K_p \ K_{p-1} \ \cdots K_1 \ K_x] \begin{bmatrix} \vec{e} \\ \vec{e} \\ \vdots \\ \vec{e}^{[p-1]} \\ \vec{\eta} \end{bmatrix} \quad (2.346)$$

or, multiplying out and substituting for  $\vec{\eta}$ , one has

$$\vec{\mu} = \vec{u}^{[p]} - \sum_{i=1}^p a_i \vec{u}^{[p-i]} = \sum_{i=1}^p K_i \vec{e}^{[p-i]} + K_x (\vec{x}^{[p]} - \sum_{i=1}^p a_i \vec{x}^{[p-i]}) \quad (2.347)$$

Then, integrating  $p$  times provides the control law for the original system as

$$\vec{u} = K_x \vec{x} + \sum_{i=1}^p \frac{(a_i(\vec{u} + K_x \vec{x}) + K_i \vec{e})}{s^i} \quad (2.348)$$

where it should be noted that the  $K_x$  term will enforce closed-loop stability to the dynamical system and the  $p$  integrators and the integrator gains provide integral error control, and the coefficients,  $a_i$ , embed the model of the reference command to be tracked. Thus, the closed-loop dynamics model is

$$\dot{\vec{x}}_{aug} = (\tilde{A} + \tilde{B}K_z)\vec{x}_{aug} + \begin{bmatrix} -I_{n_y \times n_y} \\ 0_{n_x \times n_y} \end{bmatrix} \vec{r} \quad (2.349)$$

where  $\vec{x}_{aug}$  is the **augmented state** which is related to the servomechanism state,  $\vec{z}$ , by  $p$  integrations, i.e.

$$\vec{x}_{aug} = \frac{\vec{z}}{s^p} \quad (2.350)$$

### Implicit Model-Following Control for Command Tracking

### Explicit Model-Following Control using Command Generator Tracking

### References

For more information, please refer to the following

- Lavretsky, E., and Wise, K. A., “4.4 Command Tracking and Robust Servomechanism Control,” *Robust and Adaptive Control: With Aerospace Applications*, 2nd ed., Vol. 1, Springer, London, 2024, pp. 214-230
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “5.6 Model-Following Design,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 455-463

---

# Continuous-Time LTI Optimal Control Theory

## 3.1 Introduction to Optimal Control

**Decision theory** is the determination of the optimal decision given constraints and assumptions on a decision problem. **Optimal control theory** is special case of a model-based decision problem and can be stated as deciding which controller one should select as the optimal control policy with respect to some chosen objective for the dynamical system. When the chosen objective involves the actions of other, separate decision maker(s), one has an extension of optimal control theory called **differential game theory**. Regardless, the dynamic optimization, i.e. optimization over time, is critical to optimal control theory and can be formulated for both continuous-time and discrete-time dynamical systems. This chapter focuses on the continuous-time framework and a similar framework for discrete-time dynamical systems can be found in Appendix B of this textbook.

To solve for this optimal control policy, one uses methods from **mathematical optimization**, also known as **mathematical programming**, which formulates the selection of the “best” or optimal element of a set of possible elements with respect to some criteria, i.e. objective. Typically, the optimal is defined as the decision that achieves the minimum (or maximum) with respect to the mathematical representation of the objective. These sets of possible elements can be continuous, discrete, and/or constrained. The objective in optimal control can be imposed superficially by the control designer or dictated by the real dynamic process. Furthermore, this optimization procedure provides the general **optimal control problem (OCP)** which is typically stated as a minimization as follows.

The OCP for continuous-time dynamical systems can be generally written as

$$\begin{aligned} \vec{u}_{opt}(t) &= \underset{u(t) \forall t \in [t_0, t_f]}{\operatorname{argmin}} \quad \mathcal{J}(\vec{x}, \vec{u}, t, t_0, t_f) \\ &\text{subject to:} \\ \text{continuous-time dynamics: } & \dot{\vec{x}}(t) = f(\vec{x}(t), \vec{u}(t), t) \\ \text{boundary conditions: } & e(\vec{x}(t_0), t_0, \vec{x}(t_f), t_f) = 0 \\ \text{path constraints: } & c(\vec{x}(t), \vec{u}(t), t) \leq 0 \end{aligned} \tag{3.1}$$

where  $\vec{u}^{\text{opt}}(t)$  is the **optimal control function**,  $\operatorname{argmin}$  stands for “the argument which minimizes” the following expression,  $t_0$  is the start time,  $t_f$  is the final time,  $t_f - t_0$  is the **time horizon**,  $\mathcal{J}(\vec{x}, \vec{u}, t)$  is the **cost functional**, also known as the **objective functional** or the **performance index**, which generally depends on the state, input, and time. The term **functional** is a mathematical definition for a “function of a function.”

In optimal control design, one determines what to use for the cost functional,  $\mathcal{J}$ . Next, one must show there exists an optimal solution. Then, one uses synthesis methods to solve for the optimal controller based on the chosen  $\mathcal{J}$ . The optimization is also subject to some dynamics for the state of the system, boundary conditions for the system, and some constraints on the state, inputs, and time and there may be multiple solutions to the OCP. Lastly, it should be noted often one is not necessarily interested in finding the truly optimal solution to the OCP, but an efficient near-optimal solution for complex OCPs.

Within the general OCP framework, there are different versions of the OCP that can be further characterized. First, one can optimize for finite- or infinite-time horizons typically shortened to finite- or infinite-horizon OCPs, i.e.  $t_f \neq \infty$  or  $t_f = \infty$ , respectively. Also, often the infinite-horizon OCPs may be easier to solve than finite-horizon OCPs as the optimal control policy will not depend on any specific time, but only on the state, thus becoming a *fixed-gain* control policy, which may be desirable. It should also be mentioned that finite-horizon OCPs may be re-optimized recursively by the control system forming a **receding-horizon control (RHC)**, also known as a **model predictive control (MPC)**. Second, the dynamics of the system may be linear or nonlinear where linear systems allow for simpler solvers. Third, OCP solvers may be simplified if the OCP is unconstrained or constrained. The effects of constraints on the solving method typically revolves around whether there are state constraints, input constraints, or both. Finally, the cost functional or function may be characterized as linear, quadratic, convex, or non-convex, each requiring more complex solvers. This is due to the fact that for a finite minimum cost to exist, the simplest model would be a linear cost for constrained OCPs and a quadratic cost for unconstrained OCPs. Beyond these cases, one can also make a more general distinction between convex and non-convex costs since convexity implies a local minimum is a global minimum which simplifies many numerical search methods which modern computing can perform very efficiently.

## Generalized Calculus of Variations

Consider the **classical OCP**

$$\begin{aligned} \vec{u}_{opt}(t) = \underset{u(t) \forall t \in [t_0, t_f]}{\operatorname{argmin}} \quad & \mathcal{E}(\vec{x}(t_f)) + \int_{t_0}^{t_f} \mathcal{L}(\vec{x}(t), \vec{u}(t)) dt \\ & \text{subject to:} \end{aligned} \quad (3.2)$$

continuous-time dynamics:  $\dot{\vec{x}}(t) = f(\vec{x}(t), \vec{u}(t))$

initial conditions:  $\vec{x}(t_0) - \vec{x}_0 = 0$

where  $t_f$  and  $t_0$  are fixed,  $\mathcal{E}$  is the **Mayer term**, also known as the **endpoint cost** or **terminal cost**, and  $\mathcal{L}$  is the **Lagrangian term**, also known as the **running cost** or the **instant cost** which must be nonnegative.

To solve these OCPs, one can use **generalized calculus of variations** by augmenting the Lagrangian with the **costate vector**, also known as the **adjoint vector**,  $\vec{\lambda}(t)$ , which are similar to Lagrange multiplier, but are functions of time rather than constants. These adjoint vectors allow one to form the augmented cost functional

$$\bar{\mathcal{J}} = \mathcal{E}(\vec{x}(t_f)) + \int_{t_0}^{t_f} \left( \mathcal{L}(\vec{x}(t), \vec{u}(t), t) + \vec{\lambda}^T(f(\vec{x}(t), \vec{u}(t), t) - \dot{\vec{x}}(t)) dt \right) \quad (3.3)$$

where  $\vec{\lambda}(t)$  can be *any* vector because the state dynamics require that

$$f(\vec{x}(t), \vec{u}(t)) - \dot{\vec{x}}(t) = 0 \quad (3.4)$$

holds for all time, which means  $\vec{\lambda}(t)$  is being multiplied by zero in the expression above. Furthermore, define the **Hamiltonian** as

$$\mathcal{H}(\vec{x}(t), \vec{u}(t), \vec{\lambda}(t), t) = \mathcal{L}(\vec{x}(t), \vec{u}(t)) + \vec{\lambda}^T f(\vec{x}(t), \vec{u}(t)) \quad (3.5)$$

where notably

$$\dot{\vec{x}}^T = f(\vec{x}(t), \vec{u}(t))^T = \frac{\partial \mathcal{H}}{\partial \vec{\lambda}} \quad (3.6)$$

Thus, one has

$$\bar{\mathcal{J}} = \mathcal{E}(\vec{x}(t_f)) + \int_{t_0}^{t_f} \mathcal{H} - \vec{\lambda}^T \dot{\vec{x}}(t) dt \quad (3.7)$$

Next, one can form the **variation of  $\bar{\mathcal{J}}$**  as

$$\delta \bar{\mathcal{J}} = \frac{\partial \mathcal{E}}{\partial \vec{x}} \delta \vec{x}(t_f) + \int_{t_0}^{t_f} \frac{\partial \mathcal{H}}{\partial \vec{x}} \delta \vec{x} + \frac{\partial \mathcal{H}}{\partial \vec{u}} \delta \vec{u} - \vec{\lambda}^T \delta \dot{\vec{x}} dt \quad (3.8)$$

By expanding the last term using integration by parts, one obtains

$$-\int_{t_0}^{t_f} \vec{\lambda}^T \delta \dot{\vec{x}} dt = -\vec{\lambda}^T(t_f) \delta \vec{x}(t_f) + \vec{\lambda}^T(t_0) \delta \vec{x}(t_0) + \int_{t_0}^{t_f} \dot{\vec{\lambda}}^T \delta \vec{x} dt \quad (3.9)$$

By substitution and rearrangement, one can separate  $\delta \bar{\mathcal{J}}$  into four different components as

$$\delta \bar{\mathcal{J}} = \left( \frac{\partial \mathcal{E}(t_f)}{\partial \vec{x}} - \vec{\lambda}^T(t_f) \right) \partial \vec{x}(t_f) + \vec{\lambda}^T(t_0) \partial \vec{x}(0) + \int_{t_0}^{t_f} \left( \frac{\partial \mathcal{H}}{\partial \vec{x}} + \dot{\vec{\lambda}}^T \right) \partial \vec{x} + \frac{\partial \mathcal{H}}{\partial \vec{u}} \partial \vec{u} dt \quad (3.10)$$

However, as the adjoint vectors are arbitrary, one can select them to make the coefficients of  $\partial \vec{x}(t)$  and  $\partial \vec{x}(t_f)$  equal to zero, i.e.

$$\begin{aligned} \dot{\vec{\lambda}}^T &= -\frac{\partial \mathcal{H}}{\partial \vec{x}} \\ \vec{\lambda}(t_f) &= \frac{\partial \mathcal{E}(t_f)}{\partial \vec{x}} \end{aligned} \quad (3.11)$$

where the first equation is the **costate equation**, also known as the **adjoint equation**, for the dynamics of  $\vec{\lambda}$  while the equation provides the final condition for  $\vec{\lambda}$ .

With this choice, one has

$$\delta \bar{\mathcal{J}} = \int_0^{t_f} \frac{\partial \mathcal{H}}{\partial \vec{u}} \partial \vec{u} dt \quad (3.12)$$

which for  $\mathcal{J}$  to be minimized requires  $\delta \bar{\mathcal{J}} = 0$ , thus, assuming  $\vec{u}$  is free to vary, one requires

$$\frac{\partial \mathcal{H}^T}{\partial \vec{u}} = 0 \quad (3.13)$$

However, if  $\vec{u}$  is constrained in set  $\mathcal{U}$ , one may use **Pontryagin's principle** to which replaces the previous equation with the requirement

$$\mathcal{H}(\vec{x}_{opt}(t), \vec{u}_{opt}(t), \vec{\lambda}_{opt}(t)) \leq \mathcal{H}(\vec{x}(t), \vec{u}(t), \vec{\lambda}(t)) \quad \forall t \in [t_0, t_f], \quad \vec{u} \in \mathcal{U} \quad (3.14)$$

For example, if one bounds the control simply by

$$\vec{u}_{min} \leq \vec{u} \leq \vec{u}_{max} \quad (3.15)$$

which requires for feasibility of the variation of  $\vec{u}$  at its limits which must only be in the direction allowed that one has

$$\begin{cases} \vec{u}_{opt}(t) = \vec{u}_{min} & \text{for } \frac{\partial \mathcal{H}}{\partial \vec{u}} \geq 0 \\ \vec{u}_{min} < \vec{u}_{opt}(t) < \vec{u}_{max} & \text{for } \frac{\partial \mathcal{H}}{\partial \vec{u}} = 0 \\ \vec{u}_{opt}(t) = \vec{u}_{max} & \text{for } \frac{\partial \mathcal{H}}{\partial \vec{u}} \leq 0 \end{cases} \quad (3.16)$$

It should be pointed out that Pontryagin's principle is a necessary condition, but only sufficient if  $\mathcal{L}(\vec{x}(t), \vec{u}(t), t)$  and  $f(\vec{x}(t), \vec{u}(t), t)$  are both convex in  $\vec{x}(t)$  and  $\vec{u}(t)$ . However, a necessary and sufficient condition as an alternative to Pontryagin's principle, is given by the **Hamilton-Jacobi-Bellman (HJB) equation**

$$\begin{aligned} \frac{\partial \mathcal{V}_{opt}(\vec{x}, t)}{\partial t} &= -\frac{\partial \mathcal{H}}{\partial \vec{x}} \\ \mathcal{V}_{opt}(\vec{x}(t_f), t_f) &= \mathcal{E}(\vec{x}(t_f), t_f) \end{aligned} \quad (3.17)$$

where  $\mathcal{V}(\vec{x}, t)$  is the continuous-time **cost-to-go**, also known as the **value function**, and replaces the costate or adjoint vector in the Hamiltonian expression. The **principle of optimality** states that regardless of the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision, i.e. for any  $t > t_0$

$$\mathcal{V}_{opt}(\vec{x}, t) = \min_{\vec{u}(t) \forall t \in [t, t_f]} \mathcal{J}(\vec{x}, \vec{u}, t, t_f) \quad (3.18)$$

However, the HJB equation requires the value function has a well-defined gradient and, moreover, that this gradient can be further differentiated with respect to time. Thus, the second-order partial derivatives of  $\mathcal{V}$  must exist which may not be true in general. The discrete-time version of the HJB is the Bellman equation and forms the basis of dynamic programming which is introduced in Appendix B of this textbook.

### Unconstrained LQ Optimal Control Problem

One of the fundamental OCPs is the **linear-quadratic OCP**, defined as having *linear* dynamics and a *quadratic* cost function/functional, with or without constraints. In particular, for continuous-time one has dynamics of the form

$$\dot{\vec{x}}(t) = A(t)x(t) + B(t)u(t) \quad (3.19)$$

and a cost functional of the form

$$\mathcal{J} = \frac{1}{2}x^T(t_f)Ex(t_f) + \frac{1}{2} \int_{t_0}^{t_f} \left( x^T(t)Q(t)x(t) + u^T(t)R(t)u(t) + 2x^T(t)S(t)u(t)dt \right) \quad (3.20)$$

This formulation contains **cost/weight matrices** where  $E$  is the **endpoint cost/weight matrix** or **terminal cost/weight matrix**,  $Q$  is the **state cost/weight matrix**,  $R$  is the **input cost/weight matrix**, and  $S$  is the **cross-cost/weight matrix**. The control designer selects the values of these matrices to balance the costs of large values of  $\vec{x}$  and  $\vec{u}$ . Furthermore, for the unconstrained case,  $E$ ,  $Q$ , and  $R$  are all symmetric, positive semi-definite matrices so that the cost functional is non-negative. The controller which solves the LQ OCP is called the **linear-quadratic regulator (LQR)**, i.e.  $\vec{u}^{\text{opt}}(t)$ . The term **regulator** denotes that this controller steers the system state to 0. However, to generalize this recall the robust servomechanism control introduced in the last chapter. With the servomechanism to augment the dynamics, one can easily obtain the **linear-quadratic tracker (LQT)**. Furthermore, it should be noted that  $A/F$ ,  $B/G$ ,  $Q$ ,  $R$ , and  $S$  can all vary with  $t$ .

The **unconstrained finite-horizon continuous-time LQ OCP** can be stated as

$$\begin{aligned} \vec{u}^{\text{opt}}(t) = \underset{u(t) \forall t \in [0, t_f]}{\text{argmin}} \quad & \mathcal{J} = \vec{x}^T(t_f)E\vec{x}(t_f) + \int_0^{t_f} \vec{x}^T(t)Q\vec{x}(t) + \vec{u}^T(t)R\vec{u}(t) + 2\vec{x}^T(t)S\vec{u}(t)dt \\ \text{subject to: } & \dot{\vec{x}}(t) = A\vec{x}(t) + B\vec{u}(t) \\ \text{initial condition: } & \vec{x}(0) = \vec{x}_0 \end{aligned} \quad (3.21)$$

where the unconstrained finite-horizon continuous-time LQR is the optimal control function,  $\vec{u}^{\text{opt}}(t)$ , which minimizes the quadratic cost functional,  $J$ . Recalling the generalized calculus of variations solution method, one can assign the following Hamiltonian for this continuous-time OCP.

$$\mathcal{H}(\vec{x}(t), \vec{u}(t), \vec{\lambda}(t), t) = \vec{x}^T(t)Q\vec{x}(t) + \vec{u}^T(t)R\vec{u}(t) + 2\vec{x}^T(t)S\vec{u}(t) + \vec{\lambda}^T(A\vec{x}(t) + B\vec{u}(t)) \quad (3.22)$$

which allows one to define the unconstrained condition equations as

$$\begin{cases} \dot{\vec{\lambda}}(t) &= -\frac{\partial \mathcal{H}^T}{\partial \vec{x}} = -Q \vec{x} - S \vec{u} - A^T \vec{\lambda} \\ 0 &= \frac{\partial \mathcal{H}}{\partial \vec{u}} = \vec{u}^T R + \vec{x}^T S + \vec{\lambda}^T B \\ \vec{\lambda}(t_f) &= \frac{\partial \mathcal{E}(t_f)}{\partial \vec{x}}^T = \vec{x}^T(t_f) E \end{cases} \quad (3.23)$$

for the LQ OCP.

To solve this OCP, assume the costate has a linear form, i.e.

$$\vec{\lambda}(t) = P(t) \vec{x}(t) \quad (3.24)$$

where  $P(t)$  is a symmetric matrix. Substituting this into the costate equations and including the state dynamics equation, one has

$$\begin{cases} \dot{\vec{x}} &= A \vec{x} + B \vec{u} \\ \frac{d}{dt} (P \vec{x}) = \dot{P} \vec{x} + P \dot{\vec{x}} &= -Q \vec{x} - S \vec{u} - A^T P \vec{x} \\ 0 &= \vec{u}^T R + \vec{x}^T S + \vec{x}^T P B \\ P(t_f) \vec{x}(t_f) &= E \vec{x}(t_f) \end{cases} \quad (3.25)$$

Next, substituting the first equation into the second equation results in

$$\dot{P} \vec{x} + PA \vec{x} + PB \vec{u} = -Q \vec{x} - S \vec{u} - A^T P \vec{x} \quad (3.26)$$

Then, rewriting the third equation, one has

$$\vec{u} = -R^{-1}(B^T P + S^T) \vec{x} \quad (3.27)$$

By substitution for  $\vec{u}$  into the newly derived equation, one has

$$\dot{P} \vec{x} + PA \vec{x} - PBR^{-1}(B^T P + S^T) \vec{x} = -Q \vec{x} + SR^{-1}(B^T P + S^T) \vec{x} - A^T P \vec{x} \quad (3.28)$$

By removing the common  $\vec{x}$  term and rearranging, one has

$$\dot{P} = -PA - A^T P + (PB + S) R^{-1}(B^T P + S^T) - Q \quad (3.29)$$

which is known as the **differential Riccati equation** and describes the dynamics of the continuous-time **Riccati matrix**,  $P(t)$ , and can be solved using the boundary condition on the costate, i.e.

$$P(t_f) = E \quad (3.30)$$

Thus, the Lagrangian multiplier problem has been reduced to a matrix-valued ODE which must be solved in reverse time from the end condition.

Finally, the **unconstrained finite-horizon continuous-time LQR** has the form

$$\vec{u}^{\text{opt}}(t) = -K(t) \vec{x}(t) \quad (3.31)$$

where

$$K(t) = R^{-1}(B^T P(t) + S^T) \quad (3.32)$$

which notably results in closed-loop state-space dynamics represented by

$$\dot{\vec{x}}(t) = (A - BK(t)) \vec{x}(t) \quad (3.33)$$

Finally, if one considers the **unconstrained infinite-horizon continuous-time LQ OCP**, the solution is given by the steady-state of the Riccati differential equation, i.e.

$$0 = -PA - A^T P + (PB + S) R^{-1}(B^T P + S^T) - Q \quad (3.34)$$

which is known as the **continuous-time algebraic Riccati equation (CARE)**.

### Infinite-Horizon Unconstrained Optimal Control

However, another approach to setting up an infinite-horizon OCP for this generalized feedback control system is to design a stabilizing LTI controller  $K(s)$  to minimize the input-to-output gain from  $\vec{d} \rightarrow \vec{e}$  by setting the cost functional  $J(\vec{u})$  for LTI systems as some system norm  $\|F_L(G, K)\|$ , i.e.

$$\begin{aligned} \vec{u}_{opt}(s) &= K_{opt}(s) \vec{y}(s) = \underset{u(s)}{\operatorname{argmin}} \|F_L(G, K)\| \\ &\text{subject to:} \\ \text{continuous dynamics: } & \dot{\vec{x}}(t) = \vec{x}(t) = A \vec{x}(t) + B_1 \vec{d}(t) + B_2 \vec{u}(t) \\ & \vec{e}(t) = C_1 \vec{x}(t) + D_{12} \vec{u}(t) \\ & \vec{y}(t) = C_2 \vec{x}(t) + D_{21} \vec{d}(t) \\ \text{constraints: } & K \text{ stabilizing} \end{aligned} \quad (3.35)$$

where  $F_L(G, K)$  defines a state-space representation as

$$\begin{aligned} \dot{\vec{x}}(t) &= A \vec{x}(t) + B_1 \vec{d}(t) + B_2 \vec{u}(t) \\ \vec{e}(t) &= C_1 \vec{x}(t) + D_{12} \vec{u}(t) \\ \vec{y}(t) &= C_2 \vec{x}(t) + D_{21} \vec{d}(t) \end{aligned} \quad (3.36)$$

This is often written more succinctly as

$$K_{opt}(s) = \underset{K \text{ stabilizing}}{\operatorname{argmin}} \|F_L(G, K)\| \quad (3.37)$$

This chapter of the textbook considers two types of system norms, the  $\mathcal{H}_2$ - and  $\mathcal{H}_\infty$ -norms, for this type of OCP.

## 3.2 Introductory Energy-, Time-, and Fuel-Optimal Control

Consider the **classical OCP**

$$\begin{aligned} \vec{u}_{opt}(t) = \underset{u(t) \forall t \in [t_0, t_f]}{\operatorname{argmin}} \quad \mathcal{J} = \mathcal{E}(\vec{x}(t_f), t_f) + \int_{t_0}^{t_f} \mathcal{L}(\vec{x}(t), \vec{u}(t), t) dt \\ \text{subject to:} \\ \text{continuous-time dynamics: } \dot{\vec{x}}(t) = f(\vec{x}(t), \vec{u}(t), t) \\ \text{initial conditions: } \vec{x}(t_0) - \vec{x}_0 = 0 \\ \text{control constraints: } |\vec{u}(t)| < u_{max} \end{aligned} \quad (3.38)$$

with  $t_0$  fixed and  $t_f$  free where  $\mathcal{E}$  is the endpoint cost,  $\mathcal{L}$  is the running cost or Lagrangian, and  $\mathcal{H} = \mathcal{L} + \vec{\lambda}^T f(\vec{x}(t), \vec{u}(t), t)$  is the corresponding Hamiltonian. This section will discuss some sub-types of OCPs that are applicable to many flight vehicle OCPs for guidance and control laws, namely, the energy, fuel, and time optimal control problems and will provide analysis of the solutions to these OCPs for LTI systems.

### Minimum-Energy Optimal Control Problem

A **minimum-energy OCP** is defined as

$$\begin{aligned} \vec{u}_{opt}(t) = \underset{u(t) \forall t \in [0, t_f]}{\operatorname{argmin}} \quad \mathcal{J} = \mathcal{E}(\vec{x}(t_f), t_f) + \int_0^{t_f} \vec{u}^T(t) \vec{u}(t) dt \\ \text{subject to: } \dot{\vec{x}}(t) = f(\vec{x}(t), \vec{u}(t), t) \\ \text{initial conditions: } \vec{x}(t_0) - \vec{x}_0 = 0 \\ \text{control constraints: } |\vec{u}(t)| < 1 \end{aligned} \quad (3.39)$$

where the integral term is the  $\mathcal{L}_2$ -norm of the control input signal. If one desires  $\vec{x}_f$  to be the origin and uses a quadratic terminal cost, i.e.  $\mathcal{E}(\vec{x}(t_f), t_f) = \vec{x}^T(t_f) E \vec{x}(t_f)$ , and linear dynamics, i.e.  $f(\vec{x}(t), \vec{u}(t), t) = A(t) \vec{x} + B(t) \vec{u}$ , then one has a special case of the LQR OCP with  $Q = 0$  and  $R = I$ .

Assuming that  $(A, B)$  is controllable for all  $t \in [0, t_f]$ , the solution to the minimum-energy LQR OCP is given by the differential Riccati equation

$$\dot{P}(t) = -P(t)A(t) - A^T(t)P(t) + P(t)B(t)B^T(t)P(t) \quad (3.40)$$

with endpoint condition

$$P(t_f) = E \quad (3.41)$$

where the optimal control is given by

$$u_{opt} = -B^T(t)P(t)\vec{x}(t) \quad (3.42)$$

with closed-loop state-space dynamics represented by

$$\dot{\vec{x}}(t) = \left( A(t) - B(t)B^T(t)P(t) \right) \vec{x}(t) \quad (3.43)$$

It should be noted if one requires the system to converge to some  $x(t_f) = x_f$ , then one can define

$$E = \text{diag}(\vec{E}) \quad (3.44)$$

and take the limit of the optimal control as this cost weight goes to  $\infty$ , i.e.

$$\lim_{\vec{E} \rightarrow \infty} u_{opt}(E, t) \quad (3.45)$$

Furthermore,  $\vec{u}$  may also be constrained, e.g.  $|\vec{u}_i| < u_{c,i} \forall i = 1, \dots, n_u$ . This forms a constrained optimization problem which can be solved using Pontryagin's principle for admissible control set.

### Minimum-Time Optimal Control Problem

A **minimum-time OCP** is defined as

$$\begin{aligned} \vec{u}_{opt}(t) &= \underset{u(t) \forall t \in [0, t_f]}{\operatorname{argmin}} \quad \mathcal{J} = \int_0^{t_f} dt = t_f \\ &\text{subject to: } \dot{\vec{x}}(t) = f(\vec{x}(t), \vec{u}(t), t) \\ &\text{initial conditions: } \vec{x}(t_0) - \vec{x}_0 = 0 \\ &\text{final conditions: } \vec{x}(t_f) - \vec{x}_c = 0 \\ &\text{control constraints: } \vec{u}(t) \in \mathcal{U} \end{aligned} \quad (3.46)$$

which corresponds to the Hamiltonian

$$\mathcal{H}(\vec{x}(t), \vec{u}(t), \vec{\lambda}(t), t) = 1 + \vec{\lambda}^T f(\vec{x}(t), \vec{u}(t), t) \quad (3.47)$$

This OCP generally requires use of Pontryagin's principle to solve as well as methods for determining admissible controls,  $\vec{u}(t) \in \mathcal{U}$ .

However, in this introductory section, consider linear dynamics, i.e.  $f(\vec{x}(t), \vec{u}(t), t) = A(t)\vec{x} + B(t)\vec{u}$ , a Hamiltonian

$$\mathcal{H}(\vec{x}(t), \vec{u}(t), \vec{\lambda}(t), t) = 1 + \vec{\lambda}^T A \vec{x} + \vec{\lambda}^T B \vec{u} \quad (3.48)$$

and a  $n_u$ -dimensional hypercube for independent control actuators as the admissible control set, i.e.

$$\mathcal{U} = \{\vec{u} \in \mathbb{R}^{n_u} : u_i \in [u_{i,min}, u_{i,max}], i = 1, \dots, n_u\} \quad (3.49)$$

Then, Pontryagin's principle states

$$\mathcal{H}(\vec{x}_{opt}(t), \vec{u}_{opt}(t), \vec{\lambda}_{opt}(t)) \leq \mathcal{H}(\vec{x}(t), \vec{u}(t), \vec{\lambda}(t)) \quad \forall t \in [0, t_f], \vec{u} \in \mathcal{U} \quad (3.50)$$

which can be simplified to

$$\vec{\lambda}_{opt}^T B \vec{u}_{opt} = \min_{\vec{u} \in \mathcal{U}} \vec{\lambda}^T B \vec{u} \quad \forall t \in [0, t_f] \quad (3.51)$$

Next, defining,  $\vec{b}_i$  as the  $i^{\text{th}}$  column of  $B$  and since each  $u_i$  can be chosen independently, one has

$$\sum_{i=1}^{n_u} \vec{\lambda}_{opt}^T b_i u_{i,opt} = \min_{u_{i,min} \leq u_i \leq u_{i,max}} \sum_{i=1}^{n_u} \vec{\lambda}^T \vec{b}_i u_i \quad \forall t \in [0, t_f] \quad (3.52)$$

Thus, one has the optimal control cases as

$$u_{i,opt}(t) = \begin{cases} u_{i,max} & \text{if } \vec{\lambda}_{opt}^T \vec{b}_i < 0 \\ ? & \text{if } \vec{\lambda}_{opt}^T \vec{b}_i = 0 \\ u_{i,min} & \text{if } \vec{\lambda}_{opt}^T \vec{b}_i > 0 \end{cases} \quad (3.53)$$

To solve for the ? element, recall that the costate equation is

$$\dot{\vec{\lambda}}_{opt}^T = -\frac{\partial \mathcal{H}}{\partial \vec{x}} = -A^T \vec{\lambda}_{opt} \quad (3.54)$$

or

$$\vec{\lambda}_{opt}(t) = \exp\left(A^T(t_{opt} - t)\right) \vec{\lambda}_{opt}(t_{opt}) \quad (3.55)$$

Thus,  $\vec{\lambda}_{opt}^T \vec{b}_i = \vec{\lambda}_{opt}^T(t_{opt}) \exp(A^T(t_{opt} - t)) \vec{b}_i$  is only = 0 over some time *interval* if it's zero for all  $t$  and its derivatives, i.e.

$$\vec{\lambda}_{opt}^T(t_{opt}) \vec{b}_i = \vec{\lambda}_{opt}^T(t_{opt}) A \vec{b}_i = \dots = \vec{\lambda}_{opt}^T(t_{opt}) A^{n_x-1} \vec{b}_i \quad (3.56)$$

While  $\vec{\lambda}_{opt}(t_{opt})$  cannot be 0 for  $\vec{x}(t_0) \neq \vec{x}_0$ , one requires each pair  $(A, \vec{b}_i)$  to be controllable also known as a **normal LTI system**.

Thus,  $\vec{u}_{opt}$  only takes values at the vertices of the hypercube  $\mathcal{U}$  and has a finite number of discontinuities, i.e. *switches*, between these values, and is unique. This result is known as the **bang-bang property** for linear systems. However, it can be shown that the assumption of normality for the bang-bang property of minimum time optimal control for a hypercube can be relaxed to a modified bang-bang principle for linear systems which states that while not every minimum-time optimal control is bang-bang, one can always select one that *is* bang-bang if  $\mathcal{U}$  is a convex polyhedron. Notably, one can also demonstrate a bang-bang property for normal linear-in-control systems.

The bang-bang property for LTI systems can be established without Pontryagin's principle by recalling the general solution to the LTI system as

$$\vec{x}(t) = \exp(A(t - t_0)) \vec{x}_0 + \int_{t_0}^{t_f} \exp(A(t - \tau)) B \vec{u}(\tau) d\tau \quad (3.57)$$

Then, for  $t \geq t_0$ , one has the *reachable set* from  $\vec{x}(t_0) = \vec{x}_0$  at time  $t$  as

$$\mathcal{R}^t(\vec{x}_0) = \left\{ \exp(A(t - t_0)) \vec{x}_0 + \int_{t_0}^{t_f} \exp(A(t - \tau)) B \vec{u}(\tau) d\tau : \vec{u} \in \mathcal{U}, t_0 \leq \tau \leq t \right\} \quad (3.58)$$

where

$$t_{opt} = \underset{t}{\operatorname{argmin}} \vec{x}_{opt} \in \mathcal{R}^t(\vec{x}_0) \quad (3.59)$$

Thus,  $\vec{x}_{opt}$  must occur on the boundary of  $\mathcal{R}^{t_{opt}}(\vec{x}_0)$  as if it was in the interior, one could reach it sooner. Noting that  $\mathcal{R}^{t_{opt}}(\vec{x}_0)$  is compact and convex, there exists a hyperplane that passes through  $\vec{x}_{opt}$  and contains  $\mathcal{R}^T(\vec{x}_0)$  on one side.

Choosing the normal vector to this hyperplane as  $\vec{\lambda}_{opt}^T(t_{opt})$ , one has

$$\vec{\lambda}_{opt}^T(t_{opt}) \vec{x}_{opt} \geq \vec{\lambda}_{opt}^T(t_{opt}) \vec{x} \quad \forall \vec{x} \in \mathcal{R}^{t_{opt}}(\vec{x}_0) \quad (3.60)$$

or, by definition

$$\int_{t_0}^{t_{opt}} \vec{\lambda}_{opt}^T(t_{opt}) \exp(A(t-\tau)) B \vec{u}_{opt}(\tau) d\tau \geq \int_{t_0}^{t_{opt}} \vec{\lambda}_{opt}^T(t_{opt}) \exp(A(t-\tau)) B \vec{u}(\tau) d\tau \quad (3.61)$$

$\forall \vec{u} \in \mathcal{U}$  from  $[t_0, t_{opt}]$

and noting  $\vec{\lambda}(\tau) = \exp(A^T(t_{opt} - \tau)) \vec{\lambda}_{opt}(t_{opt})$ , one has

$$\int_{t_0}^{t_{opt}} \vec{\lambda}_{opt}^T(\tau) B \vec{u}_{opt}(\tau) d\tau \geq \int_{t_0}^{t_{opt}} \vec{\lambda}_{opt}^T(\tau) B \vec{u}(\tau) d\tau \quad \forall \vec{u} \in \mathcal{U} \text{ from } [t_0, t_{opt}] \quad (3.62)$$

from which one can show the optimal control cases as previously.

### Minimum-Fuel Optimal Control Problem

A **minimum-fuel OCP** is defined as

$$\begin{aligned} \vec{u}_{opt}(t) &= \underset{u(t) \forall t \in [0, t_f]}{\operatorname{argmin}} \mathcal{J} = \int_0^{t_f} \sum_{i=1}^{n_u} |u_i(t)| dt \\ &\text{subject to: } \dot{\vec{x}}(t) = f(\vec{x}(t), \vec{u}(t), t) \\ &\text{initial conditions: } \vec{x}(t_0) - \vec{x}_0 = 0 \\ &\text{final conditions: } \vec{x}(t_f) - \vec{x}_c = 0 \\ &\text{control constraints: } |\vec{u}(t)| < 1 \end{aligned} \quad (3.63)$$

which corresponds to the Hamiltonian

$$\mathcal{H}(\vec{x}(t), \vec{u}(t), \vec{\lambda}(t), t) = \sum_{i=1}^{n_u} |\vec{u}_i(t)| + \vec{\lambda}^T f(\vec{x}(t), \vec{u}(t), t) \quad (3.64)$$

The integral is the  $\mathcal{L}_1$ -norm of the control input. Thus, a minimum-fuel optimal control is also known as  $\mathcal{L}_1$ -optimal control. This OCP generally requires use of Pontryagin's principle to solve as well as methods for determining admissible controls,  $\vec{u}(t) \in \mathcal{U}$ .

However, in this introductory section, consider linear dynamics, i.e.  $f(\vec{x}(t), \vec{u}(t), t) = A(t) \vec{x} + B(t) \vec{u}$ , a Hamiltonian

$$\mathcal{H}(\vec{x}(t), \vec{u}(t), \vec{\lambda}(t), t) = \sum_{i=1}^{n_u} |u_i(t)| + \vec{\lambda}^T A \vec{x} + \vec{\lambda}^T B \vec{u} \quad (3.65)$$

and a  $n_u$ -dimensional hypercube for independent control actuators as the admissible control set, i.e.

$$\begin{aligned} \mathcal{U} &= \{ \vec{u} \in \mathbb{R}^{n_u} : u_i \in [u_{i,min}, u_{i,max}], i = 1, \dots, n_u \} \\ u_{i,min} &< 0 \\ u_{i,max} &> 0 \end{aligned} \quad (3.66)$$

Then, Pontryagin's principle states

$$\mathcal{H}(\vec{x}_{opt}(t), \vec{u}_{opt}(t), \vec{\lambda}_{opt}(t)) \leq \mathcal{H}(\vec{x}(t), \vec{u}(t), \vec{\lambda}(t)) \quad \forall t \in [0, t_f], \vec{u} \in \mathcal{U} \quad (3.67)$$

which can be simplified to

$$\sum_{i=1}^{n_u} |u_{i,opt}| + \vec{\lambda}_{opt}^T B \vec{u}_{opt} = \min_{\vec{u} \in \mathcal{U}} \sum_{i=1}^{n_u} |u_i| + \vec{\lambda}^T B \vec{u} \quad \forall t \in [0, t_f] \quad (3.68)$$

Next, defining,  $\vec{b}_i$  as the  $i^{\text{th}}$  column of  $B$  and since each  $u_i$  can be chosen independently, one has

$$\sum_{i=1}^{n_u} |u_{i,opt}| + \vec{\lambda}_{opt}^T b_i u_{i,opt} = \min_{u_{i,min} \leq u_i \leq u_{i,max}} \sum_{i=1}^{n_u} |u_i| + \vec{\lambda}^T \vec{b}_i u_i \quad \forall t \in [0, t_f] \quad (3.69)$$

Thus, one has the optimal control cases as

$$u_{i,opt}(t) = \begin{cases} u_{i,max} & \text{if } \vec{\lambda}_{opt}^T \vec{b}_i \leq -1 \\ 0 & \text{if } -1 \leq \vec{\lambda}_{opt}^T \vec{b}_i \leq 1 \\ u_{i,min} & \text{if } \vec{\lambda}_{opt}^T \vec{b}_i \geq 1 \end{cases} \quad (3.70)$$

This result is known as the **bang-off-bang property** for LTI systems as there may exist time intervals where  $u_{i,opt}(t) = 0$  for the minimum-fuel OCP. However, in general, the exact time equation for  $\vec{u}_{opt}$  is not provided by the previous analysis as one does not directly know the dependence of the costate,  $\vec{\lambda}$ , on the state,  $\vec{x}$ , and there may be any number of minimum-fuel solutions depending on the state. Thus, minimum-fuel OCPs are more difficult to compute analytically although there are explicit solutions to second-order systems.

Lastly, due to the similar nature of minimum-time and minimum-fuel, these two conditions can also be linearly combined into a **minimum-time/fuel OCP** defined as

$$\begin{aligned} \vec{u}_{opt}(t) &= \underset{u(t) \forall t \in [0, t_f]}{\operatorname{argmin}} \quad \mathcal{J} = \int_0^{t_f} k + \sum_{i=1}^{n_u} |u_i(t)| dt \\ &\text{subject to: } \dot{\vec{x}}(t) = f(\vec{x}(t), \vec{u}(t), t) \\ &\text{initial conditions: } \vec{x}(t_0) = \vec{x}_0 \\ &\text{final conditions: } \vec{x}(t_f) = \vec{x}_c \\ &\text{control constraints: } |\vec{u}(t)| < 1 \end{aligned} \quad (3.71)$$

which corresponds to the Hamiltonian

$$\mathcal{H}(\vec{x}(t), \vec{u}(t), \vec{\lambda}(t), t) = k + \sum_{i=1}^{n_u} |\vec{u}_i(t)| + \vec{\lambda}^T f(\vec{x}(t), \vec{u}(t), t) \quad (3.72)$$

where  $k$  is the relative cost or weight of time with respect to fuel consumption. This solution for LTI systems also results in a bang-off-bang property.

### Path Planning via Dubins Path

The planning system for flight vehicles typically solves two coupled problems, namely the mission plan and the path plan. The mission planning at the most general level consists of searching for unknown targets

and/or optimally assigning targets, either as individual targets or sequences of targets, to all available flight vehicles in the mission. The selection or assignment of the targets or search paths for multiple flight vehicles typically involves the computation of a path cost for each vehicle, thus coupling the general mission planning problem with the path planning problem. In general, the optimal path can be solved via the **Euclidean shortest-path problem** which attempts to find the shortest-path within a continuously-valued Euclidean space, as opposed to a shortest-path problem based on discrete graphs which can use dynamic programming, e.g. Dijkstra's method, which naturally handles obstacles. For vehicles, the two-dimensional Euclidean shortest-path problem is also known as the **Dubins path problem** which can be solved in 2D and 3D using Pontryagin's principle with added constraints on the control.

For the two-dimensional Dubins path problem, one assumes the two-dimensional **Dubins vehicle** with dynamics

$$\dot{\vec{x}} = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\psi} \end{bmatrix} = f(\vec{x}, \vec{u}) = \begin{bmatrix} V \cos \psi \\ V \sin \psi \\ u \end{bmatrix} \quad (3.73)$$

where  $V$  is the *constant* velocity of the vehicle,  $(x, y)$  is the position of the vehicle,  $\psi$  is the heading angle, and  $u$  is the turn rate control input. The **instantaneous curvature** which is constrained by

$$\frac{-V}{R_{min}} \leq u \leq \frac{V}{R_{min}} \quad (3.74)$$

where  $R_{min}$  is the minimum radius of curvature. The objective for a Dubins path is starting at  $t = 0$  at some state  $(x_0, y_0, \psi_0)$ , one wishes to achieve some other state,  $(x_f, y_f, \psi_f)$ , at time  $t = t_f$  in the shortest amount of time.

Thus, the **2D Dubins path OCP** can be formally stated as

$$\begin{aligned} \vec{u}^{opt}(t) &= \underset{u(t) \forall t \in [0, t_f]}{\operatorname{argmin}} J = \int_0^{t_f} dt = t_f \\ &\text{subject to:} \\ &\text{dynamics } \begin{bmatrix} \dot{x} \\ i \end{bmatrix} = \begin{bmatrix} f(\vec{x}, \vec{u}) \\ 1 \end{bmatrix} \\ &\text{initial condition } \begin{bmatrix} x(0) \\ y(0) \\ \psi(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \\ \psi_0 \end{bmatrix} \\ &\text{final condition } \begin{bmatrix} x(t_f) \\ y(t_f) \\ \psi(t_f) \end{bmatrix} = \begin{bmatrix} x_f \\ y_f \\ \psi_f \end{bmatrix} \\ &\text{constraints } |u(t)| \leq \frac{V}{R_{min}} \end{aligned} \quad (3.75)$$

which corresponds to the Hamiltonian with  $\vec{x}' = \begin{bmatrix} \vec{x}^T \\ t \end{bmatrix}$  as

$$\mathcal{H}(\vec{x}'(t), \vec{u}(t), \vec{\lambda}(t)) = \lambda^T f(\vec{x}'(t), \vec{u}(t), \vec{\lambda}(t)) \quad (3.76)$$

Then, the costate equations can be written as

$$\dot{\vec{\lambda}} = \begin{bmatrix} \dot{p} \\ \dot{q} \\ \dot{\beta} \\ \dot{e} \end{bmatrix} = -\frac{\partial \mathcal{H}}{\partial \vec{x}'} = -\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -V \sin \psi & V \cos \psi & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p \\ q \\ \beta \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ pV \sin \psi - qV \cos \psi \\ 0 \end{bmatrix} \quad (3.77)$$

with final values for the costate vector as

$$\vec{\lambda}(t_f) = \begin{bmatrix} p(t_f) \\ q(t_f) \\ \beta(t_f) \\ e(t_f) \end{bmatrix} = \frac{\partial \mathcal{E}^T}{\partial \vec{x}'} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (3.78)$$

and with arbitrary, but not all zero, initial values.

Thus,  $p$ ,  $q$ , and  $e$  are constant on  $[0, t_f]$ . Thus, by defining  $r = \sqrt{p^2 + q^2} \geq 0$ , one can write  $p = r \cos \phi$  and  $q = r \sin \phi$  which determines an angle  $\phi \in [0, 2\pi)$  such that  $\tan \phi = \frac{q}{p}$  which allows one to rewrite the third costate equation as

$$\dot{\beta} = rV \sin(\psi - \phi) \quad (3.79)$$

Furthermore, one has for the bounded control for the minimum-time OCP,  $\frac{-V}{R_{min}} \leq u \leq \frac{V}{R_{min}}$ , one requires that

$$\begin{cases} u(t) = \frac{-V}{R_{min}} & \text{for } \beta < 0 \\ u(t) = 0 & \text{for } \beta = 0 \\ u(t) = \frac{V}{R_{min}} & \text{for } \beta > 0 \end{cases} \quad (3.80)$$

Hence, if  $\beta = 0$ , then  $\dot{\beta} = 0$  and  $\psi = \phi$  or  $\psi = \phi + \pi$  and the path is a line segment with direction  $\phi$ . Otherwise if  $\beta \neq 0$ , then  $u$  must take its maximum or minimum values, i.e.  $u = \pm \frac{V}{R_{min}}$  and the path is an arc of circle of radius  $R$ . This demonstrates that any optimal path is the concatenation of arcs of circles of radius  $R$  and line segments all parallel to some fixed direction  $\phi$ .

## References

For more information, please refer to the following

- D. Liberzon, “4.4.2 Bang-bang principle for linear systems,” in *Calculus of Variations and Optimal Control Theory: A Concise Introduction*, Princeton University Press, 2009
- R. Sarkar, D. U. Patil, and I. N. Kar, “Computation of Time-fuel Optimal Control for a Class of LTI System,” in *Proceedings of the 2019 Fifth Indian Control Conference (ICC)*, January 2019
- J.-D. Boissonnat, A. Cerezo, and J. Leblond, “Shortest path of bounded curvature in the plane,” in *Proceedings of the 1992 IEEE International Conference on Robotics and Automation*, May 1992

### 3.3 Convex Optimization in LTI Control

#### Convexity

Many OCPs can be set up as convex optimization problems which uses the concept of convexity. To explain this concept, first consider two points in vector space  $\mathcal{V}$ , i.e.  $\vec{v}_1 \in \mathcal{V}$  and  $\vec{v}_2 \in \mathcal{V}$ , then the line segment  $L(\vec{v}_1, \vec{v}_2)$  between them contains the set of points defined by

$$L(\vec{v}_1, \vec{v}_2) = \{\vec{v} \in \mathcal{V} : \vec{v} = \mu \vec{v}_1 + (1 - \mu) \vec{v}_2 \text{ for some } \mu \in [0, 1]\} \quad (3.81)$$

Next, let  $Q \subset \mathcal{V}$  be nonempty, then  $Q$  is a **convex set** of  $\mathcal{V}$  if for any  $\vec{v}_1, \vec{v}_2 \in Q$ ,  $L(\vec{v}_1, \vec{v}_2) \subset Q$ . One can also consider the expression  $\vec{v} = \mu \vec{v}_1 + (1 - \mu) \vec{v}_2$  as a weighed average, i.e. if  $\mu_1, \mu_2 \in [0, 1]$  with  $\mu_1 + \mu_2 = 1$ , then  $\vec{v} = \mu_1 \vec{v}_1 + \mu_2 \vec{v}_2$  which can be generalized to an average of  $n$  points as  $\vec{v} = \mu_1 \vec{v}_1 + \dots + \mu_n \vec{v}_n$  with  $\mu_1, \dots, \mu_n \in [0, 1]$  and  $\mu_1 + \dots + \mu_n = 1$ .

Furthermore, the set comprised of all weighted averages of the points  $\vec{v}_1, \dots, \vec{v}_n$  is called its convex hull,  $co(\{\vec{v}_1, \dots, \vec{v}_n\})$ . More generally, given a set  $Q$  one can define its **convex hull**,  $co(Q)$ , by the set  $\{\vec{v} \in \mathcal{V} : \text{there exists } n \text{ and } \vec{v}_1, \dots, \vec{v}_n \in Q \text{ such that } \vec{v} \in co(\{\vec{v}_1, \dots, \vec{v}_n\})\}$ , i.e. the convex hull of  $Q$  is the collection of all possible weighted averages of points in  $Q$ . Furthermore, one can show

- the subset  $Q \subset co(Q)$  is satisfied
- the convex hull  $co(Q)$  is convex
- $co(Q) = co(co(Q))$
- a set  $Q$  is convex if and only if  $co(Q) = Q$

where it should also be noted by definition, the intersection of convex sets is always convex.

Lastly, a set  $Q \subset \mathcal{V}$  is called a **cone** if it is closed under positive scalar multiplication, i.e. if

$$\vec{v} \in Q \text{ implies } t \vec{v} \in Q \text{ for every } t > 0 \quad (3.82)$$

thus, subspaces are cones, but cones are a broader set, e.g. the half-line  $C_v = \{\alpha \vec{v} : \alpha > 0\}$  for a fixed vector  $\vec{v}$ . Furthermore, a **convex cone** are cones closed under addition, i.e.  $\vec{v}_1, \vec{v}_2 \in Q$  implies  $\vec{v}_1 + \vec{v}_2 \in Q$ .

#### Cone Programming

Many generalized OCPs can be formulated as **cone programs (CF)**  
which have an order  $n$ .

#### Semidefinite Programming

Many generalized OCPs can be formulated as **semidefinite programs (SDP)**, a type of convex optimization which use LMI constraints on the solution space  $X$ , e.g. ARIs. The general form of an SDP can be formulated as

$$\begin{aligned} X^{opt} &= \underset{X \in \mathcal{X}}{\operatorname{argmin}} \quad c(X) \\ \text{subject to: } F(X) &\leq Q \\ &X \in \mathcal{X} \end{aligned} \quad (3.83)$$

where  $c(X)$  is a *linear* functional on the vector space  $\mathcal{X}$ , and is also known as a type of **linear objective problem**.

Note that for this SDP, the LMI constraint must be feasible. Thus, the simplest SDP may be formulated as a **feasibility problem** if there exists  $X \in \mathcal{X}$  satisfying  $F(X) < Q$  which can also be recast as the linear objective problem

$$\begin{aligned} J &= \inf t \\ \text{subject to: } &F(X) - tI \leq Q \end{aligned} \tag{3.84}$$

where if and only if  $J < 0$ , then the LMI  $F(X) < Q$  is feasible. Thus, the feasibility question can be made a part of the SDP and the focus of solving SDPs is on solving linear objective problems with LMI constraints.

To derive solutions for SDPs, note that the set  $C = \{X \in \mathcal{X} : F(X) < Q\}$  is a **convex set** in  $\mathcal{X}$ . As a proof, let  $X_1, X_2 \in C$  which means one satisfies  $F(X_1) < Q$  and  $F(X_2) < Q$ . Then, considering any point  $X_3 \in L(X_1, X_2)$ , i.e.  $X_3 = \mu X_1 + (1 - \mu)X_2$  for some value  $\mu \in [0, 1]$ , using the linearity of  $F()$ , one has

$$F(X_3) = \mu F(X_1) + (1 - \mu)F(X_2) < \mu Q + (1 - \mu)Q = Q \tag{3.85}$$

where the inequality follows from the fact that positive definite matrices are convex cones. Therefore,  $X_3 \in C$ .

This property generalizes to any convex optimization problem and ensures these problems can be solved globally, not just locally, by numerical search algorithms. At first glance, one may think that the minimum if it exists must lie on the boundary of the feasible set which would restrict the search to the boundary which is the case for **linear programming**. However, the boundary in general SDPs is complicated, thus, **interior point methods** are favored in practice. As an example convex optimization method, suppose one currently has the point  $X_n$  in the feasible set. An immediate consequence is that one only needs to keep the set

$$\{X \in C : c(X) \leq c(X_n)\} \tag{3.86}$$

for the remaining search for the global minimum. This amounts to setting  $C$  with a half-space, thus one can progressively shrink the feasibility region to zero, provided one is able to successively generate a “good” feasible point, e.g.  $X_{n+1}$ .

Many optimization methods are based on this principle with one of the simplest known as the **ellipsoid algorithm** that alternates between “cutting” and bounding the resulting set by an ellipsoid where  $X_{n+1}$  would be the center of such ellipsoid. More efficient methods for SDPs are based on **barrier functions** to impose the feasibility constraint. Here the idea is to minimize the function

$$c(X) + \alpha\phi(X) \tag{3.87}$$

where  $\alpha > 0$  and the barrier function  $\phi(X)$  is convex and approaches infinity on the boundary of the feasible set, e.g. for the set  $C$

$$\phi(X) = -\log(\det[Q - F(X)]) \tag{3.88}$$

can serve as a barrier function. Thus, provided one starts from a feasible point, the minimization can be globally solved by unconstrained optimization methods, e.g. Newton’s method. By successively reducing the weight of the barrier function, an iteration is produced which can be shown to converge to the global minimum with computational complexity of polynomial growth with problem size characterized by the dimension of  $\mathcal{X}$  and constraint set  $\mathbb{H}^n$ . Additional details on convex optimization algorithms are beyond the scope of this textbook.

### Bounded Real Lemma

As will be shown, solutions to infinite-horizon OCPs can be found by setting up algebraic Riccati equations (ARE) of the general form

$$A^T P + PA + Q + PRP = 0 \quad (3.89)$$

where typically, one is required to find a particular solution  $P = P^T \in \mathbb{R}^{n_x \times n_x}$  such that  $A + RP$  is stable, i.e. all its eigenvalues have strict negative real part. For such problems, one can define the Hamiltonian matrix of the ARE of  $(A, Q, R)$  as

$$H = \begin{bmatrix} A & R \\ -Q & -A^T \end{bmatrix} \quad (3.90)$$

Thus, the ARE becomes

$$\begin{bmatrix} P & -I \end{bmatrix} H \begin{bmatrix} I \\ P \end{bmatrix} = 0 \quad (3.91)$$

where the eigenvalue decomposition of  $H$  is key to solving the ARE. In this case, one can state that if  $A$ ,  $Q = Q^T$  and  $R = R^T$  are given,  $H$  has no purely imaginary axis eigenvalues, and  $R \geq 0$  or  $R \leq 0$ , and  $(A, R)$  is stabilizable, then the ARE of  $(A, Q, R)$  has a unique solution  $P = P^T$  such that  $A + RP$  is stable.

However, for more advanced control synthesis, one can instead use **algebraic Riccati inequality (ARI)** which is defined as

$$A^T P + PA + Q + PRP < 0 \quad (3.92)$$

which is a type of linear matrix inequality which will be discussed in the following section. For example, consider the inequality

$$A^T P + PA + C^T C + \gamma^{-2} PBB^T P < 0 \quad (3.93)$$

which is quadratic in  $P \in \mathbb{S}^n > 0$ . However, rewriting this as

$$\begin{bmatrix} A^T P + PA + C^T C + \gamma^{-2} PBB^T P & 0 \\ 0 & -\gamma^{-2} I \end{bmatrix} < 0 \quad (3.94)$$

by the Schur Complement Lemma, one can form the equivalent LMI for  $P$ .

$$\begin{bmatrix} A^T P + PA + C^T C & PB \\ B^T P & -\gamma^2 I \end{bmatrix} < 0 \quad (3.95)$$

or

$$\begin{bmatrix} A^T P + PA & PB \\ B^T P & \gamma^{-2} I \end{bmatrix} < \begin{bmatrix} -C^T C & 0 \\ 0 & 0 \end{bmatrix} \quad (3.96)$$

which is an LMI in  $P$  as the left side can be assigned as  $F(P) : \mathbb{S}^n \rightarrow \mathbb{S}^n$  and the right side can be assigned as  $Q \in \mathbb{S}$ .

In this case, one can relate solutions to AREs and ARIs for OCPs. One of the important lemmas between AREs/ARIs is known as the **Bounded Real Lemma**, also known as the **Kalman-Yacubovich-Popov (KYP) lemma**, which can be stated as follows. Consider the following LTI system,  $F_L(G, K)$ :

$$\begin{aligned} \vec{x}(t) &= A_L \vec{x}(t) + B_1 \vec{d}(t) \\ \vec{e}(t) &= C_1 \vec{x}(t) \end{aligned} \quad (3.97)$$

and let  $\gamma > 0$  be given. Then, the following three statements are equivalent.

1.  $F_L(G, K)$  is stable, i.e.  $A_L$  is stable, and  $\|F_L(G, K)\|_\infty < \gamma^2$
2. There exists a unique  $P_1 \geq 0$  such that  $A_L + \gamma^{-2}P_1B_1B_1^TP_1$  is stable and satisfies the ARE:

$$A^T P_1 + P_1 A + C_1^T C_1 + \gamma^{-2} P_1 B_1 B_1^T P_1 = 0 \quad (3.98)$$

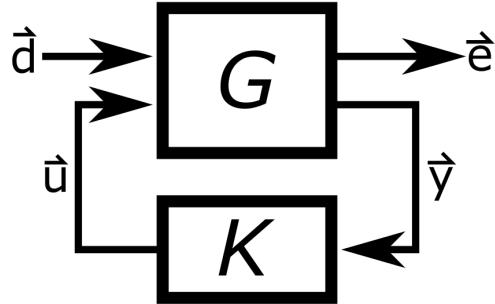
3. There exists  $P_2 > 0$  satisfying the strict ARI:

$$A^T P_2 + P_2 A + C_1^T C_1 + \gamma^{-2} P_2 B_1 B_1^T P_2 < 0 \quad (3.99)$$

where to solve ARIs for  $P_2$  requires **semidefinite programming (SDP)**, a class of convex optimization, which will be discussed in the following subsection and then applied to both the  $\mathcal{H}_2$  and  $\mathcal{H}_\infty$  OCPS in later sections.

### Stabilizing State Feedback as LMI

Consider the generalized feedback control system



with generalized plant,  $G(s)$ :

$$\begin{aligned} \dot{\vec{x}}(t) &= A \vec{x}(t) + B_1 \vec{d}(t) + B_2 \vec{u}(t) \\ \vec{e}(t) &= C_1 \vec{x}(t) + D_{11} \vec{d}(t) + D_{12} \vec{u}(t) \end{aligned} \quad (3.100)$$

and a state feedback control policy for  $K(s)$  as

$$\vec{u}(t) = D_K \vec{x}(t) \quad (3.101)$$

which results in a closed-loop system

$$\begin{aligned} \dot{\vec{x}}(t) &= (A + B_2 D_K) \vec{x}(t) + B_1 \vec{d}(t) \\ \vec{e}(t) &= (C_1 + D_{12} D_K) \vec{x}(t) + D_{11} \vec{d}(t) \end{aligned} \quad (3.102)$$

which is stable if and only if  $A_L = A + B_2 D_K$  is stable.

To construct an LMI for constructing a stabilizing  $D_K$ , one can enforce this stable condition by a matrix Lyapunov inequality

$$A_L P + P A_L^T < 0 \quad (3.103)$$

which has a solution  $P > 0$ . Substituting for  $A_L$ , one has

$$(A + B_2 D_K)P + P(A + B_2 D_K)^T < 0 \quad (3.104)$$

which can be expanded as

$$AP + PA^T + B_2(D_K P) + (PD_K^T)B_2^T < 0 \quad (3.105)$$

Then, substituting  $Y = D_K P$ , then one has the LMI in  $P$  and  $Y$  as

$$\begin{bmatrix} A & B_2 \end{bmatrix} \begin{bmatrix} P \\ Y \end{bmatrix} - \begin{bmatrix} X & Y^T \end{bmatrix} \begin{bmatrix} A^T \\ B_2^T \end{bmatrix} < 0 \quad (3.106)$$

which can be solved for  $P$  and  $Y$  and one can obtain the stabilizing feedback controller as  $D_K = YP^{-1}$ . It should be noted that this process can be expanded to general LMI characterizations for non-state feedback controllers  $D_K$ .

## 3.4 $\mathcal{H}_2$ Optimal Control

### Unconstrained Continuous-Time $\mathcal{H}_2$ OCP

The objective of  $\mathcal{H}_2$  optimal control is to find  $K$  that minimizes the  $\mathcal{H}_2$ -norm of the generalized feedback control system, i.e.

$$K^{opt} = \underset{K \text{ stabilizing}}{\operatorname{argmin}} \|F_L(G, K)\|_2 \quad (3.107)$$

where  $F_L(G, K)$  defines a state-space model as

$$\begin{aligned} \dot{\vec{x}}(t) &= A\vec{x}(t) + B_1\vec{d}(t) + B_2\vec{u}(t) \\ \vec{e}(t) &= C_1\vec{x}(t) + D_{12}\vec{u}(t) \\ \vec{y}(t) &= C_2\vec{x}(t) + D_{21}\vec{d}(t) \end{aligned} \quad (3.108)$$

where  $D_{11} = 0$  for a finite  $\mathcal{H}_2$ -norm and  $D_{22} = 0$  without loss of generality for ease of notation. Recall that for a solution to the  $\mathcal{H}_2$  OCP to exist,  $(A, B_2)$  must be stabilizable and  $(A, C_2)$  must be detectable.

There are two groups to which all  $\mathcal{H}_2$  OCPs can be divided, regular and singular. A **regular  $\mathcal{H}_2$  OCP** additionally assumes that (1) for the LTI plant state-space model  $(A, B_2, C_1, D_{12})$  there are no zeros on the  $j\omega$  axis and  $(D_{12}^T D_{12})^{-1}$  exists, i.e. injective, and (2) for the LTI plant state-space model  $(A, B_1, C_2, D_{21})$  there are no zeros on the  $j\omega$  axis and  $(D_{21} D_{21}^T)^{-1}$  exists, i.e. surjective. For regular  $\mathcal{H}_\infty$  OCPs one can find a *unique* solution, whereas a sub-optimal solution must be found for singular  $\mathcal{H}_2$  OCPs using a modification of the semidefinite programming approach shown here.

Furthermore, note that by definition of the  $\mathcal{H}_2$ -norm in the time domain with  $\vec{d}(t)$  constrained as impulse inputs, one can alternatively state the  $\mathcal{H}_2$  cost function as

$$\mathcal{J}(\vec{x}, \vec{u}) = \frac{1}{2} \int_0^\infty \vec{e}^T \vec{e} d\tau \quad (3.109)$$

and by substitution, one has

$$\mathcal{J}(\vec{x}, \vec{u}) = \frac{1}{2} \int_0^\infty (C_1 \vec{x}(t) + D_{12} \vec{u}(t))^T (C_1 \vec{x}(t) + D_{12} \vec{u}(t)) d\tau \quad (3.110)$$

$$\mathcal{J}(\vec{x}, \vec{u}) = \frac{1}{2} \int_0^\infty \vec{x}(t)^T C_1^T C_1 \vec{x}(t) + 2\vec{x}(t)^T C_1^T D_{12} \vec{u}(t) + \vec{u}(t)^T D_{12}^T D_{12} \vec{u}(t) d\tau \quad (3.111)$$

which is simply a more generalized form of the infinite-horizon continuous-time LQ OCP with  $Q = C_1^T C_1$ ,  $S = C_1^T D_{12}$ , and  $R = D_{12}^T D_{12}$ .

Thus, the infinite-horizon LQR OCP is the state feedback  $\mathcal{H}_2$  OCP, i.e.  $C_2 = I$ ,  $B_1 = \text{diag}(\vec{x}_0)$  and  $D_{21} = 0$ . This chapter of the textbook focuses on the synthesis for state feedback  $\mathcal{H}_2$  OCP for which one can simplify the problem to

$$\vec{y}(t) = \vec{x} \quad (3.112)$$

and

$$K : \vec{u} = D_k \vec{x} \quad (3.113)$$

and is always a regular  $\mathcal{H}_2$  OCP. However, output feedback  $\mathcal{H}_2$  OCPs may be regular or singular. However, the observer feedback  $\mathcal{H}_2$  OCP as a regular type allows one to apply the separation principle to the problem to obtain the LQR and the **linear-quadratic estimator (LQE)**. Also, it should be noted that when  $\vec{d}$  are modeled as Gaussian noise, this type of problem is also known as the **linear-quadratic-Gaussian (LQG) OCP**, a fundamental OCP for stochastic dynamical systems, and is addressed in a subsequent chapters.

### ARE Synthesis for State Feedback $\mathcal{H}_2$ OCP

Recall the solution to the finite-horizon continuous-time LQR OCP used the **Riccati differential equation**, i.e.

$$\dot{P} = -PA - A^T P + (PB + S) R^{-1} (B^T P + S^T) - Q \quad (3.114)$$

Furthermore, if one considers the **unconstrained infinite-horizon continuous-time LQ OCP** which sets  $t_f = \infty$ , the **unconstrained infinite-horizon continuous-time LQR** requires the steady-state solution of the Riccati differential equation, i.e.

$$0 = PA + A^T P - (PB + S) R^{-1} (B^T P + S^T) + Q \quad (3.115)$$

which is also known as the **continuous algebraic Riccati equation (CARE)** which can be solved using standard algorithms. Thus, the optimal control is given by

$$\vec{u}(t) = D_K \vec{x}(t) = -R^{-1} (B^T P + S^T) \vec{x}(t) \quad (3.116)$$

which notably results in closed-loop dynamics represented by

$$\dot{\vec{x}}(t) = (A + BD_K) \vec{x}(t) \quad (3.117)$$

### SDP Synthesis for State Feedback $\mathcal{H}_2$ OCP

To setup the  $\mathcal{H}_2$  OCP as a SDP, one requires the following result. Consider a strictly proper state-space LTI system,  $G$ , i.e.  $D = 0$ . Then, the state matrix,  $A$ , is stable and  $\|G\|_2 < 1$  if and only if, there exists a  $W \in \mathbb{S}^{n_x}$  such that

$$\text{Tr} [CWC^T] < 1 \quad (3.118)$$

and

$$AW + WA^T + BB^T < 0 \quad (3.119)$$

As a necessary proof, assume  $A$  is stable and  $\|G\|_2 < 1$ , then, by definition of the  $\mathcal{H}_2$ -norm in terms of the controllability gramian  $W_C$ , one has

$$\|G\|_2 = \text{Tr} [CW_C C^T] < 1 \quad (3.120)$$

Next, consider the perturbed expression for  $\epsilon > 0$

$$W(\epsilon) = \int_0^\infty e^{At} (BB^T + \epsilon I) e^{At} dt \quad (3.121)$$

where it should be noted that  $W(\epsilon) > 0$  for  $\epsilon > 0$ . Furthermore, as  $W(\epsilon)$  is a continuous function of  $\epsilon$  and  $W(\epsilon) = W_C$  for  $\epsilon = 0$ . By continuity, one has  $\text{Tr} [CWC^T] < 1$  for some  $\epsilon > 0$ . Thus, this matrix satisfies the matrix Lyapunov equation

$$AW + WA^T + (BB^T + \epsilon I) = 0 \quad (3.122)$$

or

$$AW + WA^T + BB^T = -\epsilon I < 0 \quad (3.123)$$

As a sufficient proof, assume  $W > 0$  such that

$$\text{Tr} [CW_C C^T] < 1 \quad (3.124)$$

and

$$AW + WA^T + BB^T < 0 \quad (3.125)$$

Thus, as  $BB^T > 0$ ,  $AW + WA^T < 0$  so that  $A$  is stable. Recalling the controllability gramian satisfies

$$AW_C + W_C A^T + BB^T = 0 \quad (3.126)$$

and using a technical result that solutions of the Lyapunov equality lower bound solutions of the corresponding Lyapunov inequality, one has

$$W_C \leq W \quad (3.127)$$

which implies

$$\text{Tr} [CW_C C^T] \leq \text{Tr} [CWC^T] < 1 \quad (3.128)$$

or, by definition

$$\|G\|_2 < 1 \quad (3.129)$$

Thus, by the previous  $\mathcal{H}_2$ -norm and stability lemma with an additional scaling argument with  $\gamma$ , one can state that there exists a  $D_K \in \mathbb{R}^{n_u \times n_y}$  such that  $A_L$  is stable and satisfies  $\|F_L(G, K)\|_2 < \gamma$  if and only if there exists  $P > 0$  and  $D_K$  such that

$$A_L P + PA_L^T + B_L B_L^T < 0 \quad (3.130)$$

and

$$\text{Tr} [C_L P C_L^T] < \gamma^2 \quad (3.131)$$

Substituting for the closed-loop matrices, one has

$$(A + B_2 D_K)P + P(A + B_2 D_K)^T + B_1 B_1^T < 0 \quad (3.132)$$

and

$$\text{Tr} [(C_1 + D_{12}D_K)P(C_1 + D_{12}D_K)^T] < \gamma^2 \quad (3.133)$$

which is a nonlinear matrix inequality in  $(P, D_K)$  due to the bilinear terms.

However, one can define a change of variable from  $D_K$  to

$$Q = D_K P \quad (3.134)$$

which allows one to write

$$\begin{bmatrix} A & B_2 \end{bmatrix} \begin{bmatrix} P \\ Q \end{bmatrix} + \begin{bmatrix} P & Q \end{bmatrix} \begin{bmatrix} A^T \\ B_2^T \end{bmatrix} + B_1 B_1^T < 0 \quad (3.135)$$

which is now an LMI in variables  $(P, Q)$ , however

$$\text{Tr} [(C_1 + D_{12}QP^{-1})P(C_1 + D_{12}QP^{-1})^T] < \gamma^2 \quad (3.136)$$

or

$$\text{Tr} [(C_1 P + D_{12}Q)P^{-1}(C_1 P + D_{12}Q)^T] < \gamma^2 \quad (3.137)$$

is not a convex condition, but one can convexify this inequality by introducing a **slack variable**,  $R \in \mathbb{S}^{n_e}$ , based on the following linear algebra fact: if  $M_1 \leq M_2$ , then  $\text{Tr}[M_1] \leq \text{Tr}[M_2]$ .

In this case, one can write the trace inequality above as two separate inequalities

$$\begin{aligned} (C_1 P + D_{12}Q)P^{-1}(C_1 P + D_{12}Q)^T &< R \\ \text{Tr}[R] &< \gamma^2 \end{aligned} \quad (3.138)$$

Then, by subtracting  $R$  from both sides and using the Schur complement lemma, one has

$$\begin{bmatrix} R & C_1 P + D_{12}Q \\ (C_1 P + D_{12}Q)^T & P \end{bmatrix} > 0 \quad (3.139)$$

$$\text{Tr}[R] < \gamma^2$$

Thus, one can state the  $\mathcal{H}_2$  OCP as an SDP in  $(P, Q, R, \gamma)$ , i.e.

$$\begin{aligned} (P, Q, R, \gamma)^{opt} = & \underset{P \in \mathbb{S}^{n_x}, Q \in \mathbb{R}^{n_u \times n_x}, R \in \mathbb{S}^{n_e}, \gamma > 0}{\text{argmin}} \gamma \\ \text{subject to: } & \begin{bmatrix} A & B_2 \end{bmatrix} \begin{bmatrix} P \\ Q \end{bmatrix} + \begin{bmatrix} P & Q \end{bmatrix} \begin{bmatrix} A^T \\ B_2^T \end{bmatrix} + B_1 B_1^T < 0 \\ & - \begin{bmatrix} R & C_1 P + D_{12}Q \\ (C_1 P + D_{12}Q)^T & P \end{bmatrix} < 0 \\ & \text{Tr}[R] - \gamma^2 < 0 \end{aligned} \quad (3.140)$$

with the  $\mathcal{H}_2$  optimal controller,  $K^{opt}$ , as  $\vec{u}(t) = D_K \vec{x}(t)$  reconstructed with

$$D_K = QP^{-1} \quad (3.141)$$

## Cost Matrix Design for $\mathcal{H}_2$ Control

Control designers using the LQR method must select the  $Q$ ,  $R$ , and  $S$  cost matrices which implicitly affect the optimal controller. Thus, any implementation using the LQR method should use some useful guidelines. First, recall that one requires  $Q$  and  $R$  to be symmetric and positive semi-definite for the Lagrangian to be non-negative. Secondly, for an easier design process, one typically sets  $S = 0$  and uses one of the four following methods for selecting  $Q$  and  $R$ .

The first method uses a relative cost,  $\rho$ , between the state and input which sets

$$Q = I \quad R = \rho^2 I \quad (3.142)$$

where the cost functional has become

$$\mathcal{J} = \int_0^\infty \vec{x}^T \vec{x} + \rho^2 \vec{u}^T \vec{u} \quad (3.143)$$

which can be said to balance the  $\mathcal{L}_2$ -norm of the state and input through  $\rho$ . By varying  $\rho$  from  $0 \rightarrow \infty$ , one can sweep through the different values of  $\rho$  to find a satisfactory response. One analysis plot using this sweeping method is  $\rho$  versus  $\mathcal{J}$  which is called the **optimal tradeoff curve** to identify the lowest cost overall. Another analysis plot would be a root locus as a function of  $\rho$ .

The second method uses a relative cost,  $\rho$ , between output and input which sets

$$Q = C^T C \quad R = \rho^2 I \quad (3.144)$$

where one has incorporated an output equation with no feedthrough term,  $\vec{y} = C \vec{x}$ , in order to balance the  $\mathcal{L}_2$ -norm of the output with the input. With this method, the cost functional has become

$$\mathcal{J} = \int_0^\infty \vec{y}^T \vec{y} + \rho^2 \vec{u}^T \vec{u} \quad (3.145)$$

which can be said to balance the  $\mathcal{L}_2$ -norm of output and input through  $\rho$ . Here one can also vary  $\rho$  from  $0 \rightarrow \infty$  to find satisfactory response from the possible options.

The third method uses individual diagonal costs in addition to the relative cost,  $\rho$ , which sets

$$Q = \begin{bmatrix} q_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & q_n \end{bmatrix} \quad R = \rho^2 \begin{bmatrix} r_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_p \end{bmatrix} \quad (3.146)$$

where each  $q_i$  and  $r_i$  are selected to normalize the state and input for “equal” levels of error or effort, respectively. This can also be thought of as a weighted  $\mathcal{L}_2$ -norm of the state and input, and can also be extended for a weighted  $\mathcal{L}_2$ -norm of the output and input using the no-feedthrough model,  $\vec{y} = C \vec{x}$ . It should be noted that the normalization is typically based on units. As an example, consider that an error of 5 m/s in  $x_1$  is *as poor as* an error of 3° in  $x_2$ . Then, by setting

$$q_1 = \left(\frac{1}{5}\right)^2 \quad (3.147)$$

and

$$q_2 = \left(\frac{1}{3}\right)^2 \quad (3.148)$$

one has normalized  $q_1 x_1^2 = 1$  and  $q_2 x_2^2 = 1$  for comparable levels of error. Here again, one can vary  $\rho$  from  $0 \rightarrow \infty$  to find a satisfactory response. Here, choosing the diagonal costs may require additional trial and error if no simple method gives a satisfactory controller and is the primary task for the control designer.

The fourth method uses **Parseval's theorem** which converts scalar quadratic functions in the time domain to the frequency domain using Fourier transforms, i.e.

$$\mathcal{J} = \int_0^\infty \vec{x}^T(t) Q \vec{x}(t) + \vec{u}^T(t) R \vec{u}(t) dt = \frac{1}{2\pi} \int_{-\infty}^\infty \vec{x}^T(-j\omega) Q \vec{x}(j\omega) + \vec{u}^T(-j\omega) R \vec{u}(j\omega) d\omega \quad (3.149)$$

where  $\vec{x}(j\omega)$  is the continuous-time Fourier transform of  $\vec{x}(t)$  and  $\vec{u}(j\omega)$  is the continuous-time Fourier transform of  $\vec{u}(t)$ . Here, the matrices  $Q$  and  $R$  can be made into functions of frequency,  $\omega$ , and forms a sort of loop-shaping design method.

In any of these cases, the  $Q$  and  $R$  matrices of the quadratic cost functional can be related to the eigenvalues of the closed-loop state dynamics for the  $\mathcal{H}_2$  controller, i.e.

$$\dot{\vec{x}} = (A + BD_K) \vec{x} = A_L \vec{x} \quad (3.150)$$

by considering the Hamiltonian matrix,  $H$ , given for continuous-time by

$$H = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix} \quad (3.151)$$

which contains  $2n_x$  eigenvalues.  $n_x$  of these are the eigenvalues of  $A_L$  with negative real parts and  $n_x$  have positive real parts, unstable, but stable backward in time. To show, recall the characteristic equation of the closed-loop whose zeros are the eigenvalues can be defined as

$$\phi_{cl} = \det [sI - A - BD_K] \quad (3.152)$$

then, by the above fact, one has

$$\phi_{cl}(s)\phi_{cl}(-s) = \det [sI - H] \quad (3.153)$$

$$\phi_{cl}(s)\phi_{cl}(-s) = \det \begin{bmatrix} sI - A & BR^{-1}B^T \\ Q & sI + A^T \end{bmatrix} \quad (3.154)$$

$$\phi_{cl}(s)\phi_{cl}(-s) = \det \begin{bmatrix} sI - A & BR^{-1}B^T \\ 0 & (sI + A^T) - Q(sI - A)^{-1}BR^{-1}B^T \end{bmatrix} \quad (3.155)$$

and multiplying the first row by  $-Q(sI - A)^{-1}$  and adding to the second row, one has

$$\phi_{cl}(s)\phi_{cl}(-s) = \det [sI - A] \det [(sI + A^T) - Q(sI - A)^{-1}BR^{-1}B^T] \quad (3.156)$$

$$\phi_{cl}(s)\phi_{cl}(-s) = \phi(s) \det \left[ (sI + A^T) \left( I - (sI + A^T)^{-1}Q(sI - A)^{-1}BR^{-1}B^T \right) \right] \quad (3.157)$$

$$\phi_{cl}(s)\phi_{cl}(-s) = \phi(s) \det [(sI + A^T)] \det [I - (sI + A^T)^{-1}Q(sI - A)^{-1}BR^{-1}B^T] \quad (3.158)$$

Next, factoring  $Q = Q_1^T Q_1$  and  $BR - 1B^T = R_1 R_1^T$ , one has

$$\phi_{cl}(s)\phi_{cl}(-s) = \phi(s)\phi(-s) \det [I - (sI + A^T)^{-1} Q_1^T Q_1 (sI - A)^{-1} R_1 R_1^T] \quad (3.159)$$

and using the identity,  $\det[I - AB] = \det[I - BA]$ , one has

$$\phi_{cl}(s)\phi_{cl}(-s) = \phi(s)\phi(-s) \det [I - Q_1(sI - A)^{-1} R_1 R_1^T (sI + A^T)^{-1} Q_1^T] \quad (3.160)$$

Then, letting  $H_1 = Q_1(sI - A)^{-1} R_1$ , one has

$$\phi_{cl}(s)\phi_{cl}(-s) = \phi(s)\phi(-s) \det [I - H_1(s)H_1^T(-s)] \quad (3.161)$$

Finally, considering a cost functional with relative cost,  $\rho^2$

$$\mathcal{J} = \int_0^\infty \vec{x}^T Q \vec{x} + \rho^2 \vec{u}^T R \vec{u} \quad (3.162)$$

one can show that the poles of the Hamiltonian, i.e. the zeros of  $\phi(s)\phi(-s) \det [I - H_1(s)H_1^T(-s)]$ , are also the zeros of  $\phi(s)\phi(-s) \det [\rho I - H_1(s)H_1^T(-s)]$  which show that as  $\rho \rightarrow 0$ , some of the roots will go to infinity and those that stay finite approach the zeros of  $H_1(s)$  and these finite zeros dictate the dynamic response of the LQR. As  $\rho \rightarrow \infty$ , the roots of  $\phi_{cl}(s)$  are the  $n_x$  stable roots of  $\phi(s)\phi(-s)$ . Hence, one can see that shaping the zeros of  $H_1(s)$  is crucial to the design of the LQR controller.

## 3.5 $\mathcal{H}_\infty$ Optimal Control

### Unconstrained Continuous-Time $\mathcal{H}_\infty$ OCP

The design goal of  $\mathcal{H}_\infty$  optimal control is to find a stabilizing  $K$  that minimizes the  $\mathcal{H}_\infty$ -norm of the generalized feedback control system, i.e.

$$K^{opt} = \underset{\substack{K \text{ stabilizing}}}{\operatorname{argmin}} \|F_L(G, K)\|_\infty \quad (3.163)$$

where  $F_L(G, K)$  defines a state-space model as

$$\begin{aligned} \dot{\vec{x}}(t) &= A\vec{x}(t) + B_1\vec{d}(t) + B_2\vec{u}(t) \\ \vec{e}(t) &= C_1\vec{x}(t) + D_{11}\vec{d}(t) + D_{12}\vec{u}(t) \\ \vec{y}(t) &= C_2\vec{x}(t) + D_{21}\vec{d}(t) + D_{22}\vec{u}(t) \end{aligned} \quad (3.164)$$

Recalling the definition of the  $\mathcal{H}_\infty$ -norm, one can recast this OCP in a min-max OCP framework as

$$K^{opt}, \vec{d}^{opt} = \underset{\substack{K \text{ stabilizing}}}{\operatorname{argmin}} \underset{0 \neq \|\vec{d}\|_2 \leq \infty, \vec{x}(0)=0}{\operatorname{argmax}} \frac{\|\vec{e}\|_2^2}{\|\vec{d}\|_2^2} \quad (3.165)$$

which can be a difficult problem to solve for general  $F_L(G, K)$ .

Thus, one can restate the problem by noting that, one has for all  $\vec{d}$  with  $\|\vec{d}\|_2 < \infty$

$$\|F_L(G, K)\|_\infty \geq \frac{\|\vec{e}\|_2}{\|\vec{d}\|_2} \quad (3.166)$$

Thus, one method to solve the  $\mathcal{H}_\infty$  OCP is to bound  $\mathcal{H}_\infty$ -norm by some  $\gamma > 0$ , i.e.

$$\|F_L(G, K)\|_\infty^2 \leq \gamma^2 \quad (3.167)$$

Then, one has

$$\gamma^2 \|\vec{d}\|_2^2 \geq \|\vec{e}\|_2^2 \quad (3.168)$$

$$\|\vec{e}\|_2^2 - \gamma^2 \|\vec{d}\|_2^2 \leq 0 \quad (3.169)$$

which will equal zero for some worst case  $\vec{d}$  and  $\gamma = \gamma_{min} = \|F_L(G, K)\|_\infty$ .

Thus, the  $\mathcal{H}_\infty$  OCP can be reformulated as solving for the optimal controller *and* maximizing disturbance that solves the following constrained min-max OCP.

$$\begin{aligned} K^{opt}, \vec{d}^{opt} &= \underset{K \text{ stabilizing}}{\operatorname{argmin}} \underset{\vec{d}}{\operatorname{argmax}} \mathcal{J}(K, \vec{d}) \\ \text{subject to : } \gamma &\geq \|F_L(G, K)\|_\infty \end{aligned} \quad (3.170)$$

with cost functional

$$\mathcal{J}(K, \vec{d}) = \frac{1}{2} \int_0^{t_f} \vec{e}^T \vec{e} - \gamma^2 \vec{d}^T \vec{d} d\tau \quad (3.171)$$

where notably  $\vec{x}_0 = 0$ ,  $t_f$  is given and may be finite in this reformulation, and  $\vec{x}(t_f)$  is free to vary in the optimization, though one can include a terminal cost on  $\vec{x}(t_f)$  if desired.

If one assumes  $C_2 = I$ ,  $D_{21} = D_{22} = 0$ , for state feedback  $\mathcal{H}_\infty$  optimal control,  $\vec{u}(t) = D_K(t) \vec{x}(t)$ , then this cost functional becomes

$$\mathcal{J}(\vec{u}, \vec{d}) = \frac{1}{2} \int_0^{t_f} [C_1 \vec{x} + D_{12} \vec{u} + D_{11} \vec{d}]^T [C_1 \vec{x} + D_{12} \vec{u} + D_{11} \vec{d}] - \gamma^2 \vec{d}^T \vec{d} d\tau \quad (3.172)$$

$$\begin{aligned} \mathcal{J}(\vec{u}, \vec{d}) &= \frac{1}{2} \int_0^{t_f} \vec{x}^T C_1^T C_1 \vec{x} + 2 \vec{x}^T [C_1^T D_{12} \quad C_1^T D_{11}] \begin{bmatrix} \vec{u} \\ \vec{d} \end{bmatrix} \\ &\quad + \begin{bmatrix} \vec{u} \\ \vec{d} \end{bmatrix}^T \begin{bmatrix} D_{12}^T D_{12} & D_{12}^T D_{11} \\ D_{11}^T D_{12} & D_{11}^T D_{11} - \gamma^2 I \end{bmatrix} \begin{bmatrix} \vec{u} \\ \vec{d} \end{bmatrix} d\tau \end{aligned} \quad (3.173)$$

which is close to a quadratic cost functional except for the  $\gamma^2$  term which is to be minimized as part of the optimization, and thus is not truly a quadratic cost.

### ARE Synthesis for State Feedback $\mathcal{H}_\infty$ OCP

However, assuming one has a constant upper bound  $\gamma$ , then one has the sub-problem to find the  $\vec{d}$  and  $\vec{u}$  which maximize and minimize the cost functional to obtain  $D_K(t)$ . In this way, one can define the LQR sub-problem with

$$Q = C_1^T C_1 \quad (3.174)$$

$$S = \begin{bmatrix} C_1^T D_{12} & C_1^T D_{11} \end{bmatrix} \quad (3.175)$$

$$R = \begin{bmatrix} D_{11}^T D_{12} & D_{11}^T d_2 \\ D_{11}^T D_{12} & D_{11}^T D_{11} - \gamma^2 I \end{bmatrix} \quad (3.176)$$

$$\tilde{B} = \begin{bmatrix} B_2 & B_1 \end{bmatrix} \quad (3.177)$$

and

$$\tilde{u} = \begin{bmatrix} \vec{u} \\ d \end{bmatrix} \quad (3.178)$$

Then, one has the LQR cost functional

$$\mathcal{J}(\tilde{u}) = \frac{1}{2} \int_0^{t_f} \vec{x}^T Q \vec{x} + 2 \vec{x}^T S \tilde{u} + \tilde{u}^T R \tilde{u} d\tau \quad (3.179)$$

which is solved by the differential Riccati equation

$$-\dot{P} = PA + A^T P + Q - (P \tilde{B} + S) R^{-1} (\tilde{B}^T P + S^T) \quad (3.180)$$

with the optimal  $\tilde{u}^{opt}$  given by

$$\tilde{u} = -R^{-1} (\tilde{B}^T P(t) + S^T) \vec{x} \quad (3.181)$$

and the optimal control sequence  $\vec{u}$  is

$$\begin{aligned} \vec{u} &= [I_{n_u \times n_u} \ 0] \tilde{u} \\ &= -[I_{n_u \times n_u} \ 0] R^{-1} (\tilde{B}^T P(t) + S^T) \vec{x} \\ &= D_K(t) \vec{x} \end{aligned} \quad (3.182)$$

For  $t_f \rightarrow \infty$ ,  $D_K$  is a fixed-gain controller obtained by the Riccati matrix  $P \geq 0$  that solves the CARE

$$0 = PA + A^T P + Q - (P \tilde{B} + S) R^{-1} (\tilde{B}^T P + S^T) \quad (3.183)$$

which can be rewritten as

$$PA + A^T P + C_1^T C_1 - \begin{bmatrix} B_2^T P + D_{12}^T C_1 \\ B_1^T P + D_{11}^T C_1 \end{bmatrix}^T \begin{bmatrix} D_{12}^T D_{12} & D_{12}^T D_{11} \\ D_{11}^T D_{12} & D_{11}^T D_{11} - \gamma^2 I \end{bmatrix} \begin{bmatrix} B_2^T P + D_{12}^T C_1 \\ B_1^T P + D_{11}^T C_1 \end{bmatrix} = 0 \quad (3.184)$$

Note that for this LQR sub-problem to be well-posed, one must assume that for the LTI plant state-space model  $(A, B_2, C_1, D_{12})$  there are no zeros on the  $j\omega$  axis,  $(A, B_2)$  is stabilizable,  $(A, C_1)$  is detectable, and  $(D_{12}^T D_{12})^{-1}$  exists (i.e. injective). However, choosing a  $\gamma$  to solve the  $\mathcal{H}_\infty$  OCP is still required for this LQR solution method and is typically done through  **$\gamma$ -iteration**

A simple iteration procedure can be performed as a bisection search as

1. Initialize  $\gamma$  larger than the anticipated optimal  $\gamma$  for binary search

- Form LQR cost matrices using  $\gamma$
- Solve continuous-time algebraic Riccati equation (CARE) for matrix  $P$
- If  $P > 0$  and  $(A - BD_K)$  Hurwitz:

- Decrease  $\gamma$  by bisection (until convergence threshold)
- Else:
  - Increase  $\gamma$  by bisection

2. Convergence to  $\gamma_{min}$  to form  $D_K$

Note that care must be taken as  $\gamma$  approaches  $\gamma_{min}$  as  $R$  typically becomes ill-conditioned. It is also typically prudent to slightly increase  $\gamma$  from  $\gamma_{min}$  to reduce feedback gain magnitudes and improve the accuracy of the numerical solver for the CARE.

### SDP Synthesis for State Feedback $\mathcal{H}_\infty$ OCP

The Bounded Real Lemma states that  $A_L$  is Hurwitz, and  $\|F_L(G, K)\|_\infty < \gamma^2$  if and only if there exists  $P > 0$  satisfying the strict ARI:

$$A_L^T P + P A_L + C_L^T C_L + \gamma^{-2} P B_L B_L^T P < 0 \quad (3.185)$$

which can be written as

$$\begin{bmatrix} (A + B_2 D_K)^T P + P(A + B_2 D_K) & PB_1 \\ B_1^T P & \gamma^{-2} I \end{bmatrix} + \begin{bmatrix} (C_1 + D_{12} D_K)^T \\ 0 \end{bmatrix} \begin{bmatrix} C_1 + D_{12} D_K & 0 \end{bmatrix} < 0 \quad (3.186)$$

which is a nonlinear matrix inequality in  $(P, D_K)$  due to the bilinear and quadratic terms. However, one can use the **symmetric congruence transformation**

$$\begin{bmatrix} P^{-1} & 0 \\ 0 & I \end{bmatrix} \quad (3.187)$$

which can be left and right multiplied to get

$$\begin{bmatrix} P^{-1}(A + B_2 D_K)^T + (A + B_2 D_K)P^{-1} & B_1 \\ B_1^T & \gamma^{-2} I \end{bmatrix} + \begin{bmatrix} P^{-1}(C_1 + D_{12} D_K)^T \\ 0 \end{bmatrix} \begin{bmatrix} (C_1 + D_{12} D_K)P^{-1} & 0 \end{bmatrix} < 0 \quad (3.188)$$

Next, defining a change of variables as  $Q = P^{-1}$  and  $R = D_K Q$ , one has alternatively

$$\begin{bmatrix} QA^T + AQ + R^T B_2^T + B_2 R & B_1 \\ B_1^T & \gamma^{-2} I \end{bmatrix} + \begin{bmatrix} (C_1 Q + D_{12} R)^T \\ 0 \end{bmatrix} \begin{bmatrix} C_1 Q + D_{12} R & 0 \end{bmatrix} < 0 \quad (3.189)$$

which is still not an LMI due to the quadratic term,  $(C_1 Q + D_{12} R)^T (C_1 Q + D_{12} R)$ . However, by the Schur Complement Lemma, one can write

$$\begin{bmatrix} QA^T + AQ + R^T B_2^T + B_2 R & B_1 & (C_1 Q + D_{12} R)^T \\ B_1^T & \gamma^{-2} I & 0 \\ C_1 Q + D_{12} R & 0 & -I \end{bmatrix} < 0 \quad (3.190)$$

Thus, one can state the  $\mathcal{H}_\infty$  OCP as an SDP in  $(Q, R, \gamma)$ , i.e.

$$(Q, R, \gamma)^{opt} = \underset{\substack{Q \in \mathbb{S}^{n_x}, R \in \mathbb{R}^{n_u \times n_x}, \gamma > 0}}{\operatorname{argmin}} \gamma$$

subject to: 
$$\begin{bmatrix} QA^T + AQ + R^T B_2^T + B_2 R & B_1 & (C_1 Q + D_{12} R)^T \\ B_1^T & -\gamma^{-2} I & 0 \\ C_1 Q + D_{12} R & 0 & -I \end{bmatrix} < 0$$

$$-Q < 0$$
(3.191)

with the state feedback  $\mathcal{H}_\infty$  optimal controller,  $K^{opt}$ , as  $\vec{u}(t) = D_K \vec{x}(t)$  reconstructed with

$$D_K = RQ^{-1} \quad (3.192)$$

### Signal-Weighted $\mathcal{H}_\infty$ Optimal Control Design

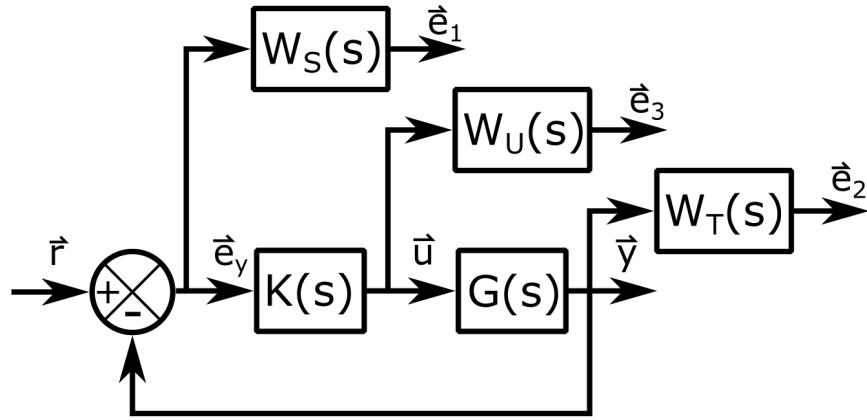
Recall in classical control, one can design the feedback control policy,  $K(s)$ , by shaping the frequency response of the loop gain,  $|L(j\omega)|$ , to meet robust control system performance and robust stability requirements, i.e. loop-shaping. Note that in SISO loop-shaping,  $L(s) = G(s)K(s)$  is the open-loop transfer function with singular value  $\sigma(L) = \bar{\sigma}(L) = |L|$  that can be related to the error/sensitivity transfer function, i.e.  $S(s) = (1 + L(s))^{-1}$ , and the closed-loop/complementary sensitivity transfer function,  $T(s) = L(s)(1 + L(s))^{-1}$ , where  $S(s) + T(s) = 1$  must hold. Here, one shapes the loop gain through selecting the *control stages* of  $K(s)$  to set the crossover frequency, increase the gain at low frequencies, decrease the gain at high frequencies, and reduce the slope of the crossover region to achieve good classical stability margins. Furthermore, to check the control effort requires one to also assess that  $K(s)S(s)$  separately from the loop-shaping procedure.

This frequency-domain robust control design can be generalized to MIMO LTI systems through  $\mathcal{H}_\infty$  optimal control and to meet robust MIMO system performance and robust stability requirements. For MIMO LTI systems, instead of using separate control stage transfer functions for the controller, one can use different *weight transfer functions* on the different signals which make up the both the generalized disturbances,  $\vec{d}$ , and generalized errors,  $\vec{e}$ , for the generalized LTI feedback control system  $F_L(G, K)$  so that  $F_L(G, K)$  is stable and

$$\left\| \frac{\vec{e}}{\vec{d}} \right\|_\infty = \|F_L(G, K)\|_\infty \leq \gamma \quad (3.193)$$

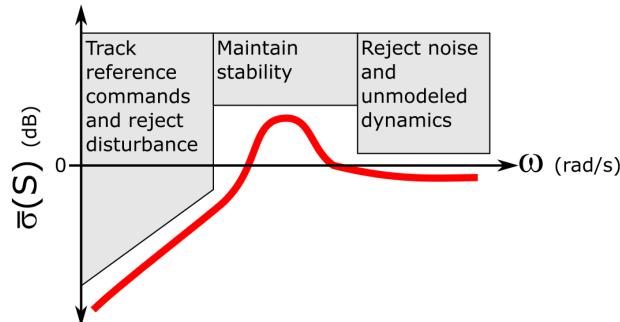
In general, these weight transfer functions are selected as functions of frequency which allow one to normalize and balance the various robust stability and performance requirements across the expected frequency ranges, e.g. reference commands, plant disturbances, control effort, error tracking, model-following, state regulation. In some cases, one can choose the weight transfer functions such that if  $\gamma = 1$ , then all robust stability and performance requirements are met. Thus,  $\mathcal{H}_\infty$  optimal control design is quite varied and is typically problem-specific.

However, an introductory method for a **single degree-of-freedom LTI feedback control system** is **mixed-sensitivity  $\mathcal{H}_\infty$  loop-shaping** which forms the following signal-weighted LTI feedback control system:

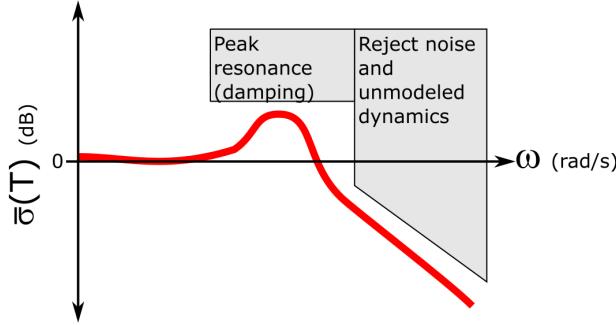


where one chooses the weights,  $W_S$ ,  $W_u$ , and  $W_T$  on the signals,  $\vec{e}_y$ ,  $u(s)$ , and  $\vec{y}$ , respectively, to form the generalized error vector,  $\vec{e}$  as three vectors,  $\vec{e}_1$ ,  $\vec{e}_2$ , and  $\vec{e}_3$ . Note that  $\vec{e}_y$  is the tracking error and  $\vec{r}$  is the only generalized disturbance.

The error or sensitivity transfer function matrix is  $S(s) = (I + K(s)G(s))^{-1}$  and should generally follow the shape



and the closed-loop or complementary sensitivity matrix is  $T(s) = (I + K(s)G(s))^{-1}L(s)$  and should generally follow the shape



At low frequency,  $S(s)$  needs to be small as  $S(s)$  is the transfer function for the tracking error and  $T(s)$  is near unity or 0 dB as  $S(s) + T(s) = 1$ . At high frequencies,  $T(s)$  needs to roll off to reject high-frequency noise and unmodeled dynamics and  $S(s)$  is near unity or 0 dB as  $S(s) + T(s) = 1$ . Furthermore, the near-singularity of  $S^{-1}(s) = I + K(s)G(s)$  is measured by the minimum of  $\underline{\sigma}(I + K(j\omega)G(j\omega))$  across all frequencies which equates to the maximum or peak value of  $S(j\omega)$ , i.e.  $\|S\|_\infty$ . At the peak of  $T(j\omega)$ ,  $\|T\|_\infty$  denotes the closed-loop system's peak resonance at the frequency for which inputs are most amplified. Thus  $\|T\|_\infty$  is analogous to the idea of damping for second-order systems, i.e. the dominant poles are close to the  $j\omega$ -axis at this frequency.

With this frequency-dependent weights framework, the  $\mathcal{H}_\infty$  optimal control synthesis will provide  $\gamma_{min}$  such that to minimize

$$\begin{aligned} \|W_S S\|_\infty &\leq \gamma \\ \|W_u K S\|_\infty &\leq \gamma \\ \|W_T T\|_\infty &\leq \gamma \end{aligned} \tag{3.194}$$

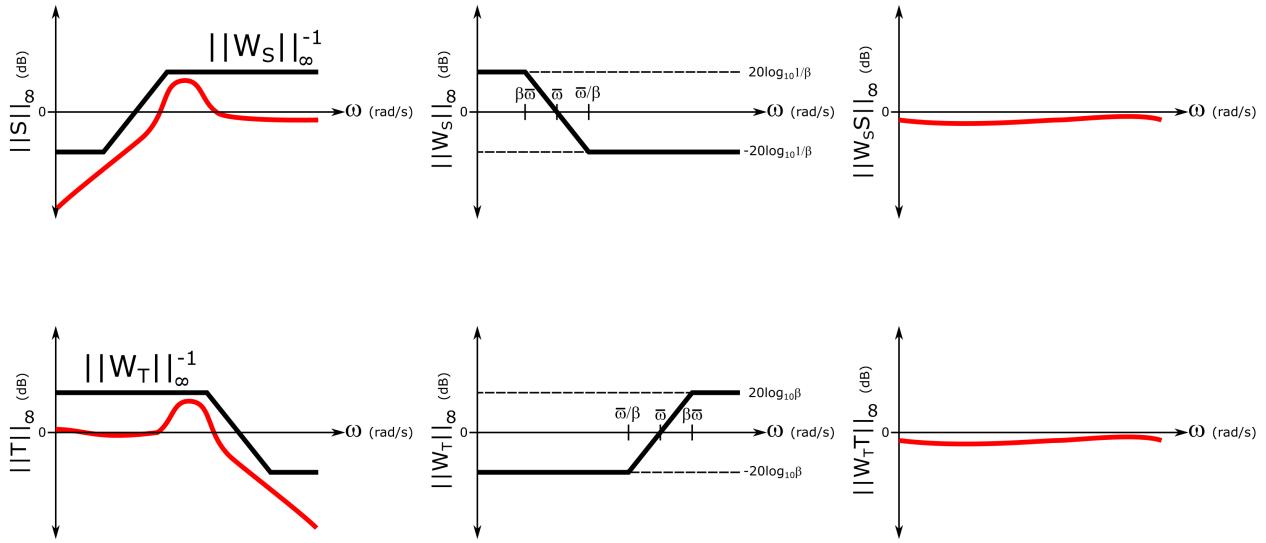
which equates to

$$\begin{aligned} \|S\|_\infty &\leq \gamma |W_S^{-1}| \\ \|K S\|_\infty &\leq \gamma |W_u^{-1}| \\ \|T\|_\infty &\leq \gamma |W_T^{-1}| \end{aligned} \tag{3.195}$$

where one typically desires to minimize the control effort,  $\|u\|$ , across all frequencies to ensure that any actuators are not position or rate saturated. Thus, one typically sets a constant weight on  $W_u$ . Here the selection of the  $W_S$  and  $W_T$  weighting filters are first-order lag and lead transfer functions, i.e.

$$W(s) = \frac{\beta s + \bar{\omega}}{s + \beta \bar{\omega}} \tag{3.196}$$

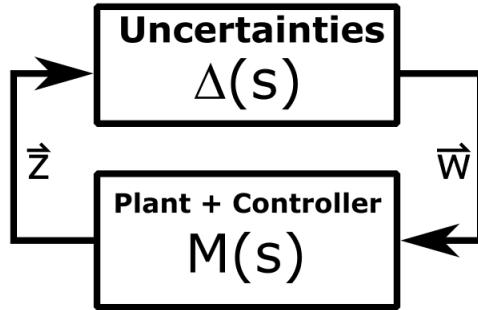
where for  $0 < \beta < 1$ , this corresponds to a lag control (e.g.  $W_S$ ) and  $\beta > 1$  corresponds to a lead control (e.g.  $W_T$ ). This corresponds to the following diagrams:



The degree of difficulty in choosing these weighting filters can be comparable to choosing the proportional gain, low-frequency boost, and lead control stages in SISO loop-shaping design. Similarly, design iteration with the weighting filters is used in conjunction with the  $\mathcal{H}_\infty$  optimal control synthesis to perform mixed-sensitivity  $\mathcal{H}_\infty$  loop-shaping control design.

### $\mu$ -Synthesis via D-K Iteration

To incorporate robust stability margins directly in the optimal control framework, recall the generalized  $\Delta M$  framework as



where the structured singular value (SSV)  $\mu_\Delta$  can be used to compute an upper and lower bound for the inverse of the minimum possible  $\bar{\sigma}\Delta$  which causes the  $\Delta M$  model to become unstable. This is related to the maximum singular value of  $M$  by

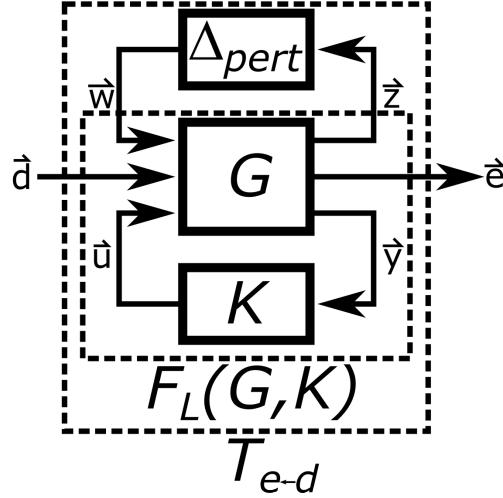
$$\bar{\rho}(M) \leq \mu_\Delta(M) \leq \bar{\sigma}(M) \quad (3.197)$$

which is numerically approximated by

$$\max_Q \lambda(QM) \leq \mu_\Delta(M) \leq \inf_D \bar{\sigma}(DMD^{-1}) \quad (3.198)$$

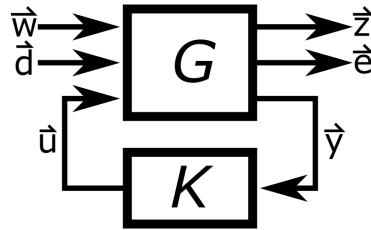
where the frequency-dependent  $D$  matrices, which commute with  $\Delta$  (i.e.  $D\Delta = \Delta D$ ), are called the  **$D$  scalings**.

This  $\Delta M$  framework can be combined with the  $\mathcal{H}_\infty$  optimal control framework as the general block diagram



where  $\vec{d}$  and  $\vec{e}$  are the generalized disturbance and generalized error,  $\vec{w}$  and  $\vec{z}$  are the perturbation input and output vectors to some plant perturbation uncertainty,  $\Delta_{pert}$ , and  $\vec{u}$  and  $\vec{y}$  are the control input to the plant and plant output to the controller vectors as previously. Note that the lower LFT,  $F_L(G, K)$ , is the equivalent to  $M$  in the  $\Delta M$  robust analysis model.

Here, one can then set up the model  $F_L(G, K)$  as



where one has for the augmented uncertainty block matrix,  $\Delta$ , in the  $\Delta M$  robust analysis

$$\Delta = \begin{bmatrix} \Delta_{pert} & 0 \\ 0 & \Delta_F \end{bmatrix} \quad (3.199)$$

which takes into account both input and output vectors in the  $\mu$ -synthesis model as  $\vec{e} = \Delta_F \vec{d}$ . Thus, the  $\mu$ -**synthesis** OCP attempts to minimize over all stabilizing controllers,  $K$ , the peak value of  $\mu_\Delta$  of the closed-loop transfer function, i.e.

$$K^{opt} = \underset{K \text{ stabilizing}}{\operatorname{argmin}} \max_{\omega} \mu_\Delta(F_L(G, K)(j\omega)) \quad (3.200)$$

To derive a tractable approximation for this  $\mu$ -synthesis optimization known as  $D - K$  iteration, recall the upper bound definition for  $\mu_\Delta$ , one can rewrite this minimization of the maximum possible  $\mu_\Delta$  as

$$K^{opt} = \underset{K \text{ stabilizing}}{\operatorname{argmin}} \max_{\omega} \min_{D_\omega} \bar{\sigma}\left(D_\omega F_L(G, K)(j\omega) D_\omega^{-1}\right) \quad (3.201)$$

where  $D_\omega$  is chosen from all possible scalings independently as every  $\omega$ . Then, rearranging, one has

$$K^{opt} = \underset{K \text{ stabilizing}}{\operatorname{argmin}} \min_{D_\omega} \max_{\omega} \bar{\sigma}\left(D_\omega F_L(G, K)(j\omega) D_\omega^{-1}\right) \quad (3.202)$$

or

$$K^{opt} = \underset{K \text{ stabilizing}}{\operatorname{argmin}} \min_{D_\omega} \|D_\omega F_L(G, K) D_\omega^{-1}\|_\infty \quad (3.203)$$

Thus, one can see that this optimization can be constructed as minimizing two different parameters,  $D$  and  $K$ , which is done as an iterative procedure, i.e. holding  $D$  as fixed and finding the optimal  $K$  using  $\mathcal{H}_\infty$ , then holding  $K$  fixed and finding the optimal  $D$  that minimizes the transformed  $\mathcal{H}_\infty$ -norm. Note that this  $D$  procedure serves as an approximation for maximizing the  $\mu_\Delta$  upper bound, i.e. maximizing the “distance” to singularity/instability of  $F_L(G, K)$  for the unstructured uncertainty  $\Delta$  (i.e. fully complex block matrix as defined previously). This approximation is typically very close. However, it should also be noted that  $D - K$  iteration is not guaranteed to converge to a global or even a local minimum which is a serious shortcoming of this approach, but has been shown to work well in many different flight vehicle control problems including highly flexible airplanes, missile autopilots, modern fighter airplanes, and the space shuttle flight control system.

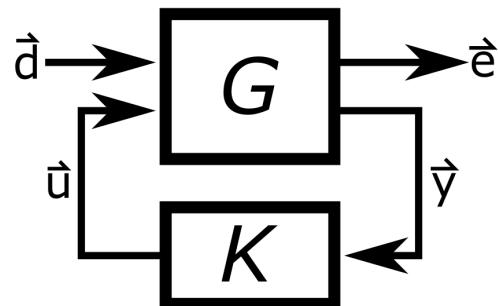
---

# Discrete-Time Linear Control Theory

## 4.1 Discrete-Time Linear Feedback Control Systems

### Discrete-Time Generalized LTI Feedback Control System

Consider the generalized LTI feedback control system architecture



where the vector input,  $\vec{d} \in \mathbb{R}^{n_d}$ , to the generalized plant is known as the **generalized disturbance** which typically includes reference commands,  $\vec{r}$ , process noise,  $\vec{w}$ , and measurement noise,  $\vec{v}$ . The vector output,  $\vec{e} \in \mathbb{R}^{n_e}$ , from the generalized plant is known as the **generalized error** which typically includes at least the weighted tracking error and the weighted control effort. The other vectors are the **generalized control input** vector  $\vec{u} \in \mathbb{R}^{n_u}$  and the **generalized output!discrete-time**,  $\vec{y} \in \mathbb{R}^{n_y}$ .

Just as for continuous-time, the LTI state-space model for the generalized plant,  $G$ , assuming **loop-**

**shifting**, can be written as

$$\begin{aligned}\vec{x}[k+1] &= F\vec{x}[k] + [G_1 \quad G_2] \begin{bmatrix} \vec{d}[k] \\ \vec{u}[k] \end{bmatrix} \\ \begin{bmatrix} \vec{e}[k] \\ \vec{y}[k] \end{bmatrix} &= \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} \vec{x}[k] + \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & 0 \end{bmatrix} \begin{bmatrix} \vec{d}[k] \\ \vec{u}[k] \end{bmatrix}\end{aligned}\tag{4.1}$$

with the **generalized LTI feedback controller**,  $K$ , designed as

$$\begin{aligned}\vec{x}_K[k+1] &= F_K \vec{x}_K[k] + G_K \vec{y}[k] \\ \vec{u}[k] &= H_K \vec{x}_K[k] + D_K \vec{y}[k]\end{aligned}\tag{4.2}$$

and the interconnection of  $G$  and  $K$  is well-posed if and only if  $(I - D_{22}D_K)^{-1}$  exists. In the same way, one can obtain the discrete-time closed-loop LTI state-space system model as

$$\begin{aligned}\begin{bmatrix} \vec{x}[k+1] \\ \vec{x}_K[k+1] \end{bmatrix} &= \begin{bmatrix} F + G_2 D_K H_2 & G_2 H_K \\ G_K H_2 & F_K \end{bmatrix} \begin{bmatrix} \vec{x}[k] \\ \vec{x}_K[k] \end{bmatrix} + \begin{bmatrix} G_1 + G_2 D_K D_{21} \\ G_K D_{21} \end{bmatrix} \vec{d}[k] \\ \vec{e}[k] &= \begin{bmatrix} H_1 + D_{12} D_K H_2 & D_{12} H_K \end{bmatrix} \begin{bmatrix} \vec{x}[k] \\ \vec{x}_K[k] \end{bmatrix} + (D_{11} + D_{12} D_K D_{21}) \vec{d}[k]\end{aligned}\tag{4.3}$$

with the definitions for the closed-loop state matrix,  $F_L$ , as

$$F_L = \begin{bmatrix} F + G_2 D_K H_2 & G_2 H_K \\ G_K H_2 & F_K \end{bmatrix}\tag{4.4}$$

the closed-loop input matrix,  $G_L$ , as

$$G_L = \begin{bmatrix} G_1 + G_2 D_K D_{21} \\ G_K D_{21} \end{bmatrix}\tag{4.5}$$

the closed-loop output matrix,  $H_L$ , as

$$H_L = \begin{bmatrix} H_1 + D_{12} D_K H_2 & D_{12} H_K \end{bmatrix}\tag{4.6}$$

and the closed-loop feedthrough matrix,  $D_L$ , as

$$D_L = D_{11} + D_{12} D_K D_{21}\tag{4.7}$$

which for stability of the generalized feedback control system, one requires that  $F_L$  is stable.

### Discrete-Time Output and State Feedback Control

If  $F_K = G_K = H_K = 0$ , one has **static-controller feedback**, a type of **fixed-gain controller**. The general case is also known as a.k.a. **output feedback control**, where one has an output feedback controller as

$$\vec{u}[k] = D_K \vec{y}[k]\tag{4.8}$$

which results in a simplified closed-loop LTI state-space model

$$\begin{aligned}\vec{x}[k+1] &= F_L \vec{x}[k] + G_L \vec{d}[k] \\ \vec{e}[k] &= H_L \vec{x}[k] + D_L \vec{d}[k]\end{aligned}\tag{4.9}$$

which results in a closed-loop state matrix,  $F_L$ , as

$$F_L = F + G_2 D_K H_{12}\tag{4.10}$$

a closed-loop input matrix,  $G_L$ , as

$$G_L = G_1 + G_2 D_K D_{21}\tag{4.11}$$

a closed-loop output matrix,  $H_L$ , as

$$H_L = H_1 + D_{12} D_K H_2\tag{4.12}$$

and a closed-loop feedthrough matrix,  $D_L$ , as

$$D_L = D_{11} + D_{12} D_K D_{21}\tag{4.13}$$

Thus, an output feedback control system is stable if and only if  $F + G_2 D_K H_{12}$  is stable.

A special case of fixed-gain feedback control is **state feedback control**, i.e.  $H_2 = I$ ,  $D_{21} = 0$ , and state feedback controller

$$\vec{u}[k] = D_K \vec{x}[k]\tag{4.14}$$

with closed-loop state matrix,  $F_L$ , is

$$F_L = F + G_2 D_K\tag{4.15}$$

the closed-loop input matrix,  $G_L$ , is

$$G_L = G_1\tag{4.16}$$

the closed-loop output matrix,  $H_L$ , as

$$H_L = H_1 + D_{12} D_K\tag{4.17}$$

and the closed-loop feedthrough matrix,  $D_L$ , as

$$D_L = D_{11}\tag{4.18}$$

Thus, a state feedback control system is stable if and only if  $F + G_2 D_K$  is stable.

### Discrete-Time Observer Feedback Control

The general case of **dynamic-controller feedback** is also known as discrete-time **observer feedback control**, one defines the controller state as the **state estimate**,  $\hat{\vec{x}}[k]$ , i.e.

$$\vec{x}_K[k] = \hat{\vec{x}}[k]\tag{4.19}$$

where an observer is designed to form  $\hat{\vec{x}}[k]$  and uses this to form the control input as

$$\vec{u}[k] = -K \hat{\vec{x}}[k]\tag{4.20}$$

Here, an **open-loop observer** could be formed based on the linear state-space model for continuous-time, assuming the disturbances  $\vec{d}$  are unknown, as

$$\hat{\vec{x}}[k+1] = F\hat{\vec{x}}[k] + G_2\vec{u}[k] \quad (4.21)$$

However, for feedback control, one receives the output signal from the output equation, and forms a **closed-loop observer** for discrete-time as

$$\begin{aligned} \hat{\vec{x}}[k+1] &= F\hat{\vec{x}}(t) + G_2\vec{u}[k] + L(\vec{y}[k] - \hat{\vec{y}}[k]) \\ \hat{\vec{y}}[k] &= H_2\hat{\vec{x}}[k] \end{aligned} \quad (4.22)$$

where  $\hat{\vec{y}}$  is the output estimate based on the output equation model and  $L$  is the **Luenberger observer matrix**. Thus, with  $\vec{u}[k] = -K\hat{\vec{x}}[k]$ , one can form the continuous-time **observer feedback control system** as the generalized LTI feedback controller

$$\begin{aligned} \hat{\vec{x}}[k+1] &= (F - G_2K - LH_2)\hat{\vec{x}}[k] + L\vec{y}(t) \\ \vec{u}[k] &= -K\hat{\vec{x}}(t) \end{aligned} \quad (4.23)$$

where

$$F_K = F - G_2K - LH_2 \quad (4.24)$$

$$G_K = L \quad (4.25)$$

$$H_K = -K \quad (4.26)$$

$$D_K = 0 \quad (4.27)$$

Thus, one has the closed-loop dynamics

$$\begin{aligned} \begin{bmatrix} \vec{x}[k+1] \\ \hat{\vec{x}}[k+1] \end{bmatrix} &= \begin{bmatrix} F & -G_2K \\ LH_2 & F - G_2K - LH_2 \end{bmatrix} \begin{bmatrix} \vec{x}[k] \\ \hat{\vec{x}}[k] \end{bmatrix} + \begin{bmatrix} G_1 \\ LD_{21} \end{bmatrix} \vec{d}[k] \\ \vec{e}[k] &= [H_1 \quad -D_{12}K] \begin{bmatrix} \vec{x}[k] \\ \hat{\vec{x}}[k] \end{bmatrix} + D_{11}\vec{d}[k] \end{aligned} \quad (4.28)$$

However, to better assess the stability of the closed-loop system, consider the **state error**,  $\vec{e}_x[k]$ , defined as

$$\vec{e}_x[k] = \vec{x}[k] - \hat{\vec{x}}[k] \quad (4.29)$$

Then, after the same algebra as for continuous-time, one has

$$\begin{aligned} \begin{bmatrix} \vec{x}[k+1] \\ \vec{e}_x[k+1] \end{bmatrix} &= \begin{bmatrix} (F - G_2K) & G_2K \\ 0 & F - LH_2 \end{bmatrix} \begin{bmatrix} \vec{x}[k] \\ \vec{e}_x[k] \end{bmatrix} + \begin{bmatrix} G_1 \\ G_1 - LD_{21} \end{bmatrix} \vec{d}[k] \\ \vec{e}[k] &= [(H_1 - D_{12}K) \quad D_{12}K] \begin{bmatrix} \vec{x}[k] \\ \vec{e}_x[k] \end{bmatrix} + D_{11}\vec{d}[k] \end{aligned} \quad (4.30)$$

Thus, a discrete-time observer feedback control system is stable if and only if  $F - G_2K$  and  $F - LH_2$  are stable, i.e., one has the same **separation principle** for discrete-time observer feedback control.

Lastly, it should be noted that the Luenberger observer is also known as an  $\alpha - \beta$  **filter** if

1.  $F$ 

$\mathbb{R}^{2 \times 2}$  with all main diagonal terms equal to 1 and upper diagonal terms equal to  $\Delta t$ .

2.  $H = [10]$ 3.  $L = [\alpha\beta]^T$ 4.  $u$  is known or set to zero.5.  $G$  has non-zero gain term as last element if  $u$  is non-zero.

## Controllability and Observability of Discrete-Time Linear Systems

Recall the discrete-time state controllability matrix

$$C = [G_{k-1} \quad F_{k-1}G_{k-2} \quad \cdots \quad F_{k-1} \cdots F_1 G_0] \quad (4.31)$$

and for LTI system, one has

$$C = [G \quad FG \quad \cdots \quad F^k G] \quad (4.32)$$

which has the same structure as continuous-time. Thus, the same logic applies that any  $\vec{x}[k]$  can be reached if and only if  $C$  has full rank. Similarly one can define the **controllability Gramian** for discrete-time as

$$W_C[k] = \sum_{k=0}^{\infty} \Phi(0, k) G_k G_k^T \Phi(0, k)^T \quad (4.33)$$

where one can show if and only if  $W[k]$  full rank for *any*  $t > 0$  or  $k > 0$ , then the system is state controllable. Note that  $W_C$  is  $n_x \times n_x$  matrix, thus the degree of controllability possible by assessing eigenvalue decomposition of  $W_C$ .

**State observability** is defined as: if, for some finite  $N > 0$ , inputs  $\vec{u}[k]$ , and outputs  $\vec{y}[k]$  with  $0 < k \leq N$ , the initial state  $\vec{x}[0]$  can be determined, then one can observe the system's *past* initial state. One can quantify this concept without loss of generality given the input and output sequences,  $\vec{u}[k]$  and  $\vec{y}[k]$  for  $k = 0, 1, \dots, n - 1$ , and the discrete-time linear output equation

$$\vec{y}[k] = H_k \vec{x}[k] + D_k \vec{u}[k] \quad (4.34)$$

which can be rearranged as

$$\vec{y}[k] - D_k \vec{u}[k] = H_k \vec{x}[k] \quad (4.35)$$

Thus, one has the following sequence of output equations

$$\vec{y}[0] - D_0 \vec{u}[0] = H_0 \vec{x}[0] \quad (4.36)$$

$$\vec{y}[1] - D_1 \vec{u}[1] = H_1 \vec{x}[1] = H_1 (F_0 \vec{x}[0] + G_0 \vec{u}[0]) = H_1 F_0 \vec{x}[0] + H_1 G_0 \vec{u}[0] \quad (4.37)$$

and

$$\vec{y}[2] - D_2 \vec{u}[2] = H_2 \vec{x}[2] = H_2 (F_1 \vec{x}[1] + G_1 \vec{u}[1]) \quad (4.38)$$

or more simply

$$\vec{y}[2] - D_2 \vec{u}[2] = H_2 F_1 (F_0 \vec{x}[0] + G_0 \vec{u}[0]) + H_2 G_1 \vec{u}[1] \quad (4.39)$$

$$\vec{y}[2] - D_2 \vec{u}[2] = H_2 F_1 F_0 \vec{x}[0] + H_2 F_1 G_0 \vec{u}[0] + H_2 G_1 \vec{u}[1] \quad (4.40)$$

Thus, by extension one has

$$\vec{y}[n-1] - D_{n-1} \vec{u}[n-1] = H_{n-1} F_{n-2} \cdots F_0 \vec{x}[0] + H_{n-1} F_{n-2} \cdots F_1 G_0 \vec{u}[0] + \cdots + H_{n-1} G_{n-2} \vec{u}[n-2] \quad (4.41)$$

and rearranging as

$$\begin{bmatrix} \vec{y}[0] \\ \vec{y}[1] \\ \vdots \\ \vec{y}[n-2] \\ \vec{y}[n-1] \end{bmatrix} - \begin{bmatrix} D_0 & 0 & \cdots & 0 & 0 \\ HG_0 & D_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ HF^{n-3}G & HF^{n-4}G & \cdots & D_{n-2} & 0 \\ H_{n-1}F^{n-2}G_0 & H_{n-1}F^{n-3}G_1 & \cdots & H_{n-1}G_{n-2} & D_{n-1} \end{bmatrix} \begin{bmatrix} \vec{u}[0] \\ \vec{u}[1] \\ \vdots \\ \vec{u}[n-2] \\ \vec{u}[n-1] \end{bmatrix} = \begin{bmatrix} H_0 \\ H_1 F_0 \\ \vdots \\ H_{n-1} F_{n-2} \cdots F_0 \end{bmatrix} \vec{x}[0] \quad (4.42)$$

which has the form

$$\vec{y} = \mathcal{H} \vec{x}[0] \quad (4.43)$$

where  $\mathcal{H}$  is the **observability matrix**. Thus, by inspection one can see that *any*  $\vec{x}[0]$  can be observed if and only if **observability matrix**  $\mathcal{H}$  has full row rank, i.e.  $\text{rank} = n_x$ . The initial state can be computed by

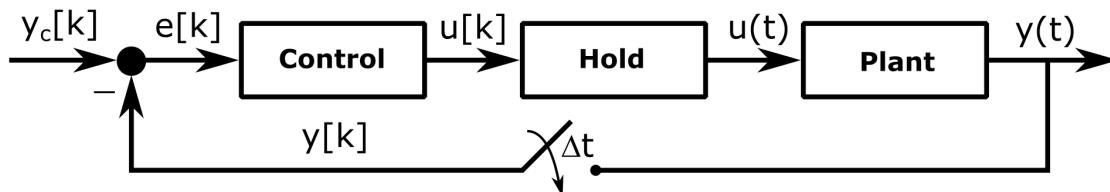
$$\vec{x}(0) = (\mathcal{H}^T \mathcal{H})^{-1} \mathcal{H}^T \vec{y} \quad (4.44)$$

where  $(\mathcal{H}^T \mathcal{H})^{-1} \mathcal{H}^T$  is known as the **pseudoinverse** of  $\mathcal{H}$  which exists if  $\mathcal{H}$  has full row rank.

### Hybrid-Time Feedback Control Systems

Before modern computers, flight controller design was performed using standard circuit components to implement continuous-time control laws, i.e. “classical” control. However, modern flight control systems are implemented by microprocessors or digital computers where one has a digital control system where time and values are discrete. Thus, there are additional considerations that must be taken for hardware implementation of the control system concepts discussed in this textbook. The additional effects of discrete-values in truly digital systems is beyond the scope of this textbook, but generally has little effect if the numerical precision of the digital computer is high enough relative to the system dynamics.

One approach is to use the continuous-time control design methods in this course and then convert the continuous-time control system to a discrete-time control system, i.e., **discretization**. However, one may also directly design the discrete-time feedback control system by modeling the discrete-time effects with the continuous-time plant while using discrete-time control methods for the controller. This is called a **hybrid-time feedback control system**. In general, the block diagram for a **hybrid-time feedback control system** can be considered as the following.



This system notably consists of the following hybrid processing steps in addition to the continuous-time plant.

- **Sampling:** The output of the system,  $y(t)$  is updated in the digital system every  $\Delta t$  forming the  $k^{\text{th}}$  sampled output,  $y[k]$ , where  $k = 1, 2, \dots$  is the time step, i.e.  $t = k\Delta t$ . This is typically performed by an analog-to-digital converter (ADC).
- **Control Update:** The control input at time step  $k$ ,  $u[k]$ , is updated using a digital computer. This is typically performed by a microprocessor with a “real-time” operating system (RTOS).
- **Hold:** The updated control input,  $u[k]$ , is converted into a continuous-time signal  $u(t)$  for the physical (e.g. electro-mechanical) system. This is typically performed by a digital-to-analog converter (DAC).

It should be noted that the signal processing from system to system will also cause some small time delays that should be modeled in the full system analysis and typically would also include an actuator in the plant model. Furthermore, if one is simply discretizing a continuous-time control law to discrete-time without this additional analysis, one must ensure that the Nyquist frequency,  $\Delta t/2$ , is much faster than any relevant dynamics. However, if this is not that case, one must consider the real continuous-time effects of the digital feedback control system primarily including the sampling and hold steps.

## 4.2 Discrete-Time Linear-Quadratic Regulator

### Discrete-Time Systems and Dynamic Programming

The OCP for discrete-time dynamical systems can be generally written as

$$\vec{u}^{opt}[k] = \underset{\vec{u}[k] \text{ for } k=0, \dots, N-1}{\operatorname{argmin}} \quad \mathcal{J} = \mathcal{E}(\vec{x}[N], N) + \sum_{k=0}^{N-1} \mathcal{L}(\vec{x}[k], \vec{u}[k], k)$$

subject to:

discrete dynamics:  $\vec{x}[k+1] = f(\vec{x}[k], \vec{u}[k], k)$   
initial condition:  $\vec{x}[0] = \vec{x}_0$   
constraints:  $c(\vec{x}[k], \vec{u}[k], k) \leq 0$

(4.45)

where  $\vec{u}^{opt}[k]$  is the **optimal control sequence**,  $N$  is the final time step, also known as the discrete-time **time horizon** since  $k$  starts at zero without loss of generality), and  $\mathcal{J}$  is the **objective function**, also known as the **cost function**. Comparing this discrete-time OCP to the continuous-time OCP, the integration became a summation and the differential equation became a difference equation.

A key concept in optimization over time is one can perform the optimization in stages. In essence, one is balancing the lowest possible cost at the present stage against the impact this would have for costs at future stages. The optimal control action minimizes the sum of the cost incurred at the current stage and the least total cost that can be incurred from all subsequent stages, consequent on this decision. This is known as the **principle of optimality** which states that from any point on an optimal trajectory, the remaining trajectory

is optimal for the corresponding problem initiated at that point. Formally for discrete-time OCPs, this can be stated by defining the **cost-to-go function** at time step  $k$  as

$$\mathcal{J}_k(\vec{x}[k]) = \mathcal{E}(\vec{x}[N]) + \sum_{i=k}^{N-1} \mathcal{L}(\vec{x}[i], \vec{u}[i]) \quad (4.46)$$

which has a corresponding sub-optimization problem for the minimum cost-to-go,  $\mathcal{V}_k(\vec{x}[k])$ , also known as the **value function**, as

$$\mathcal{V}_k(\vec{x}[k]) = \min_{\vec{u}[i] \text{ for } i=k, \dots, N-1} \mathcal{J}_k = \mathcal{E}(\vec{x}[N], N) + \sum_{i=k}^{N-1} \mathcal{L}(\vec{x}[i], \vec{u}[i]) \quad (4.47)$$

subject to the dynamics, initial conditions, and constraints.

Considering  $\mathcal{V}_k$  as a function of  $k$  and  $\vec{x}[k]$  with the following properties

$$\mathcal{V}_N(\vec{x}[N]) = \mathcal{E}(\vec{x}[N], N) \quad (4.48)$$

and

$$\mathcal{V}_0(\vec{x}[0]) = \min_{\vec{u}[k] \text{ for } k=0, \dots, N-1} \mathcal{J} \quad (4.49)$$

Furthermore, this formalization allows one to write the sub-optimization as a recursive relationship

$$\mathcal{V}_k(\vec{x}[k]) = \min_{\vec{u}[i] \text{ for } i=k, \dots, N-1} \mathcal{L}(\vec{x}[k], \vec{u}[k]) + \mathcal{V}_{k+1}(\vec{x}[k+1]) \quad (4.50)$$

which is the OCP form of the **Bellman equation**, also known as the **optimality equation**. Thus, one can begin at  $\mathcal{V}_N(\vec{x}[N])$  and calculate the value function at previous time steps by working backwards using the Bellman equation and finally obtaining  $\mathcal{V}_0(\vec{x}[0])$  as the value of the optimal cost. Then, the optimal control sequence,  $\vec{u}^{\text{opt}}[k]$ , can then be recovered by tracing back the calculations already performed for the value function.

The Bellman equation serves as the fundamental result of dynamic programming (DP) and is often referred to as the **dynamic programming equation**. By definition, **dynamic programming (DP)** solves an optimization problem by recursively solving simpler sub-problems which make up the overall problem. In general, not all optimization problems allow dynamic programming methods, however optimization problems over a sequence of time steps often allow recursive sub-problems to be nested inside the overall problem. This time sequencing is the origin of the term “dynamic” in dynamic programming. Mathematically, these sub-problems done by defining a sequence of value functions, as shown previously for discrete-time OCPs. The value function at any time step is valued based on future time steps and thus can be used to compute the minimum cost-to-go function.

## Unconstrained Discrete-Time LQR

The **unconstrained finite-horizon discrete-time LQ OCP** can be stated as

$$\begin{aligned} \vec{u}^{\text{opt}}[k] &= \underset{\vec{u}[k] \text{ for } k=0, \dots, N-1}{\operatorname{argmin}} \mathcal{J} = \vec{x}[N]^T E \vec{x}[N] + \sum_{k=0}^{N-1} \vec{x}[k]^T Q \vec{x}[k] + \vec{u}[k]^T R \vec{u}[k] + 2 \vec{x}[k]^T S \vec{u}[k] \\ &\text{subject to: } \vec{x}[k+1] = F \vec{x}[k] + G \vec{u}[k] \\ &\text{initial condition: } \vec{x}[0] = \vec{x}_0 \end{aligned} \quad (4.51)$$

where the unconstrained finite-horizon discrete-time LQR is the optimal control sequence,  $\vec{u}^{\text{opt}}[k]$ , which minimizes the quadratic cost function,  $\mathcal{J}$ .

Recalling the dynamic programming solution method, one can define the following value function for this discrete-time OCP:

$$\mathcal{V}_k(\vec{x}[k]) = \min_{\vec{u}[k] \text{ for } k=0, \dots, N-1} \vec{x}_N^T E \vec{x}_N + \sum_{\tau=k}^{N-1} \vec{x}_\tau^T Q \vec{x}_\tau + \vec{u}_\tau^T R \vec{u}_\tau + 2 \vec{x}_\tau^T S \vec{u}_\tau \quad (4.52)$$

which has the boundary condition

$$\mathcal{V}_k(\vec{x}) = \vec{x}^T E \vec{x} \quad (4.53)$$

Moreover, the incurred cost for the LQ OCP can be defined as

$$\mathcal{L}(\vec{x}[k], \vec{u}[k]) = \vec{x}^T Q \vec{x} + \vec{u}^T R \vec{u} \quad (4.54)$$

and the cost-to-go from time step  $k$  due to “next” state governed by the linear dynamics of the LQR OCP can be identified as

$$\mathcal{V}_{k+1}(\vec{x}[k+1]) = \mathcal{V}_{k+1}(F \vec{x}[k] + G \vec{u}[k]) \quad (4.55)$$

which, by the Bellman equation, one has

$$\mathcal{V}_k(\vec{x}[k]) = \min_{\vec{u}} \left( \vec{x}^T Q \vec{x} + \vec{u}^T R \vec{u} + 2 \vec{x}^T S \vec{u} + \mathcal{V}_{k+1}(F \vec{x} + G \vec{u}) \right) \quad (4.56)$$

To solve this OCP, assume the value function has a quadratic form, i.e.

$$\mathcal{V}_k(\vec{x}) = \vec{x}^T P[k] \vec{x} \quad (4.57)$$

with  $P[k]$  being a symmetric and positive semi-definite matrix and with the condition

$$P[N] = E \quad (4.58)$$

and by substitution into the Bellman equation, one has

$$\mathcal{V}_{k-1}(\vec{x}) = \vec{x}^T Q \vec{x} + \vec{u}^T R \vec{u} + 2 \vec{x}^T S \vec{u} + (F \vec{x} + G \vec{u})^T P[k] (F \vec{x} + G \vec{u}) \quad (4.59)$$

and setting

$$\frac{d\mathcal{V}_{k-1}}{d\vec{u}} = 0 \quad (4.60)$$

one has

$$2 \vec{u}^{\text{opt}}{}^T R + 2 \vec{x}^T S + 2(F \vec{x} + G \vec{u}^{\text{opt}})^T P[k] G = 0 \quad (4.61)$$

Dividing by 2 and rearranging, one has

$$\vec{u}^{\text{opt}}{}^T R + \vec{u}^{\text{opt}}{}^T G^T P[k] G = -\vec{x}^T (F^T P[k] G + S) \quad (4.62)$$

and taking the transpose

$$R \vec{u}^{\text{opt}} + G^T P[k] G \vec{u}^{\text{opt}} = -(G^T P[k] F + S^T) \vec{x} \quad (4.63)$$

and the inverse

$$\vec{u}^{\text{opt}} = - \left( G^T P[k] G + R \right)^{-1} \left( G^T P[k] F + S^T \right) \vec{x} \quad (4.64)$$

one obtains the optimal control input for a single time step and can be summarized for entire sequence as

$$\vec{u}^{\text{opt}}[k] = -K[k]x[k] \quad (4.65)$$

with

$$K[k] = \left( G^T P[k] G + R \right)^{-1} \left( G^T P[k] F + S^T \right) \quad (4.66)$$

where one still must solve for  $P[k]$  to get an explicit solution.

First, using  $\vec{u}^{\text{opt}}$  and

$$2\vec{x}^T S \vec{u}^{\text{opt}} = \vec{x}^T S \vec{u}^{\text{opt}} + \vec{u}^{\text{opt}}^T S \vec{x} \quad (4.67)$$

one can rewrite the Bellman equation as

$$\mathcal{V}_{k-1}(\vec{x}) = \vec{x}^T Q \vec{x} + \vec{u}^{\text{opt}}^T R \vec{u}^{\text{opt}} + \vec{x}^T S \vec{u}^{\text{opt}} + \vec{u}^{\text{opt}}^T S \vec{x} + (F \vec{x} + G \vec{u}^{\text{opt}})^T P[k] (F \vec{x} + G \vec{u}^{\text{opt}}) \quad (4.68)$$

and distributing out the quadratic components, one has four terms

$$\begin{aligned} \mathcal{V}_{k-1}(\vec{x}) &= \vec{x}^T \left( Q + F^T P[k] F \right) \vec{x} \\ &\quad + \vec{u}^{\text{opt}}^T \left( G^T P[k] G + R \right) \vec{u}^{\text{opt}} \\ &\quad + \vec{x}^T \left( F^T P[k] G + S \right) \vec{u}^{\text{opt}} \\ &\quad + \vec{u}^{\text{opt}}^T \left( G^T P[k] F + S^T \right) \vec{x} \end{aligned} \quad (4.69)$$

Next, one can substitute for  $\vec{u}^{\text{opt}}$  from the solution earlier and by using the equivalent

$$\vec{u}^{\text{opt}}^T = -\vec{x}^T \left( F^T P[k] G + S \right) \left( G^T P[k] G + R \right)^{-1} \quad (4.70)$$

the Bellman equation becomes

$$\begin{aligned} \mathcal{V}_{k-1}(\vec{x}) &= \vec{x}^T \left( Q + F^T P[k] F \right) \vec{x} \\ &\quad + \vec{x}^T \left( F^T P[k] G + S \right) \left( G^T P[k] G + R \right)^{-1} \\ &\quad \left( G^T P[k] F + S^T \right) \vec{x} \\ &\quad - \vec{x}^T \left( F^T P[k] G + S \right) \left( G^T P[k] G + R \right)^{-1} \\ &\quad \left( G^T P[k] F + S^T \right) \vec{x} \\ &\quad - \vec{x}^T \left( F^T P[k] G + S \right) \left( G^T P[k] G + R \right)^{-1} \\ &\quad \left( G^T P[k] F + S^T \right) \vec{x} \end{aligned} \quad (4.71)$$

and by simplifying

$$\begin{aligned}\mathcal{V}_{k-1}(\vec{x}) &= \vec{x}^T P[k-1] \vec{x} = \vec{x}^T \left[ F^T P[k] F - \left( F^T P[k] G + S \right) \right. \\ &\quad \left. \left( G^T P[k] G + R \right)^{-1} \left( G^T P[k] F + S^T \right) \right] \vec{x}\end{aligned}\tag{4.72}$$

Finally, the quadratic form for the value function provides a recursive equation for  $P[k-1]$  given  $P[k]$  as

$$P[k-1] = F^T P[k] F + Q - \left( F^T P[k] G + S \right) \left( G^T P[k] G + R \right)^{-1} \left( G^T P[k] F + S^T \right)\tag{4.73}$$

which is known as the **dynamic Riccati equation** which is a backwards recursion formula to solve for the discrete-time **Riccati matrix**,  $P[k]$ , over some time horizon  $N$  starting from  $P[N] = E$ .

Finally, the **unconstrained finite-horizon discrete-time LQR** has the form

$$\vec{u}^{\text{opt}} = -K[k] \vec{x}[k] \quad \forall k\tag{4.74}$$

where

$$K[k] = \left( G^T P[k+1] G + R \right)^{-1} \left( G^T P[k+1] F + S^T \right)\tag{4.75}$$

which notably results in closed-loop state-space dynamics represented by

$$\vec{x}[k+1] = (F - GK[k]) \vec{x}[k]\tag{4.76}$$

Finally, if one considers the **unconstrained infinite-horizon discrete-time LQ OCP**, the solution is given by the steady-state of the dynamic Riccati equation, i.e.

$$P = F^T P F + Q - \left( F^T P G + S \right) \left( G^T P G + R \right)^{-1} \left( G^T P F + S^T \right)\tag{4.77}$$

which is known as the **discrete-time algebraic Riccati equation (DARE)**.

It should be noted that if one has LTV dynamics and/or time-varying cost matrices, the dynamic Riccati equation still applies. However, a steady-state solution will exist if for  $N \gg 1$ , the difference between  $P[N] - P[N-1] \rightarrow 0$  as  $t \rightarrow 0$ . Otherwise, the discrete-time LTV LQR cannot only be implemented for an infinite-horizon. Typically, in order to accomplish this steady-state, one must design  $Q[k]$  and  $R[k]$  to have some structure.

### 4.3 Iterative and Extended Linear-Quadratic Regulator

For nonlinear and/or non-quadratic discrete-time OCPs, one can often form locally-optimal control sequences using approximation methods for the OCP. Iterative LQR (ILQR) and extended LQR (ELQR) are two common LQR-based methods derived that use forwards-backwards iterative procedures with linearization for the dynamics and/or quadratization methods for the cost functions. It should also be pointed out that these methods also appear in the iterative extended Kalman filter and the iterative extended Kalman smoother discussed later in this textbook for nonlinear Bayesian state estimation.

To this end, consider the unconstrained, nonlinear, non-quadratic finite-horizon discrete-time OCP written as

$$\begin{aligned} \vec{u}_k^{opt} = \underset{\vec{u}_k \text{ for } k=0, \dots, N-1}{\operatorname{argmin}} \quad \mathcal{J} = \mathcal{E}(\vec{x}_N) + \sum_{k=0}^{N-1} \mathcal{L}_k(\vec{x}_k, \vec{u}_k) \\ \text{subject to:} \\ \text{discrete dynamics: } \vec{x}_{k+1} = f_k(\vec{x}_k, \vec{u}_k) \\ \text{initial condition: } \vec{x}_0 \end{aligned} \quad (4.78)$$

where the solution will provide a locally optimal control sequence of the form

$$\vec{u}_k = \pi_k(\vec{x}_k) \quad k = 0, \dots, N-1 \quad (4.79)$$

which has a definable inverse or “backward” control sequence

$$\vec{u}_k = \tilde{\pi}_k(\vec{x}_{k+1}) \quad k = 0, \dots, N-1 \quad (4.80)$$

as opposed to the forward control law.

Consider a nominal trajectory defined as  $\hat{\vec{x}}_k$  with  $k = 0, \dots, N$  with a nominal control input as  $\hat{\vec{u}}_k$  for  $k = 0, \dots, N-1$  where the state trajectory is computed via the (forward) dynamics as

$$\hat{\vec{x}}_{k+1} = f_k(\hat{\vec{x}}_k, \hat{\vec{u}}_k) \quad (4.81)$$

and the control input is computed with the current chosen (forward) control law,  $\pi_k()$ , as

$$\hat{\vec{u}}_k = \pi_k(\hat{\vec{x}}_{k+1}) \quad (4.82)$$

or, alternatively, the state trajectory is computed via the backward dynamics,  $\tilde{f}_k()$  as

$$\hat{\vec{x}}_k = \tilde{f}_k(\hat{\vec{x}}_{k+1}, \hat{\vec{u}}_k) \quad (4.83)$$

and the control input is computed with the current chosen backward control law,  $\tilde{\pi}_k()$ , as

$$\hat{\vec{u}}_k = \tilde{\pi}_k(\hat{\vec{x}}_{k+1}) \quad (4.84)$$

In the I/ELQR approaches, assume one can quadratize the local cost function as

$$\delta \mathcal{J}_k = \frac{1}{2} \vec{x}_N^T E \vec{x}_N + \vec{x}_N^T \vec{e} + \sum_{k=0}^{N-1} \frac{1}{2} \left[ \begin{array}{c} \vec{x}_k \\ \vec{u}_k \end{array} \right]^T \left[ \begin{array}{cc} Q_k & S_k^T \\ S_k & R_k \end{array} \right] \left[ \begin{array}{c} \vec{x}_k \\ \vec{u}_k \end{array} \right] + \left[ \begin{array}{c} \vec{q}_k \\ \vec{r}_k \end{array} \right]^T \left[ \begin{array}{c} \vec{q}_k \\ \vec{r}_k \end{array} \right] \quad (4.85)$$

where

$$E = \left. \frac{\partial^2 \mathcal{E}}{\partial \vec{x}_k \partial \vec{x}_k} \right|_{\vec{x}_N = \hat{\vec{x}}_N} \quad (4.86)$$

$$\vec{e} = \left. \frac{\partial \mathcal{E}}{\partial \vec{x}_k} \right|_{\vec{x}_N = \hat{\vec{x}}_N} - E \hat{\vec{x}}_N \quad (4.87)$$

$$\begin{bmatrix} Q_k & S_k^T \\ S_k & R_k \end{bmatrix} = \frac{\partial^2 \mathcal{L}_k}{\partial \begin{bmatrix} \vec{x}_k \\ \vec{u}_k \end{bmatrix} \partial \begin{bmatrix} \vec{x}_k \\ \vec{u}_k \end{bmatrix}} \Bigg|_{\vec{x}_k = \hat{\vec{x}}_k, \vec{u}_k = \hat{\vec{u}}_k} \quad (4.88)$$

and

$$\begin{bmatrix} \vec{q}_k \\ \vec{r}_k \end{bmatrix} = \frac{\partial \mathcal{L}_k}{\partial \begin{bmatrix} \vec{x}_k \\ \vec{u}_k \end{bmatrix}} \Bigg|_{\vec{x}_k = \hat{\vec{x}}_k, \vec{u}_k = \hat{\vec{u}}_k} - \begin{bmatrix} Q_k & S_k^T \\ S_k & R_k \end{bmatrix} \begin{bmatrix} \hat{\vec{x}}_k \\ \hat{\vec{u}}_k \end{bmatrix} \quad (4.89)$$

where one must enforce  $E > 0$ ,  $Q_k \geq 0$  and  $R_k > 0$  for all  $k$ .

In addition, for the I/ELQR approaches, assume one can linearize the forward dynamics with

$$F_k = \frac{\partial f_k}{\partial \vec{x}_k} \Bigg|_{\vec{x}_k = \hat{\vec{x}}_k, \vec{u}_k = \hat{\vec{u}}_k} \quad (4.90)$$

and

$$G_k = \frac{\partial f_k}{\partial \vec{u}_k} \Bigg|_{\vec{x}_k = \hat{\vec{x}}_k, \vec{u}_k = \hat{\vec{u}}_k} \quad (4.91)$$

with a linearization error of

$$\delta \vec{x}_{k+1} = \hat{\vec{x}}_{k+1} - F_k \hat{\vec{x}}_k - G_k \hat{\vec{u}}_k \quad (4.92)$$

This provides an affine forward dynamics model as

$$\hat{\vec{x}}_{k+1} = F_k \hat{\vec{x}}_k + G_k \hat{\vec{u}}_k + \delta \vec{x}_{k+1} \quad (4.93)$$

Lastly, for ELQR, assume one can linearize the backward dynamics with

$$\tilde{F}_k = \frac{\partial \tilde{f}_k}{\partial \vec{x}_{k+1}} \Bigg|_{\vec{x}_{k+1} = \hat{\vec{x}}_{k+1}, \vec{u}_k = \hat{\vec{u}}_k} \quad (4.94)$$

and

$$\tilde{G}_k = \frac{\partial \tilde{f}_k}{\partial \vec{u}_k} \Bigg|_{\vec{x}_k = \hat{\vec{x}}_k, \vec{u}_k = \hat{\vec{u}}_k} \quad (4.95)$$

with a linearization error of

$$\delta \tilde{\vec{x}}_k = \hat{\vec{x}}_k - \tilde{F}_k \hat{\vec{x}}_{k+1} - \tilde{G}_k \hat{\vec{u}}_k \quad (4.96)$$

This provides an affine backward dynamics model as

$$\hat{\vec{x}}_k = \tilde{F}_k \hat{\vec{x}}_{k+1} + \tilde{G}_k \hat{\vec{u}}_k + \delta \tilde{\vec{x}}_k \quad (4.97)$$

### Backward LQR Synthesis via Cost-to-Go Functions

With this linearization-quadratization formulation in mind, consider the case where one has an affine forward dynamics model as

$$\vec{x}_{k+1} = F_k \vec{x}_k + G_k \vec{u}_k + \delta \vec{x}_{k+1} \quad (4.98)$$

and the cost function is quadratic, i.e.,

$$\mathcal{J} = \frac{1}{2} \vec{x}_N^T E \vec{x}_N + \vec{x}_N^T \vec{e} + \sum_{k=0}^{N-1} \frac{1}{2} \begin{bmatrix} \vec{x}_k \\ \vec{u}_k \end{bmatrix}^T \begin{bmatrix} Q_k & S_k^T \\ S_k & R_k \end{bmatrix} \begin{bmatrix} \vec{x}_k \\ \vec{u}_k \end{bmatrix} + \begin{bmatrix} \vec{x}_k \\ \vec{u}_k \end{bmatrix}^T \begin{bmatrix} \vec{q}_k \\ \vec{r}_k \end{bmatrix} \quad (4.99)$$

Then the optimal solution, similar to the previous derivation of the standard LQR, can be synthesized by solving for the minimum cost-to-go function,  $\mathcal{V}_k$ , which solves the optimization problem

$$\mathcal{V}_k(\vec{x}_k) = \min_{\vec{u}[k] \text{ for } k=0, \dots, N-1} \mathcal{V}_{k+1}(\vec{x}_{k+1}) + \frac{1}{2} \begin{bmatrix} \vec{x}_k \\ \vec{u}_k \end{bmatrix}^T \begin{bmatrix} Q_k & S_k^T \\ S_k & R_k \end{bmatrix} \begin{bmatrix} \vec{x}_k \\ \vec{u}_k \end{bmatrix} + \begin{bmatrix} \vec{x}_k \\ \vec{u}_k \end{bmatrix}^T \begin{bmatrix} \vec{q}_k \\ \vec{r}_k \end{bmatrix} \quad (4.100)$$

with final condition

$$\mathcal{V}_N(\vec{x}_N) = \frac{1}{2} \vec{x}_N^T E \vec{x}_N + \vec{x}_N^T \vec{e} \quad (4.101)$$

The solution to this general backwards LQR is obtained by assuming  $\mathcal{V}_k$  has the general quadratic form of

$$\mathcal{V}_k(\vec{x}_k) = \frac{1}{2} \vec{x}_k^T P_k \vec{x}_k + \vec{x}_k^T \vec{p}_k + c \quad (4.102)$$

where  $c$  is a constant. Then, the optimal  $P_k > 0$  can be shown to be computed backwards via the Riccati difference equation

$$\begin{aligned} P_k = & F_k^T P_{k+1} F_k + Q_k \\ & - (G_k^T P_{k+1} F_k + S_k)^T \\ & (G_k^T P_{k+1} G_k + R_k)^{-1} \\ & (G_k^T P_{k+1} F_k + S_k) \end{aligned} \quad (4.103)$$

and the optimal  $\vec{p}_k$  is computed backwards via the recursive equation

$$\begin{aligned} \vec{p}_k = & F_k^T \vec{p}_{k+1} + F_k^T P_{k+1} \delta \vec{x}_k + \vec{q}_k \\ & - (G_k^T P_{k+1} F_k + S_k)^T (G_k^T P_{k+1} G_k + R_k)^{-1} \\ & (G_k^T \vec{p}_{k+1} + G_k^T P_{k+1} \delta \vec{x}_k + \vec{r}_k) \end{aligned} \quad (4.104)$$

with initial values,  $P_N = E$ , and  $\vec{p}_N = \vec{e}$ . Finally, the backward LQR optimal control law has the linear form

$$\begin{aligned} \vec{u}_k = \pi(\vec{x}_k) = & - (G_k^T P_{k+1} G_k + R_k)^{-1} \left( (G_k^T P_{k+1} F_k + S_k) \vec{x}_k \right. \\ & \left. + (G_k^T \vec{p}_{k+1} + G_k^T P_{k+1} \delta \vec{x}_k + \vec{r}_k) \right) \end{aligned} \quad (4.105)$$

## Iterative LQR Synthesis

The **iterative linear-quadratic regulator (ILQR)** solves the backwards LQR solution iteratively for the linearized dynamics and quadratized local cost function provided one has an initial nominal state trajectory, i.e.,  $(\hat{\vec{x}}_{k,0}, \hat{\vec{u}}_{k,0})$  for  $k = 0, \dots, N - 1$ . If no initial nominal state trajectory is known, then one often sets  $\hat{\vec{u}}_{k,0} = \vec{0}$  for  $k = 0, \dots, N - 1$  and solves for  $\hat{\vec{x}}_{k,0}$  using the forward dynamics. The  $i^{\text{th}}$  iteration of the backwards LQR provides a new nominal locally optimal control law

$$\hat{\vec{u}}_{k,i} = \pi_{k,i}(\hat{\vec{x}}_{k,i}) \quad (4.106)$$

where one ends the iterative procedure for some convergence criteria, e.g.,

$$\sum_{k=0}^{N-1} \|\hat{\vec{u}}_{k,i} - \hat{\vec{u}}_{k,i-1}\|_2 < \tau \quad (4.107)$$

where  $\tau$  is the convergence threshold. If the control has not converged, then one computes the  $i+1^{\text{th}}$  iteration, by first computing the  $i^{\text{th}}$  nominal state trajectory for  $k = 0, \dots, N - 1$  forwards in time as

$$\hat{\vec{x}}_{k+1,i} = f_{k,i}(\hat{\vec{x}}_{k,i}, \pi_{k,i}(\hat{\vec{x}}_{k,i})) \quad (4.108)$$

Next, quadratizing and linearizing about  $\hat{\vec{x}}_{k,i}$  and  $\hat{\vec{u}}_{k,i}$  to obtain  $E_{i+1}$ ,  $\vec{e}_{i+1}$ ,  $Q_{k,i+1}$ ,  $R_{k,i+1}$ ,  $S_{k,i+1}$ ,  $\vec{q}_{k,i+1}$ ,  $\vec{r}_{k,i+1}$ ,  $F_{k,i+1}$ , and  $G_{k,i+1}$ . Finally, resolving the backward LQR to obtain  $\pi_{k,i+1}$  and checking for convergence.

## Forward LQR Synthesis via Cost-to-Come Functions

With the backward linearization-quadratization formulation in mind, consider the case where one has an affine backward dynamics model as

$$\vec{x}_k = \tilde{F}_k \vec{x}_k + \tilde{G}_k \vec{u}_k + \delta \vec{x}_k \quad (4.109)$$

the cost function is the same as before. However, the optimal solution, similar to the previous derivation of the backward LQR, can be synthesized by assuming the minimum **cost-to-come function**, also known as the **forward value function**,  $\tilde{\mathcal{V}}_k$ , which an initial condition of

$$\mathcal{V}_0(\vec{x}_0) = \vec{0} \quad (4.110)$$

and a general quadratic form

$$\tilde{\mathcal{V}}_k(\vec{x}_k) = \frac{1}{2} \vec{x}_k^T \tilde{P}_k \vec{x}_k + \vec{x}_k^T \tilde{P}_k \vec{x}_k + \tilde{c} \quad (4.111)$$

where  $\tilde{c}$  is a constant.

Then, the optimal  $\tilde{P}_k > 0$  is computed forwards via the Riccati difference equation

$$\begin{aligned} \tilde{P}_{k+1} = & \tilde{F}_k^T (\tilde{P}_k + Q_k) \tilde{F}_k \\ & - \left( \tilde{G}_k^T (\tilde{P}_k + Q_k) \tilde{F}_k + S_k \tilde{F}_k \right)^T \\ & \left( \tilde{G}_k^T (\tilde{P}_k + Q_k) \tilde{G}_k + R_k + S_k \tilde{G}_k + \tilde{G}_k^T S_k^T \right)^{-1} \\ & \left( \tilde{G}_k^T (\tilde{P}_k + Q_k) \tilde{F}_k + S_k \tilde{F}_k \right) \end{aligned} \quad (4.112)$$

and  $\tilde{\vec{p}}_k$  is computed forwards via the recursive equation

$$\begin{aligned}\tilde{\vec{p}}_{k+1} = & \tilde{F}_k^T(\tilde{\vec{p}}_k + \vec{q}_k) + \tilde{F}_k^T(\tilde{P}_k + Q_k)\delta\tilde{\vec{x}}_k \\ & \left(\tilde{G}_k^T(\tilde{P}_k + Q_k)\tilde{F}_k + S_k\tilde{F}_k\right)^T \\ & \left(\tilde{G}_k^T(\tilde{P}_k + Q_k)\tilde{G}_k + R_k + S_k\tilde{G}_k + \tilde{G}^T S_k^T\right)^{-1} \\ & \left(\tilde{G}_k^T(\tilde{\vec{p}}_k + \vec{q}_k) + (\tilde{G}_k^T(\tilde{P}_k + Q_k) + S_k)\delta\tilde{\vec{x}}_k + \vec{r}_k\right)\end{aligned}\quad (4.113)$$

with initial values,  $\tilde{P}_0 = [0]$ , and  $\tilde{\vec{p}}_0 = \vec{0}$ .

Finally, the forward LQR optimal control law has a linear form as

$$\begin{aligned}\vec{u}_k = \tilde{\pi}(\vec{x}_{k+1}) = & -\left(\tilde{G}_k^T(\tilde{P}_k + Q_k)\tilde{G}_k + R_k + S_k\tilde{G}_k + \tilde{G}^T S_k^T\right)^{-1} \\ & \left(\left(\tilde{G}_k^T(\tilde{P}_k + Q_k)\tilde{F}_k + S_k\tilde{F}_k\right)\vec{x}_{k+1}\right. \\ & \left.+\left(\tilde{G}_k^T(\tilde{\vec{p}}_k + \vec{q}_k) + (\tilde{G}_k^T(\tilde{P}_k + Q_k) + S_k)\delta\tilde{\vec{x}}_k + \vec{r}_k\right)\right)\end{aligned}\quad (4.114)$$

## Extended LQR Synthesis

The **extended linear-quadratic regulator (ELQR)** which, like the ILQR, linearizes the forward dynamics and quadratizes the local cost function in the backward iteration of the cost-to-go functions, but, instead of recomputing the forward dynamics, the ELQR linearizes the backward dynamics, quadratizes the local cost function, and computes the forward LQR in the forwards iteration of the cost-to-come functions. This requires twice as many linearization and quadratization computations than the ILQR and twice as much memory for the dynamics for forward and backward, but can provide significantly less iterations than the ILQR for certain types of nonlinear and non-quadratic OCPs.

To optimally combine the two procedures of a backward and forward LQR for an accurate linearization/quadratization point, the Extended LQR takes advantage of the minimum-cost sequence of states and controls at time step  $k$  as the total-cost,  $\mathcal{V}_k^{tot}(\vec{x}_k)$ , defined as the sum of the minimum cost-to-go and cost-to-come functions, i.e.,

$$\mathcal{V}_k^{tot}(\vec{x}_k) = \mathcal{V}_k(\vec{x}_k) + \tilde{\mathcal{V}}_k(\vec{x}_k) = \frac{1}{2}\vec{x}_k^T(P_k + \tilde{P}_k)\vec{x}_k + \vec{x}_k^T(\vec{p}_k + \tilde{\vec{p}}_k) + (c + \tilde{c})\quad (4.115)$$

Then, the state trajectory for  $k = 1, \dots, N$  for which the  $\mathcal{V}_k^{tot}(\vec{x}_k)$  is minimum-cost can be written as simply

$$\hat{\vec{x}}_k = \underset{\vec{x}_k}{\operatorname{argmin}} \mathcal{V}_k^{tot}(\vec{x}_k) = -(P_k + \tilde{P}_k)^{-1}(\vec{p}_k + \tilde{\vec{p}}_k)\quad (4.116)$$

Then, the associated forward and backward control laws are given by

$$\hat{\vec{x}}_{k+1} = f_k(\hat{\vec{x}}_k, \pi_k(\hat{\vec{x}}_k))\quad (4.117)$$

and

$$\hat{\vec{x}}_k = \tilde{f}_k(\hat{\vec{x}}_{k+1}, \tilde{\pi}_k(\hat{\vec{x}}_{k+1})) \quad (4.118)$$

where  $\hat{\vec{u}}_k = \pi_k(\hat{\vec{x}}_k) = \tilde{\pi}_k(\hat{\vec{x}}_{k+1})$ . This can be understood as the LQR-dual of the fixed-interval Kalman smoother (FI-KS) discussed later in this textbook on state estimation, or as **LQR-smoothing**.

Each iteration of the ELQR is performed as follows. First, one computes the backward nominal control input using the previous forward LQR control law as

$$\hat{\vec{u}}_{k,i\leftarrow} = \tilde{\pi}_{k,i-1}(\hat{\vec{x}}_{k+1,i\leftarrow}) \quad (4.119)$$

simultaneously with the backward nominal state trajectory using the backwards dynamics as

$$\hat{\vec{x}}_{k,i\leftarrow} = \tilde{f}_{k,i}(\hat{\vec{x}}_{k+1,i\leftarrow}, \hat{\vec{u}}_{k,i\leftarrow}) \quad (4.120)$$

which can be initialized at  $i = 1$  according to some prior information or by assuming  $\tilde{\pi}_{k,0} = \vec{0}$  for all  $k$ , and  $\hat{\vec{x}}_{N,0\leftarrow} = 0$ . Next, one linearizes the forward dynamics and quadratizes the local cost function about  $\hat{\vec{x}}_{k,i\leftarrow}$  and  $\hat{\vec{u}}_{k,i\leftarrow}$ . Then, one computes the backward LQR control law as  $\pi_{k,i}(\hat{\vec{x}}_k)$ .

This backward LQR control law provides the forward control input as

$$\hat{\vec{u}}_{k,i\rightarrow} = \pi_{k,i}(\hat{\vec{x}}_{k,i\rightarrow}) \quad (4.121)$$

simultaneously with the forward nominal state trajectory using the forward dynamics as

$$\hat{\vec{x}}_{k+1,i\rightarrow} = f_{k,i}(\hat{\vec{x}}_{k,i\rightarrow}) \quad (4.122)$$

Next, one linearizes the backwards dynamics and quadratizes the local cost function about  $\hat{\vec{x}}_{k,i\rightarrow}$  and  $\hat{\vec{u}}_{k,i\rightarrow}$ . Then, one computes the forward LQR control as  $\tilde{\pi}_{k,i}(\hat{\vec{x}}_{k+1})$  up until  $k = N - 1$  with  $\tilde{P}_0 = [0]$  and  $\tilde{\vec{p}} = \vec{0}$ . In addition, the initial quadratization point for time step  $k$  for the next backward LQR pass is calculated using the minimum-cost state formula, i.e.,

$$\hat{\vec{x}}_N = -(P_N + \tilde{P}_N)^{-1}(\vec{p}_N + \tilde{\vec{p}}_N) \quad (4.123)$$

Finally, one can check for convergence at either pass based on  $\pi$  and  $\tilde{\pi}$

## References

For more information, please refer to the following

- Li, W. and Todorov, E., “Iterative Linear Quadratic Regulator Design for Nonlinear Biological Movement Systems,” in *Proceedings of the 1st International Conference on Informatics in Control, Automation and Robotics*, Setubal, Portugal, 2004
- Van Den Burg, J., “Extended LQR: Locally-Optimal Feedback Control for Systems with Non-Linear Dynamics and Non-Quadratic Cost,” in *Inaba, M., Corke, P. (eds) Robotics Research: Springer Tracts in Advanced Robotics*, Vol. 114. 2016

## 4.4 Receding-Horizon Linear-Quadratic Regulator

Another sub-optimal approach to solving discrete-time OCPs is **receding horizon control (RHC)**, also known as **model predictive control (MPC)**, which optimizes the control input over a **control horizon** of length  $M < N$  and re-optimizes at *every* time step of the control process,  $k$ , instead of optimizing over the entire time horizon  $N$ , also known as the **prediction horizon** in RHC. This approach also is more conducive for constrained OCPs as it can enforces state and input constraints remain true over the prediction horizon. Thus, the name RHC derives from the control and prediction horizons receding with each time step while the name MPC derives from the optimization method using a dynamics model to predict the state forward in time for the OCP solved at every time step.

The general receding horizon OCP can be constructed as

$$\begin{aligned} \vec{u}[0]^{opt}, \dots, \vec{u}[M-1]^{opt} = \underset{\vec{u}}{\operatorname{argmin}} \quad J = \mathcal{E}(\vec{x}[M]) + \sum_{k=0}^{M-1} \mathcal{L}(\vec{x}[k], \vec{u}[k]) \\ \text{subject to: } \vec{x}[k+1] = f(\vec{x}[k], \vec{u}[k]) \\ \text{initial condition: } \vec{x}_0 \\ \text{state constraints: } \vec{x}[k] \in \mathcal{X} \quad \forall 0 \leq k \leq N \\ \text{input constraints: } \vec{u}[k] \in \mathcal{U} \quad \forall 0 \leq k \leq M \end{aligned} \tag{4.124}$$

It is important to note that the state constraints are enforced over the prediction horizon and the input constraints are enforced over the control horizon. In addition, for  $M \leq k \leq N$ , the control input  $u[k]$ , must be specified. RHC often is computed in real-time and cannot be precomputed as a known controller. Thus, RHC serves as an open-loop control scheme since it recomputes its action at every time step and does not explicitly provide a control law for the feedback loop, though it does reuse the “new” state as the “new” initial in its re-optimization of the problem every time step. This feature of RHC may be especially desirable if the real process dynamics are not exactly represented in the dynamics model, especially nonlinearities. Thus, RHC continues to be used in many circumstances.

When the receding horizon OCP has linear dynamics, a quadratic cost, and minimum and/or maximum input and state constraints, one has a constrained receding horizon LQ OCP which can be stated as

$$\begin{aligned} \vec{u}_0^{opt}, \dots, \vec{u}_{M-1}^{opt} = \underset{\vec{u}}{\operatorname{argmin}} \quad J = \vec{x}^T[M]E\vec{x}[M] \\ + \sum_{k=0}^{M-1} \vec{x}^T[k]Q\vec{x}[k] \\ + \vec{u}^T[k]R\vec{u}[k] \\ \text{subject to: } \vec{x}[k+1] = F\vec{x}[k] + G\vec{u}[k] \\ \text{initial condition: } \vec{x}_0 \\ \text{state constraints: } \vec{x}_{min} \leq \vec{x}[k] \leq \vec{x}_{max} \quad \forall 0 \leq k \leq N \\ \text{input constraints: } \vec{u}_{min} \leq \vec{u}[k] \leq \vec{u}_{max} \quad \forall 0 \leq k \leq M \end{aligned} \tag{4.125}$$

for which the optimal control sequence is the receding horizon linear-quadratic regulator (RHLQR)

To formulate the RHLQR solution for an LTI state-space system with a quadratic cost function with constraints, one typically uses quadratic programming on the entire control horizon. Thus, based on the

discrete-time linear dynamics and initial condition, one can write out the general solution of the state as a “large” linear function, i.e.

$$\begin{bmatrix} \vec{x}[0] \\ \vdots \\ \vec{x}[M] \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ G & 0 & 0 & \cdots & 0 \\ FG & G & 0 & \cdots & 0 \\ F^2G & FG & G & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F^{M-1}G & F^{M-2}G & F^{M-3}G & \cdots & G \end{bmatrix} \begin{bmatrix} \vec{u}[0] \\ \vdots \\ \vec{u}[M-1] \end{bmatrix} + \begin{bmatrix} I \\ F \\ \vdots \\ F^M \end{bmatrix} \vec{x}_0 \quad (4.126)$$

which one can rewrite succinctly by defining the above vectors and matrices as

$$\vec{x} = \mathbf{G}\vec{u} + \mathbf{F}\vec{x}_0 = \mathbf{G}\vec{u} + \vec{x}_0 \quad (4.127)$$

Then, using these new terms and the following block diagonal matrices

$$\mathbf{Q} = \begin{bmatrix} Q & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & Q & 0 \\ 0 & \cdots & 0 & E \end{bmatrix} \quad (4.128)$$

and

$$\mathbf{R} = \begin{bmatrix} R & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & R \end{bmatrix} \quad (4.129)$$

the cost function for the discrete-time LQ OCP can be rewritten using vector notation as

$$J = \vec{x}^T \mathbf{Q} \vec{x} + \vec{u}^T \mathbf{R} \vec{u} \quad (4.130)$$

and, by substitution for  $\vec{x}$ , one has

$$J = (\mathbf{G}\vec{u} + \vec{x}_0)^T \mathbf{Q} (\mathbf{G}\vec{u} + \vec{x}_0) + \vec{u}^T \mathbf{R} \vec{u} \quad (4.131)$$

$$J = \vec{x}_0^T \mathbf{Q} \vec{x}_0 + 2\vec{x}_0^T \mathbf{Q} \mathbf{G} \vec{v} + \vec{v}^T \mathbf{G}^T \mathbf{Q} \mathbf{G} \vec{v} + \vec{v}^T \mathbf{R} \vec{v} \quad (4.132)$$

Finally, noting that the first term can be ignored since it is not affected by the choice of the argument  $\vec{v}$ , one can assign

$$\tilde{\mathbf{Q}} = \mathbf{G}^T \mathbf{Q} \mathbf{G} + \mathbf{R} \quad (4.133)$$

and

$$c^T = \vec{x}_0^T \mathbf{Q} \mathbf{G} \quad (4.134)$$

to rewrite the cost function in the form of the QP cost function.

$$J = \frac{1}{2} \vec{u}^T \tilde{\mathbf{Q}} \vec{u} + c^T \vec{u} \quad (4.135)$$

By inspection, one can see that the only parameter that changes in this optimization from one time step to the next is the term  $\vec{x}_0$  which changes with  $\vec{x}_0$  which is the current state at each subsequent time step as one moves forward in time.

In order to put the inequality constraints into the QP form above, stacking techniques can be used as follows. For the input constraints,

$$\begin{bmatrix} \vec{u}_{min} \\ \vdots \\ \vec{u}_{min} \end{bmatrix} \leq I\vec{u} \leq \begin{bmatrix} \vec{u}_{max} \\ \vdots \\ \vec{u}_{max} \end{bmatrix} \quad (4.136)$$

for the state constraints for  $1 \leq k \leq M$ ,

$$\begin{bmatrix} \vec{x}_{min} \\ \vdots \\ \vec{x}_{min} \end{bmatrix} - \vec{x}_0 \leq \mathbf{G}\vec{u} \leq \begin{bmatrix} \vec{x}_{max} \\ \vdots \\ \vec{x}_{max} \end{bmatrix} - \vec{x}_0 \quad (4.137)$$

For the state constraints for  $M + 1 \leq k \leq N$ ,

$$\begin{bmatrix} \vec{x}_{min} \\ \vdots \\ \vec{x}_{min} \end{bmatrix} \leq \begin{bmatrix} \vec{x}[M+1] \\ \vdots \\ \vec{x}[N] \end{bmatrix} \leq \begin{bmatrix} \vec{x}_{max} \\ \vdots \\ \vec{x}_{max} \end{bmatrix} \quad (4.138)$$

where the stacked states can be written as

$$\begin{bmatrix} \vec{x}[M+1] \\ \vdots \\ \vec{x}[N] \end{bmatrix} = \begin{bmatrix} F \\ \vdots \\ F^{N-M} \end{bmatrix} \vec{x}_M \quad (4.139)$$

and using the solution for the state at time step  $M$

$$\vec{x}_M = F^M \vec{x}_0 + [F^{M-1}G \quad \cdots \quad G] \vec{v} \quad (4.140)$$

one can write

$$\begin{bmatrix} \vec{x}[M+1] \\ \vdots \\ \vec{x}[N] \end{bmatrix} = \begin{bmatrix} F \\ \vdots \\ F^{N-M} \end{bmatrix} (F^M \vec{x}_0 + [F^{M-1}G \quad \cdots \quad G] \vec{u}) \quad (4.141)$$

Then, after substitution and rearrangement, the state constraints for  $M + 1 \leq k \leq N$  can finally be put into the QP form as

$$\begin{bmatrix} \vec{x}_{min} \\ \vdots \\ \vec{x}_{min} \end{bmatrix} - \begin{bmatrix} F \\ \vdots \\ F^{N-M} \end{bmatrix} F^M \vec{x}_0 \leq \begin{bmatrix} F \\ \vdots \\ F^{N-M} \end{bmatrix} [F^{M-1}G \quad \cdots \quad G] \vec{u} \leq \begin{bmatrix} \vec{x}_{max} \\ \vdots \\ \vec{x}_{max} \end{bmatrix} - \begin{bmatrix} F \\ \vdots \\ F^{N-M} \end{bmatrix} F^M \vec{x}_0 \quad (4.142)$$

It should also be noted that rate constraints on the inputs can also be incorporated into the QP problem form in a similar fashion, the minimum and maximum values for  $\vec{x}$  and  $\vec{u}$  can be dependent on  $k$ , and equality constraints can be represented by choosing the element of  $\vec{x}_{min}$  equal to the same element in  $\vec{x}_{max}$  for any particular  $k$ .

---

# Non-LTI Systems and Adaptive Control Theory

## 5.1 Time-Varying Systems Theory

A **time-varying dynamical system** is given generally by

$$\begin{aligned}\dot{\vec{x}}(t) &= f(t, \vec{x}, \vec{u}) \\ \vec{y}(t) &= h(t, \vec{x}, \vec{u})\end{aligned}\tag{5.1}$$

which can also be unforced, i.e.

$$\begin{aligned}\dot{\vec{x}}(t) &= f(t, \vec{x}) \\ \vec{y}(t) &= h(t, \vec{x})\end{aligned}\tag{5.2}$$

Suppose that one initializes at time  $t_0 \geq 0$  the state of a dynamical system as some  $\vec{x}(t_0) = \vec{x}_0 \in \mathbb{R}^n$ . Then, one has an **initial value problem (IVP)**, also known as the **Cauchy problem**, which may have many solutions, one unique solution, or no existing solution. Contrary to LTI dynamical systems, the existence and uniqueness of solutions to IVPs for non-LTI dynamical systems is not always guaranteed and often one can only provide sufficient conditions for existence and uniqueness.

The **Cauchy-Peano Theorem** states that sufficient conditions for the IVP to admit a solution (not necessarily unique) if, for some  $T > 0$ , some  $\epsilon > 0$ , and  $f(t, \vec{x}) : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  is continuous in a closed region, i.e.

$$B = \{(t, \vec{x}) : |t - t_0| \leq T, \|\vec{x} - \vec{x}_0\| \leq \epsilon\} \subseteq \mathbb{R} \times \mathbb{R}^n\tag{5.3}$$

then there exists  $t_0 < t_1 \leq T$  such that the IVP has at least one continuously differentiable solution  $\vec{x}(t)$  on the interval  $[t_0, T]$ . The assumed continuity of  $f(t, \vec{x})$  in its arguments,  $t$  and  $\vec{x}$  ensures that there is at least one solution of the IVP. Note that this theorem does not guarantee the uniqueness of the solution.

The key constraint that yields uniqueness is the **Lipschitz condition** whereby  $f(t, \vec{x})$  satisfies the inequality

$$\|f(t, \vec{x}) - f(t, \vec{y})\| \leq L \|\vec{x} - \vec{y}\|\tag{5.4}$$

for all  $(t, \vec{x})$  and  $(t, \vec{y})$  in some “neighborhood” of  $(t_0, \vec{x}_0)$  with a finite constant  $L > 0$ . A theorem for sufficient conditions for the IVP to admit the *local* existence and uniqueness of a solution states that if, for some  $\epsilon > 0$ ,  $f(t, \vec{x}) : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  is piece-wise continuous in  $t$  and satisfies the Lipschitz condition, i.e.

$$\forall \vec{x}, \vec{y} \in B = \{\vec{x} \in \mathbb{R}^n : \|\vec{x} - \vec{x}_0\| \leq \epsilon\}, \forall t \in [t_0, t_1] \quad (5.5)$$

then, there exists some  $\delta > 0$  such that the IVP for the state equation  $\dot{\vec{x}} = f(t, \vec{x})$  with  $\vec{x}(t_0) = \vec{x}_0$  has a unique solution over  $[t_0, t_0 + \delta]$ . A theorem for sufficient conditions for the IVP to admit the *global* existence and uniqueness of a solution states that if, for some finite  $L > 0$  and  $f(t, \vec{x}) : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  is piece-wise continuous in  $t$  and is globally Lipschitz in  $\vec{x}$ , i.e.

$$\|f(t, \vec{x}) - f(t, \vec{y})\| \leq L\|\vec{x} - \vec{y}\|, \forall \vec{x}, \vec{y} \in \mathbb{R}^n, \forall t \in [t_0, t_1] \quad (5.6)$$

then, the IVP has a unique solution over  $[t_0, t_1]$  where the final time  $t_1$  may be arbitrarily large.

## Equilibrium Points for Time-Varying Systems

For unforced, time-varying dynamics functions  $f : [0, \infty) \times D \rightarrow \mathbb{R}^n$  as piece-wise continuous in  $t$  and locally Lipschitz in  $\vec{x}$  and with domain  $D \subset \mathbb{R}^n$  that contains the origin  $\vec{x} = \vec{0}$ . Then, the origin in  $\mathbb{R}^n$  is an **equilibrium point** for the unforced, time-varying system at  $t_0 = 0$  if

$$f(t, \vec{0}) = \vec{0}, \forall t \geq 0 \quad (5.7)$$

Without loss of generality, this result can be extended beyond the origin and initial time by defining a nonzero vector  $\vec{x} \in \mathbb{R}^n$  to be an equilibrium point of  $\dot{\vec{x}} = f(t, \vec{x})$  at a nonzero initial time  $t = t_0$ , thus

$$f(t, \vec{x}) = \vec{0}, \forall t \geq t_0 \quad (5.8)$$

then defining the new time as  $\tau = t - t_0$  and new state as  $\vec{z}(\tau) = \vec{x}(\tau + t_0) - \vec{x}$ , one has for the new system dynamics

$$\frac{d\vec{z}(\tau)}{d\tau} = \frac{d\vec{x}(\tau + t_0)}{dt} = f(\tau + t_0, \vec{z}(\tau) + \vec{x}) = g(\tau, \vec{z}(\tau)) \quad (5.9)$$

which notably has the property  $g(0, \vec{0}) = f(t_0, \vec{x}) = 0$ , i.e., one can shift the equilibrium point to the origin and initial time to zero.

Furthermore, suppose one has a state trajectory  $\vec{x}(t)$  that starts at  $t = t_0$ , i.e.

$$\frac{d\vec{x}(t)}{dt} = f(t, \vec{x}(t)), \forall t \geq t_0 \quad (5.10)$$

then again defining the new time as  $\tau = t - t_0$  and the new state this time as  $\vec{z}(\tau) = \vec{x}(\tau + t_0) - \vec{x}(\tau + t_0)$ , one has for the new system dynamics

$$\begin{aligned} \frac{d\vec{z}(\tau)}{d\tau} &= \frac{d\vec{x}(\tau + t_0)}{dt} - \frac{d\vec{x}(\tau + t_0)}{dt} \\ &= f(\tau + t_0, \vec{z}(\tau) + \vec{x}(\tau + t_0)) - f(\tau + t_0, \vec{x}(\tau + t_0)) = g(\tau, \vec{z}(\tau)) \end{aligned} \quad (5.11)$$

which notably also has the property  $g(0, \vec{0}) = \vec{0}$ , i.e., one can shift the state trajectory to the origin and initial time to zero. Consequently, these new dynamics around the origin as an equilibrium point while starting at  $t_0$  allows one to determine the original system behavior around the original nonzero equilibrium  $\vec{x}$ , i.e. one can assess the system relative dynamics with respect to any time-dependent trajectory  $\vec{x}(t)$ , starting at an arbitrary initial time  $t_0 \geq 0$ . This is key for stability analysis of equilibrium points.

## Lyapunov Stability Theory for Time-Varying Systems

Lyapunov stability theory provides an analysis tool for assessing the behavior of system trajectories near equilibrium points, but *without* explicit computation of these trajectories. System stability can be interpreted as a continuity of the system trajectories, with respect to initial conditions, over an *infinite* time interval. This infinite time interval highlights the primary notion of stability as a continuity property of Lipschitz-continuous differential equations holding infinitely in time. Formally, this can be stated by letting  $\vec{x}(t, \vec{x}_0)$  define a unique solution of  $\dot{\vec{x}} = f(t, \vec{x})$  with initial condition  $\vec{x}(t_0) = \vec{x}_0$  which exists on a finite, possibly open-ended interval  $[t_0, T]$ . The continuity property of  $\vec{x}(t, \vec{x}_0)$  due to changes in  $\vec{x}_0$  can be stated as follows. Given any constant  $\epsilon > 0$ , there must exist a sufficiently small constant  $\delta > 0$  such that for all perturbed initial conditions  $\vec{x}_0 + \Delta\vec{x}_0$  with  $\|\Delta\vec{x}_0\| \leq \delta$ , the corresponding perturbed solution  $\vec{x}(t, \vec{x}_0 + \Delta\vec{x}_0)$  deviates from the original  $\vec{x}_0$  by no more than  $\epsilon$ , i.e.  $\|\vec{x}(t, \vec{x}_0 + \Delta\vec{x}_0) - \vec{x}(t, \vec{x}_0)\| \leq \epsilon$ , for all  $t_0 \leq t < T$ .

The **Lyapunov stability of an equilibrium point**,  $\vec{x} = \vec{0}$ , i.e. the origin, for time-varying unforced dynamics as *stable* if for any  $\epsilon > 0$  and  $t \geq 0$  there exists some  $\delta(\epsilon, t_0) > 0$  such that for all initial conditions  $\|\vec{x}_0\| < \delta$  and for all  $t \geq t_0 \geq 0$ , the corresponding system trajectories are bounded, i.e.  $\|\vec{x}(t)\| < \epsilon$ , otherwise it is *unstable*. In essence, Lyapunov stability of an equilibrium point  $\vec{x}$  means that given an outer “hyper-sphere”  $B_\epsilon = \{\vec{x} \in \mathbb{R}^n : \|\vec{x}\| \leq \epsilon\}$ , one can find an inner “hyper-sphere”  $B_\delta = \{\vec{x} \in \mathbb{R}^n : \|\vec{x}\| \leq \delta\}$ , such that any trajectory that starts inside  $B_\delta$  will evolve inside  $B_\epsilon$  for *all* future times. Nonlinear dynamical systems may display completely different behavior in various domains of the state. Thus, one must distinguish between local and global stability. In particular, the equilibrium point,  $\vec{x} = \vec{0}$ , i.e. the origin, has **global stability** if it is stable and  $\lim_{\epsilon \rightarrow \infty} \delta(\epsilon, t_0) = \infty$ , i.e. the trajectory will not deviate “too far” from any arbitrary initial conditions. System trajectories of time-varying dynamical systems depend on initial time  $t_0$  and the stability of an equilibrium point for time-varying systems may depend on  $t_0$ . Thus, the equilibrium point has **uniform stability** if it is stable and  $\delta$  does not depend on  $t_0$ .

Beyond these, one can also define the following stricter senses of stability. The equilibrium point,  $\vec{x} = \vec{0}$ , i.e. the origin, has **asymptotic stability** if it is stable and there exists a constant  $c = c(t_0) > 0$  such that  $\vec{x} \rightarrow \vec{0}$  as  $t \rightarrow \infty$  for all  $\|\vec{x}_0\| \leq c$ . The equilibrium point,  $\vec{x} = \vec{0}$ , i.e. the origin, has **uniform asymptotic stability** if it is uniformly stable and there exists a constant  $c > 0$  independent of  $t_0$  such that  $\vec{x} \rightarrow \vec{0}$  as  $t \rightarrow \infty$  for all  $\|\vec{x}_0\| \leq c$  uniformly in  $t_0$ . The equilibrium point,  $\vec{x} = \vec{0}$ , i.e. the origin, has **global uniform asymptotic stability** if it is uniformly asymptotically stable and  $\lim_{\epsilon \rightarrow \infty} \delta(\epsilon) = \infty$ . Of note, uniform asymptotic stability is typically a highly desirable property of control system design as these systems are able to maintain their closed-loop performance in the presence of state perturbations and disturbances.

In his dissertation on the stability of motion, Lyapunov developed two theorems known as the indirect and direct methods for assessing the stability of nominal solutions to dynamical systems governed by a finite number of coupled ODEs. Both methods provide verifiable sufficient conditions for stability of a nominal trajectory without an explicit knowledge of the system solutions. **Lyapunov's indirect method** states that one can determine the stability of an equilibrium point, i.e. the origin, for a nonlinear, time-invariant  $n$ -dimensional systems by linearizing the system dynamics about the equilibrium point. This method has been utilized in previous systems theory discussion in this textbook. Notably, **Lyapunov's direct method** requires the concepts of positive and negative definite functions and the time derivative of a scalar function along the state trajectories of a differential equation, i.e. it's possible solutions.

A scalar function  $V(\vec{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$  of a vector argument  $\vec{x} \in \mathbb{R}^n$  is called **locally positive definite** if  $V(\vec{0}) = 0$  and there exists a constant  $\epsilon > 0$  such that  $V > 0$  for all  $\vec{x} \in \mathbb{R}^n$  in the neighborhood of the origin,

i.e.  $B_\epsilon = \{\vec{x} \in \mathbb{R}^n : \|\vec{x}\| \leq \epsilon\}$ , where if  $\epsilon = \infty$ , then  $V(\vec{x})$  is **globally positive definite**.

A scalar function  $V(\vec{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$  of a vector argument  $\vec{x} \in \mathbb{R}^n$  is called **locally positive semi-definite** if  $V(\vec{0}) = 0$  and there exists a constant  $\epsilon > 0$  such that  $V \geq 0$  for all  $\vec{x} \in \mathbb{R}^n$  in the neighborhood of the origin, i.e.  $B_\epsilon = \{\vec{x} \in \mathbb{R}^n : \|\vec{x}\| \leq \epsilon\}$ , where if  $\epsilon = \infty$ , then  $V(\vec{x})$  is **globally positive semi-definite**.

A scalar function  $V(\vec{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$  of a vector argument  $\vec{x} \in \mathbb{R}^n$  is called **locally negative definite** if  $V(\vec{0}) = 0$  and there exists a constant  $\epsilon > 0$  such that  $V < 0$  for all  $\vec{x} \in \mathbb{R}^n$  in the neighborhood of the origin, i.e.  $B_\epsilon = \{\vec{x} \in \mathbb{R}^n : \|\vec{x}\| \leq \epsilon\}$ , where if  $\epsilon = \infty$ , then  $V(\vec{x})$  is **globally negative definite**.

A scalar function  $V(\vec{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$  of a vector argument  $\vec{x} \in \mathbb{R}^n$  is called **locally negative semi-definite** if  $V(\vec{0}) = 0$  and there exists a constant  $\epsilon > 0$  such that  $V \leq 0$  for all  $\vec{x} \in \mathbb{R}^n$  in the neighborhood of the origin, i.e.  $B_\epsilon = \{\vec{x} \in \mathbb{R}^n : \|\vec{x}\| \leq \epsilon\}$ , where if  $\epsilon = \infty$ , then  $V(\vec{x})$  is **globally negative semi-definite**.

Given a scalar continuously differentiable function  $V(\vec{x})$  where  $\vec{x}(t) \in \mathbb{R}^n$  represents a time-varying trajectory of the time-varying system. Then, the time derivative of  $V(\vec{x}(t))$  along the system solution  $\vec{x}(t) = [x_1(t), \dots, x_n(t)]^T$  as

$$\dot{V}(\vec{x}) = \sum_{i=1}^n \frac{\partial V}{\partial x_i} \dot{x}_i = \sum_{i=1}^n \frac{\partial V}{\partial x_i} f_i(t, \vec{x}) = \nabla V(\vec{x}) f(t, \vec{x}) \quad (5.12)$$

where  $\nabla V(\vec{x}) = [\frac{\partial V}{\partial x_1}, \dots, \frac{\partial V}{\partial x_n}]$  is the row vector gradient of  $V(\vec{x})$  with respect to  $\vec{x}$ . Note that the time derivative of  $V(\vec{x})$  depends not only on the function  $V(\vec{x})$  but also on the system dynamics under consideration. Changing the latter while keeping the same  $V(\vec{x})$  may result in a different  $\dot{V}(\vec{x})$ . Next, to state Lyapunov's direct method, let  $\vec{x} = \vec{0} \in \mathbb{R}^n$  be an equilibrium point for the time-varying dynamics, whose initial conditions are drawn from a domain  $D \subset \mathbb{R}^n$  with  $\vec{x} \in D$  and  $t_0 = 0$ .

Then, if on the domain  $D$ , there exists a continuously differentiable locally positive definite function  $V(\vec{x}) : D \rightarrow \mathbb{R}$ , whose time derivative along the system trajectories is locally negative semi-definite, i.e.

$$\dot{V}(\vec{x}) = \nabla V(\vec{x}) f(t, \vec{x}) \leq 0 \quad (5.13)$$

for all  $t \geq 0$  and for all  $\vec{x} \in D$ , then  $\vec{x} = \vec{0}$  is locally uniformly stable. Furthermore, if  $\dot{V}(\vec{x}) < 0$  for all  $t \geq 0$ , i.e. the time derivative along the system trajectories is locally negative definite, then  $\vec{x} = \vec{0}$  is locally uniformly asymptotically stable. Here any locally positive definite  $V(\vec{x})$  is called a **Lyapunov function candidate** and if it satisfies the time derivative condition it is called a **Lyapunov function**. Note though that the existence of a Lyapunov function is sufficient to claim uniform stability for the equilibrium point, if one cannot be found, nothing can be stated about the stability of the equilibrium point. Furthermore, it should be noted that Lyapunov functions are not unique.

The Lyapunov function can be viewed as an “energy-like” function for testing the stability of a system. If the values of  $V$  do not increase along the system trajectories, then the origin is uniformly stable. If  $V$  strictly decreases, then, in addition, the system trajectories will approach the origin asymptotically. Lastly, note that the uniform asymptotic stability requires a subset of  $D$  known as the **region of attraction**, i.e. starting there the system solutions will converge to the origin. If the region of attraction of a uniformly asymptotically stable equilibrium is  $\mathbb{R}^n$ , then the equilibrium is said to be globally uniformly asymptotically stable.

Next, define the property  $\lim_{\|\vec{x}\| \rightarrow \infty} V(\vec{x}) = \infty$  where  $V(\vec{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$  as a **radially unbounded** Lyapunov function candidate. In addition, define  $V_c = \{\vec{x} \in \mathbb{R}^n : V(\vec{x}) = c\}$  as a **level set**, i.e.  $V_c$  has a constant value  $c$ , of a radially unbounded Lyapunov function candidate  $V(\vec{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ , and define  $\Omega_c = \{\vec{x} \in \mathbb{R}^n : V(\vec{x}) \leq c\}$  as the union of the interior set of  $V_c$  and  $V_c$  itself.

Then, consider a converging sequence  $\lim_{k \rightarrow \infty} \vec{x}_k = \vec{a}$  with all  $\vec{x}$  from  $\Omega_c$ , then the limit point  $\vec{a}$  must also be in  $\Omega_c$ . Here,  $\Omega_c$  is a **closed set** because  $V(\vec{x})$  is continuous on  $\mathbb{R}^n$  and  $V(\vec{x}) \leq c$  for all  $k = 1, 2, \dots$ , thus  $c \geq \lim_{k \rightarrow \infty} V(\vec{x}_k) = V(\vec{a})$  and  $\vec{a} \in \Omega_c$  which shows every converging sequence in  $\Omega_c$  has its limit point in the same set, i.e. closed.  $\Omega_c$  is also a **bounded set**. If it was not, then there must exist a sequence of points  $\{\vec{x}_k\} \in \Omega_c$  whose limit is  $\infty$ . However, since  $V(\vec{x})$  is continuous and radially unbounded, then  $c \geq \lim_{k \rightarrow \infty} V(\vec{x}_k) = \infty$ , which is a contradiction. Thus, since  $\Omega_c$  is closed, bounded, and belongs to  $\mathbb{R}^n$ , it is a **compact set** which allows one to state the following **Krasovskii-LaSalle theorem**. If  $\vec{x} = \vec{0}$  is an equilibrium point of  $\dot{\vec{x}} = f(t, \vec{x})$  and  $V(\vec{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$  is a radially unbounded Lyapunov function, then  $\vec{x}$  is a globally uniformly asymptotically stable equilibrium point. Note that a simple example of radially unbounded Lyapunov function candidates include the quadratic form  $V(\vec{x}) = \vec{x}^T P \vec{x}$  where  $P$  is a symmetric positive definite matrix, i.e.  $P = P^T > 0$ .

### Linear, Time-Varying Systems

A linear, time-varying (LTV) systems are represented by the following linear state-space model

$$\begin{aligned}\dot{\vec{x}}(t) &= A(t)\vec{x}(t) + B(t)\vec{u}(t) \\ \vec{y}(t) &= C(t)\vec{x}(t) + D(t)\vec{u}(t)\end{aligned}\tag{5.14}$$

Similar to LTI systems, the general solution of a LTV system can be written as

$$\vec{x}(t) = \Phi(t, t_0)\vec{x}(t_0) + \int_{t_0}^t \Phi(t, \tau)B(\tau)\vec{u}(\tau)d\tau\tag{5.15}$$

where  $\Phi(t, t_0)$  is the **state-transition matrix** of LTV system and can be given by the **Peano-Baker series** form

$$\Phi(t, t_0) = I + \int_{t_0}^t A(\sigma_1)d\sigma_1 + \int_{t_0}^t A(\sigma_1) \int_{t_0}^{\sigma_1} A(\sigma_2)d\sigma_2 d\sigma_1 + \dots\tag{5.16}$$

which converges uniformly and absolutely to a solution that exists and is unique. The  $\vec{\beta}(t)$  typically also has restricted admissible trajectories for any particular realization.

For Lyapunov stability of LPV systems, one can use the Lyapunov candidate function with  $P > 0$

$$V(\vec{x}) = \vec{x}^T P \vec{x}\tag{5.17}$$

which has time derivative

$$\dot{V}(\vec{x}) = \dot{\vec{x}}^T P \vec{x} + \vec{x}^T P \dot{\vec{x}}\tag{5.18}$$

and evaluated along the state trajectories, one has

$$\dot{V}(\vec{x}) = \vec{x}^T \left( A^T(t)P + PA(t) \right) \vec{x}\tag{5.19}$$

which is negative definite, if and only if

$$A^T(t)P + PA(t) < 0 \quad \forall t\tag{5.20}$$

Define the **linear, time-varying controllability Gramian** using the state-transition matrix as

$$W_C(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t_0, t) B(t) B(t)^T \Phi(t_0, t)^T dt \quad (5.21)$$

where  $W_C(t_0, t_1)$  is symmetric, positive semi-definite, and satisfies

$$W_C(t_0, t_1) = W_C(t_0, t) + \Phi(t_0, t) W(t, t_1) \Phi(t_0, t)^T \quad (5.22)$$

If there exists some  $t_1$  such that  $W_C(t_0, t_1)$  is full rank, then the LTV system is state controllable.

Define the **linear, time-varying observability Gramian** using the state-transition matrix as

$$W_O(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t, t_0)^T C(t)^T C(t) \Phi(t, t_0) dt \quad (5.23)$$

where  $W_O(t_0, t_1)$  is symmetric, positive semi-definite, and satisfies

$$W_O(t_0, t_1) = W_O(t_0, t) + \Phi(t, t_0)^T W(t, t_1) \Phi(t, t_0) \quad (5.24)$$

If there exists some  $t_1$  such that  $W_O(t_0, t_1)$  is full rank, then the LTV system is observable.

## References

For more information, please refer to the following

•

## 5.2 Gain-Scheduled Adaptive Control

As mentioned previously, a common control design approach for nonlinear systems using linear control theory is known as **gain-scheduled adaptive control** which can be divided into two different approaches, divide-and-conquer and linear, parameter-varying (LPV). This section will introduce these design approaches for a nonlinear state-space system

$$\begin{aligned} \dot{\vec{x}} &= f(\vec{x}, \vec{u}) = A\vec{x} + B\vec{u} + \tilde{f}(\vec{\beta}) \\ \vec{y} &= h(\vec{x}, \vec{u}) = C\vec{x} + D\vec{u} + \tilde{h}(\vec{\beta}) \end{aligned} \quad (5.25)$$

where the scheduling parameter,  $\vec{\beta}(\vec{x}, \vec{u})$ , denotes the nonlinear dependence of the dynamics on the state and input with  $\nabla_{\vec{x}} \vec{\rho}$  and  $\nabla_{\vec{u}} \vec{\rho}$  constant. Notably, setting  $\vec{\rho} = [\vec{x}^T \vec{u}^T]^T$  provides all states and inputs to be included. Associated with this nonlinear system and an equilibrium trajectory  $(\vec{x}_e, \vec{y}_e, \vec{u}_e)$ , one can obtain an **equilibrium linearization**

$$\begin{aligned} \Delta \dot{\vec{x}}(t) &= \nabla_{\vec{x}} f(\vec{x}_e, \vec{u}_e) \Delta \vec{x}(t) + \nabla_{\vec{u}} f(\vec{x}_e, \vec{u}_e) \Delta \vec{u}(t) \\ \Delta \vec{y}(t) &= \nabla_{\vec{x}} h(\vec{x}_e, \vec{u}_e) \Delta \vec{x}(t) + \nabla_{\vec{u}} h(\vec{x}_e, \vec{u}_e) \Delta \vec{u}(t) \end{aligned} \quad (5.26)$$

with

$$\begin{aligned}\Delta \vec{x} &= \vec{x} - \bar{\vec{x}} \\ \Delta \vec{y} &= \vec{y} - \bar{\vec{y}} \\ \Delta \vec{u} &= \vec{u} - \bar{\vec{u}}\end{aligned}\tag{5.27}$$

Alternatively, one can differentiate the nonlinear system to obtain

$$\begin{aligned}\dot{\vec{x}} &= \vec{z} \\ \dot{\vec{z}} &= (A + \nabla_{\vec{\beta}} \tilde{f}(\vec{\beta})) \vec{z} + (B + \nabla_{\vec{\beta}} \tilde{f}(\vec{\beta})) \dot{\vec{u}} \\ \dot{\vec{y}} &= (C + \nabla_{\vec{\beta}} \tilde{h}(\vec{\beta})) \vec{z} + (D + \nabla_{\vec{\beta}} \tilde{h}(\vec{\beta})) \dot{\vec{u}}\end{aligned}\tag{5.28}$$

which provides a **velocity-based linearization** for a local operating point  $\bar{\vec{\rho}}$  as the “frozen” form

$$\begin{aligned}\dot{\vec{x}} &= \vec{z} \\ \dot{\vec{z}} &= (A + \nabla_{\vec{\beta}} \tilde{f}(\bar{\vec{\beta}})) \vec{z} + (B + \nabla_{\vec{\beta}} \tilde{f}(\bar{\vec{\beta}})) \dot{\vec{u}} \\ \dot{\vec{y}} &= (C + \nabla_{\vec{\beta}} \tilde{h}(\bar{\vec{\beta}})) \vec{z} + (D + \nabla_{\vec{\beta}} \tilde{h}(\bar{\vec{\beta}})) \dot{\vec{u}}\end{aligned}\tag{5.29}$$

which remains a local approximation about some  $\bar{\vec{\beta}}$ . However, this family of linearizations captures the entire dynamics of the nonlinear system and its solutions can be pieced together to approximate the nonlinear dynamics to an arbitrary degree of accuracy.

When equilibrium linearization produces an LTI system, necessary and sufficient conditions for stability are well-known. However, when equilibrium linearization produces an LTV system, frozen-time theory is widely used to establish stability conditions. Specifically, the linearized time-varying system is guaranteed provided the time variation of  $\nabla_{\vec{x}} f(\vec{x}, \vec{u})(t)$  is sufficiently slow, e.g., for a sufficiently small  $\epsilon$ , one must have

$$\sup_{t \geq 0} \left\| \frac{d}{dt} \nabla_{\vec{x}} f(\vec{x}, \vec{u}) \right\| < \epsilon\tag{5.30}$$

Furthermore, it can be shown that the LTV system inherits the worst-case stability robustness of the family of frozen-time LTV systems provided the rate of variation is sufficiently slow. It is important to note that frozen-time theory tends to be conservative as it establishes a sufficient condition for stability.

Frozen-input theory allows one to show the nonlinear system is locally bounded-input, bounded-output (BIBO) stable in the vicinity near equilibrium or everywhere provided the members of the family of equilibrium linearizations or velocity-based linearizations are uniformly stable and the rate of variation is sufficiently slow. This rate of variation requirement typically imposes a certain class of initial conditions and forcing inputs with a certain rate of variation itself. Importantly, if the rate of variation is sufficiently slow, the nonlinear system also inherits the stability robustness of the family of linearizations to smooth, finite-dimensional, nonlinear perturbations. Another important technical requirement of stability analysis for equilibrium linearization is that equilibrium points are smoothly parameterized by the control input. Another technical requirement of stability analysis for velocity-based linearization is that unboundedness of the state,  $\vec{x}$ , must imply the state rate,  $\vec{z}$  is unbounded assuming the input  $\vec{u}$  is bounded. Notably, the velocity-based result involves no restriction to near-equilibrium operation and for nonlinear systems where the slow variation is automatically satisfied.

## Divide-and-Conquer Gain-Scheduled Control Design

In **divide-and-conquer gain-scheduled control**, one utilizes linearization to obtain a family of linear systems parameterized by  $\vec{\beta}$ . For the classical case, one utilizes the equilibrium linearization about a set of equilibrium points while off-equilibrium gain-scheduling can be achieved via velocity-based linearization.

The design procedure of classical gain-scheduling control design can be outlined as follows:

1. Choose the equilibrium points parameterized by a set of  $\vec{\beta}$
2. Determine the family of equilibrium linearizations based on the set of  $\vec{\beta}$
3. Synthesize a family of LTI controllers for each equilibrium linearization system with constant  $\vec{\beta}$  and some input vector to the controller, e.g., may be  $\vec{e} = \vec{y} - \bar{\vec{y}}$ 
  - For smooth gain-scheduling, compatible structure of LTI controllers must be chosen for smooth interpolation
  - Practical interpolation is linear interpolation of matrix elements or gain, poles, and zeros of controller transfer functions
  - Number of members of family increased until satisfactory obtained for interpolated points
4. Implement a nonlinear controller based on the family of LTI controllers using one of three approaches
  - Classical
    - Implement controller input and output transformations between nonlinear system and linear controller
    - Substitute  $\vec{\beta}$  for  $\vec{\beta}$  in the family of local linear controllers, may not be continuous function of  $\vec{\beta}$
  - Local linear equivalence
    - Nonlinear controller selected with linearization at each equilibrium point matching the relevant member of the family of linear controllers with agreement between equilibrium inputs and outputs
    - Existence conditions typically exist for controllers which contain integral action
  - Extended local linear equivalence
    - Chooses nonlinear controller satisfying local linear equivalence and maximizing neighborhood of accurate controller dynamics

The design procedure of velocity-based gain-scheduling control design can be outlined as follows:

1. Choose a set of operation points parameterized by a set of  $\vec{\beta}$
2. Determine the family of velocity-based linearizations based on a set of  $\vec{\beta}$
3. Synthesize a family of velocity-based linear controllers
  - Performed with LTI control for each  $\vec{\beta}$
  - Performed with LPV methods below
4. Realize a nonlinear controller with the velocity-based linearization family
  - Owing to differentiation and integration operations, the order of the velocity-based controller may be greater than a direct representation

## LPV Gain-Scheduling Control Design

In **linear, parameter-varying (LPV)** gain-scheduling, one utilizes the following LPV state-space model

$$\begin{aligned}\dot{\vec{x}}(t) &= A(\vec{\beta}(t))\vec{x}(t) + B(\vec{\beta}(t))\vec{u}(t) \\ \vec{y}(t) &= C(\vec{\beta}(t))\vec{x}(t) + D(\vec{\beta}(t))\vec{u}(t)\end{aligned}\quad (5.31)$$

where  $\vec{\beta}(t)$  is the **parameter vector** that must be a continuously differentiable function of time and the state-space matrices are also continuous functions of  $\vec{\beta}$ .

This is typically at least restricted by some upper and lower bounds, e.g.  $\vec{\beta}_U$  and  $\vec{\beta}_L$ , respectively, and may also be rate restricted by some upper and lower bounds, e.g.  $\dot{\vec{\beta}}_U$  and  $\dot{\vec{\beta}}_L$ . Thus, one can write the set of admissible trajectories for  $\vec{\beta}$ ,  $\mathcal{B}$ , as

$$\mathcal{B} = \left\{ \vec{\beta}(t) : \mathbb{R}_+ \rightarrow \mathbb{R}^{n_\beta} \text{ such that: } \vec{\beta}(t) \text{ continuously differentiable,} \right. \\ \left. \vec{\beta}_L \leq \vec{\beta}(t) \leq \vec{\beta}_U \forall t \geq 0, \quad \dot{\vec{\beta}}_L \leq \dot{\vec{\beta}}(t) \leq \dot{\vec{\beta}}_U \forall t \geq 0 \right\} \quad (5.32)$$

However, more advanced approaches for LPV systems have been proposed based on the specific structure of the state-space matrices, e.g., linear or rational dependence on  $\vec{\beta}$ . For arbitrary dependence, one can generally specify the performance of an LPV system,  $G_\beta$  in terms of its maximum possible induced  $\mathcal{L}_{2 \leftarrow 2}$  gain from input  $\vec{u}$  to output  $\vec{y}$ , i.e.

$$\|G_\beta\|_{2 \leftarrow 2} = \max_{\vec{\beta} \in \mathcal{B}, 0 \neq \|\vec{u}\|_2 \leq \infty, \vec{x}(0)=0} \frac{\|\vec{y}\|_2}{\|\vec{u}\|_2} \quad (5.33)$$

As  $G_\beta$  is time-varying, one *cannot* interpret this as an  $\mathcal{H}_\infty$ -norm in the frequency domain. However, one can use a **generalized Bounded Real Lemma** to derive sufficient conditions for an upper bound on  $\|G_\beta\|_{2 \leftarrow 2}$ .

In particular, a rate-unbounded LPV system is globally uniformly asymptotically stable and  $\|G_\beta\|_{2 \leftarrow 2} \leq \gamma$  if there exists  $P > 0$  such that  $\forall \vec{\beta} \in [\vec{\beta}_L, \vec{\beta}_U]$ , one has

$$\begin{bmatrix} A^T(\vec{\beta})P + PA(\vec{\beta}) & PB(\vec{\beta}) \\ B^T(\vec{\beta})P & -I \end{bmatrix} + \frac{1}{\gamma^2} \begin{bmatrix} C^T(\vec{\beta}) \\ D^T(\vec{\beta}) \end{bmatrix} [C(\vec{\beta}) \quad D(\vec{\beta})] < 0 \quad (5.34)$$

where the exponential stability notion by noting the definition is held by the upper left block of the first term. To also show that  $\|G_\beta\|_{2 \leftarrow 2} \leq \gamma$  note that one can left and right multiply by the stacked state and input vectors,  $[\vec{x}^T \quad \vec{u}^T]^T$ , to obtain

$$\vec{x}^T P \vec{x} + \vec{x}^T P \dot{\vec{x}} + \frac{1}{\gamma^2} \vec{y}^T \vec{y} - \vec{u}^T \vec{u} \leq 0 \quad (5.35)$$

$$\dot{V} + \frac{1}{\gamma^2} \vec{y}^T \vec{y} - \vec{u}^T \vec{u} \leq 0 \quad (5.36)$$

and integrating from  $t = 0$  to  $t = t_f$ , one has

$$V(\vec{x}(t_f)) - V(\vec{x}(0)) + \frac{1}{\gamma^2} \int_0^{t_f} \vec{y}^T \vec{y} dt - \int_0^{t_f} \vec{u}^T \vec{u} dt \leq 0 \quad (5.37)$$

where  $V(\vec{x}(t_f)) > 0$  and  $V(\vec{x}(0)) = 0$ , thus

$$\int_0^{t_f} \vec{y}^T \vec{y} dt \leq \gamma^2 \int_0^{t_f} \vec{u}^T \vec{u} dt \quad (5.38)$$

Then, as  $t_f \rightarrow \infty$ , one has

$$\|\vec{y}\|_2^2 \leq \gamma^2 \|\vec{u}\|_2^2 \quad (5.39)$$

$$\|G_\beta\|_{2 \leftarrow 2} \leq \gamma \quad (5.40)$$

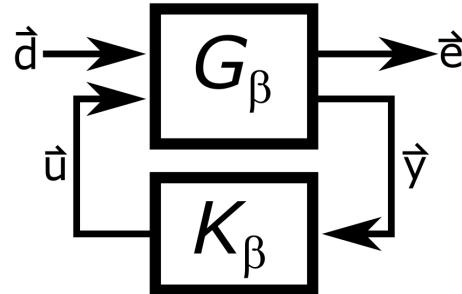
The matrix inequality is a parameterized LMI condition, i.e. one LMI for each value of  $\vec{\beta} \in \mathcal{B}$ . In practice, this infinite collection of LMI conditions is approximated by enforcing them only on a finite grid of points. Then, the finite dimensional LMI conditions can be directly obtained to bound  $\|G_\beta\|_{2 \leftarrow 2}$  without approximation if the state matrices have a rational dependence on  $\vec{\beta}$ .

For rate-bounded LPV systems, one can alter the matrix inequality with a slight variation on the Lyapunov theory, which requires one to check  $2^{n_\beta}$  LMI conditions evaluated at the endpoints defined by the hypercube of the elements of  $\vec{\beta}_U$  and  $\vec{\beta}_L$ . Furthermore, in this case, one also needs to search over the infinitely dimensional space of functions,  $P(\beta)$ , which must be restricted to a finite dimensional subspace. In practice, one specifies a collection of scalar basis functions,  $g_i(\vec{\beta})$ , and use a linear combination of these basis functions

$$P(\vec{\beta}) = \sum_{i=1}^N g_i(\vec{\beta}) P_i \quad (5.41)$$

where  $P_i$  are symmetric matrices that form the finite collection of decision variables. Often polynomials are chosen as basis functions.

Consider the following LPV feedback control system



with LPV plant

$$\begin{aligned} \dot{\vec{x}}(t) &= A(\vec{\beta}) \vec{x}(t) + [B_1(\vec{\beta}) \quad B_2(\vec{\beta})] \begin{bmatrix} \vec{d}(t) \\ \vec{u}(t) \end{bmatrix} \\ \begin{bmatrix} \vec{e}(t) \\ \vec{y}(t) \end{bmatrix} &= \begin{bmatrix} C_1(\vec{\beta}) \\ I \end{bmatrix} \vec{x}(t) + \begin{bmatrix} 0 & D_{12}(\vec{\beta}) \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \vec{d}(t) \\ \vec{u}(t) \end{bmatrix} \end{aligned} \quad (5.42)$$

While optimal LPV controllers can also be designed for output feedback and for observer feedback, consider the state feedback LPV controller

$$\vec{u} = D_K(\vec{\beta}) \vec{x} \quad (5.43)$$

This results in a closed-loop LPV system

$$\begin{aligned}\dot{\vec{x}}(t) &= \left( A(\vec{\beta}) + B_2(\vec{\beta})D_K(\vec{\beta}) \right) \vec{x}(t) + B_1(\vec{\beta})\vec{d}(t) \\ \vec{e}(t) &= \left( C_1(\vec{\beta}) + D_{12}(\vec{\beta})D_K(\vec{\beta}) \right) \vec{x}(t)\end{aligned}\quad (5.44)$$

Then, by the generalized bounded real lemma, the closed-loop LPV system is internally stable and  $\|G_\beta\|_{2 \leftarrow 2} \leq \gamma$  if there exists  $P > 0$  such that  $\forall \vec{\beta} \in [\vec{\beta}_L, \vec{\beta}_U]$ , one has

$$\begin{bmatrix} A_L^T(\vec{\beta})P + PA_L(\vec{\beta}) & PB_L(\vec{\beta}) \\ B_L^T(\vec{\beta})P & -I \end{bmatrix} + \frac{1}{\gamma^2} \begin{bmatrix} C_L^T(\vec{\beta}) \\ D_L^T(\vec{\beta}) \end{bmatrix} [C_L(\vec{\beta}) \quad D_L(\vec{\beta})] < 0 \quad (5.45)$$

which can be written as the following using a change of variables

$$Q = P^{-1} \quad (5.46)$$

and

$$R(\vec{\beta}) = D_K(\vec{\beta})Q \quad (5.47)$$

and the Schur Complement Lemma to get

$$\begin{bmatrix} QA^T(\vec{\beta}) + A(\vec{\beta})Q + R^T(\vec{\beta})B_2^T(\vec{\beta}) + B_2(\vec{\beta})R(\vec{\beta}) & B_1(\vec{\beta}) & (C_1(\vec{\beta})Q + D_{12}(\vec{\beta})R(\vec{\beta}))^T \\ B_1^T(\vec{\beta}) & \gamma^{-2}I & 0 \\ C_1(\vec{\beta})Q + D_{12}(\vec{\beta})R(\vec{\beta}) & 0 & -I \end{bmatrix} < 0 \quad (5.48)$$

Thus, one can state the state feedback LPV OCP as an SDP in  $(Q, R, \gamma)$ , i.e.

$$\begin{aligned}(Q, R(\vec{\beta}), \gamma)^{opt} &= \underset{Q \in \mathbb{S}^{n_x}, R \in \mathbb{R}^{n_u \times n_x}, \gamma > 0}{\operatorname{argmin}} \gamma \\ \text{subject to: } &\begin{bmatrix} QA^T(\vec{\beta}) + A(\vec{\beta})Q + R^T(\vec{\beta})B_2^T(\vec{\beta}) + B_2(\vec{\beta})R(\vec{\beta}) & B_1(\vec{\beta}) & (C_1(\vec{\beta})Q + D_{12}(\vec{\beta})R(\vec{\beta}))^T \\ B_1^T(\vec{\beta}) & -\gamma^{-2}I & 0 \\ C_1(\vec{\beta})Q + D_{12}(\vec{\beta})R(\vec{\beta}) & 0 & -I \end{bmatrix} < 0 \\ Q > 0\end{aligned}\quad (5.49)$$

with the state feedback LPV optimal controller as  $\vec{u}(t) = D_K(\vec{\beta})\vec{x}(t)$  reconstructed with

$$D_K(\vec{\beta}) = R(\vec{\beta})Q^{-1} \quad (5.50)$$

which must be approximately solved by gridding over  $[\vec{\beta}_L, \vec{\beta}_U]$ .

## References

For more information, please refer to the following

- Leith, D. J. and Leithead, W.E. (2000) "Survey of Gain-Scheduling Analysis & Design," *International Journal of Control*, 73(11), pp. 1001–1025

## 5.3 Linear-in-Control Systems and Dynamic Inversion

### Linear-In-Control Dynamical System

Furthermore, flight vehicles are not inherently LTI systems and therefore one often linearizes the nonlinear, time-invariant dynamics over a grid of flight conditions in order to design the MIMO LTI state feedback controllers at those conditions and interpolate control gains, an adaptive technique known as gain-scheduling. This technique necessitates smooth transitions between flight conditions such that the linearized model is a good approximation to the dynamics and robustness to the neglected higher-order terms (HOT) in the dynamics about those flight conditions. However, during some flight maneuvers, one may not be able to sufficiently neglect the nonlinear dynamics for the flight vehicles. An alternative approach to gain-scheduling adaptive control is to use **dynamic inversion (DI) control** which transforms a nonlinear, time-invariant dynamical system into an LTI system via inversion which can then be controlled using a MIMO LTI controller.

Dynamic inversion control will assume a square system, i.e.  $n_u = n_y$ ,  $\vec{x}$  is available for feedback, then a **linear-in-control dynamical system**, also known as a **control-affine dynamical system**

$$\begin{aligned}\dot{\vec{x}} &= f(\vec{x}) + g(\vec{x})\vec{u} \\ \vec{y} &= h(\vec{x})\end{aligned}\tag{5.51}$$

which is suitable for many flight vehicle dynamics, and that one desires to control the output  $\vec{y}(t)$  to track a commanded reference signal,  $\vec{r}(t)$ . Thus,  $\vec{y}(t)$  is also known as the **controlled variable** in this context and the tracking error,  $\vec{e}$ , is defined as

$$\vec{e}(t) = \vec{r}(t) - \vec{y}(t)\tag{5.52}$$

This type of dynamic inversion differentiates the controlled variable,  $\vec{y}(t)$ , until the control,  $\vec{u}(t)$ , appears in the expression for the derivative. Thus, this type of dynamic inversion control is also known as **input-output feedback linearization control** and will be presented here first for linear, time-invariant systems as an introduction to the concept and linear-in-control systems afterwards where it shows its true usefulness.

Lastly, it is important to note that this method relies heavily on the differentiation of the dynamical system which may not be suitably robust to model uncertainties. To overcome this result, nonlinear dynamic inversion is often used alongside adaptive control methods for linear-in-control systems to account for **matched uncertainties** in the dynamics model, i.e. those that can be canceled by choosing some  $\vec{u}(t)$ , as opposed to **unmatched uncertainties**, i.e. those that can't be canceled.

The following subsections will introduce these basic control designs which can be extended to be robust to unmatched uncertainties. An exhaustive discussion of these control methods including stability and robustness is beyond the scope of this textbook and are typically addressed in nonlinear systems and/or adaptive control courses.

### LTI Dynamic Inversion Control

As an introduction to dynamic inversion control, consider the LTI system

$$\begin{aligned}\dot{\vec{x}} &= A\vec{x} + B\vec{u} \\ \vec{y} &= C\vec{x}\end{aligned}\tag{5.53}$$

where the derivative of  $\vec{y}$  can be shown to be

$$\dot{\vec{y}} = C \dot{\vec{x}} = CA \vec{x} + CB \vec{u} \quad (5.54)$$

Then, one can define the **inner feedback linearization loop** as the control law

$$\vec{u} = (CB)^{-1} (-CA \vec{x} + \dot{\vec{r}} + \vec{v}) \quad (5.55)$$

where  $\vec{v}$  is the **virtual control input**.

Substituting this control law for  $\dot{\vec{y}}$ , one has

$$\dot{\vec{y}} = CA \vec{x} + CB (CB)^{-1} (-CA \vec{x} + \dot{\vec{r}} + \vec{v}) \quad (5.56)$$

$$\dot{\vec{y}} = CA \vec{x} - CA \vec{x} + \dot{\vec{r}} + \vec{v} \quad (5.57)$$

$$\vec{v} = \dot{\vec{y}} - \dot{\vec{r}} \quad (5.58)$$

or as the **error dynamics**

$$\dot{\vec{e}} = -\vec{v} \quad (5.59)$$

which has  $n_y$  poles at  $s = 0$ . Thus, selecting  $\vec{u}$  in this way allowed one to cancel the  $CA \vec{x}$  term and relate it to the tracking error without the  $CB$  term, making the system from  $\vec{v}$  to  $\vec{y}$  appear like a linear system with poles at the origin.

Then, one can design the **outer tracking loop** as a state feedback control law,  $K$ , using MIMO LTI control on the tracking error,  $\vec{e}$ , i.e.

$$\vec{v} = K \vec{e} \quad (5.60)$$

which provides error dynamics

$$\dot{\vec{e}} = -K \vec{e} \quad (5.61)$$

and is stable if and only if  $K > 0$ , i.e. positive definite, and is typically a diagonal matrix to keep the control channels in the outer-loop decoupled. Thus, the overall **LTI dynamic inversion (LDI) controller** is given by

$$\vec{u} = (CB)^{-1} (-CA \vec{x} + \dot{\vec{r}} + K \vec{e}) \quad (5.62)$$

which sums the feedback linearization loop output, the LTI control gain on the tracking error, and a feedforward term on the reference derivative  $\dot{\vec{r}}$ .

Furthermore, substituting the LDI controller into the state equation, one has closed-loop system dynamics as

$$\begin{aligned} \dot{\vec{x}} &= A \vec{x} + B (CB)^{-1} (-CA \vec{x} + \dot{\vec{r}} + \vec{v}) \\ \dot{\vec{x}} &= \left( I - B (CB)^{-1} \right) A \vec{x} + B (CB)^{-1} (\dot{\vec{r}} + \vec{v}) \end{aligned} \quad (5.63)$$

To analyze the stability of this system, one must not only understand the error dynamics, i.e.  $\dot{\vec{e}}$ , but also the **zero dynamics**, i.e.  $\vec{y}(t) = 0$  or  $\vec{v} = -\dot{\vec{r}}$  such that

$$\dot{\vec{x}} = \left( I - B (CB)^{-1} \right) A \vec{x} = A_z \vec{x} \quad (5.64)$$

Note that the dimension of the  $\vec{e}$  is  $n_y < n_x$ , thus, the remaining  $n_x - n_y$  system poles must be LHP stable for the dynamic inversion controller to be stable. These poles are unobservable by selecting the controlled variable  $\vec{y} = C\vec{x}$  as the term  $I - B(CB)^{-1}$  can be shown as the projection on the null space of  $C$  along the range of  $B$ . Thus,  $(I - B(CB)^{-1})A = A_z$  describes the dynamics in the null space of  $C$  and in range perpendicular of  $B$ . This demonstrates that designing  $C$ , i.e. the controlled variables of the state, is crucial to the stability of LTI dynamic inversion control.

Lastly, it should be noted that this design assumed  $CB$  was invertible and non-zero. However, one can note that the  $n^{\text{th}}$  derivative of  $\vec{y}$  as

$$\vec{y}^{[n]} = CA^{n+1}\vec{x} + C \begin{bmatrix} A^n B & \cdots & AB & B \end{bmatrix} \begin{bmatrix} \vec{u} \\ \vdots \\ \vec{u}^{[n-1]} \\ \vec{u}^{[n]} \end{bmatrix} \quad (5.65)$$

Thus, if  $CB$  is zero, one can simply take the smallest  $n^{\text{th}}$  derivative which has a non-zero  $CA^nB$  term, i.e.

$$\vec{y}^{[n]} = CA^{n+1}\vec{x} + CA^nB\vec{u} \quad (5.66)$$

where one can form the dynamic inversion controller as

$$\vec{u} = \left( CA^{n-1}B \right)^{-1} \left( -CA^n\vec{x} + \vec{r}^{[n]} + \vec{v} \right) \quad (5.67)$$

which provides

$$\vec{e}^{[n]} = -\vec{v} \quad (5.68)$$

which has  $n_y \times n$  poles at the origin.

However, if  $CB$  is non-zero and singular, the system must be controllable and taking  $n = n_x$ , one can form the controllability matrix as

$$C = \begin{bmatrix} A^{n_x} B & \cdots & AB & B \end{bmatrix} \quad (5.69)$$

which has rank  $n_x$ . However, the  $C$  matrix has rank  $n_u$ , thus one cannot use a strict inverse transformation. However, defining the **pseudoinverse** of  $CC$  as

$$(CC)^+ = (CC)^T \left( (CC)(CC)^T \right)^{-1} \quad (5.70)$$

one can form the dynamic inversion controller as

$$\begin{bmatrix} \vec{u} \\ \vdots \\ \vec{u}^{[n_x-1]} \\ \vec{u}^{[n_x]} \end{bmatrix} = (CC)^+ \left( -CA^{n_x}\vec{x} + \vec{r}^{[n_x]} + \vec{v} \right) \quad (5.71)$$

which provides

$$\vec{e}^{[n_x]} = -\vec{v} \quad (5.72)$$

which has  $n_y \times n_x$  poles at the origin due to the pseudoinverse.

In both cases, one can design the **outer tracking loop** as a state feedback control law,  $K$ , using MIMO LTI control on the tracking error,  $\vec{e}$ , i.e.

$$\vec{v} = K_{n-1} \vec{e}^{[n-1]} + \cdots + K_0 \vec{e} \quad (5.73)$$

which provides error dynamics

$$\vec{e}^{[n]} + K_{n-1} \vec{e}^{[n-1]} + \cdots + K_0 \vec{e} = 0 \quad (5.74)$$

where the gain matrices,  $K_i$  for  $i = 0, \dots, n_x$  are selected to make the error dynamics stable. Notably, these cases require one to compute the tracking error and its  $n$  derivatives. Furthermore, for  $CB$  as non-zero and singular, one must also use a dynamical system to extract  $\vec{u}(t)$  from its derivative vector  $[\vec{u} \ \dots \ \vec{u}^{[n_x-1]} \ \vec{u}^{[n_x]}]^T$ .

### Nonlinear Dynamic Inversion Control

Next, consider the linear-in-control, time-invariant system

$$\begin{aligned} \dot{\vec{x}} &= f(\vec{x}) + g(\vec{x}) \vec{u} \\ \vec{y} &= h(\vec{x}) \end{aligned} \quad (5.75)$$

where the derivative of  $\vec{y}$  can be shown to be

$$\dot{\vec{y}} = \frac{\partial h}{\partial \vec{x}} \dot{\vec{x}} = \frac{\partial h}{\partial \vec{x}} f(\vec{x}) + \frac{\partial h}{\partial \vec{x}} g(\vec{x}) \vec{u} \quad (5.76)$$

where  $\frac{\partial h}{\partial \vec{x}} f(\vec{x})$  is known as the **Lie derivative** of  $h(\vec{x})$  along  $f(\vec{x})$ . Then, one can define the inner **feedback linearization loop**

$$\vec{u} = \left( \frac{\partial h}{\partial \vec{x}} g(\vec{x}) \right)^{-1} \left( -\frac{\partial h}{\partial \vec{x}} f(\vec{x}) + \dot{\vec{r}} + \vec{v} \right) \quad (5.77)$$

with **virtual control input**,  $\vec{v}(t)$ .

Substituting this control law for  $\dot{\vec{y}}$ , one obtains the **error dynamics**

$$\dot{\vec{e}} = -\vec{v} \quad (5.78)$$

Then, one can design the **outer tracking loop** as a state feedback control law,  $K$ , using MIMO LTI control on the tracking error,  $\vec{e}$ , i.e.

$$\vec{v} = K \vec{e} \quad (5.79)$$

which provides error dynamics

$$\dot{\vec{e}} = -K \vec{e} \quad (5.80)$$

and is stable if and only if  $K > 0$ , i.e. positive definite, and is typically a diagonal matrix to keep the control channels in the outer-loop decoupled.

Thus, the overall **nonlinear dynamic inversion (NDI) controller** is given by

$$\vec{u} = \left( \frac{\partial h}{\partial \vec{x}} g(\vec{x}) \right)^{-1} \left( -\frac{\partial h}{\partial \vec{x}} f(\vec{x}) + \dot{\vec{r}} + K \vec{e} \right) \quad (5.81)$$

which sums the feedback linearization loop output, the LTI control gain on the tracking error, and a feedforward term on the reference derivative  $\dot{\vec{r}}$ . Notably  $\vec{u}$  requires one to know the nonlinear dynamics or to be able to use lookup tables to compute  $\frac{\partial h}{\partial \vec{x}} f(\vec{x})$  and  $\frac{\partial h}{\partial \vec{x}} g(\vec{x})$ . Furthermore, similar to the case where  $CB$  is singular or zero for LDI, if  $\frac{\partial h}{\partial \vec{x}} g(\vec{x})$  is singular, one must use successive Lie derivatives to form suitable feedback linearization controllers.

Furthermore, substituting the NDI controller into the state equation, one has closed-loop system dynamics as

$$\begin{aligned} \dot{\vec{x}} &= f(\vec{x}) + g(\vec{x}) \left( \frac{\partial h}{\partial \vec{x}} g(\vec{x}) \right)^{-1} \left( -\frac{\partial h}{\partial \vec{x}} f(\vec{x}) + \dot{\vec{r}} + K \vec{e} \right) \\ \dot{\vec{x}} &= (I - g(\vec{x})) \left( \frac{\partial h}{\partial \vec{x}} g(\vec{x}) \right)^{-1} \left( -\frac{\partial h}{\partial \vec{x}} \right) f(\vec{x}) + g(\vec{x}) \left( \frac{\partial h}{\partial \vec{x}} g(\vec{x}) \right)^{-1} (\dot{\vec{r}} + \vec{v}) \end{aligned} \quad (5.82)$$

with **zero dynamics**, i.e.  $\vec{y}(t) = 0$  or  $\vec{v} = -\dot{\vec{r}}$  such that

$$\dot{\vec{x}} = (I - g(\vec{x})) \left( \frac{\partial h}{\partial \vec{x}} g(\vec{x}) \right)^{-1} \left( -\frac{\partial h}{\partial \vec{x}} \right) f(\vec{x}) \quad (5.83)$$

which can be linearized at specific flight conditions to check the suitability of the controlled variable,  $\vec{y} = h(\vec{x})$ , or can be simulated at different initial conditions and checked.

### Incremental Nonlinear Dynamic Inversion

To improve the robustness to model uncertainty in the NDI control strategy, one can alternatively design an **incremental nonlinear dynamic inversion (INDI) controller** instead of a full NDI which calculates the required *change* to the control input as opposed to the full control input. Thus, a full system model is not required, only a local part of the model, making INDI more robust.

### Adaptive Control with Nonlinear Dynamic Inversion

Another control strategy to improve the performance of NDI to model uncertainty is to use **NDI-Based adaptive control**. Here, the reference signal,  $\vec{r}$ , and its derivative,  $\dot{\vec{r}}$ , are defined via a specified dynamical system called the **reference model** which is modeled to track the plant state, i.e. regulate the tracking error  $\vec{e} = \vec{r} - \vec{x}$  to zero, and the control system incorporates an additional adaptive feedback loop to account for the matched uncertainties in the dynamics.

As an example of NDI-based adaptive control, consider a linear-in-control dynamical system with additive uncertainty  $\Delta(\vec{x})$ , one has

$$\dot{\vec{x}} = f(\vec{x}) + g(\vec{x}) \vec{u} + \Delta(\vec{x}, \vec{u}) \quad (5.84)$$

where one forms the NDI-based adaptive controller as

$$\vec{u} = g(\vec{x})^{-1} \left( -f(\vec{x}) + \dot{\vec{r}} + K \vec{e} - \hat{\Delta}(\vec{x}, \vec{u}, \vec{\theta}) \right) \quad (5.85)$$

where  $\hat{\Delta}$  is the new adaptive control signal that depends on  $\vec{x}$ ,  $\vec{u}$ , and internal design parameters,  $\vec{\theta}$ , and is nominally used to cancel the uncertainty  $\Delta(\vec{x})$ . Thus, in this case, the error dynamics for the model reference are

$$\dot{\vec{e}} = \dot{\vec{r}} - \dot{\vec{x}} \quad (5.86)$$

$$\dot{\vec{e}} = \dot{\vec{r}} - f(\vec{x}) - g(\vec{x})\vec{u} - \Delta(\vec{x}, \vec{u}) \quad (5.87)$$

$$\dot{\vec{e}} = \dot{\vec{r}} - f(\vec{x}) - \left( -f(\vec{x}) + \dot{\vec{r}} + K\vec{e} - \hat{\Delta}(\vec{x}, \vec{u}, \vec{\theta}) \right) - \Delta(\vec{x}) \quad (5.88)$$

$$\dot{\vec{e}} = -K\vec{e} + \hat{\Delta}(\vec{x}, \vec{u}, \vec{\theta}) - \Delta(\vec{x}, \vec{u}) \quad (5.89)$$

where one ideally desires to design the adaptive element  $\hat{\Delta}(\vec{x}, \vec{u}, \vec{\theta})$  to cancel the uncertainty,  $\Delta(\vec{x})$ , and design  $K$  to stabilize the  $n_x$  system poles. However, the stability and performance of such a system is dependent on the nature of  $\Delta(\vec{x}, \vec{u})$  and the state trajectory  $\vec{x}(t)$  where often  $\hat{\Delta}(\vec{x}, \vec{u}, \vec{\theta})$  may not even converge to  $\Delta(\vec{x}, \vec{u})$ , but the system is still stable. Furthermore, adaptive neural networks have been used to perform the uncertainty cancellation term  $\hat{\Delta}(\vec{x}, \vec{u})$  and adaptive control can be performed with and without NDI. In the following sections, adaptive control will be discussed further.

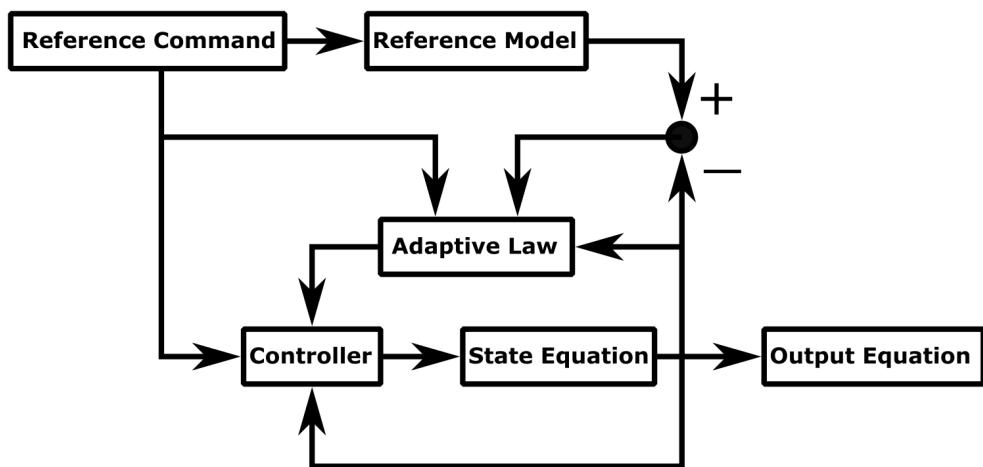
## References

For more information, please refer to the following

- 

## 5.4 Model Reference Adaptive Control

A generic block diagram for a system operating with Model Reference Adaptive Control (MRAC) is



In essence, an MRAC system consists of a controller whose gains are updated online using an adaptive law. The adaptive law is a function of the system output, the reference command, and the tracking error,

i.e. the difference between the system output and the reference model output which specifies the desired trajectories for the system to follow. Finally, the controller computes its commands based on the reference command, the system output, and the online adjusted parameters from the adaptive law. Per design, the adaptive controller forces the system output to follow the desired reference commands while operating in the presence of system parameter uncertainties. The main objective of the controller is to maintain consistent performance of the closed-loop system in the presence of uncertainties and unknown variations in the system parameters. When the true system parameters are unknown, one may attempt to estimate control gains online using available measurements. This approach is referred to as **direct adaptive control**. Alternatively, the gains can be estimated online by solving system design equations that relate the uncertain parameters to the measured signals in the system. This is referred to as **indirect adaptive control**. MRAC systems can be designed using either approach as well as combining both approaches as **hybrid adaptive control**.

MRAC considers the model-following command tracking algorithm for continuous-time dynamical systems with constant vector of uncertain parameters,  $\vec{\theta}$ , bounded environmental disturbances,  $\vec{\eta}$ , i.e.

$$\begin{aligned}\dot{\vec{x}} &= f(t, \vec{x}, \vec{u}, \vec{\theta}, \vec{\eta}) \\ \vec{y} &= h(t, \vec{x}, \vec{u}, \vec{\theta}, \vec{\eta})\end{aligned}\tag{5.90}$$

where  $\vec{x}$  is available for full state feedback control, i.e.,  $\vec{y}(t) = \vec{x}(t)$ . Then, the command tracking control problem involves designing the control input  $\vec{u}$  so that the regulated output  $\vec{y}(t)$  tracks a given bounded reference signal  $\vec{r}(t) \in \mathbb{R}^n$  in the presence of system uncertainties,  $\vec{\theta}$  and environmental disturbances,  $\vec{\eta}(t)$ . Specifically, one must find  $\vec{u}(t)$  such that the command tracking error

$$\vec{e} = \vec{y}(t) - \vec{r}(t)\tag{5.91}$$

becomes sufficiently small as  $t \rightarrow \infty$  and all the signals in the corresponding closed-loop system remain uniformly bounded in time. Note that if one could achieve  $\vec{e}(t) \rightarrow 0$  as  $t \rightarrow \infty$ , then asymptotic command tracking would be achieved. However, this may not be feasible, in which case, the control objective would be to achieve uniform ultimate boundedness of the command tracking error, i.e.

$$\|\vec{e}\| \leq \epsilon, \quad \forall t \geq T\tag{5.92}$$

where  $\epsilon > 0$  is the desired tracking tolerance and  $T$  is some finite time.

### Direct MRAC

This subsection will consider the extension of the direct MRAC design to a **linear-in-control dynamical system**

$$\dot{\vec{x}} = A\vec{x} + B\Lambda(\vec{u} + f(\vec{x}))\tag{5.93}$$

where  $\vec{x} \in \mathbb{R}^n$  is the system state,  $\vec{u} \in \mathbb{R}^p$  is the control input,  $B \in \mathbb{R}^{n \times p}$  is the known control matrix,  $A \in \mathbb{R}^{n \times n}$  is the unknown state matrix, and  $\Lambda \in \mathbb{R}^{p \times p}$  is the unknown diagonal matrix of control uncertainties with diagonal elements  $\lambda_i > 0$ . This uncertainty can be due to modeling errors or model control failures. It is assumed that the pair  $(A, B\Lambda)$  is controllable. In addition, the unknown nonlinear vector function  $f(\vec{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^n$  represents the system matched uncertainty where each individual component  $f_i(\vec{x})$  of

$f(\vec{x})$  can be written as a linear combination of  $N$  known locally Lipschitz-continuous basis function  $\phi(\vec{x})$  with unknown constant coefficients, i.e.

$$f(\vec{x}) = \vec{\theta}^T \Phi(\vec{x}) \quad (5.94)$$

where  $\vec{\theta} \in \mathbb{R}^N$  is constant unknown matrix and  $\Phi(\vec{x}) = [\phi_1(\vec{x}) \dots \phi_N(\vec{x})]^T$  is the known regressor vector.

A direct MIMO state feedback MRAC is required to force the system state  $\vec{x}$  to globally uniformly asymptotically track the reference model state  $\vec{x}_{ref}$  given by the **reference model**

$$\dot{\vec{x}}_{ref} = A_{ref} \vec{x}_{ref} + B_{ref} \vec{r}(t) \quad (5.95)$$

where  $A_{ref} \in \mathbb{R}^n$  is chosen such that all eigenvalues are in the LHP,  $B_{ref} \in \mathbb{R}^{n \times m}$  is the reference input matrix, and  $\vec{r}(t) \in \mathbb{R}^P$  is the external, bounded reference command. One also requires that during tracking in an MRAC system all signals in the closed-loop system remain uniformly bounded. Note that the reference command has the same dimension as the plant control input. Thus, given any bounded command  $\vec{r}(t)$ , the control input  $\vec{u}(t)$  needs to be chosen such that the state tracking error,  $\vec{e}(t) = \vec{x}(t) - \vec{x}_{ref}(t)$ , globally uniformly asymptotically tends to zero, i.e.

$$\lim_{t \rightarrow \infty} \|\vec{x}(t) - \vec{x}_{ref}\| = 0 \quad (5.96)$$

First, note that if matrices  $A$  and  $\Lambda$  were known, one can apply the ideal fixed-gain control law as

$$\vec{u} = K_x^T \vec{x} + K_r^T \vec{r} - \vec{\theta}^T \Phi(\vec{x}) \quad (5.97)$$

and obtain the closed-loop system

$$\dot{\vec{x}} = (A + B\Lambda K_x^T) \vec{x} + B\Lambda K_r^T \vec{r} \quad (5.98)$$

Comparing this with the desired reference dynamics, for the existence of a controller of the ideal fixed-gain form, the ideal unknown control gains,  $K_x$  and  $K_r$ , must satisfy the **matching conditions**

$$\begin{aligned} A + B\Lambda K_x^T &= A_{ref} \\ B\Lambda K_r^T &= B_{ref} \end{aligned} \quad (5.99)$$

Assuming these hold, by inspection, the ideal fixed-gain control law will result in a closed-loop system which is exactly the same as the reference model. Consequently, for any bounded reference signal input,  $\vec{r}(t)$ , the fixed-gain controller provides global uniform asymptotic tracking performance. It is important to note that given a general  $A$ ,  $B$ ,  $\Lambda$ ,  $A_{ref}$  and  $B_{ref}$ , there is no guarantee that the ideal MIMO gains  $K_x$  and  $K_r$  exist such that the matching conditions are satisfied and the chosen adaptive control law may not be able to meet the design objective. However, in practice, the structure of  $A$  is known and the reference model matrices  $A_{ref}$  and  $B_{ref}$  are chosen so that the system has at least one ideal solution for  $K_x$  and  $K_r$ .

Assuming  $K_x$  and  $K_r$  do exist, consider the following control law based on the ideal fixed-gain control law, i.e.

$$\vec{u} = \hat{K}_x^T \vec{x} + \hat{K}_r^T \vec{r} - \hat{\theta}^T \Phi(\vec{x}) \quad (5.100)$$

where  $\hat{K}_x \in \mathbb{R}^{n \times p}$ ,  $\hat{K}_r \in \mathbb{R}^{p \times P}$ , and  $\hat{\theta} \in \mathbb{R}^{N \times n}$  are the estimates of the ideal unknown matrices  $K_x$ ,  $K_r$  and  $\vec{\theta}$ , respectively, based on Lyapunov stability analysis. Substituting this control law into the plant dynamics results in the closed-loop system dynamics as

$$\dot{\vec{x}} = (A + B\Lambda \hat{K}_x^T) \vec{x} + B\Lambda \left( \hat{K}_r^T \vec{r} - (\hat{\theta} - \vec{\theta})^T \Phi(\vec{x}) \right) \quad (5.101)$$

and subtracting the reference model from these closed-loop system dynamics, one has the closed-loop tracking error dynamics as

$$\dot{\vec{e}} = (A + B\Lambda\hat{K}_x^T)\vec{x} + B\Lambda\left(\hat{K}_r^T\vec{r} - (\hat{\theta} - \vec{\theta})^T\Phi(\vec{x})\right) - A_{ref}\vec{x}_{ref} - B_{ref}\vec{r} \quad (5.102)$$

Including the matching conditions, one has

$$\begin{aligned} \dot{\vec{e}} &= (A_{ref} + B\Lambda(\hat{K}_x - K_x))\vec{x} - A_{ref}\vec{x}_{ref} + B\Lambda(\hat{K}_r - K_r)\vec{r} - B\Lambda(\hat{\theta} - \vec{\theta})^T\Phi(\vec{x}) \\ \dot{\vec{e}} &= A_{ref}\vec{e} + B\Lambda\left((\hat{K}_x - K_x)^T\vec{x} + (\hat{K}_r - K_r)^T\vec{r} - (\hat{\theta} - \vec{\theta})^T\Phi(\vec{x})\right) \end{aligned} \quad (5.103)$$

where defining  $\Delta K_x = \hat{K}_x$ ,  $\Delta K_r = \hat{K}_r - K_r$ , and  $\Delta \vec{\theta} = \hat{\theta} - \vec{\theta}$  as the parameter estimation errors, one has

$$\dot{\vec{e}} = A_{ref}\vec{e} + B\Lambda\left(\Delta K_x^T\vec{x} + \Delta K_r^T\vec{r} - \Delta \vec{\theta}^T\Phi(\vec{x})\right) \quad (5.104)$$

Next, one can define the rates of adaptation as  $\Gamma_x = \Gamma_x^T > 0$ ,  $\Gamma_r = \Gamma_r^T > 0$ , and  $\Gamma_\theta = \Gamma_\theta^T > 0$ . Then, consider a globally radially unbounded quadratic Lyapunov function candidate in the form

$$V(\vec{e}, \Delta K_x, \Delta K_r, \Delta \vec{\theta}) = \vec{e}^T P \vec{e} + \text{Tr}\left(\left(\Delta K_x^T \Gamma_x^{-1} \Delta K_x + \Delta K_r^T \Gamma_r^{-1} \Delta K_r + \Delta \vec{\theta}^T \Gamma_\theta^{-1} \Delta \vec{\theta}\right) \Lambda\right) \quad (5.105)$$

where  $P = P^T > 0$  satisfies the **algebraic Lyapunov equation**

$$PA_{ref} + A_{ref}^T P = -Q \quad (5.106)$$

for some  $Q = Q^T > 0$ . Then, the time derivative of  $V$  evaluated along the state trajectories can be written as

$$\dot{V} = \dot{\vec{e}}^T P \vec{e} + \vec{e}^T P \dot{\vec{e}} + 2\text{Tr}\left(\left(\Delta K_x^T \Gamma_x^{-1} \dot{\hat{K}}_x + \Delta K_r^T \Gamma_r^{-1} \dot{\hat{K}}_r + \Delta \vec{\theta}^T \Gamma_\theta^{-1} \dot{\hat{\theta}}\right) \Lambda\right) \quad (5.107)$$

$$\begin{aligned} \dot{V} &= \left(A_{ref}\vec{e} + B\Lambda(\Delta K_x^T\vec{x} + \Delta K_r^T\vec{r} - \Delta \vec{\theta}^T\Phi(\vec{x}))\right)^T P \vec{e} \\ &\quad + \vec{e}^T P \left(A_{ref}\vec{e} + B\Lambda(\Delta K_x^T\vec{x} + \Delta K_r^T\vec{r} - \Delta \vec{\theta}^T\Phi(\vec{x}))\right) \\ &\quad + 2\text{Tr}\left(\left(\Delta K_x^T \Gamma_x^{-1} \dot{\hat{K}}_x + \Delta K_r^T \Gamma_r^{-1} \dot{\hat{K}}_r + \Delta \vec{\theta}^T \Gamma_\theta^{-1} \dot{\hat{\theta}}\right) \Lambda\right) \end{aligned} \quad (5.108)$$

$$\begin{aligned} \dot{V} &= \vec{e}^T (A_{ref}P + PA_{ref})\vec{e} + 2\vec{e}^T P B \Lambda (\Delta K_x^T \vec{x} + \Delta K_r^T \vec{r} - \Delta \vec{\theta}^T \Phi(\vec{x})) \\ &\quad + 2\text{Tr}\left(\left(\Delta K_x^T \Gamma_x^{-1} \dot{\hat{K}}_x + \Delta K_r^T \Gamma_r^{-1} \dot{\hat{K}}_r + \Delta \vec{\theta}^T \Gamma_\theta^{-1} \dot{\hat{\theta}}\right) \Lambda\right) \end{aligned} \quad (5.109)$$

$$\begin{aligned} \dot{V} &= -\vec{e}^T Q \vec{e} + \left(2\vec{e}^T P B \Lambda \Delta K_x^T \vec{x} + 2\text{Tr}\left(\Delta K_x^T \Gamma_x^{-1} \dot{\hat{K}}_x \Lambda\right)\right) \\ &\quad + \left(2\vec{e}^T P B \Lambda \Delta K_r^T \vec{r} + 2\text{Tr}\left(\Delta K_r^T \Gamma_r^{-1} \dot{\hat{K}}_r \Lambda\right)\right) \\ &\quad + \left(-2\vec{e}^T P B \Lambda \Delta \vec{\theta}^T \Phi(\vec{x}) + 2\text{Tr}\left(\Delta \vec{\theta}^T \Gamma_\theta^{-1} \dot{\hat{\theta}} \Lambda\right)\right) \end{aligned} \quad (5.110)$$

Next, one can use the vector trace identity  $\vec{a}^T \vec{b} = \text{Tr}(\vec{b} \vec{a}^T)$  to form

$$\begin{aligned} \dot{V} &= -\vec{e}^T Q \vec{e} + 2\text{Tr}\left(\Delta K_x^T \left(\Gamma_x^{-1} \dot{\hat{K}}_x + \vec{x} \vec{e}^T P B\right) \Lambda\right) \\ &\quad + 2\text{Tr}\left(\Delta K_r^T \left(\Gamma_r^{-1} \dot{\hat{K}}_r + \vec{r} \vec{e}^T P B\right) \Lambda\right) + 2\text{Tr}\left(\Delta \vec{\theta}^T \left(\Gamma_\theta^{-1} \dot{\hat{\theta}} - \Phi(\vec{x}) \vec{e}^T P B\right) \Lambda\right) \end{aligned} \quad (5.111)$$

Thus, if the **Direct MIMO adaptive laws** are chosen as

$$\begin{aligned}\dot{\hat{K}}_x &= -\Gamma_x \vec{x} \vec{e}^T P B \\ \dot{\hat{K}}_r &= -\Gamma_r \vec{r}(t) \vec{e}^T P B \\ \dot{\hat{\theta}} &= \Gamma_\theta \Phi(\vec{x}) \vec{e}^T P B\end{aligned}\tag{5.112}$$

Then,

$$\dot{V} = -\vec{e}^T Q \vec{e} \leq 0\tag{5.113}$$

Therefore, the closed-loop error dynamics are uniformly stable. So, the tracking error  $\vec{e}(t)$  and the parameter estimation errors  $\Delta K_x(t)$ ,  $\Delta K_r(t)$ , and  $\Delta \theta(t)$  are uniformly bounded and so are the parameter estimates  $\hat{K}_x(t)$ ,  $\hat{K}_r(t)$ , and  $\hat{\theta}(t)$ . Since  $\vec{r}(t)$  is bounded and  $A_{ref}$  has all LHP eigenvalues, then  $\vec{x}_{ref}$  and  $\dot{\vec{x}}_{ref}$  are bounded. Hence, the system state  $\vec{x}(t)$  is uniformly bounded, the control input  $\vec{u}(t)$  is bounded, and  $\vec{x}(t)$  is bounded, and thus,  $\vec{e}(t)$  is bounded. Furthermore, the second time derivative of  $V(t)$

$$\ddot{V} = -2\vec{e}^T Q \dot{\vec{e}}\tag{5.114}$$

is bounded, and so  $\dot{V}(t)$  is uniformly continuous. Since, in addition,  $V(t)$  is lower bounded and  $\dot{V}(t) \leq 0$ , then using Barbalat's lemma  $\lim_{t \rightarrow \infty} \dot{V}(t) = 0$ . Thus, it has been formally proven that the state tracking error  $\vec{e}(t)$  tends to the origin globally, uniformly, and asymptotically, i.e.

$$\lim_{t \rightarrow \infty} \|\vec{x}(t) - \vec{x}_{ref}(t)\| = 0\tag{5.115}$$

Lastly, note that the tuning knobs in the MIMO case are  $\Gamma_x$ ,  $\Gamma_r$ , and  $\Gamma_\theta$  as before, but also  $Q$  for the algebraic Lyapunov equation, all of which must be symmetric, positive definite matrices.

## $\mathcal{L}_1$ Adaptive Control

### References

For more information, please refer to the following

-

## **Part II**

# **Aerospace Vehicle Dynamics and Control Systems**

---

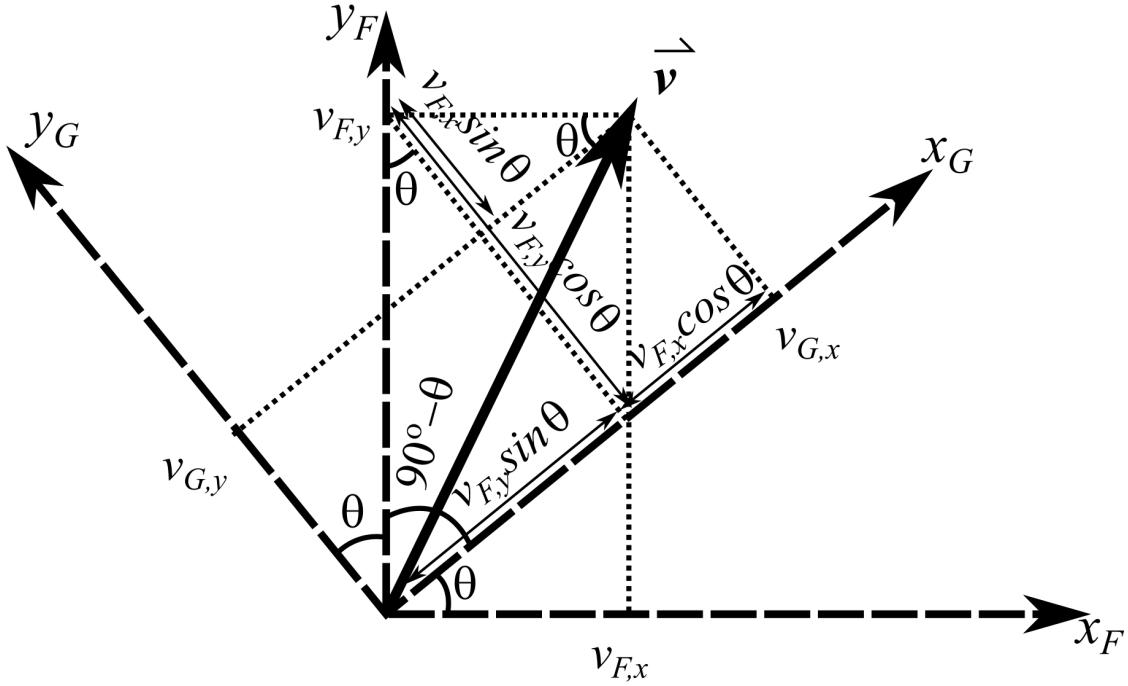
# Body Dynamics

## 6.1 Reference Frames and Transformations

In order to derive the dynamical system models for aerospace vehicles, one uses various coordinate systems centered at different origins, known as **reference frames**. A three-dimensional reference frame  $F$  is uniquely described by its origin point  $O_F$  and three Cartesian coordinate axes  $x_F - y_F - z_F$ . Furthermore, when discussing vehicle dynamics, one must be able to transform vector quantities, e.g., position, velocity, acceleration, forces, moments, between different reference frames. This section will discuss the matrix/vector transformations for representing two- and three-dimensional vectors referenced to the same origin, also known as **reference frame rotation**, different origins, but with the same axes, also known as **reference frame translation**, and different origins and different axes. Notably, reference frame translation and rotation are related, but different concepts than the active translation and rotation of vectors.

### Two-Dimensional Reference Frame Rotation

For two-dimensional rotations, consider the vector  $\vec{v}$  in frames  $F$  and  $G$ , expressed as  $\vec{v}_F = [v_{F,x} \ v_{F,y}]^T$  and  $\vec{v}_G = [v_{G,x} \ v_{G,y}]^T$  with the same origin.



By trigonometry,

$$\vec{v}_G = \begin{bmatrix} v_{G,x} \\ v_{G,y} \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} v_{F,x} \\ v_{F,y} \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \vec{v}_F \quad (6.1)$$

or

$$\vec{v}_G = \begin{bmatrix} v_{G,x} \\ v_{G,y} \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} v_{F,x} \\ v_{F,y} \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \vec{v}_F \quad (6.2)$$

which can be defined as

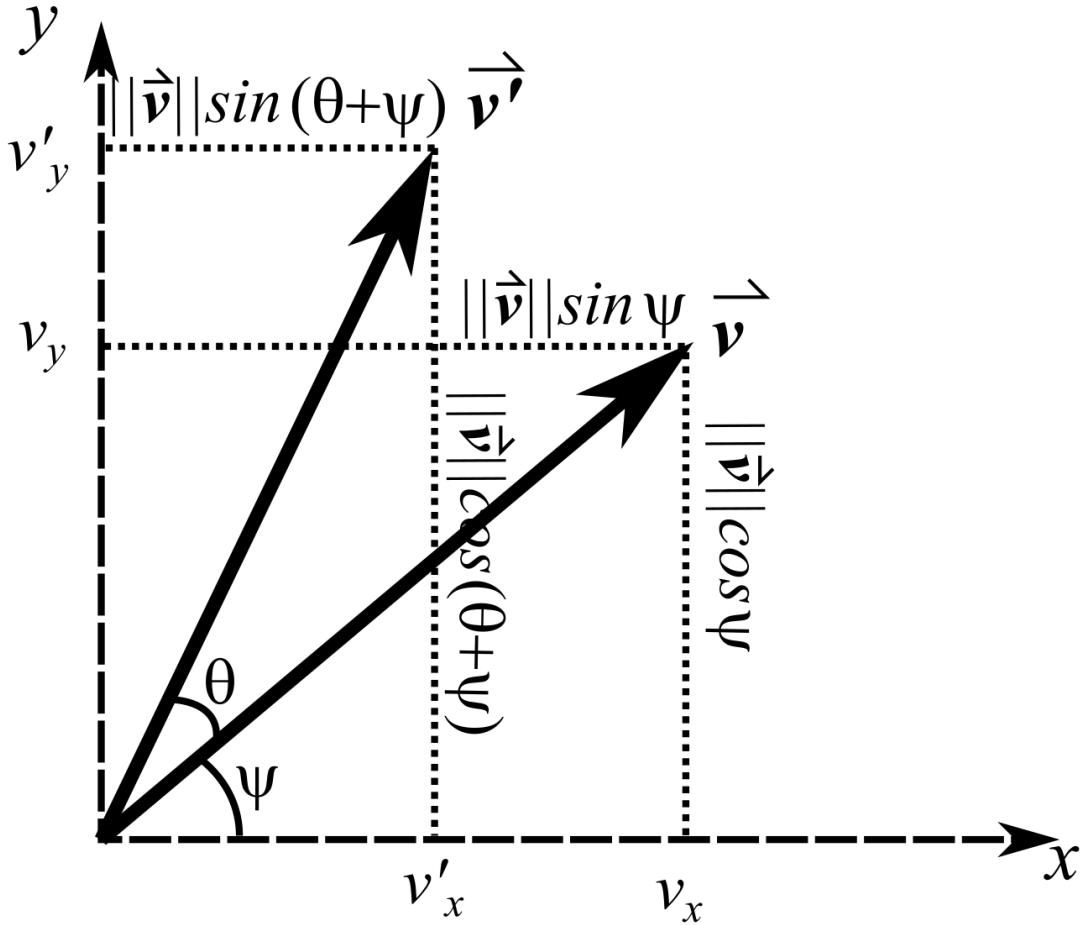
$$\vec{v}_G = C_{G \leftarrow F}(\theta) \vec{v}_F \quad (6.3)$$

where  $C_{G \leftarrow F}(\theta)$  is the **passive rotation matrix** from frame  $F$  to  $G$  by the angle  $\theta$ . It is also known as the **two-dimensional direction cosine matrix (DCM)** originates from the identity,  $\sin \theta = \cos(90^\circ - \theta)$  and  $-\sin \theta = \cos(90^\circ + \theta)$ , where one can rewrite

$$C_{G \leftarrow F} = \begin{bmatrix} \cos \theta & \cos(90^\circ - \theta) \\ \cos(\theta - 90^\circ) & \cos \theta \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \quad (6.4)$$

where each element,  $C_{i,j}$ , corresponds to the cosine of the angle between the  $i$ -axis of the  $G$  frame and the  $j$ -axis of the  $F$  frame. This can also be seen by inspection of the previous figure.

The DCM is opposed to the **active rotation matrix** of the vector  $\vec{v}$  by an angle  $\theta$  to obtain a new vector  $\vec{v}'$ .



In this case, an active 2D rotation would be given by

$$\vec{v}' = \begin{bmatrix} \|\vec{v}\| \cos(\psi + \theta) \\ \|\vec{v}\| \sin(\psi + \theta) \end{bmatrix} \quad (6.5)$$

By the trigonometric addition rules, one has

$$\vec{v}' = \begin{bmatrix} \|\vec{v}\| \cos \psi \cos \theta - \|\vec{v}\| \sin \psi \sin \theta \\ \|\vec{v}\| \sin \psi \cos \theta - \|\vec{v}\| \cos \psi \sin \theta \end{bmatrix} \quad (6.6)$$

and by definition of the components of  $\vec{v}$ , one has

$$\vec{v}' = \begin{bmatrix} v_x \cos \theta - v_y \sin \theta \\ v_x \sin \theta + v_y \cos \theta \end{bmatrix} \quad (6.7)$$

which implies the active 2D rotation matrix definition

$$\vec{v}' = R(\theta) \vec{v} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \vec{v} = C^T(\theta) \vec{v} \quad (6.8)$$

so it is important to note the context of “rotation matrix” when reading sources on rotational kinematics.

To transform back, one can form the inverse DCM by using the negative angle, i.e.

$$\vec{v}_F = \begin{bmatrix} \cos(-\theta) & \sin(-\theta) \\ -\sin(-\theta) & \cos(-\theta) \end{bmatrix} \vec{v}_G = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \vec{v}_G \quad (6.9)$$

or

$$\vec{v}_F = C_{F \leftarrow G}(-\theta) \vec{p}_G \quad (6.10)$$

which by inspection can also be seen to be

$$\vec{v}_F = C_{G \leftarrow F}^T(\theta) \vec{p}_G \quad (6.11)$$

which implies the property

$$C_{G \leftarrow F}^{-1}(\theta) = C_{G \leftarrow F}^T(\theta) \quad (6.12)$$

for direction cosine matrices. Thus, it is important to reference the angle between two frames, typically as  $\theta_{G \leftarrow F}$  as the rotation angle of frame  $G$  with respect to frame  $F$  and  $\theta_{F \leftarrow G}$  as the angle of frame  $F$  with respect to frame  $G$  where  $\theta_{F \leftarrow G} = -\theta_{G \leftarrow F}$ .

Suppose frame  $G$  is rotating in time. Then, the time derivative of the two-dimensional DCM is given by

$$\dot{C}_{G \leftarrow F}(t) = \frac{d}{dt} \begin{pmatrix} \cos \theta_{G \leftarrow F}(t) & \sin \theta_{G \leftarrow F}(t) \\ -\sin \theta_{G \leftarrow F}(t) & \cos \theta_{G \leftarrow F}(t) \end{pmatrix} = \begin{bmatrix} -\sin \theta_{G \leftarrow F}(t) \dot{\theta}_{G \leftarrow F}(t) & \cos \theta_{G \leftarrow F}(t) \dot{\theta}_{G \leftarrow F}(t) \\ -\cos \theta_{G \leftarrow F}(t) \dot{\theta}_{G \leftarrow F}(t) & -\sin \theta_{G \leftarrow F}(t) \dot{\theta}_{G \leftarrow F}(t) \end{bmatrix} \quad (6.13)$$

and defining the quantity  $\omega_{G \leftarrow F}(t) = \dot{\theta}_{G \leftarrow F}(t)$  as the **two-dimensional angular velocity** of frame  $G$  with respect to  $F$ , one has

$$\dot{C}_{G \leftarrow F} = \begin{bmatrix} 0 & -\omega_{G \leftarrow F}(t) \\ \omega_{G \leftarrow F}(t) & 0 \end{bmatrix} \begin{bmatrix} \cos \theta(t) & \sin \theta(t) \\ -\sin \theta(t) & \cos \theta(t) \end{bmatrix} = \begin{bmatrix} \cos \theta(t) & \sin \theta(t) \\ -\sin \theta(t) & \cos \theta(t) \end{bmatrix} \begin{bmatrix} 0 & \omega_{G \leftarrow F}(t) \\ -\omega_{G \leftarrow F}(t) & 0 \end{bmatrix} \quad (6.14)$$

Notably, the angular velocity also be differentiated again to obtain  $\alpha_{G \leftarrow F}(t) = \dot{\omega}_{G \leftarrow F}(t) = \ddot{\theta}_{G \leftarrow F}(t)$ .

### Three-Dimensional Rotation

For three-dimensional rotations, one has several different methods for specifying the rotation between two different reference frames. One direct representation is the **axis-angle** representation defined by **Euler's rotation theorem** states that any rotating frame possesses an instantaneous axis of rotation. Thus, the first representation specifies the instantaneous axis of rotation, also known as the **Euler axis**, as  $\vec{e} = [\cos(e_x) \cos(e_y) \cos(e_z)]^T$ , of unit length as well as the **rotation angle**,  $\theta_{G \leftarrow F}$ , about that axis from frame  $F$  to frame  $G$ . Note that in two-dimensions, the Euler axis was the normal to the two-dimensional plane, i.e., a hypothetical “z-axis” where  $\vec{e} = \vec{e}_z$ , and the angle was  $\theta_{G \leftarrow F}$ . However, in three-dimensions, the Euler axis is arbitrary and can be expressed in any reference frame, e.g.,  $H$ , though typically frame  $H$  will be either frame  $F$  or frame  $G$ . The Euler axis and rotation angle can be combined into the **Euler vector**, also known as the **passive rotation vector**, **orientation vector**, or **attitude vector**, denoted as  $\vec{\theta}_{G \leftarrow F, H} = \theta_{G \leftarrow F} \vec{e}_H$  and described as “the Euler vector of frame  $G$  relative to frame  $F$  expressed in frame  $H$ .”

Furthermore, as the Euler vector quantity may change in time, the **three-dimensional angular velocity** of frame  $G$  relative to frame  $F$  and expressed in frame  $H$  is defined as

$$\vec{\omega}_{G \leftarrow F, H} = \dot{\vec{\theta}}_{G \leftarrow F, H} = \omega_{G \leftarrow F} \vec{e}_H \quad (6.15)$$

Likewise, the **three-dimensional angular acceleration** of frame  $G$  relative to frame  $F$  and expressed in frame  $H$  is defined as

$$\vec{\alpha}_{G \leftarrow F, H} = \ddot{\vec{\theta}}_{G \leftarrow F, H} = \alpha_{G \leftarrow F} \vec{e}_H + +\omega_{G \leftarrow F} \dot{\vec{e}}_H \quad (6.16)$$

Similar to the two-dimensional DCM, another representation is to use the **three-dimensional direction cosine matrix (DCM)**,  $C_{G \leftarrow F}$ , i.e.,

$$\vec{v}_G = C_{G \leftarrow F} \vec{v}_F \quad (6.17)$$

which is often denoted component-wise as

$$C_{G \leftarrow F} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \quad (6.18)$$

where each row corresponds to the unit vector for the  $x$ -,  $y$ -, and  $z$ -axis for the  $G$  expressed in  $F$  frame coordinates, respectively. Also, each column corresponds to the unit vector for the  $x$ -,  $y$ -, and  $z$ -axis for the  $F$  expressed in  $G$  frame coordinates, respectively. Each element,  $C_{i,j}$ , corresponds to the cosine of the angle between the  $i$ -axis of the  $G$  frame and the  $j$ -axis of the  $F$  frame.

To derive the time derivative of the 3D DCM, consider the time derivative of any arbitrary vector  $\vec{v}$  in the stationary frame  $F$ , one has

$$\dot{\vec{v}}_F = \dot{C}_{F \leftarrow G} \vec{v}_G + C_{F \leftarrow G} \dot{\vec{v}}_G \quad (6.19)$$

where if  $\vec{v}$  rotates with frame  $G$ , i.e., it is fixed in the rotating frame  $G$ , i.e.,  $\dot{\vec{v}}_G = 0$ , one has

$$\dot{\vec{v}}_F = \dot{C}_{F \leftarrow G} \vec{v}_G \quad (6.20)$$

However, the relationship between the rate of change of the instantaneous axis of rotation and the “velocity” of point represented by  $\vec{v}$  is given by

$$\dot{\vec{v}}_F = [\vec{\omega}_{G \leftarrow F, F}] \times \vec{v}_F \quad (6.21)$$

where  $[\vec{\omega}_{G \leftarrow F, F}] \times$  is the cross-product matrix of the angular velocity of frame  $G$  with respect to frame  $F$  expressed in frame  $F$ . Thus, one has

$$\dot{C}_{F \leftarrow G} \vec{v}_G = [\vec{\omega}_{G \leftarrow F, F}] \times \vec{v}_F = [\vec{\omega}_{G \leftarrow F, F}] \times C_{F \leftarrow G} \vec{v}_G \quad (6.22)$$

or

$$\dot{C}_{F \leftarrow G} = [\vec{\omega}_{G \leftarrow F, F}] \times C_{F \leftarrow G} \quad (6.23)$$

Then, noting the following property for the cross-product matrix

$$[\vec{\omega}_{G \leftarrow F, F}] \times = [C_{F \leftarrow G} \vec{\omega}_{G \leftarrow F, G}] \times = C_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}] \times C_{F \leftarrow G}^T \quad (6.24)$$

and

$$[\vec{\omega}_{G \leftarrow F, F}] \times^T = -[\vec{\omega}_{G \leftarrow F, F}] \times \quad (6.25)$$

one has the following formulas

$$\begin{aligned}\dot{C}_{F \leftarrow G} &= [\vec{\omega}_{G \leftarrow F, F}] \times C_{F \leftarrow G} = C_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}] \times \\ \dot{C}_{G \leftarrow F} &= -[\vec{\omega}_{G \leftarrow F, G}] \times C_{G \leftarrow F} = -C_{G \leftarrow F} [\vec{\omega}_{G \leftarrow F, F}] \times\end{aligned}\quad (6.26)$$

A third representation uses three **basic rotation matrices** defined as

$$C_1(\theta_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta_x & \sin \theta_x \\ 0 & -\sin \theta_x & \cos \theta_x \end{bmatrix} \quad (6.27)$$

$$C_2(\theta_y) = \begin{bmatrix} \cos \theta_y & 0 & -\sin \theta_y \\ 0 & 1 & 0 \\ \sin \theta_y & 0 & \cos \theta_y \end{bmatrix} \quad (6.28)$$

$$C_3(\theta_z) = \begin{bmatrix} \cos \theta_z & \sin \theta_z & 0 \\ -\sin \theta_z & \cos \theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (6.29)$$

which make use of rotations about each primary coordinate axis. Euler proved that every possible rotation in three dimensions, i.e., the rotation angle and Euler axis, can be represented by three sequential rotations which are known as the **Euler angles**. For aerospace vehicle applications, one typically defines the Euler angles for a 3 – 1 – 3 rotation sequence defined by the formula:

$$\vec{v}_G = C_3(\theta_{z*}) C_1(\theta_x) C_3(\theta_z) \vec{v}_F \quad (6.30)$$

or a 3 – 2 – 1 rotation sequence defined by the formula:

$$\vec{v}_G = C_1(\theta_x) C_2(\theta_y) C_3(\theta_z) \vec{v}_F \quad (6.31)$$

where  $G$  is the frame rotating with respect to  $F$ . Here  $\theta_{z*} \in [0, 2\pi]$ ,  $\theta_x \in [0, \pi]$ , and  $\theta_z \in [0, 2\pi]$  are the 3 – 1 – 3 Euler angles, also known as the **classic Euler angles** or **proper Euler angles**, and  $\theta_x \in [0, 2\pi]$ ,  $\theta_y \in [-\pi, \pi]$ , and  $\theta_z \in [0, 2\pi]$  are the 3 – 2 – 1 Euler angles, also known as the **yaw-pitch-roll (ypr) Euler angles** or **Tait-Bryan Euler angles**.

The DCM is related explicitly to the 3 – 1 – 3 Euler angles by the formulas:

$$C_{G \leftarrow F} = C_3(\theta_{z*}) C_1(\theta_x) C_3(\theta_z) \quad (6.32)$$

$$C_{G \leftarrow F} = \begin{bmatrix} \cos \theta_z \cos \theta_{z*} - \cos \theta_x \sin \theta_z \sin \theta_{z*} & \sin \theta_z \cos \theta_{z*} + \cos \theta_x \cos \theta_z \sin \theta_{z*} & \sin \theta_x \sin \theta_{z*} \\ -\cos \theta_z \sin \theta_{z*} - \cos \theta_x \sin \theta_z \cos \theta_{z*} & -\sin \theta_z \sin \theta_{z*} + \cos \theta_x \cos \theta_z \cos \theta_{z*} & \sin \theta_x \cos \theta_{z*} \\ \sin \theta_x \sin \theta_z & -\sin \theta_x \cos \theta_z & \cos \theta_x \end{bmatrix} \quad (6.33)$$

and

$$\begin{bmatrix} \theta_{z*} \\ \theta_x \\ \theta_z \end{bmatrix} = \begin{bmatrix} \arctan \frac{C_{13}}{C_{23}} \\ \arccos C_{33} \\ \arctan \frac{C_{31}}{-C_{32}} \end{bmatrix} \quad (6.34)$$

The DCM is related explicitly to the 3 – 2 – 1 Euler angles by the formulas:

$$C_{G \leftarrow F} = C_1(\theta_x) C_2(\theta_y) C_3(\theta_z) \quad (6.35)$$

$$C_{G \leftarrow F} = \begin{bmatrix} \cos \theta_y \cos \theta_z & \cos \theta_y \sin \theta_z & -\sin \theta_y \\ \sin \theta_x \sin \theta_y \cos \theta_z - \cos \theta_x \sin \theta_z & \sin \theta_x \sin \theta_y \sin \theta_z + \cos \theta_x \cos \theta_z & \sin \theta_x \cos \theta_y \\ \cos \theta_x \sin \theta_y \cos \theta_z + \sin \theta_x \sin \theta_z & \cos \theta_x \sin \theta_y \sin \theta_z - \sin \theta_x \cos \theta_z & \cos \theta_x \cos \theta_y \end{bmatrix} \quad (6.36)$$

and

$$\begin{bmatrix} \theta_x \\ \theta_y \\ \theta_z \end{bmatrix} = \begin{bmatrix} \arctan \frac{C_{23}}{C_{33}} \\ -\arcsin C_{13} \\ \arctan \frac{C_{12}}{C_{11}} \end{bmatrix} \quad (6.37)$$

It should be noted that here the  $\arctan()$  functions must return 4-quadrant solution. Also, if  $\theta_y = \pm 90^\circ$ , then the  $3 - 2 - 1$  Euler angle transformation to the DCM loses a degree-of-freedom. This is known as the **Euler angle ambiguity** that must be dealt with in some flight vehicle applications, e.g., switching Euler angle conventions. Euler angles will be used throughout this textbook as they encode the attitude as three parameters, i.e., the three degrees-of-freedom (3-DOF) for the rotation of a rigid body.

In the use of the DCM for dynamics, one often uses the small angle approximation for sine and cosine to derive a linearized expression for the  $3 - 1 - 3$  Euler angles, as

$$C_{G \leftarrow F} = \begin{bmatrix} 1 - \theta_z \theta_{z*} & \theta_z + \theta_{z*} & \theta_x \theta_{z*} \\ -\theta_{z*} - \theta_z & -\theta_z \theta_{z*} - 1 & \theta_x \\ \theta_x \theta_z & -\theta_x & 1 \end{bmatrix} \quad (6.38)$$

and for the  $3 - 2 - 1$  Euler angles as

$$C_{G \leftarrow F} = \begin{bmatrix} 1 & \theta_z & -\theta_y \\ \theta_x \theta_y - \theta_z & \theta_x \theta_y \theta_z + 1 & \theta_x \\ \theta_y + \theta_x \theta_z & \theta_y \theta_z - \theta_x & 1 \end{bmatrix} \quad (6.39)$$

Discarding the second-order terms, one has for the  $3 - 1 - 3$  Euler angles

$$C_{G \leftarrow F} = \begin{bmatrix} 1 & \theta_z + \theta_{z*} & 0 \\ -\theta_{z*} - \theta_z & -1 & \theta_x \\ 0 & -\theta_x & 1 \end{bmatrix} = I - \begin{bmatrix} \theta_x \\ 0 \\ \theta_z + \theta_{z*} \end{bmatrix}_x \quad (6.40)$$

and for the  $3 - 2 - 1$  Euler angles

$$C_{G \leftarrow F} = \begin{bmatrix} 1 & \theta_z & -\theta_y \\ -\theta_z & 1 & \theta_x \\ \theta_y & -\theta_x & 1 \end{bmatrix} = I - \begin{bmatrix} \theta_x \\ \theta_y \\ \theta_z \end{bmatrix}_x \quad (6.41)$$

By definition, the basic rotation matrices have derivatives

$$\dot{C}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\sin \theta_x & \cos \theta_x \\ 0 & -\cos \theta_x & -\sin \theta_x \end{bmatrix} \dot{\theta}_x = - \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta_x & \sin \theta_x \\ 0 & -\sin \theta_x & \cos \theta_x \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\dot{\theta}_x \\ 0 & \dot{\theta}_x & 0 \end{bmatrix} = -C_1 \begin{bmatrix} \dot{\theta}_x \\ 0 \\ 0 \end{bmatrix}_x \quad (6.42)$$

$$\dot{C}_2 = \begin{bmatrix} -\sin \theta_y & 0 & \cos \theta_y \\ 0 & 0 & 0 \\ -\cos \theta_y & 0 & -\sin \theta_y \end{bmatrix} \dot{\theta}_y = - \begin{bmatrix} \cos \theta_y & 0 & -\sin \theta_y \\ 0 & 1 & 0 \\ \sin \theta_y & 0 & \cos \theta_y \end{bmatrix} \begin{bmatrix} 0 & 0 & \dot{\theta}_y \\ 0 & 0 & 0 \\ -\dot{\theta}_y & 0 & 0 \end{bmatrix} = -C_2 \begin{bmatrix} 0 \\ \dot{\theta}_y \\ 0 \end{bmatrix}_x \quad (6.43)$$

$$\dot{C}_3 = \begin{bmatrix} -\sin \theta_z & \cos \theta_z & 0 \\ -\cos \theta_z & -\sin \theta_z & 0 \\ 0 & 0 & 0 \end{bmatrix} \dot{\theta}_z = - \begin{bmatrix} \cos \theta_z & \sin \theta_z & 0 \\ -\sin \theta_z & \cos \theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -\dot{\theta}_z & 0 \\ \dot{\theta}_z & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = -C_3 \begin{bmatrix} 0 \\ 0 \\ \dot{\theta}_z \end{bmatrix} \quad (6.44)$$

Thus, by the product rule and the definition of the intermediate axes of rotation, the derivatives of the 3 – 1 – 3 Euler angles are related explicitly to the angular velocity by the

$$\vec{\omega}_{G \leftarrow F, G} = C_3(\theta_{z*}) \begin{bmatrix} 0 \\ 0 \\ \dot{\theta}_{z*} \end{bmatrix} + C_3(\theta_{z*})C_1(\theta_x) \begin{bmatrix} \dot{\theta}_x \\ 0 \\ 0 \end{bmatrix} + C_3(\theta_{z*})C_1(\theta_x)C_3(\theta_z) \begin{bmatrix} 0 \\ 0 \\ \dot{\theta}_z \end{bmatrix} \quad (6.45)$$

which result in the following formulas:

$$\vec{\omega}_{G \leftarrow F, G} = \begin{bmatrix} \sin \theta_x \sin \theta_z & \cos \theta_z & 0 \\ \sin \theta_x \cos \theta_z & -\sin \theta_z & 0 \\ \cos \theta_x & 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{\theta}_{z*} \\ \dot{\theta}_x \\ \dot{\theta}_z \end{bmatrix} \quad (6.46)$$

and

$$\begin{bmatrix} \dot{\theta}_{z*} \\ \dot{\theta}_x \\ \dot{\theta}_z \end{bmatrix} = \begin{bmatrix} \sin \theta_z \csc \theta_x & \cos \theta_z \csc \theta_x & 0 \\ \cos \theta_z & -\sin \theta_z & 0 \\ -\sin \theta_z \cot \theta_x & -\cos \theta_z \cot \theta_x & 1 \end{bmatrix} \vec{\omega}_{G \leftarrow F, G} \quad (6.47)$$

The derivatives of the 3 – 2 – 1 Euler angles are related explicitly to the angular velocity by

$$\vec{\omega}_{G \leftarrow F, G} = C_1(\theta_x) \begin{bmatrix} \dot{\theta}_x \\ 0 \\ 0 \end{bmatrix} + C_1(\theta_x)C_2(\theta_y) \begin{bmatrix} 0 \\ \dot{\theta}_y \\ 0 \end{bmatrix} + C_1(\theta_x)C_2(\theta_y)C_3(\theta_z) \begin{bmatrix} 0 \\ 0 \\ \dot{\theta}_z \end{bmatrix} \quad (6.48)$$

which results in

$$\vec{\omega}_{G \leftarrow F, G} = \begin{bmatrix} 1 & 0 & -\sin \theta_x \\ 0 & \cos \theta_x & \sin \theta_x \cos \theta_y \\ 0 & -\sin \theta_x & \cos \theta_x \cos \theta_y \end{bmatrix} \begin{bmatrix} \dot{\theta}_x \\ \dot{\theta}_y \\ \dot{\theta}_z \end{bmatrix} \quad (6.49)$$

and can be solved for the inverse conversion as

$$\begin{bmatrix} \dot{\theta}_x \\ \dot{\theta}_y \\ \dot{\theta}_z \end{bmatrix} = \begin{bmatrix} 1 & \sin \theta_x \tan \theta_y & \cos \phi \tan \theta_y \\ 0 & \cos \theta_x & -\sin \theta_x \\ 0 & \sin \theta_x \sec \theta_y & \cos \theta_x \sec \theta_y \end{bmatrix} \vec{\omega}_{G \leftarrow F, G} \quad (6.50)$$

where it should be noted that these conversion matrices are not transposes of each other as for DCMs.

## Reference Frame Translations and Rotations

For two- and three-dimensional translations without rotation, one simply can add the offset vector between the origins of frames  $F$  and  $G$ , i.e.,

$$\vec{v}_G = \vec{v}_F + \vec{o}_{G \leftarrow F} \quad (6.51)$$

where  $\vec{o}_{G \leftarrow F}$  is the offset of frame  $G$ 's origin relative to frame  $F$ 's origin in coordinates consistent with the identical axes. Similarly,

$$\vec{v}_F = \vec{v}_G + \vec{o}_{F \leftarrow G} \quad (6.52)$$

where  $\vec{o}_{F \leftarrow G} = -\vec{o}_{G \leftarrow F}$  is the offset of frame  $F$ 's origin relative to frame  $G$ 's origin. Thus, for any general two- and three-dimensional frame transformation involving rotation and translation between two frames  $F$  and  $G$ , one can first perform a rotation to align the frame axes, then translate the origin. In terms of the two- or three-dimensional DCM for the axes rotation component, one can write

$$\vec{v}_G = C_{G \leftarrow F} \vec{v}_F + \vec{o}_{G \leftarrow F, G} = C_{G \leftarrow F} (\vec{v}_F + \vec{o}_{F \leftarrow G, F}) \quad (6.53)$$

and

$$\vec{v}_F = C_{F \leftarrow G} \vec{v}_G + \vec{o}_{F \leftarrow G, F} = C_{F \leftarrow G} (\vec{v}_G + \vec{o}_{F \leftarrow G, G}) \quad (6.54)$$

where notably, the offset vector is written in different axes dependent on the direction of the transformation.

Furthermore, using the properties of matrices, one can also use homogeneous coordinates and augmented matrices to perform the matrix multiplication and vector addition as a single matrix multiplication which may be desirable for computational efficiency especially when many coordinate frames are used. **Homogeneous coordinates** append a 1 to the two- or three-dimensional vector. This allows one to rewrite the frame transformation as

$$\begin{bmatrix} \vec{v}_G \\ 1 \end{bmatrix} = \begin{bmatrix} C_{G \leftarrow F} & \vec{o}_{G \leftarrow F, G} \\ \vec{0}_{1 \times \bullet} & 1 \end{bmatrix} \begin{bmatrix} \vec{v}_F \\ 1 \end{bmatrix} \quad (6.55)$$

and

$$\begin{bmatrix} \vec{v}_F \\ 1 \end{bmatrix} = \begin{bmatrix} C_{F \leftarrow G} & \vec{o}_{F \leftarrow G, F} \\ \vec{0}_{1 \times \bullet} & 1 \end{bmatrix} \begin{bmatrix} \vec{v}_G \\ 1 \end{bmatrix} \quad (6.56)$$

where  $\bullet$  is either 2 or 3 corresponding to the dimension of  $\vec{v}$ .

## Euler Symmetric Parameters

A fourth mathematically convenient and vectorized representation of the Euler axis and rotation angle is the second representation known as the **Euler symmetric parameters**, also known as a **passive rotation quaternion**, which can be defined in several different ways resulting in different rigid-body kinematics. For aerospace vehicles, two different conventions are common, the Jet Propulsion Laboratory (JPL) and the Hamilton. This textbook presents the Hamilton convention.

$$\vec{q} = \begin{bmatrix} q_w \\ q_x \\ q_y \\ q_z \end{bmatrix} = \begin{bmatrix} \cos(\theta/2) \\ \sin(\theta/2) \cos(e_x) \\ \sin(\theta/2) \cos(e_y) \\ \sin(\theta/2) \cos(e_z) \end{bmatrix} \quad (6.57)$$

where  $\vec{q}$  is the unit quaternion

$$\|\vec{q}\|_2 = \sqrt{q_w^2 + q_x^2 + q_y^2 + q_z^2} = 1 \quad (6.58)$$

The Euler parameters to axis-angle conversion is given by

$$\theta = 2 \arccos \vec{q}_w \quad (6.59)$$

$$\vec{e} = \frac{\tilde{\vec{q}}}{\|\tilde{\vec{q}}\|_2} \quad (6.60)$$

where

$$\tilde{\vec{q}} = \begin{bmatrix} q_x \\ q_y \\ q_z \end{bmatrix} \quad (6.61)$$

The Euler parameters to DCM conversion is given by

$$C_{G \leftarrow F} = \begin{bmatrix} q_w^2 + q_x^2 - q_y^2 - q_z^2 & 2(q_x q_y + q_w q_z) & 2(q_x q_z - q_w q_y) \\ 2(q_x q_y - q_w q_z) & q_w^2 - q_x^2 + q_y^2 - q_z^2 & 2(q_y q_z + q_w q_x) \\ 2(q_x q_z + q_w q_y) & 2(q_y q_z - q_w q_x) & q_w^2 - q_x^2 - q_y^2 + q_z^2 \end{bmatrix} \quad (6.62)$$

The DCM to Euler parameters conversion is given by any of four equivalent methods

$$\begin{bmatrix} q_w \\ q_x \\ q_y \\ q_z \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\sqrt{1+C_{11}+C_{22}+C_{33}} \\ \frac{1}{4q_w}(C_{32}-C_{23}) \\ \frac{1}{4q_w}(C_{13}-C_{31}) \\ \frac{1}{4q_w}(C_{21}-C_{12}) \end{bmatrix} \quad (6.63)$$

$$\begin{bmatrix} q_w \\ q_x \\ q_y \\ q_z \end{bmatrix} = \begin{bmatrix} \frac{1}{4q_x}(C_{32}-C_{23}) \\ \frac{1}{2}\sqrt{1+C_{11}-C_{22}-C_{33}} \\ \frac{1}{4q_x}(C_{21}+C_{12}) \\ \frac{1}{4q_x}(C_{13}+C_{31}) \end{bmatrix} \quad (6.64)$$

$$\begin{bmatrix} q_w \\ q_x \\ q_y \\ q_z \end{bmatrix} = \begin{bmatrix} \frac{1}{4q_y}(C_{13}-C_{31}) \\ \frac{1}{4q_y}(C_{21}+C_{12}) \\ \frac{1}{2}\sqrt{1-C_{11}+C_{22}-C_{33}} \\ \frac{1}{4q_y}(C_{32}+C_{23}) \end{bmatrix} \quad (6.65)$$

and

$$\begin{bmatrix} q_w \\ q_x \\ q_y \\ q_z \end{bmatrix} = \begin{bmatrix} \frac{1}{4q_z}(C_{21}-C_{12}) \\ \frac{1}{4q_z}(C_{13}+C_{31}) \\ \frac{1}{4q_z}(C_{32}+C_{23}) \\ \frac{1}{2}\sqrt{1-C_{11}-C_{22}+C_{33}} \end{bmatrix} \quad (6.66)$$

where usually use formula with largest element, i.e.,  $q_w$ ,  $q_x$ ,  $q_y$ , and  $q_z$ , for numerical precision of division operation.

The Euler parameters to Euler angles conversion is given by

$$\begin{bmatrix} \theta_x \\ \theta_y \\ \theta_z \end{bmatrix} = \begin{bmatrix} \arctan\left(\frac{2(q_w q_x + q_y q_z)}{1 - 2(q_x^2 + q_y^2)}\right) \\ \arcsin\left(2(q_w q_y - q_x q_z)\right) \\ \arctan\left(\frac{2(q_w q_z + q_x q_y)}{1 - 2(q_y^2 + q_z^2)}\right) \end{bmatrix} \quad (6.67)$$

and the Euler angles to Euler parameters conversion is given by

$$\vec{q} = \begin{bmatrix} \cos \frac{\theta_x}{2} \cos \frac{\theta_y}{2} \cos \frac{\theta_z}{2} + \sin \frac{\theta_x}{2} \sin \frac{\theta_y}{2} \sin \frac{\theta_z}{2} \\ \sin \frac{\theta_x}{2} \cos \frac{\theta_y}{2} \cos \frac{\theta_z}{2} - \cos \frac{\theta_x}{2} \sin \frac{\theta_y}{2} \sin \frac{\theta_z}{2} \\ \cos \frac{\theta_x}{2} \sin \frac{\theta_y}{2} \cos \frac{\theta_z}{2} + \sin \frac{\theta_x}{2} \cos \frac{\theta_y}{2} \sin \frac{\theta_z}{2} \\ \cos \frac{\theta_x}{2} \cos \frac{\theta_y}{2} \sin \frac{\theta_z}{2} - \sin \frac{\theta_x}{2} \sin \frac{\theta_y}{2} \cos \frac{\theta_z}{2} \end{bmatrix} \quad (6.68)$$

## References

For more information, please refer to the following

- Mebius, J. E., “Derivation of the Euler-Rodrigues Formula for Three-Dimensional Rotations from the General Formula for Four-Dimensional Rotations,” at <https://arxiv.org/abs/math/0701759>, 2007
- Nelson, R. C., “3.3 Orientation and Position of the Airplane,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 101-103
- Schmidt, D. K., “1.3 Vectors, Coordinate Transformations, and Direction-Cosine Matrices,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 4-10
- Sola, J., “Quaternion Kinematics for the Error-State Kalman Filter,” at <https://arxiv.org/abs/1711.02508> 2017
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “1.3 Matrix Operations on Vector Coordinates,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 8-15
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “1.4 Rotational Kinematics,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 16-20
- Zhao, S., “Time Derivative of Rotation Matrices: A Tutorial,” at <https://arxiv.org/abs/1609.06088>, 2016

## 6.2 Point-Mass Dynamics

In **point-mass dynamics**, also known as **particle dynamics**, one models a body by a single position in space,  $\vec{x}$ , i.e., a point, with some mass,  $m$  which characterizes the body’s **inertia** or resistance to motion. As no body is truly a singular point in space, point-mass dynamics use the **center of mass**, i.e., the average position of the mass distribution, as the point for inferring the motion of a physical body. The motion of a point-mass is governed by Newton’s laws of motion, known as **point-mass kinetics**, and requires the use of calculus to relate position to velocity and acceleration, known as **point-mass kinematics**. In this section, and frame  $F$  is an inertial reference frame, i.e., its axes are non-rotating and its origin is non-accelerating, and frame  $G$  is an rotating and/or accelerating reference frame.

### Point-Mass Kinematics

Consider if the position of the point-mass and the reference frame for representation of that position are changing in time, then one can represent **inertial point-mass velocity**,  $\vec{v}(t)$ , and the **inertial point-mass acceleration**,  $\vec{a}(t)$ , in an inertial frame  $F$  or a rotating and/or accelerating frame  $G$  at a dynamic rotation angle  $\theta_{G \leftarrow F}(t)$  and a dynamic offset vector  $\vec{o}_{G \leftarrow F}(t)$ . For an inertial frame  $F$ , one has the following for the first and second time derivative of the position coordinates

$$\vec{v}_F(t) = \dot{\vec{x}}_F \quad (6.69)$$

and

$$\vec{a}_F(t) = \dot{\vec{v}}_F = \ddot{\vec{x}}_F \quad (6.70)$$

For a rotating and/or accelerating frame  $G$ , note the inertial velocity of the point-mass can be written as

$$\vec{v}_F(t) = C_{F \leftarrow G} \vec{v}_G(t) = \vec{v}_{\vec{\sigma}_{G \leftarrow F}, F} + \frac{d}{dt} (C_{F \leftarrow G}(t) \vec{x}_G(t)) \quad (6.71)$$

where  $\vec{v}_{o,G \leftarrow F,F}$  is the **inertial** velocity of the origin of frame  $G$  relative to frame  $F$ . By the product rule, one has

$$C_{F \leftarrow G} \vec{v}_G(t) = \vec{v}_{\vec{\sigma}_{G \leftarrow F}, F} + C_{F \leftarrow G} \dot{\vec{x}}_G(t) + \dot{C}_{F \leftarrow G}(t) \vec{x}_G(t) \quad (6.72)$$

where  $\dot{\vec{x}}_G$  is the **apparent linear velocity** in the frame  $G$  and can be considered the partial time derivative of the vector elements, i.e.,

$$\dot{\vec{x}}_G(t) = \frac{\partial}{\partial t} \vec{x}_G(t) \quad (6.73)$$

By the definition of the DCM derivative, one has

$$C_{F \leftarrow G} \vec{v}_G(t) = \vec{v}_{\vec{\sigma}_{G \leftarrow F}, F} + C_{F \leftarrow G} \dot{\vec{x}}_G(t) + C_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}]_{\times} \vec{x}_G(t) \quad (6.74)$$

where  $\vec{\omega}_{G \leftarrow F, G} = \dot{\vec{\theta}}_{G \leftarrow F, G}$  is the angular velocity of frame  $G$  with respect to frame  $F$  expressed in frame  $G$ . Finally, by multiplying by  $C_{G \leftarrow F}^T = C_{G \leftarrow F}^{-1}$  on both sides, one has

$$\vec{v}_G(t) = \vec{v}_{\vec{\sigma}_{G \leftarrow F}, G} + \dot{\vec{x}}_G(t) + [\vec{\omega}_{G \leftarrow F, G}(t)]_{\times} \vec{x}_G(t) \quad (6.75)$$

where  $\vec{v}_{\vec{\sigma}_{G \leftarrow F}, G}$  is the **inertial** velocity of the offset vector of frame  $G$  relative to frame  $F$  expressed in frame  $G$  coordinates.

Thus, the inertial acceleration in the rotating and/or accelerating frame  $G$  can be written as

$$\vec{a}_G(t) = \vec{a}_{\vec{\sigma}_{G \leftarrow F}, G} + \dot{\vec{v}}_G(t) + [\vec{\omega}_{G \leftarrow F, G}]_{\times} \vec{v}_G(t) \quad (6.76)$$

where  $\vec{a}_{\vec{\sigma}_{G \leftarrow F}, G}$  is the **inertial** acceleration of the offset vector of frame  $G$  relative to frame  $F$  expressed in frame  $G$  coordinates. Working from the inertial acceleration, one has

$$\vec{a}_F(t) = \vec{a}_{\vec{\sigma}_{G \leftarrow F}, F} + \frac{d^2}{dt^2} (C_{F \leftarrow G}(t) \vec{x}_G(t)) \quad (6.77)$$

$$C_{F \leftarrow G} \vec{a}_G(t) = \vec{a}_{\vec{\sigma}_{G \leftarrow F}, F} + \frac{d}{dt} (C_{F \leftarrow G} \dot{\vec{x}}_G(t) + \dot{C}_{F \leftarrow G}(t) \vec{x}_G(t)) \quad (6.78)$$

$$C_{F \leftarrow G} \vec{a}_G(t) = \vec{a}_{\vec{\sigma}_{G \leftarrow F}, F} + C_{F \leftarrow G} \ddot{\vec{x}}_G(t) + \dot{C}_{F \leftarrow G}(t) \dot{\vec{x}}_G(t) + \ddot{C}_{F \leftarrow G}(t) \vec{x}_G(t) + \dot{C}_{F \leftarrow G}(t) \dot{\vec{x}}_G(t) \quad (6.79)$$

$$C_{F \leftarrow G} \vec{a}_G(t) = \vec{a}_{\vec{\sigma}_{G \leftarrow F}, F} + C_{F \leftarrow G} \ddot{\vec{x}}_G(t) + 2\dot{C}_{F \leftarrow G} \dot{\vec{x}}_G(t) + \frac{d}{dt} \dot{C}_{F \leftarrow G} \vec{x}_G(t) \quad (6.80)$$

By the definition of the DCM derivative, one has

$$C_{F \leftarrow G} \vec{a}_G(t) = \vec{a}_{\vec{\sigma}_{G \leftarrow F}, F} + C_{F \leftarrow G} \ddot{\vec{x}}_G(t) + 2C_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}]_{\times} \dot{\vec{x}}_G(t) + \frac{d}{dt} (C_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}]_{\times}) \vec{x}_G(t) \quad (6.81)$$

$$C_{F \leftarrow G} \vec{a}_G(t) = \vec{a}_{\vec{o}_{G \leftarrow F}, F} + C_{F \leftarrow G} \ddot{\vec{x}}_G(t) + 2C_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}] \times \dot{\vec{x}}_G(t) + C_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}] \times \vec{x}_G(t) + \dot{C}_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}] \times \vec{x}_G(t) \quad (6.82)$$

$$C_{F \leftarrow G} \vec{a}_G(t) = \vec{a}_{\vec{o}_{G \leftarrow F}, F} + C_{F \leftarrow G} \ddot{\vec{x}}_G(t) + 2C_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}] \times \dot{\vec{x}}_G(t) + C_{F \leftarrow G} [\vec{a}_{G \leftarrow F, G}] \times \vec{x}_G(t) + C_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}] \times [\vec{\omega}_{G \leftarrow F, G}] \times \vec{x}_G(t) \quad (6.83)$$

where  $\vec{a}_{G \leftarrow F, G} = \dot{\vec{\omega}}_{G \leftarrow F, G}$  is the angular acceleration of frame  $G$  with respect to frame  $F$  expressed in frame  $G$  and  $\dot{\vec{x}}_G$  is the **apparent linear velocity** in the frame  $G$  and can be considered the partial time derivative of the vector elements, i.e.,

$$\ddot{\vec{x}}_G(t) = \frac{\partial^2}{\partial t^2} \vec{x}_G(t) \quad (6.84)$$

Finally, by multiplying by  $C_{G \leftarrow F}^T = C_{G \leftarrow F}^{-1}$  on both sides, one has

$$\vec{a}_G(t) = \vec{a}_{\vec{o}_{G \leftarrow F}, G} + \ddot{\vec{x}}_G(t) + 2[\vec{\omega}_{G \leftarrow F, G}] \times \dot{\vec{x}}_G(t) + [\vec{a}_{G \leftarrow F, G}] \times \vec{x}_G(t) + [\vec{\omega}_{G \leftarrow F, G}] \times [\vec{\omega}_{G \leftarrow F, G}] \times \vec{x}_G(t) \quad (6.85)$$

where  $\vec{a}_{\vec{o}_{G \leftarrow F}, F}$  is the **inertial** acceleration of the offset vector of frame  $G$  relative to frame  $F$  expressed in frame  $F$  coordinates,

$$2[\vec{\omega}_{G \leftarrow F, G}] \times \dot{\vec{x}}_G(t) \quad (6.86)$$

is the **Coriolis acceleration**,

$$[\vec{a}_{G \leftarrow F, G}] \times \vec{x}_G(t) \quad (6.87)$$

is the **Euler acceleration**, and

$$[\vec{\omega}_{G \leftarrow F, G}] \times [\vec{\omega}_{G \leftarrow F, G}] \times \vec{x}_G(t) \quad (6.88)$$

is the **centrifugal acceleration**. Collectively, these are known as **fictitious accelerations** and produce **fictitious forces**.

## Point-Mass Kinetics

Point-mass kinetics are represented using the **Newton's second law** governing the translation of the point-mass which relates the forces acting on the point-mass to the momentum. This results in the vector-valued differential equation in an inertial frame  $F$  known as **Newton equation of motion (EOM)**

$$\sum \vec{F}_F = \frac{d}{dt} \vec{p}_F = \frac{d}{dt} (m \vec{v}_F) = \dot{m} \vec{v}_F + m \dot{\vec{v}}_F = \dot{m} \vec{v}_F + m \vec{a}_F = \dot{m} \dot{\vec{x}}_F + m \ddot{\vec{x}}_F \quad (6.89)$$

where  $\sum \vec{F}_F$  is the total force on the point-mass in frame  $F$  coordinates,  $\vec{p}_F$  is the momentum of the point-mass in frame  $F$  coordinates, and  $\vec{v}_F$  is the velocity of the point-mass in frame  $F$  coordinates. If the mass is constant over time, one has

$$\sum \vec{F}_F = m \vec{a}_F = m \dot{\vec{v}}_F = m \ddot{\vec{x}}_F \quad (6.90)$$

Furthermore, one can define the **net force impulse**,  $\vec{I}_F$ , as

$$\vec{I}_F = \int_{t_1}^{t_2} \sum \vec{F}_F dt = \Delta \vec{p}_F \quad (6.91)$$

where  $\Delta \vec{p}_F = m(t_2) \vec{v}_F(t_2) - m(t_1) \vec{v}_F(t_1)$  is the change in momentum. Notably, if the mass is constant, one has

$$\Delta \vec{v}_F = \frac{\vec{I}_F}{m} \quad (6.92)$$

where  $\Delta \vec{v}_F = \vec{v}_F(t_2) - \vec{v}_F(t_1)$ . If  $\sum \vec{F}_F$  is also constant

$$\Delta \vec{v}_F = \frac{\sum \vec{F}_F \Delta t}{m} \quad (6.93)$$

where  $\Delta t = t_2 - t_1$ .

In addition, one can define the moment of the net force on point-mass  $m$  about some point  $O$  as

$$\vec{M}_{O,F} = [\vec{r}_F] \times \vec{F}_F = [\vec{r}_F] \times (\dot{m} \vec{v}_F + m \vec{v}_F) = [\vec{r}_F] \times (\dot{m} \vec{x}_F + m \vec{\ddot{x}}_F) \quad (6.94)$$

where  $\vec{r}_F$  is the position of the point-mass relative to point  $O$ . The angular momentum of point-mass  $m$  about some point  $O$  as

$$\vec{H}_{O,F} = [\vec{r}_F] \times m \vec{v}_F = [\vec{r}_F] \times m \vec{\dot{x}}_F \quad (6.95)$$

which provides

$$\vec{M}_{O,F} = [\vec{r}_F] \times \dot{m} \vec{v}_F + \vec{H}_{O,F} = [\vec{r}_F] \times \dot{m} \vec{\dot{x}}_F + \vec{H}_{O,F} \quad (6.96)$$

which are related via

$$\int_{t_1}^{t_2} \vec{M}_{O,F} dt = \Delta \vec{H}_{O,F} + \int_{t_1}^{t_2} [\vec{r}_F] \times \dot{m} \vec{v}_F dt \quad (6.97)$$

where  $\int_{t_1}^{t_2} \vec{M}_{O,F} dt$  is the **net angular impulse** of a force  $F$ . Notably, if the mass is constant over time, one has

$$\int_{t_1}^{t_2} \vec{M}_{O,F} dt = \Delta \vec{H}_{O,F} \quad (6.98)$$

If one is using a rotating and/or accelerating reference frame  $G$ , the Newton EOM becomes

$$\sum \vec{F}_G = \frac{d}{dt} \vec{p}_G = \frac{d}{dt} (m \vec{v}_G) = \dot{m} \vec{v}_G + m (\vec{a}_{\vec{\sigma}_{G \leftarrow F}, G} + \vec{\dot{v}}_G + [\vec{\omega}_{G \leftarrow F, G}] \times \vec{v}_G) \quad (6.99)$$

and if one assumes the mass is constant over time, one has

$$\sum \vec{F}_G = m (\vec{a}_{\vec{\sigma}_{G \leftarrow F}, G} + \vec{\dot{v}}_G + [\vec{\omega}_{G \leftarrow F, G}] \times \vec{v}_G) \quad (6.100)$$

where  $\sum \vec{F}_G$  is the total force acting on the point-mass in frame  $G$  coordinates,  $\vec{v}_G$  is the inertial velocity of the point-mass in frame  $G$  coordinates, and  $\vec{\omega}_{G \leftarrow F, G}$  is the angular velocity of the rotating/accelerating reference frame  $G$  relative to the inertial reference frame  $F$  written in frame  $G$  coordinates. Notably, one could also substitute  $\vec{x}_G$  and its derivatives instead of  $\vec{v}_G$  into the Newton EOM especially if some forces depend on the position. This provides

$$\begin{aligned} \sum \vec{F}_G = & \dot{m} (\vec{v}_{\vec{\sigma}_{G \leftarrow F}, G} + \vec{\dot{x}}_G(t) + [\vec{\omega}_{G \leftarrow F, G}(t)] \times \vec{x}_G(t)) \\ & + m (\vec{a}_{\vec{\sigma}_{G \leftarrow F}, G} + \vec{\ddot{x}}_G + 2[\vec{\omega}_{G \leftarrow F, G}] \times \vec{\dot{x}}_G + [\vec{\omega}_{G \leftarrow F, G}] \times [\vec{\omega}_{G \leftarrow F, G}] \times \vec{x}_G) \end{aligned} \quad (6.101)$$

and if one assumes the mass is constant over time, one has

$$\sum \vec{F}_G = m (\vec{a}_{\vec{o}_{G \leftarrow F}, G} + \ddot{\vec{x}}_G(t) + 2[\vec{\omega}_{G \leftarrow F, G}] \times \dot{\vec{x}}_G + [\vec{\omega}_{G \leftarrow F, G}] \times [\vec{\omega}_{G \leftarrow F, G}] \times \vec{x}_G) \quad (6.102)$$

However, if not, one typically separates the position and velocity kinematics relationship as a supplemental equation, e.g.,

$$\dot{\vec{x}}_F = \vec{v}_F(t) = C_{F \leftarrow G}(t) \vec{v}_G(t) \quad (6.103)$$

which could be numerically integrated for a point-mass trajectory over time in the inertial reference frame. Notably, to solve for  $C_{F \leftarrow G}(t)$ , one requires an additional supplemental equation for the relationship between the angular velocity and the attitude representation, e.g., the DCM derivative equation

$$\dot{C}_{F \leftarrow G} = C_{F \leftarrow G} [\vec{\omega}_{G \leftarrow F, G}] \times \quad (6.104)$$

the 3 – 1 – 3 Euler angle rates equation

$$\begin{bmatrix} \dot{\theta}_{z^*} \\ \dot{\theta}_x \\ \dot{\theta}_z \end{bmatrix} = \begin{bmatrix} \sin \theta_z \csc \theta_x & \cos \theta_z \csc \theta_x & 0 \\ \cos \theta_z & -\sin \theta_z & 0 \\ -\sin \theta_z \cot \theta_x & -\cos \theta_z \cot \theta_x & 1 \end{bmatrix} \vec{\omega}_{G \leftarrow F, G} \quad (6.105)$$

or the 3 – 2 – 1 Euler angle rates equation

$$\begin{bmatrix} \dot{\theta}_x \\ \dot{\theta}_y \\ \dot{\theta}_z \end{bmatrix} = \begin{bmatrix} 1 & \sin \theta_x \tan \theta_y & \cos \phi \tan \theta_y \\ 0 & \cos \theta_x & -\sin \theta_x \\ 0 & \sin \theta_x \sec \theta_y & \cos \theta_x \sec \theta_y \end{bmatrix} \vec{\omega}_{G \leftarrow F, G} \quad (6.106)$$

## References

For more information, please refer to the following

- Curtis, H. D., “1.3 Kinematics,” *Orbital Mechanics for Engineering Students*, 4th ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2021, pp. 9-14
- Curtis, H. D., “1.5 Newton’s law of motion,” *Orbital Mechanics for Engineering Students*, 4th ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2021, pp. 17-21
- Curtis, H. D., “1.6 Time derivatives of moving vectors,” *Orbital Mechanics for Engineering Students*, 4th ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2021, pp. 21-26
- Curtis, H. D., “1.7 Relative motion,” *Orbital Mechanics for Engineering Students*, 4th ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2021, pp. 26-34
- Schmidt, D. K., “1.5 Newton’s Second Law,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 14-18
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “1.5 Translational Kinematics,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 20-23

### 6.3 Rigid-Body Dynamics

Rigid-body dynamics are represented using the **Newton-Euler equations of motion (EOM)** which adds Euler's equation governing rotation which relates moments acting on the rigid body to the angular momentum. This results in the vector-valued differential equations:

$$\begin{aligned}\sum \vec{F} &= \frac{d}{dt} \vec{p}_G \\ \sum \vec{M} &= \frac{d}{dt} \vec{H}_G\end{aligned}\tag{6.107}$$

where  $\sum \vec{M}$  is the total moment on the rigid body, and  $\vec{H}_G$  is the angular momentum about the center of mass  $G$ . It should be noted that the first equation of the Newton-Euler EOMs is called the **translation equation** and is equivalent to the point-mass equation while the second is called the **rotation equation**.

Since a rigid body can be represented as a continuous mass distribution, these quantities can be computed as integrals over the volume  $V$  by the following equations

$$\begin{aligned}\int_V \vec{f} \rho dV &= \frac{d}{dt} \int_V \vec{v} \rho dV \\ \int_V \vec{r} \times \vec{f} \rho dV &= \frac{d}{dt} \int_V \vec{r} \times \vec{v} \rho dV\end{aligned}\tag{6.108}$$

where  $\vec{f}$  is the forces acting on the rigid body per unit mass,  $dm = \rho dV$  is an infinitesimal mass element of the body with density  $\rho$ ,  $\vec{v}$  is the velocity of that mass element, and  $\vec{r}$  is the position vector of the mass element with respect to the origin of an inertial reference frame.

To get around these integrals, one can represent the static distribution of mass through the **inertia matrix**, also known as the **inertia tensor**, which is composed of the **moments of inertia** and the **products of inertia** about the body-fixed frame coordinates axes  $x_B - y_B - z_B$ , i.e.

$$I_G = \begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{bmatrix},\tag{6.109}$$

where

$$I_{xx} = \int_V (y_B^2 + z_B^2) \rho dV\tag{6.110}$$

$$I_{yy} = \int_V (x_B^2 + z_B^2) \rho dV\tag{6.111}$$

$$I_{zz} = \int_V (x_B^2 + y_B^2) \rho dV\tag{6.112}$$

$$I_{xy} = I_{yx} = \int_V x_B y_B \rho dV\tag{6.113}$$

$$I_{xz} = I_{zx} = \int_V x_B z_B \rho dV\tag{6.114}$$

$$I_{yz} = I_{zy} = \int_V y_B z_B \rho dV \quad (6.115)$$

This introductory chapter on rigid-body dynamics will also use the **constant-mass approximation** which assumes the mass distribution of the rigid aerospace vehicle does not change. However, for non-electric powered aerospace vehicles, engine fuel consumption will cause mass to change at some rate. This mass rate may be slow in the case of aircraft engines and is often neglected, but for rocket engines the mass rate is crucial in deriving appropriate EOMs. This and other mass effects will be addressed in the next chapter of this textbook on advanced rigid aerospace vehicle dynamics.

Thus, if one assumes that the total mass  $m$  and inertia matrix are constant over time and one defines the reference frame as a body-fixed frame, i.e. attached to the rigid structure, then the Newton-Euler EOMs written in the rotating reference frame become

$$\begin{aligned} \sum \vec{F}_B &= m (\dot{\vec{v}}_B + [\vec{\omega}_{B/I,B}] \times \vec{v}_B) \\ \sum \vec{M}_B &= I_G \dot{\vec{\omega}}_{B/I,B} + [\vec{\omega}_{B/I,B}] \times I_G \vec{\omega}_{B/I,B} \end{aligned} \quad (6.116)$$

where  $\sum \vec{F}_B$  is the total force acting on the center of mass  $G$  in the body-fixed frame,  $\vec{v}_B$  is the linear velocity of the center of mass,  $\sum \vec{M}_B$  is the total moment acting about the center of mass  $G$  in the body-fixed frame, and  $\vec{\omega}_{B/I,B}$  is the angular velocity of the body-fixed frame relative to the inertial frame written in body-frame coordinates.

For the body-fixed frame linear and angular velocity components, one typically uses the following representations

$$\vec{v}_B = [u \ v \ w]^T \quad (6.117)$$

and

$$\vec{\omega}_{B/I,B} = [p \ q \ r]^T \quad (6.118)$$

which allows one to write out the translation equation using a skew-symmetric matrix as

$$\sum \vec{F}_B = m \left( \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \begin{bmatrix} 0 & -r & q \\ r & 0 & -p \\ -q & p & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \right) \quad (6.119)$$

$$\sum \vec{F}_B = m \begin{bmatrix} \dot{u} + qw - rv \\ \dot{v} + ru - pw \\ \dot{w} + pv - qu \end{bmatrix} \quad (6.120)$$

and the rotation equation using a skew-symmetric matrix as

$$\sum \vec{M}_B = \begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{bmatrix} \begin{bmatrix} \dot{p} \\ \dot{q} \\ \dot{r} \end{bmatrix} + \begin{bmatrix} 0 & -r & q \\ r & 0 & -p \\ -q & p & 0 \end{bmatrix} \begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix} \quad (6.121)$$

$$\sum \vec{M}_B = \begin{bmatrix} I_{xx}\dot{p} - I_{xy}\dot{q} - I_{xz}\dot{r} \\ -I_{xy}\dot{p} + I_{yy}\dot{q} - I_{yz}\dot{r} \\ -I_{xz}\dot{p} - I_{yz}\dot{q} + I_{zz}\dot{r} \end{bmatrix} + \begin{bmatrix} 0 & -r & q \\ r & 0 & -p \\ -q & p & 0 \end{bmatrix} \begin{bmatrix} I_{xx}p - I_{xy}q - I_{xz}r \\ -I_{xy}p + I_{yy}q - I_{yz}r \\ -I_{xz}p - I_{yz}q + I_{zz}r \end{bmatrix} \quad (6.122)$$

$$\sum \vec{M}_B = \begin{bmatrix} I_{xx}\dot{p} + (I_{zz} - I_{yy})qr - I_{xy}(\dot{q} - pr) - I_{xz}(\dot{r} + pq) - I_{yz}(q^2 - r^2) \\ I_{yy}\dot{q} + (I_{xx} - I_{zz})pr - I_{xy}(\dot{p} + qr) - I_{xz}(r^2 - p^2) - I_{yz}(\dot{r} - pq) \\ I_{zz}\dot{r} + (I_{yy} - I_{xx})pq - I_{xy}(p^2 - q^2) - I_{xz}(\dot{p} - qr) - I_{yz}(\dot{q} + pr) \end{bmatrix} \quad (6.123)$$

Thus, for any rigid body in a rotating body-fixed frame, the Newton-Euler EOMs are

$$\begin{aligned} \sum \vec{F}_B &= m \begin{bmatrix} \dot{u} + qw - rv \\ \dot{v} + ru - pw \\ \dot{w} + pv - qu \end{bmatrix} \\ \sum \vec{M}_B &= \begin{bmatrix} I_{xx}\dot{p} + (I_{zz} - I_{yy})qr - I_{xy}(\dot{q} - pr) - I_{xz}(\dot{r} + pq) + I_{yz}(r^2 - q^2) \\ I_{yy}\dot{q} + (I_{xx} - I_{zz})pr - I_{xy}(\dot{p} + qr) + I_{xz}(p^2 - r^2) - I_{yz}(\dot{r} - pq) \\ I_{zz}\dot{r} + (I_{yy} - I_{xx})pq + I_{xy}(q^2 - p^2) - I_{xz}(\dot{p} - qr) - I_{yz}(\dot{q} + pr) \end{bmatrix} \end{aligned} \quad (6.124)$$

These represent six coupled nonlinear ODEs with 6 free variables,  $u$ ,  $v$ ,  $w$ ,  $p$ ,  $q$ ,  $r$ , thus forming the **six degree-of-freedom (6-DOF) equations of motion (EOM)**.

It should be noted that the directions of the forces and moments for aerospace vehicles are typically expressed well in different frames than the body-fixed frame. Thus, these terms will generally depend on the relative passive rotations and the representation used, e.g., the DCM, the Euler symmetric parameters, or the Euler angles, all of which change as a function of the angular velocity resulting in supplementary differential equations relating the angular velocity to the passive rotations. This textbook will use either the 3 – 1 – 3 or 3 – 2 – 1 Euler angles to represent these dynamics.

Thus, if the Euler angles are used to represent the attitude of the body-fixed frame relative to an “inertial” LVLH/navigation frame, one has the LVLH/navigation-to-body-fixed frame 3 – 2 – 1 Euler angle rates as

$$\begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} 1 & \sin \phi \tan \theta & \cos \phi \tan \theta \\ 0 & \cos \phi & -\sin \phi \\ 0 & \sin \phi \sec \theta & \cos \phi \sec \theta \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix} \quad (6.125)$$

where  $\vec{\omega} = [p \ q \ r]^T$  is the angular velocity of the body-fixed frame relative to the navigation or LVLH frame. However, if the inertial angular acceleration is defined for a rotating Earth-frame, then the rotation equation of motion remains the same, but the LVLH-to-body-fixed frame 3 – 2 – 1 Euler angle rates as the *supplemental equation*,

$$\begin{bmatrix} \dot{\phi}_L \\ \dot{\theta}_L \\ \dot{\psi}_L \end{bmatrix} = \begin{bmatrix} 1 & \sin \phi_L \tan \theta_L & \cos \phi_L \tan \theta_L \\ 0 & \cos \phi_L & -\sin \phi_L \\ 0 & \sin \phi_L \sec \theta_L & \cos \phi_L \sec \theta_L \end{bmatrix} \left( \begin{bmatrix} p \\ q \\ r \end{bmatrix} - \begin{bmatrix} p_{L/I,B} \\ q_{L/I,B} \\ r_{L/I,B} \end{bmatrix} \right) \quad (6.126)$$

or the navigation-to-body-fixed frame 3 – 2 – 1 Euler angle rates as

$$\begin{bmatrix} \dot{\phi}_N \\ \dot{\theta}_N \\ \dot{\psi}_N \end{bmatrix} = \begin{bmatrix} 1 & \sin \phi_N \tan \theta_N & \cos \phi_N \tan \theta_N \\ 0 & \cos \phi_N & -\sin \phi_N \\ 0 & \sin \phi_N \sec \theta_N & \cos \phi_N \sec \theta_N \end{bmatrix} \left( \begin{bmatrix} p \\ q \\ r \end{bmatrix} - \begin{bmatrix} p_{N/I,B} \\ q_{N/I,B} \\ r_{N/I,B} \end{bmatrix} \right) \quad (6.127)$$

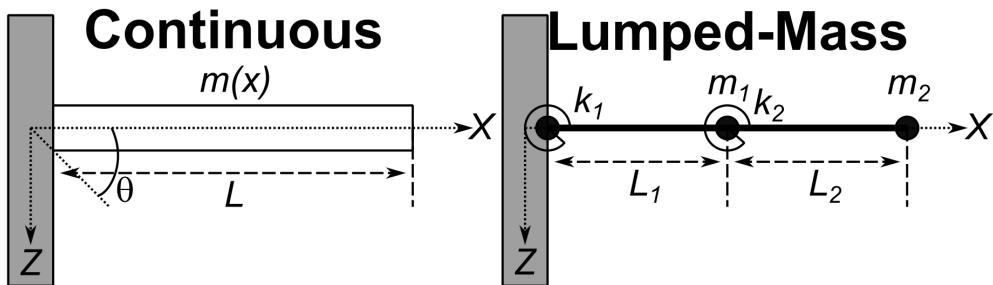
where  $\vec{\omega} = [p \ q \ r]^T$  is the angular velocity of the body-fixed frame relative to the ECI frame.

## 6.4 Elastic-Body Dynamics

As noted previously, introductory FDC assumes that rigid-body dynamics can be used to model the equations of motion for flight vehicles. However, in reality, flight vehicle structures are not rigid, but have some elasticity or flexibility in the structure. Thus, this chapter will look at the additional modeling of structural vibrations in the flight vehicle equations of motion, highlighting elastic airplane dynamics.

### Fixed-Beam Vibration Problem

To contextualize this topic, consider the following two figures for a fixed-beam vibration problem



In the left figure of the fixed-beam, one has a continuous deformable body with some mass distribution,  $m(x)$ , as a function of the horizontal coordinate,  $x$ . It can be shown that the partial differential equation (PDE) governing the vertical deformation of the beam,  $Z$ , is given by

$$\frac{\partial^2}{\partial x^2} \left( EI(x) \frac{\partial^2 Z(x, t)}{\partial x^2} \right) + m(x) \frac{\partial^2 Z(x, t)}{\partial t^2} = 0 \quad (6.128)$$

where  $E$  is the elastic modulus of the beam material and  $I$  is the area moment of inertia of the beam cross-section about its neutral axis. Using the separation of variables technique, one can write the solution to this PDE as

$$Z(x, t) = \sum_{i=1}^{\infty} v_i(x) \eta_i(t) \quad (6.129)$$

which is an infinite sum of terms, each consisting of a purely space-dependent function  $v_i(x)$  and a purely time-dependent function,  $\eta_i(t)$ . The functions  $v_i(x)$  are called the **mode shapes**, also known as the **eigenfunctions**, and the functions  $\eta_i(t)$  are called the **modal coordinates**. As the solution is an infinite sum, the beam-vibration problem is called an **infinite-dimensional problem**.

To obtain a finite-dimensional approximation to this infinite-dimensional problem, one may simplify the continuous beam model as a discrete mass model with  $i = 1, \dots, n$  particles, a.k.a. **lumped-mass model**, where each mass particle,  $m_i$ , has an associated spring stiffness,  $k_i$ , and different particles are connected by massless rigid rods of lengths  $L_i$ . It should be noted that the solution to a vibration problem using a finite-element analysis (FEA) will result in a lumped-mass approximation. Thus, lumped-mass approximations are always used for real, complex, flight vehicle structures in FDC. As such, in the right figure of the beam-vibration example, the continuous beam has been approximated by two masses, two springs, and two rods.

Note that if one desired to have a better approximation, one would simply add more lumped-masses to the beam.

In Lagrangian mechanics, one can solve for the equations of motion for these lumped-mass systems using the Lagrangian  $L = T - U$  where  $T$  is the kinetic energy of the system and  $U$  is the potential (or strain) energy of the system. Then, with no external forces acting on the system, the calculus of variations provides the **Euler-Lagrange equation of motion**

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{q}_i} \right) - \frac{\partial T}{\partial q_i} + \frac{\partial U}{\partial q_i} = 0 \quad (6.130)$$

where  $q_i$  is the  $i$ -th generalized coordinate used to describe the system. These **generalized coordinates** can include both physical and nonphysical coordinates and will be discussed next lecture in detail. Note that this course will make continually use of the Euler-Lagrange equation in the treatment of elastic-body dynamics and will be invoked without proof as its derivation and properties are beyond the scope of this course and typically introduced in graduate level dynamics course.

For the simple beam-vibration example previously, possible coordinates to describe the beam's motion could be the transverse displacements of the masses  $Z_i(t)$ , or the angular displacements of the masses,  $\theta_i(t)$ , which are related using small angle approximation by

$$\begin{bmatrix} \dot{Z}_1 \\ \dot{Z}_2 \end{bmatrix} = \begin{bmatrix} L_1 & 0 \\ L_1 + L_2 & L_2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \quad (6.131)$$

Recalling systems-of-particles dynamics, the kinetic energy of the beam can be written as

$$T = \frac{1}{2} \left( m_1 \dot{Z}_1^2 + m_2 \dot{Z}_2^2 \right) = \frac{1}{2} \begin{bmatrix} \dot{Z}_1 \\ \dot{Z}_2 \end{bmatrix}^T \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \begin{bmatrix} \dot{Z}_1 \\ \dot{Z}_2 \end{bmatrix} \quad (6.132)$$

and the potential energy of the beam can be written as

$$U = \frac{1}{2} \left( k_1 \theta_1^2 + k_2 \theta_2^2 \right) = \frac{1}{2} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}^T \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \quad (6.133)$$

Then, using Lagrange's equation, one has

$$\begin{bmatrix} m_1^2 L_1^2 + m_2 (L_1 + L_2)^2 & m_1 L_2 (L_1 + L_2) \\ m_1 L_2 (L_1 + L_2) & m_2 L_2^2 \end{bmatrix} \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \end{bmatrix} + \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (6.134)$$

which demonstrates the general matrix-vector differential form for *all* lumped-mass vibration problems

$$M \ddot{\vec{q}} + K \vec{q} = \vec{0} \quad (6.135)$$

where  $M$  is the **mass matrix** and  $K$  is the **stiffness matrix**, both of which will be real, symmetric matrices and  $M$  will always be positive-definite, i.e. have strictly positive eigenvalues. Thus, a vibration problem can be completely described by selecting the generalized coordinates, setting up the initial conditions, and finding the mass and stiffness matrices.

## Modal Analysis of Lumped-Mass Vibration

Furthermore, this general lumped-mass vibration problem can be rewritten as

$$\ddot{\vec{q}} + D\vec{q} = \vec{0} \quad (6.136)$$

where  $D = M^{-1}K$  is the **dynamic matrix** and always exists since  $M$  is positive-definite. Then, using the standard transformation

$$\vec{q} = \Psi\vec{\eta} \quad (6.137)$$

where  $\Psi$  is the **modal matrix** of  $D$  consisting of its  $n$  eigenvectors,  $\vec{v}_i$  where  $i = 1, \dots, n$ . Furthermore, one can write

$$\Psi^{-1}D\Psi = \Lambda \quad (6.138)$$

where  $\Lambda$  is a diagonal matrix of the  $n$  eigenvalues of  $D$ ,  $\lambda_i$  where  $i = 1, \dots, n$ .

Recall that the eigenvectors satisfy the equation

$$(\lambda_i I - D) \vec{v}_i = \vec{0} \quad (6.139)$$

and

$$\Psi = [\vec{v}_1 | \vec{v}_2 | \cdots | \vec{v}_n] \quad (6.140)$$

Thus, the general lumped-mass vibration problem becomes

$$\ddot{\vec{\eta}} + \Lambda\vec{\eta} = \vec{0} \quad (6.141)$$

whose  $n$  differential equations are now independent, i.e.

$$\ddot{\eta}_i + \lambda_i \eta_i = 0 \quad (6.142)$$

for  $i = 1, \dots, n$  where each has a solution

$$\eta_i(t) = A_i \cos(\sqrt{\lambda_i}t + \Gamma_i) \quad (6.143)$$

where the constants of integration,  $A_i$  and  $\Gamma_i$ , depend on initial conditions. Thus, for general lumped-mass vibration problems, one can find  $n$  natural modes each oscillating at natural frequencies  $\omega_i = \sqrt{\lambda_i}$ . Furthermore, from the definition of  $\Psi$ , one has

$$\vec{q}(t) = [\vec{v}_1 | \vec{v}_2 | \cdots | \vec{v}_n] \vec{\eta}(t) = \sum_{i=1}^n \vec{v}_i \eta_i(t) \quad (6.144)$$

where each modal response  $\eta_i$  contributes to the system response through the eigenvectors or mode shapes. As eigenvectors have arbitrary magnitude, they are typically normalized to a unit length, unity displacement of a selected element, or unity generalized mass as will be shown.

Recall that by definition of  $D$  and  $\vec{v}_i$ , one has

$$(\lambda_i I - M^{-1}K) \vec{v}_i = 0 \quad (6.145)$$

or

$$\lambda_i M \vec{v}_i = K \vec{v}_i \quad (6.146)$$

Thus, if one multiplies by another eigenvector such that

$$\lambda_i \vec{v}_j^T M \vec{v}_i = v_j^T K \vec{v}_i \quad (6.147)$$

and by the same process, one also has

$$\lambda_j \vec{v}_i^T M \vec{v}_j = v_i^T K \vec{v}_j \quad (6.148)$$

Finally by noting that  $M$  and  $K$  are symmetric, one can further write that

$$(\lambda_i - \lambda_j) \vec{v}_j^T M \vec{v}_i = 0 \quad (6.149)$$

and

$$(\lambda_i - \lambda_j) \vec{v}_j^T K \vec{v}_i = 0 \quad (6.150)$$

which means that if  $\lambda_i \neq \lambda_j \forall i \neq j$ , then the **orthogonality property** holds for the restrained lumped-mass modes, i.e.

$$\vec{v}_j^T M \vec{v}_i = 0, i \neq j \quad (6.151)$$

and

$$\vec{v}_j^T K \vec{v}_i = 0, i \neq j \quad (6.152)$$

Furthermore if  $i = j$ , one then can define the **i-th generalized mass**

$$\mathcal{M}_i = \vec{v}_i^T M \vec{v}_i \quad (6.153)$$

and **i-th generalized stiffness**

$$\mathcal{K}_i = \vec{v}_i^T K \vec{v}_i \quad (6.154)$$

Finally, by this orthogonality property, the definition of  $\Psi$ , and defining  $\mathcal{M} = \text{diag}[\mathcal{M}_1, \dots, \mathcal{M}_n]$  and  $\mathcal{K} = \text{diag}[\mathcal{K}_1, \dots, \mathcal{K}_n]$ , one can alternatively write the lumped-mass vibration equations of motion as

$$\Psi^{-1} M \ddot{\Psi} \vec{\eta} + \Psi^{-1} K \Psi \vec{\eta} = \vec{0} \quad (6.155)$$

$$\mathcal{M} \ddot{\vec{\eta}} + \mathcal{K} \vec{\eta} = \vec{0} \quad (6.156)$$

or

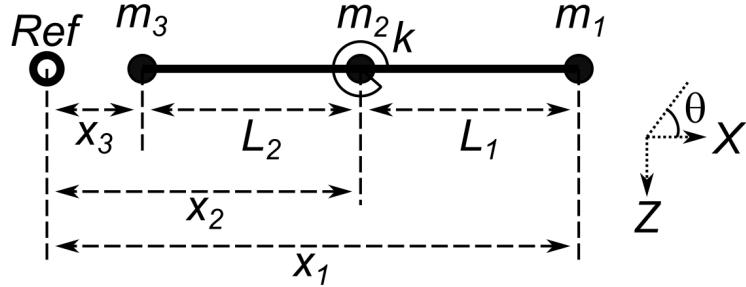
$$\ddot{\vec{\eta}} + \mathcal{M}^{-1} \mathcal{K} \vec{\eta} = \vec{0} \quad (6.157)$$

which demonstrates that

$$\lambda_i = \frac{\mathcal{M}_i}{\mathcal{K}_i} \quad (6.158)$$

### Unrestrained Lumped-Mass Model

Consider the following unrestrained three-lumped-mass system which is free to translate and rotate



where first only the vertical displacement  $Z$  will be analyzed. Note that the bending displacement of the beam occurs by the relative deflection angle  $\theta$  between the lines for rods 1 and 2. This deflection angle can be shown using small angle approximation to be

$$\theta = \frac{Z_1 - Z_2}{x_1 - x_2} - \frac{Z_2 - Z_3}{x_2 - x_3} = \left[ \frac{1}{x_1 - x_2} \quad -\frac{1}{x_1 - x_2} \quad -\frac{1}{x_2 - x_3} \quad \frac{1}{x_2 - x_3} \right] \vec{Z} = C \vec{Z} \quad (6.159)$$

where  $C$  is a constraint matrix that relates the beam-displacement coordinates.

Similar to the restrained beam, the kinetic energy of the beam can be written as

$$T = \frac{1}{2} \left( m_1 \dot{Z}_1^2 + m_2 \dot{Z}_2^2 + m_3 \dot{Z}_3^2 \right) = \frac{1}{2} \begin{bmatrix} \dot{Z}_1 \\ \dot{Z}_2 \\ \dot{Z}_3 \end{bmatrix}^T \begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix} \begin{bmatrix} \dot{Z}_1 \\ \dot{Z}_2 \\ \dot{Z}_3 \end{bmatrix} = \frac{1}{2} \dot{\vec{Z}}^T M \dot{\vec{Z}} \quad (6.160)$$

and the potential energy of the beam can be written as

$$U = \frac{1}{2} k \theta^2 = \frac{1}{2} \theta^T k \theta = \frac{1}{2} \vec{Z}^T C^T k C \vec{Z} = \frac{1}{2} \vec{Z}^T K_c \vec{Z} \quad (6.161)$$

where  $K_c$  is the **constrained stiffness matrix**. Then, again using Euler-Lagrange's equation, one has

$$M \ddot{\vec{Z}} + K_c \dot{\vec{Z}} = \vec{0} \quad (6.162)$$

or defining  $D_c = M^{-1} K_c$  as the **constrained dynamic matrix**, one has

$$\ddot{\vec{Z}} + D_c \dot{\vec{Z}} = \vec{0} \quad (6.163)$$

Continuing with the modal analysis one has

$$\lambda_i \vec{v}_i = D_c \vec{v}_i \quad (6.164)$$

or for two eigenvalues/eigenvectors

$$\lambda_i M \vec{v}_i = K_c \vec{v}_i \quad (6.165)$$

$$\lambda_j M \vec{v}_j = K_c \vec{v}_j \quad (6.166)$$

and repeating the previous equations as  $M$  and  $K_c$  are still symmetric, one has

$$(\lambda_i - \lambda_j) \vec{v}_j^T M \vec{v}_i = 0 \quad (6.167)$$

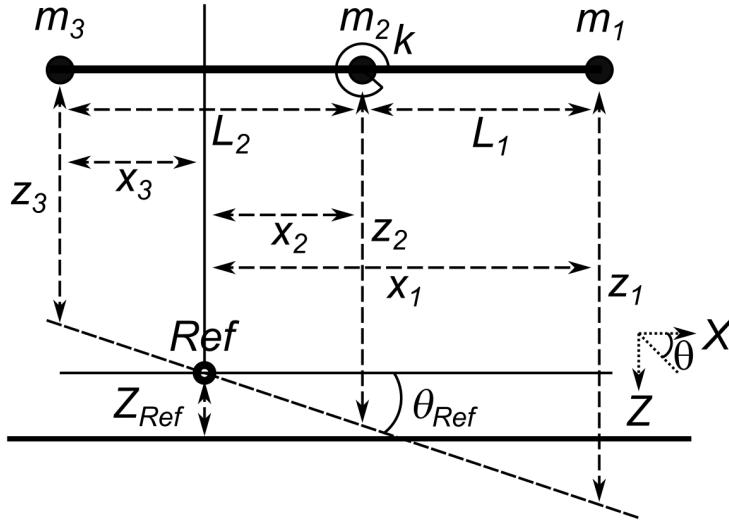
and

$$(\lambda_i - \lambda_j) \vec{v}_j^T K_c \vec{v}_i = 0 \quad (6.168)$$

However, for the unrestrained beam, two of the eigenvalues of  $D$  can be shown to be 0 and hence equal due to the existence of two rigid-body degrees-of-freedom (DOF), vertical translation and rotation of the *entire* beam. Thus, this system will have two rigid-body modes and a single vibration mode corresponding to the non-zero eigenvalue and associated eigenvector.

Thus, further work must be done to describe the entire elastic-body motion in terms of mutually orthogonal or **normal** modes. From linear algebra, if a matrix has repeated eigenvalues, *any* linear combination of the eigenvectors associated with the repeated eigenvalues are also eigenvectors of the given matrix. This fact can be used to obtain mutually orthogonal modes for unrestrained bodies.

Before this is derived further, consider an alternate approach that will consider the rigid-body degrees-of-freedom more directly, in particular consider the total, i.e., inertial, vertical position of the masses are referenced to some arbitrary reference axis, as demonstrated in the following three-lumped-mass example



where  $Ref$  is the arbitrary reference point for the positional coordinates. Here the total vertical displacements of the lumped-masses with the small angle approximation are

$$\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} = \begin{bmatrix} Z_{Ref} \\ Z_{Ref} \\ Z_{Ref} \end{bmatrix} + \begin{bmatrix} x_1 \\ x_2 \\ -x_3 \end{bmatrix} \theta_{Ref} + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \quad \vec{Z} = \vec{1} Z_{Ref} + \vec{x} \theta_{Ref} + \vec{z} \quad (6.169)$$

Using Lagrange's equation for the equations of motion, the kinetic energy can be written as

$$T = \frac{1}{2} \dot{\vec{Z}}^T M \dot{\vec{Z}} \quad (6.170)$$

and the potential (or strain) energy can be written as

$$U = \frac{1}{2} \theta^T k \theta \quad (6.171)$$

As before, one can model the angular displacement using the small angle approximation as

$$\theta = \frac{z_1 - z_2}{x_1 - x_2} - \frac{z_2 - z_3}{x_2 + x_3} = \left[ \frac{1}{x_1 - x_2} \quad \left( -\frac{1}{x_1 - x_2} - \frac{1}{x_2 + x_3} \right) \quad \frac{1}{x_2 + x_3} \right] \vec{z} = C \vec{z} \quad (6.172)$$

Thus,

$$U = \frac{1}{2} \vec{z}^T C^T k C \vec{z} = \frac{1}{2} \vec{z}^T K_c \vec{z} \quad (6.173)$$

where it should be noted that  $T$  is a function of the inertial velocities, while  $U$  is a function of the relative displacements. This will need to be resolved before using Lagrange's equation on the generalized coordinates.

Next, two additional constraints must be imposed. First, in the absence of external forces and moments, the Newton-Euler equations of motion dictate that both the translational and rotational momenta must be constant. Taking this arbitrary constant to zero results in two constraints

$$m_1 \dot{Z}_1 + m_2 \dot{Z}_2 + m_3 \dot{Z}_3 = \vec{1}^T M \dot{\vec{Z}} = 0 \quad (6.174)$$

and

$$m_1 x_1 \dot{Z}_1 + m_2 x_2 \dot{Z}_2 - m_3 x_3 \dot{Z}_3 = \vec{x}^T M \dot{\vec{Z}} = 0 \quad (6.175)$$

These two constraints imply that  $\dot{\vec{Z}}$  must be orthogonal (with respect to  $M$ ) to the vectors  $\vec{1}$  and  $\vec{x}$ . Therefore, one can define  $\vec{Z}_c$  and  $\vec{z}$  as the absolute and relative vertical displacement velocity which satisfy these constraints, respectively. Furthermore, from this analysis one can infer that  $\vec{1}$  and  $\vec{x}$  may be appropriate rigid-body mode shapes. If so, then these two shapes must be mutually orthogonal (with respect to  $M$ ), i.e.

$$m_1 x_1 + m_2 x_2 - m_3 x_3 = \vec{1}^T M \vec{x} = 0 \quad (6.176)$$

which is equivalent to setting the reference point,  $Ref$ , at the center of mass of the beam, point  $G$ . Lastly, recall that the total mass of the beam can be written as

$$M_{tot} = \vec{1}^T M \vec{1} \quad (6.177)$$

and the moment of inertia of the beam about its center of mass,  $G$ , can be written as

$$I_G = \sum_{i=1}^3 m_i x_i^2 = \vec{x}^T M \vec{x} \quad (6.178)$$

Then, rewriting Equation 6.169 in terms of constrained displacements, one has

$$\vec{Z}_c = \vec{1} Z_{Ref} + \vec{x} \theta_{Ref} + \vec{z}_c \quad (6.179)$$

from which if one can invoke the constraints and relative motion  $\vec{z}_c$  in terms of mutually orthogonal modal responses, then the desired solution to the vibration problem will be derived.

To that end, differentiating Equation 6.179 with respect to time and using the momenta constraints, one has

$$\vec{1}^T M \dot{\vec{Z}}_c = \vec{1}^T M [\vec{1} \dot{Z}_{Ref} + \vec{x} \dot{\theta}_{Ref} + \dot{\vec{z}}_c] = 0 \quad (6.180)$$

and

$$\vec{x}^T M \dot{\vec{Z}}_c = \vec{x}^T M [\vec{1} \dot{Z}_{Ref} + \vec{x} \dot{\theta}_{Ref} + \dot{\vec{z}}_c] = 0 \quad (6.181)$$

Then, noting the total mass and moment of inertia equations and center of mass constraint, one has

$$\dot{Z}_{Ref} = -\frac{1}{M_{tot}} \vec{1}^T M \dot{\vec{z}}_c \quad (6.182)$$

and

$$\dot{\theta}_{Ref} = \frac{1}{I_G} \vec{x}^T M \dot{\vec{z}}_c \quad (6.183)$$

Finally, applying these two constraints, one can write that constrained total velocities as functions of the constrained relative velocities as

$$\dot{\vec{Z}}_c = \left( I_{3 \times 3} - \frac{1}{M_{tot}} \vec{1} \vec{1}^T M - \frac{1}{I_G} \vec{x} \vec{x}^T M \right) \dot{\vec{z}}_c = \Xi \dot{\vec{z}}_c \quad (6.184)$$

Then, the kinetic energy in terms of the constrained relative velocities can be written as

$$T = \frac{1}{2} \dot{\vec{z}}_c^T \Xi^T M \Xi \dot{\vec{z}}_c = \frac{1}{2} \dot{\vec{z}}_c^T M_c \dot{\vec{z}}_c \quad (6.185)$$

and the potential (or strain) energy in terms of the constrained relative displacements can be written as

$$U = \frac{1}{2} \vec{z}_c^T K_c \vec{z}_c \quad (6.186)$$

Then utilizing  $\vec{z}_c$  as the generalized coordinates  $\vec{q}$  in Lagrange's equation in vector form, i.e.

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{\vec{q}}} \right) - \frac{\partial T}{\partial \vec{q}} + \frac{\partial U}{\partial \vec{q}} = \vec{0}^T \quad (6.187)$$

one has for the unrestrained beam-vibration equations of motion

$$M_c \ddot{\vec{z}}_c + K_c \vec{z}_c = \vec{0} \quad (6.188)$$

where the constrained mass matrix  $M_c$  is now singular and its inverse does not exist. Thus, to solve for the mode shapes and vibration frequencies, one must solve the **generalized eigenvalue problem**, i.e.

$$(\lambda_i M_c - K_c) \vec{v}_i = 0 \quad (6.189)$$

## Modal Analysis of Generalized Eigenvalue Problem

With the unrestrained beam, the modal analysis of the generalized eigenvalue problem will provide a model for incorporating vibration modes into the rigid-body dynamics to produce elastic-body dynamics. Recall the generalized eigenvalue problem from the previous section

$$\lambda_i M_c \vec{v}_i = K_c \vec{v}_i \quad (6.190)$$

So, as before, consider two  $i, j$  pairs of generalized eigenvalues and eigenvectors

$$(\lambda_i - \lambda_j) \vec{v}_i^T M_c \vec{v}_j = 0 \quad (6.191)$$

However, when  $i \neq j$  and the eigenvalues are distinct, e.g. not both zero, this equation assures that the two associated eigenvectors are orthogonal with respect to the *constrained* mass matrix, not the mass matrix  $M$ , as required. From the definition of  $M_c$ , one can write

$$(\lambda_i - \lambda_j) \vec{v}_i^T \Xi^T M \Xi \vec{v}_j = 0 \quad (6.192)$$

or defining the **constrained eigenvectors** as

$$\vec{v}_{c,i} = \Xi \vec{v}_i \quad (6.193)$$

Thus, when  $i \neq j$  and the two eigenvalues are distinct,  $\vec{v}_{c,i}$  are mutually orthogonal with respect to the mass matrix  $M$ . Note that for the unrestrained beam example, two of the transformed generalized eigenvectors will be equal to  $\vec{0}$  due to the two rigid-body modes and one will satisfy the three orthogonal constraints, the single vibration mode shape  $\vec{v}_{vib}$ . For an  $n$ -lumped-mass beam, one can extend this to the relative displacement equation for  $n - 2$  vibration modes

$$\vec{z}_c(t) = \sum_{i=1}^{n-2} \vec{v}_{vib,i} A_i \cos(\omega_{vib,i} t + \Gamma_i) \quad (6.194)$$

Note that all *vibration* mode shapes will be mutually orthogonal with respect to the mass matrix. To prove that these are also orthogonal to the rigid-body mode shapes,  $\vec{l}$  and  $\vec{x}$ , first consider the relative motion  $\vec{z}_c$  as a function of the original eigenvectors.

$$\vec{z}_c(t) = \sum_{i=1}^n \vec{v}_i \eta_i(t) \quad (6.195)$$

Next, recall the orthogonality constraints for the  $\dot{\vec{Z}}_c$  relative to  $\vec{l}$  and  $\vec{x}$ , i.e.

$$\vec{l}^T M \dot{\vec{Z}}_c = 0 \quad (6.196)$$

and

$$\vec{x}^T M \dot{\vec{Z}}_c = 0 \quad (6.197)$$

as well as the relation

$$\dot{\vec{Z}}_c = \Xi \dot{\vec{z}}_c \quad (6.198)$$

Thus, one has

$$\vec{1}^T M \vec{Z}_c = \vec{1}^T M \Xi \vec{z}_c = \vec{1}^T M \Xi \sum_{i=1}^n \vec{v}_i \dot{\eta}_i(t) = 0 \quad (6.199)$$

and

$$\vec{x}^T M \vec{Z}_c = \vec{x}^T M \Xi \vec{z}_c = \vec{x}^T M \Xi \sum_{i=1}^n \vec{v}_i \dot{\eta}_i(t) = 0 \quad (6.200)$$

Finally, by inspection, this requires

$$\vec{1}^T M \Xi \vec{v}_i = \vec{1}^T M \vec{v}_{c,i} = 0 \quad \forall i \quad (6.201)$$

and

$$\vec{x}^T M \Xi \vec{v}_i = \vec{x}^T M \vec{v}_{c,i} = 0 \quad \forall i \quad (6.202)$$

These equations signify that all the constrained eigenvectors are orthogonal to  $\vec{x}$  with respect to  $M$ . Thus, these constrained eigenvectors must be the desired orthogonal vibration mode shapes. Furthermore, all these eigenvectors (including  $\vec{1}$  and  $\vec{x}$ ) are mutually orthogonal with respect to  $M$ .

The key idea for orthogonality among the modal decomposition is that the physical responses of the unrestrained beam can be expressed in terms of the linear combination of *mutually orthogonal* modes, i.e.

$$\vec{q}(t) = \sum_{i=1}^n \vec{v}_{c,i} \eta_i = \vec{1} \eta_n + \vec{x} \eta_{n-1} + \sum_{i=1}^{n-2} \vec{v}_{vib,i} \eta_i \quad (6.203)$$

where the rigid-body and vibration modes derived previously are also known as the **normal modes**. Then, the unrestrained three-lumped mass model had the following model for the vertical displacements (with the small angle approximation)

$$\begin{aligned} \vec{Z}(t) &= \vec{1} Z_{Ref}(t) + \vec{x} \theta_{Ref}(t) + \vec{z}_{vib}(t) \\ \vec{Z}(t) &= \vec{1} Z_{Ref}(t) + \vec{x} \theta_{Ref}(t) + \sum_{i=1}^{n-2} \vec{v}_{vib,i} \eta_i(t) \end{aligned} \quad (6.204)$$

Thus, with this consideration, one can form generalized coordinates for deriving the equation of motion for the vertical displacement of an elastic-body as

$$\begin{aligned} \vec{Z} &= [\vec{1} \quad \vec{x} \quad \vec{v}_{vib,1} \quad \cdots \quad \vec{v}_{vib,n-2}] \begin{bmatrix} Z_{Ref} \\ \theta_{Ref} \\ \eta_1 \\ \vdots \\ \eta_{n-2} \end{bmatrix} \\ &= \Psi \vec{q} \end{aligned} \quad (6.205)$$

Then, the kinetic energy of the beam can be rewritten

$$\begin{aligned} T &= \frac{1}{2} \vec{Z}^T M \vec{Z} = \frac{1}{2} \vec{q}^T \Psi^T M \Psi \vec{q} \\ &= \frac{1}{2} M_{tot} \vec{Z}_{Ref}^2 + \frac{1}{2} I_G \theta_{Ref}^2 + \frac{1}{2} \vec{\eta}_{vib}^T \mathcal{M}_{vib} \vec{\eta}_{vib} \\ &= \frac{1}{2} \vec{q}^T \mathcal{M} \vec{q} \end{aligned} \quad (6.206)$$

where the rigid-body kinetic energy terms and elastic kinetic energy terms notably linearly combine, the **generalized mass matrix** is

$$\mathcal{M} = \begin{bmatrix} M_{tot} & 0 & 0 \\ 0 & I_G & 0 \\ 0 & 0 & \mathcal{M}_{vib} \end{bmatrix}, \quad (6.207)$$

and

$$\vec{\eta}_{vib} = [\eta_1 \ \cdots \ \eta_{n-2}]^T. \quad (6.208)$$

The potential (or strain) energy can be rewritten as

$$U = \frac{1}{2} \vec{z}^T K_c \vec{z} = \frac{1}{2} \vec{\eta}_{vib}^T \Psi_{vib}^T K_c \Psi_{vib} \vec{\eta}_{vib} = \frac{1}{2} \vec{\eta}_{vib}^T \mathcal{K}_{vib} \vec{\eta}_{vib} \quad (6.209)$$

where  $\mathcal{K}_{vib}$  is the generalized stiffness matrix and  $\Psi_{vib}$  is the vibration modal matrix, i.e.

$$\Psi_{vib} = [\vec{v}_{vib,1} \ \cdots \ \vec{v}_{vib,n-2}] \quad (6.210)$$

Finally, using Lagrange's equation for these energy expressions, one has

$$\mathcal{M} \ddot{\vec{q}} + \mathcal{K} \vec{q} = \begin{bmatrix} M_{tot} & 0 & 0 \\ 0 & I_G & 0 \\ 0 & 0 & \mathcal{M}_{vib} \end{bmatrix} \ddot{\vec{q}} + \begin{bmatrix} 0 & 0 \\ 0 & \mathcal{K}_{vib} \end{bmatrix} \vec{q} = 0 \quad (6.211)$$

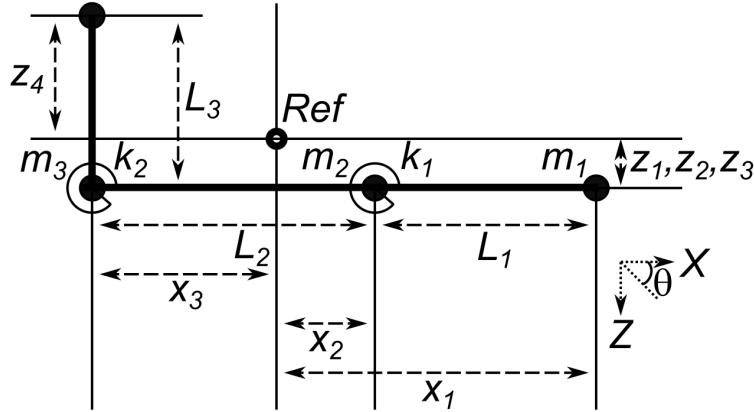
which can alternatively be written as

$$\begin{aligned} M_{tot} \ddot{Z}_{Ref} &= 0 \\ I_G \ddot{\theta}_{Ref} &= 0 \\ \mathcal{M}_{vib} \ddot{\vec{\eta}}_{vib} + \mathcal{K}_{vib} \vec{\eta}_{vib} &= 0 \end{aligned} \quad (6.212)$$

for the free response of the unrestrained beam's EOMs which will be fundamental for the elastic-body flight dynamics.

### Multi-Directional Elastic Motion

To extend these results to multi-directional motion requires more generalized vectors and matrices as each element of a mode shape/eigenvector can only correspond to direction of motion. To demonstrate this, consider the following bi-directional example of a unforced 2D truss.



where the directions of motion are  $X$  and  $Z$  and the reference point,  $Ref$  is the center of mass of the truss. The kinetic energy of the truss can be written as

$$\begin{aligned} T &= \frac{1}{2} [m_1(\dot{X}_1^2 + \dot{Z}_1^2) + m_2(\dot{X}_2^2 + \dot{Z}_2^2) + m_3(\dot{X}_3^2 + \dot{Z}_3^2) + m_4(\dot{X}_4^2 + \dot{Z}_4^2)] \\ &= \frac{1}{2} \begin{bmatrix} \dot{\vec{X}}^T & \dot{\vec{Z}}^T \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & M \end{bmatrix} \begin{bmatrix} \dot{\vec{X}} \\ \dot{\vec{Z}} \end{bmatrix} \end{aligned} \quad (6.213)$$

The potential (or strain) energy of the truss can be written as

$$U = \frac{1}{2} k_1 \theta_1^2 + \frac{1}{2} k_2 \theta_2^2 = \frac{1}{2} [\theta_1 \quad \theta_2] \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \quad (6.214)$$

where these two deflections  $\theta_1$  &  $\theta_2$  are the relative angular displacements between rods 1 and 2 & 2 and 3, respectively. From the geometry of the truss structure (and small angle approximations), one can form the following geometric constraints for these relative angular displacements as

$$\begin{aligned} \theta_1 &= \frac{Z_1 - Z_2}{x_1 - x_2} - \frac{Z_2 - Z_3}{x_2 + x_3} \\ &= \begin{bmatrix} \frac{1}{x_1 - x_2} & \left( \frac{-1}{x_1 - x_2} - \frac{1}{x_2 + x_3} \right) & \frac{1}{x_2 + x_3} & 0 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} \\ &= C_1 \vec{Z} \end{aligned} \quad (6.215)$$

and

$$\begin{aligned}\theta_2 &= \frac{Z_3 - Z_2}{x_2 + x_3} - \frac{X_3 - X_4}{z_3 + z_4} \\ &= \begin{bmatrix} 0 & 0 & \frac{-1}{z_3+z_4} & \frac{1}{z_3+z_4} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} + \begin{bmatrix} 0 & \left(\frac{-1}{x_2+x_3} - \frac{1}{x_2+x_3}\right) & \frac{1}{x_2+x_3} & 0 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} \\ &= [C_2 \quad C_3] \begin{bmatrix} \vec{X} \\ \vec{Z} \end{bmatrix}\end{aligned}\quad (6.216)$$

Furthermore, assuming equal vibration displacements for the colinear masses, one has

$$\begin{aligned}X_1 &= X_2 = X_3 \\ \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_3 \\ X_4 \end{bmatrix} \\ \vec{X} &= C'_x \vec{X}'\end{aligned}\quad (6.217)$$

$$\begin{aligned}Z_3 &= Z_4 \\ \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} \\ \vec{Z} &= C'_z \vec{Z}'\end{aligned}\quad (6.218)$$

where the degrees of freedom of the truss has been reduced using these constraint matrices, resulting in the modified coordinate vectors,  $\vec{X}'$  and  $\vec{Z}'$ .

With these constraints, the kinetic energy of the truss can be rewritten as

$$\begin{aligned}T &= \frac{1}{2} \begin{bmatrix} \dot{\vec{X}}'^T & \dot{\vec{Z}}'^T \end{bmatrix} \begin{bmatrix} C'_x & 0 \\ 0 & C'_z \end{bmatrix}^T \begin{bmatrix} M & 0 \\ 0 & M \end{bmatrix} \begin{bmatrix} C'_x & 0 \\ 0 & C'_z \end{bmatrix} \begin{bmatrix} \dot{\vec{X}}' \\ \dot{\vec{Z}}' \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} \dot{\vec{X}}'^T & \dot{\vec{Z}}'^T \end{bmatrix} [MM]' \begin{bmatrix} \dot{\vec{X}}' \\ \dot{\vec{Z}}' \end{bmatrix}\end{aligned}\quad (6.219)$$

and the potential (or strain) energy of the truss can be rewritten as

$$\begin{aligned}
 U &= \frac{1}{2} [\vec{X}^T \quad \vec{Z}^T] \begin{bmatrix} 0 & C_1 \\ C_2 & C_3 \end{bmatrix}^T \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix} \begin{bmatrix} 0 & C_1 \\ C_2 & C_3 \end{bmatrix} [\vec{X}] \\
 &= \frac{1}{2} [\vec{X}^T \quad \vec{Z}^T] K_c \begin{bmatrix} \vec{X} \\ \vec{Z} \end{bmatrix} \\
 &= \frac{1}{2} [\vec{X}'^T \quad \vec{Z}'^T] \begin{bmatrix} C'_x & 0 \\ 0 & C'_z \end{bmatrix}^T K_c \begin{bmatrix} C'_x & 0 \\ 0 & C'_z \end{bmatrix} [\vec{X}'] \\
 &= \frac{1}{2} [\vec{X}'^T \quad \vec{Z}'^T] K'_c \begin{bmatrix} \vec{X}' \\ \vec{Z}' \end{bmatrix}
 \end{aligned} \tag{6.220}$$

Finally, solving Lagrange's equation using these newly defined modified constrained mass matrix  $[MM]'$  and modified constrained stiffness matrix  $K'_c$ , one has

$$[MM]' \begin{bmatrix} \ddot{\vec{X}} \\ \ddot{\vec{Z}} \end{bmatrix} + K'_c \begin{bmatrix} \vec{X} \\ \vec{Z} \end{bmatrix} = 0 \tag{6.221}$$

which has a modified constrained dynamic matrix  $D'_c = [MM']^{-1}K'_c$ . Thus, the same modal analysis for mutually orthogonal modes and coordinates previously performed can also be done for multi-directional motion as long as the “proper” mass and stiffness matrices are used.

### Forced Motion and Virtual Work

To consider the addition of external forces on the lumped-mass systems, consider Lagrange's equation with external forces applied

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{\vec{q}}} \right) - \frac{\partial T}{\partial \vec{q}} + \frac{\partial U}{\partial \vec{q}} = \vec{Q} \tag{6.222}$$

where  $\vec{Q}$  is the generalized force, i.e.

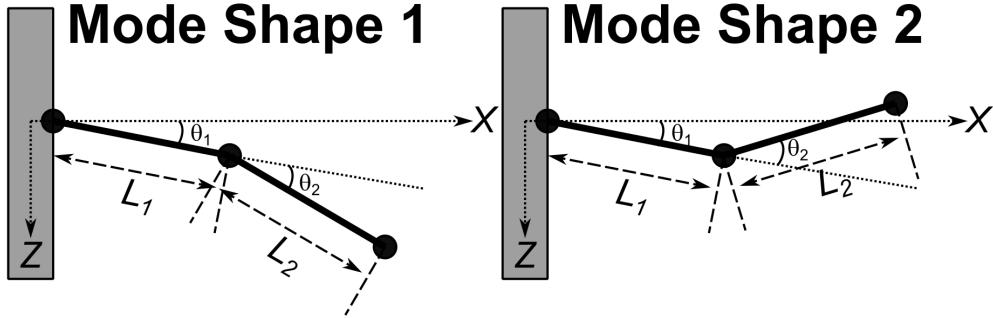
$$\vec{Q}^T = \frac{\partial \delta W}{\partial \delta \vec{q}} \tag{6.223}$$

where  $\delta \vec{q}$  is a virtual displacement of the generalized coordinates  $\vec{q}$  and the **virtual work** is defined as

$$\begin{aligned}
 \delta W &= \sum_{i=1}^m \vec{F}_i \cdot \delta \vec{d}_i \\
 \delta W &= [\vec{F}_1^T \quad \dots \quad \vec{F}_m^T]^T \begin{bmatrix} \delta \vec{d}_1 \\ \vdots \\ \delta \vec{d}_m \end{bmatrix}
 \end{aligned} \tag{6.224}$$

where  $\delta \vec{d}_i$  is the virtual displacement of the point of application of force  $F_i$ .

As an example, consider the vertical displacement of two-lumped-mass restrained beam shown here with its two mode shapes.



Defining  $\hat{k}$  as the unit vector for the  $Z$  direction, one can define the vertical forces as

$$\vec{F}_1 = F_1 \hat{k} \quad \vec{F}_2 = F_2 \hat{k} \quad (6.225)$$

and the virtual physical displacements at the points of application as

$$\begin{aligned} \delta \vec{d}_1 &= \delta Z_1 \hat{k} \\ \delta \vec{d}_2 &= \delta Z_2 \hat{k} \end{aligned} \quad (6.226)$$

Thus,

$$\begin{aligned} \delta W &= F_1 \delta Z_1 + F_2 \delta Z_2 \\ \delta W &= [F_1 \quad F_2] \begin{bmatrix} \delta Z_1 \\ \delta Z_2 \end{bmatrix} \end{aligned} \quad (6.227)$$

where

$$\begin{bmatrix} \delta Z_1 \\ \delta Z_2 \end{bmatrix} = \begin{bmatrix} L_1 & 0 \\ L_1 + L_2 & L_2 \end{bmatrix} \begin{bmatrix} \delta \theta_1 \\ \delta \theta_2 \end{bmatrix} \quad (6.228)$$

Now, assuming that the unforced (i.e. “free”) vibration problem has been solved in terms of  $\theta_1$  and  $\theta_2$ , then the elements of the free-vibration mode shapes would correspond to angular displacements and the virtual displacements  $\delta \theta_1$  and  $\delta \theta_2$  can be expressed in terms of these mode shapes and two vibration modal coordinates,  $\eta_1$  and  $\eta_2$  as

$$\begin{aligned} \begin{bmatrix} \delta \theta_1 \\ \delta \theta_2 \end{bmatrix} &= [\vec{v}_1 \quad \vec{v}_2] \begin{bmatrix} \delta \eta_1 \\ \delta \eta_2 \end{bmatrix} \\ \delta \vec{\theta} &= \Psi \delta \vec{\eta} \end{aligned} \quad (6.229)$$

Then, by substitution the virtual work can be written as

$$\begin{aligned} \delta W &= [F_1 \quad F_2] \begin{bmatrix} L_1 & 0 \\ L_1 + L_2 & L_2 \end{bmatrix} \Psi \delta \vec{\eta} \\ \delta W &= \vec{\mathcal{F}}^T \Psi \delta \vec{\eta} \end{aligned} \quad (6.230)$$

where  $\vec{\mathcal{F}}$  is the vector of applied forces relative to the virtual displacements. Finally, recall one can write the kinetic energy in terms of the modal coordinates as

$$T = \frac{1}{2} \dot{\theta}^T M \dot{\theta} = \frac{1}{2} \dot{\eta}^T \Psi^T M \Psi \dot{\eta} = \frac{1}{2} \dot{\eta}^T \mathcal{M} \dot{\eta} \quad (6.231)$$

and the potential (or strain) energy in terms of the modal coordinates as

$$U = \frac{1}{2} \vec{\theta}^T K \vec{\theta} = \frac{1}{2} \vec{\eta}^T \Psi^T K \Psi \vec{\eta} = \frac{1}{2} \vec{\eta}^T \mathcal{K} \vec{\eta} \quad (6.232)$$

Then, using Lagrange's equation with external forces applied results in

$$\mathcal{M} \ddot{\vec{\eta}} + \mathcal{K} \vec{\eta} = Q = \Psi^T \vec{\mathcal{F}} \quad (6.233)$$

Furthermore, for the unrestrained three-lumped-mass beam as described earlier, one can derive a similar expression which results in the form

$$\begin{aligned} M_{tot} \ddot{\vec{Z}}_{Ref} &= F_1 + F_2 + F_3 = \vec{1}^T \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} \\ I_G \ddot{\vec{\theta}}_{Ref} &= F_1 x_1 + F_2 x_2 - F_3 x_3 = \vec{x}^T \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} \\ \mathcal{M}_{vib} \ddot{\vec{\eta}}_{vib} + \mathcal{K}_{vib} \vec{\eta}_{vib} &= Q = \Psi_{vib}^T \vec{\mathcal{F}} = \vec{v}_{vib}^T \vec{\mathcal{F}} \end{aligned} \quad (6.234)$$

for the forced response of the unrestrained beam's EOMs which will be fundamental for the elastic-body flight dynamics. Recall that the *Ref* here is the center of mass of the lumped-mass system. This can easily be extended to *n*-lumped-mass systems.

## References

For more information, please refer to the following

- Schmidt, D. K., “Chapter 3 Structural Vibrations - A ‘Just-In-Time Tutorial,’” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 83-127

---

# General Aerospace Vehicle Dynamics

## 7.1 Aerospace Vehicle Reference Frames and Rotations

### Celestial and Earth-Centered Frames

Flight vehicles operate either in or outside of the atmospheres of celestial bodies which have their own gravitational motion. Although general relativity implies that there are no true inertial frames around gravitating bodies, the use of the **International Celestial Reference Frame (ICRF)** allows one to define the positions of astronomical objects as the ICRF is based on hundreds of extra-galactic radio sources, mostly quasars, distributed around the entire sky. The ICRF does not exhibit any measurable angular rotation since the extragalactic sources used to define the ICRF are so far away and appear stationary to any sensing technology. However, their positions can be measured very accurately by **Very Long Baseline Interferometry (VLBI)**. The positions of most are known to 1 milliarcsecond or better. Its origin is at the barycenter of the Solar System, with axes that are intended to “show no global rotation with respect to a set of distant extragalactic objects.”

In flight dynamics, one often desires to understand a vehicle’s motion relative to the celestial body whose gravitational force is forcing an orbit or resisting flight on the vehicle. As Earth is the most orbited celestial body, this textbook considers only specific **Earth-centered frames** which can be generalized for the sun-centered frame, also known as **heliocentric frame**, any other planetary-centered frame, or a moon-centered frame. These can be of two types, inertial or attached to the surface of the celestial body. The relationship between the ICRF and planetary-centered frames is related through the **ecliptic plane**, i.e., the plane of the planet’s orbit around Sun, and the planet’s equatorial plane, i.e., the plane containing the planet’s equator, whose normal direction coincides with planet’s average axis of rotation, not its instantaneous rotational axis because of the planet’s wobble. This axis passes through the planet’s surface at the point called the **true north pole** which is not the same as the magnetic north pole. With these planes in mind, one can define the planet’s **obliquity of ecliptic**,  $\epsilon$ , which is the angle between the normals of ecliptic and equatorial planes, e.g.,  $\epsilon_{Earth} \approx 23.4^\circ$ . In addition, one can define,  $\gamma$ , as the line of intersection between the planet’s equatorial plane and ecliptic plane on the vernal equinox, i.e., the first day of spring in the northern hemisphere, e.g.,

$\gamma_{Earth} = \Upsilon$  is the **first point of Aries**, also known as **vernal point**. Notably, and 4000 years ago,  $\gamma$  was located in constellation “Aries the Ram,” but now is in Pisces due to the precession of the equinoxes.

The first frame is an **Earth-centered, inertial (ECI) frame**, where a common standard is the **geocentric celestial reference frame (GCRF)** which has axes consistent with the ICRF and is denoted with the subscript  $I$ , which is useful for spacecraft orbiting around Earth. The GCRF frame is defined as

- The origin is the Earth’s center of mass.
- The  $x_I$ -axis coincides with the first point of Aries.
- The  $y_I$ -axis is orthogonal to both the  $x_I$ - and  $z_I$ -axes according to right-hand-rule.
- The  $z_I$ -axis passes through the **true north pole**.

The traditional name for the **spherical coordinates** used for the ECI frame are declination,  $\delta$ , right ascension,  $\alpha$ , and radius,  $r$ . An **ephemeris** is a table of  $(\alpha, \delta)$  coordinates of astronomical objects and artificial satellites as a function of time. These coordinates depends on location of vernal equinox at given time or epoch, typically the **J2000 epoch**, i.e., January 1, 2000, 12 h Universal time (UT). It should be noted that the axes of an ECI frame can be rotated w.r.t. the GCRF, e.g., the perifocal frame, which will be discussed later for orbital vehicles.

The second frame is an **Earth-centered, Earth-fixed (ECEF) frame** (subscript  $E$ ) is attached to the Earth’s surface, thereby it rotates with the Earth’s surface, i.e. the coordinates of a point on the surface of the Earth do not change. This frame is defined as

- The origin is the Earth’s center of mass.
- The  $x_E$ -axis passes through intersection of the prime meridian and the equator, located just south of west Africa.
- The  $y_E$ -axis is orthogonal to both  $x_E$ - and  $z_E$ -axes according to right-hand-rule located just south of India along the equator.
- The  $z_E$ -axis passes through **true north pole**.

The traditional name for the **spherical coordinates** used for the ECEF frame are **geocentric latitude**,  $\ell$ , **longitude**,  $\lambda$ , and **geocentric altitude**,  $h$  (**LLA**) where the geocentric altitude is defined relative to **Earth’s mean radius**,  $\bar{R}_E = 6,371,000$  m. This frame essentially models the Earth as a sphere for the purposes of the altitude and latitude coordinates. For the ECEF geocentric coordinates the **prime meridian** is defined as  $0^\circ$  geocentric longitude and the **equator** is defined as  $0^\circ$  geocentric latitude.

However, it should be noted that the surface of the Earth is not perfectly spherical. It is actually not a fully geometrically realizable shape at all and its surface and center of mass changes with time. Thus, the discipline of **geodesy** has developed which is the study of Earth’s shape. This science typically uses two different approximations for the Earth’s shape, the geoid and the reference ellipsoid. The **geoid** which is an idealized equilibrium surface of the Earth’s gravitational potential which varies according to its crust formation. A common geodesy standard for ECEF frames are **International Terrestrial Reference Frames (ITRFs)**.

A **reference ellipsoid** approximates the geoid and has an equatorial radius,  $R_e$ , longer than its polar radius,  $R_p$ , a shape also known as an **oblate spheroid**. The traditional name for the **ellipsoidal coordinates** used for the reference ellipsoid are **geodetic latitude**,  $\ell$ , **longitude**,  $\lambda$ , and **altitude**,  $h$  (**LLA**). The **prime meridian** is defined as  $0^\circ$  geodetic longitude. The **equator** is defined as  $0^\circ$  geodetic latitude. **Mean sea level (MSL)** is defined as 0 altitude and can be considered the ideal continuous surface of the ocean in the absence of currents and air pressure variations and whose surface continues under the continental masses. Notably, ellipsoidal coordinates are defined according to ellipsoidal trigonometry. Thus, because these are ellipsoidal as opposed to spherical, the latitude angle does not define the angle between the Earth's center and a point on the surface and altitude does not always intersect the Earth's center of mass, but both have an analytical offset dependent on the eccentricity of the ellipsoid. However, geodetic and geocentric longitude is identical.

A common model for the reference ellipsoid is the **World Geodetic System (WGS)**. The **WGS84 parameters** define the following constants for Earth. The ellipsoidal **equatorial radius** is given as

$$R_e = 6378137 \text{ m} \quad (7.1)$$

The ellipsoidal **flattening** is given as

$$f = \frac{1}{298.257223563} \quad (7.2)$$

The mean **Earth angular rate (EAR)** is given as

$$\omega_E = 7292115 \times 10^{-11} \text{ rad/s} \quad (7.3)$$

The **Earth gravitational parameter** for GPS is given as

$$\mu_E = 3.9860050 \times 10^{14} \text{ m}^2/\text{s}^2 \quad (7.4)$$

Notably, one can derive the ellipsoidal **polar radius** as

$$R_p = R_e(1 - f) = 6,356,752.3 \text{ m} \quad (7.5)$$

the ellipsoidal **eccentricity** of Earth is

$$e_E = \sqrt{f(2 - f)} = 0.081819190842622 \quad (7.6)$$

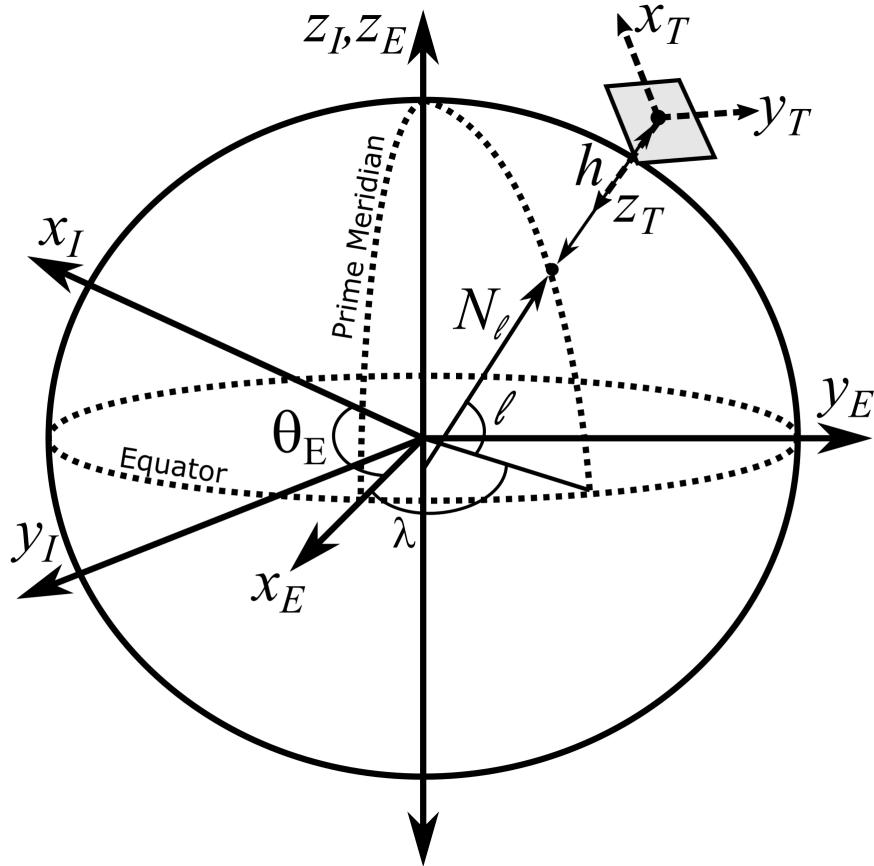
and the angular velocity of the ECEF frame relative to the ECI frame as

$$\omega_{E \leftarrow I, I} = \omega_{E \leftarrow I, E} = \begin{bmatrix} 0 \\ 0 \\ \omega_E \end{bmatrix} \quad (7.7)$$

An important aspect of the ECEF reference ellipsoid is the description of the **local tangent plane (LTP)** or the horizon plane at a reference latitude,  $\ell_0$ , longitude,  $\lambda_0$ , and altitude,  $h_0$ . By considering the LTP and its normal direction, a third frame is a **topocentric-horizon frame** (subscript  $T$ ) with its origin as a point on the Earth's surface and the coordinate axes are along the east-west direction, i.e., tangent to the latitude parallels, the north-south direction, i.e., tangent to the longitudinal meridians, and the up-down direction,

i.e., normal to the LTP. The choice for these must satisfy a right-hand rule reference frame, e.g., typical aerospace topocentric-horizon frames are north-east-down (NED) or east-north-up (ENU) for the  $x_T$ -axis,  $y_T$ -axis, and  $z_T$ -axis directions, respectively.

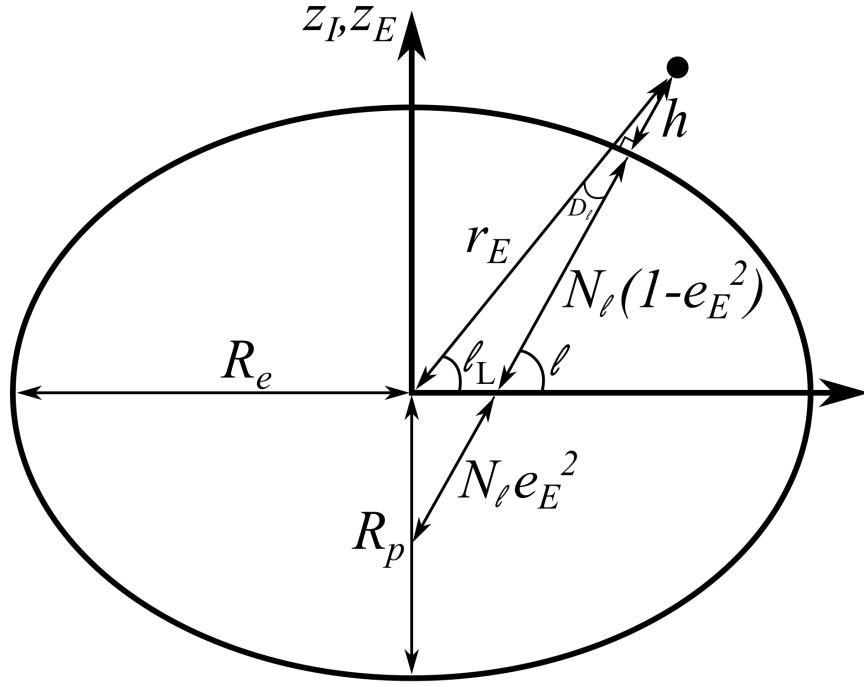
The relationship between the ECI frame, the ECEF frame, geodetic coordinates, the LTP, and the topocentric-horizon frame as shown in the following diagram.



where  $N_\ell$  is the **prime-vertical radius of curvature**, also known as the **transverse radius of curvature** or **east-west radius of curvature**, of the reference ellipsoid and depends on latitude

$$N_\ell = \frac{R_e}{\sqrt{1 - e_E^2 \sin^2 \ell}} \quad (7.8)$$

To further see the difference between geodetic and geocentric latitude, consider the following diagram.



where  $r_E = \bar{R}_e + h_E$  is the geocentric radius,  $N_\ell$  has notably been split into two convenient sections, and  $D_\ell$  is the **deviation of the normal** related to the geodetic and geocentric latitudes by

$$D_\ell = \ell - \ell_E \quad (7.9)$$

Lastly, using

$$n_\ell = \frac{e_E^2 N_\ell}{N_\ell + h} \quad (7.10)$$

one can show

$$\tan D_\ell = \frac{n_\ell \sin \ell \cos \ell}{1 - n_\ell \sin^2 \ell} = \frac{n_\ell \sin \ell_E \cos \ell_E}{1 - n_\ell \cos^2 \ell_E} \quad (7.11)$$

$$\sin \ell_E = \frac{(1 - n_\ell) \sin \ell}{\sqrt{1 - n_\ell(2 - n_\ell) \sin^2 \ell}} \quad (7.12)$$

$$\cos \ell_E = \frac{\cos \ell}{\sqrt{1 - n_\ell(2 - n_\ell) \sin^2 \ell}} \quad (7.13)$$

$$\tan \ell_E = (1 - n_\ell) \tan \ell \quad (7.14)$$

Lastly, the geocentric distance is related to the geodetic height via

$$r_E = (N_\ell + h) \sqrt{1 - n_\ell(2 - n_\ell) \sin^2 \ell} \quad (7.15)$$

## Vehicle-Centered Frames

The **vehicle-centered inertial frame** (subscript  $I^*$ ) is typically used for describing the attitude of a vehicle relative to the ICRF or ECI frame. This frame is defined as follows:

- The origin is the aerospace vehicle's center of mass.
- $x_{I^*}$ ,  $y_{I^*}$ , and  $z_{I^*}$ -axes are the same as the ICRF or obtained from a constant DCM transformation from the ICRF axes.

The **navigation frame** is typically used for the describing the long-term motion of the vehicle relative to the Earth's surface as represented by the reference ellipsoid. This frame is defined as follows:

- The origin is the aerospace vehicle's center of mass.
- $x_N$ ,  $y_N$ ,  $z_N$ -axes: the same as the instantaneous topocentric-horizon frame

Thus, the navigation frame is typically defined as either an **East-North-Up (ENU)** or a **North-East-Down (NED)** navigation frame and is only differentiated from the “instantaneous” topocentric-horizon frame by a constant offset vector along the  $z$ -axis direction. NED is traditionally used for aerospace vehicles as objects of interest are typically below aerospace vehicles and is used throughout this textbook with subscript  $N$ , while a ENU navigation frame will use a subscript  $ENU$ .

The **geodetic frame** (subscript  $G$ ) is an alternative to the navigation frame and is often used for orbiting satellites. This frame is defined as follows:

- The origin is the aerospace vehicle's center of mass.
- $x_G$ -axis is orthogonal to  $y_G$ - and  $z_G$ -axes forming right-handed coordinate frame.
- $y_G$ -axis is directed along the angular momentum vector of the aerospace vehicle, e.g. orthogonal to the orbital plane.
- $z_G$ -axis is directed along the Down direction of instantaneous LTP, i.e., orthogonal to the LTP.

The nadir-pointing **local-vertical, local-horizontal (LVLH) frame** (subscript  $L$ ), also known as the **local-vertical, local-horizontal, normal (LVLH-N) frame** or the **Hill frame**. This frame is defined as follows:

- The origin is the aerospace vehicle's center of mass.
- The  $x_L$ -axis, also known as the **local-horizontal axis**, is directed along the nominal direction of motion, orthogonal to  $y_L$ - and  $z_L$ -axes forming a right-handed coordinate frame.
- The  $y_L$ -axis is directed along the angular momentum vector of the vehicle, e.g., orthogonal to the orbital plane.
- The  $z_L$ -axis, also known as the **local-vertical axis**, is directed towards the center of the Earth.

Notably, some sources use a zenith-pointing LVLH frame**local-vertical, local-horizontal frame!zenith-pointing**, i.e.,

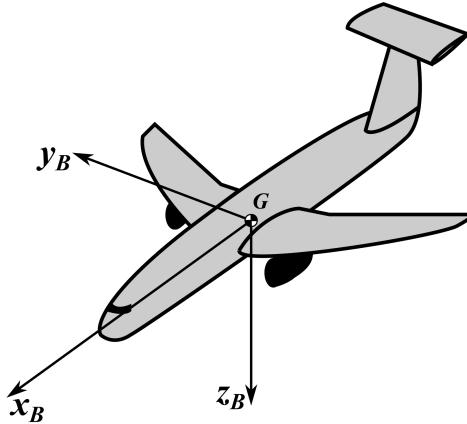
- The origin is the aerospace vehicle's center of mass.
- The  $x_L$ -axis, also known as the **local-horizontal axis**, is directed along the nominal direction of motion, orthogonal to  $x_L$ - and  $z_L$ -axes forming a right-handed coordinate frame.
- The  $y_L$ -axis, also known as the **local-vertical axis**, is directed away the center of the Earth.
- The  $z_L$ -axis, is directed along the angular momentum vector of the vehicle, e.g., orthogonal to the orbital plane.

However, this textbook will exclusively use the nadir-pointing LVLH frame.

The **body-fixed frame** (subscript  $B$ ) is attached to the rigid body of the aerospace vehicle, thus it is ideal for geometric configuration and structural modeling. This frame is defined as follows:

- The origin is the aerospace vehicle's center of mass.
- The  $x_B$  axis, known as the **longitudinal axis**, points out the front of the aerospace vehicle, typically "along" the nominal path of travel.
- The  $y_B$  axis, known as the **lateral axis**, points out the right side of the aerospace vehicle.
- The  $z_B$  axis points is orthogonal to  $x_B$ - and  $z_B$ -axes forming a right-handed coordinate frame.

This frame is depicted for an airplane in the following figure.



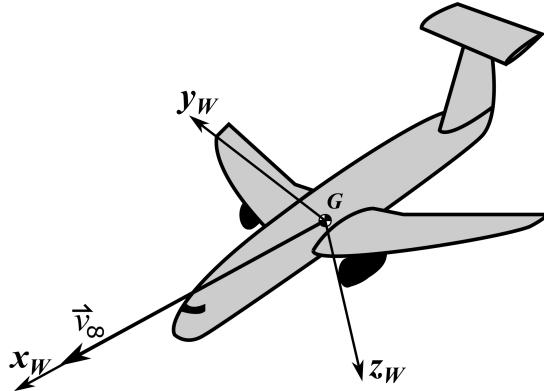
where point  $G$  is the vehicle's center of mass.

The **wind frame** (subscript  $W$ ) relates the **free-stream airflow** that an aerospace vehicle encounters as it flies, thus it is ideal for aerodynamics modeling and is particularly useful for aircraft dynamics and control. This frame is defined as follows:

- The origin is the aerospace vehicle's center of gravity.
- The  $x_W$  axis is colinear with vehicle's velocity relative to the air mass that the vehicle is traveling through, i.e., the **free-stream velocity**,  $\vec{v}_\infty$ .

- The  $z_W$  axis is in the plane of symmetry of the aerospace vehicle, positive below the aerospace vehicle.
- The  $y_W$  axis is orthogonal to both.

This frame is depicted for an airplane in the following figure.



The **velocity-turn-climb (VTC) frame** (subscript  $VTC$ ) relates the aerospace vehicle's velocity to the Earth's surface directly beneath the vehicle. This frame is defined as follows:

- The origin is the aerospace vehicle's center of gravity.
- The  $x_{VTC}$  axis, known as the **velocity axis**, is colinear with the velocity of the aerospace vehicle with respect to the Earth's surface, i.e., the **groundspeed** vector,  $\vec{v}_g$ .
- The  $y_{VTC}$  axis, known as the **turn axis**, is parallel to the instantaneous LTP, perpendicular to the velocity.
- The  $z_{VTC}$  axis, known as the **climb axis**, is orthogonal to  $x_B$ - and  $z_B$ -axes forming a right-handed coordinate frame

Thus, the unit vectors for the  $VTC$  frame can be defined

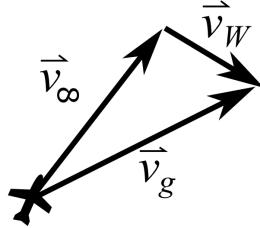
$$\begin{aligned}\vec{e}_{x_{VTC}} &= \frac{\vec{v}_g}{\|\vec{v}_g\|_2} \\ \vec{e}_{y_{VTC}} &= [\vec{e}_h] \times \vec{e}_{x_{VTC}} \\ \vec{e}_{z_{VTC}} &= [\vec{e}_{x_{VTC}}] \times \vec{e}_{y_{VTC}}\end{aligned}\tag{7.16}$$

where  $\vec{e}_h$  is the “Up” unit vector normal to the instantaneous LTP, e.g.  $[0 \ 0 \ 1]^T$  in ENU.

Importantly, the free-stream velocity,  $\vec{v}_\infty$ , is also known as the **airspeed** vector. The airspeed is different than the groundspeed vector if there is any **wind speed** vector,  $\vec{v}_w$ , i.e., the velocity of the air mass relative to the Earth's surface. These are related via the **wind triangle** equation

$$\vec{v}_g = \vec{v}_\infty + \vec{v}_w\tag{7.17}$$

which can be visualized in 2D as a triangle



### Aerospace Vehicle Reference Frame Rotations

The 3 Euler angle for describing the ECEF frame relative to the ECI frame is the Earth rotation angle (ERA),  $\theta_E$ . Thus, a vector expressed in ECI frame coordinates,  $\vec{v}_I$ , can be expressed as a vector in ECEF frame coordinates,  $\vec{v}_E$ , through the sequence

$$\vec{v}_E = C_3(\theta_E) \vec{v}_I \quad (7.18)$$

$$\vec{v}_E = C_{E \leftarrow I} \vec{v}_I \quad (7.19)$$

and the DCM from the ECI frame to the ECEF frame is given by

$$C_{E \leftarrow I} = \begin{bmatrix} \cos \theta_E & \sin \theta_E & 0 \\ -\sin \theta_E & \cos \theta_E & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (7.20)$$

Notably,  $\theta_E = \omega_E t$  where  $\omega_E$  is the **Earth angular rate (ERA)** and  $t$  is a reference time, e.g., J2000 epoch.

The 3 – 2 Euler angles for describing the NED navigation frame relative to the ECEF frame are the **longitude**  $\lambda$ , and, **geodetic latitude**,  $\ell$ . Thus, a vector expressed in vehicle-centered inertial frame coordinates,  $\vec{v}_E$ , can be expressed as a vector in NED navigation frame coordinates,  $\vec{v}_N$ , through the sequence

$$\vec{v}_N = C_2(-\ell - \pi/2) C_3(\lambda) \vec{v}_E \quad (7.21)$$

$$\vec{v}_N = C_{N \leftarrow E} \vec{v}_E \quad (7.22)$$

and the DCM from the ECEF frame to the NED navigation frame is given by

$$C_{N \leftarrow E} = C_2(-\ell - \pi/2) C_3(\lambda) = \begin{bmatrix} -\sin \ell \cos \lambda & -\sin \ell \sin \lambda & \cos \ell \\ -\sin \lambda & \cos \lambda & 0 \\ -\cos \ell \cos \lambda & -\cos \ell \sin \lambda & -\sin \ell \end{bmatrix} \quad (7.23)$$

Furthermore, by inspection of Figure ?? of the  $x_N$ - $z_N$  plane, the offset vector in the navigation frame can be broken into two vectors in the NED navigation frame denoted in gray as

$$\vec{o}_{E \leftarrow N, N} = \begin{bmatrix} N_\ell e_E^2 \cos \ell \sin \lambda \\ 0 \\ N_\ell e_E^2 \cos^2 \ell \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ N_\ell(1 - e_E^2) + h \end{bmatrix} \quad (7.24)$$

which can be combined into

$$\vec{o}_{E \leftarrow N, N} = \begin{bmatrix} N_\ell e_E^2 \cos \ell \sin \ell \\ 0 \\ N_\ell (1 - e_E^2 \sin^2 \ell) + h \end{bmatrix} \quad (7.25)$$

which can be transformed to ECEF coordinates and reversing the direction of the offset vector, one has

$$\vec{o}_{N \leftarrow E, E} = - \begin{bmatrix} -\sin \ell \cos \lambda & -\sin \lambda & -\cos \ell \cos \lambda \\ -\sin \ell \sin \lambda & \cos \lambda & -\cos \ell \sin \lambda \\ \cos \ell & 0 & -\sin \ell \end{bmatrix} \begin{bmatrix} N_\ell e_E^2 \cos \ell \sin \ell \\ 0 \\ (N_\ell + h) - N_\ell e_E^2 \sin^2 \ell \end{bmatrix} \quad (7.26)$$

multiplying and noting  $\sin^2 \ell = 1 - \cos^2 \ell$ , one has

$$\vec{o}_{N \leftarrow E, E} = \begin{bmatrix} N_\ell e_E^2 \cos \ell \sin^2 \ell \cos \lambda + (N_\ell + h) \cos \ell \cos \lambda - N_\ell e_E^2 \sin^2 \ell \cos \ell \cos \lambda \\ N_\ell e_E^2 \cos \ell \sin^2 \ell \sin \lambda + (N_\ell + h) \cos \ell \sin \lambda - N_\ell e_E^2 \sin^2 \ell \cos \ell \sin \lambda \\ N_\ell e_E^2 \cos^2 \ell \sin \ell + (N_\ell + h) \sin \ell - N_\ell e_E^2 (1 - \cos^2 \ell) \sin \ell \end{bmatrix} \quad (7.27)$$

or

$$\vec{o}_{N \leftarrow E, E} = \begin{bmatrix} (N_\ell + h) \cos \ell \cos \lambda \\ (N_\ell + h) \sin \lambda \cos \ell \\ (N_\ell (1 - e_E^2) + h) \sin \ell \end{bmatrix} \quad (7.28)$$

The  $3 - 1 - 3$  Euler angles for describing the body-fixed frame relative to the vehicle-centered, inertial frame are the **precession angle**,  $\phi_p$ , **nutation angle**,  $\theta_n$ , and **spin angle**,  $\psi_s$ , where the subscripts have been added to differentiate these Euler angles from the  $3 - 2 - 1$  navigation-to-body frame Euler angles, roll, pitch, and yaw. Thus, a vector expressed in vehicle-centered inertial frame coordinates,  $\vec{v}_{I*}$ , can be expressed as a vector in body-fixed frame coordinates,  $\vec{v}_B$ , through the sequence

$$\vec{v}_B = C_3(\psi_s) C_1(\theta_n) C_3(\phi_p) \vec{v}_{I*} \quad (7.29)$$

$$\vec{v}_B = C_{B \leftarrow I*} \vec{v}_{I*} \quad (7.30)$$

and the DCM for the body-fixed frame relative to the vehicle-centered, inertial frame can be computed via

$$C_{B \leftarrow I*} = \begin{bmatrix} \cos \phi_p \cos \psi_s - \cos \theta_n \sin \phi_p \sin \psi_s & \sin \phi_p \cos \psi_s - \cos \theta_n \cos \phi_p \sin \psi_s & \sin \theta_n \sin \psi_s \\ -\cos \phi_p \sin \psi_s - \cos \theta_n \sin \phi_p \cos \psi_s & -\sin \phi_p \sin \psi_s - \cos \theta_n \cos \phi_p \cos \psi_s & \sin \theta_n \cos \psi_s \\ \sin \theta_n \sin \psi_s & -\sin \theta_n \cos \psi_s & \cos \theta_n \end{bmatrix} \quad (7.31)$$

The  $3 - 2 - 1$  Euler angles for describing the body-fixed frame relative to the navigation or LVLH frame are the **roll angle**,  $\phi$ , **pitch angle**,  $\theta$ , and **yaw angle**,  $\psi$ . Thus, a vector expressed in NED navigation or nadir-pointing LVLH frame coordinates,  $\vec{v}_{N \text{ or } L}$ , can be expressed as a vector in body-fixed frame coordinates,  $\vec{v}_B$ , through the sequence

$$\vec{v}_B = C_1(\phi) C_2(\theta) C_3(\psi) \vec{v}_{N \text{ or } L} \quad (7.32)$$

$$\vec{v}_B = C_{B \leftarrow N \text{ or } L} \vec{v}_{N \text{ or } L} \quad (7.33)$$

and the DCM for the body-fixed frame relative to the navigation or nadir-pointing LVLH frame can be computed via

$$C_{B \leftarrow N \text{ or } L} = \begin{bmatrix} \cos \theta \cos \psi & \cos \theta \sin \psi & -\sin \theta \\ \sin \phi \sin \theta \cos \psi - \cos \phi \sin \psi & \sin \phi \sin \theta \sin \psi + \cos \phi \cos \psi & \sin \phi \cos \theta \\ \cos \phi \sin \theta \cos \psi + \sin \phi \sin \psi & \cos \phi \sin \theta \sin \psi - \sin \phi \cos \psi & \cos \phi \cos \theta \end{bmatrix} \quad (7.34)$$

These Euler angles are used to represent the orientation or **attitude** of the aircraft as they provide a description of how the rigid body of the aircraft is currently oriented in 3D space. It should be noted that the yaw angle can be arbitrarily set as subsequent rotations in yaw can be performed before the other angles.

The  $3 - 2 - 1$  Euler angles for describing the wind frame relative to the navigation or LVLH frame are the **bank angle**,  $\mu$ , **flight-path angle**,  $\gamma$ , and **heading angle**,  $\sigma$ . Thus, a vector expressed in NED navigation frame coordinates,  $\vec{v}_N$  or  $L$ , can be expressed as a vector in wind frame coordinates,  $\vec{v}_W$ , through the sequence

$$\vec{v}_W = C_1(\mu)C_2(\gamma)C_3(\sigma)\vec{v}_{N \text{ or } L} \quad (7.35)$$

$$\vec{v}_W = C_{W \leftarrow N \text{ or } L}\vec{v}_{N \text{ or } L} \quad (7.36)$$

and the DCM for the wind frame relative to the navigation or LVLH frame can be computed via

$$C_{W \leftarrow N \text{ or } L} = \begin{bmatrix} \cos \gamma \cos \sigma & \cos \gamma \sin \sigma & -\sin \gamma \\ \sin \mu \sin \gamma \cos \sigma - \cos \mu \sin \sigma & \sin \mu \sin \gamma \sin \sigma + \cos \mu \cos \sigma & \sin \mu \cos \gamma \\ \cos \mu \sin \gamma \cos \sigma + \sin \mu \sin \sigma & \cos \mu \sin \gamma \sin \sigma - \sin \mu \cos \sigma & \cos \mu \cos \gamma \end{bmatrix} \quad (7.37)$$

It should be noted that the heading angle can be arbitrarily set as subsequent rotations in heading can be performed before the other navigation-to-body-fixed frame Euler angles, but is typically referenced to north.

The  $3 - 2$  Euler angles for describing the body-fixed frame relative to the wind frame are the **angle of attack**,  $\alpha$ , and the **sideslip angle**,  $\beta$ . Thus, a vector expressed in wind frame coordinates,  $\vec{v}_W$ , can be expressed as a vector in body-fixed frame coordinates,  $\vec{v}_B$ , through the sequence

$$\vec{v}_B = C_2(\alpha)C_3(-\beta)\vec{v}_W \quad (7.38)$$

$$\vec{v}_B = C_{B \leftarrow W}\vec{v}_W \quad (7.39)$$

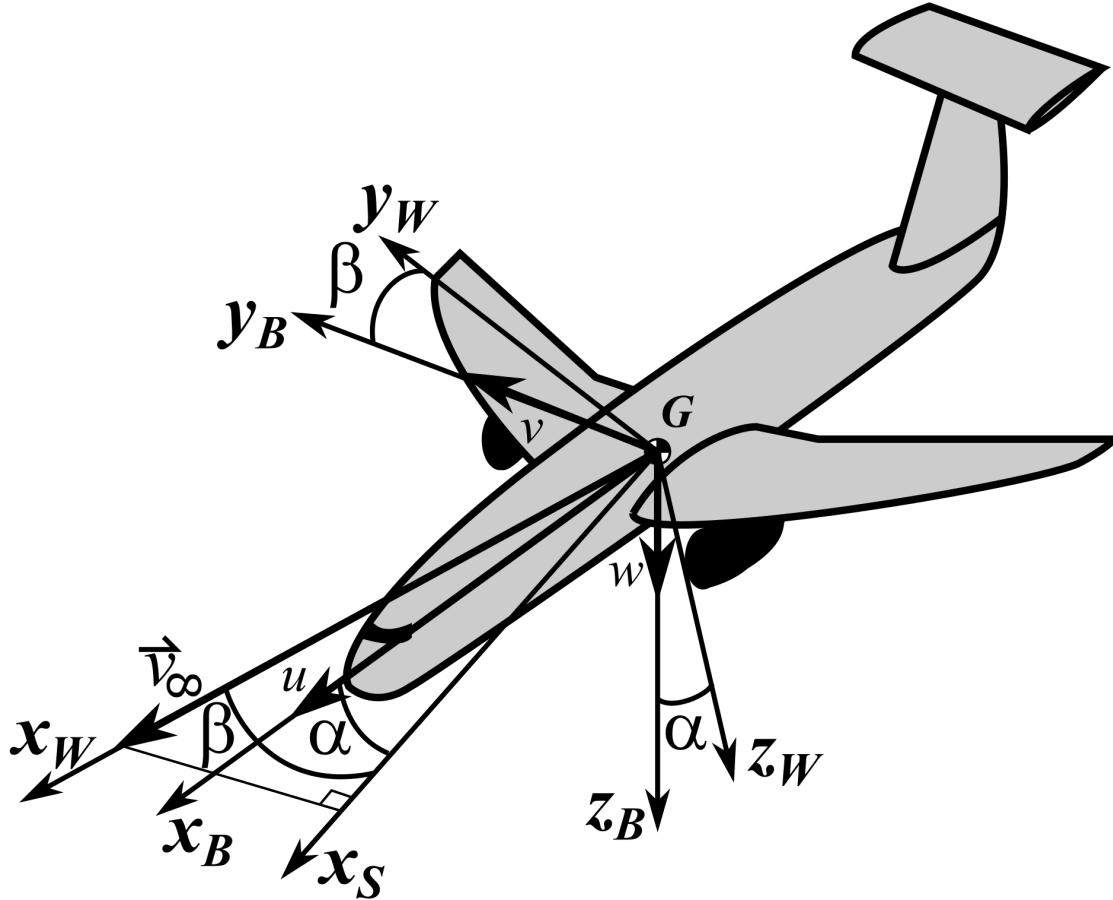
and the DCM for the body-fixed frame relative to the wind frame can be computed via

$$\vec{v}_B = \begin{bmatrix} \cos \alpha \cos \beta & -\cos \alpha \sin \beta & -\sin \alpha \\ \sin \beta & \cos \beta & 0 \\ \sin \alpha \cos \beta & -\sin \alpha \sin \beta & \cos \alpha \end{bmatrix} \vec{v}_W \quad (7.40)$$

Of particular importance for this rotation is the representation of the airspeed vector,  $\vec{v}_\infty$ , with magnitude  $v_\infty$  expressed in the body-fixed frame as the components

$$\vec{v}_{\infty, B} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} \quad (7.41)$$

which can be visualized in the following diagram



where  $x_S$  is the  $x$ -axis of the **stability frame** and will be discussed in more detail in rigid-body airplane dynamics.

From the previous equations, one has

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} \cos \alpha \cos \beta & -\cos \alpha \sin \beta & -\sin \alpha \\ \sin \beta & \cos \beta & 0 \\ \sin \alpha \cos \beta & -\sin \alpha \sin \beta & \cos \alpha \end{bmatrix} \begin{bmatrix} v_\infty \\ 0 \\ 0 \end{bmatrix} \quad (7.42)$$

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} v_\infty \cos \alpha \cos \beta \\ v_\infty \sin \beta \\ v_\infty \sin \alpha \cos \beta \end{bmatrix} \quad (7.43)$$

Dividing the first row by the third row, one has

$$\frac{u}{w} = \frac{v_\infty \cos \alpha \cos \beta}{v_\infty \sin \alpha \cos \beta} \quad (7.44)$$

$$\frac{u}{w} = \frac{\cos \alpha}{\sin \alpha} \quad (7.45)$$

$$\frac{u}{w} = \frac{1}{\tan \alpha} \quad (7.46)$$

which provides the relationship between the angle of attack and the components of the airspeed vector as

$$\alpha = \tan^{-1} \frac{w}{u} \quad (7.47)$$

While isolating the second row, one has

$$v = v_\infty \sin \beta \quad (7.48)$$

which provides the relationship between the sideslip angle and the components of the airspeed vector as

$$\beta = \sin^{-1} \frac{v}{v_\infty} \quad (7.49)$$

Moreover, it should be pointed out that there is an implicit relationship between these sets of Euler angles, e.g.

$$C_{W \leftarrow N \text{ or } L} = C_{W \leftarrow B} C_{B \leftarrow N \text{ or } L} \quad (7.50)$$

which, in general, results in complicated trigonometric equations.

The 3 – 2 Euler angles for describing the VTC frame relative to the navigation frame are the **elevation angle**,  $\epsilon$ , and **heading angle**,  $\sigma$ . Thus, a vector expressed in ENU navigation frame coordinates,  $\vec{v}_{ENU}$ , can be expressed as a vector in wind frame coordinates,  $\vec{v}_{VTC}$ , through the sequence

$$\vec{v}_{VTC} = C_2(\gamma) C_3(\sigma) \vec{v}_{ENU} \quad (7.51)$$

$$\vec{v}_{VTC} = C_{VTC \leftarrow N} \vec{v}_{ENU} \quad (7.52)$$

and the DCM for the wind frame relative to the navigation frame can be computed via

$$C_{VTC \leftarrow ENU} = \begin{bmatrix} \cos \epsilon \cos \sigma & \cos \epsilon \sin \sigma & -\sin \epsilon \\ -\cos \epsilon & \sin \sigma & 0 \\ \sin \epsilon \cos \sigma & \sin \epsilon \sin \sigma & \cos \gamma \end{bmatrix} \quad (7.53)$$

Of particular importance for this rotation is the representation of the groundspeed vector,  $\vec{v}_g$ , with magnitude  $\|\vec{v}_g\|_2$  expressed in the ENU navigation frame as the components

$$\vec{v}_{g,ENU} = \begin{bmatrix} u_g \\ v_g \\ w_g \end{bmatrix} \quad (7.54)$$

From the previous equations, one has

$$\begin{bmatrix} u_g \\ v_g \\ w_g \end{bmatrix} = \begin{bmatrix} \cos \epsilon \cos \sigma & -\cos \epsilon & \sin \epsilon \cos \sigma \\ \cos \epsilon \sin \sigma & \sin \sigma & \sin \epsilon \sin \sigma \\ -\sin \epsilon & 0 & \cos \gamma \end{bmatrix} \begin{bmatrix} \|\vec{v}_g\|_2 \\ 0 \\ 0 \end{bmatrix} \quad (7.55)$$

$$\begin{bmatrix} u_g \\ v_g \\ w_g \end{bmatrix} = \begin{bmatrix} \|\vec{v}_g\|_2 \cos \epsilon \cos \sigma \\ \|\vec{v}_g\|_2 \cos \epsilon \sin \sigma \\ -\|\vec{v}_g\|_2 \sin \epsilon \end{bmatrix} \quad (7.56)$$

Dividing the first row by the second row, one has

$$\frac{u_g}{v_g} = \frac{\|\vec{v}_g\|_2 \cos \epsilon \cos \sigma}{\|\vec{v}_g\|_2 \cos \epsilon \sin \sigma} \quad (7.57)$$

$$\frac{u_g}{v_g} = \frac{\cos \sigma}{\sin \sigma} \quad (7.58)$$

$$\frac{u_g}{v_g} = \frac{1}{\tan \sigma} \quad (7.59)$$

which provides the relationship between the heading angle and the components of the groundspeed vector as

$$\sigma = \tan^{-1} \frac{v_g}{u_g} \quad (7.60)$$

While isolating the third row, one has

$$w_g = -v_g \sin \epsilon \quad (7.61)$$

which provides the relationship between the elevation angle and the components of the groundspeed vector as

$$\epsilon = \sin^{-1} \frac{-w_g}{\|\vec{v}_g\|_2} \quad (7.62)$$

## References

For more information, please refer to the following

- Curtis, H. D., “9.9 Euler Angles,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 448-459
- Curtis, H. D., “9.10 Yaw, Pitch, and Roll Angles,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 459-463
- Nelson, R. C., “3.3 Orientation and Position of the Airplane,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 101-103
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “1.4 Rotational Kinematics,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 16-20
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “1.6 Geodesy, Coordinate Systems, Gravity,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 23-34

## 7.2 Point-Mass Aerospace Vehicle Dynamics

Recall in classical mechanics, for any point-mass in a rotating body-fixed frame, the Newton EOMs are

$$\sum \vec{F}_B = m \begin{bmatrix} \dot{u} + qw - rv \\ \dot{v} + ru - pw \\ \dot{w} + pv - qu \end{bmatrix} \quad (7.63)$$

where for aerospace vehicles the primary forces to consider are the gravitational,  $\vec{F}_g$ , propulsive,  $\vec{F}_p$ , aerodynamic,  $\vec{F}_a$ , and radiation pressure,  $\vec{F}_r$ , i.e.,

$$\sum \vec{F} = \vec{F}_g + \vec{F}_p + \vec{F}_a + \vec{F}_r \quad (7.64)$$

whereas ground vehicles and water vehicles experience ground forces and hydrodynamic forces, respectively.

Fundamentally, aerospace vehicles always encounter gravitational forces. **Newton's law of gravitation** for point and spherical celestial bodies states that the **gravitational force**,  $F_g$ , also known as the **weight**,  $W$ , is a function of the distance  $r$  between the centers of mass for the vehicle  $m$  and a celestial body  $M$ , i.e.

$$F_g = W = \frac{GMm}{r^2} = \frac{\mu m}{r^2} \quad (7.65)$$

where  $G$  is the gravitational constant, i.e.  $6.674 \times 10^{-11} \text{ m}^3 \text{kg}^{-1} \text{s}^{-2}$  and  $\mu = GM$  is the **standard gravitational parameter** of the celestial body. It should be noted that though a celestial body's gravity field typically also contains variation, but this may be neglected in the dynamics modeling in the design of different aerospace vehicle systems.

A high-fidelity ellipsoidal-Earth gravity model can be obtained by expanding Newton's law of gravitation to include the **first zonal harmonic** of the Earth's gravitation field,  $J_2$ , due to the ellipsoidal nature of the geoid, which yields

$$\vec{G}_{E \text{ or } I} = -\frac{\mu_E}{\|\vec{x}_{E \text{ or } I}\|_2^3} \begin{bmatrix} \left(1 + \frac{3}{2} \left(\frac{R_e}{\|\vec{x}_{E \text{ or } I}\|_2}\right)^2 J_2 (1 - 5 \sin^2 \ell)\right) x_{E \text{ or } I} \\ \left(1 + \frac{3}{2} \left(\frac{R_e}{\|\vec{x}_{E \text{ or } I}\|_2}\right)^2 J_2 (1 - 5 \sin^2 \ell)\right) y_{E \text{ or } I} \\ \left(1 + \frac{3}{2} \left(\frac{R_e}{\|\vec{x}_{E \text{ or } I}\|_2}\right)^2 J_2 (3 - 5 \sin^2 \ell)\right) z_{E \text{ or } I} \end{bmatrix} \quad (7.66)$$

where  $\vec{x}_{E \text{ or } I} = [x_{E \text{ or } I} \ y_{E \text{ or } I} \ z_{E \text{ or } I}]^T$  is the position of the flight vehicle's center of mass in the ECI or ECEF frame,  $R_e = 6378137.0 \text{ m}$  is Earth's equatorial radius, **Earth's gravitational parameter** is

$$\mu_E = GM_E = 3986004.418 \times 10^8 \text{ m}^3/\text{s}^2 \quad (7.67)$$

and the second-order term by the **WGS84  $J_2$  parameter**

$$J_2 = -\sqrt{5} \bar{C}_{2,0} = 0.001082626684 \quad (7.68)$$

where  $\bar{C}_{2,0}$  is the actual specified parameter in the WGS84 model. This is part of **Earth Gravitational Models (EGM)** published by the National Geospatial-Intelligence Agency (NGA).

In the ECI frame, this gravitational attraction is the gravitational force on the aerospace vehicle, i.e.,

$$\vec{F}_{g,I} = m \vec{g}_I = m \vec{G}_I \quad (7.69)$$

where  $\vec{g}$  is the **gravity vector** present in aerospace vehicle EOMs. In the ECEF frame, this gravitational attraction minus the centripetal acceleration of the rotating Earth makes up the gravitational force on the aerospace vehicle as

$$\vec{F}_{g,E} = m \vec{g}_E = m \left( \vec{G}_E - [\vec{\omega}_{E \leftarrow I}]_{\times} [\vec{\omega}_{E \leftarrow I}]_{\times} \vec{x}_E \right) \quad (7.70)$$

Notably, this ellipsoidal-Earth gravity model can be used to form a first-order approximation of gravity as a function of altitude  $h$  using average values for the derivatives of  $h$  and  $\ell$  with respect to  $\|\vec{x}_E\|_2$ , i.e., the distance from the center of the Earth in ECEF coordinates. This is typically the case for using the standard atmosphere models introduced later by choosing these derivatives for  $\ell = 45^\circ$ .

For aircraft dynamics, i.e.,  $h < 100$  km, a lower-fidelity ellipsoidal-Earth gravity model may suffice which assumes the gravity vector acts straight downward in the navigation, geodetic, or LVLH frame, i.e.

$$\vec{F}_{g,N \text{ or } G \text{ or } L} = m \vec{g}_{N \text{ or } G \text{ or } L} \approx \begin{bmatrix} 0 \\ 0 \\ mg(\ell, h) \end{bmatrix} \quad (7.71)$$

and will be a function of latitude  $\ell$  and altitude  $h$ . A model for the latitude variation of gravity at MSL is based the **WGS84 Ellipsoidal Gravity Formula**, a refinement of the **Somigliana gravity model**, defined as

$$g_0(\ell) = g_e \left( \frac{1 + k \sin^2 \ell}{\sqrt{1 - e^2 \sin^2 \ell}} \right) \quad (7.72)$$

where  $e^2 = 1 - (R_p/R_e)^2 \approx 0.00669437999014$  is the Earth's eccentricity squared,  $k = \frac{R_p g_p - R_e g_e}{R_e g_e} \approx 0.001931853$  is a formula constant,  $R_e = 6378137.0$  m is Earth's equatorial radius,  $R_p = 6356752.3$  m is Earth's polar radius,  $g_e \approx 9.7803253359$  m/s<sup>2</sup> is the acceleration due to gravity at the equator, and  $g_p = 9.8321849378$  m/s<sup>2</sup> is the acceleration due to gravity at the poles.

Then, one can include the ellipsoidal height dependence via the **Geodetic Reference System (GRS) 1967 model** as

$$g(\ell, h) = g_0(\ell) \left( 1 - (k_1 - k_2 \sin^2 \ell)h + 3k_3 h^2 \right) \quad (7.73)$$

where

$$\begin{aligned} k_1 &= \frac{2(1 + f + \frac{\omega_{E \leftarrow I}^2 R_e^2 R_p}{\mu_E})}{R_e} = 3.15704 \times 10^{-7} \text{ m}^{-1} \\ k_2 &= \frac{4f}{R_e} = 2.10269 \times 10^{-9} \text{ m}^{-1} \\ k_3 &= \frac{3}{R_e^2} = 7.37452 \times 10^{-14} \text{ m}^{-2} \end{aligned} \quad (7.74)$$

Alternatively, for a spherical-Earth or flat-Earth gravity model, one can model the variation of gravity with respect to the altitude *above* MSL, also known as the **free air correction (FAC)**, as

$$g(h) = g_0 \left( \frac{\bar{R}_E}{\bar{R}_E + h} \right)^2 \quad (7.75)$$

where  $g_0 = 9.80665 \text{ m/s}^2$  is the **standard acceleration due to gravity** and  $\bar{R}_E = 6,371,000 \text{ m}$  is **Earth's mean radius**. As a linear approximation for  $h \ll \bar{R}_E$ , one has

$$g(h) \approx g_0 - \frac{2g_0}{\bar{R}_E^2} h = g_0 - 3.086 \times 10^{-6} h \quad (7.76)$$

To overcome gravity and obtain sustained flight, aerospace vehicles must have propulsion systems which produce the propulsive forces and moments. These systems typically use either rotating propellers, rotors, and/or reaction engines, e.g. jet, rocket, or ion, which are affixed to the vehicle's structure. Thus, one typically defines the **propulsive force** vector,  $\vec{F}_p$ , and the **propulsive moment** vector,  $\vec{M}_p$ , in the body frame as a function of **thrust** vectors,  $\vec{T}$ . The propulsion system is primarily used to translate the entire vehicle; however, some propulsion systems are used to steer aerospace vehicles, a design known as **thrust vectoring**. This textbook will not consider propulsion system modeling in detail, but will assume that one can define a dynamical system model for the propulsion that may be dependent on the flight conditions, e.g. airspeed and air density, and can then be approximated by first- or second-order dynamics, similar to other actuators. It should also be noted that rotating propellers and rotors create a changing mass distribution due to their rotation.

The aerodynamic forces and moments are due to the air pressure distribution around the aerospace vehicle. The **aerodynamic force** vector,  $\vec{F}_a$ , resist/assist translation through the air mass while the **aerodynamic moment** vector,  $\vec{M}_a$ , cause the vehicle to rotate due to the varying pressure distribution over the body. The aerodynamic forces and moments are the primary factor in airborne vehicles and are often negligible for spaceborne vehicles, except at some low planetary orbits. These forces and moments for fixed-wing aircraft are introduced using conventional analytical models in an appendix of this textbook using basic aerodynamic theory. It should be noted that the determination of these models for general aircraft is studied as **aircraft system identification (SID)** which typically employs **optimal parameter estimation** and is discussed in latter parts of this textbook.

The radiation pressure forces and moments are due to the radiation pressure distribution around the aerospace vehicle due to the exchange of momentum between the vehicle and the electromagnetic waves that it absorbs or reflects. The **radiation pressure force** vector,  $\vec{F}_r$ , also known as the **force of light**, resists translation through space while the **radiation pressure moment** vector,  $\vec{M}_r$ , causes the vehicle to rotate due to the varying radiation pressure distribution over the spacecraft. The radiation pressure forces and moments are a factor in spaceborne vehicle dynamics and are negligible for airborne vehicles. For spaceborne vehicles, significant radiation pressure is generated by "nearby" celestial bodies, e.g., the Sun, Moon, Earth. The solar, lunar, and Earth radiation pressures are considered disturbances in the analysis of spacecraft dynamics and control in this textbook. Notably for artificial satellites, the aerodynamic drag force is larger than the solar radiation pressure force below roughly 800 km, although as the aerodynamic drag force depends on the density of the upper atmosphere, which is related to solar activity, the exact height at they are equivalent varies depending on the solar cycle.

### Point-Mass Aerospace Vehicle Dynamics in ECEF Frame

For some aerospace vehicle dynamics, one may simply use the ECI frame which allows for the inertial kinetics equations of motion for point-masses if one neglects the centripetal acceleration of the Earth about the Sun nor the Sun about the Milky Way's center which is typical. This provides the inertial velocity in ECI

frame coordinates as

$$\vec{v}_I = \dot{\vec{x}}_I = \begin{bmatrix} \dot{x}_I \\ \dot{y}_I \\ \dot{z}_I \end{bmatrix} \quad (7.77)$$

the inertial acceleration in ECI frame coordinates as

$$\vec{a}_I = \ddot{\vec{x}}_I = \begin{bmatrix} \ddot{x}_I \\ \ddot{y}_I \\ \ddot{z}_I \end{bmatrix} \quad (7.78)$$

Then, the point-mass equation of motion for aerospace vehicles in ECI coordinates as

$$\vec{F}_{g,I} + \vec{F}_{p,I} + \vec{F}_{a,I} + \vec{F}_{r,I} + \vec{F}_{\dot{m},I} = m \begin{bmatrix} \ddot{x}_I \\ \ddot{y}_I \\ \ddot{z}_I \end{bmatrix} \quad (7.79)$$

where

$$\vec{F}_{\dot{m},I} = -\dot{m} \begin{bmatrix} \dot{x}_I \\ \dot{y}_I \\ \dot{z}_I \end{bmatrix} \quad (7.80)$$

$\vec{F}_{\dot{m},I}$  is the additional “fictitious force” due to the mass rate. However, for other aerospace vehicles it may be useful to use the ECEF, NED navigation, or other vehicle-centered frame for point-mass analysis.

### Point-Mass Aerospace Vehicle Dynamics in ECEF Frame

First, recall that the ECEF and ECI frames share the same origin, one has  $\vec{\omega}_{E \leftarrow I} = 0$ ,  $\dot{\vec{\omega}}_{E \leftarrow I} = 0$ , and  $\ddot{\vec{\omega}}_{E \leftarrow I} = 0$ . Second, recall the angular velocity of the ECEF frame relative to the ECI frame is given by

$$\omega_{E \leftarrow I,I} = \omega_{E \leftarrow I,E} = \begin{bmatrix} 0 \\ 0 \\ \omega_E \end{bmatrix} \quad (7.81)$$

where  $\omega_E$  is Earth's mean angular rate and is constant, i.e.,  $\dot{\omega}_E = 0$  rad/s<sup>2</sup>. This allows one to write the inertial velocity in ECEF frame coordinates as

$$\vec{v}_E = \dot{\vec{x}}_E + [\vec{\omega}_{E \leftarrow I,E}] \times \vec{x}_E \quad (7.82)$$

By substitution, one has

$$\vec{v}_E = \begin{bmatrix} v_{x,E} \\ v_{y,E} \\ v_{z,E} \end{bmatrix} = \begin{bmatrix} \dot{x}_E \\ \dot{y}_E \\ \dot{z}_E \end{bmatrix} + \begin{bmatrix} -\omega_E y_E \\ \omega_E x_E \\ 0 \end{bmatrix} \quad (7.83)$$

and the inertial point-mass acceleration in ECEF frame coordinates as

$$\vec{a}_E = \ddot{\vec{x}}_E + 2[\vec{\omega}_{E \leftarrow I,E}] \times \dot{\vec{x}}_E + [\vec{\omega}_{E \leftarrow I,E}] \times [\vec{\omega}_{E \leftarrow I,E}] \times \vec{x}_E \quad (7.84)$$

$$\vec{a}_E = \begin{bmatrix} a_{x,E} \\ a_{y,E} \\ a_{z,E} \end{bmatrix} = \begin{bmatrix} \ddot{x}_E \\ \ddot{y}_E \\ \ddot{z}_E \end{bmatrix} + \begin{bmatrix} -2\omega_E \dot{y}_E - \omega_E^2 x_E \\ 2\omega_E \dot{x}_E - \omega_E^2 y_E \\ 0 \end{bmatrix} \quad (7.85)$$

Then, one has the point-mass equation of motion for aerospace vehicles in ECEF coordinates as

$$\vec{F}_{g,E} + \vec{F}_{p,E} + \vec{F}_{a,E} + \vec{F}_{r,E} + \vec{F}_{\dot{m},E} + \vec{F}_{Earth,E} = m \begin{bmatrix} \ddot{x}_E \\ \ddot{y}_E \\ \ddot{z}_E \end{bmatrix} \quad (7.86)$$

where

$$\vec{F}_{\dot{m},E} = -\dot{m} \left( \begin{bmatrix} \dot{x}_E \\ \dot{y}_E \\ \dot{z}_E \end{bmatrix} + \begin{bmatrix} -\omega_E y_E \\ \omega_E x_E \\ 0 \end{bmatrix} \right) \quad (7.87)$$

and

$$\vec{F}_{Earth,E} = -m \begin{bmatrix} 2\omega_E \dot{y}_E + \omega_E^2 x_E \\ -2\omega_E \ddot{x}_E + \omega_E^2 y_E \\ 0 \end{bmatrix} \quad (7.88)$$

$\vec{F}_{\dot{m},E}$  and  $\vec{F}_{Earth,E}$  are the additional “fictitious forces” due to the mass rate and Earth’s rotation, respectively.

### Point-Mass Aerospace Vehicle Dynamics in Navigation Frame

First, recall that the NED navigation frame’s origin is defined as the vehicle’s position. Therefore, the “position” of a point-mass within the NED navigation frame coordinates is

$$\vec{x}_N(t) = \vec{0} \quad \forall t \quad (7.89)$$

Therefore, the apparent velocity and acceleration within the reference frame are also zero, i.e.,

$$\dot{\vec{x}}_N(t) = \ddot{\vec{x}}_N(t) = \vec{0} \quad \forall t \quad (7.90)$$

Thus, the inertial velocity and acceleration equations in NED navigation frame coordinates can be written in terms of the offset vector’s inertial velocity and acceleration as

$$\vec{v}_N = \vec{v}_{\vec{o}_{N \leftarrow I, N}} \quad (7.91)$$

and

$$\vec{a}_N = \vec{a}_{\vec{o}_{N \leftarrow I, N}} \quad (7.92)$$

However, the offset vector’s inertial velocity and acceleration are known in ECEF coordinates, i.e.,  $\vec{o}_{N \leftarrow I, E}$  which requires the use of the ECEF frame point-mass equations, i.e.,

$$\vec{v}_E = \dot{\vec{o}}_{N \leftarrow I, E} + [\vec{\omega}_{E \leftarrow I, E}] \times \vec{o}_{N \leftarrow I, E} \quad (7.93)$$

and

$$\vec{a}_E = \vec{o}_{N \leftarrow I, E} + 2[\vec{\omega}_{E \leftarrow I, E}] \times \dot{\vec{o}}_{N \leftarrow I, E} + [\vec{\omega}_{E \leftarrow I, E}] \times [\vec{\omega}_{E \leftarrow I, E}] \times \vec{o}_{N \leftarrow I, E} \quad (7.94)$$

To convert these to the desired NED navigation frame coordinates, one can multiply by  $C_{N \leftarrow E}$  to obtain the inertial velocity and acceleration as

$$\vec{v}_N = \dot{\vec{o}}_{N \leftarrow I, N} + [\vec{\omega}_{E \leftarrow I, N}] \times \vec{o}_{N \leftarrow I, N} \quad (7.95)$$

with the slightly altered definition of the first term as

$$\dot{\vec{o}}_{N \leftarrow I, N} = C_{N \leftarrow E} \dot{\vec{o}}_{N \leftarrow I, E} \quad (7.96)$$

where it should be noted  $\dot{\vec{o}}_{N \leftarrow I, N}$  can be considered the “apparent” velocity of the point-mass in NED navigation frame coordinates when the point-mass is moving.

Furthermore,

$$\vec{a}_N = \frac{d}{dt} (C_{N \leftarrow E} \dot{\vec{o}}_{N \leftarrow I, E}) + 2[\vec{\omega}_{E \leftarrow I, N}] \times \dot{\vec{o}}_{N \leftarrow I, N} + C_{N \leftarrow E} [\vec{\omega}_{E \leftarrow I, E}] \times [\vec{\omega}_{E \leftarrow I, E}] \times \vec{o}_{N \leftarrow I, E} \quad (7.97)$$

to be consistent with the definition of  $\vec{o}_{N \leftarrow I, N}$ . Thus, one has

$$\vec{a}_N = \ddot{\vec{o}}_{N \leftarrow I, N} + [\vec{\omega}_{N \leftarrow E, N}] \times \dot{\vec{o}}_{N \leftarrow I, N} + 2[\vec{\omega}_{E \leftarrow I, N}] \times \dot{\vec{o}}_{N \leftarrow I, N} + C_{N \leftarrow E} [\vec{\omega}_{E \leftarrow I, E}] \times [\vec{\omega}_{E \leftarrow I, E}] \times \vec{o}_{N \leftarrow I, E} \quad (7.98)$$

$$\vec{a}_N = \ddot{\vec{o}}_{N \leftarrow I, N} + [\vec{\omega}_{N \leftarrow E, N} + 2\vec{\omega}_{E \leftarrow I, N}] \times \dot{\vec{o}}_{N \leftarrow I, N} + C_{N \leftarrow E} [\vec{\omega}_{E \leftarrow I, E}] \times [\vec{\omega}_{E \leftarrow I, E}] \times \vec{o}_{N \leftarrow I, E} \quad (7.99)$$

or, recalling  $\vec{\omega}_{N \leftarrow I} = \vec{\omega}_{N \leftarrow E} + \vec{\omega}_{E \leftarrow I}$ , one alternatively has

$$\vec{a}_N = \ddot{\vec{o}}_{N \leftarrow I, N} + [\vec{\omega}_{N \leftarrow I, N} + \vec{\omega}_{E \leftarrow I, N}] \times \dot{\vec{o}}_{N \leftarrow I, N} + C_{N \leftarrow E} [\vec{\omega}_{E \leftarrow I, E}] \times [\vec{\omega}_{E \leftarrow I, E}] \times \vec{o}_{N \leftarrow I, E} \quad (7.100)$$

where it should be noted  $\ddot{\vec{o}}_{N \leftarrow I, N}$  can be considered the “apparent” acceleration of the point-mass in NED navigation frame coordinates when the point-mass is accelerating.

Next, recall for  $\vec{\omega}_{E \leftarrow I, N}$ , one has

$$\vec{\omega}_{E \leftarrow I, N} = C_{N \leftarrow E} \vec{\omega}_{E \leftarrow I, E} = \begin{bmatrix} -\sin \ell \cos \lambda & -\sin \ell \sin \lambda & \cos \ell \\ -\sin \lambda & \cos \lambda & 0 \\ -\cos \ell \cos \lambda & -\cos \ell \sin \lambda & -\sin \ell \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \omega_E \end{bmatrix} = \begin{bmatrix} \omega_E \cos \ell \\ 0 \\ -\omega_E \sin \ell \end{bmatrix} \quad (7.101)$$

For  $\vec{\omega}_{N \leftarrow E, N}$ , known as the **transport rate**, one can use the definition of the derivative of a rotation matrix, i.e.,

$$\dot{C}_{N \leftarrow E} = [\vec{\omega}_{N \leftarrow E, N}] \times C_{N \leftarrow E} \quad (7.102)$$

Thus, defining,

$$\vec{\omega}_{N \leftarrow E, N} = \begin{bmatrix} \omega_{N \leftarrow E, n} \\ \omega_{N \leftarrow E, e} \\ \omega_{N \leftarrow E, d} \end{bmatrix} \quad (7.103)$$

one has

$$\begin{aligned} & \begin{bmatrix} \dot{\lambda} \sin \ell \sin \lambda - \dot{\ell} \cos \ell \cos \lambda & -\dot{\lambda} \sin \ell \cos \lambda - \dot{\ell} \cos \ell \sin \lambda & -\dot{\ell} \sin \ell \\ -\dot{\lambda} \cos \lambda & -\dot{\lambda} \sin \lambda & 0 \\ \dot{\lambda} \cos \ell \sin \lambda + \dot{\ell} \sin \ell \cos \lambda & -\dot{\lambda} \cos \ell \cos \lambda + \dot{\ell} \sin \ell \sin \lambda & -\dot{\ell} \cos \ell \end{bmatrix} \\ &= \begin{bmatrix} 0 & -\omega_{N \leftarrow E, d} & \omega_{N \leftarrow E, e} \\ \omega_{N \leftarrow E, d} & 0 & -\omega_{N \leftarrow E, n} \\ -\omega_{N \leftarrow E, e} & \omega_{N \leftarrow E, n} & 0 \end{bmatrix} \begin{bmatrix} -\sin \ell \cos \lambda & -\sin \ell \sin \lambda & \cos \ell \\ -\sin \lambda & \cos \lambda & 0 \\ -\cos \ell \cos \lambda & -\cos \ell \sin \lambda & -\sin \ell \end{bmatrix} \end{aligned} \quad (7.104)$$

Computing the (1,1), (2,1), and (3,3) elements, one has

$$\begin{bmatrix} \dot{\lambda} \sin \ell \sin \lambda - \dot{\ell} \cos \ell \cos \lambda \\ -\dot{\lambda} \cos \lambda \\ -\dot{\ell} \cos \ell \end{bmatrix} = \begin{bmatrix} -\omega_{N \leftarrow E, d} (-\sin \ell) + \omega_{N \leftarrow E, e} (-\cos \ell \cos \lambda) \\ \omega_{N \leftarrow E, d} (-\sin \ell \cos \lambda) + -\omega_{N \leftarrow E, n} (-\cos \ell \cos \lambda) \\ \omega_{N \leftarrow E, e} (-\cos \ell) \end{bmatrix} \quad (7.105)$$

Substituting the third equation into the second and simplifying, one has

$$\begin{bmatrix} -\dot{\lambda} \sin \ell \\ -\dot{\ell} \sin \ell \\ -\dot{\lambda} \end{bmatrix} = \begin{bmatrix} \omega_{N \leftarrow E, d} \\ -\omega_{N \leftarrow E, d} \sin \ell + \omega_{N \leftarrow E, n} \cos \ell \\ -\omega_{N \leftarrow E, e} \sin \ell \end{bmatrix} \quad (7.106)$$

Substituting the first into the second, one has

$$\vec{\omega}_{N \leftarrow E, N} = \begin{bmatrix} \dot{\lambda} \cos \ell \\ -\dot{\ell} \\ -\dot{\lambda} \sin \ell \end{bmatrix} \quad (7.107)$$

Combining, one also has

$$\vec{\omega}_{N \leftarrow I, N} = \begin{bmatrix} (\omega_E + \dot{\lambda}) \cos \ell \\ -\dot{\ell} \\ -(\omega_E + \dot{\lambda}) \sin \ell \end{bmatrix} \quad (7.108)$$

Then, recall the offset vector for the navigation frame relative to the ECEF/ECI frame can be expressed in the NED navigation frame as

$$\vec{o}_{N \leftarrow I, N} = \begin{bmatrix} -N_\ell e_E^2 \cos \ell \sin \ell \\ 0 \\ -N_\ell (1 - e_E^2 \sin^2 \ell) - h \end{bmatrix} \quad (7.109)$$

or in ECEF coordinates as

$$\vec{o}_{N \leftarrow I, E} = \begin{bmatrix} (N_\ell + h) \cos \lambda \cos \ell \\ (N_\ell + h) \sin \lambda \cos \ell \\ (N_\ell (1 - e_E^2) + h) \sin \ell \end{bmatrix} \quad (7.110)$$

For geodetic coordinates, one can show the following derivative relationships

$$\frac{d}{dt} ((N_\ell + h) \cos \ell) = -(M_\ell + h) \sin \ell \quad (7.111)$$

and

$$\frac{d}{dt} ((N_\ell (1 - e_E^2) + h) \sin \ell) = (M_\ell + h) \cos \ell \quad (7.112)$$

where  $M_\ell$  is the **meridian radius of curvature**, also known as the **north-south radius of curvature**, as

$$M_\ell = \frac{R_e (1 - e_E^2)}{(1 - e_E^2 \sin^2 \ell)^{3/2}} \quad (7.113)$$

Thus, the offset vector's derivative in ECEF coordinates can be written using geodetic coordinates as

$$\dot{\vec{o}}_{N \leftarrow I, E} = \begin{bmatrix} -\dot{\ell}(M_\ell + h) \sin \ell \cos \lambda - \dot{\lambda}(N_\ell + h) \cos \ell \sin \lambda + \dot{h} \cos \ell \cos \lambda \\ -\dot{\ell}(M_\ell + h) \sin \ell \sin \lambda + \dot{\lambda}(N_\ell + h) \cos \ell \cos \lambda + \dot{h} \cos \ell \sin \lambda \\ \dot{\ell}(M_\ell + h) \cos \ell + \dot{h} \sin \ell \end{bmatrix} \quad (7.114)$$

or written in NED navigation frame coordinates, one has

$$\dot{\vec{o}}_{N \leftarrow I, N} = C_{N \leftarrow E} \dot{\vec{o}}_{N \leftarrow I, E} = \begin{bmatrix} -\sin \ell \cos \lambda & -\sin \ell \sin \lambda & \cos \ell \\ -\sin \lambda & \cos \lambda & 0 \\ -\cos \ell \cos \lambda & -\cos \ell \sin \lambda & -\sin \ell \end{bmatrix} \begin{bmatrix} -\dot{\ell}(M_\ell + h) \sin \ell \cos \lambda - \dot{\lambda}(N_\ell + h) \cos \ell \sin \lambda + \dot{h} \cos \ell \cos \lambda \\ -\dot{\ell}(M_\ell + h) \sin \ell \sin \lambda + \dot{\lambda}(N_\ell + h) \cos \ell \cos \lambda + \dot{h} \cos \ell \sin \lambda \\ \dot{\ell}(M_\ell + h) \cos \ell + \dot{h} \sin \ell \end{bmatrix} \quad (7.115)$$

which provides

$$\dot{\vec{o}}_{N \leftarrow I, N} = \begin{bmatrix} \dot{\ell}(M_\ell + h) \\ \dot{\lambda}(N_\ell + h) \cos \ell \\ -\dot{h} \end{bmatrix} \quad (7.116)$$

Notably, can be rearranged to obtain the **geodetic rates**

$$\begin{bmatrix} \dot{\ell} \\ \dot{\lambda} \\ \dot{h} \end{bmatrix} = \begin{bmatrix} \frac{1}{M_\ell + h} & 0 & 0 \\ 0 & \frac{1}{(N_\ell + h) \cos \ell} & 0 \\ 0 & 0 & -1 \end{bmatrix} \dot{\vec{o}}_{N \leftarrow E, N} \quad (7.117)$$

Furthermore, defining

$$\dot{\vec{o}}_{N \leftarrow I, N} = \begin{bmatrix} \dot{o}_{N \leftarrow I, n} \\ \dot{o}_{N \leftarrow I, e} \\ \dot{o}_{N \leftarrow I, d} \end{bmatrix} \quad (7.118)$$

One has

$$\begin{bmatrix} \dot{\ell} \\ \dot{\lambda} \\ \dot{h} \end{bmatrix} = \begin{bmatrix} \frac{\dot{o}_{N \leftarrow I, n}}{M_\ell + h} \\ \frac{\dot{o}_{N \leftarrow I, e}}{(N_\ell + h) \cos \ell} \\ -\dot{o}_{N \leftarrow E, d} \end{bmatrix} \quad (7.119)$$

If one assumes a spherical-Earth model, i.e.,  $N_\ell = M_\ell = \bar{R}_E$ , one obtains the **geocentric rates**

$$\begin{bmatrix} \dot{\ell} \\ \dot{\lambda} \\ \dot{h} \end{bmatrix} = \begin{bmatrix} \frac{1}{\bar{R}_E + h} & 0 & 0 \\ 0 & \frac{1}{(\bar{R}_E + h) \cos \ell} & 0 \\ 0 & 0 & -1 \end{bmatrix} \dot{\vec{o}}_{N \leftarrow E, N} \quad (7.120)$$

and

$$\begin{bmatrix} \dot{\ell} \\ \dot{\lambda} \\ \dot{h} \end{bmatrix} = \begin{bmatrix} \frac{\dot{o}_{N \leftarrow E, n}}{\bar{R}_E + h} \\ \frac{\dot{o}_{N \leftarrow E, e}}{(\bar{R}_E + h) \cos \ell} \\ -\dot{o}_{N \leftarrow E, d} \end{bmatrix} \quad (7.121)$$

where this spherical-Earth NED navigation frame points directly towards the Earth's center. Thus, for a spherical-Earth model, one has  $\dot{o}_{N \leftarrow E, d} = -(\bar{R}_E + h)$ .

Substituting these geodetic rates into the angular velocity for the navigation frame relative to the ECEF frame, one has

$$\vec{\omega}_{N \leftarrow E, N} = \begin{bmatrix} \frac{\dot{\phi}_{y, N \leftarrow I, N}}{(N_\ell + h)} \\ -\frac{\dot{\phi}_{x, N \leftarrow I, N}}{M_\ell + h} \\ -\frac{\dot{\phi}_{y, N \leftarrow I, N}}{(N_\ell + h)} \tan \ell \end{bmatrix} \quad (7.122)$$

and defining

$$\ddot{\vec{o}}_{N \leftarrow I, N} = \begin{bmatrix} \ddot{o}_{N \leftarrow I, n} \\ \ddot{o}_{N \leftarrow I, e} \\ \ddot{o}_{N \leftarrow I, d} \end{bmatrix} \quad (7.123)$$

By substitution, one has the inertial velocity in NED navigation frame coordinates as

$$\vec{v}_N = \begin{bmatrix} \dot{o}_{N \leftarrow I, n} \\ \dot{o}_{N \leftarrow I, e} \\ \dot{o}_{N \leftarrow I, d} \end{bmatrix} + \begin{bmatrix} \omega_E \cos \ell \\ 0 \\ -\omega_E \sin \ell \end{bmatrix} \times \begin{bmatrix} -N_\ell e_E^2 \cos \ell \sin \ell \\ 0 \\ -N_\ell (1 - e_E^2 \sin^2 \ell) - h \end{bmatrix} \quad (7.124)$$

$$\vec{v}_N = \begin{bmatrix} \dot{o}_{N \leftarrow I, n} \\ \dot{o}_{N \leftarrow I, e} \\ \dot{o}_{N \leftarrow I, d} \end{bmatrix} + \begin{bmatrix} 0 \\ \omega_E N_\ell e_E^2 \cos \ell \sin^2 \ell + \omega_E (N_\ell (1 - e_E^2 \sin^2 \ell) + h) \cos \ell \\ 0 \end{bmatrix} \quad (7.125)$$

$$\vec{v}_N = \begin{bmatrix} \dot{o}_{N \leftarrow I, n} \\ \dot{o}_{N \leftarrow I, e} \\ \dot{o}_{N \leftarrow I, d} \end{bmatrix} + \begin{bmatrix} 0 \\ \omega_E (N_\ell + h) \cos \ell \\ 0 \end{bmatrix} \quad (7.126)$$

By substitution, one has the inertial acceleration in NED navigation frame coordinates as

$$\begin{aligned} \vec{a}_N &= \begin{bmatrix} \ddot{o}_{N \leftarrow I, n} \\ \ddot{o}_{N \leftarrow I, e} \\ \ddot{o}_{N \leftarrow I, d} \end{bmatrix} + \left[ \begin{bmatrix} \frac{\dot{\phi}_{y, N \leftarrow I, N}}{(N_\ell + h)} \\ -\frac{\dot{\phi}_{x, N \leftarrow I, N}}{M_\ell + h} \\ -\frac{\dot{\phi}_{y, N \leftarrow I, N}}{(N_\ell + h)} \tan \ell \end{bmatrix} + 2 \begin{bmatrix} \omega_E \cos \ell \\ 0 \\ -\omega_E \sin \ell \end{bmatrix} \right] \times \begin{bmatrix} \dot{o}_{N \leftarrow I, n} \\ \dot{o}_{N \leftarrow I, e} \\ \dot{o}_{N \leftarrow I, d} \end{bmatrix} \\ &\quad + \begin{bmatrix} -\sin \ell \cos \lambda & -\sin \ell \sin \lambda & \cos \ell \\ -\sin \lambda & \cos \lambda & 0 \\ -\cos \ell \cos \lambda & -\cos \ell \sin \lambda & -\sin \ell \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \omega_E \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \\ \omega_E \end{bmatrix} \times \begin{bmatrix} (N_\ell + h) \cos \lambda \cos \ell \\ (N_\ell + h) \sin \lambda \cos \ell \\ (N_\ell (1 - e_E^2) + h) \sin \ell \end{bmatrix} \end{aligned} \quad (7.127)$$

$$\begin{aligned} \vec{a}_N &= \begin{bmatrix} \ddot{o}_{N \leftarrow I, n} \\ \ddot{o}_{N \leftarrow I, e} \\ \ddot{o}_{N \leftarrow I, d} \end{bmatrix} + \left[ \begin{bmatrix} \frac{\dot{o}_{N \leftarrow I, e}^2 \tan \ell}{N_\ell + h} - \frac{\dot{o}_{N \leftarrow I, n} \dot{o}_{N \leftarrow I, d}}{M_\ell + h} \\ -\frac{\dot{o}_{N \leftarrow I, e} (\dot{o}_{N \leftarrow I, d} + \dot{o}_{N \leftarrow I, n} \tan \ell)}{N_\ell + h} \\ \frac{\dot{o}_{N \leftarrow I, n}^2}{M_\ell + h} + \frac{\dot{o}_{N \leftarrow I, e}^2}{N_\ell + h} \end{bmatrix} + 2\omega_E \begin{bmatrix} \sin \ell \dot{o}_{N \leftarrow I, e} \\ (\sin \ell \dot{o}_{N \leftarrow I, n} - \cos \ell \dot{o}_{N \leftarrow I, d}) \\ \cos \ell \dot{o}_{N \leftarrow I, e} \end{bmatrix} \right] \\ &\quad + \begin{bmatrix} -\sin \ell \cos \lambda & -\sin \ell \sin \lambda & \cos \ell \\ -\sin \lambda & \cos \lambda & 0 \\ -\cos \ell \cos \lambda & -\cos \ell \sin \lambda & -\sin \ell \end{bmatrix} \begin{bmatrix} -\omega_E^2 (N_\ell + h) \cos \lambda \cos \ell \\ -\omega_E^2 (N_\ell + h) \sin \lambda \cos \ell \\ 0 \end{bmatrix} \end{aligned} \quad (7.128)$$

Finally, one can write the inertial acceleration in NED navigation frame coordinates as

$$\vec{a}_N = \begin{bmatrix} \ddot{o}_{N \leftarrow I, n} \\ \ddot{o}_{N \leftarrow I, e} \\ \ddot{o}_{N \leftarrow I, d} \end{bmatrix} + \begin{bmatrix} \frac{\dot{o}_{N \leftarrow I, e}^2 \tan \ell}{N_\ell + h} - \frac{\dot{o}_{N \leftarrow I, n} \dot{o}_{N \leftarrow I, d}}{M_\ell + h} \\ -\frac{\dot{o}_{N \leftarrow I, e} (\dot{o}_{N \leftarrow I, d} + \dot{o}_{N \leftarrow I, n} \tan \ell)}{N_\ell + h} \\ \frac{\dot{o}_{N \leftarrow I, n}^2}{M_\ell + h} + \frac{\dot{o}_{N \leftarrow I, e}^2}{N_\ell + h} \end{bmatrix} + 2\omega_E \begin{bmatrix} \sin \ell \dot{o}_{N \leftarrow I, e} \\ (\sin \ell \dot{o}_{N \leftarrow I, n} - \cos \ell \dot{o}_{N \leftarrow I, d}) \\ \cos \ell \dot{o}_{N \leftarrow I, e} \end{bmatrix} + \omega_E^2 \begin{bmatrix} (N_\ell + h) \sin \ell \cos \ell \\ 0 \\ (N_\ell + h) \cos^2 \ell \end{bmatrix} \quad (7.129)$$

If one assumes a spherical-Earth model, i.e.,  $N_\ell = M_\ell = \bar{R}_E$ , one has

$$\vec{v}_N = \begin{bmatrix} \dot{o}_{N \leftarrow I, n} \\ \dot{o}_{N \leftarrow I, e} \\ \dot{o}_{N \leftarrow I, d} \end{bmatrix} + \begin{bmatrix} 0 \\ \omega_E (\bar{R}_E + h) \cos \ell \\ 0 \end{bmatrix} \quad (7.130)$$

and

$$\vec{a}_N = \begin{bmatrix} \ddot{o}_{N \leftarrow I, n} \\ \ddot{o}_{N \leftarrow I, e} \\ \ddot{o}_{N \leftarrow I, d} \end{bmatrix} + \frac{1}{\bar{R}_E + h} \begin{bmatrix} \dot{o}_{N \leftarrow I, e}^2 \tan \ell - \dot{o}_{N \leftarrow I, n} \dot{o}_{N \leftarrow I, d} \\ -\dot{o}_{N \leftarrow E, e} (\dot{o}_{N \leftarrow I, d} + \dot{o}_{N \leftarrow I, n} \tan \ell) \\ \dot{o}_{N \leftarrow E, n}^2 + \dot{o}_{N \leftarrow I, e}^2 \end{bmatrix} + 2\omega_E \begin{bmatrix} \sin \ell \dot{o}_{N \leftarrow I, e} \\ \sin \ell \dot{o}_{N \leftarrow I, n} - \cos \ell \dot{o}_{N \leftarrow I, d} \\ \cos \ell \dot{o}_{N \leftarrow I, e} \end{bmatrix} + \omega_E^2 \begin{bmatrix} (\bar{R}_E + h) \sin \ell \cos \ell \\ 0 \\ (\bar{R}_E + h) \cos^2 \ell \end{bmatrix} \quad (7.131)$$

where this spherical-Earth NED navigation frame points directly towards the Earth's center. Lastly, if one assumes a flat-Earth model, i.e.,  $N_\ell \rightarrow \infty$ ,  $M_\ell \rightarrow \infty$ , and  $\omega_E = 0^\circ/\text{s}$ , one has

$$\vec{v}_N = \begin{bmatrix} \dot{o}_{N \leftarrow I, n} \\ \dot{o}_{N \leftarrow I, e} \\ \dot{o}_{N \leftarrow I, d} \end{bmatrix} \quad (7.132)$$

and

$$\vec{a}_N = \begin{bmatrix} \ddot{o}_{N \leftarrow I, n} \\ \ddot{o}_{N \leftarrow I, e} \\ \ddot{o}_{N \leftarrow I, d} \end{bmatrix} \quad (7.133)$$

Then, one has the point-mass equation of motion for aerospace vehicles in NED navigation frame coordinates as

$$\vec{F}_{g,N} + \vec{F}_{p,N} + \vec{F}_{a,N} + \vec{F}_{r,N} + \vec{F}_{\dot{m},N} + \vec{F}_{Earth,N} = m \begin{bmatrix} \ddot{o}_{N \leftarrow I, n} \\ \ddot{o}_{N \leftarrow I, e} \\ \ddot{o}_{N \leftarrow I, d} \end{bmatrix} \quad (7.134)$$

where one has for an ellipsoidal-Earth model

$$\vec{F}_{\dot{m},N} = -\dot{m} \left( \begin{bmatrix} \dot{o}_{N \leftarrow I, n} \\ \dot{o}_{N \leftarrow I, e} \\ \dot{o}_{N \leftarrow I, d} \end{bmatrix} + \begin{bmatrix} 0 \\ \omega_E (N_\ell + h) \cos \ell \\ 0 \end{bmatrix} \right) \quad (7.135)$$

and

$$\vec{F}_{Earth,N} = -m \left( \begin{bmatrix} \frac{\dot{\phi}_{N \leftarrow I,e}^2 \tan \ell}{N_\ell - z_{N \leftarrow I,N}} - \frac{\dot{\phi}_{N \leftarrow I,n} \dot{\phi}_{N \leftarrow I,d}}{M_\ell - z_{N \leftarrow I,N}} \\ -\frac{\dot{\phi}_{N \leftarrow I,e} (\dot{\phi}_{N \leftarrow I,d} + \dot{\phi}_{N \leftarrow I,n} \tan \ell)}{N_\ell - z_{N \leftarrow I,N}} \\ \frac{\dot{\phi}_{N \leftarrow I,n}^2}{M_\ell + h} + \frac{\dot{\phi}_{N \leftarrow I,e}^2}{N_\ell - z_{N \leftarrow I,N}} \end{bmatrix} + 2\omega_E \begin{bmatrix} \sin \ell \dot{\phi}_{N \leftarrow I,e} \\ (\sin \ell \dot{\phi}_{N \leftarrow I,n} - \cos \ell \dot{\phi}_{N \leftarrow I,d}) \\ \cos \ell \dot{\phi}_{N \leftarrow I,e} \end{bmatrix} + \omega_E^2 \begin{bmatrix} (N_\ell + h) \sin \ell \cos \ell \\ 0 \\ (N_\ell + h) \cos^2 \ell \end{bmatrix} \right) \quad (7.136)$$

for a spherical-Earth model

$$\vec{F}_{\dot{m},N} = -\dot{m} \left( \begin{bmatrix} \dot{\phi}_{N \leftarrow I,n} \\ \dot{\phi}_{N \leftarrow I,e} \\ \dot{\phi}_{N \leftarrow I,d} \end{bmatrix} + \begin{bmatrix} 0 \\ \omega_E (\bar{R}_E + h) \cos \ell \\ 0 \end{bmatrix} \right) \quad (7.137)$$

and

$$\begin{aligned} \vec{F}_{Earth,N} = & -m \left( \frac{1}{\bar{R}_E + h} \begin{bmatrix} \dot{\phi}_{N \leftarrow I,e}^2 \tan \ell - \dot{\phi}_{N \leftarrow I,n} \dot{\phi}_{N \leftarrow I,d} \\ -\dot{\phi}_{N \leftarrow I,e} (\dot{\phi}_{N \leftarrow I,d} + \dot{\phi}_{N \leftarrow I,n} \tan \ell) \\ \dot{\phi}_{N \leftarrow I,n}^2 + \dot{\phi}_{N \leftarrow I,e}^2 \end{bmatrix} \right. \\ & \left. + 2\omega_E \begin{bmatrix} \sin \ell \dot{\phi}_{N \leftarrow I,e} \\ \sin \ell \dot{\phi}_{N \leftarrow I,n} - \cos \ell \dot{\phi}_{N \leftarrow I,d} \\ \cos \ell \dot{\phi}_{N \leftarrow I,e} \end{bmatrix} + \omega_E^2 \begin{bmatrix} (\bar{R}_E + h) \sin \ell \cos \ell \\ 0 \\ (\bar{R}_E + h) \cos^2 \ell \end{bmatrix} \right) \end{aligned} \quad (7.138)$$

and, for a flat-Earth model

$$\vec{F}_{\dot{m},N} = -\dot{m} \begin{bmatrix} \dot{\phi}_{N \leftarrow I,n} \\ \dot{\phi}_{N \leftarrow I,e} \\ \dot{\phi}_{N \leftarrow I,d} \end{bmatrix} \quad (7.139)$$

and

$$\vec{F}_{Earth,N} = \vec{0} \quad (7.140)$$

$\vec{F}_{\dot{m},N}$  and  $\vec{F}_{Earth,N}$  are the additional “fictitious forces” due to the mass rate and Earth’s rotation and curvature, respectively. Notably, these forces include geodetic coordinates and their rates, which requires the equation for the offset vector in terms of geodetic coordinates and the geodetic rates equation to solve these equations of motion for the position of the point-mass.

### Point-Mass Aerospace Vehicle Dynamics in Other Vehicle-Centered Frames

Lastly, from these equations for the navigation frame offset, one can also compute the inertial velocity in any other vehicle-centered reference frame  $\bullet$  from

$$\vec{v}_\bullet = C_{\bullet \leftarrow N} \vec{v}_N = \dot{\vec{\sigma}}_{N \leftarrow I, \bullet} + [\vec{\omega}_{E \leftarrow I, \bullet}]_\times \vec{\sigma}_{N \leftarrow I, \bullet} \quad (7.141)$$

with the slightly altered definition of the first term as

$$\dot{\vec{\sigma}}_{N \leftarrow I, \bullet} = C_{\bullet \leftarrow N} C_{N \leftarrow E} \dot{\vec{\sigma}}_{N \leftarrow I, E} \quad (7.142)$$

where it should be noted  $\dot{\vec{\sigma}}_{N \leftarrow I, \bullet}$  can be considered the “apparent” velocity of the point-mass in the vehicle-centered  $\bullet$  frame coordinates when the point-mass is moving.

Furthermore, one has

$$\vec{a}_{\bullet} = C_{\bullet \leftarrow N} \vec{a}_N = \ddot{\vec{o}}_{N \leftarrow I, \bullet} + [\vec{\omega}_{\bullet \leftarrow N, \bullet} + \vec{\omega}_{N \leftarrow I, \bullet} + \vec{\omega}_{E \leftarrow I, \bullet}] \times \dot{\vec{o}}_{N \leftarrow I, \bullet} + C_{N \leftarrow E} [\vec{\omega}_{E \leftarrow I, E}] \times [\vec{\omega}_{E \leftarrow I, E}] \times \vec{x}_E \quad (7.143)$$

where it should be noted  $\ddot{\vec{o}}_{N \leftarrow I, N}$  can be considered the “apparent” acceleration of the point-mass in frame  $\bullet$  coordinates when the point-mass is accelerating. Then, one has the point-mass equation of motion for aerospace vehicles in frame  $\bullet$  coordinates as

$$\vec{F}_{g, \bullet} + \vec{F}_{p, \bullet} + \vec{F}_{a, \bullet} + \vec{F}_{r, \bullet} + C_{\bullet \leftarrow N} \vec{F}_{m, N} + C_{\bullet \leftarrow N} \vec{F}_{Earth, N} = m (\ddot{\vec{o}}_{N \leftarrow E, \bullet} + [\vec{\omega}_{\bullet \leftarrow N, \bullet}] \times \dot{\vec{o}}_{N \leftarrow I, \bullet}) \quad (7.144)$$

which simply accounts for the rotation of the axes of frame  $\bullet$  relative to the NED navigation frame separate from the fictitious forces due to the mass rate and Earth’s rotation. This separation allows for easy their inclusion or neglect in the use of the point-mass equations of motion for aerospace vehicles. Notably, the supplemental relationship between the DCM,  $C_{\bullet \leftarrow N}$ , and the angular velocity,  $\vec{\omega}_{\bullet \leftarrow N, \bullet}$ , is required in this formulation.

## References

For more information, please refer to the following

- Curtis, H. D., “1.7 Relative motion,” *Orbital Mechanics for Engineering Students*, 4th ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2021, pp. 26-34
- Schmidt, D. K., “2.6 Effects of Spherical, Rotating Earth,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 65-75
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “1.6 Geodesy, Coordinate Systems, Gravity,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 23-34
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “1.8 Advanced Topics,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 44-58

## 7.3 Rigid-Body Aerospace Vehicle Dynamics

Aerospace vehicle propulsion systems produce the **propulsive moment** vector,  $\vec{M}_p$ , in the body-fixed frame as a function of **thrust** vectors,  $\vec{T}$ , and the location of the nozzle, propeller, or and/or rotor(s) relative to the center of mass of the vehicle. Some propulsion systems are used to steer aerospace vehicles, a design known as **thrust vectoring**. It should also be noted that rotating propellers and rotors create a changing mass distribution due to their rotation and are discussed in this section on this effect. The aerodynamic forces and moments are due to the air pressure distribution around the aerospace vehicle. The **aerodynamic moment** vector,  $\vec{M}_a$ , causes the vehicle to rotate due to the varying pressure distribution over the body. The radiation pressure moments are due to the radiation pressure distribution around the aerospace vehicle due to the exchange of momentum between the vehicle and the electromagnetic waves that it absorbs or reflects. The **radiation pressure moment** vector,  $\vec{M}_r$ , causes the vehicle to rotate due to the varying radiation pressure distribution over the spacecraft.

### Earth-Centered Rigid-Body Aerospace Vehicle Dynamics

Defining the velocity of the vehicle's center of mass in body-fixed frame coordinates, one has

$$\dot{\vec{x}}_{B/E,L} = C_{L \leftarrow B} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \quad (7.145)$$

$$\begin{bmatrix} \dot{\ell}_E \\ \dot{\lambda} \\ \dot{r}_E \end{bmatrix} = \begin{bmatrix} \frac{1}{r_E} & 0 & 0 \\ 0 & \frac{1}{r_E \cos \ell_E} & 0 \\ 0 & 0 & -1 \end{bmatrix} C_{L \leftarrow B} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \quad (7.146)$$

or

$$\dot{\vec{x}}_{B/E,N} = C_{N \leftarrow B} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \quad (7.147)$$

$$\begin{bmatrix} \dot{\ell} \\ \dot{\lambda} \\ \dot{h} \end{bmatrix} = \begin{bmatrix} \frac{1}{M_\ell + h} & 0 & 0 \\ 0 & \frac{1}{(N_\ell + h) \cos \ell} & 0 \\ 0 & 0 & -1 \end{bmatrix} C_{N \leftarrow B} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \quad (7.148)$$

The angular velocity effect of the rotating Earth will appear as two additional angular velocity terms for the ECI and ECEF frames, i.e.

$$\vec{\omega}_{B/I} = \vec{\omega}_{B/L} + \vec{\omega}_{L/E} + \vec{\omega}_{E/I} = \vec{\omega}_{B/N} + \vec{\omega}_{N/E} + \vec{\omega}_{E/I} \quad (7.149)$$

and in body-fixed frame coordinates as

$$\vec{\omega}_{E/I,B} = C_{B \leftarrow L}(\phi_L, \theta_L, \psi_L) \begin{bmatrix} \omega_E \cos \ell_E \\ 0 \\ -\omega_E \sin \ell_E \end{bmatrix} = C_{B \leftarrow N}(\phi, \theta, \psi) \begin{bmatrix} \omega_E \cos \ell \\ 0 \\ -\omega_E \sin \ell \end{bmatrix} \quad (7.150)$$

where  $(\phi_L, \theta_L, \psi_L)$  are the 3–2–1 Euler angles from the LVLH frame to the body-fixed frame and  $(\phi, \theta, \psi)$  are the 3–2–1 Euler angles from the navigation frame to the body-fixed frame. Note that for flat-Earth and spherical-Earth models, these are the same.

Recall that

$$\vec{\omega}_{B/L,B} = \begin{bmatrix} p_{B/L,B} \\ q_{B/L,B} \\ r_{B/L,B} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\sin \theta_L \\ 0 & \cos \phi_L & -\sin \phi_L \cos \theta_L \\ 0 & -\sin \phi_L & \cos \phi_L \cos \theta_L \end{bmatrix} \begin{bmatrix} \dot{\phi}_L \\ \dot{\theta}_L \\ \dot{\psi}_L \end{bmatrix} \quad (7.151)$$

and

$$\vec{\omega}_{B/N,B} = \begin{bmatrix} p_{B/N,B} \\ q_{B/N,B} \\ r_{B/N,B} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\sin \theta \\ 0 & \cos \phi & -\sin \phi \cos \theta \\ 0 & -\sin \phi & \cos \phi \cos \theta \end{bmatrix} \begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} \quad (7.152)$$

which are equivalent for a spherical-Earth model and is the “inertial” angular velocity for a flat-Earth model.

Thus, to account for the rotation of the Earth, one must redefine the new inertial angular velocity coordinates as

$$\vec{\omega}_{B/I,B} = \begin{bmatrix} 1 & 0 & -\sin \theta_L \\ 0 & \cos \phi_L & -\sin \phi_L \cos \theta_L \\ 0 & -\sin \phi_L & \cos \phi_L \cos \theta_L \end{bmatrix} \begin{bmatrix} \dot{\phi}_L \\ \dot{\theta}_L \\ \dot{\psi}_L \end{bmatrix} + C_{B \leftarrow L}(\phi_L, \theta_L, \psi_L) \begin{bmatrix} (\omega_E + \lambda) \cos \ell_E \\ -\dot{\ell}_E \\ -(\omega_E + \lambda) \sin \ell_E \end{bmatrix}$$

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} = \begin{bmatrix} p_{B/L,B} \\ q_{B/L,B} \\ r_{B/L,B} \end{bmatrix} + \begin{bmatrix} p_{L/I,B} \\ q_{L/I,B} \\ r_{L/I,B} \end{bmatrix} \quad (7.153)$$

or

$$\vec{\omega}_{B/I,B} = \begin{bmatrix} 1 & 0 & -\sin \theta \\ 0 & \cos \phi & -\sin \phi \cos \theta \\ 0 & -\sin \phi & \cos \phi \cos \theta \end{bmatrix} \begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} + C_{B \leftarrow N}(\phi, \theta, \psi) \begin{bmatrix} (\omega_E + \lambda) \cos \ell \\ -\dot{\ell} \\ -(\omega_E + \lambda) \sin \ell \end{bmatrix}$$

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} = \begin{bmatrix} p_{B/N,B} \\ q_{B/N,B} \\ r_{B/N,B} \end{bmatrix} + \begin{bmatrix} p_{N/I,B} \\ q_{N/I,B} \\ r_{N/I,B} \end{bmatrix} \quad (7.154)$$

which can be transformed to body-fixed frame coordinates as

$$\ddot{\vec{x}}_{B/I,B} = \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \begin{bmatrix} p + p_{E/I,B} \\ q + q_{E/I,B} \\ r + r_{E/I,B} \end{bmatrix} \times \begin{bmatrix} u \\ v \\ w \end{bmatrix} + [\vec{\omega}_{E/I,B}] \times \begin{bmatrix} u \\ v \\ w \end{bmatrix} + C_{B \leftarrow E} \begin{bmatrix} -\omega_E^2 & 0 & 0 \\ 0 & -\omega_E^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \vec{x}_{B/E,E} \quad (7.155)$$

where

$$\vec{\omega}_{E/I,B} = \begin{bmatrix} p_{E/I,B} \\ q_{E/I,B} \\ r_{E/I,B} \end{bmatrix} \quad (7.156)$$

$$\sum \vec{F}_B - \begin{bmatrix} q_{E/I,B}w - r_{E/I,B}v \\ r_{E/I,B}u - p_{E/I,B}w \\ p_{E/I,B}v - q_{E/I,B}u \end{bmatrix} - C_{B \leftarrow E} \begin{bmatrix} -\omega_E^2 & 0 & 0 \\ 0 & -\omega_E^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \vec{x}_{B/E,E} = \begin{bmatrix} \dot{u} + qw - rv \\ \dot{v} + ru - pw \\ \dot{w} + pv - qu \end{bmatrix} \quad (7.157)$$

where

$$\begin{bmatrix} p_{E/I,B} \\ q_{E/I,B} \\ r_{E/I,B} \end{bmatrix} = C_{B \leftarrow L}(\phi_L, \theta_L, \psi_L) \begin{bmatrix} \omega_E \cos \ell_E \\ 0 \\ -\omega_E \sin \ell_E \end{bmatrix} = C_{B \leftarrow N}(\phi, \theta, \psi) \begin{bmatrix} \omega_E \cos \ell \\ 0 \\ -\omega_E \sin \ell \end{bmatrix} \quad (7.158)$$

and

$$\vec{x}_{B/E,E} = \begin{bmatrix} r_E \cos \ell_E \cos \lambda \\ r_E \cos \ell_E \sin \lambda \\ r_E \sin \ell_E \end{bmatrix} = \begin{bmatrix} (N_\ell + h) \cos \lambda \cos \ell \\ (N_\ell + h) \sin \lambda \cos \ell \\ (N_\ell(1 - e_E^2) + h) \sin \ell \end{bmatrix} \quad (7.159)$$

Of particular note, if one assumes a **spherical-Earth model**, i.e.  $\ell_E = \ell$  and  $r_E = \bar{R}_E + h$  where  $\bar{R}_E$  is the **mean radius of the Earth**, defined as 6,366,707.0195 m by the WGS 84, one has the simpler equations

$$\begin{bmatrix} \dot{\ell} \\ \dot{\lambda} \\ \dot{h} \end{bmatrix} = \begin{bmatrix} \frac{1}{\bar{R}_E + h} & 0 & 0 \\ 0 & \frac{1}{(\bar{R}_E + h) \cos \ell} & 0 \\ 0 & 0 & -1 \end{bmatrix} C_{L \leftarrow B} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \quad (7.160)$$

and

$$\sum \vec{F}_B - \begin{bmatrix} q_{E/I,BW} - r_{E/I,BV} \\ r_{E/I,Bu} - p_{E/I,BW} \\ p_{E/I,Bv} - q_{E/I,Bu} \end{bmatrix} - C_{B \leftarrow N} \begin{bmatrix} \omega_E^2 (\bar{R}_E + h) \cos \lambda \sin \lambda \\ 0 \\ \omega_E^2 (\bar{R}_E + h) \cos^2 \lambda \end{bmatrix} = \begin{bmatrix} \dot{u} + qw - rv \\ \dot{v} + ru - pw \\ \dot{w} + pv - qu \end{bmatrix} \quad (7.161)$$

### Symmetrically Rotating Mass Effect

The presence of a mass rotating about its center of mass, i.e., symmetrically rotating, can be shown to have no effect on the translation equations of motion of a vehicle. However, the presence of a mass rotating about its center of mass does have an effect on the rotation equations of motion of a vehicle as this rotation directly contributes to the total angular momentum of the vehicle. To show this, consider the angular momentum of the vehicle as

$$\begin{bmatrix} I_{xx}L \\ I_{yy}M \\ I_{zz}N \end{bmatrix} = \dot{\vec{H}}_N = \dot{\vec{H}}_B + \vec{\omega}_{B \leftarrow N} \times \vec{H}_B \quad (7.162)$$

where  $\vec{H}_B$  consists of two components, the rigid vehicle's angular momentum (with the mass of the propeller disk) and the propeller's angular momentum, i.e.

$$\vec{H}_B = \vec{H}_{rig,B} + \vec{H}_{prop,B} \quad (7.163)$$

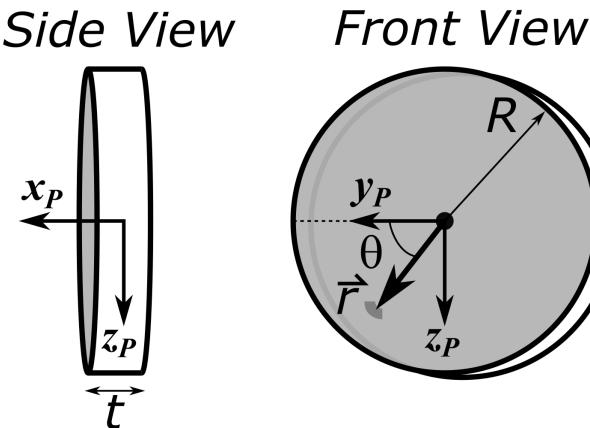
where

$$\vec{H}_{rig,B} = I_G \times \vec{\omega}_{B/N} \quad (7.164)$$

and the general form for the propeller angular momentum is

$$\vec{H}_{prop,B} = \int_{Vol} \vec{x}_B \times \dot{\vec{x}}_B \rho_V dV \quad (7.165)$$

where  $\vec{x}_B$  is the radial position of a mass element  $\rho_V dV$  with respect to the center of mass. Next, one can idealize the propeller as a rotating disk with radius  $R$  and constant thickness  $t$  as



where the  $x_P - y_P - z_P$  axes denotes the propeller-fixed frame centered at the disk's center (subscript  $P$ ), the mass element volume at radius  $r$  is

$$dV = t(r d\theta) dr \quad (7.166)$$

and the disk has a radial mass distribution matching that of the propeller, i.e. setting the disk density,  $\rho_{disk}$ , to cover the entire disk volume, but with equivalent mass. Then, one can write the mass element velocity as

$$\vec{x}_{B'} = \vec{x}_P + \vec{\omega}_{B'/P} \times \vec{x}_P \quad (7.167)$$

where  $B'$  here is a body-fixed frame centered at the propeller with potentially different center and axes than  $B$ . Assuming the propeller disk is rigid, this can be rewritten as

$$\begin{aligned} \vec{x}_{B'} &= \omega_{B'/P} \times \vec{x}_P = \begin{bmatrix} \omega_{prop} \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ r \cos \theta \\ r \sin \theta \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ \omega_{prop} r \sin \theta \\ -\omega_{prop} r \cos \theta \end{bmatrix} \end{aligned} \quad (7.168)$$

Thus, one has for the integrand of the angular velocity

$$\begin{aligned} \vec{x}_{B'} \times \rho_V \vec{x}_{B'} &= \rho_V \begin{bmatrix} 0 \\ r \cos \theta \\ r \sin \theta \end{bmatrix} \times \begin{bmatrix} 0 \\ \omega_{prop} r \sin \theta \\ -\omega_{prop} r \cos \theta \end{bmatrix} \\ &= \begin{bmatrix} \omega_{prop} r^2 \\ 0 \\ 0 \end{bmatrix} \end{aligned} \quad (7.169)$$

and the angular momentum of the propeller disk in body frame coordinates is

$$\begin{aligned} \vec{H}_{prop,B'} &= \begin{bmatrix} \int_0^R \int_0^{2\pi} \omega_{prop} r^2 \rho_{disk} (r t d\theta dr) \\ 0 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \omega_{prop} 2\pi t \int_0^R r^3 \rho_{disk} dr \\ 0 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \omega_{prop} I_{prop} \\ 0 \\ 0 \end{bmatrix} \end{aligned} \quad (7.170)$$

where  $I_{prop}$  is the moment of inertia of the propeller about its center of mass.

Next, returning to the propeller angular momentum in the vehicle body frame centered at the center of mass, one has

$$\vec{H}_{prop,B} = \begin{bmatrix} h_{x,prop} \\ h_{y,prop} \\ h_{z,prop} \end{bmatrix} = C_{B \leftarrow B'} \begin{bmatrix} \omega_{prop} I_{prop} \\ 0 \\ 0 \end{bmatrix} \quad (7.171)$$

where the DCM,  $C_{B \leftarrow B'}$ , will depend on the orientation of the propeller relative to the body frame. For example, if the  $B'$  at some positive rotation angle,  $\tau_P$ , about the  $y_B$  axis, then

$$\vec{H}_{prop,B} = \begin{bmatrix} \omega_{prop} I_{prop} \cos \tau_P \\ 0 \\ -\omega_{prop} I_{prop} \sin \tau_P \end{bmatrix} \quad (7.172)$$

Then, by substituting the additive angular momentum terms into the angular momentum differential equation, one has

$$\begin{bmatrix} I_{xx}L \\ I_{yy}M \\ I_{zz}N \end{bmatrix} = \left( \dot{\vec{H}}_{rig,B} + \vec{H}_{prop,B} \right) + \vec{\omega}_{B \leftarrow N} \times \left( \vec{H}_{rig,B} + \vec{H}_{prop,B} \right) \quad (7.173)$$

$$\begin{bmatrix} I_{xx}L \\ I_{yy}M \\ I_{zz}N \end{bmatrix} = \left( \dot{\vec{H}}_{rig,B} + \vec{\omega}_{B \leftarrow N} \times \vec{H}_{rig,B} \right) + \left( \dot{\vec{H}}_{prop,B} + \vec{\omega}_{B \leftarrow N} \times \vec{H}_{prop,B} \right) \quad (7.174)$$

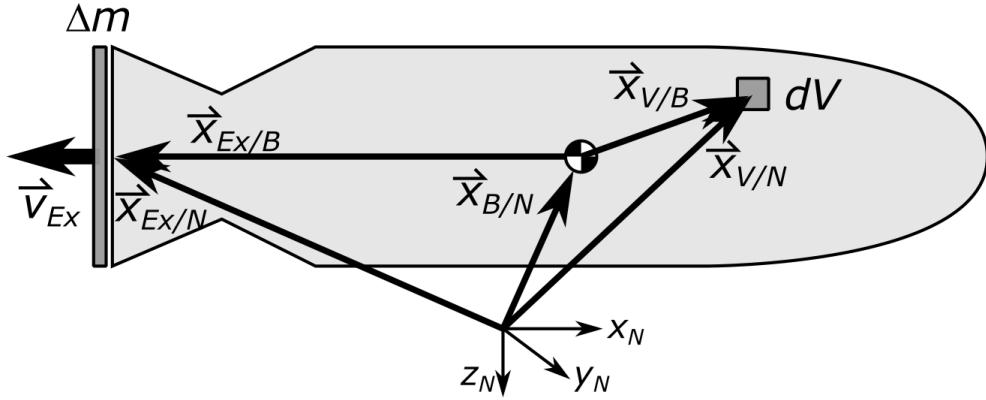
$$\begin{bmatrix} L \\ M \\ N \end{bmatrix} = \begin{bmatrix} \dot{p} + \frac{I_{zz}-I_{yy}}{I_{xx}} qr - \frac{I_{xz}}{I_{xx}} (\dot{r} + pq) \\ \dot{q} + \frac{I_{xx}-I_{zz}}{I_{yy}} pr - \frac{I_{xz}}{I_{yy}} (r^2 - p^2) \\ \dot{r} + \frac{I_{yy}-I_{xx}}{I_{zz}} pq - \frac{I_{xz}}{I_{zz}} (\dot{p} - qr) \end{bmatrix} + \begin{bmatrix} \frac{1}{I_{xx}} (\dot{h}_{x,prop} - rh_{y,prop} + qh_{z,prop}) \\ \frac{1}{I_{yy}} (\dot{h}_{y,prop} + rh_{x,prop} - ph_{z,prop}) \\ \frac{1}{I_{zz}} (\dot{h}_{z,prop} - qh_{x,prop} + ph_{y,prop}) \end{bmatrix} \quad (7.175)$$

Thus, the effect of adding additional rotating masses simply results in more additive terms which are a function of additional states, e.g. the propeller angular rate would be a control input that would affect the  $\dot{p}$ ,  $\dot{q}$  and  $\dot{r}$  equations whether in the nonlinear state-space EOMs or in the LTI state-space EOMs. Note that additional rotating masses can easily be added to this model, though for an even number of rotating masses, often each is spun in opposite directions to counteract this contribution.

### Variable Mass Effect

Another mass effect is associated with the production of propulsive thrust through expelling mass. While this includes combustion jet engines that combust expelled fuel in air-breathing vehicles, the variable mass effects are typically only significant with rocket engines where the expelled mass includes the fuel *and* oxidizer for the combustion. Note that this derivation will consider the each of these position vectors as represented in inertial navigation frame  $N$  coordinates unless otherwise noted.

Thus, to this end, consider the position vectors of the various elements of a rocket as shown below



where the rocket's inertial *instantaneous* body frame center, i.e. the center of mass, is  $\vec{x}_{B/N}$ , the mass element of the vehicle,  $\rho_V dV$ , with position vectors relative to the body-fixed and navigation frames as  $\vec{x}_{V/B}$  and  $\vec{x}_{V/N}$ , and the expelled mass  $\Delta m$  has position vectors relative to the body-fixed and navigation frames as  $\vec{x}_{Ex/B}$  and  $\vec{x}_{Ex/N}$  with  $\vec{v}_{Ex}$  as the **exit velocity** of  $\Delta m$  relative to the vehicle whose magnitude is often specified for rocket engines by

$$\|\vec{v}_{Ex}\|_2^2 = I_{sp} g_0 \quad (7.176)$$

where  $I_{sp}$  is the specific impulse of the rocket engine and  $g_0$  is the standard acceleration due to gravity.

At time  $t$ , one has for the linear momentum of the variable mass vehicle in the inertial navigation frame

$$\vec{P}_N(t) = \int_{Vol} \rho_V \dot{\vec{x}}_{V/N} dV \quad (7.177)$$

while at time  $t + \Delta t$ , the linear momentum is

$$\vec{P}_N(t + \Delta t) = \int_{Vol} \rho_V \dot{\vec{x}}_{V/N} + \Delta \dot{\vec{x}}_{V/N} \rho_V dV + \Delta m (\dot{\vec{x}}_{Ex/N} + \Delta \dot{\vec{x}}_{Ex/N}) \quad (7.178)$$

where the last term represents the linear momentum balance due to expelled mass, i.e.  $\Delta m < 0$  for expelled mass. Then taking the limit as  $\Delta t \rightarrow 0$ , one has

$$\dot{\vec{P}}_N = \int_{Vol} \frac{d}{dt} (\rho_V \dot{\vec{x}}_{V/N}) dV + \dot{m} \dot{\vec{x}}_{Ex/N} \quad (7.179)$$

Comparing this expression with Newton's translation equation of motion, one can see that the total rate of change in the translational momentum may be rewritten as

$$\dot{\vec{P}}_N = \int_{Vol} \rho_V \vec{g} dV + \int_{Area} d\vec{F}_{ext} + \dot{m} \dot{\vec{x}}_{Ex/N} \quad (7.180)$$

where  $dF_{ext}$  is the external force acting at some infinitesimal surface area, i.e. pressure forces. Note that the inertial velocity of the expelled mass  $dm$  is

$$\dot{\vec{x}}_{Ex/N} = \vec{v}_{Ex} + \vec{v}_{B/N} + \vec{\omega}_{B/N} \times \vec{x}_{Ex/B} \quad (7.181)$$

Thus,

$$\dot{\vec{P}}_N = \int_{Vol} \rho_V \vec{g} dV + \int_{Area} d\vec{F}_{ext} + \dot{m} (\vec{v}_{Ex} + \vec{v}_{B/N} + \vec{\omega}_{B/N} \times \vec{x}_{Ex/B}) \quad (7.182)$$

Furthermore, from the definition of the center of mass at time  $t$ , one has

$$m \vec{x}_{B/N} = \int_{Vol} \rho_V \vec{x}_{V/N} dV \quad (7.183)$$

and at time  $t + \Delta t$ , one has

$$m \vec{x}_{B/N} + \Delta m \vec{x}_{B/N} = \int_{Vol} \rho_V (\vec{x}_{V/N} + \Delta \vec{x}_{V/N}) dV + \Delta m (\vec{x}_{Ex/N} + \Delta \vec{x}_{Ex/N}) \quad (7.184)$$

then as  $\Delta t \rightarrow 0$ , one has

$$\frac{d}{dt} (m \vec{x}_{B/N}) = \vec{P}_N(t) + \dot{m} \vec{x}_{Ex/N} \quad (7.185)$$

where notably  $\dot{m} < 0$ . However, one can rewrite the position of the expelled mass  $dm$  as

$$\vec{x}_{Ex/N} = \vec{x}_{B/N} + \vec{x}_{Ex/B} \quad (7.186)$$

Thus, by the chain rule, one has

$$\dot{m} \vec{x}_{B/N} + m \vec{x}_{B/N} = \vec{P}_N(t) + \dot{m} (\vec{x}_{B/N} + \vec{x}_{Ex/B}) \quad (7.187)$$

Therefore, the translational momentum can be rewritten as

$$\vec{P}_N(t) = m \vec{x}_{B/N} - \dot{m} \vec{x}_{Ex/B} \quad (7.188)$$

and differentiating results in

$$\dot{\vec{P}}_N = m \ddot{\vec{x}}_{B/N} + \dot{m} (\dot{\vec{x}}_{B/N} - \dot{\vec{x}}_{Ex/B}) - \ddot{m} \vec{x}_{Ex/B} \quad (7.189)$$

where the inertial velocity of expelled mass  $dm$  is

$$\dot{\vec{x}}_{Ex/N} = \vec{v}_{B/N} + \vec{\omega}_{B/N} \times \vec{x}_{Ex/B} + \vec{v}_{Ex} \quad (7.190)$$

Combining Equations 7.182 and 7.189, one has

$$\begin{aligned} & m \ddot{\vec{x}}_{B/N} + \dot{m} (\dot{\vec{x}}_{B/N} - \dot{\vec{x}}_{Ex/B}) - \ddot{m} \vec{x}_{Ex/B} \\ &= \int_{Vol} \rho_V \vec{g} dV + \int_{Area} d\vec{F}_{ext} + \dot{m} (\vec{v}_{B/N} + \dot{\vec{x}}_{Ex/B} + \vec{v}_{Ex}) \end{aligned} \quad (7.191)$$

Then, taking gravity to be constant over the volume, defining the total aerodynamic force as

$$\vec{F}_{a,B} = \int_{Body Area} d\vec{F}_{ext} \quad (7.192)$$

and defining the propulsive thrust force to be acting forward on the vehicle as

$$\vec{F}_{p,B} = \dot{m} \vec{v}_{Ex} + \int_{Exit Area} d\vec{F}_{ext} \quad (7.193)$$

where  $\int_{Exit\ Area} d\vec{F}_{ext}$  is known as the **pressure thrust**, one has the translational equation of motion for a variable mass vehicle as

$$m \dot{\vec{v}}_{B/N,B} = \vec{F}_{g,B} + \vec{F}_{p,B} + \vec{F}_{a,B} + 2\dot{m} \dot{\vec{x}}_{Ex/B,B} + \ddot{m} \vec{x}_{Ex/B,B} \quad (7.194)$$

or in terms of body-fixed frame coordinates

$$m \begin{bmatrix} u - qw + rv \\ v - ru + pw \\ w - pv + qu \end{bmatrix} = \vec{F}_{g,B} + \vec{F}_{p,B} + \vec{F}_{a,B} + 2\dot{m} (\dot{\vec{x}}_{Ex/B,B} + \vec{\omega}_{B/N,B} \times \vec{x}_{Ex/B,B}) + \ddot{m} \vec{x}_{Ex/B,B} \quad (7.195)$$

Thus, accounting for the variable mass results in two additive terms as apparent forces in the translational equation of motion. Note that  $\dot{\vec{x}}_{Ex/B,B}$  changes with time as the expelled mass will alter the location of the center of mass, i.e. the origin of the body frame. These terms are often neglected due to  $\vec{x}_{Ex/B}$  being small magnitude relative to the propellant exit velocity and the mass rate is roughly constant.

Continuing the rigid-body EOM analysis, at time  $t$ , one has for the inertial rotational momentum of the variable mass vehicle

$$\vec{H}_N(t) = \int_{Vol} \vec{x}_{V/N} \times \rho_V \dot{\vec{x}}_{V/N} dV \quad (7.196)$$

while at time  $t + \Delta t$ , the rotational momentum is

$$\begin{aligned} \vec{H}_N(t + \Delta t) &= \int_{Vol} (\vec{x}_{V/N} + \Delta \vec{x}_{V/N}) \times \rho_V (\dot{\vec{x}}_{V/N} + \Delta \dot{\vec{x}}_{V/N}) dV \\ &\quad + (\vec{x}_{Ex/N} + \Delta \vec{x}_{Ex/N}) \times \Delta m (\dot{\vec{x}}_{Ex/N} + \Delta \dot{\vec{x}}_{Ex/N}) \end{aligned} \quad (7.197)$$

represents the change in the vehicle's rotational momentum due to the change in mass, i.e.  $\Delta m < 0$  for expelled mass. Then taking the limit as  $\Delta t \rightarrow 0$  (and neglecting higher order  $\Delta$  terms), one has

$$\dot{\vec{H}}_N = \int_{Vol} \frac{d}{dt} (\vec{x}_{V/N} \times \rho_V \dot{\vec{x}}_{V/N}) dV + \vec{x}_{Ex/N} \times \dot{m} \dot{\vec{x}}_{Ex/N} \quad (7.198)$$

Comparing this expression with Euler's rotational equation of motion, one can see that the total rate of change in the rotational momentum may be rewritten as

$$\dot{\vec{H}}_N = \int_{Vol} \vec{x}_{V/N} \times \rho_V \vec{g} dV + \int_{Area} \vec{x}_{V/N} \times d\vec{F}_{ext} + \vec{x}_{Ex/N} \dot{m} \dot{\vec{x}}_{Ex/N} \quad (7.199)$$

Next, noting  $\vec{x}_{V/N} = \vec{x}_{B/N} + \vec{x}_{V/B}$ , one has

$$\begin{aligned} \vec{H}_N &= \vec{x}_{B/N} \times \int_{Vol} \rho_V \dot{\vec{x}}_{V/N} dV + \int_{Vol} \vec{x}_{V/B} \times \rho_V \dot{\vec{x}}_{V/B} \\ &= \vec{x}_{B/N} \times \int_{Vol} \rho_V \dot{\vec{x}}_{V/N} dV + \vec{H}_{B,N} \end{aligned} \quad (7.200)$$

where the first term is the angular momentum of the body-fixed frame and  $\vec{H}_{B,N}$  is the angular momentum of the vehicle in navigation frame coordinates. Recalling

$$\vec{P}_N(t) = \int_{Vol} \rho_V \dot{\vec{x}}_{V/N} dV \quad (7.201)$$

and differentiating with respect to the navigation frame, one has

$$\dot{\vec{H}}_N = \dot{\vec{x}}_{B/N} \times \vec{P}_N + \vec{x}_{B/N} \times \dot{\vec{P}}_N + \dot{\vec{H}}_{B,N} \quad (7.202)$$

Then, substituting for  $\vec{P}_N$  and  $\dot{\vec{P}}_N$  from before, one has

$$\dot{\vec{H}}_N = -\dot{\vec{x}}_{B/N} \times \dot{m} \vec{x}_{Ex/B} + \vec{x}_{B/N} \times \left( \int_{Vol} \rho_V \vec{g} dV + \int_{Area} d\vec{F}_{ext} + \dot{m} \vec{x}_{Ex/N} \right) + \dot{\vec{H}}_{B,N} \quad (7.203)$$

Finally, equating Equations 7.199 and 7.203, one has

$$\begin{aligned} \dot{\vec{H}}_{B,N} &= \dot{\vec{x}}_{B/N} \times \dot{m} \vec{x}_{Ex/B} + \vec{x}_{B/N} \times \left( \int_{Vol} \rho_V \vec{g} dV + \int_{Area} d\vec{F}_{ext} + \dot{m} \vec{x}_{Ex/N} \right) \\ &= \int_{Vol} \vec{x}_{V/N} \times \rho_V \vec{g} dV + \int_{Area} \vec{x}_{V/N} \times d\vec{F}_{ext} + \vec{x}_{Ex/N} \dot{m} \vec{x}_{Ex/N} \end{aligned} \quad (7.204)$$

and noting that the first mass moment about the center of mass is zero for no gravity gradient modeling and

$$\begin{aligned} \vec{x}_{V/N} &= \vec{x}_{B/N} + \vec{x}_{V/B} \\ \vec{x}_{Ex/N} &= \vec{x}_{B/N} + \vec{x}_{Ex/B} \\ \dot{\vec{x}}_{Ex/N} &= \vec{v}_{Ex} + \vec{x}_{E/B} + \dot{\vec{x}}_{B/N} \end{aligned} \quad (7.205)$$

one can rewrite the angular momentum of the variable mass vehicle about its center of mass as

$$\dot{\vec{H}}_{B,N} = \int_{Area} \vec{x}_{V/B} \times d\vec{F}_{ext} + \vec{x}_{Ex/B} \times \dot{m} (\vec{x}_{E/B} + \vec{v}_{Ex}) \quad (7.206)$$

which after separating the external moments into aerodynamic and propulsive contributions, one has

$$\dot{\vec{H}}_{B,N} = \vec{M}_{a,N} + \vec{M}_{p,N} + \vec{x}_{Ex/B} \times \dot{m} (\vec{x}_{E/B,B} + \vec{\omega}_{B/N} \times \vec{x}_{Ex/B}) \quad (7.207)$$

$$\dot{\vec{H}}_{B,N} = C_{N \leftarrow B} \begin{bmatrix} I_{xx}L \\ I_{yy}M \\ I_{zz}N \end{bmatrix} + \vec{x}_{Ex/B} \times \dot{m} (\vec{x}_{E/B,B} + \vec{\omega}_{B/N} \times \vec{x}_{Ex/B}) \quad (7.208)$$

where  $\vec{M}_p = \vec{x}_{Ex/B} \times \vec{F}_p$  and the additive triple product is known as the **jet-damping effect** because it tends to act against the vehicle's angular motion. Note that this is proportional to the vehicle's angular velocity, the square of the distance from the nozzle exit to the center of mass, and the mass flow rate. Thus, this damping is more significant for maneuvering, long vehicles with large mass flow rates, e.g., launch vehicles.

## References

For more information, please refer to the following

- Curtis, H. D., "9.3 Equations of Translational Motion," *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 408-410

- Curtis, H. D., “9.4 Equations of Rotational Motion,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 420-414
- Curtis, H. D., “9.5 Moments of Inertia,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 414-435
- Nelson, R. C., “3.2 Derivation of Rigid Body Equations of Motion,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 49-434
- Nelson, R. C., “3.3 Orientation and Position of the Airplane,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 101-103
- Schmidt, D. K., “2.3 Reference and Perturbation Equations - Flat Earth,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 42-48
- Sidi, M. J., “4.2 Angular Momentum and the Inertia Matrix,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 88-90
- Sidi, M. J., “4.4 Moment-of-Inertia Matrix in Selected Axis Frames,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 90-95
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “1.7 Rigid-Body Dynamics,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 34-43
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “2.3 Aircraft Forces and Moments,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 75-101
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “2.5 The Nonlinear Aircraft Model,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 108-116

## 7.4 Elastic-Body Aerospace Vehicle Dynamics

The perspective in this course on elastic-body dynamics is the ability to study the effects of elastic deformation on the flight dynamics of a vehicle, not to study purely structural/aeroelastic phenomena such as flutter or divergence. Thus, only the lower frequency modes are typically of interest for addition to the rigid-body modes developed in 6-DOF flight dynamics. Though all real vehicles are elastic, in many cases, this type of vibration analysis may only be necessary to check if the rigid-body assumption can be made for the modeling of the flight vehicle. In general, larger flight vehicles typically require some elastic modeling as the natural frequencies will be lower and more likely to interact with the rigid-body or flight controller modes of the vehicle.

The treatment for elastic vehicles will continue to make use of Lagrangian mechanics and energy concepts which will be stated here without proof, as typically these are developed in a graduate level dynamics course.

These concepts will allow the derivation of important results for the reference frame requirements for elastic vehicles, namely the mean axes, which will be developed in this lecture. Previous sections have derived generalized coordinates to model vibration problems where the vibration modes are orthogonal to the rigid-body modes and demonstrated this for some simple lumped-mass models to assist in visualizing the construction of these elastic-body dynamic models. As such, these types of coordinate frames satisfy the mean-axis constraints which will be used for the mean-axes body-fixed frame derived in the equations of motion for describing elastic-body flight dynamics in the subsequent sections. Before developing the concept of mean-axes, first assume that the navigation frame  $N$  is inertial, i.e., the “flat-Earth” model.

## Lagrangian Energies for Vehicles

Let the kinetic energy of the entire vehicle be

$$T = \frac{1}{2} \int_{Vol} \vec{v}_N \cdot \vec{v}_N \rho_V dV \quad (7.209)$$

where  $\vec{v}_N$  is the inertial velocity of a mass element of the vehicle, i.e.

$$\vec{v}_N = \vec{v}_{B/N,N} + \vec{v}_B + \vec{\omega}_{B/N} \times \vec{x}_B \quad (7.210)$$

where  $\vec{v}_{B/N}$  is the velocity of the body-fixed frame relative to the navigation frame,  $\vec{v}_B$  is the velocity of the mass element in the body-fixed frame,  $\vec{\omega}_{B/N}$  is the angular velocity of the body-fixed frame relative to the navigation frame, and  $\vec{x}_B$  is the position of the mass element in the body-fixed frame, i.e. relative to the center of mass. In addition, one has

$$\vec{x}_N = \vec{x}_B + \vec{x}_{B/N} \quad (7.211)$$

where  $\vec{x}_{B/N}$  is the position of the origin of the body-fixed frame relative to the navigation frame. Then, by substitution, one has

$$\begin{aligned} T = \frac{1}{2} \int_{Vol} & [ \vec{v}_{B/N,N} \cdot \vec{v}_{B/N,N} \\ & + 2\vec{v}_{B/N,N} \cdot \vec{v}_B \\ & + 2(\vec{v}_{B/N,N} + \vec{v}_B) \cdot (\vec{\omega}_{B/N} \times \vec{x}_B) \\ & + \vec{v}_B \cdot \vec{v}_B \\ & + (\vec{\omega}_{B/N} \times \vec{x}_B) \cdot (\vec{\omega}_{B/N} \times \vec{x}_B) ] \rho_V dV \end{aligned} \quad (7.212)$$

Furthermore, the potential energy of the vehicle includes gravitational potential energy  $U_g$  and elastic strain energy  $U_e$ . The gravitational potential energy for the vehicle can be modeled as

$$\begin{aligned} U_g &= - \int_{Vol} \vec{g} \cdot \vec{x}_N \rho_V dV \\ U_g &= - \int_{Vol} \vec{g} \cdot (\vec{x}_B + \vec{x}_{B/N}) \rho_V dV \end{aligned} \quad (7.213)$$

where  $\vec{g}$  is the acceleration due to gravity. The elastic strain energy is the energy stored in an elastic structure due to its deformation resulting from some applied force. The strain energy is the negative of the work done

on the structure by the applied force, and the work is the force acting over a displacement  $\vec{d}_e$ . Thus, the position of a mass element of the vehicle can be represented as

$$\vec{x} = \vec{x}_{r-b} + \vec{d}_e \quad (7.214)$$

where  $\vec{x}_{r-b}$  is the position of the mass element in terms of its undeformed or rigid-body position. Furthermore, since  $\vec{x}_{r-b,B}$  is invariant with respect to the body-fixed frame, one has

$$\vec{v}_B = \dot{\vec{d}}_{e,B} \quad (7.215)$$

Thus, using D'Alembert's principle to express the force on a mass element in terms of the mass of the element and its acceleration, one has for the elastic strain energy

$$U_e = -\frac{1}{2} \int_{Vol} \ddot{\vec{d}}_{e,B} \cdot \vec{d}_{e,B} \rho_V dV \quad (7.216)$$

### Mean-Axes Body Frame

For rigid vehicle dynamics, the selection of body-fixed frame axes was arbitrary with the only requirement being the body-fixed frame origin as the vehicle's center of mass. Such a construction decoupled the rotation mode and the translation mode of the dynamics. For elastic vehicle dynamics, one must further consider the existence of any number of vibration modes which results in *additional* requirements for a body-fixed frame to exhibit decoupled dynamic modes, namely the **mean-axes constraints** which define coordinate axes about which the relative translational and angular momenta (about the center of mass) due to elastic vibrations are zero, i.e.

$$\int_{Vol} \vec{d}_{e,B} \rho_V dV = \int_{Vol} \vec{x}_B \times \vec{d}_{e,B} \rho_V dV = \vec{0} \quad (7.217)$$

This "special" **mean-axes body-fixed frame** can be shown to always exist for an elastic-body which will be assumed here and not proven. By substitution for these quantities from the relations in the previous section, one has

$$\int_{Vol} \frac{d}{dt} (\vec{x}_{r-b,B} + \vec{d}_{e,B}) \rho_V dV = \vec{0} \quad (7.218)$$

and

$$\int_{Vol} \vec{x}_{r-b,B} \times \vec{d}_{e,B} \rho_V dV + \int_{Vol} \vec{d}_{e,B} \times \vec{d}_{e,B} \rho_V dV = \vec{0} \quad (7.219)$$

which if the elastic displacement,  $\vec{d}_{e,B}$ , is sufficiently small such that only linear effects are considered, then one can neglect the moment from the elastic displacements, and one has the *practical mean-axes constraints* written as

$$\int_{Vol} \vec{d}_{e,B} \rho_V dV = \int_{Vol} \vec{x}_{r-b} \times \vec{d}_{e,B} \rho_V dV = \vec{0} \quad (7.220)$$

which are analogous to the modal orthogonality constraints for the lumped-mass systems considered earlier. Thus, while the mean-axes will be used for theoretical development, in practice, one uses the practical mean-axes constraints to confirm the selected axes are mean-axes through mutual orthogonality among all modes.

To demonstrate this, assume a free-vibration analysis has been performed as previously described yielding the  $n$  free-vibration mutually orthogonal mode shapes,  $\vec{v}_i$ , and frequencies,  $\omega_i = \sqrt{\lambda_i}$ , (i.e. eigenvectors and eigenvalues) including both the rigid-body and elastic modes. Then, the elastic vibrations can be expressed as

$$\vec{d}_{e,B} = \sum_{i=1}^n \vec{v}_i(\vec{x})\eta_i(t) \quad (7.221)$$

where  $\eta_i(t)$  is the generalized coordinate associated with the  $i$ -th vibration mode. In general, each mode shape,  $\vec{v}_i(\vec{x})$ , is a vector with components defined in the body-fixed frame with each component a function of the  $\vec{x}$  location on the *undeformed* structure. With this analysis, the *practical mean-axes constraints* can be rewritten as

$$\int_{Vol} \vec{d}_{e,B} \rho_V dV = \sum_{i=1}^n \dot{\eta}_i(t) \left( \int_{Vol} \vec{v}_i(\vec{x}) \rho_V dV \right) = 0 \quad (7.222)$$

and

$$\int_{Vol} \vec{x}_{r-b} \times \vec{d}_{e,B} \rho_V dV = \sum_{i=1}^n \dot{\eta}_i(t) \left( \int_{Vol} \vec{x}_{r-b} \times \vec{v}_i(\vec{x}) \rho_V dV \right) = 0 \quad (7.223)$$

which are satisfied as the integrals inside the parentheses above correspond to the momenta conservation requirements and the selected vibration modes are, by design, orthogonal to the rigid-body translation and rotation modes (with respect to the mass distribution here).

### Mean-Axis Constraints on Energies

Now, one can apply the *mean-axis constraints* to the energies to greatly simplify these equations. For the first term of the kinetic energy, as the velocity of the center of mass is independent of the volume, one has

$$\begin{aligned} \int_{Vol} \vec{v}_{B/N,N} \cdot \vec{v}_{B/N,N} \rho_V dV &= \vec{v}_{B/N,N} \cdot \vec{v}_{B/N,N} \int_{Vol} \rho_V dV \\ &= m \vec{v}_{B/N,N} \cdot \vec{v}_{B/N,N} \end{aligned} \quad (7.224)$$

where  $m$  is the total mass of the vehicle. The second term of the kinetic energy becomes zero as

$$\int_{Vol} \vec{v}_{B/N,N} \cdot \vec{d}_{e,B} \rho_V dV = \vec{v}_{B/N,N} \cdot \int_{Vol} \vec{d}_{e,B} \rho_V dV = 0 \quad (7.225)$$

For the third term of the kinetic energy, one has

$$\int_{Vol} (\vec{v}_{B/N,N} + \vec{d}_{e,B}) \cdot (\vec{\omega}_{B/N} \times \vec{x}_B) \rho_V dV = \vec{v}_{B/N,N} \cdot \left( \vec{\omega}_{B/N} \times \int_{Vol} \vec{x}_B \rho_V dV \right) + \int_{Vol} \vec{d}_{e,B} \cdot (\vec{\omega}_{B/N} \times \vec{x}_B) \rho_V dV \quad (7.226)$$

recalling the requirement for the origin of the mean-axis body-fixed frame be the instantaneous center of mass states that

$$\int_{Vol} \vec{x}_B \rho_V dV = 0 \quad (7.227)$$

Thus, one has

$$\int_{Vol} (\vec{v}_{B/N,N} + \vec{d}_{e,B}) \cdot (\vec{\omega}_{B/N} \times \vec{x}_B) \rho_V dV = \left( \int_{Vol} \vec{x}_B \times \vec{v}_{B/N} \rho_V dV \right) \cdot \vec{\omega}_{B/N} = 0 \quad (7.228)$$

The fourth term of the kinetic energy can be rewritten using the mode shapes and generalized coordinate summations for the displacement *velocity*, i.e.

$$\begin{aligned} \int_{Vol} \vec{v}_B \cdot \vec{v}_B \rho_V dV &= \frac{1}{2} \int_{Vol} \left( \sum_{i=1}^n \vec{v}_i(\vec{x}) \dot{\eta}_i \cdot \sum_{i=j}^n \vec{v}_j(\vec{x}) \dot{\eta}_j \right) \rho_V dV \\ &= \int_{Vol} \left( \sum_{i=1}^n \vec{v}_i(\vec{x}) \cdot \vec{v}_i(\vec{x}) \dot{\eta}_i^2 \right) \rho_V dV \\ &= \sum_{i=1}^n \mathcal{M}_i \dot{\eta}_i^2 \end{aligned} \quad (7.229)$$

where  $\mathcal{M}_i$  is the *i*-th generalized mass of the *i*-th vibration mode. This simplification is due to the mutual orthogonality of the vibration modes, i.e.

$$\int_{Vol} \vec{v}_i \cdot \vec{v}_j \rho_V dV = \begin{cases} 0 & i \neq j \\ \mathcal{M}_i & i = j \end{cases} \quad (7.230)$$

The fifth term of the kinetic energy can be rewritten in terms of the inertia matrix as

$$\int_{Vol} (\vec{\omega}_{B/N} \times \vec{x}_B) \cdot (\vec{\omega}_{B/N} \times \vec{x}_B) \rho_V dV = \vec{\omega}_{B/N}^T I_G \vec{\omega}_{B/N} \quad (7.231)$$

Thus, the kinetic energy can be rewritten as

$$T = \frac{1}{2} m \vec{v}_{B/N,N}^T \vec{v}_{B/N,N} + \frac{1}{2} \vec{\omega}_{B/N}^T I_G \vec{\omega}_{B/N} + \frac{1}{2} \sum_{i=1}^n \mathcal{M}_i \dot{\eta}_i^2 \quad (7.232)$$

Next, applying the *mean-axis constraints* to the gravitational potential energy, one has

$$\begin{aligned} U_g &= - \int_{Vol} \vec{g} \cdot (\vec{x}_B + \vec{x}_{B/N}) \rho_V dV \\ U_g &= - \vec{g} \cdot \int_{Vol} \vec{x}_B \rho_V dV - \vec{g} \cdot \vec{x}_{B/N} \int_{Vol} \rho_V dV \\ U_g &= -m \vec{g}^T \vec{x}_{B/N} \end{aligned} \quad (7.233)$$

Finally, applying the *mean-axis constraints* to the elastic strain energy, one has

$$U_e = -\frac{1}{2} \int_{Vol} \sum_{i=1}^n \vec{v}_i(\vec{x}) \ddot{\eta}_i(t) \cdot \sum_{i=1}^n \vec{v}_i(\vec{x}) \eta_i(t) \rho_V dV \quad (7.234)$$

and due to the orthogonality of the modes, one has

$$U_e = -\frac{1}{2} \int_{Vol} \sum_{i=1}^n \vec{v}_i(\vec{x}) \cdot \vec{v}_i(\vec{x}) \ddot{\eta}_i(t) \eta_i(t) \rho_V dV \quad (7.235)$$

Then, recalling the sinusoidal solution for the modal coordinates, i.e.

$$\eta_i(t) = A_i \cos(\omega_i t + \Gamma_i) \quad (7.236)$$

one has

$$U_e = -\frac{1}{2} \sum_{i=1}^n \omega_i^2 \eta_i^2(t) \mathcal{M}_i \quad (7.237)$$

## Generalized Coordinate Selection

Before applying Lagrange's equation to the energies to obtain the elastic vehicle equations of motion, one must select suitable generalized coordinates. First, one can select the inertial position of the body-fixed frame's origin (i.e. the center of mass) as the coordinates

$$\vec{x}_{B/N} = \begin{bmatrix} x_{B/N} \\ y_{B/N} \\ z_{B/N} \end{bmatrix} \quad (7.238)$$

which was represented in rigid vehicle dynamics without the referencing slash which is appropriate for rigid bodies as the relative position of any point on the body can be represented with the location of the center of mass *and* the attitude of the vehicle. Though for elastic vehicles this is no longer the case, the  $B/N$  subscript will be dropped to easily compare the elastic vehicle dynamics with the rigid-body dynamics, i.e.

$$\vec{x}_{B/N,N} = \begin{bmatrix} x, N \\ y, N \\ z, N \end{bmatrix} \quad (7.239)$$

will be the selected generalized coordinates. Also recall that  $z_N = -h$  for the flat-Earth approximation. Thus, the velocity of the body-fixed frame's origin expressed in the navigation frame is

$$\vec{v}_{B/N,N} = \begin{bmatrix} \dot{x}_N \\ \dot{y}_N \\ \dot{z}_N \end{bmatrix} \quad (7.240)$$

and the body-fixed frame axes as

$$\vec{v}_{B/N,B} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} \quad (7.241)$$

Secondly, one can select the angular velocity of the body-fixed frame relative to the navigation frame as

$$\vec{\omega}_{B/N} = \begin{bmatrix} p \\ q \\ r \end{bmatrix} \quad (7.242)$$

which can also be expressed in either the body-fixed frame or the navigation frame (with an appropriate subscript). The generalized coordinates for this derivation will use the body-fixed frame axes. Thirdly, one can select the 3-2-1 Euler angles  $\psi, \theta, \phi$  to define the orientation of the mean-axes body-fixed frame with respect to the navigation frame. Recall that the Euler angles are related to the angular velocity by the equations

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\sin \theta \\ 0 & \cos \phi & -\sin \phi \cos \theta \\ 0 & -\sin \phi & \cos \phi \cos \theta \end{bmatrix} \begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} \quad (7.243)$$

$$\begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} 1 & \sin \phi \tan \theta & \cos \phi \tan \theta \\ 0 & \cos \phi & -\sin \phi \\ 0 & \sin \phi \sec \theta & \cos \phi \sec \theta \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix} \quad (7.244)$$

Thus, one can select the generalized coordinates as

$$\vec{q} = [x \ y \ z \ \phi \ \theta \ \psi \ \eta_i, i = 1, 2, \dots]^T \quad (7.245)$$

which is the same as for rigid vehicle dynamics with the addition of the modal coordinates corresponding to the mutually orthogonal mode shapes.

Using these definitions, one can write the kinetic energy of the elastic vehicle as

$$T = \frac{1}{2}m [\dot{x}_N \ \dot{y}_N \ \dot{z}_N] \begin{bmatrix} \dot{x}_N \\ \dot{y}_N \\ \dot{z}_N \end{bmatrix} + \frac{1}{2} [p \ q \ r] I_G \begin{bmatrix} p \\ q \\ r \end{bmatrix} + \frac{1}{2} \sum_{i=1}^n \mathcal{M}_i \dot{\eta}_i^2 \quad (7.246)$$

the gravitational potential energy as

$$U_g = -mgz_N = mgh \quad (7.247)$$

and the elastic strain energy remains the same.

$$U_e = -\frac{1}{2} \sum_{i=1}^n \omega_i^2 \eta_i^2(t) \mathcal{M}_i \quad (7.248)$$

Using Lagrange's equation in vector form, i.e.,

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{\vec{q}}} \right) - \frac{\partial T}{\partial \vec{q}} + \frac{\partial U}{\partial \vec{q}} = \vec{Q}^T = \frac{\partial \delta W}{\partial \delta \vec{q}} \quad (7.249)$$

one can derive the equations of motion for elastic flight vehicles in three coupled vector equations of motion: translation, rotation, and vibration.

### Translation Equations of Motion

For a constant-mass vehicle, consider the inertial center of mass coordinates,  $\vec{x}_N = [x_N \ y_N \ z_N]^T$ , chosen as the translation generalized coordinates, and applying Lagrange's equation for the translational kinetic energy, i.e.

$$T_{tran} = \frac{1}{2}m [\dot{x}_N \ \dot{y}_N \ \dot{z}_N] \begin{bmatrix} \dot{x}_N \\ \dot{y}_N \\ \dot{z}_N \end{bmatrix} \quad (7.250)$$

one has

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{\vec{x}}_N} \right) - \frac{\partial T}{\partial \vec{x}_N} = m \begin{bmatrix} \ddot{x}_N \\ \ddot{y}_N \\ \ddot{z}_N \end{bmatrix} = m \ddot{\vec{x}}_N \quad (7.251)$$

and for the gravitational potential energy, one has

$$\frac{\partial U_g}{\partial \vec{x}_N} = \begin{bmatrix} 0 \\ 0 \\ -mg \end{bmatrix} \quad (7.252)$$

Including the generalized forces in the navigation frame  $\vec{Q}_N$ , one has the following translation EOMs

$$\ddot{\vec{x}}_N = \begin{bmatrix} \ddot{x}_N \\ \ddot{y}_N \\ \ddot{z}_N \end{bmatrix} = \begin{bmatrix} \frac{Q_{x,N}}{m} \\ \frac{Q_{y,N}}{m} \\ \frac{Q_{z,N}}{m} + g \end{bmatrix} \quad (7.253)$$

However, for flight dynamics, one typically desires to convert this to the body frame accelerations and velocities, i.e. recalling the conversion of the velocity by

$$\ddot{\vec{x}}_N = \ddot{\vec{x}}_B + \vec{\omega}_{B/N} \times \dot{\vec{x}}_B \quad (7.254)$$

one has using the definitions for the body frame linear and angular velocity components

$$\begin{aligned} \ddot{\vec{x}}_N &= \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \begin{bmatrix} p \\ q \\ r \end{bmatrix} \times \begin{bmatrix} u \\ v \\ w \end{bmatrix} \\ &= \begin{bmatrix} \dot{u} - rv + qw \\ \dot{v} + ru - wp \\ \dot{w} - qu + pv \end{bmatrix} \end{aligned} \quad (7.255)$$

For the generalized forces for flight vehicle, recall that the net force (besides gravitational) are the net aerodynamic and propulsive which are defined in the body frame as

$$\vec{F}_{a,B} + \vec{F}_{p,B} = \begin{bmatrix} mX \\ mY \\ mZ \end{bmatrix} \quad (7.256)$$

The virtual work,  $\delta W_F$ , is done by these forces as virtual displacements in the body frame, i.e.  $\delta \vec{x}_B = [\delta x_B \ \delta y_B \ \delta z_B]^T$ , which can be written as

$$\delta W_F = [mX \ mY \ mZ] \begin{bmatrix} \delta x_B \\ \delta y_B \\ \delta z_B \end{bmatrix} \quad (7.257)$$

In terms of the selected generalized coordinates, one can use the DCM from  $N \rightarrow B$  which is a function of the Euler angles as follows:

$$\delta W_F = [mX \ mY \ mZ] C_{B \leftarrow N}(\phi, \theta, \psi) \begin{bmatrix} \delta x_N \\ \delta y_N \\ \delta z_N \end{bmatrix} \quad (7.258)$$

Thus the generalized forces become

$$\begin{aligned} \begin{bmatrix} Q_{x,N} \\ Q_{y,N} \\ Q_{z,N} \end{bmatrix} &= \frac{\partial \delta W_F}{\partial \delta \vec{x}_N} = [mX \ mY \ mZ] C_{B \leftarrow N}(\phi, \theta, \psi) \\ &= C_{B \leftarrow N}^T(\phi, \theta, \psi) \begin{bmatrix} mX \\ mY \\ mZ \end{bmatrix} \end{aligned} \quad (7.259)$$

which can be rewritten in body frame coordinates as simply

$$\begin{bmatrix} Q_{x,B} \\ Q_{y,B} \\ Q_{z,B} \end{bmatrix} = C_{B \leftarrow N}(\phi, \theta, \psi) C_{B \leftarrow N}^T(\phi, \theta, \psi) \begin{bmatrix} mX \\ mY \\ mZ \end{bmatrix} = \begin{bmatrix} mX \\ mY \\ mZ \end{bmatrix} \quad (7.260)$$

as expected. Lastly, for the gravitational force, one has in the body frame coordinates

$$\begin{bmatrix} -mg \sin \theta \\ mg \cos \theta \sin \phi \\ mg \cos \theta \cos \phi \end{bmatrix} = C_{B \leftarrow N}(\phi, \theta, \psi) \begin{bmatrix} 0 \\ 0 \\ mg \end{bmatrix} \quad (7.261)$$

Thus, one has the same rigid vehicle translation equations of motion for elastic vehicle translation equations of motion, i.e.

$$\begin{bmatrix} \dot{u} - rv + qw \\ \dot{v} + ru - wp \\ \dot{w} - qu + pv \end{bmatrix} = \begin{bmatrix} X - g \sin \theta \\ Y + g \cos \theta \sin \phi \\ Z + g \cos \theta \cos \phi \end{bmatrix} \quad (7.262)$$

and any elastic deformation effects will enter by the aerodynamic and propulsive forces.

### Rotation Equations of Motion

For a constant-mass vehicle, consider the Euler angles,  $\vec{q}_\perp = [\phi \ \theta \ \psi]^T$ , chosen as the rotation generalized coordinates. However as the Euler angles represent three sequential rotations, the relationship between the body frame angular velocity (which appears in the kinetic energy) and the Euler angles is

$$\vec{\omega}_{B/N} = \begin{bmatrix} p \\ q \\ r \end{bmatrix} = C_\omega(\vec{q}_\perp) \dot{\vec{q}}_\perp = \begin{bmatrix} 1 & 0 & -\sin \phi \\ 0 & \cos \phi & \cos \theta \sin \phi \\ 0 & -\sin \phi & \cos \theta \cos \phi \end{bmatrix} \begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} \quad (7.263)$$

Thus, one has

$$\frac{\partial \vec{\omega}_{B/N}}{\partial \dot{\vec{q}}_\perp} = C_\omega \quad (7.264)$$

Next, note that the rotational kinetic energy becomes

$$T_{rot} = \frac{1}{2} \vec{\omega}_{B/N}^T I_G \vec{\omega}_{B/N} \quad (7.265)$$

Thus, one has

$$\frac{\partial T}{\partial \vec{\omega}_{B/N}} = \vec{\omega}_{B/N}^T I_G \quad (7.266)$$

Furthermore, one can show

$$\frac{\partial \vec{\omega}_{B/N}}{\partial \vec{q}_\perp} = \begin{bmatrix} 0 & -\dot{\psi} \cos \theta & 0 \\ \dot{\psi} \cos \theta \cos \phi - \dot{\theta} \sin \phi & -\dot{\psi} \sin \theta \sin \phi & 0 \\ -\dot{\psi} \cos \theta \sin \phi - \dot{\theta} \cos \phi & -\dot{\psi} \sin \theta \cos \phi & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\dot{\psi} \cos \theta & 0 \\ r & -\dot{\psi} \sin \theta \sin \phi & 0 \\ -q & -\dot{\psi} \sin \theta \cos \phi & 0 \end{bmatrix} \quad (7.267)$$

Then, applying Lagrange's equation for the kinetic energy, one has

$$\frac{d}{dt} \frac{\partial T}{\partial \dot{\vec{q}}_\perp} - \frac{\partial T}{\partial \vec{q}_\perp} = \frac{d}{dt} \left( \frac{\partial T}{\partial \omega_{B/N}} \frac{\partial \omega_{B/N}}{\partial \dot{\vec{q}}_\perp} \right) - \frac{\partial T}{\partial \omega_{B/N}} \frac{\partial \vec{\omega}}{\partial \vec{q}_\perp} \quad (7.268)$$

which after the matrix multiplications and some algebra, it can be shown

$$\frac{d}{dt} \frac{\partial T}{\partial \dot{\vec{q}}_\perp} - \frac{\partial T}{\partial \vec{q}_\perp} = C_\omega^T (I_G \dot{\omega}_{B/N} + \vec{\omega}_{B/N} \times I_G \vec{\omega}_{B/N}) = \vec{Q}_\perp \quad (7.269)$$

where the virtual work associated with the net moment on the flight vehicle in the body frame can be related to the virtual angular displacements simply by

$$\delta W_M = [I_{xx}L \quad I_{yy}M \quad I_{zz}N] \begin{bmatrix} \delta\phi \\ \delta\theta \\ \delta\psi \end{bmatrix} \quad (7.270)$$

where the infinitesimal rotations due to the angular displacements are related to the Euler angles themselves, by

$$\begin{bmatrix} \delta\phi \\ \delta\theta \\ \delta\psi \end{bmatrix} = C_\omega \vec{q}_\perp \quad (7.271)$$

Thus, in terms of the generalized coordinates

$$\vec{Q}_\perp^T = \frac{\partial \delta W_M}{\partial \vec{q}_\perp} = [I_{xx}L \quad I_{yy}M \quad I_{zz}N] C_\omega \quad (7.272)$$

$$\vec{Q}_\perp = C_\omega^T \begin{bmatrix} I_{xx}L \\ I_{yy}M \\ I_{zz}N \end{bmatrix} \quad (7.273)$$

Thus, one has the same rigid vehicle rotation equations of motion for elastic vehicle rotation equations of motion, i.e.

$$I_G \dot{\omega}_{B/N} + \vec{\omega}_{B/N} \times I_G \vec{\omega}_{B/N} = \begin{bmatrix} I_{xx}L \\ I_{yy}M \\ I_{zz}N \end{bmatrix} \quad (7.274)$$

which for  $I_{xy} = I_{yz} = 0$  can be written out as

$$\begin{bmatrix} \dot{p} + \frac{I_{zz}-I_{yy}}{I_{xx}} qr - \frac{I_{xz}}{I_{xx}} (\dot{r} + pq) \\ \dot{q} + \frac{I_{xx}-I_{zz}}{I_{yy}} pr - \frac{I_{xy}}{I_{yy}} (r^2 - p^2) \\ \dot{r} + \frac{I_{yy}-I_{xx}}{I_{zz}} pq - \frac{I_{xz}}{I_{zz}} (\dot{p} - qr) \end{bmatrix} = \begin{bmatrix} L \\ M \\ N \end{bmatrix} \quad (7.275)$$

and any elastic deformation effects will enter by the aerodynamic and propulsive moments.

## Vibration Equations of Motion

For a constant-mass vehicle, consider the  $n$  vibration coordinates,  $\vec{\eta} = [\eta_1 \cdots \eta_n]^T$ , chosen as the vibration generalized coordinates. Applying Lagrange's equation for the vibration kinetic energy and elastic strain energy for each individual modal coordinate, one has

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{\eta}_i} \right) - \frac{\partial T}{\partial \eta_i} + \frac{\partial U_e}{\partial \eta_i} = Q_i = \frac{\partial \delta W}{\partial \delta \eta_i} \quad (7.276)$$

one has  $n$  equations of motion for the vibration coordinates of the form

$$\ddot{\eta}_i + \omega_i^2 \eta_i = \frac{Q_i}{M_i} \quad i = 1, \dots, n \quad (7.277)$$

where  $M$  is the generalized mass and  $Q_i$  is the generalized force, each associated with the  $i$ -th vibration mode.

For aerodynamics modeling for the generalized mass, let the external pressure distribution acting on the surface of the vehicle's structure at point  $\vec{x}_B$  be defined as  $\vec{P}(\vec{x}_B)$ . By construction, the integral of this pressure distribution will result in the aerodynamic and propulsive forces and moments  $X, Y, Z, L, M, N$ . Then, the local elastic deformation of the structure can be written in terms of the modes as

$$\delta d_e(\vec{x}_B) = \sum_{i=1}^n \vec{v}_i \delta \eta_i(t) \quad (7.278)$$

Thus, the virtual work done due to the pressure  $\vec{P}$  located at  $\vec{x}_B$  on the structure is

$$d\delta W_P = \vec{P}(\vec{x}_B) \cdot \sum_{i=1}^n \vec{v}_i(\vec{x}_B) \delta \eta_i(t) dS \quad (7.279)$$

where  $dS$  is the infinitesimal surface area over which the pressure applies. Thus, the total virtual work is

$$\begin{aligned} \delta W_P &= \int_{Area} \vec{P}(\vec{x}_B) \cdot \sum_{i=1}^n \vec{v}_i(\vec{x}_B) \delta \eta_i(t) dS \\ &= \sum_{i=1}^n \int_{Area} \vec{P}(\vec{x}_B) \cdot \vec{v}_i(\vec{x}_B) dS \delta \eta_i(t) \end{aligned} \quad (7.280)$$

Finally, the  $n$  vibration equations of motion become

$$\ddot{\eta}_i + \omega_i^2 \eta_i = \frac{1}{M_i} \int_{Area} \vec{P}(\vec{x}_B) \cdot \vec{v}_i(\vec{x}_B) dS \quad i = 1, \dots, n \quad (7.281)$$

which connects the aerodynamic pressure with the structural deformations.

### Point Motion on Elastic Vehicle

With these three sets of equations of motion, one is now able to describe the motion of any point on the elastic flight vehicle which is a combination of the rigid body motion *and* the vibration motion. Specifically, the position of any point on the elastic vehicle as defined in the navigation frame is a linear combination of the body frame's origin (i.e. the center of mass), the rigid body position of the point,  $\vec{x}_{r-b}$ , and the elastic displacement,  $d_e$ , i.e.

$$\vec{x}_N = \vec{x}_{B/N} + \vec{x}_{r-b} + d_e(\vec{x}_B, t) \quad (7.282)$$

or

$$\vec{x}_N = \vec{x}_{B/N} + \vec{x}_{r-b} + \sum_{i=1}^n \vec{v}_i(\vec{x}_B) \eta_i(t) \quad (7.283)$$

where each of these terms can be determined by the elastic vehicle equations of motion (assuming one has a solution for the free-vibration problem for the mode shapes). Note that here the position of the instantaneous center of mass is governed by the rigid body translation equations of motion which typically uses the body frame coordinates for the velocity of the body frame, thus, one should recall

$$\dot{\vec{x}}_{B/N} = \begin{bmatrix} \dot{x}_N \\ \dot{y}_N \\ \dot{z}_N \end{bmatrix} = C_{N \leftarrow B} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \quad (7.284)$$

based on the final equations given previously.

### Elastic Aerospace Vehicle Dynamics

Thus, the elastic aerospace vehicle equations of motion can be written together as

$$\begin{bmatrix} \dot{u} + qw - rv \\ \dot{v} + ru - pw \\ \dot{w} + pv - qu \\ \dot{p} + \frac{I_{zz} - I_{yy}}{I_{xx}} qr - \frac{I_{xz}}{I_{xx}} (\dot{r} + pq) \\ \dot{q} + \frac{I_{xx} - I_{zz}}{I_{yy}} pr - \frac{I_{xz}}{I_{yy}} (r^2 - p^2) \\ \dot{r} + \frac{I_{yy} - I_{xx}}{I_{zz}} pq - \frac{I_{xz}}{I_{zz}} (\dot{p} - qr) \end{bmatrix} = \begin{bmatrix} X - g \sin \theta \\ Y + g \cos \theta \sin \phi \\ Z + g \cos \theta \cos \phi \\ L \\ M \\ N \end{bmatrix} \quad (7.285)$$

$$\ddot{\eta}_i + 2\zeta_i \omega_i \dot{\eta}_i + \omega_i^2 \eta_i = \frac{Q_i}{M_i}, \quad i = 1, \dots, n$$

And defining the linear and angular velocity of the body frame as the rigid state vector

$$\vec{x}_{rig} = [u \quad v \quad w \quad p \quad q \quad r]^T \quad (7.286)$$

the modal coordinates and modal coordinate rates as the vibration state vector

$$\vec{x}_{vib} = [\eta_1 \quad \dots \quad \eta_n \quad \dot{\eta}_1 \quad \dots \quad \dot{\eta}_n]^T \quad (7.287)$$

With these definitions, the nonlinear equations of motion for an elastic flight vehicle can be rewritten as

$$\begin{aligned} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -\frac{I_{xz}}{I_{zz}} \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -\frac{I_{xz}}{I_{zz}} & 0 & 1 \end{bmatrix} \dot{\vec{x}}_{rig} = & \begin{bmatrix} rv - qw - g \sin \theta \\ pw - ru + g \cos \theta \sin \phi \\ qu - pv + g \cos \theta \cos \phi \\ \frac{I_{yy}-I_{zz}}{I_{xx}} qr + \frac{I_{xz}}{I_{xx}} pq \\ \frac{I_{zz}-I_{xx}}{I_{yy}} pr + \frac{I_{xz}}{I_{yy}} (r^2 - p^2) \\ \frac{I_{xx}-I_{yy}}{I_{zz}} pq - \frac{I_{xz}}{I_{zz}} qr \end{bmatrix} + \begin{bmatrix} X \\ Y \\ Z \\ L \\ M \\ N \end{bmatrix} \\ \dot{\vec{x}}_{vib} = & \begin{bmatrix} 0_{n \times n} & I_{n \times n} \\ -\Omega^2 & -2\Omega_\zeta \end{bmatrix} \vec{x}_{vib} + \begin{bmatrix} \vec{0}_n \\ \frac{\Omega_1}{M_1} \\ \vdots \\ \frac{\Omega_n}{M_n} \end{bmatrix} \end{aligned} \quad (7.288)$$

where

$$\Omega^2 = \begin{bmatrix} \omega_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_n^2 \end{bmatrix} \quad (7.289)$$

and

$$\Omega_\zeta = \begin{bmatrix} \zeta_1 \omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \zeta_n \omega_n \end{bmatrix} \quad (7.290)$$

where  $\zeta$  has been added as potential damping terms to the equations of motion here.

Then, recalling that the aerodynamic forces and moments and the generalized forces can both be modeled as linear functions of the rigid states, vibration states, and control inputs, one can define the following terms:

$$\mathcal{I} = \begin{bmatrix} 1 & 0 & -X_{\dot{w}} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 - Z_{\dot{w}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -\frac{I_{xz}}{I_{zz}} \\ 0 & 0 & -M_{\dot{w}} & 0 & 1 & 0 \\ 0 & 0 & 0 & -\frac{I_{xz}}{I_{zz}} & 0 & 1 \end{bmatrix} \quad (7.291)$$

(note if  $I_{xz} = 0$  and one ignores  $\dot{w}$  effects,  $\mathcal{I}$  will be a  $6 \times 6$  identity matrix),

$$\mathcal{M} = \begin{bmatrix} M_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & M_n \end{bmatrix} \quad (7.292)$$

$$f_{rig}(\vec{x}_{rig}, \theta, \phi) = \mathcal{I}^{-1} \begin{bmatrix} rv - qw - g \sin \theta + X_0 \\ pw - ru + g \cos \theta \sin \phi + Y_0 \\ qu - pv + g \cos \theta \cos \phi + Z_0 \\ \frac{I_{xx}-I_{zz}}{I_{xx}} qr + \frac{I_{xz}}{I_{xx}} pq + L_0 \\ \frac{I_{zz}-I_{xx}}{I_{yy}} pr + \frac{I_{xz}}{I_{yy}} (r^2 - p^2) + M_0 \\ \frac{I_{xx}-I_{yy}}{I_{zz}} pq - \frac{I_{xz}}{I_{zz}} qr + N_0 \end{bmatrix} \quad (7.293)$$

$$\mathcal{A}_{rig \leftarrow rig} = \mathcal{I}^{-1} \begin{bmatrix} X_u & 0 & X_w & 0 & X_q & 0 \\ 0 & Y_v & 0 & Y_p & 0 & Y_r \\ Z_u & 0 & Z_w & 0 & Z_q & 0 \\ 0 & L_v & 0 & L_p & 0 & L_r \\ M_u & 0 & M_w & 0 & M_q & 0 \\ 0 & N_v & 0 & N_p & 0 & N_r \end{bmatrix} \quad (7.294)$$

$$\mathcal{A}_{rig \leftarrow \eta} = \mathcal{I}^{-1} \begin{bmatrix} X_{\eta_1} & \cdots & X_{\eta_n} \\ Y_{\eta_1} & \cdots & Y_{\eta_n} \\ Z_{\eta_1} & \cdots & Z_{\eta_n} \\ L_{\eta_1} & \cdots & L_{\eta_n} \\ M_{\eta_1} & \cdots & M_{\eta_n} \\ N_{\eta_1} & \cdots & N_{\eta_n} \end{bmatrix} \quad (7.295)$$

$$\mathcal{A}_{rig \leftarrow \dot{\eta}} = \mathcal{I}^{-1} \begin{bmatrix} X_{\dot{\eta}_1} & \cdots & X_{\dot{\eta}_n} \\ Y_{\dot{\eta}_1} & \cdots & Y_{\dot{\eta}_n} \\ Z_{\dot{\eta}_1} & \cdots & Z_{\dot{\eta}_n} \\ L_{\dot{\eta}_1} & \cdots & L_{\dot{\eta}_n} \\ M_{\dot{\eta}_1} & \cdots & M_{\dot{\eta}_n} \\ N_{\dot{\eta}_1} & \cdots & N_{\dot{\eta}_n} \end{bmatrix} \quad (7.296)$$

$$\mathcal{B}_{rig} = \mathcal{I}^{-1} \begin{bmatrix} 0 & X_{\delta_e} & 0 & X_{\delta_t} \\ 0 & 0 & Y_{\delta_r} & 0 \\ 0 & Z_{\delta_e} & 0 & Z_{\delta_t} \\ L_{\delta_a} & 0 & L_{\delta_r} & 0 \\ 0 & M_{\delta_e} & 0 & M_{\delta_t} \\ N_{\delta_a} & 0 & N_{\delta_r} & 0 \end{bmatrix} \quad (7.297)$$

$$A_{vib \leftarrow rig} = \mathcal{M}^{-1} \begin{bmatrix} Q_{1_u} & Q_{1_v} & Q_{1_w} & Q_{1_p} & Q_{1_q} & Q_{1_r} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Q_{n_u} & Q_{n_v} & Q_{n_w} & Q_{n_p} & Q_{n_q} & Q_{n_r} \end{bmatrix} \quad (7.298)$$

$$A_{vib \leftarrow \eta} = \mathcal{M}^{-1} \begin{bmatrix} Q_{1_{\eta_1}} & \cdots & Q_{1_{\eta_n}} \\ \vdots & \ddots & \vdots \\ Q_{n_{\eta_1}} & \cdots & Q_{n_{\eta_n}} \end{bmatrix} - \Omega^2 \quad (7.299)$$

$$A_{vib \leftarrow \dot{\eta}} = \mathcal{M}^{-1} \begin{bmatrix} Q_{1_{\dot{\eta}_1}} & \cdots & Q_{1_{\dot{\eta}_n}} \\ \vdots & \ddots & \vdots \\ Q_{n_{\dot{\eta}_1}} & \cdots & Q_{n_{\dot{\eta}_n}} \end{bmatrix} - 2\Omega_\zeta \quad (7.300)$$

and

$$B_{vib} = \mathcal{M}^{-1} \begin{bmatrix} Q_{1_{\delta_a}} & Q_{1_{\delta_e}} & Q_{1_{\delta_r}} & Q_{1_{\delta_t}} \\ \vdots & \vdots & \vdots & \vdots \\ Q_{n_{\delta_a}} & Q_{n_{\delta_e}} & Q_{n_{\delta_r}} & Q_{n_{\delta_t}} \end{bmatrix} \quad (7.301)$$

With these definitions, one may finally rewrite the elastic flight vehicle equations of motion in state-space form as

$$\begin{aligned} \dot{\vec{x}}_{rig} &= f_{rig}(\vec{x}_{rig}, \phi, \theta) + \mathcal{A}_{rig \leftarrow rig} \vec{x}_{rig} + [\mathcal{A}_{rig \leftarrow \eta} \quad \mathcal{A}_{rig \leftarrow \dot{\eta}}] \vec{x}_{vib} + \mathcal{B}_{rig} \vec{u} \\ \dot{\vec{x}}_{vib} &= \begin{bmatrix} 0_{n \times 6} \\ A_{vib \leftarrow rig} \end{bmatrix} \vec{x}_{rig} + \begin{bmatrix} 0_{n \times n} & I_{n \times n} \\ A_{vib \leftarrow \eta} & A_{vib \leftarrow \dot{\eta}} \end{bmatrix} \vec{x}_{vib} + \begin{bmatrix} 0_{n \times 4} \\ B_{vib} \end{bmatrix} \vec{u} \end{aligned} \quad (7.302)$$

and it should be noted that here,  $v$ ,  $w$ , and  $\dot{w}$  have been used in place of  $\beta$ ,  $\alpha$ , and  $\dot{\alpha}$ , respectively, though these could be replaced by the linear approximations and coefficient conversions if required. Note also that one would also require the supplemental Euler angle equation to relate  $p$ ,  $q$ , and  $r$  to  $\dot{\phi}$  and  $\dot{\theta}$  to complete the state-space formulation, i.e.

$$\begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} 1 & \sin \phi \tan \theta & \cos \phi \tan \theta \\ 0 & \cos \phi & -\sin \phi \\ 0 & \sin \phi \sec \theta & \cos \phi \sec \theta \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix} \quad (7.303)$$

Furthermore, as shown previously for the rigid flight vehicle dynamics, one can linearize the elastic flight vehicle EOMs for easier analysis, simulation, and design. As  $f_{rig}(\vec{x}_{rig}, \theta, \phi)$  and the Euler angle equations are the only nonlinear term of the EOMs, but as these terms are rigid vehicle specific, one can simply use a rigid flight vehicle linearization method to form the LTI state-space system as

$$\begin{bmatrix} \Delta \vec{x}_{rig} \\ \Delta \vec{x}_{eul} \\ \Delta \vec{x}_{vib} \end{bmatrix} = \begin{bmatrix} A_{rig \leftarrow rig} & A_{rig \leftarrow eul} & A_{rig \leftarrow vib} \\ A_{eul \leftarrow rig} & A_{eul \leftarrow eul} & 0_{3 \times 2n} \\ A_{vib \leftarrow rig} & 0_{2n \times 3} & A_{vib \leftarrow vib} \end{bmatrix} \begin{bmatrix} \Delta \vec{x}_{rig} \\ \Delta \vec{x}_{eul} \\ \Delta \vec{x}_{vib} \end{bmatrix} + \begin{bmatrix} B_{rig} \\ 0_{3 \times 4} \\ B_{vib} \end{bmatrix} \Delta \vec{u} \quad (7.304)$$

where

$$\Delta \vec{x}_{eul} = \begin{bmatrix} \Delta \phi \\ \Delta \theta \\ \Delta \psi \end{bmatrix} \quad (7.305)$$

$$A_{vib \leftarrow vib} = \begin{bmatrix} 0_{n \times n} & I_{n \times n} \\ A_{vib \leftarrow \eta} & A_{vib \leftarrow \dot{\eta}} \end{bmatrix} \quad (7.306)$$

$A_{rig \leftarrow rig}$  has been altered from  $\mathcal{A}_{rig \leftarrow rig}$  due to the linearization of  $f_{rig}(\vec{x}_{rig}, \theta, \phi)$  with respect to  $\vec{x}_{rig}$ ,  $A_{rig \leftarrow eul}$  results from the linearization of  $f_{rig}(\vec{x}_{rig}, \theta, \phi)$  with respect to  $\vec{x}_{eul}$ , and  $A_{eul \leftarrow rig}$  and  $A_{eul \leftarrow eul}$  result from the linearization of the supplemental Euler angle equation. The explicit computation of this linearized state matrix for the rigid vehicle about a general trim/reference flight condition will be addressed in more detail later.

Note that one can also use the reference and perturbation form for the modal coordinates as

$$\eta_i(t) = \bar{\eta}_i + \Delta \eta_i(t) \quad (7.307)$$

and for the pressure distribution as

$$\vec{P}(\vec{x}) = \bar{P}(\vec{x}) + \Delta\vec{P}(\vec{x}) \quad (7.308)$$

Thus, for the perturbation or linearized vibration equation of motion, one has

$$\Delta\ddot{\eta}_i + \omega_i^2\Delta\eta_i = \frac{1}{M} \int_{Area} \Delta\vec{P}(\vec{x}_B) \cdot \vec{v}(\vec{x}_B) dS, \quad i = 1, \dots, n \quad (7.309)$$

## References

For more information, please refer to the following

- Schmidt, D. K., “Chapter 4: Equations of Motion for Elastic Vehicles,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 128-155

## 7.5 Atmospheric Effects on Aerospace Vehicle Dynamics

Many aerospace vehicles fly through the atmosphere during all or part of their flight and must be traversed to get spacecraft into space. Thus, aerodynamics due to the atmosphere are vital to understanding aerospace vehicle dynamics and are fundamentally governed by the conservation of kinetic, potential, and internal energy in a flow through **Bernoulli's equation** which can be written for steady, inviscid flow as

$$\frac{\|\vec{v}\|^2}{2} + \Phi(h) + \int \frac{dP}{\rho(P)} = \text{constant along a streamline} \quad (7.310)$$

where  $\|\vec{v}\|$  is flow speed at a point along the streamline,  $\Phi(h)$  is the geopotential due to the changing altitude,  $h$ , of the point along the streamline,  $P$  is the pressure at a point along the streamline,  $\rho(P)$  is fluid density along the streamline and can be considered a function of pressure, i.e., a **compressible flow**, or not, i.e., an **incompressible flow**. For static air, i.e.,  $\vec{v} = \vec{0}$ , one obtains the **aerostatic equation** as

$$\Phi(h) + \int \frac{dP}{\rho(P)} = \text{constant along a streamline} \quad (7.311)$$

The **geopotential** is the specific gravitational potential energy at a specific altitude  $h$  and is calculated as

$$\Phi(h) = \int_0^h \|\vec{g}(h)\| dh \quad (7.312)$$

where the altitude here must be integrated along the line-of-force of gravity and its value will depend on the gravity model used. This allows one to write the aerostatic equation in differential form as

$$dP = -\|\vec{g}(h)\| \rho dh \quad (7.313)$$

For an incompressible flow, one has

$$\frac{\|\vec{v}\|^2}{2} + \Phi(h) + \frac{P}{\rho} = \text{constant along a streamline} \quad (7.314)$$

For an adiabatic compressible flow, i.e.,

$$\frac{P}{\rho^\gamma} = \text{constant} \quad (7.315)$$

one has

$$\frac{\|\vec{v}\|^2}{2} + \Phi(h) + \frac{\gamma}{\gamma - 1} \left( \frac{P}{\rho} - \frac{P_0}{\rho_0} \right) = \text{constant along a streamline} \quad (7.316)$$

where  $P_0$  is the reference pressure,  $\rho_0$  is the reference density, and  $\gamma$  is the adiabatic index formed as the ratio of the specific heat at constant pressure to the specific heat at constant volume.  $\gamma$  for dry air is considered as 1.4 due to primary diatomic gases  $N_2$  and  $O_2$ .

For the aerodynamics produced within the atmosphere, one uses the subscript  $\infty$  to represent the “free-stream” atmospheric conditions, i.e., where the atmosphere exhibits an undisturbed uniform flow. In this case,  $\rho_\infty$  is the **air density**,  $\|\vec{v}_\infty\|$  is the **airspeed**, i.e., the magnitude of the velocity of the vehicle relative to the atmosphere, and  $P_\infty$  is the **atmospheric static pressure**, also known as the atmosphere’s **aerostatic pressure**. Thus, aerodynamic models for aerospace vehicle dynamics require modeling both the free-stream atmospheric conditions and the relative speed of the vehicle with respect to the atmosphere which may be moving due to the presence of wind according to the wind triangle, i.e.,

$$\vec{v}_g = \vec{v}_\infty + \vec{v}_{wind} \quad (7.317)$$

where  $\vec{v}_g$  is the **groundspeed vector** or the velocity of the vehicle with respect to the Earth’s surface and  $\vec{v}_{wind}$  i.e., the velocity of the atmosphere relative to the ground, i.e., the wind speed vector. However, lower-fidelity aerodynamics models often assume the no-wind condition, e.g., due to the high speed of the aerospace vehicle. Under the **no-wind condition**  $\vec{v}_w = \vec{0}$  and the airspeed vector is the same as the groundspeed vector.

Furthermore, in aerospace vehicle dynamics, a streamline does not vary significantly in altitude to affect the gravitational potential during interaction with the vehicle. This negligible-gravity assumption provides the following simplified Bernoulli’s equation for an atmospheric incompressible flow over an aerospace vehicle

$$\frac{\|\vec{v}_\infty\|^2}{2} + \frac{P_\infty}{\rho_\infty} = P_t \quad (7.318)$$

where  $P_t$  is the atmospheric **total pressure**, also known as the atmospheric **stagnation pressure**. Furthermore, defining  $Q_\infty$  as the atmospheric **incompressible dynamic pressure**, also known as the **incompressible impact pressure** given by

$$Q_\infty = \frac{1}{2} \rho_\infty \|\vec{v}_\infty\|^2 \quad (7.319)$$

one has

$$P_\infty + Q_\infty = P_t \quad (7.320)$$

This negligible-gravity assumption also provides the following simplified Bernoulli’s equation for an atmospheric adiabatic compressible flow over an aerospace vehicle

$$\frac{\|\vec{v}\|^2}{2} + \left( \frac{\gamma}{\gamma - 1} \right) \frac{P_\infty}{\rho_\infty} = \left( \frac{\gamma}{\gamma - 1} \right) \frac{P_t}{\rho_t} \quad (7.321)$$

which can be rewritten as

$$\frac{\gamma - 1}{\gamma} \frac{P_\infty}{\rho_\infty} \frac{\|\vec{v}\|^2}{2} + 1 = \frac{\rho_\infty}{\rho_t} \frac{P_t}{P_\infty} \quad (7.322)$$

where one also has the following adiabatic flow condition as

$$\frac{P_t}{\rho_t^\gamma} = \frac{P_\infty}{\rho_\infty^\gamma} \quad (7.323)$$

$$\left( \frac{\rho_\infty}{\rho_t} \right)^\gamma = \frac{P_t}{P_\infty} \quad (7.324)$$

or

$$\frac{\rho_\infty}{\rho_t} = \left( \frac{P_\infty}{P_t} \right)^{\frac{1}{\gamma}} \quad (7.325)$$

Thus, by substitution, one has

$$\frac{\gamma - 1}{\gamma} \frac{P_\infty}{\rho_\infty} \frac{\|\vec{v}\|^2}{2} + 1 = \left( \frac{P_\infty}{P_t} \right)^{\frac{1}{\gamma}} \frac{P_t}{P_\infty} \quad (7.326)$$

and substituting for the speed of sound, i.e.,

$$c_s = \sqrt{\frac{dP}{d\rho}} = \sqrt{\gamma \frac{P_\infty}{\rho_\infty}} \quad (7.327)$$

one has

$$\frac{\gamma - 1}{2} \frac{\|\vec{v}\|^2}{c_s^2} + 1 = \left( \frac{P_\infty}{P_t} \right)^{\frac{\gamma-1}{\gamma}} \quad (7.328)$$

or

$$P_t = P_\infty \left( 1 + \frac{\gamma - 1}{2} M^2 \right)^{\frac{\gamma}{\gamma-1}} \quad (7.329)$$

where  $M$  is the **Mach number**

$$M = \frac{\|\vec{v}\|^2}{c_s^2} \quad (7.330)$$

which provides a means to compute pressures for compressible flow along a streamline as a function of the non-dimensionalized Mach number. Notably, if  $M > 1$ , one must account for pressure jumps across shock waves.

For an aerospace vehicle's equations of motion, one must resolve this atmospheric pressure distribution about the vehicle into an aerodynamic force vector at the center of pressure where if this center of pressure is not at the center of mass for the vehicle, then an aerodynamic moment vector will also be produced. Generally, an aerodynamic force vector on the vehicle is non-dimensionalized, resolved into two fundamental forces, and considered a function of three non-dimensionalized "similarity parameters" as

$$\frac{\vec{F}_a}{Q_\infty S_{ref}} = \frac{\vec{D} + \vec{L}}{Q_\infty S_{ref}} = f(\mathcal{R}, \mathcal{F}, M) \quad (7.331)$$

where  $\vec{D}$  is the **drag force**,  $\vec{L}$  is the **lift force**,  $S_{ref}$  is some **reference area** for the aerospace vehicle,  $\mathcal{R}$  is the **Reynolds number** which governs the magnitude of the friction drag, e.g., laminar or turbulent flow conditions, and  $\mathcal{F}$  is the **Froude number** which governs the ground effect on the aerodynamic force. This normalization of the drag and lift forces naturally leads to the models

$$\vec{D} = Q_\infty S_{ref} C_D \vec{e}_{-\vec{v}} \quad (7.332)$$

and

$$\vec{L} = Q_\infty S_{ref} C_L \vec{e}_{\perp \vec{v}} \quad (7.333)$$

where  $C_D$  is the **drag coefficient**,  $C_L$  is the **lift coefficient**,  $\vec{e}_{-\vec{v}}$  is the unit vector in the opposite direction of the vehicle's velocity, and  $\vec{e}_{\perp \vec{v}}$  is a unit vector perpendicular to the velocity along the lift force direction. Notably, this lift force is sometimes split into two forces perpendicular to each other and the drag force depending on the dynamics model. This split between lift and drag is due to their different aerodynamic effects on the vehicle's motion as the drag force is the resistance generated by the atmosphere to resist the vehicle's motion through it while the lift force is generated by specific structures on the aerospace vehicle for the purposes of changing its direction through the atmosphere, e.g., overcoming gravity and flying. Thus,  $Q_\infty$  is a vital parameter in aerodynamics modeling and is fundamentally related to the velocity of the vehicle and the atmospheric static pressure and air density. Importantly, it is often computed for compressible flows by substituting for  $\rho_\infty$  in terms of  $\mathcal{M}$  as

$$Q_\infty = \frac{1}{2} \gamma P_\infty \mathcal{M}^2 \quad (7.334)$$

### Reference Atmosphere Models

For the atmospheric static pressure and air density, one must use a reference atmosphere model of these do not account for variations in barometric conditions, e.g., due to weather and water vapor, but serves as a reference for deriving a “static” set of atmospheric conditions that approximate the expected conditions at various altitudes above the Earth’s surface consistent with idealized fluid mechanics and historical atmospheric data.

There are three primary reference atmosphere models used for the free-stream conditions in aerospace vehicle dynamics which are presented here without the  $\infty$  subscript for a less cluttered notation. The first are the **standard atmosphere models** which use three governing equations. The first relates the free-stream pressure to the density and temperature via the **perfect-gas equation**, also known as the **ideal gas equation**,

$$P = \rho \frac{R^*}{M} T \quad (7.335)$$

where  $R^* = 8.31446261815324 \text{ J/(mol}\cdot^\circ\text{K)}$  is the **universal gas constant**,  $T$  is the atmospheric temperature, and  $M$  is the mean molecular weight of the atmosphere. Notably, assigning a reference point as

$$P_0 = \rho_0 \frac{R^*}{M_0} T_0 \quad (7.336)$$

one has the ratios

$$\frac{P}{P_0} = \frac{\rho M_0 T}{\rho_0 M T_0} \quad (7.337)$$

This allows one to define the **molecular-scale temperature** as

$$T_M = \frac{M_0}{M} T \quad (7.338)$$

and with  $T_0 = T_{M,0}$ , one has

$$\frac{P}{P_0} = \frac{\rho T_M}{\rho_0 T_{M,0}} \quad (7.339)$$

where  $M_0/M = 1$  and  $T_M = T$  up to geodetic altitudes of 90 km.

The second equation is a linear temperature variation with altitude for different layers of the atmosphere, i.e.,

$$\begin{aligned} T &= T_0 + L_T(h - h_0)h < 90 \text{ km} \\ T_M &= T_{M,0} + L'_T(h - h_0)h > 90 \text{ km} \end{aligned} \quad (7.340)$$

where  $L_T$  or  $L'_T$  is the **temperature lapse rate**,  $h$  is the geodetic altitude, and  $\mathbf{h}$  is the **geopotential altitude** determined as the ratio of the geopotential,  $\Phi(h)$ , to the standard acceleration due to gravity at mean sea level (MSL), i.e., the specific gravitational potential energy at

$$\mathbf{h} = \frac{\Phi(h)}{g_0} \|\vec{g}\| \quad (7.341)$$

where  $g_0$  is the acceleration due to gravity at mean sea level (MSL) and  $h$  is the geodetic altitude. Here a layer is called **isothermal** if  $L_T = 0$  or  $L'_T = 0$ .

The standard atmosphere's tabulated data assumes the geodetic altitude can be approximated by the geocentric altitude, which provides the following conversion from geopotential altitude to geocentric altitude via the differential equation

$$g_0 d\mathbf{h} = \|\vec{g}(h)\| dh \quad (7.342)$$

which provides the equation

$$\mathbf{h} - \mathbf{h}_0 = \frac{1}{g_0} \int_{h_0}^h \|\vec{g}(h)\| dh \quad (7.343)$$

where  $g(h)$  for the standard atmosphere models is taken as a first-order approximation of the ellipsoidal-Earth gravity vector starting at MSL and  $\ell = 45^\circ$  as a function of  $h$ . Notably, if one assumes the spherical-Earth gravity model for  $\vec{g}(h)$ , one obtains the approximation

$$\mathbf{h} \approx \frac{\bar{R}_E}{\bar{R}_E + h} h \quad (7.344)$$

where  $\bar{R}_E$  is the Earth's mean radius where  $h = \mathbf{h} = 0$ .

The third is the differential form of the aerostatic equation, i.e.,

$$dP = -\|\vec{g}(h)\| \rho dh = -g_0 \rho dh \quad (7.345)$$

Dividing the aerostatic equation by the perfect-gas equation in terms of  $h$ , one has

$$\frac{dP}{P} = -\frac{gM}{R^*T} dh == -\frac{gM_0}{R^*T_M} dh \quad (7.346)$$

where  $g = \|\vec{g}(h)\|$  and in terms of  $\mathbf{h}$ , one has

$$\frac{dP}{P} = -\frac{g_0 M}{R^* T} dh \quad (7.347)$$

For  $h < 90$  km, one has  $M = M_0$  and a linear variation of  $T = T_M$  with  $\mathbf{h}$ . For  $L_T = 0$ , Equation 7.347 can be integrated from  $P_0, \mathbf{h}_0$  to  $P, \mathbf{h}$  as

$$\frac{P}{P_0} = \exp\left(-\frac{g_0 M_0}{R^* T_0} (\mathbf{h} - \mathbf{h}_0)\right) \quad (7.348)$$

and related to density as

$$\frac{\rho}{\rho_0} = \exp\left(-\frac{g_0 M_0}{R^* T_0} (\mathbf{h} - \mathbf{h}_0)\right) \quad (7.349)$$

and for  $L_T \neq 0$ , one can substitute for  $d\mathbf{h}$  to obtain

$$\frac{dP}{P} = -\frac{g_0 M_0}{R^* (T_0 + L_T(\mathbf{h} - \mathbf{h}_0))} \frac{dT}{L_T} \quad (7.350)$$

which can be integrated from  $P_0, T_0$  to  $P, T$  as

$$\frac{P}{P_0} = \left(\frac{T}{T_0}\right)^{-\frac{-g_0 M}{L_T R^*}} \quad (7.351)$$

or

$$\frac{P}{P_0} = \left(\frac{T_0}{T_0 + L_T(\mathbf{h} - \mathbf{h}_0)}\right)^{\frac{-g_0 M}{L_T R^*}} \quad (7.352)$$

and related to density as

$$\frac{\rho}{\rho_0} = \left(\frac{T_0}{T_0 + L_T(\mathbf{h} - \mathbf{h}_0)}\right)^{\frac{-g_0 M}{L_T R^*} + 1} \quad (7.353)$$

For  $h > 90$  km one has a linear variation of  $T_M$  with  $h$ . For  $L'_T = 0$ , Equation 7.346 can be integrated from  $P_0, h_0$  to  $P, h$  as

$$\frac{P}{P_0} = \exp\left(-\frac{M_0}{R^* T_{M,0}} \int_{h_0}^h g(h) dh\right) \quad (7.354)$$

and related to density as

$$\frac{\rho}{\rho_0} = \exp\left(-\frac{M_0}{R^* T_{M,0}} \int_{h_0}^h g(h) dh\right) \quad (7.355)$$

and for  $L'_T \neq 0$ , one can substitute for  $dh$  to obtain

$$\frac{dP}{P} = -\frac{g_0 M_0}{R^* (T_0 + L'_T(h - h_0))} \frac{dT}{L'_T} \quad (7.356)$$

which can be integrated from  $P_0, T_{M,0}$  to  $P, T_M$  as

$$\frac{P}{P_0} = \exp\left(-\frac{M_0}{R^* L'_T} \int_{h_0}^h \frac{g(h)}{\frac{T_0}{L'_T} + (h - h_0)} dh\right) \quad (7.357)$$

and related to density as

$$\frac{\rho}{\rho_0} = \frac{T_{M,0}}{T_M} \exp\left(-\frac{M_0}{R^* L'_T} \int_{h_0}^h \frac{g(h)}{\frac{T_0}{L'_T} + (h - h_0)} dh\right) \quad (7.358)$$

For the **United States Standard Atmosphere 1962 (USSA62)** used by the Federal Aviation Administration (FAA), one has

Atmospheric Layer	Lower Geopotential Altitude, $h_0$ (km)	Lapse Rate $L_T$ ( $^{\circ}\text{K}/\text{km}$ )	Temperature $T_0$ ( $^{\circ}\text{K}$ )	Mean Molecular Weight of Air $M$ (g/mol)
Troposphere	0	-6.5	288.15	28.9644
Tropopause	11	0	216.63	28.9644
Stratosphere 1	20	1	216.65	28.9644
Stratosphere 2	32	2.8	228.65	28.9644
Stratopause	47	0	270.65	28.9644
Mesosphere 1	52	-2	270.65	28.9644
Mesosphere 2	61	-4	252.65	28.9644
Mesopause	79	0	180.65	28.9644

and the mesopause is assumed to end at  $h_0 = 88.743$  km or  $h = 90$  km. Notably, for these layers the mean molecular weight of air is constant, but for the next layers, this value changes.

Atmospheric Layer	Lower Geodetic Altitude, $h_0$ (km)	Lapse Rate $L'_T$ ( $^{\circ}\text{K}/\text{km}$ )	Molecular-Scale Temperature $T_{M,0}$ ( $^{\circ}\text{K}$ )	Mean Molecular Weight of Air $M$
Thermosphere 1	90	3	180.65	28.9644
Thermosphere 2	100	5	288.15	28.88
Thermosphere 3	110	10	288.15	28.56
Thermosphere 4	120	20	288.15	28.07
Thermosphere 5	150	15	288.15	26.92
Thermosphere 6	160	10	288.15	26.66
Thermosphere 7	170	7	288.15	26.40
Thermosphere 8	190	5	288.15	25.85
Thermosphere 9	230	4	288.15	24.70
Thermosphere 10	300	3.3	288.15	22.66
Thermosphere 11	400	2.6	288.15	19.94
Thermosphere 12	500	1.7	288.15	17.94
Thermosphere 13	600	1.1	288.15	16.84
Thermopause	700	-	288.15	16.17

where the exosphere continues the Earth's atmosphere from the thermopause, also known as the exobase, to approximately 10,000 km. In the exosphere, barometric conditions no longer apply and is sometimes not considered a strict part of the atmosphere, but as an atmosphere-like volume with essentially no molecular collisions due to its extremely low density. The USSA1962 notably agrees with the **International Standard Atmosphere** and **International Civil Aviation Organization (ICAO) Standard Atmosphere** for  $h = [0, 32]$  km. As an aside, the speed of sound as a function of altitude is typically computed from the molecular-scale temperature of the standard atmosphere as

$$c_s = \sqrt{\gamma_{air} \frac{R^*}{M_0} T_M} \quad (7.359)$$

The second reference atmospheric model is to approximate multiple layers of the atmospheric as isothermal and with a constant mean molecular weight of air,  $M$ . This results in multiple exponential functions for static pressure and air density with the following forms

$$\frac{P}{P_0} = \exp\left(-\frac{g_0 M}{R^* T}(h - h_0)\right) \quad (7.360)$$

$$\frac{\rho}{\rho_0} = \exp\left(-\frac{g_0 M}{R^* T}(h - h_0)\right) \quad (7.361)$$

where  $h_0$  is the lower geopotential altitude of each layer and  $T$  is chosen as the mean of the temperature variation within the layer to be approximated and will exhibit discrete jumps between layers.

The third reference atmospheric model is the **exponential atmosphere model** which approximates the entire atmospheric as isothermal and with a constant mean molecular weight of air,  $M$ . This results in a single exponential function for static pressure and air density as

$$\frac{P}{P_0} = \exp\left(-\frac{g_0 M}{R^* T} h\right) \quad (7.362)$$

$$\frac{\rho}{\rho_0} = \exp\left(-\frac{g_0 M}{R^* T} h\right) \quad (7.363)$$

where  $h = 0$  at MSL. Here  $T$  can be chosen as the mean of the temperature variation given by full standard atmosphere model up to some desired height. This model is typically a good approximation up to an altitude of 140 km which is the extent that reentry trajectories are significantly affected by the atmospheric aerodynamic force.

## Wind Effects and Models

In deriving the equations of motion in introductory FDC, one typically uses the stability frame for coordinated flight where the velocity vector is colinear with the body-fixed frame  $x_B$ -axis, i.e.  $\bar{u} = v_a$ , and one can use approximations for  $\Delta v$  and  $\Delta w$  by  $\beta$  and  $\alpha$ . However, in the presence of wind, this relationship is more complex due to the wind triangle which can be expressed with the relevant frames as

$$\vec{v}_{B/N} = \vec{v}_{B/W} + \vec{v}_{W/N} \quad (7.364)$$

where  $\vec{v}_{B/N} = \vec{v}_g$  is the groundspeed vector, i.e. the velocity of the body-fixed frame relative to the navigation frame,  $\vec{v}_{B/W} = \vec{v}_\infty$  is the airspeed vector, i.e., the velocity of the body-fixed frame relative to the atmosphere, and  $\vec{v}_{W/N} = \vec{v}_{wind}$  is the wind speed vector, i.e. the velocity of the air mass relative to the “inertial” navigation frame. In FDC, as the aerodynamic forces and moments are typically a function of the airspeed velocity instead of the groundspeed vector, it is often more useful to use the airspeed velocity vector as part of the state vector in the equations of motion.

First, utilizing navigation frame and body-fixed frame coordinates, one can write

$$\vec{v}_{B/N,N} = C_{N \leftarrow B} \vec{v}_{\infty,B} + \vec{v}_{W/N,N} \quad (7.365)$$

and differentiating, one has

$$\dot{\vec{v}}_{B/N,N} = C_{N \leftarrow B} \dot{\vec{v}}_{\infty,B} + C_{N \leftarrow B} [\vec{\omega}_{B/N,B}]_\times \vec{v}_{\infty,B} + \dot{\vec{v}}_{W/N,N} \quad (7.366)$$

Next, recall the sum of forces in “inertial” navigation frame coordinates can be written as

$$\vec{F}_{a,N} + \vec{F}_{p,N} + \vec{F}_{g,N} = m \dot{\vec{v}}_{B/N,N} \quad (7.367)$$

Then, substituting and converting to body-fixed frame coordinates, one has

$$\vec{F}_{a,B} + \vec{F}_{p,B} + m C_{B \leftarrow N} \vec{g}_N = m \left( \dot{\vec{v}}_{\infty,B} + [\vec{\omega}_{B/N,B}]_{\times} \vec{v}_{\infty,B} + C_{B \leftarrow N} \dot{\vec{v}}_{W/N,N} \right) \quad (7.368)$$

or via mass normalization and using  $\vec{v}_{\infty,B} = [u \ v \ w]^T$

$$\begin{bmatrix} X - g \sin \theta \\ Y + g \cos \theta \sin \phi \\ Z + g \cos \theta \cos \phi \end{bmatrix} = \begin{bmatrix} \dot{u} + qw - rv \\ \dot{v} + ru - pw \\ \dot{w} + pv - qu \end{bmatrix} + \begin{bmatrix} \dot{u}_{W/N,N} \\ \dot{v}_{W/N,N} \\ \dot{w}_{W/N,N} \end{bmatrix} \quad (7.369)$$

Note that if one has **steady wind**, i.e.  $\dot{v}_{W/N,N} = 0$ , then these equations have the same mathematical form as the no-wind translation EOMs. Furthermore, one can integrate these velocity equations of motion to find the vehicle’s inertial position as

$$\vec{x}_{B/N,N} = \int \vec{v}_{B/N,N} dt = \int C_{N \leftarrow B}(t) \vec{v}_{\infty,B}(t) + \vec{v}_{W/N,N} dt \quad (7.370)$$

Thus, one can simply use the airspeed vector components in the EOMs which would add a positional offset that grows linearly with time due to the steady wind.

However, wind is not a steady phenomenon. A **wind shear** is a variation of the wind vector with respect to position, typically decoupled into vertical and horizontal coordinates. A vertical wind shear describes the wind vector varying with altitude and a horizontal wind shear describes the wind vector varying along some horizontal distance. For aerospace vehicle dynamics and control, a vertical wind shear becomes important during certain flight phases, e.g., take-off, climb, descent, approach, and landing. While specific wind shear profiles caused by topography, frontal systems, and thunderstorms are typically obtained via numerical studies, one basic logarithmic model for a vertical wind shear profile based on a point measurement can be approximated for 3 ft.  $< h <$  1000 ft. as

$$\bar{u}_{ws}(h) = \|\vec{v}_{w,20}\|_2 \frac{\log\left(\frac{h}{z_0}\right)}{\log\left(\frac{20}{z_0}\right)} \quad (7.371)$$

where  $\bar{u}_{ws}$  is the mean horizontal wind speed in the navigation frame,  $\vec{v}_{w,20}$  is the measured horizontal wind speed at a height above ground level of 20 ft., and  $z_0 = 0.15$  ft. is used for take-off, climb, approach, and landing flight phases and  $z_0 = 2$  ft. for cruise.

For assessing the effects of vertical wind shear on the modal analysis, one typically utilizes the relationship between the forward flight velocity gradient and the altitude as some constant  $du/dh$ . Then, one can augment

the longitudinal state dynamics as

$$\begin{bmatrix} \Delta \dot{u} \\ \Delta \dot{\alpha} \\ \Delta \dot{q} \\ \Delta \dot{\theta} \\ \Delta \dot{h} \end{bmatrix} = \begin{bmatrix} X_u & X_\alpha & 0 & -g \cos \bar{\theta} & -X_u \left( \frac{du}{dh} \right) \\ \frac{Z_u}{\bar{u} - Z_\alpha} & \frac{Z_\alpha}{\bar{u} - Z_\alpha} & \frac{\bar{u} + Z_q}{\bar{u} - Z_\alpha} & -\frac{g}{\bar{u} - Z_\alpha} \sin \bar{\theta} & -\left( \frac{Z_u}{\bar{u} - Z_\alpha} \right) \left( \frac{du}{dh} \right) \\ M_u + M_\alpha \frac{Z_u}{\bar{u} - Z_\alpha} & M_\alpha + M_\dot{\alpha} \frac{Z_\alpha}{\bar{u} - Z_\alpha} & M_q + M_\dot{\alpha} \frac{\bar{u} + Z_q}{\bar{u} - Z_\alpha} & -M_\dot{\alpha} \frac{g}{\bar{u} - Z_\alpha} \sin \bar{\theta} & -\left( M_u + M_\alpha \frac{Z_u}{\bar{u} - Z_\alpha} \right) \left( \frac{du}{dh} \right) \\ 0 & 0 & 1 & 0 & 0 \\ \sin \bar{\theta} & \bar{u} \cos \bar{\theta} & 0 & -\bar{u} \cos \bar{\theta} & 0 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \\ \Delta \theta \\ \Delta h \end{bmatrix} + \begin{bmatrix} 0 & X_T \\ \frac{Z_{\delta_e}}{\bar{u} - Z_\alpha} & \frac{Z_T}{\bar{u} - Z_\alpha} \\ M_{\delta_e} + M_\alpha \frac{Z_{\delta_e}}{\bar{u} - Z_\alpha} & M_T + M_\alpha \frac{Z_T}{\bar{u} - Z_\alpha} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \delta_e \\ \Delta T \end{bmatrix} \quad (7.372)$$

For assessing the effects of **gusts**, i.e., the **unsteady wind component**, on the velocity EOMs as opposed to the constant steady wind component, one typically assumes that the unsteady wind component is a zero-mean **random gust** which take on some random properties of varying magnitude denoted as  $\vec{v}_g$ . For simplicity, assume one has no steady-wind, i.e. one can redefine the body-fixed frame velocity as

$$\vec{v}_{B/N} = \vec{v}_\infty + \vec{v}_g \quad (7.373)$$

which uses both  $\vec{v}_\infty = [u \ v \ w]^T$  and  $\vec{v}_g = [u_g \ v_g \ w_g]^T$  in the state vector for the EOMs. Thus, by component one has

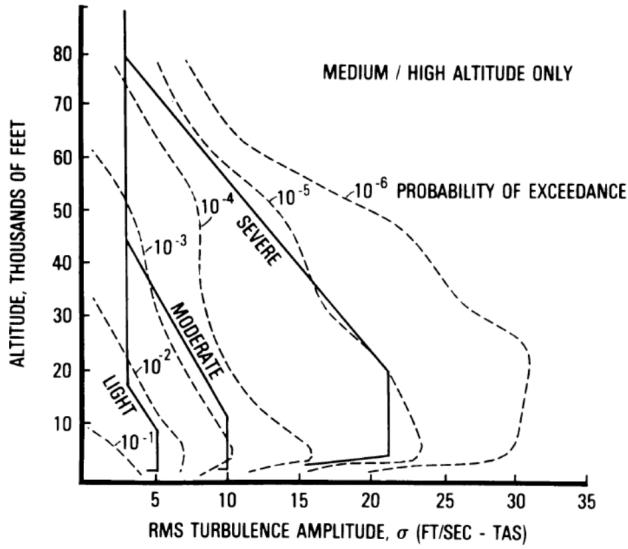
$$\begin{bmatrix} u_{tot} \\ v_{tot} \\ w_{tot} \end{bmatrix} = \begin{bmatrix} u + u_g \\ v + v_g \\ w + w_g \end{bmatrix} \quad (7.374)$$

where  $u$ ,  $v$ , and  $w$  implicitly model the velocity relative to a hypothetically *still* air mass. Two of the most widely used random gust models are the **Dryden gust model** and the **von Kármán gust model**. The Dryden gust state equations are provided here for addition to the stability frame coordinates as

$$\begin{bmatrix} \dot{u}_g(t) \\ \dot{v}_g(t) \\ \dot{v}_{g_1}(t) \\ \dot{w}_g(t) \\ \dot{w}_{g_1}(t) \end{bmatrix} = \begin{bmatrix} -\frac{\bar{u}}{L_u} & 0 & 0 & 0 & 0 \\ 0 & -\frac{\bar{u}}{L_u} & \sigma_v (1 - \sqrt{3}) \left( \frac{\bar{u}}{L_v} \right)^{3/2} & 0 & 0 \\ 0 & 0 & -\frac{\bar{u}}{L_v} & 0 & 0 \\ 0 & 0 & 0 & -\frac{\bar{u}}{L_w} & \sigma_w (1 - \sqrt{3}) \left( \frac{\bar{u}}{L_w} \right)^{3/2} \\ 0 & 0 & 0 & 0 & -\frac{\bar{u}_a}{L_w} \end{bmatrix} \begin{bmatrix} u_g(t) \\ v_g(t) \\ v_{g_1}(t) \\ w_g(t) \\ w_{g_1}(t) \end{bmatrix} + \begin{bmatrix} \sigma_u \left( \frac{2\bar{u}}{\pi L_u} \right)^{1/2} \\ \sigma_v \left( \frac{3\bar{u}}{L_v} \right)^{1/2} \\ 1 \\ \sigma_w \left( \frac{3\bar{u}}{L_w} \right)^{1/2} \\ 1 \end{bmatrix} \vec{n}(t) \quad (7.375)$$

$$\dot{\vec{x}}_g = A_g \vec{x}_g + B_g n$$

where the driving function,  $n(t)$ , is a zero-mean, additive white Gaussian noise (AWGN) of unit intensity across all five channels,  $[L_u \ L_v \ L_w]$  are  $[h \ 145h^{1/3} \ 145h^{1/3}]$  for  $h < 1750$  ft and  $[1750 \ 1750 \ 1750]$  for  $h \geq 1750$  ft, and  $\sigma_u$ ,  $\sigma_v$ , and  $\sigma_w$  are the standard deviations of the gusts, or **RMS gust intensities** which can be obtained from data such as “MIL-F-8785C Military Specification: Flying Qualities of Piloted Airplanes” which provides the plot of three levels of RMS gust intensities for different altitudes: *light*, *moderate*, and *severe*.



Note that here the **probability of exceedance** is the probability that the RMS gust intensity would exceed the value shown on the curves at that altitude. Also note that for numerical simulations,  $n(t)$  can be approximated as continuous over a short time step of size  $\Delta t$ , thus approximating  $n(t)$  by a random sequence  $n_i$  which are zero-mean Gaussian with variance  $1/\Delta t$ .

Lastly, using the wind frame Euler angles, i.e. the angle of attack and sideslip angles, to transform the instantaneous velocity magnitude,  $v_\infty$ , to body-fixed frame coordinates, one has

$$\begin{bmatrix} v_\infty \cos \alpha_{tot} \cos \beta_{tot} \\ v_\infty \sin \beta_{tot} \\ v_\infty \sin \alpha_{tot} \cos \beta_{tot} \end{bmatrix} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} + \begin{bmatrix} u_g \\ v_g \\ w_g \end{bmatrix} \quad (7.376)$$

From this equation, one can form the following relationships. First, the magnitude equation can be written as

$$v_\infty^2 = (u + u_g)^2 + (v + v_g)^2 + (w + w_g)^2 \quad (7.377)$$

Second, taking the third row divided by the first row, one has

$$\tan \alpha_{tot} = \frac{w + w_g}{u + u_g} \quad (7.378)$$

Third, the second row can be rewritten as

$$\sin \beta_{tot} = \frac{v + v_g}{v_\infty} \quad (7.379)$$

For small angles and  $u_w \ll u$ , one can write

$$\alpha_{tot} \approx \frac{w}{u} + \frac{w_g}{u} = \alpha + \alpha_g \quad (7.380)$$

$$\beta_{tot} \approx \frac{v}{v_\infty} + \frac{v_g}{v_a} = \beta + \beta_g \quad (7.381)$$

and similarly

$$\dot{\alpha}_{tot} \approx \dot{\alpha} + \dot{\alpha}_g \quad (7.382)$$

which are often used in linear flight dynamics instead of  $v_{tot}$  and  $w_{tot}$ .

With these relationships and a simplifying assumption, it is easy to model the aerodynamic forces and moments using the total velocity components. The assumption is that the axial and lateral gusts velocities have a negligible variation across the flight vehicle compared to the global variations of the gusts relative to the surface of the Earth. Note that this does not assume anything about the vertical gusts due to the presence of  $\dot{\alpha}$ . Then, one may simply use the addition formulas for  $u + u_g$ ,  $v + v_g$  (or  $\approx \beta + \beta_g$ ),  $w + w_g$  (or  $\approx \alpha + \alpha_g$ ), and  $\dot{w} + \dot{w}_g$  (or  $\approx \dot{\alpha} + \dot{\alpha}_g$  for computing the stability derivative contributions with respect to  $u$ ,  $v$  (or  $\beta$ ),  $w$  (or  $\alpha$ ), and  $\dot{w}$  (or  $\dot{\alpha}$ ) for  $X$ ,  $Y$ ,  $Z$ ,  $L$ ,  $M$ , and  $N$ .

Furthermore, for the linearized flight dynamics, if one can uses the linear Dryden gust model above combined with the nominal LTI state-space state equation, i.e.

$$\Delta \dot{\vec{x}} = A \Delta \vec{x} + B \Delta \vec{u} \quad (7.383)$$

where  $\Delta \vec{x} = [\Delta u \ \Delta \beta \ \Delta \alpha \ \Delta p \ \Delta q \ \Delta r \ \Delta \phi \ \Delta \theta \ \Delta \psi]^T$  and  $\Delta \vec{u} = [\Delta \delta_a \ \Delta \delta_e \ \Delta \delta_r \ \Delta \delta_t]^T$ , then one can form the following augmented LTI state equation

$$\begin{bmatrix} \Delta \dot{\vec{x}} \\ \dot{\vec{x}}_g \end{bmatrix} = \begin{bmatrix} A & G_g C_g \\ 0 & A_g \end{bmatrix} \begin{bmatrix} \Delta \vec{x} \\ \vec{x}_g \end{bmatrix} + \begin{bmatrix} B & G_g D_g \\ 0 & B_g \end{bmatrix} \begin{bmatrix} \Delta \vec{u} \\ n \end{bmatrix} \quad (7.384)$$

where the gust output matrix and feedthrough matrix can be constructed as

$$C_g = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\bar{u}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\bar{u}} & 0 \\ 0 & 0 & 0 & -\frac{1}{L_w} & \sigma_w (1 - \sqrt{3}) \left( \frac{\bar{u}}{L_w^3} \right)^{1/2} \end{bmatrix} \quad (7.385)$$

and

$$D_g = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \sigma_w \left( \frac{3}{L_w \bar{u}} \right)^{1/2} \end{bmatrix} \quad (7.386)$$

to output  $u_g$ ,  $\beta_g$ ,  $\alpha_g$ , and  $\dot{\alpha}_g$  from the gust state vector  $\vec{x}_g = [u_g \ v_g \ v_{g1} \ w_g \ w_{g1}]^T$ , and the mapping matrix from these gust outputs to the stability derivatives

$$G_g = \begin{bmatrix} X_u + \frac{X_\alpha Z_u}{\bar{u} - Z_\alpha} & 0 & X_\alpha + \frac{X_{\dot{\alpha}} Z_\alpha}{\bar{u} - Z_{\dot{\alpha}}} & X_{\dot{\alpha}} + \frac{X_{\dot{\alpha}} Z_{\dot{\alpha}}}{\bar{u} - Z_{\dot{\alpha}}} \\ 0 & \frac{Y_\beta}{\bar{u}} & 0 & 0 \\ \frac{Z_u}{\bar{u} - Z_\alpha} & 0 & \frac{Z_\alpha}{\bar{u} - Z_{\dot{\alpha}}} & \frac{Z_{\dot{\alpha}}}{\bar{u} - Z_{\dot{\alpha}}} \\ 0 & L_\beta^* & 0 & 0 \\ M_u + \frac{M_\alpha Z_u}{\bar{u} - Z_\alpha} & 0 & M_\alpha + \frac{M_{\dot{\alpha}} Z_\alpha}{\bar{u} - Z_{\dot{\alpha}}} & M_{\dot{\alpha}} + \frac{M_{\dot{\alpha}} Z_{\dot{\alpha}}}{\bar{u} - Z_{\dot{\alpha}}} \\ 0 & N_\beta^* & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (7.387)$$

Alternatively, this can also be used with the nonlinear state equation as an additive linear portion.

## References

For more information, please refer to the following

- Anderson Jr., J. D., “Chapter 3: The Standard Atmosphere,” *Introduction to Flight*, 8th ed., Vol. 1, McGraw-Hill, New York, 2016, pp. 110-133
- Anonymous, *U.S. Standard Atmosphere, 1962*, U.S. Government Printing Office, Washington D.C., 1962
- Curtis, H. D., “10.4 Atmospheric Drag,” *Orbital Mechanics for Engineering Students*, 4th ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2021, pp. 483-487
- Nelson, R. C., “6.6 Wind Shear,” *Flight Stability and Automatic Control*, 2nd ed., Vol. 1, McGraw-Hill, New York, 1998, pp. 229-232
- Schmidt, D. K., “Appendix A Properties of the Atmosphere,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 809-813
- Schmidt, D. K., “Appendix C Models of Atmospheric Turbulence,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 839-851

---

# General Aerospace Vehicle Guidance and Control Systems

## 8.1 Aerospace Vehicle Actuation Systems

The actuation system is the primary limiting factor for the inner-loop control systems of flight vehicles. The control surfaces for modern aircraft are fly-by-wire (FBW), i.e., the vehicles convert electrical signals into mechanical movements, either linear or rotary in order to control the vehicle. These fly-by-wire actuators can be electro-mechanical, electro-hydraulic, or electro-hydrostatic control actuators. For the propulsion systems, one typically uses

### Electro-Mechanical Actuator Models

The electro-mechanical dynamical system for a brushless direct current (BLDC) electric motor can be modeled using Kirchoff's second law of circuits and Euler's second law of motion as following two equations:

$$\begin{aligned} L_m \dot{i}_m &= v_m - R_m i_m - v_{emf} \\ I_m \dot{\omega}_m &= \sum M = M_{emf,m} - M_{D,m} - M_{f,m} \end{aligned} \quad (8.1)$$

where  $i_m$  is the motor current,  $L_m$  is the inductance of the motor coils,  $v_m$  is the motor voltage,  $R_m$  is the resistance of the motor coils or **armature resistance**,  $I_m$  is the *axial* polar moment of inertia of the motor,  $M_m$  is the sum of moments on the motor, including the electromotion,  $M_{emf,m}$ , drag loading,  $M_{D,m}$ , motor friction loading,  $M_{f,m}$ , and  $v_{emf}$  is the electromotive force which is proportional to the motor angular velocity

$$v_{emf} = k_{emf} \omega_m \quad (8.2)$$

where  $k_{emf}$  is the **electromotive force (EMF) constant**. Lorentz's law which describes that the EMF on a coil in a magnetic field results in a moment produced in proportion to the motor current, i.e.,

$$M_{emf,m} = k_m i_m \quad (8.3)$$

where  $k_m$  is the **motor moment constant**, also known as the **armature constant**. Assuming the motor friction is negligible, i.e.  $M_{f,m} \approx 0$ , one has a second-order nonlinear dynamics equation

$$\begin{aligned}\dot{i}_m &= -\frac{R_m}{L_m}i_m + \frac{1}{L_m}v_m - \frac{k_{emf}}{L_m}\omega_m \\ \dot{\omega}_m &= \frac{k_m}{I_m}i_m - \frac{1}{I_m}M_{D,m}\end{aligned}\quad (8.4)$$

Often, one can assume a low inductance BLDC electric motor, i.e.,  $L_m \ll 1$  and  $\dot{i}_m \approx 0$ , which leads to a first-order approximation of the dynamics by including only the mechanical mode as

$$\begin{aligned}0 &= v_m - R_m i_m - k_{emf} \omega_m \\ \dot{\omega}_m &= \frac{k_m}{I_m} i_m - \frac{1}{I_m} M_{D,m}\end{aligned}\quad (8.5)$$

and by substitution for  $i_m$ , one obtains

$$\dot{\omega}_m = \frac{k_m}{R_m I_m} v_m - \frac{k_m k_{emf}}{R_m I_m} \omega_m - \frac{1}{I_m} M_{D,m}\quad (8.6)$$

### Electric Geared Propeller

The gearbox relationship from the motor to the propeller can be modeled by two equations:

$$\begin{aligned}\omega_m &= r_g \omega_p \\ M_m &= \frac{1}{\eta_g r_g} M_p\end{aligned}\quad (8.7)$$

where  $r_g$  is the gearbox reduction ratio,  $\eta_g$  is the gearbox efficiency (typically 80 – 90%),  $\omega_m$  is the angular velocity of the motor,  $\omega_p$  is the angular velocity of the propeller,  $M_m$  is the load moment that the motor experiences, and  $M_p$  is the load moment that the propeller experiences. Note that ideally the power must be conserved as

$$M_m \omega_m = M_p \omega_p\quad (8.8)$$

but the efficiency comes into play due to the heat loss. Lastly, it should be noted that the total polar moment of inertia experienced by the motor,  $I_t$ , combines the motor shaft moment of inertia,  $I_m$ , with the propeller moment of inertia,  $I_p$ , through the gearbox reduction ratio as

$$I_t = I_m + \frac{I_p}{r_g^2}\quad (8.9)$$

The propeller dynamics are primarily governed by the propulsive thrust,  $T_p$ , and drag-induced aerodynamic moment produced due to the angular velocity of the propeller, i.e.

$$\begin{aligned}T_p &= C_T \omega_p^2 \\ M_{D,p} &= C_{dm} \omega_p^2\end{aligned}\quad (8.10)$$

where  $C_T$  is the thrust coefficient and  $C_{dm}$  is the drag-moment coefficient. Note that one can control the thrust directly through  $\omega_m$  as

$$T_p = \frac{C_T}{r_g^2} \omega_m^2 \quad (8.11)$$

By substitution of the second propeller equation into the first gearbox equation, one has

$$M_{D,p} = \frac{C_{dm}}{r_g^2} \omega_m^2 \quad (8.12)$$

Then, by substitution into the second gearbox equation, one has

$$M_{D,m} = \frac{C_{dm}}{\eta_g r_g^3} \omega_m^2 \quad (8.13)$$

Finally, substituting  $M_{D,m}$  into the motor dynamics equation, the combined second-order **propeller-gearbox-motor plant** can be written as the nonlinear dynamical system

$$\begin{bmatrix} \dot{i}_m \\ \dot{\omega}_m \end{bmatrix} = \begin{bmatrix} -\frac{R_m}{L_m} i_m - \frac{k_{emf}}{L_m} \omega_m + \frac{1}{L_m} u \\ \frac{k_m}{I_t} i_m - \frac{C_{dm}}{\eta_g r_g^3 I_t} \omega_m^2 \end{bmatrix} \quad (8.14)$$

$$y = \frac{C_T}{r_g^2} \omega_m^2$$

where  $u$  is the motor voltage input,  $v_m$ , and  $y$  is propulsive thrust output,  $T_p$ .

Assuming a low inductance BLDC electric motor leads to a first-order approximation of the dynamics by including only the mechanical mode as

$$\dot{\omega}_m = -\frac{k_m k_{emf}}{R_m I_t} \omega_m - \frac{C_{dm}}{\eta_g r_g^3 I_t} \omega_m^2 + \frac{k_m}{R_m I_t} u \quad (8.15)$$

$$y = \frac{C_T}{r_g^2} \omega_m^2$$

Notably, both of these systems must be linearized about a trim condition for  $[\bar{i}_m \ \bar{\omega}_m]^T$  to obtain a second- or first-order LTI system.

## Momentum Exchange Devices

Spacecraft often use **momentum exchange devices (MED)**, i.e., an axisymmetric rotational device, to actuate control inputs by affecting the total angular momentum of the spacecraft. MEDs fundamentally contain a **flywheel**, also known as an **inertia wheel**, which can be used to store or transfer angular momentum to affect the vehicle's total angular momentum. **Momentum wheels** are flywheels designed to operate with some nonzero momentum and **reaction wheels** are flywheels designed to operate with zero momentum. These MEDs are typically connected rigidly to the satellite, but their flywheels are driven by electric motors to produce the necessary moment to the spacecraft.

To apply a moment to the spacecraft about some axis  $s$ , a moment in the opposite direction must be produced by the MED, i.e.,  $\dot{H}_s = -\dot{H}_w$  or

$$I_s \omega_s = -I_w \omega_w \quad (8.16)$$

where  $I_s$  is the moment of inertia of the spacecraft,  $\omega_s$  is the angular velocity of the spacecraft,  $I_w$  is the moment of inertia of the wheel, and  $\omega_w$  is the angular velocity of the wheel. Furthermore, the relative angular velocity between the wheel and the spacecraft,  $\omega_{rel}$ , provides damping moment of the wheel as

$$M_{D,w} = C_d \omega_{rel} = C_d (\omega_w - \omega_s) \quad (8.17)$$

where  $C_d$  is the viscosity damping coefficient sensed by the wheel.

With this relative angular velocity between the wheel and the spacecraft which is also the relative velocity between the rotor and stator of the electric motor, one has the dynamics for a single MED as the LTI state-space model

$$\begin{aligned} \begin{bmatrix} \dot{i}_m \\ \dot{\omega}_w \\ \dot{\omega}_s \end{bmatrix} &= \begin{bmatrix} -\frac{R_m}{L_m} & -\frac{k_{emf}}{L_m} & \frac{k_{emf}}{L_m} \\ \frac{k_m}{I_w} & -\frac{C_d}{I_w} & \frac{C_d}{I_w} \\ 0 & -\frac{1}{I_s} & 0 \end{bmatrix} \begin{bmatrix} i_m \\ \omega_w \\ \omega_s \end{bmatrix} + \begin{bmatrix} \frac{1}{L_m} \\ 0 \\ 0 \end{bmatrix} v_m \\ \dot{H}_w &= [k_m \quad -C_d \quad C_d] \begin{bmatrix} i_m \\ \omega_w \\ \omega_s \end{bmatrix} \end{aligned} \quad (8.18)$$

where  $\vec{x} = [i_m \ \omega_w \ \omega_s]^T$ ,  $u = v_m$  is the wheel's input voltage, and  $y = \dot{H}_w$  is the wheel's angular momentum rate. Assuming a low inductance BLDC electric motor leads to a second-order approximation of the MED dynamics by including only the mechanical mode of the wheel as

$$\begin{aligned} \begin{bmatrix} \dot{\omega}_w \\ \dot{\omega}_s \end{bmatrix} &= \begin{bmatrix} -\frac{k_m k_{emf}}{R_w I_w} - \frac{C_d}{I_w} & \frac{k_m k_{emf}}{R_w I_w} + \frac{C_d}{I_w} \\ -\frac{1}{I_s} & 0 \end{bmatrix} \begin{bmatrix} \omega_w \\ \omega_s \end{bmatrix} + \begin{bmatrix} \frac{k_m}{R_w I_w} \\ 0 \end{bmatrix} v_m \\ \dot{H}_w &= \left[ -\left( \frac{k_m k_{emf}}{R_w} + C_d \right) \quad \left( \frac{k_m k_{emf}}{R_w} + C_d \right) \right] \begin{bmatrix} \omega_w \\ \omega_s \end{bmatrix} + \frac{k_m}{R_w} v_m \end{aligned} \quad (8.19)$$

However, for control purposes, one typically commands a moment,  $M_c$ , to the wheel which can be related to the input voltage,  $v_m$ , as

$$\dot{v}_m = \frac{1}{k_m} M_c - \frac{1}{R_w} (K v_m - k_{emf} (\omega_w - \omega_s)) \quad (8.20)$$

where  $K$  is a control gain to be designed. This results in an LTI state-space model as

$$\begin{aligned} \begin{bmatrix} \dot{\omega}_w \\ \dot{\omega}_s \\ \dot{v}_m \end{bmatrix} &= \begin{bmatrix} -\frac{k_m k_{emf}}{R_w I_w} - \frac{C_d}{I_w} & \frac{k_m k_{emf}}{R_w I_w} + \frac{C_d}{I_w} & \frac{k_m}{R_w I_w} \\ -\frac{1}{I_s} & 0 & 0 \\ \frac{k_{emf}}{R_w} & -\frac{k_{emf}}{R_w} & -\frac{K}{R_w} \end{bmatrix} \begin{bmatrix} \omega_w \\ \omega_s \\ v_m \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{k_m} \end{bmatrix} M_c \\ \dot{H}_w &= -\left( \frac{k_m k_{emf}}{R_w} + C_d \right) \omega_w + \left( \frac{k_m k_{emf}}{R_w} + C_d \right) \omega_s + \frac{k_m}{R_w} v_m \end{aligned} \quad (8.21)$$

Next, assuming  $C_d \approx 0$ , one can show the transfer function between the commanded moment

$$\frac{\dot{H}_w}{M_c} = \frac{\frac{K}{R_m}}{s + \frac{K}{R_m} \left( 1 + \frac{k_m k_{emf}}{KI_w} \right)} \quad (8.22)$$

Thus, if one chooses  $K \gg k_m k_{emf}/I_w$ , then, one obtains the first-order approximation of the MED actuator as

$$\frac{\dot{H}_w}{M_c} = \frac{\frac{K}{R_m}}{s + \frac{K}{R_m}} \quad (8.23)$$

Often, the bandwidth of an MED would be greater than 100 Hz and can be neglected in the satellite attitude dynamics.

For multiple wheels, suppose one has a spacecraft of mass  $m_0$  with center of mass  $G$  and  $N$  wheels with angular velocities,  $\omega_1, \dots, \omega_N$ , masses  $m_1, \dots, m_N$ , and center of masses  $G_1, \dots, G_N$ . The mass of the entire system is

$$m = m_0 + \sum_{i=1}^N m_i \quad (8.24)$$

and the angular momentum of the entire system about the center of mass is

$$\vec{H}_G = \vec{H}_{G,v} + \vec{H}_w \quad (8.25)$$

where  $\vec{H}_{G,v}$  is the total angular momentum of the rigid body comprising the spacecraft and all the wheel masses concentrated at their centers of mass with common angular momentum,  $\vec{\omega}$ , and  $\vec{H}_w$  is the net angular momentum of the  $N$  flywheels about each of their center of masses.

Thus,

$$\vec{H}_{G,v} = I_{G,v} \vec{\omega} = \left( I_{G,0} + \sum_{i=1}^N I_{G,m_i} \right) \vec{\omega} \quad (8.26)$$

where  $I_{G,0}$  is the inertia tensor of the satellite body and  $I_{G,m_i}$  is the inertia tensor of the condensed point-mass of the  $i^{\text{th}}$  flywheel located at  $(x_i, y_i, z_i)$  in the body-fixed frame, i.e.

$$I_{G,m_i} = \begin{bmatrix} m_i(x_i^2 + z_i^2) & -m_i x_i y_i & -m_i x_i z_i \\ -m_i x_i y_i & m_i(y_i^2 + z_i^2) & -m_i y_i z_i \\ -m_i x_i z_i & -m_i y_i z_i & m_i(x_i^2 + y_i^2) \end{bmatrix} \quad (8.27)$$

Both of these are constant with respect to the body frame. Also, for the flywheels' net angular momentum

$$\vec{H}_w = \sum_{i=1}^N \vec{H}_{G_i} = \sum_{i=1}^N I_{G_i} \vec{\omega}_i \quad (8.28)$$

where  $\vec{H}_{G_i}$  is the angular momentum for the  $i^{\text{th}}$  flywheel and  $I_{G_i}$  is the inertia tensor for the  $i^{\text{th}}$  flywheel about its center mass,  $G_i$ , along axes parallel to the body-fixed frame. Notably,  $I_{G_i}$  may vary with time if the flywheels' can pivot on gimbals.

Altogether, for the wheel-equipped satellite attitude dynamics, one has

$$\vec{M}_{G,d} + \vec{M}_{G,c} = \dot{\vec{H}}_{G,v} + [\vec{\omega}] \times \vec{H}_{G,v} + \dot{\vec{H}}_w + [\vec{\omega}] \times \vec{H}_w \quad (8.29)$$

And for torque-free motion, one has from the conservation of angular momentum about the vehicle's center of mass

$$\dot{\vec{H}}_{G,v} + [\vec{\omega}] \times \vec{H}_{G,v} + \dot{\vec{H}}_w + [\vec{\omega}] \times \vec{H}_w = 0 \quad (8.30)$$

or

$$\vec{H}_{G,v} + \vec{H}_w = \mathbf{c} \quad (8.31)$$

where  $\mathbf{c}$  is a constant.

A **control moment gyro** (CMG), also known as a **gyrotorquer**, uses a single- or double-gimballed momentum wheel spinning at a constant rate. These gyros spin at several thousand revolutions per minute (rpm). At each gimbal axis are placed electric motors called **torquers** which exert moments on the satellite when tilting the gimbal normal to the gimbal axis.

Assuming the angular momentum is directed totally along the spin axis, one has for the  $i^{\text{th}}$  CMG

$$\vec{H}_{G_i} = I_{G_i} \vec{\omega}_{w,i} \quad (8.32)$$

where  $\vec{\omega}_{w,i}$  is the absolute angular velocity of the  $i^{\text{th}}$  spinning flywheel, i.e.,

$$\vec{\omega}_{w,i} = \vec{\omega} + \vec{\omega}_{p,i} + \vec{\omega}_{n,i} + \vec{\omega}_{s,i} \quad (8.33)$$

where  $\vec{\omega}$  is the angular velocity of the vehicle to which the CMG is attached,  $\vec{\omega}_{p,i}$  is the gyro's precession rate,  $\vec{\omega}_{n,i}$  is the gyro's nutation rate, and  $\vec{\omega}_{s,i}$  is the gyro's spin rate. However, CMGs are designed such that the  $\|\vec{\omega}_{s,i}\|_2$  is three or more orders of magnitude greater than the other three components.

Thus, for the axisymmetric flywheel in the CMG-fixed frame  $C$ , one has the simplified

$$\vec{H}_{G_i,C}^{(i)} = I_{G_i} \vec{\omega}_{s,i} = \begin{bmatrix} 0 \\ 0 \\ I_{3,i} \omega_{s,i} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ H_i \end{bmatrix} \quad (8.34)$$

Then, relative to the body-fixed frame of the satellite, one has

$$\vec{H}_{G_i,B}^{(i)} = \begin{bmatrix} H_i \sin \theta_i \cos \phi_i \\ H_i \sin \theta_i \sin \phi_i \\ H_i \cos \theta_i \end{bmatrix} \quad (8.35)$$

and for the attitude dynamics of a satellite with  $N$  CMGs, one has

$$\sum \vec{M}_G = \dot{\vec{H}}_{G,v} + [\vec{\omega}] \times \vec{H}_{G,v} + \sum_{i=1}^N \dot{\vec{H}}_{G_i} + [\vec{\omega}] \times \vec{H}_i \quad (8.36)$$

$$\begin{aligned} \sum \vec{M}_G &= \begin{bmatrix} I_1 p \\ I_2 q \\ I_3 r \end{bmatrix} + \begin{bmatrix} (I_3 - I_2) qr \\ (I_1 - I_3) pr \\ (I_2 - I_1) pq \end{bmatrix} \\ &\quad + \sum_{i=1}^N \left( \begin{bmatrix} \sin \theta_i \cos \phi_i \dot{H}_i + H_i \cos \theta_i \cos \phi_i \dot{\phi}_i - H_i \sin \theta_i \sin \phi_i \dot{\phi}_i \\ \sin \theta_i \sin \phi_i \dot{H}_i + H_i \cos \theta_i \sin \phi_i \dot{\phi}_i - H_i \sin \theta_i \cos \phi_i \dot{\phi}_i \\ \dot{H}_i \cos \theta_i - H_i \sin \theta_i \dot{\phi}_i \end{bmatrix} + \begin{bmatrix} q \cos \theta_i - r \sin \theta_i \sin \phi_i \\ r \sin \theta_i \cos \phi_i - p \cos \theta_i \\ p \sin \theta_i \sin \phi_i - q \sin \theta_i \cos \phi_i \end{bmatrix} H_i \right) \end{aligned} \quad (8.37)$$

$$\begin{aligned} \sum \vec{M}_G = & \begin{bmatrix} I_1 p + (I_3 - I_2) qr \\ I_2 q + (I_1 - I_3) pr \\ I_3 r + (I_2 - I_1) pq \end{bmatrix} \\ & + \sum_{i=1}^N \left( \begin{bmatrix} \sin \theta_i \cos \phi_i \dot{H}_i + H_i \cos \theta_i \cos \phi_i \dot{\theta}_i - H_i \sin \theta_i \sin \phi_i \dot{\phi}_i + (H_i \cos \theta_i) q - (H_i \sin \theta_i \sin \phi_i) r \\ \sin \theta_i \sin \phi_i \dot{H}_i + H_i \cos \theta_i \sin \phi_i \dot{\theta}_i - H_i \sin \theta_i \cos \phi_i \dot{\phi}_i + (H_i \sin \theta_i \cos \phi_i) r - (H_i \cos \theta_i) p \\ \dot{H}_i \cos \theta_i - H_i \sin \theta_i \dot{\theta}_i + p(H_i \sin \theta_i \sin \phi_i) - (H_i \sin \theta_i \cos \phi_i) q \end{bmatrix} \right) \end{aligned} \quad (8.38)$$

where the CMGs noticeably add to the coupling angular rates based on the CMG angles,  $\theta_i$  and  $\phi_i$ .

### Magnetorquers

A **magnetorquer**, also known as a **magnetic torquer**, is a set of electromagnets designed to produce a rotationally asymmetric magnetic field over an extended area of the spacecraft that is controlled with current flow through the coils. This produces a magnetic dipole,  $\vec{m}$ , given by

$$\vec{m} = ni\vec{A} \quad (8.39)$$

where  $n$  is the number of turns of the coil,  $i$  is the input current, and  $\vec{A}$  is the vector area of the coil. The electromagnets are rigidly attached to the spacecraft so that any magnetic force they exert on the magnetic field intensity of the orbited body,  $\vec{B}$ , will lead to a magnetic reverse force and result in mechanical torque,  $\vec{M}$ , given by

$$\vec{M} = \begin{bmatrix} M_x \\ M_y \\ M_z \end{bmatrix} = [\vec{m}] \times \vec{B} = -[\vec{B}] \times \vec{M} = \begin{bmatrix} 0 & B_z & -B_y \\ -B_z & 0 & B_x \\ B_y & -B_x & 0 \end{bmatrix} \begin{bmatrix} m_x \\ m_y \\ m_z \end{bmatrix} \quad (8.40)$$

In control systems, one would command a particular  $\vec{M}_c$ , but the skew-symmetric matrix is singular, so one typically replaces one of the electromagnets with a reaction wheel, e.g., for the  $y$ -axis, one obtains the relationship

$$\begin{bmatrix} M_x \\ M_y \\ M_z \end{bmatrix} = \begin{bmatrix} 0 & 0 & -B_y \\ -B_z & 1 & B_x \\ B_y & 0 & 0 \end{bmatrix} \begin{bmatrix} m_x \\ \dot{H}_{w,y} \\ m_z \end{bmatrix} \quad (8.41)$$

which has an inverse between the moment commands and the applied moments as

$$\begin{bmatrix} m_x \\ \dot{H}_{w,y} \\ m_z \end{bmatrix} = \frac{1}{B_y^2} \begin{bmatrix} 0 & 0 & B_y \\ B_x B_y & B_y^2 & B_y B_z \\ -B_y & 0 & 0 \end{bmatrix} \begin{bmatrix} M_{x,c} \\ M_{y,c} \\ M_{z,c} \end{bmatrix} \quad (8.42)$$

### Thruster Systems

A **thruster** is used to apply a pure thrust force to a spacecraft for orbital maneuvers. Thrusters can be designed using rocket or electric thrusters. Rocket thrusters include solid and liquid rocket engines while electric thrusters include (1) ion thrusters which use magnetism or static electricity to push ions out of the spacecraft, (2) Hall effect thrusters which use electricity to accelerate and ionize a neutral gas, (3) electrospray thrusters which use a high electric field to emit positively or negatively charged particles from an electrically

conductive liquid, (4) an arcjet, a magnetoplasmadynamic thruster which use the Lorentz force to generate thrust, (5) pulsed plasma thrusters, and (6) resistojet which use a resistor to heat a non-reactive fluid, which is then expelled through a nozzle. Regardless of the technology, a thruster allows one to directly change the linear momentum,  $\Delta \vec{P}$ , impulsively, i.e.,

$$\Delta \vec{P} = \int_0^{\Delta t} \vec{T} dt \quad (8.43)$$

where  $\Delta t$  is the time length of the impulse and  $\vec{T}$  is the thrust produced by the thruster.

A **thruster pair** is used to apply a pure moment to the spinning spacecraft. Thruster pairs are small thrusters mounted in principal planes, i.e., planes normal to principal axes, passing through the center of mass of the spacecraft. Thruster pairs are also accompanied by another thruster pair of rockets pointing in the opposite direction to exert a moment in the opposite sense. A thruster pair allows one to directly change the angular momentum,  $\Delta \vec{H}_G$ , impulsively, i.e.

$$\Delta \vec{H}_G = \int_0^{\Delta t} \vec{M}_G dt \quad (8.44)$$

where  $\Delta t$  is the time length of the impulse,  $\vec{M}$  is the moment generated by the thruster pair, and  $\Delta \vec{H}_G$  is typically applied normal to the spin axis, i.e., the nominal axis for  $H_G$ .

Suppose the thruster pairs are positioned at some  $\vec{r}_T$  and  $-\vec{r}_T$  to the center of mass and if  $\vec{T}$  is their individual thrust, one has for the change in angular momentum after some brief time interval

$$\Delta \vec{H}_G = \int_0^{\Delta t} [\vec{r}_T]_{\times} \vec{T} \Delta t + [-\vec{r}_T]_{\times} (-\vec{T}) \Delta t \quad (8.45)$$

or

$$\Delta \vec{H}_G = 2[\vec{r}_T]_{\times} \vec{T} \Delta t \quad (8.46)$$

## References

For more information, please refer to the following

- Close, C. M., Frederick, D. K., and Newell, J. C., “Modeling and Analysis of Dynamic Systems,” 3rd ed., Vol. 1, John Wiley & Sons, 2002, pp. 347-351
- Curtis, H. D., “10.7 Attitude Control Thrusters,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 506-509
- Nelson, R. C., “8.3 Control Surface Actuator,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 288-292
- Sidi, M. J., “7.3.1 Model of a Momentum Exchange Device,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 161-164
- Sidi, M. J., “7.4.1 Basic Magnetic Torque Control Equation,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 185-186
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “4.6 Autopilots,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 322-344

## 8.2 Aerospace Vehicle Guidance and Control Systems

### Guidance and Control Systems

For MIMO systems, such as aerospace vehicles, one may use state-space methods to develop suitable controllers, typically with optimal, adaptive, and/or robust control considerations. However, for many systems with only a few number of inputs and/or outputs, the different outputs that one desires to control may have responses on much different time scales which allow SISO methods to be applied at these different time scales through cascade control.

For example, consider the LTI systems represented by the following transfer functions:

$$G_{fast}(s) = \frac{y_{fast}(s)}{u(s)} \quad (8.47)$$

and

$$G_{slow}(s) = \frac{y_{slow}(s)}{u(s)} \quad (8.48)$$

Next, assuming that a change in  $y_{fast}$  naturally causes a change in  $y_{slow}$  with the relationship, i.e.

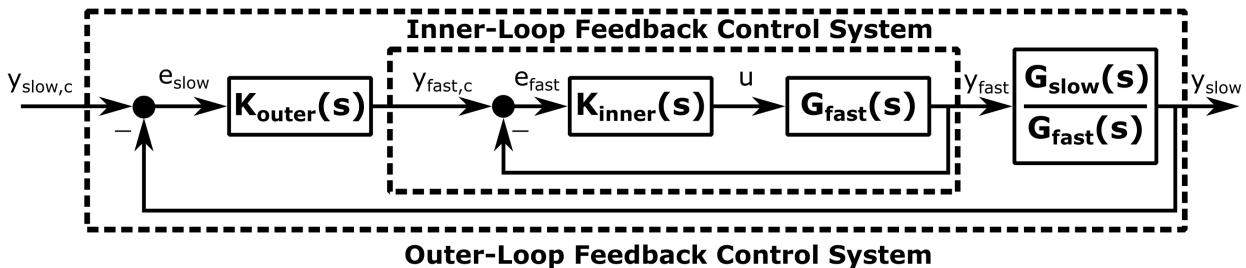
$$\frac{y_{slow}(s)}{y_{fast}(s)} = \frac{G_{slow}(s)}{G_{fast}(s)} \quad (8.49)$$

one has the fact that

$$\frac{y_{slow}(s)}{u(s)} = \frac{G_{slow}(s)}{G_{fast}(s)} G_{fast}(s) \quad (8.50)$$

Then, by assuming that  $y_{fast}$  and  $y_{slow}$  are related to one another through at least one integration, one knows that the denominator of  $\frac{G_{slow}(s)}{G_{fast}(s)}$  is at least order one higher in  $s$  than the numerator.

With this in mind, consider the following block diagram which uses two linked feedback control systems, i.e. an outer feedback control system and an inner feedback control system.



where this technique is known as **cascade control**, **nested-loop control**, or **inner-outer loop control**. Although this control design may seem complicated, with proper loop-shaping, this type of feedback control system can be simpler to design than a MIMO system while also being robust with SISO LTI system stability margins.

For cascade control, one designs  $K_{inner}(s)$  such that  $y_{fast}$  tracks  $y_{fast,c}$  for all frequencies up to  $\omega_{c,inner}$ . Then, note that this makes the inner-loop feedback control system have a transfer function from  $y_{fast,c}$  to  $y_{fast}$  as

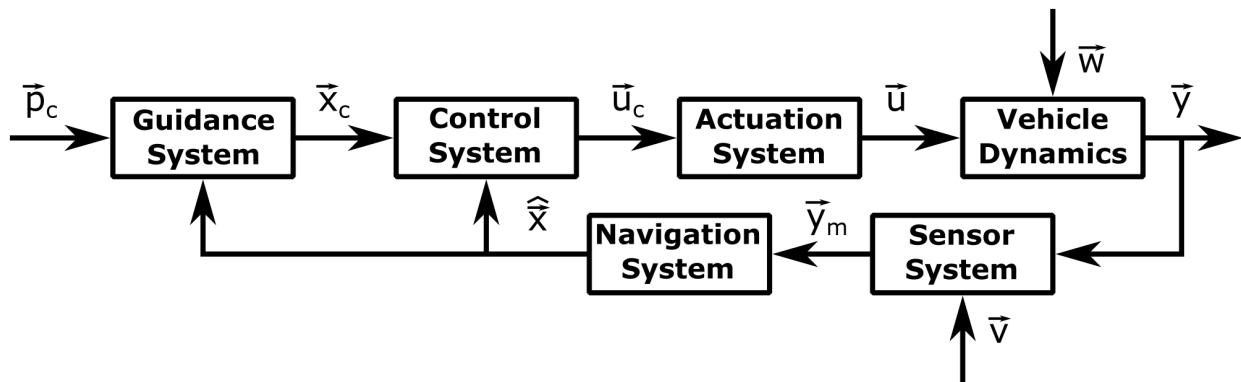
$$\frac{y_{fast}(s)}{y_{fast,c}(s)} \approx \frac{\omega_{c,inner}}{s + \omega_{c,inner}} \quad (8.51)$$

which for  $\omega < \omega_{c,inner}$ , this transfer function will be approximately unity below the inner-loop crossover frequency  $\omega_{c,inner}$ , i.e. the inner-loop bandwidth. This fact makes the preliminary design of the outer-loop much easier where assuming Equation 8.51, one can design  $K_{outer}(s)$  such that  $y_{slow}$  tracks  $y_{slow,c}$  for all frequencies up to  $\omega_{c,outer}$  which takes advantage of the simpler form of  $\frac{G_{slow}(s)}{G_{fast}(s)}$  as opposed to  $G_{slow}(s)u(s)$  coupled with  $G_{fast}(s)u(s)$ . This two-step approach thus can be seen as much simpler than a one-step coupled design approach.

However, this cascade control method requires that there is sufficient **crossover frequency separation** between the inner- and outer-loops, i.e.  $\omega_{c,outer} < \omega_{c,inner}$  by some amount such that the assumption of Equation 8.51 applies. For systems where there is already a natural frequency separation between  $G_{outer}(s)$  and  $G_{inner}(s)$ , this can be physically realizable, but must be conserved when performing the loop-shaping of the inner- and outer-loops. In particular, standard loop-shaping of  $L_{inner} = G_{inner}K_{inner}$  and  $L_{outer} = G_{outer}K_{outer}$  will set  $|L_{inner}|$  and  $|L_{outer}|$  large for  $\omega < \omega_{c,inner}$  and  $\omega < \omega_{c,outer}$ , respectively, and small for  $\omega > \omega_{c,inner}$  and  $\omega > \omega_{c,outer}$ , respectively. Thus, by designing  $\omega_{c,outer} < \omega_{c,inner}$  closing the outer-loop around the inner-loop will only slightly modify  $|L_{inner}|$  for  $\omega < \omega_{c,inner}$  where the magnitude was already large. In a similar manner, this type of logic can also be used extended to demonstrating that cascade design will have little to no affect on the inner- and outer-loop bandwidths and stability margins using loop-shaping. It should also be noted that in this way, multiple inner-loops can be cascaded which is typical for vehicle feedback control systems.

## Guidance, Navigation, and Control Systems

The entire control system for vehicles are also known as **guidance, navigation, and control (GNC)** systems which refers to the three traditional sub-systems used to control the motion of vehicles. The block diagram of a typical GNC system is shown as follows.



where the vehicle state vector,  $\vec{x}$ , typically includes the position, velocity, and attitude of the vehicle. Here, the vehicle's state,  $\vec{x}$ , evolves in time according to its natural dynamics, the control inputs,  $\vec{u}$ , and any disturbances,  $\vec{w}$ , from the environment. It should be noted that for manual control of aerospace vehicles, the guidance and control systems can be considered as the human operator and not a computer system.

At the base of guidance, navigation, and control systems is the necessity of the pilot or autopilot to use a **sensor system** to sense the vehicle's state vector. In addition, a corresponding **navigation system** is typically necessary to process, i.e. "filter," the raw signals obtained by the sensor system. Furthermore, when only certain aspects of the vehicle dynamics, i.e. the outputs  $\vec{y}$ , are measurable by the sensor system, the navigation system estimates the state vector,  $\hat{x}$ , that is required for the guidance and control systems to operate. Thus, **navigation** can be defined as **vehicle (self-)state estimation**. This part of the textbook introduces some sensor systems for direct attitude determination while later parts of this textbook further develop the theory and design of full navigation systems for aerospace vehicles.

Given the current vehicle state vector, the **guidance system** determines the commanded vehicle state vector,  $\vec{x}_c$ , for following some reference path or trajectory,  $\vec{p}_{traj}$ , e.g. determines a commanded velocity, attitude, and acceleration. Notably, this system is often simply controlled by a human operator and does not use automatic control algorithms. However, as the input to the operator is the available human sensory information and/or an information display from sensors, as it would be for electronic information from the navigation system, the operator is still using the basic logic of GNC reasoning. The **control system** achieves the commanded state,  $\vec{x}_c$ , while also maintaining vehicle dynamic stability and typically uses either automatic or semi-automatic control. Furthermore, as the control system include the direct manipulation of the forces applied to the vehicle through some sort of physical phenomena, one should also consider the effects of any **actuation system** on the GNC system. Coupled with the vehicle dynamics, these actuator dynamics play a vital role in the control system design for vehicles.

The vast majority of aerospace vehicles use a cascade control scheme for guidance and control which takes advantage of the physics-based relationships between position, velocity, and acceleration as well as angular versus linear states, both of which naturally have frequency separation required for cascade control. As such the different control loops for vehicles have been given different names: planning, guidance, and control. From a feedback control perspective for aerospace vehicles, planning roughly corresponds to position or trajectory feedback control, guidance to velocity feedback control, and control to attitude feedback control directly through the actuation system. The actuation system is the primary limiting factor for the inner-loop control systems of aerospace vehicles. It should also be noted that some aerospace vehicles may also have a SAS to alter the modal characteristics of the vehicle. In many cases, the reference trajectory may simply be commands for speed, altitude, heading. More generally, the planned trajectory may be a line in space and the system may be some form of **line-following guidance law**. Multiple lines may connect together to form a sequence of **waypoints**. Then, the switching from each line connecting waypoints is performed by the planning system of the aerospace vehicle as it assesses the current position estimate of the aerospace vehicle and often uses some success and feasibility criteria checks for reaching each waypoint.

The reference path or trajectory,  $\vec{p}_c$ , determined using **planning** at the mission level and the path level. Using the most general definition of a **plan**, i.e. a task for a vehicle that transports a payload, a vehicle must travel from current location to one (or more) designated target(s). The determination of the target is known as **mission planning** and the determination of the path to the target is known as **path planning**, also known as **trajectory planning**. As such, one can also consider the operation of autonomous aerospace vehicles at a level higher than GNC and is generally performed by a **planning system** that determines the

current designated target and/or the reference path or trajectory that the vehicle should follow from its current location to the target. This stage is often completely manually derived and pre-programmed without real-time signal inputs, but autonomous planning of aerospace vehicles is also possible. Lastly, if one must reach a dynamic target or targets, an additional **target tracking system** will be necessary to accomplish the flight plan and typically will feed the target state to the guidance system. Furthermore, if more than one target specified, a planning system must have feedback from the navigation and target tracking system to assess when to switch from targets. Later parts of this textbook further develop the theory and design of planning and tracking systems for aerospace vehicles.

Beyond this general framework, it should be noted that sometimes the guidance and control systems are combined into one feedback control system which is possible when state-space methods are used for feedback control system design to provide faster system responses in aerospace vehicle dynamics than is possible with a traditional cascade control design. It should also be noted that when *designing* the guidance and control systems, one is typically using linearized vehicle dynamics models. Thus, aerospace vehicles often use some form of **adaptive control** which alters the guidance and control laws based on current operating conditions. A basic example of this is **gain scheduling** where the guidance and control laws or “gains” change according to the “scheduled” trim conditions. However, when employing gain scheduling, one typically requires that the transition from one scheduled trim condition to another is relatively “smooth,” i.e. the nonlinear dynamics are not significantly excited by the control inputs during the transitions between trim conditions. This is typically dealt with by designing sufficient stability robustness for each LTI controller to satisfy this requirement. As part of the gain scheduling of aircraft, one typically requires the estimation of the current flight conditions in real-time. This is often performed by an onboard sensor system known as an **air data system (ADS)** which provide measurements of a aerospace vehicle’s surrounding air mass, collectively known as the **air data**, typically quantified as the airspeed, angle of attack, sideslip angle, and perhaps the altitude and rate of climb. The airspeed, angle of attack, and sideslip angle are also known as the **air data triplet**.

## References

For more information, please refer to the following

- Schmidt, D. K., “12.2 Inner and Outer Loops, and Frequency Separation,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 703-706
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “4.6 Autopilots,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 322-344

## 8.3 Intercept and Rendezvous Guidance Systems

The **intercept guidance problem**, also known as **pursuit-evasion guidance problem**, is defined

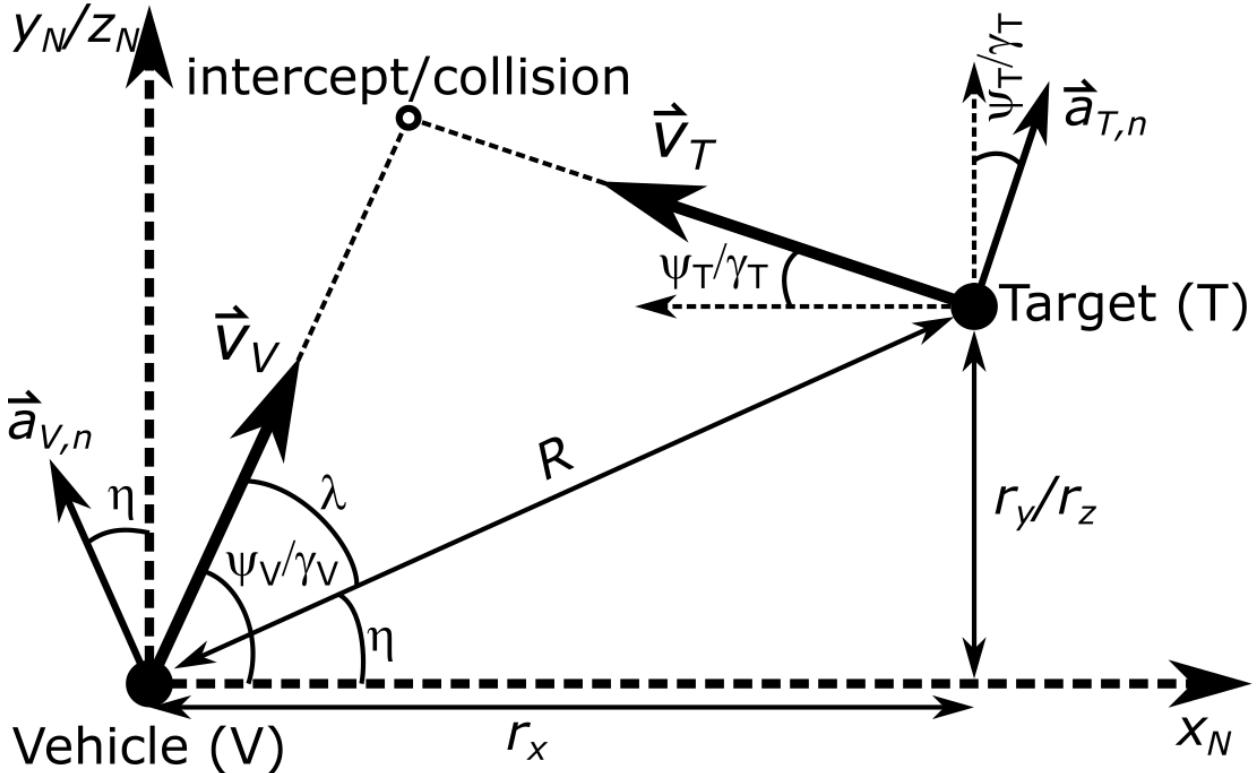
The **rendezvous guidance problem** is defined as

**mid-course guidance**

**terminal guidance**

### Planar Terminal Guidance

While flight vehicles operate in three-dimensional space, most intercept and rendezvous guidance laws for terminal guidance devise and implement planar guidance laws in each of the two maneuver planes, e.g. longitudinal and lateral-directional. This problem is depicted in the following diagram



where  $\psi_V/\gamma_V$  is the heading/flight-path angle of the vehicle,  $\psi_T/\gamma_T$  is the heading/flight-path angle of the target,  $\vec{v}_V$  is the velocity of the vehicle,  $\vec{v}_T$  is the velocity of the target,  $r_x$  is the  $x_N$  relative position,  $r_y/r_z$  is the relative  $y_N/z_N$  position,  $R$  is the range-to-target,  $\eta$  is the **line-of-sight (LOS) angle**,  $\vec{a}_{V,n}$  is the acceleration of the vehicle normal to the LOS,  $\vec{a}_{T,n}$  is the acceleration of the target normal to its velocity, and  $\lambda$  is the **lead angle** between the LOS vector and the vehicle velocity vector which must be positive and  $\dot{R} < 0$  for an intercept to exist. Determining this lead angle is a key purpose of the guidance law. In this case, if the vehicle actively reduces and holds  $r_y/r_z$  to zero, then  $r_x$  will continue to decrease until collision occurs. Thus, regulating  $r_y/r_z$  is the key analysis required for planar intercept.

In Cartesian form, the intercept kinematics can be stated as the target-vehicle relative position,  $\vec{r}$ ,

$$\vec{r} = \vec{r}_T - \vec{r}_V = \begin{bmatrix} r_x \\ r_y \end{bmatrix} = \begin{bmatrix} R \cos \eta \\ R \sin \eta \end{bmatrix} \quad (8.52)$$

the relative velocity

$$\vec{v} = \vec{v}_T - \vec{v}_V = \begin{bmatrix} v_x \\ v_y \end{bmatrix} = \begin{bmatrix} -\|\vec{v}_T\|_2 \cos \gamma_T - \|\vec{v}_V\|_2 \cos(\lambda + \eta) \\ \|\vec{v}_T\|_2 \sin \gamma_T - \|\vec{v}_V\|_2 \sin(\lambda + \eta) \end{bmatrix} \quad (8.53)$$

and the relative acceleration

$$\vec{a} = \vec{a}_T - \vec{a}_V = \begin{bmatrix} a_x \\ a_y \end{bmatrix} = \begin{bmatrix} \|\vec{a}_{T,n}\|_2 \sin \gamma_T + \|\vec{a}_{V,n}\|_2 \sin \eta \\ \|\vec{a}_{T,n}\|_2 \cos \gamma_T - \|\vec{a}_{V,n}\|_2 \cos \eta \end{bmatrix} \quad (8.54)$$

These are clearly nonlinear equations.

However, with small angle approximations for  $\gamma_T$  and  $\eta$ , i.e. near-collision course conditions, one has

$$r_y \approx R\eta \quad (8.55)$$

and

$$a_y \approx a_{T,n} - a_{V,n} \quad (8.56)$$

which allows the kinematics

$$v_y = \int \|\vec{a}_{T,n}\|_2 - \|\vec{a}_{V,n}\|_2 \quad (8.57)$$

and

$$\ddot{r}_y = \int \int \|\vec{a}_{T,n}\|_2 - \|\vec{a}_{V,n}\|_2 \quad (8.58)$$

which can be written as an augmented LTI state-space system

$$\begin{bmatrix} \dot{r}_y \\ \dot{v}_y \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} r_y \\ v_y \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} \|\vec{a}_{V,n}\|_2 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|\vec{a}_{T,n}\|_2 \quad (8.59)$$

$$\vec{y} = \begin{bmatrix} r_y \\ v_y \end{bmatrix}$$

which can be used to form different linear-quadratic guidance laws depending on what additional assumptions are made. Note that if  $r_y$  and  $v_y$  are not both available for feedback, then one can employ the separation principle and use a parallel guidance filter to estimate  $r_y$  and  $v_y$ .

## Constant-Velocity Guidance Laws

Assuming the vehicle and target speeds,  $\|\vec{v}_V\|_2$  and  $\|\vec{v}_T\|_2$ , are constant, the target is non-maneuvering, i.e.  $\|\vec{a}_{T,n}\|_2 = 0$ , and the vehicle responds instantaneously to an acceleration command,  $a_{V,c} = \|\vec{a}_{V,n}\|_2$ , one can form the following LQ minimum-energy OCP as

$$\begin{aligned} \vec{u}^{\text{opt}}(t) &= \underset{u(t) \forall t \in [0, t_f]}{\text{argmin}} \quad \mathcal{J} = \vec{x}^T(t_f) \begin{bmatrix} E_r & 0 \\ 0 & E_v \end{bmatrix} \vec{x}(t_f) + \int_0^{t_f} \vec{u}^T(t) \vec{u}(t) dt \\ \text{subject to: } \dot{\vec{x}}(t) &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \vec{x}(t) + \begin{bmatrix} 0 \\ -1 \end{bmatrix} \vec{u}(t) \\ \text{initial condition: } \vec{x}(0) & \end{aligned} \quad (8.60)$$

where  $\vec{x} = [r_y \ v_y]^T$ ,  $u = a_{V,c}$ ,  $E_r$  is the cost on the relative range at time  $t_f$ , and  $E_v$  is the cost on the relative velocity at time  $t_f$ . When  $E_r \rightarrow \infty$  and  $E_v = 0$ , one has an **intercept problem** and when  $E_r \rightarrow \infty$  and  $E_v \rightarrow \infty$ , one has an **rendezvous problem**.

The solution to this OCP is given by the differential Riccati equation

$$\dot{P} = -PA - A^T P + PBB^T P \quad (8.61)$$

with endpoint condition

$$P(t_f) = \begin{bmatrix} E_r & 0 \\ 0 & E_v \end{bmatrix} \quad (8.62)$$

where the optimal control is given by

$$u^{opt}(t) = -B^T P(t) \vec{x} \quad (8.63)$$

which can be solved as

$$u^{opt}(t) = \frac{3}{t_{go}^2} \left[ \frac{\left(1 + \frac{1}{2}E_v t_{go}\right) r_y(t) + \left(1 + \frac{1}{3}E_v t_{go} + \frac{E_v}{E_r t_{go}^2}\right) v_y(t) t_{go}}{1 + \frac{3}{E_r t_{go}^3} (1 + E_v t_{go}) + \frac{E_v t_{go}}{4}} \right] \quad (8.64)$$

where  $t_{go} = t_f - t$ .

This can be expressed as the following structure

$$u^{opt}(t) = \frac{\tilde{N}(E_r, E_v, t_{go})}{t_{go}^2} Z(r_y, v_y, E_r, E_v, t_{go}) \quad (8.65)$$

with

$$\tilde{N}(E_r, E_v, t_{go}) = \frac{3}{1 + \frac{3}{E_r t_{go}^3} (1 + E_v t_{go}) + \frac{E_v t_{go}}{4}} \quad (8.66)$$

as the **effective navigation ratio** and

$$Z(r_y, v_y, E_r, E_v, t_{go}) = \left(1 + \frac{1}{2}E_v t_{go}\right) r_y(t) + \left(1 + \frac{1}{3}E_v t_{go} + \frac{E_v}{E_r t_{go}^2}\right) v_y(t) t_{go} \quad (8.67)$$

The **proportional navigation guidance (PN) guidance** occurs for the intercept problem, i.e. if  $E_v = 0$  and as  $E_r \rightarrow \infty$ , one has

$$u_{PN}(t) = \frac{3}{t_{go}^2} [r_y(t) + v_y(t)t_{go}] \quad (8.68)$$

where  $\tilde{N}_{PN} = 3$  and  $Z = ZEM_{PN} = r_y(t) + v_y(t)t_{go}$  which is often referred to as the **zero-effort-miss**, i.e. the current miss distance that would result if the vehicle and the target did not maneuver over the time period  $[t, t_f]$ , which, in practice, is estimated and fed to the controller by a guidance filter. Likewise, the **rendezvous (REN) guidance** law occurs for the rendezvous problem, i.e. as  $E_v, E_r \rightarrow \infty$ , one has

$$u_{REN}(t) = \frac{6}{t_{go}^2} \left[ r_y(t) + \frac{2}{3}v_y(t)t_{go} \right] \quad (8.69)$$

As an aside, recall that

$$r_y \approx R\lambda \quad (8.70)$$

$$v_y = \dot{R}\lambda + R\dot{\lambda} \quad (8.71)$$

where  $R = -\dot{R}t_{go}$ . Thus, substituting for  $\lambda$  and  $R$ , one has

$$v_y = \dot{R} \frac{r_y}{-\dot{R}t_{go}} - \dot{R}t_{go}\dot{\lambda} \quad (8.72)$$

and rearranging, one has

$$\dot{R}\dot{\lambda} = \frac{r_y(t) + v_y(t)t_{go}}{t_{go}^2} \quad (8.73)$$

and the PN guidance law can be alternatively expressed as

$$u_{PN}(t) = -\tilde{N}\dot{R}\dot{\lambda} \quad (8.74)$$

where  $\tilde{N} = 3$  is the “energy-optimal” PN gain as shown previously.

### Constant-Acceleration Guidance Law

Next, assume that the target is undergoing a constant acceleration,  $\|\vec{a}_{T,n}\|_2 = a_{T,n,y}$ , which can be added to the state and estimated. Then, one can form the following LQ minimum-energy OCP as

$$\begin{aligned} \vec{u}^{opt}(t) &= \underset{u(t) \forall t \in [0, t_f]}{\operatorname{argmin}} \quad \mathcal{J} = \vec{x}^T(t_f) \begin{bmatrix} E_r & 0 & 0 \\ 0 & E_v & 0 \\ 0 & 0 & 0 \end{bmatrix} \vec{x}(t_f) + \int_0^{t_f} \vec{u}^T(t) \vec{u}(t) dt \\ \text{subject to: } \vec{x}(t) &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \vec{x}(t) + \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} \vec{u}(t) \\ \text{initial condition: } \vec{x}(0) & \end{aligned} \quad (8.75)$$

where  $\vec{x} = [r_y \ v_y \ a_{T,n,y}]^T$  and  $u = \|\vec{a}_{V,n}\|_2$ .

This provides the following general solution

$$u^{opt}(t) = \frac{3}{t_{go}^2} \left[ \frac{\left(1 + \frac{1}{2}E_v t_{go}\right)r_y(t) + \left(1 + \frac{1}{3}E_v t_{go} + \frac{E_v}{E_r t_{go}^2}\right)v_y(t)t_{go} + \frac{1}{2}\left(1 + \frac{1}{6}E_v t_{go} + \frac{2E_v}{E_r t_{go}^2}\right)a_{T,n,y}(t)t_{go}^2}{1 + \frac{3}{E_r t_{go}^3}(1 + E_v t_{go}) + \frac{E_v t_{go}}{4}} \right] \quad (8.76)$$

which show that the only difference is the additional term in the numerator which is time-varying gain multiplying the target acceleration state.

The **augmented proportional navigation guidance (APN) guidance** occurs for the intercept problem, i.e. if  $E_v = 0$  and as  $E_r \rightarrow \infty$ , one has

$$u_{APN}(t) = \frac{3}{t_{go}^2} \left[ r_y(t) + v_y(t)t_{go} + \frac{1}{2}a_{T,n,y}t_{go}^2 \right] \quad (8.77)$$

where  $\tilde{N}_{APN} = 3$  and  $Z = ZEM_{APN} = r_y(t) + v_y(t)t_{go} + \frac{1}{2}a_{T,n,y}t_{go}^2$ . Thus, the only change in the APN from PN is a change in the ZEM estimate. As such, the APN guidance law will only perform better than PN if the vehicle is able to sufficiently estimate the target normal acceleration.

Similarly, the **augmented rendezvous (AREN) guidance** law occurs for the rendezvous problem, i.e. as  $E_v, E_r \rightarrow \infty$ , one has

$$u_{AREN}(t) = \frac{6}{t_{go}^2} \left[ r_y(t) + \frac{2}{3} v_y(t) t_{go} \right] + a_{T,n,y} \quad (8.78)$$

which is the same as REN, but adds a direct cancellation of the target maneuver in the acceleration command.

Furthermore, consider the case where the vehicle does not respond immediately to an acceleration command, but experiences some lag modeled as a first-order transfer function

$$\frac{\|\vec{a}_{V,n}\|_2}{a_{V,c}} = \frac{\omega}{s + \omega} \quad (8.79)$$

Then, one can form the following LQ minimum-energy OCP as

$$\begin{aligned} \vec{u}^{opt}(t) &= \underset{u(t) \forall t \in [0, t_f]}{\operatorname{argmin}} \quad \mathcal{J} = \vec{x}^T(t_f) \begin{bmatrix} E_r & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \vec{x}(t_f) + \int_0^{t_f} \vec{u}^T(t) \vec{u}(t) dt \\ \text{subject to: } \dot{\vec{x}}(t) &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\omega \end{bmatrix} \vec{x}(t) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \omega \end{bmatrix} \vec{u}(t) \quad (8.80) \\ \text{initial condition: } \vec{x}(0) \end{aligned}$$

where  $\vec{x} = [r_y \ v_y \ a_{T,n,y} \ \|\vec{a}_{V,n}\|_2]^T$  and  $u = a_{V,c}$ .

This provides the following general solution

$$u^{opt}(t) = \frac{6\omega^2 t_{go}^2 (\omega t_{go} + e^{-\omega t_{go}} - 1)}{t_{go}^2} \left[ \frac{r_y(t) + v_y(t) t_{go} + \frac{1}{2} a_{T,n,y}(t) t_{go}^2 - \frac{1}{\omega^2} (\omega t_{go} + e^{-\omega t_{go}} - 1) \|\vec{a}_{V,n}\|_2}{\frac{6\omega^3}{E_r} + 3 + 6\omega t_{go} - 6\omega^2 t_{go}^2 + 2\omega^3 t_{go}^3 - 12\omega t_{go} e^{-\omega t_{go}} - 3e^{-2\omega t_{go}}} \right] \quad (8.81)$$

which is considerably more complex than the non-ideal vehicle response to an acceleration command. The **optimal guidance law (OGL)** occurs for the intercept problem, i.e. as  $E_v \rightarrow \infty$ , one has

$$u_{OGL}(t) = \frac{6\omega^2 t_{go}^2 (\omega t_{go} + e^{-\omega t_{go}} - 1)}{t_{go}^2} \left[ \frac{r_y(t) + v_y(t) t_{go} - \frac{1}{\omega^2} (\omega t_{go} + e^{-\omega t_{go}} - 1) \|\vec{a}_{V,n}\|_2}{3 + 6\omega t_{go} - 6\omega^2 t_{go}^2 + 2\omega^3 t_{go}^3 - 12\omega t_{go} e^{-\omega t_{go}} - 3e^{-2\omega t_{go}}} \right] \quad (8.82)$$

where the time-varying effective navigation ratio is

$$\tilde{N}_{OGL} = \frac{6\omega^2 t_{go}^2 (\omega t_{go} + e^{-\omega t_{go}} - 1)}{3 + 6\omega t_{go} - 6\omega^2 t_{go}^2 + 2\omega^3 t_{go}^3 - 12\omega t_{go} e^{-\omega t_{go}} - 3e^{-2\omega t_{go}}} \quad (8.83)$$

and zero-effort-miss is

$$ZEM_{OGL} = r_y(t) + v_y(t) t_{go} - \frac{1}{\omega^2} (\omega t_{go} + e^{-\omega t_{go}} - 1) \|\vec{a}_{V,n}\|_2 \quad (8.84)$$

which simply adds another term to the  $ZEM_{APN}$  based on the response lag.

### Constant-Jerk Guidance Laws

Finally, assume that the target is undergoing a constant jerk,  $j_{T,n,y}$ , which can be added to the state along with the target acceleration and estimated. Then, one can form the following LQ minimum-energy OCP as

$$\vec{u}^{\text{opt}}(t) = \underset{u(t) \forall t \in [0, t_f]}{\text{argmin}} \quad \mathcal{J} = \vec{x}^T(t_f) \begin{bmatrix} E_r & 0 & 0 & 0 \\ 0 & E_v & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \vec{x}(t_f) + \int_0^{t_f} \vec{u}^T(t) \vec{u}(t) dt$$

subject to:  $\dot{\vec{x}}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \vec{x}(t) + \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \end{bmatrix} \vec{u}(t)$

initial condition:  $\vec{x}(0)$

(8.85)

where  $\vec{x} = [r_y \ v_y \ a_{T,n,y} \ j_{T,n,y}]^T$  and  $u = \|\vec{a}_{V,n}\|_2$ .

This provides the following general solution

$$u^{\text{opt}}(t) = \frac{3}{t_{go}^2} \left[ \frac{\left(1 + \frac{1}{2}E_v t_{go}\right)r_y(t) + \left(1 + \frac{1}{3}E_v t_{go} + \frac{E_v}{E_r t_{go}^2}\right)v_y(t)t_{go} + \frac{1}{2}\left(1 + \frac{1}{6}E_v t_{go} + \frac{2E_v}{E_r t_{go}^2}\right)a_{T,n,y}(t)t_{go}^2 + \frac{1}{6}\left(1 + \frac{3E_v}{E_r t_{go}^2}\right)j_{T,n,y}(t)t_{go}^3}{1 + \frac{3}{E_r t_{go}^3}(1 + E_v t_{go}) + \frac{E_v t_{go}}{4}} \right]$$
(8.86)

which show that the only difference is the additional term in the numerator which is time-varying gain multiplying the target acceleration state.

The **extended proportional navigation guidance (EPN) guidance** occurs for the intercept problem, i.e. if  $E_v = 0$  and as  $E_r \rightarrow \infty$ , one has

$$u_{EPN}(t) = \frac{3}{t_{go}^2} \left[ r_y(t) + v_y(t)t_{go} + \frac{1}{2}a_{T,n,y}t_{go}^2 + \frac{1}{6}j_{T,n,y}(t)t_{go}^3 \right]$$
(8.87)

where  $\tilde{N}_{EPN} = 3$  and  $Z = ZEM_{EPN} = r_y(t) + v_y(t)t_{go} + \frac{1}{2}a_{T,n,y}t_{go}^2 + \frac{1}{6}j_{T,n,y}(t)t_{go}^3$ . Thus, the only change in the EPN from APN and PN is a change in the ZEM estimate. As such, the EPN guidance law will only perform better if the vehicle is able to sufficiently estimate the target normal acceleration *and* its jerk.

### $L_1$ Guidance Law

### References

For more information, please refer to the following:

- Palumbo, N. F., Blauwkamp, R. A., and Lloyd, J. M., “Modern Homing Missile Guidance Theory and Techniques,” in *John Hopkins APL Technical Digest*, Vol. 29, No. 1, 2010,
- Park, S., Deyst, J., and How, J. P., “A New Nonlinear Guidance Logic for Trajectory Tracking,” in

## 8.4 Obstacle Avoidance Guidance Systems

**gradient vector field (GVF)** using **potential field (PF)** and **virtual force field (VFF)** guidance

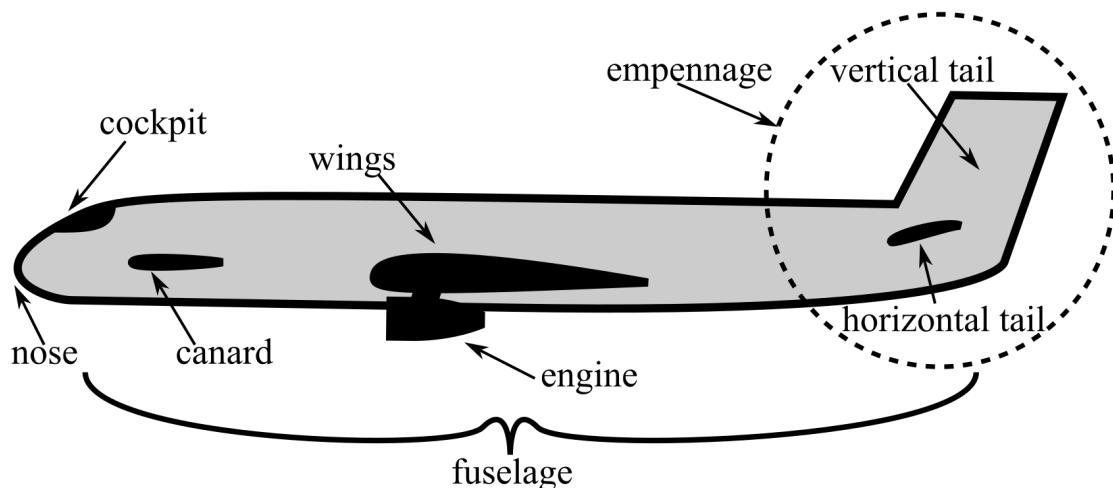
---

# Airplane Dynamics and Control Systems

## 9.1 Introduction to Fixed-Wing Vehicles

### Fixed-Wing Vehicle Anatomy

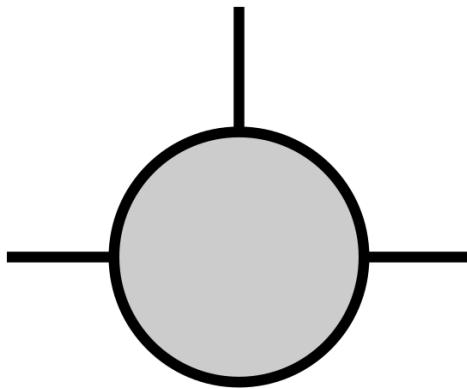
The basic components of a conventional **airplane**, i.e. a powered fixed-wing aircraft, are shown in the following diagram



The **nose** is the front of the airplane and houses the **cockpit** where the pilot(s) and/or other operators are located. The **fuselage** is the tubular structure of the airplane that houses the payload, fuel/batteries, and **avionics**, i.e. the aviation electronics. The **wings** extend horizontally out of the fuselage and generate the majority of the lift force to overcome the vehicle's weight and fly. Some airplanes have two wings, either as **tandem wings** which are located as front-and-back wings or as stacked wings called a **biplane**. The

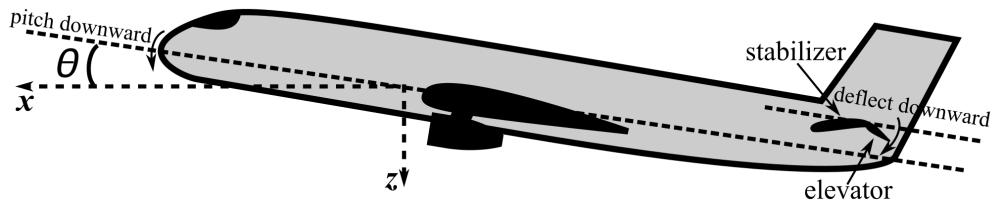
**engines** produce the thrust forces, with a possible **propeller** creating a swirling slipstream using rotating airfoil-section blades, which are controlled by the pilot. These can be mounted on the wings, empennage, or nose.

The **empennage** is the rear section of an airplane and creates aerodynamic effects to stabilize and steer the airplane in the direction of flight. This airplane shows the **conventional tail**.



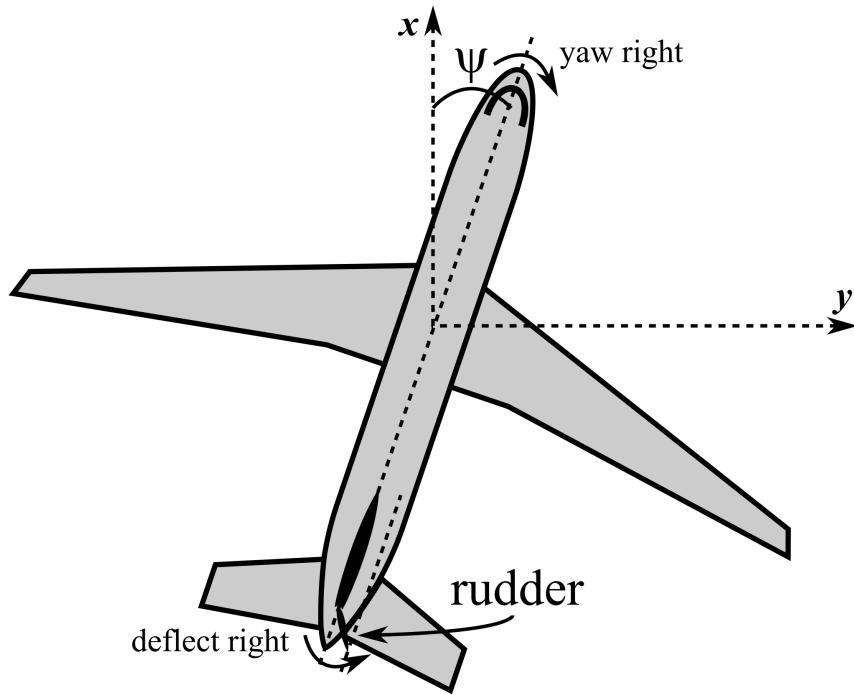
The **vertical tail** is composed of a static **vertical stabilizer** and a dynamic **rudder** which produce a horizontal left or right side force to maintain vertical stability and steer the airplane. The **horizontal tail** is composed of a static **horizontal stabilizer** and a dynamic **elevator** which produce a vertical up or down lift force to maintain horizontal stability and steer the airplane. Some airplanes may be **tailless** which do not have any other horizontal lifting surface besides its main wing, i.e. no horizontal tail, canard or tandem wing. They may or may not have a vertical tail. These airplanes use compound ailerons/elevator control surfaces called **elevons** or **tailerons**. Furthermore, a **flying wing** is a tailless airplane that also does not have a distinct fuselage. Some airplanes allow the entire horizontal tail to move which forms a compound horizontal stabilizer/elevator called a **stabilator**. Some airplanes may use a **canard**, i.e. a forward horizontal stabilizer and elevator, either in addition or instead of a rear horizontal tail.

Fixed-wing aircraft generally have three major control surfaces that allow the pilot or autopilot to affect the aerodynamic forces on the aircraft thereby altering its motion in a controlled manner. The **elevator** is mounted on trailing edge of horizontal tail and is used as the primary pitch angle,  $\theta$ , control input as shown in the following graphic where a deflection angle,  $\delta_e$ , upwards pushes the airplane's nose up.

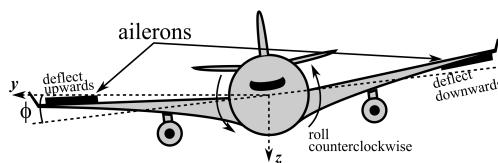


The **rudder** is mounted on trailing edge of the vertical tail and is used as the primary yaw angle,  $\psi$ ,

control input as shown in the following graphic where a deflection angle,  $\delta_r$ , right pushes airplane's nose right.

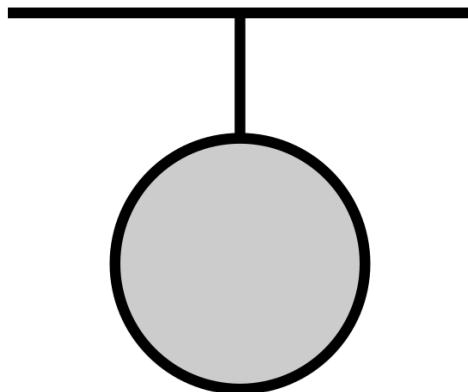


The **ailerons** are a differential pair of surfaces, i.e. if one deflects up, the other deflects down, mounted on trailing edge of wings near the tips and are used as the primary roll angle,  $\phi$ , control input as shown in the following graphic where the airplane rolls to the side that deflects up at some deflection angle,  $\delta_a$ .

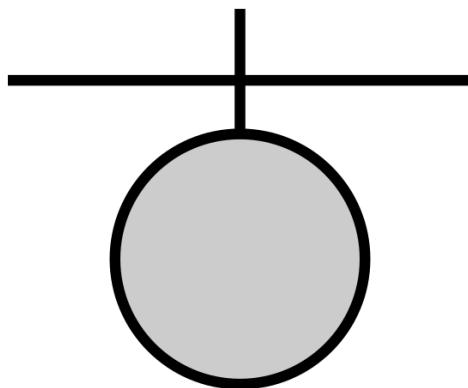


Depending on the tail, some airplanes may have multiple of these control surfaces, e.g. H-tails have two rudders.

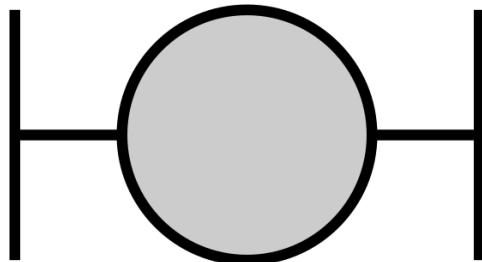
There are also a wide variety of empennage configurations, e.g. a **T-tail**



a **cruciform tail**

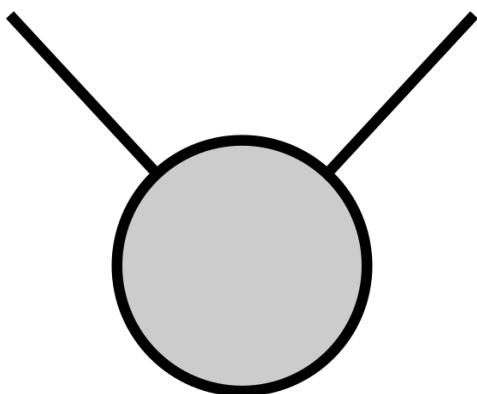


or an **H-tail**, also known as a **twin tail**.

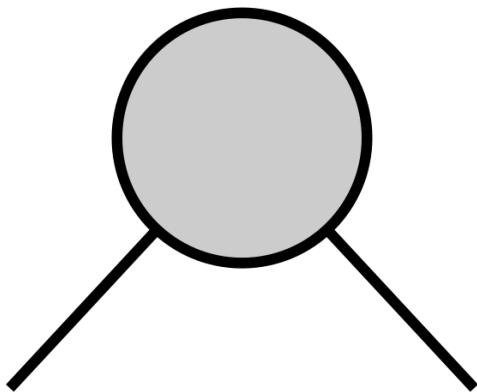


These twin tails may be mounted on the fuselage, two booms extending from the wing, or the wings themselves.

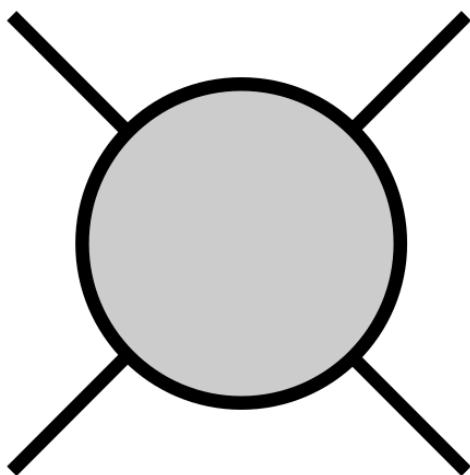
In addition, one may also use a compound horizontal/vertical tail empennage, e.g. a **V-tail**



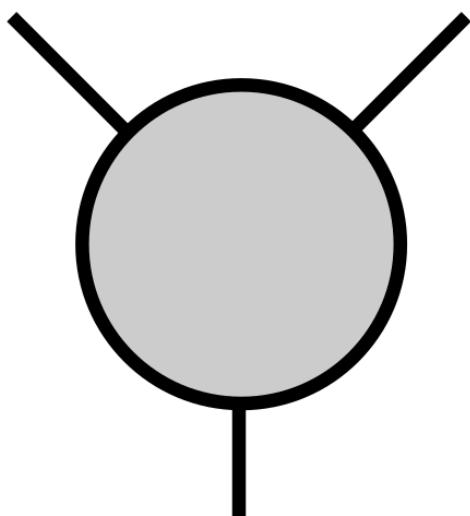
**inverted V-tail**



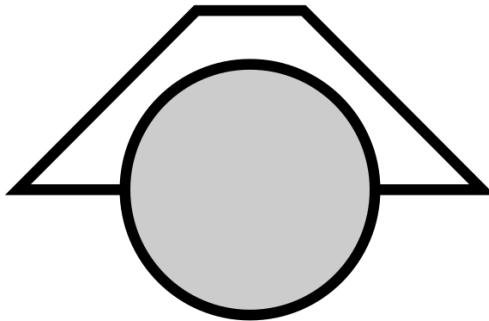
**X-tail**



**Y-tail**



or an **A-tail**

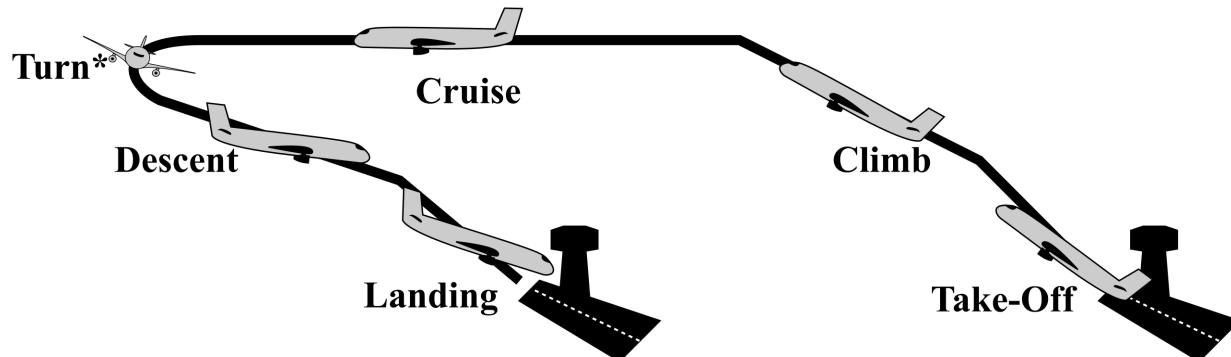


which are typically mounted on twin booms. These empennage configurations notably use compound rudder and elevator control surfaces called a **ruddervator**.

Many airplanes also have secondary control surfaces that can be controlled to change the airplane aerodynamics and thus alter its motion. These are typically grouped into two different categories. **Spoilers** are hinged flat plates attached to wing facing oncoming airflow. These “spoil” the lift on the wing when rotated up from the surface of the wing and thus are typically used as speed brakes, but sometimes can be used for roll control with or without ailerons. **Slats** and **flaps** are mechanized leading and trailing edges of wings, respectively, that can be extended/retracted from the nominal wing shape, thereby changing the airfoil cross-sectional shape and effectively increasing/decreasing the lift potential of the wing with higher/lower induced drag. As such, these are primarily extended during takeoff and landing and are not dynamically active control surfaces, i.e. they are either extended or retracted for different flight conditions and tasks.

### Phases of Flight for Fixed-Wing Vehicle

The five phases of aircraft flight can generally be described as takeoff, ascent/climb, cruise, descent/approach, landing as shown in the following graphic.



where each phase could include maneuvers such as coordinated turns. The aircraft will experience

different flight conditions and may be configured differently at each phase, e.g. for takeoff and landing of airplanes, flaps may be deployed and the landing gear will be extended down, which will directly impact the vehicle dynamics.

A **flight profile** is a graphical timeline of the operational characteristics, configurations, and velocities of an aerospace vehicle along its flight path in a specific phase of flight, e.g., one may have profiles for takeoff and climb, approach/descent, and landing. Flight profiles can also depict maneuvers like steep turns.

Notably, **cruise missiles** are fixed-wing aircraft that achieve flight through aerodynamic lift for the majority of their flight plan and have flight phases of **boost**, mid-course or cruise, and **terminal**.

## 9.2 Point-Mass Dynamics for Airplanes

To achieve and control atmospheric flight, fixed-wing aircraft generate aerodynamic forces from several different lifting surfaces for which one must define the aerodynamic forces for the entire aircraft. Similar to finite-wing theory (see appendix A.1), one can model a fixed-wing aircraft's total aerodynamic force vector,  $\vec{F}_a$ , in the wind frame as

$$\vec{F}_{a,W} = \begin{bmatrix} -D \\ S \\ -L \end{bmatrix} \quad (9.1)$$

where  $D$  is the **drag force** for the entire fixed-wing aircraft,  $S$  is the **side force** for the entire fixed-wing aircraft, and  $L$  is the **lift force** for the entire fixed-wing aircraft.

Here, one typically models each of these by their vehicle coefficients, i.e.

$$L = Q_\infty S_w C_L \quad (9.2)$$

$$S = Q_\infty S_w C_S \quad (9.3)$$

$$D = Q_\infty S_w C_D \quad (9.4)$$

where, one can model the drag coefficient as the sum of the parasitic drag, induced drag, and potentially wave drag, i.e.,

$$C_D = C_{D,p} + C_{D,i} + C_{D,w} = C_{D,0} + k C_L^2 \quad (9.5)$$

For subsonic flight,  $C_{D,w} = 0$ ,  $C_{D,0} = C_{D,p}$ , and

$$k = k_w = \frac{1}{\pi e_{eff} AR_w} \quad (9.6)$$

where  $e_{eff}$  is the **effective Oswald's efficiency** which can depend on the Mach number and  $C_L$  near and above the drag-divergence Mach number. For supersonic flight,

$$C_{D,0} = C_{D,p} + \frac{49(t/\bar{c})^2}{\sqrt{\mathcal{M}^2 - 1}} \quad (9.7)$$

$$k = k_i + k_w = \frac{1}{\pi e_{eff} AR_w} + \frac{AR_w(\mathcal{M}^2 - 1)}{4AR_w\sqrt{\mathcal{M}^2 - 1} - 2} \cos \Lambda_{LE} \quad (9.8)$$

while in the transonic region past  $\mathcal{M} = 1$ ,  $k$  transitions slowly from  $k_i$  to  $k_i + k_w$ .

Notably, these wind frame aerodynamic forces can be rotated to the body-fixed frame by

$$\vec{F}_{a,B} = C_{B \leftarrow W} \begin{bmatrix} -D \\ S \\ -L \end{bmatrix} \quad (9.9)$$

$$\vec{F}_{a,B} = \begin{bmatrix} \cos \alpha \cos \beta & -\cos \alpha \sin \beta & -\sin \alpha \\ \sin \beta & \cos \beta & 0 \\ \sin \alpha \cos \beta & -\sin \alpha \sin \beta & \cos \alpha \end{bmatrix} \begin{bmatrix} -D \\ S \\ -L \end{bmatrix} \quad (9.10)$$

or

$$\vec{F}_{a,B} = \begin{bmatrix} -D \cos \alpha \cos \beta - S \cos \alpha \sin \beta + L \sin \alpha \\ S \cos \beta - D \sin \beta \\ -D \sin \alpha \cos \beta - S \sin \alpha \sin \beta - L \cos \alpha \end{bmatrix} \quad (9.11)$$

In addition,  $S$  here without a subscript should not be confused with the surface area of a lifting surface which always has a specifying subscript with it in this textbook.

For airplanes, one typically models the propulsive force as

$$\vec{F}_{p,B} = \begin{bmatrix} T \cos \theta_T \\ 0 \\ T \sin \theta_T \end{bmatrix} \quad (9.12)$$

where  $T$  is the **thrust force** and  $\theta_T$  is a potential offset angle with respect to the  $x_B$ -axis of the body-fixed frame. Moreover, a propulsive moment may also be present nominally about the  $y_B$ -axis as

$$\vec{M}_{p,B} = \begin{bmatrix} 0 \\ T(z_T \cos \theta_T - x_T \sin \theta_T) \\ 0 \end{bmatrix} \quad (9.13)$$

where  $(x_T, z_T)$  denotes the location of the thrust force in the  $x_B - z_B$  plane. It should be noted that often  $\theta_T = 0^\circ$  or approximately and  $T$  is generally a function of the airspeed, altitude, and throttle setting.

For this point-mass model, one can use the wind frame for the translation equation, i.e.

$$\sum \vec{F}_W = \frac{d}{dt}(m \vec{v}_W) = m(\dot{\vec{v}}_W + \vec{\omega}_{W/N} \times \vec{v}_W) \quad (9.14)$$

Next, defining the angular velocity of the wind frame relative to the navigation as

$$\vec{\omega}_{W/N} = [p_W \quad q_W \quad r_W]^T \quad (9.15)$$

and recalling, by definition

$$\vec{v}_W = \begin{bmatrix} v_\infty \\ 0 \\ 0 \end{bmatrix} \quad (9.16)$$

one has

$$\vec{F}_{a,W} + \vec{F}_{p,W} + \vec{F}_{g,W} = \begin{bmatrix} m \dot{v}_\infty \\ 0 \\ 0 \end{bmatrix} + m \begin{bmatrix} p_W \\ q_W \\ r_W \end{bmatrix} \times \begin{bmatrix} v_\infty \\ 0 \\ 0 \end{bmatrix} \quad (9.17)$$

Furthermore, one can model the gravitational force as

$$\vec{F}_{g,W} = C_{W \leftarrow N}(\mu, \gamma, \sigma) \begin{bmatrix} 0 \\ 0 \\ mg \end{bmatrix} \quad (9.18)$$

the propulsive force assuming that  $\theta_T = 0$  as

$$\vec{F}_{p,W} = C_{W \leftarrow B}(\alpha, \beta) \begin{bmatrix} T \\ 0 \\ 0 \end{bmatrix} \quad (9.19)$$

and the aerodynamic force as

$$\vec{F}_{a,W} = \begin{bmatrix} -D \\ S \\ -L \end{bmatrix} \quad (9.20)$$

Thus, one has

$$\begin{bmatrix} -D \\ S \\ -L \end{bmatrix} + \begin{bmatrix} T \cos \alpha \cos \beta \\ -T \cos \alpha \sin \beta \\ -T \sin \alpha \end{bmatrix} + \begin{bmatrix} -mg \sin \gamma \\ mg \sin \mu \cos \gamma \\ mg \cos \mu \cos \gamma \end{bmatrix} = \begin{bmatrix} \dot{v}_\infty \\ mv_\infty r_W \\ -mv_\infty q_W \end{bmatrix} \quad (9.21)$$

Then, substituting for the wind frame angular velocity components with the navigation-to-wind frame Euler angles, i.e.

$$\begin{bmatrix} p_W \\ q_W \\ r_W \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\sin \gamma \\ 0 & \cos \mu & \sin \mu \cos \gamma \\ 0 & -\sin \mu & \cos \mu \cos \gamma \end{bmatrix} \begin{bmatrix} \dot{\mu} \\ \dot{\gamma} \\ \dot{\sigma} \end{bmatrix} \quad (9.22)$$

Rearranging, one has

$$\begin{bmatrix} -D + T \cos \alpha \cos \beta - mg \sin \gamma \\ S - T \cos \alpha \sin \beta + mg \sin \mu \cos \gamma \\ L + T \sin \alpha - mg \cos \mu \cos \gamma \end{bmatrix} = \begin{bmatrix} m\dot{v}_\infty \\ mv_\infty (\dot{\sigma} \cos \mu \cos \gamma - \dot{\gamma} \sin \mu) \\ v_\infty (\dot{\gamma} \cos \mu + \dot{\sigma} \sin \mu \cos \gamma) \end{bmatrix} \quad (9.23)$$

where one can add control inputs to this model through the thrust  $T$  and the lift  $L$ . Lastly, one can also include the navigation frame velocity components as

$$\begin{bmatrix} \dot{x}_N \\ \dot{y}_N \\ \dot{h} \end{bmatrix} = \begin{bmatrix} v_\infty \cos \gamma \cos \sigma \\ v_\infty \cos \gamma \sin \sigma \\ v_\infty \sin \gamma \end{bmatrix} \quad (9.24)$$

where  $-\dot{h} = \dot{z}_N$  is the altitude rate instead of the down rate.

With these dynamics, one can analyze the **point-mass equilibrium flight conditions**, also known as the **point-mass steady-flight conditions** which dictates  $\dot{v}_\infty = \dot{\gamma} = 0$ . Then, the forces and trim states of the airplane can be related as

$$\begin{bmatrix} -\bar{D} + \bar{T} \cos \bar{\alpha} \cos \bar{\beta} - mg \sin \bar{\gamma} \\ \bar{S} - \bar{T} \cos \bar{\alpha} \sin \bar{\beta} + mg \sin \bar{\mu} \cos \bar{\gamma} \\ -\bar{T} \sin \bar{\alpha} - \bar{L} + mg \cos \bar{\mu} \cos \bar{\gamma} \end{bmatrix} = \begin{bmatrix} 0 \\ m\bar{v}_\infty \dot{\sigma} \cos \bar{\mu} \cos \bar{\gamma} \\ m\bar{v}_\infty \dot{\sigma} \sin \bar{\mu} \cos \bar{\gamma} \end{bmatrix} \quad (9.25)$$

Of particular note is the **coordinated steady-flight condition** when  $\bar{S} = 0$  and  $\bar{\beta} = 0^\circ$ . Then, by substitution, one has

$$\begin{bmatrix} -\bar{D} + \bar{T} \cos \bar{\alpha} - mg \sin \bar{\gamma} \\ mg \sin \bar{\mu} \cos \bar{\gamma} \\ -\bar{T} \sin \bar{\alpha} - \bar{L} + mg \cos \bar{\mu} \cos \bar{\gamma} \end{bmatrix} = \begin{bmatrix} 0 \\ m\bar{v}_\infty \dot{\sigma} \cos \bar{\mu} \cos \bar{\gamma} \\ m\bar{v}_\infty \dot{\sigma} \sin \bar{\mu} \cos \bar{\gamma} \end{bmatrix} \quad (9.26)$$

Finally, defining the **instantaneous radius of curvature** in the navigation frame,  $R_c$ , as

$$R_c = \frac{\bar{v}_\infty \cos \bar{\gamma}}{\dot{\sigma}} \quad (9.27)$$

one has the alternative **point-mass steady-flight equations**

$$\begin{bmatrix} \bar{T} \cos \bar{\alpha} - \bar{D} - mg \sin \bar{\gamma} \\ mg \sin \bar{\mu} \cos \bar{\gamma} \\ -\bar{T} \sin \bar{\alpha} - \bar{L} + mg \cos \bar{\mu} \cos \bar{\gamma} \end{bmatrix} = \begin{bmatrix} 0 \\ m \frac{(\bar{v}_\infty \cos \bar{\gamma})^2}{R_c} \cos \bar{\mu} \\ m \frac{(\bar{v}_\infty \cos \bar{\gamma})^2}{R_c} \sin \bar{\mu} \end{bmatrix} \quad (9.28)$$

It should be noted that from the second equation, i.e.

$$mg \sin \bar{\mu} \cos \bar{\gamma} = m \frac{(\bar{v}_\infty \cos \bar{\gamma})^2}{R} \cos \bar{\mu} \quad (9.29)$$

one can see that for any non-zero bank angle

$$\dot{\sigma} = \frac{g \tan \bar{\mu}}{\bar{v}_\infty} \quad (9.30)$$

and

$$R_c = \frac{\bar{v}_\infty^2 \cos \bar{\gamma}}{g |\tan \bar{\mu}|} \quad (9.31)$$

The most general maneuver described by the point-mass steady-flight equations for an airplane is a steady climbing or descending coordinated turn and are primarily controlled by altering the lift, thrust, and bank angle of the airplane. The trajectory the airplane flies during this maneuver is a helix about the  $z_N$ -axis and a circular projection on the  $x_N - y_N$  plane. Three special cases of this steady-flight maneuver are straight climbs/descents, level turns, and straight-and-level where **straight flight** occurs when  $\dot{\sigma} = \bar{\mu} = 0^\circ$  and **level flight** occurs when  $\bar{\gamma} = 0^\circ$ . As the lift and drag are primarily a function of the steady-state air density, airspeed, and angle of attack, the point-mass steady-flight equations can be considered as a balance of six conditions: the altitude (affects air density), bank angle, flight-path angle, angle of attack, airspeed, and thrust for a given airplane's aerodynamic and mass properties. However, once one considers the airplane as a rigid body, the additional moment equations must also be balanced to ensure that the airplane remains at the prescribed steady-flight conditions. As these additional moment equations can be altered by the control inputs to the ailerons, rudder, and elevator, one typically refers to this moment balance as **trimming the airplane**, which will be considered in the subsequent chapter.

## References

For more information, please refer to the following:

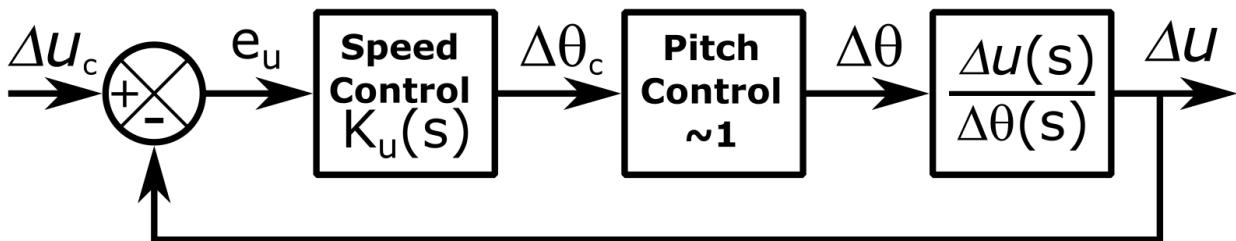
- Saarlas, M., “Appendix B: On the Drag Coefficient,” *Aircraft Performance*, 1st ed., Vol. 1, John Wiley & Sons, 2007, pp. 260-264
- Schmidt, D. K., “2.7 Point-Mass Performance Equations,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 76-80

## 9.3 Airplane Hold and Landing Guidance Systems

### Airplane Hold Guidance Systems

Though airplanes are inherently MIMO systems, due to the standard steady-flight conditions for the majority of airplanes, the longitudinal and lateral-directional feedback control systems are typically designed separately and operated in parallel using the different control inputs for controlling the inner-loops of the airplane. In particular, the elevator and throttle are used for the longitudinal feedback control system and the rudder and ailerons are used for the lateral-directional feedback control systems. The inner-loops in both cases are first for attitude control, i.e. pitch angle (or angle of attack or pitch rate) and roll angle, respectively, which are the fundamental attitude angles to control as they directly affect the magnitude and direction of the airplane’s lift vector, respectively. Closed around the inner-loop attitude controllers are the outer-loops guidance laws for both longitudinal and lateral which follow the planned reference trajectory for which the simplest are holds on speed, altitude, and heading. For long-distance flying, one typically prescribes the line as the shortest distance from one location to another, also known as a **great circle** line, as it follows the curvature of the Earth’s reference ellipsoid.

One option for the longitudinal outer-loop guidance system for an airplane is a **speed hold** which is typically employed during climbing flight under air traffic control. This guidance loop uses the speed control law,  $K_u(s)$ , the pitch inner-loop control system, and the transfer function from  $\Delta\theta \rightarrow \Delta u$  for the outer-loop plant as shown in the following block diagram.

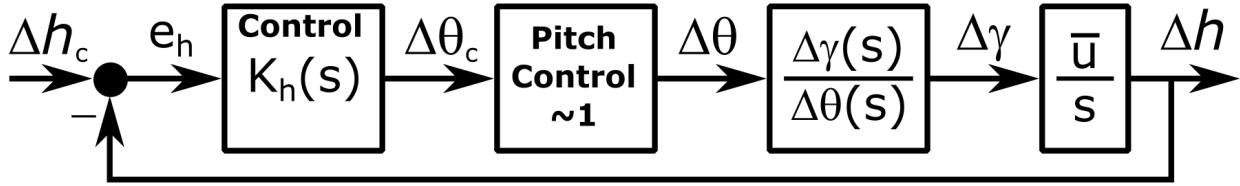


As the inner-loop of this speed hold control loop uses the elevator, one can form the outer-loop plant as

$$\frac{\Delta u(s)}{\Delta \theta(s)} = \frac{\text{num}\left(\frac{\Delta u(s)}{\Delta \delta_e(s)}\right)}{\text{num}\left(\frac{\Delta \theta(s)}{\Delta \delta_e(s)}\right)} \quad (9.32)$$

where  $\text{num}(G(s))$  represents the numerator of the transfer function  $G(s)$ . To form this model, one may use the transfer functions  $\frac{\Delta u(s)}{\Delta \delta_e(s)}$  and  $\frac{\Delta \theta(s)}{\Delta \delta_e(s)}$  for the vehicle alone as long as the crossover frequency separation for the inner- and outer-loops is maintained. Typically, one also includes the pitch inner-loop control system once it has been designed. It should be noted that  $K_u(s)$  or  $\frac{\Delta u(s)}{\Delta \theta(s)}$  may need to be negative as a positive change in pitch  $\Delta\theta$  may cause a reduction in flight velocity  $u$ .

Another option for the longitudinal outer-loop guidance system for an airplane is an **altitude hold** which is typically employed during a cruise flight condition at a specific cruise velocity specified by the throttle setting. This guidance loop uses the altitude control law,  $K_h(s)$ , the pitch inner-loop control system, and the transfer function from  $\Delta\theta \rightarrow \Delta\gamma$  for the outer-loop plant as shown in the following block diagram.



It should be noted that the transfer function from  $\Delta\theta \rightarrow \Delta\gamma$  can be calculated using the small angle approximation for the Euler angles as

$$\Delta\gamma(s) = \Delta\theta(s) - \Delta\alpha(s) \quad (9.33)$$

which provides

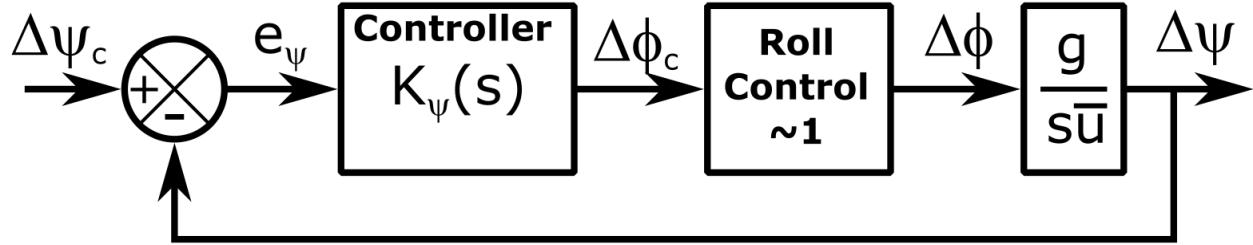
$$\frac{\Delta\gamma(s)}{\Delta\theta(s)} = 1 - \frac{\Delta\alpha(s)}{\Delta\theta(s)} \quad (9.34)$$

or

$$\frac{\Delta\gamma(s)}{\Delta\theta(s)} = 1 - \frac{\text{num}\left(\frac{\Delta\alpha(s)}{\Delta\delta_e(s)}\right)}{\text{num}\left(\frac{\Delta\theta(s)}{\Delta\delta_e(s)}\right)} \quad (9.35)$$

To form this model, one may again use the transfer functions  $\frac{\alpha(s)}{\delta_e(s)}$  and  $\frac{\theta(s)}{\delta_e(s)}$  for the vehicle alone as long as the crossover frequency separation for the inner- and outer-loops is maintained. Typically, one also includes the pitch inner-loop control system once it has been designed. It should also be noted that here it is required that an increase in pitch angle  $\theta$  must produce a steady-state flight path angle  $\gamma$ , i.e. the airplane must be “on the front side of the power curve.”

One option for the lateral-directional outer-loop guidance system for an airplane is a **heading hold** which is typically employed at a specific cruise velocity. This guidance loop uses the heading control law,  $K_\psi(s)$ , the roll inner-loop control system, and the transfer function from  $\Delta\phi \rightarrow \Delta\psi$  for the outer-loop plant as shown in the following block diagram.



It should be noted that this heading hold assumes that the heading and yaw are the same, i.e.  $\bar{\beta} = 0^\circ$  so  $\sigma = \psi$ , and that  $\mu \approx \phi$  which is true for the linearized lateral-directional dynamics. Furthermore, the transfer function from  $\Delta\phi \rightarrow \Delta\psi$  can be calculated using the steady flight relationship for the bank angle and the rate of change of heading/yaw, i.e.

$$\dot{\sigma} = \frac{g}{v_\infty} \tan \mu \quad (9.36)$$

which, for the small angle approximation  $\tan \mu = \Delta\mu$  and using a stability frame for linearization, i.e.  $v_\infty = \bar{u}$ , one has

$$\Delta\psi(s) = \frac{g}{s\bar{u}}\Delta\phi(s) \quad (9.37)$$

after angle substitutions. It should be noted that often the roll angle must be hard limited for particular airplane missions. Thus, often one includes a **limiter** for the commanded roll which limits the maximum and minimum values to some specified values. Such an inclusion in the feedback control system will introduce a nonlinearity in the system, but generally reduces the speed of response for fast maneuvers. Performances of such systems is typically assessed through the nonlinear simulations of the feedback control systems initially designed with linear models.

# Precision Approach Guidance Systems

In commercial aviation, guidance systems have been developed to assist an airplane during the approach and landing phases of an airplane's flight as these phases require the most accurate navigation information and precision guidance especially when visibility is impaired for pilots and an autopilot is engaged, also known as a **precision approach**. These systems are designed to supply this information using radio signals that are interpreted by specialized equipment onboard the airplane which supply a heading correction that the pilot or autopilot must make to continue on the planned descent trajectory to that airport's runway. Thus, these systems are maintained by individual airports. Thus, the reference trajectories generated for these guidance systems are constant, predefined trajectories based on the topography of the area, the layout of the airport's runways, and regulations concerning safety for airplane descent.

An important parameter in precision approaches is the **decision height (DH)** or **decision altitude (DA)**. Both of which are defined as the specified lowest height/altitude in the approach descent at which, if the required **runway visual reference (RVR)** to continue the approach is not visible to the pilot, the pilot must initiate a missed approach maneuver and reroute to try the approach again. A decision height is measured *above ground level (AGL)* while a decision altitude is measured above *mean sea level (MSL)*. The specific values for the DH/DA at a given airport are established with intention to allow a pilot sufficient time to safely

re-configure an airplane to climb and execute the missed approach procedures while avoiding terrain and obstacles. A DH/DA denotes the height/altitude in which a missed approach procedure must be started, it does not preclude the airplane from descending below the prescribed DH/DA.

For precision approach guidance, there are four categories which allow these decision heights to be lowered. To allow these precision approaches, two primary technologies serve commercial airplanes as both

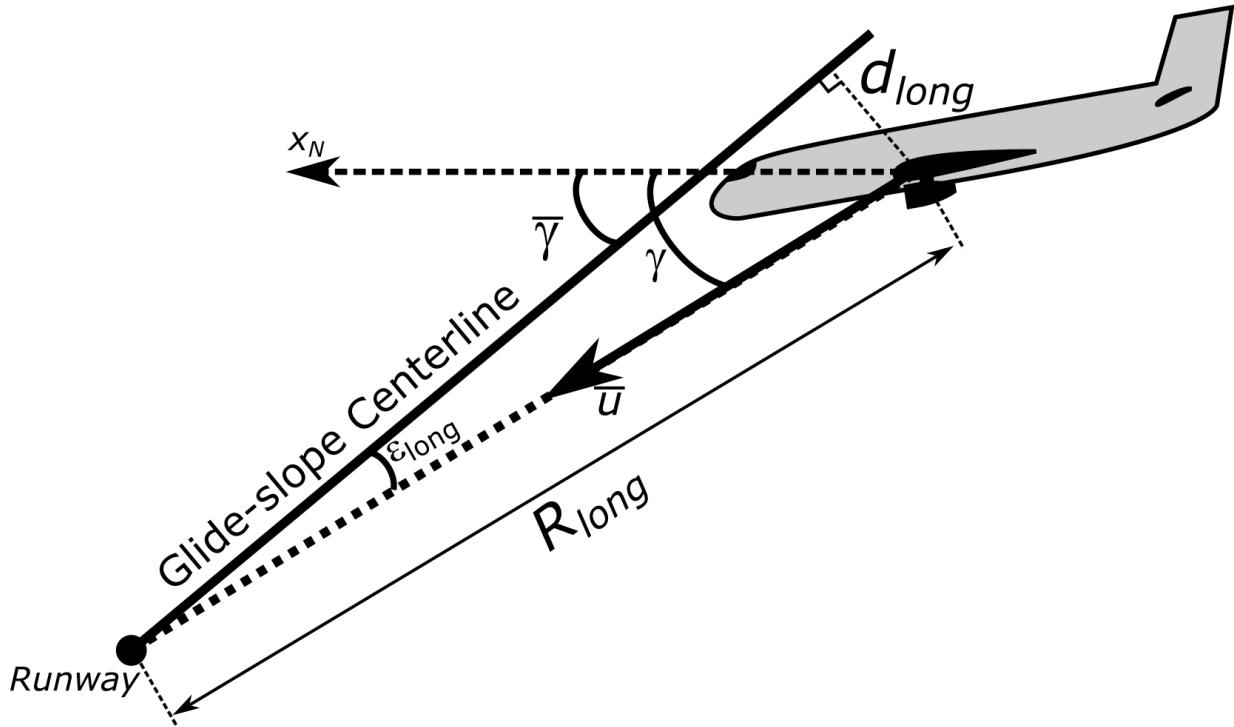
Category	DH	RVR
I	> 60 m	> 550 m
II	30-60 m	> 350 m
III A	< 60 m	> 200 m
III B	< 15 m	> 50 m
III C	none	none

navigational and guidance aids.

One of the earliest technologies for automatic guidance and is still in use today is the **Instrument Landing System (ILS)** which used radionavigation to provide airplane with horizontal and vertical guidance information during approach and landing. At certain fixed points, it also provides the distance to the reference point of landing. An ILS uses two signals: a **localizer** and a **glide-slope** which indicate the correction that the pilot or autopilot must make. These are displayed to pilots as a point plotted on 2 coordinate axes whose origin denotes no error, i.e., if the airplane is to the left of the reference trajectory, the point is to the right of the origin and if the airplane is above the reference trajectory, then the point is below the origin. The glide-slope path for an ILS is typically defined between 2-6° below the horizontal. Occasionally a modified ILS called an **Instrument Guidance System (IGS)**, also known as a **Localizer-type Directional Aid (LDA)** in the United States, must be used for non-straight approaches into airports with certain topographical features which prevent normal operation. As part of this system, there are also typically marker beacons or distance measuring equipment. **Marker beacons** which provide distance to the runway information at setpoints along the approach. **Distance measuring equipment (DME)** is also a common system available at airports that provides pilots or autopilots with a continuous distance to runway measurement.

Due to the rapid development of Global Navigation Satellite Systems (GNSS), in particular the Global Positioning System (GPS) run by the United States Department of Defense, the most recent advances in automatic guidance technologies for aircraft include the following technologies which enhance the basic capabilities of GNSS by reducing the errors found in the GNSS signals thereby improving the accuracy of the navigation solution. The need and details for these technologies will be discussed in later parts of this textbook. The Ground-Based Augmentation System (GBAS), which for GPS is also known as Local Area Augmentation System (LAAS), also augments the GNSS measurements in order to provide enhanced levels of service to support automatic guidance information during all phases of approach, landing, departure, and surface operations within radio distance. GBAS is anticipated to replace ILS in the future.

For precision approach guidance, an airplane is to follow a “straight-line” path in the longitudinal plane for the glide-slope as shown



where  $d_{long}$  is the longitudinal position deviation,  $\epsilon_{long}$  is the longitudinal angular deviation,  $R_{long}$  is the longitudinal range to intercept,  $\bar{\gamma}$  is the reference glide-slope angle.

Given this figure, the longitudinal deviations can be related by

$$\sin \epsilon_{long} = \frac{d_{long}}{R_{long}} \quad (9.38)$$

which for small angular deviations is

$$\epsilon_{long} = \frac{d_{long}}{R_{long}} \quad (9.39)$$

Furthermore, as the longitudinal angular deviation is related to the flight path angle by

$$\epsilon_{long} = \bar{\gamma} - \gamma = -\Delta\gamma \quad (9.40)$$

Then, by rearrangement and substitution, one has

$$d_{long} = -R_{long}\Delta\gamma \quad (9.41)$$

then taking the derivative, one has

$$\dot{d}_{long} = -\dot{R}_{long}\Delta\gamma \quad (9.42)$$

and the instantaneous range rate is

$$\dot{R}_{long} = -\bar{u} \quad (9.43)$$

one has for the measured angular deviation

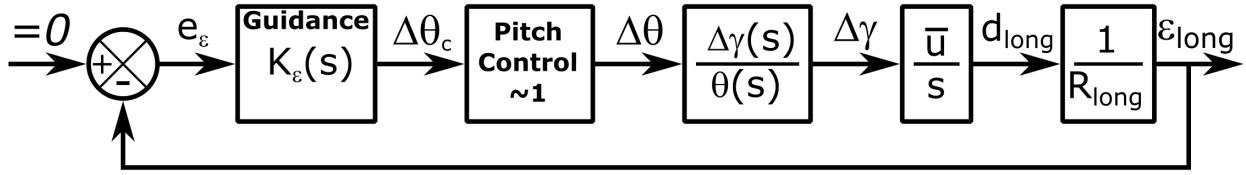
$$\dot{\epsilon}_{long} \approx \frac{\bar{u}}{R_{lat}} \Delta\gamma \quad (9.44)$$

which implies the simple transfer function of

$$\frac{\epsilon_{long}(s)}{\Delta\gamma(s)} \approx \frac{\bar{u}}{Rs} \quad (9.45)$$

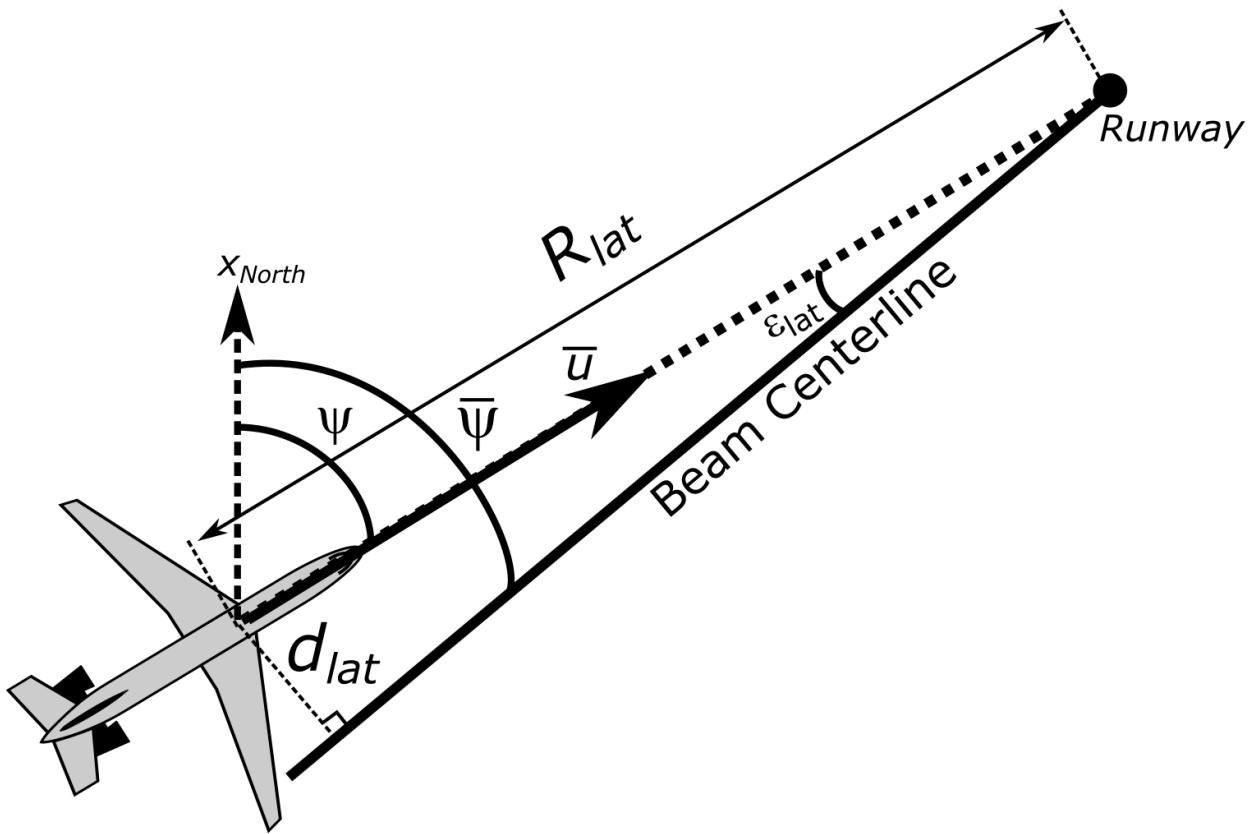
where it should be noted that this also assumes that the body-fixed reference frame is the stability frame, i.e.  $\bar{\alpha} = 0$ , thus  $\bar{\gamma} = \bar{\theta}$ .

Thus, one can form a block diagram of the **glide-slope guidance system** as



which is very similar to the altitude hold feedback loop with an additional outer-loop guidance for reducing the angular deviation from a reference flight-path to zero which scales inversely with the distance  $R_{long}$  from the runway.

For precision approach guidance, the airplane follows a “straight-line” path in lateral-directional plane for the localizer as shown



where  $d_{lat}$  is the lateral-directional position deviation,  $\epsilon_{lat}$  is the lateral-directional angular deviation,  $R_{lat}$  is the lateral-directional range to intercept,  $\bar{\psi}$  is the reference beam centerline for the localizer.

Given this figure, the lateral-directional deviations can be related by

$$\sin \epsilon_{lat} = \frac{d_{lat}}{R_{lat}} \quad (9.46)$$

which for small angular deviations is

$$\epsilon_{lat} = \frac{d_{lat}}{R_{lat}} \quad (9.47)$$

Furthermore, as the lateral angular deviation is related to the heading angle by

$$\epsilon_{lat} = \psi - \bar{\psi} \quad (9.48)$$

Then, by rearrangement and substitution, one has

$$d_{lat} = -R_{lat} (\psi - \bar{\psi}) \quad (9.49)$$

then taking the derivative, one has

$$\dot{d}_{lat} = -\dot{R}_{lat} (\psi - \bar{\psi}) \quad (9.50)$$

and the instantaneous range rate is

$$\dot{R}_{lat} = -\bar{u} \quad (9.51)$$

one has for the measured angular deviation

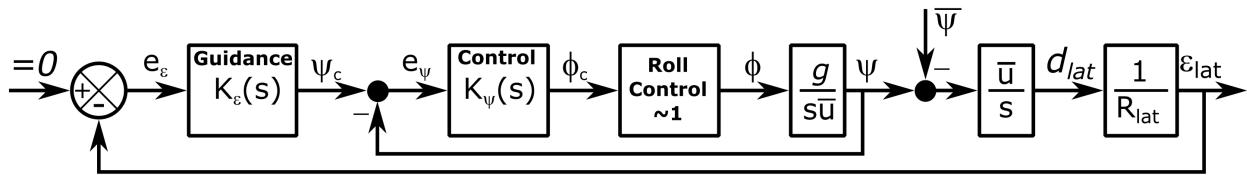
$$\dot{\epsilon}_{lat} \approx \frac{\bar{u}}{R_{lat}} (\psi - \bar{\psi}) \quad (9.52)$$

or in the Laplace domain

$$\frac{\epsilon_{lat}(s)}{\Psi(s) - \bar{\Psi}(s)} \approx \frac{\bar{u}}{Rs} \quad (9.53)$$

where it should be noted that this also assumes that the body-fixed reference frame is the stability frame, i.e.  $\vec{v} = u$  and that coordinated flight is occurring, i.e.  $\bar{\beta} = 0^\circ$ .

Thus, one can form a block diagram of the **localizer guidance system** as



which essentially uses a heading hold feedback loop with an additional outer-loop guidance for reducing the angular deviation from a reference heading to zero which scales inversely with the distance  $R_{lat}$  from the runway.

## Flare Guidance System

### References

For more information, please refer to the following

- Nelson, R. C., “8.4 Displacement Autopilot,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 292-312
- Nelson, R. C., “8.6 Instrument Landing,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 314-318
- Schmidt, D. K., “12.5 Response Holds,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 708-737
- Schmidt, D. K., “12.6 Path Guidance - ILS Couplers and VOR Homing,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 753-771
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “4.6 Autopilots,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 322-344

## 9.4 Airplane Trim Analysis

By convention, the aerodynamic and propulsive forces along the  $x_B$ -,  $y_B$ -, and  $z_B$ -axes are combined and normalized by the mass of the aircraft,  $m$ , and denoted by  $X$ ,  $Y$ , and  $Z$ , respectively. This infers the following force equation for aircraft as

$$\vec{F}_{p,B} + \vec{F}_{a,B} = \begin{bmatrix} mX \\ mY \\ mZ \end{bmatrix} \quad (9.54)$$

or

$$\frac{1}{m} (\vec{F}_{p,B} + \vec{F}_{a,B}) = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (9.55)$$

As an example, one could model this for airplanes as

$$\frac{1}{m} \left( \begin{bmatrix} T \cos \theta_T \\ 0 \\ T \sin \theta_T \end{bmatrix} + \begin{bmatrix} -D \cos \alpha \cos \beta - S \cos \alpha \sin \beta + L \sin \alpha \\ S \cos \beta - D \sin \beta \\ -D \sin \alpha \cos \beta - S \sin \alpha \sin \beta - L \cos \alpha \end{bmatrix} \right) = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (9.56)$$

Furthermore, the aerodynamic and propulsive moments about the  $x_B$ -,  $y_B$ -, and  $z_B$ -axes are conventionally normalized by the moments of inertia,  $I_{xx}$ ,  $I_{yy}$ , and  $I_{zz}$ , about that axis, i.e.  $x_B$ ,  $y_B$ , and  $z_B$ , and denoted by  $L$ ,  $M$ , and  $N$ , respectively. This infers the following moment equation for aircraft as

$$\vec{M}_{p,B} + \vec{M}_{a,B} = \begin{bmatrix} I_{xx}L_{roll} \\ I_{yy}M \\ I_{zz}N \end{bmatrix} \quad (9.57)$$

$$\vec{M}_{p,B} + \vec{M}_{a,B} = \begin{bmatrix} I_{xx} & 0 & 0 \\ 0 & I_{yy} & 0 \\ 0 & 0 & I_{zz} \end{bmatrix} \begin{bmatrix} L_{roll} \\ M \\ N \end{bmatrix} \quad (9.58)$$

As an example, one could model this for airplanes as

$$\begin{bmatrix} I_{xx}^{-1} & 0 & 0 \\ 0 & I_{yy}^{-1} & 0 \\ 0 & 0 & I_{zz}^{-1} \end{bmatrix} \left( \begin{bmatrix} 0 \\ T(z_T \cos \theta_T - x_T \sin \theta_T) \\ 0 \end{bmatrix} + \vec{M}_{a,B} \right) = \begin{bmatrix} L_{roll} \\ M \\ N \end{bmatrix} \quad (9.59)$$

Furthermore, the vast majority of aircraft are designed as symmetric in the  $x_B - z_B$  plane, i.e. the left side of the aircraft is reflected to the right side, thus the inertia matrix can be simplified to

$$I_G = \begin{bmatrix} I_{xx} & 0 & -I_{xz} \\ 0 & I_{yy} & 0 \\ -I_{xz} & 0 & I_{zz} \end{bmatrix} \quad (9.60)$$

where often  $I_{xz}$  is also neglected due to its relatively small magnitude.

For aircraft, the convention is to assume the 3–2–1 Euler angle representation for relating the navigation frame to the body-fixed frame for modeling the gravitational force. Assuming a flat- or spherical-Earth approximation, one can mode the gravity as

$$\vec{F}_{g,B} = C_{B \leftarrow N} \vec{F}_{g,N} = \begin{bmatrix} \cos \theta \cos \psi & \cos \theta \sin \psi & -\sin \theta \\ \sin \phi \sin \theta \cos \psi - \cos \phi \sin \psi & \sin \phi \sin \theta \sin \psi + \cos \phi \cos \psi & \sin \phi \cos \theta \\ \cos \phi \sin \theta \cos \psi + \sin \phi \sin \psi & \cos \phi \sin \theta \sin \psi - \sin \phi \cos \psi & \cos \phi \cos \theta \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ mg \end{bmatrix} \quad (9.61)$$

$$C_{B \leftarrow N} \vec{F}_{g,N} = mg \begin{bmatrix} -\sin \theta \\ \sin \phi \cos \theta \\ \cos \phi \cos \theta \end{bmatrix} \quad (9.62)$$

which results in

$$\begin{aligned} \vec{F}_{a,B} + \vec{F}_{p,B} + mg \begin{bmatrix} -\sin \theta \\ \sin \phi \cos \theta \\ \cos \phi \cos \theta \end{bmatrix} &= m \begin{bmatrix} \dot{u} + qw - rv \\ \dot{v} + ru - pw \\ \dot{w} + pv - qu \end{bmatrix} \\ \vec{M}_{a,B} + \vec{M}_{p,B} &= \begin{bmatrix} I_{xx}\dot{p} + (I_{zz} - I_{yy})qr - I_{xy}(\dot{q} - pr) - I_{xz}(\dot{r} + pq) + I_{yz}(r^2 - q^2) \\ I_{yy}\dot{q} + (I_{xx} - I_{zz})pr - I_{xy}(\dot{p} + qr) + I_{xz}(p^2 - r^2) - I_{yz}(\dot{r} - pq) \\ I_{zz}\dot{r} + (I_{yy} - I_{xx})pq + I_{xy}(q^2 - p^2) - I_{xz}(\dot{p} - qr) - I_{yz}(\dot{q} + pr) \end{bmatrix} \end{aligned} \quad (9.63)$$

which notably only requires that eight states be known to calculate the derivatives at any instant, namely  $u$ ,  $v$ ,  $w$ ,  $p$ ,  $q$ ,  $r$ ,  $\phi$ , and  $\theta$ , while  $\psi$  is simply a derived parameter from  $p$ ,  $q$ , and  $r$ .

Finally, with these substitutions and zeroing the solar forces and moments, one obtains the **rigid aircraft equations of motion**

$$\begin{aligned} \begin{bmatrix} X - g \sin \theta \\ Y + g \sin \phi \cos \theta \\ Z - g \cos \phi \cos \theta \end{bmatrix} &= \begin{bmatrix} \dot{u} + qw - rv \\ \dot{v} + ru - pw \\ \dot{w} + pv - qu \end{bmatrix} \\ \begin{bmatrix} L \\ M \\ N \end{bmatrix} &= \begin{bmatrix} \dot{p} + \frac{I_{zz} - I_{yy}}{I_{xx}} qr - \frac{I_{xz}}{I_{xx}} (\dot{r} + pq) \\ \dot{q} + \frac{I_{xx} - I_{zz}}{I_{yy}} pr + \frac{I_{xz}}{I_{yy}} (p^2 - r^2) \\ \dot{r} + \frac{I_{yy} - I_{xx}}{I_{zz}} pq - \frac{I_{xz}}{I_{zz}} (\dot{p} - qr) \end{bmatrix} \end{aligned} \quad (9.64)$$

This introductory chapter of the textbook will use the **no-wind approximation** to introduce rigid aircraft dynamics. Although with wind speeds significantly close to the nominal aircraft airspeed and/or long distance flight analysis, the no-wind approximation may not be suitable. The additional effects of both steady and unsteady wind on the aircraft dynamics is presented in the next chapter of this textbook on advanced rigid aerospace vehicle dynamics. As an alternative to this form of the EOMs with the no-wind approximation, then  $\vec{v}_g = \vec{v}_\infty$ , and one may also substitute for the  $y_B$ -axis and  $z_B$ -axis velocity terms

$$v = u \tan \beta \quad (9.65)$$

$$w = u \sin \alpha \quad (9.66)$$

as well as their derivatives

$$\dot{v} = \dot{u} \tan \beta + \dot{\beta} u \sec^2 \beta \quad (9.67)$$

$$\dot{w} = \dot{u} \sin \alpha + \dot{\alpha} u \cos \alpha \quad (9.68)$$

in order to obtain the **alternative rigid aircraft EOMs**

$$\begin{aligned} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \vec{F}_G &= \begin{bmatrix} \dot{u} + qu \sin \alpha - ru \tan \beta \\ \dot{u} \tan \beta + \dot{\beta} u \sec^2 \beta + ru - pu \sin \alpha \\ \dot{u} \sin \alpha + \dot{\alpha} u \cos \alpha + pu \tan \beta - qu \end{bmatrix} \\ \begin{bmatrix} L \\ M \\ N \end{bmatrix} &= \begin{bmatrix} \dot{p} + \frac{I_{zz}-I_{yy}}{I_{xx}} qr - \frac{I_{xz}}{I_{xx}} (\dot{r} + pq) \\ \dot{q} + \frac{I_{xx}-I_{zz}}{I_{yy}} pr - \frac{I_{xz}}{I_{yy}} (r^2 - p^2) \\ \dot{r} + \frac{I_{yy}-I_{xx}}{I_{zz}} pq - \frac{I_{xz}}{I_{zz}} (\dot{p} - qr) \end{bmatrix} \end{aligned} \quad (9.69)$$

Lastly, one is typically also interested in the position of the rigid body in the navigation frame. Letting  $[\dot{x}_N \dot{y}_N \dot{z}_N]^T$  represent the velocity in the navigation frame, these quantities can be related to the previous body-fixed frame by the simple rotation

$$\begin{bmatrix} \dot{x}_N \\ \dot{y}_N \\ \dot{z}_N \end{bmatrix} = C_{N \leftarrow B} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \quad (9.70)$$

or

$$\begin{bmatrix} \dot{x}_N \\ \dot{y}_N \\ \dot{z}_N \end{bmatrix} = \begin{bmatrix} \cos \theta \cos \psi & \sin \phi \sin \theta \cos \psi - \cos \phi \sin \psi & \cos \phi \sin \theta \cos \psi + \sin \phi \sin \psi \\ \cos \theta \sin \psi & \sin \phi \sin \theta \sin \psi + \cos \phi \cos \psi & \cos \phi \sin \theta \sin \psi - \sin \phi \cos \psi \\ -\sin \theta & \sin \phi \cos \theta & \cos \phi \cos \theta \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \quad (9.71)$$

where notably one can also use  $\dot{z}_N = -\dot{h}$ . Using this representation in the navigation frame, it is then possible to integrate the linear velocity to obtain the trajectory of the aircraft for a flat-Earth model,  $\vec{x}_N(t)$ . The integration of the velocity typically is done numerically because  $\phi$ ,  $\theta$ , and  $\psi$  are also functions of time.

### Rigid-Body Steady-Flight Equations

When modeling an airplane as a rigid-body, the **rigid-body equilibrium flight conditions**, also known as the **rigid-body steady-flight conditions**, by definition, occur when the state variables in the rigid airplane EOMs are constant, i.e.

$$\dot{u} = \dot{\alpha} = \dot{\beta} = \dot{p} = \dot{q} = \dot{r} = \dot{\phi} = \dot{\theta} = 0 \quad (9.72)$$

which imply that the steady-flight conditions solve the **rigid-body steady-flight equations** in the body-fixed frame as

$$\begin{bmatrix} \bar{X} - g \sin \bar{\theta} \\ \bar{Y} + g \sin \bar{\phi} \cos \bar{\theta} \\ \bar{Z} - g \cos \bar{\phi} \cos \bar{\theta} \\ \bar{L}_{roll} \\ \bar{M} \\ \bar{N} \end{bmatrix} = \begin{bmatrix} \bar{q}\bar{u} \sin \bar{\alpha} - \bar{r}\bar{u} \tan \bar{\beta} \\ \bar{r}\bar{u} - \bar{p}\bar{u} \sin \bar{\alpha} \\ \bar{p}\bar{u} \tan \bar{\beta} - \bar{q}\bar{u} \\ \frac{I_{zz}-I_{yy}}{I_{xx}} \bar{q}\bar{r} - \frac{I_{xz}}{I_{xx}} \bar{p}\bar{q} \\ \frac{I_{xx}-I_{zz}}{I_{yy}} \bar{p}\bar{r} + \frac{I_{xz}}{I_{yy}} (\bar{p}^2 - \bar{r}^2) \\ \frac{I_{yy}-I_{xx}}{I_{zz}} \bar{p}\bar{q} + \frac{I_{xz}}{I_{zz}} \bar{q}\bar{r} \end{bmatrix} \quad (9.73)$$

and the supplemental kinematic equations given as

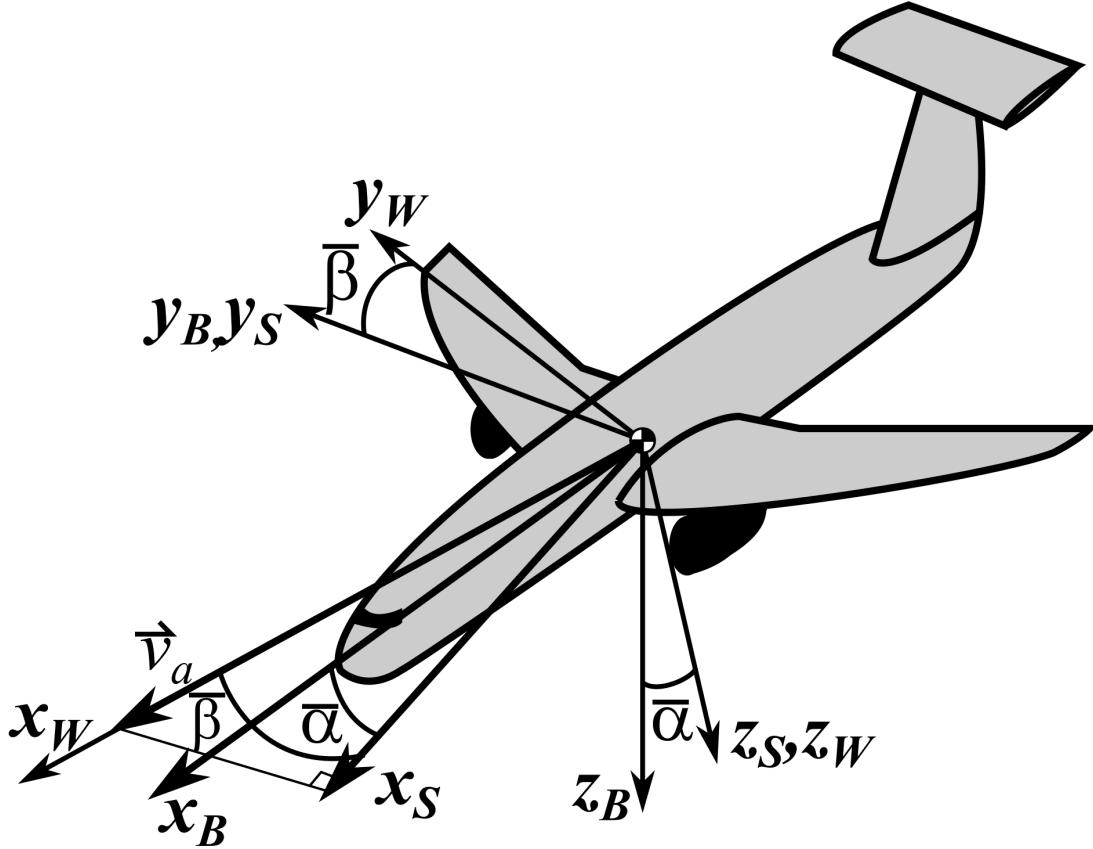
$$\begin{bmatrix} 0 \\ 0 \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} \bar{p} + \bar{q} \sin \bar{\phi} \tan \bar{\theta} + \bar{r} \cos \bar{\phi} \tan \bar{\theta} \\ \bar{q} \cos \bar{\phi} - \bar{r} \sin \bar{\phi} \\ \bar{q} \sin \bar{\phi} \sec \bar{\theta} + \bar{r} \cos \bar{\phi} \sec \bar{\theta} \end{bmatrix} \quad (9.74)$$

or

$$\begin{bmatrix} \bar{p} \\ \bar{q} \\ \bar{r} \end{bmatrix} = \begin{bmatrix} -\dot{\psi} \sin \bar{\theta} \\ \dot{\psi} \sin \bar{\phi} \cos \bar{\theta} \\ \dot{\psi} \cos \bar{\phi} \cos \bar{\theta} \end{bmatrix} \quad (9.75)$$

where the  $\bar{u}$ ,  $\bar{\beta}$ ,  $\bar{\alpha}$ ,  $\bar{p}$ ,  $\bar{q}$ ,  $\bar{r}$ ,  $\bar{\phi}$ ,  $\bar{\theta}$ , and  $\dot{\psi}$  are the steady-flight conditions while  $\bar{X}$ ,  $\bar{Y}$ ,  $\bar{Z}$ ,  $\bar{L}$ ,  $\bar{M}$ , and  $\bar{N}$  are the aerodynamic and propulsive forces and moments at steady-flight and are functions of the steady-flight conditions. Alternatively, one may use the thrust, lift, side, and drag forces at steady flight,  $\bar{T}$ ,  $\bar{L}$ ,  $\bar{S}$ , and  $\bar{D}$ , respectively, instead of  $\bar{X}$ ,  $\bar{Y}$ , and  $\bar{Z}$ . It should also be noted that the gravitational acceleration,  $g$ , and air density,  $\rho$ , will also vary as a function of altitude,  $h$ , which further requires that a strictly steady-flight condition would be at a constant altitude. However, as these variations occur slowly, one typically assumes these are constant for analyzing different “steady-flight” maneuvers.

In addition, recall that previous modeling defined the aerodynamic and propulsive forces in a body-fixed frame affixed to fuselage, also known as a **fuselage frame**. However, for the linearized rigid airplane dynamics, it is common to use an alternative body-fixed frame known as the **stability frame** (subscript  $S$ ). In particular, the stability frame is related to the fuselage frame through a rotation of  $\bar{\alpha}$  about the  $y_B$ -axis as shown in the following figure.



Note that by being defined by  $\bar{\alpha}$ , *different* stability frames are defined for *different* steady-flight conditions. However, the stability frame rotates with the airplane body as the perturbed states vary, not remaining fixed to the free-stream velocity vector as for the wind frame. Furthermore, it should be noted that  $\theta = \gamma$  if  $\theta$  describes the pitch angle from  $N$  to  $S$ .

It is important to note that here the normalized force and moment vector elements,  $X$ ,  $Y$ ,  $Z$ ,  $L_{roll}$ ,  $M$ , and  $N$ , are now defined in the stability frame, but can be related to the fuselage-fixed body frame  $B$  using a  $C_2$  basic rotation matrix about  $\bar{\alpha}$ , i.e.

$$\vec{v}_B = C_2(\bar{\alpha}) \vec{v}_S \quad \vec{v}_S = C_2^T(\bar{\alpha}) \vec{v}_B \quad (9.76)$$

where  $\vec{v}$  is some vector. In addition, the **inertia matrix in the stability frame**,  $I_S$ , is related to the inertia matrix in fuselage-fixed frame,  $I_B$ , through the transformation

$$I_B = C_2(\bar{\alpha}) I_S C_2^T(\bar{\alpha}) \quad I_S = C_2^T(\bar{\alpha}) I_B C_2(\bar{\alpha}) \quad (9.77)$$

For straight steady-flight, i.e.,  $\dot{\psi} = 0^\circ/\text{s}$ , represented in the stability frame, i.e.,  $\bar{\alpha} = 0^\circ$ , one has

$$\begin{bmatrix} \bar{p} \\ \bar{q} \\ \bar{r} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (9.78)$$

Thus, with the substitution  $\bar{\theta} = \bar{\gamma}$  which holds in the stability frame, one has

$$\begin{bmatrix} \bar{X} - g \sin \bar{\gamma} \\ \bar{Y} + g \sin \bar{\phi} \cos \bar{\gamma} \\ \bar{Z} - g \cos \bar{\phi} \cos \bar{\gamma} \\ \bar{L}_{roll} \\ \bar{M} \\ \bar{N} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (9.79)$$

Next, substituting for the propulsive and aerodynamic forces separately in the stability frame, one has

$$\begin{bmatrix} -\bar{D} \cos \bar{\beta} - \bar{S} \sin \bar{\beta} + \bar{T} \cos(\theta_T + \bar{\alpha}) \\ \bar{S} \cos \bar{\beta} - \bar{D} \sin \bar{\beta} \\ \bar{L} + \bar{T} \sin(\theta_T + \bar{\alpha}) \\ \bar{L}_{a,S} \\ \bar{M}_{a,S} + \bar{T}(z_T \cos(\theta_T + \bar{\alpha}) - x_T \sin(\theta_T + \bar{\alpha})) \\ \bar{N}_{a,S} + \bar{N}_{p,S} \end{bmatrix} = \begin{bmatrix} mg \sin \bar{\gamma} \\ -mg \sin \bar{\phi} \cos \bar{\gamma} \\ mg \cos \bar{\phi} \cos \bar{\gamma} \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (9.80)$$

where  $\bar{N}_{p,S} \neq 0$  has been included in the case of an **engine-out flight condition** which occurs for multiple-engine airplanes. These are typically separated into the **longitudinal straight steady-flight equations**

$$\begin{bmatrix} -\bar{D} \cos \bar{\beta} - \bar{S} \sin \bar{\beta} + \bar{T} \cos(\theta_T + \bar{\alpha}) \\ \bar{L} + \bar{T} \sin(\theta_T + \bar{\alpha}) \\ \bar{M}_{a,S} + \bar{T}(z_T \cos(\theta_T + \bar{\alpha}) - x_T \sin(\theta_T + \bar{\alpha})) \end{bmatrix} = \begin{bmatrix} mg \sin \bar{\gamma} \\ mg \cos \bar{\phi} \cos \bar{\gamma} \\ 0 \end{bmatrix} \quad (9.81)$$

and the **lateral-directional straight steady-flight equations**

$$\begin{bmatrix} \bar{S} \cos \bar{\beta} - \bar{D} \sin \bar{\beta} \\ \bar{L}_{a,S} \\ \bar{N}_{a,S} + \bar{N}_{p,S} \end{bmatrix} = \begin{bmatrix} -mg \sin \bar{\phi} \cos \bar{\gamma} \\ 0 \\ 0 \end{bmatrix} \quad (9.82)$$

For another simplified analysis, assume coordinated **level** turns, i.e.  $\bar{\gamma} = 0$ , the supplemental kinematic equations as

$$\begin{bmatrix} \bar{p} \\ \bar{q} \\ \bar{r} \end{bmatrix} = \begin{bmatrix} 0 \\ \dot{\psi} \sin \bar{\phi} \\ \dot{\psi} \cos \bar{\phi} \end{bmatrix} \quad (9.83)$$

where notably the steady-state pitch rate,  $\bar{q} \neq 0$ , which is due to the non-zero roll angle of the airplane required to turn. Furthermore, substituting for the aerodynamic and propulsive forces and moments, one has

$$\begin{bmatrix} -\bar{D} \cos \bar{\beta} - \bar{S} \sin \bar{\beta} + \bar{T} \cos(\theta_T + \bar{\alpha}) \\ \bar{S} \cos \bar{\beta} - \bar{D} \sin \bar{\beta} + mg \sin \bar{\phi} \\ -\bar{L} - \bar{T} \sin(\theta_T + \bar{\alpha}) + mg \cos \bar{\phi} \\ \bar{L}_{a,S} \\ \bar{M}_{a,S} + \bar{T}(z_T \cos \theta_T - x_T \sin \theta_T) \\ \bar{N}_{a,S} \end{bmatrix} = \begin{bmatrix} -m\bar{r}\bar{u} \tan \bar{\beta} \\ m\bar{r}\bar{u} \\ -m\bar{q}\bar{u} \\ (I_{zz} - I_{yy}) \bar{q}\bar{r} \\ -I_{xz}\bar{r}^2 \\ I_{xz}\bar{q}\bar{r} \end{bmatrix} \quad (9.84)$$

which assumes the thrust force is symmetric with respect to the  $x_B - z_B$  plane of the vehicle.

Next, assume that the lateral aerodynamic force is zero, i.e.

$$\bar{S} \cos \bar{\beta} - \bar{D} \sin \bar{\beta} = 0 \quad (9.85)$$

Then, the second and third equations for the lateral and vertical forces state

$$\begin{bmatrix} mg \sin \bar{\phi} \\ \bar{L} + \bar{T} \sin(\theta_T + \bar{\alpha}) - mg \cos \bar{\phi} \end{bmatrix} = \begin{bmatrix} m\bar{r}\bar{u} \\ m\bar{q}\bar{u} \end{bmatrix} \quad (9.86)$$

and substituting in terms of  $\dot{\psi}$ , one has

$$\begin{bmatrix} mg \sin \bar{\phi} \\ \bar{L} + \bar{T} \sin(\theta_T + \bar{\alpha}) - mg \cos \bar{\phi} \end{bmatrix} = \begin{bmatrix} m\bar{u} (\dot{\psi} \cos \bar{\phi}) \\ m\bar{u} (\dot{\psi} \sin \bar{\phi}) \end{bmatrix} \quad (9.87)$$

The first equation can be rewritten as

$$mg \tan \bar{\phi} = m\bar{u}\dot{\psi} \quad (9.88)$$

where one can note

$$\dot{\psi} = \frac{g}{\bar{u}} \tan \bar{\phi} \quad (9.89)$$

and substituting this into the second equation, one has

$$\bar{L} + \bar{T} \sin(\theta_T + \bar{\alpha}) = mg (\cos \bar{\phi} + \tan \bar{\phi} \sin \bar{\phi}) \quad (9.90)$$

and multiplying by  $\cos \bar{\phi}$ , one has

$$\bar{L} + \bar{T} \sin(\theta_T + \bar{\alpha}) \cos \bar{\phi} = mg (\cos^2 \bar{\phi} + \sin^2 \bar{\phi}) \quad (9.91)$$

$$\bar{L} + \bar{T} \sin(\theta_T + \bar{\alpha}) \cos \bar{\phi} = mg \quad (9.92)$$

With this in mind, one can define the dimensionless **normal load factor**,  $n$ , as

$$n(mg) = L + T \sin(\theta_T + \alpha) \quad (9.93)$$

which is typically referred to in terms of “g’s.” Thus, one has

$$\bar{n}(mg) = \frac{1}{\cos \bar{\phi}} \quad (9.94)$$

which is often specified instead of  $\bar{\phi}$  for steady turns. However, one does require the direction of the turn to be specified in this case. Note that in wings-level flight, i.e.,  $\bar{\phi} = 0$ ,  $\bar{n} = 1$  g.

Lastly, this can also be used to denote the pitch and yaw rates as

$$\begin{bmatrix} \bar{q} \\ \bar{r} \end{bmatrix} = \begin{bmatrix} \frac{g}{\bar{u}} \tan \bar{\phi} \sin \bar{\phi} \\ \frac{g}{\bar{u}} \sin \bar{\phi} \end{bmatrix} \quad (9.95)$$

or

$$\begin{bmatrix} \bar{q} \\ \bar{r} \end{bmatrix} = \begin{bmatrix} \frac{g}{\bar{u}} \left( \bar{n} - \frac{1}{\bar{n}} \right) \\ \pm \frac{g}{\bar{u}\bar{n}} \sqrt{\bar{n}^2 - 1} \end{bmatrix} \quad (9.96)$$

### Straight Steady-Flight Trim Analysis

For simplified analysis, assume **wings-level**, i.e.  $\bar{\phi} = 0^\circ$  and **coordinated flight**, i.e.  $\bar{\beta} = 0^\circ$  and  $\bar{N}_{p,S} = 0$ , then one has only three non-trivial equations

$$\begin{bmatrix} -\bar{D} + \bar{T} \cos(\theta_T + \bar{\alpha}) \\ \bar{L} + \bar{T} \sin(\theta_T + \bar{\alpha}) \\ \bar{M}_a + \bar{T}(z_T \cos(\theta_T + \bar{\alpha}) - x_T \sin(\theta_T + \bar{\alpha})) \end{bmatrix} = \begin{bmatrix} mg \sin \bar{\gamma} \\ mg \cos \bar{\gamma} \\ 0 \end{bmatrix} \quad (9.97)$$

which describe the longitudinal trim of the airplane. Next, one can express the aerodynamic forces in terms of the coefficients,  $C_L$ ,  $C_D$ , and  $C_m$ , as

$$\begin{bmatrix} -Q_\infty S_w \bar{C}_D + \bar{T} \cos(\theta_T + \bar{\alpha}) \\ Q_\infty S_w \bar{C}_L + \bar{T} \sin(\theta_T + \bar{\alpha}) \\ Q_\infty S_w \bar{c}_w \bar{C}_m + \bar{T}(z_T \cos(\theta_T + \bar{\alpha}) - x_T \sin(\theta_T + \bar{\alpha})) \end{bmatrix} = \begin{bmatrix} mg \sin \bar{\gamma} \\ mg \cos \bar{\gamma} \\ 0 \end{bmatrix} \quad (9.98)$$

Furthermore, for a simplified analysis of these equations, one can assume linear relationships for these aerodynamic coefficients with respect to the trim angle of attack,  $\bar{\alpha}$ , and trim elevator deflection,  $\bar{\delta}_e$ , i.e.

$$\bar{C}_D = C_{D_0} + C_{D_\alpha} \bar{\alpha} + C_{D_{\delta_e}} \bar{\delta}_e \quad (9.99)$$

$$\bar{C}_L = C_{L_0} + C_{L_\alpha} \bar{\alpha} + C_{L_{\delta_e}} \bar{\delta}_e \quad (9.100)$$

and

$$\bar{C}_m = C_{m_0} + C_{m_\alpha} \bar{\alpha} + C_{m_{\delta_e}} \bar{\delta}_e \quad (9.101)$$

Then, by substitution, one has

$$\begin{bmatrix} -Q_\infty S_w (C_{D_0} + C_{D_\alpha} \bar{\alpha} + C_{D_{\delta_e}} \bar{\delta}_e) + \bar{T} \cos(\theta_T + \bar{\alpha}) \\ Q_\infty S_w (C_{L_0} + C_{L_\alpha} \bar{\alpha} + C_{L_{\delta_e}} \bar{\delta}_e) + \bar{T} \sin(\theta_T + \bar{\alpha}) \\ Q_\infty S_w \bar{c}_w (C_{m_0} + C_{m_\alpha} \bar{\alpha} + C_{m_{\delta_e}} \bar{\delta}_e) + \bar{T} \frac{z_T \cos(\theta_T + \bar{\alpha}) - x_T \sin(\theta_T + \bar{\alpha})}{Q_\infty S_w \bar{c}_w} \end{bmatrix} = \begin{bmatrix} mg \sin \bar{\gamma} \\ mg \cos \bar{\gamma} \\ 0 \end{bmatrix} \quad (9.102)$$

Thus, for a given  $Q_\infty$ , i.e. a given airspeed and altitude, and flight-path angle,  $\bar{\gamma}$ , these equations determine the three unknowns for trimming the airplane, i.e.  $\bar{\alpha}$ ,  $\bar{\delta}_e$ , and  $\bar{T}$ . This can be solved using numerical methods. Two analytical methods to determine approximate trim values is to assume  $\bar{L} \gg \bar{T} \sin(\theta_T + \bar{\alpha})$  or  $\sin(\theta_T + \bar{\alpha}) \approx 0$  as well as  $\cos(\theta_T + \bar{\alpha}) \approx 1$ , then one has

$$\begin{bmatrix} -C_{D_\alpha} & -C_{D_{\delta_e}} & \frac{1}{Q_\infty S_w} \\ C_{L_\alpha} & C_{L_{\delta_e}} & 0 \\ C_{m_\alpha} & C_{m_{\delta_e}} & \frac{z_T}{Q_\infty S_w \bar{c}_w} \end{bmatrix} \begin{bmatrix} \bar{\alpha} \\ \bar{\delta}_e \\ \bar{T} \end{bmatrix} \approx \begin{bmatrix} \frac{mg}{Q_\infty S_w} \sin \bar{\gamma} + C_{D_0} \\ \frac{mg}{Q_\infty S_w} \cos \bar{\gamma} - C_{L_0} \\ -C_{m_0} \end{bmatrix} \quad (9.103)$$

which can be solved by multiplying both sides by the inverse matrix on the left side. However, if one alternatively assumes  $z_T \cos \theta_T - x_T \sin \theta_T \approx 0$ , then one has a second method via

$$\begin{bmatrix} -C_{D_\alpha} & -C_{D_{\delta_e}} & \frac{\cos(\theta_T + \bar{\alpha})}{Q_\infty S_w} \\ C_{L_\alpha} & C_{L_{\delta_e}} & 0 \\ C_{m_\alpha} & C_{m_{\delta_e}} & 0 \end{bmatrix} \begin{bmatrix} \bar{\alpha} \\ \bar{\delta}_e \\ \bar{T} \end{bmatrix} \approx \begin{bmatrix} \frac{mg}{Q_\infty S_w} \sin \bar{\gamma} + C_{D_0} \\ \frac{mg}{Q_\infty S_w} \cos \bar{\gamma} - C_{L_0} \\ -C_{m_0} \end{bmatrix} \quad (9.104)$$

which allows one to decouple  $\bar{\alpha}$  and  $\bar{\delta}_e$  from  $\bar{T}$ . This results in the analytical solution

$$\begin{aligned}\bar{\alpha} &\approx \frac{\left(\frac{mg}{Q_\infty S_w} \cos \bar{\gamma} - C_{L_0}\right) C_{m_{\delta_e}} + C_{m_0} C_{L_{\delta_e}}}{C_{L_\alpha} C_{m_{\delta_e}} - C_{m_\alpha} C_{L_{\delta_e}}} \\ \bar{\delta}_e &\approx -\frac{\left(\frac{mg}{Q_\infty S_w} \cos \bar{\gamma} - C_{L_0}\right) C_{m_\alpha} + C_{m_0} C_{L_\alpha}}{C_{L_\alpha} C_{m_{\delta_e}} - C_{m_\alpha} C_{L_{\delta_e}}}\end{aligned}\quad (9.105)$$

which is only a function of the lift and  $M$ -moment coefficients. Then, one can determine the trim thrust via the original equation

$$\bar{T} \approx \frac{mg \sin \bar{\gamma} + (C_{D_0} + C_{D_\alpha} \bar{\alpha} + C_{D_{\delta_e}} \bar{\delta}_e) Q_\infty S_w}{\cos(\theta_T + \bar{\alpha})} \quad (9.106)$$

which notably is a function of the drag coefficients and the trim angle of attack and elevator deflection.

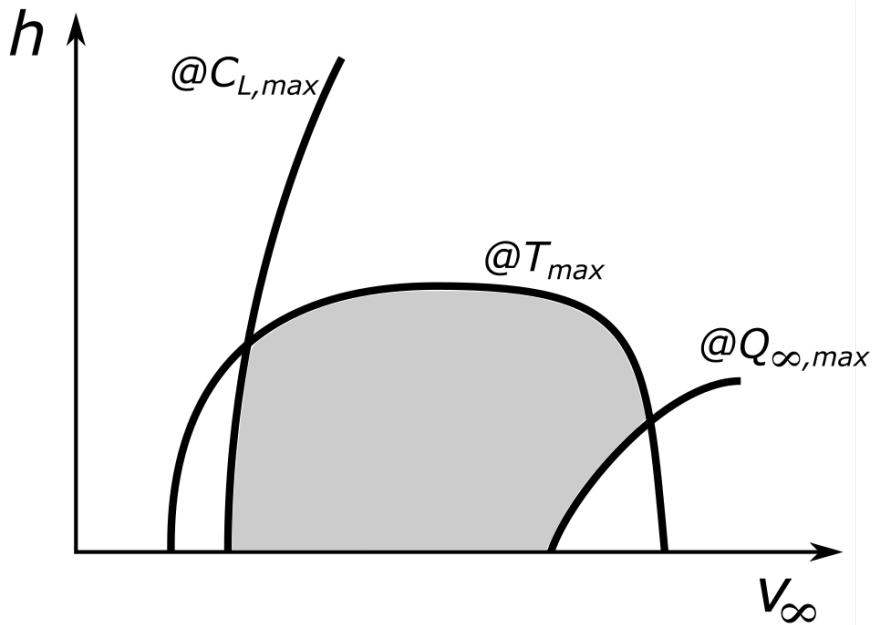
Lastly, recall that

$$\bar{L} = Q_\infty S_w \bar{C}_L \quad (9.107)$$

or

$$\bar{L} = Q_\infty S_w (C_{L_0} + C_{L_\alpha} \bar{\alpha} + C_{L_{\delta_e}} \bar{\delta}_e) \quad (9.108)$$

Thus, as these three trim values are coupled, the maximum limit on thrust,  $T_{max}$ , is the limiting factor of achievable lift via maximizing dynamic pressure,  $Q_{\infty,max}$ , and the maximum limit on elevator deflection,  $\delta_{e,max}$  or angle of attack  $\alpha_{max}$  is the limiting factor of achievable lift via maximizing the lift coefficient,  $C_{L,max}$ . These limits for straight-and-level flight, i.e.  $\bar{\gamma} = 0$ , define the vehicle's **flight envelope plot** which depicts the three curves of  $T_{max}$ ,  $C_{L,max}$ , and  $Q_{\infty,max}$  for airspeed,  $v_\infty$  versus altitude,  $h$ .



where the center region defines the possible steady-flight conditions for the airplane.

One can express the lateral-directional aerodynamic forces and moments in terms of the coefficients,  $C_S$ ,  $C_l$ , and  $C_n$ , as

$$\begin{bmatrix} Q_\infty S_w \bar{C}_S \cos \bar{\beta} - Q_\infty S_w \bar{C}_D \sin \bar{\beta} \\ Q_\infty S_w \bar{C}_l \\ Q_\infty S_w \bar{C}_n + \bar{N}_{p,S} \end{bmatrix} = \begin{bmatrix} -mg \sin \bar{\phi} \cos \bar{\gamma} \\ 0 \\ 0 \end{bmatrix} \quad (9.109)$$

Furthermore, for a simplified analysis of these equations, one can assume linear relationships for these aerodynamic coefficients with respect to the trim sideslip angle,  $\bar{\beta}$ , trim aileron deflection  $\bar{\delta}_a$ , and trim rudder deflection,  $\bar{\delta}_r$ , i.e.

$$\bar{C}_S = C_{S\beta} \bar{\beta} + C_{S\delta_a} \bar{\delta}_a + C_{S\delta_r} \bar{\delta}_r \quad (9.110)$$

$$\bar{C}_l = C_{l\beta} \bar{\beta} + C_{l\delta_a} \bar{\delta}_a + C_{l\delta_r} \bar{\delta}_r \quad (9.111)$$

and

$$\bar{C}_n = C_{n\beta} \bar{\beta} + C_{n\delta_a} \bar{\delta}_a + C_{n\delta_r} \bar{\delta}_r \quad (9.112)$$

Then, by substitution, one has

$$\begin{bmatrix} Q_\infty S_w (C_{S\beta} \bar{\beta} + C_{S\delta_a} \bar{\delta}_a + C_{S\delta_r} \bar{\delta}_r) \cos \bar{\beta} - Q_\infty S_w \bar{C}_D \sin \bar{\beta} \\ Q_\infty S_w b_w (C_{l\beta} \bar{\beta} + C_{l\delta_a} \bar{\delta}_a + C_{l\delta_r} \bar{\delta}_r) \\ Q_\infty S_w b_w (C_{n\beta} \bar{\beta} + C_{n\delta_a} \bar{\delta}_a + C_{n\delta_r} \bar{\delta}_r) + \bar{N}_{p,S} \end{bmatrix} = \begin{bmatrix} -mg \sin \bar{\phi} \cos \bar{\gamma} \\ 0 \\ 0 \end{bmatrix} \quad (9.113)$$

Then, assuming  $\bar{\beta}$  is small, i.e.  $\sin \bar{\beta} \approx \bar{\beta}$  and  $\cos \bar{\beta} \approx 1$ , one has

$$\begin{bmatrix} \bar{C}_{S\beta} - \bar{C}_D & C_{S\delta_a} & C_{S\delta_r} \\ C_{l\beta} & C_{l\delta_a} & C_{l\delta_r} \\ C_{n\beta} & C_{n\delta_a} & C_{n\delta_r} \end{bmatrix} \begin{bmatrix} \bar{\beta} \\ \bar{\delta}_a \\ \bar{\delta}_r \end{bmatrix} = \begin{bmatrix} \frac{mg}{Q_\infty S_w} \cos \bar{\gamma} \sin \bar{\phi} \\ 0 \\ -\frac{\bar{N}_{p,S}}{Q_\infty S_w b_w} \end{bmatrix} \quad (9.114)$$

which can be solved by multiplying both sides by the inverse matrix on the left side. Notably, the maximum allowable roll angle,  $\phi_{max}$ , plays a role in possible trim states for the airplane.

However, if one is required to balance the engine-out moment,  $\bar{N}_{p,S}$  only using the rudder, then the required rudder deflection must be

$$\bar{\delta}_r = -\frac{\bar{N}_{p,S}}{Q_\infty S_w b_w C_{n\delta_r}} \quad (9.115)$$

which by setting the rudder deflection to its limit,  $\bar{\delta}_r = \delta_{r,max}$ , one has the **minimum control airspeed**,  $v_{mc}$ , i.e.

$$v_{mc} = \sqrt{-\frac{\bar{N}_{p,S}}{\frac{1}{2} \rho S_w b_w C_{n\delta_r} \delta_{r,max}}} \quad (9.116)$$

which is typically used in sizing the vertical tail and rudder.

### Level Turn Trim Analysis

With these derivations in mind which notably assume

$$\bar{S} \cos \bar{\beta} - \bar{D} \sin \bar{\beta} = 0 \quad (9.117)$$

and expressing the lateral-directional aerodynamic forces and moments in terms of the coefficients,  $C_S$ ,  $C_l$ , and  $C_n$ , one has the following lateral-directional trim conditions for level turns

$$\begin{bmatrix} Q_\infty S_w \bar{C}_S \cos \bar{\beta} - Q_\infty S_w \bar{C}_D \sin \bar{\beta} \\ Q_\infty S_w \bar{C}_l \\ Q_\infty S_w \bar{C}_n \end{bmatrix} = \begin{bmatrix} 0 \\ (I_{zz} - I_{yy}) \bar{q} \bar{r} \\ I_{xz} \bar{q} \bar{r} \end{bmatrix} \quad (9.118)$$

and substituting for  $\bar{r}$  and  $\bar{q}$ , one has

$$\begin{bmatrix} Q_\infty S_w \bar{C}_S \cos \bar{\beta} - Q_\infty S_w \bar{C}_D \sin \bar{\beta} \\ Q_\infty S_w \bar{C}_l \\ Q_\infty S_w \bar{C}_n \end{bmatrix} = \begin{bmatrix} 0 \\ (I_{zz} - I_{yy}) \left(\frac{g}{\bar{u}}\right)^2 \tan \bar{\phi} \sin^2 \bar{\phi} \\ I_{xz} \left(\frac{g}{\bar{u}}\right)^2 \tan \bar{\phi} \sin^2 \bar{\phi} \end{bmatrix} \quad (9.119)$$

Furthermore, for a simplified analysis of these equations, one can assume linear relationships for these aerodynamic coefficients with respect to the trim drag coefficient,  $\bar{C}_D$ , trim side coefficient,  $\bar{C}_S$ , trim sideslip angle,  $\bar{\beta}$ , trim yaw rate,  $\bar{r}$ , trim aileron deflection  $\bar{\delta}_a$ , and trim rudder deflection,  $\bar{\delta}_r$ , i.e.

$$\bar{C}_S = C_{S_\beta} \bar{\beta} + C_{S_r} \bar{r} + C_{S_{\delta_a}} \bar{\delta}_a + C_{S_{\delta_r}} \bar{\delta}_r \quad (9.120)$$

$$\bar{C}_l = C_{l_\beta} \bar{\beta} + C_{l_r} \bar{r} + C_{l_{\delta_a}} \bar{\delta}_a + C_{l_{\delta_r}} \bar{\delta}_r \quad (9.121)$$

and

$$\bar{C}_n = C_{n_\beta} \bar{\beta} + C_{n_r} \bar{r} + C_{n_{\delta_a}} \bar{\delta}_a + C_{n_{\delta_r}} \bar{\delta}_r \quad (9.122)$$

Then, by substitution for the coefficients and  $\bar{r}$  in terms of  $\bar{\phi}$ , one has

$$\begin{bmatrix} Q_\infty S_w \left( C_{S_\beta} \bar{\beta} + C_{S_r} \frac{g}{\bar{u}} \sin \bar{\phi} + C_{S_{\delta_a}} \bar{\delta}_a + C_{S_{\delta_r}} \bar{\delta}_r \right) \cos \bar{\beta} - Q_\infty S_w \bar{C}_D \sin \bar{\beta} \\ Q_\infty S_w b_w \left( C_{l_\beta} \bar{\beta} + C_{l_r} \frac{g}{\bar{u}} \sin \bar{\phi} + C_{l_{\delta_a}} \bar{\delta}_a + C_{l_{\delta_r}} \bar{\delta}_r \right) \\ Q_\infty S_w b_w \left( C_{n_\beta} \bar{\beta} + C_{n_r} \frac{g}{\bar{u}} \sin \bar{\phi} + C_{n_{\delta_a}} \bar{\delta}_a + C_{n_{\delta_r}} \bar{\delta}_r \right) \end{bmatrix} = \begin{bmatrix} 0 \\ (I_{zz} - I_{yy}) \left(\frac{g}{\bar{u}}\right)^2 \tan \bar{\phi} \sin^2 \bar{\phi} \\ I_{xz} \left(\frac{g}{\bar{u}}\right)^2 \tan \bar{\phi} \sin^2 \bar{\phi} \end{bmatrix} \quad (9.123)$$

Thus, for a given  $Q_\infty$ , i.e. a given airspeed  $\bar{u}$  and altitude, and roll angle,  $\bar{\phi}$  or normal load factor  $\bar{n}$ , these equations determine the three unknowns for the lateral-directional trim conditions, i.e.  $\bar{\beta}$ ,  $\bar{\delta}_a$ , and  $\bar{\delta}_r$ . This can be solved using numerical methods. However, it does require that  $\bar{C}_D$  has already been obtained from the longitudinal trim analysis. However, the longitudinal analysis depends on the trim sideslip angle,  $\bar{\beta}$ , thus one typically must iterate between these trim computations with the initial attempt typically assuming  $\bar{C}_D \approx 0$ .

An analytical approach is to assume  $\bar{\beta}$  is small, i.e.  $\sin \bar{\beta} \approx \bar{\beta}$  and  $\cos \bar{\beta} \approx 1$ , then, one has

$$\begin{bmatrix} \bar{C}_{S_\beta} - \bar{C}_D & C_{S_{\delta_a}} & C_{S_{\delta_r}} \\ C_{l_\beta} & C_{l_{\delta_a}} & C_{l_{\delta_r}} \\ C_{n_\beta} & C_{n_{\delta_a}} & C_{n_{\delta_r}} \end{bmatrix} \begin{bmatrix} \bar{\beta} \\ \bar{\delta}_a \\ \bar{\delta}_r \end{bmatrix} = \begin{bmatrix} -C_{S_r} \frac{g}{\bar{u}} \sin \bar{\phi} \\ \left(\frac{I_{zz} - I_{yy}}{Q_\infty S_w b_w}\right) \left(\frac{g}{\bar{u}}\right)^2 \tan \bar{\phi} \sin^2 \bar{\phi} - C_{l_r} \frac{g}{\bar{u}} \sin \bar{\phi} \\ \frac{I_{xz}}{Q_\infty S_w b_w} \left(\frac{g}{\bar{u}}\right)^2 \tan \bar{\phi} \sin^2 \bar{\phi} - C_{n_r} \frac{g}{\bar{u}} \sin \bar{\phi} \end{bmatrix} \quad (9.124)$$

or, in terms of the trim normal load factor, one has

$$\begin{bmatrix} \bar{C}_{S_\beta} - \bar{C}_D & C_{S_{\delta_a}} & C_{S_{\delta_r}} \\ C_{l_\beta} & C_{l_{\delta_a}} & C_{l_{\delta_r}} \\ C_{n_\beta} & C_{n_{\delta_a}} & C_{n_{\delta_r}} \end{bmatrix} \begin{bmatrix} \bar{\beta} \\ \bar{\delta}_a \\ \bar{\delta}_r \end{bmatrix} = \begin{bmatrix} \mp C_{S_r} \pm \frac{g}{\bar{u}\bar{n}} \sqrt{\bar{n}^2 - 1} \\ \pm \left( \frac{I_{zz} - I_{yy}}{Q_\infty S_w b_w} \right) \left( \frac{g}{\bar{u}} \right)^2 \left( 1 - \frac{1}{\bar{n}^2} \right) \sqrt{\bar{n}^2 - 1} \mp C_{l_r} \frac{g}{\bar{u}\bar{n}} \sqrt{\bar{n}^2 - 1} \\ \pm \frac{I_{xz}}{Q_\infty S_w b_w} \left( \frac{g}{\bar{u}} \right)^2 \left( 1 - \frac{1}{\bar{n}^2} \right) \sqrt{\bar{n}^2 - 1} \mp C_{n_r} \frac{g}{\bar{u}\bar{n}} \sqrt{\bar{n}^2 - 1} \end{bmatrix} \quad (9.125)$$

which either can be solved by multiplying both sides by the inverse matrix on the left side.

Thus, having solved for the trim sideslip angle,  $\bar{\beta}$ , trim side force coefficient,  $\bar{C}_S$ , trim aileron deflection,  $\bar{\delta}_a$ , and trim rudder deflection  $\bar{\delta}_r$ , one has the following longitudinal trim conditions for level turns

$$\begin{bmatrix} -\bar{D} \cos \bar{\beta} - \bar{S} \sin \bar{\beta} + \bar{T} \cos(\theta_T + \bar{\alpha}) \\ -\bar{L} - \bar{T} \sin(\theta_T + \bar{\alpha}) + mg \cos \bar{\phi} \\ \bar{M}_a + \bar{T}(z_T \cos(\theta_T + \bar{\alpha}) - x_T \sin(\theta_T + \bar{\alpha})) \end{bmatrix} = \begin{bmatrix} -m\bar{r}\bar{u} \tan \bar{\beta} \\ -m\bar{q}\bar{u} \\ -I_{xz}\bar{r}^2 \end{bmatrix} \quad (9.126)$$

Next, one can express the longitudinal aerodynamic forces and moments in terms of the coefficients,  $C_L$ ,  $C_S$ ,  $C_D$ , and  $C_m$ , as

$$\begin{bmatrix} -Q_\infty S_w \bar{C}_D \cos \bar{\beta} - Q_\infty S_w \bar{C}_S \sin \bar{\beta} + \bar{T} \cos(\theta_T + \bar{\alpha}) \\ Q_\infty S_w \bar{C}_L + \bar{T} \sin(\theta_T + \bar{\alpha}) \\ Q_\infty S_w \bar{C}_m + \bar{T}(z_T \cos(\theta_T + \bar{\alpha}) - x_T \sin(\theta_T + \bar{\alpha})) \end{bmatrix} = \begin{bmatrix} -m\bar{r}\bar{u} \tan \bar{\beta} \\ -m\bar{q}\bar{u} \\ -I_{xz}\bar{r}^2 \end{bmatrix} \quad (9.127)$$

Furthermore, for a simplified analysis of these equations, one can assume linear relationships for  $C_L$ ,  $C_D$ , and  $C_m$  with respect to the trim angle of attack,  $\bar{\alpha}$ , trim pitch rate,  $\bar{q}$ , trim elevator deflection,  $\bar{\delta}_e$ , trim aileron deflection,  $\bar{\delta}_a$ , and trim rudder deflection  $\bar{\delta}_r$ , i.e.,

$$\bar{C}_D = C_{D_0} + C_{D_\alpha} \bar{\alpha} + C_{D_q} \bar{q} + C_{D_{\delta_e}} \bar{\delta}_e + C_{D_{\delta_a}} \bar{\delta}_a + C_{D_{\delta_r}} \bar{\delta}_r \quad (9.128)$$

$$\bar{C}_L = C_{L_0} + C_{L_\alpha} \bar{\alpha} + C_{L_q} \bar{q} + C_{L_{\delta_e}} \bar{\delta}_e \quad (9.129)$$

and

$$\bar{C}_m = C_{m_0} + C_{m_\alpha} \bar{\alpha} + C_{m_q} \bar{q} + C_{m_{\delta_e}} \bar{\delta}_e \quad (9.130)$$

Then, by substitution, one has

$$\begin{bmatrix} -Q_\infty S_w \left( C_{D_0} + C_{D_\alpha} \bar{\alpha} + C_{D_q} \bar{q} + C_{D_{\delta_e}} \bar{\delta}_e + C_{D_{\delta_a}} \bar{\delta}_a + C_{D_{\delta_r}} \bar{\delta}_r \right) \cos \bar{\beta} - Q_\infty S_w \bar{C}_S \sin \bar{\beta} + \bar{T} \cos(\theta_T + \bar{\alpha}) \\ Q_\infty S_w \left( C_{L_0} + C_{L_\alpha} \bar{\alpha} + C_{L_q} \bar{q} + C_{L_{\delta_e}} \bar{\delta}_e \right) + \bar{T} \sin(\theta_T + \bar{\alpha}) \\ Q_\infty S_w \bar{C}_m \left( C_{m_0} + C_{m_\alpha} \bar{\alpha} + C_{m_q} \bar{q} + C_{m_{\delta_e}} \bar{\delta}_e \right) + \bar{T} \frac{z_T \cos(\theta_T + \bar{\alpha}) - x_T \sin(\theta_T + \bar{\alpha})}{Q_\infty S_w \bar{C}_w} \end{bmatrix} = \begin{bmatrix} -m\bar{r}\bar{u} \tan \bar{\beta} \\ -m\bar{q}\bar{u} \\ -I_{xz}\bar{r}^2 \end{bmatrix} \quad (9.131)$$

Thus, for a given  $Q_\infty$ , i.e. a given airspeed  $\bar{u}$  and altitude, and roll angle,  $\bar{\phi}$  or normal load factor  $\bar{n}$ , these equations determine the three unknowns for the longitudinal trim conditions, i.e.  $\bar{\alpha}$ ,  $\bar{\delta}_e$ , and  $\bar{T}$ . This can be solved using numerical methods.

An analytical approach is to assume  $\bar{\beta}$  is small, i.e.  $\sin \bar{\beta} \approx \bar{\beta}$  and  $\cos \bar{\beta} \approx 1$ ,  $\cos(\theta_T + \bar{\alpha}) = 1$ , and  $\sin(\theta_T + \bar{\alpha}) = 0$ , then, one has

$$\begin{bmatrix} -C_{D_\alpha} & -C_{D_{\delta_e}} & \frac{1}{Q_\infty S_w} \\ C_{L_\alpha} & C_{L_{\delta_e}} & 0 \\ C_{m_\alpha} & C_{m_{\delta_e}} & \frac{z_T}{Q_\infty S_w \bar{C}_w} \end{bmatrix} \begin{bmatrix} \bar{\alpha} \\ \bar{\delta}_e \\ \bar{T} \end{bmatrix} = \begin{bmatrix} \left( \bar{C}_S - \frac{mg}{Q_\infty S_w} \sin \bar{\phi} \right) \bar{\beta} + C_{D_0} + C_{D_q} \frac{g}{\bar{u}} \tan \bar{\phi} \sin \bar{\phi} + C_{D_{\delta_a}} \bar{\delta}_a + C_{D_{\delta_r}} \bar{\delta}_r \\ \frac{mg}{Q_\infty S_w \cos \bar{\phi}} - C_{L_0} - C_{L_q} \frac{g}{\bar{u}} \tan \bar{\phi} \sin \bar{\phi} \\ -I_{xz} \left( \frac{g}{\bar{u}} \right)^2 \sin^2 \bar{\phi} - C_{m_0} - C_{m_q} \frac{g}{\bar{u}} \tan \bar{\phi} \sin \bar{\phi} \end{bmatrix} \quad (9.132)$$

or, in terms of the trim normal load factor, one has

$$\begin{bmatrix} -C_{D_\alpha} & -C_{D_{\delta_e}} & \frac{1}{Q_\infty S_w} \\ C_{L_\alpha} & C_{L_{\delta_e}} & 0 \\ C_{m_\alpha} & C_{m_{\delta_e}} & \frac{z_T}{Q_\infty S_w \bar{c}_w} \end{bmatrix} \begin{bmatrix} \bar{\alpha} \\ \bar{\delta}_e \\ \bar{T} \end{bmatrix} = \begin{bmatrix} \left( \bar{C}_S - \frac{mg}{Q_\infty S_w} \sqrt{1 - \frac{1}{\bar{n}^2}} \right) \bar{\beta} + C_{D_0} + C_{D_q} \frac{g}{\bar{u}} \left( 1 - \frac{1}{\bar{n}^2} \right) + C_{D_{\delta_a}} \bar{\delta}_a + C_{D_{\delta_r}} \bar{\delta}_r \\ \frac{mg}{Q_\infty S_w} - C_{L_0} - C_{L_q} \frac{g}{\bar{u}} \left( \bar{n} - \frac{1}{\bar{n}} \right) \\ -I_{xz} \left( \frac{g}{\bar{u}} \right)^2 \left( 1 - \frac{1}{\bar{n}^2} \right) - C_{m_0} - C_{m_q} \frac{g}{\bar{u}} \left( \bar{n} - \frac{1}{\bar{n}} \right) \end{bmatrix} \quad (9.133)$$

which either can be solved by multiplying both sides by the inverse matrix on the left side.

However, if one alternatively assumes  $z_T \cos \theta_T - x_T \sin \theta_T \approx 0$ , then one has a second method via

$$\begin{bmatrix} -C_{D_\alpha} & -C_{D_{\delta_e}} & \frac{\cos(\theta_T + \bar{\alpha})}{Q_\infty S_w} \\ C_{L_\alpha} & C_{L_{\delta_e}} & 0 \\ C_{m_\alpha} & C_{m_{\delta_e}} & 0 \end{bmatrix} \begin{bmatrix} \bar{\alpha} \\ \bar{\delta}_e \\ \bar{T} \end{bmatrix} = \begin{bmatrix} \left( \bar{C}_S - \frac{mg}{Q_\infty S_w} \sqrt{1 - \frac{1}{\bar{n}^2}} \right) \bar{\beta} + C_{D_0} + C_{D_q} \frac{g}{\bar{u}} \left( 1 - \frac{1}{\bar{n}^2} \right) + C_{D_{\delta_a}} \bar{\delta}_a + C_{D_{\delta_r}} \bar{\delta}_r \\ \bar{n} \frac{mg}{Q_\infty S_w} - C_{L_0} - C_{L_q} \frac{g}{\bar{u}} \left( \bar{n} - \frac{1}{\bar{n}} \right) \\ -I_{xz} \left( \frac{g}{\bar{u}} \right)^2 \left( 1 - \frac{1}{\bar{n}^2} \right) - C_{m_0} - C_{m_q} \frac{g}{\bar{u}} \left( \bar{n} - \frac{1}{\bar{n}} \right) \end{bmatrix} \quad (9.134)$$

which allows one to decouple  $\bar{\alpha}$  and  $\bar{\delta}_e$  from  $\bar{T}$ . This results in the analytical solution

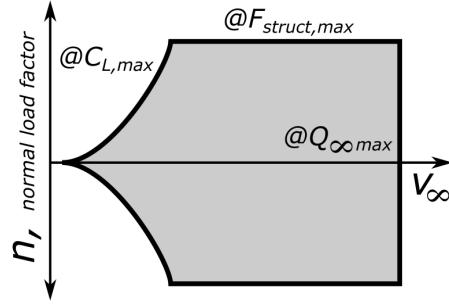
$$\begin{aligned} \bar{\alpha} &\approx \frac{C_1(n) C_{m_{\delta_e}} - C_2(n) C_{L_{\delta_e}}}{C_{L_\alpha} C_{m_{\delta_e}} - C_{m_\alpha} C_{L_{\delta_e}}} \\ \bar{\delta}_e &\approx -\frac{C_1(n) C_{m_\alpha} - C_2(n) C_{L_\alpha}}{C_{L_\alpha} C_{m_{\delta_e}} - C_{m_\alpha} C_{L_{\delta_e}}} \\ C_1(n) &= \bar{n} \frac{mg}{Q_\infty S_w} - C_{L_0} - C_{L_q} \frac{g}{\bar{u}} \left( \bar{n} - \frac{1}{\bar{n}} \right) \\ C_2(n) &= -I_{xz} \left( \frac{g}{\bar{u}} \right)^2 \left( 1 - \frac{1}{\bar{n}^2} \right) - C_{m_0} - C_{m_q} \frac{g}{\bar{u}} \left( \bar{n} - \frac{1}{\bar{n}} \right) \end{aligned} \quad (9.135)$$

which is only a function of the normal load factor and the lift and  $M$ -moment coefficients. Then, one can determine the trim thrust via the drag equation

$$\bar{T} = \frac{-mg \sqrt{1 - \frac{1}{\bar{n}^2}} \tan \bar{\beta} + Q_\infty S_w \bar{C}_D \cos \bar{\beta} + Q_\infty S_w \bar{C}_S \sin \bar{\beta}}{\cos(\theta_T + \bar{\alpha})} \quad (9.136)$$

which notably is a function of the drag coefficients and the trim angle of attack and elevator deflection.

As for the flight envelope, the lift coefficient,  $C_{L,max}$ , dynamic pressure,  $Q_{\infty,max}$ , and structural limits,  $F_{struct,max}$ , for level, turning flight define the vehicle's **v-n plot** plotting airspeed,  $v_\infty$ , versus the normal load factor,  $n$ , as



where  $C_{L,max}$  results from either the maximum angle of attack or the maximum elevator deflection,  $Q_{\infty,max}$  is often due to a flutter limit, and the center region defines the possible steady-flight conditions for the airplane.

## References

For more information, please refer to the following

- Schmidt, D. K., “9.1 Equilibrium Reference Conditions,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 481-485
- Schmidt, D. K., “9.3 Analysis of Steady Rectilinear Flight,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 500-523

## 9.5 Rigid Airplane Dynamics and Stability

For linearized flight dynamics, it is useful to form the following approximate relationships between the vehicle frame Euler angles using the small angle approximations for sin and cos. For Equation 7.50, this becomes

$$\begin{bmatrix} 1 & \sigma & -\gamma \\ \mu\gamma - \sigma & \mu\gamma + 1 & \mu \\ \gamma + \mu\sigma & \gamma\mu - \mu & 1 \end{bmatrix} \approx \begin{bmatrix} 1 & \beta & \alpha \\ -\beta & 1 & 0 \\ -\alpha & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \psi & -\theta \\ \phi\theta - \psi & \phi\theta + 1 & \phi \\ \theta + \phi\psi & \theta\psi - \phi & 1 \end{bmatrix} \quad (9.137)$$

By computing the element in the second row and third column, one has

$$\mu \approx \theta\beta + \phi \quad (9.138)$$

and discarding higher-order terms, one has

$$\mu \approx \phi \quad (9.139)$$

By computing the element in the first row and third column

$$-\gamma \approx -\theta + \beta\phi + \alpha \quad (9.140)$$

and discarding higher-order terms, one has

$$\gamma \approx \theta - \alpha \quad (9.141)$$

By computing the element in the first row and second column, one has

$$\sigma \approx \psi + \beta + \phi\theta(1 + \beta) - \psi\theta^2 \quad (9.142)$$

and discarding higher-order terms, one has

$$\sigma \approx \psi + \beta \quad (9.143)$$

It should be noted that small  $\sigma$  and  $\psi$  here correspond to small *changes* in the *nominal* heading and yaw as these both can be set to an arbitrary reference direction in the rotation sequence.

### Decoupled Rigid Airplane Linearized Dynamics

This section details the linearization of a conventional airplane dynamics equations about straight flight in the stability frame with coordinated flight, i.e.  $\bar{\beta} = 0$ , and wings-level flight, i.e.  $\bar{\phi} = 0$ . Before beginning the derivation, it should be pointed out the airplane linearized, time-invariant state equation will use states  $u$ ,  $\alpha$ ,  $\beta$ ,  $p$ ,  $q$ ,  $r$ ,  $\phi$ , and  $\theta$ , and inputs  $\delta_a$ ,  $\delta_e$ ,  $\delta_r$ ,  $T$  in perturbed form with leading  $\Delta$ 's. The output equation will not be addressed in this section as it depends on the use of the dynamics model.

First, recall the airplane 6-DOF equations of motion as

$$\begin{bmatrix} X - g \sin \theta \\ Y + g \cos \theta \sin \phi \\ Z + g \cos \theta \cos \phi \\ L \\ M \\ N \end{bmatrix} = \begin{bmatrix} \dot{u} + qu \sin \alpha - ru \tan \beta \\ \dot{u} \tan \beta + \dot{\beta}u \sec^2 \beta + ru - pu \sin \alpha \\ \dot{u} \sin \alpha + \dot{\alpha}u \cos \alpha + pu \tan \beta - qu \\ \dot{p} + \frac{I_{zz} - I_{yy}}{I_{xx}} qr - \frac{I_{xz}}{I_{xx}} (\dot{r} + pq) \\ \dot{q} + \frac{I_{xx} - I_{zz}}{I_{yy}} pr - \frac{I_{xz}}{I_{yy}} (r^2 - p^2) \\ \dot{r} + \frac{I_{yy} - I_{xx}}{I_{zz}} pq - \frac{I_{xz}}{I_{zz}} (\dot{p} - qr) \end{bmatrix} \quad (9.144)$$

Next, for straight, coordinated, wings-level steady-flight, i.e.  $\dot{\psi} = \bar{\beta} = \bar{\phi} = 0$ , one can decouple these into the linearized longitudinal and lateral-directional EOMs using the trim and perturbed states, forces, and moments as

$$\begin{bmatrix} \bar{X} + \Delta X - g \sin(\bar{\theta} + \Delta\theta) \\ \bar{Z} + \Delta Z + g \cos(\bar{\theta} + \Delta\theta) \cos \Delta\phi \\ \bar{M} + \Delta M \end{bmatrix} = \begin{bmatrix} \Delta \dot{u} + \Delta q (\bar{u} + \Delta u) \sin \Delta\alpha \\ \Delta \dot{u} \sin \Delta\alpha + \Delta \dot{\alpha} (\bar{u} + \Delta u) \cos \Delta\alpha - \Delta q (\bar{u} + \Delta u) \\ \Delta \dot{q} \end{bmatrix} \quad (9.145)$$

and

$$\begin{bmatrix} \bar{Y} + \Delta Y + g \cos \bar{\theta} \sin \Delta\phi \\ \bar{L} + \Delta L \\ \bar{N} + \Delta N \end{bmatrix} = \begin{bmatrix} \Delta \dot{\beta} \bar{u} \sec^2 \Delta\beta + \bar{u} (\Delta r) \\ \Delta \dot{p} - \frac{I_{xz}}{I_{xx}} \Delta \dot{r} \\ \Delta \dot{r} - \frac{I_{xz}}{I_{zz}} \Delta \dot{p} \end{bmatrix} \quad (9.146)$$

Then, using the trigonometric addition formulas and the small angle approximation, i.e.

$$\sin(\bar{a} + \Delta a) = \sin \bar{a} \cos \Delta a + \cos \bar{a} \sin \Delta a = \sin \bar{a} + \cos \bar{a} \Delta a \quad (9.147)$$

and

$$\cos(\bar{a} + \Delta a) = \cos \bar{a} \cos \Delta a - \sin \bar{a} \sin \Delta a = \cos \bar{a} - \sin \bar{a} \Delta a \quad (9.148)$$

for  $\theta, \phi, \alpha$ , and  $\beta$  where  $\bar{\phi} = \bar{\alpha} = \bar{\beta} = 0$  has already been assumed, one has

$$\begin{bmatrix} \bar{X} + \Delta X - g \sin \bar{\theta} - g \cos \bar{\theta} \Delta \theta \\ \bar{Z} + \Delta Z - g \sin \bar{\theta} \Delta \theta \\ \bar{M} + \Delta M \end{bmatrix} = \begin{bmatrix} \Delta \dot{u} + \Delta q (\bar{u} + \Delta u) \Delta \alpha \\ \Delta \dot{u} \Delta \alpha + \Delta \dot{\alpha} (\bar{u} + \Delta u) - \Delta q (\bar{u} + \Delta u) \\ \Delta \dot{q} \end{bmatrix} \quad (9.149)$$

and

$$\begin{bmatrix} \bar{Y} + \Delta Y - g \cos \bar{\theta} \Delta \phi \\ \bar{L} + \Delta L \\ \bar{N} + \Delta N \end{bmatrix} = \begin{bmatrix} \Delta \dot{\beta} \bar{u} + \Delta r \bar{u} \\ \Delta \dot{p} - \frac{I_{xz}}{I_{xx}} \Delta \dot{r} \\ \Delta \dot{r} - \frac{I_{xz}}{I_{zz}} \Delta \dot{p} \end{bmatrix} \quad (9.150)$$

These can be further linearized by eliminating higher-order terms of the perturbations and separating out the perturbation terms, i.e.

$$\begin{bmatrix} \bar{X} - g \sin \bar{\theta} \\ \bar{Z} \\ \bar{M} \end{bmatrix} + \begin{bmatrix} \Delta X \\ \Delta Z \\ \Delta M \end{bmatrix} + \begin{bmatrix} -g \cos \bar{\theta} \\ -g \sin \bar{\theta} \\ 0 \end{bmatrix} \Delta \theta = \begin{bmatrix} \Delta \dot{u} \\ \bar{u} \Delta \dot{\alpha} - \bar{u} \Delta q \\ \Delta \dot{q} \end{bmatrix} \quad (9.151)$$

and

$$\begin{bmatrix} \bar{Y} \\ \bar{L} \\ \bar{N} \end{bmatrix} + \begin{bmatrix} \Delta Y \\ \Delta L \\ \Delta N \end{bmatrix} + \begin{bmatrix} -g \cos \bar{\theta} \\ 0 \\ 0 \end{bmatrix} \Delta \phi = \begin{bmatrix} \Delta \dot{\beta} \bar{u} + \Delta r \bar{u} \\ \Delta \dot{p} - \frac{I_{xz}}{I_{xx}} \Delta \dot{r} \\ \Delta \dot{r} - \frac{I_{xz}}{I_{zz}} \Delta \dot{p} \end{bmatrix} \quad (9.152)$$

Next, by definition of straight steady-flight conditions for which

$$\begin{aligned} \bar{X} - g \sin \bar{\theta} &= 0 \\ \bar{Y} &= 0 \\ \bar{Z} + g \cos \bar{\theta} &= 0 \\ \bar{L} &= 0 \\ \bar{M} &= 0 \\ \bar{N} &= 0 \end{aligned} \quad (9.153)$$

due to  $\vec{\omega}_{S/N} = 0$ , the linearized models can be rewritten using matrices and the states as

$$\begin{bmatrix} \Delta X \\ \Delta Z \\ \Delta M \end{bmatrix} + \begin{bmatrix} -g \cos \bar{\theta} \\ -g \sin \bar{\theta} \\ 0 \end{bmatrix} \Delta \theta + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \bar{u} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \bar{u} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta \dot{u} \\ \Delta \dot{\alpha} \\ \Delta \dot{q} \end{bmatrix} \quad (9.154)$$

and

$$\begin{bmatrix} \Delta Y \\ \Delta L \\ \Delta N \end{bmatrix} + \begin{bmatrix} -g \cos \bar{\theta} \\ 0 \\ 0 \end{bmatrix} \Delta \phi - \begin{bmatrix} 0 & 0 & \bar{u} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \beta \\ \Delta p \\ \Delta r \end{bmatrix} = \begin{bmatrix} \bar{u} & 0 & 0 \\ 0 & 1 & -\frac{I_{xz}}{I_{xx}} \\ 0 & -\frac{I_{xz}}{I_{xx}} & 1 \end{bmatrix} \begin{bmatrix} \Delta \dot{\beta} \\ \Delta \dot{p} \\ \Delta \dot{r} \end{bmatrix} \quad (9.155)$$

As a general airplane modeling principle, the perturbed aerodynamic and propulsive forces and moments are modeled as two sets

$$\begin{bmatrix} \Delta X \\ \Delta Z \\ \Delta M \end{bmatrix} = \begin{bmatrix} X_{\dot{u}} & X_{\dot{\alpha}} & X_{\dot{q}} \\ Z_{\dot{u}} & Z_{\dot{\alpha}} & Z_{\dot{q}} \\ M_{\dot{u}} & M_{\dot{\alpha}} & M_{\dot{q}} \end{bmatrix} \begin{bmatrix} \Delta \dot{u} \\ \Delta \dot{\alpha} \\ \Delta \dot{q} \end{bmatrix} + \begin{bmatrix} X_u & X_\alpha & X_q \\ Z_u & Z_\alpha & Z_q \\ M_u & M_\alpha & M_q \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \end{bmatrix} + \begin{bmatrix} X_{\delta_e} & X_{\delta_t} \\ Z_{\delta_e} & Z_{\delta_t} \\ M_{\delta_e} & M_{\delta_t} \end{bmatrix} \begin{bmatrix} \Delta \delta_e \\ \Delta \delta_t \end{bmatrix} \quad (9.156)$$

and

$$\begin{bmatrix} \Delta Y \\ \Delta L \\ \Delta N \end{bmatrix} = \begin{bmatrix} Y_{\dot{\beta}} & Y_{\dot{p}} & Y_{\dot{r}} \\ L_{\dot{\beta}} & L_{\dot{p}} & L_{\dot{r}} \\ N_{\dot{\beta}} & N_{\dot{p}} & N_{\dot{r}} \end{bmatrix} \begin{bmatrix} \Delta \dot{\beta} \\ \Delta \dot{p} \\ \Delta \dot{r} \end{bmatrix} + \begin{bmatrix} Y_{\beta} & Y_p & Y_r \\ L_{\beta} & L_p & L_r \\ N_{\beta} & N_p & N_r \end{bmatrix} \begin{bmatrix} \Delta \beta \\ \Delta p \\ \Delta r \end{bmatrix} + \begin{bmatrix} Y_{\delta_a} & Y_{\delta_r} \\ L_{\delta_a} & L_{\delta_r} \\ N_{\delta_a} & N_{\delta_r} \end{bmatrix} \begin{bmatrix} \Delta \delta_a \\ \Delta \delta_r \end{bmatrix} \quad (9.157)$$

where the coefficients of the perturbed states and inputs inside these matrices are called the **stability and control derivatives** and correspond to the Jacobian partial derivative terms about trimmed steady-flight.

As these derivatives generally change with an airplane's trim conditions, one typically calculates tables of these derivatives at many steady-flight conditions using wind tunnels tests, flight tests, and/or computational fluid dynamics (CFD). Methods for determining these derivatives from such data fall under the discipline of **airplane system identification**, in particular, optimal parameter estimation, a topic addressed later in this textbook. In this textbook, the linearized dynamics derivation for airplanes will assume the following stability and control derivatives dominate the perturbed forces and moments

$$\begin{bmatrix} \Delta X \\ \Delta Z \\ \Delta M \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & Z_{\dot{\alpha}} & 0 \\ 0 & M_{\dot{\alpha}} & 0 \end{bmatrix} \begin{bmatrix} \Delta \dot{u} \\ \Delta \dot{\alpha} \\ \Delta \dot{q} \end{bmatrix} + \begin{bmatrix} X_u & X_\alpha & 0 \\ Z_u & Z_\alpha & Z_q \\ M_u & M_\alpha & M_q \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \end{bmatrix} + \begin{bmatrix} 0 & X_{\delta_t} \\ Z_{\delta_t} & Z_{\delta_t} \\ M_{\delta_t} & M_{\delta_t} \end{bmatrix} \begin{bmatrix} \Delta \delta_e \\ \Delta \delta_t \end{bmatrix} \quad (9.158)$$

and

$$\begin{bmatrix} \Delta Y \\ \Delta L \\ \Delta N \end{bmatrix} = \begin{bmatrix} Y_{\beta} & Y_p & Y_r \\ L_{\beta} & L_p & L_r \\ N_{\beta} & N_p & N_r \end{bmatrix} \begin{bmatrix} \Delta \beta \\ \Delta p \\ \Delta r \end{bmatrix} + \begin{bmatrix} 0 & Y_{\delta_r} \\ L_{\delta_a} & L_{\delta_r} \\ N_{\delta_a} & N_{\delta_r} \end{bmatrix} \begin{bmatrix} \Delta \delta_a \\ \Delta \delta_r \end{bmatrix} \quad (9.159)$$

Thus, substituting for the perturbed aerodynamic and propulsive forces and moments as assumed in this course, one has

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & Z_{\dot{\alpha}} & 0 \\ 0 & M_{\dot{\alpha}} & 0 \end{bmatrix} \begin{bmatrix} \Delta \dot{u} \\ \Delta \dot{\alpha} \\ \Delta \dot{q} \end{bmatrix} + \begin{bmatrix} X_u & X_\alpha & 0 \\ Z_u & Z_\alpha & Z_q \\ M_u & M_\alpha & M_q \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \end{bmatrix} + \begin{bmatrix} 0 & X_{\delta_t} \\ Z_{\delta_t} & Z_{\delta_t} \\ M_{\delta_t} & M_{\delta_t} \end{bmatrix} \begin{bmatrix} \Delta \delta_e \\ \Delta \delta_t \end{bmatrix} \quad (9.160)$$

$$+ \begin{bmatrix} -g \cos \bar{\theta} \\ -g \sin \bar{\theta} \\ 0 \end{bmatrix} \Delta \theta + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \bar{u} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \bar{u} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta \dot{u} \\ \Delta \dot{\alpha} \\ \Delta \dot{q} \end{bmatrix}$$

and

$$\begin{bmatrix} Y_{\beta} & Y_p & Y_r \\ L_{\beta} & L_p & L_r \\ N_{\beta} & N_p & N_r \end{bmatrix} \begin{bmatrix} \Delta \beta \\ \Delta p \\ \Delta r \end{bmatrix} + \begin{bmatrix} 0 & Y_{\delta_r} \\ L_{\delta_a} & L_{\delta_r} \\ N_{\delta_a} & N_{\delta_r} \end{bmatrix} \begin{bmatrix} \Delta \delta_a \\ \Delta \delta_r \end{bmatrix} \quad (9.161)$$

$$+ \begin{bmatrix} -g \cos \bar{\theta} \\ 0 \\ 0 \end{bmatrix} \Delta \phi - \begin{bmatrix} 0 & 0 & \bar{u} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \beta \\ \Delta p \\ \Delta r \end{bmatrix} = \begin{bmatrix} \bar{u} & 0 & 0 \\ 0 & 1 & -\frac{I_{xz}}{I_{xx}} \\ 0 & -\frac{I_{xz}}{I_{xx}} & 1 \end{bmatrix} \begin{bmatrix} \Delta \dot{\beta} \\ \Delta \dot{p} \\ \Delta \dot{r} \end{bmatrix}$$

Next, combining matrices of similar terms and reversing the sides of the equations, one has

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \bar{u} - Z_{\dot{\alpha}} & 0 \\ 0 & -M_{\dot{\alpha}} & 1 \end{bmatrix} \begin{bmatrix} \Delta \dot{u} \\ \Delta \dot{\alpha} \\ \Delta \dot{q} \end{bmatrix} = \begin{bmatrix} X_u & X_\alpha & 0 & -g \cos \bar{\theta} \\ Z_u & Z_\alpha & \bar{u} + Z_q & -g \sin \bar{\theta} \\ M_u & M_\alpha & M_q & 0 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \end{bmatrix} + \begin{bmatrix} 0 & X_{\delta_t} \\ Z_{\delta_t} & Z_{\delta_t} \\ M_{\delta_t} & M_{\delta_t} \end{bmatrix} \begin{bmatrix} \Delta \delta_e \\ \Delta \delta_t \end{bmatrix} \quad (9.162)$$

and

$$\begin{bmatrix} \bar{u} & 0 & 0 \\ 0 & 1 & -\frac{I_{xz}}{I_{xx}} \\ 0 & -\frac{I_{xz}}{I_{xx}} & 1 \end{bmatrix} \begin{bmatrix} \Delta\dot{\beta} \\ \Delta\dot{p} \\ \Delta\dot{r} \end{bmatrix} = \begin{bmatrix} Y_\beta & Y_p & Y_r - \bar{u} & -g \cos \bar{\theta} \\ L_\beta & L_p & L_r & 0 \\ N_\beta & N_p & N_r & 0 \end{bmatrix} \begin{bmatrix} \Delta\beta \\ \Delta p \\ \Delta r \\ \Delta\phi \end{bmatrix} + \begin{bmatrix} 0 & Y_{\delta_r} \\ L_{\delta_a} & L_{\delta_r} \\ N_{\delta_a} & N_{\delta_r} \end{bmatrix} \begin{bmatrix} \Delta\delta_a \\ \Delta\delta_r \end{bmatrix} \quad (9.163)$$

Then, recalling for small angles

$$\Delta\dot{\theta} = \Delta q \quad (9.164)$$

and

$$\Delta\dot{\phi} = \Delta p + \tan \bar{\theta} \Delta r \quad (9.165)$$

one has

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \bar{u} - Z_{\dot{\alpha}} & 0 & 0 \\ 0 & -M_{\dot{\alpha}} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta\dot{u} \\ \Delta\dot{\alpha} \\ \Delta\dot{q} \\ \Delta\dot{\theta} \end{bmatrix} = \begin{bmatrix} X_u & X_\alpha & 0 & -g \cos \bar{\theta} \\ Z_u & Z_\alpha & \bar{u} + Z_q & -g \sin \bar{\theta} \\ M_u & M_\alpha & M_q & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta\alpha \\ \Delta q \\ \Delta\theta \end{bmatrix} + \begin{bmatrix} 0 & X_{\delta_t} \\ Z_{\delta_e} & Z_{\delta_t} \\ M_{\delta_e} & M_{\delta_t} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta\delta_e \\ \Delta\delta_t \end{bmatrix} \quad (9.166)$$

and

$$\begin{bmatrix} \bar{u} & 0 & 0 & 0 \\ 0 & 1 & -\frac{I_{xz}}{I_{xx}} & 0 \\ 0 & -\frac{I_{xz}}{I_{xx}} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta\dot{\beta} \\ \Delta\dot{p} \\ \Delta\dot{r} \\ \Delta\dot{\phi} \end{bmatrix} = \begin{bmatrix} Y_\beta & Y_p & Y_r - \bar{u} & -g \cos \bar{\theta} \\ L_\beta & L_p & L_r & 0 \\ N_\beta & N_p & N_r & 0 \\ 0 & 1 & \tan \bar{\theta} & 0 \end{bmatrix} \begin{bmatrix} \Delta\beta \\ \Delta p \\ \Delta r \\ \Delta\phi \end{bmatrix} + \begin{bmatrix} 0 & Y_{\delta_r} \\ L_{\delta_a} & L_{\delta_r} \\ N_{\delta_a} & N_{\delta_r} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta\delta_a \\ \Delta\delta_r \end{bmatrix} \quad (9.167)$$

Finally, taking the inverse matrices of the left side, one has the **linearized longitudinal rigid airplane dynamics**

$$\begin{bmatrix} \Delta\dot{u} \\ \Delta\dot{\alpha} \\ \Delta\dot{q} \\ \Delta\dot{\theta} \end{bmatrix} = \begin{bmatrix} X_u & X_\alpha & 0 & -g \cos \bar{\theta} \\ \frac{Z_u}{\bar{u} - Z_{\dot{\alpha}}} & \frac{Z_\alpha}{\bar{u} - Z_{\dot{\alpha}}} & \frac{\bar{u} + Z_q}{\bar{u} - Z_{\dot{\alpha}}} & -\frac{g}{\bar{u} - Z_{\dot{\alpha}}} \sin \bar{\theta} \\ M_u + M_{\dot{\alpha}} \frac{Z_u}{\bar{u} - Z_{\dot{\alpha}}} & M_\alpha + M_{\dot{\alpha}} \frac{Z_\alpha}{\bar{u} - Z_{\dot{\alpha}}} & M_q + M_{\dot{\alpha}} \frac{\bar{u} + Z_q}{\bar{u} - Z_{\dot{\alpha}}} & -M_{\dot{\alpha}} \frac{g}{\bar{u} - Z_{\dot{\alpha}}} \sin \bar{\theta} \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta\alpha \\ \Delta q \\ \Delta\theta \end{bmatrix} + \begin{bmatrix} 0 & X_{\delta_t} \\ \frac{Z_{\delta_e}}{\bar{u} - Z_{\dot{\alpha}}} & \frac{Z_{\delta_t}}{\bar{u} - Z_{\dot{\alpha}}} \\ M_{\delta_e} + M_{\dot{\alpha}} \frac{Z_{\delta_e}}{\bar{u} - Z_{\dot{\alpha}}} & M_{\delta_t} + M_{\dot{\alpha}} \frac{Z_{\delta_t}}{\bar{u} - Z_{\dot{\alpha}}} \end{bmatrix} \begin{bmatrix} \Delta\delta_e \\ \Delta\delta_t \end{bmatrix} \quad (9.168)$$

and one has the **linearized lateral-directional rigid airplane dynamics**

$$\begin{bmatrix} \Delta\dot{\beta} \\ \Delta\dot{p} \\ \Delta\dot{r} \\ \Delta\dot{\phi} \end{bmatrix} = \begin{bmatrix} \frac{Y_\beta}{\bar{u}} & \frac{Y_p}{\bar{u}} & \frac{Y_r}{\bar{u}} - 1 & \frac{g}{\bar{u}} \cos \bar{\theta} \\ L_\beta^* & L_p^* & L_r^* & 0 \\ N_\beta^* & N_p^* & N_r^* & 0 \\ 0 & 1 & \tan \bar{\theta} & 0 \end{bmatrix} \begin{bmatrix} \Delta\beta \\ \Delta p \\ \Delta r \\ \Delta\phi \end{bmatrix} + \begin{bmatrix} 0 & \frac{Y_{\delta_r}}{\bar{u}} \\ L_{\delta_a}^* & L_{\delta_r}^* \\ N_{\delta_a}^* & N_{\delta_r}^* \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta\delta_a \\ \Delta\delta_r \end{bmatrix} \quad (9.169)$$

where

$$L_{\bullet}^* = \frac{L_{\bullet} + N_{\bullet} \frac{I_{xz}}{I_{zz}}}{1 - \frac{I_{xz}^2}{I_{xx} I_{zz}}} \quad (9.170)$$

and

$$N_{\bullet}^* = \frac{N_{\bullet} + L_{\bullet} \frac{I_{xz}}{I_{zz}}}{1 - \frac{I_{xz}^2}{I_{xx} I_{zz}}} \quad (9.171)$$

for  $\bullet = \beta, p, r, \delta_a$ , and  $\delta_r$  due to the coupling of  $L$  and  $N$  for non-zero  $I_{xz}$ . Furthermore, note that if  $I_{xz} = 0$ , then  $L_{\bullet}^* = L_{\bullet}$  and  $N_{\bullet}^* = N_{\bullet}$ .

First, note that this derivation results in two sets of fourth-order LTI state-space systems. Secondly, it should be noted that one may also form these EOMs using the  $v$  and  $w$  states instead of  $\beta$  and  $\alpha$ , respectively. Third, it should also be noted that often these EOMs are extended to fifth-order systems by including the perturbed altitude,  $\Delta h$ , as a perturbed longitudinal state and the perturbed yaw,  $\Delta\psi$ , as a perturbed lateral-directional state which are simply related to the other states for wings-level, coordinated steady-flight by

$$\Delta\dot{h} = \sin\bar{\theta}\Delta u + \bar{u}\cos\bar{\theta}(\Delta\theta - \Delta\alpha) \quad (9.172)$$

and

$$\Delta\dot{\psi} = \sec\bar{\theta}\Delta r \quad (9.173)$$

and do not have any direct control derivatives. Lastly, note that these derived state-space systems do not have an explicit output equation whose form will depend on which states should be used as “outputs,” e.g. in a control system.

Another important output that is sometimes used in the output equation is the acceleration vector at some position along the rigid airplane,  $\vec{p}_S = [x_p \ y_p \ z_p]$  in stability frame coordinates, for which, in general

$$\vec{a}_p = \dot{\vec{v}}_S + [\vec{\omega}_{S/N}]_{\times} \vec{v}_S + [\vec{\alpha}_{S/N}]_{\times} \vec{p}_S + [\vec{\omega}_{S/N}]_{\times} [\vec{\omega}_{S/N}]_{\times} \vec{p}_S \quad (9.174)$$

However, after linearizing and for  $\bar{\beta} = 0$  and  $\bar{\phi} = 0$ , one has

$$\begin{bmatrix} \Delta a_{p,x} \\ \Delta a_{p,y} \\ \Delta a_{p,z} \end{bmatrix} = \begin{bmatrix} \Delta\dot{u} + \bar{w}\Delta q + z_p\Delta\dot{q} - y_p\Delta\dot{r} \\ \Delta\dot{v} + \bar{u}\Delta r - \bar{w}\Delta p + x_p\Delta\dot{r} - z_p\Delta\dot{p} \\ \Delta\dot{w} - \bar{u}\Delta q + y_p\Delta\dot{p} - x_p\Delta\dot{q} \end{bmatrix} \quad (9.175)$$

which can be used for the output equation for an accelerometer placed at  $\vec{p}$ .

Lastly,  $X, Y, Z, L_{roll}, M$ , and  $N$  can be written with aerodynamic coefficients for airplanes as

$$X = \frac{Q_{\infty} S_w}{m} C_X \quad (9.176)$$

$$Y = \frac{Q_{\infty} S_w}{m} C_Y \quad (9.177)$$

$$Z = \frac{Q_{\infty} S_w}{m} C_Z \quad (9.178)$$

$$L_{roll} = \frac{Q_{\infty} S_w b_w}{I_{xx}} C_l \quad (9.179)$$

$$M = \frac{Q_\infty S_w \bar{c}_w}{I_{yy}} C_m \quad (9.180)$$

and

$$N = \frac{Q_\infty S_w b_w}{I_{zz}} C_n \quad (9.181)$$

where lowercase letters are used for  $L$ ,  $M$ , and  $N$  coefficients as  $C_L$  is already used for the lift coefficient.

All the aerodynamic coefficients in this section can be modeled using aircraft system identification (SID) via analytical equations, e.g. the **component build-up model** for conventional airplanes presented in appendix A, or estimated using wind tunnel and/or flight test data. Thus, the overall aircraft aerodynamic forces and moments are not simple to compute for arbitrary flight conditions as the aerodynamic coefficients are generally functions of the local airspeed, angle of attack, and sideslip angle at each lifting surface, their geometric layout, as well as the control inputs, and the linear and angular velocities of the vehicle. Thus, one typically linearizes the rigid aircraft dynamics about equilibrium flight conditions in order to analyze its response characteristics and design suitable control systems.

### Coupled Rigid Airplane Linearized Dynamics

For non-coordinated, non-wings-level steady-flight conditions, including turning flight, the linearized longitudinal and lateral-directional state-space models cannot be decoupled, but become coupled. Often this coupling is weak and can be ignored in the control design. However, under certain trim conditions, e.g., high angles of attack or sideslip, the fully coupled equations are necessary in order to assess the coupling effects of the control inputs on all the airplane states. This coupling also typically uses coupled stability and control derivatives, e.g.  $X_\beta$ ,  $Y_\alpha$ ,  $L_\alpha^*$ ,  $N_\alpha^*$ , and  $X_{\delta_r}$ . This section presents the coupled airplane linearized dynamics using the **polynomial-matrix model** for more clarity in the expressions, i.e.,

$$P(s) \vec{y}(s) = Q(s) \vec{u}(s) \quad (9.182)$$

where the system transfer function matrix is given by

$$[G(s)] = P^{-1}(s)Q(s) \quad (9.183)$$

For the coupled polynomial-matrix model, one has

$$\begin{bmatrix} P_{long}(s) & P_{long-lat}(s) \\ P_{lat-long}(s) & P_{lat}(s) \end{bmatrix} \begin{bmatrix} \vec{y}_{long}(s) \\ \vec{y}_{lat}(s) \end{bmatrix} = \begin{bmatrix} Q_{long}(s) & Q_{long-lat}(s) \\ Q_{lat-long}(s) & Q_{lat}(s) \end{bmatrix} \begin{bmatrix} \vec{u}_{long}(s) \\ \vec{u}_{lat}(s) \end{bmatrix} \quad (9.184)$$

where the output vectors are

$$\vec{y}_{long}(s) = \begin{bmatrix} \Delta u(s) \\ \Delta \alpha(s) \\ \Delta q(s) \\ \Delta \theta(s) \end{bmatrix} \quad (9.185)$$

$$\vec{y}_{lat}(s) = \begin{bmatrix} \Delta \beta(s) \\ \Delta p(s) \\ \Delta r(s) \\ \Delta \phi(s) \\ \Delta \psi(s) \end{bmatrix} \quad (9.186)$$

the input vectors are

$$\vec{u}_{long}(s) = \begin{bmatrix} \Delta\delta_e(s) \\ \Delta\delta_T(s) \end{bmatrix} \quad (9.187)$$

$$\vec{u}_{lat}(s) = \begin{bmatrix} \Delta\delta_a(s) \\ \Delta\delta_r(s) \end{bmatrix} \quad (9.188)$$

the diagonal polynomial-matrices are

$$P_{long}(s) = \begin{bmatrix} s - X_u & -X_\alpha + \bar{q}\bar{u} & \bar{w} - X_q & g \cos \bar{\theta} \\ -Z_u - \bar{q} & (\bar{u} - Z_{\dot{\alpha}})s - Z_\alpha & -Z_q - \bar{u} & g \sin \bar{\theta} \cos \bar{\phi} \\ -M_u & -M_{\dot{\alpha}}s - M_\alpha & s - M_q & 0 \\ 0 & 0 & -\cos \bar{\phi} & s \end{bmatrix} \quad (9.189)$$

$$Q_{long}(s) = \begin{bmatrix} X_{\delta_e} & X_{\delta_T} \\ Z_{\delta_e} & Z_{\delta_T} \\ M_{\delta_e} & M_{\delta_T} \\ 0 & 0 \end{bmatrix} \quad (9.190)$$

$$P_{lat}(s) = \begin{bmatrix} \bar{u}s - Y_\beta & -Y_p - \bar{w} & \bar{u} - Y_r & -g \cos \bar{\theta} \cos \bar{\phi} & 0 \\ -L_\beta^* & s - L_p^* - C_1\bar{q} & -L_r^* - C_2\bar{q} & 0 & 0 \\ -N_\beta^* & -N_p^* - C_3\bar{q} & s - N_r^* + C_1\bar{q} & 0 & 0 \\ 0 & 1 & \tan \bar{\theta} \cos \bar{\phi} & -s + (\bar{q} \cos \bar{\phi} - \bar{r} \sin \bar{\phi}) \tan \bar{\theta} & 0 \\ 0 & 0 & \cos \bar{\phi} & \bar{q} \cos \bar{\phi} - \bar{r} \sin \bar{\phi} & -s \cos \bar{\phi} \end{bmatrix} \quad (9.191)$$

where

$$L_\bullet^* = \left( L_\bullet + N_\bullet \frac{I_{xz}}{I_{zz}} \right) D_{xz} \quad (9.192)$$

$$N_\bullet^* = \left( N_\bullet + L_\bullet \frac{I_{xz}}{I_{zz}} \right) D_{xz} \quad (9.193)$$

$$D_{xz} = \left( 1 - \frac{I_{xz}^2}{I_{xx} I_{zz}} \right)^{-1} \quad (9.194)$$

for  $\bullet = \beta, p, r, \delta_a$ , and  $\delta_r$  due to the coupling of  $L$  and  $N$  for non-zero  $I_{xz}$ , and

$$C_1 = (I_{xx} - I_{yy} + I_{zz}) I_{xz} I_{xx}^{-1} I_{zz}^{-1} D_{xz} \quad (9.195)$$

$$C_2 = \left( I_{yy} - I_{zz} + \frac{I_{xz}^2}{I_{zz}} \right) I_{xx}^{-1} D_{xz} \quad (9.196)$$

$$C_3 = \left( I_{xx} - I_{yy} + \frac{I_{xz}^2}{I_{xx}} \right) I_{zz}^{-1} D_{xz} \quad (9.197)$$

$$Q_{lat}(s) = \begin{bmatrix} Y_{\delta_a} & Y_{\delta_r} \\ L_{\delta_a}^* & L_{\delta_r}^* \\ N_{\delta_a}^* & N_{\delta_r}^* \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad (9.198)$$

and the off-diagonal polynomial-matrices are

$$P_{long-lat}(s) = \begin{bmatrix} -\bar{r}\bar{u} - X_\beta & 0 & -\bar{\beta}\bar{u} & 0 & 0 \\ \bar{p}\bar{u} & \bar{\beta}\bar{u} & 0 & g \cos \bar{\theta} \sin \bar{\phi} & 0 \\ 0 & (I_{xx} - I_{zz}) I_{yy}^{-1} \bar{r} + 2I_{xz} I_{yy}^{-1} \bar{p} & (I_{xx} - I_{zz}) I_{yy}^{-1} \bar{p} - 2I_{xz} I_{yy}^{-1} \bar{r} & 0 & 0 \\ 0 & 0 & \sin \bar{\phi} & \dot{\psi} \cos \bar{\theta} & 0 \end{bmatrix} \quad (9.199)$$

$$Q_{long-lat}(s) = \begin{bmatrix} 0 & X_{\delta_r} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad (9.200)$$

$$P_{lat-long}(s) = \begin{bmatrix} \bar{r} & \bar{p}\bar{u} - Y_\alpha & 0 & g \sin \bar{\theta} \sin \bar{\phi} \\ 0 & -L_\alpha^* & I_{zz}^{-1} D_{xz} C_4 + I_{xz} I_{xx}^{-1} I_{zz}^{-1} D_{xz} C_5 & 0 \\ 0 & -N_\alpha^* & I_{xx}^{-1} D_{xz} C_5 + I_{xz} I_{xx}^{-1} I_{zz}^{-1} D_{xz} C_4 & 0 \\ 0 & 0 & \tan \bar{\theta} \sin \bar{\phi} & \bar{q} \sin \bar{\phi} + \bar{r} \cos \bar{\phi} + \dot{\psi} \sin \bar{\theta} \tan \bar{\theta} \\ 0 & 0 & \sin \bar{\phi} & \dot{\psi} \sin \bar{\theta} \end{bmatrix} \quad (9.201)$$

and

$$C_4 = (I_{zz} - I_{yy}) \bar{r} - I_{xz} \bar{p} \quad (9.202)$$

$$C_5 = (I_{yy} - I_{xx}) \bar{p} + I_{xz} \bar{r} \quad (9.203)$$

$$Q_{lat-long}(s) = 0_{5 \times 2} \quad (9.204)$$

### Longitudinal Modes and Stability

The fourth-order linearized longitudinal dynamics for rigid airplanes in straight-and-level flight as

$$\begin{bmatrix} \Delta \bar{u} \\ \Delta \dot{\alpha} \\ \Delta \dot{q} \\ \Delta \dot{\theta} \end{bmatrix} = \begin{bmatrix} X_u & X_\alpha & 0 & -g \\ \frac{Z_u}{\bar{u} - Z_\alpha} & \frac{Z_\alpha}{\bar{u} - Z_\alpha} & \frac{\bar{u} + Z_q}{\bar{u} - Z_\alpha} & 0 \\ M_u + M_\alpha \frac{Z_u}{\bar{u} - Z_\alpha} & M_\alpha + M_\dot{\alpha} \frac{Z_\alpha}{\bar{u} - Z_\alpha} & M_q + M_\dot{\alpha} \frac{\bar{u} + Z_q}{\bar{u} - Z_\alpha} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \\ \Delta \theta \end{bmatrix} + \begin{bmatrix} 0 & X_{\delta_T} \\ \frac{Z_{\delta_e}}{\bar{u} - Z_\alpha} & \frac{Z_{\delta_T}}{\bar{u} - Z_\alpha} \\ M_{\delta_e} + M_\alpha \frac{Z_{\delta_e}}{\bar{u} - Z_\alpha} & M_{\delta_T} + M_\alpha \frac{Z_{\delta_T}}{\bar{u} - Z_\alpha} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \delta_e \\ \Delta \delta_T \end{bmatrix} \quad (9.205)$$

where  $\Delta w \approx \bar{u} \Delta \alpha$  may also be substituted with

$$M_\alpha = \bar{u} M_w, \quad Z_\alpha = \bar{u} Z_w, \quad M_\dot{\alpha} = \bar{u} M_{\dot{w}} \quad (9.206)$$

There are conventionally two oscillatory modes that dominate the response in the longitudinal plane: the **short-period mode** and the **phugoid mode** derived from the Greek words, *phuge* and *eidos*, for “flight-like.” The phugoid mode is also known as the **long-period mode** as it is typically oscillatory though it may separate into two real modes for some airplanes. Recall from Lyapunov stability theory, an airplane is **longitudinally stable** if and only if both modes are stable which occurs when the real part of the eigenvalues of the state matrix  $A$  are negative, i.e. the modes are in the left-half plane (LHP) of the complex plane.

The short-period mode can be approximated by assuming  $\Delta u = 0$ . Then, one has the second-order **short-period mode approximation** as the characteristic polynomial

$$\phi_{S-P}(\lambda) = \lambda^2 + \left( -\frac{Z_\alpha}{\bar{u} - Z_{\dot{\alpha}}} - M_q - M_{\dot{\alpha}} \frac{\bar{u} + Z_q}{\bar{u} - Z_{\dot{\alpha}}} \right) \lambda + \left( M_q \frac{Z_\alpha}{\bar{u} - Z_{\dot{\alpha}}} - M_\alpha \frac{\bar{u} + Z_q}{\bar{u} - Z_{\dot{\alpha}}} \right) \quad (9.207)$$

The short-period mode primarily affects  $\Delta\alpha$  and the short-term behavior of  $\Delta\theta$ . For airplane design, one typically balances a desire for a high natural frequency, i.e. a quick response to input, but also heavy damping, i.e. little overshoot. Increasing the short-period damping is done through a SAS called a **pitch damper**.

The phugoid mode can be approximated by assuming  $\Delta\dot{\alpha} = \Delta\dot{q} = 0$ . Then, one has the second-order **phugoid mode approximation** as the characteristic polynomial

$$\phi_{L-P}(\lambda) = \lambda^2 + \left( -X_u + X_\alpha \frac{Z_u M_q - M_u (\bar{u} + Z_q)}{Z_\alpha M_q - M_\alpha (\bar{u} + Z_q)} \right) \lambda + \left( g \frac{Z_u M_\alpha - Z_\alpha M_u}{Z_\alpha M_q - M_\alpha (\bar{u} + Z_q)} \right) \quad (9.208)$$

The phugoid mode primarily affects  $\Delta u$  and the long-term behavior of  $\Delta\theta$  or  $\Delta h$ . Because the phugoid mode is typically highly oscillatory, most applications require it to be corrected during flight, often manually, but these corrections can be fatiguing for pilots if the damping ratio is too low. In this case, one can use a SAS using pitch feedback. Lastly, it should be noted that phugoid mode can be considered as the gradual interchange of kinetic and potential energy through the varying velocity and altitude, respectively.

## Lateral-Directional Modes and Stability

The fourth-order linearized lateral-directional dynamics for rigid airplanes in straight-and-level flight as

$$\begin{bmatrix} \Delta\dot{\beta} \\ \Delta\dot{p} \\ \Delta\dot{r} \\ \Delta\dot{\phi} \end{bmatrix} = \begin{bmatrix} \frac{Y_\beta}{\bar{u}} & \frac{Y_p}{\bar{u}} & \frac{Y_r}{\bar{u}} - 1 & \frac{g}{\bar{u}} \cos \bar{\theta} \\ L_\beta^* & L_p^* & L_r^* & 0 \\ N_\beta^* & N_p^* & N_r^* & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta\beta \\ \Delta p \\ \Delta r \\ \Delta\phi \end{bmatrix} + \begin{bmatrix} 0 & \frac{Y_{\delta_r}}{\bar{u}} \\ L_{\delta_a}^* & L_{\delta_r}^* \\ N_{\delta_a}^* & N_{\delta_r}^* \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta\delta_a \\ \Delta\delta_r \end{bmatrix} \quad (9.209)$$

where  $\Delta v \approx \bar{u}\Delta\beta$  may also be substituted with

$$Y_\beta = \bar{u}Y_v, \quad L_\beta^* = \bar{u}L_v^*, \quad N_\beta^* = \bar{u}N_v^* \quad (9.210)$$

There are conventionally two exponentially decaying modes and one oscillatory mode that dominate the responses in the lateral-directional planes: the **roll mode**, the **spiral mode**, and the **dutch-roll mode**.

However, for some airplanes the **roll-spiral mode** may combine into a single oscillatory mode. The motion corresponding to the dutch-roll mode was named for its side-to-side motion which alludes to dutch figure skaters using repetitive right-and-left skating on the outer edge of their skates to maintain speed. Recall from Lyapunov stability theory, an airplane is **lateral-directionally stable** if and only if all three modes are stable which occurs when the real part of the eigenvalues of the state matrix  $A$  are negative, i.e. the modes are in the left-half plane (LHP) of the complex plane.

The roll mode can be approximated by assuming a pure rolling equation of motion. Then, one has the first-order **roll mode approximation** as the characteristic polynomial

$$\phi_R(\lambda) = \lambda - L_p^* \quad (9.211)$$

The roll mode primarily affects the short-term behavior of  $p$ . The value of  $L_p$  depends primarily on the size of the wing and tail surfaces. Typically this mode decays rapidly compared to the other lateral-directional modes, albeit sometimes too quickly. Reducing the roll mode time constant is done through a SAS called a **roll damper**.

The dutch-roll mode can be approximated by assuming  $Y_p = Y_r = L_r^* = N_p^* = g/\bar{u} = 0$ . Then, one has the second-order **dutch-roll mode approximation** as the characteristic polynomial

$$\phi_{D-R}(\lambda) = \lambda^2 + \left( -N_r^* - \frac{Y_\beta}{\bar{u}} \right) \lambda + \left( N_\beta^* + N_r^* \frac{Y_\beta}{\bar{u}} \right) \quad (9.212)$$

The dutch-roll mode primarily affects  $\Delta\beta$ , the long-term behavior of  $p$ , and the short-term behavior of  $\Delta\phi$ . For airplane design, one typically balances a desire for a high natural frequency, i.e. a quick response to input, but also heavy damping, i.e. little overshoot. Increasing the dutch-roll damping is done through a SAS called a **yaw damper**. In addition, a roll damper can be used in parallel to increase the effectiveness of the yaw damper.

The spiral mode can be approximated by assuming  $\dot{\beta} = \dot{p} = \dot{r} = Y_p = Y_r = 0$ . Then, one has the first-order **spiral mode approximation** as

$$\phi_S(\lambda) = \lambda + g \frac{L_\beta^* N_r^* - L_r^* N_\beta^*}{Y_\beta (L_r^* N_p^* - N_r^* L_p^*) + \bar{u} (L_\beta^* N_p^* - N_\beta^* L_p^*)} \quad (9.213)$$

The spiral mode primarily affects the long-term behavior of  $\Delta\phi$ . Typically this mode decays slowly compared to the other lateral-directional modes and may even be unstable, requiring the pilot to make corrections or a SAS added to the control system.

Lastly, it should be noted that airplane design characteristics affect the spiral and dutch-roll modes in opposite ways. In particular, increasing the dihedral effect makes the dutch-roll mode less stable and the spiral mode more stable. Conversely, increasing the directional stability makes the dutch-roll mode more stable and the spiral mode less stable. Thus, often one must use roll/yaw dampers to sufficiently stabilize both the spiral and dutch roll modes.

## Flying Qualities

When designing aircraft, the **flying qualities**, also known as the **handling qualities**, i.e. the handling of the aircraft by a pilot, are closely related to the modal characteristics of the aircraft. However, as pilots are generally charged with performing various tasks or **missions**, the flying qualities are generally specified by airplane pilots according to the following three subjective levels:

- **Level 1** (Good): Flying qualities clearly adequate for the mission flight phase.
- **Level 2** (Acceptable): Flying qualities adequate to accomplish the mission flight phase, but with some increase in pilot workload and/or degradation in mission effectiveness or both.
- **Level 3** (Poor): Flying qualities such that the airplane can be controlled safely, but pilot workload is excessive and/or mission effectiveness is inadequate or both.

which also depend on the three generalized flight phase categories:

- **Category A**: nonterminal flight phases that require rapid maneuvering, precision tracking, or highly accurate flight-path control.
- **Category B**: nonterminal flight phases that are normally accomplished using gradual maneuvers and without precision tracking, although accurate flight-path control may be required.
- **Category C**: terminal flight phases that are normally accomplished using gradual maneuvers and usually require accurate flight-path control.

Notably, for Level 3 flying qualities, Category A flight phases can be terminated safely and Category B and C flight phases can be completed.

Lastly, another important part of assessing these flying qualities is by class of aircraft. For airplanes these are generally classified as

- **Class I**: Small, light airplanes
- **Class II**: Medium-weight, low-to-medium maneuverability airplanes
- **Class III**: Large, heavy, low-to-medium maneuverability airplanes
- **Class IV**: High-maneuverability airplanes

The following table of modal characteristics for different flying quality levels and flight phase categories for different classes of airplanes.

Mode	Level	Category	Class	Characteristic(s)
Phugoid	1	All	All	$\zeta > 0.04$
	2	All	All	$\zeta > 0$
	3	All	All	$T > 55s$
Short-Period	1	A and C	All	$0.35 \leq \zeta \leq 1.3$
		B	All	$0.3 \leq \zeta \leq 2.0$
	2	A and C	All	$0.25 \leq \zeta \leq 2.0$
		B	All	$0.2 \leq \zeta \leq 2.0$
	3	All	All	$0.15 \leq \zeta$
Roll	1	A and C	I, IV II, III	$\tau \leq 1.0$ sec $\tau \leq 1.4$ sec
		B	All	$\tau \leq 1.4$ sec
	2	A and C	I, IV II, III	$\tau \leq 1.4$ sec $\tau \leq 3.0$ sec
		B	All	$\tau \leq 3.0$ sec
	3	All	All	$\tau \leq 10$ sec
Spiral	1	A	I, IV II, III	Doubling amplitude $\geq 12$ sec Doubling amplitude $\geq 20$ sec
		B and C	All	Doubling amplitude $\geq 20$ sec
	2	All	All	Doubling amplitude $\geq 12$ sec
	3	All	All	Doubling amplitude $\geq 4$ sec
Dutch-Roll	1	A	I, IV II, III	$\zeta\omega_n \geq 0.35$ rad/s, $\zeta \geq 0.19$ , $\omega_n > 1.0$ rad/s $\zeta\omega_n \geq 0.35$ rad/s, $\zeta \geq 0.19$ , $\omega_n > 0.4$ rad/s
		B	All	$\zeta\omega_n \geq 0.15$ rad/s, $\zeta \geq 0.08$ , $\omega_n > 0.4$ rad/s
		C	I, II-C, IV II-L, III	$\zeta\omega_n \geq 0.15$ rad/s, $\zeta \geq 0.08$ , $\omega_n > 1.0$ rad/s $\zeta\omega_n \geq 0.15$ rad/s, $\zeta \geq 0.08$ , $\omega_n > 0.4$ rad/s
	2	All	All	$\zeta\omega_n \geq 0.05$ rad/s, $\zeta \geq 0.02$ , $\omega_n > 0.4$ rad/s
	3	All	All	$\zeta \geq 0.02$ , $\omega_n \geq 0.4$ rad/s

where C and L denote carrier- or land-based airplanes. Note that the spiral mode requirements are for the doubling of amplitude for a potentially *unstable* spiral mode.

Lastly, it is important to note that the **Cooper-Harper Rating Scale (CHRS)** is another standard used by pilots and flight test engineers to rate the flying qualities of the airplane design as a whole. The CHRS scale goes from 1 to 10 with lower numbers corresponding to better flying qualities. The description for each rating is shown in the following table.

Pilot Rating	Aircraft Characteristic	Demand of Pilot	Overall Assessment
1	Excellent, highly desirable	Pilot compensation not a factor for desired performance	Good
2	Good, negligible deficiencies	Pilot compensation not a factor for desired performance	Good
3	Fair, some mildly unpleasant deficiencies	Minimal pilot compensation required for desired performance	Good
4	Minor, but annoying deficiencies	Desired performance requires moderate pilot compensation	Acceptable
5	Moderately objectionable deficiencies	Adequate performance requires considerable pilot compensation	Acceptable
6	Very objectionable, but tolerable deficiencies	Adequate performance requires extensive pilot compensation	Acceptable
7	Major deficiencies	Adequate performance not attainable with maximum tolerable compensation	Poor
8	Major deficiencies	Considerable pilot compensation is required for control	Poor
9	Major deficiencies	Intense pilot compensation is required for control	Poor
10	Major deficiencies	Control will be lost during some portion of required operation	Unacceptable

## References

For more information, please refer to the following

- Nelson, R. C., “3.5 Small-Disturbance Theory,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 104-108
- Nelson, R. C., “4.4 Stick-Fixed Longitudinal Motion,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 147-151
- Nelson, R. C., “4.5 Longitudinal Approximations,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 152-162
- Nelson, R. C., “4.7 Flying Qualities,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 164-168
- Nelson, R. C., “5.4 Lateral-Directional Equations of Motion,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 193-203
- Nelson, R. C., “5.5 Lateral Flying Qualities,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 203-204

- Schmidt, D. K., “10.2 Linear Flight-Dynamics Perturbation,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 481-485
- Schmidt, D. K., “10.3 Decoupled Longitudinal and Lateral-Directional Linear Models,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 500-523
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “2.6 Linear Models and the Stability Derivatives,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 116-137
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “4.2 Aircraft Rigid-Body Modes,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 257-274
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “4.3 The Handling-Qualities Requirements,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 274-287

## 9.6 Elastic Airplane Dynamics

### Dynamic-Elastic Airplane Effects

Previous sections have provided the background for developing the elastic vibration equations of motion (EOMs) alongside the rigid-body EOMs. These showed that the vibration modes have their own set of equations of motion and enter the rigid body EOMs only through elastic effects on the aerodynamic and propulsive forces and moments. These considerations can be studied as **dynamic-elastic effects**, which models the elastic effects on these forces and moments in the rigid body EOMs and the vibration dynamics, or the **static-elastic effects** which *only* consider the elastic deformation effects on the forces and moments in the rigid body EOMs and not do include the vibration dynamics. This section will consider the dynamic-elastic effects modeling for both the nonlinear forms of airplane dynamics while the subsequent section will discuss the linearized dynamic-elastic effects and the static-elastic effects modeling for airplane dynamics.

To this end, recall the rigid airplane equations of motion

$$\begin{bmatrix} X - g \sin \theta \\ Y + g \cos \theta \sin \phi \\ Z + g \cos \theta \cos \phi \\ L \\ M \\ N \end{bmatrix} = \begin{bmatrix} \dot{u} + qw - rv \\ \dot{v} + ru - pw \\ \dot{w} + pv - qu \\ \dot{p} + \frac{I_{zz} - I_{yy}}{I_{xx}} qr - \frac{I_{xz}}{I_{xx}} (\dot{r} + pq) \\ \dot{q} + \frac{I_{xx} - I_{zz}}{I_{yy}} pr - \frac{I_{xz}}{I_{yy}} (r^2 - p^2) \\ \dot{r} + \frac{I_{yy} - I_{xx}}{I_{zz}} pq - \frac{I_{xz}}{I_{zz}} (\dot{p} - qr) \end{bmatrix} \quad (9.214)$$

where the body frame forces can alternatively be written using the thrust,  $\vec{T}$ , and the wind frame aerodynamic

forces: lift  $L$ , side  $S$ , and drag  $D$ , as

$$\begin{aligned} m \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} &= \vec{T} + C_{B \leftarrow W}(\alpha, \beta) \begin{bmatrix} -D \\ S \\ -L \end{bmatrix} \\ m \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} &= \vec{T} + \begin{bmatrix} \cos \alpha \cos \beta & -\cos \alpha \sin \beta & -\sin \alpha \\ \sin \beta & \cos \beta & 0 \\ \sin \alpha \cos \beta & -\sin \alpha \sin \beta & \cos \alpha \end{bmatrix} \begin{bmatrix} -D \\ S \\ -L \end{bmatrix} \end{aligned} \quad (9.215)$$

where the dynamic-elastic effects can be alternatively applied to  $L$ ,  $S$ , and  $D$  instead of  $X$ ,  $Y$ ,  $Z$ . In addition, recall that one can model these aerodynamic and propulsive forces using stability and control derivatives/coefficients (where  $\delta_T = T$ ). The same approach can be used to add coefficients for each modal coordinate and modal coordinate rate, i.e.

$$\begin{aligned} \begin{bmatrix} X \\ Z \\ M \end{bmatrix} &= \begin{bmatrix} X_0 \\ Z_0 \\ M_0 \end{bmatrix} + \begin{bmatrix} 0 & X_{\dot{\alpha}} & 0 \\ 0 & Z_{\dot{\alpha}} & 0 \\ 0 & M_{\dot{\alpha}} & 0 \end{bmatrix} \begin{bmatrix} \dot{u} \\ \dot{\alpha} \\ \dot{q} \end{bmatrix} + \begin{bmatrix} X_u & X_\alpha & X_q \\ Z_u & Z_\alpha & Z_q \\ M_u & M_\alpha & M_q \end{bmatrix} \begin{bmatrix} u \\ \alpha \\ q \end{bmatrix} + \begin{bmatrix} X_{\delta_e} & X_{\delta_t} \\ Z_{\delta_e} & Z_{\delta_t} \\ M_{\delta_e} & M_{\delta_t} \end{bmatrix} \begin{bmatrix} \delta_e \\ \delta_t \end{bmatrix} \\ &+ \begin{bmatrix} X_{\dot{\eta}_1} & \dots & X_{\dot{\eta}_1} \\ Z_{\dot{\eta}_1} & \dots & Z_{\dot{\eta}_1} \\ M_{\dot{\eta}_1} & \dots & M_{\dot{\eta}_1} \end{bmatrix} \begin{bmatrix} \dot{\eta}_1 \\ \vdots \\ \dot{\eta}_n \end{bmatrix} + \begin{bmatrix} X_{\eta_1} & \dots & X_{\eta_1} \\ Z_{\eta_1} & \dots & Z_{\eta_1} \\ M_{\eta_1} & \dots & M_{\eta_1} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \end{bmatrix} \end{aligned} \quad (9.216)$$

and

$$\begin{aligned} \begin{bmatrix} Y \\ L \\ N \end{bmatrix} &= \begin{bmatrix} Y_0 \\ L_0 \\ N_0 \end{bmatrix} + \begin{bmatrix} Y_\beta & Y_p & Y_r \\ L_\beta & L_p & L_r \\ N_\beta & N_p & N_r \end{bmatrix} \begin{bmatrix} \beta \\ p \\ r \end{bmatrix} + \begin{bmatrix} Y_{\delta_a} & Y_{\delta_r} \\ L_{\delta_a} & L_{\delta_r} \\ N_{\delta_a} & N_{\delta_r} \end{bmatrix} \begin{bmatrix} \delta_a \\ \delta_r \end{bmatrix} \\ &+ \begin{bmatrix} Y_{\dot{\eta}_1} & \dots & Y_{\dot{\eta}_1} \\ L_{\dot{\eta}_1} & \dots & L_{\dot{\eta}_1} \\ N_{\dot{\eta}_1} & \dots & N_{\dot{\eta}_1} \end{bmatrix} \begin{bmatrix} \dot{\eta}_1 \\ \vdots \\ \dot{\eta}_n \end{bmatrix} + \begin{bmatrix} Y_{\eta_1} & \dots & Y_{\eta_1} \\ L_{\eta_1} & \dots & L_{\eta_1} \\ N_{\eta_1} & \dots & N_{\eta_1} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \end{bmatrix} \end{aligned} \quad (9.217)$$

Furthermore, the conversions from coefficient to derivative can be shown to be as follows. Recall that  $Q_\infty = \frac{1}{2}\rho_\infty \bar{v}_\infty^2$  is the dynamic pressure with trimmed airspeed  $\bar{v}_\infty = \sqrt{u^2 + v^2 + w^2}$ .

$\bullet$	$X_\bullet$	$Z_\bullet$	$M_\bullet$
$u$	$\frac{Q_\infty S_w}{m \bar{v}_\infty} C_{X_u}$	$\frac{Q_\infty S_w}{m \bar{v}_\infty} C_{Z_u}$	$\frac{Q_\infty S_w \bar{c}_w}{I_{yy} \bar{v}_\infty} C_{m_u}$
$\alpha$	$\frac{Q_\infty S_w}{m} C_{X_\alpha}$	$\frac{Q_\infty S_w}{m} C_{Z_\alpha}$	$\frac{Q_\infty S_w \bar{c}_w}{I_{yy}} C_{m_\alpha}$
$q$	$\frac{Q_\infty S_w \bar{c}_w}{2m \bar{v}_\infty} C_{Z_q}$	$\frac{Q_\infty S_w \bar{c}_w}{2m \bar{v}_\infty} C_{Z_q}$	$\frac{Q_\infty S_w \bar{c}_w^2}{2I_{yy} \bar{v}_\infty} C_{m_q}$

$\dot{\alpha}$	$\frac{Q_\infty S_w \bar{c}_w}{2m\bar{v}_\infty} C_{X_{\dot{\alpha}}}$	$\frac{Q_\infty S_w \bar{c}_w}{2m\bar{v}_\infty} C_{Z_{\dot{\alpha}}}$	$\frac{Q_\infty S_w \bar{c}_w^2}{2I_{yy}\bar{v}_\infty} C_{m_{\dot{\alpha}}}$
$\delta_e$	$\frac{Q_\infty S_w}{m} C_{X_{\delta_e}}$	$\frac{Q_\infty S_w}{m} C_{Z_{\delta_e}}$	$\frac{Q_\infty S_w \bar{c}_w}{I_{yy}} C_{m_{\delta_e}}$
$\delta_t$	$\frac{Q_\infty S_w}{m} C_{X_{\delta_t}}$	$\frac{Q_\infty S_w}{m} C_{Z_{\delta_t}}$	$\frac{Q_\infty S_w \bar{c}_w}{I_{yy}} C_{m_{\delta_t}}$
$\eta_i$	$\frac{Q_\infty S_w}{m} C_{X_{\eta_i}}$	$\frac{Q_\infty S_w}{m} C_{Z_{\eta_i}}$	$\frac{Q_\infty S_w \bar{c}_w}{I_{yy}} C_{m_{\eta_i}}$
$\dot{\eta}_i$	$\frac{Q_\infty S_w}{m\bar{v}_\infty} C_{X_{\dot{\eta}_i}}$	$\frac{Q_\infty S_w}{m\bar{v}_\infty} C_{Z_{\dot{\eta}_i}}$	$\frac{Q_\infty S_w \bar{c}_w}{I_{yy}\bar{v}_\infty} C_{m_{\dot{\eta}_i}}$

$\bullet$	$Y_\bullet$	$L_\bullet$	$N_\bullet$
$\beta$	$\frac{Q_\infty S_w}{m} C_{Y_\beta}$	$\frac{Q_\infty S_w b_w}{I_{xx}} C_{l_\beta}$	$\frac{Q_\infty S_w b_w}{I_{zz}} C_{n_\beta}$
$p$	$\frac{Q_\infty S_w b_w}{2m\bar{v}_\infty} C_{Y_p}$	$\frac{Q_\infty S_w b_w^2}{2I_{xx}\bar{v}_\infty} C_{l_p}$	$\frac{Q_\infty S_w b_w^2}{2I_{zz}\bar{v}_\infty} C_{n_p}$
$r$	$\frac{Q_\infty S_w b_w}{2m\bar{v}_\infty} C_{Y_r}$	$\frac{Q_\infty S_w b_w^2}{2I_{xx}\bar{v}_\infty} C_{l_r}$	$\frac{Q_\infty S_w b_w^2}{2I_{zz}\bar{v}_\infty} C_{n_r}$
$\delta_a$	$\frac{Q_\infty S_w}{m} C_{Y_{\delta_a}}$	$\frac{Q_\infty S_w b_w}{I_{xx}} C_{l_{\delta_a}}$	$\frac{Q_\infty S_w b_w}{I_{zz}} C_{n_{\delta_a}}$
$\delta_r$	$\frac{Q_\infty S_w}{m} C_{Y_{\delta_r}}$	$\frac{Q_\infty S_w b_w}{I_{xx}} C_{l_{\delta_r}}$	$\frac{Q_\infty S_w b_w}{I_{zz}} C_{n_{\delta_r}}$
$\eta_i$	$\frac{Q_\infty S_w}{m} C_{Y_{\eta_i}}$	$\frac{Q_\infty S_w b_w}{I_{xx}} C_{l_{\eta_i}}$	$\frac{Q_\infty S_w b_w}{I_{zz}} C_{n_{\eta_i}}$
$\dot{\eta}_i$	$\frac{Q_\infty S_w}{m\bar{v}_\infty} C_{Y_{\dot{\eta}_i}}$	$\frac{Q_\infty S_w b_w}{I_{xx}\bar{v}_\infty} C_{n_{\dot{\eta}_i}}$	$\frac{Q_\infty S_w b_w}{I_{zz}\bar{v}_\infty} C_{n_{\dot{\eta}_i}}$

Lastly, one must also include the  $n$  vibration LTI ODEs where typically one includes some level of damping for each mode,  $\zeta_i$ , whose value is typically assessed from matching analytical modeling with experimental data (usually quite low, e.g. 0.02). Thus, one has

$$\ddot{\eta}_i + 2\zeta_i \omega_i \dot{\eta}_i + \omega_i^2 \eta_i = \frac{Q_i}{M_i}, \quad i = 1, \dots, n \quad (9.218)$$

where the generalized forces can be modeled as linear equations of the states, i.e.

$$Q_i = Q_{i_0} + [Q_{i_u} \ Q_{i_\beta} \ Q_{i_\alpha} \ Q_{i_p} \ Q_{i_q} \ Q_{i_r}] \begin{bmatrix} u \\ \beta \\ \alpha \\ p \\ q \\ r \end{bmatrix} + [Q_{i_{\delta_a}} \ Q_{i_{\delta_e}} \ Q_{i_{\delta_r}} \ Q_{i_{\delta_t}}] \begin{bmatrix} \delta_a \\ \delta_e \\ \delta_r \\ \delta_t \end{bmatrix} + [Q_{i_{\dot{\eta}_1}} \ \cdots \ Q_{i_{\dot{\eta}_n}}] \begin{bmatrix} \dot{\eta}_1 \\ \vdots \\ \dot{\eta}_n \end{bmatrix} + [Q_{i_{\eta_1}} \ \cdots \ Q_{i_{\eta_n}}] \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \end{bmatrix} \quad (9.219)$$

where

$$Q_{i_\bullet} = Q_\infty S_w \bar{c}_w C_{Q_{i_\bullet}} \quad (9.220)$$

for  $\bullet = \alpha, \beta, \delta_a, \delta_e, \delta_r, \delta_T, \eta_j$  for  $j = 1, \dots, n$ , and

$$Q_{i_\bullet} = \frac{Q_\infty S_w \bar{c}_w}{\bar{v}_\infty} C_{Q_{i_\bullet}} \quad (9.221)$$

for  $\bullet = u, p, q, r, \dot{\eta}_j$  for  $j = 1, \dots, n$ .

### Static-Elastic Airplane Effects

First, recall the elastic flight vehicle equations of motion in state-space form as

$$\begin{aligned} \vec{x}_{rig} &= f_{rig}(\vec{x}_{rig}, \phi, \theta) + \mathcal{A}_{rig \leftarrow rig} \vec{x}_{rig} + [\mathcal{A}_{rig \leftarrow \eta} \ \mathcal{A}_{rig \leftarrow \dot{\eta}}] \vec{x}_{vib} + \mathcal{B}_{rig} \vec{u} \\ \vec{x}_{vib} &= \begin{bmatrix} 0_{n \times 6} \\ A_{vib \leftarrow rig} \end{bmatrix} \vec{x}_{rig} + \begin{bmatrix} 0_{n \times n} & I_{n \times n} \\ A_{vib \leftarrow \eta} & A_{vib \leftarrow \dot{\eta}} \end{bmatrix} \vec{x}_{vib} + \begin{bmatrix} 0_{n \times 4} \\ B_{vib} \end{bmatrix} \vec{u} \end{aligned} \quad (9.222)$$

To consider only the static-elastic effects, i.e., deformation equilibrium, one can set all  $\dot{\eta}_i = \ddot{\eta}_i = 0 \forall i$ . Then, one can solve for the **static-elastic modal coordinates**,

$$\bar{\eta} = [\bar{\eta}_1 \ \cdots \ \bar{\eta}_n] \quad (9.223)$$

in terms of the rigid vehicle state and control inputs.

Using the vibration equation of motion, i.e.,

$$\vec{0} = \begin{bmatrix} 0_{n \times 6} \\ A_{vib \leftarrow rig} \end{bmatrix} \vec{x}_{rig} + \begin{bmatrix} 0_{n \times n} & I_{n \times n} \\ A_{vib \leftarrow \eta} & A_{vib \leftarrow \dot{\eta}} \end{bmatrix} \begin{bmatrix} \bar{\eta} \\ \vec{0} \end{bmatrix} + \begin{bmatrix} 0_{n \times 4} \\ B_{vib} \end{bmatrix} \vec{u} \quad (9.224)$$

for which, the non-trivial portion states:

$$\vec{0} = A_{vib \leftarrow rig} \vec{x}_{rig} + A_{vib \leftarrow \eta} \bar{\eta} + B_{vib} \vec{u} \quad (9.225)$$

or, finally, the **static-elastic constraint**

$$\bar{\eta} = -A_{vib \leftarrow \eta}^{-1} (A_{vib \leftarrow rig} \vec{x}_{rig} + B_{vib} \vec{u}) \quad (9.226)$$

Using the rigid-body equation of motion, i.e.,

$$\dot{\vec{x}}_{rig} = f_{rig}(\vec{x}_{rig}, \phi, \theta) + \mathcal{A}_{rig \leftarrow rig} \vec{x}_{rig} + [\mathcal{A}_{rig \leftarrow \eta} \quad \mathcal{A}_{rig \leftarrow \dot{\eta}}] \begin{bmatrix} \bar{\eta} \\ 0 \end{bmatrix} + \mathcal{B}_{rig} \vec{u} \quad (9.227)$$

Finally, by back-substitution, one has for the **static-elastic rigid vehicle EOMs**

$$\begin{aligned} \dot{\vec{x}}_{rig} = & f_{rig}(\vec{x}_{rig}, \phi, \theta) + \left( \mathcal{A}_{rig \leftarrow rig} - \mathcal{A}_{rig \leftarrow \eta} A_{vib \leftarrow \eta}^{-1} A_{vib \leftarrow rig} \right) \vec{x}_{rig} \\ & + \left( B_{rig} - \mathcal{A}_{rig \leftarrow \eta} \mathcal{A}_{vib \leftarrow \eta}^{-1} B_{vib} \right) \vec{u} \end{aligned} \quad (9.228)$$

which is a process known as **residualization** of the vibration degrees-of-freedom into the new matrices of static-elastic stability and control derivatives/coefficients, i.e., the elements of

$(\mathcal{A}_{rig \leftarrow rig} - \mathcal{A}_{rig \leftarrow \eta} A_{vib \leftarrow \eta}^{-1} A_{vib \leftarrow rig})$  and  $(B_{rig} - \mathcal{A}_{rig \leftarrow \eta} A_{vib \leftarrow \eta}^{-1} B_{vib})$ . It is important to note that, in general, these residualized static-elastic derivatives/coefficients depend on the flight conditions as these directly affect the loads on the vehicle's structure, especially the dynamic pressure. Furthermore, if the aerodynamic forces and moments are not truly linear, then one must use numerical techniques to find the static-elastic modal coordinates.

### Linearized Elastic Airplane Dynamics

As opposed to rigid body modeling, the linearized equations of motion for elastic airplanes typically use the fuselage body frame (subscript  $F$ ) instead of the stability body frame (subscript  $S$ ) for developing the vibration and dynamic-elastic coefficients. Thus, if one has developed the linearized rigid airplane in the stability frame, one must first transform the perturbed rigid body aerodynamic and propulsive forces and moments from the stability frame to the fuselage frame as

$$\begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix}_F = \begin{bmatrix} \cos \bar{\alpha} & 0 & -\sin \bar{\alpha} \\ 0 & 1 & 0 \\ \sin \bar{\alpha} & 0 & \cos \bar{\alpha} \end{bmatrix} \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix}_S \quad (9.229)$$

and

$$\begin{bmatrix} \Delta L \\ \Delta M \\ \Delta N \end{bmatrix}_F = \begin{bmatrix} \cos \bar{\alpha} & 0 & -\sin \bar{\alpha} \\ 0 & 1 & 0 \\ \sin \bar{\alpha} & 0 & \cos \bar{\alpha} \end{bmatrix} \begin{bmatrix} \Delta L \\ \Delta M \\ \Delta N \end{bmatrix}_S \quad (9.230)$$

Having redefined these terms, one may use the linearized equations of motion in the fuselage frame as opposed to the stability frame. However, in this case  $\bar{\alpha}$  may not equal 0, thus, the linearized equations become

$$\begin{aligned} \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} + g \begin{bmatrix} -\cos \bar{\theta} & 0 & 0 \\ -\sin \bar{\theta} \sin \bar{\phi} & \cos \bar{\theta} \cos \bar{\phi} & 0 \\ \sin \bar{\theta} \cos \bar{\phi} & \cos \bar{\theta} \sin \bar{\phi} & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \phi \\ \psi \end{bmatrix} = & \begin{bmatrix} \Delta \dot{u} \\ \Delta \dot{v} \\ \Delta \dot{w} \end{bmatrix} \\ + \begin{bmatrix} 0 & -\bar{r} & \bar{q} \\ \bar{r} & 0 & -\bar{p} \\ -\bar{q} & \bar{p} & 0 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta v \\ \Delta w \end{bmatrix} + \begin{bmatrix} 0 & \bar{w} & -\bar{v} \\ -\bar{w} & 0 & \bar{u} \\ \bar{v} & -\bar{u} & 0 \end{bmatrix} \begin{bmatrix} \Delta p \\ \Delta q \\ \Delta r \end{bmatrix} \end{aligned} \quad (9.231)$$

and

$$\begin{bmatrix} \Delta L \\ \Delta M \\ \Delta N \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\frac{I_{xz}}{I_{xx}} \\ 0 & 1 & 0 \\ -\frac{I_{xz}}{I_{zz}} & 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta \dot{p} \\ \Delta \dot{q} \\ \Delta \dot{r} \end{bmatrix} + \begin{bmatrix} -\frac{I_{xz}}{I_{xx}} \bar{q} & -\frac{I_{xz}}{I_{xx}} \bar{p} + \frac{I_{zz}-I_{yy}}{I_{xx}} \bar{r} & \frac{I_{zz}-I_{yy}}{I_{xx}} \bar{q} \\ \frac{I_{xx}-I_{zz}}{I_{yy}} \bar{r} + 2 \frac{I_{xz}}{I_{yy}} \bar{p} & 1 & \frac{I_{xx}-I_{zz}}{I_{yy}} \bar{p} - 2 \frac{I_{xz}}{I_{yy}} \bar{r} \\ \frac{I_{yy}-I_{xx}}{I_{zz}} \bar{q} & \frac{I_{xz}}{I_{zz}} \bar{r} + \frac{I_{yy}-I_{xx}}{I_{zz}} \bar{p} & \frac{I_{xz}}{I_{zz}} \bar{q} \end{bmatrix} \begin{bmatrix} \Delta p \\ \Delta q \\ \Delta r \end{bmatrix} \quad (9.232)$$

where the perturbed forces and moments can be modeled as

$$\begin{bmatrix} \Delta X \\ \Delta Z \\ \Delta M \end{bmatrix} = \begin{bmatrix} 0 & X_{\dot{w}} & 0 \\ 0 & Z_{\dot{w}} & 0 \\ 0 & M_{\dot{w}} & 0 \end{bmatrix} \begin{bmatrix} \Delta \dot{u} \\ \Delta \dot{w} \\ \Delta \dot{q} \end{bmatrix} + \begin{bmatrix} X_u & X_w & X_q \\ Z_u & Z_w & Z_q \\ M_u & M_w & M_q \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta w \\ \Delta q \end{bmatrix} + \begin{bmatrix} X_{\delta_e} & X_{\delta_t} \\ Z_{\delta_e} & Z_{\delta_t} \\ M_{\delta_e} & M_{\delta_t} \end{bmatrix} \begin{bmatrix} \Delta \delta_e \\ \Delta \delta_t \end{bmatrix} + \begin{bmatrix} X_{\dot{\eta}_1} & \dots & X_{\dot{\eta}_n} \\ Z_{\dot{\eta}_1} & \dots & Z_{\dot{\eta}_n} \\ M_{\dot{\eta}_1} & \dots & M_{\dot{\eta}_n} \end{bmatrix} \begin{bmatrix} \Delta \dot{\eta}_1 \\ \vdots \\ \Delta \dot{\eta}_n \end{bmatrix} + \begin{bmatrix} X_{\eta_1} & \dots & X_{\eta_n} \\ Z_{\eta_1} & \dots & Z_{\eta_n} \\ M_{\eta_1} & \dots & M_{\eta_n} \end{bmatrix} \begin{bmatrix} \Delta \eta_1 \\ \vdots \\ \Delta \eta_n \end{bmatrix} \quad (9.233)$$

and

$$\begin{bmatrix} \Delta Y \\ \Delta L \\ \Delta N \end{bmatrix} = \begin{bmatrix} Y_v & Y_p & Y_r \\ L_v & L_p & L_r \\ N_v & N_p & N_r \end{bmatrix} \begin{bmatrix} \Delta v \\ \Delta p \\ \Delta r \end{bmatrix} + \begin{bmatrix} Y_{\delta_a} & Y_{\delta_r} \\ L_{\delta_a} & L_{\delta_r} \\ N_{\delta_a} & N_{\delta_r} \end{bmatrix} \begin{bmatrix} \Delta \delta_a \\ \Delta \delta_r \end{bmatrix} + \begin{bmatrix} Y_{\dot{\eta}_1} & \dots & Y_{\dot{\eta}_n} \\ L_{\dot{\eta}_1} & \dots & L_{\dot{\eta}_n} \\ N_{\dot{\eta}_1} & \dots & N_{\dot{\eta}_n} \end{bmatrix} \begin{bmatrix} \Delta \dot{\eta}_1 \\ \vdots \\ \Delta \dot{\eta}_n \end{bmatrix} + \begin{bmatrix} Y_{\eta_1} & \dots & Y_{\eta_n} \\ L_{\eta_1} & \dots & L_{\eta_n} \\ N_{\eta_1} & \dots & N_{\eta_n} \end{bmatrix} \begin{bmatrix} \Delta \eta_1 \\ \vdots \\ \Delta \eta_n \end{bmatrix} \quad (9.234)$$

Note that alternatively one may also use the angle of attack and sideslip angle to make the substitutions

$$\Delta w = \bar{u} \Delta \alpha \quad (9.235)$$

and

$$\Delta v = \bar{v}_{\infty} \Delta \beta \quad (9.236)$$

under the small angle approximation and no-wind assumption. Furthermore, if  $\bar{v} = \bar{\phi} = \bar{p} = \bar{q} = \bar{r} = 0$ , then one can decouple the dynamics into the longitudinal and lateral-directional.

Lastly, one must also include the linearized vibration equations. However, as these are already linearly modeled, one can simply write

$$\Delta \ddot{\eta}_i + 2\zeta_i \omega_i \Delta \dot{\eta}_i + \omega_i^2 \Delta \eta_i = \frac{\Delta Q_i}{M_i}, \quad i = 1, \dots, n \quad (9.237)$$

where the generalized forces can be modeled as linear equations of the states, i.e.

$$\begin{aligned} \Delta Q_i = & [Q_{i_u} \quad Q_{i_\beta} \quad Q_{i_\alpha} \quad Q_{i_p} \quad Q_{i_q} \quad Q_{i_r}] \begin{bmatrix} \Delta u \\ \Delta \beta \\ \Delta \alpha \\ \Delta p \\ \Delta q \\ \Delta r \end{bmatrix} + [Q_{i_{\delta_a}} \quad Q_{i_{\delta_e}} \quad Q_{i_{\delta_r}} \quad Q_{i_{\delta_t}}] \begin{bmatrix} \Delta \delta_a \\ \Delta \delta_e \\ \Delta \delta_r \\ \Delta \delta_t \end{bmatrix} \\ & + [Q_{i_{\dot{\eta}_1}} \quad \cdots \quad Q_{i_{\dot{\eta}_n}}] \begin{bmatrix} \Delta \dot{\eta}_1 \\ \vdots \\ \Delta \dot{\eta}_n \end{bmatrix} + [Q_{i_{\eta_1}} \quad \cdots \quad Q_{i_{\eta_n}}] \begin{bmatrix} \Delta \eta_1 \\ \vdots \\ \Delta \eta_n \end{bmatrix} \end{aligned} \quad (9.238)$$

For an explicit example of a linearized elastic flight vehicle equation of motion, consider the simpler case where the trim flight condition is straight-and-level flight and the longitudinal and lateral-directional EOMs can be decoupled. Furthermore, assume that the vibration modes can also be decoupled between the longitudinal and lateral-directional and  $\dot{\alpha}$  derivatives are zero. Then, the longitudinal rigid airplane LTI state-space model

$$\begin{bmatrix} \Delta \dot{u} \\ \Delta \dot{\alpha} \\ \Delta \dot{q} \\ \Delta \dot{\theta} \end{bmatrix} = \begin{bmatrix} X_u & X_\alpha & X_q & -g \\ \frac{Z_u}{\bar{u}} & \frac{Z_\alpha}{\bar{u}} & 1 + \frac{Z_q}{\bar{u}} & 0 \\ M_u & M_\alpha & M_q & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \\ \Delta \theta \end{bmatrix} + \begin{bmatrix} X_{\delta_e} & X_{\delta_t} \\ \frac{Z_{\delta_e}}{\bar{u}} & \frac{Z_{\delta_t}}{\bar{u}} \\ M_{\delta_e} & M_{\delta_t} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \delta_e \\ \Delta \delta_t \end{bmatrix} \quad (9.239)$$

which models the following portions of Equation 7.304 as

$$\begin{bmatrix} \Delta \dot{\vec{x}}_{rig} \\ \Delta \dot{\vec{x}}_{eul} \end{bmatrix} = \begin{bmatrix} A_{rig \leftarrow rig} & A_{rig \leftarrow eul} \\ A_{eul \leftarrow rig} & A_{eul \leftarrow eul} \end{bmatrix} \begin{bmatrix} \Delta \vec{x}_{rig} \\ \Delta \vec{x}_{eul} \end{bmatrix} + \begin{bmatrix} B_{rig} \\ 0 \end{bmatrix} \Delta \vec{u} \quad (9.240)$$

Note that here  $\Delta \alpha$  has been used in place of  $\Delta w$ . Then, one may form the other matrices as

$$A_{rig \leftarrow vib} = \begin{bmatrix} X_{\eta_1} & \cdots & X_{\eta_n} & X_{\dot{\eta}_1} & \cdots & X_{\dot{\eta}_n} \\ \frac{Z_{\eta_1}}{\bar{u}} & \cdots & \frac{Z_{\eta_n}}{\bar{u}} & \frac{Z_{\dot{\eta}_1}}{\bar{u}} & \cdots & \frac{Z_{\dot{\eta}_n}}{\bar{u}} \\ M_{\eta_1} & \cdots & M_{\eta_n} & M_{\dot{\eta}_1} & \cdots & M_{\dot{\eta}_n} \end{bmatrix} \quad (9.241)$$

and defining the **elastic stability and control derivative** for the state or input  $\bullet$  as

$$\Xi_{i\bullet} = \frac{Q_{i\bullet}}{\mathcal{M}_i} \quad (9.242)$$

one can write the vibration state and input sub-matrices as

$$A_{vib \leftarrow rig} = \begin{bmatrix} \Xi_{1_u} & \Xi_{1_\alpha} & \Xi_{1_q} \\ \vdots & \vdots & \vdots \\ \Xi_{n_u} & \Xi_{n_\alpha} & \Xi_{n_q} \end{bmatrix} \quad (9.243)$$

$$A_{vib \leftarrow \eta} = \begin{bmatrix} \Xi_{1\eta_1} & \cdots & \Xi_{1\eta_n} \\ \vdots & \ddots & \vdots \\ \Xi_{n\eta_1} & \cdots & \Xi_{n\eta_n} \end{bmatrix} - \Omega^2 \quad (9.244)$$

$$A_{vib \leftarrow \dot{\eta}} = \begin{bmatrix} \Xi_{1\dot{\eta}_1} & \cdots & \Xi_{1\dot{\eta}_n} \\ \vdots & \ddots & \vdots \\ \Xi_{n\dot{\eta}_1} & \cdots & \Xi_{n\dot{\eta}_n} \end{bmatrix} - 2\Omega_\zeta \quad (9.245)$$

and

$$B_{vib} = \begin{bmatrix} \Xi_{1\delta_e} & \Xi_{1\delta_t} \\ \vdots & \vdots \\ \Xi_{n\delta_e} & \Xi_{n\delta_t} \end{bmatrix} \quad (9.246)$$

Thus, in the end, one has

$$\begin{bmatrix} \Delta\dot{u} \\ \Delta\dot{\alpha} \\ \Delta\dot{q} \\ \Delta\dot{\theta} \\ \Delta\dot{\eta}_1 \\ \vdots \\ \Delta\dot{\eta}_n \\ \Delta\ddot{\eta}_1 \\ \vdots \\ \Delta\ddot{\eta}_n \end{bmatrix} = \begin{bmatrix} X_u & X_\alpha & X_q & -g & X_{\eta_1} & \cdots & X_{\eta_n} & X_{\dot{\eta}_1} & \cdots & X_{\dot{\eta}_n} \\ \frac{Z_u}{\bar{u}} & \frac{Z_\alpha}{\bar{u}} & 1 + \frac{Z_q}{\bar{u}} & 0 & \frac{Z_{\eta_1}}{\bar{u}} & \cdots & \frac{Z_{\eta_n}}{\bar{u}} & \frac{Z_{\dot{\eta}_1}}{\bar{u}} & \cdots & \frac{Z_{\dot{\eta}_n}}{\bar{u}} \\ M_u & M_\alpha & M_q & 0 & M_{\eta_1} & \cdots & M_{\eta_n} & M_{\dot{\eta}_1} & \cdots & M_{\dot{\eta}_n} \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \\ \Xi_{1u} & \Xi_{1\alpha} & \Xi_{1q} & 0 & \Xi_{1\eta_1} - \omega_1^2 & \cdots & \Xi_{1\eta_n} & \Xi_{1\dot{\eta}_1} - 2\zeta_1\omega_1 & \cdots & \Xi_{1\dot{\eta}_n} \\ \vdots & \vdots & \vdots & 0 & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Xi_{nu} & \Xi_{n\alpha} & \Xi_{nq} & 0 & \Xi_{n\eta_1} & \cdots & \Xi_{n\eta_n} - \omega_n^2 & \Xi_{n\dot{\eta}_1} & \cdots & \Xi_{n\dot{\eta}_n} - 2\zeta_n\omega_n \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \\ \Delta \theta \\ \Delta \eta_1 \\ \vdots \\ \Delta \eta_n \\ \Delta \dot{\eta}_1 \\ \vdots \\ \Delta \dot{\eta}_n \end{bmatrix} + \begin{bmatrix} X_{\delta_e} & X_{\delta_t} \\ \frac{Z_{\delta_e}}{\bar{u}} & \frac{Z_{\delta_t}}{\bar{u}} \\ M_{\delta_e} & M_{\delta_t} \\ \vec{0}_{n+1} & \vec{0}_{n+1} \\ \Xi_{1\delta_e} & \Xi_{1\delta_t} \\ \vdots & \vdots \\ \Xi_{n\delta_e} & \Xi_{n\delta_t} \end{bmatrix} \begin{bmatrix} \Delta \delta_e \\ \Delta \delta_t \end{bmatrix} \quad (9.247)$$

## References

For more information, please refer to the following

- Schmidt, D. K., “Chapter 7: Effects of Elastic Deformation on the Forces and Moments,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 323-393
- Schmidt, D. K., “10.10 On the Flight Dynamics of Flexible Vehicles,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 621-626

## 9.7 Airplane Attitude Control Systems

### Airplane Attitude Plant Model

The design of an airplane's inner-loop attitude control systems for a particular steady-flight condition typically use the LTI longitudinal state-space model approximation, i.e.

$$\begin{aligned} \begin{bmatrix} \Delta\dot{u} \\ \Delta\dot{\alpha} \\ \Delta\dot{q} \\ \Delta\dot{\theta} \end{bmatrix} &= A_{long} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \\ \Delta \theta \end{bmatrix} + B_{long} \begin{bmatrix} \Delta \delta_e \\ \Delta \delta_T \end{bmatrix} \\ \Delta \vec{y}_{long} &= C_{long} \begin{bmatrix} \Delta u \\ \Delta \alpha \\ \Delta q \\ \Delta \theta \end{bmatrix} \end{aligned} \quad (9.248)$$

and the LTI lateral-directional state-space model approximation, i.e.

$$\begin{aligned} \begin{bmatrix} \Delta\dot{\beta} \\ \Delta\dot{p} \\ \Delta\dot{r} \\ \Delta\dot{\phi} \end{bmatrix} &= A_{lat} \begin{bmatrix} \Delta \beta \\ \Delta p \\ \Delta r \\ \Delta \phi \end{bmatrix} + B_{lat} \begin{bmatrix} \Delta \delta_a \\ \Delta \delta_r \end{bmatrix} \\ \Delta \vec{y}_{lat} &= C_{lat} \begin{bmatrix} \Delta \beta \\ \Delta p \\ \Delta r \\ \Delta \phi \end{bmatrix} \end{aligned} \quad (9.249)$$

where  $C_{long}$  and  $C_{lat}$  can be chosen for any state of interest.

Thus, the transfer function matrix for the longitudinal and lateral-directional inputs and outputs can be computed, respectively, as

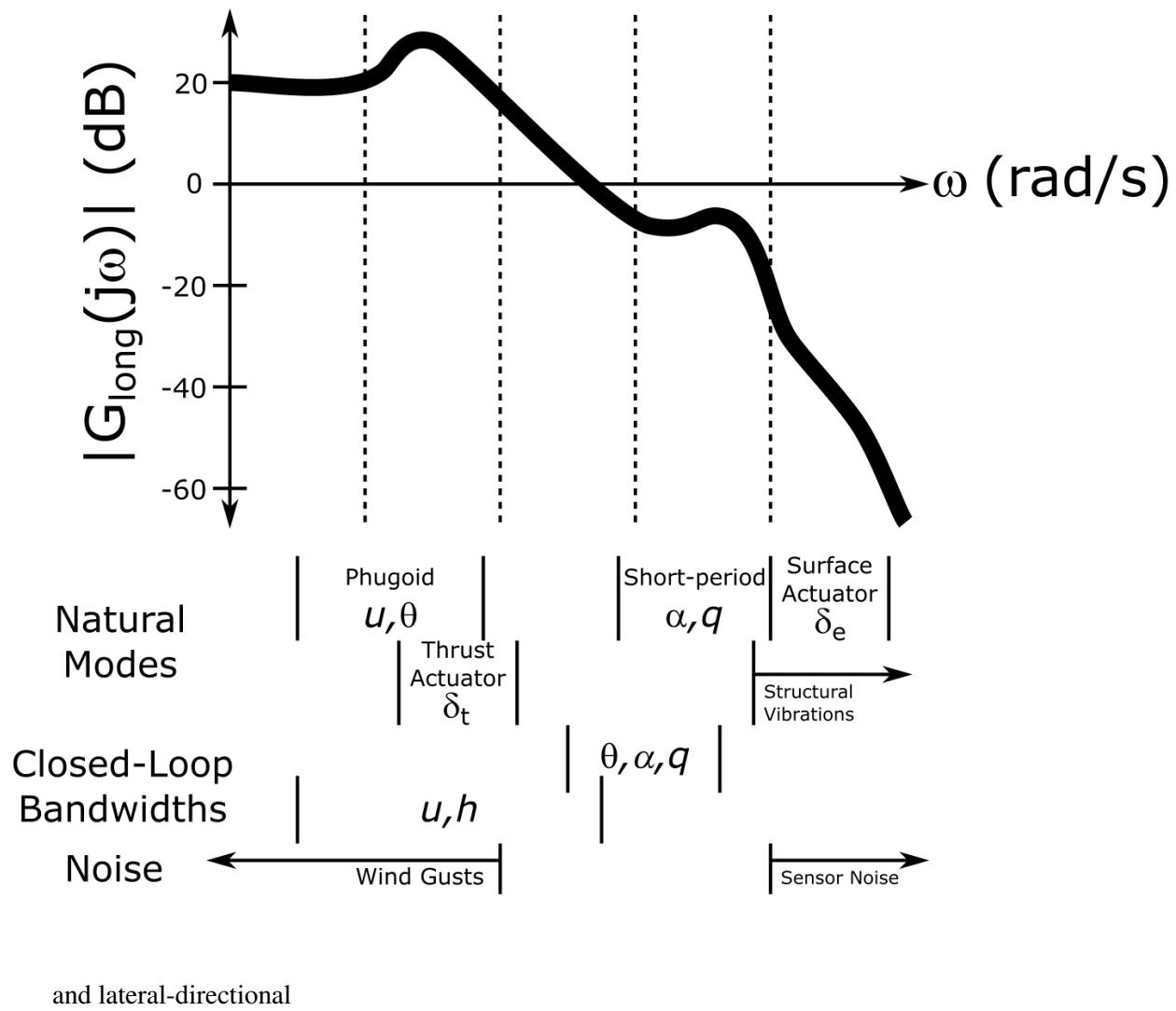
$$\Delta \vec{y}_{long}(s) = C_{long}(sI_{4 \times 4} - A_{long})^{-1} B_{long} \begin{bmatrix} \Delta \delta_e(s) \\ \Delta \delta_T(s) \end{bmatrix} \quad (9.250)$$

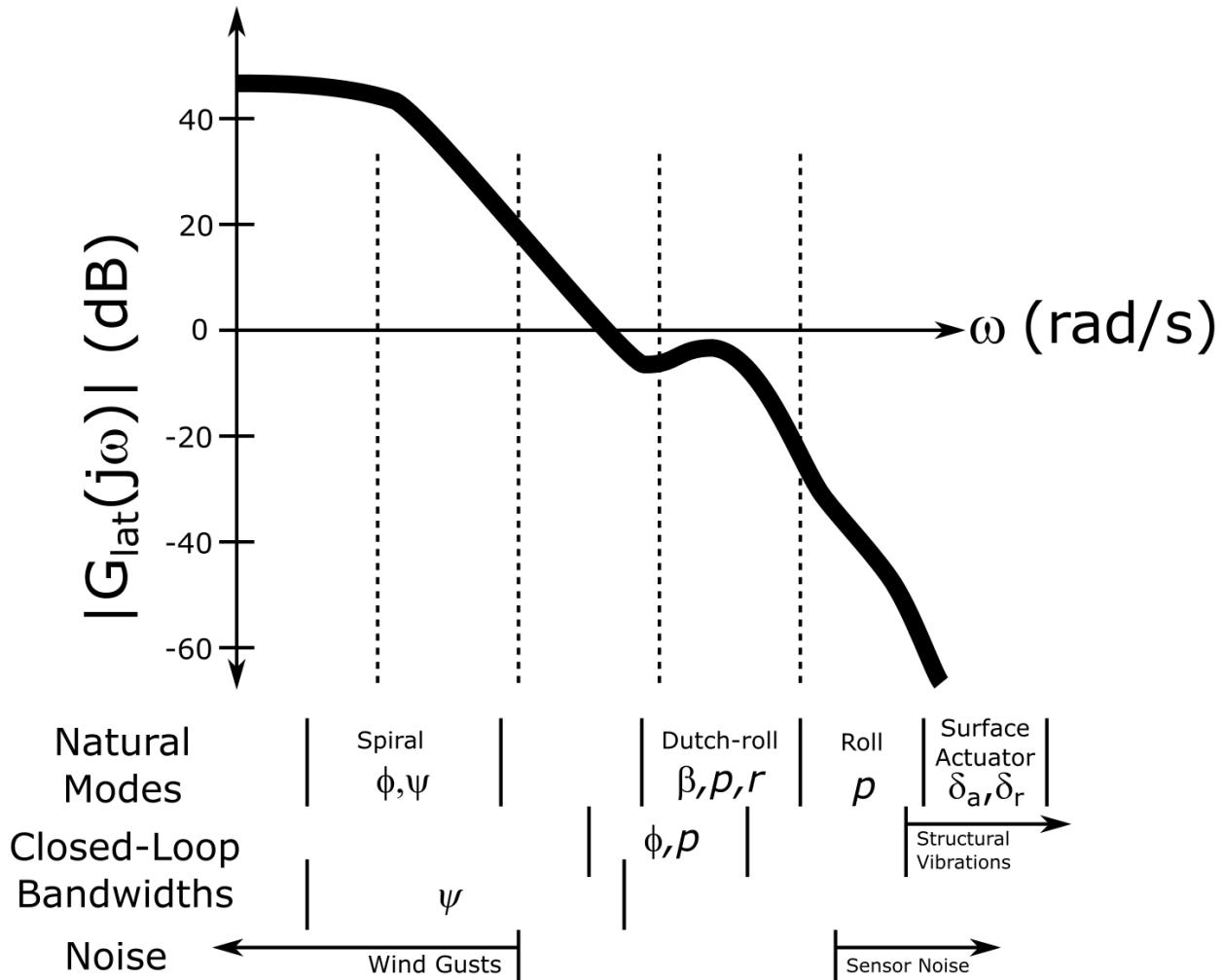
and

$$\Delta \vec{y}_{lat}(s) = C_{lat}(sI_{4 \times 4} - A_{lat})^{-1} B_{lat} \begin{bmatrix} \Delta \delta_a(s) \\ \Delta \delta_r(s) \end{bmatrix} \quad (9.251)$$

where  $(sI_{4 \times 4} - A_{long})^{-1}$  provides the characteristic polynomial for the short-period and phugoid modes and  $(sI_{4 \times 4} - A_{lat})^{-1}$  provides the characteristic polynomial for the roll, dutch-roll, and spiral modes.

The general frequency response for these dynamic modes, the actuation modes, the wind disturbances, structural vibrations, the modes of the feedback control systems, and the frequency bands of the sensor noise are shown generally below for the longitudinal





It should be noted that the numerators of transfer functions to the states  $\Delta u(s)$ ,  $\Delta \alpha(s)$ ,  $\Delta q(s)$ ,  $\Delta \beta(s)$ ,  $\Delta p(s)$ , and  $\Delta r(s)$  are third-order polynomials in  $s$  while  $\Delta \theta(s)$  and  $\Delta \phi(s)$  are second order polynomials in  $s$  as  $\Delta q(s) = s\Delta \theta(s)$  and  $\Delta p(s) = s\Delta \phi(s)$  for straight-and-level flight. Furthermore, though the linearized and decoupled transfer functions are used for airplane feedback control system design, it should be mentioned that the control inputs will generally affect *all* airplane states and should be simulated with the nonlinear airplane model once the guidance and control system design has been completed.

In addition, one may also use the following transfer function relationships via the small angle approximations in feedback control design. For the flight path angle, one has

$$\Delta \gamma(s) = \Delta \theta(s) - \Delta \alpha(s) \quad (9.252)$$

For altitude  $h$ , one has

$$\Delta h(s) = \frac{\bar{u}}{s} \Delta \gamma(s) \quad (9.253)$$

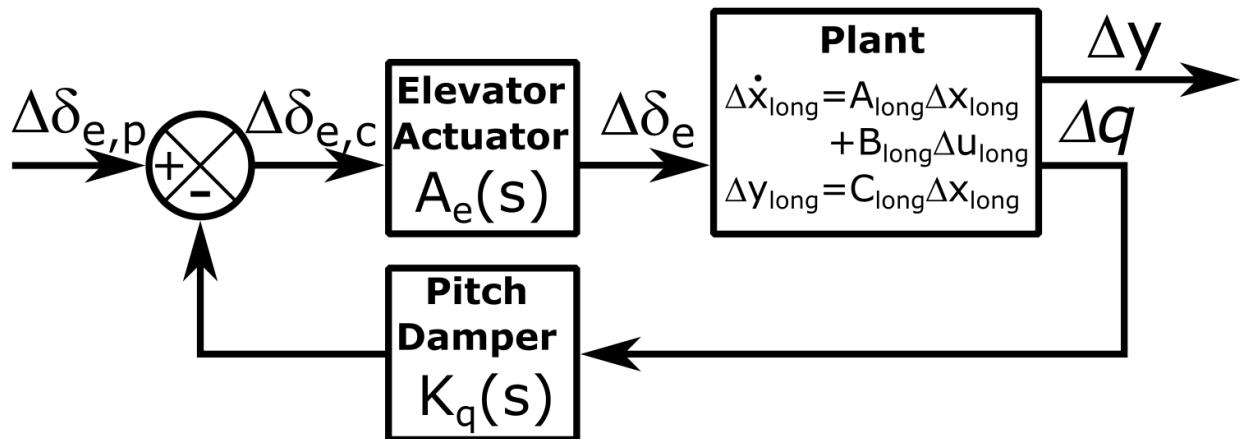
For the yaw angle, one has

$$\Delta\psi(s) = \frac{1}{s} \Delta r(s) \quad (9.254)$$

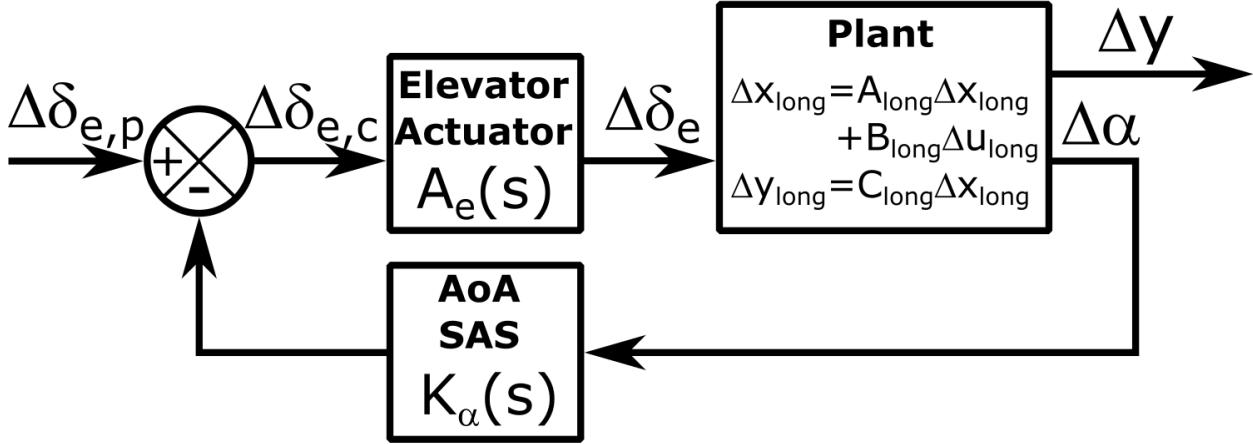
### Airplane Stability Augmentation System Design

**Stability augmentation systems (SAS)** for airplanes are designed for the airplane to be controlled by a pilot (subscript  $p$  on the command), but the inherent modal stability or damping ratios are not within suitable flying qualities for the pilot. Thus, a SAS is used to augment the airplane dynamics to achieve certain stability or damping ratios for the airplane modes. These SAS are generally single-output feedback systems that use a single gain term,  $K_s$ , which changes the location of the system poles. To choose the value of  $K_s$ , one can use a **root locus** plot, i.e. a plot of the roots of the closed-loop characteristic polynomial in the complex plane as a function of  $K_s$ .

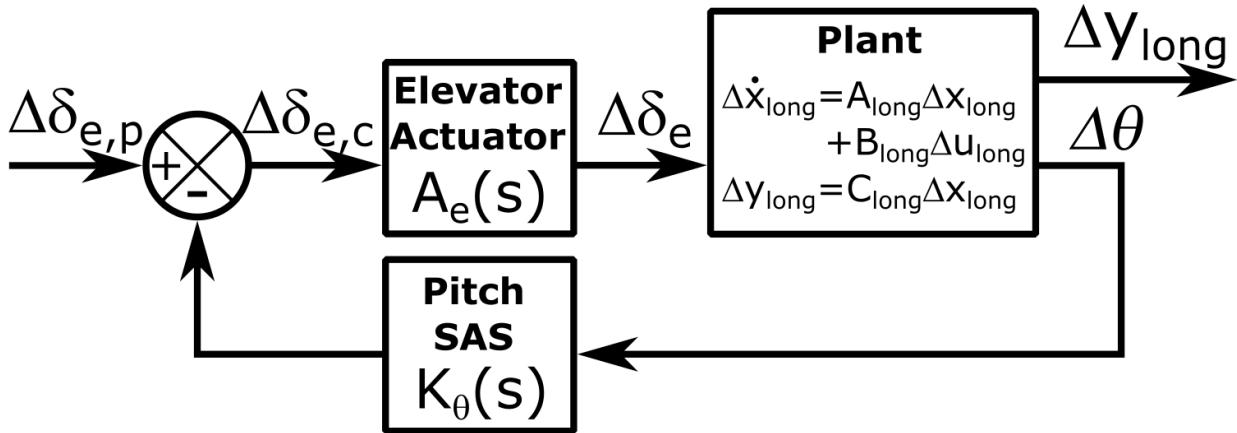
A **pitch damper** is used to reduce the damping ratio of the short-period mode of an airplane. A pitch damper,  $K_q(s)$  uses a proportional gain,  $K_q$ , on the pitch rate,  $q$ , subtracted from the elevator deflection,  $\delta_e$ . This SAS can be designed using a linearized plant as shown



An **angle of attack (AoA) SAS** is used to increase the frequency and stabilize the short-period mode of an airplane. An angle of attack SAS,  $K_\alpha(s)$  uses a proportional gain,  $K_\alpha$ , on the angle of attack,  $\alpha$ , subtracted from the elevator deflection,  $\delta_e$ . This SAS can be designed using a linearized plant as shown



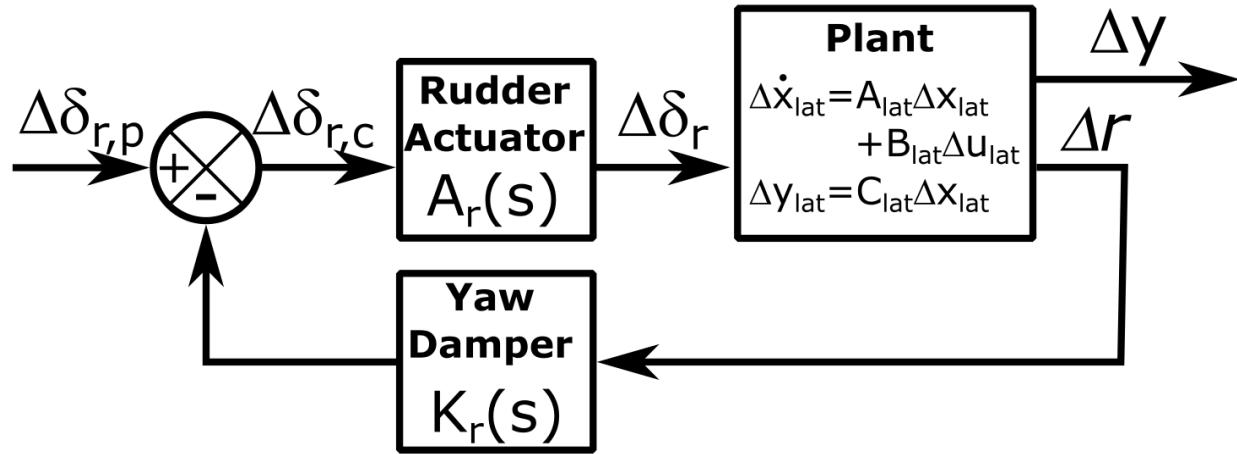
A **pitch SAS** is used to increase the natural frequency and stabilize the phugoid mode of an airplane. A pitch SAS,  $K_\theta(s)$  uses a proportional gain,  $K_\theta$ , on the pitch,  $\theta$ , subtracted from the elevator deflection,  $\delta_e$ . This SAS can be designed using a linearized plant as shown



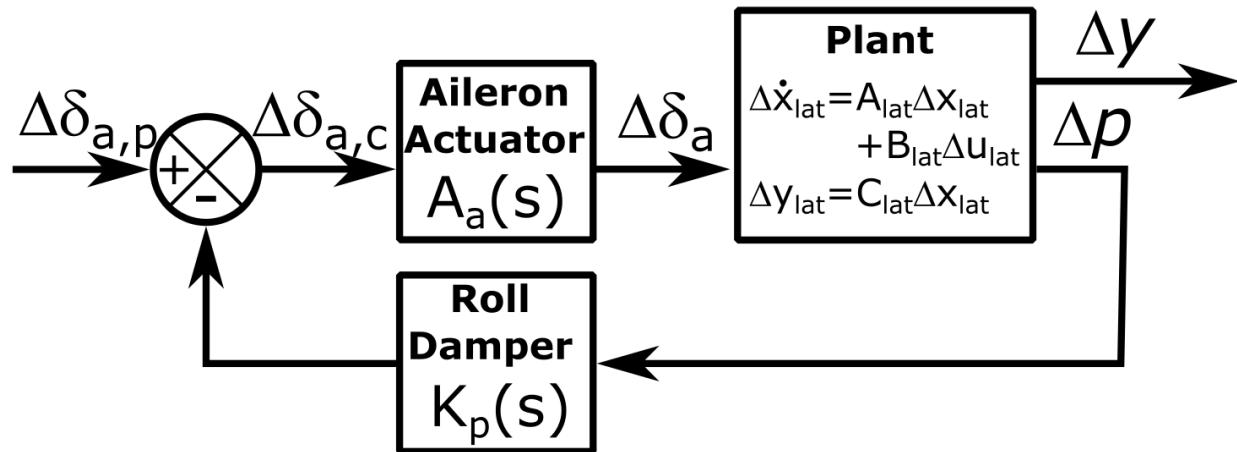
A **yaw damper** is used to increasing the damping ratio of the dutch-roll mode of an airplane. A first-order yaw damper feeds back the yaw rate,  $r$ , to the system composed of two stages

$$K_r(s) = K_r \frac{\omega_w s}{s + \omega_w} \quad (9.255)$$

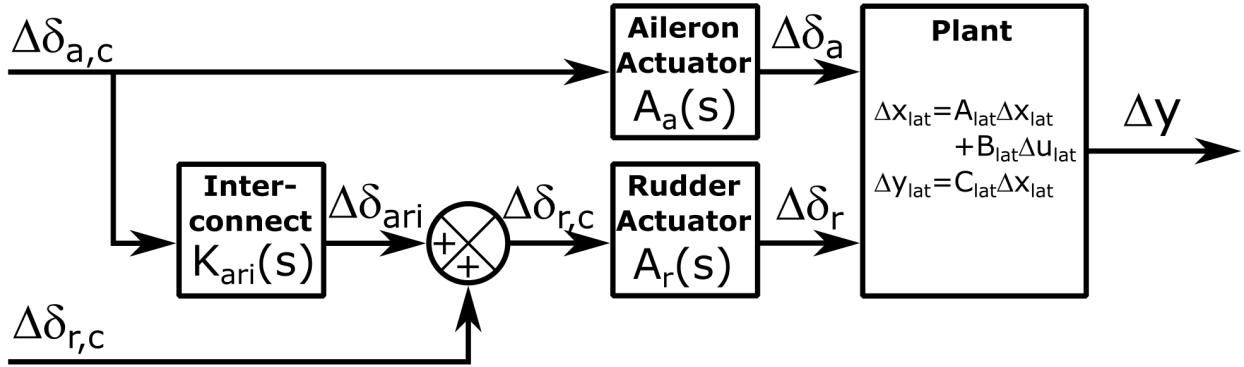
where  $K_r$  is the yaw damper gain and  $\omega_w$  is the washout frequency of the first-order high-pass filter which turns off the yaw damper for sustained turns where the steady yaw rate is not zero and is typically selected well below the dutch-roll natural frequency, e.g., a decade below. The effect of this first-order filter “washes out” the yaw damper for low frequency signals, i.e. the yaw damper only affects  $\omega \geq \omega_w$ . The control input is subtracted from the rudder deflection,  $\delta_r$ . This SAS can be designed using a linearized plant as shown



A **roll damper** is used to reduce the roll mode time constant which primarily affects  $\Delta p$ . A roll damper,  $K_p(s)$  uses a proportional gain,  $K_p$ , on the roll rate,  $p$ , subtracted from the aileron deflection,  $\delta_a$ . Thus, a roll damper can also be used to increase the effectiveness of a yaw damper as it feeds back  $\Delta p$  to the SAS. This SAS can be designed using a linearized plant as shown



As opposed to the longitudinal control inputs, the lateral-directional attitude control is performed by two control surface deflections which have strong coupling effects on the dynamics. For example, aileron deflections often cause an undesirable excitation of the dutch roll mode and/or an adverse yawing moment,  $N_{\delta_a}$ . Thus, many aircraft typically use some sort of **aileron-rudder interconnect (ARI)** to reduce these effects.



A simple way to select  $K_{ari}$  is to cancel the adverse yaw with the rudder, i.e. forcing the effective  $N_{\delta_a} \approx 0$  by setting the yawing moment due to rudder deflection equal to the negative of the yawing moment due to aileron deflection or in mathematical terms

$$N_{\delta_r}^* \Delta\delta_r = -N_{\delta_a}^* \Delta\delta_a \quad (9.256)$$

or letting

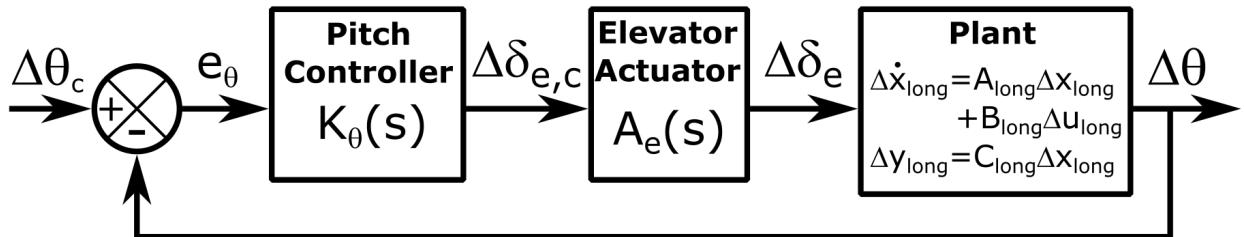
$$\Delta\delta_r = -\frac{N_{\delta_a}^*}{N_{\delta_r}^*} \Delta\delta_a \quad (9.257)$$

one has

$$K_{ari} = -\frac{N_{\delta_a}^*}{N_{\delta_r}^*} \quad (9.258)$$

### Introductory Airplane Attitude Control System Design

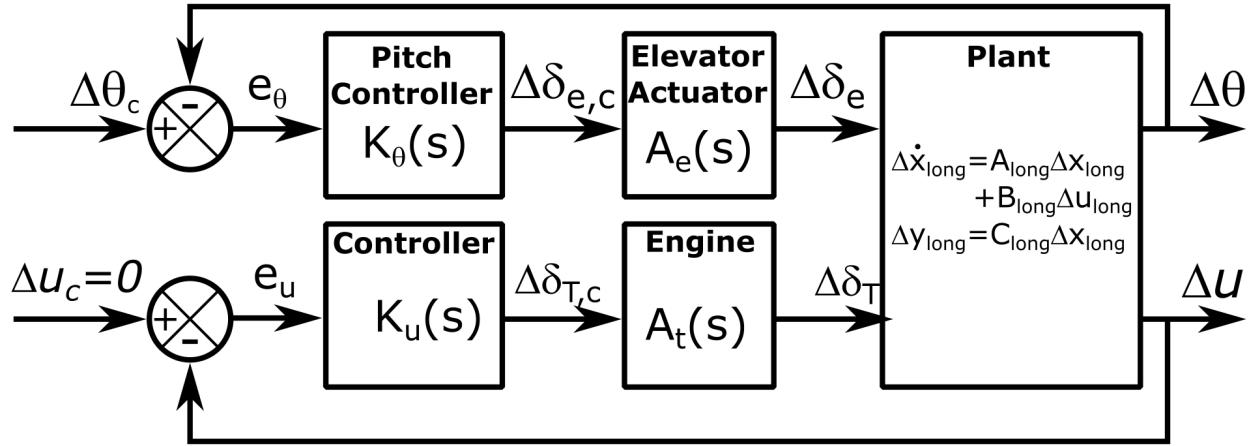
The design of the longitudinal inner-loop control system for an airplane typically uses the pitch control law,  $K_\theta(s)$ , with the longitudinal LTI plant model as shown in the following block diagram.



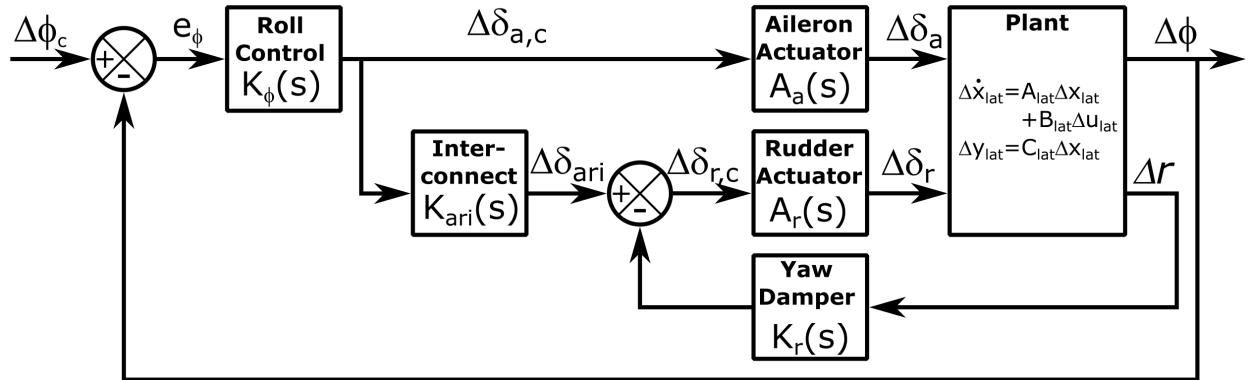
Other alternatives for the inner-loop include the pitch rate  $\Delta q$  and the angle of attack  $\Delta\alpha$  as the single output for the system to track with the inner-loop. Note that a pitch damper can also be the “derivative stage” for a PID controller on  $\theta$ .

When designing the pitch attitude controller, an important aspect to assess is whether an increase in  $\Delta\theta_c$  leads to an increase or decrease in the flight-path angle  $\Delta\gamma$  as  $t \rightarrow \infty$ . If the airplane is on the “front side

of the power curve,” i.e. a reduction in airspeed,  $v_\infty$ , requires less power to maintain level flight, then an increase in  $\Delta\theta_c$  will lead to a positive steady-state  $\Delta\gamma$ . However, if the airplane is on the “back side of the power curve,” i.e. a reduction in airspeed,  $\Delta u$ , requires more power to maintain level flight, then an increase in  $\Delta\theta_c$  will lead to a negative steady-state  $\Delta\gamma$ . To overcome this coupling between  $\Delta\theta_c$ ,  $\Delta\gamma$ , and  $v_\infty$ , one can use  $\delta_T$  in an outer-loop **auto-throttle** control system to keep the velocity at some  $\Delta u_c$  to account for the change in overall airspeed,  $v_\infty$ , due to a command in  $\Delta\theta_c$  as shown



The lateral-directional inner-loop control system for an airplane typically uses the roll control law,  $K_\phi(s)$ , the yaw damper,  $K_r(s)$ , the ARI,  $K_{ARI}(s)$ , and the lateral-directional LTI plant model as shown in the following block diagram.



This type of roll controller can be used to command a coordinated turn, i.e.  $\beta = 0^\circ$  with  $\Delta\phi_c \neq 0^\circ$  through the use of the ARI. It should be noted that the washout filter stage must be included for the case of sustained turns where the yaw rate is not zero. Without it, the yaw rate feedback (i.e. damper) would tend to “fight” the turn by maintaining a rudder deflection opposite to that desired for the turn. Considering these additions to a roll angle feedback control system, one typically must first form the plant model  $G_\phi(s)$  using

the yaw damper before loop-shaping  $L_\phi(s) = G_\phi(s)K_\phi(s)$ . It should be mentioned that though using an ARI is one method for controlling a coordinated turn, i.e.  $\beta = 0^\circ$  with  $\Delta\phi_c \neq 0^\circ$ , one may also consider feeding back to the rudder, either  $\beta$ ,  $a_y$ , or  $r$ .

Lastly, when performing a coordinated turn, one must also generate the suitable longitudinal control inputs to maintain altitude, a concept called **turn compensation**. Primarily, by rotating the lift vector of the airplane, the vertical component of lift must still counteract the weight for constant altitude. Thus, as an airplane enters the turn, the angle of attack must be increased which requires an appropriate elevator deflection given the flight velocity and bank angle. One simple method is to compute the yaw rate and roll angle to generate the necessary commanded pitch rate for a pitch or pitch rate feedback control system, i.e.

$$q_c = r \tan \phi \quad (9.259)$$

where these values are for the entire airplane, not the perturbed states.

### Structural-Mode Control

**Structural-mode control (SMC)** is the use of additional stages in the flight control law to account for the effects of vibrations or structural-modes at frequencies close to the rigid-body dynamics that cannot simply be removed using low-pass filtering, e.g., if the modal frequencies are within a factor of ten of the rigid-body modes. An important aspect of the feedback control for these systems is the placement of inertial sensors, i.e., accelerometers and gyroscopes, at proper locations along the structure that provide mode-displacement or mode-slope measurements, respectively, to the feedback control system. With this information, one can implement passive structural-mode control which simply filters these measurements or active structural-mode control which utilizes specialized actuators with the sensors.

**Passive SMC** uses additional mechanical or physical devices to increase effectiveness of the structural modes to reject disturbances. This is typically used in conjunction with **structural-mode filtering** which, in a similar fashion to a low-pass filtering of high-frequency structural-modes, one can use a targeted band-stop filter or a notch filter to target the band at which the structural-mode is strongly excited by the actuators. This provides a method to mitigate structural-mode effects from disturbances and control inputs.

A more direct way to mitigate structural-mode effects is through **active SMC** which utilizes co-located actuators and sensors at proper locations on the structure to measure the modal acceleration and/or mode-slope rate via accelerometer and/or a rate gyroscope and use feedback to regulate the structural-mode excitation across the frequency band of the structural mode. This design approach utilizes a dampening effect from a force or moment applied to an object in proportion to and in the opposite direction from the linear or angular velocity of the object. Notably, the linear velocity must be calculated via the integration of the measured acceleration. Importantly, if the force/moment and measured velocity are at the same location on the flexible structure then proper phasing will always be present, i.e., the feedback will occur at the “right time,” not causing the elastic system to become unstable. To avoid exciting the rigid-body modes, one typically must use a high-pass filter to target the structural modes only.

### Total Energy Control System Design

In classical airplane attitude control design, one uses SISO principles to design the pitch attitude controller using the elevator input, the roll attitude controller using the aileron control input, sideslip attitude controller

using the rudder control input, and the airspeed controller using the thrust input. However, this decoupled approach is sub-optimal as these SISO systems do not coordinate their control actions to manage the airplane energy or heading for a commanded trajectory. In particular, the airspeed controller does not incorporate the important state variable, the flight-path angle, in determining the required thrust while the pitch attitude controller does not know the vertical velocity nor its limits. Furthermore, the roll and sideslip angles are interrelated to the yaw and heading angles, one is required to use a yaw damper and turn compensation for coordinated turns in the inner-loop as well as thrust asymmetry compensation in the event of an engine-out condition.

As the lift force for airplanes is affected by both speed and angle of attack, the fundamental physics in longitudinal control should be energy management of the flight vehicle. Also, one can use the small angle dynamics to provide processed error signals that decouple the yaw and sideslip before providing the command to the ailerons and rudder actuator using the fundamental physics of the lateral and directional angles of the flight vehicle. Thus, modern fixed-wing vehicles often use total energy and total heading control systems for MIMO attitude control.

From the point-mass perspective with the constant-wind assumption in straight flight, the total energy of the flight vehicle,  $\mathcal{E}$ , is given by the sum of the kinetic and potential energy, i.e.

$$\mathcal{E} = mg \left( \frac{1}{2} \frac{v_\infty^2 + v_w^2}{g} + h \right) \quad (9.260)$$

where  $m$  is the mass,  $v_\infty$  is the airspeed,  $v_w$  is the constant wind,  $g$  is the assumed constant acceleration due to gravity, and  $h$  is the altitude. Similarly, the Lagrangian,  $\mathcal{L}$ , is given by the difference of the kinetic and potential energy, i.e.

$$\mathcal{L} = mg \left( \frac{1}{2} \frac{v_\infty^2 + v_w^2}{g} - h \right) \quad (9.261)$$

Differentiating with respect to time, one has

$$\dot{\mathcal{E}} = mg \left( \frac{v_\infty \dot{v}_\infty}{g} + \dot{h} \right) \quad (9.262)$$

and

$$\dot{\mathcal{L}} = mg \left( \frac{v_\infty \dot{v}_\infty}{g} - \dot{h} \right) \quad (9.263)$$

Rearranging and substituting for  $\dot{h} \approx \bar{v}_\infty \gamma$  for the small angle approximation, one can rewrite these equations as

$$\dot{\tilde{\mathcal{E}}} = \frac{\dot{\mathcal{E}}}{mg v_\infty} = \frac{\dot{v}_\infty}{g} + \gamma \quad (9.264)$$

and

$$\dot{\tilde{\mathcal{L}}} = \frac{\dot{\mathcal{L}}}{mg v_\infty} = \frac{\dot{v}_\infty}{g} - \gamma \quad (9.265)$$

where  $\dot{\tilde{\mathcal{E}}}$  is the normalized total energy and  $\dot{\tilde{\mathcal{L}}}$  is the normalized Lagrangian.

Next, recall the  $x_W$ -axis force balance is given by

$$-D + T \cos \alpha \cos \beta - mg \sin \gamma = m \dot{v}_\infty \quad (9.266)$$

which for small angles, one has

$$T = mg \left( \frac{\dot{v}_\infty}{g} + \gamma \right) + D \quad (9.267)$$

For some steady-flight conditions, one has

$$\bar{T} + \Delta T = mg \left( \frac{\dot{v}_\infty}{g} + \bar{\gamma} + \Delta \gamma \right) + \bar{D} - \Delta D \quad (9.268)$$

where assuming  $\bar{T} \approx mg\bar{\gamma} + \bar{D}$  for trim and  $\Delta D \approx 0$ , one has

$$\frac{\Delta T}{mg} = \delta_T = \frac{\dot{v}_\infty}{g} + \Delta \gamma = \frac{\Delta \dot{\mathcal{E}}}{mg v_\infty} \quad (9.269)$$

which demonstrates that the normalized thrust,  $\delta_T$ , can be used effectively to change the rate of the total energy for the system.

Finally, recall that for coordinated, straight-flight, one has

$$\gamma = \theta - \alpha \quad (9.270)$$

Thus, the normalized Lagrangian can be written as

$$\dot{\mathcal{L}} = \frac{\dot{v}_\infty}{g} + \alpha - \theta \quad (9.271)$$

which demonstrates that the pitch,  $\theta$ , can be used effectively to change the rate of the normalized Lagrangian. In TECS, the elevator deflection,  $\delta_e$ , provides an additional SISO inner-loop attitude control for the pitch.

In a total energy control system (TECS) design, the planning system provides the TECS a commanded airspeed acceleration in g's,  $\frac{\dot{v}_{\infty,c}}{g}$ , and a commanded flight-path angle,  $\gamma_c$ . This is related to the a commanded total energy rate

$$\dot{\mathcal{E}}_c = \frac{\dot{v}_{\infty,c}}{g} + \gamma_c \quad (9.272)$$

and a commanded Lagrangian rate

$$\dot{\mathcal{L}}_c = \frac{\dot{v}_{\infty,c}}{g} - \gamma_c \quad (9.273)$$

which make up the reference signal to track, i.e.

$$\vec{r} = \begin{bmatrix} \dot{\mathcal{E}}_c \\ \dot{\mathcal{L}}_c \end{bmatrix} = \begin{bmatrix} \frac{\dot{v}_{\infty,c}}{g} + \gamma_c \\ \frac{\dot{v}_{\infty,c}}{g} - \gamma_c \end{bmatrix} \quad (9.274)$$

with output

$$\vec{y} = \begin{bmatrix} \dot{\mathcal{E}} \\ \dot{\mathcal{L}} \end{bmatrix} = \begin{bmatrix} \frac{\dot{v}_\infty}{g} + \gamma \\ \frac{\dot{v}_\infty}{g} - \gamma \end{bmatrix} \quad (9.275)$$

with tracking error

$$\vec{e} = \vec{y} - \vec{r} \quad (9.276)$$

Next, consider that one desires to track constant step inputs to  $\vec{r}$  and reject constant disturbances,  $\vec{w}$ , e.g. the additive drag, one can use the servomechanism state-space augmentation with  $p = 1$  and  $a_1 = 0$  as

$$\vec{z} = \begin{bmatrix} \vec{e} \\ \dot{\vec{y}} \end{bmatrix} \quad (9.277)$$

and

$$\vec{u} = \dot{\vec{u}} \quad (9.278)$$

where one can design a feedback control law on the servomechanism as

$$\vec{\mu} = K_z \vec{z} \quad (9.279)$$

or

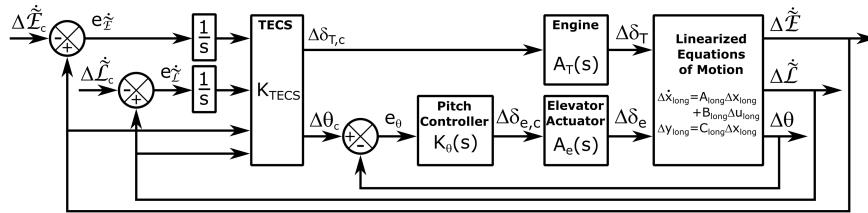
$$\vec{u} = [K_e \quad K_y] \left[ \int \frac{\vec{e}}{\vec{y}} \right] \quad (9.280)$$

where  $K_e \in \mathbb{R}^{n_u \times n_y}$  and  $K_y \in \mathbb{R}^{n_u \times n_y}$ .

For the TECS, one simplifies this general structure into four non-zero and four zero elements of the gain matrix, i.e.

$$\begin{bmatrix} \delta_T \\ \theta_c \end{bmatrix} = \begin{bmatrix} K_{e,\dot{\mathcal{E}}} & 0 & K_{y,\dot{\mathcal{E}}} & 0 \\ 0 & K_{e,\dot{\mathcal{L}}} & 0 & K_{y,\dot{\mathcal{L}}} \end{bmatrix} \begin{bmatrix} \int (\dot{\mathcal{E}} - \dot{\mathcal{E}}_c) \\ \int (\dot{\mathcal{L}} - \dot{\mathcal{L}}_c) \\ \dot{\mathcal{E}} \\ \dot{\mathcal{L}} \end{bmatrix} = K_{TECS} \begin{bmatrix} \int (\dot{\mathcal{E}}_c - \dot{\mathcal{E}}) \\ \int (\dot{\mathcal{L}}_c - \dot{\mathcal{L}}) \\ \dot{\mathcal{E}} \\ \dot{\mathcal{L}} \end{bmatrix} \quad (9.281)$$

Thus, a TECS uses the two proportional gains,  $K_{y,\dot{\mathcal{E}}}$  and  $K_{y,\dot{\mathcal{L}}}$ , to stabilize the system dynamics and two integral gains,  $K_{e,\dot{\mathcal{E}}}$  and  $K_{e,\dot{\mathcal{L}}}$ , to reject disturbances and achieve zero steady-state error to step commands. These gains as the matrix,  $K_z$ , can be designed using linear control methods, i.e. eigenvalue placement or optimal control. Thus, for control design, one can model a TECS system as the block diagram



where  $\Delta\dot{\tilde{F}}_c$  and  $\Delta\dot{\tilde{L}}_c$  can be computed directly from  $\Delta\gamma_c$ ,  $\Delta v_{\infty,c}$ , and  $\dot{v}_{\infty,c}$  and  $\Delta\dot{\tilde{E}}$  and  $\Delta\dot{\tilde{L}}$  can be computed directly from  $\Delta\gamma$ ,  $\Delta v_\infty$ , and  $\dot{v}_\infty$ .

The  $\Delta\theta_c$  is passed to a SISO inner-loop controller for  $\delta_{e,c}$ , as

$$\Delta\delta_{e,c}(s) = K_\theta(s)\Delta\theta_c(s) \quad (9.282)$$

which is typically designed first to track the  $\Delta\theta_c$  signal faster than the outer TECS loop and improve the stability characteristics of the short-period dynamics. For control design of the TECS outer-loop, the linearized plant will correspond to some LTI state-space model as

$$\begin{aligned}\dot{\vec{x}}_{cl} &= A_{cl}\vec{x}_{cl} + B_{cl}\vec{u} \\ \vec{y} &= C_{cl}\vec{x}_{cl} + D_{cl}\vec{u}\end{aligned}\quad (9.283)$$

where  $\vec{x}_{cl}$  includes the state of the vehicle, elevator actuator, engine, and pitch controller inner-loop closure,  $\vec{u} = [\Delta\delta_T \ \Delta\theta_c]^T$  and  $\vec{y} = [\Delta\dot{\mathcal{E}} \ \Delta\dot{\mathcal{L}}]^T$ .

Next, defining the augmented state as

$$\vec{x}_{aug} = \begin{bmatrix} \int e_{\dot{\mathcal{E}}} \\ \int e_{\dot{\mathcal{L}}} \\ \vec{x}_{cl} \end{bmatrix} \quad (9.284)$$

with output feedback

$$\vec{u} = K_{TECS} (C_{aug}\vec{x}_{aug} + D_{aug}\vec{u}_{aug}) \quad (9.285)$$

one has closed-loop dynamics

$$\begin{aligned}\dot{\vec{x}}_{aug} &= \left( A_{aug} + B_{aug}K_{TECS}(I_2 + D_{aug}K_{TECS})^{-1}C_{aug} \right) \vec{x}_{aug} + \begin{bmatrix} -I_2 \\ 0 \end{bmatrix} \vec{r} \\ \vec{y}_{aug} &= C_{aug}\vec{x}_{aug}\end{aligned}\quad (9.286)$$

where

$$A_{aug} = \begin{bmatrix} 0 & C_{cl} \\ 0 & A_{cl} \end{bmatrix} \quad (9.287)$$

$$B_{aug} = \begin{bmatrix} D_{cl} \\ B_{cl} \end{bmatrix} \quad (9.288)$$

$$C_{aug} = \begin{bmatrix} I_2 & 0 \\ 0 & C_{cl} \end{bmatrix} \quad (9.289)$$

and

$$D_{aug} = \begin{bmatrix} 0 \\ D_{cl} \end{bmatrix} \quad (9.290)$$

## References

For more information, please refer to the following

- Faleiro, L. F., and Lambergts, A. A., “Analysis and Tuning of a ‘Total Energy Control System’ Control Law Using Eigenstructure Assignment,” *Aerospace Science and Technology*, Elsevier, Paris, 1999, pp. 127-140
- Nelson, R. C., “8.2 Aircraft Transfer Functions,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 283-288

- Nelson, R. C., “8.4 Displacement Autopilot,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 292-312
- Nelson, R. C., “8.5 Stability Augmentation,” *Flight Stability and Automatic Control*, 2nd ed., Vol 1., McGraw-Hill, New York, 1997, pp. 312-314
- Schmidt, D. K., “12.3 The Flight-Dynamics Frequency Spectra,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 706-708
- Schmidt, D. K., “12.4 Attitude Control,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 708-737
- Schmidt, D. K., “12.7 Elastic Effects and Structural-Mode Control,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 772-786
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “4.4 Stability Augmentation,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 287-303
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “4.5 The Handling-Qualities Requirements,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 303-322
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “4.6 Autopilots,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 322-344

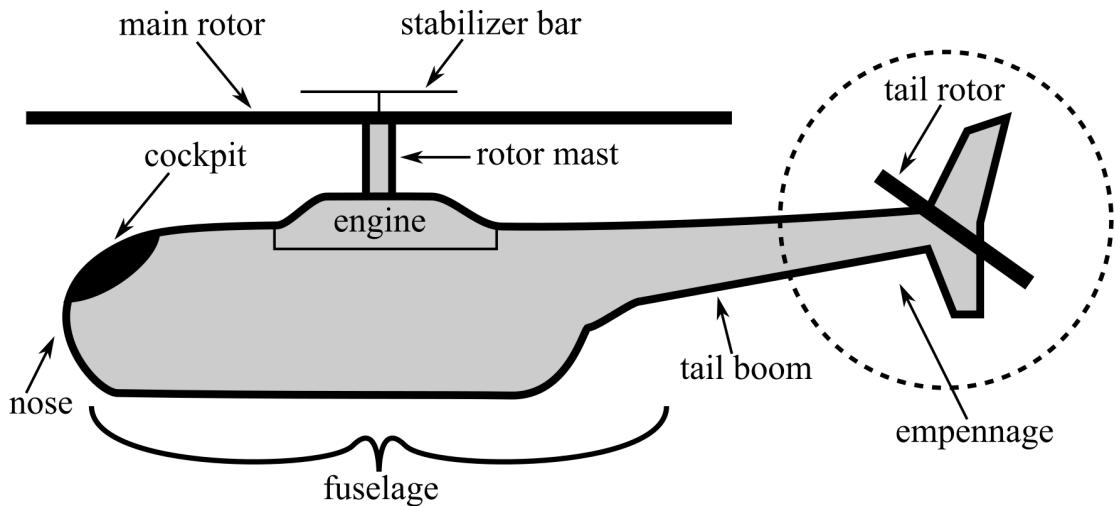
---

# Helicopter Dynamics and Control Systems

## 10.1 Introduction to Helicopters

### Rotary-Wing Vehicle Anatomy

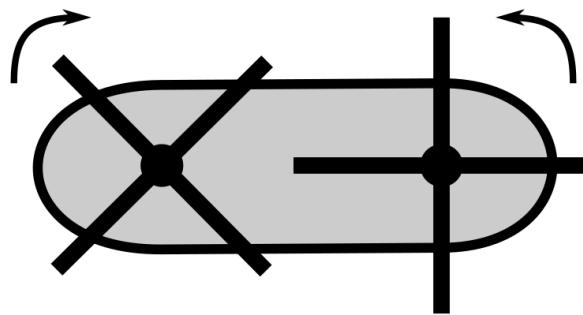
The basic components of a conventional single-rotor **helicopter**, i.e., powered rotary-wing aircraft without propellers, are shown in the following diagram



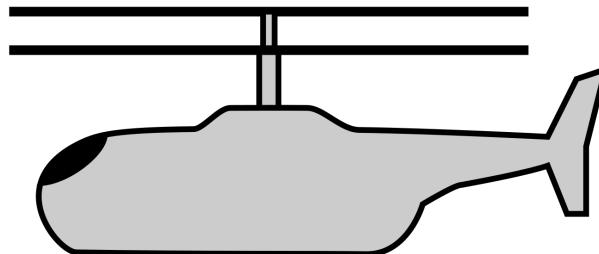
The **nose** is the front of the helicopter and houses the **cockpit** where the pilot(s) and/or other operators are located. The **fuselage** is the structure of the helicopter beneath the main rotor that houses the payload, fuel/batteries, and avionics. The **main rotor** provides lift as the blades rotate to overcome the vehicle's weight and fly. The lift can be changed by the rotor's angular velocity or the rotor's angle of attack, either

**collectively or cyclically**, i.e. at different parts of its rotation cycle. This angle of attack is acutated through a **swashplate** mounted on the **rotor mast** or through rotor servo flaps. This adjustment to the relative blade angle of attack allows for forward and lateral movement of the helicopter. The optional **stabilizer bar** sits above and across the main rotor blade and dampens unwanted vibrations in the main rotor, helping to stabilize the helicopter in all flight conditions. The second **tail rotor** allows for directional control of the helicopter and is required in order to negate the angular momentum produced by the main rotor blades on the helicopter.

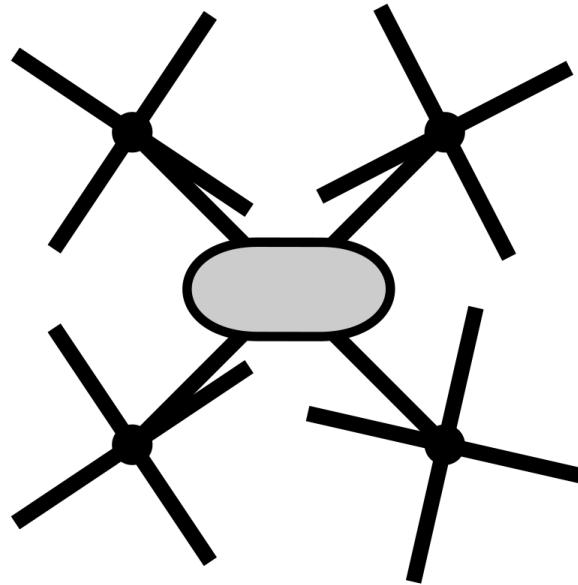
Thus, helicopters are designed to cancel out the rotary-wing angular momentum through tail rotors as shown or with **counter-rotating rotors**, i.e. pairs of rotors which rotate in opposite directions in order to counter each other's angular momentum, e.g. tandem rotor configuration



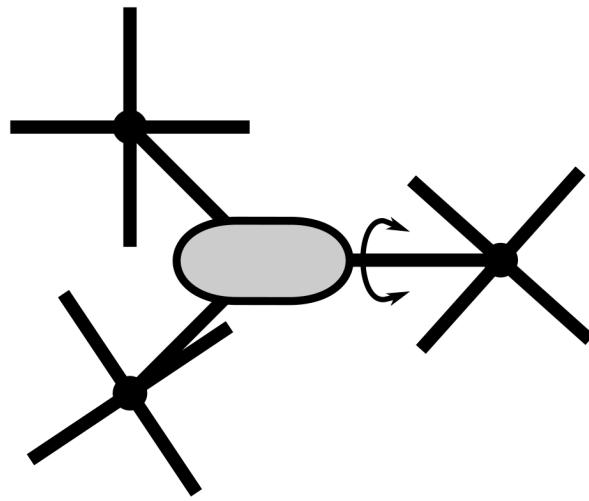
coaxial rotors configuration



or multirotor ( $\geq 3$ ) configurations, e.g. a quadrotor or 4-rotor configuration

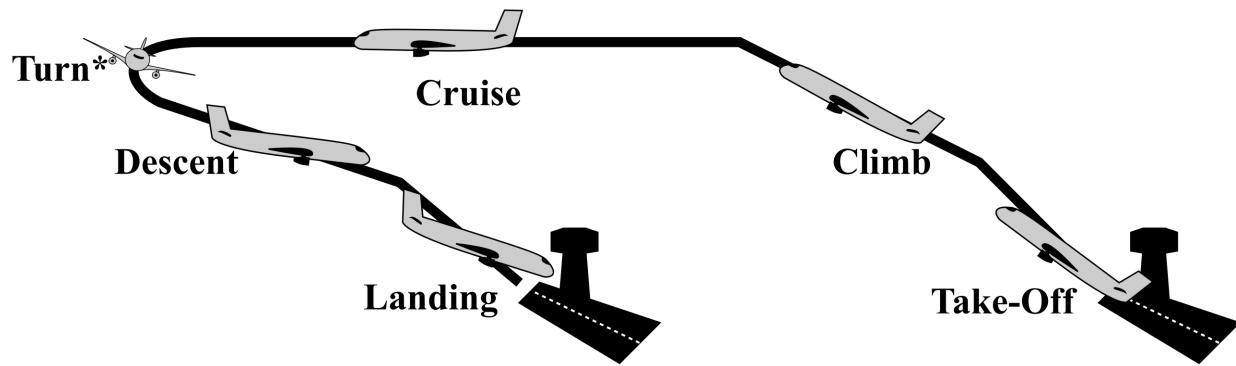


Notably, with  $\geq 4$  rotors, one does not require any rotor angle of attack control, as the angular velocities of each rotor allow for total motion control. For a **tricopter**, a 3-rotor configuration, one of the rotors is typically allowed to rotate about its arm to provide a thrust force with some horizontal component.



### Phases of Flight

The five phases of aircraft flight can generally be described as takeoff, climb, cruise, approach/descent, landing as shown in the following graphic.



where each phase could include maneuvers such as coordinated turns. The aircraft will experience different flight conditions and may be configured differently at each phase, e.g. for takeoff and landing of airplanes, flaps may be deployed and the landing gear will be extended down, which will directly impact the vehicle dynamics.

## 10.2 Point-Mass Dynamics for Helicopters

### Momentum Theory

### Point-Mass Helicopter Steady-Flight

To achieve and control atmospheric flight, rotary-wing aircraft generate thrust aerodynamic forces

### References

For more information, please refer to the following

- Venkatesan, C., “Helicopter Trim (or Equilibrium) Analysis,” *Fundamentals of Helicopter Dynamics*, CRC Press, Florida, 2015, pp. 117-121

## 10.3 Rigid Helicopter Dynamics and Stability

### Traditional Helicopter Flight Dynamics

### Multi-Rotor Helicopter Flight Dynamics

### References

For more information, please refer to the following

- Venkatesan, C., “Helicopter Trim (or Equilibrium) Analysis,” *Fundamentals of Helicopter Dynamics*, CRC Press, Florida, 2015, pp. 117-121

## **10.4 Helicopter Attitude Control Systems**

**Traditional Helicopter Attitude Control Systems**

**Multi-Rotor Helicopter Attitude Control Systems**

---

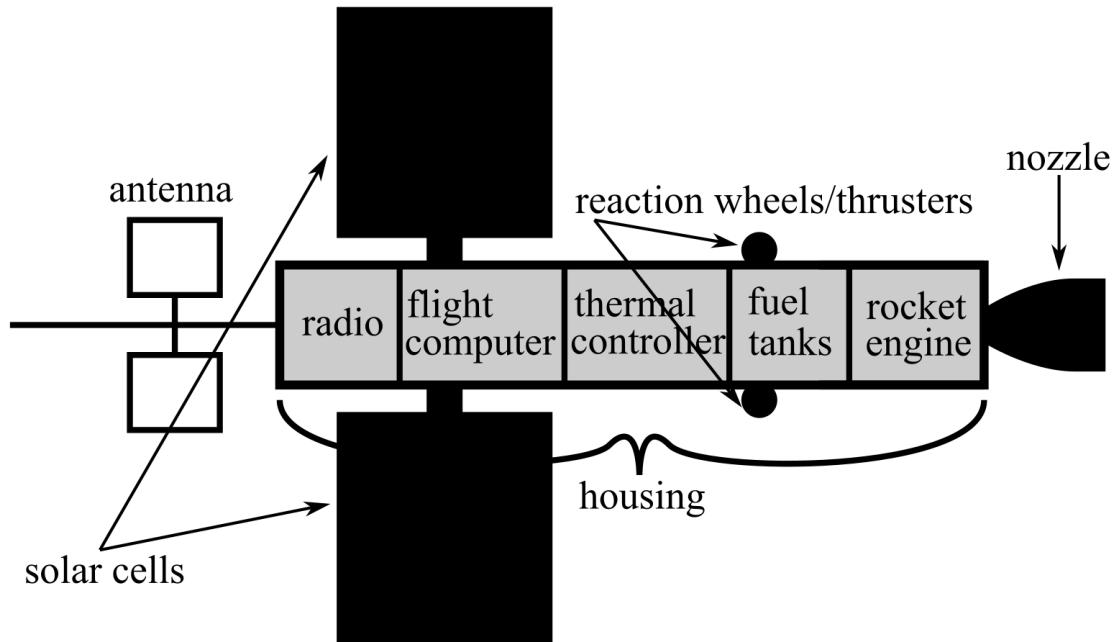
# Orbital Vehicle Dynamics and Control Systems

## 11.1 Introduction to Orbital Vehicles

Space flight can be divided in two types, **orbital** and **sub-orbital**, also known as **ballistic**. Notably, ballistic flight can also be completely atmospheric flight and not a combination of atmospheric and space flight and is treated in the following chapter. In general, spacecraft can be any vehicle that flies beyond the atmosphere which provides a variety of designs, but can typically be separated into orbital and ballistic spacecraft which is the case for this textbook. **Satellites** are objects that travel in **orbits** around celestial bodies. Satellites can be natural, e.g. moons, planets, or **artificial satellites**, also known as **orbital vehicles**.

### Orbital Vehicle Anatomy

The basic components of a conventional orbital vehicle design are shown in the following diagram



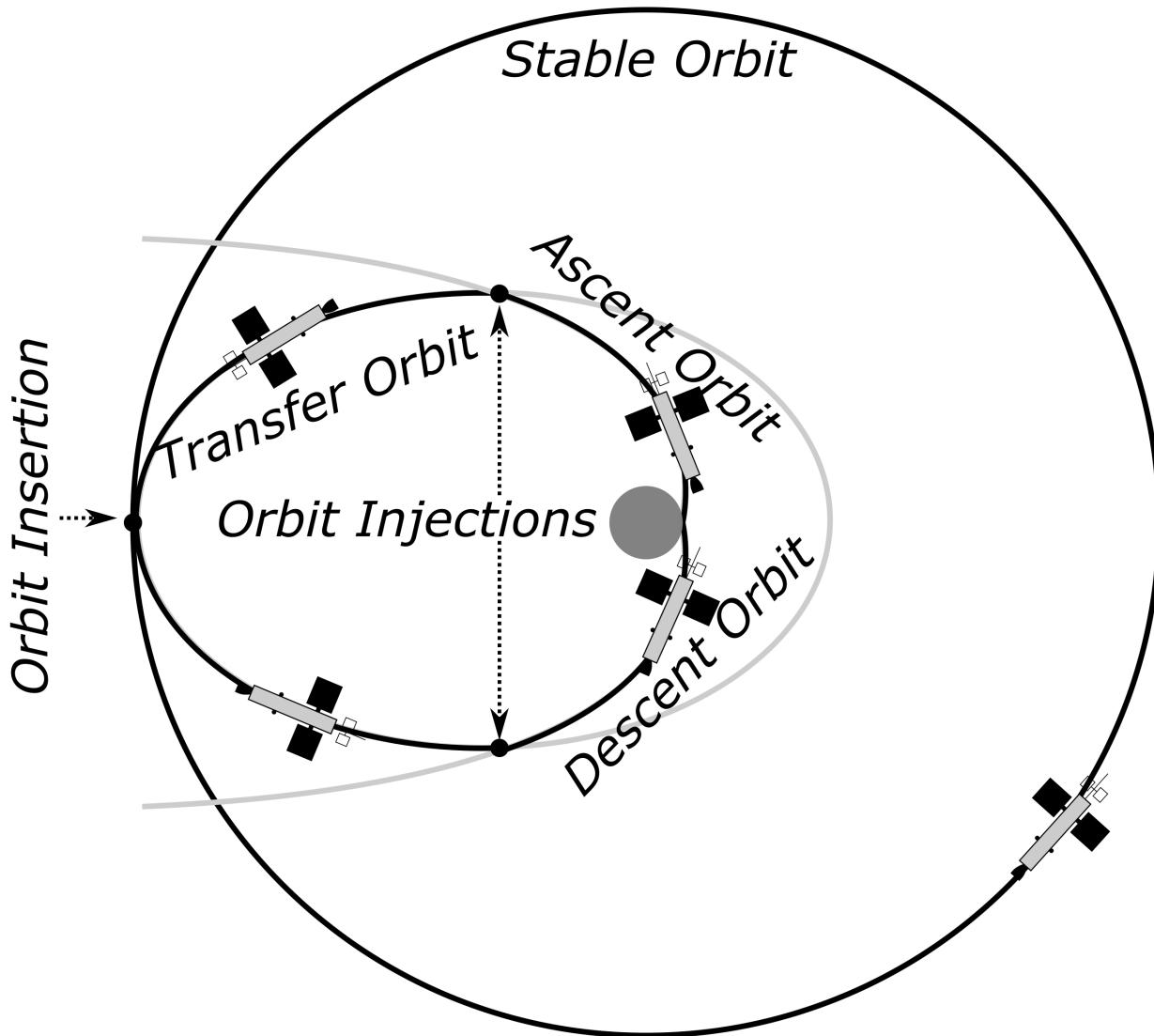
The **housing** houses the rocket engine, fuel tanks, thermal controller, flight computer, the **radio** which communicates between the orbital vehicle and the ground control station and/or other orbital vehicles, and the **antenna** which converts the electrical signal to radio waves and vice versa. **Reaction wheels or thrusters** are placed strategically on the orbital vehicle housing to provide control inputs for maintaining orbit stabilization. Finally, the orbital vehicle is powered using batteries and **solar cells** mounted on the orbital vehicle.

### Phases of Orbital Flight

Orbital flight consists of a sequence of any number of orbits and planned orbital maneuvers that connect these sequences. When not undergoing an orbital maneuver, a spacecraft is in **coast**. An **orbital maneuver**, also known as a **burn**, is the use of propulsion system and/or a gravitational body to change the orbit of a spacecraft and can be characterized as either impulsive, non-impulsive, or gravitational. An **impulsive maneuver** uses the burn to generate a change in velocity almost instantaneously. A **non-impulsive maneuver**, also known as a **finite burn or low-thrust maneuver**, uses the burn to change the momentum slowly over long time by expending a low amount of propulsion. An **insertion maneuver** occurs when the orbital vehicle enters the destination orbit while an **injection maneuver** occurs when the orbital vehicle enters a transfer orbit. A **fly-by maneuver**, also known as an **assist maneuver**, uses non-solar gravitational bodies to alter the solar orbit which can use a burn or not which are called a **powered fly-by** or **Oberth maneuver** and a **gravity assist maneuver**, respectively. An **inclination maneuver** changes an orbital vehicle's inclination angle and an **phasing maneuver** adjusts a spacecraft's true anomaly. A **rendezvous maneuver** is the coordination of two orbital vehicles, called the "chaser" and the "target," arriving at nominally the same position and velocity and approaching to a very close relative distance. Often, these rendezvous maneuvers are accompanied by other coordinated operations in close proximity, as well as docking and undocking. Thus, these are often

grouped together as **rendezvous, proximity operations, docking, and undocking (RPODU)** phases for orbital vehicles.

An example orbital flight plan for an Earth-centered mission is shown in the following graphic composed of an ascent orbit, a transfer orbit, a destination orbit, a descent orbit, an injection maneuver, and an insertion maneuver.



## 11.2 Point-Mass Dynamics for Orbital Vehicles

To achieve space flight, a spacecraft must have enough momentum to maintain an orbit about a celestial body. If this is achieved, **orbital mechanics** are used to describe the classical mechanics derivations when

modeling spacecraft as point-masses about a celestial body. Spacecraft dynamics & control will assume that  $M_{Earth} \gg m$  and the **two-body approximation**, i.e., that Earth is the only gravitational body affecting the spacecraft and thus, the ECI frame is an inertial reference frame. Then, in the absence of aerodynamic forces, the equation of motion of the spacecraft into **coast**, i.e., no propulsion, can be modeled by the force of gravity acting towards the center of the Earth, i.e.

$$\ddot{\vec{r}}_I = \frac{F_g}{m} \frac{\vec{r}_I}{\|\vec{r}_I\|_2^2} = -\frac{\mu}{\|\vec{r}_I\|_2^3} \vec{r}_I \quad (11.1)$$

where  $\vec{r}_I$  is the position vector of the spacecraft relative to the center of the Earth in ECI coordinates. This second-order equation of a three-dimensional vector can be solved explicitly given six initial conditions, e.g.  $\dot{\vec{r}}_0 = \vec{v}_0$  and  $\vec{r}_0$ . Thus, if one knows the position and velocity of a spacecraft at any point, one can determine the reference trajectory of that spacecraft. Furthermore, one is typically interested in the **orbital angular momentum** defined as

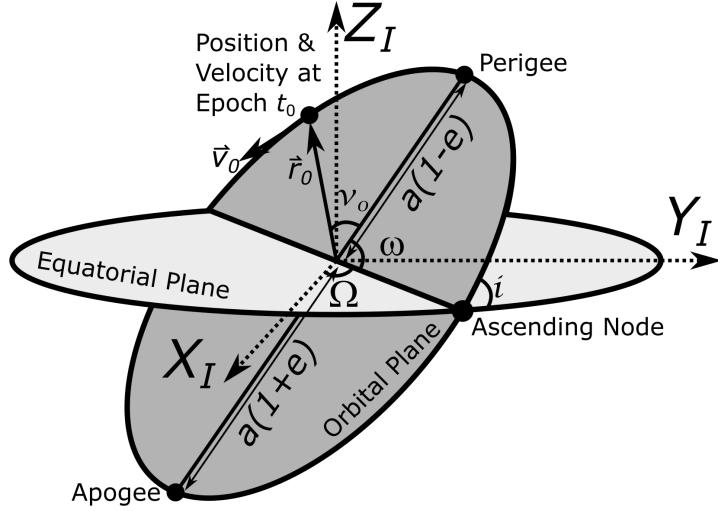
$$\vec{h}_I = [\vec{r}_I] \times \vec{v}_I \quad (11.2)$$

which by taking the cross product of the equation of motion with the angular momentum on both sides and integrating to obtain a constant of integration, one can show that the

$$\frac{[\vec{v}_I] \times \vec{h}_I}{\mu} - \frac{\vec{r}_I}{\|\vec{r}_I\|_2} = \vec{e} \quad (11.3)$$

where  $\vec{e}$  is called the **eccentricity vector** and whose magnitude,  $e = \|\vec{e}\|_2$  determines the type of orbit for the spacecraft. Namely, if  $0 \leq e < 1$ , the spacecraft is in a *stable* elliptical orbit and if  $e \geq 1$ , the spacecraft is in a hyperbolic or parabolic orbit and will eventually escape the gravitational attraction of Earth, i.e. it is not a stable trajectory around Earth.

For spacecraft in an elliptical orbit, instead of specifying the position and velocity at some **epoch**, or reference time,  $t_0$ , one can represent their reference trajectory about a celestial body by six **orbital elements**, also known as the **Keplerian elements**, which are depicted about that celestial body as



and defined as:

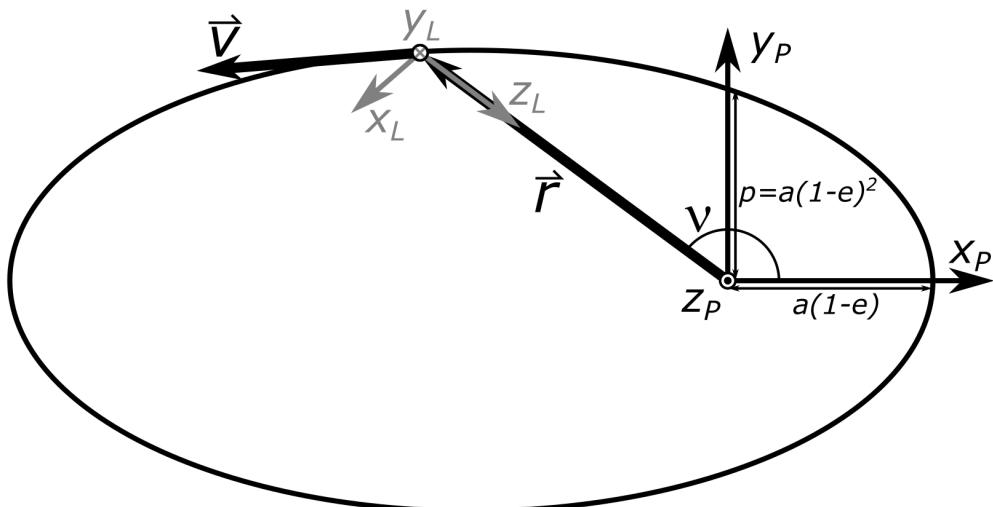
- the **eccentricity**,  $0 < e < 1$ , which is a measure of the elongation of the elliptical orbit compared to a circular orbit ( $e = 0$ );
- the **semi-major axis**,  $a$ , which is half the distance between the **periapsis** and **apoapsis**, i.e. the closest and further points from the focus corresponding to the planet's center of gravity;
- the **inclination**,  $i$ , which is the angle between the reference and orbital planes at the ascending node where the satellite moves from beneath the equatorial plane to above;
- the **longitude of ascending node**,  $\Omega$ , which is the angle relative to the reference direction where the orbital plane intersects the reference plane;
- the **argument of periapsis**,  $\omega$ , which is the angle in the orbital plane from the ascending node to periapsis; and
- the **true anomaly**,  $v_0$ , which is the angular coordinate at some **epoch**, or time,  $t_0$ .

For **geocentric orbits**, i.e. the celestial body is Earth,  $\omega$  is also known as the **argument of perigee** and  $\Omega$  is the **right ascension of the ascending node** if the reference plane is the equatorial plane and the reference direction is the “first point of Aries”, i.e., the ICRF/GCRF  $x$ -axis.

The orbital plane allows one to formally define the **perifocal frame** (subscript  $P$ ), which is defined as

- the origin is at the focus corresponding to the celestial body's center of mass;
- the  $x_P$ -axis is colinear with a line connecting the two foci pointing to the periapsis;
- the  $y_P$ -axis is orthogonal to  $x_P$ - and  $z_P$ -axes by the right hand rule; and
- the  $z_P$ -axis is normal to the orbital plane in the direction of the orbital angular momentum vector.

This frame can be depicted as



The  $3 - 1 - 3$  Euler angles for describing the perifocal frame relative to the ECI frame are the **longitude of the ascending node**,  $\Omega$ , **inclination**,  $i$ , and **argument of periapsis**,  $\omega$ . Thus, a vector expressed in ECI coordinates,  $\vec{v}_I$ , can be expressed as a vector in perifocal frame coordinates,  $\vec{v}_P$ , through the sequence

$$\vec{v}_P = C_3(\omega)C_1(i)C_3(\Omega)\vec{v}_I \quad (11.4)$$

$$\vec{v}_P = C_{P \leftarrow I}\vec{v}_I \quad (11.5)$$

and the DCM for the body-fixed frame relative to the vehicle-centered, inertial frame can be computed via

$$C_{P \leftarrow I} = \begin{bmatrix} \cos \omega \cos \Omega - \cos i \sin \omega \sin \Omega & \sin \omega \cos \Omega - \cos i \cos \omega \sin \Omega & \sin i \sin \Omega \\ -\cos \omega \sin \Omega - \cos i \sin \omega \cos \Omega & -\sin \omega \sin \Omega - \cos i \cos \omega \cos \Omega & \sin i \cos \Omega \\ \sin i \sin \Omega & -\sin i \cos \Omega & \cos i \end{bmatrix} \quad (11.6)$$

The offset vector between the nadir-pointing LVLH frame and the perifocal frame is  $\vec{o}_{LVLH \leftarrow P} = \vec{r}_P$ .

The  $3 - 1$  Euler angles for describing the LVLH frame of the satellite relative to the perifocal frame are the

$$\vec{v}_L = C_1(-90^\circ)C_3(\nu + 90^\circ)\vec{v}_P \quad (11.7)$$

$$\vec{v}_L = C_{L \leftarrow P}\vec{v}_P \quad (11.8)$$

The DCM for the perifocal frame relative to the nadir-pointing LVLH frame of the satellite is given by

$$C_{L \leftarrow P} = \begin{bmatrix} -\sin \nu & \cos \nu & 0 \\ 0 & 0 & -1 \\ -\cos \nu & -\sin \nu & 0 \end{bmatrix} \quad (11.9)$$

Thus, if the perifocal frame is constant, i.e.  $\dot{a} = \dot{e} = \dot{i} = \dot{\Omega} = \dot{\omega} = 0$ , the position of the satellite in the perifocal frame is given by

$$\vec{r}_P = \begin{bmatrix} \frac{p \cos \nu}{1+e \cos \nu} \\ \frac{p \sin \nu}{1+e \cos \nu} \\ 0 \end{bmatrix} \quad (11.10)$$

where  $p$  is the **semi-latus rectum**

$$p = a(1 - e^2) \quad (11.11)$$

the velocity of the satellite in the perifocal frame is given by

$$\vec{v}_P = \begin{bmatrix} -\sqrt{\frac{\mu}{p}} \sin \nu \\ \sqrt{\frac{\mu}{p}}(e + \cos \nu) \\ 0 \end{bmatrix} \quad (11.12)$$

the angular velocity of the LVLH frame with respect to the perifocal frame can be expressed as

$$\vec{\omega}_{L \leftarrow P, L} = C_1(90^\circ)C_2(-180^\circ)[\vec{r}_P] \times \vec{v}_P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \frac{n(1+e \cos \nu)^2}{(1-e^2)^{3/2}} \end{bmatrix} \quad (11.13)$$

or

$$\vec{\omega}_{L \leftarrow P, L} = \begin{bmatrix} 0 \\ -\omega_O \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{n(1+e \cos \nu)^2}{(1-e^2)^{3/2}} \\ 0 \end{bmatrix} \quad (11.14)$$

where  $\omega_O$  is the **orbital angular velocity** and  $n$  is the **orbital mean motion** defined as

$$n = \frac{\mu}{a^3} \quad (11.15)$$

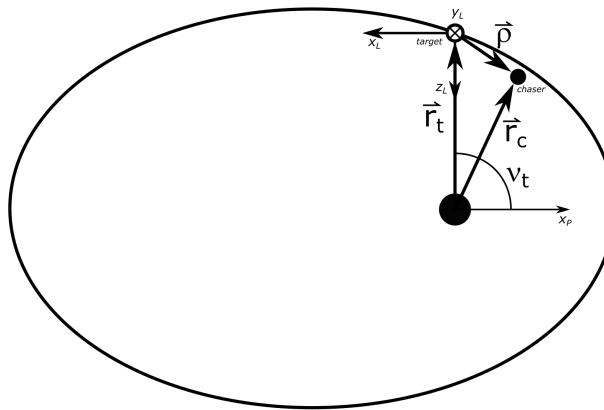
Lastly, the angular acceleration of the LVLH frame with respect to the perifocal frame can be expressed as

$$\vec{\alpha}_{L \leftarrow P, L} = \begin{bmatrix} 0 \\ 0 \\ \alpha_O \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \frac{-2\mu e \sin \nu}{(\frac{p}{1+e \cos \nu})^3} \end{bmatrix} \quad (11.16)$$

where  $\alpha_O$  is the **orbital angular acceleration**. Note that if one has a circular orbit of radius  $a$ , i.e.,  $e = 0$ , then  $\omega_O = n = \frac{\mu}{a^3}$  and  $\alpha_O = 0$ .

## Relative Orbital Dynamics

In many instances for spacecraft dynamics and control, one is interested in modeling the dynamics between multiple spacecraft, e.g., a **rendezvous and proximity operation (RPO)**. To that end, consider the following simplified model of two satellites operating in proximity, i.e. a simplified **three-body problem**, involving a **chaser spacecraft** (subscript  $c$ ) and a **target spacecraft** (subscript  $t$ ) on an elliptical orbit. One can represent this relative motion in the LVLH reference frame for the target spacecraft as shown in the following figure.



In this context, this frame is sometimes known as the Clohessy-Wiltshire (CW) or Hill frame (HF). This provides the relationship

$$\vec{\rho} = -\vec{r}_t + \vec{r}_c \quad (11.17)$$

Then, using Newton's Law of Gravitation, assuming equal external forces on the target and chaser, one has **relative orbital dynamics** of the chaser in an inertial reference frame as

$$\frac{d^2}{dt^2} \vec{\rho} = \frac{\mu}{\|\vec{r}_t\|_2^3} \vec{r}_t - \frac{\mu}{\|\vec{r}_t + \vec{\rho}\|_2^3} (\vec{r}_t + \vec{\rho}) + \vec{a}_c \quad (11.18)$$

where  $\vec{a}_c$  are the acceleration due to thrust forces of the chaser. Simplifying, one has

$$\frac{d^2}{dt^2} \vec{\rho} = \frac{\mu}{\|\vec{r}_t\|_2^3} \vec{r}_t - \frac{\mu}{(\|\vec{r}_t\|_2^2 + 2\vec{r}_t \cdot \vec{\rho} + \|\vec{\rho}\|_2^2)^{3/2}} (\vec{r}_t + \vec{\rho}) + \vec{a}_c \quad (11.19)$$

$$\frac{d^2}{dt^2} \vec{\rho} = \frac{\mu}{\|\vec{r}_t\|_2^3} \left( \vec{r}_t - \frac{\vec{r}_t + \vec{\rho}}{\left(1 + \frac{2\vec{r}_t \cdot \vec{\rho}}{\|\vec{r}_t\|_2^2} + \frac{\|\vec{\rho}\|_2^2}{\|\vec{r}_t\|_2^2}\right)^{3/2}} \right) + \vec{a}_c \quad (11.20)$$

which for  $\vec{\rho} \ll \vec{r}_t$ , one has the linearized equation

$$\frac{d^2}{dt^2} \vec{\rho} = \frac{\mu}{\|\vec{r}_t\|_2^3} \left( \vec{r}_t - \frac{\vec{r}_t + \vec{\rho}}{\left(1 + \frac{2\vec{r}_t \cdot \vec{\rho}}{\|\vec{r}_t\|_2^2}\right)^{3/2}} \right) + \vec{a}_c \quad (11.21)$$

which can be further simplified using the binomial theorem

$$\frac{d^2}{dt^2} \vec{\rho} = \frac{\mu}{\|\vec{r}_t\|_2^3} \left( \vec{r}_t - \left(1 - 3\frac{\vec{r}_t \cdot \vec{\rho}}{\|\vec{r}_t\|_2^2}\right) (\vec{r}_t + \vec{\rho}) \right) + \vec{a}_c \quad (11.22)$$

and simplifying and dropping higher-order terms in  $\vec{\rho}$ , one has

$$\frac{d^2}{dt^2} \vec{\rho} = -\frac{\mu}{\|\vec{r}_t\|_2^3} \left( \vec{\rho} - \left(3\frac{\vec{r}_t \cdot \vec{\rho}}{\|\vec{r}_t\|_2^2}\right) \vec{r}_t \right) + \vec{a}_c \quad (11.23)$$

If using the LVLH frame for defining  $\vec{\rho}$  one has

$$\ddot{\vec{\rho}} + 2[\vec{\omega}_{L/I}] \times \dot{\vec{\rho}} + [\vec{\alpha}_{L/I}] \times \vec{\rho} + [\vec{\omega}_{L/I}] \times [\vec{\omega}_{L/I}] \times \vec{\rho} = -\frac{\mu}{\|\vec{r}_t\|_2^3} \left( \vec{\rho} - \left(3\frac{\vec{r}_t \cdot \vec{\rho}}{\|\vec{r}_t\|_2^2}\right) \vec{r}_t \right) + \vec{a}_c \quad (11.24)$$

Next, one can denote the target position relative to the celestial body in the target's LVLH frame as

$$\vec{r}_t = \begin{bmatrix} 0 \\ 0 \\ -\|\vec{r}_t\|_2 \end{bmatrix} \quad (11.25)$$

the relative position of the chaser in the target's LVLH frame as

$$\vec{\rho} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad (11.26)$$

the target's orbital angular velocity in the target's LVLH frame as

$$\vec{\omega}_{L/I} = \begin{bmatrix} 0 \\ -\omega_t \\ 0 \end{bmatrix} \quad (11.27)$$

and the acceleration of the chaser as the input

$$\vec{a}_c = \vec{u} \quad (11.28)$$

Then, one has for the linearized dynamics

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} + 2 \begin{bmatrix} 0 \\ -\omega_t \\ 0 \end{bmatrix} \times \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} + \begin{bmatrix} 0 \\ -\omega_t \\ 0 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} 0 \\ -\omega_t \\ 0 \end{bmatrix} \times \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = -\frac{\mu}{\|\vec{r}_t\|_2^3} \left( \begin{bmatrix} x \\ y \\ z \end{bmatrix} - \left( \frac{3}{\|\vec{r}_t\|_2^2} \right) \left( \begin{bmatrix} 0 \\ 0 \\ -\|\vec{r}_t\|_2 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} \right) \begin{bmatrix} 0 \\ 0 \\ -\|\vec{r}_t\|_2 \end{bmatrix} \right) + \vec{u} \quad (11.29)$$

Performing the matrix multiplications, one has

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} + 2 \begin{bmatrix} -\omega_t \dot{z} \\ 0 \\ \omega_t \dot{x} \end{bmatrix} + \begin{bmatrix} -\dot{\omega}_t z \\ 0 \\ \dot{\omega}_t x \end{bmatrix} + \begin{bmatrix} -\omega_t^2 x \\ 0 \\ -\omega_t^2 z \end{bmatrix} = -\frac{\mu}{\|\vec{r}_t\|_2^3} \left( \begin{bmatrix} x \\ y \\ z \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 3z \end{bmatrix} \right) + \vec{u} \quad (11.30)$$

Finally, defining the constant

$$k = \frac{\mu}{h_t^{3/2}} \quad (11.31)$$

where  $h_t$  is the target's constant orbital angular momentum and substituting for  $\frac{\mu}{\|\vec{r}_t\|_2^3} = k\omega_t^{3/2}$ , one has the **linearized relative orbital dynamics**

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = \begin{bmatrix} (\omega_t^2 - k\omega_t^{3/2})x + 2\omega_t \dot{z} + \dot{\omega}_t z \\ -k\omega_t^{3/2}y \\ (\omega_t^2 + 2k\omega_t^{3/2})z - 2\omega_t \dot{x} - \dot{\omega}_t x \end{bmatrix} + \vec{u} \quad (11.32)$$

which are linear 6-DOF equations in  $\vec{r}$  and time-varying due to the  $\omega_t$  parameter.

## Relative Orbital Dynamics for Circular Orbit

The time-invariant solution to the linearized relative orbital dynamics equations occurs if one assumes a circular orbit for the target spacecraft. In this case,  $\dot{\omega}_t = 0$  and

$$\omega_t = n_t = \sqrt{\frac{\mu}{\|\vec{r}_t\|_2^3}} = \sqrt{k\omega_t^{3/2}} \quad (11.33)$$

where  $n_t$  is known as the target's **mean motion**. By substitution, one obtains the **Clohessy-Wiltshire (CW) equations** for artificial satellites or the **Hill equations** for natural satellites, in the target's LVLH frame as

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = \begin{bmatrix} 2n_t \dot{z} \\ -n_t^2 y \\ 3n_t^2 z - 2n_t \dot{x} \end{bmatrix} + \vec{u} \quad (11.34)$$

which can be rearranged to obtain the continuous-time LTI state-space representation as

$$\dot{\vec{x}} = A\vec{x} + B\vec{u} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2n_t \\ 0 & -n_t^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3n_t^2 & -2n_t & 0 & 0 \end{bmatrix} \vec{x} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \vec{u} \quad (11.35)$$

where

$$\vec{x} = \begin{bmatrix} \vec{\rho} \\ \dot{\vec{\rho}} \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \\ \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} \quad (11.36)$$

These are also known as the Clohessy-Wiltshire-Hill (CWH) and Hill-Clohessy-Wiltshire (HCW) equations. These equations can be solved analytically. However, the explicit result is left to the references.

### Relative Orbital Dynamics for Elliptical Orbit

Another useful solution to the linearized relative orbital dynamics equations occurs if one changes the independent variable from time  $t$  to the target's true anomaly,  $\nu_t$ , which is possible as the true anomaly is a monotonically increasing variable with time. In this case, the first and second derivatives of a variable  $v$  with respect to time,  $t$ , becomes

$$\frac{dv}{dt} = \frac{dv}{d\nu_t} \frac{d\nu_t}{dt} = \omega_t \frac{dv}{d\nu_t} \quad (11.37)$$

and

$$\frac{d^2v}{dt^2} = \omega_t \frac{d}{d\nu_t} \left( \omega_t \frac{dv}{d\nu_t} \right) = \omega_t \frac{d\omega_t}{d\nu_t} \frac{dv}{d\nu_t} + \nu_t^2 \frac{d^2v}{d\nu_t^2} \quad (11.38)$$

Defining  $dv/d\nu_t = v'$ , one has

$$\dot{v} = \omega_t v' \quad (11.39)$$

and

$$\ddot{v} = \omega_t \omega'_t v' + \omega_t^2 v'' \quad (11.40)$$

With these equations, one can rewrite the linearized relative orbital dynamics as

$$\begin{bmatrix} \omega_t \omega'_t x' + \omega_t^2 x'' \\ \omega_t \omega'_t y' + \omega_t^2 y'' \\ \omega_t \omega'_t z' + \omega_t^2 z'' \end{bmatrix} = \begin{bmatrix} \left( \omega_t^2 - k\omega_t^{3/2} \right) x + 2\omega_t^2 z' + \omega_t \omega'_t z \\ -k\omega_t^{3/2} y \\ \left( \omega_t^2 + 2k\omega_t^{3/2} \right) z - 2\omega_t^2 x' - \omega_t \omega'_t x \end{bmatrix} + \vec{u} \quad (11.41)$$

Next, one can define  $\varsigma = 1 + e_t \cos \nu_t$  where  $e_t$  is the target's eccentricity which provides the relationships

$$\omega_t = k^2 \varsigma^2 \quad (11.42)$$

and

$$\omega_t' = 2k^2 \varsigma \varsigma' = -2k^2 e_t \sin \nu_t \varsigma \quad (11.43)$$

Then, by substitution, one has

$$\begin{bmatrix} -2k^4 e_t \sin \nu_t \varsigma^3 x' + k^4 \varsigma^4 x'' \\ -2k^4 e_t \sin \nu_t \varsigma^3 y' + k^4 \varsigma^4 y'' \\ -2k^4 e_t \sin \nu_t \varsigma^3 z' + k^4 \varsigma^4 z'' \end{bmatrix} = \begin{bmatrix} (k^4 \varsigma^4 - k^4 \varsigma^3) x + 2k^4 \varsigma^4 z' - 2k^4 e_t \sin \nu_t \varsigma^3 z \\ -k^4 \varsigma^3 y \\ (k^4 \varsigma^4 + 2k^4 \varsigma^3) z - 2k^4 \varsigma^4 x' + 2k^4 e_t \sin \nu_t \varsigma^3 x \end{bmatrix} + \vec{u} \quad (11.44)$$

and dividing each equation by  $k^4 \varsigma^3$ , one has

$$\begin{bmatrix} \varsigma x'' - 2e_t \sin \nu_t x' \\ \varsigma y'' - 2e_t \sin \nu_t y' \\ \varsigma z'' - 2e_t \sin \nu_t z' \end{bmatrix} = \begin{bmatrix} (\varsigma - 1) x + 2\varsigma z' - 2e_t \sin \nu_t z \\ -y \\ (\varsigma + 2) z - 2\varsigma x' + 2e_t \sin \nu_t x \end{bmatrix} + \frac{1}{k^4 \varsigma^3} \vec{u} \quad (11.45)$$

or by back-substitution for  $\varsigma = 1 + e \cos \nu_t$ , one obtains

$$\begin{bmatrix} \varsigma x'' - 2e_t \sin \nu_t x' - e \cos \nu_t x \\ -\varsigma y'' - 2e_t \sin \nu_t y' \\ \varsigma z'' - 2e_t \sin \nu_t z' - e_t \cos \nu_t z \end{bmatrix} = \begin{bmatrix} 2\varsigma z' - 2e_t \sin \nu_t z \\ -y \\ 3z - 2\varsigma x' + 2e_t \sin \nu_t x \end{bmatrix} + \frac{1}{k^4 \varsigma^3} \vec{u} \quad (11.46)$$

Finally, transforming the relative coordinates as

$$\tilde{\rho} = \begin{bmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{bmatrix} = \begin{bmatrix} \varsigma x \\ \varsigma y \\ \varsigma z \end{bmatrix} \quad (11.47)$$

with first-order and second-order derivatives as

$$\begin{bmatrix} \tilde{x}' \\ \tilde{y}' \\ \tilde{z}' \end{bmatrix} = \begin{bmatrix} \varsigma x' - e_t \sin \nu_t x \\ \varsigma y' - e_t \sin \nu_t y \\ \varsigma z' - e_t \sin \nu_t z \end{bmatrix} \quad (11.48)$$

and

$$\begin{bmatrix} \tilde{x}'' \\ \tilde{y}'' \\ \tilde{z}'' \end{bmatrix} = \begin{bmatrix} \varsigma x'' - 2e_t \sin \nu_t x' - e_t \cos \nu_t x \\ \varsigma y'' - 2e_t \sin \nu_t y' - e_t \cos \nu_t y \\ \varsigma z'' - 2e_t \sin \nu_t z' - e_t \cos \nu_t z \end{bmatrix} \quad (11.49)$$

one obtains the **Tschauner-Hempel (TH) equations** as

$$\begin{bmatrix} \tilde{x}'' \\ \tilde{y}'' \\ \tilde{z}'' \end{bmatrix} = \begin{bmatrix} 2\tilde{z}' \\ -\tilde{y} \\ \frac{3}{1+e_t \cos \nu_t} \tilde{z} - 2\tilde{x}' \end{bmatrix} + \frac{1}{k^4(1+e_t \cos \nu_t)^3} \vec{u} \quad (11.50)$$

which can be rearranged into the continuous-time LTV state-space representation with “time” as  $\nu_t$  as

$$\tilde{x}' = A(\nu_t) \tilde{x} + B(\nu_t) \vec{u} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{1+e_t \cos \nu_t} & -2 & 0 & 0 \end{bmatrix} \tilde{x} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{1}{k^4(1+e_t \cos \nu_t)^3} & 0 & 0 \\ 0 & \frac{1}{k^4(1+e_t \cos \nu_t)^3} & 0 \\ 0 & 0 & \frac{1}{k^4(1+e_t \cos \nu_t)^3} \end{bmatrix} \vec{u} \quad (11.51)$$

where

$$\tilde{\vec{x}} = \begin{bmatrix} \tilde{\vec{r}} \\ \tilde{\vec{p}}' \end{bmatrix} = \begin{bmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \\ \tilde{x}' \\ \tilde{y}' \\ \tilde{z}' \end{bmatrix} \quad (11.52)$$

These equations can also be solved analytically. However, the explicit result is left to the references, e.g., the solution given by Yamanaka and Ankerson provides a solution for the state transition matrix for all eccentricities  $0 \leq e_t \leq 1$ .

## References

For more information, please refer to the following

- Curtis, H. D., “2.4 Angular Momentum and the Orbit Formulas,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 42-50
- Curtis, H. D., “2.7 Elliptical Orbits ( $0 < e < 1$ )”, *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 76-80
- Curtis, H. D., “4.6 Transformation between Geocentric Equatorial and Perifocal Frames,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 172-176
- Curtis, H. D., “7.2 Relative Motion in Orbit,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 316-322
- Curtis, H. D., “7.3 Linearization of the Equations of Relative Motion in Orbit,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 322-324
- Curtis, H. D., “7.4 Clohessy-Wiltshire Equations,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 324-329
- Li, Y., Liu, X., and Zing, G., “Discrete-time LQ optimal control of satellite formations in elliptical orbits based on feedback linearization,” in *Acta Astronautica*, Vol. 83, 2013
- Pesce, V., Colagrossi, A., and Silvestrini, S., “Chapter Two - Reference Systems and Planetary Models,” *Modern Spacecraft Guidance, Navigation, and Control*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2023, pp. 45-75
- Sullivan, J., Grimberg, S., and D’Amico, S., “Comprehensive Survey and Assessment of Spacecraft Relative Motion Dynamics Models,” in *Journal of Guidance, Control, and Dynamics*, Vol. 40, No. 8, 2017
- Yamanaka, K., and Ankerson, F., “New State Transition Matrix for Relative Motion on an Arbitrary Elliptical Orbit,” in *Journal of Guidance, Control, and Dynamics*, Vol. 25, No. 1, 2002

## 11.3 Orbital Maneuvers

**non-impulsive maneuvers**

**Impulsive Maneuvers**

A standard Hohmann transfer

- A generalized Hohmann transfer
- A bielliptic Hohmann transfer
- A phasing maneuver
- A non-Hohmann transfer
- A apse line rotation
- A plane change maneuver

## References

For more information, please refer to the following

- Curtis, H. D., “Chapter 6 Orbital Maneuvers,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 287-334

## 11.4 Rigid Orbital Vehicle Dynamics and Stability

As spacecraft are typically designed in a wide variety of shapes, the inertia tensor varies depending on where one places the body-fixed frame axes. The **principal body-fixed frame** is defined as the axes for which the inertia tensor is diagonal, i.e.

$$I_G = \begin{bmatrix} I_{xx} & 0 & 0 \\ 0 & I_{yy} & 0 \\ 0 & 0 & I_{zz} \end{bmatrix} \quad (11.53)$$

where  $I_{xx}$ ,  $I_{yy}$ , and  $I_{zz}$  are the **principal moments of inertia** where the largest to smallest in magnitude are known as the major, intermediate, and minor moments of inertia and axes. Note that given another body-fixed frame with inertia tensor,  $I'_G$ , one can use the eigenvalue decomposition of the *symmetric* matrix  $I'_G$  to form  $I_G$  by

$$I_G = V I'_G V^T \quad (11.54)$$

$$\begin{bmatrix} I_{xx} & 0 & 0 \\ 0 & I_{yy} & 0 \\ 0 & 0 & I_{zz} \end{bmatrix} = [\vec{v}_1 \quad \vec{v}_2 \quad \vec{v}_3] I'_G \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vec{v}_3^T \end{bmatrix} \quad (11.55)$$

where  $I_{xx}$ ,  $I_{yy}$ , and  $I_{zz}$  are the eigenvalues of  $I'_G$  with corresponding right eigenvectors,  $\vec{v}_1$ ,  $\vec{v}_2$ ,  $\vec{v}_3$ , respectively.

Thus, for the angular momentum of the satellite in principal body-fixed frame coordinates, one has

$$\vec{H}_{G,B} = I_G \vec{\omega}_{B/I,B} = \begin{bmatrix} I_{xx}p \\ I_{yy}q \\ I_{zz}r \end{bmatrix} \quad (11.56)$$

Finally, assuming principal body-fixed frame for the Newton-Euler equations of motion, one obtains the **satellite attitude equations of motion**

$$\vec{M}_B = \begin{bmatrix} I_{xx}\dot{p} + (I_{zz} - I_{yy})qr \\ I_{yy}\dot{q} + (I_{xx} - I_{zz})pr \\ I_{zz}\dot{r} + (I_{yy} - I_{xx})pq \end{bmatrix} \quad (11.57)$$

Also, note that the translation of the satellite is only governed by orbital mechanics as described in the inertial frame and is decoupled from the attitude dynamics. Here the inertial velocity can be rotated to body-fixed frame if one wishes to express the apparent body-fixed frame velocities, typically using the orbital elements as described previously and the angular velocities from the attitude dynamics.

### Rotating Axisymmetric Spacecraft Dynamics in Torque-Free Motion

If one assumes  $\vec{M}_{g,B} = 0$ , i.e., the spacecraft is undergoing **torque-free motion**, one has

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \frac{d}{dt} \vec{H}_G = \begin{bmatrix} I_{xx}\dot{p} + (I_{zz} - I_{yy})qr \\ I_{yy}\dot{q} + (I_{xx} - I_{zz})pr \\ I_{zz}\dot{r} + (I_{yy} - I_{xx})pq \end{bmatrix} \quad (11.58)$$

Without loss of generality, one can assume  $\vec{H}_G$  defines the  $z_I$ -axis. Then, by definition of the 3 – 1 – 3 Euler angles, the nutation angle,  $\theta_n$ , defines the angle between the  $z_B$ -axis and  $\vec{H}_G$ . Thus, one can state using the dot product

$$\cos \theta_n = \frac{\vec{H}_G}{\|\vec{H}_G\|_2} \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \frac{1}{\|\vec{H}_G\|_2} \begin{bmatrix} I_{xx}p \\ I_{yy}q \\ I_{zz}r \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (11.59)$$

Thus, one has

$$\cos \theta_n = \frac{1}{\|\vec{H}_G\|_2} I_{zz}r \quad (11.60)$$

or

$$\|\vec{H}_G\|_2 = \frac{I_{zz}r}{\cos \theta_n} \quad (11.61)$$

Taking the derivative with respect to time and assuming torque-free motion, one has

$$\frac{d \cos \theta_n}{dt} = \frac{1}{\|\vec{H}_G\|_2} \begin{bmatrix} I_{xx}p \\ I_{yy}q \\ I_{zz}r \end{bmatrix} \cdot \left( [\vec{\omega}_{B/I,B}] \times \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) \quad (11.62)$$

$$-\sin \theta_n \dot{\theta}_n = \frac{1}{\|\vec{H}_G\|_2} \begin{bmatrix} I_{xx}p \\ I_{yy}q \\ I_{zz}r \end{bmatrix} \cdot \begin{bmatrix} q \\ -p \\ 0 \end{bmatrix} \quad (11.63)$$

or

$$\dot{\theta}_n = \omega_n = -\frac{(I_{xx} - I_{yy})pq}{\|\vec{H}_G\|_2 \sin \theta_n} \quad (11.64)$$

Thus, the first aspect to note for torque-free motion is that the nutation rate,  $\omega_n$ , vanishes only if  $I_{xx} = I_{yy}$ , i.e., the  $z_B$ -axis is an axis of “rotational symmetry” as the placement of the  $x_B$ - and  $y_B$ -axes can be switched for the principal body-fixed frame and the spacecraft is **axisymmetric** about  $z_B$ .

For this introductory look at torque-free motion, assume  $I_{xx} = I_{yy}$ , then the torque-free motion equations become

$$\sum \vec{M}_B = \frac{d}{dt} \vec{H}_G = \begin{bmatrix} I_{xx}\dot{p} + (I_{zz} - I_{xx})qr \\ I_{xx}\dot{q} + (I_{xx} - I_{zz})pr \\ I_{zz}\dot{r} \end{bmatrix} = 0 \quad (11.65)$$

Thus, the spin rate is constant, i.e.  $r = \bar{r}$  and

$$\|\vec{H}_G\|_2 = \frac{I_{zz}\bar{r}}{\cos \theta_n} \quad (11.66)$$

Next, define the *temporary* angular rate,  $\omega_*$ , as

$$\omega_* = \frac{I_{xx} - I_{zz}}{I_{xx}} \bar{r} \quad (11.67)$$

which is related to another angular rate as will be shown.

This allows one to rewrite the torque-free motion dynamics as two coupled differential equations

$$\begin{aligned} \dot{p} - \omega_* q &= 0 \\ \dot{q} + \omega_* p &= 0 \end{aligned} \quad (11.68)$$

Taking the Laplace transform and separating variables, one has the eigenvalue equation

$$\begin{bmatrix} s & -\omega_* \\ \omega_* & s \end{bmatrix} \begin{bmatrix} p(s) \\ q(s) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (11.69)$$

with left-side determinant, or characteristic polynomial, as

$$s^2 + \omega_*^2 = 0 \quad (11.70)$$

which has two purely imaginary roots or system poles.

This equation has the well-known solution

$$\begin{aligned} p &= \Omega \sin \omega_* t \\ q &= \Omega \cos \omega_* t \end{aligned} \quad (11.71)$$

which corresponds to

$$\vec{\omega} = \begin{bmatrix} \Omega \sin \omega_* t \\ \Omega \cos \omega_* t \\ \bar{r} \end{bmatrix} \quad (11.72)$$

which describes the sweeping out of the **space cone** at constant nutation angle,  $\theta_n$ , about  $\vec{H}_G$  and constant height,  $\bar{r}$ , above the  $x_B - y_B$  plane with a circular base of radius  $\Omega$  about the  $z_B$ -axis.

Next, consider the description of this motion using the 3 – 1 – 3 Euler angles. Substituting these expressions into the 3 – 1 – 3 Euler angle rate equations, i.e.

$$\begin{bmatrix} \omega_p \\ \omega_n \\ \omega_s \end{bmatrix} = \begin{bmatrix} \sin \psi_s \csc \theta_n & \cos \psi_s \csc \theta_n & 0 \\ \cos \psi_s & -\sin \psi_s & 0 \\ -\sin \psi_s \cot \theta_n & -\cos \psi_s \cot \theta_n & 1 \end{bmatrix} \begin{bmatrix} \Omega \sin \omega_* t \\ \Omega \cos \omega_* t \\ \bar{r} \end{bmatrix} \quad (11.73)$$

one has

$$\begin{bmatrix} \omega_p \\ \omega_n \\ \omega_s \end{bmatrix} = \begin{bmatrix} \Omega \sin \omega_* t \sin \psi_s \csc \theta_n + \Omega \cos \omega_* t \cos \psi_s \csc \theta_n \\ \Omega \sin \omega_* t \cos \psi_s - \Omega \cos \omega_* t \sin \psi_s \\ -\Omega \sin \omega_* t \sin \psi_s \cot \theta_n - \Omega \cos \omega_* t \cos \psi_s \cot \theta_n + \bar{r} \end{bmatrix} \quad (11.74)$$

By trigonometric identities, one has

$$\begin{bmatrix} \omega_p \\ \omega_n \\ \omega_s \end{bmatrix} = \begin{bmatrix} \Omega \csc \theta_n \cos (\omega_* t - \psi_s) \\ \Omega (\sin \omega_* t - \psi_s) \\ \bar{r} - \Omega \cot \theta_n \cos (\omega_* t - \psi_s) \end{bmatrix} \quad (11.75)$$

However, as  $\omega_n = 0$  for  $I_{xx} = I_{yy}$ , one has  $\sin \omega_* t - \psi_s = 0$  and  $\cos (\omega_* t - \psi_s) = 1$ , which results in

$$\begin{bmatrix} \omega_p \\ \omega_n \\ \omega_s \end{bmatrix} = \begin{bmatrix} \Omega \csc \theta_n \\ 0 \\ \bar{r} - \Omega \cot \theta_n \end{bmatrix} \quad (11.76)$$

Furthermore, one must have

$$\omega_s = \dot{\psi}_s = \omega_* = \frac{I_{xx} - I_{zz}}{I_{xx}} \bar{r} \quad (11.77)$$

and

$$\vec{H}_{G,B} = \begin{bmatrix} I_{xx} \Omega \sin \omega_s t \\ I_{xx} \Omega \cos \omega_s t \\ I_{zz} \bar{r} \end{bmatrix} \quad (11.78)$$

However, by back-substitution into the third component yields

$$\Omega = \frac{I_{zz}}{I_{xx}} \bar{r} \tan \theta_n \quad (11.79)$$

Substituting for  $\Omega$  into the first component of the angular velocity yields

$$\bar{r} = \frac{I_{xx}}{I_{zz}} \omega_p \cos \theta_n \quad (11.80)$$

and, substituting this into the angular momentum magnitude yields

$$\|\vec{H}_G\|_2 = I_{xx} \omega_p \quad (11.81)$$

Lastly, substituting for  $\bar{r}$  in the equation for  $\omega_s$ , one has

$$\omega_s = \frac{I_{xx} - I_{zz}}{I_{zz}} \cos \theta_n \omega_p \quad (11.82)$$

Thus, for a **oblate body**, i.e.  $I_{xx} < I_{zz}$ , then  $\omega_p$  has the opposite sign as  $\omega_s$  and one has a **retrograde precession**. For a **prolate body**, i.e.  $I_{xx} > I_{zz}$ , then  $\omega_p$  has the same sign as  $\omega_s$  and one has a **prograde procession**.  $\gamma$  is the **wobble angle** defined as the angle between  $\vec{\omega}_{B/I,B}$  and the  $z_B$ -axis, i.e.

$$\cos \gamma = \frac{r}{\|\vec{\omega}_{B/I,B}\|} = \frac{\bar{r}}{\sqrt{\Omega^2 \sin^2(\omega_* t) + \Omega^2 \cos^2(\omega_* t) + \bar{r}^2}} = \frac{\bar{r}}{\sqrt{\Omega^2 + \bar{r}^2}} \quad (11.83)$$

Substituting for  $\Omega$ , one has

$$\cos \gamma = \frac{\omega_0}{\sqrt{\left(\frac{I_{zz}}{I_{xx}} \bar{r} \tan \theta_n\right)^2}} = \frac{I_{xx}}{\sqrt{I_{xx}^2 + I_{zz}^2 \tan^2 \theta_n}} \quad (11.84)$$

and using trigonometric identities, one has

$$\cos \gamma = \frac{\cos \theta_n}{\sqrt{\frac{I_{zz}^2}{I_{xx}^2} + \left(1 - \frac{I_{zz}^2}{I_{xx}^2}\right) \cos^2 \theta_n}} \quad (11.85)$$

which shows that  $\gamma > \theta$  for prograde precession and  $\gamma < \theta$  for retrograde precession.

### Stability of Rotating Axisymmetric Spacecraft

Consider the equilibrium condition where the angular velocity is about only the spin axis, i.e., without loss of generality, the  $z_B$ -axis. Thus,  $r(t) = \bar{r}$  for all  $t > 0$ ,  $\bar{p} = 0^\circ$ , and  $\bar{q} = 0^\circ$ , which trivially solves the torque-free equations of motion. Then, using the perturbation notation, the rotating axisymmetric spacecraft dynamics in torque-free motion are given by

$$\begin{bmatrix} I_{xx} \Delta \dot{p} + (I_{zz} - I_{yy}) (\bar{r} + \Delta r) \Delta q \\ I_{yy} \Delta \dot{q} + (I_{xx} - I_{zz}) (\bar{r} + \Delta r) \Delta p \\ I_{zz} \Delta \dot{r} + (I_{yy} - I_{xx}) \Delta p \Delta q \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (11.86)$$

Then, keeping only the first-order terms, one has

$$\begin{bmatrix} I_{xx} \Delta \dot{p} + (I_{zz} - I_{yy}) \bar{r} \Delta q \\ I_{yy} \Delta \dot{q} + (I_{xx} - I_{zz}) \bar{r} \Delta p \\ I_{zz} \Delta \dot{r} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (11.87)$$

which implies  $\Delta \dot{r} = \Delta \dot{r}_0$  is decoupled from the other perturbations.

Thus, one has

$$\begin{bmatrix} \Delta \dot{p} + \frac{(I_{zz} - I_{yy})}{I_{xx}} \bar{r} \Delta q \\ \Delta \dot{q} + \frac{(I_{xx} - I_{zz})}{I_{yy}} \bar{r} \Delta p \\ \Delta \dot{r} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (11.88)$$

Taking the Laplace transform and separating variables, one has the eigenvalue equation

$$\begin{bmatrix} s & \frac{(I_{zz} - I_{yy})}{I_{xx}} \bar{r} \\ \frac{(I_{xx} - I_{zz})}{I_{yy}} \bar{r} & s \end{bmatrix} \begin{bmatrix} \Delta p(s) \\ \Delta q(s) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (11.89)$$

with determinant, or characteristic equation, as

$$s^2 - \frac{(I_{xx} - I_{zz})(I_{zz} - I_{yy})}{I_{xx}I_{yy}}\bar{r}^2 = 0 \quad (11.90)$$

Thus, by this analysis, if the spin axis is either the major axis, i.e.  $I_{zz} > I_{xx}$  and  $I_{zz} > I_{yy}$ , or if the spin axis is the minor axis, i.e.  $I_{zz} < I_{xx}$  and  $I_{zz} < I_{yy}$ , then the system poles are purely imaginary and the motion is sinusoidal and marginally stable as there is no damping in the torque-free motion model. However, if the spin axis is the intermediate axis, i.e.  $I_{xx} > I_{zz} > I_{yy}$  or  $I_{yy} > I_{zz} > I_{xx}$ , then the motion is unstable due to a RHP system pole.

However, in reality, all spacecrafts experience some degree of flexibility which leads to additional stability analysis of the spin axis as the major or minor axis due to the energy dissipation into the structural vibrations. Consider a spacecraft with  $I_{xx} = I_{yy}$  and rotational kinetic energy

$$T_R = \frac{1}{2}I_{xx}p^2 + \frac{1}{2}I_{xx}q^2 + \frac{1}{2}I_{zz}r^2 \quad (11.91)$$

Defining  $\omega_{\perp}^2 = p^2 + q^2$  as the perpendicular angular momentum to the spin axis, one has

$$T_R = \frac{1}{2}I_{xx}\omega_{\perp}^2 + \frac{1}{2}I_{zz}r^2 \quad (11.92)$$

and differentiating with respect to time, one has

$$\dot{T}_R = \frac{1}{2}I_{xx}\frac{d\omega_{\perp}^2}{dt} + I_{zz}r\dot{r} \quad (11.93)$$

where  $\dot{T}_R < 0$  for all spacecrafts due to the energy dissipation.

Then, recall the angular momentum of a spacecraft with  $I_{xx} = I_{yy}$  is

$$\vec{H}_{G,B} = \begin{bmatrix} I_{xx}p \\ I_{xx}q \\ I_{zz}r \end{bmatrix} \quad (11.94)$$

with magnitude

$$\|\vec{H}_G\|_2^2 = I_{xx}^2(p^2 + q^2) + I_{zz}^2r^2 = I_{xx}^2\omega_{\perp}^2 + I_{zz}^2r^2 \quad (11.95)$$

Differentiating the angular momentum magnitude with respect to time yields

$$\frac{d\|\vec{H}_G\|_2^2}{dt} = I_{xx}^2\frac{d\omega_{\perp}^2}{dt} + 2I_{zz}^2r\dot{r} = 0 \quad (11.96)$$

which holds for torque-free motion. Thus, one has

$$\dot{r} = -\frac{I_{xx}^2}{2I_{zz}^2r}\frac{d\omega_{\perp}^2}{dt} \quad (11.97)$$

Finally, substituting this expression into the energy dissipation equation, one obtains

$$\dot{T}_R = \frac{1}{2}I_{xx}\frac{d\omega_{\perp}^2}{dt} + I_{zz}r\left(-\frac{I_{xx}^2}{2I_{zz}^2r}\frac{d\omega_{\perp}^2}{dt}\right) \quad (11.98)$$

$$\dot{T}_R = \left( \frac{I_{xx}I_{zz}}{2I_{zz}} - \frac{I_{xx}^2}{2I_{zz}} \right) \frac{d\omega_{\perp}^2}{dt} \quad (11.99)$$

and rearranging, one has

$$\frac{d\omega_{\perp}^2}{dt} = \frac{2I_{zz}}{I_{xx}(I_{zz} - I_{xx})} \dot{T}_R \quad (11.100)$$

This equation shows that if one has an oblate spacecraft, i.e.  $I_{zz} > I_{xx}$ , then  $\frac{d\omega_{\perp}^2}{dt} < 0$  and any perturbation in  $p$  or  $q$  will asymptotically decay to 0. Thus, the spin is asymptotically stable. However, if one has a prolate spacecraft, i.e.  $I_{zz} < I_{xx}$ , then  $\frac{d\omega_{\perp}^2}{dt} > 0$  and the spin is asymptotically unstable. Thus, if a spacecraft is to be **single-spin stabilized**, it must be an oblate spinner. Notably, a **nutation damper** can be used to passively or actively increase this rate of energy dissipation for increased stability. These are discussed in a later section of this textbook.

### Linearized Attitude Dynamics about Circular Orbit

Consider the satellite attitude equations of motion as

$$\vec{M}_B = \begin{bmatrix} I_{xx}\dot{p} + (I_{zz} - I_{yy})qr \\ I_{yy}\dot{q} + (I_{xx} - I_{zz})pr \\ I_{zz}\dot{r} + (I_{yy} - I_{xx})pq \end{bmatrix} \quad (11.101)$$

where, one can define the following for the inertial angular velocity

$$\vec{\omega}_{B/P,B} = \begin{bmatrix} p \\ q \\ r \end{bmatrix} = \vec{\omega}_{B/L,B} + \vec{\omega}_{L/P,B} \quad (11.102)$$

where one the orbital angular velocity  $\omega_O = -\frac{\mu}{a^3}$  for a circular orbit explicitly enters the dynamics equation via

$$\vec{\omega}_{L/P,L} = \begin{bmatrix} 0 \\ -\omega_O \\ 0 \end{bmatrix} \quad (11.103)$$

Thus, one has

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} = \vec{\omega}_{B/L,B} + C_{B \leftarrow L} \begin{bmatrix} 0 \\ -\omega_O \\ 0 \end{bmatrix} \quad (11.104)$$

which can be written in terms of the 3 – 2 – 1 Euler angles of the body-fixed frame relative to the LVLH frame and the corresponding Euler angle rates as

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\sin\phi \\ 0 & \cos\phi & \sin\phi\cos\theta \\ 0 & -\sin\phi & \cos\phi\cos\theta \end{bmatrix} \begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} + \begin{bmatrix} \cos\theta\cos\psi & \cos\theta\sin\psi & -\sin\theta \\ \sin\phi\sin\theta\cos\psi - \cos\phi\sin\psi & \sin\phi\sin\theta\sin\psi + \cos\phi\cos\psi & \sin\phi\cos\theta \\ \cos\phi\sin\theta\cos\psi + \sin\phi\sin\psi & \cos\phi\sin\theta\sin\psi - \sin\phi\cos\psi & \cos\phi\cos\theta \end{bmatrix} \begin{bmatrix} 0 \\ -\omega_O \\ 0 \end{bmatrix} \quad (11.105)$$

This angular velocity can be linearized for small Euler angles *and* small Euler angle rates using the small angle approximation which results in

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} \approx \begin{bmatrix} \dot{\phi} - \phi\dot{\psi} \\ \dot{\theta} + \sin\phi\dot{\psi} \\ \dot{\psi} - \phi\dot{\theta} \end{bmatrix} + \begin{bmatrix} 1 & \psi & -\theta \\ -\psi & 1 & \phi \\ \theta & -\phi & 1 \end{bmatrix} \begin{bmatrix} 0 \\ -\omega_O \\ 0 \end{bmatrix} \quad (11.106)$$

and discarding higher-order terms, one has

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} \approx \begin{bmatrix} \dot{\phi} - \psi\omega_O \\ \dot{\theta} - \omega_O \\ \dot{\psi} + \phi\omega_O \end{bmatrix} \quad (11.107)$$

and, by differentiation,

$$\begin{bmatrix} \dot{p} \\ \dot{q} \\ \dot{r} \end{bmatrix} \approx \begin{bmatrix} \ddot{\phi} - \omega_O\dot{\psi} \\ \ddot{\theta} \\ \ddot{\psi} + \omega_O\dot{\phi} \end{bmatrix} \quad (11.108)$$

Finally, by substitution, one has

$$\vec{M}_B = \begin{bmatrix} I_{xx}\ddot{\phi} - I_{xx}\omega_O\dot{\psi} + (I_{zz} - I_{yy})(\dot{\theta} - \omega_O)(\dot{\psi} + \phi\omega_O) \\ I_{yy}\ddot{\theta} + (I_{xx} - I_{zz})(\dot{\phi} - \psi\omega_O)(\dot{\psi} + \phi\omega_O) \\ I_{zz}\ddot{\psi} - I_{zz}\omega_O\dot{\phi} + (I_{yy} - I_{xx})(\dot{\phi} - \psi\omega_O)(\dot{\theta} - \omega_O) \end{bmatrix} \quad (11.109)$$

and discarding higher-order terms, one has

$$\vec{M}_B = \begin{bmatrix} I_{xx}\ddot{\phi} - \omega_O(I_{xx} + I_{zz} - I_{yy})\dot{\psi} + \omega_O^2(I_{zz} - I_{yy})\phi \\ I_{yy}\ddot{\theta} \\ I_{zz}\ddot{\psi} + \omega_O(I_{xx} + I_{zz} - I_{yy})\dot{\phi} + \omega_O^2(I_{yy} - I_{zz})\psi \end{bmatrix} \quad (11.110)$$

Note that the linearized pitch dynamics are decoupled from the linearized roll and yaw dynamics.

For non-spinning satellites, one typically considers the effects of the gravity-gradient moment,  $\vec{M}_g$ , separate from other moment disturbances,  $\vec{M}_d$ , and control inputs,  $\vec{M}_c$ , which can be defined for a rigid-body as

$$\vec{M}_{g,B} = \int [\vec{x}_B] \times d\vec{F}_G = \int_V [\vec{x}_B] \times \left( -\frac{\mu\rho dV}{\|\vec{r}_m\|_2^3} \vec{r}_m \right) \quad (11.111)$$

where  $dm = \rho dV$  is an infinitesimal mass element of the body with density  $\rho$ ,  $\vec{r}_m = \vec{r}_{P,L} + \vec{x}_B$  is the position vector of the mass element with respect to the origin of the perifocal frame, and  $\vec{r}_{P,B}$  is the position of the center of mass of the rigid-body with respect to the origin of the perifocal frame expressed in body-fixed frame coordinates, i.e., for a circular orbit

$$\vec{r}_{P,B} = C_{B \leftarrow L} \begin{bmatrix} 0 \\ 0 \\ -a \end{bmatrix} = \begin{bmatrix} a \sin \theta \\ -a \sin \phi \cos \theta \\ -a \cos \phi \cos \theta \end{bmatrix} \quad (11.112)$$

With the assumption  $\vec{x}_B \ll \vec{r}_P$ , one can approximate the radial distance cubed by the truncated Taylor series

$$\frac{1}{\|\vec{r}_m\|_2^3} \approx \frac{1}{a^3} \left( 1 - \frac{3\vec{r}_P \cdot \vec{x}_B}{a^2} \right) \quad (11.113)$$

which provides

$$\vec{M}_{g,B} = \frac{3\mu}{a^5} \int_V (\vec{r}_P \cdot \vec{x}_B) ([\vec{x}_B] \times \vec{r}_P) \rho dV \quad (11.114)$$

and by the use of the principal axes, all cross-products of inertia's are zero, which results in

$$\vec{M}_{g,B} \approx \frac{3\mu}{a^5} \begin{bmatrix} r_{P,y} r_{P,z} \left( \int y_B^2 \rho dV - \int z_B^2 \rho dV \right) \\ r_{P,x} r_{P,z} \left( \int z_B^2 \rho dV - \int x_B^2 \rho dV \right) \\ r_{P,x} r_{P,y} \left( \int x_B^2 \rho dV - \int y_B^2 \rho dV \right) \end{bmatrix} \quad (11.115)$$

Substituting for the moment of inertia integrals and the components of  $\vec{r}_P$ , one has

$$\vec{M}_{g,B} \approx \frac{3\mu}{a^5} \begin{bmatrix} (-a \sin \phi \cos \theta)(-a \cos \phi \cos \theta)(I_{zz} - I_{yy}) \\ (a \sin \theta)(-a \cos \phi \cos \theta)(I_{xx} - I_{zz}) \\ (a \sin \theta)(-a \sin \phi \cos \theta)(I_{yy} - I_{xx}) \end{bmatrix} \quad (11.116)$$

which simplifies to

$$\vec{M}_{g,B} \approx \frac{3\mu}{2a^3} \begin{bmatrix} (I_{zz} - I_{yy}) \sin(2\phi) \cos^2(\theta) \\ (I_{zz} - I_{xx}) \sin(2\theta) \cos \phi \\ (I_{xx} - I_{yy}) \sin(2\theta) \sin \phi \end{bmatrix} \quad (11.117)$$

Using the small-angle approximation, one has

$$\vec{M}_{g,B} = \frac{3\omega_O^2}{2} \begin{bmatrix} 2\phi(I_{zz} - I_{yy}) \\ 2\theta(I_{zz} - I_{xx}) \\ 2\theta\phi(I_{xx} - I_{yy}) \end{bmatrix} \quad (11.118)$$

and discarding higher-order terms, one has the **linearized gravity-gradient moment** vector

$$\vec{M}_{g,B} = \begin{bmatrix} 3\omega_O^2(I_{zz} - I_{yy})\phi \\ 3\omega_O^2(I_{zz} - I_{xx})\theta \\ 0 \end{bmatrix} \quad (11.119)$$

Thus, by substitution, for the **linearized gravity-gradient attitude dynamics**, one has

$$\begin{bmatrix} M_{d,x} + M_{c,x} \\ M_{d,y} + M_{c,y} \\ M_{d,z} + M_{c,z} \end{bmatrix} + \begin{bmatrix} 3\omega_O^2(I_{zz} - I_{yy})\phi \\ \omega_O^2(I_{zz} - I_{xx})\theta \\ 0 \end{bmatrix} = \begin{bmatrix} I_{xx}\ddot{\phi} - \omega_O(I_{xx} + I_{zz} - I_{yy})\dot{\psi} + \omega_O^2(I_{zz} - I_{yy})\phi \\ I_{yy}\ddot{\theta} + 3\omega_O^2(I_{xx} - I_{zz})\theta \\ I_{zz}\ddot{\psi} + \omega_O(I_{xx} + I_{zz} - I_{yy})\dot{\phi} + \omega_O^2(I_{yy} - I_{zz})\psi \end{bmatrix} \quad (11.120)$$

or

$$\begin{bmatrix} M_{d,x} + M_{c,x} \\ M_{d,y} + M_{c,y} \\ M_{d,z} + M_{c,z} \end{bmatrix} = \begin{bmatrix} I_{xx}\ddot{\phi} + 4\omega_O^2(I_{yy} - I_{zz})\phi - \omega_O(I_{xx} + I_{zz} - I_{yy})\dot{\psi} \\ I_{yy}\ddot{\theta} + 3\omega_O^2(I_{xx} - I_{zz})\theta \\ I_{zz}\ddot{\psi} + \omega_O^2(I_{yy} - I_{xx})\psi + \omega_O(I_{xx} + I_{zz} - I_{yy})\dot{\phi} \end{bmatrix} \quad (11.121)$$

## Stability of Gravity-Gradient Satellite Dynamics

To analyze the stability of the linearized gravity-gradient attitude dynamics, first, define the moments of inertia ratios as

$$\sigma_{xx} = \frac{I_{yy} - I_{zz}}{I_{xx}} = \frac{\int(x^2 + z^2)dm - \int(x^2 + y^2)dm}{\int(y^2 + z^2)dm} = \frac{\int(z^2 - y^2)dm}{\int(z^2 + y^2)dm} \quad (11.122)$$

$$\sigma_{yy} = \frac{I_{xx} - I_{zz}}{I_{yy}} = \frac{\int(y^2 + z^2)dm - \int(x^2 + y^2)dm}{\int(x^2 + z^2)dm} = \frac{\int(z^2 - x^2)dm}{\int(z^2 + x^2)dm} \quad (11.123)$$

$$\sigma_{zz} = \frac{I_{yy} - I_{xx}}{I_{zz}} = \frac{\int(x^2 + z^2)dm - \int(y^2 + z^2)dm}{\int(x^2 + y^2)dm} = \frac{\int(x^2 - y^2)dm}{\int(x^2 + y^2)dm} \quad (11.124)$$

Thus,

$$|\sigma_{xx}| < 1 \quad \& \quad |\sigma_{yy}| < 1 \quad \& \quad |\sigma_{zz}| < 1 \quad (11.125)$$

Ignoring the disturbances, the linearized pitch state-space dynamics can be rearranged as

$$\begin{bmatrix} \dot{\theta} \\ \ddot{\theta} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -3\omega_O^2 \sigma_{yy} & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} 0 \\ I_{yy}^{-1} \end{bmatrix} M_{c,y} \quad (11.126)$$

with characteristic equation

$$\lambda^2 + 3\omega_O^2 \sigma_{yy} = 0 \quad (11.127)$$

and two roots

$$\lambda_{1,2} = \pm \omega_O \sqrt{-3\sigma_{yy}} \quad (11.128)$$

which are either purely imaginary or one positive real and one negative real. Thus, the gravity-gradient pitch dynamics are marginally stable if  $\sigma_{yy} > 0$  or  $I_{xx} > I_{zz}$ , otherwise it is unstable.

Furthermore, recalling that  $I_{yy} < I_{xx} + I_{zz}$  must also hold and multiplying by  $I_{xx} - I_{zz} > 0$  for pitch dynamics stability, one has

$$I_{yy}I_{xx} - I_{yy}I_{zz} < I_{xx}^2 - I_{zz}^2 \quad (11.129)$$

$$I_{xx}I_{yy} - I_{xx}^2 < I_{yy}I_{zz} - I_{zz}^2 \quad (11.130)$$

$$I_{xy}(I_{yy} - I_{xx}) < I_{zz}(I_{yy} - I_{zz}) \quad (11.131)$$

$$\frac{I_{yy} - I_{zz}}{I_{xx}} < \frac{I_{yy} - I_{zz}}{I_{xx}} \quad (11.132)$$

or, by definition, one has the alternative criterion for pitch dynamics stability

$$\sigma_{zz} < \sigma_{xx} \quad (11.133)$$

Ignoring the disturbances, the linearized roll-yaw state-space dynamics can be rearranged as

$$\begin{bmatrix} \dot{\phi} \\ \dot{\psi} \\ \ddot{\phi} \\ \ddot{\psi} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -4\omega_O^2 \sigma_{xx} & 0 & 0 & \omega_O(1 - \sigma_{xx}) \\ 0 & -\omega_O^2 \sigma_{zz} & -\omega_O(1 - \sigma_{zz}) & 0 \end{bmatrix} \begin{bmatrix} \phi \\ \psi \\ \dot{\phi} \\ \dot{\psi} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ I_{xx}^{-1} & 0 \\ 0 & I_{zz}^{-1} \end{bmatrix} \begin{bmatrix} M_{c,x} \\ M_{c,z} \end{bmatrix} \quad (11.134)$$

with characteristic equation

$$\lambda^4 + (3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1)\omega_O^2\lambda^2 + 4\sigma_{xx}\sigma_{zz}\omega_O^4 = 0 \quad (11.135)$$

and four roots

$$\lambda_{1,2,3,4} = \pm\omega_O\sqrt{\frac{1}{2}\left(-(3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1) \pm \sqrt{(3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1)^2 - 16\sigma_{xx}\sigma_{zz}}\right)} \quad (11.136)$$

with  $\lambda_1 = -\lambda_2$  and  $\lambda_3 = -\lambda_4$ .

Thus, for all roots to not have any positive real parts, then all  $\lambda$ 's must be purely imaginary. This requires

$$-(3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1) \pm \sqrt{(3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1)^2 - 16\sigma_{xx}\sigma_{zz}} \quad (11.137)$$

be a purely real negative number for both terms. This requires the following three criteria

$$\begin{aligned} -(3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1) &< 0 \\ (3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1)^2 - 16\sigma_{xx}\sigma_{zz} &> 0 \\ -(3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1) &< \sqrt{(3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1)^2 - 16\sigma_{xx}\sigma_{zz}} \end{aligned} \quad (11.138)$$

$$\begin{aligned} 3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1 &> 0 \\ (3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1)^2 &> 16\sigma_{xx}\sigma_{zz} \\ (3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1)^2 &> (3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1)^2 - 16\sigma_{xx}\sigma_{zz} \end{aligned} \quad (11.139)$$

$$\begin{aligned} 3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1 &> 0 \\ 3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1 &> 4\sqrt{\sigma_{xx}\sigma_{zz}} \\ (3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1)^2 &> (3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1)^2 - 16\sigma_{xx}\sigma_{zz} \end{aligned} \quad (11.140)$$

$$\begin{aligned} 3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1 &> 0 \\ 3\sigma_{xx} + \sigma_{xx}\sigma_{zz} + 1 &> 4\sqrt{\sigma_{xx}\sigma_{zz}} \\ \sigma_{xx}\sigma_{zz} &> 0 \end{aligned} \quad (11.141)$$

By inspection, the third criterion demonstrates that satisfying the second criterion naturally satisfies the first criterion. Furthermore, squaring the second criterion on both sides, one has

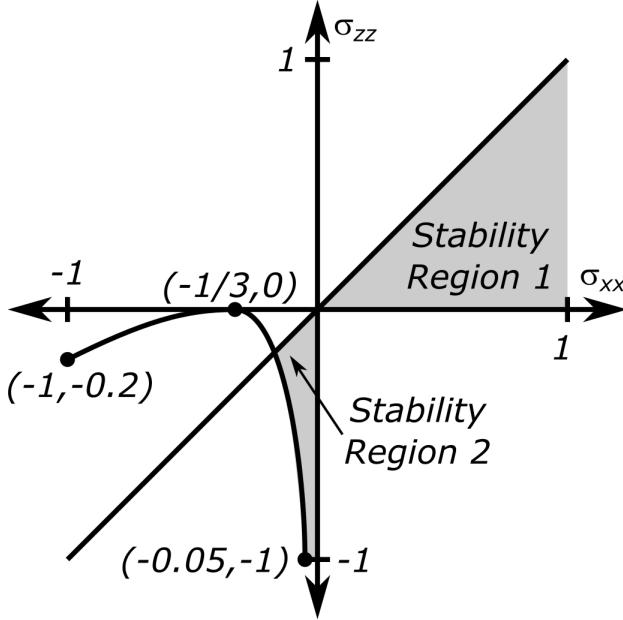
$$9\sigma_{xx}^2 + \sigma_{xx}^2\sigma_{zz}^2 + 1 + 6\sigma_{xx}^2\sigma_{zz} + 2\sigma_{xx}\sigma_{zz} + 6\sigma_{xx} > 16\sigma_{xx}\sigma_{zz} \quad (11.142)$$

Thus, the gravity-gradient roll-yaw dynamics are marginally stable if and only if

$$\begin{aligned} (\sigma_{zz}^2 + 6\sigma_{zz} + 9)\sigma_{xx}^2 + (-14\sigma_{zz} + 6)\sigma_{xx} + 1 &> 0 \\ \sigma_{xx}\sigma_{zz} &> 0 \end{aligned} \quad (11.143)$$

Finally, recalling the pitch dynamics stability criterion,  $\sigma_{zz} < \sigma_{xx}$ , one can define two **gravity-gradient stability regions** for marginal stability of the gravity-gradient satellite dynamics. The first is

$$\sigma_{xx} > 0, \quad \sigma_{zz} > 0, \quad \sigma_{zz} < \sigma_{xx} \quad (11.144)$$



and the second can be stated by

$$\sigma_{xx} < 0, \quad \sigma_{zz} < 0, \quad \sigma_{zz} < \sigma_{xx}, \quad (\sigma_{zz}^2 + 6\sigma_{zz} + 9)\sigma_{xx}^2 + (-14\sigma_{zz} + 6)\sigma_{xx} + 1 > 0 \quad (11.145)$$

These two stability regions are plotted in the following figure. Stability region 2 is notably a very small region and is seldom used owing to practical structural difficulties. It is also important to note that as the linearized dynamics are only marginally stable, passive and/or active damping is typically necessary for gravity-gradient-stabilized satellites.

By definition of the moment of inertia ratios, stability region 1 equivalently requires

$$I_{yy} - I_{zz} > 0, \quad I_{yy} - I_{xx} > 0, \quad I_{xx} > I_{zz} \quad (11.146)$$

and incorporating the requirements due to the definition of the moments of inertia, one has

$$I_{xx} + I_{zz} > I_{yy} > I_{xx} > I_{zz} \quad (11.147)$$

which imposes strict structural designs on a gravity-gradient-stabilized satellite in region 1. By definition of the moment of inertia ratios, stability region 2 equivalently requires

$$I_{yy} - I_{zz} < 0, \quad I_{yy} - I_{xx} < 0, \quad I_{xx} > I_{zz} \quad (11.148)$$

and incorporating the requirements due to the definition of the moments of inertia, one has

$$I_{yy} + I_{zz} > I_{xx} > I_{zz} > I_{yy}, \quad I_{yy} - I_{zz} \ll I_{xx} \quad (11.149)$$

which imposes even stricter criteria on a gravity-gradient-stabilized satellite in region 2 than region 1 which is why region 1 is preferred.

## References

For more information, please refer to the following

- Curtis, H. D., “9.4 Equations of Rotational Motion,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 420-414
- Curtis, H. D., “9.5 Moments of Inertia,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 414-435
- Curtis, H. D., “9.6 Euler’s Equations,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 435-440
- Curtis, H. D., “10.2 Torque-Free Motion,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 476-486
- Curtis, H. D., “10.3 Stability of Torque-Free Motion,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 486-491
- Curtis, H. D., “10.10 Gravity-Gradient Stabilization,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 530-540
- Hibbeler, R. C., “21.6 Torque-Free Motion,” *Dynamics*, 15th ed., Vol. 1, Pearson Inc., 2022, pp. 606-609
- Sidi, M. J., “4.5 Euler’s Moment Equations,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 95-98
- Sidi, M. J., “4.6 Characteristics of Rotational Motion of a Spinning Body,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 98-100
- Sidi, M. J., “4.8 Attitude Dynamics Equations of Motion for a Nonspinning Spacecraft,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 107-110
- Sidi, M. J., “5.3 Gravity Gradient Attitude Control,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 114-129
- Sidi, M. J., “6.2 Attitude Spin Stabilization during the ΔV Stage,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 132-135

## 11.5 Orbital Vehicle Attitude Control Systems

In order to control the attitude of a satellite, one must change the dynamics of the angular velocity in some way. As the angular velocity is directly related to the angular momentum, one requires control devices to change the angular momentum without a change in mass to affect the attitude dynamics. These stabilization approaches use either thruster pairs, passive reaction wheels, or active reaction and momentum wheels to

stabilize a axisymmetric, spinning spacecraft which is known as **spin stabilization**. This section will discuss the use of a single device for nutation control of the spinning spacecraft or a dual-spin configuration known as the gyrostat. The subsequent chapter will discuss stabilization and control of satellites using more advanced momentum-biased systems with momentum wheels.

### Thruster-Based Nutation Control

Recall that for spin-stabilized spacecraft rotating about their major axis of inertia are marginally stable while those rotating about their minor axis are unstable. However, stability can be achieved through the use of **thruster-based active nutation control (TB-ANC)** which implement a band-pass filter centered on the nutation frequency,  $\lambda_n$ , to extract that component from the measured angular velocity components for feedback with one or both of the transverse axes, e.g.,  $x_B$  and/or  $y_B$  with  $z_B$  as the spin axis. Here the control law would be expressed as

$$\begin{aligned} M_{x,c} &= -T_c \Delta \text{sign}(\lambda_{n,x}) \\ M_{y,c} &= -T_c \Delta \text{sign}(\lambda_{n,y}) \end{aligned} \quad (11.150)$$

Importantly, time delays and dead zones in the system hardware are typically vital to the final design of these TB-ANC systems.

To determine the nutation frequency for a particular disturbing moment, one can assume an axisymmetric spacecraft rotating about the  $z_B$ -axis with some disturbing moment,  $M_d$ , acting only about the  $x_B$ -axis, then, the linearized Euler equations for a spinning satellite undergoing nutation can be written as

$$\begin{aligned} H_x(t) &= I_{xx} p(t) = \frac{M_d}{\bar{r} \left( \frac{I_{zz}}{I_{xx}} - 1 \right)} \sin(\lambda_n t) \\ H_y(t) &= I_{yy} q(t) = I_{xx} \omega_x(t) = \frac{M_d}{\bar{r} \left( \frac{I_{zz}}{I_{xx}} - 1 \right)} (1 - \cos(\lambda_n t)) \\ \bar{r} &= \text{constant} \end{aligned} \quad (11.151)$$

where the **nutation frequency** is given by

$$\lambda_n = \bar{r} \frac{(I_{zz} - I_{xx})}{I_{xx}} = \bar{r} \left( \frac{I_{zz}}{I_{xx}} - 1 \right) \quad (11.152)$$

Next, the total momentum in the  $x_B$ - $y_B$  plane perpendicular to  $z_B$  is

$$H_\perp(t) = \sqrt{H_x^2 + H_y^2} = \frac{\sqrt{2} M_d}{\bar{r} \left( \frac{I_{zz}}{I_{xx}} - 1 \right)} \sqrt{1 - \cos(\lambda_n t)} = \frac{2 M_d}{\bar{r} \left( \frac{I_{zz}}{I_{xx}} - 1 \right)} \left| \sin \left( \frac{\lambda_n t}{2} \right) \right| \quad (11.153)$$

Then, recalling the relationship between the nutation angle and the perpendicular angular momentum, one has

$$\tan \theta_n = \frac{H_\perp}{H_z} = \frac{2 M_d}{I_{zz} \bar{r}^2 \left( \frac{I_{zz}}{I_{xx}} - 1 \right)} \left| \sin \left( \frac{\lambda_n t}{2} \right) \right| \quad (11.154)$$

For small nutation angles, one has

$$\theta_n \approx \frac{\sqrt{2}M_d}{I_{zz}\bar{r}^2 \left( \frac{I_{zz}}{I_{xx}} - 1 \right)} \left| \sin \left( \frac{\lambda_n t}{2} \right) \right| \quad (11.155)$$

Thus, the magnitude of the nutation is inversely proportional to  $\bar{r}^2$  which costs fuel to increase for TB-ANC systems.

### Reaction Wheel-Based Nutation Control

Reaction **wheel-based nutation control** systems utilize a single reaction wheel with some moment inertia,  $I_w$ , and angular velocity,  $\omega_w$ , to add to the total angular momentum of the spacecraft and wheel system as

$$\vec{H}_G = I_G \vec{\omega}_{B/I} = \begin{bmatrix} I_{xx}p \\ (I_{yy} + I_w)q + I_w\omega_w \\ I_{zz}r \end{bmatrix} \quad (11.156)$$

Assuming  $I_{xx} \gg I_w$ ,  $I_{xx} = I_{yy}$ , then one has

$$\vec{H}_G \approx \begin{bmatrix} I_{xx}p \\ I_{xx}q + I_w\omega_w \\ I_{zz}r \end{bmatrix} \quad (11.157)$$

Using Euler's equation, one has

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \frac{d}{dt} \vec{H}_G = \begin{bmatrix} I_{xx}\dot{p} + (I_{zz} - I_{xx})qr - I_w r \omega_w \\ I_{xx}\dot{q} + I_w \dot{\omega}_w + (I_{xx} - I_{zz})pr \\ I_{zz}\dot{r} + I_w p \omega_w \end{bmatrix} \quad (11.158)$$

or assuming  $I_{zz} \gg I_w$  and using the nutation frequency,  $\lambda_n$ , and inertia ratio,  $\epsilon_w = I_w/I_{xx}$ , one has

$$\begin{bmatrix} \dot{p} + \lambda_n q - \epsilon_w \bar{r} \omega_w \\ \dot{q} + \epsilon_w \dot{\omega}_w - \lambda_n p \\ \dot{r} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (11.159)$$

where  $\dot{r} = 0$  infers the constant  $r = \bar{r}$ .

**Wheel-based passive nutation control (WB-PNC)** systems use a damper wheel immersed in viscous fluid. In this case, there is no external moment on the reaction wheel and its momentum can be modeled as

$$M_w = I_w(\dot{q} + \dot{\omega}_w) + C_d \omega_w = 0 \quad (11.160)$$

where  $C_d < 0$  is the damping coefficient of the liquid. Then, one obtains the  $p$  and  $q$  dynamics as

$$\begin{bmatrix} \dot{p} + \lambda_n q - \epsilon_w \bar{r} \omega_w \\ \dot{q} + \epsilon_w \dot{\omega}_w - \lambda_n p \\ \dot{q} + \dot{\omega}_w + \frac{C_d}{I_w} \omega_w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (11.161)$$

or, in the Laplace domain, one has

$$\begin{bmatrix} s & \lambda_n & -\epsilon_w \bar{r} \\ s & -\lambda_n & \epsilon_w s \\ 0 & s & s + \frac{C_d}{I_w} \end{bmatrix} \begin{bmatrix} p \\ q \\ \omega_w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (11.162)$$

which corresponds to a characteristic equation given by the determinant of the right matrix as

$$(1 - \epsilon_w)s^3 + \frac{C_d}{I_w}s^2 + (\lambda_n^2 - \lambda_n \epsilon_w \bar{r})s + \frac{C_d}{I_w}\lambda_n^2 = 0 \quad (11.163)$$

This can be rearranged as

$$1 + \frac{C_d}{I_w(1 - \epsilon_w)} \frac{s^2 + \lambda_n^2}{s(s^2 + \lambda_1^2)} \quad (11.164)$$

where

$$\lambda_1 = \lambda_n \sqrt{1 + \frac{I_{zz}I_w}{(I_{zz} - I_{xx})(I_{xx} - I_w)}} \quad (11.165)$$

Thus, if

$$\begin{aligned} I_{xx} > I_{zz} &\rightarrow \lambda_n > \lambda_1 \\ I_{xx} < I_{zz} &\rightarrow \lambda_n < \lambda_1 \end{aligned} \quad (11.166)$$

From an analysis of the roots of this characteristic equation, a spinning satellite with a WB-PNC system is nutationally stable if and only if  $I_{xx} < I_{zz}$ , i.e., the spin axis must be the major axis.

**Wheel-based active nutation control (WB-ANC)** systems use a damper wheel that is driven by a low inductance BLDC motor with constant current. In this case, there is no external moment on the reaction wheel and its equation for the *total* angular velocity of the reaction wheel can be modeled as

$$\dot{\omega}_w + \dot{q} = \frac{k_m}{R_m I_w} v_m - \frac{k_m k_{emf}}{R_m I_w} \omega_w \quad (11.167)$$

where  $v_w$  is the input voltage to the reaction wheel,  $R_m$  is the **armature resistance**,  $I_w$  is the reaction wheel's moment of inertia,  $k_{emf}$  is the **electromotive force (EMF) constant**, and  $k_m$  is the **armature constant**. Then, by substitution for  $\dot{\omega}_w$ , one obtains the  $p$  and  $q$  dynamics as

$$\begin{bmatrix} \dot{p} + \lambda_n q - \epsilon_w \bar{r} \omega_w \\ \dot{q} + \epsilon_w \dot{\omega}_w - \lambda_n p \\ \dot{\omega}_m + \dot{q} + \frac{k_m k_{emf}}{R_m I_w} \omega_w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \frac{k_m}{R_m I_w} \end{bmatrix} v_m \quad (11.168)$$

or, in the Laplace domain, one has

$$\begin{bmatrix} s & \lambda_n & -\epsilon_w \bar{r} \\ s & -\lambda_n & \epsilon_w s \\ 0 & s & s + \frac{k_m k_{emf}}{R_m I_w} \end{bmatrix} \begin{bmatrix} p \\ q \\ \omega_w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \frac{k_m}{R_m I_w} \end{bmatrix} v_m \quad (11.169)$$

With a simple proportional controller for the voltage as a function of one of the spacecraft's perpendicular angular velocity components, e.g.,  $p$ , one can write

$$v_m = K_p p \quad (11.170)$$

Then, from an analysis of the roots of the closed-loop system, a spinning satellite with a WB-ANC system is nutationally stable for both  $I_{xx} < I_{zz}$  and  $I_{xx} > I_{zz}$ , i.e., the spin axis can be the major or minor axis.

### Dual-Spin Gyrostat Stabilization Systems

A **gyrostat stabilization** system uses an external prolate momentum device, called the rotor, joined to a *smaller* axisymmetric platform along a colinear spin axis at a bearing. As a rigid body, the platform and rotor have the same perpendicular angular velocity,  $\vec{\omega}_\perp$ , but have different components of angular velocity, along the spin axis  $\hat{k}_B$ , where the platform's spin rate is designed much slower than the rotor's spin rate. Also, notably the bearing must have an electric motor integrated in order to overcome the frictional resistance which would eventually cause the platform to match the rotor.

Under torque-free motion, the gyrostat attitude dynamics, taking the axisymmetric platform as the "momentum wheel" with angular velocity,  $\omega_1$  along the spin axis and motor-driven axisymmetric rotor as the body-fixed frame for the vehicle with angular velocity  $\vec{\omega}$ , one has

$$\dot{\vec{H}}_{G,v} + [\vec{\omega}] \times \vec{H}_{G,v} + \dot{\vec{H}}_{G_1}^1 + [\vec{\omega}] \times \vec{H}_{G_1}^1 = 0 \quad (11.171)$$

The total angular momentum about  $G$  is

$$\vec{H}_{G,v} = (I_G + I_{G,m_1}) \vec{\omega} \quad (11.172)$$

where  $I_G$  is the inertia tensor of the rotor about the center of mass of the platform-rotor system,  $G$ , and  $I_{G,m_1}$  is the inertia tensor of the concentrated mass of the platform about  $G$ . The angular momentum of the platform about its own center of mass is

$$\vec{H}_{G_1} = I_{G_1} \vec{\omega}_1 \quad (11.173)$$

The relative derivatives of the angular momentum are given

$$\dot{\vec{H}}_{G,v} = (I_G + I_{G,m_1}) \dot{\vec{\omega}} \quad (11.174)$$

and

$$\dot{\vec{H}}_{G_1} = I_{G_1} \dot{\vec{\omega}}_1 \quad (11.175)$$

as both the platform and rotor are axisymmetric. Thus, the gyrostat attitude dynamics can be written as

$$(I_{G,v} + I_{G,m_1}) \dot{\vec{\omega}} + [\vec{\omega}] \times (I_{G,v} + I_{G,m_1}) \vec{\omega} + I_{G_1} \dot{\vec{\omega}}_1 + [\vec{\omega}] \times I_{G_1} \vec{\omega}_1 \quad (11.176)$$

Then, defining the relative angular velocity of the platform to the rotor as  $\omega_{rel}$ , and the common  $\vec{\omega}_\perp$ , one

$$\vec{\omega}_{rel} = \vec{\omega}_1 - \vec{\omega} = \begin{bmatrix} p \\ q \\ r + \omega_{rel} \end{bmatrix} - \begin{bmatrix} p \\ q \\ r \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \omega_{rel} \end{bmatrix} \quad (11.177)$$

one has

$$(I_{G,v} + I_{G,1}) \dot{\vec{\omega}} + [\vec{\omega}] \times (I_{G,v} + I_{G,1}) \vec{\omega} + I_{G_1} \vec{\omega}_{rel} + [\vec{\omega}] \times I_{G_1} \vec{\omega}_{rel} \quad (11.178)$$

where  $I_{G,1} = I_{G,m_1} + I_{G_1}$  by the parallel axis formula.

This expression can be expanded by defining the components of the inertia tensor for the rotor about the system center of mass as

$$I_{G,v} = \begin{bmatrix} I_{xx,v} & 0 & 0 \\ 0 & I_{xx,v} & 0 \\ 0 & 0 & I_{zz,v} \end{bmatrix} \quad (11.179)$$

the components of the inertia tensor for the platform about the system center of mass as

$$I_{G,1} = \begin{bmatrix} I_{xx,1} & 0 & 0 \\ 0 & I_{xx,1} & 0 \\ 0 & 0 & I_{zz,1} \end{bmatrix} \quad (11.180)$$

and the components of the inertia tensor for the platform about its own center of mass as

$$I_{G_1} = \begin{bmatrix} I_{1_{xx}} & 0 & 0 \\ 0 & I_{1_{xx}} & 0 \\ 0 & 0 & I_{1_{zz}} \end{bmatrix} \quad (11.181)$$

where  $I_{zz,1} = I_{1_{zz}}$  as  $G$  and  $G_p$  lie on the same  $z_B$ -axis.

With these definitions and the definitions, one has

$$\begin{bmatrix} (I_{xx,v} + I_{xx,1}) \dot{p} \\ (I_{xx,v} + I_{xx,1}) \dot{q} \\ (I_{zz,v} + I_{zz,1}) \dot{r} \end{bmatrix} + \begin{bmatrix} (I_{zz,v} - I_{xx,v} + I_{zz,1} - I_{xx,1}) qr \\ (I_{xx,v} - I_{zz,v} + I_{xx,1} - I_{zz,1}) pr \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ I_{zz,1} \dot{\omega}_{rel} \end{bmatrix} + \begin{bmatrix} I_{zz,1} q \omega_{rel} \\ -I_{zz,1} p \omega_{rel} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (11.182)$$

or

$$\begin{bmatrix} I_{xx} \dot{p} + (I_{zz} - I_{xx}) qr \\ I_{xx} \dot{q} + (I_{xx} - I_{zz}) pr \\ I_{zz} \dot{r} \end{bmatrix} + \begin{bmatrix} I_{zz,1} q \omega_{rel} \\ -I_{zz,1} p \omega_{rel} \\ I_{zz,1} \dot{\omega}_{rel} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (11.183)$$

where  $I_{xx} = I_{xx,v} + I_{xx,1}$  and  $I_{zz} = I_{zz,v} + I_{zz,1}$ . These are three differential equations in four states,  $p$ ,  $q$ ,  $r$ , and  $\omega_{rel}$ . As mentioned, for dual-spin stabilization the electric motor at the bearing is used to set  $\omega_{rel}$ , thus  $\dot{\omega}_{rel} = 0$ . Thus,  $r = \bar{r}$  and one can change  $p$  and  $q$  as function of the relative spin rate  $\omega_{rel}$ .

Similar to the single-spinner, one can analyze this dual-spinner using an energy dissipation model. For the entire gyrostat, the angular momentum is

$$\vec{H}_G = \begin{bmatrix} I_{xx,v} p \\ I_{xx,v} q \\ I_{zz,v} r \end{bmatrix} + \begin{bmatrix} I_{xx,1} p \\ I_{xx,1} q \\ I_{zz,1}(r + \omega_{rel}) \end{bmatrix} \quad (11.184)$$

or with  $r_1 = r + \omega_{rel}$ , one has

$$\vec{H}_G = \begin{bmatrix} I_{xx} p \\ I_{xx} q \\ I_{zz,v} r + I_{zz,1} r_1 \end{bmatrix} \quad (11.185)$$

with magnitude

$$\|\vec{H}_G\|_2^2 = I_{xx}\omega_\perp^2 + (I_{zz,v}r + I_{zz,1}r_1)^2 \quad (11.186)$$

where  $\omega_\perp^2 = p^2 + q^2$ . For torque-free motion, one has

$$\frac{d\|\vec{H}_G\|_2^2}{dt} = 0 = I_{xx}\frac{d\omega_\perp^2}{dt} + 2(I_{zz,v}r + I_{zz,1}r_1)(I_{zz,v}\dot{r} + I_{zz,1}\dot{r}_1) \quad (11.187)$$

or

$$\frac{d\omega_\perp^2}{dt} = -\frac{2}{I_{xx}^2}(I_{zz,v}r + I_{zz,1}r_1)(I_{zz,v}\dot{r} + I_{zz,1}\dot{r}_1) \quad (11.188)$$

In addition, the total rotational kinetic energy of the gyrostat

$$T_R = T_{R,0} + T_{R,1} = \frac{1}{2}I_{xx}\omega_\perp^2 + \frac{1}{2}I_{zz,v}r^2 + \frac{1}{2}I_{zz,1}r_1^2 \quad (11.189)$$

where  $T_{R,0}$  is the energy of the rotor and  $T_{R,1}$  is the energy of the platform. Differentiating this with respect to time and rearranging, one has

$$\frac{d\omega_\perp^2}{dt} = \frac{2}{I_{xx}^2}(\dot{T}_{R,0} + \dot{T}_{R,1} - I_{zz,v}r\dot{r} - I_{zz,1}r_1\dot{r}_1) \quad (11.190)$$

Equating this with the expression from the angular momentum, one has

$$\frac{2}{I_{xx}^2}(\dot{T}_{R,0} + \dot{T}_{R,1} - I_{zz,v}r\dot{r} - I_{zz,1}r_1\dot{r}_1) = -\frac{2}{I_{xx}^2}(I_{zz,v}r + I_{zz,1}r_1)(I_{zz,v}\dot{r} + I_{zz,1}\dot{r}_1) \quad (11.191)$$

Rearranging, one has

$$\dot{T}_{R,0} + \dot{T}_{R,1} = \frac{I_{zz,v}}{I_{xx}}((I_{xx} - I_{zz,v})r - I_{zz,1}r_1)\dot{r} + \frac{I_{zz,1}}{I_{xx}}((I_{xx} - I_{zz,1})r_1 - I_{zz,v}r)\dot{r}_1 \quad (11.192)$$

Equating the two terms to the energy dissipation, one has

$$\dot{r} = \frac{I_{xx}}{I_{zz,v}} \frac{\dot{T}_{R,0}}{(I_{xx} - I_{zz,v})r - I_{zz,1}r_1} \quad (11.193)$$

and

$$\dot{r}_1 = \frac{I_{xx}}{I_{zz,1}} \frac{\dot{T}_{R,1}}{(I_{xx} - I_{zz,1})r_1 - I_{zz,v}r} \quad (11.194)$$

which allows one to rewrite

$$\frac{d\omega_\perp^2}{dt} = \frac{2}{I_{xx}} \left( \frac{\dot{T}_{R,0}}{I_{zz,1}\frac{r_1}{r} - (I_{xx} - I_{zz,v})r} + \frac{\dot{T}_{R,1}}{I_{zz,v} - (I_{xx} - I_{zz,1})\frac{r_1}{r}} \right) \left( I_{zz,v} + I_{zz,1}\frac{r_1}{r} \right) \quad (11.195)$$

An important class of dual-spin stabilized satellites exists where  $\frac{r_1}{r} \approx 0$ , e.g., a geosynchronous satellite

$$\frac{r_1}{r} = \frac{2\pi rad/day}{2\pi rad/s} \approx 10^{-5} \quad (11.196)$$

and for interplanetary spacecraft,  $r_1 = 0$ . Here, one has the simpler

$$\frac{d\omega_\perp^2}{dt} = \frac{2}{I_{xx}} \left( \dot{T}_{R,0} + \frac{I_{zz,v}}{I_{zz,v} - I_{xx}} \dot{T}_{R,1} \right) \quad (11.197)$$

Thus, as  $\dot{T}_{R,0} < 0$  and  $\dot{T}_{R,1} < 0$ , if the rotor is oblate, i.e.  $I_{zz,v} > I_{xx}$ , then the satellite is unconditionally stable. However, if the rotor is prolate, i.e.  $I_{zz,v} < I_{xx}$ , then the satellite is stable, if and only if

$$\dot{T}_{R,0} < \frac{I_{zz,v}}{I_{xx} - I_{zz,v}} \dot{T}_{R,1} \quad (11.198)$$

where the platform dissipation rate is typically augmented by passive nutation control systems.

### Small Euler Angle Maneuvers

While the primary purpose of a satellite attitude control system is to stabilize the attitude of the satellite against external disturbances, the secondary purpose is to perform attitude maneuvers for the system. Typically for small attitude changes, one uses the Euler angle errors for the feedback control system while for large attitude changes, one uses Euler symmetric parameter errors for the feedback control system to improve the tracking performance.

Many of these maneuvers occur in close proximity to an orbited body, e.g., Earth observations, the gravity-gradient equations of motion are relevant for operation. To this end, recall the linearized gravity-gradient equations of motion along the principal axes, one has

$$\begin{bmatrix} M_{d,x} + M_{c,x} \\ M_{d,y} + M_{c,y} \\ M_{d,z} + M_{c,z} \end{bmatrix} = \begin{bmatrix} I_{xx}\ddot{\phi} + 4\omega_O^2(I_{yy} - I_{zz})\phi - \omega_O(I_{xx} + I_{zz} - I_{yy})\dot{\psi} \\ I_{yy}\ddot{\theta} + 3\omega_O^2(I_{xx} - I_{zz})\theta \\ I_{zz}\ddot{\psi} + \omega_O^2(I_{yy} - I_{xx})\psi + \omega_O(I_{xx} + I_{zz} - I_{yy})\dot{\phi} \end{bmatrix} \quad (11.199)$$

which has decoupled pitch and roll-yaw dynamics that can be written in state-space form as

$$\begin{bmatrix} \dot{\theta} \\ \ddot{\theta} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -3\omega_O^2(I_{xx} - I_{zz})I_{yy}^{-1} & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} 0 \\ I_{yy}^{-1} \end{bmatrix} M_{c,y} \quad (11.200)$$

and

$$\begin{bmatrix} \dot{\phi} \\ \ddot{\phi} \\ \ddot{\psi} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -4\omega_O^2(I_{yy} - I_{zz})I_{xx}^{-1} & 0 & 0 & \omega_O(1 - (I_{yy} - I_{zz})I_{xx}^{-1}) \\ 0 & -\omega_O^2(I_{yy} - I_{xx})I_{zz}^{-1} & -\omega_O(1 - (I_{yy} - I_{xx})I_{zz}^{-1}) & 0 \end{bmatrix} \begin{bmatrix} \phi \\ \psi \\ \dot{\phi} \\ \dot{\psi} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ I_{xx}^{-1} & 0 \\ 0 & I_{zz}^{-1} \end{bmatrix} \begin{bmatrix} M_{c,x} \\ M_{c,z} \end{bmatrix} \quad (11.201)$$

Notably, if one neglects the gravity-gradient moment, one obtains the simplified attitude dynamics as three decoupled SISO systems that act as double integrators, i.e.,

$$\frac{\phi(s)}{M_{c,x}(s)} = \frac{1}{I_{xx}s^2} \quad (11.202)$$

$$\frac{\theta(s)}{M_{c,y}(s)} = \frac{1}{I_{yy}s^2} \quad (11.203)$$

and

$$\frac{\psi(s)}{M_{c,z}(s)} = \frac{1}{I_{zz}s^2} \quad (11.204)$$

SISO satellite attitude feedback control laws utilize some form of the PID control strategies, e.g., the PI with rate feedback. In addition, the use of feedforward elements are used instead of integrators to reduce the steady-state error under certain known disturbances.

### Coning Maneuvers

A **coning maneuver** is used to reorient the spin axis by some angle  $\theta_c$ . Consider a spinning satellite with some initial  $\vec{H}_{G,0}$  about the spin axis,  $z_B$ . A thruster pair impulsively fires creating some  $\Delta H_{G,1}$  normal to the spin axis. This induces a precession, or coning, of the satellite about an axis at an angle  $\theta_c/2$  to  $H_{G,0}$ . Following the torque-free motion model, the precession rate after the impulse is given by

$$\omega_p = \frac{I_{zz}\omega_s}{I_{xx} - I_{zz}} \sec\left(\frac{\theta_c}{2}\right) \quad (11.205)$$

Then, after precessing  $\pi$  radians, the thruster pair impulsively fires creating some  $\Delta H_{G,2}$  in the same direction relative to the spacecraft as  $\Delta H_{G,1}$  with  $\|\Delta \vec{H}_{G,2}\|_2 = \|\Delta \vec{H}_{G,1}\|_2$ , stabilizing the spin vector in the commanded reorientation,  $\theta$ . The total  $\Delta H$  for the single coning maneuver is

$$\Delta H_{tot} = \|\Delta \vec{H}_{G,1}\|_2 + \|\Delta \vec{H}_{G,2}\|_2 = 2\|\vec{H}_{G,0}\|_2 \tan\left(\frac{\theta_c}{2}\right) \quad (11.206)$$

The time required for this single coning maneuver,  $t_1$ , is found by dividing the precession angle by the rate, i.e.

$$t_1 = \frac{\pi}{\omega_p} = \pi \frac{I_{xx} - I_{zz}}{I_{zz}\omega_s} \cos\left(\frac{\theta_c}{2}\right) \quad (11.207)$$

However, as propellant expenditure is reflected in the magnitude of the individual angular momentum increments, one can reduce the amount of fully expended fuel by a sequence of  $N$  small coning maneuvers rather than one large maneuver. For these  $N$  small coning maneuvers, one has

$$\Delta H_{tot} = 2N\|\vec{H}_{G,0}\|_2 \tan\left(\frac{\theta_c}{2N}\right) \quad (11.208)$$

However, for large  $N$ , one has by the small angle approximation

$$\Delta H_{tot} \approx 2N\|\vec{H}_{G,0}\|_2 \left(\frac{\theta_c}{2N}\right) \approx \|\vec{H}_{G,0}\|_2 \theta_c \quad (11.209)$$

However, the time is increased to perform the reorientation, i.e.

$$t_N = N\pi \frac{I_{xx} - I_{zz}}{I_{zz}\omega_s} \cos\left(\frac{\theta_c}{2N}\right) \quad (11.210)$$

or in ratio form, one has by the small angle approximation

$$\frac{t_N}{t_1} = N \frac{\cos\left(\frac{\theta_c}{2N}\right)}{\cos\left(\frac{\theta_c}{2}\right)} \approx \frac{N}{\cos\left(\frac{\theta_c}{2}\right)} \quad (11.211)$$

### Momentum-Biased Attitude Control

**Momentum-biased satellites** are dual-spin stabilized satellites that do not consist of two external components, e.g., platform and rotor, but use an internal momentum wheel to stabilize a satellite in three axes. For these systems, the momentum bias stabilizes the out-of-plane axis rotation while the moment-generating capabilities of the wheel stabilize the in-plane axes rotations.

To this end, consider the linearized gravity-gradient equations of motion where additional moments in a circular orbit provide a total angular momentum of

$$\vec{H} = \vec{H}_B + \vec{H}_w = I_G \begin{bmatrix} p \\ q \\ r \end{bmatrix} + \begin{bmatrix} 0 \\ H_{w,y} \\ 0 \end{bmatrix} \quad (11.212)$$

where  $\vec{H}_B$  is the rigid-body angular momentum and  $\vec{H}_w$  is the combined wheel angular momentum. Momentum-biased satellites implement a constant momentum bias,  $\bar{H}_{w,y}$ , in the  $y_B$ -axis of the satellite with angular momentum which can also be accelerated or decelerated. In this case, the Euler equation of motion can be written as

$$\vec{M}_{d,B} + \vec{M}_{c,B} + \vec{M}_{g,B} = \begin{bmatrix} I_{xx}\dot{p} + (I_{zz} - I_{yy})qr \\ I_{yy}\dot{q} + (I_{xx} - I_{zz})pr \\ I_{zz}\dot{r} + (I_{yy} - I_{xx})pq \end{bmatrix} + \begin{bmatrix} -r\bar{H}_{w,y} \\ \dot{H}_{w,y} \\ p\bar{H}_{w,y} \end{bmatrix} \quad (11.213)$$

where  $\vec{M}_{d,B}$  is the disturbance moment vector,  $\vec{M}_{c,B}$  is the control moment vector, and  $\vec{M}_{g,B}$  is the gravity-gradient moment vector.

Furthermore, recalling the small angle approximations for the angular velocity as

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} \approx \begin{bmatrix} \dot{\phi} - \psi\omega_O \\ \dot{\theta} - \omega_O \\ \dot{\psi} + \phi\omega_O \end{bmatrix} \quad (11.214)$$

and the gravity-gradient moment vector as

$$\vec{M}_{g,B} = \begin{bmatrix} 3\omega_O^2(I_{zz} - I_{yy})\phi \\ 3\omega_O^2(I_{zz} - I_{xx})\theta \\ 0 \end{bmatrix} \quad (11.215)$$

one obtains the augmented linearized gravity-gradient equations of motion as

$$\begin{bmatrix} M_{d,x} + M_{c,x} \\ M_{d,y} + M_{c,y} \\ M_{d,z} + M_{c,z} \end{bmatrix} = \begin{bmatrix} I_{xx}\ddot{\phi} + (4\omega_O^2(I_{yy} - I_{zz}) - \omega_O\bar{H}_{w,y})\phi - (\omega_O(I_{xx} + I_{zz} - I_{yy}) + \bar{H}_{w,y})\dot{\psi} \\ I_{yy}\ddot{\theta} + 3\omega_O^2(I_{xx} - I_{zz})\theta + \bar{H}_{w,y} \\ I_{zz}\ddot{\psi} + (\omega_O^2(I_{yy} - I_{xx}) - \omega_O\bar{H}_{w,y})\psi + (\omega_O(I_{xx} + I_{zz} - I_{yy}) + \bar{H}_{w,y})\dot{\phi} \end{bmatrix} \quad (11.216)$$

which has decoupled pitch and roll-yaw dynamics.

Thus, for small Euler angles, one can rewrite these equations in state-space representation for the pitch dynamics as

$$\begin{bmatrix} \dot{\theta} \\ \ddot{\theta} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -3\omega_O^2(I_{xx} - I_{zz})I_{yy}^{-1} & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ I_{yy}^{-1} & I_{yy}^{-1} \end{bmatrix} \begin{bmatrix} \dot{H}_{w,y} \\ M_{c,y} \end{bmatrix} \quad (11.217)$$

and the roll-yaw dynamics as

$$\begin{bmatrix} \dot{\phi} \\ \dot{\psi} \\ \ddot{\phi} \\ \ddot{\psi} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ (-4\omega_O^2(I_{yy} - I_{zz}) + \omega_O \bar{H}_{w,y}) I_{xx}^{-1} & 0 & 0 \\ 0 & (-\omega_O^2(I_{yy} - I_{xx}) + \omega_O \bar{H}_{w,y}) I_{zz}^{-1} & (-\omega_O(I_{xx} + I_{zz} - I_{yy}) - \bar{H}_{w,y}) I_{zz}^{-1} \end{bmatrix} \begin{bmatrix} \phi \\ \psi \\ \dot{\phi} \\ \dot{\psi} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ I_{xx}^{-1} & 0 \\ 0 & I_{zz}^{-1} \end{bmatrix} \begin{bmatrix} M_{c,x} \\ M_{c,z} \end{bmatrix} \quad (11.218)$$

which include  $\dot{H}_{w,y}$  for inertia wheel control and  $\bar{H}_{w,y}$  due to the momentum bias.

To assess the inherent stability of a momentum-biased satellite, the pitch dynamics are marginally stable for  $I_{xx} > I_{zz}$  and unstable for  $I_{xx} < I_{zz}$  while the roll-yaw dynamics can be approximated by assuming that  $4\omega_O^2(I_{yy} - I_{zz}) \approx 0$ ,  $\omega_O(I_{xx} + I_{zz} - I_{yy}) \approx 0$ , and  $\omega_O^2(I_{yy} - I_{xx}) \approx 0$ , one has a characteristic equation as

$$\lambda^4 + \left( -\omega_O \bar{H}_{w,y} \left( I_{xx}^{-1} + I_{yy}^{-1} \right) + \bar{H}_{w,y}^2 I_{xx}^{-1} I_{zz}^{-1} \right) \lambda^2 + \bar{H}_{w,y}^2 I_{xx}^{-1} I_{zz}^{-1} \omega_O^2 = 0 \quad (11.219)$$

Furthermore, if one assumes  $\bar{H}_{w,y}^2 \gg \omega_O^2 I_{xx} I_{zz}$  and  $\bar{H}_{w,y} \gg \omega_O(I_{xx} + I_{zz})$ , then one has the simplified characteristic equation

$$\left( \lambda^2 + \omega_O^2 \right) \left( \lambda^2 + \bar{H}_{w,y} I_{xx}^{-1} I_{zz}^{-1} \right) = 0 \quad (11.220)$$

which consist of two purely imaginary conjugate pairs of poles, one at the orbital frequency,  $\omega_O$ , and the other at the nutation frequency,  $\omega_n$ , of the satellite which is proportional to the momentum bias of the satellite. Thus, the momentum bias itself is not enough to stabilize the roll-yaw dynamics.

Two different approaches can be used to stabilize the roll-yaw dynamics of a momentum-biased satellite. The first utilizes roll and yaw angles and rates in feedback, e.g., for a simple PD control law, one has

$$\begin{aligned} \frac{M_{c,x}(s)}{\phi(s)} &= K_{d,\phi}s + K_{p,\phi} \\ \frac{M_{c,z}(s)}{\psi(s)} &= K_{d,\psi}s + K_{p,\psi} \end{aligned} \quad (11.221)$$

The second utilizes roll-only angles and rates in feedback with a cross-coupled control law, e.g., for a simple PD control law, one has

$$\begin{aligned} \frac{M_{c,x}(s)}{\phi(s)} &= K_{d,\phi}s + K_{p,\phi} \\ \frac{M_{c,z}(s)}{\phi(s)} &= -K_\psi \end{aligned} \quad (11.222)$$

With this development of the momentum-biased satellite and the consideration of attitude maneuvers where one may not be using the principal axes, the fully coupled linearized dynamics may be applicable for the control design using MIMO linear control methods, e.g.,  $H_2$  and  $H_\infty$ . Considering the momentum of the rigid spacecraft,  $\vec{H} = [H_x \ H_y \ H_z]^T$  and the momentum of the MEDs  $\vec{H}_w = [H_{w,x} \ H_{w,y} \ H_{w,z}]^T$ , the generalized attitude equations of motion are

$$\vec{M}_d + \vec{M}_c = \dot{\vec{H}} + \dot{\vec{H}}_w + [\vec{\omega}_{B/I}] \times (H + H_w) \quad (11.223)$$

where  $\vec{M}_d$  is the disturbance input,  $\vec{M}_c$  is the control input, and  $\vec{\omega}_{B/I} = [\omega_x \ \omega_y \ \omega_z]^T$ . By component, one has

$$\begin{bmatrix} M_{d,x} + M_{c,x} \\ M_{d,y} + M_{c,y} \\ M_{d,z} + M_{c,z} \end{bmatrix} = \begin{bmatrix} \dot{H}_x + \omega_y H_z - \omega_z H_y \\ \dot{H}_y + \omega_z H_x - \omega_x H_z \\ \dot{H}_z + \omega_x H_y - \omega_y H_x \end{bmatrix} + \begin{bmatrix} \dot{H}_{w,x} + \omega_y H_{w,z} - \omega_z H_{w,y} \\ \dot{H}_{w,y} + \omega_z H_{w,x} - \omega_x H_{w,z} \\ \dot{H}_{w,z} + \omega_x H_{w,y} - \omega_y H_{w,x} \end{bmatrix} \quad (11.224)$$

Next, by substitution for small Euler angles and the orbital angular velocity,  $\omega_O$ , as well as  $H_{w,y} = \bar{H}_{w,y}$  for momentum-biased satellites, one has the coupled linearized attitude dynamics equation of motion as

$$\begin{bmatrix} M_{d,x} + M_{c,x} \\ M_{d,y} + M_{c,y} \\ M_{d,z} + M_{c,z} \end{bmatrix} = \begin{bmatrix} I_{xx}\ddot{\phi} + 4\omega_O^2(I_{yy} - I_{zz})\phi - \omega_O(I_{xx} + I_{zz} - I_{yy})\dot{\psi} + \dot{H}_{w,x} - \omega_O H_{w,z} - \bar{H}_{w,y}\dot{\psi} - \omega_O \bar{H}_{w,y}\phi - I_{xy}\ddot{\theta} - I_{xz}\ddot{\psi} \\ I_{yy}\ddot{\theta} + 3\omega_O^2(I_{xx} - I_{zz})\theta + \dot{H}_{w,y} - I_{xy}\ddot{\phi} + \omega_O^2 I_{xy}\phi + 2\omega_O I_{xy}\dot{\psi} - I_{yz}\ddot{\psi} - 2\omega_O I_{yz}\dot{\phi} + \omega_O^2 I_{yz}\phi \\ I_{zz}\ddot{\psi} + \omega_O^2(I_{yy} - I_{xx})\psi + \omega_O(I_{xx} + I_{zz} - I_{yy})\dot{\phi} + \dot{H}_{w,z} + \omega_O H_{w,z} + \bar{H}_{w,y}\dot{\phi} - \omega_O \bar{H}_{w,y}\psi - I_{yz}\ddot{\theta} - I_{xz}\ddot{\phi} \end{bmatrix} \quad (11.225)$$

which has coupled pitch, roll, and yaw dynamics and results in a ninth-order LTI system with state vector  $\vec{x} = [\dot{\phi} \ \dot{\theta} \ \dot{\psi} \ \phi \ \theta \ \psi \ H_{w,x} \ H_{w,z}]^T$ , and six inputs as  $\vec{u} = [M_{c,x} \ M_{c,y} \ M_{c,z} \ \dot{H}_{w,x} \ \dot{H}_{w,y} \ \dot{H}_{w,z}]^T$  which includes both potential external moment inputs and internal reaction wheel inputs.

## References

For more information, please refer to the following

- Curtis, H. D., “10.4 Dual-Spin Spacecraft,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 476-486
- Curtis, H. D., “10.6 Coning Maneuver,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 503-505
- Schmidt, D. K., “12.7 Elastic Effects and Structural-Mode Control,” *Modern Flight Dynamics*, 1st ed., Vol. 1, McGraw-Hill, New York, 2012, pp. 772-786
- Sidi, M. J., “4.8 Attitude Dynamics Equations of Motion for a Nonspinning Satellite,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 107-111
- Sidi, M. J., “6.2 Attitude Spin Stabilization during the  $\Delta v$  Stage,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 132-134
- Sidi, M. J., “6.3 Active Nutation Control,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 135-137

- Sidi, M. J., “6.6 Single-Spin Stabilization,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 144-148
- Sidi, M. J., “6.7 Dual-Spin Stabilization,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 148-151
- Sidi, M. J., “7.2 Equations for Basic Control Laws,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 152-159
- Sidi, M. J., “8.2 Stabilization without Active Control,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 210-214
- Sidi, M. J., “8.3 Stabilization with Active Control,” *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 2000, pp. 215-222

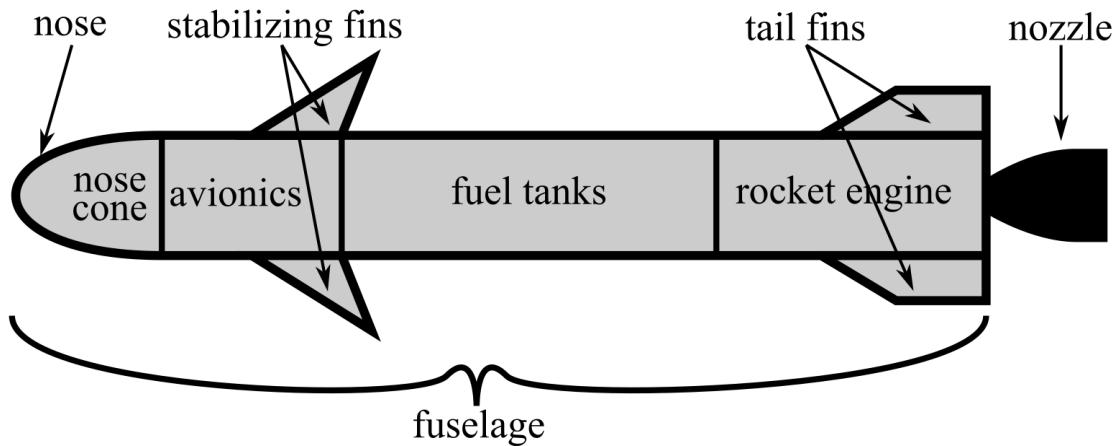
# Ballistic Vehicle Dynamics and Control Systems

## 12.1 Introduction to Orbital and Ballistic Vehicles

Space flight can be divided in two types, orbital and sub-orbital, also known as **ballistic**. Notably, sub-orbital and ballistic flight can also be completely atmospheric flight and not a combination of atmospheric and space flight. As a space flight vehicle has either first travel through the atmosphere as a hybrid flight vehicle or be transported through the atmosphere into space before being released, these two types of vehicles are typically analyzed together. **Satellites** are objects that travel in **orbits** around celestial bodies. Satellites can be natural, e.g. moons, planets, or **artificial satellites**, also known as **orbital vehicles**. In general, spacecraft can be any vehicle that flies beyond the atmosphere which provides a variety of designs.

### Ballistic/Sub-Orbital Vehicle Anatomy

The basic components of a conventional ballistic/sub-orbital vehicle design are shown in the following diagram

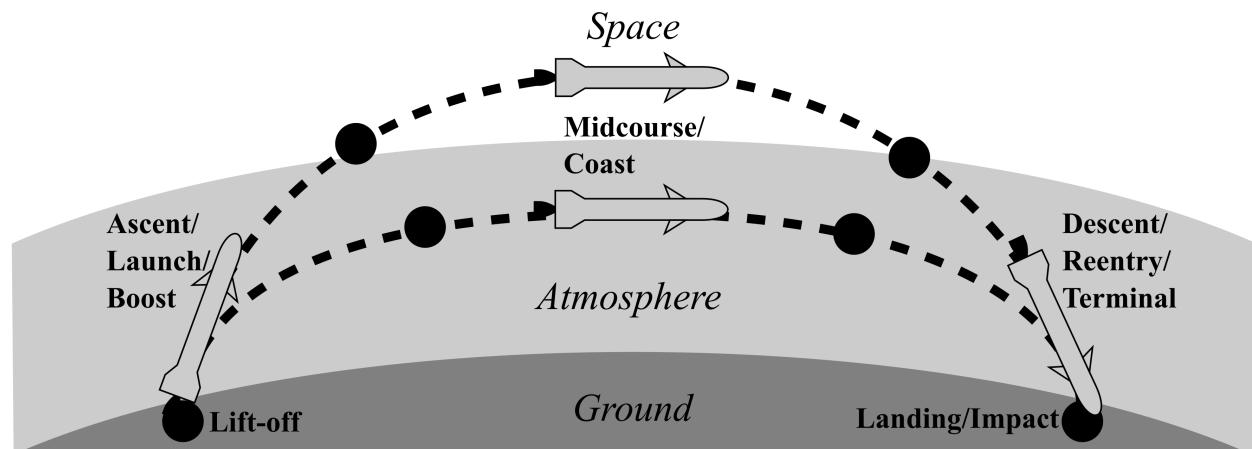


The **nose cone** is the front of the ballistic vehicle and houses the **fairing** typically for the payload which may include crew. The **fuselage** is the tubular structure of the ballistic vehicle that houses the avionics, fuel, oxidizer, and rocket engine. The **fins**, both stabilizing and tail, produce an up or down lift force to steer the ballistic vehicle. Some ballistic vehicles may have only tail fins. These fins also may have control surfaces on them called **rudders**. The **rocket engine** projects its reaction mass through the **nozzle** to produce the thrust force to overcome the ballistic vehicle's weight and fly. Some rocket-powered vehicle nozzle's may have thrust vectoring capabilities for control as well.

Notably, some ballistic vehicles are designed to travel to and from the surface of celestial bodies without an atmosphere, e.g., the Moon, which would change the design of the vehicle considerably.

### Phases of Ballistic Flight

The five typical phases of ballistic flight are **lift-off**; **ascent**, also known as **launch** or **boost**; **mid-course**, also known as **coast** if performing exoatmospheric flight; **descent**, **terminal** or **reentry** if it performed exoatmospheric flight, and **landing** or **impact**.



Notably, modern **ballistic missiles** may exhibit more elaborate flight plans, e.g., **skip** phase(s) of intermediate atmospheric exit and reentry phases and **glide** phase(s) of long unpowered atmospheric flight where lift is generated to maintain the longer flight time.

## 12.2 Point-Mass Dynamics for Ballistic Vehicles

**Coast**

**Boost**

**Reentry**

**References**

For more information, please refer to the following

- Curtis, H. D., “2.4 Angular Momentum and the Orbit Formulas,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 42-50
- Curtis, H. D., “2.7 Elliptical Orbits ( $0 < e < 1$ )”, *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 76-80
- Curtis, H. D., “4.6 Transformation between Geocentric Equatorial and Perifocal Frames,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 172-176
- Curtis, H. D., “7.2 Relative Motion in Orbit,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 316-322
- Curtis, H. D., “7.3 Linearization of the Equations of Relative Motion in Orbit,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 322-324
- Curtis, H. D., “7.4 Clohessy-Wiltshire Equations,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 324-329
- Li, Y., Liu, X., and Zing, G., “Discrete-time LQ optimal control of satellite formations in elliptical orbits based on feedback linearization,” in *Acta Astronautica*, Vol. 83, 2013
- Pesce, V., Colagrossi, A., and Silvestrini, S., “Chapter Two - Reference Systems and Planetary Models,” *Modern Spacecraft Guidance, Navigation, and Control*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2023, pp. 45-75
- Sullivan, J., Grimberg, S., and D’Amico, S., “Comprehensive Survey and Assessment of Spacecraft Relative Motion Dynamics Models,” in *Journal of Guidance, Control, and Dynamics*, Vol. 40, No. 8, 2017
- Yamanaka, K., and Ankersen, F., “New State Transition Matrix for Relative Motion on an Arbitrary Elliptical Orbit,” in *Journal of Guidance, Control, and Dynamics*, Vol. 25, No. 1, 2002

## 12.3 Ascent and Descent Guidance Systems

### Gravity Turn

A **gravity turn** or **zero-lift turn** is a maneuver used for spacecraft ascending into or descending from an orbit around a celestial body that uses gravity to steer the vehicle onto its planned trajectory. It offers two main advantages over a trajectory steered solely through the vehicle's own thrust. First, the thrust is not used to change the launch vehicle's direction, but only to accelerate the vehicle into orbit. Second, during the initial ascent phase the launch vehicle can maintain low or even zero angle of attack which minimizes transverse aerodynamic stress thereby requiring a lighter weight.

A launch vehicle begins its flight by flying straight up, gaining both vertical speed and altitude. During this initial launch phase, gravity acts directly against the thrust of the rocket, lowering its vertical acceleration which is known as **gravity drag** which can be minimized by executing the next launch phase, the pitchover maneuver, which is carried out while the vertical velocity is small to avoid large aerodynamic loads on the vehicle during the maneuver. The **pitchover maneuver** either deflects aerodynamic control surfaces or gimbals the rocket engine to direct some of its thrust to one side to create a net moment thereby turning the vehicle so that it no longer points vertically.

The simplest analysis of a gravity turn trajectory in inertial frame coordinates, neglecting air resistance, can be computed from the following equation

$$m \frac{d^2 \vec{x}}{dt^2} = \vec{T} - \begin{bmatrix} 0 \\ 0 \\ mg \end{bmatrix} \quad (12.1)$$

where if one constrains the thrust to point in the direction of the velocity, one

$$\begin{aligned} \ddot{\vec{x}} &= g(n - \cos \beta) \\ \dot{x} \beta \beta &= g \sin \beta \end{aligned} \quad (12.2)$$

where  $\beta$  is the angle between the velocity vector and the up direction

$$\beta = \cos^{-1} \left( \frac{\dot{\vec{x}}}{\|\dot{\vec{x}}\|_2} \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) \quad (12.3)$$

and  $n$  is the thrust-to-weight ratio defined as

$$n = \frac{\vec{T}}{mg} \quad (12.4)$$

If one assumes a constant acceleration due to gravity and a constant thrust-to-weight ratio, these equations can be solved analytically. However, numerical integration is more typical which provides the initial planned trajectory for a launch vehicle that may be computed as an open-loop guidance system or it may be regulated with a closed-loop guidance system to reject aerodynamic disturbances and thrust variations while maintaining the planned trajectory.

## Ascent Guidance

The **ascent phase**, also known as the **launch phase** or **boost phase**, is the first flight phase of a launch vehicle, ballistic missile, or hypersonic glide vehicle. The vast majority of **powered explicit guidance (PEG)** is

## Descent Guidance

The **descent phase**, also known as the **terminal phase**, is the final flight phase of a spacecraft leading to touchdown or impact. The descent phase of a spacecraft may be powered, i.e., thrust is produced by a propulsion system, or unpowered, no thrust is produced.

### Powered Descent Guidance PDG

**Apollo Powered Descent Guidance (APDG)** Ping Lu

## References

For more information, please refer to the following:

- Acikmese, B. and Ploen, S. R., “Convex Programming Approach to Powered Descent Guidance for Mars Landing,” *Journal of Guidance, Control, and Dynamics*, Vol. 30, No. 5, 2007, pp. 1353-1366
- Curtis, H. D., “13.2 Equations of Motion,” *Orbital Mechanics for Engineering Students*, 4th ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 706-708
- Dueri, D., Acikmese, B., Scharf, D. P., and Harris, M. W., “Customized Real-Time Interior-Point Methods for Onboard Powered-Descent Guidance,” *Journal of Guidance, Control, and Dynamics*, Vol. 40, No. 2, 2017, pp. 197-212
- Lu, P., “Augmented Apollo Powered Descent Guidance,” *Journal of Guidance, Control, and Dynamics*, Vol. 42, No. 3, 2019, pp. 447-457

## 12.4 Rigid Ballistic Vehicle Dynamics and Stability

A **projectile** is a vehicle that is initially propelled by the application of an external force called boost, and then moves freely under the influence of gravity and air resistance. **Ballistics** is the study of projectile dynamics concerned with their launching, flight behavior, and impact effects. Thus, projectiles are also known as **ballistic bodies**. For **ballistic vehicles** are projectiles that carry a payload.

### Magnus Force and Moment

### Projectile Linear Theory

## References

For more information, please refer to the following:

- Dykes, J. W., “Projectile Linear Theory for Aerodynamically Asymmetric Projectiles,” PhD Thesis, Georgia Institute of Technology, 2011
- McCoy, R. L., “Modern Exterior Ballistics: The Launch and Flight Dynamics of Symmetric Projectiles,” Schiffer Military History, Atglen, PA, 2012
- Ollerenshaw, D. and Costello, M., “Model predictive control of a direct fire projectile equipped with canards,” in *Journal of Dynamic Systems, Measurement, and Control*, vol. 130, 2008

## 12.5 Ballistic Vehicle Attitude Control Systems

## **Part III**

# **Probability and Perception Theory**

---

# Probability Theory

## 13.1 Introduction to Probability Theory

This chapter presents an overview of important concepts from probability theory for the purposes of perception systems. For brevity, it is devoid of proofs for the axioms, theorems, and concepts presented. The interested reader is encouraged to use the reference materials and other resources for more in-depth studies of probability theory which will provide further insight into the design and analysis of aerospace vehicle perception systems beyond this textbook. This chapter also introduces more set theory definitions which are reviewed in section A.1 in the appendix.

### Frequentist Probability Theory

One justification of probability is based on **frequentist probability theory**. Here one defines the **outcomes** as all possible results of a “yet-to-be-performed” experiment. The probabilities of occurrence can be assigned to sets of its potential values known as **events**. To assign probabilities to events from a frequentist viewpoint, one would perform an experiment  $N$  times with some number of possible outcomes corresponding to events  $E_1, E_2, \dots$ , count the number of times each event occurred, i.e.,  $N_1, N_2, \dots$ . Then, forming ratios of the number of occurrences of each events, i.e., the **relative frequencies**, one has

$$\frac{N_1}{N}, \frac{N_2}{N}, \dots \quad (13.1)$$

Then, a frequentist would assign the probabilities of each event based on the concept of **statistical regularity** in that these relative frequencies converge to some limiting value which provides the basis for the frequentist definition of probability as

$$\mathbb{P}(E_1) = \lim_{N \rightarrow \infty} \frac{N_1}{N}, \mathbb{P}(E_2) = \lim_{N \rightarrow \infty} \frac{N_2}{N}, \dots \quad (13.2)$$

As an example, consider rolling two dice. This experiment would have the following outcomes:

$$\begin{aligned}
 & (1, 1) \\
 & (1, \cdot) \\
 & (1, 6) \\
 & (2, 1) \\
 & (2, \cdot) \\
 & (2, 6) \\
 & (\cdot, \cdot) \\
 & (6, 6)
 \end{aligned} \tag{13.3}$$

which mathematically provides thirty-six possible outcomes for this experiment. Considering the random variable as the sum of the two dice, and the event  $E_1$  would be the sum of two dice being equal to 2, which has one outcome. Similarly, the event  $E_2$  could be assigned as the sum of dice begin equal to 3 which has two outcomes. This could be continued for all eleven possible events. Note that if the two dice are **fair**, i.e., each outcome is equally probable, then  $\mathbb{P}(E_1) = \frac{1}{36}$ ,  $\mathbb{P}(E_2) = \frac{2}{36} = \frac{1}{18}$ ,  $\dots$ .

### Axiomatic Probability Theory

Another justification of probability is Kolmogorov's **axiomatic probability theory** which relies on set theory and measure theory. In this theory, all possible outcomes of some non-empty phenomenon make up the set  $\Omega$  known as the **outcome space**, also known as the **possibility space** or the **sample space**. All events make up the **event space**,  $\mathcal{E}$ . The set function,  $\mathbb{P}$ , is a measure that returns any event's probability, i.e., it satisfies the first axiom

$$\mathbb{P}(E) \geq 0 \in \mathbb{R} \quad \forall E \in \mathcal{E} \tag{13.4}$$

with unit measure as the second axiom

$$\mathbb{P}(\Omega) = 1 \tag{13.5}$$

and for any mutually exclusive or pairwise disjoint events,  $E_1, E_2, \dots$ , it also satisfies the third axiom

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i) \tag{13.6}$$

which is known as  **$\sigma$ -additivity**.

From these three axioms, one can directly infer the following five corollaries for the probability of the empty set as

$$\mathbb{P}(\emptyset) = 0 \tag{13.7}$$

the probability of finite disjoint unions

$$\mathbb{P}\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n \mathbb{P}(E_i) \tag{13.8}$$

the probability of the complement as

$$\mathbb{P}(E^c) = 1 - \mathbb{P}(E) \quad (13.9)$$

the monotonicity of  $\mathbb{P}$  as

$$\text{if } E_1 \subseteq E_2 \text{ then } \mathbb{P}(E_1) \leq \mathbb{P}(E_2) \quad (13.10)$$

and the **addition law of probability** as

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2) \quad (13.11)$$

where  $\mathbb{P}(E_1 \cap E_2) = 0$  if events  $E_1$  and  $E_2$  are **mutually exclusive**. This final corollary can be extended to any number of sets as the **inclusion-exclusion principle**

$$\mathbb{P}\left(\bigcup_{i=1}^n E_i\right) = \sum_{k=1}^n \left( (-1)^{k-1} \sum_{I \subseteq \{1, \dots, n\}, |I|=k} \mathbb{P}\left(\bigcap_{i \in I} E_i\right) \right) \quad (13.12)$$

where the last sum iterates over all subsets  $I$  of the indices  $1, \dots, n$  which contain exactly  $k$  elements. It can also be noted that any **measure** is defined by axioms 1 and 3 and corollary 1 while using axiom 2 instead of corollary 1 is specific to probability measures.

In set theory, the use of these three axioms and their corollaries infer that the ordered triplet  $(\Omega, \mathcal{E}, \mathbb{P})$  forms a **probability space** and the collection,  $\mathcal{A}$ , of all possible events  $E_i \in \mathcal{E}$  is a  $\sigma$ -algebra on  $\Omega$ , i.e., it contains the empty set and is closed under complements and countable unions. The formulation of this definition is important for formalizing probability measures for uncountably infinite sample spaces,  $\Omega$ , where one must “construct” a countably infinite collection of events, e.g., the real number line which is done through the use of Borel sets. A **Borel set** is any set in a topological space that can be formed from open sets through the operations of countable union, countable intersection, and relative complement. For certain  $\Omega$ , the collection of all Borel sets on  $\Omega$  forms a Borel  $\sigma$ -algebra on  $\Omega$  which is the smallest  $\sigma$ -algebra containing all open sets, e.g., the Borel  $\sigma$ -algebra on  $\mathbb{R}$  is the smallest  $\sigma$ -algebra on  $\mathbb{R}$  that contains all possible *intervals* of the real numbers. Lastly, it should be noted that  $\mathbb{P}$  is defined only for events and need not be defined for outcomes. However, for countable outcome spaces  $\Omega$ , one can define an **elementary event**, also known as a **sample point** as an event that contains only a single outcome. Thus, allowing outcomes to be assigned probabilities indirectly for countable  $\Omega$ . Further details on this measure-theoretic basis for formal definitions of probability, especially for continuous random variables, is beyond the scope of this textbook, but is mentioned to provide readers with some exposure to formal ways of understanding the peculiarities of probability modeling.

## Fundamental Concepts in Probability Theory for Perception

An important concept in **conditional probability** is defined as the probability that the event  $E_1$  occurs given that event  $E_2$  has occurred which is denoted as

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} \text{ if } \mathbb{P}(E_2) \neq 0 \quad (13.13)$$

This can also be rearranged as

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1|E_2)\mathbb{P}(E_2) \quad (13.14)$$

Then, considering the conditional probability of  $E_2$  given  $E_1$  has occurred, one obtains **Bayes' theorem**, also known as **Bayes' law** or **Bayes' rule**, for two events as

$$\mathbb{P}(E_2|E_1) = \frac{\mathbb{P}(E_1|E_2)\mathbb{P}(E_2)}{\mathbb{P}(E_1)} \quad (13.15)$$

where  $\mathbb{P}(E_2)$  is called the **prior probability**,  $\mathbb{P}(E_2|E_1)$  is known as the **posterior probability**,  $\mathbb{P}(E_1|E_2)$  is known as the **likelihood**. Notably the quotient  $\mathbb{P}(E_1|E_2)/\mathbb{P}(E_1)$  is referred to as the **support** event  $E_1$  provides for event  $E_2$ .

Another important concept is to consider the event  $E_1$  occurring with event  $E_2$  or its complement  $E_2^c$ . First, note that by definition of the complement

$$E_1 = (E_1 \cap E_2) \cup (E_1 \cap E_2^c) \quad (13.16)$$

Thus, since  $E_2$  and  $E_2^c$  are a disjoint union, one has

$$\mathbb{P}(E_1) = \mathbb{P}(E_1 \cap E_2) + \mathbb{P}(E_1 \cap E_2^c) \quad (13.17)$$

or, in terms of conditional probabilities

$$\mathbb{P}(E_1) = \mathbb{P}(E_1|E_2)\mathbb{P}(E_2) + \mathbb{P}(E_1|E_2^c)\mathbb{P}(E_2^c) \quad (13.18)$$

This concept can be extended to any number of sets,  $E_2, \dots, E_n$ , which partition the entire outcome space into  $n$  disjoint events. This allows one to state the **law of total probability (LTP)** for event  $E_1$  conditioned on all other events  $E_i$  for  $i = 2, \dots, n$  as

$$\mathbb{P}(E_1) = \sum_{i=2}^n \mathbb{P}(E_1 \cap E_i) = \sum_{i=2}^n \mathbb{P}(E_1|E_i)\mathbb{P}(E_i) \quad (13.19)$$

The LTP also provides additional insight into Bayes theorem via the substitution for  $\mathbb{P}(E_1)$  for two events,  $E_1$  and  $E_2$ , as

$$\mathbb{P}(E_2|E_1) = \frac{\mathbb{P}(E_1|E_2)\mathbb{P}(E_2)}{\mathbb{P}(E_1|E_2)\mathbb{P}(E_2) + \mathbb{P}(E_1|E_2^c)\mathbb{P}(E_2^c)} \quad (13.20)$$

or for more than two events,  $E_2, \dots, E_n$ , as

$$\mathbb{P}(E_i|E_1) = \frac{\mathbb{P}(E_1|E_i)\mathbb{P}(E_i)}{\sum_{i=2}^n \mathbb{P}(E_1|E_i)\mathbb{P}(E_i)} \quad i = 2, \dots, n \quad (13.21)$$

Note that Bayes' theorem allows one to calculate probabilities of events as more model-based information is obtained via sensing another related event, e.g., if one has modeled the prior probability for events  $E_2, \dots, E_n$ , then observes event  $E_1$ , one can use that information to compute the posterior probability for events  $E_2, \dots, E_n$ , if one has a *probabilistic models* for their probabilities conditioned on  $E_1$ .

Lastly, with these concepts in mind, one can define **independence** between events  $E_1$  and  $E_2$  as

$$\begin{aligned} \mathbb{P}(E_1) &= \mathbb{P}(E_1|E_2) = \mathbb{P}(E_1|E_2^c) \\ \mathbb{P}(E_2) &= \mathbb{P}(E_2|E_1) = \mathbb{P}(E_2|E_1^c) \end{aligned} \quad (13.22)$$

which implies the relationships

$$\begin{aligned}\mathbb{P}(E_1 \cap E_2) &= \mathbb{P}(E_1)\mathbb{P}(E_2) \\ \mathbb{P}(E_1 \cup E_2) &= \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1)\mathbb{P}(E_2)\end{aligned}\tag{13.23}$$

It is important to note that independent is a property of events  $E_1$ ,  $E_2$ , and the measure  $\mathbb{P}$  while disjoint is a property only of the events  $E_1$  and  $E_2$ . Also, it can be shown that if  $E_1$  and  $E_2$  are independent, then  $E_1$  and  $E_2^c$ ,  $E_1^c$  and  $E_2$ , and  $E_1^c$  and  $E_2^c$  are independent.

This concept can be extended to more than two events,  $E_i$ ,  $i = 1, \dots, n$ , as **mutually independent** events if

$$\mathbb{P}\left(\bigcap_{i \in \mathcal{I}} E_i\right) = \prod_{i \in \mathcal{I}} \mathbb{P}(E_i)\tag{13.24}$$

for all possible finite subsets of events  $\mathcal{I}$ . However, if the previous equations holds only if  $|\mathcal{I}|$  is restricted to 2, then events  $E_i$  are only **pairwise independent**.

## References

For more information, please refer to the following

- Gubner, J.A., “Chapter 1 Introduction to Probability,” in *Probability and Random Processes for Electrical and Computer Engineers*, 1st Ed., Cambridge University Press, 2006, pp. 1-47
- Simon, D., “2.1 Probability,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 49-53

## 13.2 Random Variables

### Random Variables and Probability Distributions

A **random variable**,  $X$ , is a measurable function which returns values that *depend* on the outcomes of a random phenomenon, i.e.,  $X : \Omega \rightarrow \mathbb{S}$  where  $\mathbb{S}$  is called the **state space**, e.g.,  $\mathbb{R}$ . This textbook will represent random variables by the uppercase letters, e.g.  $X$ , and the values or the **realizations**, by lowercase letters, e.g.,  $x$ . Thus,  $X = x$  denotes the random variable  $X$  is equal to the value  $x$ , which differentiates between the random variables and their particular realizations. A **discrete random variable** depends on outcomes that take on any value in a *countable* outcome set.

Thus, for discrete random variables, one can assign probabilities for individual outcomes,  $x_i$ , for all  $i \in \mathbb{N}$ . A **continuous random variable** depends on outcomes that take on any value in an uncountable outcome set. Thus, one cannot assign probabilities for the exact values of individual outcomes. Thus, though exact values still occur as part of the outcome space, one has the property  $\mathbb{P}(X = x) = 0$ . However, one can assign probabilities to *intervals* of values in the uncountable set as the events as mentioned in the previous section. These definitions can be shown to infer the following different probability distributions models for discrete and continuous random variables. Further details on the connections between random variables and axiomatic probability theory is beyond the scope of this textbook.

For modeling discrete random variables, one typically uses the **probability mass functions (PMFs)** defined for  $x \in \mathbb{S} \subset \mathbb{R}$  as

$$p_X(x) = \mathbb{P}(X = x) \quad (13.25)$$

which must also satisfy that the probabilities of every possible event sum up to 100%, i.e.

$$\sum_i p_X(x_i) = \sum_{i=1}^{\infty} \mathbb{P}(X = x_i) = 1 \quad (13.26)$$

thus, an event must always occur. A related function is the **probability-generating function (PGF)**

$$G_X(z) = \sum_{x=0}^{\infty} p_X(x) z^x \quad (13.27)$$

which exists for  $x \in \mathbb{S} \subset \mathbb{R}$ . One can see that the PGF is equivalent to the  $z$ -transform of the PMF. Here  $p_X(x)$  can be obtained from the PGF by taking the derivatives of  $G_X$  via the formula

$$p_X(x) = \mathbb{P}(X = x) = \frac{G^{[x]}(0)}{x!} \quad (13.28)$$

For modeling continuous random variables, one typically uses **cumulative distribution functions (CDFs)** defined as

$$F_X(x) = \mathbb{P}(X \leq x) \quad (13.29)$$

From the properties of probability theory, the CDF is always a monotonically non-decreasing function, i.e., the accumulated probabilities as one moves from  $-\infty$  to  $\infty$  can never go down as an event cannot have a “negative” probability. Furthermore, the limits of the CDF must be

$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad (13.30)$$

and

$$\lim_{x \rightarrow \infty} F_X(x) = 1 \quad (13.31)$$

thus, as one approaches  $-\infty$ , the probability that  $X \leq -\infty$  should vanish and as one approaches  $\infty$ , the probability that  $X \leq \infty$  should approach 1. Furthermore, using the CDF, one can form the probability for  $X$  taking any realization between values  $a$  and  $b$  as

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) \quad (13.32)$$

For absolutely continuous CDFs, one also typically considers the **probability density function (PDF)** of a continuous random variable defined as

$$f_X(x) = \frac{dF_X(x)}{dx} \quad (13.33)$$

or

$$F_X(x) = \int_{-\infty}^x f_X(\zeta) d\zeta \quad (13.34)$$

For  $f_X(x)$  to be a properly defined PDF, the following equation must hold

$$\int_{-\infty}^{\infty} f_X(x) dx = 1 \quad (13.35)$$

It should be noted that one can intuitively think of the PDF as an infinitesimal PMF, i.e.

$$f_X(x) \approx \mathbb{P}(x < X \leq x + dx) \quad (13.36)$$

Two related functions are the **moment-generating function (MGF)**

$$M_X(s) = \int_{-\infty}^{\infty} f_X(x) e^{sx} dx \quad (13.37)$$

and the **characteristic function (CF)**

$$\varphi_X(\omega) = \int_{-\infty}^{\infty} f_X(x) e^{i\omega x} dx \quad (13.38)$$

where one can see that the MGF and CF are equivalent to the Laplace and Fourier transforms of the PDF, respectively. However, the MGF may not exist for all probability distributions while the CF always exists.

Lastly, the **complementary cumulative distribution function (CCDF)**, also known as the **tail distribution** or **exceedance**, is defined as

$$\bar{F}_X(x) = 1 - F_X(x) = \mathbb{P}(X \geq x) \quad (13.39)$$

Sometimes, this is combined into the **folded cumulative distribution function**

$$\tilde{F}_X(x) = F_X(x) 1_{\{F_X(x) \leq 0.5\}} + \bar{F}_X(x) 1_{\{F_X(x) > 0.5\}} \quad (13.40)$$

Depending on the properties of  $F_X(x)$ , the inverse of the CDF can be defined differently. If  $F_X(x)$  strictly increasing and continuous,  $F_X^{-1}(p)$   $p \in [0, 1]$  is the **inverse distribution function**, a.k.a. the **quantile function**. Otherwise, if there is no *unique* inverse, the inverse CDF can be defined as its **generalized inverse distribution function (GIDF)**, given by

$$F_X^{-1}(p) = \inf\{x \in \mathbb{R} : F_X(x) \geq p\} \quad \forall p \in [0, 1] \quad (13.41)$$

which preserves the inverse CDF properties

- $F_X^{-1}$ : non-decreasing function
- $F_X^{-1}(xF(x)) \leq x$
- $F_X(F_X^{-1}(p)) \geq p$
- $F_X^{-1}(p) \leq x$  if and only if  $p \leq F_X(x)$

One estimate of the CDF of a random variable is to take  $N$  independent samples. Then, one can state formally that if  $X_1, \dots, X_N$  are independent, identically distributed (IID) real random numbers with CDF,  $F_X(x)$ , then the **empirical distribution function (EDF)**,  $\hat{F}_{X,N}(x)$ , is given by the number of elements in sample  $\leq x$  divided by number of samples,  $N$ , i.e.

$$\hat{F}_{X,N}(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{X_i \leq x}(x_i) \quad (13.42)$$

where  $\mathbf{1}_A(x)$  is the **indicator** of event  $A$ , i.e.

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \quad (13.43)$$

One can quantify the statistical difference between the EDF and the CDF via the **Kolmogorov-Smirnov (KS) statistic** which is defined for a given CDF,  $F_X(x)$ , as

$$D_n = \sup_x |\hat{F}_{X,N}(x) - F_X(x)| \quad (13.44)$$

Intuitively, the KS statistic takes the largest absolute difference between the two distribution functions across all  $x$  values. By the **Gilvenko-Cantelli theorem** as  $N \rightarrow \infty$ , the EDF converges almost surely to the CDF, i.e.,

$$\hat{F}_{X,N}(x) \rightarrow F_X(x) \quad (13.45)$$

Lastly, the **Dvoretzky-Kiefer-Wolfowitz (DKW) inequality** uses the KS statistic to form a confidence interval for the CDF given an EDF through the inequality

$$\mathbb{P}\left(\sup_x |\hat{F}_{X,N}(x) - F_X(x)| > \epsilon\right) \leq 2 \exp(-2N\epsilon^2) \quad (13.46)$$

A **finite mixture distribution** is defined as a distribution that is a weighted sum of a finite  $N$  number of probability distributions, i.e., a discrete finite mixture distributions has a PMF as

$$p_X(x) = \sum_{i=1}^N w_i p_i(x) \quad (13.47)$$

and a continuous finite mixture distribution has a CDF as

$$F_X(x) = \sum_{i=1}^N w_i F_i(x) \quad (13.48)$$

where  $p_i(x)$  is the  $i^{\text{th}}$  PMF term,  $F_i(x)$  is the  $i^{\text{th}}$  CDF term, and  $w_i \geq 0$  is the  $i^{\text{th}}$  weight with

$$\sum_{i=1}^N w_i = 1 \quad (13.49)$$

A **mixed random variable** allows some countable subset of outcomes to be assigned specific probabilities as part of an entire uncountable outcome space. These are formed as a mixture of both continuous and discrete probability distributions where the discrete values are modeled using the Dirac delta function, e.g.,  $F_i(x) = w_i\delta(x)$ .

A **scale mixture random variable** is a random variable,  $Y$ , of the form

$$Y = \sqrt{Z} X \quad (13.50)$$

with  $Z > 0$  independent from  $X$ . Here the distribution  $X|(Z = z)$  is defined as the primary distribution and  $Z$  is the **scale parameter** or **generating variate**. Thus,

$$f_Y(x) = \int_0^\infty f_{X|Z}(x|z)f_Z(z)dz \quad (13.51)$$

is well-defined for a scale mixture.

### Statistics of Random Variables

Furthermore, one often desires to compute some characteristic of the values a random variable may take, known as a **statistic**. There are many different types of statistics; however, many common statistics are defined using the expectation operator  $\mathbb{E}[\bullet]$ , which for the discrete random variable  $X$  is defined as

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i p_X(x_i) \quad (13.52)$$

and for continuous random variable  $X$  as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx \quad (13.53)$$

which notably can also be computed for any function of a random variable. One of the most common functions are the **moments** of a random variable, where the  $i^{\text{th}}$  moment of a discrete random variable  $X$  is defined as

$$\mathbb{E}[X^i] = \sum_{j=1}^{\infty} x_j^i p_X(x_j) \quad (13.54)$$

and for continuous random variable  $X$  as

$$\mathbb{E}[X^i] = \int_{-\infty}^{\infty} x^i f_X(x) dx \quad (13.55)$$

It should be noted that this allows one to redefine the MGF for random variables as

$$\begin{aligned} M_X(s) &= \mathbb{E}[e^{sx}] \\ &= 1 + s\mathbb{E}[X] + \frac{s^2\mathbb{E}[X^2]}{2!} + \frac{s^3\mathbb{E}[X^3]}{3!} + \dots \end{aligned} \quad (13.56)$$

and for  $Y = mX + b$  where  $m$  and  $b$  are scalars, one has

$$\begin{aligned} M_{mX+b}(s) &= \mathbb{E}[e^{s(mx+b)}] \\ M_{mX+b}(s) &= \mathbb{E}[e^{smx} e^{bs}] \\ M_Y(s) &= e^{bs} M_X(ms) \end{aligned} \quad (13.57)$$

In addition, the  $i^{\text{th}}$  **central moment** of a discrete random variable  $X$  is defined as

$$\mathbb{E}[(X - \mu_X)^i] = \sum_{j=1}^{\infty} (x_j - \mu_X)^i p_X(x_j) \quad (13.58)$$

and for continuous random variable  $X$  as

$$\mathbb{E}[(X - \mu_X)^i] = \int_{-\infty}^{\infty} (x - \mu_X)^i f_X(x) dx \quad (13.59)$$

The  $i^{\text{th}}$  **standardized moment** of a discrete random variable  $X$  is defined as

$$\mathbb{E}\left[\left(\frac{X - \mu_X}{\sigma_x}\right)^i\right] = \sum_{j=1}^{\infty} \left(\frac{x_j - \mu_X}{\sigma_x}\right)^i p_X(x_j) \quad (13.60)$$

and for continuous random variable  $X$  as

$$\mathbb{E}\left[\left(\frac{X - \mu_X}{\sigma_x}\right)^i\right] = \int_{-\infty}^{\infty} \left(\frac{x - \mu_X}{\sigma_x}\right)^i f_X(x) dx \quad (13.61)$$

Four moments of random variables have particular names. These are the **mean**, denoted by  $\mu_X$ , which is the first moment, the **variance**, denoted by  $\sigma_x^2$ , which is second central moment, the **skewness** which is the third standardized moment, and the **kurtosis** which is the fourth standardized moment. Three other common statistics of a random variable  $X$  are the **median**, denoted by  $m_x$ , and defined by the equation

$$m_X = F_X^{-1}(0.5) \quad (13.62)$$

the **mode**, denoted by  $Mo_x$ , and defined by

$$Mo_X = \operatorname{argmax}_x f_X(x) \quad (13.63)$$

and the **standard deviation**, denoted by  $\sigma_X$ , and defined by

$$\sigma_X = \sqrt{\sigma_x^2} \quad (13.64)$$

## Fundamental Discrete Probability Distributions

The **Poisson distribution**, denoted by  $X \sim Pois(\lambda)$ , has a PMF of the form

$$p_X(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad (13.65)$$

with  $\lambda > 0$  is the Poisson rate and can be used to model the number of events,  $x$ , that may occur in some time interval. The mean and variance of the Poisson distribution is  $\lambda$ .

The **Bernoulli distribution**, denoted by  $X \sim Ber(r)$ , has a PMF of the form

$$p_X(x; r) = \begin{cases} 1 - r & \text{if } X = x_1 \\ r & \text{if } X = x_2 \end{cases} \quad (13.66)$$

and can be used to model event with only two outcomes, e.g., existence or non-existence.

The **multi-Bernoulli distribution**, denoted by  $X \sim MBer(r_1, \dots, r_N)$ , is given by the union of  $N$  independent Bernoulli-distributed random variables and has a PMF of the form

$$p_X(x; N, p) = \begin{cases} \prod_{j=1}^N (1 - r_j) & \text{if } X_1 = \dots = X_N = x_1 \\ \vdots & \vdots \\ \prod_{j=1}^N (1 - r_j) \sum_{1 \leq i_1 \neq \dots \neq i_n \leq N} \prod_{j=1}^n \frac{r_{i_j}}{1 - r_{i_j}} & \text{if } X_{i_1} = \dots = X_{i_n} = x_1, X_{i_{n+1}} = \dots = X_{i_N} = x_2 \\ \vdots & \vdots \\ \prod_{j=1}^N r_{i_j} & \text{if } X_1 = \dots = X_N = x_2 \end{cases} \quad (13.67)$$

where the sum for each intermediate outcome is taken over all permutations of  $n \leq N$  of the  $N$  constituent Bernoulli-distributed random variables. The numerator is the probability density that the Bernoulli components with indices  $i_1, \dots, i_n$  generate realizations  $x_2$ . The leading constant cancels with the divisors inside the sum to give the probability that the leftover Bernoulli components with indices  $i_{n+1}, \dots, i_N$  generate realizations  $x_1$ . This can be used to model  $N$  simultaneous events with only two outcomes, e.g.,  $N$  existences or non-existences.

## Fundamental Continuous Univariate Probability Distributions

The **Gaussian distribution**, also known as the **normal distribution**, denoted by  $X \sim \mathcal{N}(\mu, \sigma^2)$ , has a PDF of the form

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right] \quad (13.68)$$

where  $\mu$  is the mean and  $\sigma^2$  is the variance. The **standard normal distribution** has  $\mu = 0$  and  $\sigma^2 = 1$  which provides the simpler PDF

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} y^2 \right] \quad (13.69)$$

and thereby rewrite  $f_X(x; \mu, \sigma^2)$  in terms of  $f_Y(y)$  as

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sigma} f_Y \left( \frac{x - \mu}{\sigma} \right) \quad (13.70)$$

It should be noted that while an analytical form exists for the PDF of a Gaussian, the CDF doesn't not admit an analytical solution and must be numerically approximated.

The mean of a Gaussian distribution is also the median and mode. The skewness of a Gaussian is 0 and the kurtosis is 3. The prevalence of the Gaussian distribution leads to the definition of the **excess kurtosis** as the kurtosis minus 3. The Gaussian distribution typically is used for computing **confidence regions** as the probability that a realization  $x$  of  $X \sim \mathcal{N}(\mu, \sigma^2)$  lies within some interval around the mean. Typically values are

$$\Pr(\mu - \sigma < X \leq \mu + \sigma) = 0.68 \quad (13.71)$$

$$\Pr(\mu - 2\sigma < X \leq \mu + 2\sigma) = 0.95 \quad (13.72)$$

$$\Pr(\mu - 3\sigma < X \leq \mu + 3\sigma) = 0.997 \quad (13.73)$$

The **Cauchy distribution**, denoted by  $X \sim \text{Cauchy}(x_0, \gamma)$ , has a PDF of the form

$$f_X(x; x_0, \gamma) = \left( \pi \gamma \left[ 1 + \left( \frac{x - x_0}{\gamma} \right)^2 \right] \right)^{-1} \quad (13.74)$$

where  $x_0$  is the location and  $\gamma$  is the scale. Notably, the mean, variance, skewness, and kurtosis for the Cauchy are undefined. However,  $x_0$  is the median and the mode. The **standard Cauchy distribution** occurs for  $x_0 = 0$  and  $\gamma = 1$ . Notably, if  $X \sim \mathcal{N}(0, 1)$  and  $Y \sim \mathcal{N}(0, 1)$ , then  $X/Y \sim \text{Cauchy}(0, 1)$

The **Pareto distribution**, denoted by  $X \sim \text{Pareto}(\gamma, \beta)$ , has a PDF of the form

$$f_X(x; \gamma, \beta) = \begin{cases} \frac{1}{\beta} \left( 1 + \frac{\gamma}{\beta} x \right)^{-\frac{\gamma+1}{\gamma}} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (13.75)$$

where  $\gamma \geq 0$  is the shape and  $\beta > 0$  is the scale. For  $\gamma = 0$ , one has

$$f_X(x; 0, \beta) = \begin{cases} e^{-\frac{x}{\beta}} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (13.76)$$

The Pareto distribution is used to model heavy-tailed distribution,  $F_X$ , due to its appearance in **Extreme Value Theory (EVT)** as a limiting distribution, as the threshold  $u \rightarrow \infty$ , of its conditional excess distribution function (CEDF) defined as

$$F_u(y) = \mathbb{P}(X - u \leq y | X > u) = \frac{F_X(u + y) - F_X(u)}{1 - F_X(u)} \quad (13.77)$$

A **heavy-tailed distribution** is defined as a probability distributions whose tails are not exponentially bounded, i.e.,

$$\lim_{x \rightarrow \infty} e^{tx} \bar{F}_X(x) = \infty \quad \forall t > 0 \quad (13.78)$$

The **Levy  $\alpha$ -stable ( $\alpha$ S) distribution**, denoted by  $X \sim S(\alpha, \beta, c, \mu)$ , has a CF of the form

$$\varphi_{\vec{X}}(\omega; \alpha, \beta, c, \mu) = \exp(i\mu\omega - |c\omega|^\alpha(1 - i\beta\text{sign}(\omega)\Phi)) \quad (13.79)$$

where

$$\Phi = \begin{cases} \tan\left(\frac{\pi\alpha}{2}\right) & \alpha \neq 1 \\ -\frac{2}{\pi} \log|\omega| & \alpha = 1 \end{cases} \quad (13.80)$$

where  $0 < \alpha \leq 2$  is the **stability index**, also known as the **tail index** or **characteristic exponent**,  $-1 \leq \beta \leq 1$  is the skewness,  $c > 0$  is the scale parameter, and  $\mu \in \mathbb{R}$  is the shift parameter. The **symmetric  $\alpha$ -stable (SaS) distribution** occurs for  $\beta = 0$ . Notably, the Gaussian distribution occurs for  $\alpha = 2$  and  $\beta = 0$ , the **Cauchy distribution** occurs for  $\alpha = 1$  and  $\beta = 0$ , and the **Levy distribution** occurs for  $\alpha = 0.5$  and  $\beta = 1$ . Other than these three, no analytical form exists for the CDF or PDF.

Furthermore, the variance, skewness, and kurtosis are only defined for  $\alpha = 2$  and the mean is not defined for  $\alpha \leq 1$ . The reason for the name is that  $\alpha$ -stable distributions are closed under convolution for a fixed value of  $\alpha$ , i.e., sums of independent  $\alpha$ -stable distributions remain  $\alpha$ -stable if  $\alpha$  is constant, albeit  $\beta$ ,  $\gamma$ , and  $\delta$  may change.

An important theorem in perception is the **generalized central limit theorem** which states the sum of IID random variables with distributions having infinite variance, i.e., power-law “Pareto” tails, decreasing as  $|x|^{-\alpha-1}$  where  $0 < \alpha \leq 2$ , will tend to an  $\alpha$ -stable distribution  $S(\alpha, 0, c, 0)$  as the number of summands grows. If the symmetric distributions have finite variance, i.e.,  $\alpha > 2$ , then the **central limit theorem** states the sum converges to a Gaussian distribution, i.e., a stable distribution with stability parameter equal to 2.

The (**student's**)  **$t$  distribution**, denoted by  $X \sim t(\nu)$ , has a PDF of the form

$$f_X(x; \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (13.81)$$

where  $\nu > 0$  is the degrees of freedom and  $\Gamma$  is the gamma function

$$\Gamma(n) = (n-1)! \quad (13.82)$$

Notably, the **standard Cauchy distribution** occurs for  $\nu = 1$  and the standard normal distribution occurs as  $\nu \rightarrow \infty$ . Thus, the  $t$  distribution can be used to model heavier tails than the standard normal distribution. The mean of the  $t$  distribution is 0 for  $\nu > 1$  and undefined otherwise. The variance of the  $t$  distribution is  $\nu(\nu-2)^{-1}$  for  $\nu > 1$  and undefined otherwise. The skewness of the  $t$  distribution is 0 for  $\nu > 3$  and undefined otherwise. The excess kurtosis of the  $t$  distribution is  $6(\nu-4)^{-1}$  for  $\nu > 4$ ,  $\infty$  for  $2 < \nu \leq 4$ , and undefined otherwise.

The  **$\Gamma$  distribution**, denoted by  $X \sim \Gamma(\alpha, \theta)$ , has a PDF of the form

$$f_X(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-\beta x} \beta^\alpha}{\Gamma(\alpha)} \quad (13.83)$$

where  $\alpha > 0$  is the shape parameter and  $\beta > 0$  is the rate parameter. Two special cases of the  $\Gamma$  distribution are used in perception systems. The **exponential distribution**, denoted by  $X \sim Exp(\lambda)$ , occurs when  $X \sim \Gamma(1, \lambda)$  has a PDF of the form

$$f_X(x; \beta) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (13.84)$$

where  $\lambda > 0$  is the rate parameter. The mean of the exponential is  $\beta$  and the variance is  $\beta^2$ . The **central  $\chi^2$  distribution**, denoted by  $X \sim \chi^2(\nu)$ , occurs when  $X \sim \Gamma(k/2, 2)$  and has a PDF of the form

$$f_X(x; \nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\nu/2-1}e^{-x/2} \quad (13.85)$$

where  $\nu > 0$  is the degrees of freedom. The **noncentral  $\chi^2$  distribution**, denoted by  $X \sim \chi^2(\nu, \lambda)$ , has a PDF of the form

$$f_X(x; \nu, \lambda) = \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^i}{i!} f_Y(x; \nu + 2i) \quad (13.86)$$

where  $f_Y(x; \nu + 2i)$  is a central  $\chi^2$  distribution and  $\lambda > 0$  is the non-centrality parameter. The **scaled inverse- $\chi^2$  distribution**, denoted by  $X \sim I\chi^2(\tau^2, \nu)$ , has a PDF of the form

$$f_X(x; \tau^2, \nu) = \frac{(\tau^2\nu/2)^{\nu/2}}{\Gamma(k/2)x^{k/2-1}} \exp\left(-\frac{\nu\tau^2}{2x}\right) \quad (13.87)$$

where  $\tau^2$  is the scale parameter and  $\nu$  is the degrees of freedom. If  $X \sim \chi^2(k)$ , then  $X^{-1} \sim I\chi^2(1/k, k)$ .

The  $\chi^2$  distributions are used in detection and analysis of perception systems. An important property of the  $\chi^2$  distribution if one has the random variable,  $Y$ , given by

$$Y = \sum_{i=1}^k X_i^2 \quad (13.88)$$

where  $X_1, \dots, X_k$  are independent, normally distributed random variables with parameters  $\mu_i$  and  $\sigma_i = 1$  for  $i = 1, \dots, k$ , then  $Y \sim \chi^2(k, \lambda)$  with

$$\lambda = \sum_{i=1}^k \mu_i^2 \quad (13.89)$$

Furthermore, if  $\mu_1 = \dots = \mu_k = 0$ , then  $Y \sim \chi^2(k)$ .

The **Laplace distribution**, also known as the **double exponential distribution**, denoted by  $X \sim Laplace(\mu, \beta)$ , has a PDF of the form

$$f_X(x; \mu, \beta) = \frac{1}{2\beta} \exp\left(-\frac{|x - \mu|}{\beta}\right) \quad (13.90)$$

where  $\mu$  is the location parameter and  $\beta > 0$  is the scale parameter. The **K distribution**, denoted by  $X \sim K(\alpha, \lambda)$ , has a PDF of the form

$$f_X(x; \alpha, \lambda) = \frac{2\lambda}{\Gamma(\alpha)} (\sqrt{\lambda x})^{\alpha-1} K_{\alpha-1}(2\sqrt{\lambda x}) \quad (13.91)$$

where  $K_{\alpha-1}$  is the modified Bessel function of the second kind. The  $K$  distribution is used in radar, electro-optical, and sonar sensor modeling.

## References

For more information, please refer to the following

- Gubner, J.A., “Chapter 2 Introduction to Discrete Random Variables,” in *Probability and Random Processes for Electrical and Computer Engineers*, 1st Ed., Cambridge University Press, 2006, pp. 1-47
- Gubner, J.A., “Chapter 3 More About Discrete Random Variables,” in *Probability and Random Processes for Electrical and Computer Engineers*, 1st Ed., Cambridge University Press, 2006, pp. 1-47
- Gubner, J.A., “Chapter 4 Continuous Random Variables,” in *Probability and Random Processes for Electrical and Computer Engineers*, 1st Ed., Cambridge University Press, 2006, pp. 1-47
- Gubner, J.A., “Chapter 5 Cumulative Distributions Functions and Their Applications,” in *Probability and Random Processes for Electrical and Computer Engineers*, 1st Ed., Cambridge University Press, 2006, pp. 1-47
- Nolan, J. P., “1 Basic Properties of Univariate Stable Distributions” in *Stable Distributions – Models for Heavy Tailed Data*, 2017, pp. 3-22
- Simon, D., “2.2 Random Variables,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 53-59
- Simon, D., “2.3 Transformations of Random Variables,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 59-61

### 13.3 Random Vectors

When considering a finite collection of  $n$  random variables, one has a **random vector** with dimension  $n$  defined as

$$\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \quad (13.92)$$

with a corresponding **realization vector** that can be defined as

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad (13.93)$$

## Probability Functions of Random Vectors

To model the probabilities and relationships between the random elements of a discrete random vector, one typically uses the **joint probability mass function (joint PMF)**, defined as

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) \quad (13.94)$$

which can be written in vector form as

$$p_{\vec{X}}(\vec{x}) = \mathbb{P}(\vec{X} = \vec{x}) \quad (13.95)$$

To model the probabilities and relationships between the random elements of a continuous random vector, one typically uses the **joint cumulative distribution function (joint CDF)**, defined as

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) \quad (13.96)$$

which can be written in vector form as

$$F_{\vec{X}}(\vec{x}) = \mathbb{P}(\vec{X} \leq \vec{x}) \quad (13.97)$$

Furthermore, by taking the  $n$  partial derivatives of the joint CDF, one can define the **joint probability density function (joint PDF)** as

$$f_{\vec{X}}(\vec{x}) = \frac{\partial^n F_{\vec{X}}(\vec{x})}{\partial x_1 \dots \partial x_n} \quad (13.98)$$

which can be considered intuitively as

$$\mathbb{P}(x_1 < X_1 \leq x_1 + dx_1, \dots, x_n < X_n \leq x_n + dx_n) \quad (13.99)$$

Reversing this relationship, one can also relate the joint CDF and joint PDF by

$$F_{\vec{X}}(\vec{x}) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f_{\vec{X}}(x_1, \dots, x_n) dx_1 \dots dx_n \quad (13.100)$$

where for any PDF, including a joint PDF, the following is true.

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\vec{X}}(x_1, \dots, x_n) dx_1 \dots dx_n = 1 \quad (13.101)$$

Similarly to random variables, three related functions are the **probability-generating function (PGF)**

$$G_{\vec{X}}(\vec{z}) = \sum_{\vec{x}=0}^{\infty} p_{\vec{X}}(\vec{x}) z_1^{x_1} \dots z_n^{x_n} \quad (13.102)$$

the **moment-generating function (MGF)**

$$M_{\vec{X}}(\vec{s}) = \mathbb{E}[\exp \vec{s}^T \vec{x}] \quad (13.103)$$

and the **characteristic function (CF)**

$$\varphi_{\vec{X}}(\vec{\omega}) = \mathbb{E}[\exp i\vec{\omega}^T \vec{x}] \quad (13.104)$$

The **empirical distribution function (EDF)** for random vectors,  $\hat{F}_{\vec{X},N}(\vec{x})$ , can be defined analogously to the random variable case, i.e.,

$$\hat{F}_{\vec{X},N}(\vec{x}) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{X_i \leq x}(x_i) \quad (13.105)$$

and the multivariate **Dvoretzky-Kiefer-Wolfowitz (DKW) inequality** uses the KS statistic to form a confidence interval for the joint CDF given an EDF through the inequality

$$\mathbb{P} \left( \sup_x |\hat{F}_{\vec{X},N}(\vec{x}) - F_{\vec{X}}(\vec{x})| > \epsilon \right) \leq 2n_x \exp(-2N\epsilon^2) \quad (13.106)$$

### Bayes' Rule for Random Vectors

For discrete random vectors, the **marginal PDF** can be defined for any element of  $\vec{X}$  as

$$p_{X_i}(x) = \sum_{x_1 \cdots x_{i-1}, x_i=x, x_{i+1} \cdots x_n} p_{\vec{X}}(\vec{x}) \quad (13.107)$$

which can also be generalized to be any sub-vector. Thus, the general marginal PMF also leads to the random vector definition for the **conditional PMF** of  $\vec{X}$  as the probability of  $\vec{X} = \vec{x}$  given  $\vec{Y} = \vec{y}$ , i.e.,

$$p_{\vec{X}|\vec{Y}}(\vec{x}|\vec{y}) = \frac{p_{\vec{X},\vec{Y}}(\vec{x},\vec{y})}{p_{\vec{Y}}(\vec{y})} \quad (13.108)$$

Extending Bayes' rule for discrete random variables,  $\vec{X}$  and  $\vec{Y}$ , one has

$$p_{\vec{X}|\vec{Y}}(\vec{x}|\vec{y}) = \frac{p_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})p_{\vec{X}}(\vec{x})}{p_{\vec{Y}}(\vec{y})} \quad (13.109)$$

Extending the **law of total probability (LTP)** for discrete random variables, one can relate marginal and conditional PMFs as

$$p_{\vec{Y}}(\vec{y}) = \sum_{\vec{x}} p_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})p_{\vec{X}}(\vec{x}) \quad (13.110)$$

Note that by using the LTP for  $p_{\vec{Y}}(\vec{y})$ , one also has for Bayes' Rule

$$p_{\vec{X}|\vec{Y}}(\vec{x}|\vec{y}) = \frac{p_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})p_{\vec{X}}(\vec{x})}{\sum_{\vec{x}} p_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})p_{\vec{X}}(\vec{x})} \quad (13.111)$$

where the denominator essentially serves as a normalizing factor for the product  $p_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})p_{\vec{X}}(\vec{x})$  as  $\sum_{\vec{x}} p_{\vec{X}|\vec{Y}}(\vec{x}|\vec{y}) = 1$  by definition of a PMF.

For continuous random vectors, the **marginal PDF** can be defined for any element of  $\vec{X}$  as

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\vec{X}}(\vec{x}) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n \quad (13.112)$$

which can also be generalized to be any sub-vector. Thus, the general marginal PDF also leads to the random vector definition for the **conditional PDF** of  $\vec{X}$  as the probability of  $\vec{X} = \vec{x}$  given  $\vec{Y} = \vec{y}$ , i.e.,

$$f_{\vec{X}|\vec{Y}}(\vec{x}|\vec{y}) = \frac{f_{\vec{X},\vec{Y}}(\vec{x},\vec{y})}{f_{\vec{Y}}(\vec{y})} \quad (13.113)$$

Extending Bayes' rule for continuous random variables,  $\vec{X}$  and  $\vec{Y}$ , one has

$$f_{\vec{X}|\vec{Y}}(\vec{x}|\vec{y}) = \frac{f_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})f_{\vec{X}}(\vec{x})}{f_{\vec{Y}}(\vec{y})} \quad (13.114)$$

Extending the **law of total probability (LTP)** for continuous random variables, one can relate marginal and conditional PDFs as

$$f_{\vec{Y}}(\vec{y}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\vec{Y}|\vec{X}}(\vec{y}|\vec{\zeta})f_{\vec{X}}(\vec{\zeta})d\vec{\zeta} \quad (13.115)$$

Note that by using the LTP for  $f_{\vec{Y}}(\vec{y})$ , one also has for Bayes' Rule

$$f_{\vec{X}|\vec{Y}}(\vec{x}|\vec{y}) = \frac{f_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})f_{\vec{X}}(\vec{x})}{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\vec{Y}|\vec{X}}(\vec{y}|\vec{\zeta})f_{\vec{X}}(\vec{\zeta})d\vec{\zeta}} \quad (13.116)$$

where the denominator essentially serves as a normalizing factor for the product  $f_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})f_{\vec{X}}(\vec{x})$  as  $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\vec{Y}|\vec{X}}(\vec{y}|\vec{x})f_{\vec{X}}(\vec{x})dx = 1$  by definition of a PDF.

## Functions of Random Vectors

In perception systems, one often uses functions of random vectors, in particular, invertible functions,  $h$ , of random vectors, e.g.

$$\vec{Y} = h(\vec{X}) \quad (13.117)$$

For discrete random variables, the PMF of  $\vec{Y}$  in terms of the PMF of  $\vec{X}$  is

$$p_{\vec{Y}}(\vec{y}) = \mathbb{P}(\vec{Y} = \vec{y}) = \mathbb{P}(h(\vec{X}) = \vec{y}) = \mathbb{P}(\vec{X} = h^{-1}(\vec{y})) = p_{\vec{X}}(h^{-1}(\vec{y})) \quad (13.118)$$

and for continuous random variables, the CDF of  $\vec{Y}$  in terms of the CDF of  $\vec{X}$  is

$$F_{\vec{Y}}(\vec{y}) = \mathbb{P}(\vec{Y} \leq \vec{y}) = \mathbb{P}(h(\vec{X}) \leq \vec{y}) = \begin{cases} \mathbb{P}(\vec{X} \leq h^{-1}(\vec{y})) = F_{\vec{X}}(h^{-1}(\vec{y})) & h^{-1} \text{ increasing} \\ \mathbb{P}(\vec{X} \geq h^{-1}(\vec{y})) = 1 - F_{\vec{X}}(h^{-1}(\vec{y})) & h^{-1} \text{ decreasing} \end{cases} \quad (13.119)$$

which by the definition of the CDF and derivatives, one can show that the PDF of  $\vec{Y}$  given  $\vec{X}$  is

$$f_{\vec{Y}}(\vec{y}) = \left[ \frac{f_{\vec{X}}(\vec{x})}{\left| \det \frac{\partial h(\vec{x})}{\partial \vec{x}} \right|} \right]_{\vec{x}=h^{-1}(\vec{y})} \quad (13.120)$$

which is generally difficult to handle directly for general functions, but is approximated using many different methods.

However, a closed-form solution exists if  $h()$  is the **affine transformation**

$$\vec{Y} = h(\vec{x}) = H\vec{X} + \vec{b} \quad (13.121)$$

Thus,

$$h^{-1}(\vec{y}) = H^{-1}(\vec{y} - \vec{b}) \quad (13.122)$$

and

$$\frac{\partial h(\vec{x})}{\partial \vec{x}} = H \quad (13.123)$$

then, the PDF of  $\vec{Y}$  can be written as

$$f_{\vec{Y}}(\vec{y}) = \frac{f_{\vec{X}}(H^{-1}(\vec{y} - \vec{b}))}{|\det H|} \quad (13.124)$$

Another important result for perception is that the PMF of the sum of two independent discrete random vectors  $\vec{Z} = \vec{X} + \vec{Y}$  is given by

$$p_{\vec{Z}}(\vec{z}) = \mathbb{P}(\vec{Z} = \vec{z}) = \sum_{\vec{x}} \mathbb{P}(\vec{Z} = \vec{z} | \vec{X} = \vec{x}) \mathbb{P}(\vec{X} = \vec{x}) \quad (13.125)$$

which by substitution for the event that  $\vec{Z} = \vec{z} | \vec{X} = \vec{x}$  as  $\vec{Y} = \vec{z} - \vec{x}$  since  $\vec{Y} = \vec{Z} - \vec{X}$  is

$$p_{\vec{Z}}(\vec{z}) = \sum_{\vec{x}} \mathbb{P}(\vec{Y} = \vec{z} - \vec{x}) \mathbb{P}(\vec{X} = \vec{x}) \quad (13.126)$$

for independent discrete random variables  $\vec{X}$  and  $\vec{Y}$ , one has

$$p_{\vec{Z}}(\vec{z}) = \sum_{\vec{x}} p_{\vec{Y}}(\vec{z} - \vec{x}) p_{\vec{X}}(\vec{x}) \quad (13.127)$$

which is a convolution summation between the PMFs of  $\vec{X}$  and  $\vec{Y}$ . Thus, due to the properties of the  $z$ -transform, this can be written in terms of the PGFs as

$$G_{\vec{Z}}(\vec{z}) = G_{\vec{X}}(\vec{z}) G_{\vec{Y}}(\vec{z}) \quad (13.128)$$

Likewise, the sum of two independent continuous random variables  $\vec{Z} = \vec{X} + \vec{Y}$  is given by

$$F_{\vec{Z}}(\vec{z}) = \mathbb{P}(\vec{Z} \leq \vec{z}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} F_{\vec{Z}|\vec{X}}(\vec{z} | \vec{x}) f_{\vec{X}}(\vec{x}) d\vec{x} \quad (13.129)$$

which by substitution for the event that  $\vec{Z} \leq \vec{z} | \vec{X} = \vec{x}$  as  $\vec{Y} \leq \vec{z} - \vec{x}$  since  $\vec{Y} = \vec{Z} - \vec{X}$  is

$$F_{\vec{Z}}(\vec{z}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} F_{\vec{Y}}(\vec{z} - \vec{x}) f_{\vec{X}}(\vec{x}) d\vec{x} \quad (13.130)$$

or, in similar fashion, one can obtain

$$F_{\vec{Z}}(\vec{z}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} F_{\vec{X}}(\vec{z} - \vec{y}) f_{\vec{Y}}(\vec{y}) d\vec{y} \quad (13.131)$$

Also, taking the derivative with respect to  $\vec{z}$ , one has

$$f_{\vec{Z}}(\vec{z}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\vec{Y}}(\vec{z} - \vec{x}) f_{\vec{X}}(\vec{x}) d\vec{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\vec{X}}(\vec{z} - \vec{y}) f_{\vec{Y}}(\vec{y}) d\vec{y} \quad (13.132)$$

which is a convolution integral between the PDFs of independent discrete random variables  $\vec{X}$  and  $\vec{Y}$ . Thus, due to the properties of the Fourier transform, this can be written in terms of the CFs as

$$\varphi_{\vec{Z}}(\vec{z}) = \varphi_{\vec{X}}(\vec{z}) \varphi_{\vec{Y}}(\vec{z}) \quad (13.133)$$

### Statistics of Random Vectors

To characterize the statistics of different random vectors, one can use the expectation operator which operates on each individual random variable within the vector, i.e.

$$\mathbb{E}[\vec{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix} \quad (13.134)$$

For discrete random vectors

$$\mathbb{E}[\vec{X}] = \sum_{\vec{x}} \vec{x} p_{\vec{X}}(\vec{x}) \quad (13.135)$$

and for continuous random vectors, one has

$$\mathbb{E}[\vec{X}] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \vec{x} f_{\vec{X}}(\vec{x}) d\vec{x} \quad (13.136)$$

Thus, the multivariate analog for the first moment of  $\vec{X}$  is the **mean of a random vector** is simply

$$\mu_{\vec{X}} = \begin{bmatrix} \mu_{X_1} \\ \vdots \\ \mu_{X_n} \end{bmatrix} \quad (13.137)$$

However, higher moments of random vectors, one must consider the relationship *between* the random variables within the vector. Thus, the multivariate analog of the second moment of  $\vec{X}$  is the **correlation matrix** defined as

$$R_{XX} = \mathbb{E}[\vec{X} \vec{X}^T] = \begin{bmatrix} \mathbb{E}[X_1^2] & \cdots & \mathbb{E}[X_1 X_n] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[X_n X_1] & \cdots & \mathbb{E}[X_n^2] \end{bmatrix} \quad (13.138)$$

while the second centralized moment of  $\vec{X}$  is the **covariance matrix**, also known as the **variance-covariance matrix**, defined as

$$\begin{aligned} C_{XX} &= \mathbb{E} [(\vec{X} - \mu_{\vec{X}})(\vec{X} - \mu_{\vec{X}})^T] = \begin{bmatrix} \mathbb{E} [(X_1 - \mu_{X_1})^2] & \cdots & \mathbb{E} [(X_1 - \mu_{X_1})(X_n - \mu_{X_n})] \\ \vdots & \ddots & \vdots \\ \mathbb{E} [(X_n - \mu_{X_n})(X_1 - \mu_{X_1})] & \cdots & \mathbb{E} [(X_n - \mu_{X_n})^2] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{n,1} \\ \vdots & \ddots & \vdots \\ \sigma_{1,n} & \cdots & \sigma_n^2 \end{bmatrix} \end{aligned} \quad (13.139)$$

which is a symmetric positive definite matrix. It can easily be shown that

$$R_{XX} = C_{XX} + \mu_{\vec{X}}\mu_{\vec{X}}^T \quad (13.140)$$

The individual elements of the covariance matrix are called the variances on the diagonal, i.e.

$$\sigma_i^2 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_{X_i})^2 f_{\vec{X}}(x_1, \dots, x_n) dx_1 \dots dx_n \quad (13.141)$$

and the covariances on the off-diagonal, i.e.

$$\sigma_{i,j} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_{X_i})(x_j - \mu_{X_j}) f_{\vec{X}}(x_1, \dots, x_n) dx_1 \dots dx_n \quad (13.142)$$

Lastly, the inverse of the covariance is known as the **precision matrix**.

Furthermore, if one relates the correlation for each covariance element in  $C_{XX}$ , one has

$$\rho_{i,j} = \rho_{i,j} \sigma_i \sigma_j \quad (13.143)$$

where  $\rho_{i,j} \in [-1, 1]$  is called the **correlation coefficient** between  $X_i$  and  $X_j$ . Thus, the elements of  $\vec{X}$  are **uncorrelated** if  $\rho_{i,j} = \sigma_{i,j} = 0$ , or in terms of expectations

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i] \mathbb{E}[X_j] \quad \forall i \neq j \quad (13.144)$$

If the elements of  $\vec{X}$  are uncorrelated for all  $i \neq j \rightarrow$ , then the correlation matrix is a diagonal matrix. Furthermore, the elements of  $\vec{X}$  are **pairwise independent** if

$$p_{X_i, X_j}(x_i, x_j) = p_{X_i}(x_i)p_{X_j}(x_j) \quad \text{or} \quad F_{X_i, X_j}(x_i, x_j) = F_{X_i}(x_i)F_{X_j}(x_j) \quad \forall i \neq j \quad (13.145)$$

It can be shown that independence implies uncorrelation, however, it does not hold vice versa.

It should be noted that this concepts of covariance, correlation, and independence can be extended to *different* random vectors, e.g.  $\vec{X}$  and  $\vec{Y}$ . The correlation matrix can be generalized to the **cross-correlation matrix** defined as

$$R_{XY} = \mathbb{E} [\vec{X} \vec{Y}^T] \quad (13.146)$$

The covariance matrix can be generalized to the **cross-covariance matrix** defined as

$$C_{XY} = \mathbb{E} [(\vec{X} - \mu_{\vec{X}})(\vec{Y} - \mu_{\vec{Y}})^T] = R_{XY} - \mu_{\vec{X}}\mu_{\vec{Y}}^T \quad (13.147)$$

Independence between  $\vec{X}$  and  $\vec{Y}$  occurs if

$$p_{\vec{X}, \vec{Y}}(\vec{x}, \vec{y}) = p_{\vec{X}}(\vec{x})p_{\vec{Y}}(\vec{y}) \quad \text{or} \quad F_{\vec{X}, \vec{Y}}(\vec{x}, \vec{y}) = F_{\vec{X}}(\vec{x})F_{\vec{Y}}(\vec{y}) \quad \forall \vec{x}, \vec{y} \quad (13.148)$$

**Statistical distances**, denoted as  $D(X, Y)$ , are used to determine the degree of similarity between two statistical objects  $X$  and  $Y$ , i.e., between realizations or between probability distributions. Distances are important to perception system design for determining an algorithm's performance, e.g., the error of estimation with respect to some truth source, and for comparing different algorithms with each other. A statistical distance,  $D(X, Y)$  is called a **metric** if it satisfies the following conditions:

- 1 Non-negativity:  $D(X, Y) \geq 0$
- 2 Identity of indiscernibles:  $D(X, Y) = 0$  if and only if  $X = Y$
- 3 Symmetry:  $D(X, Y) = D(Y, X)$
- 4 Triangle inequality:  $D(X, Z) \leq D(X, Y) + D(Y, Z)$

where a **pseudometric** does not satisfy 2, a **quasimetric** does not satisfy 3, a **semimetric** does not satisfy 4, and a **divergence** satisfies 1 and 2.

One common metric for perception systems is the **Mahalanobis distance** which can be defined between the probability distribution,  $X$  on  $\mathbb{R}^n$ , with mean  $\vec{\mu}_x$  and covariance  $P$ , and the probability distribution,  $Y$  on  $\mathbb{R}^n$ , with mean  $\vec{\mu}_y$  and the same covariance,  $P$ , as

$$D_M(X, Y) = \|\vec{\mu}_x - \vec{\mu}_y\|_{P^{-1}} = \sqrt{(\vec{\mu}_x - \vec{\mu}_y)P^{-1}(\vec{\mu}_x - \vec{\mu}_y)} \quad (13.149)$$

or, alternatively, between the realization  $\vec{x} \in \mathbb{R}^n$  and the probability distribution,  $Y$ , as

$$D_M(\vec{x}, Y) = \|\vec{x} - \vec{\mu}_y\|_{P^{-1}} = \sqrt{(\vec{x} - \vec{\mu}_y)P^{-1}(\vec{x} - \vec{\mu}_y)} \quad (13.150)$$

One common divergence for perception systems is the **Kullback-Leibler (KL) divergence**, also known as the **relative entropy** or **information gain**, which can be defined for the discrete probability distribution,  $X$  on  $\mathcal{X}$ , with PMF  $p_{\vec{X}}(\vec{x})$ , and the discrete probability distribution,  $Y$  on  $\mathcal{X}$ , with PMF  $p_{\vec{Y}}(\vec{y})$ , as

$$D_{KL}(X||Y) = \sum_{\vec{x} \in \mathcal{X}} p_{\vec{X}}(\vec{x}) \log \left( \frac{p_{\vec{X}}(\vec{x})}{p_{\vec{Y}}(\vec{x})} \right) \quad (13.151)$$

or, alternatively, between the continuous probability distribution,  $X$  on  $\mathbb{R}^n$ , with PDF  $f_{\vec{X}}(\vec{x})$ , and the continuous probability distribution,  $Y$  on  $\mathbb{R}^n$ , with PDF  $f_{\vec{Y}}(\vec{x})$ , as

$$D_{KL}(X||Y) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\vec{X}}(\vec{x}) \log \left( \frac{f_{\vec{X}}(\vec{x})}{f_{\vec{Y}}(\vec{x})} \right) d\vec{x} \quad (13.152)$$

This is typically referred to as the “divergence of  $X$  from  $Y$ ” or the “information gain from  $X$  over  $Y$ .” In a Bayesian context,  $Y$  is the prior distribution and  $X$  is the posterior distribution.

## Samples of Random Vectors

A sample realization of a random variable or vector is known as a **random variate**. Thus, generating samples of random variables is known as **random variate generation** which can be uniform or non-uniform, also known as true **random number generation (RNG)** or **pseudo-random number generator (PRNG)**, respectively. Both of these types of generation are used in control and perception system modeling and simulations. True RNGs depend on some natural phenomenon that can be sampled using hardware, e.g., thermal noise, shot noise, jitter, and metastability of electronic circuits. However, in many applications, one can approximate the statistical nature of RNG using PRNG algorithms.

The basic method of PRNG utilizes the **uniform distribution**, represented by  $U \sim \mathcal{U}(a, b)$ , which has a CDF of the form

$$F_X(u; a, b) = \begin{cases} 0 & u < a \\ \frac{u-a}{b-a} & u \in [a, b] \\ 1 & x > b \text{ otherwise} \end{cases} \quad (13.153)$$

and PDF of the form

$$f_X(u; a, b) = \begin{cases} \frac{1}{b-a} & u \in [a, b] \\ 0 & \text{otherwise} \end{cases} \quad (13.154)$$

and can be used to model events when all values in the range  $[a, b]$  are equally likely. Then, for generating a random variate,  $x$ , from an arbitrary probability distribution with CDF,  $F_X(x)$ , one can use **inverse transform sampling**, also known as the **Smirnov transform**, by generating a random variate,  $u$ , from  $U \sim \mathcal{U}(0, 1)$  and computing

$$x = F_X^{-1}(u) \quad (13.155)$$

where  $F_X^{-1}$  is the generalized inverse CDF.

Notably, the inverse transform sampling requires an analytical expression or numerical tables for the CDF. Thus, an alternative sampling method is **rejection sampling**, also known as **acceptance-rejection sampling**, where one generates a random variate,  $y$ , from **proposal distribution** with PDF,  $f_Y$ , and a random variate,  $u$ , from  $U \sim \mathcal{U}(0, 1)$  and “accept”  $y$  as drawn from  $X$  with PDF,  $f_X$ , if

$$u < \frac{f_X(y)}{M f_Y(y)} \quad (13.156)$$

where  $1 \geq M < \infty$  is chosen to satisfy  $f_X(y) \leq M f_Y(y)$  for all values of  $x$  and a sample will be “accepted” after  $M$  iterations on average. Notably, the support of  $Y$  must include the support of  $X$ . In practice, a lower value of  $M$  is preferred as there will be fewer rejected samples; however, this requires that  $f_Y(y)$  should resemble  $f_X(y)$  in some fashion while also satisfying its bounding property.

For sampling random vectors of dimension  $n$ , one can take advantage of **Sklar's theorem** which states that every joint CDF can be decomposed into its marginal CDFs and a **copula**, i.e.,

$$F_{\vec{X}}(\vec{x}) = C(F_{X_1}(x_1), \dots, F_{X_n}(x_n)) \quad (13.157)$$

where  $C : [0, 1]^n \rightarrow [0, 1]$  is an  $n$ -dimensional copula which is a specific type of joint CDF. Furthermore, if the joint PDF is available, one also has

$$f_{\vec{X}}(\vec{x}) = c(f_{X_1}(x_1), \dots, f_{X_n}(x_n)) \quad (13.158)$$

where  $c$  is the density of the  $n$ -dimensional copula. Thus, by sampling each marginal distribution via inverse transform or rejection sampling, one can simply compute the output of the joint CDF via the copula. Intuitively, one can think of the copula as directly modeling the dependence between random variables. Often, copulas are used in estimating the distributions of random vectors by separating the marginal estimation problems and the copula estimation problem.

### Fundamental Continuous Multivariate Probability Distributions

The **multivariate Gaussian distribution**, denoted as  $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ , has a joint PDF of the form

$$f_{\vec{X}}(\vec{x}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{0.5}} \exp \left[ -\frac{1}{2} (\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}) \right] \quad (13.159)$$

where  $\vec{\mu}$  is the mean and  $\Sigma$  is the covariance and completely characterize the random vector. The PDF can be integrated to obtain the joint Gaussian CDF for which there is no analytical solution. Typically this integration is approximated using numerical methods. A useful property of the multivariate Gaussians is that independence and uncorrelatedness are equivalent, i.e. one implies the other.

Another useful property of the multivariate Gaussian is that Gaussian marginal PDFs are also Gaussian distributed and can be formed by simply dropping the unnecessary elements from  $\vec{\mu}$  and  $\Sigma$ . As an example, consider the 2-dimensional case, i.e.,

$$\vec{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad (13.160)$$

with

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad (13.161)$$

and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix} \quad (13.162)$$

Then, the marginal distribution of  $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$  where one has dropped  $\mu_2, \sigma_2, \sigma_{1,2}$  and the marginal distribution of  $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$  where one has dropped  $\mu_1, \sigma_1, \sigma_{1,2}$ .

A third useful property of random vectors is for independent multivariate Gaussians  $\vec{X} \sim \mathcal{N}(\vec{\mu}_X, \Sigma_X)$  and  $\vec{Y} \sim \mathcal{N}(\vec{\mu}_Y, \Sigma_Y)$ , their sum is defined as

$$\vec{Z} = \vec{X} + \vec{Y} \quad (13.163)$$

can be shown to be  $\vec{Z} \sim \mathcal{N}(\vec{\mu}_X + \vec{\mu}_Y, \Sigma_X + \Sigma_Y)$ . However, it should be noted that this is *not* true for dependent Gaussians. Similarly, the affine transformation for multivariate Gaussians,  $\vec{Y} = H\vec{X} + \vec{b}$ , results in a distribution  $\vec{Y} \sim \mathcal{N}(\vec{b} + H\vec{\mu}, H\Sigma H^T)$ . Lastly, the **confidence region** for multivariate Gaussian which consists of all vectors,  $\vec{x}$ , which satisfy

$$D_M(\vec{x}, X) = (\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}) \leq F_{\chi_n^2}^{-1}(p) \quad (13.164)$$

where  $F_{\chi_n^2}(p)$  is the central chi-squared CDF with  $n$  degrees of freedom at the probability  $p$  where  $n$  is the dimension of  $\vec{x}$ . This results from the fact that the sum of squared Gaussians is  $\chi^2$  distributed.

Furthermore, by using the eigenvalue decomposition of the covariance, i.e.

$$\Sigma = U\Lambda^{1/2}(U\Lambda^{1/2})^T \quad (13.165)$$

for the multivariate Gaussian  $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ . Then, by the definition of the PDF, one can show

$$\vec{X} \sim \vec{\mu} + U\mathcal{N}(0, \Lambda) \quad (13.166)$$

or

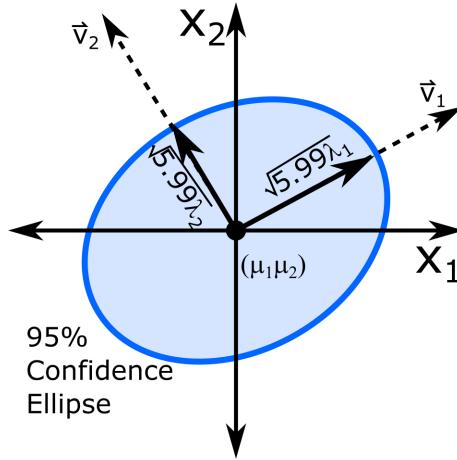
$$\vec{X} \sim \vec{\mu} + U\Lambda^{1/2}\mathcal{N}(0, I) \quad (13.167)$$

where the value of  $\Lambda^{1/2}$  define the principle standard deviations and the eigenvector columns of  $U$  define the principle axes of the confidence region.

As an example, assume that for a two-dimensional Gaussian random variable, i.e.  $\vec{X} = [X_1 \ X_2]^T \sim \mathcal{N}(\vec{\mu}, \Sigma)$  with  $\vec{\mu} = [\mu_1 \ \mu_2]^T$ , and one can decompose  $\Sigma$  into  $V = [\vec{v}_1 \ \vec{v}_2]$  where  $\Lambda = \text{diag}(\lambda_1, \lambda_2)$ . Then, consider the 95% confidence level, i.e.

$$F_{\chi^2(2)}^{-1}(0.95) = 5.99 = \left(\frac{x_1}{\sqrt{\lambda_1}}\right)^2 + \left(\frac{x_2}{\sqrt{\lambda_2}}\right)^2 \quad (13.168)$$

where the left side defines the **confidence ellipse** along the eigenvector principle axes. The length of each ellipse axis is  $2\sqrt{5.99\lambda_1}$  and  $2\sqrt{5.99\lambda_2}$ . Thus, a 95% error ellipse can be drawn as



The **Gaussian scale mixture (GSM) distribution** is defined for the random vector  $\vec{Y}$  as

$$\vec{Y} = \vec{\mu} + \sqrt{Z}\vec{X} \quad (13.169)$$

where  $\vec{X}|(Z = z) \sim \mathcal{N}(\mu, z^2\Sigma)$  and  $Z > 0$  is the univariate scale parameter and is independent from  $X$ .

$$f_Y(x) = \int_0^\infty f_X(x; \mu, z\Sigma) f_Z(z) dz \quad (13.170)$$

With this representation, one can show that the **multivariate symmetric  $\alpha$ -stable (MS $\alpha$ S) distribution** occurs for  $Z \sim S(\alpha/2, 1, 1, 0)$ , the **multivariate  $t$  distribution** occurs for  $Z \sim \Gamma(\nu/2, 2/\nu)$ , and the **multivariate  $K$  (MK) distribution** occurs for  $Z \sim \Gamma(k, \theta)$ .

The **Wishart distribution**, denoted as  $\Lambda \sim \mathcal{W}(\Phi, \nu)$ , has a joint PDF of the  $p(p+1)/2$  elements  $\Lambda_{i,j}$  for  $i \leq j$  of the form

$$f_\Lambda(\Lambda; \Phi, \nu) = \frac{1}{2^{\nu n/2} |\Phi|^{\nu/2} \Gamma_p(\nu/2)} \det(X)^{(\nu-p-1)/2} \exp\left(-\frac{1}{2} \text{Tr}(\Phi^{-1} \Lambda)\right) \quad (13.171)$$

where  $\Lambda \in \mathbb{R}^{p \times p}$  must be symmetric, positive semi-definite,  $\Phi \in \mathbb{R}^{p \times p}$  must be symmetric positive definite, and  $\nu \geq p-1$  is the degrees of freedom. If one has  $N$  samples of a random vector  $\vec{Y} \sim \mathcal{N}(\vec{0}, \Sigma)$ ,  $\vec{y}_1, \dots, \vec{y}_N$  where one can form

$$Y = [\vec{y}_1 \ \cdots \ \vec{y}_N] \quad (13.172)$$

Then,  $X = YY^T \sim \mathcal{W}(\Sigma, N)$ .

The **inverse-Wishart (IW) distribution**, denoted as  $\Sigma \sim \mathcal{IW}(\Phi, \nu)$  has a joint PDF of the form

$$f_\Sigma(\Sigma; \Phi, \nu) = \frac{|\Phi|^{\nu/2}}{2^{\nu p/2} \Gamma_p(\nu/2)} \det(\Sigma)^{-(\nu+p)/2+1} \exp\left(-\frac{1}{2} \text{Tr}(\Phi \Sigma^{-1})\right) \quad (13.173)$$

where  $\Sigma \in \mathbb{R}^{p \times p} > 0$  must be symmetric, positive semi-definite,  $\Phi \in \mathbb{R}^{p \times p} > 0$  must be symmetric positive definite,  $\nu > p-1$  is the degrees of freedom,  $\Gamma_p(a)$  is the **multivariate Gamma function**, i.e.,

$$\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma(a + (1-j)/2) \quad (13.174)$$

where if  $\Sigma \sim \mathcal{IW}(\Phi, \nu)$ , then  $\Sigma^{-1} \sim \mathcal{W}(\Phi^{-1}, \nu)$ . Notably, the mode of  $\mathcal{IW}(\Phi, \nu)$  is  $\Phi(\nu+p+1)^{-1}$  and for  $\nu > p+1$ , the mean of  $\mathcal{IW}(\Phi, \nu)$  is  $\Phi(\nu-p-1)^{-1}$ .

Then, similar to a GSM, the **Gaussian-inverse-Wishart (GIW) distribution**

$$f_{\vec{x}, \Sigma}(\vec{x}, \Sigma; \vec{\mu}, \lambda, \Phi, \nu) = f_N(\vec{x}; \vec{\mu}, \lambda^{-1} \Sigma) f_{\mathcal{IW}}(\Sigma; \Phi, \nu) \quad (13.175)$$

where  $\Sigma$  is the covariance of the conditional Gaussian distribution of  $\vec{x} \in \mathbb{R}^n$ . For univariate mean  $\mu$  and variance  $\sigma^2$ , one can define the **Gaussian-inverse- $\chi^2$  (GIW) distribution**

$$f_{x, \sigma^2}(x, \sigma^2; \mu, \lambda, \sigma^2, \nu) = f_N(x; \mu, \lambda^{-1} \sigma^2) f_{\mathcal{I}\chi^2}(\sigma^2; \tau^2, \nu) \quad (13.176)$$

These distributions are used in perception systems for modeling unknown covariances matrices with known and unknown means in Bayesian estimation.

## References

For more information, please refer to the following

- Eltoft, T., Kim, T., and Lee, T.-W. “Multivariate Scale Mixture of Gaussians Modeling,” in *Independent Component Analysis and Blind Signal Separation*, 2006, pp. 799-806

- Naaman, M., “On the Tight Constant in the Multivariate Dvoretzky–Kiefer–Wolfowitz Inequality,” in *Statistics and Probability Letters*, Vol. 173, 2021, pp. 1–8.
- Simon, D., “2.4 Multiple Random Variables,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 61–67

## 13.4 Random Processes and Dynamical Systems

When considering a collection of random variables or vectors that are indexed by some mathematical set, one has a **random process**, also known as a **stochastic process** defined as

$$\{\vec{X}(t) : t \in \mathbb{T}\} \quad (13.177)$$

where each random vector  $\vec{X}(t)$  is uniquely associated with an element,  $t$ , in the **index set**,  $\mathbb{T}$ , where  $\mathbb{T}$  is often *time*, but can also be a spatial quantity, e.g., images. Furthermore, the values that the random process can take are known as the **state space**,  $\mathbb{S}$ , and can be either a continuous or discrete state space. Another important distinction in random processes is the **cardinality classification** where a random process is called continuous-time if  $\mathbb{T}$  is an uncountable index set and discrete-time if  $\mathbb{T}$  is a countable index set. A discrete-time random process is also known as a **random sequence** and can be distinguished from a continuous-time random process by square brackets, i.e.,  $\vec{X}[k]$ . However, for this section, () will be used for both to explain common concepts.

A **sample** of a random process,  $\vec{X}$ , is a *single* outcome at  $t$ . For time-indexed processes, a sample is also known as a **time sample**. In contrast, the outcome of the combined samples through the index set  $\mathbb{T}$  is the **realization** of the random process, also known as the **sample function**. For time-indexed processes, a realization is also known as the **sample path** or **sample trajectory**. When discussing random processes, the difference between two samples with the random process is called the **increment** whose probability model describes how a random process can change over a certain time period. For any times  $t_1 \leq t_2$ , the increment is represented by the difference in the indexed random vectors as  $\vec{X}(t_2) - \vec{X}(t_1)$ .

To describe the probabilities and relationships between elements of a random process from indices  $t \in \mathbb{T}$ , for a discrete-time random process, one can define the **joint PMF** now indexed by  $t$  as

$$p_{\vec{X}}(\vec{x}, t) = \mathbb{P}(\vec{X}(t) = \vec{x}) \quad (13.178)$$

and for a continuous-time random processes, one can define the **joint CDF** now indexed by  $t$  as

$$F_{\vec{X}}(\vec{x}, t) = \mathbb{P}(\vec{X}(t) \leq \vec{x}) \quad \forall t \in \mathbb{T} \quad (13.179)$$

as well as the **joint PDF** now indexed by  $t$  as

$$f_{\vec{X}}(\vec{x}, t) = \frac{dF_{\vec{X}}(\vec{x}, t)}{d\vec{x}} \quad (13.180)$$

Thus, one can see that the dynamic or spatial nature of the random variable or vector can now be modeled explicitly in the probability functions. The nature of this dynamic/spatial dependence allows for different properties of random processes to be modeled and analyzed beyond random vector theory.

## Statistics of Random Processes

To characterize the statistics of a random process with respect to the index  $t \in \mathbb{T}$ , one can define the **expectation function** which also defines the **mean function**,  $\mu_{\vec{X}}(t)$ , as

$$\mathbb{E} [\vec{X}(t)] = \mu_{\vec{X}}(t) = \int_{-\infty}^{\infty} \vec{x} f_{\vec{X}}(\vec{x}, t) d\vec{x} \quad (13.181)$$

The **auto-correlation function** between any two samples is defined as

$$R_{\vec{X}\vec{X}}(t_1, t_2) = \mathbb{E} [\vec{X}(t_1) \vec{X}^T(t_2)] \quad (13.182)$$

and it should be noted that for any  $t$ ,  $t_1$ , and  $t_2$

$$R_{XX}(t, t) = \mathbb{E} [X(t)^2] \geq 0 \quad (13.183)$$

and

$$|R_{XX}(t_1, t_2)| \leq \sqrt{R_{XX}(t_1, t_1) R_{XX}(t_2, t_2)} \quad (13.184)$$

The **auto-covariance function** between any two samples is defined as

$$C_{\vec{X}\vec{X}}(t_1, t_2) = \mathbb{E} \left[ (\vec{X}(t_1) - \mu_{\vec{X}}(t_1)) (\vec{X}(t_2) - \mu_{\vec{X}}(t_2))^T \right] \quad (13.185)$$

and it can be shown that

$$C_{\vec{X}\vec{X}}(t_1, t_2) = R_{\vec{X}\vec{X}}(t_1, t_2) - \mu_{\vec{X}}(t_1) \mu_{\vec{X}}^T(t_2) \quad (13.186)$$

and a random process,  $\vec{X}(t)$ , is **homoscedastic** if  $C_{\vec{X}\vec{X}}(t_1, t_2) = C_{\vec{X}\vec{X}}$  is constant and finite for all  $t_1, t_2 \in \mathbb{T}$ .

For any two random processes,  $\vec{X}(t)$  and  $\vec{Y}(t)$ , the **cross-correlation function** between two samples is defined as

$$R_{\vec{X}\vec{Y}}(t_1, t_2) = \mathbb{E} [\vec{X}(t_1) \vec{Y}(t_2)] \quad (13.187)$$

and the **cross-covariance function** between two samples is represented by

$$C_{\vec{X}\vec{Y}}(t_1, t_2) = \mathbb{E} \left[ (\vec{X}(t_1) - \mu_{\vec{X}}(t_1)) (\vec{Y}(t_2) - \mu_{\vec{Y}}(t_2)) \right] \quad (13.188)$$

or

$$C_{\vec{X}\vec{Y}}(t_1, t_2) = R_{\vec{X}\vec{Y}}(t_1, t_2) - \mu_{\vec{X}}(t_1) \mu_{\vec{Y}}^T(t_2) \quad (13.189)$$

## Properties of Random Processes

A random process,  $\vec{X}(t)$ , is **independent** if  $\forall N \in \mathbb{N}$  and every set of indices,  $t_1, \dots, t_N \in T$

$$p_{\vec{X}}(\vec{x}, t, \dots, t_N) = p_{\vec{X}}(\vec{x}, t_1) \cdots p_{\vec{X}}(\vec{x}, t_N) \quad \forall \vec{x} \in \mathbb{R}^n \quad (13.190)$$

for a discrete-time process or

$$F_{\vec{X}}(x, t_1, \dots, t_N) = F_{\vec{X}}(x, t_1) \cdots F_{\vec{X}}(x, t_N) \quad \forall \vec{x} \in \mathbb{R}^n \quad (13.191)$$

for a continuous-time process. A random process,  $\vec{X}(t)$ , has **independent increments** if and only if for every  $N \in \mathbb{N}$  and any choice  $t_0, t_1, t_2, \dots, t_N \in \mathbb{T}$  with  $t_0 < t_1 < t_2 < \dots < t_N$ , the random vectors  $\vec{X}(t_1) - \vec{X}(t_0)$ ,  $\vec{X}(t_2) - \vec{X}(t_1)$ , ...,  $\vec{X}(t_N) - \vec{X}(t_{N-1})$  are independent.

A random process,  $\vec{X}(t)$ , is **strict-sense stationary (SSS)** if each random vector indexed in the process is identically distributed  $\forall N \in \mathbb{N} < \infty$ . Intuitively, this means that as time passes, the probability distribution for a single sample of a SSS random process remains constant which can often be difficult to prove about a process in reality. A weaker notion of stationarity is **wide-sense stationarity (WSS)**, also known as **covariance stationarity**. The random process,  $\vec{X}(t)$ , is WSS if  $\mu_{\vec{X}}(t) = \mu_{\vec{X}}$  does not depend on  $t$ , each  $\vec{X}(t)$  has finite correlation, and the covariance of the random variables  $\vec{X}(t)$  and  $\vec{X}(t + \tau)$  depends only on increment length  $\tau \forall t \in \mathbb{T}$ . Similar to independence implying uncorrelation, SSS implies WSS, but not vice versa. Thus, for WSS random processes, one can analyze the auto-covariance and auto-correlation functions, e.g.,  $C_{\vec{X}\vec{X}}(\tau)$ ,  $R_{\vec{X}\vec{X}}(\tau)$ , as time-invariant, i.e., only dependent on the difference  $\tau = t_2 - t_1$ . Notably, for a WSS process,  $\vec{X}(t)$ , one has

$$R_{\vec{X}\vec{X}}(\tau) = R_{\vec{X}\vec{X}}(-\tau) \quad (13.192)$$

and in the scalar case

$$|R_{XX}(\tau)| \leq R_{XX}(0) \quad (13.193)$$

With these definitions in mind, a WSS random process,  $\vec{X}(t)$ , is a **white noise process** if its mean function  $\mu_{\vec{X}} = \vec{0}$  and its auto-correlation function  $R_{\vec{X}\vec{X}}(\tau) = \Sigma\delta(\tau) < \infty$  where  $\delta(\tau)$  is the **Dirac delta** function. If  $\vec{X}(t)$  is also SSS, then **strong-sense white noise process**. If the discrete-time process  $\vec{X}[k]$  is heteroskedastic, i.e.,  $\Sigma$  varies with time  $k$ , then  $\vec{X}[k]$  is **weak white noise**. However, if a WSS random process is not white noise, then it is **colored noise**. Furthermore, the WSS definition allows one to compute the Fourier transform of  $R_{\vec{X}\vec{X}}(\tau)$  with respect to  $\tau$  as the **power spectral density (PSD)** which for continuous-time random processes is

$$S_{\vec{X}\vec{X}}(\omega) = \int_{-\infty}^{\infty} R_{\vec{X}\vec{X}}(\tau) e^{-j\tau\omega} d\tau \quad (13.194)$$

and for discrete-time random processes is

$$S_{\vec{X}\vec{X}}(\omega) = \sum_{\tau=-\infty}^{\infty} R_{\vec{X}\vec{X}}(\tau) e^{-j\omega\tau} \quad (13.195)$$

Recalling the definition of white noise, if  $\vec{X}(t)$  is a (weakly) white noise process, then  $S_{\vec{X}\vec{X}}(\omega)$  is constant. By the Fourier inversion theorem, for continuous-time random processes, one has

$$R_{\vec{X}\vec{X}}(\tau) = \int_{-\infty}^{\infty} S_{\vec{X}\vec{X}}(f) e^{j\omega\tau} d\omega \quad (13.196)$$

and for discrete-time random processes, one has

$$R_{\vec{X}\vec{X}}(\tau) = \frac{1}{2\pi} \sum_{\omega=-\pi}^{\pi} S_{\vec{X}\vec{X}}(f) e^{j\omega\tau} \quad (13.197)$$

It should be noted that the term “power spectral density” comes from its appearance in the **expected average power**,  $P_{XX} = \mathbb{E}[X(t)^2]$ , of a scalar WSS random process  $X(t)$  and is simply  $R_{XX}(0)$  which is the maximum

of  $R_{XX}(\tau)$ . Thus, a discrete-time white noise process has finite power, but a continuous-time white noise process has “infinite” expected average power, an impossible process in reality.

Two random processes,  $\vec{X}(t)$  and  $\vec{Y}(t)$ , with the same index set,  $t \in \mathbb{T}$ , are **mutually independent** if  $\forall N \in \mathbb{N}$  and every set of indices,  $t_1, \dots, t_N \in T$ ,  $[\vec{X}(t_1) \dots \vec{X}(t_N)]^T$  and  $[\vec{Y}(t_1) \dots \vec{Y}(t_N)]^T$  are independent. Similarly, two random processes,  $\vec{X}(t)$  and  $\vec{Y}(t)$ , with the same index set,  $t \in \mathbb{T}$ , are **uncorrelated** if

$$C_{\vec{X}\vec{Y}}(t_1, t_2) = \mathbb{E}[(\vec{X}(t_1) - \mu_{\vec{X}}(t_1))(\vec{Y}(t_2) - \mu_{\vec{Y}}(t_2))^T] = 0 \quad \forall t_1, t_2 \in \mathbb{T} \quad (13.198)$$

In addition, two random processes,  $\vec{X}(t)$  and  $\vec{Y}(t)$ , with the same index set,  $t \in \mathbb{T}$ , are **orthogonal** if

$$R_{\vec{X}\vec{Y}}(t_1, t_2) = \mathbb{E}[\vec{X}(t_1)\vec{Y}(t_2)^T] = 0 \quad \forall t_1, t_2 \in \mathbb{T} \quad (13.199)$$

$\vec{X}(t)$  and  $\vec{Y}(t)$  are **joint wide-sense stationary (WSS)** if they have the same index set,  $\mathbb{T}$ ,  $\mu_{\vec{X}}(t) = \mu_{\vec{X}}$  and  $\mu_{\vec{Y}}(t) = \mu_{\vec{Y}}$  do not depend on  $t$ , each  $\vec{X}(t)$  and  $\vec{Y}(t)$  have finite correlation, and the covariance of the random vectors  $\vec{X}(t)$  and  $\vec{Y}(t + \tau)$  depends only on increment length  $\tau \forall t \in \mathbb{T}$ .

## Fundamental Random Processes for Perception

The **Markov property** for random processes exists when the current sample of a random process depends only on the immediately previous sample, i.e., the sample at any time is conditionally independent of past values, and future behavior conditionally depends only on the present. Thus, given an initial realization, one can solve for the *probabilities* of achieving any other state at any point in the future. A random process with the Markov property is called a **Markov process**, also known as a **memoryless processes**. An example of a Markov process is a random process with independent increments, but this is not necessary. A Markov process with a countable state space  $S$ , is known as a **Markov chain**. A continuous-time Markov chain is also known as a **Markov jump process**. In Markov processes, the sample at any time is called the **state** and the change in the current state to a future state is called a **state transition**.

For discrete-time Markov chains, the countable number of states,  $\vec{x}_i \forall i \in \mathbb{N}$  have associated **one-step state transition probabilities** of changing from one state to any other possible future state in the state space  $\mathbb{S}$ , i.e.,

$$P_{i,j}[k] = \mathbb{P}(\vec{X}[k+1] = \vec{x}_j | \vec{X}[k] = \vec{x}_i) \quad (13.200)$$

where

$$\sum_j P_{i,j}[k] = 1 \quad (13.201)$$

The (possibly infinite-dimensional) **state transition matrix**,  $P[k]$ , has the element  $P_{i,j}[k]$  at the  $i^{\text{th}}$  row and the  $j^{\text{th}}$  column.

If  $P[k] = P \forall k$ , then the Markov chain is **time-homogeneous** and the  **$m$ -step state transition probabilities** defined as

$$P_{i,j}^{(m)} = \mathbb{P}(\vec{X}[k+m] = \vec{x}_j | \vec{X}[k] = \vec{x}_i) \quad (13.202)$$

are stationary and satisfy the **Chapman-Kolmogorov equation (CKE)**

$$P_{i,j}^{(n+m)} = \sum_k P_{i,k}^{(n)} P_{k,j}^{(m)} \quad (13.203)$$

or for the state matrix,  $P^{(n+m)}$  containing all  $P_{i,j}^{(n+m)}$ , one has

$$P^{n+m} = P^n P^m \quad (13.204)$$

Consider the initial condition  $v_j = \mathbb{P}(\vec{X}[0] = \vec{x}_j)$ .

Then, by the LTP, one has

$$\rho_j^{(n)} = \sum_i \mathbb{P}(\vec{X}[k+n] = \vec{x}_j | \vec{X}[k] = \vec{x}_i) \mathbb{P}(\vec{X}[0] = \vec{x}_i) = \sum_i P_{i,j}^{(n)} v_i \quad (13.205)$$

Then, by defining  $\vec{\rho}^{(n)} = [\rho_1^{(n)}, \rho_2^{(n)}, \dots]^T$  and  $\vec{v} = [v_1, v_2, \dots]^T$ , one has

$$(\vec{\rho}^{(n)})^T = \vec{v}^T P^n \quad (13.206)$$

where if there exists a PMF  $\vec{\pi}$  such that

$$(\vec{\rho}^{(n)})^T = \vec{\pi}^T = \vec{\pi}^T P \quad (13.207)$$

Then,  $\vec{\pi}$  is a **stationary distribution** and notably defines a left eigenvector of  $P$  with normalization  $\sum_i \vec{\pi} = 1$ .

For Markov jump processes, the countable number of states,  $\vec{x}_i \forall i \in \mathbb{N}$  have associated **transition probabilities**

$$P_{i,j}(t, s) = \mathbb{P}(\vec{X}(t) = \vec{x}_j | \vec{X}(s) = \vec{x}_i) \quad (13.208)$$

where for time-homogeneous Markov jump process

$$P_{i,j}(t) = \mathbb{P}(\vec{X}(t) = \vec{x}_j | \vec{X}(0) = \vec{x}_i) \quad (13.209)$$

which provides a **Chapman-Kolmogorov equation (CKE)** as

$$P_{i,j}(t+s) = \sum_k P_{i,k}(t) P_{k,j}(s) \quad (13.210)$$

The **transition rate**

$$F_{i,j} = \lim_{\Delta t \downarrow 0} \frac{P_{i,j}(\Delta t)}{\Delta t} \quad (13.211)$$

The **sojourn time**, also known as the **holding time**, is the time spent in one state before jumping to the next state, given as

$$\exp\left(-\lim_{\Delta t \downarrow 0} \frac{P_{i,i}(\Delta t) - 1}{\Delta t}\right) = \exp(-F_{i,i}) \quad (13.212)$$

A Markov jump process  $X(t)$  with transition rate  $F_{i,i+1} = \lambda$  is a **Poisson process**, also known as a **counting process** if it has the following properties:

- 1  $X(0) = 0$ ;
- 2  $X(t)$  continuous in  $t \geq 0$ ;
- 3  $X(t)$  has independent increments; and

4  $X(t)$  has Poisson increments, i.e.,

$$X(t + \tau) - X(t) \sim \text{Pois}(\lambda\tau) \quad (13.213)$$

Sojourn time for Poisson process is  $e^\lambda$ .

Another important type of continuous- or discrete-time random process is the **Gaussian process**,  $\vec{X}(t)$ , which for every finite set of indices  $t_1, \dots, t_N \in \mathbb{T}$ , the finite vector  $[\vec{X}(t_1), \dots, \vec{X}(t_N)]^T$  is distributed as a multivariate Gaussian random vector. This also implies that every linear combination of  $\vec{X}(t_i)$  is a multivariate Gaussian. An important property of a WSS Gaussian process is that it is also guaranteed to be SSS since the covariance only depends on separation, not individual time values. Furthermore, WSS Gaussian processes are completely defined by their auto-covariance function.  $X(t)$  is a **white Gaussian noise (WGN) process** if it is both a Gaussian and a white noise process.  $\vec{X}(t)$  is a **Gauss-Markov process** if it is both a Gaussian and a Markov process.

The non-stationary Gauss-Markov process,  $X(t)$ , is a continuous-time **Wiener process**, also known as the **Brownian motion**, if it has the following properties:

- 1  $X(0) = 0$ ;
- 2  $X(t)$  continuous in  $t \geq 0$ ;
- 3  $X(t)$  has independent increments; and
- 4  $X(t)$  has Gaussian increments, i.e.,

$$X(t + \tau) - X(t) \sim \mathcal{N}(0, \tau\sigma^2) \quad (13.214)$$

If  $\sigma^2 = 1$ , then it is a **standard Wiener process** and is the integral of a scalar WGN process.

The only stationary Gauss-Markov process,  $X(t)$ , is a continuous-time **Ornstein-Uhlenbeck process** with an auto-covariance function as

$$C_{XX}(\tau) = \frac{\sigma^2}{2\theta} \exp(-\theta|\tau|) \quad (13.215)$$

where  $\sigma > 0$  is the standard deviation of the noise fluctuations and  $\theta > 0$  is the drift which captures how much two points  $X(t_1)$  and  $X(t_2)$  influence each other. The Ornstein-Uhlenbeck process can be defined via a stochastic differential equation as

$$dX(t) = -\theta X(t)dt + \sigma dW(t) \quad (13.216)$$

where  $W(t)$  denotes a standard Wiener process. An Ornstein-Uhlenbeck process is a mean-reverting process as  $\theta$  is constrained to be positive and connects the idea of stationarity of random processes to stability for stochastic differential equations. If  $\theta = 0$ , the equation would represent a Wiener process with variance  $\sigma^2$ .

$X[k]$  is a discrete-time  $n^{\text{th}}$ -order **auto-regressive (AR) process**, denoted  $AR(n)$ , if

$$X[k] + a_1X[k - 1] + \dots + a_nX[k - n] = \epsilon[k] \quad (13.217)$$

where  $\epsilon[k]$  is a weakly white noise process, i.e., an uncorrelated sequence of zero-mean random variables with common variance  $\sigma^2 = \mathbb{E}[\epsilon[k]]$ . Notably,  $AR(1)$  is a Markov process and can be considered the discrete-time analogue of the Ornstein-Uhlenbeck process if  $|a_1| < 1$  and one has white Gaussian noise. However, one can extend the Markov property to an arbitrary number of past states using the  **$n^{\text{th}}$ -order Markov property** for random processes exists when the current sample of a random process depends only on the previous  $n$  samples or by the  $n$  derivatives of the state. In this context, a first-order Markov process is the traditional Markov process, but multiple time steps allow for  $AR(n)$  processes to be  $n^{\text{th}}$ -order Markov processes.

Similarly,  $X[k]$  is a discrete-time  **$p^{\text{th}}$ -order moving-average (MA) process**, denoted  $MA(p)$ , if

$$X[k] = \mathbb{E}[X[k]] + \epsilon[k] + \dots + b_p \epsilon[k-p] \quad (13.218)$$

where  $\epsilon[k], \dots, \epsilon[k-p]$  are weakly white noise process, i.e., uncorrelated sequences of zero-mean random variables with common variance  $\sigma_0^2, \dots, \sigma_p^2$ .

$X[k]$  is a discrete-time  **$(n^{\text{th}}, p^{\text{th}})$ -order auto-regressive moving-average (ARMA) process**, denoted  $ARMA(n, p)$ , if

$$X[k] + a_1 X[k-1] + \dots + a_n X[k-n] = \epsilon[k] + \dots + b_p \epsilon[k-p] \quad (13.219)$$

where  $\epsilon[k], \dots, \epsilon[k-p]$  are weakly white noise process, i.e., uncorrelated sequences of zero-mean random variables with common variance  $\sigma_0^2, \dots, \sigma_p^2$ . Notably, an ARMA process can be interpreted as an infinite impulse response filter applied to white noise.

Similar to deterministic differential and difference equations, ARMA random processes can be generalized to vectorized continuous-space Markov processes, e.g., **vector auto-regressive moving-average (VARMA) process**, where the order  $n$  is the dimension of the state vector that may be modeled as discrete- or continuous-time and the dimension  $p$  of the process noise appears in the input or output matrices. This generalization results in the stochastic state-space model presented in the next section.

## Random Dynamical Systems

A dynamical system is a **random dynamical system** if from a given initial state, the same inputs may produce different outputs. Random dynamical systems are also sometimes referred to as **stochastic dynamical systems** as “stochastic” refers to the modeling approach and “random” refers to the phenomena, but these two terms are often used interchangeably. For random dynamical systems it is typical to use a **stochastic state-space model** which can be defined for continuous-time as

$$\begin{aligned} \dot{\vec{x}}(t) &= f(\vec{x}(t), \vec{u}(t), \vec{w}(t), t) \\ \vec{y}(t) &= h(\vec{x}(t), \vec{v}(t), t) \end{aligned} \quad (13.220)$$

for discrete-time as

$$\begin{aligned} \vec{x}[k+1] &= f(\vec{x}[k], \vec{u}[k], \vec{w}[k], k) \\ \vec{y}[k] &= h(\vec{x}[k], \vec{v}[k], k) \end{aligned} \quad (13.221)$$

and for hybrid-time as

$$\begin{aligned} \dot{\vec{x}}(t) &= f(\vec{x}(t), \vec{u}(t), \vec{w}(t), t) \\ \vec{y}[k] &= h(\vec{x}[k], \vec{v}[k], k) \end{aligned} \quad (13.222)$$

where  $t$  is the time,  $k$  is the time step,  $\Delta t = t/k$  is the time interval, and  $\vec{w}$  and  $\vec{v}$  are the **process noise** and the **measurement noise**, respectively. In these models, the first stochastic differential or difference function is called the **process equation**, also known as the **state equation**, as it models the state of the random process and the second algebraic function is called the **measurement equation**, also known as the **observation equation**, as it models a random vector transformation of the process state to the observed measurement.

Notably, one can also define a **linear stochastic state-space model** for continuous-time as

$$\begin{aligned}\dot{\vec{x}}(t) &= A(t)\vec{x}(t) + B(t)\vec{u}(t) + L(t)\vec{w}(t) \\ \vec{y}(t) &= C(t)\vec{x}(t) + M(t)\vec{v}(t)\end{aligned}\quad (13.223)$$

for discrete-time as

$$\begin{aligned}\vec{x}[k+1] &= F[k]\vec{x}[k] + G[k]\vec{u}[k] + L[k]\vec{w}[k] \\ \vec{y}[k] &= H[k]\vec{x}[k] + M[k]\vec{v}[k]\end{aligned}\quad (13.224)$$

and for hybrid-time as

$$\begin{aligned}\dot{\vec{x}}(t) &= A(t)\vec{x}(t) + B(t)\vec{u}(t) + L(t)\vec{w}(t) \\ \vec{y}[k] &= H[k]\vec{x}[k] + M[k]\vec{v}[k]\end{aligned}\quad (13.225)$$

where the  $L$  is the process noise matrix and  $M$  is the measurement noise matrix.

A **continuous-time stochastic hybrid-state-space model** is defined as

$$\begin{aligned}\dot{\vec{x}}(t) &= f(\vec{x}(t), \mathcal{S}(t), \vec{w}(t), t) \\ \vec{y}(t) &= h(\vec{x}(t), \mathcal{S}(t), \vec{v}(t), t)\end{aligned}\quad (13.226)$$

A **discrete-time stochastic hybrid-state-space model** is defined as

$$\begin{aligned}\vec{x}[k+1] &= f_k(\vec{x}[k], \mathcal{S}[k], \vec{w}[k], k) \\ \vec{y}[k] &= h_k(\vec{x}[k], \mathcal{S}[k], \vec{v}[k], k)\end{aligned}\quad (13.227)$$

A **hybrid-time stochastic hybrid-state-space model** is defined as

$$\begin{aligned}\dot{\vec{x}}(t) &= f(\vec{x}(t), \mathcal{S}(t), \vec{w}(t), t) \\ \vec{y}[k] &= h_k(\vec{x}[k], \mathcal{S}[k], \vec{v}[k], k)\end{aligned}\quad (13.228)$$

where  $\vec{x}$  is the **base state** which varies continuously,  $\mathcal{S} \in \mathbb{S}$  is the **modal state** which is a jump process, i.e., it may either jump discretely or stay constant,  $\vec{y}$  is the measurement,  $\vec{w}$  is the process noise, and  $\vec{v}$  is the measurement noise. Here, the whole state,  $\vec{\xi} = (\vec{x}, \mathcal{S})$ , is a **hybrid process**, and is a **Markov jump system** if  $\mathcal{S}$  is a Markov chain/jump process. Often,  $\mathcal{S}$  is assumed to be a time-homogeneous Markov process.

A **jump-linear system** for continuous-time is

$$\begin{aligned}\dot{\vec{x}}(t) &= A(\mathcal{S}(t))\vec{x}(t) + B(\mathcal{S}(t))\vec{u}(t) + L(\mathcal{S}(t))\vec{w}(t) \\ \vec{y}(t) &= C(\mathcal{S}(t))\vec{x}(t) + M(\mathcal{S}(t))\vec{v}(t)\end{aligned}\quad (13.229)$$

and for discrete-time is

$$\begin{aligned}\vec{x}[k+1] &= F(\mathcal{S}[k])\vec{x}[k] + G(\mathcal{S}[k])\vec{u}[k] + L(\mathcal{S}[k])\vec{w}[k] \\ \vec{y}[k] &= H(\mathcal{S}[k])\vec{x}[k] + M(\mathcal{S}[k])\vec{v}[k]\end{aligned}\quad (13.230)$$

where if  $S$  is a Markov jump process, this is a **Markov jump-linear system (MJLS)**. This system is generally nonlinear, but may be considered a type of linear, parameter-varying (LPV) system where  $S$  is the parameter.

Thus, stochastic state-space models are functions of the random processes/sequences and require an initial prior *distribution* for  $\vec{x}(0)$  or  $\vec{x}[0]$  to solve. It is typical to assume in discrete-time  $\vec{w}[k]$  and  $\vec{v}[k]$  are white noise processes, i.e., they are zero-mean with no auto-correlation for different  $k$ . In this case, the process equation of the discrete-time linear stochastic state-space model is also known as a **vector auto-regressive (VAR) process**. Also, it is typical to assume that for continuous-time,  $\vec{w}(t)$  and  $\vec{v}(t)$  having independent “increments” for infinitesimal  $d\vec{w}(t)$  or  $d\vec{v}(t)$ , i.e., are  $n^{\text{th}}$ -order Markov processes. Under these assumptions of the  $n^{\text{th}}$ -order Markov property for  $\vec{X}$  where one does not observe the state  $\vec{X}$  directly, but only “knows”  $\vec{u}$  and  $\vec{y}$ , these stochastic state-space systems are types of **Hidden Markov Models (HMM)**. Lastly, one often assumes that  $x(0)$ ,  $\vec{w}$ , and  $\vec{v}$  are independent of each other as well, or at least uncorrelated. For the simplest model of the stochastic noise, assume that  $\vec{w}[k]$ ,  $\vec{v}[k]$ ,  $d\vec{w}(t)$ , and  $d\vec{v}(t)$  are vectorized additive white Gaussian noises (AWGN), i.e.,  $\vec{w}(t)$  and  $\vec{v}(t)$  are vectorized additive Wiener processes as integrals of the white Gaussian noise process. This type of model is often referred to as “natural” noise because it arises in several cases, e.g., thermal vibrations of atoms in conductors, black-body radiation, and deep space signals.

### Discretized Stochastic State-Space Model

Consider a discrete-time process model with identity state transition matrix,  $F_k = I$ , as

$$\vec{x}_k = \vec{x}_{k-1} + \vec{w}_{k-1} \quad (13.231)$$

with  $\vec{W}_k$  as a discrete-time white noise process, i.e., zero-mean with covariance  $Q_{d-t}$ , and initial condition  $\vec{x}_0 = 0$  which has the solution

$$\vec{x}_k = \vec{w}_0 + \cdots + \vec{w}_k \quad (13.232)$$

known as a **discrete-time random walk**. If  $\vec{W}_k \sim \mathcal{N}(0, Q_{d-t})$ , then this is a **discrete-time Gaussian random walk**.

A discrete-time random walk has a mean given by

$$\mathbb{E}[\vec{x}_k] = \mathbb{E}[\vec{w}_0 + \cdots + \vec{w}_k] = 0 \quad (13.233)$$

and a covariance given by

$$\mathbb{E}[\vec{x}_k \vec{x}_k^T] = \mathbb{E}[(\vec{w}_0 + \cdots + \vec{w}_k)(\vec{w}_0 + \cdots + \vec{w}_k)^T] \quad (13.234)$$

$$\mathbb{E}[\vec{x}_k \vec{x}_k^T] = \mathbb{E}[\vec{w}_0 \vec{w}_0^T] + \cdots + \mathbb{E}[\vec{w}_k \vec{w}_k^T] \quad (13.235)$$

$$\mathbb{E}[\vec{x}_k \vec{x}_k^T] = kQ_{d-t} \quad (13.236)$$

Next, consider the continuous-time equivalence with  $t = k\Delta t$  which provides a covariance

$$\mathbb{E}[\vec{x}(t) \vec{x}^T(t)] = \mathbb{E}[\vec{x}_k \vec{x}_k^T] = kQ_{d-t} \quad (13.237)$$

Thus, one can consider  $\vec{X}(t)$  as the analogous continuous-time random walk process represented as the stochastic differential equation

$$\dot{\vec{x}}(t) = \vec{w}(t) \quad (13.238)$$

with  $\vec{w}(t)$  as continuous-time white noise with a mean function given by

$$\mathbb{E}[\vec{w}(t)] = 0 \quad (13.239)$$

and an auto-covariance function given by

$$\mathbb{E}[\vec{w}(t)\vec{w}^T(\tau)] = \frac{Q_{d-t}}{\Delta t}\delta(t-\tau) = Q_{c-t}\delta(t-\tau) \quad (13.240)$$

where  $\delta(t-\tau)$  is the Dirac delta function,  $Q_{c-t}$  is the continuous-time process noise covariance, and  $Q_{d-t} = Q_{c-t}\Delta t$  is the **discretized process noise covariance**. Thus,  $\vec{W}(t)$  is infinitely correlated with itself at  $t = \tau$  as defined previously. Note that if  $\vec{X}_k$  is a Gaussian random walk, then  $\vec{X}(t)$  is a Wiener process of its limit as the sampling interval,  $\Delta t$ , approaches zero.

One can show this is the continuous-time equivalent of the discrete-time random walk by assessing the covariance of  $\vec{X}(t)$ , i.e.,

$$\mathbb{E}[\vec{x}(t)\vec{x}^T(t)] = \mathbb{E}\left[\int_0^t \dot{\vec{x}}(\zeta)d\zeta \int_0^t \dot{\vec{x}}^T(\tau)d\tau\right] \quad (13.241)$$

$$\mathbb{E}[\vec{x}(t)\vec{x}^T(t)] = \int_0^t \int_0^t \mathbb{E}[\vec{w}(\zeta)\vec{w}^T(\tau)]d\zeta d\tau \quad (13.242)$$

$$\mathbb{E}[\vec{x}(t)\vec{x}^T(t)] = \int_0^t \int_0^t \frac{Q_{d-t}}{\Delta t}\delta(\zeta-\tau)d\zeta d\tau \quad (13.243)$$

$$\mathbb{E}[\vec{x}(t)\vec{x}^T(t)] = \int_0^t \frac{Q_{d-t}}{\Delta t}d\tau \quad (13.244)$$

$$\mathbb{E}[\vec{x}(t)\vec{x}^T(t)] = \frac{Q_{d-t}t}{\Delta t} \quad (13.245)$$

$$\mathbb{E}[\vec{x}(t)\vec{x}^T(t)] = kQ_{d-t} \quad (13.246)$$

Next, consider a discrete-time hidden constant-state model with an identity state transition matrix  $F_k = I$ , no process noise with  $Q_k = 0$ , and an identity measurement matrix,  $H_k = I$  as

$$\begin{aligned} \vec{x}_k &= \vec{x}_{k-1} \\ \vec{y}_k &= \vec{x}_k + \vec{v}_k \end{aligned} \quad (13.247)$$

with  $\vec{V}_k$  as a discrete-time white noise process, i.e., zero-mean with covariance  $R_{d-t}$ . For the Kalman filter, the one-step posterior state covariance, one has

$$\begin{aligned} K_k &= P_{k-1|k-1} (P_{k-1|k-1} + R_{d-t})^{-1} \\ P_{k|k} &= (I - K_k) P_{k-1|k-1} \end{aligned} \quad (13.248)$$

Combining, one has

$$P_{k|k} = (P_{k-1|k-1} + R_{d-t} - P_{k-1|k-1}) (P_{k-1|k-1} + R_{d-t})^{-1} P_{k-1|k-1} \quad (13.249)$$

$$P_{k|k} = R (P_{k-1|k-1} + R_{d-t})^{-1} P_{k-1|k-1} \quad (13.250)$$

or, beginning from  $P_0$  at  $k = 0$ , one has

$$P_{k|k} = R(kP_0 + R_{d-t})^{-1} P_0 \quad (13.251)$$

$$\lim_{P_0 \rightarrow \infty} P_{k|k} = \frac{R_{d-t}}{k} = \frac{R_{d-t}\Delta t}{t_k} \quad (13.252)$$

where the state covariance will be independent of the sample time for some constant continuous-time measurement noise covariance,

$$R_{c-t} = R_{d-t}\Delta t \quad (13.253)$$

This implies the analogous continuous-time process represented as the algebraic measurement equation

$$\vec{y}(t) = \vec{x}(t) + \vec{v}(t) \quad (13.254)$$

with  $\vec{v}(t)$  as a continuous-time white noise process with mean function given by

$$\mathbb{E}[\vec{v}(t)] = 0 \quad (13.255)$$

and an auto-covariance function given by

$$\mathbb{E}[\vec{v}(t)\vec{v}^T(\tau)] = R_{d-t}\Delta t\delta(t-\tau) = R_{c-t}\delta(t-\tau) \quad (13.256)$$

where  $\delta(t-\tau)$  is the Dirac delta function,  $R_{c-t}$  is the continuous-time process noise covariance, and  $R_{d-t} = R_{c-t}/\Delta t$  is the **discretized measurement noise covariance**.

Finally, consider a continuous-time LTI stochastic state-space model

$$\begin{aligned} \vec{x}(t) &= A\vec{x}(t) + B\vec{u}(t) + \vec{w}(t) \\ \vec{y}(t) &= C\vec{x}(t) + \vec{v}(t) \end{aligned} \quad (13.257)$$

with zero-mean process noise  $\vec{W}(t)$  and covariance,  $Q_{c-t}$ , and zero-mean measurement noise  $\vec{V}(t)$  and covariance,  $R_{c-t}$ . One can compute the **discretized LTI stochastic state-space model** with time interval  $\Delta t = t/k$  as

$$\begin{aligned} \vec{x}_k &= F\vec{x}_{k-1} + G\vec{u}_{k-1} + \vec{w}_{k-1} \\ \vec{y}_k &= H\vec{x}_k + \vec{v}_k \end{aligned} \quad (13.258)$$

with zero-mean process noise  $\vec{W}_k$  and covariance,  $Q_{d-t} = Q_{c-t}\Delta t$ , zero-mean measurement noise  $\vec{V}_k$  and covariance,  $R_{d-t} = R_{c-t}/\Delta t$ , discretized state transition matrix

$$F = \exp(A\Delta t) \quad (13.259)$$

discretized input matrix

$$G = \exp(A\Delta t) \int_0^{\Delta t} \exp(-A\tau) d\tau B \quad (13.260)$$

or if  $A^{-1}$  exists, one has

$$G = \exp(A\Delta t) (I - \exp(-A\Delta t)) A^{-1} B = (\exp(A\Delta t) - I) A^{-1} B \quad (13.261)$$

and measurement matrix

$$H = C \quad (13.262)$$

Furthermore, for  $\Delta t \ll 1$ , one has the first-order approximation

$$F \approx I + A\Delta t \quad (13.263)$$

$$G \approx B\Delta t \quad (13.264)$$

## Sensor Modeling for Perception Systems

**Perception systems** are used to achieve the **perception** of the dynamical system state and/or its environment via sensor systems and information fusion where the sensor systems may be single or multi-sensor. A **sensor** is devices that detects or measures an object, area, phenomenon, or process in the system state or the external environment and produces an output signal. Perception systems process the single or multi-sensor output signals into relevant information for planning and control which may include multi-sensor information fusion. As modern sensor systems are primarily implemented with digital computer systems, this textbook will exclusively use discrete-time models for perception systems.

One classification of sensor type is for sensors that detect or measure an object, area, phenomenon, or process in the system state directly are known as **proprioceptive** as opposed to sensors that detect or measure an object, area, phenomenon, or process in the external environment state directly and the system state indirectly are known as **exteroceptive**. A subset of exteroceptive sensing is the additional field of **remote sensing** defined as the science and techniques of obtaining information about an object, area, phenomenon, or process acquired by a sensor that is *not in contact* with the object, area, phenomenon, or process under investigation. Another classification of sensor type is for a **corrective sensor** that directly or indirectly measure the system state itself as opposed to a **predictive sensor** that directly or indirectly measures the changes in the system state.

For corrective sensor data at time step  $k$ ,  $\vec{y}[k]$ , one can typically model its probability distribution based on a sample path of  $n$  previous time steps and the current sample, i.e.,  $\vec{x}[k-n], \dots, \vec{x}[k]$ , through a conditional PDF

$$f_{\vec{Y}|\vec{X}}(\vec{y}[k]|\vec{x}[k], \dots, \vec{x}[k-n]) \quad (13.265)$$

However, if the measurement is only independent of past states, i.e.,

$$f_{\vec{Y}|\vec{X}}(\vec{y}[k]|\vec{x}[k]) \quad (13.266)$$

then one can alternatively represent this sensor as the measurement equation in a stochastic state-space model, i.e.

$$\vec{y}[k] = h_k(\vec{x}[k], \vec{v}[k]) \quad (13.267)$$

where  $\vec{v}[k]$  is the stochastic **corrective sensor error** at time step  $k$ .

For predictive sensor data at the previous time step  $k-1$ ,  $\vec{u}[k-1]$ , one can relate its “state-transition” information with the current state,  $\vec{x}[k]$ , and the *previous*  $n$  states,  $\vec{x}[k-1], \dots, \vec{x}[k-n]$ , through a conditional PDF

$$f_{\vec{X}}(\vec{x}[k]|\vec{x}[k-1], \vec{u}[k-1], \dots, \vec{x}[k-n], \vec{u}[k-n]) \quad (13.268)$$

However, if the state has a  $n_x^{\text{th}}$ -order Markov property, i.e.,

$$f_{\vec{X}}(\vec{x}[k] | \vec{x}[k-1], \vec{u}[k-1]) \quad (13.269)$$

then one can alternatively represent as the process equation in a stochastic state-space model, i.e.,

$$\vec{x}[k] = f_{k-1}(\vec{x}[k-1], \vec{u}[k-1], \vec{w}[k-1]) \quad (13.270)$$

where  $\vec{w}[k-1]$  is the stochastic **predictive sensor error** at time step  $k-1$ .

## References

For more information, please refer to the following

- Gubner, J.A., “10 Introduction to Random Processes,” in *Probability and Random Processes for Electrical and Computer Engineers*, 1st Ed., Cambridge University Press, 2006, pp. 383-442
- Gubner, J.A., “11 Advanced Concepts in Random Processes,” in *Probability and Random Processes for Electrical and Computer Engineers*, 1st Ed., Cambridge University Press, 2006, pp. 443-475
- Gubner, J.A., “12 Introduction to Markov Chains,” in *Probability and Random Processes for Electrical and Computer Engineers*, 1st Ed., Cambridge University Press, 2006, pp. 476-516
- Gubner, J.A., “15.5 ARMA processes,” in *Probability and Random Processes for Electrical and Computer Engineers*, 1st Ed., Cambridge University Press, 2006, pp. 606-608
- Sarkka, S., “4.1 Probabilistic state space models,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 51-53
- Simon, D., “2.5 Stochastic Processes,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 68-74
- Simon, D., “4.1 Discrete-time systems,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 107-111
- Simon, D., “4.2 Sampled-data systems,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 111-114
- Simon, D., “4.3 Continuous-time systems,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 114-117
- Simon, D., “8.1 Discrete-Time and Continuous-Time White Noise,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 230-232

## 13.5 Random Finite Sets

The underlying mathematical foundation for the general multi-sensor, multi-target problems one encountered in target tracking systems involves the mathematical foundation of point process theory which has been adapted to be “engineering-friendly” by Mahler during the 1990s and 2000s into random finite set theory which can be mechanized for multi-target problems using **finite set statistics (FISST)**, a term which Mahler coined. RFS extends Bayesian definitions for single-target filtering, by defining multi-target state spaces and measurement spaces, models, and measurement and Markov densities based on RFS definitions. A **random finite set (RFS)  $\mathbf{X}$**  is a finite-set-valued random variable which can be described by a PMF and a family of joint PDFs. The PMF characterizes the **cardinality** of  $\mathbf{X}$ , denoted by  $|\mathbf{X}|$ , i.e. the number of elements in the set. For a given cardinality, an appropriate PDF from the family characterizes the joint distribution of the elements of  $\mathbf{X}$ . Let  $\mathcal{X} = \mathbb{R}^{n_x}$  denote the single-target state space whose elements are vectors of the form  $\vec{x} = [x_1 \dots x_{n_x}]^T$  where  $x_1, \dots, x_{n_x}$  in the set of real numbers, e.g. position, velocity. Thus, the state of a multi-target system is a finite set of state vectors, i.e.  $\mathbf{x} = \{\vec{x}_1, \dots, \vec{x}_N\}$  where  $N$  is the number of targets and  $\vec{x}_1, \dots, \vec{x}_N$  are the individual states. The multi-target state space is the class of all finite subsets of  $\mathcal{X}$ , i.e.  $\mathcal{F}(\mathcal{X})$ . This allows one to define the random state set as a random finite set,  $\mathbf{X}$ .

The **RFS density** of an RFS  $\mathbf{X} \in \mathcal{F}(\mathcal{X})$  is any non-negative function  $f_{\mathbf{X}}(\mathbf{x})$  which is defined by the **set integral** in a region  $\mathcal{S} \subseteq \mathcal{X}$  as

$$\Pr(\mathbf{x} \subseteq \mathcal{S}) = \int_{\mathcal{S}} f_{\mathbf{X}}(\mathbf{x}) \delta \mathbf{x} = \sum_{i=0}^{\infty} \frac{1}{i!} \int_{\mathcal{S}^i} f_{\mathbf{X}}(\{\vec{x}_1, \dots, \vec{x}_i\}) d(\vec{x}_1 \cdots \vec{x}_i) \quad (13.271)$$

where  $f_{\mathbf{X}}(\emptyset)$  for  $i = 0$ . It should be noted that if  $\mathcal{S} = \mathcal{X}$ , then the  $N^{\text{th}}$  term for  $N = 0, 1, \dots$  is the probability that there are  $N$  objects. Thus, the **cardinality distribution** of an RFS  $\mathbf{X} \in \mathcal{F}(\mathcal{X})$  is defined as

$$\varrho(N) = \Pr(|\mathbf{X}| = N) \quad (13.272)$$

Notably, if  $\mathbb{X}$  and  $\mathbb{Y}$  are two independent RFSs and  $\mathbb{Z} = \mathbb{X} \cup \mathbb{Y}$ , then

$$f_{\mathbb{Z}}(\mathbf{z}) = \sum_{\mathbf{x} \subseteq \mathbf{z}} f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{z} - \mathbf{x}) \quad (13.273)$$

which is analogous to the result that the PDF of the sum of two independent discrete random variables  $Z = X + Y$  is

$$p_Z(s) = \sum_x p_X(x) p_Y(z - x) \quad (13.274)$$

To incorporate object identity, each state  $\vec{x} \in \mathbb{X}$  can be augmented with a unique label  $\ell \in \mathbb{L} = \{\alpha_i : i \in \mathbb{N}\}$  where  $\mathbb{N}$  denotes the set of natural numbers and  $\alpha_i$ 's are distinct. Let  $\mathcal{L} : \mathbb{X} \times \mathbb{L} \rightarrow \mathbb{L}$  be the projection  $\mathcal{L}((\vec{x}, \ell)) = \ell$ , then a *labeled* finite set,  $\underline{\mathbf{x}}$  of  $\mathbb{X} \times \mathbb{L}$ , has distinct labels if and only if  $\underline{\mathbf{x}}$  and its labels  $\mathcal{L}(\underline{\mathbf{x}}) = \{\mathcal{L}(\vec{x}) : \vec{x} \in \underline{\mathbf{x}}\}$  have the same cardinality, i.e.

$$\Delta(\underline{\mathbf{x}}) = \delta_{|\underline{\mathbf{x}}|}(|\mathcal{L}(\underline{\mathbf{x}})|) = 1 \quad (13.275)$$

where  $\Delta(\underline{\mathbf{x}})$  is the **distinct label indicator** as only distinct sets of labels will provide the correct cardinality for the  $N$  elements of  $\underline{\mathbf{x}}$ .

Furthermore, the **inclusion function**, a generalization of the indicator function, can be defined as

$$\mathbf{1}_{\mathbf{y}}(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \subseteq \mathbf{y} \\ 0, & \text{otherwise} \end{cases} \quad (13.276)$$

and one can define the mixed integral for a labeled single-target state as

$$\int f(\underline{\mathbf{x}}) d\underline{\mathbf{x}} = \sum_{\ell \in \mathbb{L}} \int_{\mathbb{X}} f((\vec{x}, \ell)) d\vec{x} \quad (13.277)$$

Lastly, recall that

$$h^{\mathbf{x}} = \prod_{\vec{x} \in \mathbf{x}} h(\vec{x}) \quad (13.278)$$

with  $h(\emptyset) = 1$  by convention.

A **labeled random finite set (LRFS)**,  $\underline{\mathbf{X}}$  with state space  $\mathbb{X}$  and discrete label space  $\mathbb{L}$  is an RFS on  $\mathbb{X} \times \mathbb{L}$  such that each realization has distinct labels. Notably, the unlabeled version of a LRFS is obtained by simply discarding the labels and the cardinality distribution of an LRFS is the same as its unlabeled version. Furthermore, the set integral for the function  $f : \mathcal{F}(\mathbb{X} \times \mathbb{L}) \rightarrow \mathbb{R}$  is given by

$$\int f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}) \delta \underline{\mathbf{x}} = \sum_{i=0}^{\infty} \frac{1}{i!} \sum_{(\ell_1, \dots, \ell_i) \in \mathbb{L}^i} \int_{\mathbb{X}^i} f_{\mathbf{X}}(\{(\vec{x}_1, \ell_1), \dots, (\vec{x}_i, \ell_i)\}) d(\vec{x}_1 \cdots \vec{x}_i) \quad (13.279)$$

### Probability Generating Functionals of RFS

Let  $\mathbf{O}$  be an RFS of objects in some space  $\mathcal{X}$  and given a measurable  $\mathcal{S}$  of  $\mathcal{X}$ , let

$$\mathbf{1}_{\mathcal{S}}(\vec{x}) = \begin{cases} 1 & \text{if } \vec{x} \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases} \quad (13.280)$$

Then, the **belief mass function (BMF)** can be rewritten as the expected value of the indicator function of  $\mathbf{O}$  over the subset  $\mathcal{S}$ , i.e.

$$\beta_{\mathbf{O}}(\mathcal{S}) = \prod_{i=1}^N \mathbf{1}_{\mathcal{S}}(\vec{x}_i) f_{\mathbf{O}}(\mathbf{x}) \delta \mathbf{x} = \int \mathbf{1}_{\mathcal{S}}^{\mathbf{O}} f_{\mathbf{O}}(\mathbf{x}) \delta \mathbf{x} \quad (13.281)$$

Then, by defining for any finite subset,  $\mathbf{x}$  of  $\mathcal{X}$  and any real-valued function of  $\mathbf{x}$ ,

$$h^{\mathbf{x}} = \begin{cases} 1 & \text{if } \mathbf{x} = \emptyset \\ \prod_{i=1}^N h(\vec{x}_i) & \text{if } \mathbf{x} = \{\vec{x}_1, \dots, \vec{x}_N\} \text{ with } \vec{x}_1, \dots, \vec{x}_N \text{ distinct} \end{cases} \quad (13.282)$$

one can generalize the concept of the BMF by replacing  $\mathbf{1}_{\mathcal{S}}(\vec{x})$  with any  $h(\vec{x})$  such that

$$h(\vec{x}) = h_0(\vec{x}) + w_1 \delta_{\vec{w}_1}(\vec{x}) + \cdots + w_{n_x} \delta_{\vec{w}_{n_x}}(\vec{x}) \quad (13.283)$$

where  $0 \leq h_0(\vec{x}) \leq 1$  has no units of measurement,  $\delta_w(\vec{x})$  is the Dirac delta function,  $\vec{w}_1, \dots, \vec{w}_{n_x}$  are fixed distinct elements of  $\mathcal{X}$ , and  $w_1, \dots, w_{n_x}$  have the same units of measurement as  $\vec{x}$ .

Then, one can define the **probability generating functional (PGFL)** of  $\mathbf{X}$  as

$$G_{\mathbf{O}}[h] = \int h^{\mathbf{x}} f_{\mathbf{O}}(\mathbf{x}) \delta \mathbf{x} \quad (13.284)$$

and

$$\beta_{\mathbf{O}}(\mathcal{S}) = G_{\mathbf{O}}[\mathbf{1}_{\mathcal{S}}] \quad (13.285)$$

The PGFL is well-defined and finite valued because  $f_{\mathbf{O}}(\{\vec{x}_1, \dots, \vec{x}_i, \dots, \vec{x}_j, \dots, \vec{x}_N\}) = 0$  whenever  $\vec{x}_i = \vec{x}_j$  for  $i \neq j$ , so undefined products of  $\delta_u(\vec{y})^2$  do not occur. The intuitive meaning of the PGFL is as follows. Let  $X$  be a single-target state space,  $\mathbf{O} = \mathbf{x}$ , a random finite subset of  $X$ , and  $0 \leq h(\vec{x}) \leq 1$ , so that  $h(\vec{x})$  can be interpreted as the probability of detection for the surveillance volume (SV) or field-of-view (FOV) of a sensor. Then, it can be shown that  $G_{\mathbf{x}}[h]$  is the probability that  $\mathbf{x}$  is contained in the surveillance volume. Furthermore, as  $h(\vec{x})$  is a fuzzy membership function on  $\mathbb{R}^{n_x}$ ,  $G_{\mathbf{x}}[h]$  is a generalization of the BMF from hard sets  $X$  to fuzzy subsets  $h$ .

Importantly,

$$G_Z[h] = G_{\mathbf{X}}[h]G_Y[h] \quad (13.286)$$

which is analogous to the result that the PGF of the sum of two independent random variables  $Z = X + Y$  is

$$G_Z(s) = G_X(s)G_Y(s) \quad (13.287)$$

### Moments of Random Finite Sets

For an RFS  $\mathbf{X}$  with RFS density  $f_{\mathbf{X}}(\vec{x})$ , the **RFS moment density** is defined as

$$D_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X}}(\mathbf{x} \cup \mathbf{w}) \delta \mathbf{w} \quad (13.288)$$

where  $D_{\mathbf{X}}(\emptyset) = 1$  and the set integral is well-defined. An intuitive interpretation of the moment density is for any  $\mathbf{x} = \{\vec{x}_1, \dots, \vec{x}_N\}$ ,  $D_{\mathbf{X}}(\mathbf{x})$  is the probability density that  $N$  of the targets in  $\mathbf{X}$  have states  $\vec{x}_1, \dots, \vec{x}_N$ . The  $N^{\text{th}}$ -order **RFS moment density** is defined as

$$D_{\mathbf{X}}(\{\vec{x}_1, \dots, \vec{x}_N\}) \quad (13.289)$$

Furthermore, the **first-order RFS moment density** is defined as

$$\nu(\vec{x}) = D_{\mathbf{X}}(\{\vec{x}\}) = \int f_{\mathbf{X}}(\{\vec{x}\} \cup \mathbf{w}) \delta \mathbf{w} = \int \delta_{\mathbf{x}}(\vec{x}) f_{\mathbf{X}}(\vec{x}) \delta \mathbf{x} \quad (13.290)$$

which holds by setting  $h(\vec{x}) = \delta_{\mathbf{X}}(\vec{x})$ . This is also known as the **probability hypothesis density (PHD)** or **intensity function** of  $\mathbf{X}$ , so named because one can show that for any subset  $\mathcal{S} \subseteq X$

$$\hat{N}(\mathcal{S}) = \mathbb{E}[|\mathbf{X} \cap \mathcal{S}|] = \int_{\mathcal{S}} \nu(\vec{x}) d\vec{x} \quad (13.291)$$

Intuitively, the local maxima of the PHD are points in  $X$  with the highest local concentration of expected number of elements. Intuitively, one can use  $\hat{N} = \mathbb{E}[|\mathbf{X}|]$  as the estimated number of targets and the locations of the  $\hat{N}$  highest local maxima of the PHD as the estimated target states.

## Fundamental RFS Distributions

A **IID cluster RFS**,  $\mathbf{X}$ , has elements IID according to a PDF  $f_{\vec{X}}(\vec{x})$  and is completely characterized by  $\varrho$  and  $f$ . Its RFS density is defined as

$$f_{\mathbf{X}}(\{\vec{x}_1, \dots, \vec{x}_N\}) = N! \varrho(N) \prod_{i=1}^N f_{\vec{X}}(\vec{x}_i) \quad (13.292)$$

with  $f_{\mathbf{X}}(\emptyset) = \varrho(0)$ . A **Poisson RFS**, also known as a **Poisson Point Process (PPP) RFS** is a special case of an IID cluster RFS with a Poisson cardinality distribution, i.e.,

$$\varrho(N) = \Pr(|\mathbf{x}| = N) = e^{-\lambda} \frac{\lambda^N}{N!} \quad (13.293)$$

where mean value,  $\hat{N} = \lambda$ , also known as the Poisson rate. Thus, the Poisson RFS density is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \exp\left(-\int \lambda(x) dx\right) = e^{-\lambda} \lambda^N \prod_{j=1}^N f_{\vec{X}}(\vec{x}_j) \quad (13.294)$$

where, by definition of the set integral, one has

$$\begin{aligned} \int f_{\mathbf{X}}\{\vec{x}\} \delta \mathbf{x} &= \sum_{i=0}^{\infty} \frac{1}{i!} \int f_{\mathbf{X}}(\{\vec{x}_1, \dots, \vec{x}_i\}) d(\vec{x}_1 \cdots \vec{x}_i) \\ &= \sum_{i=0}^{\infty} \frac{1}{i!} \int_{S^i} e^{-\lambda} \lambda^i \prod_{j=1}^i f_{\vec{X}}(\vec{x}_j) d(\vec{x}_1 \cdots \vec{x}_i) \\ &= e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} \\ &= e^{-\lambda} e^{\lambda} = 1 \end{aligned} \quad (13.295)$$

Furthermore, by definition of the PHD  $v(\vec{x})$  as

$$v(\vec{x}) = \int f_{\mathbf{X}}(\vec{x} \cup \mathbf{w}) \delta \mathbf{w} \quad (13.296)$$

which, by definition of the set integral, one has

$$\begin{aligned} v(\vec{x}) &= \sum_{i=0}^{\infty} \frac{1}{i!} \int_{S^i} f_{\mathbf{X}}(\{\vec{x}, \vec{w}, \dots, \vec{w}_i\}) d\vec{x} \\ &= e^{-\lambda} f_{\vec{X}}(\vec{x}) \sum_{i=0}^{\infty} \lambda^{i+1} i! \int_{S^i} f_{\mathbf{X}}(\{\vec{w}, \dots, \vec{w}_i\}) d(\vec{w}_1 \cdots \vec{w}_i) \\ &= \lambda f_{\vec{X}}(\vec{x}) \end{aligned} \quad (13.297)$$

Thus, a Poisson RFS with cardinality parameter,  $\lambda$ , and single-target PDF,  $f_{\vec{X}}(\vec{x})$ , as one has

$$f_{\vec{X}}(\vec{x}) = \frac{v(\vec{x})}{\lambda} \quad (13.298)$$

$$f_{\mathbf{X}}(\mathbf{x}) = e^{-\lambda} \prod_{j=1}^N \frac{\nu(\vec{x}_j)}{\lambda} \quad (13.299)$$

and

$$\hat{N} = \lambda = \int_{\mathcal{X}} \nu(\vec{x}) d\vec{x} \quad (13.300)$$

and the PGFL is given by

$$G_{\mathbf{X}}[h] = \exp \quad (13.301)$$

**A Bernoulli RFS**,  $\mathbf{X}$  has probability  $1 - r$  of being empty, and probability  $r$  of being a singleton whose element is distributed according to a PDF  $f_{\vec{X}}(\vec{x})$ . The RFS density of the Bernoulli RFS is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} 1 - r & \mathbf{x} = \emptyset \\ rf_{\vec{X}}(\vec{x}) & \mathbf{x} = \{\vec{x}\} \end{cases} \quad (13.302)$$

Furthermore, a **multi-Bernoulli RFS** is a union of independent Bernoulli RFSs. The multi-Bernoulli RFS density is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} 1 - r & \mathbf{x} = \emptyset \\ rf_{\vec{X}}(\vec{x}) & \mathbf{x} = \{\vec{x}\} \end{cases} \quad (13.303)$$

and the PGFL is given by

$$G_{\mathbf{X}}[h] = \quad (13.304)$$

Lastly, a **multi-Bernoulli mixture RFS** is a weighted sum of independent multi-Bernoulli RFSs. The multi-Bernoulli mixture RFS density is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} 1 - r & \mathbf{x} = \emptyset \\ rf_{\vec{X}}(\vec{x}) & \mathbf{x} = \{\vec{x}\} \end{cases} \quad (13.305)$$

and the PGFL is given by

$$G_{\mathbf{X}}[h] = \quad (13.306)$$

**A labeled IID cluster RFS**,  $\underline{\mathbf{X}}$ , has an RFS density given by

$$f_{\underline{\mathbf{X}}}(\{(\vec{x}_1, \ell_1), \dots, (\vec{x}_N, \ell_N)\}) = \delta_{\{\alpha_1, \dots, \alpha_N\}}(\{\ell_1, \dots, \ell_N\}) \varrho(N) \prod_{i=1}^N \frac{\nu(\vec{x}_i)}{\langle \nu, 1 \rangle} \quad (13.307)$$

where  $\mathbb{L}(N) = \{\alpha_i \in \mathbb{L}\}_{i=1}^N$ ,  $\nu$  is the PHD of the unlabeled states, and  $\varrho(N)$  is the cardinality distribution. For a **labeled Poisson RFS**, one has

$$\varrho(N) = e^{-\lambda} \frac{\lambda^N}{N!} \quad (13.308)$$

where  $\lambda = \langle \nu, 1 \rangle$ .

The **labeled multi-Bernoulli (LMB) RFS**,  $\underline{\mathbf{X}}$ , has a finite parameter set  $(r^{(\zeta)}, f_{\vec{X}}^{(\zeta)}) : \zeta \in \Psi$  where the labels in  $\mathbb{L}$  are “assigned” to non-empty Bernoulli components, i.e. if  $(r^{(\zeta)}, f_{\vec{X}}^{(\zeta)})$  is non-empty, then the label is given by  $\alpha(\zeta)$  where  $\alpha : \Psi \rightarrow \mathbb{L}$  is a 1-1 mapping. Then, the multi-target FISST density is given by

$$f_{\underline{\mathbf{X}}}(\{(\vec{x}_1, \ell_1), \dots, (\vec{x}_N, \ell_N)\}) = \delta_N(|\{\ell_1, \dots, \ell_N\}|) \prod_{\zeta \in \Psi} (1 - r^{(\zeta)}) \prod_{j=1}^N \frac{1_{\alpha(\Psi)}(\ell_j) r^{(\alpha^{-1}(\ell_j))} f_{\vec{X}}^{(\alpha^{-1}(\ell_j))}(\vec{x}_j)}{1 - r^{(\alpha^{-1}(\ell_j))}} \quad (13.309)$$

which can be written alternatively as

$$f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}) = \Delta(\underline{\mathbf{x}}) 1_{\alpha(\Psi)}(\mathcal{L}(\underline{\mathbf{x}})) [\Phi(\vec{x}; \cdot)]^{\underline{\mathbf{x}}} \quad (13.310)$$

where

$$\Phi(\vec{x}; \zeta) = \begin{cases} 1 - r^{(\zeta)}, & \text{if } \alpha(\zeta) \notin \mathcal{L}(\underline{\mathbf{x}}) \\ r^{(\zeta)} f_{\vec{X}}^{(\zeta)}(\vec{x}) & \text{if } (\vec{x}, \alpha(\zeta)) \in \underline{\mathbf{x}} \end{cases} \quad (13.311)$$

With these two LRFS in mind, the **generalized labeled multi-Bernoulli (GLMB) RFS** can be defined as

$$f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}) = \Delta(\underline{\mathbf{x}}) \sum_{\xi \in \Xi} w^{(\xi)}(\mathcal{L}(\underline{\mathbf{x}})) [f_{\vec{X}}^{(\xi)}]^{\underline{\mathbf{x}}} \quad (13.312)$$

where  $\Xi$  is a discrete index set,  $w^{(\xi)}(L)$  satisfies

$$\sum_{L \subseteq \mathbb{L}} \sum_{\xi \in \Xi} w^{(\xi)}(L) = 1 \quad (13.313)$$

and  $f_{\vec{X}}^{(\xi)}$  satisfies

$$\int f_{\vec{X}}^{(\xi)}(\vec{x}, \ell) d\vec{x} = 1 \quad (13.314)$$

A GLMB can be interpreted as a mixture of multi-target exponentials where each term consists of a weight  $w^{(\xi)}(\mathcal{L}(\underline{\mathbf{x}}))$  that only depends on the labels of the multi-target state, and a multi-target exponential  $[f_{\vec{X}}^{(\xi)}]^{\underline{\mathbf{x}}}$  that depends on the entire multi-object state. The points of a GLMB RFS are statistically dependent.

Notably, the PHD of the unlabeled version of the GLMB RFS is

$$\nu(\vec{x}) = \sum_{\xi \in \Xi} \sum_{\ell \in \mathbb{L}} f_{\vec{X}}^{(\xi)}(\vec{x}, \ell) \sum_{L \subseteq \mathbb{L}} 1_L(\ell) w^{(\xi)}(L) \quad (13.315)$$

and the cardinality distribution of a GLMB RFS is given by

$$\varrho(N) = \sum_{L \in \mathcal{F}_N(\mathbb{L})} \sum_{\xi \in \Xi} w^{(\xi)}(L) \quad (13.316)$$

where  $\mathcal{F}_N(\mathbb{L})$  denotes the subsets of  $\mathbb{L}$  with exactly  $N$  elements.

Thus, the GLMB RFS is flexible enough to approximate any labeled RFS density with matching PHD and cardinality distribution, including the labeled Poisson RFS which sets

$$\begin{aligned} w^{(\xi)}(L) &= \delta_{\mathbb{L}(|L|)}(L) \varrho(|L|) \\ f_{\vec{X}}^{(\xi)}(\vec{x}, \ell) &= \frac{\nu(\vec{x})}{\langle \nu, 1 \rangle} \end{aligned} \quad (13.317)$$

and the LMB RFS which sets

$$\begin{aligned} w^{(\xi)}(L) &= \prod_{i \in \mathbb{L}} (1 - r^{(i)}) \prod_{\ell \in L} \frac{1_{\mathbb{L}}(\ell) r^{(\ell)}}{1 - r^{(\ell)}} \\ f_{\underline{\vec{X}}}^{(\xi)}(\vec{x}, \ell) &= f_{\underline{\vec{X}}}^{(\ell)}(\vec{x}, \ell) \end{aligned} \quad (13.318)$$

both of which have an index space  $\Xi$  with one element, thus, not requiring the  $(\xi)$  superscript.

## References

---

# Optimal Parameter Estimation and Detection Theory

## 14.1 Introduction to Optimal Parameter Estimation

**Decision theory** is the determination of the optimal decision given constraints and assumptions on a decision problem. **Parameter estimation** is one special case of a decision problem and can be stated as deciding which parameter(s) one should select and the general **parameter estimation problem** can be considered as follows. Let  $\vec{y}$  be samples of a random vector  $\vec{Y}$  which depend on some *uncertain* parameter vector,  $\vec{\beta}$ . As such, this probabilistic relationship can be described using a conditional PDF,  $f_{\vec{Y}|\vec{\beta}}(\vec{y}|\vec{\beta})$ , or as a **likelihood function**  $\mathcal{L}(\vec{\beta}|\vec{y})$  which are functionally equivalent, but view the relationships from different viewpoints. Using this information, one generally forms some mathematical law called the **parameter estimator** of  $\vec{\beta}$  based on  $\vec{y}$  and is typically denoted by  $\hat{\vec{\beta}}(\vec{y})$ . An **optimal parameter estimator** is used when one forms the mathematical law through an optimization with respect to a chosen attribute or statistic of  $f_{\vec{Y}|\vec{\beta}}(\vec{y}|\vec{\beta})$ .

In addition, a **statistical inference** problem is a special case of the decision problem and can be stated as deciding which probability distribution one should select to optimally represent observed data,  $\vec{y}$ , among some **candidate set** of candidate probability distributions. Typically this selection is based on some mathematical law which generally be either fully parametric, non-parametric, or semi-parametric. Fully parametric statistical inference assumes  $\vec{y}$  are *fully* described by statistical models, e.g., PDFs/likelihoods, involving only a finite number of unknown parameters,  $\vec{\beta}$ . Thus, the fully parametric statistical inference problem can often be formulated as a type of optimal parameter estimation problem for the PDF parameters,  $\hat{\vec{\beta}}$ .

### Maximum Likelihood Estimator

An example of an optimal parameter estimator is the **maximum likelihood estimator (MLE)** which can be stated as

$$\hat{\vec{\beta}}_{MLE} = \underset{\vec{\beta}}{\operatorname{argmax}} \mathcal{L}(\vec{\beta}|\vec{y}) \quad (14.1)$$

which is often computed using the log-likelihood as

$$\hat{\vec{\beta}}_{MLE} = \underset{\hat{\vec{\beta}}}{\operatorname{argmin}} -\ln \mathcal{L}(\vec{\beta} | \vec{y}) = \underset{\hat{\vec{\beta}}}{\operatorname{argmin}} \quad (14.2)$$

The MLE exists if the Jacobian of the log-likelihood is the zero row vector. For some probability models, one can directly compute  $\hat{\vec{\beta}}_{MLE}$  through solving the gradient equation

$$\nabla \ln \mathcal{L}(\hat{\vec{\beta}}_i | \vec{y}) = \frac{\partial \ln \mathcal{L}(\vec{\beta} | \vec{y})}{\partial \vec{\beta}} = 0 \quad (14.3)$$

where the gradient is also known as the **score**,  $\vec{s}_i(\hat{\vec{\beta}})$ , for iteration  $i$ . However, in general, no closed-form solution to the MLE problem is known and can only be found via numerical optimization methods.

The most common optimization methods for MLE are iterative procedures of the form:

$$\hat{\vec{\beta}}_{i+1} = \hat{\vec{\beta}}_i + \eta_i \vec{d}_i(\hat{\vec{\beta}}) \quad (14.4)$$

where  $\eta_i$  is the step length, i.e. **learning rate** and  $\vec{d}_i$  is the **descent direction**. For choosing these values, one can use the Gradient Descent method where  $\eta_i$  is chosen small enough for convergence and  $\vec{d}_i = \nabla \ln \mathcal{L}(\hat{\vec{\beta}}_i | \vec{y})$ . One can also use the Newton-Raphson method where  $\eta_i = 1$  and  $\vec{d}_i = -H_i^{-1}(\hat{\vec{\beta}}) \vec{s}_i(\hat{\vec{\beta}})$ . Then,  $H_i^{-1}(\hat{\vec{\beta}})$ : inverse of Hessian matrix of log-likelihood function, i.e.,

$$H(\hat{\vec{\beta}}) = -\frac{\partial^2 \ln \mathcal{L}(\vec{\beta} | \vec{y})}{\partial \vec{\beta}^2} \quad (14.5)$$

Quasi-Newton methods which attempt to approximate the Hessian matrix are the Davidson-Fletcher-Powell (DFP) formula, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, and the Fisher's Scoring which replaces the explicit Hessian with **Fisher information matrix**, i.e., the covariance of the score as

$$\mathcal{I}(\vec{\beta}) = \mathbb{E}\left[\left(\frac{\partial \ln \mathcal{L}(\vec{\beta} | \vec{y})}{\partial \vec{\beta}}\right)\left(\frac{\partial \ln \mathcal{L}(\vec{\beta} | \vec{y})}{\partial \vec{\beta}}\right)^T\right] \quad (14.6)$$

For a single observed sample,  $\vec{y}$ , the MLE has no guaranteed performance. However, the MLE has some nice properties as the number of repeated samples approaches infinity, i.e. for repeated  $\vec{y}_k$  with  $k = 1, 2, 3, \dots$ . Namely, if each sample is **independent and identically distributed (IID)**, then the combined likelihood can be written simply as a product of the conditional PDFs or likelihoods as

$$\hat{\vec{\beta}}_{MLE} = \underset{\hat{\vec{\beta}}}{\operatorname{argmin}} \prod_{k=1}^N \mathcal{L}(\vec{\beta} | \vec{y}_k) \quad (14.7)$$

or, in terms of the log-likelihood, as

$$\hat{\vec{\beta}}_{MLE} = \underset{\hat{\vec{\beta}}}{\operatorname{argmin}} - \sum_{k=1}^N \ln \mathcal{L}(\vec{\beta} | \vec{y}_k) \quad (14.8)$$

It can be shown that as  $N \rightarrow \infty$ ,  $\hat{\vec{\beta}}_{MLE}$  converges in probability to  $\vec{\beta}$ , a property known as **estimator consistency**.

## Method of Moments

An alternative to the MLE for IID samples which characterize a conditional PDF,  $f_Y(y|\vec{\beta})$ , is the **method of moments (MOM) estimator** which can be used for estimating the  $k$  unknown parameters of the distribution,  $\vec{\beta}$ .

In basic MOM, one can form the  $k$  moments of  $Y$  as

$$\begin{bmatrix} m_1 \\ \vdots \\ m_k \end{bmatrix} = \begin{bmatrix} \mathbb{E}[Y] \\ \vdots \\ \mathbb{E}[Y^k] \end{bmatrix} = \begin{bmatrix} g_1(\beta_1, \dots, \beta_k) \\ \vdots \\ g_k(\beta_1, \dots, \beta_k) \end{bmatrix} \quad (14.9)$$

Next, suppose that  $N$  samples are drawn of  $Y$ , denoted as  $y_i$  for  $i = 1, \dots, N$ , then one can form the empirical moments

$$\begin{bmatrix} \hat{m}_1 \\ \vdots \\ \hat{m}_k \end{bmatrix} = \begin{bmatrix} \frac{1}{N} \sum_{i=1}^N y_i \\ \vdots \\ \frac{1}{N} \sum_{i=1}^N y_i \end{bmatrix} \quad (14.10)$$

Finally, the basic MOM estimator solves the equations

$$\begin{bmatrix} \hat{m}_1 \\ \vdots \\ \hat{m}_k \end{bmatrix} = \begin{bmatrix} g_1(\hat{\beta}_1, \dots, \hat{\beta}_k) \\ \vdots \\ g_k(\hat{\beta}_1, \dots, \hat{\beta}_k) \end{bmatrix} \quad (14.11)$$

An alternative to the basic moments of distributions are **probability weighted moments (PWM)** defined for three parameters  $(p, r, s)$  as

$$\vec{M}(p, r, s) = \mathbb{E}[Y^p F_Y(y)^r (1 - F_Y(y))^s] \quad (14.12)$$

which can be used to yield a MOM estimator. A special case of PWMs are the **L-moments** where the  $r^{\text{th}}$  L-moment,  $\lambda_r$ , can be defined as

$$\lambda_r = r^{-1} \sum_{j=0}^{r-1} (-1)^j \frac{(r-1)!}{(r-j-1)! j!} \mathbb{E}[Y_{r-j:r}] \quad (14.13)$$

where  $Y_{k:N}$  denotes the  $k^{\text{th}}$  **order statistic**, i.e. the  $k^{\text{th}}$  smallest value, in the  $N$  IID samples. Furthermore, the empirical L-moments can be defined as

$$\hat{\lambda}_r = r^{-1} \frac{(N-r)! r!}{N!} \sum_{y_1 < \dots < y_j < \dots < y_r} (-1)^{r-j} \frac{(r-1)!}{(r-j-1)! j!} y_j \quad (14.14)$$

where one has summed over the  $r$ -element subset of the samples  $\{y_1 < \dots < y_j < \dots < y_r\}$ , thereby averaging by dividing by the binomial coefficient.

In Generalized MOM, one must be able to form a vector-valued moment conditions,  $g(\vec{\beta})$  for the moment functions,  $\vec{m} \in \mathbb{R}^k$ , of the conditional PDF, i.e.

$$\vec{m}(\vec{\beta}) = \mathbb{E}[g(\vec{y}, \vec{\beta})] = 0 \quad (14.15)$$

Next, let the empirical moments,  $\hat{m}(\vec{\beta})$ , be defined as

$$\hat{m}(\vec{\beta}) = \frac{1}{N} \sum_{i=1}^N g(\vec{y}_i, \vec{\beta}) \quad (14.16)$$

Then, the MOM estimator can be stated as

$$\hat{\vec{\beta}}_{MOM} = \operatorname{argmin}_{\hat{\vec{\beta}}} \left( \frac{1}{N} \sum_{k=1}^N g(\vec{y}_k, \vec{\theta}) \right)^T \left( \frac{1}{N} \sum_{k=1}^N g(\vec{y}_k, \vec{\beta}) \right) \quad (14.17)$$

where this norm minimization attempts to look for a  $\theta$  which makes  $\hat{m}$  as close to zero as possible which should be the case by the law of large numbers (for  $N$ ).

### Estimator Error Optimality

One important characteristic of any parameter estimator is the **estimator error**, i.e.  $\hat{\vec{\beta}} - \vec{\beta}$ , which is a random vector that can generally have *positive or negative* values. Thus, a common statistic for any estimator considers the expectation of the square of the error, also known as the **mean square error (MSE)** of the estimator, defined as

$$\text{MSE}(\hat{\vec{\beta}}) = \mathbb{E} \left[ (\hat{\vec{\beta}} - \vec{\beta})^T (\hat{\vec{\beta}} - \vec{\beta}) \right] \quad (14.18)$$

Thus, an optimal parameter estimator may be formed as

$$\hat{\vec{\beta}}_{MMSE} = \operatorname{argmin}_{\hat{\vec{\beta}}} \mathbb{E} \left[ (\hat{\vec{\beta}} - \vec{\beta})^T (\hat{\vec{\beta}} - \vec{\beta}) \right] \quad (14.19)$$

which is known as the **minimum MSE (MMSE) estimator**. For the MLE, it can be shown that as  $N \rightarrow \infty$ , the MLE becomes the MMSE.

Another statistic of an estimator is the **estimator mean**, i.e.  $\mathbb{E}[\hat{\vec{\beta}}]$ , which may or may not match the actual parameter  $\vec{\beta}$ . Thus, one often may consider the **estimator bias**,  $\vec{b}(\hat{\vec{\beta}})$  defined as

$$\vec{b}(\hat{\vec{\beta}}) = \mathbb{E}[\hat{\vec{\beta}}] - \vec{\beta} \quad (14.20)$$

Thus, if for some  $\hat{\vec{\beta}}$ ,  $\mathbb{E}[\hat{\vec{\beta}}] = \vec{\beta}$ , then  $\hat{\vec{\beta}}$  is **unbiased**. A third statistic of an estimator is the **estimator variance**, i.e.

$$\text{Var}(\hat{\vec{\beta}}) = \mathbb{E} \left[ (\hat{\vec{\beta}} - \mathbb{E}[\hat{\vec{\beta}}])^T (\hat{\vec{\beta}} - \mathbb{E}[\hat{\vec{\beta}}]) \right] \quad (14.21)$$

With these two characteristics in mind, an optimal parameter estimator may be formed as

$$\hat{\vec{\beta}}_{MVUE} = \operatorname{argmin}_{\substack{\text{unbiased } \vec{\beta}}} \mathbb{E} \left[ (\hat{\vec{\beta}} - \mathbb{E}[\hat{\vec{\beta}}])^T (\hat{\vec{\beta}} - \mathbb{E}[\hat{\vec{\beta}}]) \right] \quad (14.22)$$

which is known as the **minimum variance unbiased estimator (MVUE)**.

Next, note that the MSE of  $\hat{\vec{\beta}}$  can be written as

$$\text{MSE}(\hat{\vec{\beta}}) = \mathbb{E} \left[ \hat{\vec{\beta}}^T \hat{\vec{\beta}} - \hat{\vec{\beta}}^T \vec{\beta} - \vec{\beta}^T \hat{\vec{\beta}} - \vec{\beta}^T \vec{\beta} \right] \quad (14.23)$$

$$\text{MSE}(\hat{\beta}) = \mathbb{E}[\hat{\beta}^T \hat{\beta}] - \mathbb{E}[2\hat{\beta}^T \vec{\beta}] - \mathbb{E}[\vec{\beta}^T \vec{\beta}] \quad (14.24)$$

or

$$\text{MSE}(\hat{\beta}) = \mathbb{E}[\hat{\beta}^T \hat{\beta}] - 2\mathbb{E}[\hat{\beta}^T] \vec{\beta} - \vec{\beta}^T \vec{\beta} \quad (14.25)$$

Also, note that the square of the bias can be written as

$$\vec{b}(\hat{\beta})^2 = \mathbb{E}[\hat{\beta}] \mathbb{E}[\hat{\beta}]^T - 2\mathbb{E}[\hat{\beta}^T] \vec{\beta} - \vec{\beta}^T \vec{\beta} \quad (14.26)$$

and that the variance can be written as

$$\text{Var}(\hat{\beta}) = \mathbb{E} \left[ \hat{\beta}^T \hat{\beta} - 2\mathbb{E}[\hat{\beta}]^T \hat{\beta} - \mathbb{E}[\hat{\beta}]^T \mathbb{E}[\hat{\beta}] \right] \quad (14.27)$$

or

$$\text{Var}(\hat{\beta}) = \mathbb{E} \left[ \hat{\beta}^T \hat{\beta} \right] - 2\mathbb{E}[\hat{\beta}]^T \mathbb{E}[\hat{\beta}] - \mathbb{E}[\hat{\beta}]^T \mathbb{E}[\hat{\beta}] \quad (14.28)$$

Thus, by comparing the MSE, variance, and bias, one can show that

$$\text{MSE}(\hat{\beta}) = \text{Var}(\hat{\beta}) + \vec{b}(\hat{\beta})^2 \quad (14.29)$$

which demonstrates that the MVUE is equivalent to the MMSE among unbiased estimators.

Lastly, the **Cramer-Rao Bound (CRB)** states that the lower bound on the covariance of any *unbiased* estimator of  $\vec{\beta}$  can be shown to be the inverse of Fisher information matrix, i.e.

$$\text{Cov}(\hat{\beta}) \geq \mathcal{I}^{-1}(\vec{\beta}) \quad (14.30)$$

where an estimator which achieves the CRB is known as an **efficient estimator**. Furthermore, under certain conditions, it can also be shown that the MLE is asymptotically efficient and its estimator error converges in distribution to a normal distribution, i.e.

$$\sqrt{N}(\hat{\beta}) - \vec{\beta} \rightarrow \mathcal{N}(0, \mathcal{I}^{-1}) \quad (14.31)$$

## Linear Parameter Estimators

In many cases, one desires to use a **linear parameter estimator**, i.e.

$$\hat{\beta} = L \vec{y} \quad (14.32)$$

where  $L$  is the **estimator gain matrix**. Then, for a MVUE that is also linear, one can form the optimization

$$\hat{\beta}_{BLUE} = \underset{\text{unbiased, linear } \hat{\beta}}{\operatorname{argmin}} \mathbb{E} \left[ (\hat{\beta} - \mathbb{E}[\hat{\beta}])^T (\hat{\beta} - \mathbb{E}[\hat{\beta}]) \right] \quad (14.33)$$

which is known as the **best linear unbiased estimator (BLUE)**.

As an example, consider a linear measurement, observation, or regression model, i.e.

$$\vec{y} = X \vec{\beta} + \vec{\epsilon} \quad (14.34)$$

where  $X$  is known as the **measurement matrix**, also known as the **observation matrix**, **regression matrix**, or the **design matrix**, and  $\vec{\epsilon}$  is zero-mean **measurement error**, also known as the **observation error** or **regression error**, i.e.,  $\mathbb{E}[\vec{\epsilon}] = 0$ , which is an arbitrary assumption since one can simply form a new linear regression model,  $\vec{y}'$ , with zero-mean regression error,  $\vec{\epsilon}'$ , as

$$\vec{y}' = \vec{y} - \mathbb{E}[\vec{\epsilon}] = X\vec{\beta} + \vec{\epsilon} - \mathbb{E}[\vec{\epsilon}] = X\vec{\beta} + \vec{\epsilon}' \quad (14.35)$$

Furthermore, assume that the covariance of  $\vec{\epsilon}$  is

$$\mathbb{E}[\vec{\epsilon}\vec{\epsilon}^T] = \sigma^2 I \quad (14.36)$$

With this model, the mean of  $\vec{y}$  can be shown to be

$$\mathbb{E}[\vec{y}] = \mathbb{E}[X\vec{\beta} + \vec{\epsilon}] \quad (14.37)$$

$$\mathbb{E}[\vec{y}] = \mathbb{E}[X\vec{\beta}] + \mathbb{E}[\vec{\epsilon}] \quad (14.38)$$

$$\mathbb{E}[\vec{y}] = X\vec{\beta} \quad (14.39)$$

which demonstrates that this linear model is modeling the mean of samples as a linear combination of the unknown parameters.

The covariance of  $\vec{y}$  can be shown to be

$$\text{Cov}(\vec{y}) = \mathbb{E}[(\vec{y} - \mathbb{E}[\vec{y}])(\vec{y} - \mathbb{E}[\vec{y}])^T] \quad (14.40)$$

$$\text{Cov}(\vec{y}) = \mathbb{E}[(X\vec{\beta} + \vec{\epsilon} - X\vec{\beta})(X\vec{\beta} + \vec{\epsilon} - X\vec{\beta})^T] \quad (14.41)$$

$$\text{Cov}(\vec{y}) = \mathbb{E}[\vec{\epsilon}\vec{\epsilon}^T] \quad (14.42)$$

which assumes that the samples have variance modeled by the additive error random vector.

If one assumes that

$$\text{Cov}(\vec{y}) = \sigma^2 I \quad (14.43)$$

Then, it can be shown that the BLUE is given by

$$\hat{\vec{\beta}}_{BLUE} = (X^T X)^{-1} X^T \vec{y} \quad (14.44)$$

with covariance

$$\text{Cov}(\hat{\vec{\beta}}_{BLUE}) = \text{Cov}((X^T X)^{-1} X^T \vec{y}) \quad (14.45)$$

$$\text{Cov}(\hat{\vec{\beta}}_{BLUE}) = (X^T X)^{-1} X^T \text{Cov}(\vec{y}) X (X^T X)^{-1} \quad (14.46)$$

$$\text{Cov}(\hat{\vec{\beta}}_{BLUE}) = (X^T X)^{-1} X^T (\sigma^2 I) X (X^T X)^{-1} \quad (14.47)$$

$$\text{Cov}(\hat{\vec{\beta}}_{BLUE}) = \sigma^2 (X^T X)^{-1} \quad (14.48)$$

As a proof, assume a different linear unbiased estimator exists with minimum variance, i.e.

$$\tilde{\vec{\beta}} = \left( (X^T X)^{-1} X^T + D \right) \vec{y} \quad (14.49)$$

where  $D$  is some non-zero matrix. Then, computing expectation of this estimator allows one to find the condition for which  $\tilde{\beta}$  is unbiased, i.e.

$$\mathbb{E} \left[ \tilde{\beta} \right] = \mathbb{E} \left[ \left( (X^T X)^{-1} X^T + D \right) \vec{y} \right] \quad (14.50)$$

Substituting for  $\vec{y}$

$$\mathbb{E} \left[ \tilde{\beta} \right] = \mathbb{E} \left[ \left( (X^T X)^{-1} X^T + D \right) (X \beta + \epsilon) \right] \quad (14.51)$$

Distributing the expectation for the random vector term

$$\mathbb{E} \left[ \tilde{\beta} \right] = \left( (X^T X)^{-1} X^T + D \right) X \beta + \mathbb{E} \left[ \left( (X^T X)^{-1} X^T + D \right) \epsilon \right] \quad (14.52)$$

which can be simplified since  $\epsilon$  is zero-mean as

$$\mathbb{E} \left[ \tilde{\beta} \right] = \left( (X^T X)^{-1} X^T + D \right) X \beta \quad (14.53)$$

and rearranging

$$\mathbb{E} \left[ \tilde{\beta} \right] = (X^T X)^{-1} X^T X \beta + D X \beta \quad (14.54)$$

or

$$\mathbb{E} \left[ \tilde{\beta} \right] = (I + D X) \beta \quad (14.55)$$

which is unbiased only if  $D X = 0$ .

Then, one can inspect the covariance to see if this new estimator can potentially have a lower variance

$$\text{Cov} \left( \tilde{\beta} \right) = \text{Cov} (L \vec{y}) \quad (14.56)$$

$$\text{Cov} \left( \tilde{\beta} \right) = L \text{Cov} (\vec{y}) L^T \quad (14.57)$$

$$\text{Cov} \left( \tilde{\beta} \right) = \sigma^2 L L^T \quad (14.58)$$

Substituting the potential linear estimator for  $L$

$$\text{Cov} \left( \tilde{\beta} \right) = \sigma^2 \left( (X^T X)^{-1} X^T + D \right) \left( X (X^T X)^{-1} + D^T \right) \quad (14.59)$$

which can be distributed as

$$\text{Cov} \left( \tilde{\beta} \right) = \sigma^2 \left( (X^T X)^{-1} X^T X (X^T X)^{-1} + (X^T X)^{-1} X^T D^T + D X (X^T X)^{-1} + D D^T \right) \quad (14.60)$$

$$\text{Cov} \left( \tilde{\beta} \right) = \sigma^2 \left( (X^T X)^{-1} + (X^T X)^{-1} (D X)^T + D X (X^T X)^{-1} + D D^T \right) \quad (14.61)$$

and simplifies to the following since  $D X = 0$  must be zero

$$\text{Cov} \left( \tilde{\beta} \right) = \sigma^2 (X^T X)^{-1} + 0 + \sigma^2 D D^T \quad (14.62)$$

or

$$\text{Cov} \left( \tilde{\beta} \right) = \text{Cov} \left( \hat{\beta} \right) + \sigma^2 D D^T \quad (14.63)$$

and since  $D D^T$  is positive semi-definite for any matrix  $D$  by the properties of the transpose, one can say that

$$\text{Cov} \left( \tilde{\beta} \right) > \text{Cov} \left( \hat{\beta} \right) \quad (14.64)$$

which implies non-minimum variance for any other unbiased estimator than the BLUE as defined above.

## References

For more information, please refer to the following

- Klein, V., and Morelli, E. G., “Chapter 4 Outline of Estimation Theory,” in *Aircraft System Identification: Theory and Practice*, AIAA, 2006, pp. 75-94
- Klein, V., and Morelli, E. G., “6.1.3 Properties of Maximum Likelihood Parameter Estimates,” in *Aircraft System Identification: Theory and Practice*, AIAA, 2006, pp. 190-191

## 14.2 Batch Least-Squares Parameter Estimation

### Batch Least-Squares Regression

In optimal parameter estimation, one can alternatively select the parameters via the **regression problem** where one has  $N$  sets of independent data,  $\vec{x}[k]$  which produced dependent samples,  $\vec{y}[k]$ , where  $k = 1, \dots, N$ . In addition, consider that one desires to optimally fit this data to a chosen **regression model**, i.e.

$$\vec{y}[k] = f_k(k, \vec{x}[k], \vec{\beta}) + \vec{\epsilon}[k] \quad (14.65)$$

where  $\vec{\beta}$  contains the model parameters, also known as the **regressors**, and  $\vec{\epsilon}[k]$  is the **error** random process with zero-mean, covariances  $R[k]$ , and uncorrelated with respect to  $k$ . The **residuals** are defined as

$$\vec{r}[k] = \vec{y}[k] - f_k(k, \vec{x}[k], \vec{\beta}) \sim \vec{\epsilon}[k] \quad (14.66)$$

To cast this in vector and matrix notation, one can redefine these in as the stacked vectors

$$\vec{y} = \begin{bmatrix} \vec{y}[1] \\ \vdots \\ \vec{y}[N] \end{bmatrix} \quad (14.67)$$

$$\vec{x} = \begin{bmatrix} \vec{x}[1] \\ \vdots \\ \vec{x}[N] \end{bmatrix} \quad (14.68)$$

$$\mathbf{f}(\vec{x}, \vec{\beta}) = \begin{bmatrix} f_1(1, \vec{x}[1], \vec{\beta}) \\ \vdots \\ f_N(N, \vec{x}[N], \vec{\beta}) \end{bmatrix} \quad (14.69)$$

$$\vec{\epsilon} = \begin{bmatrix} \vec{\epsilon}[1] \\ \vdots \\ \vec{\epsilon}[N] \end{bmatrix} \quad (14.70)$$

$$\mathbf{R} = \begin{bmatrix} R[1] & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & R[N] \end{bmatrix} \quad (14.71)$$

and

$$\vec{r} = \begin{bmatrix} \vec{r}[1] \\ \vdots \\ \vec{r}[N] \end{bmatrix} \quad (14.72)$$

which results in the stacked or “batch” regression model

$$\vec{y} = \mathbf{f}(\vec{x}, \vec{\beta}) + \vec{\epsilon} \quad (14.73)$$

Choosing the optimality criterion of least-squares of the residuals, one can form the **least-squares (LS) regression problem**, which minimizes the sum of squares of the residuals, i.e.

$$\hat{\vec{\beta}}_{LS} = \underset{\vec{\beta}}{\operatorname{argmin}} \vec{r}^T \mathbf{R}^{-1} \vec{r} = \underset{\vec{\beta}}{\operatorname{argmin}} \left( \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta}) \right)^T \mathbf{R}^{-1} \left( \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta}) \right) \quad (14.74)$$

where  $\hat{\vec{\beta}}_{LS}$  is the **least-squares estimator** and  $\mathbf{R}$  is a positive definite weight matrix for each sample  $k = 1, \dots, N$ . The term “least” in *least-squares* refers to a minimization while the term “squares” refers to a summation of multiple squared terms, in particular, the residuals. It can also be pointed out that the LS estimator minimizes the Mahalanobis distance between the samples,  $\vec{y}$ , and the model distribution with unknown mean  $\mathbf{f}(\vec{x}, \vec{\beta})$  and covariance  $\mathbf{R}$ . In this case, one doesn’t explicitly require  $f_{\vec{Y}|\vec{\beta}}(\vec{y}|\vec{\beta})$ , only the mean and covariance of the regression model.

Furthermore, if the chosen regression model,  $f()$ , is linear, i.e.

$$\vec{y}[k] = X[k]\vec{\beta} + \vec{\epsilon}[k] \quad (14.75)$$

where  $X[k]$  are matrices. Then, defining the stacked matrix

$$\mathbf{X} = [X[1] \quad \cdots \quad X[N]] \quad (14.76)$$

one can rewrite the **linear least-squares problem** as

$$\hat{\vec{\beta}}_{LLS} = \underset{\vec{\beta}}{\operatorname{argmin}} \left( \vec{y} - \mathbf{X}\vec{\beta} \right)^T \mathbf{R}^{-1} \left( \vec{y} - \mathbf{X}\vec{\beta} \right) \quad (14.77)$$

where  $\hat{\vec{\beta}}_{LLS}$  is the **linear least-squares (LLS) parameter estimator**.

If  $\mathbf{R} = \sigma^2 I$ , then the LLS problem is the **ordinary least-squares (OLS) problem**. The **Gauss-Markov theorem** states that if the errors,  $\vec{\epsilon}[k]$ , are truly zero-mean, serially uncorrelated, and have constant covariances with respect to  $k$ , i.e., homoscedastic, then the OLS estimator is the BLUE. If also  $\vec{\epsilon}[k] \sim \mathcal{N}(0, \sigma^2 I)$ , then the OLS is also the MLE. If any  $R$  varies with  $[k]$ , then LLS problem is the **generalized least-squares (GLS) problem**. The corollary to Gauss-Markov theorem is that if  $\vec{\epsilon}[k]$  are zero-mean, serially uncorrelated, and heteroscedastic, then the GLS is the BLUE. If also  $\vec{\epsilon}[k] \sim \mathcal{N}(0, R[k])$ , the GLS is also the MLE. Note that if  $\mathbf{R}$  are diagonal, the GLS problem is also known as the **weighted least-squares (WLS) problem**.

## Ordinary Least-Squares Regression

If  $\mathbf{X}$  is square, then the OLS can be solved by simply taking the inverse of  $X$

$$\vec{\beta}_{OLS} = \mathbf{X}^{-1} \vec{\mathbf{y}} \quad (14.78)$$

However, if  $\mathbf{X}$  is not square, then the true inverse of  $\mathbf{X}$  does not exist, then one uses the **pseudoinverse**, also known as the **Moore-Penrose inverse**,  $\mathbf{X}^+$ , which satisfies the following properties:

- $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$
- $\mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+$
- $(\mathbf{X}\mathbf{X}^+)^* = \mathbf{X}\mathbf{X}^+$
- $(\mathbf{X}^+\mathbf{X})^* = \mathbf{X}^+\mathbf{X}$

It can be shown that the pseudoinverse exists for all matrices, but may not have a simple algebraic formula. However, if  $\mathbf{X}$  has linearly independent columns (i.e. full rank), then

$$\mathbf{X}^+ = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (14.79)$$

which is also known as the **left pseudoinverse**.

Thus, the **ordinary least-squares (OLS) parameter estimator** can be generalized as

$$\vec{\beta}_{OLS} = \mathbf{X}^+ \vec{\mathbf{y}} \quad (14.80)$$

and if  $\mathbf{X}^T \mathbf{X}$  is invertible

$$\vec{\beta}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{\mathbf{y}} \quad (14.81)$$

As a proof of the OLS solution, let

$$\mathcal{J}(\vec{\beta}) = (\vec{\mathbf{y}} - \mathbf{X}\vec{\beta})^T (\vec{\mathbf{y}} - \mathbf{X}\vec{\beta}) \quad (14.82)$$

Then, the minimum of  $\mathcal{J}(\vec{\beta})$  occurs when its gradient is equal to  $\vec{0}$ .

$$\nabla \mathcal{J}(\vec{\beta}) = 2\mathbf{X}^T (\mathbf{X}\vec{\beta}_{OLS} - \vec{\mathbf{y}}) = \vec{0}^T \quad (14.83)$$

$$\mathbf{X}^T \mathbf{X}\vec{\beta}_{OLS} - \mathbf{X}^T \vec{\mathbf{y}} = \vec{0} \quad (14.84)$$

$$\mathbf{X}^T \mathbf{X}\vec{\beta}_{OLS} = \mathbf{X}^T \vec{\mathbf{y}} \quad (14.85)$$

$$\vec{\beta}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{\mathbf{y}} \quad (14.86)$$

which exists as long as  $\mathbf{X}^T \mathbf{X}$  is left-invertible, i.e.,  $\mathbf{X}$  full rank.

Solving for  $\mathbf{X}^T \mathbf{X}\vec{\beta}^* = \mathbf{X}^T \vec{\mathbf{y}}$  by back substitution can suffer from numerical difficulties due to the matrix multiplication  $\mathbf{X}^T \mathbf{X}$  for certain matrices. Thus, most solvers use the QR decomposition of  $\mathbf{X}$  which allows one to reform  $\vec{\beta}_{OLS}$  as

$$\vec{\beta}_{OLS} = ((QR)^T QR)^{-1} (QR)^T \vec{\mathbf{y}} \quad (14.87)$$

$$\vec{\beta}_{OLS} = \left( R^T Q^T Q R \right)^{-1} R^T Q^T \vec{y} \quad (14.88)$$

$$\vec{\beta}_{OLS} = \left( R^T R \right)^{-1} R^T Q^T \vec{y} \quad (14.89)$$

$$\vec{\beta}_{OLS} = R^{-1} R^{-T} R^T Q^T \vec{y} \quad (14.90)$$

$$\vec{\beta}_{OLS} = R^{-1} Q^T \vec{y} \quad (14.91)$$

$$R \vec{\beta}_{OLS} = Q^T \vec{y} \quad (14.92)$$

Finally, it should be noted that numerical algorithms for efficiently computing the OLS solution are typically performed as follows.

1. Compute the QR decomposition on  $X$ 
  - Method typically uses Householder matrices
  - Reflects vector about some hyperplane
2. Compute  $\vec{c} = Q^T \vec{y}$
3. Solve  $R \vec{\beta}_{OLS} = \vec{c}$  for  $\vec{\beta}_{OLS}$  by back substitution

A metric that quantifies the “nearness” of  $\hat{\vec{\beta}}[k]$  to  $\vec{y}[k]$  is the **coefficient of determination**,  $R^2$ . Its definition for OLS estimation follows computing the mean value of  $\vec{y}$  as

$$\vec{\mu}_{\vec{y}} = \frac{1}{N} \sum_{k=1}^N \vec{y}[k] \quad (14.93)$$

which can be stacked  $N$  times as

$$\vec{\mu}_{N,\vec{y}} = \begin{bmatrix} \vec{\mu}_{\vec{y}} \\ \vdots \\ \vec{\mu}_{\vec{y}} \end{bmatrix} \quad (14.94)$$

Then, one can analyze the **total sum of squares**, denoted by  $SS_{tot}$ , as the total squared variations in the measured output  $\vec{y}$  from  $\vec{\mu}_{\vec{y}}$ , denoted by  $SS_{tot}$ , defined as

$$SS_{tot} = (\vec{y} - \vec{\mu}_{N,\vec{y}})^T (\vec{y} - \vec{\mu}_{N,\vec{y}}) \quad (14.95)$$

$$SS_{tot} = \vec{y}^T \vec{y} - 2 \vec{\mu}_{N,\vec{y}}^T \vec{y} + N \vec{\mu}_{\vec{y}}^T \vec{\mu}_{\vec{y}} \quad (14.96)$$

This can be compared to the **residual sum of squares** as the sum of squared variations of the residuals between each  $\vec{y}[k]$  and the estimate  $\mathbf{f}(\vec{x}, \hat{\vec{\beta}})$ , denoted by  $SS_{res}$ , and defined as

$$SS_{res} = (\vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}))^T (\vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}})) \quad (14.97)$$

For well-fit models,  $SS_{res}$  will be small relative to  $SS_{tot}$ . Specifically, the coefficient of determination,  $0 \leq R^2 \leq 1$ , represents the proportion of the variation in the measured output that is explained by the model, i.e.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (14.98)$$

where values of  $R^2$  vary from 0 to 1, where 1 represents a perfect fit to the data. Note that  $R^2$  is usually expressed as a percentage in regression analysis.

For OLS regression, this variation can be partitioned into two parts. The first is  $SS_{res}$  which for OLS is

$$SS_{res} = (\vec{y} - \mathbf{X}\hat{\beta})^T(\vec{y} - \mathbf{X}\hat{\beta}) \quad (14.99)$$

which, by expansion of the quadratic terms, one has

$$SS_{res} = \vec{y}^T \vec{y} - 2\hat{\beta}^T \mathbf{X}^T \vec{y} + \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta} \quad (14.100)$$

and by substitution for  $\hat{\beta}$ ,

$$SS_{res} = \vec{y}^T \vec{y} - 2\hat{\beta}^T \mathbf{X}^T \vec{y} + \hat{\beta}^T \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{y} \quad (14.101)$$

$$SS_{res} = \vec{y}^T \vec{y} - 2\hat{\beta}^T \mathbf{X}^T \vec{y} + \hat{\beta}^T \mathbf{X}^T \vec{y} \quad (14.102)$$

The second is the **regression sum of squares** also known as the **explained sum of squares**, denoted by  $SS_{reg}$ , and defined as the sum of squared variations of the model about the same mean value, i.e.,

$$SS_{reg} = (\mathbf{X}\hat{\beta} - \vec{\mu}_{N,\vec{y}})^T(\mathbf{X}\hat{\beta} - \vec{\mu}_{N,\vec{y}}) \quad (14.103)$$

which, by expansion of the quadratic terms, one has

$$SS_{reg} = (\mathbf{X}\hat{\beta})^T \mathbf{X}\hat{\beta} - 2\vec{\mu}_{N,\vec{y}}^T \mathbf{X}\hat{\beta} + \vec{\mu}_{N,\vec{y}}^T \vec{\mu}_{N,\vec{y}} \quad (14.104)$$

and by substitution for

$$SS_{reg} = \hat{\beta}^T \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{y} - 2\vec{\mu}_{N,\vec{y}}^T \mathbf{X}\hat{\beta} + N\vec{\mu}_{\vec{y}}^T \vec{\mu}_{\vec{y}} \quad (14.105)$$

or

$$SS_{reg} = \hat{\beta}^T \mathbf{X}^T \vec{y} - 2\vec{\mu}_{N,\vec{y}}^T \mathbf{X}\hat{\beta} + N\vec{\mu}_{\vec{y}}^T \vec{\mu}_{\vec{y}} \quad (14.106)$$

Thus, for OLS, one has

$$SS_{reg} + SS_{res} = \vec{y}^T \vec{y} - 2\vec{\mu}_{N,\vec{y}}^T \mathbf{X}\hat{\beta} + N\vec{\mu}_{\vec{y}}^T \vec{\mu}_{\vec{y}} = SS_{tot} \quad (14.107)$$

and

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}} = \frac{\hat{\beta}^T \mathbf{X}^T \vec{y} - 2\vec{\mu}_{N,\vec{y}}^T \mathbf{X}\hat{\beta} + N\vec{\mu}_{\vec{y}}^T \vec{\mu}_{\vec{y}}}{\vec{y}^T \vec{y} - 2\vec{\mu}_{N,\vec{y}}^T \mathbf{X}\hat{\beta} + N\vec{\mu}_{\vec{y}}^T \vec{\mu}_{\vec{y}}} \quad (14.108)$$

The covariance matrix of the OLS estimator error can be shown to be

$$\text{Cov}(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \mathbb{E}[\hat{\beta}])(\hat{\beta} - \mathbb{E}[\hat{\beta}])^T] \quad (14.109)$$

$$\text{Cov}(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \vec{\beta})(\hat{\beta} - \vec{\beta})^T] \quad (14.110)$$

$$\text{Cov}(\hat{\beta}) = \mathbb{E}[\left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{y} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\vec{y} - \vec{\epsilon})\right) \left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{y} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\vec{y} - \vec{\epsilon})\right)^T] \quad (14.111)$$

$$\text{Cov}(\hat{\beta}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}[\vec{\epsilon} \vec{\epsilon}^T] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \quad (14.112)$$

$$\text{Cov}(\hat{\beta}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{R} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \quad (14.113)$$

Furthermore, if the regression error is uncorrelated and has constant variance across  $k$ , i.e.,  $\mathbf{R} = \sigma^2 I$  then one has

$$\text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (14.114)$$

Recall that the diagonal elements of this covariance matrix represent the parameter variances,  $\sigma^2(\hat{\beta}_i)$ , i.e. the standard deviations squared. Furthermore, under the assumption of Gaussian distributed parameter estimate, one can form the **100(1- $\alpha$ )% confidence interval** for each parameter estimate,  $\hat{\beta}_i$  for  $i = 1, \dots, n_\beta$ , using the

$$\hat{\beta}_i + F_t^{-1}(\alpha/2, N - n_\beta) \sigma(\hat{\beta}_i) \leq \beta_i \leq \hat{\beta}_i + F_t^{-1}(1 - \alpha/2, N - n_\beta) \sigma(\hat{\beta}_i) \quad (14.115)$$

where  $F_t^{-1}$  is the inverse CDF of the  $t$ -distribution. Recall that as  $N - n_\beta \rightarrow \infty$ ,  $F_t^{-1}$  will approach the inverse standard normal distribution,  $F_N^{-1}$ , e.g.,  $F_N^{-1}(0.05/2) = -1.96 \approx -2$  and  $F_N^{-1}(1 - 0.05/2) = 1.96 \approx 2$ .

The off-diagonal elements represent the covariance between any two parameter estimates,  $\sigma^2(\hat{\beta}_i, \hat{\beta}_j)$ , and can be related to the correlation coefficient between any two parameter estimates by diving by the standard deviations of each parameter, i.e.

$$\rho(\hat{\beta}_i, \hat{\beta}_j) = \frac{\sigma^2(\hat{\beta}_i, \hat{\beta}_j)}{\sigma(\hat{\beta}_i)\sigma(\hat{\beta}_j)} \quad (14.116)$$

In the practice of least-squares regression, often one does not know if the samples are uncorrelated with respect to the time step,  $k$ . One method for checking this correlation is by computing the **residual auto-correlation function** of the observed residuals between the fitted model and the samples, i.e.

$$R_{\vec{r} \vec{r}}[j] = \frac{1}{N} \sum_{k=1}^{N-j} \vec{r}[j] \vec{r}^T[k+j] \quad (14.117)$$

for **frequency indexes**  $j = 0, \dots, N - 1$ , i.e. the different possible time separations for the residuals. As this regression approach to parameter estimation assumes the sampling process is WSS, if the residuals are completely uncorrelated, then it should be that  $R_{\vec{r} \vec{r}}[j]$  is approximately a zero matrix for  $j \neq 0$ . Although the estimated residuals will not be exactly zero, but one should expect  $100(1 - \alpha)\%$  of the  $i^{\text{th}}$  residual element would be randomly distributed within  $\pm F_t^{-1}(1 - \alpha/2, N - n_y) \sqrt{R_{r_i r_i}[0]} N$  by a Gaussian assumption on the residual auto-correlation function distribution.

## Generalized Least-Squares Regression

The GLS estimate can be considered as the OLS on scaled and “de-correlated” observations. To see this consider the Cholesky decomposition of  $\mathbf{R}$ , i.e.

$$\mathbf{R} = CC^T \quad (14.118)$$

Next, multiply the observation model by  $C^{-1}$  on both sides, i.e.

$$C^{-1}\vec{y} = C^{-1}\mathbf{X}\vec{\beta} + C^{-1}\vec{\epsilon} \quad (14.119)$$

Then, one can form a “new” model

$$\vec{y}^* = \mathbf{X}^*\vec{\beta} + \vec{\epsilon}^* \quad (14.120)$$

where

$$\vec{y}^* = C^{-1}\vec{y} \quad (14.121)$$

$$\mathbf{X}^* = C^{-1}\mathbf{X} \quad (14.122)$$

$$\vec{\epsilon}^* = C^{-1}\vec{\epsilon} \quad (14.123)$$

For this model, one can show that

$$\text{Cov}(\vec{\epsilon}^*) = \text{Cov}(C^{-1}\vec{\epsilon}) \quad (14.124)$$

$$\text{Cov}(\vec{\epsilon}^*) = C^{-1}\text{Cov}(\vec{\epsilon})C^{-T} \quad (14.125)$$

or

$$\text{Cov}(\vec{\epsilon}^*) = C^{-1}RC^{-T} \quad (14.126)$$

and substituting for  $R$  from the Cholesky decomposition, one has

$$\text{Cov}(\vec{\epsilon}^*) = C^{-1}CC^TC^{-T} \quad (14.127)$$

$$\text{Cov}(\vec{\epsilon}^*) = I \quad (14.128)$$

Thus, substituting back for the “new” parameters into the expression for  $\hat{\vec{\beta}}_{OLS}$ , one has

$$\hat{\vec{\beta}}_{GLS} = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\vec{y}^* \quad (14.129)$$

$$\hat{\vec{\beta}}_{GLS} = (\mathbf{X}^TC^{-T}C^{-1}\mathbf{X})^{-1}\mathbf{X}^TC^{-T}C^{-1}\vec{y} \quad (14.130)$$

Thus, by substitution, the **generalized least-squares (GLS) parameter estimator** is

$$\hat{\vec{\beta}}_{GLS} = (\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}^{-1}\vec{y} \quad (14.131)$$

where the covariance is given by

$$\text{Cov}(\hat{\vec{\beta}}_{GLS}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \quad (14.132)$$

## Nonlinear Batch Least-Squares Regression

For nonlinear  $f()$  and with an optimality criterion of least-squares, one can form the **nonlinear least-squares (NLS) regression problem**,  $\hat{\beta}_{NLS}$ , which minimizes the sum of squares of the residuals, i.e.

$$\hat{\beta}_{NLS} = \underset{\vec{\beta}}{\operatorname{argmin}} \vec{r}^T \mathbf{R}^{-1} \vec{r} = \underset{\vec{\beta}}{\operatorname{argmin}} \left( \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta}) \right)^T \mathbf{R}^{-1} \left( \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta}) \right) \quad (14.133)$$

where  $\hat{\beta}_{NLS}$  is the **nonlinear least-squares (NLS) estimator**. Any local minimum for this optimization, if one exists, is found by setting the derivative of the squared residuals to zero, i.e.

$$2 \vec{r}^T \frac{\partial \vec{r}}{\partial \vec{\beta}} = 0 \quad (14.134)$$

or, in terms of the regression model Jacobian,  $\frac{\partial \mathbf{f}(\vec{x}, \vec{\beta})}{\partial \vec{\beta}}$ , one has

$$-2 \vec{r}^T \frac{\partial \mathbf{f}(\vec{x}, \vec{\beta})}{\partial \vec{\beta}} = 0 \quad (14.135)$$

where the optimal estimate,  $\hat{\beta}_{NLS}$ , will be the one (or multiple) values of  $\vec{\beta}$  which accomplish this.

However, one cannot often find an analytical solution to the NLS estimation problem, thus, one must use numerical methods to find the optimal value. Such methods begin with some initial estimate,  $\hat{\beta}_0$ , and then use an iterative procedure to refine the parameter estimate, i.e.

$$\hat{\beta}_{i+1} = \hat{\beta}_i + \vec{\Delta}_i \quad (14.136)$$

where  $i$  is the iteration number and  $\vec{\Delta}_i$  is the **search vector** at iteration  $i$ . Then, after some convergence criteria for the search vector, one can approximate the NLS estimator by

$$\hat{\beta}_{NLS} \approx \hat{\beta}_i \quad \text{for } \frac{\|\vec{\Delta}_i\|}{\|\hat{\beta}_i\|} < \delta \quad (14.137)$$

where a typical value for  $\delta$  would be 0.001 which would require a precision of 0.1%. When used in this form, the NLS estimator is a type of **iterative least-squares (ILS) estimator**.

One of the simplest methods for computing  $\vec{\Delta}_i$  is the **Gauss-Newton algorithm (GNA)** which approximates the observation model by a first-order Taylor series expansion, i.e.

$$\mathbf{f}(\vec{x}, \vec{\beta}) \approx \mathbf{f}(\vec{x}, \hat{\beta}_i) + \frac{\partial \mathbf{f}(\vec{x}, \vec{\beta})}{\partial \vec{\beta}} \Big|_{\vec{\beta}=\hat{\beta}_i} \vec{\Delta}_i \vec{\beta} \quad (14.138)$$

where

$$\vec{\Delta}_i = \vec{\beta} - \hat{\beta}_i \quad (14.139)$$

Next, defining the Jacobian of the regression model as

$$\mathbf{J} = \frac{\partial \mathbf{f}(\vec{x}, \vec{\beta})}{\partial \vec{\beta}} \Big|_{\vec{\beta}=\hat{\beta}_i} \quad (14.140)$$

one can rewrite  $\vec{r} = \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta})$  as

$$\vec{r} = \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_i) + \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_i) - \mathbf{f}(\vec{x}, \vec{\beta}) \quad (14.141)$$

or

$$\vec{r} \approx \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_i) - \mathbf{J}\vec{\Delta}_i \quad (14.142)$$

Then, setting the gradient of this approximation for the squared residuals to zero, one has

$$-2\vec{r}^T \frac{\partial f}{\partial \vec{\beta}} = -2 \left( \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_i) - \mathbf{J}\vec{\Delta}_i \right)^T \mathbf{J} = 0 \quad (14.143)$$

which can be rewritten as

$$\mathbf{J}^T \left( \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_i) - \mathbf{J}\vec{\Delta}_i \right) = 0 \quad (14.144)$$

and by rearranging, one obtains the OLS solution for the search vector

$$\vec{\Delta}_i = (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \left( \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_i) \right) \quad (14.145)$$

Furthermore, if the residuals have an expected covariance matrix,  $\mathbf{R}$ , then one can alternatively use the GLS solution as

$$\vec{\Delta}_i = (\mathbf{J}^T \mathbf{R}^{-1} \mathbf{J})^{-1} \mathbf{J}^T \mathbf{R}^{-1} \left( \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_i) \right) \quad (14.146)$$

An important part of this numerical method is the initial parameter estimates and the problem of divergence. Since divergence in the GNA often occurs, one often uses the **Levenburg-Marquardt algorithm (LMA)**, also known as **damped least-squares (DLS)**, as a trust-region augmentation to the GNA. In this case, one includes a damping factor,  $\zeta_i$ , also known as **Marquardt parameter** in the search vector calculation as

$$\vec{\Delta}_i = (\mathbf{J}^T \mathbf{R}^{-1} \mathbf{J} + \zeta_i I)^{-1} \mathbf{J}^T \mathbf{R}^{-1} \left( \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_i) \right) \quad (14.147)$$

where this second term combines the GNA with the direction of **steepest gradient**, i.e.

$$\vec{\Delta}_i = \frac{1}{\zeta_i} \mathbf{J}^T \mathbf{R}^{-1} \left( \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_i) \right) \quad (14.148)$$

which the LMA approximates if  $\mathbf{J}^T \mathbf{R}^{-1} \mathbf{J} \ll \zeta_i I$ . The LMA can be further improved through Fletcher's **modified LMA**

$$\vec{\Delta}_i = \left( \mathbf{J}^T \mathbf{R}^{-1} \mathbf{J} + \zeta_i \text{diag}(\mathbf{J}^T \mathbf{R}^{-1} \mathbf{J}) \right)^{-1} \mathbf{J}^T \mathbf{R}^{-1} \left( \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_i) \right) \quad (14.149)$$

where  $\text{diag}(\mathbf{J}^T \mathbf{R}^{-1} \mathbf{J})$  selects the diagonal elements of  $\mathbf{J}^T \mathbf{R}^{-1} \mathbf{J}$ . This modification slows down convergence in the direction of small gradients by adjusting more where the gradient of  $f()$  is smaller.

In either case of LMA,  $\zeta_i$  is heuristically adjusted at each iteration which is intrinsic to the trust-region method. An effective heuristic known as **delayed gratification** consists of increasing  $\zeta_i$  by a small amount for iterations where  $\vec{r}^T \vec{r}$  increases, and decreasing  $\zeta_i$  by a large amount for iterations where  $\vec{r}^T \vec{r}$  decreases. This heuristic slows down convergence which avoids converging too quickly in the beginning of optimization. An increase by a factor of 2 and a decrease by a factor of 3 are effective for most problems. Other trust-region methods can also be used which solve a sub-problem for the optimal damping factor which are discussed in the next section.

## Constrained Batch Least-Squares Regression

**Trust-region methods**, also known as **restricted-step methods**, can be used for both unconstrained and constrained least-squares regression problems. The **constrained least-squares (CLS) regression problem** can be stated as

$$\begin{aligned}\hat{\vec{\beta}}^{CLS} &= \underset{\vec{\beta} \in X}{\operatorname{argmin}} \quad \left( \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta}) \right)^T \mathbf{R}^{-1} \left( \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta}) \right) \\ \text{subject to: } l_{\beta} &\leq \vec{\beta} \leq u_{\beta}\end{aligned}\tag{14.150}$$

where  $l_{\beta}$  and  $u_{\beta}$  are the lower and upper bounds on the parameter estimate, respectively. This section will discuss how trust-region methods solve the CLS problem at a high level.

To understand trust-region methods, consider the general minimization search problem where one desires to improve, i.e. move from  $\hat{\vec{\beta}}_i$  to some  $\hat{\vec{\beta}}_{i+1}$  which reduces the least-squares function. Trust-region methods approximate the least-squares function,  $\left( \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta}) \right)^T \mathbf{R}^{-1} \left( \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta}) \right)$  with a simpler function,  $q(\vec{y}, \vec{x}, \vec{\beta})$  in some neighborhood about  $\hat{\vec{\beta}}_i$ , then solving for the value where the least-squares function approximation does decrease, i.e. the optimal **trial step vector**,  $\vec{\Delta}_i$ , within that neighborhood, i.e. **trust-region**. However, one must also check that the obtained trial step vector can be “trusted,” i.e. that

$$\left( \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta}) \right)^T \mathbf{R}^{-1} \left( \vec{y} - \mathbf{f}(\vec{x}, \vec{\beta}) \right) < q(\vec{y}, \vec{x}, \hat{\vec{\beta}}_i + \vec{\Delta}_i)\tag{14.151}$$

where if this fails to hold, then the trust-region must be adjusted and a new trial step vector must be found. Thus, different trust-region methods use different approximations,  $q()$ , different trial step vector solvers in the trust-region, and different trust-region adjustments. Standard trust-region methods use a quadratic form for the least-squares approximation, e.g. the LMA, which defines an ellipsoidal trust-region, quadratic programming solvers, and standard trust-region adjustments similar to the delayed gratification heuristic above.

## Quadratic Programming

As an aside, the **quadratic programming (QP) problem** can be stated as

$$\begin{aligned}\vec{\beta}^{QP} &= \underset{\vec{\beta}}{\operatorname{argmin}} \quad \frac{1}{2} \vec{\beta}^T Q \vec{\beta} + c^T \vec{\beta} \\ \text{subject to: } A \vec{\beta} &\leq \vec{b}\end{aligned}\tag{14.152}$$

where the constraint is a vector-defined component-wise inequality. QP problems appear in different problems. One is minimizing a function  $f()$  in a local neighborhood about a point  $\vec{\beta}_0$ , e.g. a trust-region, one can set  $Q$  to the Hessian matrix at that point,  $H(\vec{\beta}_0)$ , and  $c$  to its gradient,  $\nabla f(\vec{\beta}_0)$ .

As a special case, when  $Q = Q^T > 0$ , a QP problem is equivalent to the LLS problem. To see this, consider the LLS minimization written as

$$\vec{\beta}_{LLS} = \underset{\vec{\beta}}{\operatorname{argmin}} \quad \left( \vec{y} - \mathbf{X} \vec{\beta} \right)^T \left( \vec{y} - \mathbf{X} \vec{\beta} \right)\tag{14.153}$$

$$\vec{\beta}_{LLS} = \underset{\vec{\beta}}{\operatorname{argmin}} \vec{\mathbf{y}}^T \vec{\mathbf{y}} - \vec{\mathbf{y}}^T \mathbf{X} \vec{\beta} - \vec{\beta}^T \mathbf{X}^T \vec{\mathbf{y}} + \vec{\beta}^T \mathbf{X}^T \mathbf{X} \vec{\beta} \quad (14.154)$$

which is equivalent to minimizing only over terms with  $\vec{\beta}$ , multiplying the expression by  $\frac{1}{2}$ , and that the order of the middle terms produces the same result, i.e.

$$\vec{\beta}_{LLS} = \underset{\vec{\beta}}{\operatorname{argmin}} \frac{1}{2} \vec{\beta}^T \mathbf{X}^T \mathbf{X} \vec{\beta} - \vec{\mathbf{y}}^T \mathbf{X} \vec{\beta} \quad (14.155)$$

which by the Cholesky decomposition  $Q = \mathbf{X}^T \mathbf{X}$  and defining  $\vec{c} = -\mathbf{X}^T \vec{\mathbf{y}}$ , one has

$$\vec{\beta}_{LLS} = \underset{\vec{\beta}}{\operatorname{argmin}} \frac{1}{2} \vec{\beta}^T Q \vec{\beta} + c^T \vec{\beta} \quad (14.156)$$

QP programming solvers can be derived explicitly for equality constraints and for  $Q > 0$  similar to the least-squares solver, i.e. the ellipsoid method solves the problem in (weakly) polynomial time. However if  $Q$  is indefinite or has even one negative eigenvalue, then the problem is NP-hard.

### Discrete-Time LQR as Batch Linear Least-Squares Problem

The unconstrained finite-horizon discrete-time LQR can be formulated as a LLS problem. Consider the unconstrained finite-horizon discrete-time LQ OCP written as

$$\begin{aligned} \vec{u}_{\text{opt}}[k] = \underset{\vec{u}[k] \text{ for } k=0, \dots, N-1}{\operatorname{argmin}} \quad & \mathcal{J} = \vec{x}^T[N] E \vec{x}[N] + \sum_{k=0}^{N-1} \vec{x}^T[k] Q \vec{x}[k] + \vec{u}^T[k] R \vec{u}[k] \\ \text{subject to: } & \vec{x}[k+1] = F \vec{x}[k] + G \vec{u}[k] \\ \text{initial condition: } & \vec{x}[0] = \vec{x}_0 \end{aligned} \quad (14.157)$$

where  $S$  has been removed from the cost function for simplification of the following derivations, although it can be included in the LLS solution for the discrete-time LQR as well.

Based on the discrete-time linear dynamics and initial condition, one can write out the general solution of the state as a “large” linear function, i.e.

$$\begin{bmatrix} \vec{x}[0] \\ \vdots \\ \vec{x}[N] \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ G & 0 & 0 & \cdots & 0 \\ FG & G & 0 & \cdots & 0 \\ F^2G & FG & G & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F^{N-1}G & F^{N-2}G & F^{N-3}G & \cdots & G \end{bmatrix} \begin{bmatrix} \vec{u}[0] \\ \vdots \\ \vec{u}[N-1] \end{bmatrix} + \begin{bmatrix} I \\ F \\ \vdots \\ F^N \end{bmatrix} \vec{x}_0 \quad (14.158)$$

which one can rewrite succinctly by defining the above vectors and matrices as

$$\vec{x} = \mathbf{G} \vec{u} + \mathbf{F} \vec{x}_0 \quad (14.159)$$

Then, using these new terms and the following block diagonal matrices

$$\mathbf{Q} = \begin{bmatrix} Q & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & Q & 0 \\ 0 & \cdots & 0 & E \end{bmatrix} \quad (14.160)$$

and

$$\mathbf{R} = \begin{bmatrix} R & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & R \end{bmatrix} \quad (14.161)$$

the cost function for the discrete-time LQ OCP can be rewritten using vector notation as

$$\mathcal{J} = \vec{\mathbf{x}}^T \mathbf{Q} \vec{\mathbf{x}} + \vec{\mathbf{u}}^T \mathbf{R} \vec{\mathbf{u}} \quad (14.162)$$

and substituting for  $\vec{\mathbf{x}}$  from before, one can write

$$\mathcal{J} = (\mathbf{G} \vec{\mathbf{u}} + \mathbf{F} \vec{x}_0)^T \mathbf{Q} (\mathbf{G} \vec{\mathbf{u}} + \mathbf{F} \vec{x}_0) + \vec{\mathbf{u}}^T \mathbf{R} \vec{\mathbf{u}} \quad (14.163)$$

By using the definition of the square root of positive semi-definite matrices, i.e.,  $Q = Q^{\frac{1}{2}}Q^{\frac{1}{2}}$ , the cost function can be rewritten as

$$\mathcal{J} = (\mathbf{G} \vec{\mathbf{u}} + \mathbf{F} \vec{x}_0)^T \mathbf{Q}^{\frac{1}{2}} \mathbf{Q}^{\frac{1}{2}} (\mathbf{G} \vec{\mathbf{u}} + \mathbf{F} \vec{x}_0) + \vec{\mathbf{u}}^T \mathbf{R}^{\frac{1}{2}} \mathbf{R}^{\frac{1}{2}} \vec{\mathbf{u}} \quad (14.164)$$

and using the definition of vector norms, one has

$$\mathcal{J} = \left\| \mathbf{Q}^{\frac{1}{2}} \mathbf{G} \vec{\mathbf{u}} + \mathbf{Q}^{\frac{1}{2}} \mathbf{F} \vec{x}_0 \right\|_2^2 + \left\| \mathbf{R}^{\frac{1}{2}} \vec{\mathbf{u}} \right\|_2^2 \quad (14.165)$$

which is a LLS problem. This form also uses the fact that  $\mathbf{Q}$  and  $\mathbf{R}$  are symmetric since  $E$ ,  $Q$ , and  $R$  are as well.

Finally, taking the gradient for  $L^2$ -norms squared with respect to  $\vec{\mathbf{u}}$ ,

$$\nabla \mathcal{J} = 2 \left( \mathbf{Q}^{\frac{1}{2}} \mathbf{G} \right)^T \left( \mathbf{Q}^{\frac{1}{2}} \mathbf{G} \vec{\mathbf{u}} + \mathbf{Q}^{\frac{1}{2}} \mathbf{F} \vec{x}_0 \right) + 2 \mathbf{R} \vec{\mathbf{u}} \quad (14.166)$$

and solving for  $\nabla \mathcal{J} = 0$ ,

$$2 \mathbf{G}^T \mathbf{Q}^{\frac{1}{2}} \mathbf{Q}^{\frac{1}{2}} \mathbf{G} \vec{\mathbf{u}}^{\text{opt}} + 2 \mathbf{G}^T \mathbf{Q}^{\frac{1}{2}} \mathbf{Q}^{\frac{1}{2}} \mathbf{F} \vec{x}_0 + 2 \mathbf{R} \vec{\mathbf{u}}^{\text{opt}} = 0 \quad (14.167)$$

$$\mathbf{G}^T \mathbf{Q} \mathbf{G} \vec{\mathbf{u}}^{\text{opt}} + \mathbf{G}^T \mathbf{Q} \mathbf{F} \vec{x}_0 + \mathbf{R} \vec{\mathbf{u}}^{\text{opt}} = 0 \quad (14.168)$$

$$(\mathbf{G}^T \mathbf{Q} \mathbf{G} + \mathbf{R}) \vec{\mathbf{u}}^{\text{opt}} = -\mathbf{G}^T \mathbf{Q} \mathbf{F} \vec{x}_0 \quad (14.169)$$

$$\vec{\mathbf{u}}_{\text{opt}} = -(\mathbf{G}^T \mathbf{Q} \mathbf{G} + \mathbf{R})^{-1} \mathbf{G}^T \mathbf{Q} \mathbf{F} \vec{x}_0 \quad (14.170)$$

which provides a method for solving the entire problem at once instead of recursively through the dynamic Riccati equation.

First, it should be noted that the recursive solution may be faster computationally due to the large matrix inversions that are necessary in the LLS approach; however the **sparsity**, i.e. the number of 0 entries, in the **Q** and **R** matrices can make this inversion faster. Second, it should also be noted that the LLS solution can be adjusted for linear time-varying (LTV) systems, i.e.  $F$ ,  $G$ ,  $E$ ,  $Q$ , and  $R$  matrices vary with  $k$ , since this will only change the values within **F**, **G**, **Q** and **R**. Third, the relationship of LLS to QP allows for straightforward computation of the constrained discrete-time LQR through QP methods.

## References

For more information, please refer to the following

- Klein, V., and Morelli, E. G., “Chapter 5.1 Ordinary Least Squares,” in *Aircraft System Identification: Theory and Practice*, AIAA, 2006, pp. 97-131
- Klein, V., and Morelli, E. G., “Chapter 5.2 Generalized Least Squares,” in *Aircraft System Identification: Theory and Practice*, AIAA, 2006, pp.132-136
- Klein, V., and Morelli, E. G., “Chapter 5.3 Nonlinear Least Squares,” in *Aircraft System Identification: Theory and Practice*, AIAA, 2006, pp. 137
- Sarkka, S., “3.1 Batch linear regression,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 27-30
- Simon, D., “3 Least squares estimation,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 131

## 14.3 Bayesian Optimal Parameter Estimation

Another approach to the optimal parameter estimation problem is from a Bayesian inference perspective. Here one considers that  $\vec{y}$  are samples of a random vector  $\vec{Y}$  which depend on some *random* parameter vector,  $\vec{\beta}$ , which has taken some realization  $\vec{\beta}$ . In this case, one must consider the joint probabilities of  $\vec{Y}$  and  $\vec{\beta}$  which can be modeled by some joint PDF,  $f_{\vec{Y}, \vec{\beta}}(\vec{y}, \vec{\beta})$ . However, as one observes the realizations of  $\vec{Y}$ , one is particularly concerned with the **posterior PDF**, also known as the *a posteriori* PDF, i.e.,  $f_{\vec{\beta}|\vec{Y}}(\vec{\beta}|\vec{y})$ , which is related to the joint PDF by

$$f_{\vec{\beta}|\vec{Y}}(\vec{\beta}|\vec{y}) = \frac{f_{\vec{Y}, \vec{\beta}}(\vec{y}, \vec{\beta})}{f_{\vec{Y}}(\vec{y})} \quad (14.171)$$

which simply restates the definition of conditional probabilities. Furthermore, one can rewrite this equation as

$$f_{\vec{\beta}|\vec{Y}}(\vec{\beta}|\vec{y}) = \frac{f_{\vec{Y}|\vec{\beta}}(\vec{y}|\vec{\beta})f_{\vec{\beta}}(\vec{\beta})}{f_{\vec{Y}}(\vec{y})} \quad (14.172)$$

which, by the law of total probability, i.e.

$$f_{\vec{Y}}(\vec{y}) = \int f_{\vec{Y}|\vec{\beta}}(\vec{y}|\vec{\beta})f_{\vec{\beta}}(\vec{\beta})d\vec{\beta} \quad (14.173)$$

one has by Bayes' rule

$$f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y}) = \frac{f_{\vec{Y}|\vec{B}}(\vec{y}|\vec{\beta})f_{\vec{B}}(\vec{\beta})}{\int f_{\vec{Y}|\vec{B}}(\vec{y}|\vec{\beta})f_{\vec{B}}(\vec{\beta})d\vec{\beta}} \quad (14.174)$$

Thus, Bayesian parameter estimation can be seen as using three sources of information, the **prior PDF**, also known as the *a priori* PDF, of the parameter vector,  $f_{\vec{B}}(\vec{\beta})$ , the likelihood function,  $f_{\vec{Y}|\vec{B}}(\vec{y}|\vec{\beta})$  or  $\mathcal{L}(\vec{\beta}|\vec{y})$ , and the **model evidence**,  $f_{\vec{Y}}(\vec{y})$ , also known as the **marginal likelihood**.

### Bayes Optimal Parameter Estimators

A formal way to define Bayes optimality is the so-called **Bayes estimator (BE)** where one minimizes the **Bayes risk** defined as the posterior expected value of a selected loss function,  $\mathcal{J}$ , i.e., for the **Bayes parameter estimator**

$$\hat{\vec{\beta}}_{BE} = \underset{\hat{\vec{\beta}}}{\operatorname{argmin}} \mathbb{E}_{\vec{B}, \vec{Y}}[\mathcal{J}(\vec{\beta}, \hat{\vec{\beta}})] \quad (14.175)$$

where the expectation is taken over the joint distribution of  $\vec{B}$  and  $\vec{Y}$ . A common optimal Bayes parameter estimator is the **maximum *a posteriori* (MAP) estimator**, i.e., the mode of the posterior,

$$\hat{\vec{\beta}}_{MAP} = \underset{\hat{\vec{\beta}}}{\operatorname{argmax}} f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y}) \quad (14.176)$$

and can be shown to select the negative Dirac delta error for the loss function, i.e.,

$$\hat{\vec{\beta}}_{MAP} = \underset{\hat{\vec{\beta}}}{\operatorname{argmin}} \mathbb{E}_{\vec{B}, \vec{Y}}[-\delta(\vec{\beta} - \hat{\vec{\beta}})] \quad (14.177)$$

which by Bayes' rule can be rewritten as

$$\hat{\vec{\beta}}_{MAP} = \underset{\vec{\beta}}{\operatorname{argmax}} \frac{f_{\vec{Y}|\vec{B}}(\vec{y}|\vec{\beta})f_{\vec{B}}(\vec{\beta})}{f_{\vec{Y}}(\vec{y})} \quad (14.178)$$

which is equivalent to

$$\hat{\vec{\beta}}_{MAP} = \underset{\vec{\beta}}{\operatorname{argmax}} f_{\vec{Y}|\vec{B}}(\vec{y}|\vec{\beta})f_{\vec{B}}(\vec{\beta}) \quad (14.179)$$

v It should be noted if the prior PDF is uniform across all values of  $\vec{\beta}$ , i.e., the prior is **non-informative** of  $\vec{\beta}$ , then, the MAP estimator is equivalent to the MLE, i.e.,

$$\hat{\vec{\beta}}_{MLE} = \underset{\vec{\beta}}{\operatorname{argmax}} f_{\vec{Y}|\vec{B}}(\vec{y}|\vec{\beta}) \quad (14.180)$$

Another common Bayes estimator is the **minimum mean-square error (MMSE) estimator** which selects the squared error for the loss function, i.e.,

$$\hat{\vec{\beta}}_{MMSE} = \underset{\hat{\vec{\beta}}}{\operatorname{argmin}} \mathbb{E}_{\vec{B}, \vec{Y}}[(\hat{\vec{\beta}} - \vec{\beta})^T(\hat{\vec{\beta}} - \vec{\beta})] \quad (14.181)$$

and, for finite means and variances, is minimized by the mean of the posterior, i.e.,

$$\hat{\vec{\beta}}_{MMSE} = \mathbb{E}_{\vec{B}|\vec{Y}}[\vec{\beta}] = \int \cdots \int \vec{\beta} f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y}) d\vec{\beta} \quad (14.182)$$

Thus, the  $\hat{\vec{\beta}}_{MMSE} = \hat{\vec{\beta}}_{MAP}$ , if  $f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y})$  is elliptically symmetric about a single peak.

Lastly, the **median (Mdn) estimator** is

$$\hat{\vec{\beta}}_{Mdn} = \inf \left\{ \hat{\vec{\beta}} : \int_{\infty}^{\hat{\vec{\beta}}} f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y}) d\vec{\beta} \geq 0.5 \right\} \quad (14.183)$$

and can be shown to select the absolute value error for the loss function, i.e.,

$$\hat{\vec{\beta}}_{Mdn} = \underset{\hat{\vec{\beta}}}{\operatorname{argmin}} \mathbb{E}_{\vec{B}, \vec{Y}}[|\hat{\vec{\beta}} - \vec{\beta}|] \quad (14.184)$$

### Conjugate Prior Parameter Estimation

Given a particular likelihood model, one can obtain a closed-form expression for the posterior distribution if the posterior and the prior distributions belong in the same distribution family. In this case, the **conjugate prior** is defined for the specific likelihood function and the known distribution parameters of the conjugate prior and corresponding posterior are the **hyper-parameters** as opposed to the estimated parameter  $\vec{\beta}$ . In this context, one may also have  $N$  samples of  $\vec{Y}$ , i.e.,  $\vec{y}_1, \dots, \vec{y}_N$ . Four common likelihood functions with their conjugate priors and corresponding posteriors for aerospace vehicle perception systems are given here without proof.

For the likelihood function given by

$$f_{Y_i|B}(y_i|\beta) = f_{Pois}(y_i; \beta) \quad (14.185)$$

with unknown rate  $\beta$ , a conjugate prior PDF for the rate is

$$f_B(\beta) = f_{\Gamma}(\beta; \alpha, \theta) \quad (14.186)$$

with the corresponding posterior PDF for the rate as

$$f_{B|Y_1, \dots, Y_N}(\beta|y_1, \dots, y_N) = f_{\Gamma}(\beta; \alpha + \sum_{i=1}^N y_i, \theta + N) \quad (14.187)$$

For the likelihood function given by

$$f_{\vec{Y}_i|\vec{\beta}}(\vec{y}_i|\vec{\beta}) = f_{\mathcal{N}}(\vec{y}_i; \vec{\beta}, \Sigma_y) \quad (14.188)$$

and denoting the **sample mean** as

$$\vec{\mu}_y = N^{-1} \sum_{i=1}^N \vec{y}_i \quad (14.189)$$

one has

$$f_{\vec{M}_y|\vec{\beta}}(\vec{\mu}_y|\vec{\beta}) = f_{\mathcal{N}}(\vec{\mu}_y; \vec{\beta}, N^{-1}\Sigma_y) \quad (14.190)$$

with unknown mean,  $\vec{\beta}$ , and known covariance,  $\Sigma_y$ , a conjugate prior PDF for the mean is

$$f_{\vec{B}}(\vec{\beta}) = f_N(\beta; \vec{\mu}_0, \Sigma_0) \quad (14.191)$$

with the corresponding posterior PDF as

$$f_{\vec{B}|\vec{Y}_1, \dots, \vec{Y}_N}(\vec{\beta} | \vec{y}_1, \dots, \vec{y}_N) = f_N(\vec{\beta}; \vec{\mu}, \Sigma) \quad (14.192)$$

with

$$\Sigma = (\Sigma_0^{-1} + N\Sigma_y^{-1})^{-1} \quad (14.193)$$

and

$$\vec{\mu} = \Sigma \left( \Sigma_0^{-1} \vec{\mu}_0 + \Sigma_y^{-1} \vec{\mu}_y \right) \quad (14.194)$$

For the likelihood function given by

$$f_{\vec{Y}_i|B}(\vec{y}_i | \beta) = f_N(\vec{y}_i; \vec{\mu}, \beta) \quad (14.195)$$

and denoting the **sample covariance** as

$$\Sigma_y = \frac{1}{N-1} \sum_{i=1}^N (\vec{y}_i - \vec{\mu}_y)(\vec{y}_i - \vec{\mu}_y)^T \quad (14.196)$$

with known mean,  $\vec{\mu}$  and unknown covariance,  $\beta$ , a conjugate prior PDF for the covariance is

$$f_B(\beta) = f_{IW}(\beta; \Phi_0, v_0) \quad (14.197)$$

with the corresponding posterior PDF as

$$f_{B|\vec{Y}_1, \dots, \vec{Y}_N}(B | \vec{y}_1, \dots, \vec{y}_N) = f_{IW}(\beta; \Phi, v_0 + N) \quad (14.198)$$

with

$$\Phi = \Phi_0 + (N-1)\Sigma_y \quad (14.199)$$

For the likelihood function given by

$$f_{\vec{Y}_i|\vec{B}_\mu, B_\Sigma}(\vec{y}_i | \vec{\beta}_\mu, \beta_\Sigma) = f_N(\vec{y}_i; \vec{\beta}_\mu, \beta_\Sigma) \quad (14.200)$$

and if  $\vec{\mu}_y$  is the sample mean and  $\Sigma_y$  as the sample scatter matrix, with unknown mean,  $\vec{\beta}_\mu \in \mathbb{R}^{n_\beta}$  and unknown covariance,  $\beta_\Sigma \in \mathbb{R}^{n_\beta \times n_\beta}$ , a conjugate prior PDF is

$$f_{\vec{B}_\mu, B_\Sigma}(\vec{\beta}_\mu, \beta_\Sigma) = f_{NIW}(\vec{\beta}_\mu, \beta_\Sigma; \vec{\mu}_0, \lambda_0, \Phi_0, v_0) \quad (14.201)$$

with the corresponding posterior PDF as

$$f_{\vec{B}_\mu, B_\Sigma|\vec{Y}_1, \dots, \vec{Y}_N}(\vec{\beta}_\mu, \beta_\Sigma | \vec{y}_1, \dots, \vec{y}_N) = f_{NIW}(\vec{\beta}_\mu, \beta_\Sigma; \vec{\mu}, \lambda_0 + N, \Phi, v_0 + N) \quad (14.202)$$

with

$$\vec{\mu} = \frac{N}{\lambda_0 + N} \vec{\mu}_y + \frac{\lambda_0}{\lambda_0 + N} \vec{\mu}_0 \quad (14.203)$$

and

$$\Phi = \Phi_0 + (N - 1)\Sigma_y + \frac{\lambda_0 N}{\lambda_0 + N}(\vec{\mu}_y - \vec{\mu}_0)(\vec{\mu}_y - \vec{\mu}_0)^T \quad (14.204)$$

Related to this case is the univariate likelihood function given by

$$f_{Y_i|B}(y_i|\beta) = f_N(y_i; \beta_\mu, \beta_\sigma^2) \quad (14.205)$$

and denoting the sample variance as

$$\sigma_y^2 = \frac{1}{N - 1} \sum_{i=1}^N (y_i - \mu_y)^2 \quad (14.206)$$

with unknown mean,  $\beta_\mu$  and unknown variance,  $\beta_\sigma^2$ , a conjugate prior PDF is

$$f_{B_\mu, B_\sigma^2}(\beta_\mu, \beta_\sigma^2) = f_{N\chi^2}(\beta_\mu, \beta_\sigma^2; \mu_0, \lambda_0, \sigma_0^2, \nu_0) \quad (14.207)$$

with the corresponding posterior PDF as

$$f_{B_\mu, B_\sigma^2 | \vec{Y}_1, \dots, \vec{Y}_N}(\beta_\mu, \beta_\sigma^2 | \vec{y}_1, \dots, \vec{y}_N) = f_{N\chi^2}(\beta_\mu, \beta_\sigma^2; \mu, \lambda_0 + N, \sigma^2, \nu_0 + N) \quad (14.208)$$

with

$$\mu = \frac{N}{\lambda_0 + N} \vec{\mu}_y + \frac{\lambda_0}{\lambda_0 + N} \vec{\mu}_0 \quad (14.209)$$

and

$$\sigma^2 = \frac{\nu_0}{\nu_0 + N} \sigma_0^2 + \frac{N - 1}{\nu_0 + N} \sigma_y^2 + \frac{\lambda_0 N}{(\lambda_0 + N)(\nu_0 + N)} (\mu_y - \mu_0)^2 \quad (14.210)$$

### Bayesian Least-Squares Regression

The **Bayesian regression problem** considers a regression model as

$$\vec{y} = f(\vec{x}, \vec{\beta}) + \vec{\epsilon} \quad (14.211)$$

with the  $\mathbb{E}[\vec{\epsilon}] = 0$  and covariance  $\mathbb{E}[\vec{\epsilon}\vec{\epsilon}^T] = R$ , but also has some *a priori* information about the possible values of  $\vec{\beta}$  represented by the prior PDF defined as  $\vec{\beta}_0 \sim f_{\vec{B}}(\vec{\beta})$  and can be regarded as user-defined prior knowledge or due to previous parameter estimates. Due to this statistical nature, the objective in Bayesian least-squares regression is to minimize the sum of the estimator errors, i.e.

$$\hat{\vec{\beta}}_{BLS} = \underset{\hat{\vec{\beta}}}{\operatorname{argmin}} \mathbb{E}_{\vec{B}, \vec{Y}} \left[ (\vec{\beta} - \hat{\vec{\beta}})^T (\vec{\beta} - \hat{\vec{\beta}}) \right] \quad (14.212)$$

where  $\hat{\vec{\beta}}_{BLS}$  is the **Bayesian least-squares (BLS) parameter estimator** and is equivalent to the MMSE estimator.

Thus, for finite means and variances, one has

$$\hat{\vec{\beta}}_{BLS} = \mathbb{E}_{\vec{B} | \vec{Y}} [\vec{\beta}] \quad (14.213)$$

BLS is also known as **recursive least-squares** as this construction allows one to recursively process sequential samples,  $\vec{y}[k]$  with  $k = 1, 2, \dots$ , and recursively update the estimate  $\hat{\beta}[k]$ . This is opposed to previous least-squares methods which require “batch” processing of the data samples using stacked vectors and matrices. The **Bayesian linear regression problem** occurs when one has a linear-Gaussian regression model

$$\vec{y} = X\vec{\beta} + \vec{\epsilon} \quad (14.214)$$

with zero-mean multivariate Gaussian errors,  $\vec{\epsilon} \sim \mathcal{N}(0, R)$ . This analysis typically has two different cases,  $R$  is either known or unknown. Both cases typically utilize conjugate priors in their Bayes recursions.

### Bayesian Linear-Gaussian Regression

When  $R$  is known, one can specify a multivariate Gaussian conjugate prior PDF,  $\vec{\beta} \sim \mathcal{N}(\vec{\beta}; \vec{\mu}_0, \Sigma_0)$  and consider a transformed sample as  $\vec{z} = X^+ \vec{y}$  which produces a transformed regression model as

$$\vec{z} = \vec{\beta} + X^+ \vec{\epsilon} \quad (14.215)$$

and a transformed likelihood function given by

$$f_{\vec{Z}|\vec{B}}(\vec{z}|\vec{\beta}) = f_N(\vec{z}; \vec{\beta}, X^+ R X^{+T}) \quad (14.216)$$

with conjugate prior

$$f_{\vec{B}}(\vec{\beta}) = f_N(\vec{\beta}; \vec{\mu}_0, \Sigma_0) \quad (14.217)$$

with corresponding transformed posterior PDF as

$$f_{\vec{B}|\vec{Z}}(\vec{\beta}|\vec{z}) = f_N(\vec{\beta}; \hat{\vec{\beta}}, \Sigma_{\vec{\beta}}) \quad (14.218)$$

with mean

$$\hat{\vec{\beta}} = \Sigma_{\vec{\beta}} \left( X^T R^{-1} X \vec{z} + \Sigma_0^{-1} \vec{\mu}_0 \right) \quad (14.219)$$

and covariance

$$\Sigma_{\vec{\beta}} = X^T R^{-1} X + \Sigma_0^{-1} \quad (14.220)$$

where as  $X^+$ ,  $R$ , and  $X^{+T}$  are invertible, one has

$$(X^+ R X^{+T})^{-1} = X^T R^{-1} X \quad (14.221)$$

Thus, by substitution to the original sample model, i.e.,  $\vec{z} = X^+ \vec{y}$ , one has

$$f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y}) = f_N(\vec{\beta}; \hat{\vec{\beta}}, \Sigma_{\vec{\beta}}) \quad (14.222)$$

and

$$\hat{\vec{\beta}} = \Sigma_{\vec{\beta}} \Sigma_0^{-1} \vec{\mu}_0 + \Sigma_{\vec{\beta}} X^T R^{-1} \vec{y} \quad (14.223)$$

By the **matrix inversion lemma**, i.e.,

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}C)^{-1}DA^{-1} = A^{-1}B(C^{-1} + DA^{-1}C)^{-1}(BC)^{-1} \quad (14.224)$$

the covariance can be rewritten as

$$\Sigma_{\vec{\beta}} = (I - \Sigma_0 X^T (X \Sigma_0 X^T + R)^{-1} X) \Sigma_0 \quad (14.225)$$

or

$$\Sigma_{\vec{\beta}} = \Sigma_0 X^T (X \Sigma_0 X^T + R)^{-1} (X^T R^{-1})^{-1} \quad (14.226)$$

and the mean can be rewritten as

$$\hat{\vec{\beta}} = \left( I - \Sigma_0 X^T (X \Sigma_0 X^T + R)^{-1} X \right) \Sigma_0 \Sigma_0^{-1} \vec{\beta}_0 + \Sigma_0 X^T (X \Sigma_0 X^T + R)^{-1} (X^T R^{-1})^{-1} X^T R^{-1} \vec{y} \quad (14.227)$$

or

$$\hat{\vec{\beta}} = \vec{\beta}_0 - \Sigma_0 X^T (X \Sigma_0 X^T + R)^{-1} (\vec{y} - X \vec{\beta}_0) \quad (14.228)$$

which provides a recursive update to the prior mean,  $\vec{\beta}_0$ , through a gain on the residual,  $\vec{y} - X \hat{\vec{\beta}}_0$ . Thus, the **BLS parameter estimator** succinctly

$$\hat{\vec{\beta}}_{BLS} = \hat{\vec{\beta}}_0 + K(\vec{y} - X \hat{\vec{\beta}}_0) \quad (14.229)$$

where

$$K = \Sigma_0 X^T (X \Sigma_0 X^T + R)^{-1} \quad (14.230)$$

and

$$\Sigma_{\vec{\beta}} = (I - K X) \Sigma_0 \quad (14.231)$$

When  $R$  is unknown, one can specify a Gaussian-inverse-Wishart conjugate prior PDF,  $\vec{\beta} \sim NIW(\vec{\beta}, R; \vec{\mu}_0, \lambda_0, \Sigma_0, \nu_0)$  and consider a transformed sample as  $\vec{z} = X^+ \vec{y}$  which produces a transformed regression model as

$$\vec{z} = \vec{\beta} + X^+ \vec{\epsilon} \quad (14.232)$$

and a transformed likelihood function given by

$$f_{\vec{Z}|\vec{\beta}_{\mu}, \Sigma_B}(\vec{z}|\vec{\beta}, \beta_{\Sigma}) = f_N(\vec{z}; \vec{\beta}_{\mu}, X^+ R X^{+T}) \quad (14.233)$$

with conjugate prior

$$f_{\vec{\beta}, \Sigma_B}(\vec{\beta}, R) = f_{NIW}(\vec{\beta}, R; \vec{\mu}_0, \lambda_0, \Sigma_0, \nu_0) \quad (14.234)$$

with corresponding transformed posterior PDF as

$$f_{\vec{\beta}_{\mu}, \Sigma_B | \vec{Z}}(\vec{\beta}, R | \vec{z}) = f_{NIW}(\vec{\beta}, R; \vec{\mu}, \lambda_0 + 1, \Phi, \nu_0 + 1) \quad (14.235)$$

with

$$\vec{\mu} = \frac{1}{\lambda_0 + 1} \vec{z} + \frac{\lambda_0}{\lambda_0 + 1} \vec{\mu}_0 \quad (14.236)$$

and

$$\Phi = \Phi_0 + \frac{\lambda_0}{\lambda_0 + 1} (\vec{z} - \vec{\mu}_0)(\vec{z} - \vec{\mu}_0)^T \quad (14.237)$$

Thus, by substitution to the original sample model, i.e.,  $\vec{z} = X^+ \vec{y}$ , one has

$$f_{\vec{\beta}_{\mu}, \Sigma_B | \vec{Y}}(\vec{\beta}, R | \vec{y}) = f_{NIW}(\vec{\beta}, R; \vec{\mu}, \lambda_0 + 1, \Phi, \nu_0 + 1) \quad (14.238)$$

with mean

$$\vec{\mu} = \frac{1}{\lambda_0 + 1} (X^T X)^{-1} X \vec{y} + \frac{\lambda_0}{\lambda_0 + 1} \vec{\mu}_0 \quad (14.239)$$

and scale

$$\Phi = \Phi_0 + \frac{\lambda_0}{\lambda_0 + 1} ((X^T X)^{-1} X \vec{y} - \vec{\mu}_0) ((X^T X)^{-1} X \vec{y} - \vec{\mu}_0)^T \quad (14.240)$$

### MAP Parameter Estimation

If one is only interested in the MAP parameter estimate for the posterior, one of the simplest approximations may be to utilize an iterative search algorithm to find nearest mode or several modes, i.e., local maximum values, of the posterior PDF without the need to calculate the posterior PDF explicitly. In this case, one can utilize a mode-finding algorithm which attempts to maximize a quadratic approximation to the log posterior density, i.e.,

$$L(\vec{\beta} + \Delta \vec{\beta}) = \log f_{\vec{B}|\vec{Y}}(\vec{\beta} + \Delta \vec{\beta} | \vec{y}) \approx L(\vec{\beta}) + \vec{h}^T(\vec{\beta}) \Delta \vec{\beta} + \frac{1}{2} \Delta \vec{\beta}^T H(\vec{\beta}) \Delta \vec{\beta} \quad (14.241)$$

where  $\vec{h}^T(\vec{\beta})$  and  $H(\vec{\beta})$  are theoretically the gradient,  $\nabla L(\vec{\beta})$  and the Hessian,  $\frac{\partial^2 L(\vec{\beta})}{\partial \vec{\beta} \partial \vec{\beta}^T}$ , but may be approximations depending on the algorithm. The **Newton-Raphson algorithm** chooses  $\vec{g}^T = \nabla L(\vec{\beta})$  and  $H = \frac{\partial L(\vec{\beta})}{\partial \vec{\beta}}$ . Quasi-newton algorithms, e.g., the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, approximate the Hessian. Finite difference methods approximate both the gradient and Hessian, e.g.,

$$g_i(\vec{\beta}) = \frac{L(\vec{\beta} + \delta_i \vec{1}_i) - L(\vec{\beta} - \delta_i \vec{1}_i)}{2\delta_i} \quad (14.242)$$

and

$$H_{i,j}(\vec{\beta}) = \frac{L(\vec{\beta} + \delta_i \vec{1}_i + \delta_j \vec{1}_j) - L(\vec{\beta} - \delta_i \vec{1}_i + \delta_j \vec{1}_j) - L(\vec{\beta} + \delta_i \vec{1}_i - \delta_j \vec{1}_j) + L(\vec{\beta} - \delta_i \vec{1}_i - \delta_j \vec{1}_j)}{4\delta_i \delta_j} \quad (14.243)$$

where  $\delta_i \ll 1$  and  $\vec{1}_i$  is the unit vector in the direction of  $\vec{\beta}_i$ .

The **mode-finding algorithm** can be written as

(1) Initialization:

(a) Choose initial value  $\hat{\vec{\beta}}(0)$

(b) Choose convergence limit  $\delta$

(c) Set  $\frac{\|\vec{\Delta}(t)\|}{\|\hat{\vec{\beta}}(t)\|} < \infty$

(2) While  $\frac{\|\vec{\Delta}(t)\|}{\|\hat{\vec{\beta}}(t)\|} > \delta$  and  $t \leq T$ :

(a) Compute  $A$  and  $B$

(b) Compute  $B(\vec{\beta})$

- (c) Set  $\hat{\vec{\beta}}(t+1) = \hat{\vec{\beta}}(t) - B^{-1}(\vec{\beta})\vec{a}(\vec{\beta})$
  - (d) Set  $t = t + 1$
- (3) Select mode of  $f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y})$  as

$$\hat{\vec{\beta}}_{mode} = \hat{\vec{\beta}}(t) \quad (14.244)$$

Repeating this mode-finding algorithm, one may generally find  $K$  modes and approximate the MAP parameter estimator as

$$\hat{\vec{\beta}}_{MAP} \approx \max_{t=1,\dots,K} \hat{\vec{\beta}}_{mode}(t) \quad (14.245)$$

After estimating  $K$  modes, one can also approximate the posterior distribution itself by a finite mixture of distributions based on each mode. The most common is the **Gaussian-mixture (GM) approximation**

$$f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y}) = \sum_{k=1}^K w_k f_N(\vec{\beta}; \hat{\vec{\beta}}_{mode}(k), V_{\vec{\beta}}) \quad (14.246)$$

where

$$V_{\vec{\beta}} = \left( -\frac{\partial^2 L(\vec{\beta})}{\partial \vec{\beta} \partial \vec{\beta}^T} \right)^{-1} \quad (14.247)$$

and enforcing

$$\sum_{k=1}^K w_k = 1 \quad (14.248)$$

The weights,  $w_k$ , are estimated based on matching  $f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y})$  or its un-normalized approximation at all  $\vec{\beta}(k)$ . An alternative for heavy-tailed distributions is the  $t$ -mixture approximation.

The mixture approximation may not be appropriate for all elements of the parameter vector. Thus, one can supplement the mixture approximation approach with an expectation-maximization (EM) algorithm which iterates between an expectation step (E-step) and a maximization step (M-step). This approach notably utilizes a partitioning of the parameter vector,  $\vec{\beta}$ , into two sub-parameter vectors, a modeled  $\vec{\gamma}$  and an unmodeled  $\vec{\phi}$ , i.e.,

$$\vec{\beta} = \begin{bmatrix} \vec{\gamma} \\ \vec{\phi} \end{bmatrix} \quad (14.249)$$

such that the analytical approximation works well for  $f_{\vec{\Gamma}, \vec{\Phi}|\vec{Y}}(\vec{\gamma}, \vec{\phi}|\vec{y})$ . Now consider the joint distribution of  $\vec{\beta}$  as

$$f_{\vec{\Gamma}, \vec{\Phi}|\vec{Y}}(\vec{\gamma}, \vec{\phi}|\vec{y}) = f_{\vec{\Gamma}|\vec{\Phi}, \vec{Y}}(\vec{\gamma}|\vec{\phi}, \vec{y}) f_{\vec{\Phi}|\vec{Y}}(\vec{\phi}|\vec{y}) \quad (14.250)$$

Next, taking the logarithm of both sides and rearranging, one has

$$\log f_{\vec{\Phi}|\vec{Y}}(\vec{\phi}|\vec{y}) = f_{\vec{\Gamma}, \vec{\Phi}|\vec{Y}}(\vec{\gamma}, \vec{\phi}|\vec{y}) - \log f_{\vec{\Gamma}|\vec{\Phi}, \vec{Y}}(\vec{\gamma}|\vec{\phi}, \vec{y}) \quad (14.251)$$

Then, by taking the expectation with respect to the sub-parameter vector  $\vec{\gamma}$  under the distribution  $f_{\vec{\Gamma}|\vec{\Phi}(t)\vec{Y}}(\vec{\gamma}|\vec{\phi}(t), \vec{y})$ , i.e., averaging over all values of  $\vec{\gamma}$ , one has

$$\log f_{\vec{\Phi}|\vec{Y}}(\vec{\phi}|\vec{y}) = \mathbb{E}_t \left[ \log f_{\vec{\Gamma}, \vec{\Phi}|\vec{Y}}(\vec{\gamma}, \vec{\phi}|\vec{y}) \right] - \mathbb{E} \left[ \log f_{\vec{\Gamma}|\vec{\Phi}, \vec{Y}}(\vec{\gamma}|\vec{\phi}, \vec{y}) \right] \quad (14.252)$$

where the second term on the right side is maximum at  $\vec{\phi}(t)$  and the first term on the right side could be maximized for a new  $\vec{\phi}(t+1)$ . This would increase the marginal PDF  $f_{\vec{\Phi}|\vec{Y}}(\vec{\phi}|\vec{y})$  at each iteration which guarantees the EM algorithm will converge to a local mode of the posterior PDF except in some special cases.

The **expectation-maximization (EM) algorithm** can be written as

(1) Initialization:

- (a) Choose initial value  $\hat{\vec{\phi}}(0)$
- (b) Choose convergence limit  $\delta$
- (c) Choose total number of iterations,  $T$

(2) While  $\frac{\|\hat{\vec{\phi}}(t) - \hat{\vec{\phi}}(t-1)\|}{\|\hat{\vec{\phi}}(t)\|} > \delta$  and  $t \leq T$ :

- (a) *E-step*: Compute

$$\mathbb{E}_t \left[ \log f_{\vec{\Gamma}, \vec{\Phi}|\vec{Y}}(\vec{\gamma}, \vec{\phi}|\vec{y}) \right] = \int \cdots \int \log \left( f_{\vec{\Gamma}, \vec{\Phi}|\vec{Y}}(\vec{\gamma}, \vec{\phi}|\vec{y}) \right) f_{\vec{\Gamma}|\vec{\Phi}, \vec{Y}}(\vec{\gamma}|\vec{\phi}(t), \vec{y}) d\vec{\gamma} \quad (14.253)$$

- (b) *M-step*: Compute

$$\hat{\vec{\phi}}(t+1) = \underset{\hat{\vec{\phi}}}{\operatorname{argmax}} \mathbb{E} \left[ \log f_{\vec{\Gamma}, \vec{\Phi}|\vec{Y}}(\vec{\gamma}, \vec{\phi}|\vec{y}) \right] \quad (14.254)$$

- (c) Set  $t = t + 1$

(3) Select EM mode estimates as

$$\hat{\vec{\phi}}_{mode} = \hat{\vec{\phi}}(t) \quad (14.255)$$

Repeating this mode-finding algorithm, one may generally find  $K$  modes and approximate the MAP parameter estimator as

$$\hat{\vec{\phi}}_{MAP} \approx \max_{t=1, \dots, K} \hat{\vec{\phi}}_{mode}(t) \quad (14.256)$$

The “hyper-parameters,”  $\vec{\phi}$ , maximization of EM as a point estimate of the MAP can be extended to **variational Bayes** methods which analytically approximate the posterior PDF by an alternate parameterized variational distribution,  $\bar{q}(\vec{\beta}|\vec{\phi})$ , with selected hyper-parameters,  $\vec{\phi}$ . This approximation is iteratively chosen such the the K-L divergence is minimized, i.e.,

$$\vec{\phi}_{opt} = \underset{\vec{\phi}}{\operatorname{argmin}} KL(\bar{q}||f_{\vec{B}|\vec{Y}}) = \underset{\vec{\phi}}{\operatorname{argmin}} - \int \cdots \int \bar{q}(\vec{\beta}|\vec{\phi}) \log \left( \frac{f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y})}{\bar{q}(\vec{\beta}|\vec{\phi})} \right) d\vec{\beta} \quad (14.257)$$

Then, the **VB parameter estimator** is given by

$$\hat{\vec{\beta}}_{VB} = \int \cdots \int g(\vec{\beta}) \bar{q}(\vec{\beta} | \vec{\phi}) d\vec{\beta} \quad (14.258)$$

For tractability, one typically assumes the variational distribution has  $M$  independent parameters, i.e.,

$$\bar{q}(\vec{\beta} | \vec{\phi}) = \prod_{j=1}^M \bar{q}_j(\beta_j | \phi_j) \quad (14.259)$$

and

$$\bar{q}_{-j}(\vec{\beta}_{-j} | \vec{\phi}_{-j}) = \bar{q}_1(\beta_1 | \phi_1) \cdots \bar{q}_{j-1}(\beta_{j-1} | \phi_{j-1}) \bar{q}_{j+1}(\beta_{j+1} | \phi_{j+1}) \cdots \bar{q}_M(\beta_M | \phi_M) \quad (14.260)$$

Then, one iteration of variational Bayes approximation updates the hyper-parameters,  $\vec{\phi}$ , so that

$$\log(\bar{q}_j(\beta_j | \phi_j)) = \int \log(f_{\vec{B} | \vec{Y}}(\vec{\beta} | \vec{y})) \bar{q}_{-j}(\vec{\beta}_{-j} | \vec{\phi}_{-j}) d\vec{\beta}_{-j} \quad (14.261)$$

which is similar to Gibbs sampling shown later except the update uses the average over the other parameters instead of the conditional PDF. This method can be shown by calculus of variations, hence the name, that this iteration converges to  $\vec{\phi}_{opt}$ . Because of the use of the logarithm of the posterior, variational Bayes works best on exponential family models with conditionally conjugate prior distributions in which case approximating variational distribution can be determined by inspection and the necessary expectations performed in closed-form.

### Numerical Integration for Parameter Estimation

In many cases, no analytical solutions exist for computing the integrals for the Bayes posterior, its modes, and/or the Bayes estimator using the expectation integral of the posterior, e.g.,

$$\hat{\vec{\beta}}_{BE} = \mathbb{E}_{\vec{B} | \vec{Y}}[g(\vec{\beta})] = \int \cdots \int g(\vec{\beta}) f_{\vec{B} | \vec{Y}}(\vec{\beta} | \vec{y}) d\vec{\beta} \quad (14.262)$$

where  $g(\vec{\beta})$  is some chosen statistic of the posterior. One approach is to approximate the posterior by an analytical approximation that provides closed-form integration, e.g., Laplace's method, or to optimally fit the posterior to a model, e.g., Variational Bayes. An alternative is to employ **numerical integration**, also known as **quadrature** methods, in which the integral is evaluated at discrete points. Numerical integration methods are divided between deterministic quadrature rule methods or stochastic Monte Carlo methods.

**Quadrature rule (QR) methods** numerically approximate the expectation integral by a weighted sum of  $M$  specified **integration-points** in the parameter space which forms the **QR parameter estimator** by

$$\hat{\vec{\beta}}_{QR} \approx \operatorname{argmin}_{\vec{\beta}} \sum_{m=1}^M w_m g(\vec{\beta}(m)) \quad (14.263)$$

where  $\vec{\beta}(m)$  is the  $m^{\text{th}}$  integration-point with corresponding weight,  $w_m$ . The choice of weights that provide a good approximation are intrinsically related to the posterior,  $f_{\vec{B}(m) | \vec{Y}}(\vec{\beta}(m) | \vec{y})$ , which may be unknown.

A typical use of QR methods is to use an iterative approach to selecting the weights or to assume a form for  $f_{\vec{\beta}(m)|\vec{Y}}(\vec{\beta}(m)|\vec{y})$ , e.g., a multivariate Gaussian with mean  $\vec{\mu}$  and covariance  $\Sigma$ . In the Gaussian case, one may use transform the integration-points to unit **sigma-points**,  $\vec{\xi}(m)$ , and corresponding weights  $w_m$  through **stochastic decoupling**, i.e.

$$\hat{\vec{\beta}}_{SP} \approx \underset{\vec{\beta}}{\operatorname{argmin}} \sum_{m=1}^M w_m g(\vec{\mu} + \Sigma^{1/2} \vec{\xi}(m)) \quad (14.264)$$

There are different methods for constructing these weights and sigma-points, e.g., the central difference, unscented transform, cubature transform, Gauss-Hermite rule, Gaussian process quadratures, etc. These methods are discussed in a following chapter on nonlinear Bayesian state estimation for which sigma-points are considered as different statistical linear regressions of statistical linearization.

**Monte Carlo (MC) methods** numerically approximate the expectation integral by a weighted sum of  $M$  IID samples of  $\vec{\beta} \sim f_{\vec{\beta}|\vec{Y}}(\vec{\beta}|\vec{y})$ , i.e.,

$$\hat{\vec{\beta}}_{MC} \approx \underset{\vec{\beta}}{\operatorname{argmin}} \frac{1}{M} \sum_{m=1}^M \mathcal{J}(\vec{\beta}(m), \hat{\vec{\beta}}) \quad (14.265)$$

where  $\vec{\beta}(m)$  is the  $m^{\text{th}}$  sample. By the law of large numbers, one can show that  $\hat{\vec{\beta}}^{MC} \rightarrow \hat{\vec{\beta}}_{BE}$  as  $M \rightarrow \infty$ . Unfortunately, it is also typically the case that one cannot draw samples directly from  $f_{\vec{\beta}|\vec{Y}}(\vec{\beta}|\vec{y})$ , so one must perform point-wise evaluations of the **target function**

$$\pi(\vec{\beta}) = f_{\vec{Y}|\vec{\beta}}(\vec{y}|\vec{\beta}) f_{\vec{\beta}}(\vec{\beta}) \quad (14.266)$$

which is notably the numerator of Bayes' Rule for the posterior PDF, but is not a proper PDF.

These MC methods all fall under three categories: rejection sampling, Markov Chain, and importance sampling, which will be summarized here. In all three cases, one uses one or more **proposal functions**,  $q(\vec{\beta})$ , or the related **proposal PDF**,  $\bar{q}(\vec{\beta})$ , where

$$\bar{q}(\vec{\beta}) = \frac{q(\vec{\beta})}{f_{\vec{Y}}(\vec{y})} \quad (14.267)$$

to generate samples from any posterior PDF,  $f_{\vec{\beta}|\vec{Y}}(\vec{\beta}|\vec{y})$ .

**Rejection sampling Monte Carlo (RSMC)** methods draw samples from a proposal PDF,  $q(\vec{\beta})$ , instead of  $f_{\vec{\beta}|\vec{Y}}(\vec{\beta}|\vec{y})$ , then the sample is either accepted or rejected based on a ratio of the two PDFs from which it can be proven that the accepted samples are then distributed according the posterior PDF. For such ratios, one must use an envelope function, i.e.  $Cq(\vec{\beta}) \geq \pi(\vec{\beta})$  for all possible  $\vec{\beta}$  where  $C$  is some constant. However, it should be noted that finding such a  $C$  for arbitrary  $f_{\vec{\beta}|\vec{Y}}(\vec{\beta}|\vec{y})$  may be difficult and the acceptance rate may be small for a large portion of the parameter space, and the number of iterations required for completion is not known *a priori* and make the RSMC very inefficient.

The RSMC algorithm can be written as

- (1) Initialization:

- (a) Choose proposal function  $q(\vec{\beta})$
  - (b) Choose required number of samples for posterior estimate:  $M$
  - (c) Find upper bound,  $C \geq \frac{\pi(\vec{\beta})}{q(\vec{\beta})}$  for all possible  $\vec{\beta}$
  - (d) Let  $t = m = 1$
- (2) While  $m < M$ :
- (a) Draw sample from  $\vec{\beta}(t) \sim \tilde{q}(\vec{\beta})$
  - (b) Draw sample from  $u \sim \mathcal{U}(0, 1)$
  - (c) If  $u \leq \frac{\pi(\vec{\beta}(t))}{Cq(\vec{\beta})(t)}$ 
    - (i) Set  $\vec{\beta}(m) = \vec{\beta}(t)$
    - (ii) Set  $m = m + 1$
  - (d) Set  $t = t + 1$ , regardless of (c)
- (3) Compute the **RSMC parameter estimator** by

|3-i

$$\hat{\vec{\beta}}^{RSMC} = \frac{1}{M} \sum_{m=1}^M g(\vec{\beta}(m)) \quad (14.268)$$

**Markov Chain Monte Carlo (MCMC)** methods produce an Markov Chain whose stationary density is the desired posterior PDF,  $f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y})$ , of which the most common are the Gibbs sampler and the Metropolis-Hastings algorithm (MHA). The **Gibbs sampler**, also known as **alternate conditional sampling**, which assumes one can choose  $M \leq n_\beta$  sub-vectors of  $\vec{\beta}$  such that one has well-defined conditional posterior PDFs for the individual sub-vectors of  $\vec{\beta}$ , i.e., for  $j = 1, \dots, M$ ,

$$q_j(\vec{\beta}) = f_{\vec{B}_j|\vec{B}_{-j}(t-1), \vec{Y}}(\vec{\beta}_j | \vec{\beta}_{-j}(t-1), \vec{y}) \quad (14.269)$$

where

$$\vec{\beta}_{-j}(t-1) = \begin{bmatrix} \vec{\beta}_1(t) \\ \vdots \\ \vec{\beta}_{1j-1}(t) \\ \vec{\beta}_{j+1}(t) \\ \vdots \\ \vec{\beta}_M(t-1) \end{bmatrix} \quad (14.270)$$

Then, by performing  $M$  sampling steps for each iteration  $t$  of the sampling process, it can be shown that one will converge to the sampling from  $f_{\vec{B}|\vec{Y}}(\vec{\beta}|\vec{y})$  as  $t \rightarrow \infty$ . This is especially useful for known mixture distributions as the true posterior distribution.

The MHA can be considered as a generalized rejection sampler whose proposal PDF depends on the result of the previous iteration, i.e., on  $\vec{\beta}(t-1)$ . Moreover, as the acceptance rate depends on  $\vec{\beta}(t-1)$ , the value is re-used whenever a candidate sample is rejected, thus the samples are no longer IID, but as the

underlying Markov Chain has the desired posterior as its limiting stationary distribution, this resampling is done with the right acceptance probability to ensure its validity. The Gibbs sampler can also be seen as a special case of the MHA where every sample is accepted due to the construction of the proposal function via  $M$  steps of sampling the  $M$  subvectors at each iteration,  $t$ .

The MHA-MCMC parameter estimator can be outlined as

(1) Initialization:

- (a) Choose proposal function  $q(\vec{\beta}^* | \vec{\beta}(t-1))$
- (b) Choose initial state  $\vec{\beta}(0)$
- (c) Choose total number of iterations,  $T$
- (d) Choose burn-in period,  $T_b$

(2) For  $t = 1, \dots, T$ :

- (a) Draw sample from  $\vec{\beta}^* \sim q(\vec{\beta}^* | \vec{\beta}(t-1))$
- (b) Draw sample from  $u \sim \mathcal{U}(0, 1)$
- (c) Compute acceptance probability:

$$\alpha(\vec{\beta}^*, \vec{\beta}(t)) = \min \left[ 1, \frac{\pi(\vec{\beta}^*) q(\vec{\beta}(t-1) | \vec{\beta}^*)}{\pi(\vec{\beta}(t-1)) q(\vec{\beta}^* | \vec{\beta}(t-1))} \right] \quad (14.271)$$

(d) If  $u \leq \alpha(\vec{\beta}(t), \vec{\beta}(t-1))$

(i) Set  $\vec{\beta}(t) = \vec{\beta}^*$

(e) Otherwise

(i) Set  $\vec{\beta}(t) = \vec{\beta}(t-1)$

(3) Compute the **MHA-MCMC parameter estimator** by

$\hat{\vec{\beta}}_{MHA-MCMC}$

$$\hat{\vec{\beta}}_{MHA-MCMC} = \frac{1}{T - T_b} \sum_{t=T_b+1}^T g(\vec{\beta}(t)) \quad (14.272)$$

**Importance Sampling Monte Carlo (ISMC)** methods draw samples from one or more proposal PDFs,  $\bar{q}(\vec{\beta})$ , however, ISMC methods accept all samples, but assign each sample a weight according to their quality in approximating the desired posterior PDF. The primary division between methods being the single or multiple proposals used in ISMC method. ISMC methods may also be non-adaptive or adaptive where the parameters of proposal(s) are adapted iteratively to better approximate the desired posterior PDF.

Thus, the standard, non-adaptive ISMC algorithm can be written as

- (1) Choose single proposal function  $\bar{q}(\vec{\beta})$
- (2) Draw  $M$  samples from  $\vec{\beta}(m) \sim \bar{q}(\vec{\beta})$  for  $m = 1, \dots, M$

(3) Compute  $M$  importance weights:

$$w(m) = \frac{\pi(\vec{\beta}(m))}{\bar{q}(\vec{\beta})(m)} \quad (14.273)$$

(4) Estimate marginal likelihood by

$$\hat{f}_{\vec{Y}} = \frac{1}{M} \sum_{m=1}^M w(m) \quad (14.274)$$

(5) Compute the **ISM**C parameter estimator by

$$\hat{\vec{\beta}}_{ISM} = \frac{1}{M \hat{f}_{\vec{Y}}} \sum_{m=1}^M w(m) g(\vec{\beta}(m)) \quad (14.275)$$

## References

For more information, please refer to the following

- Gopalan, P., Hofman, J. M. and Blei, D. M., “Scalable recommendation with hierarchical Poisson factorization,” in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI’15)*, AUAI Press, 2015, pp. 326–335
- Gelman, A., Carlin, J., Stern, H., and Rubin, D., “3.3 Normal with a conjugate prior distribution,” in *Bayesian Data Analysis*, 3rd Ed., Chapman and Hall, 2004, 2nd edition, pp. 67-69
- Gelman, A., Carlin, J., Stern, H., and Rubin, D., “3.5 Multivariate normal with known variance,” in *Bayesian Data Analysis*, 3rd Ed., Chapman and Hall, 2004, 2nd edition, pp. 71-72
- Gelman, A., Carlin, J., Stern, H., and Rubin, D., “3.6 Multivariate normal with unknown mean and variance,” in *Bayesian Data Analysis*, 3rd Ed., Chapman and Hall, 2004, pp. 72-74
- Gelman, A., Carlin, J., Stern, H., and Rubin, D., “10.1 Numerical integration,” in *Bayesian Data Analysis*, 3rd Ed., Chapman and Hall, 2004, pp. 72-74
- Gelman, A., Carlin, J., Stern, H., and Rubin, D., “10.3 Direct simulation and rejection sampling,” in *Bayesian Data Analysis*, 3rd Ed., Chapman and Hall, 2004, pp. 263-265
- Gelman, A., Carlin, J., Stern, H., and Rubin, D., “10.4 Importance sampling,” in *Bayesian Data Analysis*, 3rd Ed., Chapman and Hall, 2004, pp. 265-267
- Gelman, A., Carlin, J., Stern, H., and Rubin, D., “11.1 Gibbs sampler,” in *Bayesian Data Analysis*, 3rd Ed., Chapman and Hall, 2004, pp. 276-278

- Gelman, A., Carlin, J., Stern, H., and Rubin, D., “11.2 Metropolis and Metropolis-Hastings algorithms,” in *Bayesian Data Analysis*, 3rd Ed., Chapman and Hall, 2004, pp. 278-280
- Gelman, A., Carlin, J., Stern, H., and Rubin, D., “13.7 Variational inference,” in *Bayesian Data Analysis*, 3rd Ed., Chapman and Hall, 2004, pp. 331-338
- Sarkka, S., “2 Bayesian inference,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 17-26
- Sarkka, S., “3.2 Recursive linear regression,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 30-33
- Sarkka, S., “7.1 Monte Carlo approximations in Bayesian inference,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 116-117
- Sarkka, S., “7.2 Importance Sampling,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 117-120
- Simon, D., “3.3 Recursive least squares estimation,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 131

## 14.4 Hypothesis Testing and Optimal Detection Theory

A **statistical comparison** problem is a type of statistical inference problem where the candidate set contains a finite number of candidate models known as **hypotheses**,  $\mathcal{H}_i$ ,  $i = 0, \dots, M$ , then one has a between hypotheses, i.e., one has a **hypothesis testing problem** to select a single optimal hypothesis.  $\mathcal{H}_0$  is known as the **null hypothesis**, and  $\mathcal{H}_i$  for  $i = 1, \dots, M$  are the **alternate hypotheses**. To perform the most general hypothesis test, one observes data,  $\vec{y} \in \mathcal{Y}$ , and selects the hypothesis,  $\mathcal{H}_i$ , for which the observed data falls within its corresponding decision region,  $\mathcal{R}_i$ . Notably,  $R_i$ 's must be disjoint and form a complete partition of the observation space, i.e.,  $\bigcap_{i=0}^M \mathcal{R}_i = \emptyset$  and  $\bigcup_{i=0}^M \mathcal{R}_i = \mathcal{Y}$ .

For fully parametric hypothesis testing where the conditional PDFs/likelihoods are known for each hypothesis, i.e.,  $f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_i) = \mathcal{L}(\mathcal{H}_i|\vec{y})$ , the probabilities of correct selection of the  $i^{\text{th}}$  hypothesis,  $\gamma_i$ , known as the **confidence level** of the hypothesis, can be obtained by integrating the likelihood of that particular hypothesis over a portion of the  $\mathcal{Y}$  corresponding to its decision region, i.e.,

$$\gamma_i = \int_{R_i} \mathcal{L}(\mathcal{H}_i|\vec{y}) d\vec{y} \quad (14.276)$$

while the probability of erroneous selection of the  $i^{\text{th}}$  hypothesis,  $\alpha_i$ , known as the **significance level** of the hypothesis, is found by integrating over a portion of the  $\mathcal{Y}$  corresponding to the other decision regions, i.e.,

$$\alpha_i = 1 - \gamma_i = \sum_{i \in 0, \dots, M, i \neq j} \int_{R_i} \mathcal{L}(\mathcal{H}_i|\vec{y}) d\vec{y} \quad (14.277)$$

While these equations hold generally, most hypothesis tests compute a scalar **test statistic**,  $T(\vec{y}) \in \mathbb{R}$ , which is a function of the observed data. Here, the decision regions of  $\mathcal{R}_i \in \mathcal{Y}$  are mapped to disjoint test statistic decision regions in  $\mathbb{R}$ .

Parametric hypothesis testing problems can be classified by two primary characteristics, the number of known models and knowledge of the model parameters. A **null hypothesis test**, also known as a **goodness-of-fit test**, considers a known distribution of the test statistic under only the null hypothesis which results in choosing confidence and significance levels of only the null hypothesis being chosen. A **binary hypothesis test** considers known distributions of the test statistic under the null and one other alternative hypothesis while a  **$M$ -ary hypothesis test**, also known as a **discrimination problem** or **classification problem**, considers known distributions of the test statistic under the null and multiple alternative hypotheses. A **simple hypothesis test** considers hypotheses that are parameterized by completely known parameters,  $\vec{\beta}_i \in \vec{\mathcal{B}}_i$ , i.e., completely known conditional PDFs/likelihoods  $f_{\vec{Y}|\mathcal{H}}(\vec{y}|\vec{\beta}_i) = \mathcal{L}(\vec{\beta}_i|\vec{y})$ , while a **composite hypothesis test** considers hypotheses that are parameterized by completely or partially unknown parameters,  $\vec{\beta}_i \in \vec{\mathcal{B}}_i$ , i.e., conditional PDFs/likelihoods  $f_{\vec{Y}|\mathcal{H}}(\vec{y}|\vec{\beta}_i) = \mathcal{L}(\vec{\beta}_i|\vec{y})$ . Notably, the parameter spaces for each hypothesis,  $\vec{\mathcal{B}}_i$ , may be different. Naturally, composite hypothesis tests may also involve parameter estimation which is the subject of the following chapter in this textbook.

**Signal detection theory** is a particular version of binary and  $M$ -ary hypothesis testing theory that involves the particular decision of declaring if a *signal*,  $\vec{s}$ , i.e., an information-bearing pattern, has occurred within observed data,  $\vec{y}$ , that is corrupted by *noise*,  $\vec{n}$ , i.e., random patterns that distract from the potential information. In the general case, the null hypothesis is defined as the event of *no signal* occurring and the  $M$  alternative hypotheses are the events of a particular signal occurring from a set of  $M$  possible signals. Thus, signal detection testing is a particular “engineering” framework for working with multiple hypothesis testing. In binary detection problems, the test statistic is known as the **detector** while in  $M$ -ary detection problems, the test statistic is known as the **discriminator**.

In binary signal detection problems, a “detection” event occurs when one declares a signal occurred. The probability of a correct detection,  $\mathbb{P}_D$ , is called the **power** of the detector. A **false alarm** event, also known as a **type I error** in hypothesis testing, occurs when one declares a signal occurred when no signal actually occurred and has a probability denoted as  $\alpha = \mathbb{P}_{FA}$ . A **missed detection** event, also known as a **type II error** in hypothesis testing, occurs one declares no signal occurred when a signal did actually occur and has a probability denoted as  $\beta = \mathbb{P}_{MD} = 1 - \mathbb{P}_D$ . Here, one partitions the detector space into two with a single **detection threshold**,  $\tau$ , whose value and its associated detector is chosen to balance  $\mathbb{P}_D$  with  $\mathbb{P}_{FA}$ , i.e.,

$$\mathbb{P}_{FA} = \int_{\mathcal{R}(\tau)} f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_0) d\vec{y} \quad (14.278)$$

and

$$\mathbb{P}_{MD} = \int_{\mathcal{R}(\tau)} f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_1) d\vec{y} \quad (14.279)$$

where  $\{\vec{y} : T(\vec{y}) > \tau\}$  is the decision region for the binary detector.

As an example of a binary-simple detection test, consider univariate likelihoods for two hypotheses and a single sample,  $y$ .

one can form a binary-simple detector as

$$T(y) = y \underset{\mathcal{H}_0}{\gtrless}^{\mathcal{H}_1} \tau \quad (14.280)$$

where

$$\mathbb{P}_D = \int_{\tau}^{\infty} f_{Y|\mathcal{H}}(y|\mathcal{H}_1) dy \quad (14.281)$$

$$\alpha = \mathbb{P}_{FA} = \int_{\tau}^{\infty} f_{Y|\mathcal{H}}(y|\mathcal{H}_0) dy \quad (14.282)$$

and

$$\beta = \mathbb{P}_{MD} = \int_{-\infty}^{\tau} f_{Y|\mathcal{H}}(y|\mathcal{H}_1) dy \quad (14.283)$$

### Common Null Hypothesis Tests

The **Z-test** is an hypothesis test for which the distribution of the test statistic under the null hypothesis is modeled as a normal distribution with expected value,  $\mu_0$ , and variance,  $\sigma_0^2$ , which is known for the Z-test. Often, the test statistic for this test uses the average of the  $N$  IID random variables, e.g., it tests for the mean of the distribution, which by the central limit theorem for finite variance distributions will approach the normal distribution as  $N \rightarrow \infty$ . In this case, one can set up three different hypothesis tests

1. **Lower-tailed** where  $\mathcal{H}_0$  is  $\mu \geq \mu_0$  and  $\mathcal{H}_1$  is  $\mu \leq \mu_0$
2. **Upper-tailed** where  $\mathcal{H}_0$  is  $\mu \leq \mu_0$  and  $\mathcal{H}_1$  is  $\mu \geq \mu_0$
3. **Two-tailed** where  $\mathcal{H}_0$  is  $\mu = \mu_0$  and  $\mathcal{H}_1$  is  $\mu \neq \mu_0$

For the Z-test, the test statistic, also known as the **standard score** is calculated as

$$T(\vec{y}) = Z = \left( \frac{\hat{\mu} - \mu_0}{\sigma_0} \right) \quad (14.284)$$

where  $\hat{\mu} = N^{-1} \sum_{i=1}^N y_i$ .

Then, for the lower-tailed hypothesis test

$$F_N(T; 0, 1) \underset{\mathcal{H}_0}{\gtrless}^{\mathcal{H}_1} \alpha \quad (14.285)$$

for the upper-tailed hypothesis test

$$F_N(-T; 0, 1) \underset{\mathcal{H}_0}{\gtrless}^{\mathcal{H}_1} \alpha \quad (14.286)$$

and for the two-tailed hypothesis test

$$2F_N(-|T|; 0, 1) \underset{\mathcal{H}_0}{\gtrless}^{\mathcal{H}_1} \alpha \quad (14.287)$$

where  $F_N(x; 0, 1)$  is the CDF of the standard normal distribution, i.e. a zero-mean Gaussian with unit variance,  $\alpha$  is the significance of the test, and the double inequality denotes the selection of  $\mathcal{H}_0$  or  $\mathcal{H}_1$  depending on the values on the left side relative to  $\alpha$ .

The ***t*-test** is an hypothesis test for which the distribution of the test statistic under the null hypothesis is modeled as a *t*-distribution with degrees-of-freedom (DOF),  $v$ . Often, the test statistic for this test uses the average of  $N$  IID random variables, e.g. it tests for the mean of the distribution, which by the central limit theorem for finite variance distributions will approach the normal distribution as  $N \rightarrow \infty$ . However, if the sample set has an unknown variance and one must use the variance of the sample set as an estimate of the variance which leads to the *t*-distribution. In this case, one can set up three different hypothesis tests

1. **Lower-tailed** where  $\mathcal{H}_0$  is  $\mu \geq \mu_0$  and  $\mathcal{H}_1$  is  $\mu \leq \mu_0$
2. **Upper-tailed** where  $\mathcal{H}_0$  is  $\mu \leq \mu_0$  and  $\mathcal{H}_1$  is  $\mu \geq \mu_0$
3. **Two-tailed** where  $\mathcal{H}_0$  is  $\mu = \mu_0$  and  $\mathcal{H}_1$  is  $\mu \neq \mu_0$

For the *t*-test, the test statistic is calculated as

$$T(\vec{y}) = \frac{\hat{\mu} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}} \quad (14.288)$$

where  $\hat{\mu} = N^{-1} \sum_{i=1}^N y_i$ . This is similar to the standard score, but with an estimate of the sample variance,  $\frac{\hat{\sigma}^2}{n}$ , instead of the true variance.

Then, for the lower-tailed hypothesis test

$$F_t(t; n-1) \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \alpha \quad (14.289)$$

for the upper-tailed hypothesis test

$$F_t(-t; n-1) \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \alpha \quad (14.290)$$

and for the two-tailed hypothesis test

$$2F_t(-|t|; n-1) \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \alpha \quad (14.291)$$

where  $F_t(x; v)$  is the CDF of the *t*-distribution with  $v$  DOF and  $\alpha$  is the significance of the test.

The  **$\chi^2$ -test** is an hypothesis test for which the distribution of the test statistic under the null hypothesis is modeled as a (central)  $\chi^2$ -distribution with degrees-of-freedom (DOF),  $k$ . It can be shown that the  $\chi^2$ -distribution occurs for the sum of  $k$  squared independent standard normal random variables, i.e.,  $Y$  is  $\chi^2$  distributed with  $k$  DOF if

$$Y = \sum_{i=1}^k Z_i^2 \quad (14.292)$$

where  $Z_1, \dots, Z_k$  are standard normal random variables. Often, the test statistic for this test assumes the sample distribution is normally distributed, but one desires to test for a specific variance for the underlying distribution. In this case, one can set up a hypothesis test with  $\mathcal{H}_0$  is  $\sigma^2 \leq \sigma_0^2$  and  $\mathcal{H}_1$  is  $\sigma^2 > \sigma_0^2$ . For this case, the test statistic for the  $\chi^2$ -test is calculated as

$$T = \sum_{i=1}^N \left( \frac{y_i - \hat{\mu}}{\sigma_0} \right)^2 \quad (14.293)$$

where the mean of the sample is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N y_i \quad (14.294)$$

Then, for upper-tailed hypothesis test

$$F_{\chi^2}(T; n - 1) \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} 1 - \alpha \quad (14.295)$$

where  $F_{\chi^2}(x; k)$  is the CDF of the  $\chi^2$ -distribution with  $k$  DOF and  $\alpha$  is the significance of the test.

The **F-test** is an hypothesis test for which the distribution of the test statistic under the null hypothesis is modeled as an F-distribution with two degrees-of-freedom (DOF),  $d_1$  and  $d_2$ . The F-distribution occurs for the ratio of  $\chi^2$  random variables with DOFs,  $d_1$ , and  $d_2$ , i.e.,  $X$  is F distributed with  $d_1$  and  $d_2$  DOFs if

$$X = \frac{\frac{Y_1}{d_1}}{\frac{Y_2}{d_2}} \quad (14.296)$$

where  $Y_1$  and  $Y_2$  are  $\chi^2$  random variables with DOFs,  $d_1$  and  $d_2$ , respectively. Note that these  $\chi^2$ -distributed random variables are often two separate sets of the sum of squared independent standard normal variables. Often, the test statistic for this tests assume the sample distributions are Gaussian distributed with zero-mean and two variances,  $\sigma_1^2$  and  $\sigma_2^2$ , but one desires to test if the variances of the two sets of samples are the same. In this case, one can set up a hypothesis test with  $\mathcal{H}_0$  is  $\sigma_1^2 = \sigma_2^2$  and  $\mathcal{H}_1$  is  $\sigma_1^2 \neq \sigma_2^2$ . For this case, the test statistic for the F-test is calculated as

$$T = \frac{\frac{Y_1}{d_1}}{\frac{Y_2}{d_2}} \quad (14.297)$$

Then, for upper-tailed hypothesis test

$$F_F(T; d_1, d_2) \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} 1 - \alpha \quad (14.298)$$

where  $F_F(x; d_1, d_2)$  is the CDF of the F-distribution with  $d_1$  and  $d_2$  DOFs and  $\alpha$  is the significance of the test.

The **Kolmogorov-Smirnov test** assesses if

where  $T$  is the **Kolmogorov-Smirnov statistic**.

The **Shapiro-Wilks test** assesses if a sample set  $\vec{y} = [y_1, \dots, y_N]$  came from a normal distribution.  $T$  is the **Shapiro-Wilks statistic**

**Andersen-Darling test** assesses if

where  $T$  is the **Andersen-Darling statistic**

## Likelihood-Ratio Tests

A fundamental result in binary-simple hypothesis testing and detection theory involves the **likelihood-ratio**,  $\Lambda(\vec{y})$ , given by

$$\Lambda(\vec{y}) = \frac{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\vec{\beta}_0)}{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\vec{\beta}_1)} \quad (14.299)$$

where the null hypothesis  $\mathcal{H}_0$  defined by the parameter vector  $\vec{\beta} = \vec{\beta}_0$  is declared and has the known likelihood of  $f_{\vec{Y}|\mathcal{H}}(\vec{y}|\vec{\beta}_0)$ , and the alternative hypothesis  $\mathcal{H}_1$  defined by the parameter vector  $\vec{\beta} = \vec{\beta}_1$  is declared and has the known likelihood of  $f_{\vec{Y}|\mathcal{H}}(\vec{y}|\vec{\beta}_1)$ . The **Neyman-Pearson lemma** states that the **uniformly most powerful (UMP) test**, i.e., the detector that maximizes  $\mathbb{P}_D$  for a specified “uniform”  $\mathbb{P}_{FA} = \alpha$ , is given by the **likelihood-ratio test (LRT)**, also known as the **Wilks test**, i.e.,

$$\Lambda(\vec{y}) \stackrel{\mathcal{H}_1}{\gtrless} \stackrel{\mathcal{H}_0}{\lessdot} \tau \quad (14.300)$$

where the value for  $\tau$  is obtained from the specified probability of false alarm, i.e.,

$$\alpha = \mathbb{P}_{FA} = \int_{\vec{y}: \Lambda(\vec{y}) > \tau} f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_0) d\vec{y} \quad (14.301)$$

In addition, the power of the hypothesis test/detector is

$$\mathbb{P}_D = \int_{\vec{y}: \Lambda(\vec{y}) < \tau} f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_1) d\vec{y} \quad (14.302)$$

and the probability of missed detection is given by

$$\beta = \mathbb{P}_{MD} = \int_{\vec{y}: \Lambda(\vec{y}) < \tau} f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_1) d\vec{y} \quad (14.303)$$

Notably, for discrete random variables where the probability of  $\Lambda(\vec{y}) = \tau$  may be non-zero, one has the additional decision rule  $\Lambda(\vec{y}) = \tau$  to declare  $\mathcal{H}_1$  with probability  $q$  where the values of  $q$  and  $\tau$  are obtained from the new specified probability of false alarm, i.e.,

$$\alpha = \mathbb{P}_{FA} = \mathbb{P}(\Lambda(\vec{y}) > \tau | \mathcal{H}_0) + q \mathbb{P}(\Lambda(\vec{y}) = \tau | \mathcal{H}_0) \quad (14.304)$$

Notably, as  $\mathbb{P}_{FA}$ ,  $\mathbb{P}_D$ , and  $\mathbb{P}_{MD}$  all depend on  $\tau$ , one typically summarizes this dependence through a two-dimensional plot of the achievable  $(\mathbb{P}_D, \mathbb{P}_{FA})$  known as the **receiver-operating characteristic (ROC)** curve.

For binary-composite signal detection problems where  $\mathcal{H}_0$  is parameterized by unknown or random parameters,  $\vec{\beta}_1 \in \vec{\mathcal{B}}_0$ , and  $\mathcal{H}_1$  is parameterized by unknown or random parameters,  $\vec{\beta}_1 \in \vec{\mathcal{B}}_1$ , the likelihood-ratio can be extended to the **generalized likelihood-ratio**,  $\Lambda_G(\vec{y})$ , given by

$$\Lambda_G(\vec{y}) = \frac{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\hat{\vec{\beta}}_0, \mathcal{H}_1)}{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\hat{\vec{\beta}}_1, \mathcal{H}_1)} \quad (14.305)$$

where  $\hat{\vec{\beta}}_0$  and  $\hat{\vec{\beta}}_1$  are the maximum likelihood estimates (MLE) of the parameter vectors

$$\hat{\vec{\beta}}_0 = \underset{\hat{\vec{\beta}}}{\operatorname{argmax}} f_{\vec{Y}|\mathcal{H}}(\vec{y}|\vec{\beta}_0, \mathcal{H}_0) \quad (14.306)$$

and

$$\hat{\vec{\beta}}_1 = \underset{\hat{\vec{\beta}}}{\operatorname{argmax}} f_{\vec{Y}|\mathcal{H}}(\vec{y}|\vec{\beta}_1, \mathcal{H}_1) \quad (14.307)$$

and the **generalized likelihood-ratio test (GLRT)** can be defined as

$$\Lambda_G(\vec{y}) \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \tau \quad (14.308)$$

It should be noted that the GLRT has no optimality associated with it except it is asymptotically the UMP invariant hypothesis test where invariance means the decision with respect to one-to-one mappings of the test statistic.

An important case, consider  $N$  samples of  $y$ , i.e.,  $\vec{y} = [y_0, \dots, y_{N-1}]^T \in \mathbb{R}^N$  where the potential signal to be detected is a known sequence  $\vec{s} = [s_0, \dots, s_{N-1}]^T$  corrupted by independent and identically distributed (IID) additive white Gaussian noise with known variance  $\sigma^2$ , i.e.,

$$\begin{aligned} \mathcal{H}_0 : y_i &= n_i \quad Y_i \sim \mathcal{N}(0, \sigma^2) \quad i = 0, \dots, N-1 \\ \mathcal{H}_1 : y_i &= s_i + n_i \quad Y_i \sim \mathcal{N}(s_i, \sigma^2) \quad i = 0, \dots, N-1 \end{aligned} \quad (14.309)$$

or, in vector form, one has

$$\begin{aligned} \mathcal{H}_0 : \vec{y} &= \vec{n} \quad \vec{Y} \sim \mathcal{N}(0, \sigma^2 I_N) \\ \mathcal{H}_1 : \vec{y} &= \vec{s} + \vec{n} \quad Y_i \sim \mathcal{N}(\vec{s}, \sigma^2 I_N) \end{aligned} \quad (14.310)$$

In this case, the likelihood-ratio is given by

$$\Lambda = \frac{f_{\mathcal{N}}(\vec{y}; \vec{s}, \sigma^2 I_N)}{f_{\mathcal{N}}(\vec{y}; \vec{0}, \sigma^2 I_N)} \quad (14.311)$$

$$\Lambda = \frac{(2\pi\sigma^2)^{-N/2}}{(2\pi\sigma^2)^{-N/2}} \frac{\exp\left(\frac{-1}{2\sigma^2}(\vec{y} - \vec{s})^T(\vec{y} - \vec{s})\right)}{\exp\left(\frac{-1}{2\sigma^2}\vec{y}^T\vec{y}\right)} \quad (14.312)$$

$$\Lambda = \exp\left(\frac{-1}{2\sigma^2}((\vec{y} - \vec{s})^T(\vec{y} - \vec{s}) - \vec{y}^T\vec{y})\right) \quad (14.313)$$

Thus, the LRT becomes

$$\frac{-1}{2} \left( \vec{y}^T \vec{y} - 2\vec{s}^T \vec{y} + \vec{s}^T \vec{s} - \vec{y}^T \vec{y} \right) \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \sigma^2 \ln \tau \quad (14.314)$$

which can be written as the test statistic/detector

$$T(\vec{y}) = \vec{s}^T \vec{y} \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \tau' \quad (14.315)$$

where

$$\tau' = \sigma^2 \ln \tau + \frac{1}{2} \vec{s}^T \vec{s} \quad (14.316)$$

This form of test statistic/detector can be interpreted as the output of a discrete-time LTI system with FIR as

$$h_i = \begin{cases} s_{N-i-1} & i = 0, \dots, N-1 \\ 0 & \text{otherwise} \end{cases} \quad (14.317)$$

which is known as the **matched filter** of the signal to be detected. To see this, consider the output of this discrete-time LTI system,  $z_k$ , with input  $y_0, \dots, y_{N-1}$  at time step  $k$  given by the convolution sum

$$z_k = \sum_{j=0}^{N-1} h_{k-j} y_j = \sum_{j=1}^N s_{N-1-k+j} y_j \quad (14.318)$$

and specifically after all the samples are collected at time  $N - 1$ , one has

$$z_{N-1} = \sum_{j=1}^N s_j y_j = \vec{s}^T \vec{y} \quad (14.319)$$

In fact, it can be shown that the matched filter maximizes the signal-to-noise ratio (SNR) at time step  $N - 1$  over all possible LTI filters that are zero outside  $i = 0, \dots, N - 1$  and is given by

$$\text{SNR} = \frac{\vec{s}^T \vec{s}}{\sigma^2} \quad (14.320)$$

Furthermore, for the sum of the  $N$  IID Gaussians, one has the test statistic/detector distributed as a Gaussian with two different means dependent on the hypothesis, i.e.,

$$\begin{aligned} \mathcal{H}_0 : T(\vec{y}) &= \vec{s}^T \vec{n} \sim \mathcal{N}(\vec{0}, \sigma^2 \vec{s}^T \vec{s}) \\ \mathcal{H}_1 : T(\vec{y}) &= \vec{s}^T \vec{s} + \vec{s}^T \vec{n} \sim \mathcal{N}(\vec{s}^T \vec{s}, \sigma^2 \vec{s}^T \vec{s}) \end{aligned} \quad (14.321)$$

where one has

$$\mathbb{P}_D = 1 - F_N\left(\tau'; \vec{s}^T \vec{s}, \sigma^2 \vec{s}^T \vec{s}\right) \quad (14.322)$$

$$\mathbb{P}_{FA} = 1 - F_N\left(\tau'; \vec{0}, \sigma^2 \vec{s}^T \vec{s}\right) \quad (14.323)$$

and

$$\mathbb{P}_{MD} = F_N\left(\tau'; \vec{s}^T \vec{s}, \sigma^2 \vec{s}^T \vec{s}\right) \quad (14.324)$$

It can be shown that using the Q-function, i.e.,  $Q(\vec{x}) = 1 - Z(\vec{x}) = 1 - F_N(\vec{x}; \vec{0}, I)$ , one can write the probability of detection in terms of a given probability of false alarm as

$$\mathbb{P}_D = Q\left(Q^{-1}(\mathbb{P}_{FA}) - \sqrt{\frac{\vec{s}^T \vec{s}}{\sigma^2}}\right) \quad (14.325)$$

which demonstrates that the probability of detection for the matched filter depends directly on the SNR.

Notably, an important aspect of the matched filter is that the noise characteristics are known, e.g., the distribution type and variance. In practice, this may not be known and/or may vary with time or space. Thus, for signals that may exist within some portion of the data, an adaptive detector called a constant false alarm rate (CFAR) detector is used to estimate the noise statistics, e.g., variance, around each sample, i.e., each “cell under test (CUT)” and decide if a signal is to be declared as detected. The simplest CFAR detection scheme called a cell-averaging CFAR (CA-CFAR) algorithm computes the observed signal power around the CUT, ignoring immediately adjacent “guard cells” which may be affected by the CUT itself, and uses that

value to estimate the noise variance. This variance is then used to compute the threshold for the matched filter detector to maintain a constant  $\mathbb{P}_{FA}$ . More sophisticated CFAR algorithms use different noise distributions and corresponding statistics than Gaussian noise parameterized by some variance, e.g., detecting submarine periscopes in the presence of sea surface returns which often use a  $K$ -distribution for modeling.

Lastly, the matched filter can be generalized for known correlated noise, i.e.,  $\vec{n} \sim \mathcal{N}(\vec{0}, \Sigma)$  with  $\Sigma$ , to the **generalized matched filter**

$$T(\vec{y}) = \vec{s}^T \Sigma^{-1} \vec{y} \stackrel{\mathcal{H}_1}{\geq_{\mathcal{H}_0}} \tau' \quad (14.326)$$

with

$$\begin{aligned} \mathcal{H}_0 : T(\vec{y}) &= \vec{s}^T \Sigma^{-1} \vec{n} \sim \mathcal{N}(\vec{0}, \vec{s}^T \Sigma^{-1} \vec{s}) \\ \mathcal{H}_1 : T(\vec{y}) &= \vec{s}^T \Sigma^{-1} \vec{s} + \vec{s}^T \Sigma^{-1} \vec{n} \sim \mathcal{N}(\vec{s}^T \Sigma^{-1} \vec{s}, \vec{s}^T \Sigma^{-1} \vec{s}) \end{aligned} \quad (14.327)$$

$$\mathbb{P}_D = 1 - F_N \left( \tau'; \vec{s}^T \Sigma^{-1} \vec{s}, \vec{s}^T \Sigma^{-1} \vec{s} \right) \quad (14.328)$$

$$\mathbb{P}_{FA} = 1 - F_N \left( \tau'; \vec{0}, \vec{s}^T \Sigma^{-1} \vec{s} \right) \quad (14.329)$$

and

$$\mathbb{P}_{MD} = F_N \left( \tau'; \vec{s}^T \Sigma^{-1} \vec{s}, \vec{s}^T \Sigma^{-1} \vec{s} \right) \quad (14.330)$$

Thus, to maximize  $\mathbb{P}_D$  for any fixed  $\mathbb{P}_{FA}$ , one should make  $\vec{s}^T \Sigma^{-1} \vec{s}$  as large as possible.

## Bayes Risk Tests

The Bayesian extension to the LRTs for hypothesis testing considers assigning costs to each type of decision outcome, i.e., the cost,  $c_{i,j}$  of declaring  $\mathcal{H}_i$  when  $\mathcal{H}_j$  is true with  $i, j \in \{0, 1, \dots, M-1\}$ . Furthermore, in a Bayesian context, one may have a prior belief about the probability of each hypothesis, i.e., know  $\pi_j = \mathbb{P}(\mathcal{H}_j \text{ true})$ . Here, the **Bayes risk**, also known as the **Bayes cost** or **Bayes criterion**, for the hypothesis testing problem is given by

$$\sum_{i=0}^{M-1} \sum_{j=0}^{M-1} c_{i,j} \mathbb{P}(\text{declare } \mathcal{H}_i \& \mathcal{H}_j \text{ true}) \quad (14.331)$$

$$\sum_{i=0}^{M-1} \sum_{j=0}^{M-1} c_{i,j} \mathbb{P}(\text{declare } \mathcal{H}_i | \mathcal{H}_j \text{ true}) \mathbb{P}(\mathcal{H}_j \text{ true}) \quad (14.332)$$

$$\sum_{i=0}^{M-1} \sum_{j=0}^{M-1} c_{i,j} \pi_j \mathbb{P}(\text{declare } \mathcal{H}_i | \mathcal{H}_j \text{ true}) \quad (14.333)$$

Notably, the Bayes risk is equal to **probability of error**,  $\mathbb{P}_E = \mathbb{P}(\text{declare } \mathcal{H}_i | \mathcal{H}_j \text{ true}) + \mathbb{P}(\text{declare } \mathcal{H}_i | \mathcal{H}_j \text{ true})$  when  $c_{i,j} = 1$  for  $i \neq j$  and  $c_{i,j} = 0$  for  $i = j$ .

It can be shown that the **Bayes risk test** declares the hypothesis that minimizes Bayes risk, i.e., declare the  $i^{\text{th}}$  hypothesis  $\mathcal{H}_{i_{opt}}$  with the minimum average cost given  $\vec{y}$ , i.e.,

$$\mathcal{H}_{i_{opt}} = \underset{\mathcal{H}_i, i=0, \dots, M-1}{\operatorname{argmax}} \sum_{j=0}^{M-1} c_{i,j} p(\mathcal{H}_j | \vec{y}) = \frac{f_{\vec{Y}|\mathcal{H}}(\vec{y} | \mathcal{H}_j) \pi_j}{f_{\vec{Y}}(\vec{y})} \quad (14.334)$$

where  $p(\mathcal{H}_j|\vec{y})$  is the posterior PMF on the hypothesis given  $\vec{y}$ . Notably, if the Bayes risk is chosen as  $\mathbb{P}_E$ , then the Bayes risk test becomes the **minimum  $\mathbb{P}_E$  test**, i.e.,

$$\mathcal{H}_{i,opt} = \operatorname{argmax}_{\mathcal{H}_i, i=0, \dots, M-1} p(\mathcal{H}_i|\vec{y}) = \frac{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_i)\pi_i}{f_{\vec{Y}}(\vec{y})} \quad (14.335)$$

which is a maximum *a posteriori* (MAP) decision rule. If the prior probabilities are also all equal, one has

$$\mathcal{H}_{i,opt} = \operatorname{argmax}_{\mathcal{H}_i, i=0, \dots, M-1} f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_i) \quad (14.336)$$

which is a maximum likelihood (ML) decision rule. Similar to the GLRT, the Bayes risk test can be generalized for composite hypothesis testing by assigning prior probabilities for  $\vec{\beta}_i \in \vec{\mathcal{B}}_i$  for each hypothesis  $\mathcal{H}_i$  and substituting for the likelihoods by the equations

$$f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_i) = \int_{\vec{\mathcal{B}}_i} f_{\vec{Y}|\mathcal{H}, \vec{\mathcal{B}}}(\vec{y}|\mathcal{H}_i, \vec{\mathcal{B}}) f_{\vec{\mathcal{B}}}(\vec{\beta}_i) d\vec{\beta}_i \quad (14.337)$$

which “averages” the unknown parameters, in contrast to the ML parameters used in the GLRT.

Lastly, notably the binary Bayes risk test can be written as an LRT

$$\frac{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_1)}{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_0)} \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \tau \quad (14.338)$$

where

$$\tau = \frac{(c_{1,0} - c_{0,0})\pi_0}{(c_{0,1} - c_{1,1})\pi_1} \quad (14.339)$$

is generally chosen to take into account the decision costs and hypothesis priors. Furthermore, for the binary minimum  $\mathbb{P}_E$  test, one has

$$\frac{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_1)}{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_0)} \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \frac{\pi_0}{\pi_1} \quad (14.340)$$

and for equal prior probabilities, one has

$$\frac{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_1)}{f_{\vec{Y}|\mathcal{H}}(\vec{y}|\mathcal{H}_0)} \stackrel{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} 1 \quad (14.341)$$

i.e., the most likely hypothesis is chosen.

## Information Criterion Tests for Multiple Model Selection

### Mutual Information Criterion

**Bayesian Information Criterion (BIC)** and the **Widely Applicable Bayesian Information Criterion (WBIC)**

**Akaike Information Criterion (AIC)** and the **Widely Applicable Information Criterion (WAIC)**

**Hannan-Quinn Information Criterion (HQC)**

---

# Optimal Linear State Estimation Theory

## 15.1 Introduction to Optimal State Estimation

**Decision theory** is the determination of the optimal decision given constraints and assumptions on a decision problem. **State estimation** is one special case of a model-based decision problem and can be stated as deciding which state,  $\vec{x}$ , one should select as a function of time. However, the Bayesian state estimation problem assumes that the state,  $\vec{x}$ , is “hidden” and only observed through measurements,  $\vec{y}$ , through which Bayesian inference must be performed. With the inherent time-dependency of dynamical state models, three different types of time horizons can be considered for the type of state estimation problem and the type of state estimator. State estimation for a *future* state is known as **predicting** which utilizes a mathematical law or algorithm known as a **state predictor**. State estimation for a *current* state is known as **state filtering** which utilizes a mathematical law or algorithm known as a **state filter**. State estimation for a *past* state is known as **state smoothing** which utilizes a mathematical law or algorithm known as a **state smoother**.

In addition to this general problem for state estimation, perception systems use proprioceptive and exteroceptive sensors to observe the system and infer the state through Bayes state estimation using a model of the sensor with respect to the state and/or the state change. As every sensor exhibits some degree of inaccuracy, uncertainty, or random phenomena in its measurements, one must additionally model sensor **noise** as opposed to the modeled **signal** which are two terms based in signal processing theory. Therefore, as these noises are unknown, i.e. one cannot predict with absolute certainty what their exact values will be in the future, one instead models the noise as stochastic processes. Another perspective on the uncertainty modeling is due to the fact that with the stochastic dynamical system model, one is assuming certain aspects about the modeled process selected to represent the actual system, environment, and sensors. Any tractable sensor model cannot be exactly quantified at every instant, e.g., calibration errors or the electro-mechanical variations within an individual sensor. These sensor model errors can generally be dynamic uncertainties and contribute to the random nature of sensors.

## Bayesian State Estimation

With these concepts in mind, one typically utilizes a Bayesian approach to model-based state estimation which will be presented here for general discrete-time state and measurement process models where the subscripts denote the time step indices of the vectors with the “ $1 : N$ ” denoting grouping together all vectors with indices as  $1, \dots, 5$ . Thus, the complete state and measurement joint PDF is

$$f_{\vec{X}_{0:N}, \vec{Y}_{1:N}}(\vec{x}_{0:N}, \vec{y}_{1:N}) = f_{\vec{X}_{0:N} | \vec{Y}_{1:N}}(\vec{x}_{0:N} | \vec{y}_{1:N}) f_{\vec{Y}_{1:N}}(\vec{y}_{1:N}) \quad (15.1)$$

where “joint” denotes all time steps,  $k = 1, \dots, N$ . This can be decomposed theoretically by Bayes’ rule in the joint state posterior PDF

$$f_{\vec{X}_{0:N} | \vec{Y}_{1:N}}(\vec{x}_{0:N} | \vec{y}_{1:N}) = \frac{f_{\vec{Y}_{1:N} | \vec{X}_{0:N}}(\vec{y}_{1:N} | \vec{x}_{0:N}) f_{\vec{X}_{0:N}}(\vec{x}_{0:N})}{f_{\vec{Y}_{1:N}}(\vec{y}_{1:N})} \quad (15.2)$$

where  $f_{\vec{X}_{0:N}}(\vec{x}_{0:N})$  is the prior distribution including the previous time step given by the state process/dynamics model,  $f_{\vec{Y}_{1:N} | \vec{X}_{0:N}}(\vec{y}_{1:N} | \vec{x}_{0:N})$  is the likelihood model of the measurement process, and  $f_{\vec{X}_{0:N} | \vec{Y}_{1:N}}(\vec{x}_{0:N} | \vec{y}_{1:N})$  is the joint state posterior PDF given the observed measurement history. Recalling the discussion for Bayesian parameter estimation, if one computes the state posterior PDF, one can use different optimality criteria for the **Bayes state estimator**, e.g., the **maximum *a posteriori* (MAP)** as the maximal mode of the posterior, the **minimum mean square error (MMSE)** as the mean of the posterior for finite means and variances, and the **median estimator** as the median of the posterior.

However, this general Bayes formulation is difficult to compute useful information especially if new measurements are obtained in time  $k$  and the state should be dynamically estimated as a function of  $k$ , e.g., in a perception system. Thus, for computational tractability, one can simplify the Bayes state estimation problem with the **hidden Markov model (HMM) assumption** which assumes the state process follow an  $n_x^{\text{th}}$ -order Markov process, i.e.,  $f_{\vec{X}}(\vec{x}_k | \vec{x}_{k-1})$ , which allows one to model the joint state prior PDF as

$$f_{\vec{X}_{0:N}}(\vec{x}_{0:N}) = \left( \prod_{k=1}^N f_{\vec{X}_k | \vec{X}_{k-1}}(\vec{x}_k | \vec{x}_{k-1}) \right) f_{\vec{X}_0}(\vec{x}_0) \quad (15.3)$$

where  $f_{\vec{X}}(\vec{x}_0)$  is the **state prior PDF** at the initial time step  $k = 0$  and  $f_{\vec{X}_k | \vec{X}_{k-1}}(\vec{x}_k | \vec{x}_{k-1})$  is the **state transition PDF** from time step  $k - 1$  to  $k$ . Furthermore, as the measurement of the Markov state at time  $k$  is conditionally independent of other states and measurements, i.e.,  $f_{\vec{Y}}(\vec{y}_k | \vec{x}_k)$  is the measurement model at  $k$  that depends only on the state at that time  $k$ , one has the joint likelihood between of the state and measurements as

$$f_{\vec{Y}_{1:N} | \vec{X}_{0:N}}(\vec{y}_{1:N} | \vec{x}_{0:N}) = \prod_{k=1}^N f_{\vec{Y}_k | \vec{X}_k}(\vec{y}_k | \vec{x}_k) \quad (15.4)$$

With the HMM assumption, the joint state posterior PDF can be rewritten as

$$f_{\vec{X}_{0:N} | \vec{Y}_{1:N}}(\vec{x}_{0:N} | \vec{y}_{1:N}) = \frac{\left( \prod_{k=1}^N f_{\vec{Y}_k | \vec{X}_k}(\vec{y}_k | \vec{x}_k) f_{\vec{X}_k | \vec{X}_{k-1}}(\vec{x}_k | \vec{x}_{k-1}) \right) f_{\vec{X}}(\vec{x}_0)}{f_{\vec{Y}_{1:N}}(\vec{y}_{1:N})} \quad (15.5)$$

which could be solved similar to batch parameter estimators as stacked state trajectory over all time steps could be considered as single “parameter vector” to be estimated. However, computing the complete joint

posterior distribution is often unnecessary and inefficient for real-time perception as the number of samples will change with time. Thus, one is typically interested in one of three different posteriors of the complete joint PDF of the state and measurement, namely, the **state prediction posterior PDF**

$$f_{\vec{X}_{k+n} | \vec{Y}_{1:k}}(\vec{x}_{k+n} | \vec{y}_{1:k}) \quad (15.6)$$

with  $n > 0$  constant, the **state filtering posterior PDF**

$$f_{\vec{X}_k | \vec{Y}_{1:k}}(\vec{x}_k | \vec{y}_{1:k}) \quad (15.7)$$

and the **state smoothing posterior PDF**

$$f_{\vec{X}_k | \vec{Y}_{1:k+n}}(\vec{x}_k | \vec{y}_{1:k+n}) \quad (15.8)$$

where if  $0 < n < N - k$  and constant, one has a **fixed-lag state smoothing posterior PDF**, and if  $n = N - k$  for all  $k$  with  $N$  constant, one has a **fixed-interval state smoothing posterior**. Thus, fixed-interval smoothers can sometimes be considered the same as batch parameter estimators as the stacked state trajectory over all time steps could be considered as a single “parameter vector” to be estimated.

Recall that if with the HMM assumption, one can alternatively represent the state process,  $f_{\vec{X}_k | \vec{X}_{k-1}}(\vec{x}_k | \vec{x}_{k-1})$ , and the measurement likelihood,  $f_{\vec{Y}_k | \vec{X}_k}(\vec{y}_k | \vec{x}_k)$ , with the discrete-time stochastic state-space model

$$\begin{aligned} \vec{x}[k+1] &= f(\vec{x}[k], \vec{u}[k], \vec{w}[k], k) \\ \vec{y}[k] &= h(\vec{x}[k], \vec{v}[k], k) \end{aligned} \quad (15.9)$$

where  $\vec{w}[k]$  and  $\vec{v}[k]$  represent the process and measurement noise with some known distribution related to  $f_{\vec{X}_k | \vec{X}_{k-1}}(\vec{x}_k | \vec{x}_{k-1})$  and  $f_{\vec{Y}_k | \vec{X}_k}(\vec{y}_k | \vec{x}_k)$ .

Furthermore, one can also consider an equivalent notation and development of the HMM for continuous-time and hybrid-time Bayesian state estimation which allows similar stochastic state-space models, i.e., for continuous-time as

$$\begin{aligned} \dot{\vec{x}}(t) &= f(\vec{x}(t), \vec{u}(t), \vec{w}(t), t) \\ \vec{y}(t) &= h(\vec{x}(t), \vec{v}(t), t) \end{aligned} \quad (15.10)$$

and hybrid-time as

$$\begin{aligned} \dot{\vec{x}}(t) &= f(\vec{x}(t), \vec{u}(t), \vec{w}(t), t) \\ \vec{y}[k] &= h(\vec{x}[k], \vec{v}[k], k) \end{aligned} \quad (15.11)$$

### Bayes Predictor

The **Bayes predictor** estimates the state posterior PDF at time step  $k + n$  assuming one has knowledge of the state posterior PDF at some time step  $k$ , i.e.,

$$f_{\vec{X}_0 | \vec{Y}_{1:0}}(\vec{x}_0 | \vec{y}_{1:0}) \quad (15.12)$$

which for  $k = 0$ , one has  $f_{\vec{X}_k|\vec{Y}_{1:0}}(\vec{x}_k|\vec{y}_{1:0}) = f_{\vec{X}}(\vec{x}_0)$ . Otherwise, this could be computed using the Bayes filter or smoother equations below. Then, note that the joint state posterior PDF between future state  $\vec{X}_{k+n}, \dots, \vec{X}_{k+1}$  and the current state  $\vec{X}_k$  can be defined as

$$f_{\vec{X}_{k:k+n}|\vec{Y}_{1:k}}(\vec{x}_{k+n}, \dots, \vec{x}_k|\vec{y}_{1:k}) = \left( \prod_{i=0}^{n-1} f_{\vec{X}_{k+i+1}|\vec{X}_{k+i}, \vec{Y}_{1:k}}(\vec{x}_{k+i+1}|\vec{x}_{k+i}, \vec{y}_{1:k}) \right) f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k}) \quad (15.13)$$

and by the Markov property for the state process, one has

$$f_{\vec{X}_{k:k+n}|\vec{Y}_{1:k}}(\vec{x}_{k+n}, \dots, \vec{x}_k|\vec{y}_{1:k}) = \left( \prod_{i=0}^{n-1} f_{\vec{X}_{k+i+1}|\vec{X}_{k+i}}(\vec{x}_{k+i+1}|\vec{x}_{k+i}) \right) f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k}) \quad (15.14)$$

Thus, the **state prediction posterior PDF** at  $k+n$  can be computed by marginalizing out  $\vec{X}_k, \dots, \vec{X}_{k+n-1}$ , i.e.

$$f_{\vec{X}}(\vec{x}_{k+n}|\vec{y}_{1:k}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left( \prod_{i=0}^{n-1} f_{\vec{X}_{k+i+1}|\vec{X}_{k+i}}(\vec{x}_{k+i+1}|\vec{x}_{k+i}) \right) f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k}) d\vec{x}_k \cdots d\vec{x}_{k+n-1} \quad (15.15)$$

For  $n = 1$ , one obtains the **Chapman-Kolmogorov equation** for one-step Bayes state prediction as

$$f_{\vec{X}_{k+1}|\vec{Y}_{1:k}}(\vec{x}_{k+1}|\vec{y}_{1:k}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\vec{X}_{k+1}|\vec{X}_{k+i}}(\vec{x}_{k+1}|\vec{x}_k) f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k}) d\vec{x}_k \quad (15.16)$$

which can be implemented recursively to predict the state forward in time for any  $k+1, k+2, \dots$ . This can also be seen by the independence of the transition PDFs in the product which can be pulled out into individual sets of integrals over  $\vec{x}_k, \vec{x}_{k+1}, \dots$ .

## Bayes Filter

The **Bayes filter**, also known as the **recursive Bayes estimator**, updates or “corrects” the state prediction PDF with the currently available measurement to produce the state filtering posterior PDF. This prediction/correction framework results in two separate state filtering steps, the **prediction** and **correction** steps. This recursion assumes one has knowledge of the state filtering posterior PDF at the previous time step  $k-1$ , i.e.,

$$f_{\vec{X}_{k-1}|\vec{Y}_{1:k-1}}(\vec{x}_{k-1}|\vec{y}_{1:k-1}) \quad (15.17)$$

which for  $k = 1$ , one has  $f_{\vec{X}_0|\vec{Y}_{1:0}}(\vec{x}_0|\vec{y}_{1:0}) = f_{\vec{X}}(\vec{x}_0)$ . The **prediction step** uses the Chapman-Kolmogorov equation as

$$f_{\vec{X}_k|\vec{Y}_{1:k-1}}(\vec{x}_k|\vec{y}_{1:k-1}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\vec{X}_{k+1}|\vec{X}_{k+i}}(\vec{x}_{k+1}|\vec{x}_k) f_{\vec{X}_{k-1}|\vec{Y}_{1:k-1}}(\vec{x}_{k-1}|\vec{y}_{1:k-1}) d\vec{x}_k \quad (15.18)$$

where  $f_{\vec{X}}(\vec{x}_k|\vec{y}_{1:k-1})$  is the **state filtering prior PDF**, i.e., the state conditional PDF “prior”-measurement at  $k$ .

Recalling Bayes rule' for the state filtering posterior PDF as

$$f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k}) = \frac{f_{\vec{Y}_k|\vec{X}_k, \vec{Y}_{1:k-1}}(\vec{y}_k|\vec{x}_k, \vec{y}_{1:k-1}) f_{\vec{X}_k|\vec{Y}_{1:k-1}}(\vec{x}_k|\vec{y}_{1:k-1})}{\int \cdots \int f_{\vec{Y}_k|\vec{X}_k, \vec{Y}_{1:k-1}}(\vec{y}_k|\vec{x}_k, \vec{y}_{1:k-1}) f_{\vec{X}_k|\vec{Y}_{1:k-1}}(\vec{x}_k|\vec{y}_{1:k-1}) d\vec{x}_k} \quad (15.19)$$

which, by the conditional independence of likelihoods, the state filtering prior PDF is updated with the **correction step**

$$f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k}) = \frac{f_{\vec{Y}_k|\vec{X}_k}(\vec{y}_k|\vec{x}_k) f_{\vec{X}_k|\vec{Y}_{1:k-1}}(\vec{x}_k|\vec{y}_{1:k-1})}{\int \cdots \int f_{\vec{Y}_k|\vec{X}_k}(\vec{y}_k|\vec{x}_k) f_{\vec{X}_k|\vec{Y}_{1:k-1}}(\vec{x}_k|\vec{y}_{1:k-1}) d\vec{x}_k} \quad (15.20)$$

where  $f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k})$  is the **state filtering posterior PDF**, i.e., the state conditional PDF “post”-measurement at  $k$ .

### Fixed-Interval Bayes Smoother

The **fixed-interval Bayes smoother**, also known as the **backwards recursive Bayes estimator**, updates or “smooths” the state filtering prior and posterior PDFs with future measurements up to time step  $N$  to produce the fixed-interval state smoothing posterior PDF. This prediction-correction-smoothing framework results in three separate state smoothing steps. First, one computes the state filtering prior PDFs for all  $k = 0, \dots, N$

$$f_{\vec{X}_{k+1}|\vec{Y}_{1:k}}(\vec{x}_{k+1}|\vec{y}_{1:k}) \quad (15.21)$$

and the state filtering posterior PDFs for all  $k = 0, \dots, N$

$$f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k}) \quad (15.22)$$

which for  $k = 0$ , the state filtering posterior PDF is the state prior PDF and for  $k = N$ , the state filtering posterior PDF is the state smoothing posterior PDF at  $N$ .

The joint two-state posterior PDF for  $\vec{X}_k$  and  $\vec{X}_{k+1}$  can be written as

$$f_{\vec{X}_k, \vec{X}_{k+1}|\vec{Y}_{1:k}}(\vec{x}_k, \vec{x}_{k+1}|\vec{y}_{1:k}) = f_{\vec{X}_k|\vec{X}_{k+1}, \vec{Y}_{1:N}}(\vec{x}_k|\vec{x}_{k+1}, \vec{y}_{1:N}) f_{\vec{X}_{k+1}|\vec{Y}_{1:N}}(\vec{x}_{k+1}|\vec{y}_{1:N}) \quad (15.23)$$

and by the conditional independence of the likelihoods, one has

$$f_{\vec{X}_k, \vec{X}_{k+1}|\vec{Y}_{1:N}}(\vec{x}_k, \vec{x}_{k+1}|\vec{y}_{1:N}) = f_{\vec{X}_k|\vec{X}_{k+1}, \vec{Y}_{1:k}}(\vec{x}_k|\vec{x}_{k+1}, \vec{y}_{1:k}) f_{\vec{X}_{k+1}|\vec{Y}_{1:N}}(\vec{x}_{k+1}|\vec{y}_{1:N}) \quad (15.24)$$

By Bayes' rule for the first term, one has

$$f_{\vec{X}_k, \vec{X}_{k+1}|\vec{Y}_{1:N}}(\vec{x}_k, \vec{x}_{k+1}|\vec{y}_{1:N}) = \frac{f_{\vec{X}_{k+1}|\vec{X}_k, \vec{Y}_{1:k}}(\vec{x}_{k+1}|\vec{x}_k, \vec{y}_{1:k}) f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k})}{f_{\vec{X}_{k+1}|\vec{Y}_{1:k}}(\vec{x}_{k+1}|\vec{y}_{1:k})} f_{\vec{X}_{k+1}|\vec{Y}_{1:N}}(\vec{x}_{k+1}|\vec{y}_{1:N}) \quad (15.25)$$

By the HMM assumption, one has

$$f_{\vec{X}_k, \vec{X}_{k+1}|\vec{Y}_{1:N}}(\vec{x}_k, \vec{x}_{k+1}|\vec{y}_{1:N}) = \frac{f_{\vec{X}_{k+1}|\vec{X}_k}(\vec{x}_{k+1}|\vec{x}_k) f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k}) f_{\vec{X}_{k+1}|\vec{Y}_{1:N}}(\vec{x}_{k+1}|\vec{y}_{1:N})}{f_{\vec{X}_{k+1}|\vec{Y}_{1:k}}(\vec{x}_{k+1}|\vec{y}_{1:k})} \quad (15.26)$$

Thus, the **smoothing step** recursively updates the **state smoothing posterior PDF**,  $f_{\vec{X}_k|\vec{Y}_{1:N}}(\vec{x}_k|\vec{y}_{1:N})$ , at time step  $k$  by marginalizing out  $\vec{X}_{k+1}$ , i.e.

$$f_{\vec{X}_k|\vec{Y}_{1:N}}(\vec{x}_k|\vec{y}_{1:N}) = f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k}) \int \cdots \int \frac{f_{\vec{X}_{k+1}|\vec{X}_k}(\vec{x}_{k+1}|\vec{x}_k) f_{\vec{X}_{k+1}|\vec{Y}_{1:N}}(\vec{x}_{k+1}|\vec{y}_{1:N})}{f_{\vec{X}_{k+1}|\vec{Y}_{1:k}}(\vec{x}_{k+1}|\vec{y}_{1:k})} d\vec{x}_{k+1} \quad (15.27)$$

where  $f_{\vec{X}_k|\vec{Y}_{1:k}}(\vec{x}_k|\vec{y}_{1:k})$  is the **state filtering posterior PDF** at time step  $k$  and  $f_{\vec{X}_{k+1}|\vec{Y}_{1:k}}(\vec{x}_{k+1}|\vec{y}_{1:k})$  is the **state filtering prior PDF** at time step  $k + 1$ .

### Joint State and Parameter Estimation

In the Bayesian estimation framework, one can also incorporate **joint state and parameter estimation** which jointly estimates the unknown *static* model parameters,  $\vec{\beta}$ , and *dynamic* state,  $\vec{x}$ . Then, the joint state and parameter smoothing posterior PDF is

$$f_{\vec{X}_{0:N}, \vec{B}|\vec{Y}_{1:N}}(\vec{x}_{0:N}, \vec{\beta}|\vec{y}_{1:N}) = \frac{f_{\vec{Y}_{0:N}|\vec{X}_{0:N}, \vec{B}}(\vec{y}_{1:N}|\vec{x}_{0:N}, \vec{\beta}) f_{\vec{X}_{0:N}|\vec{B}}(\vec{x}_{0:N}|\vec{\beta}) f_{\vec{B}}(\vec{\beta})}{f_{\vec{Y}_{1:N}}(\vec{y}_{1:N})} \quad (15.28)$$

or

$$f_{\vec{X}_{0:N}, \vec{B}|\vec{Y}_{1:N}}(\vec{x}_{0:N}, \vec{\beta}|\vec{y}_{1:N}) = \frac{\left( \prod_{k=1}^N f_{\vec{Y}_k|\vec{X}_k, \vec{B}}(\vec{y}_k|\vec{x}_k, \vec{\beta}) \right) \left[ \left( \prod_{k=1}^N f_{\vec{X}_k|\vec{X}_{k-1}, \vec{B}}(\vec{x}_k|\vec{x}_{k-1}, \vec{\beta}) \right) f_{\vec{X}|\vec{B}}(\vec{x}_0|\vec{\beta}) \right] f_{\vec{B}}(\vec{\beta})}{f_{\vec{Y}_{1:N}}(\vec{y}_{1:N})} \quad (15.29)$$

One can form the **marginal state posterior PDF**

$$f_{\vec{X}_{0:N}|\vec{Y}_{1:N}}(\vec{x}_{0:N}|\vec{y}_{1:N}) = \int \cdots \int f_{\vec{X}_{0:N}, \vec{B}|\vec{Y}_{1:N}}(\vec{x}_{0:N}, \vec{\beta}|\vec{y}_{1:N}) d\vec{\beta} \quad (15.30)$$

or

$$f_{\vec{X}_{0:N}|\vec{Y}_{1:N}}(\vec{x}_{0:N}|\vec{y}_{1:N}) = \int \cdots \int f_{\vec{X}_{0:N}|\vec{B}, \vec{Y}_{1:N}}(\vec{x}_{0:N}|\vec{\beta}, \vec{y}_{1:N}) f_{\vec{B}|\vec{Y}_{1:N}}(\vec{\beta}|\vec{y}_{1:N}) d\vec{\beta} \quad (15.31)$$

Similarly one can form the **marginal parameter posterior PDF** as

$$f_{\vec{B}|\vec{Y}_{1:N}}(\vec{\beta}|\vec{y}_{1:N}) = \int \cdots \int f_{\vec{X}_{0:N}, \vec{B}|\vec{Y}_{1:N}}(\vec{x}_{0:N}, \vec{\beta}|\vec{y}_{1:N}) d\vec{x}_{0:N} \quad (15.32)$$

or

$$f_{\vec{B}|\vec{Y}_{1:N}}(\vec{\beta}|\vec{y}_{1:N}) = \int \cdots \int f_{\vec{B}|\vec{Y}_{1:N}}(\vec{\beta}|\vec{y}_{1:N}) f_{\vec{X}_{0:N}|\vec{Y}_{1:N}}(\vec{x}_{0:N}|\vec{y}_{1:N}) d\vec{x}_{0:N} \quad (15.33)$$

which can be considered as

$$f_{\vec{B}|\vec{Y}_{1:N}}(\vec{\beta}|\vec{y}_{1:N}) = \frac{f_{\vec{Y}_{1:N}|\vec{B}}(\vec{y}_{1:N}|\vec{\beta}) f_{\vec{B}}(\vec{\beta})}{f_{\vec{Y}_{1:N}}(\vec{y}_{1:N})} \quad (15.34)$$

or by the conditional independence of measurements, one has

$$f_{\vec{B}|\vec{Y}_{1:N}}(\vec{\beta}|\vec{y}_{1:N}) = \frac{\left( \prod_{k=1}^N f_{\vec{Y}_k|\vec{Y}_{1:k-1}, \vec{B}}(\vec{y}_k|\vec{y}_{1:k-1}, \vec{\beta}) \right) f_{\vec{B}}(\vec{\beta})}{f_{\vec{Y}_{1:N}}(\vec{y}_{1:N})} \quad (15.35)$$

where  $f_{\vec{Y}_k|\vec{Y}_{1:0}, \vec{B}}(\vec{y}_k|\vec{y}_{1:0}, \vec{\beta}) = f_{\vec{Y}_k|\vec{B}}(\vec{y}_k|\vec{\beta})$ . This marginal parameter posterior PDF is often easier to estimate using use with Bayes parameter estimation methods than the joint state and parameter estimation using the joint state and parameter posterior PDF.

This problem can also be written as a stochastic state-space system conditioned on the parameter vector, i.e.,

$$\begin{aligned} \vec{x}_k &= f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1}; \vec{\beta}) \\ \vec{y}_k &= h_k(\vec{x}_k, \vec{v}_k; \vec{\beta}) \end{aligned} \quad (15.36)$$

For recursive estimation, i.e. filtering, this requires coupling the state and parameter estimation problems into one optimal Bayesian filtering problem and can be approximated in one of two ways using a stochastic state-space framework.

The first is as a **dual estimation** problem which sets up two filters to run in parallel where the state-space system model for the state filter remains the same and the state-space system model for the parameter filter can be setup as

$$\begin{aligned} \vec{\beta}_k &= \vec{\beta}_{k-1} + \vec{e}_{k-1} \\ \vec{y}_k &= h_k(f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1}; \vec{\beta}_k), \vec{v}_k; \vec{\beta}_k) \end{aligned} \quad (15.37)$$

where  $\vec{e}_{k-1}$  is some additional process noise in the parameter estimate. Here the posterior state and parameter estimates,  $\vec{x}_{k|k}$  and  $\vec{\beta}_{k|k}$ , are updated in the other parallel filter after each time step.

The second is as a **joint estimation** problem which sets up one filter for a single *joint state* of the state and parameters, i.e.  $[\vec{x}^T \ \vec{\beta}^T]^T$ , also known as the **augmented state**, which results in the following **joint state-space model**, also known as the **augmented state-space model**

$$\begin{bmatrix} \vec{x}_k \\ \vec{\beta}_k \end{bmatrix} = \begin{bmatrix} f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1}; \vec{\beta}_{k-1}) \\ \vec{\beta}_{k-1} + \vec{e}_{k-1} \end{bmatrix} \quad (15.38)$$

$$\vec{y}_k = h_k(\vec{x}_k, \vec{v}_k; \vec{\beta}_k)$$

In the joint estimation case, even if one has a linear relationship between the output and the parameters and the output and the state, the joint system will be a nonlinear system which requires the use of nonlinear filters, the subject of the next chapter.

## Optimal Information Fusion

A **multi-sensor system** that processes multiple sensor output signals to produce information is said to be using **multi-sensor information fusion**, which can be data-driven and/or model-based. This textbook focuses on model-based perception system design using optimal state estimation theory and detection theory. The use of a model-based approach allows one to also robustly assess perception system risk based on model

analysis. Model-based multi-sensor information fusion relies heavily on model-based parameter and state estimation methods as this framework allows one to fuse different sensor measurements into an overall state estimate,  $\hat{x}$ , where the state is any quantity of interest to the user or the information system.

Therefore, Bayesian multi-sensor data fusion uses a **time update step** from a proprioceptive sensor as a *prediction step* in the traditional Bayes filter context and a **measurement update step** from an exteroceptive sensor as a *correction step* in the traditional Bayes filter context. Furthermore, it should be noted that any number of sensors can be used in this manner by simply employing the appropriate step of the chosen filter. One could also include a known state dynamics model in this data fusion algorithm as an additional *prediction step*. Thus, the Bayes filter equations form the **optimal information fusion** algorithm assuming the sensor models are correct. Furthermore, for truly linear models between the sensor data and the state(s) and AWGN for the sensor error models, the Kalman filter equations provide the **optimal information fusion** algorithm. However, with multi-sensor systems, one can also use a cascade of Bayesian filters and fuse the state estimates of the two or more filters, i.e., **loosely-coupled fusion** at the state estimate level as opposed to **tightly-coupled fusion** at the measurement level. When a steady-state gain is used to fuse scalar information from two sensors, this is also known as a **complementary filter**, which is typically designed in the frequency domain.

The standard Kalman filter equations assume a purely mathematical derivation from the principles of probability theory. However in many cases, the use of the Kalman filter in embedded systems may require special consideration with respect to computational resources and/or numerical precision. However, due to the linearity of the Kalman filter, one can derive a variety of alternative recursions for the prediction and correction steps of the Kalman depending on additional considerations. In addition, the memory requirements for a large number of measurements or time steps may not be feasible, especially for fixed-interval Kalman smoothing requiring memory for each state filtering prior and posterior state estimate and covariance. Another drawback of fixed-interval smoothing is the recalculation required if additional measurements are obtained and  $N$  increases, e.g., for  $M$  additional measurements,  $M$  prediction, correction, and smoothing steps must be computed and  $N$  smoothing steps must be recomputed. Thus, an alternative for smoothing in real-time applications is a fixed-lag Kalman smoother which augments the Kalman filter state and smooths the state estimate at a fixed-lag of  $k - N$  time steps behind the current measurement at  $k$ .

## References

For more information, please refer to the following

- Higgins, W. T., “A Comparison of Complementary and Kalman Filtering,” in *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 11, No. 3, 1975, pp. 321-325
- Sarkka, S., “1.3 Optimal filtering and smoothing as Bayesian inference,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 8-12
- Sarkka, S., “1.5 Parameter estimation,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 14-16
- Sarkka, S., “4.1 Probabilistic state space models,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 51-53

- Sarkka, S., “4.2 Bayesian filtering equations,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 54-56
- Sarkka, S., “8.1 Bayesian smoothing equations,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 54-55
- Sarkka, S., “12.1 Bayesian estimation of parameters in state space models,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 174-16
- Simon, D., “15.1 Bayesian state estimation,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 462-466

## 15.2 Discrete-Time Kalman Filtering and Smoothing

Consider the following multivariate Gaussian PDFs for the state transition PDF and the likelihood PDF

$$\begin{aligned} f_{\vec{X}_k|\vec{X}_{k-1}}(\vec{x}_k|\vec{x}_{k-1}) &= f_N(\vec{x}_k; F_{k-1}\vec{x}_{k-1} + G_{k-1}\vec{u}_{k-1}, Q_{k-1}) \\ f_{\vec{Y}_k|\vec{X}_k}(\vec{y}_k|\vec{x}_k) &= f_N(\vec{y}_k; H_k \vec{x}_k, R_k) \end{aligned} \quad (15.39)$$

which corresponds to the discrete-time linear stochastic state-space model with additive (weakly) white Gaussian noise (AWGN), also known as the linear-Gaussian state-space model,

$$\begin{aligned} \vec{x}_k &= F_{k-1}\vec{x}_{k-1} + G_{k-1}\vec{u}_{k-1} + \vec{w}_{k-1} \\ \vec{y}_k &= H_k \vec{x}_k + \vec{v}_k \end{aligned} \quad (15.40)$$

where the process noise is  $\vec{W}_k \sim N(0, Q_k)$  and the measurement noise is  $\vec{V}_k \sim N(0, R_k)$  which are uncorrelated with each other.

For linear models with additive white Gaussian noise (AWGN), the state filtering prior and posterior PDFs of the Bayes filter formulation remain multivariate Gaussians for both the prediction and correction steps. This can be proven directly from the Chapman-Kolmogorov equation for two multivariate Gaussians that remains a multivariate Gaussian and for Bayes rule, the multivariate Gaussian is a conjugate prior of the multivariate Gaussian likelihood. However, this section will use two properties for multivariate Gaussians in the Kalman filter derivation which illustrate the relationships between joint multivariate Gaussians, their marginals, and their conditionals.

The first property is if  $\vec{X}$  is a multivariate Gaussian distributed as

$$\vec{X} \sim N(\vec{\mu}_x, \Sigma_x) \quad (15.41)$$

and  $\vec{Y}$  is a conditional multivariate Gaussian distributed as

$$\vec{Y}|\vec{X} \sim N(M\vec{x} + \vec{b}, \Sigma_b) \quad (15.42)$$

then, one has a joint multivariate Gaussian distribution as

$$\begin{bmatrix} \vec{X} \\ \vec{Y} \end{bmatrix} \sim N\left(\begin{bmatrix} \vec{\mu}_x \\ M\vec{\mu}_x + \vec{b} \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_x M^T \\ M\Sigma_x & M\Sigma_x M^T + \Sigma_b \end{bmatrix}\right) \quad (15.43)$$

and a marginal multivariate Gaussian distribution as

$$\vec{Y} \sim \mathcal{N} \left( M\vec{\mu}_x + \vec{b}, M\Sigma_x M^T + \Sigma_b \right) \quad (15.44)$$

The second property is if  $\vec{X}$  and  $\vec{Y}$  are joint multivariate Gaussians distributed as

$$\begin{bmatrix} \vec{X} \\ \vec{Y} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \vec{\mu}_x \\ \vec{\mu}_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{bmatrix} \right) \quad (15.45)$$

then, one has marginal Gaussians distributed as

$$\vec{X} \sim \mathcal{N} (\vec{\mu}_x, \Sigma_x) \quad (15.46)$$

and

$$\vec{Y} \sim \mathcal{N} (\vec{\mu}_y, \Sigma_y) \quad (15.47)$$

and conditional multivariate Gaussians distributed as

$$\vec{X}|\vec{Y} \sim \mathcal{N} \left( \vec{\mu}_x + \Sigma_{xy}\Sigma_y^{-1}(\vec{y} - \vec{\mu}_y), \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^T \right) \quad (15.48)$$

and

$$\vec{Y}|\vec{X} \sim \mathcal{N} \left( \vec{\mu}_y + \Sigma_{xy}^T\Sigma_x^{-1}(\vec{x} - \vec{\mu}_x), \Sigma_y - \Sigma_{xy}^T\Sigma_x^{-1}\Sigma_{xy} \right) \quad (15.49)$$

### Discrete-Time Kalman Filter

The discrete-time Kalman filter recursion assumes one has knowledge of the state filtering posterior at the previous time step  $k-1$ , i.e.,

$$\vec{X}_{k-1}|\vec{Y}_{1:k-1} \sim \mathcal{N} (\vec{\mu}_{k-1|k-1}, P_{k-1|k-1}) \quad (15.50)$$

which for  $k=1$ , one has the state prior  $\vec{X}_0 \sim \mathcal{N}(\vec{\mu}_0, P_0)$ . Note that this section and the following sections on Kalman filtering has denoted these recursions by the subscript “ $i|j$ ” which denotes the hyper-parameter update at step  $i$  given measurement updates, i.e., correction steps, up to time step  $j$ . For the prediction step derivation, consider the previous posterior state distribution

$$\vec{X}_{k-1}|\vec{Y}_{1:k-1} \sim \mathcal{N} (\vec{\mu}_x, \Sigma_x) \quad (15.51)$$

and the conditional state transition distributed as

$$\vec{X}_k|\vec{X}_{k-1}, \vec{Y}_{1:k-1} \sim \mathcal{N} (F_{k-1}\vec{x}_{k-1} + G_{k-1}\vec{u}_{k-1}, Q_{k-1}) \quad (15.52)$$

Then, the first property noted previously provides the state filtering prior distribution as

$$\vec{X}_k|\vec{Y}_{1:k-1} \sim \mathcal{N} (\vec{\mu}_{k|k-1}, P_{k|k-1}) \quad (15.53)$$

where

$$\vec{\mu}_{k|k-1} = F_{k-1}\vec{\mu}_{k-1} + G_{k-1}\vec{u}_{k-1} \quad (15.54)$$

and

$$P_{k|k-1} = F_{k-1} P_{k-1} F_{k-1}^T + Q_{k-1} \quad (15.55)$$

For the correction step derivation, consider also the measurement likelihood distributed as

$$\vec{Y}_k | \vec{X}_k, \vec{Y}_{1:k-1} \sim \mathcal{N}(H_k \vec{x}_k, R_k) \quad (15.56)$$

Then, the first property noted previously provides the joint distribution of the state and measurement as

$$\vec{X}_k, \vec{Y}_k | \vec{Y}_{1:k-1} = \mathcal{N}\left(\begin{bmatrix} \vec{\mu}_{k|k-1} \\ H_k \vec{\mu}_{k|k-1} \end{bmatrix}, \begin{bmatrix} P_{k|k-1} & P_{k|k-1} H_k^T \\ H_k P_{k|k-1} & H_k P_{k|k-1} H_k^T + R_k \end{bmatrix}\right) \quad (15.57)$$

and the measurement filtering posterior distribution as

$$\vec{Y}_k | \vec{Y}_{1:k-1} \sim \mathcal{N}(H_k \vec{\mu}_{k|k-1}, S_k) \quad (15.58)$$

where

$$S_k = H_k P_{k|k-1} H_k^T + R_k \quad (15.59)$$

Also, the second property noted previously for the joint distribution further provides the state filtering posterior distribution as

$$\vec{X}_k | \vec{Y}_{1:k} \sim \mathcal{N}(\vec{\mu}_{k|k}, P_{k|k}) \quad (15.60)$$

where

$$\vec{\mu}_{k|k} = \vec{\mu}_{k|k-1} + P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + R_k)^{-1} (\vec{y}_k - H_k \vec{\mu}_{k|k-1}) \quad (15.61)$$

and

$$P_{k|k} = P_{k|k-1} - H_k P_{k|k-1} (H_k P_{k|k-1} H_k^T + R_k)^{-1} P_{k|k-1} H_k^T \quad (15.62)$$

Furthermore, as the mode, mean, and median of a multivariate Gaussian is its mean hyper-parameter, the Kalman filter assumes that the mean recursion is also the state estimate recursion, i.e.,  $\hat{\vec{x}} = \vec{\mu}$  with the state estimate covariance,  $P$ . Thus, from a Bayesian perspective, the discrete-time Kalman filter recursively updates the ‘‘hyper-parameters’’ of the multivariate Gaussian PDF model for the state filtering prior PDF as

$$f_{\vec{X}_k | \vec{Y}_{1:k-1}}(\vec{x}_k | \vec{y}_{1:k-1}) = f_{\mathcal{N}}(\vec{x}_k; \hat{\vec{x}}_{k|k-1}, P_{k|k-1}) \quad (15.63)$$

the measurement filtering posterior PDF as

$$f_{\vec{Y}_k | \vec{Y}_{1:k-1}}(\vec{y}_k | \vec{y}_{1:k-1}) = f_{\mathcal{N}}(\vec{y}_k; H_k \hat{\vec{x}}_{k|k-1}, S_k) \quad (15.64)$$

and the state filtering posterior PDF as

$$f_{\vec{X}_k | \vec{Y}_{1:k}}(\vec{x}_k | \vec{y}_{1:k}) = f_{\mathcal{N}}(\vec{x}_k; \hat{\vec{x}}_{k|k}, P_{k|k}) \quad (15.65)$$

through algebraic recursions for the means and covariances based on the assumed zero-mean noise covariances,  $Q_k$  and  $R_k$ , the state, input, and measurement matrices,  $F_k$ ,  $G_k$ , and  $H_k$ , and the observed inputs and measurements,  $\vec{u}_k$  and  $\vec{y}_k$ , as functions of the time step,  $k$ .

In summary, the discrete-time **KF initialization step** sets the state estimate and state covariance at  $k = 0$  as

$$\begin{aligned}\hat{\vec{x}}_{0|0} &= \hat{\vec{x}}_0 \\ P_{0|0} &= P_0\end{aligned}\tag{15.66}$$

The discrete-time **KF prediction step** predicts the state estimate, the state covariance, and the output using the input and state-space model parameters as

$$\begin{aligned}\hat{\vec{x}}_{k|k-1} &= F_{k-1} \hat{\vec{x}}_{k-1|k-1} + G_{k-1} \vec{u}_{k-1} \\ P_{k|k-1} &= F_{k-1} P_{k-1|k-1} F_{k-1}^T + Q_{k-1} \\ \hat{\vec{y}}_k &= H_k \hat{\vec{x}}_{k|k-1}\end{aligned}\tag{15.67}$$

The discrete-time **KF correction step** updates the state estimator and the state covariance using the measurement and state-space model parameters as

$$\begin{aligned}\tilde{\vec{y}}_k &= \vec{y}_k - \hat{\vec{y}}_k \\ S_k &= H_k P_{k|k-1} H_k^T + R_k \\ K_k &= P_{k|k-1} H_k^T S_k^{-1} \\ \hat{\vec{x}}_{k|k} &= \hat{\vec{x}}_{k|k-1} + K_k \tilde{\vec{y}}_k \\ P_{k|k} &= P_{k|k-1} - K_k S_k K_k^T\end{aligned}\tag{15.68}$$

where  $\tilde{\vec{y}}_k$  is the zero-mean **innovation process**, also known as the “pre-fit residual,” with  $S_k$  as its **innovation covariance**. By inspection, the KF correction step is the same as recursive least-squares parameter estimator while the additional prediction step provides the least-squares predictor as well. Furthermore, by extension of least-squares analysis, the Kalman filter is the MMSE, the MVUE, and the BLUE for the linear state-space model without the explicit Gaussian noise assumption. One simply only requires knowledge of exact zero-mean mean and covariances for the noise statistics and model matrices.

Furthermore, note that by definition of the posterior state covariance, one has

$$P_{k|k} = \text{Cov} \left( \vec{x}_k - \hat{\vec{x}}_{k|k} \right)\tag{15.69}$$

and by substitution for  $\hat{\vec{x}}_{k|k} = \hat{\vec{x}}_{k|k-1} + K_k \tilde{\vec{y}}_k$ ,  $\tilde{\vec{y}}_k = \vec{y}_k - H_k \hat{\vec{x}}_{k|k-1}$ , and  $\vec{y}_k = H_k \vec{x}_k + \vec{v}_k$ , one has

$$P_{k|k} = \text{Cov} \left( \vec{x}_k - \hat{\vec{x}}_{k|k-1} + K_k \left( H_k \vec{x}_k + \vec{v}_k - H_k \hat{\vec{x}}_{k|k-1} \right) \right)\tag{15.70}$$

$$P_{k|k} = \text{Cov} \left( (I - K_k H_k) (\vec{x}_k - \hat{\vec{x}}_{k|k-1}) - K_k \vec{v}_k \right)\tag{15.71}$$

$$P_{k|k} = (I - K_k H_k) \text{Cov} \left( \vec{x}_k - \hat{\vec{x}}_{k|k-1} \right) (I - K_k H_k)^T + K_k \text{Cov}(\vec{v}_k) K_k^T\tag{15.72}$$

Then, by definition, one has the **Joseph form** of the state estimate covariance posterior correction recursion

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} (I - K_k H_k)^T + K_k R_k K_k^T\tag{15.73}$$

which holds for any gain  $K_k$ .

Notably, one can expand the Joseph form out as

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} - P_{k|k-1} H_k^T K_k^T + K_k H_k P_{k|k-1} H_k^T K_k^T + K_k R_k K_k^T \quad (15.74)$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} - P_{k|k-1} H_k^T K_k^T + K_k (H_k P_{k|k-1} H_k^T + R_k) K_k^T \quad (15.75)$$

By substitution of the variance-minimizing  $k_k$  and  $S_k = S_k^T$ , one has

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} - P_{k|k-1} H_k^T S_k^{-1} H_k P_{k|k-1} + P_{k|k-1} H_k^T S_k^{-1} H_k P_{k|k-1} \quad (15.76)$$

one obtains another alternative state estimate covariance update equation as

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} \quad (15.77)$$

Furthermore, note that combining these equations provides the **one-step prior discrete-time Kalman filter** as

$$\begin{aligned} \hat{\vec{y}}_k &= H_k \hat{\vec{x}}_{k|k-1} \\ K_k &= P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + R_k)^{-1} \\ \hat{\vec{x}}_{k+1|k} &= F_k \hat{\vec{x}}_{k|k-1} + G_k \vec{u}_k + F_k K_k (\vec{y}_k - H_k \hat{\vec{x}}_{k|k-1}) \\ P_{k+1|k} &= F_k (I - K_k H_k) P_{k|k-1} F_k^T + Q_k \end{aligned} \quad (15.78)$$

and the **one-step posterior discrete-time Kalman filter** as the Luenberger observer equations for the posterior mean recursion, the Riccati difference equation for the posterior covariance recursion

$$\begin{aligned} \hat{\vec{y}}_k &= H_k (F_{k-1} \hat{\vec{x}}_{k-1|k-1} + G_{k-1} \vec{u}_{k-1}) \\ K_k &= (F_{k-1} P_{k-1|k-1} F_{k-1}^T + Q_{k-1}) H_k^T (H_k (F_{k-1} P_{k-1|k-1} F_{k-1}^T + Q_{k-1}) H_k^T + R_k)^{-1} \\ \hat{\vec{x}}_{k|k} &= F_{k-1} \hat{\vec{x}}_{k-1|k-1} + G_{k-1} \vec{u}_{k-1} + K_k (\vec{y}_k - \hat{\vec{y}}_k) \\ P_{k|k} &= (I - K_k H_k) (F_{k-1} P_{k-1|k-1} F_{k-1}^T + Q_{k-1}) \end{aligned} \quad (15.79)$$

One can consider the Kalman gain  $K_k$  as the “transformed” relative weighting between previous  $P$ ,  $Q$ , and  $R$  to balance the uncertainty one has in the process model versus the measurement model when optimally estimating the state. To see this consider the simplified univariate example with

$$P_{k-1|k-1} = \sigma_x^2, \quad F = 1, \quad H = 1, \quad R = \sigma_R^2, \quad Q = \sigma_Q^2 \quad (15.80)$$

which provides

$$P_{k|k-1} = \sigma_x^2 + \sigma_Q^2 \quad (15.81)$$

$$S_k = \sigma_x^2 + \sigma_Q^2 + \sigma_R^2 \quad (15.82)$$

and

$$K_k = \frac{\sigma_x^2 + \sigma_Q^2}{\sigma_x^2 + \sigma_Q^2 + \sigma_R^2} \quad (15.83)$$

which is large when the covariance in the process is relatively larger than the covariance of the measurement, i.e.  $\sigma_R^2 \ll \sigma_Q^2$ .

In the previous discrete-time Kalman filter recursion, one assumes two key aspects about the AWGN. The first is that the process and measurement noise are uncorrelated with each other. The second is that the process and measurement noise are uncorrelated in time, i.e., white noise. If either of these assumptions are untrue, i.e., correlated-noise or colored noise, then the standard discrete-time Kalman filter recursion must be altered for these general noise cases which are discussed in following subsections using a least-squares approach to the estimator derivations.

### Discrete-Time Fixed-Interval Kalman Smoothing

The **fixed-interval Kalman smoother (FI-KS)**, also known as the **Rauch-Tung-Striebel smoother (RTSS)**, assumes one has knowledge of the state filtering prior distribution as

$$\vec{X}_{k+1} | \vec{Y}_{1:k} \sim \mathcal{N}(\hat{\vec{x}}_{k+1|k}, P_{k+1|k}) \quad (15.84)$$

the state filtering posterior distribution as

$$\vec{X}_k | \vec{Y}_{1:k} \sim \mathcal{N}(\hat{\vec{x}}_{k|k}, P_{k|k}) \quad (15.85)$$

and the state smoothing posterior distribution as

$$\vec{X}_k | \vec{Y}_{1:N} \sim \mathcal{N}(\hat{\vec{x}}_{k+1|N}, P_{k+1|N}) \quad (15.86)$$

which for  $k + 1 = N$ , one has the initial state smoothing posterior equivalent to the final state filtering posterior, i.e.,  $\vec{X}_N | \vec{Y}_{1:N} \sim \mathcal{N}(\hat{\vec{x}}_{N|N}, P_{N|N})$ .

To this end, consider the previous posterior state distribution

$$\vec{X}_k | \vec{Y}_{1:k} \sim \mathcal{N}(\hat{\vec{x}}_{k|k}, P_{k|k}) \quad (15.87)$$

and the conditional distribution

$$\vec{X}_{k+1} | \vec{X}_k, \vec{Y}_{1:k} \sim \mathcal{N}\left(F_k \hat{\vec{x}}_k + G_k \vec{u}_k, Q_k\right) \quad (15.88)$$

Then, the first property noted previously provides the joint distribution of consecutive states as

$$\begin{bmatrix} \vec{X}_k | \vec{Y}_{1:k} \\ \vec{X}_{k+1} | \vec{X}_k, \vec{Y}_{1:k} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \hat{\vec{x}}_{k|k} \\ \hat{\vec{x}}_{k+1|k} \end{bmatrix}, \begin{bmatrix} P_{k|k} & P_{k|k} F_k^T \\ F_k P_{k|k} & P_{k|k-1} \end{bmatrix}\right) \quad (15.89)$$

By the Markov assumption, the second element can be rewritten over the entire measurement history as

$$\begin{bmatrix} \vec{X}_k | \vec{Y}_{1:k} \\ \vec{X}_{k+1} | \vec{X}_k, \vec{Y}_{1:N} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \hat{\vec{x}}_{k|k} \\ \hat{\vec{x}}_{k+1|k} \end{bmatrix}, \begin{bmatrix} P_{k|k} & P_{k|k} F_k^T \\ F_k P_{k|k} & P_{k|k-1} \end{bmatrix}\right) \quad (15.90)$$

which by the second property noted previously for the joint distribution further provides the marginal distribution as

$$\vec{X}_k | \vec{X}_{k+1}, \vec{Y}_{1:N} \sim \mathcal{N}\left(\hat{\vec{x}}_{k|k} + K_{S,k}(\vec{x}_{k+1} - \hat{\vec{x}}_{k|k}), P_{k|k} - K_{S,k} P_{k|k-1} K_{S,k}^T\right) \quad (15.91)$$

with

$$K_{S,k} = P_{k|k} F_k^T P_{k|k-1}^{-1} \quad (15.92)$$

Then, for the joint smoothed distribution for consecutive states, one has

$$\begin{bmatrix} \vec{x}_{k+1} | \vec{Y}_{1:N} \\ \vec{x}_k | \vec{Y}_{1:N} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \hat{x}_{k+1|N} \\ \hat{x}_{k|k} + K_{S,k}(\hat{x}_{k+1|N} - \hat{x}_{k+1|k}) \end{bmatrix}, \begin{bmatrix} P_{k+1|N} & P_{k+1|N} K_{S,k}^T \\ K_{S,k} P_{k+1|N} K_{S,k}^T + P_{k|k} - K_{S,k} P_{k|k-1} K_{S,k}^T \end{bmatrix} \right) \quad (15.93)$$

and by the second property noted previously provides the state smoothing posterior distributed as

$$\vec{x}_k | \vec{Y}_{1:N} \sim \mathcal{N} \left( \hat{x}_{k|k} + K_{S,k}(\hat{x}_{k+1|N} - \hat{x}_{k+1|k}), P_{k|k} + K_{S,k}(P_{k+1|N} - P_{k|k-1}) K_{S,k}^T \right) \quad (15.94)$$

where one can alternatively keep in memory the state filtering posterior mean and covariance and the input, i.e.,

$$\vec{x}_k | \vec{Y}_{1:N} \sim \mathcal{N} \left( \hat{x}_{k|k} + K_{S,k}(\hat{x}_{k+1|N} - F_k \hat{x}_{k|k} - G_k \vec{u}_k), P_{k|k} + K_{S,k}(P_{k+1|N} - F_k P_{k|k} F_k^T - Q_k) K_{S,k}^T \right) \quad (15.95)$$

Thus, the discrete-time **FI-KS smoothing step** or the **RTSS smoothing step** updates the state estimate and the state covariance using the posterior state estimate and covariance and the prior state estimate and covariance or the measurement noise covariance as

$$\begin{aligned} \tilde{x}_k &= \hat{x}_{k+1|N} - \hat{x}_{k+1|k} \\ &= \hat{x}_{k+1|N} - F_k \hat{x}_{k|k} - G_k \vec{u}_k \\ K_{S,k} &= P_{k|k} F_k^T P_{k+1|k}^{-1} \\ &= P_{k|k} F_k^T (F_k P_{k|k} F_k^T + Q_k)^{-1} \\ \hat{x}_{k|N} &= \hat{x}_{k|k} + K_{S,k} \tilde{x}_k \\ P_{k|N} &= P_{k|k} + K_{S,k}(P_{k+1|N} - P_{k+1|k}) K_{S,k}^T \\ &= P_{k|k} + K_{S,k}(P_{k+1|N} - F_k P_{k|k} F_k^T - Q_k) K_{S,k}^T \end{aligned} \quad (15.96)$$

## Sequential Kalman Filtering

One alternative to reducing the inversion computational complexity and the memory requirements for large  $n_y$  is to sequentially correct for subsets of measurements which is possible if the measurement noise covariance is block diagonal with  $r$  blocks, i.e.,

$$R_k = \begin{bmatrix} R_{k,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & R_{k,r} \end{bmatrix} \quad (15.97)$$

which invites the following partitioning of the measurement equation of the linear-Gaussian state-space model as

$$\begin{bmatrix} \vec{y}_{k,1} \\ \vdots \\ \vec{y}_{k,r} \end{bmatrix} = \begin{bmatrix} H_{k,1} \\ \vdots \\ H_{k,r} \end{bmatrix} \vec{x}_k + \begin{bmatrix} \vec{v}_{k,1} \\ \vdots \\ \vec{v}_{k,r} \end{bmatrix} \quad (15.98)$$

with  $\vec{v}_{k,i} \sim \mathcal{N}(0, R_{k,i})$  for  $i = 1, \dots, r$ .

Thus, the **discrete-time sequential Kalman filter (SKF)** algorithm has a modified correction step which is initialized with the prior state estimate and covariance

$$\begin{aligned}\hat{x}_{k|k-1,0} &= \hat{x}_{k|k-1} \\ P_{0,k|k-1} &= P_{k|k-1}\end{aligned}\tag{15.99}$$

and runs  $i = 1, \dots, r$  iterations of

$$\begin{aligned}S_{k,i} &= H_{k,i}P_{k|k-1,i-1}H_{k,i}^T + R_{k,i} \\ K_{k,i} &= P_{k|k-1,i-1}H_{k,i}^TS_{k,i}^{-1} \\ \hat{x}_{k|k-1,i} &= \hat{x}_{k|k-1,i} + K_{k,i}(\vec{y}_{k,i} - H_{k,i}\hat{x}_{k|k-1,i}) \\ P_{k|k-1,i} &= (I - K_{k,i}H_{k,i})P_{i-1,k|k-1}\end{aligned}\tag{15.100}$$

and assigns for the posterior state estimate and covariance

$$\begin{aligned}\hat{x}_{k|k} &= \hat{x}_{k|k-1,r} \\ P_{0,k|k-1} &= P_{k|k-1,r}\end{aligned}\tag{15.101}$$

Note that if  $R$  is diagonal, then no inversions are necessary. Furthermore, if  $R_k = R$ , one can obtain a Jordan form decomposition for  $\tilde{R} = V^T RV$  with  $\tilde{R}$  containing block matrices which provides the transformed measurement model

$$\tilde{\vec{y}} = \tilde{H}_k \vec{x}_k + \tilde{\vec{v}}\tag{15.102}$$

with  $\tilde{H}_k = V^T H_k$  and  $\tilde{\vec{v}} \sim \mathcal{N}(0, \tilde{R})$  which can be used in the SKF.

## Information Filtering

Another alternative recursion for the Kalman filter to change the inversion computation is to propagate the inverse of the state covariance matrix,  $P$ , i.e., the **information matrix**,  $\mathcal{I} = P^{-1}$  which forms the **information filter (IF)**. then, by substitution, the discrete-time correction step update for the information matrix is

$$\mathcal{I}_{k|k} = \mathcal{I}_{k|k-1} + H_k^T R_k^{-1} H_k\tag{15.103}$$

and the discrete-time prediction step update for the information matrix is

$$\mathcal{I}_{k|k-1} = \left( F_{k-1} \mathcal{I}_{k-1|k-1}^{-1} F_k^T + Q_{k-1} \right)^{-1}\tag{15.104}$$

Next, applying the matrix inversion lemma, i.e.,

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1}\tag{15.105}$$

where  $A = Q_{k-1}$ ,  $B = F_{k-1}$ ,  $C = \mathcal{I}^{-1}$ , and  $D = F_{k-1}^T$ , one has

$$\mathcal{I}_{k|k-1} = Q_{k-1}^{-1} - Q_{k-1}^{-1} F_{k-1} \left( \mathcal{I}_{k-1|k-1} + F_{k-1}^T Q_{k-1}^{-1} F_{k-1} \right) F_{k-1}^T Q_{k-1}^{-1}\tag{15.106}$$

The information filter equations requires a couple of  $n_x \times n_x$  inversions. Therefore, if  $n_y \gg n_x$ , it may be more efficient to use the information filter. Also, if  $R_k$  or  $Q_{k-1}$  are constant with respect to  $k$ , then one could invert those as part of the initialization process. Furthermore, if the initial uncertainty is zero, one can numerically set  $\mathcal{I}_0 = \infty$ . If the initial uncertainty is  $\infty$ , one can numerically set  $P_0 = \infty$ .

In summary, the **IF prediction step** can be written as

$$\begin{aligned}\hat{x}_{k|k-1} &= F_{k-1}\hat{x}_{k-1} + G_k \vec{u}_{k-1} \\ \mathcal{I}_{k|k-1} &= Q_{k-1}^{-1} - Q_{k-1}^{-1}F_{k-1} \left( \mathcal{I}_{k-1|k-1} + F_{k-1}^T Q_{k-1}^{-1} F_{k-1} \right) F_{k-1}^T Q_{k-1}^{-1} \\ \hat{\vec{y}}_k &= H_k \hat{x}_{k|k-1}\end{aligned}\quad (15.107)$$

and the **IF correction step** can be written as

$$\begin{aligned}\tilde{\vec{y}}_k &= \vec{y}_k - \hat{\vec{y}}_k \\ \mathcal{I}_{k|k} &= \mathcal{I}_{k|k-1} + H_k^T R_k^{-1} H_k \\ K_k &= \mathcal{I}_{k|k}^{-1} H_k^T R_k^{-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k \tilde{\vec{y}}_k\end{aligned}\quad (15.108)$$

### Square-Root Kalman Filtering

In Kalman filtering, it can be advantageous to propagate the square-root of the covariance matrix,  $P_k^{1/2}$ , as opposed to the covariance matrix,  $P_k$ . This advantage comes from the increase in numerical precision for the inversions required in the Kalman filter equations, but at the cost of additional computational resources.

To derive the prediction step for the **square-root Kalman filter (SR-KF)**, suppose one can find a  $2n_x \times 2n_x$  matrix

$$T = [T_1 \ T_2] \quad (15.109)$$

such that one could write the **SR-KF prediction step** for  $P_{k|k-1}^{T/2}$  as

$$\begin{bmatrix} P_{k|k-1}^{T/2} \\ 0 \end{bmatrix} = T \begin{bmatrix} P_{k-1|k-1}^{T/2} F^T \\ Q^{T/2} \end{bmatrix} \quad (15.110)$$

then,

$$\begin{bmatrix} P_{k|k-1}^{T/2} \\ 0 \end{bmatrix} = T_1 P_{k-1|k-1}^{T/2} F^T + T_2 Q^{T/2} \quad (15.111)$$

Furthermore, if  $T$  is prescribed as orthogonal, one has

$$T^T T = \begin{bmatrix} T_1^T \\ T_2^T \end{bmatrix} [T_1 \ T_2] = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \quad (15.112)$$

which allows one to write

$$\begin{bmatrix} P_{k|k-1}^{1/2} & 0 \end{bmatrix} \begin{bmatrix} P_{k|k-1}^{1/2} \\ 0 \end{bmatrix} = \left[ T_1 P_{k-1|k-1}^{T/2} F^T + T_2 Q^{T/2} \right]^T \left[ T_1 P_{k-1|k-1}^{T/2} F^T + T_2 Q^{T/2} \right] \quad (15.113)$$

which allows one to write

$$P_{k|k-1}^{1/2} P_{k|k-1}^{T/2} = F P_{k-1|k-1}^{1/2} T_1^T T_1 P_{k-1|k-1}^{T/2} F^T + Q^{1/2} T_2^T T_2 Q^{T/2} \quad (15.114)$$

which implies by the orthogonal properties and square-root properties that

$$P_{k|k-1} = F P_{k-1|k-1} F^T + Q \quad (15.115)$$

which is the prediction step update of the covariance and proves the square-root prediction step is equivalent. However, one requires methods for finding  $T$  and  $P_{k|k-1}^{T/2}$  matrices that solve the square-root prediction step use various methods from linear algebra, e.g. Householder, Gram-Schmidt, modified Gram-Schmidt, or Givens transformations.

For the correction step of the square-root filter, suppose one can find a  $(n_x + n_y) \times (n_x + n_y)$  matrix,

$$\tilde{T} = \begin{bmatrix} \tilde{T}_{11} & \tilde{T}_{12} \\ \tilde{T}_{21} & \tilde{T}_{22} \end{bmatrix} \quad (15.116)$$

such that one can write the **SR-KF correction step** for  $P_{k|k-1}^T$  as

$$\begin{bmatrix} (R + H P_{k|k-1} H^T)^{T/2} & \tilde{K}_k \\ 0 & P_{k|k}^{T/2} \end{bmatrix} = \tilde{T} \begin{bmatrix} R^{T/2} & 0 \\ P_{k|k-1}^T H^T & P_{k|k-1}^T \end{bmatrix} \quad (15.117)$$

where

$$\tilde{K}_k = K_k (R + H P_{k|k-1} H^T)^{T/2} \quad (15.118)$$

and the lower right matrix of the block matrix on the left side, i.e.  $P_{k|k}^{T/2}$ , cannot be found until one can solve for  $\tilde{T}$  that provides the other 3 matrices in the block matrix on the left side. Furthermore, if  $\tilde{T}$  is prescribed as orthogonal, one has

$$\tilde{T}^T \tilde{T} = \begin{bmatrix} \tilde{T}_{11}^T & \tilde{T}_{21}^T \\ \tilde{T}_{12}^T & \tilde{T}_{22}^T \end{bmatrix} \begin{bmatrix} \tilde{T}_{11} & \tilde{T}_{12} \\ \tilde{T}_{21} & \tilde{T}_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \quad (15.119)$$

which allows one to write

$$\begin{bmatrix} (R + H P_{k|k-1} H^T)^{T/2} & \tilde{K}_k \\ 0 & P_{k|k}^{T/2} \end{bmatrix} = \begin{bmatrix} \tilde{T}_{11} R^{T/2} + \tilde{T}_{12} P_{k|k-1}^{T/2} H^T & \tilde{T}_{12} P_{k|k-1}^T \\ \tilde{T}_{21} R^{T/2} + \tilde{T}_{22} P_{k|k-1}^{T/2} H^T & \tilde{T}_{22} P_{k|k-1}^T \end{bmatrix} \quad (15.120)$$

which can be written as four separate equalities by pre-multiplying by its transpose. The first column provides two equations

$$\begin{aligned} (R + H P_{k|k-1} H^T)^{1/2} (R + H P_{k|k-1} H^T)^{T/2} = \\ R^{1/2} \tilde{T}_{11}^T \tilde{T}_{11} R^{T/2} + H P_{k|k-1}^{1/2} \tilde{T}_{12} \tilde{T}_{11} R^{T/2} \\ + R^{1/2} \tilde{T}_{11}^T \tilde{T}_{12} P_{k|k-1}^{T/2} H^T + H P_{k|k-1}^{1/2} \tilde{T}_{12}^T \tilde{T}_{12} P_{k|k-1}^{T/2} H^T \end{aligned} \quad (15.121)$$

and

$$\begin{aligned} 0 = R^{1/2} \tilde{T}_{21}^T \tilde{T}_{21} R^{T/2} + R^{1/2} \tilde{T}_{21}^T \tilde{T}_{22} P_{k|k-1}^T H^T \\ + H P_{k|k-1}^{1/2} \tilde{T}_{22}^T \tilde{T}_{21} R^{T/2} + H P_{k|k-1}^{1/2} \tilde{T}_{22}^T \tilde{T}_{22} P_{k|k-1}^T H^T \end{aligned} \quad (15.122)$$

which adding together and simplifying with the transpose properties provides

$$R + HP_{k|k-1}H^T = R + HP_{k|k-1}^{1/2}P_{k|k-1}^{T/2}H^T \quad (15.123)$$

which shows the square-root property for  $P_{k|k-1}^{1/2}$ . From the second column, one has

$$\tilde{K}_k \tilde{K}_k^T = P_{k|k-1}^{1/2} \tilde{T}_{12}^T \tilde{T}_{12} P_{k|k-1}^{T/2} \quad (15.124)$$

and

$$P_{k|k}^{1/2} P_{k|k}^{T/2} = P_{k|k-1}^{1/2} \tilde{T}_{22}^T \tilde{T}_{22} P_{k|k-1}^{T/2} \quad (15.125)$$

which adding together and simplifying with the transpose properties provides

$$P_{k|k} + K_k (R + HP_{k|k-1}H^T) K_k^T = P_{k|k-1}^{1/2} P_{k|k-1}^{T/2} \quad (15.126)$$

and substituting for the Kalman gain  $K_k$  from the standard equation provides

$$P_{k|k} + P_{k|k-1}H^T K_k^T = P_{k|k-1} \quad (15.127)$$

or rearranging and using the fact that each term is symmetric, one has

$$P_{k|k} == P_{k|k-1} - K_k H P_{k|k-1} \quad (15.128)$$

which is the correction step update of the covariance and proves the square-root correction step is equivalent. However, one requires methods for finding  $\tilde{T}$  and  $P_{k|k}^{1/2}$  matrices that solve the square-root correction step.

### Discrete-Time Fixed-Lag Kalman Smoothing

The **fixed-lag Kalman smoother (FL-KS) step** recursion augments the state-space model by  $N + 1$  stacked state vectors to obtain the model

$$\begin{aligned} \begin{bmatrix} \vec{x}_{k+1} \\ \vec{x}_k \\ \vdots \\ \vec{x}_{k-N} \end{bmatrix} &= \begin{bmatrix} F_k & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix} \begin{bmatrix} \vec{x}_k \\ \vec{x}_{k-1} \\ \vdots \\ \vec{x}_{k-N-1} \end{bmatrix} + \begin{bmatrix} G_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} \vec{u}_k + \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \vec{w}_k \\ \vec{y}_k &= [H_k \ 0 \ \cdots \ 0] \begin{bmatrix} \vec{x}_k \\ \vec{x}_k \\ \vdots \\ \vec{x}_{k-N} \end{bmatrix} + \vec{v}_k \end{aligned} \quad (15.129)$$

Next, consider the one-step prior discrete-time Kalman filter recursion as

$$\begin{aligned} K_k &= P_{k|k-1}H_k^T (H_k P_{k|k-1}H_k^T + R_k)^{-1} \\ \hat{\vec{x}}_{k+1|k} &= F_k \hat{\vec{x}}_{k|k-1} + G_k \vec{u}_k + F_k K_k (\vec{y}_k - H_k \hat{\vec{x}}_{k|k-1}) \\ P_{k+1|k} &= F_k P_{k|k-1} (I - H_k^T K_k^T) F_k^T + Q_k \end{aligned} \quad (15.130)$$

Then, the one-step prior Kalman filter state estimate for the augmented state-space model can be written as

$$\begin{aligned} \begin{bmatrix} \hat{x}_{k+1|k} \\ \hat{x}_{k|k} \\ \vdots \\ \hat{x}_{k-N|k} \end{bmatrix} &= \begin{bmatrix} F_k & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix} \begin{bmatrix} \hat{x}_{k|k-1} \\ \hat{x}_{k-1|k-1} \\ \vdots \\ \hat{x}_{k-N-1|k-1} \end{bmatrix} + \begin{bmatrix} G_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} \vec{u}_k \\ &+ \begin{bmatrix} F_k & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix} \begin{bmatrix} K_{k,0} \\ K_{k,1} \\ \vdots \\ K_{k,N+1} \end{bmatrix} \left( \vec{y}_k - [H_k \quad 0 \quad \cdots \quad 0] \begin{bmatrix} \hat{x}_{k|k-1} \\ \hat{x}_{k-1|k-1} \\ \vdots \\ \hat{x}_{k-N-1|k-1} \end{bmatrix} \right) \end{aligned} \quad (15.131)$$

or

$$\begin{bmatrix} \hat{x}_{k+1|k} \\ \hat{x}_{k|k} \\ \hat{x}_{k-1|k} \\ \vdots \\ \hat{x}_{k-N|k} \end{bmatrix} = \begin{bmatrix} F_k \hat{x}_{k|k-1} + G_k \vec{u}_k + F_k K_{k,0} (\vec{y}_k - H_k \hat{x}_{k|k-1}) \\ \hat{x}_{k|k-1} + K_{k,0} (\vec{y}_k - H_k \hat{x}_{k|k-1}) \\ \hat{x}_{k-1|k-1} + K_{k,1} (\vec{y}_k - H_k \hat{x}_{k|k-1}) \\ \vdots \\ \hat{x}_{k-N|k-1} + K_{k,N} (\vec{y}_k - H_k \hat{x}_{k|k-1}) \end{bmatrix} \quad (15.132)$$

and notably the first entry for this construction provides the one-step Kalman filter state estimate equation the second entry for this construction provides the Kalman filter posterior state estimate correction,  $\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_{k,0}(\vec{y}_k - H_k \hat{x}_{k|k-1})$  with  $K_{k,0} = K_k$ , which is also used to update previous posterior states to step  $k - N$  given measurements up to time step  $k$ . The additional Kalman gains adjust the previous state estimates with the new innovation  $\vec{y}_k - H_k \hat{x}_{k|k-1}$ .

Next, the one-step prior Kalman filter stacked Kalman gains for the augmented state-space model

$$\begin{aligned} \begin{bmatrix} F_k K_{k,0} \\ K_{k,0} \\ \vdots \\ K_{k,N} \end{bmatrix} &= \begin{bmatrix} F_k & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix} \begin{bmatrix} P_{k|k-1} & P_{k,k-1|k-1}^T & \cdots & P_{k,k-N-1|k-1}^T \\ P_{k,k-1|k-1} & P_{k-1|k-1} & \cdots & P_{k,k-N-1|k-1}^T \\ \vdots & \vdots & \ddots & \vdots \\ P_{k,k-N-1|k-1} & P_{k,k-N-1|k-1} & \cdots & P_{k-N-1|k-1} \end{bmatrix} \begin{bmatrix} H_k^T \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &\quad \left( \begin{bmatrix} H_k & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} P_{k|k-1} & P_{k,k-1|k-1}^T & \cdots & P_{k,k-N-1|k-1}^T \\ P_{k,k-1|k-1} & P_{k-1|k-1} & \cdots & P_{k,k-N-1|k-1}^T \\ \vdots & \vdots & \ddots & \vdots \\ P_{k+1,k-N-1|k-1} & P_{k,k-N-1|k-1} & \cdots & P_{k-N-1|k-1} \end{bmatrix} \begin{bmatrix} H_k^T \\ 0 \\ \vdots \\ 0 \end{bmatrix} + R_k \right)^{-1} \end{aligned} \quad (15.133)$$

which simplifies to

$$\begin{bmatrix} F_k K_{k,0} \\ K_{k,0} \\ K_{k,1} \\ \vdots \\ K_{k,N} \end{bmatrix} = \begin{bmatrix} F_k P_{k|k-1} \\ P_{k|k-1} \\ P_{k,k-1|k-1} \\ \vdots \\ P_{k,k-N|k-1} \end{bmatrix} H_k^T (H_k P_{k|k-1} H_k^T + R_k)^{-1} \quad (15.134)$$

where notably, the second entry,  $K_{k,0} = P_{k|k-1}H_k^T(H_kP_{k|k-1}H_k^T + R_k)^{-1}$  is the standard Kalman gain for the correction step at time step  $k$ . The additional Kalman gains use the cross-covariance between  $k$  and  $i = 1, \dots, N+1$  previous steps.

Then, one-step prior Kalman filter state covariance for the augmented state-space model can be written as

$$\begin{bmatrix} P_{k+1|k} & P_{k+1,k|k}^T & \cdots & P_{k+1,k-N|k}^T \\ P_{k+1,k|k} & P_{k|k} & \cdots & P_{k,k-N|k}^T \\ \vdots & \vdots & \ddots & \vdots \\ P_{k+1,k-N|k} & P_{k,k-N|k} & \cdots & P_{k-N|k} \end{bmatrix} = \begin{bmatrix} F_k & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix} \begin{bmatrix} P_{k|k-1} & P_{k,k-1|k-1}^T & \cdots & P_{k,k-N-1|k-1}^T \\ P_{k,k-1|k-1} & P_{k-1|k-1} & \cdots & P_{k-1,k-N-1|k-1}^T \\ \vdots & \vdots & \ddots & \vdots \\ P_{k,k-N-1|k-1} & P_{k,k-N-1|k-1} & \cdots & P_{k-N-1|k-1} \end{bmatrix} \\ \left( \begin{bmatrix} F_k & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix} - \begin{bmatrix} F_k K_{k,0} \\ K_{k,0} \\ \vdots \\ K_{k,N+1} \end{bmatrix} \begin{bmatrix} H_k & 0 & \cdots & 0 \\ K_{k,N+1} H_k & 0 & \cdots & 0 \end{bmatrix} \right)^T \begin{bmatrix} Q_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \quad (15.135)$$

or

$$\begin{bmatrix} P_{k+1|k} & P_{k+1,k|k}^T & \cdots & P_{k+1,k-N|k}^T \\ P_{k+1,k|k} & P_{k|k} & \cdots & P_{k,k-N-1|k}^T \\ \vdots & \vdots & \ddots & \vdots \\ P_{k+1,k-N|k} & P_{k,k-N|k} & \cdots & P_{k-N|k} \end{bmatrix} = \left( \begin{bmatrix} F_k & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix} - \begin{bmatrix} F_k K_{k,0} H_k & 0 & \cdots & 0 \\ K_{k,0} H_k & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ K_{k,N+1} H_k & 0 & \cdots & 0 \end{bmatrix} \right) \\ \begin{bmatrix} P_{k|k-1} F_k^T & P_{k|k-1} & \cdots & P_{k,k-N|k-1}^T \\ P_{k,k-1|k-1} F_k^T & P_{k,k-1|k-1} & \cdots & P_{k-1,k-N|k-1}^T \\ \vdots & \vdots & \ddots & \vdots \\ P_{k,k-N-1|k-1} F_k^T & P_{k,k-N-1|k-1} & \cdots & P_{k-N,k-N-1|k-1} \end{bmatrix} \\ + \begin{bmatrix} Q_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \quad (15.136)$$

By inspection, this provides the following recursions for the first column of the covariance matrix are

$$\begin{aligned} P_{k+1|k} &= F_k P_{k|k-1} (I - H_k^T K_{k,0}) F_k^T + Q_k \\ P_{k+1,k|k} &= P_{k|k-1} (I - H_k^T K_{k,0}) F_k^T \\ P_{k+1,k-1|k} &= P_{k,k-1|k-1} (I - H_k^T K_{k,0}) F_k^T \\ &\vdots = \vdots \\ P_{k+1,k-N|k} &= P_{k,k-N|k-1} (I - H_k^T K_{k,0}) F_k^T \end{aligned} \quad (15.137)$$

and the following recursions for the diagonal of the covariance matrix

$$\begin{aligned} P_{k|k} &= P_{k|k-1} - P_{k|k-1} H_k^T K_{k,0}^T \\ P_{k-1|k} &= P_{k-1|k-1} - P_{k|k-1} H_k^T K_{k,1}^T \\ P_{k-2|k} &= P_{k-2|k-1} - P_{k,k-2|k-1} H_k^T K_{k,1}^T \\ &\vdots = \vdots \\ P_{k-N|k} &= P_{k-N|k-1} - P_{k,k-N|k-1} H_k^T K_{k,N}^T \end{aligned} \quad (15.138)$$

Thus, the **FL-KS prediction step** predicts the state estimate, the state covariance, and the output using the input and state-space model parameters as

$$\begin{aligned} K_{k,0} &= P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + R_k)^{-1} \\ \hat{x}_{k+1|k} &= F_k \hat{x}_{k|k-1} + G_k \vec{u}_k + F_k K_{k,0} (\vec{y}_k - H_k \hat{x}_{k|k-1}) \\ P_{k+1|k} &= F_k P_{k|k-1} (I - H_k^T K_{k,0}^T) F_k^T + Q_k \end{aligned} \quad (15.139)$$

and the **FL-KS smoothing step** updates the state estimate and the state covariance using the prior state estimate and covariance at time step  $k$  for  $i = i, \dots, N + 1$  as

$$\begin{aligned} K_{k,i} &= P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + R_k)^{-1} \\ \hat{x}_{k+1-i|k} &= \hat{x}_{k+2-i|k} + K_{k,i} (\vec{y}_k - \hat{y}_k) \\ P_{k+1-i|k} &= P_{k-i|k-1} - P_{k+1,k-i|k-1} H_k^T K_{k,i}^T F_k^T \\ P_{k+1,k-i|k} &= P_{k,k-1|k-1} (I - H_k^T K_{k,0}^T) F_k^T \end{aligned} \quad (15.140)$$

where  $P_{k,k|k-1} = P_{k|k-1}$  for ease of notation and, notably, the first three equations for  $i = 1$  forms the standard KF correction step at time step  $k$ .

### Discrete-Time Correlated-Noise Kalman Filtering

Consider the case where the process and measurement noise are correlated, but without colored noise, i.e.,

$$\begin{aligned} \mathbb{E}[\vec{w}_k \vec{w}_j^T] &= Q_k \delta_{k-j} \\ \mathbb{E}[\vec{v}_k \vec{v}_j^T] &= R_k \delta_{k-j} \\ \mathbb{E}[\vec{w}_k \vec{v}_j^T] &= N_k \delta_{k-j+1} \end{aligned} \quad (15.141)$$

where  $N_k$  is the **noise cross-covariance** and models the expected correlation/covariance between the zero-mean noise vectors. To derive the correlated-noise Kalman filter (CN-KF), assume the same form for the prior and posterior state estimates

$$\begin{aligned} \hat{x}_{k|k-1} &= F_{k-1} \hat{x}_{k-1|k-1} + G_{k-1} \vec{u}_{k-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k (\vec{y}_k - H_k \hat{x}_{k|k-1}) \end{aligned} \quad (15.142)$$

where the Kalman gain will be different than the standard KF as well as the posterior state covariance.

The prior state covariance by definition is

$$P_{k|k-1} = \mathbb{E} \left[ (\vec{x}_k - \hat{\vec{x}}_{k|k-1}) (\vec{x}_k - \hat{\vec{x}}_{k|k-1})^T \right] \quad (15.143)$$

By substitution for the linear state process and prior state estimate, one has

$$\begin{aligned} P_{k|k-1} &= \mathbb{E} \left[ \left( F_{k-1} \vec{x}_{k-1} + G_{k-1} \vec{u}_{k-1} + \vec{w}_{k-1} - F_{k-1} \hat{\vec{x}}_{k-1|k-1} - G_{k-1} \vec{u}_{k-1} \right) \right. \\ &\quad \left. \left( F_{k-1} \vec{x}_{k-1} + G_{k-1} \vec{u}_{k-1} + \vec{w}_{k-1} - F_{k-1} \hat{\vec{x}}_{k-1|k-1} - G_{k-1} \vec{u}_{k-1} \right)^T \right] \end{aligned} \quad (15.144)$$

$$\begin{aligned} P_{k|k-1} &= \mathbb{E} \left[ F_{k-1} (\vec{x}_k - \hat{\vec{x}}_{k-1|k-1}) (\vec{x}_k - \hat{\vec{x}}_{k-1|k-1})^T F_{k-1}^T + \vec{w}_{k-1} \vec{w}_{k-1}^T \right. \\ &\quad \left. + F_{k-1} (\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1}) \vec{w}_{k-1}^T + \vec{w}_{k-1} (\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1})^T F_{k-1}^T \right] \end{aligned} \quad (15.145)$$

By substitution of the posterior state covariance, process noise covariance, and that recalling that  $(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1})$  is uncorrelated with zero-mean  $\vec{w}_{k-1}$ , one has

$$P_{k|k-1} = F_{k-1} P_{k-1|k-1} F_{k-1}^T + Q_{k-1} \quad (15.146)$$

The posterior state covariance by definition is

$$P_{k|k} = \mathbb{E} \left[ (\vec{x}_k - \hat{\vec{x}}_{k|k}) (\vec{x}_k - \hat{\vec{x}}_{k|k})^T \right] \quad (15.147)$$

By substitution for the posterior state estimate as

$$\begin{aligned} P_{k|k} &= \mathbb{E} \left[ \left( \vec{x}_k - \hat{\vec{x}}_{k|k-1} - K_k (\vec{y}_k - H_k \hat{\vec{x}}_{k|k-1}) \right) \right. \\ &\quad \left. \left( \vec{x}_k - \hat{\vec{x}}_{k|k-1} - K_k (\vec{y}_k - H_k \hat{\vec{x}}_{k|k-1}) \right)^T \right] \end{aligned} \quad (15.148)$$

and substitution for the measurement, one has

$$\begin{aligned} P_{k|k} &= \mathbb{E} \left[ \left( \vec{x}_k - \hat{\vec{x}}_{k|k-1} - K_k (H_k \vec{x}_k + \vec{v}_k - H_k \hat{\vec{x}}_{k|k-1}) \right) \right. \\ &\quad \left. \left( \vec{x}_k - \hat{\vec{x}}_{k|k-1} - K_k (H_k \vec{x}_k + \vec{v}_k - H_k \hat{\vec{x}}_{k|k-1}) \right)^T \right] \end{aligned} \quad (15.149)$$

$$\begin{aligned} P_{k|k} &= \mathbb{E} \left[ \left( (I - K_k H_k) (\vec{x}_k - \hat{\vec{x}}_{k|k-1}) - K_k \vec{v}_k \right) \right. \\ &\quad \left. \left( (I - K_k H_k) (\vec{x}_k - \hat{\vec{x}}_{k|k-1}) - K_k \vec{v}_k \right)^T \right] \end{aligned} \quad (15.150)$$

$$P_{k|k} = \mathbb{E} \left[ (I - K_k H_k) (\vec{x}_k - \hat{\vec{x}}_{k|k-1}) (\vec{x}_k - \hat{\vec{x}}_{k|k-1})^T (I - K_k H_k)^T + K_k \vec{v}_k \vec{v}_k^T K_k^T \right. \\ \left. - K_k \vec{v}_k (\vec{x}_k - \hat{\vec{x}}_{k|k-1})^T (I - K_k H_k)^T - (I - K_k H_k) (\vec{x}_k - \hat{\vec{x}}_{k|k-1}) \vec{v}_k^T K_k^T \right] \quad (15.151)$$

By substitution for the prior covariance, measurement noise covariance, and the linear state process model from  $\vec{x}_{k-1}$  to  $\vec{x}_k$ , one has

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} (I - K_k H_k)^T + K_k R_k K_k^T \\ + K_k \mathbb{E} \left[ \vec{v}_k (F_{k-1} \vec{x}_{k-1} + G_{k-1} \vec{u}_{k-1} + \vec{w}_{k-1} - \hat{\vec{x}}_{k|k-1})^T \right] (I - K_k H_k)^T \\ - (I - K_k H_k) \mathbb{E} \left[ (F_{k-1} \vec{x}_{k-1} + G_{k-1} \vec{u}_{k-1} + \vec{w}_{k-1} - \hat{\vec{x}}_{k|k-1}) \vec{v}_k^T \right] K_k^T \quad (15.152)$$

Then, by the definition of the noise cross-covariance and using the fact that  $\vec{x}_{k-1}$ ,  $\vec{u}_{k-1}$ , and  $\hat{\vec{x}}_{k|k-1}$  are uncorrelated with zero-mean  $\vec{v}_{k-1}$ , one has

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} (I - K_k H_k)^T + K_k R_k K_k^T - K_k N_k^T (I - K_k H_k)^T - (I - K_k H_k) N_k K_k^T \quad (15.153)$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} (I - K_k H_k)^T + K_k R_k K_k^T - K_k N_k^T - N_k K_k^T + K_k (N_k^T H_k^T + H_k N_k) K_k^T \quad (15.154)$$

Thus, for the variance-minimizing Kalman gain, one must minimize the trace of the posterior state covariance. This occurs for the matrix derivative

$$\frac{\partial \text{Tr}(P_{k|k})}{\partial K_k} = -2(I - K_k H_k) P_{k|k-1} H_k^T + 2K_k R_k + 2K_k (N_k^T H_k^T + H_k N_k) - 2N_k \quad (15.155)$$

$$\frac{\partial \text{Tr}(P_{k|k})}{\partial K_k} = 2 \left[ K_k \left( H_k P_{k|k-1} H_k^T + H_k N_k + N_k^T H_k^T + R_k \right) - (P_{k|k-1} H_k^T + N_k) \right] \quad (15.156)$$

Thus, to make this zero, one requires the correlated process and measurement noise Kalman gain to be

$$K_k = (P_{k|k-1} H_k^T + N_k) \left( H_k P_{k|k-1} H_k^T + H_k N_k + N_k^T H_k^T + R_k \right)^{-1} \quad (15.157)$$

The posterior state covariance can be rearranged as

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} - P_{k|k-1} H_k^T K_k^T - K_k N_k^T - N_k K_k^T \\ + K_k (H_k P_{k|k-1} H_k^T + N_k^T H_k^T + H_k N_k + R_k) K_k^T \quad (15.158)$$

Then, substituting for  $K_k$  into the first term on the second line, one has

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} - K_k N_k^T - (P_{k|k-1} H_k^T + N_k) K_k^T \\ + (P_{k|k-1} H_k^T + N_k) \left( H_k P_{k|k-1} H_k^T + H_k N_k + N_k^T H_k^T + R_k \right)^{-1} \\ (H_k P_{k|k-1} H_k^T + N_k^T H_k^T + H_k N_k + R_k) K_k^T \quad (15.159)$$

$$\begin{aligned} P_{k|k} = & (I - K_k H_k) P_{k|k-1} - K_k N_k^T - (P_{k|k-1} H_k^T + N_k) K_k^T \\ & + (P_{k|k-1} H_k^T + N_k) \left( H_k P_{k|k-1} H_k^T + H_k N_k + N_k^T H_k^T + R_k \right)^{-1} \\ & (H_k P_{k|k-1} H_k^T + N_k^T H_k^T + H_k N_k + R_k) K_k^T \end{aligned} \quad (15.160)$$

$$\begin{aligned} P_{k|k} = & (I - K_k H_k) P_{k|k-1} - K_k N_k^T - (P_{k|k-1} H_k^T + N_k) K_k^T \\ & + (P_{k|k-1} H_k^T + N_k) K_k^T \end{aligned} \quad (15.161)$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} - K_k N_k^T \quad (15.162)$$

Similarly, the innovation covariance can be computed as

$$S_k = \mathbb{E} [(\vec{y}_k - \hat{\vec{y}}_k)(\vec{y}_k - \hat{\vec{y}}_k)^T] \quad (15.163)$$

$$S_k = \mathbb{E} [(H_k \vec{x}_k + \vec{v}_k - H_k \hat{\vec{x}}_{k|k-1})(H_k \vec{x}_k + \vec{v}_k - H_k \hat{\vec{x}}_{k|k-1})^T] \quad (15.164)$$

$$\begin{aligned} S_k = & \mathbb{E} \left[ (H_k (\vec{x}_k - \hat{\vec{x}}_{k|k-1})(\vec{x}_k - \hat{\vec{x}}_{k|k-1})^T H_k^T + \vec{v}_k \vec{v}_k^T \right. \\ & \left. + \vec{v}_k (\vec{x}_k^T - \hat{\vec{x}}_{k|k-1}^T) H_k^T + H_k (\vec{x}_k - \hat{\vec{x}}_{k|k-1}) \vec{v}_k^T \right] \end{aligned} \quad (15.165)$$

By substitution for the prior covariance, measurement noise covariance, and the linear state process model from  $\vec{x}_{k-1}$  to  $\vec{x}_k$ , one has

$$\begin{aligned} S_k = & H_k P_{k|k-1} H_k^T + R_k \\ & + \mathbb{E} \left[ \vec{v}_k (\vec{x}_{k-1}^T F_{k-1}^T + \vec{u}_{k-1}^T G_{k-1}^T + \vec{w}_{k-1}^T - \hat{\vec{x}}_{k|k-1}^T) H_k^T \right. \\ & \left. + H_k (F_{k-1} \vec{x}_{k-1} + G_{k-1} \vec{u}_{k-1} + \vec{w}_{k-1} - \hat{\vec{x}}_{k|k-1}) \vec{v}_k^T \right] \end{aligned} \quad (15.166)$$

Then, by the definition of the noise cross-covariance and using the fact that  $\vec{x}_{k-1}$ ,  $\vec{u}_{k-1}$ , and  $\hat{\vec{x}}_{k|k-1}$  are uncorrelated with zero-mean  $\vec{v}_{k-1}$ , one has

$$S_k = H_k P_{k|k-1} H_k^T + H_k N_k + N_k^T H_k^T + R_k \quad (15.167)$$

Thus, the **correlated-noise Kalman filter (CN-KF)** has a modified correction step

$$\begin{aligned} \tilde{y}_k &= \vec{y}_k - \hat{\vec{y}}_k \\ S_k &= H_k P_{k|k-1} H_k^T + H_k N_k + N_k^T H_k^T + R_k \\ K_k &= (P_{k|k-1} H_k^T + N_k) S_k^{-1} \\ \hat{\vec{x}}_{k|k} &= \hat{\vec{x}}_{k|k-1} + K_k \tilde{y}_k \\ \hat{P}_{k|k} &= (I - K_k H_k) P_{k|k-1} - K_k N_k^T \end{aligned} \quad (15.168)$$

which simply uses the modeled cross-covariance in the innovation covariance, Kalman gain, and posterior covariance recursions.

However, in some filtering problems, one may not be able to identify the cross-covariance between the information to be fused. One common method for fusing this information is known as **covariance intersection (CI)**. Here, consider information  $\vec{X}_1$  with mean  $\hat{\vec{x}}_1$  and information matrix  $P_1^{-1}$  to be fused with information  $\vec{X}_2$  with mean  $\hat{\vec{x}}_2$  and information matrix  $P_2^{-1}$ , to form the fused information  $\vec{X}_\omega$  with mean  $\hat{\vec{x}}_\omega$  and information matrix  $P_\omega^{-1}$ . If the cross-covariance between  $\vec{X}_1$  and  $\vec{X}_2$  is known as  $P_{12}$ , one has a mean and covariance update based on a generalization of the CN-KF as

$$\begin{aligned}\hat{\vec{x}}_\omega &= K_1 \hat{\vec{x}}_1 + K_2 \hat{\vec{x}}_2 \\ P_\omega &= K_1 P_1 K_1^T + K_2 P_2 K_2^T + K_1 P_{12} K_2^T + K_2 P_{12}^T K_1^T\end{aligned}\tag{15.169}$$

However, if  $P_{1,2}$  is unknown, the **covariance intersection (CI) algorithm** can be formed as a convex combination of the individual information matrices, i.e.,

$$\begin{aligned}P_\omega &= \left( \omega P_1^{-1} + (1 - \omega) P_2^{-1} \right)^{-1} \\ \hat{\vec{x}}_\omega &= P_\omega \left( \omega P_1^{-1} \hat{\vec{x}}_1 + (1 - \omega) P_2^{-1} \hat{\vec{x}}_2 \right)\end{aligned}\tag{15.170}$$

where  $\omega \in [0, 1]$  is a scalar optimization parameter that is chosen to minimize  $\text{Tr}(P_\omega)$ . The term “covariance intersection” derives from the geometrical perspective for the two-dimensional fusion case as the fused covariance ellipse is guaranteed to enclose the intersection of the two covariance ellipses with a co-located center for *any* possible cross-covariance. Furthermore, it can be shown that the CI algorithm guarantees a consistent estimate for all valid cross-covariances, i.e.,

$$P_\omega - \mathbb{E} \left[ (\vec{x} - \hat{\vec{x}}_\omega)(\vec{x} - \hat{\vec{x}}_\omega)^T \right] \geq 0\tag{15.171}$$

With this algorithm in mind, the Kalman filter can be modified for *unknown* correlation between the process and measurement noise. Thus, the **covariance intersection Kalman filter (CI-KF)** has a modified correction step

$$\begin{aligned}\tilde{y}_k &= \vec{y}_k - \hat{\vec{y}}_k \\ \omega_{opt} &= \underset{\omega \in [0, 1]}{\operatorname{argmin}} \text{Tr} \left( \omega P_{k|k-1}^{-1} + (1 - \omega) H_k^T R^{-1} H_k \right)^{-1} \\ \hat{P}_{k|k} &= \left( \omega_{opt} P_{k|k-1}^{-1} + (1 - \omega_{opt}) H_k^T R^{-1} H_k \right)^{-1} \\ K_k &= (1 - \omega_{opt}) \hat{P}_{k|k} H_k^T R^{-1} \\ \hat{\vec{x}}_{k|k} &= \hat{\vec{x}}_{k|k-1} + K_k \tilde{y}_k\end{aligned}\tag{15.172}$$

where one can use any scalar search optimization algorithm to find  $\omega_{opt}$ . Notably, this CI concept can be extended to the **generalized covariance intersection (GCI) algorithm** which fuses  $N$  information,  $\vec{X}_1, \dots, \vec{X}_N$ , with unknown correlation between all information with the possible consideration of known bounds of the correlation factors.

### Discrete-Time Colored-Noise Kalman Filtering

Consider the case where the process noise is uncorrelated with the measurement noise, but is colored, i.e.,

$$\begin{aligned}\mathbb{E} [\vec{w}_k \vec{w}_k^T] &= Q_k \\ \mathbb{E} [\vec{w}_k \vec{w}_{k-1}^T] &= \theta_{k-1} Q_k\end{aligned}\quad (15.173)$$

In this case, one can model the state process as

$$\vec{x}_k = F_{k-1} \vec{x}_{k-1} + G_{k-1} \vec{u}_{k-1} + \vec{w}_{k-1} \quad (15.174)$$

and the process noise by an additional  $n_x^{\text{th}}$ -order Markov process

$$\vec{w}_k = \theta_{k-1} \vec{w}_{k-1} + \vec{\zeta}_{k-1} \quad (15.175)$$

where  $\vec{\zeta}_{k-1}$  is zero-mean white noise with covariance  $C_{\zeta}$  that is uncorrelated with  $\vec{w}_{k-1}$ .

This construction provides the augmented state process equation as

$$\begin{bmatrix} \vec{x}_k \\ \vec{w}_k \end{bmatrix} = \begin{bmatrix} F_{k-1} & I \\ 0 & \theta_{k-1} \end{bmatrix} \begin{bmatrix} \vec{x}_{k-1} \\ \vec{w}_{k-1} \end{bmatrix} + \begin{bmatrix} G_{k-1} \\ 0 \end{bmatrix} \vec{u}_{k-1} + \begin{bmatrix} 0 \\ \vec{\zeta}_{k-1} \end{bmatrix} \quad (15.176)$$

$$\vec{x}_k^* = F_{k-1}^* \vec{x}_{k-1}^* + G_{k-1}^* \vec{u}_{k-1} + \vec{w}_{k-1}^*$$

which has augmented process noise covariance as

$$Q_{k-1}^* = \begin{bmatrix} 0 & 0 \\ 0 & C_{\zeta} \end{bmatrix} \quad (15.177)$$

Next, consider the case where the measurement noise is uncorrelated with the process noise, but is colored, i.e.,

$$\begin{aligned}\mathbb{E} [\vec{v}_k \vec{v}_k^T] &= R_k \\ \mathbb{E} [\vec{v}_k \vec{v}_{k-1}^T] &= \theta_{k-1} R_k\end{aligned}\quad (15.178)$$

In this case, one can model the measurement equation as

$$\vec{y}_k = H_k \vec{x}_k + \vec{v}_{k-1} \quad (15.179)$$

and the measurement noise by a  $n_y^{\text{th}}$ -order Markov process

$$\vec{v}_k = \theta_{k-1} \vec{v}_{k-1} + \vec{\zeta}_{k-1} \quad (15.180)$$

where  $\vec{\zeta}_{k-1}$  is zero-mean white noise with covariance  $C_{\zeta}$  that is uncorrelated with  $\vec{v}_{k-1}$ .

While one can use the same augmented state approach, this results in a singular measurement noise covariance which can cause numerical difficulties. Thus, it is preferable to use measurement differencing as

$$\Delta \vec{y}_{k+1,k} = \vec{y}_{k+1} - \theta_k \vec{y}_k \quad (15.181)$$

By substitution for the measurement equation, process equation, and measurement noise, one has

$$\Delta \vec{y}_{k+1,k} = H_{k+1}(F_k \vec{x}_k + G_k \vec{u}_k + \vec{w}_k) + \vec{v}_{k-1} - \theta_k H_k \vec{x}_k - \theta_k \vec{v}_{k-1} \quad (15.182)$$

Rearranging, one has

$$\Delta \vec{y}_{k+1,k} = H_{k+1}(F_k - \theta_k H_k + \vec{w}_k) \vec{x}_k + H_{k+1}G_k \vec{u}_k + H_{k+1}\vec{w}_k + \vec{\zeta}_k \quad (15.183)$$

which provides the alternative measurement model

$$\vec{y}_k^* = \Delta \vec{y}_{k+1,k} - H_{k+1}G_k \vec{u}_k = H_k^* \vec{x}_k + \vec{v}_k^* \quad (15.184)$$

with

$$H_k^* = H_{k+1}F_k - \theta_k H_k \quad (15.185)$$

and

$$\vec{v}_k^* = H_{k+1}\vec{w}_k + \vec{\zeta}_k \quad (15.186)$$

where the alternative noise covariance is

$$R_k^* = \mathbb{E} [\vec{v}_k^* \vec{v}_k^{*T}] = H_{k+1}Q_k H_{k+1}^T + C_\zeta \quad (15.187)$$

and alternative noise cross-covariance is

$$N_k^* = \mathbb{E} [\vec{w}_k \vec{v}_k^{*T}] = Q_k H_{k+1}^T \quad (15.188)$$

Thus, one can utilize the correlated-noise Kalman filter recursion to derive the one-step **colored-measurement-noise Kalman filter (CMN-KF)** as

$$\begin{aligned} K_k &= P_{k|1:k} H_k^{*T} (H_k^* P_{k|1:k} H_k^{*T} + R_k^*)^{-1} \\ \hat{x}_{k|1:k+1} &= \hat{x}_{k|1:k} + K_k (\vec{y}_k^* - H_k^* \hat{x}_{k|1:k}) \\ P_{k|1:k+1} &= (I - K_k H_k^*) P_{k|1:k} (I - K_k H_k^*) + K_k R_k^* K_k^T \\ C_k &= Q_k H_{k+1}^T (H_k^* P_{k|1:k} H_k^{*T} + R_k^*)^{-1} \\ \hat{x}_{k+1|1:k+1} &= F_k \hat{x}_{k|1:k+1} + C_k (\vec{y}_k^* - H_k^* \hat{x}_{k|1:k+1}) \\ P_{k+1|1:k+1} &= F_k P_{k|1:k+1} F_k^T + Q_k - C_k H_{k+1} Q_k - F_k K_k Q_k H_{k+1}^T - H_{k+1} Q_k K_k^T F_k^T \end{aligned} \quad (15.189)$$

## References

For more information, please refer to the following

- Julier, S., and Uhlmann, J. “12 General Decentralized Data Fusion with Covariance Intersection (CI),” in *Handbook of multisensor data fusion*, CRC Press LLC, 2001, pp. 12-1 to 12-25
- Reece, S., and Roberts, S., “Robust, low-bandwidth, multi-vehicle mapping,” in *8th International Conference on Information Fusion*, Vol 2, 2005, pp. 8

- Reinhardt, M., Kulkarni, S., and Hanebeck, U. D., “Generalized Covariance Intersection based on Noise Decomposition,” in *2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI)*, 2014, pp. 1-8
- Sarkka, S., “4.3 Kalman Filter,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 56-62
- Sarkka, S., “8.2 Rauch-Tung-Striebel Smoother,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 135-139
- Simon, D., “5.1 Derivation of the Discrete-Time Kalman Filter,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 124-129
- Simon, D., “6.1 Sequential Kalman filtering,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 150-155
- Simon, D., “6.2 Information filtering,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 156-158
- Simon, D., “6.3 Square root filtering,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 159-173
- Simon, D., “7.1 Correlated Process and Measurement Noise,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 184-188
- Simon, D., “7.2 Colored Process and Measurement Noise,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 188-193
- Simon, D., “7.3 Steady-State Filtering,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 193-195
- Simon, D., “9.3 Fixed-Lag Smoothing,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 274-279
- Simon, D., “9.4.2 RTS Smoothing,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 286-294

## 15.3 Continuous-Time Kalman Filtering

### Continuous-Time Kalman Filtering

With the discretized stochastic state-space model in mind, consider the discrete-time Kalman gain as

$$K_k = P_{k|k-1}H^T(HP_{k|k-1}H^T + R)^{-1} \quad (15.190)$$

By substitution for the discretized stochastic state-space model with the first-order approximation, one has

$$K_k = P_{k|k-1}C^T \left( CP_{k|k-1}C^T + \frac{R_{c-t}}{\Delta t} \right)^{-1} \quad (15.191)$$

$$\frac{K_k}{\Delta t} = P_{k|k-1} C^T \left( C P_{k|k-1} C^T \Delta t + R_{c-t} \right)^{-1} \quad (15.192)$$

$$\lim_{\Delta t \rightarrow 0} \frac{K_k}{\Delta t} = P_{k|k-1} C^T R_{c-t}^{-1} \quad (15.193)$$

Next, consider the one-step prior discrete-time Kalman filter for the state covariance

$$P_{k+1|k} = F(I - K_k H)P_{k|k-1}F^T + Q \quad (15.194)$$

By substitution for the discretized stochastic state-space model with the first-order approximation, one has

$$P_{k+1|k} = (I + A\Delta t)(I - K_k C)P_{k|k-1}(I + A\Delta t)^T + Q_{c-t}\Delta t \quad (15.195)$$

$$\begin{aligned} P_{k+1|k} = & P_{k|k-1} + AP_{k|k-1}\Delta t + P_{k|k-1}A^T\Delta t + AP_{k|k-1}A^T(\Delta t)^2 \\ & - K_k CP_{k|k-1} - AK_k CP_{k|k-1}\Delta t - K_k CP_{k|k-1}A^T\Delta t - AK_k CP_{k|k-1}A^T(\Delta t)^2 + Q_{c-t}\Delta t \end{aligned} \quad (15.196)$$

Subtracting  $P_{k|k-1}$  from both sides and dividing by  $\Delta t$ , one has

$$\begin{aligned} \frac{P_{k+1|k} - P_{k|k-1}}{\Delta t} = & AP_{k|k-1} + P_{k|k-1}A^T + AP_{k|k-1}A^T\Delta t \\ & - \frac{K_k CP_{k|k-1}}{\Delta t} - AK_k CP_{k|k-1} - K_k CP_{k|k-1}A^T - AK_k CP_{k|k-1}A^T\Delta t + Q_{c-t} \end{aligned} \quad (15.197)$$

Taking the limit with  $\lim_{\Delta t \rightarrow 0} K_k(\Delta t)^{-1} = P_{k|k-1}C^T R_{c-t}^{-1}$  in mind, one has

$$\dot{P} = \lim_{\Delta t \rightarrow 0} \frac{P_{k+1|k} - P_{k|k-1}}{\Delta t} = AP + PA^T - PC^T R_{c-t}^{-1} CP + Q_{c-t} \quad (15.198)$$

which is the forwards Riccati difference equation for the LQE and

$$K(t) = PC^T R_{c-t}^{-1} \quad (15.199)$$

Note that integrating this equation for a solution to  $P(t)$  requires only  $n_x(n_x + 1)/2$  integrations as  $P$  is symmetric.

Finally, consider the one-step prior discrete-time Kalman filter for the state estimate

$$\hat{x}_{k|k} = F\hat{x}_{k-1|k-1} + G\vec{u}_{k-1} + K_k \left( \vec{y}_k - H \left( F\hat{x}_{k-1|k-1} + G\vec{u}_{k-1} \right) \right) \quad (15.200)$$

By substitution for the discretized stochastic state-space model with the first-order approximation, one has

$$\begin{aligned} \hat{x}_{k|k} = & (I + A\Delta t)\hat{x}_{k-1|k-1} + B\Delta t\vec{u}_{k-1} \\ & + K_k \left( \vec{y}_k - C \left( I + A\Delta t \right) \hat{x}_{k-1|k-1} + B\Delta t\vec{u}_{k-1} \right) \end{aligned} \quad (15.201)$$

$$\begin{aligned} \hat{x}_{k|k} = & \hat{x}_{k-1|k-1} + A\Delta t\hat{x}_{k-1|k-1} + B\Delta t\vec{u}_{k-1} \\ & + K_k \left( \vec{y}_k - C\hat{x}_{k-1|k-1} + CA\Delta t\hat{x}_{k-1|k-1} + CB\Delta t\vec{u}_{k-1} \right) \end{aligned} \quad (15.202)$$

Subtracting  $x_{k-1|k-1}$  from both sides and dividing by  $\Delta t$ , one has

$$\begin{aligned} \frac{\hat{x}_{k|k} - \hat{x}_{k-1|k-1}}{\Delta t} &= A \hat{x}_{k-1|k-1} + B \vec{u}_{k-1} \\ &\quad + \frac{K_k}{\Delta t} \left( \vec{y}_k - C \left( I + A\Delta t \right) \hat{x}_{k-1|k-1} + B\Delta t \vec{u}_{k-1} \right) \end{aligned} \quad (15.203)$$

Taking the limit with  $\lim_{\Delta t \rightarrow 0} K_k(\Delta t)^{-1} = PC^T R_{c-t}^{-1}$  in mind, one has

$$\dot{\hat{x}} = \lim_{\Delta t \rightarrow 0} \frac{\hat{x}_{k|k} - \hat{x}_{k-1|k-1}}{\Delta t} = A \hat{x}_{k-1|k-1} + B \vec{u}_{k-1} + PC^T R_{c-t}^{-1} \left( \vec{y}_k - C \hat{x} \right) \quad (15.204)$$

In summary, these limiting arguments for the discretized stochastic state-space system provide an alternative derivation of the **continuous-time Kalman filter (KF)**, also known as the **Kalman-Bucy filter (KBF)**, also known as the **continuous-time linear-quadratic estimator (LQE)**, as the vector and matrix differential equations

$$\begin{aligned} \dot{P} &= AP + PA^T + Q_{c-t} - PC^T R_{c-t}^{-1} CP \\ K &= PC^T R_{c-t}^{-1} \\ \dot{\hat{x}}(t) &= A \hat{x}(t) + B \vec{u}(t) + K(\vec{y} - C \hat{x}(t)) \end{aligned} \quad (15.205)$$

Notably, the continuous-time **steady-state Kalman filter** uses a constant Kalman gain in the filter  $K_\infty = P_\infty C^T R_{c-t}^{-1}$  which solves the CARE and results in a continuous-time LTI filter. In some cases, the continuous-time Kalman filter converges to this steady-state condition. In this case, consider the Laplace transform of the continuous-time zero-input steady-state Kalman filter

$$s \hat{x}(s) = A \hat{x}(s) + K_\infty (\vec{y}(s) - C \hat{x}(s)) \quad (15.206)$$

$$(sI - A + K_\infty C) \hat{x}(s) = K_\infty \vec{y}(s) \quad (15.207)$$

$$\hat{x}(s) = (sI - A + K_\infty C)^{-1} K_\infty \vec{y}(s) \quad (15.208)$$

which is the transfer function matrix of the causal **Wiener filter**.

### Continuous-Time Correlated-Noise Kalman Filtering

Similarly, consider the case where the continuous-time process and measurement noise are correlated, but without colored noise, i.e.,

$$\begin{aligned} \mathbb{E}[\vec{w}(t) \vec{w}^T(\tau)] &= Q_{c-t} \delta(t - \tau) \\ \mathbb{E}[\vec{v}(t) \vec{v}^T(\tau)] &= R_{c-t} \delta(t - \tau) \\ \mathbb{E}[\vec{w}(t) \vec{v}^T(\tau)] &= N_{c-t} \delta(t - \tau) \end{aligned} \quad (15.209)$$

where  $N_{c-t}$  is the **continuous-time noise cross-covariance** and models the expected correlation/covariance between the zero-mean noise vectors.

To derive the continuous-time correlated-noise Kalman-Bucy filter (CN-KBF), consider the following Lagrange multiplier matrix,  $\Lambda$ , on the measurement equation substituted into the process equation, i.e.,

$$\dot{\vec{x}} = A\vec{x} + B\vec{u} + \vec{w} + \Lambda(\vec{y} - C\vec{x} - \vec{v}) \quad (15.210)$$

Rearranging, one has

$$\dot{\vec{x}} = (A - \Lambda C)\vec{x} + B\vec{u} + \Lambda\vec{y} + \vec{w} - \Lambda\vec{v} \quad (15.211)$$

which can be defined as transformed process

$$\dot{\vec{x}} = A^*\vec{x} + \vec{u}^* + \vec{w}^* \quad (15.212)$$

with

$$A^* = A - \Lambda C \quad (15.213)$$

$$\vec{u}^* = B\vec{u} + \Lambda\vec{y} \quad (15.214)$$

and

$$\vec{w}^* = \vec{w} - \Lambda\vec{v} \quad (15.215)$$

Then, consider the cross-covariance between this transformed process noise and the measurement noise given by

$$\mathbb{E}[\vec{w}^*(t)\vec{v}^T(\tau)] = \mathbb{E}[(\vec{w}(t) - \Lambda\vec{v}(t))\vec{v}^T(\tau)] \quad (15.216)$$

$$\mathbb{E}[\vec{w}^*(t)\vec{v}^T(\tau)] = \mathbb{E}[\vec{w}(t)\vec{v}^T(\tau)] - \Lambda\mathbb{E}[\vec{v}(t)\vec{v}^T(\tau)] \quad (15.217)$$

By definition, one has

$$\mathbb{E}[\vec{w}^*(t)\vec{v}^T(\tau)] = N_{c-t} - \Lambda R_{c-t} \quad (15.218)$$

Thus, to utilize the uncorrelated KBF recursion equation, one should choose  $\Lambda = N_{c-t}R_{c-t}^{-1}$ .

With this choice, the transformed process noise covariance is given by

$$Q_{c-t}^* = \mathbb{E}[\vec{w}^*(t)\vec{w}^{*T}(\tau)] \quad (15.219)$$

$$Q_{c-t}^* = \mathbb{E}[(\vec{w}(t) - N_{c-t}R_{c-t}^{-1}\vec{v}(t))(\vec{w}(\tau) - N_{c-t}R_{c-t}^{-1}\vec{v}(\tau))] \quad (15.220)$$

$$Q_{c-t}^* = Q_{c-t} - N_{c-t}R_{c-t}^{-1}N_{c-t}^T - N_{c-t}R_{c-t}^{-1}N_{c-t}^T + N_{c-t}R_{c-t}^{-1}N_{c-t}^T \quad (15.221)$$

$$Q_{c-t}^* = Q_{c-t} - N_{c-t}R_{c-t}^{-1}N_{c-t}^T \quad (15.222)$$

Then, the Riccati differential equation for the state covariance can be written as

$$\dot{P} = A^*P + PA^{*T} + Q_{c-t}^* - PRC^TR_{c-t}^{-1}CP \quad (15.223)$$

By substitution, one has

$$\dot{P} = (A - N_{c-t}R^{-1}C)P + P(A - NR^{-1}C)^T + Q_{c-t} - N_{c-t}R_{c-t}^{-1}N_{c-t}^T - PC^TR_{c-t}^{-1}CP \quad (15.224)$$

Defining the transformed Kalman gain as

$$K^* = PC^TR_{c-t}^{-1} + N_{c-t}R_{c-t}^{-1} = (PC^T + N_{c-t})R_{c-t}^{-1} \quad (15.225)$$

The Riccati differential equation can be rewritten as

$$\dot{P} = AP + PA^T + Q_{c-t} - K^* R_{c-t}^{-1} K^{*T} \quad (15.226)$$

Using the transformed process equation with the standard Kalman gain, one has

$$\dot{\hat{x}} = A^* \dot{\hat{x}} + \vec{u}^* + K(\vec{y} - C \hat{x}) \quad (15.227)$$

$$\dot{\hat{x}} = (A - N_{c-t} R_{c-t}^{-1} C) \dot{\hat{x}} + (B \vec{u} + N_{c-t} R_{c-t}^{-1} \vec{y}) + PC^T R_{c-t}^{-1} (\vec{y} - C \hat{x}) \quad (15.228)$$

Rearranging, one has

$$\dot{\hat{x}} = A \dot{\hat{x}} + B \vec{u} + N_{c-t} R_{c-t}^{-1} \vec{y} - N_{c-t} R_{c-t}^{-1} C \dot{\hat{x}} + PC^T R_{c-t}^{-1} (\vec{y} - C \hat{x}) \quad (15.229)$$

or

$$\dot{\hat{x}} = A \dot{\hat{x}} + B \vec{u} + K^* (\vec{y} - C \hat{x}) \quad (15.230)$$

Thus, the **correlated-noise Kalman-Bucy filter (KBF)**, also known as the **continuous-time correlated-noise Kalman filter (CT-CN-KF)**, can be summarized as

$$\begin{aligned} \dot{P} &= AP + PA^T + Q_{c-t} - (PC^T + N_{c-t})R_{c-t}^{-1}(CP + N_{c-t}^T) \\ K &= (PC^T + N_{c-t})R_{c-t}^{-1} \\ \dot{\hat{x}}(t) &= A \hat{x}(t) + B \vec{u}(t) + K(\vec{y} - C \hat{x}(t)) \end{aligned} \quad (15.231)$$

where  $\mathbb{E}[\vec{w}(t) \vec{v}^T(\tau)] = N\delta(t - \tau)$ .

## References

For more information, please refer to the following

- Simon, D., “8.2 Derivation of the Continuous-Time Kalman Filter,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 233-237
- Simon, D., “8.4 Generalizations of the Continuous-Time Filter,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 247-251

## 15.4 Multi-Modal and Heavy-Tailed Kalman Filtering

### Gaussian Sum Filter

The use of nonlinear Kalman filters for non-Gaussian distributions can also be approximated by the fact that any distribution can be approximated arbitrarily close in the  $L_1$ -sense as a Gaussian mixture (GM) distribution, or a weighted “sum” of Gaussians. The so-called **Gaussian sum filter (GSF)** approximates the true distributions of the process noise and measurement noise as a GM of  $n_N$  elements with  $n_N$  weights, means, and covariances, runs  $n_N$  parallel nonlinear Kalman filters for each element in the , and combine them for an approximately optimal estimate. Choosing  $n_N$  is a trade-off between approximation accuracy/optimality and computational cost. It should also be mentioned that for noise with heavy tails, the Gaussian

## Kalman-Levy Filtering

### References

For more information, please refer to the following

- Simon, D., “13.3 Higher-order approaches,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, pp. 410-420, 2006
- Sornette, D. and Ide, K., “The Kalman-Levy Filter,” in *Physica D: Nonlinear Phenomena*, Volume 151, No. 2-4, Elsevier, p. 142-174, 2001

## 15.5 Multiple Model Filtering

**Hybrid-state estimation** is estimation of a state with both continuous and discrete components, as one must estimate both the continuously-valued target state and the discrete-valued motion-mode.

### Multi-Model Filtering Methods

The primary approach in the presence of motion-mode uncertainty is the **multi-model (MM)** filtering method which applies to any hybrid system estimation. These methods assume a **model set**,  $\mathbb{M}$ , as possible candidates of the true mode in effect at the current time step. Thus, one can run a **elementary filter bank**, each based on a unique model  $\mathcal{M}^{(i)} \in \mathbb{M}$ . Thus, MM methods estimate the hybrid process based on the combined results of these elemental filters, providing an integrated approach to the joint model decision and state estimation problem, e.g. maneuvering target tracking. This is a semi-parametric approach as its model coverage is in between parametric and non-parametric approaches to this problem. This section will describe MM methods for MJLSs.

For a MJLS, the  $i^{\text{th}}$  model in the MM methods is defined as

$$\begin{aligned}\vec{x}_k &= F_{k-1}^{(i)} \vec{x}_{k-1} + G_{k-1}^i \vec{w}_{k-1} \\ \vec{y}_k &= H_k^{(i)} \vec{x}_k + \vec{v}_k^{(i)}\end{aligned}\tag{15.232}$$

where  $\vec{w}_{k-1}^{(i)}$  has mean  $\bar{w}_{k-1}^{(i)}$  and covariance  $Q_{k-1}^{(i)}$  and  $\vec{v}_k^{(i)}$  has mean  $\bar{v}_k^{(i)}$  and covariance  $R_k^{(i)}$ .

The superscript  $(i)$  denotes quantities pertaining to model  $\mathcal{M}^{(i)} \in \mathbb{M}$  and the jumps of the modal state are assumed to have homogeneous transition probabilities, i.e.

$$\Pr(\mathcal{M}_k^{(j)} | \mathcal{M}_{k-1}^{(i)}) = \pi_{ji} \quad \forall \mathcal{M}^{(i)}, \mathcal{M}^{(j)}, k\tag{15.233}$$

where  $\mathcal{M}_k^{(i)}$  denotes the event that model  $\mathcal{M}_k^{(i)}$  matches the modal state in effect at time step  $k$ , i.e.

$$\mathcal{M}_k^{(i)} = \{\mathcal{S}_k = \mathcal{M}^{(i)}\}\tag{15.234}$$

Similarly, the finite sequence of events are denoted as

$$\mathcal{S}_{1:k} = \{\mathcal{S}_1^{(i_1)}, \dots, \mathcal{S}_k^{(i_k)}\}\tag{15.235}$$

and

$$\mathcal{M}_{1:k} = \{\mathcal{M}_1^{(i_1)}, \dots, \mathcal{M}_k^{(i_k)}\} \quad (15.236)$$

Finally, it should be noted that the model set,  $\mathbb{M}$ , is generally a mathematical approximation of the true mode space,  $\mathbb{S}$ , and may contain fewer elements and have simpler model descriptions than the true possible set of system behaviors indicated by  $\mathbb{S}$ .

MM estimation algorithms have four key design elements. First, model set design refers to the offline selection and possible online adaptation of the model set. Second, cooperation strategy design refers to the possible pruning of unlikely model sequences, merging of “similar” model sequences, selection of the most likely model sequences, and any iterations done on these actions. Third, the filtering of the base state conditioned on some assumed mode sequence which is the only step common with non-MM algorithms. Fourth, the output process which generates overall estimates using results of all filters as well as measurements, e.g. fusing estimates, selecting the best ones. There have been three generations of MM algorithms which made different assumptions about the mode and model sets. Defining the three

- A: The true modal state is time-invariant, i.e.  $\mathcal{S}_k = \mathcal{S} \quad \forall k$ .
- B: The mode space and model set are time-variant and identical, i.e.  $\mathbb{S}_k = \mathbb{S} \quad \forall k, \mathbb{M}_k = \mathbb{M} \quad \forall k, \mathbb{S} = \mathbb{M} \quad \forall k$ .
- C: The true mode sequence is Markov or semi-Markov.

The first generation of **autonomous MM (AMM)** assumes A and B, the second generation of **cooperating multi-model (CMM)** algorithms assumes B and C, while the third generation of **variable-structure multi-model (VSMM)** assumes only C. This section will focus on CMM algorithms which typically form the building blocks of the third generation.

### Cooperating Multi-Model Filtering

In CMM, there are  $n_M^k$  possible model sequence realizations at time step  $k$  where  $n_M = |\mathbb{M}|$  is the number of possible models at each  $k$ . This allows one to represent a generic realization of a model sequence through  $k$  as

$$\mathcal{M}_{1:k}^{(i_{1:k})} = \{\mathcal{M}_1^{(i_1)}, \dots, \mathcal{M}_k^{(i_k)}\} = \{\mathcal{S}_{1:k} = \mathcal{M}^{(i_{1:k})}\} \in \mathbb{M}^k \quad (15.237)$$

where  $\mathbb{M}^k$  is the set of all such sequence events. Furthermore, one can use  $i_{1:k} \in \mathbb{M}_{1:k}$  to denote  $\mathcal{M}_{1:k}^{(i_{1:k})} \in \mathbb{M}^k$  which has  $n_M^k$  members where  $\mathbb{M}$  is the index set of  $\mathbb{M}$  and has  $n_M$  members. Notably, a single modal state sequence is denoted

$$\mathcal{S}_{1:k} = (\mathcal{S}_1, \dots, \mathcal{S}_k) \quad (15.238)$$

a single model sequence is

$$\mathcal{M}^{(i_{1:k})} = (\mathcal{M}^{i_1}, \dots, \mathcal{M}^{i_k}), \quad \mathcal{M}^{(i_n)} \in \mathbb{M} \quad (15.239)$$

To estimate the base state of hybrid system, this section will assess two different optimal estimation schemes: MMSE-CMM and MAP-CMM. These two estimators are primarily used in CMM maneuvering target tracking due to the fact that the motion-mode is never *exactly* resolved by the model used, e.g. an aircraft never exhibits an *exactly* constant turn-rate and even if it did, the turn-rate would not be *exactly* equal to one of the model turn-rates. These two estimators are robust to this inherent model mismatch as they rely on all component densities as will be shown.

Using the law of total expectation, the **CMM-MMSE base state estimator** at time step  $k$  is given by

$$\begin{aligned}\hat{\vec{x}}_{k|k}^{MMSE} &= \mathbb{E} [\vec{x}_k | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}] \\ &= \sum_{i_{1:k} \in \mathbf{M}_{1:k}} \mathbb{E} [\vec{x}_k | \vec{y}_{1:k}, \mathcal{M}_{1:k}^{(i_{1:k})}] \Pr (\mathcal{M}_{1:k}^{(i_{1:k})} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) \\ &= \sum_{i_{1:k} \in \mathbf{M}_{1:k}} \hat{\vec{x}}_{k|k}^{(i_{1:k})} \mu_{1:k}^{(i_{1:k})}\end{aligned}\quad (15.240)$$

where  $\mu_{1:k}^{(i_{1:k})} = \Pr (\mathcal{M}_{1:k}^{(i_{1:k})} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C})$  is the *a posteriori* mode sequence event probability assuming the mode sequence in effect is one and only one mode, but is possibly any one of the modes in the set  $\mathbb{M}^k$ ,  $\hat{\vec{x}}_{k|k}^{(i_{1:k})} = \mathbb{E} [\vec{x}_k | \vec{y}_{1:k}, \mathcal{M}_{1:k}^{(i_{1:k})}]$  is the conditional MMSE estimate assuming sequence  $\mathcal{M}_{1:k}^{(i_{1:k})}$  is true.

Alternatively, one can write

$$\begin{aligned}\hat{\vec{x}}_{k|k}^{MMSE} &= \sum_{i \in \mathbf{M}} \mathbb{E} [\vec{x}_k | \vec{y}_{1:k}, \mathcal{M}_k^{(i)}] \Pr (\mathcal{M}_k^{(i)} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) \\ &= \sum_{i \in \mathbf{M}} \hat{\vec{x}}_{k|k}^{(i)} \mu_k^{(i)}\end{aligned}\quad (15.241)$$

where  $\hat{\vec{x}}_{k|k}^{(i)} = \mathbb{E} [\vec{x}_k | \vec{y}_{1:k}, \mathcal{M}_k^{(i)}]$  is the conditional MMSE estimate assuming model  $\mathcal{M}_k^{(i)}$  is in effect at time step  $k$  and  $\mu_k^{(i)} = \Pr (\mathcal{M}_k^{(i)} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C})$  is a *a posteriori* mode probability that  $\mathcal{M}^{(i)}$  is in effect at time step  $k$ , in effect, lumping much of  $\hat{\vec{x}}_{k|k}^{(i)}$  as the summation over the model sequences is the same as over the current models.

It can be shown that  $\hat{\vec{x}}_{k|k}^{MMSE}$  is unbiased with the minimal conditional MSE matrix, i.e. the minimum of the base-state error covariance

$$\begin{aligned}P_{k|k} &= \text{MSE} (\hat{\vec{x}}_{k|k} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) \\ &= \sum_{i_{1:k} \in \mathbf{M}_{1:k}} \left[ \text{MSE} (\hat{\vec{x}}_{k|k}^{(i_{1:k})} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) + (\hat{\vec{x}}_{k|k}^{(i_{1:k})} - \hat{\vec{x}}_{k|k}) (\hat{\vec{x}}_{k|k}^{(i_{1:k})} - \hat{\vec{x}}_{k|k})^T \right] \mu_{1:k}^{(i_{1:k})} \\ &= \sum_{i \in \mathbf{M}} \left[ \text{MSE} (\hat{\vec{x}}_{k|k}^{(i)} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) + (\hat{\vec{x}}_{k|k}^{(i)} - \hat{\vec{x}}_{k|k}) (\hat{\vec{x}}_{k|k}^{(i)} - \hat{\vec{x}}_{k|k})^T \right] \mu_k^{(i)}\end{aligned}\quad (15.242)$$

which holds for an estimator,  $\hat{\vec{x}}_{k|k}$ .

Correspondingly, the **CMM-MMSE modal state estimator** at time step  $k$  is given by

$$\begin{aligned}
 \hat{\mathcal{M}}_{k|k} &= \mathbb{E} [\mathcal{M}_k | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}] \\
 &= \sum_{i_{1:k} \in \mathbf{M}_{1:k}} \mathbb{E} [\mathcal{M}_k | \vec{y}_{1:k}, \mathcal{M}_{1:k}^{(i_{1:k})}] \Pr (\mathcal{M}_{1:k}^{(i_{1:k})} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) \\
 &= \sum_{i_{1:k} \in \mathbf{M}_{1:k}} \mathcal{M}^{(i_{1:k})} \mu_{1:k}^{(i_{1:k})} \\
 &= \sum_{i \in \mathbf{M}} \mathbb{E} [\mathcal{M}_k | \vec{y}_{1:k}, \mathcal{M}_k^{(i)}] \Pr (\mathcal{M}_k^{(i)} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) \\
 &= \sum_{i \in \mathbf{M}} \hat{\mathcal{M}}_{k|k}^{(i)} \mu_k^{(i)}
 \end{aligned} \tag{15.243}$$

with mean square error

$$\begin{aligned}
 MSE(\hat{\mathcal{M}}_{k|k} | \vec{y}_{1:k}) &= \mathbb{E} \left[ (\mathcal{M}_k - \hat{\mathcal{M}}_{k|k}) (\mathcal{M}_k - \hat{\mathcal{M}}_{k|k})^T | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C} \right] \\
 &= \sum_{i_{1:k} \in \mathbf{M}_{1:k}} (\mathcal{M}^{(i_{1:k})} - \hat{\mathcal{M}}_{k|k}) (\mathcal{M}^{(i_{1:k})} - \hat{\mathcal{M}}_{k|k})^T \mu_{1:k}^{(i_{1:k})} \\
 &= \sum_{i \in \mathbf{M}} (\mathcal{M}^{(i)} - \hat{\mathcal{M}}_{k|k}) (\mathcal{M}^{(i)} - \hat{\mathcal{M}}_{k|k})^T \mu_k^{(i)}
 \end{aligned} \tag{15.244}$$

For the MAP estimators, one can note that the mixed PDF-PMF density function of the base state at time step  $k$  and mode sequence through time step  $k$  is

$$\begin{aligned}
 p(\vec{x}_k, \mathcal{M}_{1:k} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) &= f_{\vec{X}}(\vec{x}_k | \vec{y}_{1:k}, \mathcal{M}_{1:k}) p(\mathcal{M}_{1:k} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) \\
 &= \{f_{(i_{1:k})}(\vec{x}_k | \vec{y}_{1:k}) \mu_{1:k}^{(i_{1:k})}, i_{1:k} \in \mathbf{M}_{1:k}\} \\
 &\neq f(\vec{x}_k | \vec{y}_{1:k}, \mathcal{M}_k^{(i)}) p(\mathcal{M}_{1:k} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) \\
 &= \{f_{(i)}(\vec{x}_k | \vec{y}_{1:k}) \mu_k^{(i)}, i \in \mathbf{M}\}
 \end{aligned} \tag{15.245}$$

where  $f_{(i_{1:k})}(\vec{x}_k | \vec{y}_{1:k}) = f_{\vec{X}}(\vec{x}_k | \vec{y}_{1:k}, \mathcal{M}_{1:k})$  is the PDF assuming the mode sequence is  $\mathcal{M}_{1:k}^{(i_{1:k})}$ .

Thus, the base state has the *a posteriori* mixed density function

$$\begin{aligned}
 f_{\vec{X}}(\vec{x}_k | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) &= \sum_{i_{1:k} \in \mathbf{M}_{1:k}} f_{\vec{X}}(\vec{x}_k | \vec{y}_{1:k}, \mathcal{M}_{1:k}^{(i_{1:k})}) \Pr (\mathcal{M}_{1:k}^{(i_{1:k})} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) \\
 &= \sum_{i_{1:k} \in \mathbf{M}_{1:k}} f_{(i_{1:k})}(\vec{x}_k | \vec{y}_{1:k}) \mu_{1:k}^{(i_{1:k})} \\
 &\neq \sum_{i \in \mathbf{M}} f_{\vec{X}}(\vec{x}_k | \vec{y}_{1:k}, \mathcal{M}_k^{(i)}) \Pr (\mathcal{M}_k^{(i)} | \vec{y}_{1:k}, \mathbf{B}, \mathbf{C}) \\
 &= \sum_{i \in \mathbf{M}} f_{(i)}(\vec{x}_k | \vec{y}_{1:k}) \mu_k^{(i)}
 \end{aligned} \tag{15.246}$$

where the summation over the model sequences is no longer the same as the current models.

Thus, the **CMM-MAP base state estimator** at time step  $k$  is given by

$$\hat{\vec{x}}_{k|k}^{MAP} = \operatorname{argmax}_{\vec{x}_k} \sum_{i_{1:k} \in \mathcal{M}_{1:k}} f_{(i_{1:k})}(\vec{x}_k | \vec{y}_{1:k}) \mu_{1:k}^{(i_{1:k})} \quad (15.247)$$

and the **CMM-MAP mode sequence estimator** at time step  $k$  is given by

$$\hat{\mathcal{M}}_{1:k|k}^{MAP} = \operatorname{argmax}_{\mathcal{M}_{1:k}^{(i_{1:k})}} \{\mu_{1:k}^{(i_{1:k})}, i_{1:k} \in \mathcal{M}_{1:k}\} \quad (15.248)$$

It should be noted that while the entire mixed density function,  $f_{(i_{1:k})}(\vec{x}_k | \vec{y}_{1:k})$ , is needed to compute  $\hat{\vec{x}}_{k|k}^{MAP}$ , only the first two moments,  $\hat{\vec{x}}_{k|k}^{(i_{1:k})}$  and  $P_{k|k}^{(i_{1:k})}$ , are needed to compute  $\hat{\vec{x}}_{k|k}^{MMSE}$ , i.e. computing  $\hat{\vec{x}}_{k|k}^{MAP}$  is a maximization problem given every component density while computing  $\hat{\vec{x}}_{k|k}^{MMSE}$  is an integration problem. Regardless, since the number of possible model sequences increases exponentially with time, any brute-force implementations of the above optimal CMM-MMSE or -MAP is infeasible. Thus, cooperation strategies have been proposed to overcome this difficulty in practice. These strategies are merging of “similar” model sequences, pruning “unlikely” model sequences, randomly selecting a subset of possible model sequences, or iterative strategies. The development of each of these strategies is beyond the scope of this course. However, as an introduction to this filtering concept, the interacting multi-model filter will be presented in the following subsection.

### Interacting Multi-Model Filter

The **interacting multi-model (IMM) filter** uses  $n_M = |\mathcal{M}|$  elemental filters in its bank and consists of four steps that are run each cycle of the filtering algorithm: model-conditioned filtering, mode probability update, estimate fusion, and model-conditioned re-initialization. The model-conditioned filtering step consists of running the  $n_M$  elemental filters, e.g. EKFs, UKFs, PFs. The mode probability update uses the current innovation distribution conditioned on the past measurements to form the model likelihoods and mode probabilities. Lastly, the overall base-state estimate and covariance can be computed using the  $n_M$  filter estimates and mode probabilities. As an example, the **interacting multi-model Kalman filter (IMM-KF)** uses Kalman filters in its elemental filter bank and can be outlined as follows.

The model-conditioned re-initialization serves as the “mixing” step of the IMM filter which re-initializes all the  $n_M$  filters based on weighted sums of the probabilities of each mode, the mode transition probabilities, and all  $n_M$  filter outputs, conditioned on the measurement history. In particular, the re-initialized state for each  $i = 1, \dots, n_M$  filter is computed as

$$\begin{aligned} \hat{x}_{k-1|k-1}^{(i)} &= \mathbb{E} \left[ \vec{x}_{k-1} | \vec{y}_{1:k}, \mathcal{M}_k^{(i)} \right] \\ &= \mathbb{E} \left[ \mathbb{E} \left[ \vec{x}_{k-1} | \vec{y}_{1:k}, \mathcal{M}_k^{(i)}, \mathcal{M}_{k-1}^{(j)} \right] | \vec{y}_{1:k}, \mathcal{M}_k^{(i)} \right] \\ &= \sum_{j \in \mathcal{M}} \hat{x}_{k|k}^{(j)} \Pr \left( \mathcal{M}_{k-1}^{(j)} | \vec{y}_{1:k}, \mathcal{M}_k^{(i)} \right) \end{aligned} \quad (15.249)$$

This dependence of the re-initialization on  $\mathcal{M}_k^{(i)}$  which implies an individualized re-initialization of each filter  $i$  is the two primary benefit of IMM as opposed to other CMM filters and has been shown to be more accurate in many instances, albeit, at more computations required at each step.

The **IMM-KF mixing step** computes the following for  $i = 1, \dots, n_M$ . The  $i^{\text{th}}$  predicted mode probability is given by

$$\mu_{k|k-1}^{(i)} = \Pr\left(\mathcal{M}_k^{(i)} | \vec{y}_{1:k}\right) = \sum_{j \in \mathbb{M}} \pi_{ji} \mu_{k-1|k-1}^{(j)} \quad (15.250)$$

The  $i^{\text{th}}$  mixing weight is given by

$$\mu_{k-1}^{j|i} = \Pr\left(\mathcal{M}_k^{(i)} | \vec{y}_{1:k}\right) = \pi_{ji} \frac{\mu_{k-1|k-1}^{(j)}}{\mu_{k|k-1}^{(i)}} \quad (15.251)$$

The  $i^{\text{th}}$  mixing estimate is given by

$$\bar{x}_{k-1|k-1}^{(i)} = \mathbb{E}\left[\vec{x}_{k-1} | \mathcal{M}_k^{(i)}, \vec{y}_{1:k}\right] = \sum_{j \in \mathbb{M}} \mu_{k-1}^{j|i} \hat{x}_{k-1|k-1}^{(j)} \quad (15.252)$$

The  $i^{\text{th}}$  mixing covariance is given by

$$\bar{P}_{k-1|k-1}^{(i)} = \sum_{j \in \mathbb{M}} \mu_{k-1}^{j|i} \left[ P_{k-1|k-1}^{(j)} + \left( \bar{x}_{k-1|k-1}^{(i)} - \hat{x}_{k-1|k-1}^{(j)} \right) \left( \bar{x}_{k-1|k-1}^{(i)} - \hat{x}_{k-1|k-1}^{(j)} \right)^T \right] \quad (15.253)$$

The **IMM-KF filtering step** computes the following for  $i = 1, \dots, n_M$ . The  $i^{\text{th}}$  predicted state is given by

$$\hat{x}_{k|k-1}^{(i)} = F_{k-1|k-1}^{(i)} \bar{x}_{k-1|k-1}^{(i)} + G_{k-1|k-1}^{(i)} \bar{w}_{k-1}^{(i)} \quad (15.254)$$

The  $i^{\text{th}}$  predicted covariance is given by

$$P_{k|k-1}^{(i)} = F_{k-1|k-1}^{(i)} \bar{P}_{k-1|k-1}^{(i)} \left( F_{k-1|k-1}^{(i)} \right)^T + G_{k-1|k-1}^{(i)} Q_{k-1}^{(i)} \left( G_{k-1|k-1}^{(i)} \right)^T \quad (15.255)$$

The  $i^{\text{th}}$  innovation is given by

$$\tilde{y}_k^{(i)} = \vec{y}_k - H_k^{(i)} \hat{x}_{k|k-1}^{(i)} - \bar{v}_k^{(i)} \quad (15.256)$$

The  $i^{\text{th}}$  innovation covariance is given by

$$S_k^{(i)} = H_k^{(i)} P_{k|k-1}^{(i)} \left( H_k^{(i)} \right)^T + R_k^{(i)} \quad (15.257)$$

The  $i^{\text{th}}$  Kalman filter gain is given by

$$K_k^{(i)} = P_{k|k-1}^{(i)} \left( H_k^{(i)} \right)^T \left( S_k^{(i)} \right)^{-1} \quad (15.258)$$

The  $i^{\text{th}}$  corrected state is given by

$$\hat{x}_{k|k}^{(i)} = \hat{x}_{k|k-1}^{(i)} + K_k^{(i)} \tilde{y}_k^{(i)} \quad (15.259)$$

The  $i^{\text{th}}$  corrected covariance is given by

$$P_{k|k}^{(i)} = P_{k|k-1}^{(i)} - K_k^{(i)} S_k^{(i)} \left( K_k^{(i)} \right)^T \quad (15.260)$$

The **IMM-KF mode probability update step** computes the following for  $i = 1, \dots, n_M$ . The  $i^{\text{th}}$  model likelihood is given by

$$\mathcal{L}_k^{(i)} = f_{\tilde{Y}} \left( \tilde{y}_k^{(i)} | \mathcal{M}_k^{(i)}, \vec{y}_{1:k} \right) = f_{\mathcal{N}} \left( \tilde{y}_k^{(i)}; 0, S_k^{(i)} \right) \quad (15.261)$$

where the multivariate Gaussian model is assumed in this case. The  $i^{\text{th}}$  mode probability is given by

$$\mu_{k|k}^{(i)} = \frac{\mu_{k|k-1}^{(i)} \mathcal{L}_k^{(i)}}{\sum_{j \in M} \mu_{k|k-1}^{(j)} \mathcal{L}_k^{(j)}} \quad (15.262)$$

The **IMM-KF estimate fusion step** computes the overall base-state estimate as

$$\hat{x}_{k|k} = \sum_{i \in M} \mu_{k|k}^{(i)} \hat{x}_{k|k}^{(i)} \quad (15.263)$$

and the overall base-state covariance as

$$P_{k|k} = \sum_{i \in M} \mu_{k|k}^{(i)} \left[ P_{k|k}^{(i)} + \left( \hat{x}_{k|k} - \hat{x}_{k|k}^{(i)} \right) \left( \hat{x}_{k|k} - \hat{x}_{k|k}^{(i)} \right)^T \right] \quad (15.264)$$

which is the approximate MMSE and MAP for the Gaussian mixture assumption for this filter.

By inspection, one can see that the elemental filters could easily be extended to any Bayes-based filter, e.g. the EKF. It is also important to note that one must be able to model the modal transition probabilities,  $\pi_{ji}$ , appropriately for the IMM to work well, i.e. either constant if  $S_k$  is assumed to be a Markov process, or a sojourn-time dependent model if  $S_k$  is assumed to be a semi-Markov process.

## References

For more information, please refer to the following

- Bar-Shalom, Y., Daum, F., and Huang, J., “The Probabilistic Data Association Filter,” in *IEEE Control Systems Magazine*, Vol. 29, Issue 6, 2009, pp. 82-100
- Simon, D., “10.2 Multiple-Model Estimation,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 301-303

## 15.6 Stochastic Linear-Quadratic Optimal Control

Using observer-based feedback control for stochastic state-space systems, one can form the stochastic control systems, including stochastic eigenvalue placement and stochastic optimal control. In this context, the **fundamental stochastic optimal control problem** known as the **linear-quadratic-Gaussian (LQG) OCP** linear-quadratic Gaussian OCP defined for continuous-, discrete-, and hybrid-time. The LQG term derives from the *linear* process equation, measurement equation, and additive noise, *quadratic* due to the form of the cost function/functional, and having white *Gaussian* noise with a constant covariance matrix. For the cases below, assume the system is controllable and observable. Recall that in all three cases, the design of the state feedback and state estimator gains independently due to the **separation principle**.

The unconstrained finite-horizon continuous-time LQG OCP is

$$\begin{aligned} \vec{u}^{opt}(t) = \underset{\vec{u}(t) \text{ for } t \geq 0}{\operatorname{argmin}} \quad \mathcal{J} &= \mathbb{E} \left[ \vec{x}^T(t_f) E \vec{x}(t_f) + \int_0^{t_f} \vec{x}^T(t) Q(t) \vec{x}(t) + \vec{u}^T(t) R(t) \vec{u}(t) \right] \\ \text{subject to: } \dot{\vec{x}}(t) &= A(t) \vec{x}(t) + B(t) \vec{u}(t) + L(t) \vec{w}(t) \\ \vec{y}(t) &= C(t) \vec{x}(t) + M(t) \vec{v}(t) \\ \text{initial condition: } \vec{\mu}_0 &= \mathbb{E}[\vec{x}(0)], \Sigma_0 = \mathbb{E}[\vec{x}(0) \vec{x}^T(0)] \end{aligned} \quad (15.265)$$

where  $d\vec{w}(t)$  and  $d\vec{v}(t)$  are AWGN with covariances,  $\Sigma_{\vec{w}}$ , and  $\Sigma_{\vec{v}}$ , respectively.

To solve this OCP, assume one can use an observer-based feedback controller, i.e.

$$\vec{u}(t) = -K_{LQR}(t) \hat{\vec{x}}(t) \quad (15.266)$$

where  $K_{LQR}(t)$  is the optimal continuous-time LQR gain matrix. The initial condition of the estimator is given by an initial estimate of the state,  $\hat{\vec{x}}(t) = \vec{\mu}_0$  and initial covariance,  $\Sigma_0$ . The state estimate  $\hat{\vec{x}}(t)$  evolves according to the **continuous-time Luenberger observer equation** as

$$\dot{\hat{\vec{x}}}(t) = A(t) \hat{\vec{x}}(t) + B(t) \vec{u}(t) + K_{LQE} \left( \vec{y}(t) - C(t) \hat{\vec{x}}(t) \right) \quad (15.267)$$

For this stochastic OCP, the optimal estimator gain,  $K_{LQE}(t)$ , is the continuous-time **linear-quadratic estimator (LQE)**, also known as the **continuous-time Kalman gain**.

Recall that to find the optimal  $K_{LQR}(t)$ , one must solve the corresponding Riccati differential equation backwards-in-time for the closed-loop dynamics of the state controller, i.e.

$$\dot{P}_{LQR}(t) = -P_{LQR}(t)A(t) - A^T(t)P_{LQR}(t) + P_{LQR}(t)B(t)R^{-1}(t)B^T(t)P_{LQR}(t) - Q(t) \quad (15.268)$$

with final condition

$$P_{LQR}(t_f) = E \quad (15.269)$$

Then, solving for  $P_{LQR}(t)$  provides the optimal LQR gain as

$$K_{LQR}(t) = R^{-1}(t)B^T(t)P_{LQR}(t) \quad (15.270)$$

Similarly, to find the optimal  $K_{LQE}(t)$ , one must solve the corresponding Riccati differential equation forwards-in-time for the closed-loop dynamics of the state estimator, i.e.

$$\begin{aligned} \dot{P}_{LQE}(t) &= P_{LQE}(t)A(t) + P_{LQE}(t)A^T(t) + L(t)\Sigma_{\vec{w}}L^T(t) \\ &\quad - P_{LQE}(t)C^T(t)(M(t)\Sigma_{\vec{v}}M^T(t))^{-1}(t)C(t)P_{LQE}(t) \end{aligned} \quad (15.271)$$

with initial condition

$$P_{LQE}(0) = \Sigma_0 \quad (15.272)$$

Then, solving for  $P_{LQE}$  provides the LQE gain as

$$K_{LQE}(t) = P_{LQE}C^T(t)(M(t)\Sigma_{\vec{v}}M^T(t))^{-1} \quad (15.273)$$

Note that the unconstrained infinite-horizon continuous-time LQG OCP would require solving two CAREs and provide constant steady-state  $K_{LQR}$  and  $K_{LQE}$ .

The unconstrained finite horizon discrete-time LQG OCP is

$$\begin{aligned} \vec{u}^{opt}[k] = \underset{\vec{u}[k] \text{ for } k=0, \dots, N-1}{\operatorname{argmin}} \quad \mathcal{J} = \mathbb{E} \left[ \vec{x}^T[N] E \vec{x}[N] + \sum_{k=0}^{N-1} \vec{x}^T[k] Q[k] \vec{x}[k] + \vec{u}^T[k] R[k] \vec{u}[k] \right] \\ \text{subject to: } \vec{x}[k+1] = F[k] \vec{x}[k] + G[k] \vec{u}[k] + L[k] \vec{w}[k] \\ \vec{y}[k] = H[k] \vec{x}[k] + M[k] \vec{v}[k] \\ \text{initial condition: } \vec{\mu}_0 = \mathbb{E}[\vec{x}[k]], \Sigma_0 = \mathbb{E}[\vec{x}[k] \vec{x}^T[k]] \end{aligned} \quad (15.274)$$

where  $\vec{w}[k]$  and  $\vec{v}[k]$  are AWGN with covariances,  $\Sigma_{\vec{w}}$  and  $\Sigma_{\vec{v}}$ , respectively.

To solve this OCP, assume one can use an observer-based feedback controller, i.e.

$$\vec{u}[k] = -K_{LQR}[k] \hat{\vec{x}}[k] \quad (15.275)$$

where  $K_{LQR}[k]$  is the optimal discrete-time LQR gain matrix. The initial condition of the estimator is given by an initial estimate of the state,  $\hat{\vec{x}}[0] = \vec{\mu}_0$  and initial covariance,  $\Sigma_0$ . Recall that the state estimate  $\hat{\vec{x}}[k]$  evolves according to the **discrete-time Luenberger observer equation**

$$\hat{\vec{x}}[k+1] = F[k] \hat{\vec{x}}[k] + G[k] \vec{u}[k] + L \left( \vec{y}[k] - H[k] \hat{\vec{x}}[k] \right) \quad (15.276)$$

where for this stochastic OCP, the optimal estimator gain,  $K_{LQE}[k]$ , is the discrete-time **linear-quadratic estimator (LQE)**, also known as the **discrete-time Kalman gain**.

Recall that to find the optimal  $K_{LQR}[k]$ , one must solve the corresponding Riccati difference equation backwards-in-time for the closed-loop dynamics of the state controller, i.e.

$$\begin{aligned} P_{LQR}[k] = & F^T[k] P_{LQR}[k+1] F[k] + Q[k] \\ & - F^T[k] P_{LQR}[k+1] G[k] \left( G^T[k] P_{LQR}[k+1] G[k] + R[k] \right)^{-1} \\ & G^T[k] P_{LQR}[k+1] F[k] \end{aligned} \quad (15.277)$$

with final condition

$$P_{LQR}[k] = E \quad (15.278)$$

Then, solving for  $P_{LQR}[k]$  provides the optimal LQR gain as

$$K_{LQR}[k] = \left( G^T[k] P_{LQR}[k+1] G[k] + R[k] \right)^{-1} G^T[k] P_{LQR}[k+1] F[k] \quad (15.279)$$

Similarly, to find the optimal  $K_{LQE}[k]$ , one must solve the corresponding Riccati difference equation forwards-in-time for the closed-loop dynamics of the state estimator, i.e.

$$\begin{aligned} P_{LQE}[k+1] = & F[k] P_{LQE}[k] F^T[k] + L[k] \Sigma_{\vec{w}} L^T[k] \\ & - F[k] P_{LQE}[k] H^T[k] \left( H[k] P_{LQE}[k] H^T[k] + M[k] \Sigma_{\vec{v}} M^T[k] \right)^{-1} \\ & H[k] P_{LQE}[k] F^T[k] \end{aligned} \quad (15.280)$$

with initial condition

$$P_{LQE}[0] = \Sigma_0 \quad (15.281)$$

Then, solving for  $P_{LQE}$  provides the LQE gain as

$$K_{LQE}[k] = P_{LQE}[k]H^T[k] \left( H[k]P_{LQE}[k]H^T[k] + M[k]\Sigma_{\vec{v}}M^T[k] \right)^{-1} \quad (15.282)$$

Note, that the unconstrained infinite-horizon discrete-time LQG OCP would require solving two DAREs and provide constant steady-state  $K_{LQR}$  and  $K_{LQE}$ .

The unconstrained finite-horizon hybrid-time LQG OCP is

$$\begin{aligned} \vec{u}^{opt}(t) &= \underset{\vec{u}(t) \text{ for } t \geq 0}{\operatorname{argmin}} \mathcal{J} = \mathbb{E} \left[ \vec{x}^T(t_f)E\vec{x}(t_f) + \int_0^{t_f} \vec{x}^T(t)Q(t)\vec{x}(t) + \vec{u}^T(t)R(t)\vec{u}(t) \right] \\ \text{subject to: } &\vec{x}(t) = A(t)\vec{x}(t) + B(t)\vec{u}(t) + L(t)\vec{w}(t) \\ &\vec{y}[k] = H[k]\vec{x}[k] + M[k]\vec{v}[k] \\ \text{initial condition: } &\vec{\mu}_0 = \mathbb{E}[\vec{x}(0)], \Sigma_0 = \mathbb{E}[\vec{x}(0)\vec{x}^T(0)] \end{aligned} \quad (15.283)$$

where  $d\vec{w}(t)$  and  $\vec{v}[k]$  are AWGN with covariances,  $\Sigma_{\vec{w}}$  and  $\Sigma_{\vec{v}}$ , respectively,  $t = k\Delta t \forall k \in \mathbb{N}$ .

To solve this OCP, assume one can use an observer-based feedback controller, i.e.

$$\vec{u}(t) = -K_{LQR}(t)\hat{\vec{x}}(t) \quad (15.284)$$

where  $K_{LQR}(t)$  is the optimal continuous-time LQR gain matrix. The initial condition of the estimator is given by an initial estimate of the state,  $\hat{\vec{x}}(t) = \vec{\mu}_0$ , and initial covariance,  $\Sigma[0]$ . The state estimate  $\hat{\vec{x}}(t)$  evolves according to the **hybrid-time Luenberger observer equation**.

$$\dot{\hat{\vec{x}}}(t) = \begin{cases} A(t)\hat{\vec{x}}(t) + B(t)\vec{u}(t) + K_{LQE}[k] \left( \vec{y}[k] - H[k]\hat{\vec{x}}(t) \right) & \text{if } t = k\Delta t \forall k \in \mathbb{N} \\ A(t)\hat{\vec{x}}(t) + B(t)\vec{u}(t) & \text{otherwise} \end{cases} \quad (15.285)$$

For this stochastic OCP, the optimal estimator gain,  $K_{LQE}[k]$ , is the hybrid-time **linear-quadratic estimator (LQE)**, also known as the **hybrid-time Kalman gain**.

Recall that to find the optimal  $K_{LQR}(t)$ , one must solve the corresponding Riccati differential equation backwards-in-time for the closed-loop dynamics of the state controller, i.e.

$$\dot{P}_{LQR}(t) = -P_{LQR}(t)A(t) - A^T(t)P_{LQR}(t) + P_{LQR}(t)B(t)R^{-1}(t)B^T(t)P_{LQR}(t) - Q(t) \quad (15.286)$$

with final condition

$$P_{LQR}(t) = E \quad (15.287)$$

Then, solving for  $P_{LQR}(t)$  provides the optimal LQR gain as

$$K_{LQR}(t) = R^{-1}(t)B^T(t)P_{LQR}(t) \quad (15.288)$$

Similarly, to find the optimal  $K_{LQE}[k]$ , one must solve the corresponding  $k$  Riccati differential equations forwards-in-time for the closed-loop dynamics of the state estimator, i.e.

$$\dot{P}_{LQE}(\Delta t) = \int_0^{\Delta t} P_{LQE}(\tau)A(\tau) + P_{LQE}(\tau)A^T(\tau) + L(\tau)\Sigma_{\vec{w}}L^T(\tau)d\tau \quad \forall t \in [(k-1)\Delta t, k\Delta t] \quad (15.289)$$

with initial conditions as

$$P_{LQE}(0) = \begin{cases} \Sigma_0 & k = 0 \\ P_{LQE}[k-1] - P_{LQE}[k-1]H^T[k-1] \\ \left( H[k-1]P_{LQE}[k-1]H^T[k-1] + M[k-1]\Sigma_{\vec{v}}M^T[k-1] \right)^{-1} H[k-1]P_{LQE}[k-1] & k \geq 1 \in \mathbb{N} \end{cases} \quad (15.290)$$

Then, solving for  $P_{LQE}[k]$  as the  $k^{\text{th}}$  solution to  $\dot{P}_{LQE}(\Delta t)$  provides the LQE gain as

$$K_{LQE}[k] = P_{LQE}[k]H^T[k] \left( H[k]P_{LQE}[k]H^T[k] + M[k]\Sigma_{\vec{v}}M^T[k] \right)^{-1} \quad (15.291)$$

Note that if the state-space model is LTI, then one can compute the discretized matrices,  $F$  and  $L$ , and use those as a Riccati difference equation for solving for  $P_{LQE}[k]$  as

$$\begin{aligned} P_{LQE}[k+1] = & FP_{LQE}[k]F^T + L\Sigma_{\vec{w}}L^T \\ & - FP_{LQE}[k]H^T[k] \left( H[k]P_{LQE}[k]H^T[k] + M[k]\Sigma_{\vec{v}}M^T[k] \right)^{-1} H[k]P_{LQE}[k]F^T \end{aligned} \quad (15.292)$$

Then, the unconstrained infinite-horizon hybrid-time LQG OCP would require solving the CARE and the DARE to provide constant steady-state  $K_{LQR}$  and  $K_{LQE}$ .

To summarize these results, the LQG OCP solution is **separable** as it uses the optimal LQR gain and the Kalman gain for stochastic dynamical systems. The solution is obtained by solving the **dual** OCPs, i.e. solving the  $K_{LQE}$  Riccati equation forward-in-time from the initial condition and the  $K_{LQR}$  Riccati equation backward-in-time from the final condition. Also, note that at each step, the Kalman filter computes the state estimate  $\hat{x}$  using past  $\vec{y}$  and  $\vec{u}$  and the feedback controller generates  $\vec{u}$  using  $\hat{x}$  separately. Although no longer optimal, this framework of state feedback control separate from state estimation is considered as the primary design approach in the aerospace vehicle control and perception systems presented in this textbook even when the stochastic state-space model is nonlinear and non-Gaussian. Lastly, it should be mentioned that LQG OCPs can also be developed for MJLS, but these results are beyond the scope of this textbook.

## References

For more information, please refer to the following

- Gelb, A., Kasper, J. F., Nash, R. A., Price, C. F., and Sutherland, A. A, “4.6 Solution of the Riccati Equation,” in *Applied Optimal Estimation*, The M.I.T. Press, 1974, pp. 136-141
- Gelb, A., Kasper, J. F., Nash, R. A., Price, C. F., and Sutherland, A. A, “4.7 Statistical Steady State - The Wiener Filter,” in *Applied Optimal Estimation*, The M.I.T. Press, 1974, pp. 142-155
- Gelb, A., Kasper, J. F., Nash, R. A., Price, C. F., and Sutherland, A. A, “9.5 Optimal Control of Linear Systems,” in *Applied Optimal Estimation*, The M.I.T. Press, 1974, pp. 356-365
- Simon, D., “8.5.3 Duality,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 258-259

---

# Nonlinear Bayesian State Estimation Theory

## 16.1 Introduction to Nonlinear Bayesian State Estimation

Recall the Bayes filter is the optimal recursive solution to the general state estimation problem and can be written as prediction step for the prior PDF

$$f_{\vec{X}}(\vec{x}_k | \vec{y}_{1:k-1}) = \int_{-\infty}^{\infty} f_{\vec{X}}(\vec{x}_k | \vec{x}_{k-1}) f_{\vec{X}}(\vec{x}_{k-1} | \vec{y}_{1:k-1}) d\vec{x}_{k-1} \quad (16.1)$$

and correction step for the posterior PDF

$$f_{\vec{X}}(\vec{x}_k | \vec{y}_{1:k}) = \frac{f_{\vec{Y}}(\vec{y}_k | \vec{x}_k) f_{\vec{X}}(\vec{x}_k | \vec{y}_{1:k-1})}{\int_{-\infty}^{\infty} f_{\vec{Y}}(\vec{y}_k | \vec{x}_k) f_{\vec{X}}(\vec{x}_k | \vec{y}_{1:k-1}) d\vec{x}_k} \quad (16.2)$$

which usually only provides an exact closed-form solutions for linear-Gaussian systems, i.e., the discrete-time, continuous-time, and hybrid-time Kalman filters, and the Benes filter. Beyond these case, exact closed-form recursive solutions do not exist and one must make approximations to the optimal Bayes filter for nonlinear, non-Gaussian stochastic state-space models which are discussed in this chapter. Typically these filters are either **assumed density filters (ADF)** or **Monte-Carlo filters**.

### Discrete-Time Nonlinear Kalman Filtering

Motivated by the optimality of the discrete-time Kalman filter, one of the most common approximations is to approximate the prior and posterior distributions of the optimal Bayes filter as multivariate Gaussians. This approach is known as discrete-time **nonlinear Kalman filtering**, also known as **Gaussian filtering**, where the state mean/estimate,  $\hat{\vec{x}}_k$ , and the state covariance,  $P_k$ , are recursively updated as hyper-parameters representing the multivariate Gaussian approximation as opposed to calculating an exact posterior conditional PDF,  $f_{\vec{X}}(\vec{x}_k | \vec{y}_{1:k})$ , which may not be possible. In this context, consider the discrete-time stochastic state-

space model as

$$\begin{aligned}\vec{x}_k &= f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1}) \\ \vec{y}_k &= h_k(\vec{x}_k, \vec{v}_k)\end{aligned}\quad (16.3)$$

By definition of the mean and covariance, a discrete-time nonlinear Kalman filter has a prediction step for the prior mean given by

$$\hat{\vec{x}}_{k|k-1} = \mathbb{E}[f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1})] \quad (16.4)$$

the discrete-time prior covariance given by

$$P_{k|k-1} = \mathbb{E}[(\vec{x}_k - \hat{\vec{x}}_{k|k-1})(\vec{x}_k - \hat{\vec{x}}_{k|k-1})^T] \quad (16.5)$$

and the predicted measurement given by

$$\hat{\vec{y}}_k = \mathbb{E}[h_k(\vec{x}_k, \vec{v}_k)] \quad (16.6)$$

as well as a correction step for the measurement covariance given by

$$P_{y_k} = S_k = \mathbb{E}[(\vec{y}_k - \hat{\vec{y}}_k)(\vec{y}_k - \hat{\vec{y}}_k)^T] \quad (16.7)$$

the state-measurement cross-covariance given by

$$P_{x_k y_k} = \mathbb{E}[(\vec{x}_k - \hat{\vec{x}}_{k|k-1})(\vec{y}_k - \hat{\vec{y}}_k)^T] \quad (16.8)$$

the optimal gain given by

$$K_k = P_{x_k y_k} P_{y_k}^{-1} = P_{x_k y_k} S_k^{-1} \quad (16.9)$$

the posterior mean given by

$$\hat{\vec{x}}_{k|k} = \hat{\vec{x}}_{k|k-1} + K_k [\vec{y}_k - \hat{\vec{y}}_k] \quad (16.10)$$

and the posterior covariance given by

$$P_{k|k} = P_{k|k-1} - K_k P_{y_k} K_k^T \quad (16.11)$$

The conceptually most straightforward nonlinear Kalman filter is the extended Kalman filter (EKF) which approximates the expectation integrals for the nonlinear Kalman filter using a first-order Jacobian linearization of the nonlinear system about the current state estimate. This approach also allows one to easily apply this linearization to the continuous- and hybrid-time stochastic state-space models. However, the EKF disregards the “probabilistic spread” of the state vector as the Jacobian linearization is defined about a *single* point. Thus, this approximation can produce large errors of the approximated prior and posterior mean and covariance relative to the true prior and posterior mean and covariance which leads to poor performance and even divergence.

## 16.2 Extended Kalman Filtering and Smoothing

### Discrete-Time Extended Kalman Filter

Consider the Jacobian linearization of the process equation about the posterior mean,  $\hat{\vec{x}}_{k-1|k-1}$ , and the mean of the process noise,  $\vec{0}$ , i.e.,

$$\begin{aligned}\vec{x}_k = & f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1}) \approx f_{k-1}(\hat{\vec{x}}_{k-1|k-1}, \vec{u}_{k-1}, \vec{0}) \\ & + \frac{\partial f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1})}{\partial \vec{x}_{k-1}} \Big|_{\vec{x}_{k-1}=\hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1}=\vec{0}} (\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1}) \\ & + \frac{\partial f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1})}{\partial \vec{w}_{k-1}} \Big|_{\vec{x}_{k-1}=\hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1}=\vec{0}} (\vec{w}_{k-1} - \vec{0})\end{aligned}\quad (16.12)$$

Then, defining the linearized state matrix as

$$F_{k-1} = \frac{\partial f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1})}{\partial \vec{x}_{k-1}} \Big|_{\vec{x}_{k-1}=\hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1}=\vec{0}} \quad (16.13)$$

and the linearized process noise gain matrix

$$L_{k-1} = \frac{\partial f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1})}{\partial \vec{w}_{k-1}} \Big|_{\vec{x}_{k-1}=\hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1}=\vec{0}} \quad (16.14)$$

one has

$$\vec{x}_k \approx f_{k-1}(\hat{\vec{x}}_{k-1|k-1}, \vec{u}_{k-1}, 0) + F_{k-1}(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1}) + L_{k-1} \vec{w}_{k-1} \quad (16.15)$$

By definition, the prior state mean is given by

$$\hat{\vec{x}}_{k|k-1} = \mathbb{E}[f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1})] \quad (16.16)$$

Substituting by the approximation, one has

$$\hat{\vec{x}}_{k|k-1} \approx \mathbb{E}[f_{k-1}(\hat{\vec{x}}_{k-1|k-1}, \vec{u}_{k-1}, \vec{0}) + F_{k-1}(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1}) + L_{k-1} \vec{w}_{k-1}] \quad (16.17)$$

which by definition that  $\mathbb{E}(\hat{\vec{x}}_{k|k-1}) = \vec{x}_k$  and zero-mean noise, one has

$$\hat{\vec{x}}_{k|k-1} \approx f_{k-1}(\hat{\vec{x}}_{k-1|k-1}, \vec{u}_{k-1}, \vec{0}) \quad (16.18)$$

By definition, the prior state covariance is given by

$$P_{k|k-1} = \mathbb{E}[(\vec{x}_k - \hat{\vec{x}}_{k|k-1})(\vec{x}_k - \hat{\vec{x}}_{k|k-1})^T] \quad (16.19)$$

Substituting by the approximation for  $\vec{x}_k$  and the prior state mean, one has

$$P_{k|k-1} = \mathbb{E}[(F_{k-1}(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1}) + L_{k-1} \vec{w}_{k-1})(F_{k-1}(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1}) + L_{k-1} \vec{w}_{k-1})^T] \quad (16.20)$$

or

$$\begin{aligned}P_{k|k-1} = & F_{k-1} \mathbb{E}[(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1})(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1})^T] F_{k-1}^T + L_{k-1} \mathbb{E}[\vec{w}_{k-1} \vec{w}_{k-1}^T] L_{k-1}^T \\ & F_{k-1} \mathbb{E}[(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1}) \vec{w}_{k-1}^T] L_{k-1}^T + L_{k-1} \mathbb{E}[\vec{w}_{k-1} (\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1})^T] F_{k-1}^T\end{aligned}\quad (16.21)$$

For uncorrelated state and process noise, one has

$$P_{k|k-1} = F_{k-1} P_{k-1|k-1} F_{k-1}^T + L_{k-1} Q_{k-1} L_{k-1}^T \quad (16.22)$$

Consider the Jacobian linearization of the measurement equation about the prior mean,  $\hat{\vec{x}}_{k|k-1}$ , and the mean of the measurement noise,  $\vec{0}$ , i.e.,

$$\vec{y}_k = h_k(\vec{x}_k, \vec{v}_k) \approx h_k(\hat{\vec{x}}_{k|k-1}, \vec{0}) + \frac{\partial h_k(\vec{x}_k, \vec{v}_k)}{\partial \vec{x}_k} \Big|_{\vec{x}_k = \hat{\vec{x}}_{k|k-1}, \vec{v}_k = \vec{0}} (\vec{x}_k - \hat{\vec{x}}_{k|k-1}) + \frac{\partial h_k(\vec{x}_k, \vec{v}_k)}{\partial \vec{v}_k} \Big|_{\vec{x}_k = \hat{\vec{x}}_{k|k-1}, \vec{v}_k = \vec{0}} (\vec{v}_k - \vec{0}) \quad (16.23)$$

Then, defining the linearized output matrix as

$$H_k = \frac{\partial h_k(\vec{x}_k, \vec{v}_k)}{\partial \vec{x}_k} \Big|_{\vec{x}_k = \hat{\vec{x}}_{k|k-1}, \vec{v}_k = \vec{0}} \quad (16.24)$$

and the linearized measurement noise gain matrix

$$M_k = \frac{\partial h_k(\vec{x}_k, \vec{v}_k)}{\partial \vec{v}_k} \Big|_{\vec{x}_k = \hat{\vec{x}}_{k|k-1}, \vec{v}_k = \vec{0}} \quad (16.25)$$

one has

$$\vec{y}_k \approx h_k(\hat{\vec{x}}_{k|k-1}, \vec{0}) + H_k(\vec{x}_k - \hat{\vec{x}}_{k|k-1}) + M_k \vec{v}_k \quad (16.26)$$

By definition, the predicted measurement is given by

$$\hat{\vec{y}}_k = \mathbb{E}[h_k(\vec{x}_k, \vec{v}_k)] \quad (16.27)$$

Substituting by the approximation, one has

$$\hat{\vec{y}}_k \approx \mathbb{E}[h_k(\hat{\vec{x}}_{k|k-1}, \vec{0}) + H_k(\vec{x}_k - \hat{\vec{x}}_{k|k-1}) + M_k \vec{v}_k] \quad (16.28)$$

which by definition that  $\mathbb{E}(\hat{\vec{x}}_{k|k-1}) = \vec{x}_k$  and zero-mean noise, one has

$$\hat{\vec{y}}_k \approx h_k(\hat{\vec{x}}_{k|k-1}, \vec{0}) \quad (16.29)$$

By definition, the measurement covariance is given by

$$S_k = \mathbb{E}[(\vec{y}_k - \hat{\vec{y}}_k)(\vec{y}_k - \hat{\vec{y}}_k)^T] \quad (16.30)$$

Substituting by the approximation for  $\vec{y}_k$  and the predicted measurement, one has

$$S_k = \mathbb{E}[(H_k(\vec{x}_k - \hat{\vec{x}}_{k|k-1}) + M_k \vec{v}_k)(H_k(\vec{x}_k - \hat{\vec{x}}_{k|k-1}) + M_k \vec{v}_k)^T] \quad (16.31)$$

or

$$\begin{aligned} S_k = & H_k \mathbb{E}[(\vec{x}_k - \hat{\vec{x}}_{k|k-1})(\vec{x}_k - \hat{\vec{x}}_{k|k-1})^T] H_k^T + M_k \mathbb{E}[\vec{v}_k \vec{v}_k^T] M_k^T \\ & H_k \mathbb{E}[(\vec{x}_k - \hat{\vec{x}}_{k|k-1}) \vec{v}_k^T] M_k^T + M_k \mathbb{E}[\vec{v}_k (\vec{x}_k - \hat{\vec{x}}_{k|k-1})^T] H_k^T \end{aligned} \quad (16.32)$$

For uncorrelated state and measurement noise, one has

$$S_k = H_k P_{k|k-1} H_k^T + M_k R_{k-1} M_k^T \quad (16.33)$$

By definition, the state-measurement cross-covariance given by

$$P_{x_k y_k} = \mathbb{E} [(\vec{x}_k - \hat{\vec{x}}_{k|k-1})(\vec{y}_k - \hat{\vec{y}}_k)^T] \quad (16.34)$$

Substituting by the approximation for  $\vec{y}_k$  and the predicted measurement, one has

$$P_{x_k y_k} = \mathbb{E} [(F_{k-1}(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1}) + L_{k-1} \vec{w}_{k-1})(H_k(\vec{x}_k - \vec{x}_{k|k-1}) + M_k \vec{v}_k)^T] \quad (16.35)$$

$$\begin{aligned} P_{x_k y_k} &= \mathbb{E} [F_{k-1}(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1})(\vec{x}_k - \vec{x}_{k|k-1})] H_k^T + L_{k-1} \vec{w}_{k-1} \vec{v}_k^T M_k^T \\ &\quad + F_{k-1} \mathbb{E} [(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1}) \vec{v}_k^T] M_k^T + L_{k-1} \mathbb{E} [\vec{w}_{k-1} (\vec{x}_k - \vec{x}_{k|k-1})^T] H_k^T \end{aligned} \quad (16.36)$$

For uncorrelated state, process noise, and measurement noise, one has

$$P_{x_k y_k} = F_{k-1} \mathbb{E} [F_{k-1}(\vec{x}_{k-1} - \hat{\vec{x}}_{k-1|k-1})(\vec{x}_{k-1} - \vec{x}_{k-1|k-1})^T F_{k-1}^T] H_k^T \quad (16.37)$$

or

$$P_{x_k y_k} = P_{k|k-1} H_k^T \quad (16.38)$$

In summary, the discrete-time **EKF initialization step** sets the state estimate and state covariance at  $k = 0$  as

$$\begin{aligned} \hat{\vec{x}}_{0|0} &= \hat{\vec{x}}_0 = \mathbb{E} [\vec{x}_0] \\ P_{0|0} &= P_0 = \mathbb{E} [(\vec{x}_0 - \hat{\vec{x}}_{0|0})(\vec{x}_0 - \hat{\vec{x}}_{0|0})^T] \end{aligned} \quad (16.39)$$

The discrete-time **EKF prediction step** predicts the state estimate, the state covariance, and the output using the input, process model, and Jacobians as

$$\begin{aligned} \hat{\vec{x}}_{k|k-1} &= f(\hat{\vec{x}}_{k-1|k-1}, \vec{u}_{k-1}, \vec{0}) \\ \hat{\vec{y}}_k &= h(\hat{\vec{x}}_{k|k-1}, \vec{0}) \\ F_{k-1} &= \left[ \frac{\partial f}{\partial \vec{x}} \right]_{\vec{x}_{k-1}=\hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1}=\vec{0}} \\ L_{k-1} &= \left[ \frac{\partial f}{\partial \vec{w}} \right]_{\vec{x}_{k-1}=\hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1}=\vec{0}} \\ P_{k|k-1} &= F_{k-1} P_{k-1|k-1} F_{k-1}^T + L_{k-1} Q_{k-1} L_{k-1}^T \end{aligned} \quad (16.40)$$

The discrete-time **EKF correction step** updates the state estimate and the state covariance using the measurement model and Jacobians as

$$\begin{aligned}
 \tilde{y}_k &= \vec{y}_k - \hat{\vec{y}}_k \\
 H_k &= \left[ \frac{\partial h}{\partial \vec{x}} \right]_{\vec{x} = \hat{\vec{x}}_{k|k-1}, \vec{v} = \vec{0}} \\
 M_k &= \left[ \frac{\partial h}{\partial \vec{v}} \right]_{\vec{x} = \hat{\vec{x}}_{k|k-1}, \vec{v} = \vec{0}} \\
 S_k &= H_k P_{k|k-1} H_k^T + M_k R_k M_k^T \\
 K_k &= P_{k|k-1} H_k^T S_k^{-1} \\
 \hat{\vec{x}}_{k|k} &= \hat{\vec{x}}_{k|k-1} + K_k \tilde{y}_k \\
 P_{k|k} &= P_{k|k-1} - K_k S_k K_k^T
 \end{aligned} \tag{16.41}$$

### Continuous-Time Extended Kalman Filter

Consider the continuous-time stochastic time-invariant state-space model about some nominal trajectory

$$\begin{aligned}
 \dot{\vec{x}}_0 &= f(\vec{x}_0, \vec{u}_0, \vec{w}_0, t) \\
 \vec{y}_0 &= h(\vec{x}_0, \vec{v}_0, t)
 \end{aligned} \tag{16.42}$$

Next, define the linearized state matrix as

$$A = \left[ \frac{\partial f}{\partial \vec{x}} \right]_{\vec{x} = \vec{x}_0, \vec{w} = \vec{w}_0} \tag{16.43}$$

the linearized process noise gain matrix

$$L = \left[ \frac{\partial f}{\partial \vec{w}} \right]_{\vec{x} = \vec{x}_0, \vec{w} = \vec{w}_0} \tag{16.44}$$

the linearized output matrix as

$$C = \left[ \frac{\partial h}{\partial \vec{x}} \right]_{\vec{x} = \vec{x}_0, \vec{v} = \vec{v}_0} \tag{16.45}$$

and the linearized measurement noise gain matrix

$$M = \left[ \frac{\partial h}{\partial \vec{v}} \right]_{\vec{x} = \vec{x}_0, \vec{v} = \vec{v}_0} \tag{16.46}$$

Then, the Jacobian linearization can be written as

$$\begin{aligned}
 \dot{\vec{x}} &= f(\vec{x}, \vec{u}, \vec{w}) = f(\vec{x}_0, \vec{u}_0, \vec{w}_0) + A(\vec{x} - \vec{x}_0) + B(\vec{u} - \vec{u}_0) + L(\vec{w} - \vec{w}_0) \\
 \vec{y} &= h(\vec{x}, \vec{v}) = h(\vec{x}_0, \vec{v}_0) + C(\vec{x} - \vec{x}_0) + M(\vec{v} - \vec{v}_0)
 \end{aligned} \tag{16.47}$$

or using perturbation notation and the definitions of  $\dot{\vec{x}}_0$  and  $\vec{y}_0$ , one has

$$\begin{aligned}\Delta \dot{\vec{x}} &= A\Delta \vec{x} + B\Delta \vec{u} + L\Delta \vec{w} \\ \Delta \vec{y} &= C\Delta \vec{x} + M\Delta \vec{v}\end{aligned}\quad (16.48)$$

and assuming  $u(t)$  is perfectly known, i.e.,  $\Delta u(t) = 0$ , and  $\vec{w}_0 = \vec{v}_0 = \vec{0}$ , one can write

$$\begin{aligned}\Delta \dot{\vec{x}} &= A\Delta \vec{x} + L\vec{w} \\ \Delta \vec{y} &= C\Delta \vec{x} + M\vec{v}\end{aligned}\quad (16.49)$$

where  $\vec{w}$  and  $\vec{v}$  have covariances,  $Q$  and  $R$ , respectively.

These equations form a linear dynamical system of the state perturbation,  $\Delta \hat{\vec{x}}$ , which can be estimated using the continuous-time Kalman filter as

$$\begin{aligned}\dot{P} &= AP + PA^T - PC^T(MRM^T)^{-1}CP + LQL^T \\ K &= PC^T(MRM^T)^{-1} \\ \Delta \dot{\hat{\vec{x}}} &= A\Delta \hat{\vec{x}} + K(\Delta \vec{y} - C\Delta \hat{\vec{x}})\end{aligned}\quad (16.50)$$

and added to the nominal estimator to obtain

$$\hat{\vec{x}} = \vec{x}_0 + \Delta \hat{\vec{x}} \quad (16.51)$$

Furthermore, note that one can write

$$\dot{\vec{x}}_0 + \Delta \dot{\hat{\vec{x}}} = f(\vec{x}_0, \vec{u}_0, \vec{0}, t) + A\Delta \hat{\vec{x}} + K(\vec{y} - \vec{y}_0 - C\Delta \hat{\vec{x}}) \quad (16.52)$$

If one chooses  $\vec{x}_0(t) = \hat{\vec{x}}(t)$  so that  $\Delta \hat{\vec{x}}(t) = 0$  and  $\Delta \dot{\hat{\vec{x}}}(t) = 0$ , i.e., the linearization trajectory is equivalent to the linearized Kalman filter state estimate,  $\hat{\vec{x}}(t)$ , which also infers the nominal measurement as

$$\vec{y}_0 = h(\hat{\vec{x}}, \vec{0}, t) \quad (16.53)$$

and assume that one has some *prior* estimate of the state  $\hat{\vec{x}}_0$  with some expected uncertainty  $P_0$  where

$$\mathbb{E}[\vec{x}] = \hat{\vec{x}}_0 \quad (16.54)$$

and

$$\mathbb{E}[(\vec{x} - \hat{\vec{x}}_0)(\vec{x} - \hat{\vec{x}}_0)^T] = P_0 \quad (16.55)$$

$$\begin{aligned}\hat{\vec{x}}_{0|0} &= \hat{\vec{x}}_0 = \\ P_{0|0} &= P_0\end{aligned}\quad (16.56)$$

Then, one obtains the **continuous-time extended Kalman filter (EKF)**, also known as the **extended Kalman-Bucy filter (EKBF)**, has an initialization step as

$$\begin{aligned}\hat{\vec{x}}_{0|0} &= \hat{\vec{x}}_0 = \mathbb{E} [\vec{x}_0] \\ P_{0|0} &= P_0 = \mathbb{E} [(\vec{x}_0 - \hat{\vec{x}}_{0|0})(\vec{x}_0 - \hat{\vec{x}}_{0|0})^T] \\ A &= \left[ \frac{\partial f}{\partial \vec{x}} \right]_{\vec{x}=\hat{\vec{x}}, \vec{w}=\vec{0}} \\ L &= \left[ \frac{\partial f}{\partial \vec{w}} \right]_{\vec{x}=\hat{\vec{x}}, \vec{w}=\vec{0}} \\ C &= \left[ \frac{\partial h}{\partial \vec{x}} \right]_{\vec{x}=\hat{\vec{x}}, \vec{v}_k=\vec{0}} \\ M &= \left[ \frac{\partial h}{\partial \vec{v}} \right]_{\vec{x}=\hat{\vec{x}}, \vec{v}_k=\vec{0}}\end{aligned}\tag{16.57}$$

and differential equations as

$$\begin{aligned}\dot{P} &= AP + PA^T - PC^T(MRM^T)^{-1}CP + LQL^T \\ K &= PC^T(MRM^T)^{-1} \\ \dot{\hat{\vec{x}}} &= f(\hat{\vec{x}}, \vec{u}, \vec{0}, t) + K(\vec{y} - h(\hat{\vec{x}}, \vec{0}, t))\end{aligned}\tag{16.58}$$

### Hybrid-Time Extended Kalman Filter

Lastly, it should be noted that often in practice, one has continuous-time dynamics and discrete-time measurements, which is known as a **hybrid-time stochastic state-space model**

$$\begin{aligned}\dot{\vec{x}}(t) &= f(\vec{x}(t), \vec{u}(t), \vec{w}(t)) \\ \vec{y}_k &= h(\vec{x}_k, \vec{v}_k)\end{aligned}\tag{16.59}$$

for which one can form the **hybrid-time EKF** by utilizing the nonlinear differential equation and the Riccati differential equation for the prediction step and the discrete-time EKF equations for the correction step. The hybrid-time EKF can be considered as an alternative to discretizing  $f(\vec{x}(t), \vec{u}(t), \vec{w}(t))$  using and implementing a discrete-time EKF. This type of discretization may or may not be tractable for general nonlinear dynamical systems. Both discrete-time and hybrid-time use integrals for propagation or discretization which may require numerical methods.

In short, the hybrid-time **EKF initialization step** sets the state estimate and state covariance at  $k = 0$  as

$$\begin{aligned}\hat{\vec{x}}_{0|0} &= \hat{\vec{x}}_0 = \mathbb{E} [\vec{x}_0] \\ P_{0|0} &= P_0 = \mathbb{E} [(\vec{x}_0 - \hat{\vec{x}}_{0|0})(\vec{x}_0 - \hat{\vec{x}}_{0|0})^T]\end{aligned}\tag{16.60}$$

The hybrid-time **EKF prediction step** predicts the state estimate, the state covariance, and the output using the input, process model, and Jacobians as

$$\begin{aligned}
 \hat{\vec{x}}_{k-1}(t) &= \int_{t(k-1)}^{t(k)} f(\hat{\vec{x}}, \vec{u}, 0) \text{ with } \hat{\vec{x}}_{k-1}(t(k-1)) = \hat{\vec{x}}_{k-1|k-1} \\
 \hat{\vec{x}}_{k|k-1} &= \hat{\vec{x}}_{k-1}(t(k)) \\
 \hat{\vec{y}}_k &= h(\hat{\vec{x}}_{k|k-1}, \vec{0}) \\
 A_{k-1}(t) &= \left[ \frac{\partial f}{\partial \vec{x}} \right]_{\vec{x}=\hat{\vec{x}}_{k-1}(t), \vec{w}_{k-1}=\vec{0}} \\
 L_{k-1}(t) &= \left[ \frac{\partial f}{\partial \vec{w}} \right]_{\vec{x}=\hat{\vec{x}}_{k-1}(t), \vec{w}_{k-1}=\vec{0}} \\
 P_{k|k-1} &= P_{k-1|k-1} + \int_{t(k-1)}^{t(k)} A_{k-1}P + PA_{k-1}^T + L_{k-1}Q_{k-1}L_{k-1}^T
 \end{aligned} \tag{16.61}$$

The hybrid-time **EKF correction step** updates the state estimator and the state covariance using the measurement model and Jacobians as

$$\begin{aligned}
 \tilde{y}_k &= \vec{y}_k - \hat{\vec{y}}_k \\
 H_k &= \left[ \frac{\partial h}{\partial \vec{x}} \right]_{\vec{x}_k=\hat{\vec{x}}_{k|k-1}, \vec{v}_k=\vec{0}} \\
 M_k &= \left[ \frac{\partial h}{\partial \vec{v}} \right]_{\vec{x}_k=\hat{\vec{x}}_{k|k-1}, \vec{v}_k=\vec{0}} \\
 S_k &= H_k P_{k|k-1} H_k^T + M_k R_k M_k^T \\
 K_k &= P_{k|k-1} H_k^T S_k^{-1} \\
 \hat{\vec{x}}_{k|k} &= \hat{\vec{x}}_{k|k-1} + K_k \tilde{y}_k \\
 P_{k|k} &= P_{k|k-1} - K_k S_k K_k^T
 \end{aligned} \tag{16.62}$$

### Iterative Extended Kalman Filter

At each correction step for the discrete- and hybrid-time EKFs, the Jacobian matrices depend on the prior state estimate,  $\vec{x}_{k|k-1}$ , which, in turn, affects the Kalman gain computation, the posterior state estimate, and the posterior state covariance. An alternative EKF approach is the **iterative extended Kalman filter (IEKF)** which iterates across some fixed number of iterations, i.e., for  $i = 0, 1, \dots, N$ , for the Jacobian linearizations of the measurement matrix,  $H_{k,i}$ , the measurement noise gain matrix,  $M_{k,i}$ , the innovation covariance,  $S_{k,i}$ , and the subsequent Kalman gain,  $K_{k,i}$ , based on iterative state estimates in the correction step.

In this case, one must also update innovation using a first-order Taylor series approximation evaluated at the iterated state estimate,  $\hat{\vec{x}}_{k,i}$ , i.e.,

$$\tilde{y}_k = \vec{y}_k - h(\vec{x}_{k|k-1}, \vec{v}_k) \approx h(\vec{x}_{k,i}, \vec{v}_k) + H_{k,i}(\hat{\vec{x}}_{k|k-1} - \hat{\vec{x}}_{k,i}) \tag{16.63}$$

In summary, the discrete-time **IEKF initialization step** sets the state estimate and state covariance at  $k = 0$  as

$$\begin{aligned}\hat{\vec{x}}_{0|0} &= \hat{\vec{x}}_0 = \mathbb{E} [\vec{x}_0] \\ P_{0|0} &= P_0 = \mathbb{E} [(\vec{x}_0 - \hat{\vec{x}}_{0|0})(\vec{x}_0 - \hat{\vec{x}}_{0|0})^T]\end{aligned}\quad (16.64)$$

The discrete-time **IEKF prediction step** predicts the state estimate, the state covariance, and the output using the input, process model, and Jacobians as

$$\begin{aligned}\hat{\vec{x}}_{k|k-1} &= f_{k-1}(\hat{\vec{x}}_{k-1|k-1}, \vec{u}_{k-1}, \vec{0}) \\ F_{k-1} &= \left[ \frac{\partial f_{k-1}}{\partial \vec{x}} \right]_{\vec{x}_{k-1}=\hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1}=\vec{0}} \\ L_{k-1} &= \left[ \frac{\partial f_{k-1}}{\partial \vec{w}} \right]_{\vec{x}_{k-1}=\hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1}=\vec{0}} \\ P_{k|k-1} &= F_{k-1}P_{k-1|k-1}F_{k-1}^T + L_{k-1}Q_{k-1}L_{k-1}^T\end{aligned}\quad (16.65)$$

which could alternatively be cast as a hybrid-time IEKF prediction step.

$$\begin{aligned}\hat{\vec{x}}_{k-1}(t) &= \int_{t(k-1)}^{t(k)} f_{k-1}(\hat{\vec{x}}, \vec{u}, 0) \text{ with } \hat{\vec{x}}_{k-1}(t(k-1)) = \hat{\vec{x}}_{k-1|k-1} \\ \hat{\vec{x}}_{k|k-1} &= \hat{\vec{x}}_{k-1}(t(k)) \\ A_{k-1}(t) &= \left[ \frac{\partial f_{k-1}}{\partial \vec{x}} \right]_{\vec{x}=\hat{\vec{x}}_{k-1}(t), \vec{w}=\vec{0}} \\ L_{k-1}(t) &= \left[ \frac{\partial f_{k-1}}{\partial \vec{w}} \right]_{\vec{x}=\hat{\vec{x}}_{k-1}(t), \vec{w}=\vec{0}} \\ P_{k|k-1} &= P_{k-1|k-1} + \int_{t(k-1)}^{t(k)} A_{k-1}P + PA_{k-1}^T + L_{k-1}Q_{k-1}L_{k-1}^T\end{aligned}\quad (16.66)$$

The discrete- and hybrid-time **IEKF correction step** iteratively updates the state estimate and the state covariance using the measurement model and Jacobians for  $i = 0, 1, \dots, N$

$$\begin{aligned}H_{k,i} &= \left[ \frac{\partial h_k}{\partial \vec{x}} \right]_{\vec{x}_k=\hat{\vec{x}}_{k,i}, \vec{v}_k=\vec{0}} \\ M_{k,i} &= \left[ \frac{\partial h_k}{\partial \vec{v}} \right]_{\vec{x}_k=\hat{\vec{x}}_{k,i}, \vec{v}_k=\vec{0}} \\ S_{k,i} &= H_{k,i}P_{k|k-1}H_{k,i}^T + M_{k,i}R_kM_{k,i}^T \\ K_{k,i} &= P_{k|k-1}H_{k,i}^TS_{k,i}^{-1} \\ \tilde{y}_{k,i} &= \vec{y}_k - h_k(\hat{\vec{x}}_{k,i}, \vec{0}) - H_{k,i}(\hat{\vec{x}}_{k|k-1} - \hat{\vec{x}}_{k,i}) \\ \hat{\vec{x}}_{k,i+1} &= \hat{\vec{x}}_{k|k-1} + K_{k,i}\tilde{y}_{k,i} \\ P_{k,i+1} &= P_{k|k-1} - K_{k,i}S_{k,i}K_{k,i}^T\end{aligned}\quad (16.67)$$

where

$$\begin{aligned}\hat{\vec{x}}_{k,0} &= \hat{\vec{x}}_{k|k-1} \\ P_{k,0} &= P_{k|k-1}\end{aligned}\quad (16.68)$$

and

$$\begin{aligned}\hat{\vec{x}}_{k|k} &= \hat{\vec{x}}_{k,N+1} \\ P_{k|k} &= P_{k,N+1}\end{aligned}\quad (16.69)$$

### Innovation-Saturated Extended Kalman Filter

In practice, a sensor may experience large outliers,  $\vec{d}_k$ , which corrupt the nominal measurement model used in the filter, i.e., one can model the stochastic state-space system as

$$\begin{aligned}\vec{x}_k &= f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1}) \\ \vec{y}_k &= h_k(\vec{x}_k, \vec{v}_k) + \vec{d}_k\end{aligned}\quad (16.70)$$

where  $\vec{d}_k$  are spontaneous outliers which are difficult to detect or characterize which hinder the use of an augmented state. One simple method to detecting and rejecting outliers is to compute if the innovation for each element  $i = 1, \dots, n_y$  is within its 99% confidence bounds, i.e., one rejects the  $i^{\text{th}}$  measurement if

$$\vec{y}_{i,k} - h_{i,k}(\hat{\vec{x}}_{k|k-1}, \vec{0}) > 3\sqrt{[S_k]_{i,i}} \quad (16.71)$$

However, this simple method may not provide suitable performance to the EKF.

Another approach to mitigate the effects of these “hybrid-measurements” is through a saturation operation on the effect of the innovation on the correction step due to potential outliers. This idea leads to the **innovation-saturation extended Kalman filter** which uses a **saturation function** for each innovation  $i = 1, \dots, n_y$  as

$$\text{sat}_{\sqrt{\sigma_i}}(\tilde{y}_i) = \max\{-\sqrt{\sigma_i}, \min\{\sqrt{\sigma_i}, \tilde{y}_i\}\} \quad (16.72)$$

where  $\sigma_i > 0$  is the  $i^{\text{th}}$  **saturation bound** and the range  $[-\sqrt{\sigma_i}, \sqrt{\sigma_i}]$  defines the expected range of the innovation. With this function, one can dynamically adapt the  $i^{\text{th}}$  saturation bound with the dynamics model

$$\begin{aligned}\sigma_{i,k} &= \lambda_{1,i}\sigma_{i,k} + \gamma_{1,i}\epsilon_{i,k} \exp(-\epsilon_{i,k}) \\ \epsilon_{i,k} &= \lambda_{2,i}\epsilon_{i,k} + \gamma_{2,i}(\vec{y}_{i,k} - h_{i,k}(\hat{\vec{x}}_{k|k-1}, \vec{0}))^2\end{aligned}\quad (16.73)$$

with  $0 < \lambda_{1,i} < 1$ ,  $0 < \lambda_{2,i} < 1$ ,  $\gamma_{1,i} > 0$ , and  $\gamma_{2,i} > 0$ . These recursions serve to dynamically change  $\sigma_{i,k}$  to enable adaptation of the saturation bounds.

This formulation specifically involves a double-layer structure. The lower layer for  $\epsilon_{i,k}$  tracks the changes in the innovation signal, maintaining an appropriate level when the innovation is within normality, but becoming large when the innovation is altered by outliers. The upper layer for  $\sigma_{i,k}$  will be driven to increase for relatively small  $\epsilon_{i,k}$  when the innovation is within normality, but it will quickly reduce  $\sigma_{i,k}$  when  $\epsilon_{i,k}$  is large due to outliers. This adaptive approach for  $\sigma_{i,k}$  is intended to achieve a discernment between an outlier and a normal measurement, filtering innovations if corrupted by outliers and passing normal innovations to correct the prediction.

Thus, the discrete-time **IS-EKF initialization step** sets the state estimate and state covariance at  $k = 0$  as

$$\begin{aligned}\hat{\vec{x}}_{0|0} &= \hat{\vec{x}}_0 = \mathbb{E} [\vec{x}_0] \\ P_{0|0} &= P_0 = \mathbb{E} [(\vec{x}_0 - \hat{\vec{x}}_{0|0})(\vec{x}_0 - \hat{\vec{x}}_{0|0})^T] \\ \sigma_{i,1} &> 0 \\ \epsilon_{i,1} &> 0\end{aligned}\tag{16.74}$$

The discrete-time **IS-EKF prediction step** predicts the state estimate, the state covariance, and the output using the input, process model, and Jacobians as

$$\begin{aligned}\hat{\vec{x}}_{k|k-1} &= f(\hat{\vec{x}}_{k-1|k-1}, \vec{u}_{k-1}, \vec{0}) \\ \hat{\vec{y}}_k &= h(\hat{\vec{x}}_{k|k-1}, \vec{0}) \\ F_{k-1} &= \left[ \frac{\partial f}{\partial \vec{x}} \right]_{\vec{x}_{k-1}=\hat{\vec{x}}_{k-1|k-1}, \vec{v}_{k-1}=\vec{0}} \\ L_{k-1} &= \left[ \frac{\partial f}{\partial \vec{v}} \right]_{\vec{x}_{k-1}=\hat{\vec{x}}_{k-1|k-1}, \vec{v}_{k-1}=\vec{0}} \\ P_{k|k-1} &= F_{k-1} P_{k-1|k-1} F_{k-1}^T + L_{k-1} Q_{k-1} L_{k-1}^T\end{aligned}\tag{16.75}$$

The discrete-time **IS-EKF correction step** updates the state estimate and the state covariance using the measurement model and Jacobians as

$$\begin{aligned}\tilde{\vec{y}}_k &= \vec{y}_k - \hat{\vec{y}}_k \\ H_k &= \left[ \frac{\partial h}{\partial \vec{x}} \right]_{\vec{x}_k=\hat{\vec{x}}_{k|k-1}, \vec{v}_k=\vec{0}} \\ M_k &= \left[ \frac{\partial h}{\partial \vec{v}} \right]_{\vec{x}_k=\hat{\vec{x}}_{k|k-1}, \vec{v}_k=\vec{0}} \\ S_k &= H_k P_{k|k-1} H_k^T + M_k R_k M_k^T \\ K_k &= P_{k|k-1} H_k^T S_k^{-1} \\ \hat{\vec{x}}_{k|k} &= \hat{\vec{x}}_{k|k-1} + K_k \text{sat}_{\sigma}(\tilde{\vec{y}}_k) \\ P_{k|k} &= P_{k|k-1} - K_k S_k K_k^T \\ \sigma_{i,k+1} &= \lambda_{1,i} \sigma_{i,k} + \gamma_{1,i} \epsilon_{i,k} \exp(-\epsilon_{i,k}) \\ \epsilon_{i,k+1} &= \lambda_{2,i} \epsilon_{i,k} + \gamma_{2,i} (\vec{y}_{i,k} - h_{i,k}(\hat{\vec{x}}_{k|k-1}, \vec{0}))^2\end{aligned}\tag{16.76}$$

where

$$\text{sat}_{\sigma}(\tilde{\vec{y}}_k) = \begin{bmatrix} \vdots \\ \text{sat}_{\sqrt{\sigma_i}}(\tilde{y}_i) \\ \vdots \end{bmatrix}\tag{16.77}$$

## Second-Order Extended Kalman Filter

Recall that the EKF is based on the first-order Taylor series expansion. An alternative EKF approach is the **second-order extended Kalman filter (SO-EKF)** which truncates the multivariate Taylor series after the quadratic terms, i.e., for some equation  $g(\vec{x})$  about the mean of  $\vec{x}$ ,  $\vec{\mu}_x$ , as

$$g(\vec{x}) \approx g(\vec{\mu}_x) + G_{\vec{x}}(\vec{x} - \vec{\mu}_x) + \frac{1}{2} \sum_{i=1}^{n_g} \vec{e}_i (\vec{x} - \vec{\mu}_x)^T G_{\vec{x}\vec{x}}^{(i)} (\vec{x} - \vec{\mu}_x) \quad (16.78)$$

where  $\vec{e}_i$  is the unit vector for selecting the  $i^{\text{th}}$  element contribution for  $g(\vec{x})$ , i.e.,

$$\vec{e}_i = [0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0]^T \quad (16.79)$$

$G_x$  is the Jacobian of  $g(\vec{x})$  evaluated at  $\vec{x} = \vec{\mu}_x$  as

$$G_x = \left. \frac{\partial g}{\partial \vec{x}} \right|_{\vec{x}=\vec{\mu}_x} \quad (16.80)$$

and  $G_{xx}^{(i)}$  is the symmetric Hessian matrix of the  $g_i(\vec{x})$  equation for  $i = 1, \dots, n_g$  evaluated at  $\vec{x} = \vec{\mu}_x$  which has elements at row  $j$  and column  $k$  as

$$\left[ G_{xx}^{(i)} \right]_{j,k} = \left. \frac{\partial^2 g_i}{\partial x_j \partial x_k} \right|_{\vec{x}=\vec{\mu}_x} \quad (16.81)$$

Note that  $(\vec{x} - \vec{\mu}_x)^T G_{\vec{x}\vec{x}}^{(i)} (\vec{x} - \vec{\mu}_x)$  is a scalar quantity. Thus, the SO-EKF requires the Hessians must exist for the process and measurement equations and can be more computationally cumbersome to calculate. Notably, higher-order EKFs can also be constructed, but are not commonly used in practice as other nonlinear Kalman filters have been shown to be equivalent with less memory and computational cost.

Computing the expectation with respect to  $\vec{x}$ , one can show the mean of  $g(\vec{x})$  is approximated as

$$\mathbb{E}[g(\vec{x})] \approx g(\vec{\mu}_x) + \frac{1}{2} \sum_{i=1}^{n_g} \vec{e}_i \text{Tr} \left( G_{xx}^{(i)} P_x \right) \quad (16.82)$$

the cross-covariance of  $g(\vec{x})$  and  $\vec{x}$  is approximated as

$$\begin{aligned} \mathbb{E}[g(\vec{x})] &= \mathbb{E}[(\vec{x} - \vec{\mu}_x)(g(\vec{x}) - \mathbb{E}[g(\vec{x})])^T] \\ &\approx P_x G_{xx}^{(i)T} \end{aligned} \quad (16.83)$$

and the covariance of  $g(\vec{x})$  is approximated as

$$\begin{aligned} \text{Cov}(g(\vec{x})) &= \mathbb{E}[(g(\vec{x}) - \mathbb{E}[g(\vec{x})])(g(\vec{x}) - \mathbb{E}[g(\vec{x})])^T] \\ &\approx G_x P_x G_x^T + \frac{1}{2} \sum_{i=1}^{n_g} \sum_{j=1}^{n_g} \vec{e}_i \vec{e}_j^T \text{Tr} \left( G_{xx}^{(i)} P_x G_{xx}^{(j)} P_x \right) \end{aligned} \quad (16.84)$$

where

$$P_x = \mathbb{E} [(\vec{x} - \vec{\mu}_x)(\vec{x} - \vec{\mu}_x)^T] \quad (16.85)$$

Thus, for the SO-EKF at time step  $k$ , one requires the four Jacobians

$$\begin{aligned} F_{k-1} &= \left[ \frac{\partial f_{k-1}}{\partial \vec{x}} \right]_{\vec{x}_{k-1} = \hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1} = \vec{0}} \\ L_{k-1} &= \left[ \frac{\partial f_{k-1}}{\partial \vec{w}} \right]_{\vec{x}_{k-1} = \hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1} = \vec{0}} \\ H_k &= \left[ \frac{\partial h_k}{\partial \vec{x}} \right]_{\vec{x}_k = \hat{\vec{x}}_{k|k-1}, \vec{v}_k = \vec{0}} \\ M_k &= \left[ \frac{\partial h_k}{\partial \vec{v}} \right]_{\vec{x}_k = \hat{\vec{x}}_{k|k-1}, \vec{v}_k = \vec{0}} \end{aligned} \quad (16.86)$$

and the  $2n_x + 2n_y$  Hessians which have elements at row  $j$  and column  $k$  as

$$\begin{aligned} \left[ F_{xx,k-1}^{(i)} \right]_{j,k} &= \left. \frac{\partial^2 f_{k-1,i}}{\partial x_j \partial x_k} \right|_{\vec{x}_{k-1} = \hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1} = \vec{0}} \quad i = 1, \dots, n_x \\ \left[ L_{ww,k-1}^{(i)} \right]_{j,k} &= \left. \frac{\partial^2 f_{k-1,i}}{\partial w_j \partial w_k} \right|_{\vec{x}_{k-1} = \hat{\vec{x}}_{k-1|k-1}, \vec{w}_{k-1} = \vec{0}} \quad i = 1, \dots, n_x \\ \left[ H_{xx,k}^{(i)} \right]_{j,k} &= \left. \frac{\partial^2 h_{k,i}}{\partial x_j \partial x_k} \right|_{\vec{x}_k = \hat{\vec{x}}_{k|k-1}, \vec{v}_k = \vec{0}} \quad i = 1, \dots, n_y \\ \left[ M_{vv,k}^{(i)} \right]_{j,k} &= \left. \frac{\partial^2 h_{k,i}}{\partial v_j \partial v_k} \right|_{\vec{x}_k = \hat{\vec{x}}_{k|k-1}, \vec{v}_k = \vec{0}} \quad i = 1, \dots, n_y \end{aligned} \quad (16.87)$$

In summary, the discrete-time **SO-EKF initialization step** sets the state estimate and state covariance at  $k = 0$  as

$$\begin{aligned} \hat{\vec{x}}_{0|0} &= \hat{\vec{x}}_0 = \mathbb{E} [\vec{x}_0] \\ P_{0|0} &= P_0 = \mathbb{E} [(\vec{x}_0 - \hat{\vec{x}}_{0|0})(\vec{x}_0 - \hat{\vec{x}}_{0|0})^T] \end{aligned} \quad (16.88)$$

The discrete-time **SO-EKF prediction step** predicts the state estimate, the state covariance, and the output

using the input, process model, the Jacobians, and the Hessians as

$$\begin{aligned}
 \hat{\vec{x}}_{k|k-1} &= f_{k-1}(\hat{\vec{x}}_{k-1|k-1}, \vec{u}_{k-1}, \vec{0}) + \frac{1}{2} \sum_{i=1}^{n_x} \vec{e}_i \text{Tr} \left( F_{xx,k-1}^{(i)} P_{k-1|k-1} \right) \\
 &\quad + \frac{1}{2} \sum_{i=1}^{n_x} \vec{e}_i \text{Tr} \left( L_{ww,k-1}^{(i)} Q_{k-1} \right) \\
 P_{k|k-1} &= F_{k-1} P_{k-1|k-1} F_{k-1}^T + L_{k-1} Q_{k-1} L_{k-1}^T \\
 &\quad + \frac{1}{2} \sum_{i=1}^{n_x} \sum_{j=1}^{n_x} \vec{e}_i \vec{e}_j^T \text{Tr} \left( F_{xx,k-1}^{(i)} P_{k-1|k-1} F_{xx,k-1}^{(j)} P_{k-1|k-1} \right) \\
 &\quad + \frac{1}{2} \sum_{i=1}^{n_x} \sum_{j=1}^{n_x} \vec{e}_i \vec{e}_j^T \text{Tr} \left( L_{xx,k-1}^{(i)} Q_{k-1} L_{xx,k-1}^{(j)} Q_{k-1} \right)
 \end{aligned} \tag{16.89}$$

The discrete-time **SO-EKF correction step** updates the state estimate and the state covariance using the measurement model, the Jacobians, and the Hessians as

$$\begin{aligned}
 \tilde{y}_k &= \vec{y}_k - h_k(\hat{\vec{x}}_{k|k-1}, \vec{0}) - \frac{1}{2} \sum_{i=1}^{n_y} \vec{e}_i \text{Tr} \left( H_{ww,k}^{(i)} P_{k|k-1} \right) - \frac{1}{2} \sum_{i=1}^{n_y} \vec{e}_i \text{Tr} \left( M_{ww,k}^{(i)} Q_{k-1} \right) \\
 S_k &= H_k P_{k|k-1} H_k^T + M_k R_k M_k^T \\
 &\quad + \frac{1}{2} \sum_{i=1}^{n_y} \sum_{j=1}^{n_y} \vec{e}_i \vec{e}_j^T \text{Tr} \left( H_{xx,k}^{(i)} P_{k|k-1} H_{xx,k}^{(j)} P_{k|k-1} \right) \\
 &\quad + \frac{1}{2} \sum_{i=1}^{n_y} \sum_{j=1}^{n_y} \vec{e}_i \vec{e}_j^T \text{Tr} \left( M_{xx,k}^{(i)} R_k L_{xx,k}^{(j)} R_k \right) \\
 K_k &= P_{k|k-1} H_k^T S_k^{-1} \\
 \hat{\vec{x}}_{k|k} &= \hat{\vec{x}}_{k|k-1} + K_k \tilde{y}_k \\
 P_{k|k} &= P_{k|k-1} - K_k S_k K_k^T
 \end{aligned} \tag{16.90}$$

### Fixed-Interval Extended Kalman Smoother

Analogously to the FI-KF and RTSS, the **fixed-interval extended Kalman smoother (FI-EKS)**, also known as the **extended Rauch-Tung-Striebel smoother (ERTSS)**, assumes one has already approximated the state filtering prior distribution as

$$\vec{X}_{k+1} | \vec{Y}_{1:k} \sim \mathcal{N} \left( \hat{\vec{x}}_{k+1|k}, P_{k+1|k} \right) \tag{16.91}$$

and the state filtering posterior distribution as

$$\vec{X}_k | \vec{Y}_{1:k} \sim \mathcal{N} \left( \hat{\vec{x}}_{k|k}, P_{k|k} \right) \tag{16.92}$$

and will assume the state smoothing posterior distribution to be approximated as

$$\vec{X}_k | \vec{Y}_{1:N} \sim \mathcal{N} \left( \hat{\vec{x}}_{k+1|N}, P_{k+1|N} \right) \tag{16.93}$$

which for  $k+1 = N$ , one has the initial state smoothing posterior equivalent to the final state filtering posterior, i.e.,  $\vec{X}_N | \vec{Y}_{1:N} \sim \mathcal{N}(\hat{\vec{x}}_{N|N}, P_{N|N})$ .

Then, by similar logic to the previous discussions, the discrete-time **FI-EKS smoothing step** or the **ERTSS smoothing step** updates the state estimate and the state covariance using the posterior state estimate and covariance and the prior state estimate and covariance or the measurement noise covariance as

$$\begin{aligned}
\tilde{\vec{x}}_k &= \hat{\vec{x}}_{k+1|N} - \hat{\vec{x}}_{k+1|k} \\
&= \hat{\vec{x}}_{k+1|N} - f_k(\hat{\vec{x}}_{k|k}, \vec{u}_k, \vec{0}) \\
F_k &= \left[ \frac{\partial f_k}{\partial \vec{x}_k} \right]_{\vec{x}_k=\hat{\vec{x}}_{k|k}, \vec{w}_k=\vec{0}} \\
L_k &= \left[ \frac{\partial f_k}{\partial \vec{w}_k} \right]_{\vec{x}_k=\hat{\vec{x}}_{k|k}, \vec{w}_k=\vec{0}} \\
K_{S,k} &= P_{k|k} F_k^T P_{k+1|k}^{-1} \\
&= P_{k|k} F_k^T (F_k P_{k|k} F_k^T + L_k Q_k L_k)^{-1} \\
\hat{\vec{x}}_{k|N} &= \hat{\vec{x}}_{k|k} + K_{S,k} \tilde{\vec{x}}_k \\
P_{k|N} &= P_{k|k} + K_{S,k} (P_{k+1|N} - P_{k+1|k}) K_{S,k}^T \\
&= P_{k|k} + K_{S,k} (P_{k+1|N} - F_k P_{k|k} F_k^T - L_k Q_k L_k) K_{S,k}^T
\end{aligned} \tag{16.94}$$

## References

For more information, please refer to the following

- Fang, H., Haile, M. A., and Wang, Y. “Robust Extended Kalman Filtering for Systems With Measurement Outliers,” in *IEEE Transactions on Control Systems Technology*, 30(2), 2022, pp. 795-802
- Sarkka, S., “5.1 Taylor series expansions,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 64-69
- Sarkka, S., “5.2 Extended Kalman filter,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 69-75
- Sarkka, S., “8.2 Extended Rauch-Tung-Striebel smoother,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 69-75
- Simon, D., “13.1 The linearized Kalman filter,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 397-399
- Simon, D., “13.2 The extended Kalman filter,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 407-410
- Simon, D., “13.3 Higher-order approaches,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 410-420

## 16.3 Statistical Linearization and State Estimation

### Statistical Linearization

The previous derivations linearized a nonlinear stochastic functions,  $g(\vec{X})$ , through a Jacobian linearization about a *single* point, e.g., the mean of  $\vec{X}$  as

$$\vec{\mu}_x = \mathbb{E} [\vec{X}] \quad (16.95)$$

which provides

$$g(\vec{X}) \approx g(\vec{\mu}_x) + \frac{\partial g(\vec{X})}{\partial \vec{X}} \Big|_{\vec{X}=\vec{\mu}_x} (\vec{X} - \vec{\mu}_x) \quad (16.96)$$

As an alternative, **statistical linearization** optimally fits the linear approximation

$$g(\vec{X}) \approx \vec{b}^{opt} + A^{opt}(\vec{X} - \vec{\mu}_x) \quad (16.97)$$

that minimizes the mean-square error (MSE), i.e.,

$$A^{opt}, \vec{b}^{opt} = \underset{A, \vec{b}}{\operatorname{argmin}} \mathbb{E} \left[ \left( g(\vec{X}) - (\vec{b} + A(\vec{X} - \vec{\mu}_x)) \right)^T \left( g(\vec{X}) - (\vec{b} + A(\vec{X} - \vec{\mu}_x)) \right) \right] \quad (16.98)$$

which takes into account the uncertainty or “probabilistic spread” of the random vector,  $\vec{X}$ , about its mean through the expectation operator.

To solve this optimization, one can expand the MSE cost function as

$$\text{MSE} = \mathbb{E} [g(\vec{X})^T g(\vec{X}) - 2g(\vec{X})^T \vec{b} - 2g(\vec{X})^T A(\vec{X} - \vec{\mu}_x) + \vec{b}^T \vec{b} - 2\vec{b}^T A(\vec{X} - \vec{\mu}_x) + (\vec{X} - \vec{\mu}_x)^T A^T A(\vec{X} - \vec{\mu}_x)] \quad (16.99)$$

and performing the expectations on the final two terms, one has

$$\text{MSE} = \mathbb{E} [g(\vec{X})^T g(\vec{X}) - 2g(\vec{X})^T \vec{b} - 2g(\vec{X})^T A(\vec{X} - \vec{\mu}_x) + \vec{b}^T \vec{b}] + 0 + \text{Tr}(AP_xA^T) \quad (16.100)$$

where  $P_x$  is the covariance of  $\vec{X}$

$$P_x = \mathbb{E} [(\vec{X} - \vec{\mu}_x)(\vec{X} - \vec{\mu}_x)^T] \quad (16.101)$$

Then, computing the partial derivatives, one has

$$\frac{\partial \text{MSE}}{\partial \vec{b}} = -2\mathbb{E} [g(\vec{x})] + 2\vec{b} \quad (16.102)$$

and

$$\frac{\partial \text{MSE}}{\partial A} = \mathbb{E} [g(\vec{x})(\vec{X} - \vec{\mu}_x)^T] + 2AP_x \quad (16.103)$$

Finally, solving for the optimal values which provide zero derivatives, one has

$$\vec{b}^{opt} = \mathbb{E} [g(\vec{X})] \quad (16.104)$$

$$A^{opt} = \mathbb{E} [g(\vec{X})(\vec{X} - \vec{\mu}_x)^T] P_x^{-1} \quad (16.105)$$

Thus, the statistical linearization approximation can be written as

$$g(\vec{X}) \approx \mathbb{E}[g(\vec{X})] + \mathbb{E}[g(\vec{X})(\vec{X} - \vec{\mu}_x)^T] P_x^{-1} (\vec{X} - \vec{\mu}_x) \quad (16.106)$$

which has an exact mean, i.e.,

$$g(\vec{X}) \approx \mathbb{E}[g(\vec{X})] \quad (16.107)$$

and an approximate covariance, i.e.,

$$\text{Cov}(g(\vec{X})) = \mathbb{E}[g(\vec{X})(\vec{X} - \vec{\mu}_x)^T] P_x^{-1} \mathbb{E}[g(\vec{X})(\vec{X} - \vec{\mu}_x)^T]^T \quad (16.108)$$

Note that if  $g()$  is differentiable, then for multivariate Gaussian random vectors, one has

$$\mathbb{E}[g(\vec{X})(\vec{X} - \vec{\mu}_x)^T] = \mathbb{E}\left[\frac{\partial g(\vec{X})}{\partial \vec{X}}\right] P_x \quad (16.109)$$

which provides the covariance in terms of the expectation of the Jacobian as

$$\text{Cov}(g(\vec{X})) = \mathbb{E}\left[\frac{\partial g(\vec{X})}{\partial \vec{X}}\right] P_x \mathbb{E}\left[\frac{\partial g(\vec{X})}{\partial \vec{X}}\right]^T \quad (16.110)$$

Furthermore, if one can compute the mean expectation, then one can compute the expectation of the Jacobian as the partial derivative, i.e.,

$$\frac{\partial \mathbb{E}[g(\vec{X})]}{\partial \vec{\mu}_x} - \mathbb{E}\left[\frac{\partial g(\vec{X})}{\partial \vec{X}}\right] \quad (16.111)$$

### Statistically Linearized (Kalman) Filter

Note that the statistical linearization is useful in Kalman filtering when closed-form solutions exist for these expectations. It is important to note that because statistical linearization takes into account the mean *and* covariance of the random vector and the expectation of the nonlinear functions, the expected linearization error based on statistical linearization tends to be smaller than the Jacobian linearization. However, the statistically linearized filter only applies to multivariate Gaussian noise models. With this previous derivations in mind, the **statistically linearized (Kalman) filter (SLKF)** can be constructed similarly to the EKF. The **SLKF initialization step** sets the state estimate and state covariance at  $k = 0$  as

$$\begin{aligned} \hat{x}_{0|0} &= \hat{x}_0 = \mathbb{E}[\vec{X}_0] \\ P_{0|0} &= P_0 = \mathbb{E}[(\vec{X}_0 - \hat{x}_{0|0})(\vec{X}_0 - \hat{x}_{0|0})^T] \end{aligned} \quad (16.112)$$

The discrete-time **SLKF prediction step** predicts the state estimate and the state covariance using the input and process model and explicit expectations as

$$\begin{aligned} \hat{x}_{k|k-1} &= \mathbb{E}[f_{k-1}(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})] \\ P_{k|k-1} &= \mathbb{E}[f_{k-1}(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})(\vec{X}_{k-1} - \hat{x}_{k-1|k-1})^T] P_{k-1|k-1}^{-1} \mathbb{E}[f_{k-1}(\vec{X}, \vec{u}_{k-1}, \vec{W})(\vec{X}_{k-1} - \hat{x}_{k-1|k-1})^T]^T \\ &\quad + \mathbb{E}[f_{k-1}(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1}) \vec{W}_{k-1}^T] Q_{k-1}^{-1} \mathbb{E}[f_{k-1}(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1}) \vec{W}_{k-1}^T]^T \end{aligned} \quad (16.113)$$

where if  $f()$  is differentiable, one has

$$\begin{aligned}\hat{x}_{k|k-1} &= \mathbb{E}[f_{k-1}(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})] \\ F_{k-1} &= \mathbb{E}\left[\frac{\partial f_{k-1}(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})}{\partial \vec{X}_{k-1}}\right] \\ L_{k-1} &= \mathbb{E}\left[\frac{\partial f_{k-1}(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})}{\partial \vec{W}_{k-1}}\right] \\ P_{k|k-1} &= F_{k-1}P_{k-1|k-1}F_{k-1}^T + L_{k-1}Q_{k-1}L_{k-1}^T\end{aligned}\tag{16.114}$$

where the expectations in the prediction step are taken with respect to  $\vec{X}_{k-1} \sim \mathcal{N}(\hat{x}_{k-1|k-1}, P_{k-1|k-1})$  and  $\vec{W}_{k-1} \sim \mathcal{N}(0, Q_{k-1})$

The discrete-time **SLKF correction step** updates the state estimate and the state covariance using the measurement model and explicit expectations as

$$\begin{aligned}\hat{y}_k &= \mathbb{E}[h(\vec{X}_k, \vec{V}_k)] \\ \tilde{y}_k &= \vec{y}_k - \hat{y}_k \\ S_k &= \mathbb{E}[h_k(\vec{X}_k, \vec{V}_k)(\vec{X}_k - \hat{x}_{k|k-1})^T]P_{k|k-1}^{-1}\mathbb{E}[h_k(\vec{X}_k, \vec{V}_k)(\vec{X}_k - \hat{x}_{k|k-1})^T]^T \\ &\quad + \mathbb{E}[h_k(\vec{X}_k, \vec{V}_k))\vec{V}_k^T]R_k^{-1}\mathbb{E}[h_k(\vec{X}_k, \vec{V}_k))\vec{V}_k^T]^T \\ K_k &= \mathbb{E}[h_k(\vec{X}_k, \vec{V}_k))(\vec{X}_k - \hat{x}_{k|k-1})^T]^TS_k^{-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k\tilde{y}_k \\ P_{k|k} &= P_{k|k-1} - K_kS_kK_k^T\end{aligned}\tag{16.115}$$

where if  $h_k()$  is differentiable, one has

$$\begin{aligned}\hat{y}_k &= \mathbb{E}[h(\vec{X}_k, \vec{V}_k)] \\ \tilde{y}_k &= \vec{y}_k - \hat{y}_k \\ H_k &= \mathbb{E}\left[\frac{\partial h_k(\vec{X}_k, \vec{V}_k)}{\partial \vec{X}_k}\right] \\ M_k &= \mathbb{E}\left[\frac{\partial h_k(\vec{X}_k, \vec{V}_k)}{\partial \vec{V}_k}\right] \\ S_k &= H_kP_{k|k-1}H_k^T + M_kR_kM_k^T \\ K_k &= P_{k|k-1}H_k^TS_k^{-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k\tilde{y}_k \\ P_{k|k} &= P_{k|k-1} - K_kS_kK_k^T\end{aligned}\tag{16.116}$$

where the expectations in the correction step are taken with respect to  $\vec{X}_k \sim \mathcal{N}(\hat{x}_{k|k-1}, P_{k|k-1})$  and  $\vec{V}_k \sim \mathcal{N}(0, R_k)$

### Statistically Linearized (Kalman) Smoother

Analogously to the FI-EKF/ERTSS, the **fixed-interval statistically linearized (Kalman) smoother (FI-SLKS)**, also known as the **statistically linearized Rauch-Tung-Striebel smoother (SLRTSS)**, assumes one has already approximated the state filtering prior distribution as

$$\vec{X}_{k+1} | \vec{Y}_{1:k} \sim \mathcal{N}(\hat{\vec{x}}_{k+1|k}, P_{k+1|k}) \quad (16.117)$$

and the state filtering posterior distribution as

$$\vec{X}_k | \vec{Y}_{1:k} \sim \mathcal{N}(\hat{\vec{x}}_{k|k}, P_{k|k}) \quad (16.118)$$

and will assume the state smoothing posterior distribution to be approximated as

$$\vec{X}_k | \vec{Y}_{1:N} \sim \mathcal{N}(\hat{\vec{x}}_{k+1|N}, P_{k+1|N}) \quad (16.119)$$

which for  $k+1 = N$ , one has the initial state smoothing posterior equivalent to the final state filtering posterior, i.e.,  $\vec{X}_N | \vec{Y}_{1:N} \sim \mathcal{N}(\hat{\vec{x}}_{N|N}, P_{N|N})$ .

Then, by similar logic to the previous discussions, the discrete-time **FI-SLKS smoothing step** or the **SLRTSS smoothing step** updates the state estimate and the state covariance using the posterior state estimate and covariance and the prior state estimate and covariance or the measurement noise covariance as

$$\begin{aligned} \tilde{\vec{x}}_k &= \hat{\vec{x}}_{k+1|N} - \hat{\vec{x}}_{k+1|k} \\ &= \hat{\vec{x}}_{k+1|N} - \mathbb{E}[f_k(\vec{X}_k, \vec{u}_k, \vec{W}_{k-1})] \\ P_{k+1|k} &= \mathbb{E}\left[\frac{\partial f_k(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})}{\partial \vec{X}_{k-1}}\right] P_{k|k}^{-1} \mathbb{E}\left[\frac{\partial f_k(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})}{\partial \vec{X}_{k-1}}\right]^T \\ &\quad + \mathbb{E}\left[\frac{\partial f_k(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})}{\partial \vec{W}_{k-1}}\right] Q_k^{-1} \mathbb{E}\left[\frac{\partial f_k(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})}{\partial \vec{W}_{k-1}}\right]^T \quad (16.120) \\ K_{S,k} &= \mathbb{E}[f_k(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})(\vec{X}_k - \hat{\vec{x}}_{k|k})^T]^T P_{k+1|k}^{-1} \\ \hat{\vec{x}}_{k|N} &= \hat{\vec{x}}_{k|k} + K_{S,k} \tilde{\vec{x}}_k \\ P_{k|N} &= P_{k|k} + K_{S,k}(P_{k+1|N} - P_{k+1|k})K_{S,k}^T \end{aligned}$$

where if  $f_k()$  is differentiable, one has

$$\begin{aligned}
 \tilde{\vec{x}}_k &= \hat{\vec{x}}_{k+1|N} - \hat{\vec{x}}_{k+1|k} \\
 &= \hat{\vec{x}}_{k+1|N} - \mathbb{E} [f_k(\vec{X}_k, \vec{u}_k, \vec{W}_{k-1})] \\
 F_k &= \mathbb{E} \left[ \frac{\partial f_k(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})}{\partial \vec{X}_{k-1}} \right] \\
 L_k &= \mathbb{E} \left[ \frac{\partial f_k(\vec{X}_{k-1}, \vec{u}_{k-1}, \vec{W}_{k-1})}{\partial \vec{W}_{k-1}} \right] \\
 K_{S,k} &= P_{k|k} F_k^T P_{k+1|k}^{-1} \\
 &= P_{k|k} F_k^T (F_k P_{k|k} F_k^T + L_k Q_k L_k)^{-1} \\
 \hat{\vec{x}}_{k|N} &= \hat{\vec{x}}_{k|k} + K_{S,k} \tilde{\vec{x}}_k \\
 P_{k|N} &= P_{k|k} + K_{S,k} (P_{k+1|N} - P_{k+1|k}) K_{S,k}^T \\
 &= P_{k|k} + K_{S,k} (P_{k+1|N} - F_k P_{k|k} F_k^T - L_k Q_k L_k) K_{S,k}^T
 \end{aligned} \tag{16.121}$$

where the expectations in the smoothing step are taken with respect to  $\vec{X}_k \sim \mathcal{N}(\hat{\vec{x}}_{k|k}, P_{k|k})$  and  $\vec{W}_k \sim \mathcal{N}(0, Q_k)$ .

Furthermore, just as a Jacobian linearization is a first-order approximation of the Taylor series expansion, a statistical linearization is a first-order approximation of the Fourier-Hermite series expansion. Thus, a higher-order statistically linearized Kalman filter or smoother is known as a **Fourier-Hermite Kalman filter (FHKF)** or a **fixed-interval Fourier-Hermite Kalman smoother (FI-FHKS)**, also known as the **Fourier-Hermite Rauch-Tung-Striebel smoother (FHRTSS)**. However, the development of the FHKF and the FI-FHKS/FHRTSS is beyond the scope of this textbook. In summary, they involve expectations of higher-order derivatives with special computational properties due to the Fourier-Hermite series.

### Sigma-Point Approximation to Statistical Linearization

The statistical linearization procedure requires explicit computation of the expectation, one can approximate the statistical linearization through a linear regression problem of  $n_\sigma$  points drawn from the prior probability distribution of the random vector, known as the **sigma-points**, and the *true* nonlinear functional evaluation of the sigma-points, i.e., through **statistical linear regression**. However, the exact determination of the *sigma-points* and their corresponding regression weights must be chosen by some criterion of which there are several different methods, e.g., the central divided difference, the unscented transform, the third-order spherical cubature, and Gauss-Hermite quadrature. These are related in the next section on sigma-point Kalman filters and fixed-interval smoothers utilizing these specific methods.

Consider a nonlinear function,  $\vec{y} = g(\vec{x})$ , which can be evaluated at the  $j = 1, \dots, n_\sigma$  sigma-points  $(\vec{x}_j, \vec{y}_j)$  where  $\vec{y}_j = g(\vec{x}_j)$ . Then, one can define

$$\bar{\vec{x}} = \sum_{j=1}^{n_\sigma} w_j \vec{x}_j \tag{16.122}$$

$$\hat{P}_x = \sum_{j=1}^{n_\sigma} w_j (\vec{x}_j - \bar{\vec{x}})(\vec{x}_j - \bar{\vec{x}})^T \tag{16.123}$$

$$\bar{\mathcal{Y}} = \sum_{j=1}^{n_\sigma} w_j \vec{\mathcal{Y}}_j \quad (16.124)$$

$$\hat{P}_y = \sum_{j=1}^{n_\sigma} w_j (\vec{\mathcal{Y}}_j - \bar{\mathcal{Y}})(\vec{\mathcal{Y}}_j - \bar{\mathcal{Y}})^T \quad (16.125)$$

$$\hat{P}_{xy} = \sum_{j=1}^{n_\sigma} w_j (\vec{X}_j - \bar{X})(\vec{\mathcal{Y}}_j - \bar{\mathcal{Y}})^T \quad (16.126)$$

where  $w_j$  for  $j = 1, \dots, n_\sigma$  are a set of scalar **regression weights** such that

$$\sum_{j=1}^{n_\sigma} w_j = 1 \quad (16.127)$$

Thus, for a sigma-point approximation to statistical linearization, i.e.,

$$g(\vec{x}) \approx A\vec{X}_j + \vec{b} \quad (16.128)$$

one minimizes the mean-square errors for a *point-wise* linearization. Here, the point-wise errors are

$$g(\vec{X}_j) - (A\vec{X}_j + \vec{b}) = (\vec{\mathcal{Y}}_j) - (A\vec{X}_j + \vec{b}) = \quad (16.129)$$

This results in the following optimization problem

$$A^{opt}, \vec{b}^{opt} = \underset{A, \vec{b}}{\operatorname{argmin}} \mathbb{E} \left[ \sum_{j=1}^{n_\sigma} w_j \left( g(\vec{X}_j) - (A\vec{X}_j + \vec{b}) \right) \left( g(\vec{X}_j) - (A\vec{X}_j + \vec{b}) \right)^T \right] \quad (16.130)$$

or

$$A^{opt}, \vec{b}^{opt} = \underset{A, \vec{b}}{\operatorname{argmin}} \mathbb{E} \left[ \sum_{j=1}^{n_\sigma} w_j \left( \vec{\mathcal{Y}}_j - (A\vec{X}_j + \vec{b}) \right) \left( \vec{\mathcal{Y}}_j - (A\vec{X}_j + \vec{b}) \right)^T \right] \quad (16.131)$$

which can be solved by considering the partial derivatives of the cost function.

First, rewrite the MSE cost function as

$$\text{MSE} = \mathbb{E} \left[ \sum_{j=1}^{n_\sigma} w_j \left( (\vec{\mathcal{Y}}_j - A\vec{X}_j) - \vec{b} \right) \left( (\vec{\mathcal{Y}}_j - A\vec{X}_j) - \vec{b} \right)^T \right] \quad (16.132)$$

$$\text{MSE} = \mathbb{E} \left[ \sum_{j=1}^{n_\sigma} w_j \left( (\vec{\mathcal{Y}}_j - A\vec{X}_j)(\vec{\mathcal{Y}}_j - A\vec{X}_j)^T - 2(\vec{\mathcal{Y}}_j - A\vec{X}_j)\vec{b}^T + \vec{b}\vec{b}^T \right) \right] \quad (16.133)$$

and setting the partial derivative of the cost function with respect to  $\vec{b}$  to zero, one has

$$\mathbb{E} \left[ \sum_{j=1}^{n_\sigma} 2w_j \left( \vec{\mathcal{Y}}_j - A\vec{X}_j - \vec{b} \right) \right] = 0 \quad (16.134)$$

or, by dividing by 2 and the sigma-point statistic definitions, one has

$$\mathbb{E} \left[ \bar{\mathcal{Y}} - A\bar{\mathcal{X}} - \sum_{j=1}^{n_\sigma} w_j \vec{b} \right] = 0 \quad (16.135)$$

which, by  $\sum_{j=1}^{n_\sigma} w_j = 1$ , provides the optimal solution

$$\vec{b}^{opt} = \bar{\mathcal{Y}} - A\bar{\mathcal{X}} \quad (16.136)$$

Next, substituting this result for  $\vec{b}$  into the cost function, one has

$$A^{opt} = \operatorname{argmin} \mathbb{E} \left[ \sum_{j=1}^{n_\sigma} w_j \left( \bar{\mathcal{Y}}_j - (A\bar{\mathcal{X}}_j + \bar{\mathcal{Y}} - A\bar{\mathcal{X}}) \right) \left( \bar{\mathcal{Y}}_j - (A\bar{\mathcal{X}}_j + \bar{\mathcal{Y}} - A\bar{\mathcal{X}}) \right)^T \right] \quad (16.137)$$

$$A^{opt} = \operatorname{argmin} \mathbb{E} \left[ \sum_{j=1}^{n_\sigma} w_j \left( (\bar{\mathcal{Y}}_j - \bar{\mathcal{Y}}) - A(\bar{\mathcal{X}}_j - \bar{\mathcal{X}}) \right) \left( (\bar{\mathcal{Y}}_j - \bar{\mathcal{Y}}) - A(\bar{\mathcal{X}}_j - \bar{\mathcal{X}}) \right)^T \right] \quad (16.138)$$

or

$$A^{opt} = \operatorname{argmin} \mathbb{E} \left[ \sum_{j=1}^{n_\sigma} w_j \left( (\bar{\mathcal{Y}}_j - \bar{\mathcal{Y}})(\bar{\mathcal{Y}}_j - \bar{\mathcal{Y}})^T - 2A(\bar{\mathcal{X}}_j - \bar{\mathcal{X}})(\bar{\mathcal{Y}}_j - \bar{\mathcal{Y}}) + A(\bar{\mathcal{X}}_j - \bar{\mathcal{X}})(\bar{\mathcal{X}}_j - \bar{\mathcal{X}})^T A^T \right) \right] \quad (16.139)$$

and setting the partial derivative of the cost function with respect to  $A$  to zero, one has

$$\mathbb{E} \left[ \sum_{j=1}^{n_\sigma} 2w_j \left( A(\bar{\mathcal{X}}_j - \bar{\mathcal{X}})(\bar{\mathcal{X}}_j - \bar{\mathcal{X}})^T - (\bar{\mathcal{X}}_j - \bar{\mathcal{X}})(\bar{\mathcal{Y}}_j - \bar{\mathcal{Y}}) \right) \right] = 0 \quad (16.140)$$

or, by the sigma-point statistic definitions, one has

$$\mathbb{E} [A\hat{P}_x - \hat{P}_{xy}] = 0 \quad (16.141)$$

which provides the optimal solution

$$A^{opt} = \hat{P}_{xy}^T \hat{P}_x^{-1} \quad (16.142)$$

Substituting for  $\vec{b}$  and  $A$ , the sigma-point approximation to statistical linearization for a nonlinear function  $g()$  is

$$g(\vec{x}) \approx \hat{P}_{xy}^T \hat{P}_x^{-1} (\bar{\mathcal{X}}_j) + (\bar{\mathcal{Y}} - \hat{P}_{xy}^T \hat{P}_x^{-1} \bar{\mathcal{X}}) \quad (16.143)$$

$$g(\vec{x}) \approx \bar{\mathcal{Y}} + \hat{P}_{xy}^T \hat{P}_x^{-1} (\bar{\mathcal{X}}_j) - \bar{\mathcal{X}} \quad (16.144)$$

or, by substitution, one has

$$g(\vec{x}) \approx \sum_{j=1}^{n_\sigma} w_j g(\bar{\mathcal{X}}_j) + \left[ \left( \sum_{j=1}^{n_\sigma} w_j (\bar{\mathcal{X}}_j - \bar{\mathcal{X}}) \left( g(\bar{\mathcal{X}}_j) - \left( \sum_{j=1}^{n_\sigma} w_j g(\bar{\mathcal{X}}_j) \right) \right)^T \right)^T \left( \sum_{j=1}^{n_\sigma} w_j (\bar{\mathcal{X}}_j - \bar{\mathcal{X}})(\bar{\mathcal{X}}_j - \bar{\mathcal{X}})^T \right)^{-1} \right] \left( \bar{\mathcal{X}}_j \right) - \sum_{j=1}^{n_\sigma} w_j \bar{\mathcal{X}}_j \quad (16.145)$$

which demonstrates the numerical integration perspective for the sigma-point method for the expectation integrals of statistical linearization.

## References

For more information, please refer to the following

- Sarkka, S., “5.3 Statistical linearization,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 75-77
- Sarkka, S., “5.4 Statistically linearized filter,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 77-81
- Sarkka, S., “9.2 Statistically linearized Rauch-Tung-Striebel smoother,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 146-148

## 16.4 Sigma-Point Kalman Filtering and Smoothing

Based on the different types of sigma-point methods, the family of **sigma-point Kalman filters (SPKFs)** includes the Unscented Kalman Filter (UKF), the Central Difference Kalman Filter (CDKF), the Cubature Kalman filter (CKF), and the Gauss-Hermite Kalman filter (GHKF). Similarly, the family of **fixed-interval sigma-point Kalman smoothers (FI-SPKSs)** includes the unscented Kalman smoother (FI-UKS), the central difference Kalman smoother (FI-CDKS), the cubature Kalman smoother (FI-CKS), and the Gauss-Hermite Kalman smoother (FI-GHKS), also known as the family of **sigma-point Rauch-Tung-Striebel smoothers (SPRTSSs)** includes the unscented Rauch-Tung-Striebel smoother (URTSS), the central difference Rauch-Tung-Striebel smoother (CDRTSS), the cubature Rauch-Tung-Striebel smoother (CRTSS), and the Gauss-Hermite Rauch-Tung-Striebel smoother (GHRTSS). This section first presents the general form of an SPKF and an FI-SPKS or SPRTSS, and then defines the different sigma-point methods explicitly for the SPKF and FI-SPKS/SPRTSS. It should be noted that square-root forms of the SPKFs predict and correct the square root of the state covariance directly in Cholesky factored form, using the sigma-point methods and the following three linear algebra techniques: QR decomposition, Cholesky factor updating, and efficient pivot-based least-squares. While the square-root forms of the SPKFs do not display significant accuracy improvement, they have reduced computational complexity for certain nonlinear state-space models as well as increased numerical stability. Lastly, it should be noted that for *very* large state dimensions, the SPKFs can be computationally faster than the EKF since no Jacobian matrices must be calculated and no matrix inversion of the state dimension is necessary.

### Sigma-Point Kalman Filters

The discrete-time **SPKF initialization step** sets the initial state estimate as

$$\hat{\vec{x}}_{0|0} = \hat{\vec{x}}_0 = \mathbb{E} [\vec{x}_0] \quad (16.146)$$

the initial covariance as

$$P_{0|0} = P_0 = \mathbb{E} [(\vec{x}_0 - \hat{\vec{x}}_0)(\vec{x}_0 - \hat{\vec{x}}_0)^T] \quad (16.147)$$

The discrete-time **SPKF prediction step** first computes  $j = 1, \dots, n_{a,p}$  new **prediction augmented sigma-points** as

$$\vec{x}_{j,k-1|k-1} = \hat{\vec{x}}_{k-1|k-1}^a + \sqrt{P_{k-1|k-1}^a} \vec{\xi}_j \quad (16.148)$$

where  $\vec{\xi}_j$  depends on the sigma-point method chosen, the **SPKF process augmented state estimate** of dimension  $n_\xi = n_x + n_w$  is defined as

$$\hat{\vec{x}}_{k-1|k-1}^a = \begin{bmatrix} \hat{\vec{x}}_{k-1|k-1} \\ \vec{0} \end{bmatrix} \quad (16.149)$$

the **SPKF augmented state covariance** of dimension  $n_\xi \times n_\xi$  is defined as

$$P_{k-1|k-1}^a = \begin{bmatrix} P_{k-1|k-1} & 0 \\ 0 & Q_{k-1} \end{bmatrix} \quad (16.150)$$

Then, one uses these augmented state sigma-points to compute the prior state sigma-points, the prior state estimate, and the prior state covariance as

$$\begin{aligned} \hat{\vec{X}}_{j,k|k-1} &= f_{k-1}([\vec{X}_{j,k-1}]_{1:n_x}, \vec{u}_{k-1}, [\vec{X}_{j,k-1}]_{n_x+1:n_\xi}) \\ \hat{\vec{x}}_{k|k-1} &= \sum_{j=1}^{n_{a,p}} w_j^{(m)} \hat{\vec{x}}_{j,k|k-1} \\ P_{k|k-1} &= \sum_{j=1}^{n_{a,p}} w_j^{(c)} (\hat{\vec{x}}_{j,k|k-1} - \hat{\vec{x}}_{k|k-1})(\hat{\vec{x}}_{j,k|k-1} - \hat{\vec{x}}_{k|k-1})^T \end{aligned} \quad (16.151)$$

where  $[\vec{\bullet}]_{1:n}$  selects the sub-vector of elements 1 through  $n$  of  $\vec{\bullet}$  and  $w_j^{(m)}$  is the  $j^{\text{th}}$  mean weight and  $w_j^{(c)}$  is the  $j^{\text{th}}$  covariance weight corresponding to the  $j^{\text{th}}$  sigma-point.

The discrete-time **SPKF correction step** recomputes  $n_{a,c}$  new **correction augmented sigma-points** as

$$\vec{X}_{j,k|k-1} = \hat{\vec{x}}_{k|k-1}^a + \sqrt{P_{k|k-1}^a} \vec{\xi}_j \quad (16.152)$$

where  $\vec{\xi}_j$  depends on the sigma-point method chosen, the **SPKF process augmented state estimate** with dimension  $n_\xi = n_x + n_v$  is defined as

$$\hat{\vec{x}}_{k|k-1}^a = \begin{bmatrix} \hat{\vec{x}}_{k|k-1} \\ \vec{0} \end{bmatrix} \quad (16.153)$$

and the **SPKF augmented state covariance** with dimension  $n_\xi \times n_\xi$  is defined as

$$P_{k|k-1}^a = \begin{bmatrix} P_{k|k-1} & 0 \\ 0 & R_k \end{bmatrix} \quad (16.154)$$

Then, one uses these augmented state sigma-points to compute the measurement sigma-points, the predicted measurement, the innovation, the innovation covariance, the state-measurement cross-covariance,

the Kalman gain, the posterior state estimate, and the posterior state covariance as

$$\begin{aligned}
 \hat{\vec{y}}_{j,k} &= h_k \left( \left[ \vec{x}_{j,k|k-1}^a \right]_{1:n_x}, \left[ \vec{x}_{j,k|k-1}^a \right]_{n_x+1:n_\xi} \right) \\
 \hat{\vec{y}}_k &= \sum_{j=1}^{n_{a,c}} w_j^{(m)} \hat{\vec{y}}_{j,k} \\
 \tilde{y}_k &= \vec{y}_k - \hat{\vec{y}}_k \\
 S_k &= \sum_{j=0}^{n_{a,c}} w_j^{(c)} \left( \hat{\vec{y}}_{j,k} - \hat{\vec{y}}_k \right) \left( \hat{\vec{y}}_{j,k} - \hat{\vec{y}}_k \right)^T \\
 P_{xy} &= \sum_{j=0}^{n_{a,c}} w_j^{(c)} \left( \left[ \vec{x}_{j,k|k-1}^a \right]_{1:n_x} - \hat{x}_{k|k-1} \right) \left( \hat{\vec{y}}_{j,k} - \hat{\vec{y}}_k \right)^T \\
 K_k &= P_{xy} S_k^{-1} \\
 \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k \tilde{y}_k \\
 P_{k|k} &= P_{k|k-1} - K_k S_k K_k^T
 \end{aligned} \tag{16.155}$$

### Sigma-Point Kalman Smoother

The discrete-time **FI-SPKS smoothing step**, also known as the **SPRTSS smoothing step**, first recomputes  $n_{a,s}$  new **correction augmented sigma-points** as

$$\vec{x}_{j,k|k} = \hat{\vec{x}}_{k|k} + \sqrt{P_{k|k}} \vec{\xi}_j \tag{16.156}$$

where  $\vec{\xi}_j$  depends on the sigma-point method chosen, the **FI-SPKS/SPRTSS process augmented state estimate** with dimension  $n_\xi = n_x + n_w$  is defined as

$$\hat{\vec{x}}_{k|k}^a = \begin{bmatrix} \hat{\vec{x}}_{k|k} \\ \vec{0} \end{bmatrix} \tag{16.157}$$

and the **FI-SPKS/SPRTSS augmented state covariance** with dimension  $n_\xi \times n_\xi$  is defined as

$$P_{k|k}^a = \begin{bmatrix} P_{k|k} & 0 \\ 0 & Q_k \end{bmatrix} \tag{16.158}$$

Then, one uses the augmented state sigma-points through the exact nonlinear measurement equation to

obtain the measurement sigma-points as

$$\begin{aligned}
\hat{\vec{X}}_{j,k+1|k} &= f_k([\vec{X}_{j,k}]_{1:n_x}, \vec{u}_k, [\vec{X}_{j,k}]_{n_x+1:n_\xi}) \\
\hat{\vec{x}}_{k+1|k} &= \sum_{j=1}^{n_{a,s}} w_j^{(m)} \hat{\vec{X}}_{j,k+1|k} \\
\tilde{\vec{x}}_{k+1} &= \hat{\vec{x}}_{k+1|N} - \hat{\vec{x}}_{k+1|k} \\
P_{k+1|k} &= \sum_{j=0}^{n_{a,s}} w_j^{(c)} \left( \hat{\vec{X}}_{j,k+1|k} - \hat{\vec{x}}_{k+1|k} \right) \left( \hat{\vec{X}}_{j,k+1|k} - \hat{\vec{x}}_{k+1|k} \right)^T \\
P_{k,k+1} &= \sum_{j=0}^{n_{a,s}} w_j^{(c)} \left( [\vec{X}_{j,k}]_{1:n_x} - \hat{\vec{x}}_{k|k} \right) \left( \hat{\vec{X}}_{j,k+1|k} - \hat{\vec{x}}_{k+1|k} \right)^T \\
K_{S,k} &= P_{k,k+1} P_{k+1|k}^{-1} \\
\hat{\vec{x}}_{k|N} &= \hat{\vec{x}}_{k|k} + K_{S,k} \tilde{\vec{x}}_{k+1} \\
P_{k|N} &= P_{k|k} + K_{S,k} (P_{k+1|N} - P_{k+1|k}) K_{S,k}^T
\end{aligned} \tag{16.159}$$

## Unscented Transform

One method for determining the sigma-points and corresponding weights is such that they match the moments of the prior random state vector,  $\vec{X}$ . This is achieved by choosing the sigma-points according to some constraint equation of the form  $\eta(\{\mathcal{X}, w\}, n_\sigma, f_X(\vec{x})) = 0$  where  $\{\mathcal{X}, w\}$  is the set of all sigma-points  $X_j$  and corresponding weights  $w_j$  for  $j = 1, \dots, n_\sigma$ . It is possible to satisfy such a constraint and still have some degree of freedom in the choice of the sigma-point locations, thus, one typically also uses a cost function, e.g. some  $c(\{\mathcal{X}, w\}, n_\sigma, f_X(\vec{x}))$ , to incorporate statistical features of  $\vec{X}$  which are desirable, but do not necessarily have to be met. Taken together, one can construct this selection method as the following optimization problem.

$$\begin{aligned}
\{\vec{X}, w\} &= \operatorname{argmin} c(\{\vec{X}, w\}, n_\sigma, f_{\vec{X}}(\vec{x})) \\
\text{Subject to } &\eta(\{\vec{X}, w\}, n_\sigma, f_{\vec{X}}(\vec{x})) = 0
\end{aligned} \tag{16.160}$$

Using this general optimization, the **unscented transform (UT)** specifies that the sigma-points should be constrained to at least capture the first and second-order moments of  $\vec{X}$ . It can be shown that the *minimum* number of sigma-points required to do so is  $2n_x + 1$  and matches the first and second moments accurately for  $\vec{Y}$  and  $\hat{P}_{yy}$ . It should be noted that additional sigma-points in **higher-order unscented transforms** can be used to capture higher-order moments of  $\vec{X}$ . Consideration of these higher-order UTs is beyond the scope of this textbook.

The UKF and FI-UKS/URTSS use the UT to obtain the  $n_\sigma = 2n_\xi + 1$  sigma-points in the SPKF framework. For the first sigma-point of the UT, one has

$$\vec{\xi}_1 = \vec{0} \tag{16.161}$$

with mean weight

$$w_1^{(m)} = \frac{\lambda}{\lambda + n_\xi} \tag{16.162}$$

and covariance weight

$$w_1^{(c)} = \frac{\lambda}{\lambda + n_\xi} + (1 - \alpha^2 + \beta) \quad (16.163)$$

where  $\lambda = \alpha^2(n_\xi + \kappa) - n_\xi$  is a scaling parameter,  $\alpha$  determines the spread of the sigma-points about  $\hat{x}$  and usually set to  $1e^{-2} \leq \alpha \leq 1$ ,  $\kappa$  is a secondary scaling parameter usually set to either 0 or  $3 - n_\xi$ , and  $\beta$  is an extra degree of freedom parameter used to incorporate prior information.

For  $j = 1, \dots, n_\xi$  in the UT, one has

$$\begin{aligned} \vec{\xi}_{1+j,k-1} &= \sqrt{\lambda + n_\xi} \vec{e}_j \\ \vec{\xi}_{1+j+n_\xi,k-1} &= -\sqrt{\lambda + n_\xi} \vec{e}_j \end{aligned} \quad (16.164)$$

where  $\vec{e}_j$  is the unit vector for selecting the  $j^{\text{th}}$  element contribution for  $\vec{\xi}$ , i.e.,

$$\vec{e}_j = [0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0]^T \quad (16.165)$$

with mean and covariance weights for  $j = 1, \dots, 2n_\xi$  as

$$w_{1+j}^{(m)} = w_{1+j}^{(c)} = \frac{1}{2(\lambda + n_\xi)} \quad (16.166)$$

### Central Divided Difference

Another method to determine the sigma-points was derived through the **second-order Stirling's formula** which is computed by replacing the analytically derived first and second order derivatives in the Taylor series expansion by numerically evaluated **central divided differences (CDD)** which are defined for some nonlinear function  $g(\vec{x})$  as

$$\nabla g \approx \frac{g(\vec{x} + h\vec{\delta}_x) - g(\vec{x} - h\vec{\delta}_x)}{2h} \quad (16.167)$$

$$\nabla^2 g \approx \frac{g(\vec{x} + h\vec{\delta}_x) + g(\vec{x} - h\vec{\delta}_x) - 2g(\vec{x})}{h^2} \quad (16.168)$$

where  $h$  is the central difference half-step size. For functions of multivariate Gaussian random vectors, it should be noted that  $h = \sqrt{3}$  is optimal.

The CDKF and FI-CDKS/CDRTSS use the CDD to obtain the  $n_\sigma = 2n_\xi + 1$  sigma-points in the SPKF framework. For the first sigma-point in the CDD, one has

$$\vec{\xi}_1 = \vec{0} \quad (16.169)$$

with mean and covariance weight

$$w_1^{(m)} = w_1^{(c)} = \frac{h^2 - n_\xi}{h^2} \quad (16.170)$$

and for  $j = 1, \dots, n_\xi$  in the CDD, one has

$$\begin{aligned} \vec{\xi}_{1+j,k-1} &= h \vec{e}_j \\ \vec{\xi}_{1+j+n_\xi,k-1} &= -h \vec{e}_j \end{aligned} \quad (16.171)$$

with mean and covariance weights for  $j = 1, \dots, 2n_\xi$  as

$$w_{1+j}^{(m)} = w_{1+j}^{(c)} = \frac{1}{2h^2} \quad (16.172)$$

By inspection, one can see that the UT is equivalent to the CDD if one chooses  $h = \sqrt{3}$ ,  $\alpha = 1$ ,  $\beta = 0$ , and  $\kappa = 3 - n_\xi$ . This choice of  $\kappa$  allows for exact matching of fourth-order monomials for  $g()$ , but for  $n_\xi > 3$ , one has negative weights for the first sigma-point which can make the UKF and the CDKF with  $h = \sqrt{3}$  unstable.

### Third-Order Spherical Cubature

A third method to determine the sigma-points is through **third-order spherical cubature integration rule** which forms a sigma-point approximation of the Gaussian expectation integral for a nonlinear function  $g(\vec{x})$  as

$$\int_{-\infty}^{\infty} \cdots g(\vec{\xi}) \mathcal{N}(\vec{\xi}; \vec{0}, I) d\vec{\xi} \approx w \sum_{i=1}^{2n_\xi} g(c \vec{\xi}_i) \quad (16.173)$$

where  $\vec{\xi}_i$  belong to the symmetric set [1] with generator  $[1, 0, \dots, 0]^T$ , i.e.,

$$[1] = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \dots \right\} \quad (16.174)$$

and  $w$  and  $c$  are chosen to make the integration exact up to third-order monomials for  $g()$ , i.e.,

$$w = \frac{1}{2n_\xi} \quad (16.175)$$

and

$$c = \sqrt{n_\xi} \quad (16.176)$$

The CKF and FI-CKS/CRTSS use the TOSCR to obtain the  $n_\sigma = 2n_\xi$  sigma-points in the SPKF framework. For  $j = 1, \dots, n_\xi$  in the TOSCR, one has

$$\begin{aligned} \vec{\xi}_{1+j, k-1} &= \sqrt{n_\xi} \vec{e}_j \\ \vec{\xi}_{1+j+n_\xi, k-1} &= -\sqrt{n_\xi} \vec{e}_j \end{aligned} \quad (16.177)$$

with mean and covariance weights for  $j = 1, \dots, 2n_\xi$  as

$$w_{1+j}^{(m)} = w_{1+j}^{(c)} = \frac{1}{2n_\xi} \quad (16.178)$$

By inspection, one can see that the TOSCR is equivalent to the UT if one chooses  $\alpha = 1$  and  $\beta = \kappa = 0$  which notably discards the first sigma-point in the UT, i.e., sets the weights to zero which was the original version of the UT.

### Gauss-Hermite Quadrature

Related to the spherical cubature, a fourth method to determine the sigma-points is through the **Gauss-Hermite quadrature rule (GHQR)** which forms a  $p^{\text{th}}$ -order sigma-point approximation of the univariate Gaussian expectation integral for a function  $g(x)$  as

$$\int_{-\infty}^{\infty} g(\xi) \mathcal{N}(\xi; 0, 1) d\xi \approx \sum_{i=1}^p w_i g(\xi_i) \quad (16.179)$$

where the sigma-points  $\xi_i$  are chosen to enforce this quadrature rule exactly for polynomials up to order  $2p - 1$ . It can be shown that this corresponds to choosing  $\xi_i$  as the roots of the  $p^{\text{th}}$ -order Hermite polynomial,  $H_p(\xi)$ , given as

$$H_p(\xi) = (-1)^p \exp\left(\frac{\xi^2}{2}\right) \frac{d^p}{d\xi^p} \exp\left(\frac{-\xi^2}{2}\right) \quad (16.180)$$

which can also be computed through the recursion

$$\begin{aligned} H_0(\xi) &= 1 \\ H_1(\xi) &= \xi \\ H_2(\xi) &= \xi^2 - 1 \\ &\vdots = \vdots \\ H_{p+1}(\xi) &= \xi H_p(\xi) - p H_{p-1}(\xi) \end{aligned} \quad (16.181)$$

In practice, these are typically calculated as the eigenvalues of a suitable tri-diagonal matrix.

Then, noting through the transformation of the Gaussian expectation for  $X \sim \mathcal{N}(\mu, P)$ , one has for the GHQR

$$\int_{-\infty}^{\infty} g(x) \mathcal{N}(x; \mu, P) dx = \int_{-\infty}^{\infty} g(\mu + \sqrt{P}\xi) \mathcal{N}(\xi; 0, 1) d\xi \approx \sum_{i=1}^p w_i g(\mu + \sqrt{P}\xi_i) \quad (16.182)$$

where the unit sigma-points,  $\xi_i$ ,  $i = 1, \dots, p$ , are the roots of the  $p^{\text{th}}$ -order Hermite polynomial and corresponding weights as

$$w_i = \frac{p!}{p^2 [H_{p-1}(\xi_i)]^2} \quad (16.183)$$

Then, for the multivariate random vectors, one can generalize to the  $p^{\text{th}}$ -order **Gauss-Hermite cubature rule (GHCR)** for a  $n_x$ -dimensional Gaussian random vector  $\vec{X} \sim \mathcal{N}(\vec{\mu}, P)$  as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\vec{x}) \mathcal{N}(\vec{x}; \vec{\mu}, P) d\vec{x} \approx \sum_{k=1}^{n_x} w_{\mathcal{I}_k} g(\vec{\mu} + \sqrt{P} \vec{\xi}_{\mathcal{I}_k}) \quad (16.184)$$

where

$$\mathcal{I}_k = \{i_1, \dots, i_{n_x}\} \quad (16.185)$$

$$\vec{\xi}_{I_k} = \begin{bmatrix} \xi_{i_1} \\ \vdots \\ \xi_{i_n} \end{bmatrix} \quad (16.186)$$

and the weights are

$$w_{I_k} = w_{i_1} \cdots w_{i_n} = \frac{p!}{p^2[H_{p-1}(\xi_{i_1})]^2} \cdots \frac{p!}{p^2[H_{p-1}(\xi_{i_n})]^2} \quad (16.187)$$

and  $\xi_{i_k}$ ,  $i_k = 1, \dots, p$  for all possible sets of index combinations,  $k = 1, \dots, n_x$ , which are the  $i_k^{\text{th}}$  root of the  $p^{\text{th}}$ -order Hermite polynomial and  $w_{i_k}$  are the corresponding weights.

Thus, the GHKF and FI-GHKS/GHRTSS, also known as the **quadrature Kalman filter (QKF)**, the **fixed-interval quadrature Kalman smoother (FI-QKS)** or the **quadrature Rauch-Tung-Striebel smoother (QRTSS)**, use the  $p^{\text{th}}$ -order GHCR to obtain the  $p^{n_\xi}$  sigma-points in the SPKF framework. The GHCR uses  $p^{n_\xi}$  sigma-points in order to approximate higher-order polynomials while the UT/CDD/TOSCR only approximate monomials up to the third-order, although by choosing  $\kappa = 3 - n_\xi$  and  $\beta = 2$ , the UT can match fourth-order monomials. This makes the  $p^{\text{th}}$ -order GHCR covariance approximation exact up to  $p^{\text{th}}$ -order monomials for  $g()$  while for the UT/CDD/TOSCR, the covariance approximations are exact for first-order monomials for  $g()$  and maybe the second-order monomials for  $g()$  for UT if  $\kappa = 3 - n_\xi$  and  $\beta = 2$ .

## References

For more information, please refer to the following

- Sarkka, S., “5.6 Unscented Kalman filter,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 86-92
- Sarkka, S., “6.4 Gauss-Hermite Kalman filter,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 103-106
- Sarkka, S., “6.6 Cubature Kalman filter,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 110-114
- Sarkka, S., “9.3 Unscented Rauch-Tung-Striebel smoother,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 144-146
- Sarkka, S., “10.1 General Gaussian Rauch-Tung-Striebel smoother,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 154-155
- Sarkka, S., “10.2 Gauss-Hermite Rauch-Tung-Striebel smoother,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 155-156
- Sarkka, S., “10.3 Cubature Rauch-Tung-Striebel smoother,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 156-159
- Simon, D., “14.3 Unscented Kalman filtering,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 447-457

- Van Der Merwe, R., “Sigma-Point Kalman Filters for Probabilistic Inference in Dynamic State-Space Models,” in *Oregon Health & Science University*, 2004
- Wu, Y., Wu, M., and Hu, D., “A Numerical-Integration Perspective on Gaussian Filters,” in *IEEE Transactions on Signal Processing*, 2006, pp. 2910-2921

## 16.5 Error-State Kalman Filtering

### Error-State Kalman Filtering

In **error-state Kalman filtering** at each time step  $k$ , one considers the **true-state vector**,  $\vec{x}$ , as composed of the **nominal-state vector**,  $\bar{x}$ , and the **error-state vector**, i.e.

$$\vec{x} = \bar{x} \oplus \delta \vec{x} \quad (16.188)$$

where  $\oplus$  is any general composition where each state vector can be written as some continuous-time stochastic state-space model

$$\begin{aligned} \dot{\vec{x}} &= f(\vec{x}, \vec{u}, \vec{w}) \\ \vec{y} &= h(\vec{x}, \vec{v}) \end{aligned} \quad (16.189)$$

where, because of the potential composition of the state vector, the state dynamics for the nominal-state is

$$\dot{\bar{x}} = \bar{f}(\bar{x}, \vec{u}, \vec{w}) \quad (16.190)$$

and the state dynamics for the error-state is

$$\dot{\delta \vec{x}} = f_\delta(\bar{x}, \delta \vec{x}, \vec{u}, \vec{w}) \quad (16.191)$$

where  $\vec{u}$  is the proprioceptive sensor measurement or control input,  $\vec{v}$  is the exteroceptive sensor measurement,  $\vec{W}$  is a zero-mean stationary noise process with power spectral density,  $S$ , and  $\vec{W}$  is a zero-mean stationary noise process with covariance,  $R$ .

From control inputs or proprioceptive sensor measurements at time steps  $k - 1$  and  $k$ , the **error-state prediction step**, also known as the **error-state time update step**, estimates the nominal-state,  $\hat{\bar{x}}$  using an integration of its dynamics equation, i.e., the continuous-time update for the nominal-state estimate is

$$\hat{\bar{x}}_{k|k-1} = \hat{\bar{x}}_{k-1|k-1} + \int_{t(k-1)}^{t(k)} \bar{f}(\hat{\bar{x}}, \vec{u}, 0) \quad (16.192)$$

and the discrete-time update for the nominal-state estimate is

$$\hat{\bar{x}}_{k|k-1} = \bar{f}(\bar{x}_{k-1|k-1}, \vec{u}_{k-1}, 0) \quad (16.193)$$

Notably, the error-state estimate implicitly remains zero as any constant *known* bias should be encompassed in the nominal-state dynamics in the prediction step. However, the covariance of the error-state must be estimated from time step  $k - 1$  to  $k$  for use with the **error-state correction step**, also known as the **error-state measurement update step**, using a Kalman filter framework.

After the error-state correction step, there will be a non-zero error-state estimate due to the correction step incorporating the measurement at time step  $k$ . Then, one must perform a **nominal-state update step** by updating the nominal-state estimate,  $\hat{x}_{k|k-1}$ , with the error-state estimate,  $\delta\hat{\vec{x}}_k$ , through its composition, i.e.

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} \oplus \delta\hat{\vec{x}}_k \quad (16.194)$$

where  $\hat{x}_{k|k}$  is the updated nominal-state estimate due to the error-state estimate due to the measurement at time step  $k$ . Then, as the error-state estimate has been “injected” into the nominal-state estimate, the error-state estimate is updated to zero, i.e.,

$$\delta\hat{\vec{x}}'_k = \vec{0} \quad (16.195)$$

which can be formally defined relative to the true error-state via the relationship

$$\delta\vec{x}' = g(\delta\vec{x}) = \delta\vec{x} \ominus \bar{x} \quad (16.196)$$

where  $\ominus$  stands for the composition inverse of  $\oplus$  which “removes” the error-state estimate from the true error-state as it has been “injected” into the nominal-state estimate. By definition of the composition inverse, the error-state covariance,  $P_{k|k}$ , is updated as

$$P'_{k|k} = \left( \frac{\partial g}{\partial \delta\vec{x}} \right)_{\delta\vec{x}=\delta\hat{\vec{x}}_k} P_{k|k} \left( \frac{\partial g}{\partial \delta\vec{x}} \right)^T_{\delta\vec{x}=\delta\hat{\vec{x}}_k} \quad (16.197)$$

where  $P_{k|k}$  would be obtained using a nonlinear Kalman filter after the error-state covariance prediction and correction steps.

### Error-State Extended Kalman Filter

The **error-state extended Kalman filter (ES-EKF)** utilizes the Jacobian linearization approach of the EKF in the estimation of the error-state and its covariance.

For the error-state covariance update in the prediction step at time step  $k$ , the **ES-EKF prediction step** uses the linearized state matrix as

$$A_{k-1} = \left[ \frac{\partial f_\delta}{\partial \delta\vec{x}} \right]_{\bar{x}=\hat{x}_k, \delta\vec{x}=0, \vec{u}=\vec{u}_{k-1}, \vec{w}=0} \quad (16.198)$$

and the linearized process noise gain matrix

$$L_{k-1} = \left[ \frac{\partial f_\delta}{\partial \vec{w}} \right]_{\bar{x}=\hat{x}_k, \delta\vec{x}=\vec{0}, \vec{u}=\vec{u}_{k-1}, \vec{w}=0} \quad (16.199)$$

to form the discretized state matrix with sampling time interval,  $\Delta t$ , from time step  $k - 1$  to  $k$ , as

$$F_{k-1} = \exp(A_{k-1}\Delta t) \quad (16.200)$$

and discretized noise covariance  $Q_{k-1}$  from the PSD,  $S$ , of the process noise  $\vec{W}$  as

$$Q_{k-1} = \int_0^{\Delta t} \exp(A_{k-1}\tau) L_{k-1} S L_{k-1}^T \exp(A_{k-1}\tau)^T d\tau \quad (16.201)$$

It should be noted that  $Q_{k-1}$  can be efficiently computed using the following representation

$$\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} = \exp \left( \begin{bmatrix} -A_{k-1} & L_{k-1} S L_{k-1}^T \\ 0 & A_{k-1}^T \end{bmatrix} \Delta t \right) \quad (16.202)$$

and then multiplying

$$Q_{k-1} = E_{22}^T E_{12} \quad (16.203)$$

where notably, a first-order approximation of matrix exponential would provide

$$Q_{k-1} = (I + \Delta t A_{k-1}) \left( \Delta t L_{k-1} S L_{k-1}^T \right) \quad (16.204)$$

Finally, one can form the prior error-state covariance as

$$P_{k|k-1} = F_{k-1} P_{k-1|k-1} F_{k-1}^T + Q_{k-1} \quad (16.205)$$

At time step  $k$ , the **ES-EKF correction step** compares an exteroceptive sensor measurement,  $\vec{y}_k$ , with the expected measurement based on the nominal-state, i.e.,

$$\hat{\vec{y}}_k = h(\hat{\vec{x}}_{k|k-1}, \vec{v}_k) \quad (16.206)$$

where  $\vec{v}_k$  is white measurement noise with covariance  $R_k$ . For the ES-EKF, one computes the innovation as

$$\delta \vec{y}_k = \vec{y}_k - \hat{\vec{y}}_k \quad (16.207)$$

the linearized output matrix as

$$H_k = \left[ \frac{\partial h}{\partial \vec{x}} \right]_{\vec{x}=\hat{\vec{x}}_k, \vec{v}=0} \quad (16.208)$$

and the linearized measurement noise gain matrix

$$M_k = \left[ \frac{\partial h}{\partial \vec{v}} \right]_{\vec{x}=\hat{\vec{x}}_k, \vec{v}=0} \quad (16.209)$$

to form the approximate innovation covariance noise covariance

$$S_k = H_k P_{k|k-1} H_k^T + M_k R_k M_k^T \quad (16.210)$$

and the sub-optimal Kalman gain as

$$K_k = P_{k|k-1} H_k^T S_k^{-1} \quad (16.211)$$

Finally, one can form the posterior error-state mean estimate as

$$\hat{\vec{x}}_k = K_k \delta \vec{y}_k \quad (16.212)$$

and the approximate posterior error-state covariance as

$$P_{k|k} = P_{k|k-1} - K_k S_k K_k^T \quad (16.213)$$

The next step in the recursion uses this error-state estimated to update the nominal-state in the state update step.

Lastly, it should be noted that the **multiplicative extended Kalman filter (MEKF)** is a special case of the ES-EKF when the true-state composition between the nominal-state and the error-state is multiplicative, e.g., in attitude state estimation.

### Error-State Sigma-Point Kalman Filtering

The **error-state sigma-point Kalman filter (ES-SPKF)** utilizes the sigma-point approach of the SPKF in the estimation of the error-state and its covariance.

### References

For more information, please refer to the following

- Madyastha, V., Ravindra, V., Srinath Mallikarjunan, S., and Goyal, A., “Extended Kalman Filter vs. Error State Kalman Filter for Aircraft Attitude Estimation,”” in *AIAA Guidance, Navigation, and Control Conference*, 2011
- Sola, J., “Quaternion kinematics for the error-state Kalman filter,” arXiv:1711.02508, 2017

## 16.6 Equivariant Kalman Filtering

### References

For more information, please refer to the following

- Barrau, A., and Bonnabel, S., “The invariant extended Kalman filter as a stable observer,” in *IEEE Transactions on Automatic Control*, 62(4), 2017, pp. 1797–1812

## 16.7 Variational Bayes Kalman Filtering

Recall that in the previous discrete-time Kalman filter recursion, one typically assumes a linear stochastic state-space model as

$$\begin{aligned}\vec{x}_k &= F_{k-1} \vec{x}_{k-1} + G_{k-1} \vec{u}_{k-1} + \vec{w}_{k-1} \\ \vec{y}_k &= H_k \vec{x}_k + \vec{v}_k\end{aligned}\tag{16.214}$$

with white noise  $\vec{W}_k \sim \mathcal{N}(0, Q_k)$  and  $\vec{V}_k \sim \mathcal{N}(0, R_k)$  which are uncorrelated with each other.

### Joint State and Noise Parameter Kalman Filter

### Variational Bayes Adaptive Kalman Filter

### GSM Filter

## 16.8 Particle Filtering

Nonlinear Kalman filters assume a multivariate Gaussian posterior conditional PDF for  $\vec{X}_k | \vec{y}_{1:k}$ . Thus, this can be a poor approximation of the optimal Bayes filter when  $\vec{X}_k | \vec{y}_{1:k}$  is **multi-modal**, i.e. there are multiple values for the maximum of  $f_{\vec{X}_k | \vec{Y}_{1:k}}(\vec{x} | \vec{y}_{1:k})$ , and/or  $\vec{X}_k | \vec{Y}_{1:k}$  is **heavy-tailed**, i.e. one cannot bound  $\vec{X}_k | \vec{Y}_{1:k}$  by an exponential function as  $\vec{x} \rightarrow \infty$ .

While advanced filters exist for modeling different approximations for the posterior conditional PDFs in the Bayes filter context, the most general approach is to recursively update the posterior distribution using sequential Monte Carlo (SMC) methods by approximating the posterior conditional PDF by a set of weighted samples, i.e. “particles,” without making any analytical assumptions about its form. Thus, one can consider these particles as a PMF approximation to the PDF. Any filter which uses a particle approximation is known as a **particle filter (PF)** and can be used for state estimation for nonlinear, non-Gaussian stochastic state-space systems. This section will introduce the basic concepts used in particle filters.

## Sequential Importance Sampling

The primary method in particle filtering is the use of **sequential importance sampling (SIS)** as an extension of Monte Carlo methods of estimation. Recall that importance sampling represents a PDF  $f_{\vec{X}}(\vec{x})$  by an empirical approximation

$$f_{\vec{X}}(\vec{x}) \approx \hat{f}_{\vec{X}}(\vec{x}) = \sum_{j=1}^{n_p} w^{(i)} \delta(\vec{x} - \vec{x}^{(i)}) \quad (16.215)$$

where  $\delta(\bullet)$  is the Dirac delta function and  $\vec{x}^{(i)}$  for  $i = 1, \dots, n_p$  are the particles with corresponding **importance weights**,  $w^{(i)}$ , which are drawn from some related, easy-to-sample **importance function**,  $\pi(\vec{x})$ , also known as the **proposal PDF**. With this definition, the importance weights are given by

$$w^{(i)} = \frac{\frac{f_{\vec{X}}(\vec{x}^{(i)})}{\pi(\vec{x}^{(i)})}}{\sum_{i=1}^{n_p} \frac{f_{\vec{X}}(\vec{x}^{(i)})}{\pi(\vec{x}^{(i)})}} \quad (16.216)$$

Furthermore, the expectation of the function,  $g(\vec{x})$ , can be approximated using this PMF approximation, e.g.

$$\mathbb{E}[g(\vec{x})] = \int g(\vec{x}) f_{\vec{X}}(\vec{x}) d\vec{x} \approx \sum_{i=1}^{n_p} w^{(i)} g(\vec{x}^{(i)}) \quad (16.217)$$

In the Bayes filter context with states  $\vec{x}_{1:k}$  and measurements,  $\vec{y}_{1:k}$ , if one assumes HMM and conditional independence between the state and measurements, then one can write the importance function as a recursive relationship

$$\pi(\vec{x}_{0:k} | \vec{y}_{1:k}) = \pi(\vec{x}_{1:k-1} | \vec{y}_{1:k-1}) \pi(\vec{x}_k | \vec{x}_{0:k-1}, \vec{y}_{1:k}) \quad (16.218)$$

which allows the recursive computation of the importance weights as

$$w_k^{(i)} \propto \frac{f_{\vec{Y}_k | \vec{X}_k}(\vec{y}_k | \vec{x}_k^{(i)}) f_{\vec{X}}(\vec{x}_k^{(i)} | \vec{x}_{k-1}^{(i)})}{\pi(\vec{x}_k^{(i)} | \vec{x}_{0:k-1}^{(i)}, \vec{y}_{1:k})} w_{k-1}^{(i)} \quad (16.219)$$

which allows one to sequentially update the importance weights given an appropriate choice of importance function,  $\pi(\vec{x}_k | \vec{x}_{0:k-1}, \vec{y}_{1:k})$ .

For the Bayes filter, recall that one can sample from this importance function, evaluate the likelihood function,  $f_{\vec{Y}_k | \vec{X}_k}(\vec{y}_k | \vec{x}_k)$ , and evaluate the state transition PDF,  $f_{\vec{X}}(\vec{x}_k | \vec{x}_{k-1})$ , one can iteratively compute

the importance weights if one generates an prior set of particles and importance weights. This procedure allows one to compute

$$\mathbb{E}[g(\vec{x})] \approx \frac{\frac{1}{n_p} \sum_{i=1}^{n_p} w_k^{(i)} g(\vec{x}_k^{(i)})}{\frac{1}{n_p} \sum_{i=1}^{n_p} w_k^{(i)}} = \sum_{i=1}^{n_p} \tilde{w}_k^{(i)} g(\vec{x}_k^{(i)}) \quad (16.220)$$

with **normalized importance weights**

$$\tilde{w}_k^{(i)} \propto \frac{w_k^{(i)}}{\sum_{i=1}^{n_p} w_k^{(i)}} \quad (16.221)$$

where it should be noted that this estimate asymptotically converges if the mean and variance of  $g(\vec{x}_k)$  and  $w_k$  exist and are bounded and if the support of the importance function includes the support of the posterior PDF. Thus, as  $n_p \rightarrow \infty$ , the posterior conditional PDF of the Bayes filter can be approximately by the PMF estimate

$$\hat{f}_{\vec{X}_k | \vec{Y}_{1:k}}(\vec{x}_k | \vec{y}_{1:k}) \approx \sum_{i=1}^{n_p} \tilde{w}^{(i)} \delta(\vec{x}_k - \vec{x}_k^{(i)}) \quad (16.222)$$

where the approximation improves as the number of particles is increased.

Then, note that one can then form a state estimate based on a Bayes estimator, e.g., the MAP state estimate

$$\hat{x}_{PF,MAP} = \vec{x}^{(i)} \text{ with } i \text{ chosen as } \operatorname{argmax}_i \tilde{w}^{(i)} \quad (16.223)$$

or the MMSE estimate

$$\hat{x}_{PF,MMSE} = \sum_{i=1}^{n_p} \tilde{w}^{(i)} \vec{x}_k^{(i)} \quad (16.224)$$

and one can also estimate other statistics of the posterior conditional PDF, e.g. the covariance as

$$\hat{P}_{PF} = \sum_{i=1}^{n_p} \tilde{w}^{(i)} \left( \vec{x}_k^{(i)} - \hat{x}_{PF,MMSE} \right) \left( \vec{x}_k^{(i)} - \hat{x}_{PF,MMSE} \right)^T \quad (16.225)$$

## Particle Filter Implementation

The most important design criteria in particle filter implementation is the choice of the importance PDF. The optimal importance function which minimizes the variance on the importance weights is given by the true conditional PDF given the *entire* previous state history and all observations, i.e.,

$$\pi^{opt}(\vec{x}_k | \vec{x}_{0:k-1}, \vec{y}_{1:k}) = f_{\vec{X}_k | \vec{X}_{k-1}, \vec{Y}_k}(\vec{x}_k | \vec{x}_{0:k-1}, \vec{y}_{1:k}) \quad (16.226)$$

However, sampling from this conditional PDF may be impractical for arbitrary PDFs for dynamical systems. Thus, other practical choices are usually made for the importance PDF in particle filter implementation.

The simplest choice is the state transition conditional PDF, i.e.

$$\pi(\vec{x}_k | \vec{x}_{0:k-1}, \vec{y}_{1:k}) = f_{\vec{X}_k | \vec{X}_{k-1}}(\vec{x}_k | \vec{x}_{k-1}) \quad (16.227)$$

which allows one to form the importance weight recursion as simply

$$w_k^{(i)} \propto f_{\vec{Y}_k | \vec{X}_k}(\vec{y}_k | \vec{x}_k^{(i)}) w_{k-1}^{(i)} \quad (16.228)$$

where  $f_{\vec{Y}_k|\vec{X}_k}(\vec{y}_k|\vec{x}_k^{(i)})$  is the likelihood function.

For example, if the state equation has zero-mean AWGN,  $\vec{w}_k$ , with covariance,  $Q_{k-1}$ , i.e.

$$\vec{x}_k = f(\vec{x}_{k-1}, \vec{u}_{k-1}) + \vec{w}_{k-1} \quad (16.229)$$

then one can choose

$$\pi(\vec{x}_k|\vec{x}_{0:k-1}, \vec{y}_{1:k}) = f_N(\vec{x}_k; f_{k-1}(\vec{x}_{k-1}, \vec{u}_{k-1}), Q_{k-1}) \quad (16.230)$$

and if the measurement equation has zero-mean AWGN,  $\vec{v}_k$ , with covariance,  $R_k$ , i.e.

$$\vec{y}_k = h(\vec{x}_k) + \vec{v}_k \quad (16.231)$$

then one can choose

$$f_{\vec{Y}_k|\vec{X}_k^{(i)}}(\vec{y}_k|\vec{x}_k^{(i)}) = f_N(\vec{y}_k; h(\vec{x}_k), R_k) \quad (16.232)$$

In addition, one of the primary disadvantages of implementing the SIS algorithm discussed previously is that the variance of the importance weights increases stochastically over time. Typically, after a few iterations, most of the weights tend towards zero, effectively being removed from the particle set due to their numerically insignificance weight. To avoid this **degeneracy problem** of the particles, a resampling stage is used to eliminate particles with low importance weights and multiply samples with high importance weights. These methods are collectively known as **sequential importance sampling with resampling (SIS/R)** methods, also known as the shortened **sequential importance resampling (SIR)** or **sampling importance resampling (SIR)**. Practical methods for this resampling step are another implementation choice in particle filters. Regardless of the chosen resampling method, one should obtain  $n_p$  particles,  $\tilde{w}^{(m)+}$ , distributed approximately according to the posterior PDF.

The simplest method for resampling is the **multinomial resampling** which can be summarized as

$$\vec{x}^{(i)} = \vec{x}^{(j)} \quad \text{with probability } \tilde{w}^{(i)} \quad (16.233)$$

where it can be shown that the ensemble PDF of the resampled particles in this method tends to the posterior conditional PDF as  $n_p \rightarrow \infty$ . Thus, one has the approximation

$$f_{\vec{X}_k|\vec{Y}_k}(\vec{x}_k|\vec{y}_{1:k}) \approx \hat{f}_{X|Y}(\vec{x}_k|\vec{y}_{1:k}) = \frac{1}{n_p} \sum_{i=1}^{n_p} \delta(\vec{x}_k - \vec{x}_k^{(i)}) \quad (16.234)$$

where the PDF has been approximated by number of copies instead of the importance weights. Thus, one no longer recursively updates the weights, but has the model

$$w_k^{(i)} \propto \frac{f_{\vec{Y}_k|\vec{X}_k}(\vec{y}_k|\vec{x}_k^{(i)}) f_{\vec{X}_k|\vec{X}_{k-1}}(\vec{x}_k^{(i)}|\vec{x}_{k-1}^{(i)})}{\pi(\vec{x}_k^{(i)}|\vec{x}_{0:k-1}^{(i)}, \vec{y}_{1:k})} \quad (16.235)$$

This can be written out as the following algorithm.

For  $i = 1, \dots, n_p$

Sample  $r \sim \mathcal{U}[0, 1]$

If  $\sum_{m=1}^{j-1} \tilde{w}_k^{(m)} < r$  but  $\sum_{m=1}^j \tilde{w}_k^{(m)} \geq r$ , set  $\vec{x}^{(i)} = \vec{x}^{(j)}$

**Residual resampling**, also known as **remainder resampling**, attempts to decrease the variance of the SIR by keeping copies of particles based on rounding down the expected number of particles based on the importance weights, then only resampling based on the remainders from all particles. This can be written out as the following algorithm.

Set  $n_c = 1$

Set  $n_r = n_p$

For  $i = 1, \dots, n_p$

    For  $j = 1, \dots, \text{floor}(n_p \tilde{w}_k^{(i)})$

        Set  $\vec{x}^{(n_c)} = \vec{x}^{(i)}$

$n_c = n_c + 1$

$n_r = n_r - \text{floor}(n_p \tilde{w}_k^{(i)})$

For  $i = 1, \dots, n_r$

    Sample  $r \sim \mathcal{U}[0, 1]$

    If  $\sum_{m=1}^{j-1} n_p \tilde{w}_k^{(m)} - \text{floor}(n_p \tilde{w}_k^{(i)}) < r$  but  $\sum_{m=1}^j n_p \tilde{w}_k^{(m)} - \text{floor}(n_p \tilde{w}_k^{(i)}) \geq r$ , set  $\vec{x}^{(n_c)} = \vec{x}^{(j)}$

$n_c = n_c + 1$

**Stratified resampling** pre-partitions the  $[0, 1]$  interval into  $n_p$  disjoint sets and samples from these. Then, one performs the resampling, i.e.

For  $i = 1, \dots, n_p$

    Sample  $r(i) \sim \mathcal{U}[(i-1)/n_p, i/n_p]$

    If  $\sum_{m=1}^{j-1} \tilde{w}_k^{(m)} < r(i)$  but  $\sum_{m=1}^j \tilde{w}_k^{(m)} \geq r(i)$ , set  $\vec{x}^{(i)} = \vec{x}^{(j)}$

This approach can be even further simplified by **systematic resampling**, a.k.a. **low-variance resampling (LVR)**, also known as the **stochastic universal sampler**, which samples a *single* random number in order to resample from the  $n_p$  disjoint sets, i.e.

Sample  $r \sim \mathcal{U}[0, 1/n_p]$

For  $i = 1, \dots, n_p$

    If  $\sum_{m=1}^{j-1} \tilde{w}_k^{(m)} < r + (i-1)/n_p$  but  $\sum_{m=1}^j \tilde{w}_k^{(m)} \geq r + (i-1)/n_p$ , set  $\vec{x}^{(i)} = \vec{x}^{(j)}$

In practice, any of the residual, stratified, or systematic resampling methods have similar performance, but are all better than the simple multinomial resampling method.

## Bootstrap Particle Filter

The **bootstrap particle filter (BPF)** uses the state transition conditional PDF as the importance function and one of the four resampling methods and can be outlined as follows:

First, the **BPF initialization step** samples the initial  $n_p$  particles from an initial distribution as

$$\vec{x}_0^{(i)} \sim f_{\vec{X}_0}(\vec{x}_0) \quad i = 1, \dots, n_p \quad (16.236)$$

Then, the recursive **BPF SIR step** performs the sequential importance resampling as

S Sample  $n_p$  particles from the importance function.

For  $i = 1, \dots, n_p$ , sample the state transition PDF with  $\vec{W}_{k-1}$  as the random variate

$$\vec{x}_k^{(i)} \sim f_{\vec{X}_k|\vec{X}_{k-1}}(\vec{x}_k | \vec{x}_{k-1|k-1}^{(i)}) = f_{k-1}(\vec{x}_{k-1|k-1}^{(i)}, \vec{u}_{k-1}, \vec{W}_{k-1}) \quad (16.237)$$

I Compute the Importance weights.

For  $i = 1, \dots, n_p$ , compute the relative importance weights by

$$w_k^{(i)} = f_{\vec{Y}_k|\vec{X}_k}(\vec{y}_k | \vec{x}_{k|k-1}^{(i)}) \quad (16.238)$$

For  $i = 1, \dots, n_p$ , compute the normalized importance weights by

$$\tilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{i=1}^{n_p} w_k^{(i)}} \quad (16.239)$$

R Resample  $\vec{x}_k^{(i)}$  using the multinomial, residual, stratified, or systematic method with weights,  $\tilde{w}_k^{(i)}$

Finally, the **BPF state extraction step** extracts the chosen statistics on the estimated posterior distribution of  $\vec{X}_k$ . e.g., the mean

$$\hat{\vec{x}}_{PF} = \frac{1}{n_p} \sum_{i=1}^{n_p} \vec{x}_k^{(i)} \quad (16.240)$$

and covariance

$$\hat{P}_{PF} = \frac{1}{n_p - 1} \sum_{i=1}^{n_p} \left( \vec{x}_k^{(i)} - \hat{\vec{x}}_{PF} \right) \left( \vec{x}_k^{(i)} - \hat{\vec{x}}_{PF} \right)^T \quad (16.241)$$

## Particle Depletion and Mitigation

Furthermore, since the resampling step favors the selection of particles with higher weights, particles with low weights may be eliminated altogether after the resampling while others may have many copies made, a phenomenon known as **particle depletion**, also known as **sample impoverishment**. Therefore, resampling methods typically use additional steps to counteract particle depletion.

One simple method is known as **roughening** which adds zero-mean white noise,  $n^{(i)}$ , typically Gaussian, to each particle after the resampling process. Here, the  $m^{\text{th}}$  element of the vector  $\vec{x}_{k|k}^{(i)}$  with  $m = 1, \dots, n_x$  can be modified by an additional sampling procedure

$$\vec{x}_{k|k}^{(i)}(m) = \vec{x}_{k|k}^{(i)} + \sqrt{\alpha M(m)n_p^{-1/n_x}} n^{(i)} \quad (16.242)$$

where  $\alpha$  is a tuning parameter, e.g.  $\alpha = 0.2$ , that specifies the amount of jitter added to each particle,  $M(m)$  is the maximum difference between elements of the particle vectors before roughening, i.e.

$$M(m) = \max_{i,j} |\vec{x}_{k|k}^{(i)}(m) - \vec{x}_{k|k}^{(j)}(m)| \quad (16.243)$$

and  $n^{(i)} \sim \mathcal{N}(0, I)$ .

Another form of roughening on the prior particles generated from the importance function is known as **prior editing**. This involves the rejection and resampling of a prior sample,  $\vec{x}_{k|k-1}^{(i)}$ , before the resampling procedure if it falls in a region of the state space with a small importance weight  $\tilde{w}_k^{(i)}$ . If the prior sample is rejected, then it is roughened as many times as necessary until it is in a region of significant probability, e.g. within  $6\sigma_y$  of  $\vec{y}_k - h(\vec{x}_{k|k-1}^{(i)})$ .

A more robust version of prior editing can be achieved through the use of a **Markov chain Monte Carlo (MCMC) sampling**. This approach moves the prior particle,  $\vec{x}_{k|k-1}^{(i)}$ , to a new randomly generated state,  $\vec{x}_{k|k-1}^{(i)'}$  if a uniformly distributed random number is less than an **acceptance probability**,  $\alpha$ . While there are different methods for computing this acceptance probability, it is generally computed as the probability that the prior particle is *consistent* with the measurement, relative to the probability that new randomly generated state is *consistent* with the measurement. In the **Metropolis-Hastings algorithm**, the acceptance probability is given by

$$\alpha = \min \left[ 1, \frac{f_{\vec{Y}_k | \vec{X}_k}(\vec{y}_k | \vec{x}_{k|k-1}^{(i)'}) f_x(\vec{x}_{k|k-1}^{(i)'} | \vec{x}_{k-1|k-1}^{(i)})}{f_{\vec{Y}_k | \vec{X}_k}(\vec{y}_k | \vec{x}_{k|k-1}^{(i)}) f_x(\vec{x}_{k|k-1}^{(i)} | \vec{x}_{k-1|k-1}^{(i)})} \right] \quad (16.244)$$

where the first fraction is the ratio of the likelihood conditioned on the new particle to the likelihood conditioned on the old particle. The second fraction is the ratio of the probability of the new particle transitioning to the old particle transitioning. The acceptance probability is the product of these two fractions which increases as the probability of the new particle increases. The old prior particle,  $\vec{x}_{k|k-1}^{(i)}$ , is changed to the new particle,  $\vec{x}_{k|k-1}^{(i)'}$ , if the old particle has a low probability of being selected with the resampling step, thereby maintaining diversity in the particles that result from the resampling step.

## Regularized and Adaptive Resampling

Instead of importance resampling via multinomial, residual, stratified, or systematic methods, an alternative formulation is the **regularized resampling** which resamples from a continuously-valued approximation of

the posterior conditional PDF instead of sampling from the PMF approximation, i.e., instead of resampling from

$$\hat{f}_{\vec{X}_k|\vec{Y}_k}(\vec{x}_k|\vec{y}_k) = \sum_{i=1}^{n_p} \tilde{w}_k^{(i)} \delta(\vec{x}_k - \vec{x}_k^{(i)}) \quad (16.245)$$

one resamples from

$$\hat{f}_{\vec{X}_k|\vec{Y}_k}(\vec{x}_k|\vec{y}_k) = \sum_{i=1}^{n_p} \tilde{w}_k^{(i)} \mathcal{K}_h(\vec{x}_k - \vec{x}_k^{(i)}) \quad (16.246)$$

where  $\mathcal{K}_h(\bullet)$  is a scaled kernel PDF, i.e.

$$\mathcal{K}_h(\vec{x}_k - \vec{x}_k^{(i)}) = h^{-n_x} \mathcal{K}\left(\frac{\vec{x}_k - \vec{x}_k^{(i)}}{h}\right) \quad (16.247)$$

where  $h$  is the positive scalar kernel bandwidth, and  $\mathcal{K}(\bullet)$  is a **kernel PDF** with zero-mean and finite variance.

$K(\bullet)$  and  $h$  are typically chosen to minimize a measure of the error between the true posterior PDF and the approximation, i.e.

$$\mathcal{K}_{opt}(\vec{x}), h_{opt} = \underset{\mathcal{K}, h}{\operatorname{argmin}} \int \cdots \int \left( \hat{f}_{\vec{X}_k|\vec{Y}_k}(\vec{x}_k|\vec{y}_k) - f_{\vec{X}_k|\vec{Y}_k}(\vec{x}_k|\vec{y}_k) \right)^2 d\vec{x} \quad (16.248)$$

In the case of equal weights, i.e.  $\tilde{w}_k^{(i)} = n_p^{-1}$  for all  $i$ , the optimal kernel is the **Epanechnikov kernel**

$$\mathcal{K}_{opt}(\vec{x}) = \begin{cases} \frac{n_x+2}{2V_n} (1 - \|\vec{x}\|_2^2) & \text{if } \|\vec{x}\|_2 < 1 \\ 0 & \text{otherwise} \end{cases} \quad (16.249)$$

where  $V_n$  is the volume of the  $n_x$ -dimensional unit hypersphere which can be shown for  $n_x = 1$ , i.e., a line,  $V_n = 2$ , for  $n_x = 2$ , i.e., a circle,  $V_n = \pi$ , for  $n_x = 3$ , i.e., a sphere,  $V_n = 4\pi/3$ , and for  $n_x \geq 3$ ,  $V_n = 2\pi V_{n-2}/n_x$ . Furthermore, if  $f_{\vec{X}_k|\vec{Y}_k}(\vec{x}_k|\vec{y}_k)$  is a multivariate Gaussian with an identity matrix, then the optimal bandwidth is

$$h_{opt} = \left( 8V_n^{-1}(n_x + 4)(2\sqrt{\pi})^{n_x} \right)^{1/(n_x+4)} n_p^{-1/(n_x+4)} \quad (16.250)$$

where, to handle multimodal PDF's one should use  $h = h_{opt}/2$ .

Although these choices are optimal for only equal weights and a Gaussian PDF, they are often used in practice and the bandwidth,  $h$ , is often treated as a tuning parameter for the regularized resampling. Note that the regularization produces a continuous approximation as a weighted sum of these kernel PDFs with the normalized weights. Furthermore, since this regularization procedure assumes  $f_{\vec{X}_k|\vec{Y}_k}(\vec{x}_k|\vec{y}_k)$  has unity variance, one should numerically compute the covariance of  $\vec{x}_k^{(i)}$  at each time step, i.e.,  $P_k$ . Then, as a kernel, one should use

$$\mathcal{K}_h(\vec{x}) = \left( \det P_k^{1/2} \right)^{-1} h^{-n_x} \mathcal{K}_{opt} \left( \frac{P_k^{-1/2} \vec{x}}{h} \right) \quad (16.251)$$

As the resampling step of particle filtering can be computationally expensive, often one can employ **adaptive resampling** which monitors the “effective” number of particles in the SIS approach at time step  $k$ ,  $\hat{n}_{p,k}^{eff}$ , which can be estimated from the variance of the importance weights, i.e.,

$$\hat{n}_{p,k}^{eff} = \frac{1}{\sum_{i=1}^{n_p} (\tilde{w}_k^{(i)})^2} \quad (16.252)$$

and implements a resampling step at time step  $k$  if

$$\frac{\hat{n}_{p,k}^{eff}}{n_p} < \tau_{eff} \quad (16.253)$$

where  $\tau_{eff}$  is the **effective number threshold** for the test statistic,  $\hat{n}_{p,k}^{eff}/n_p$ , e.g., 0.1.

When implementing adaptive resampling, one must recursively update the importance weights for  $i = 1, \dots, n_p$  as

$$w_k^{(i)} = \frac{f_{\vec{Y}_k|\vec{X}_k}(\vec{y}_k|\vec{x}_k^{(i)}) f_{\vec{X}_k}(\vec{x}_k^{(i)}|\vec{x}_{k-1}^{(i)})}{\pi(\vec{x}_k^{(i)}|\vec{x}_{0:k-1}^{(i)}, \vec{y}_{1:k})} \tilde{w}_{k-1}^{(i)} \quad (16.254)$$

and then for  $i = 1, \dots, n_p$ , compute the normalized importance weights by

$$\tilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{i=1}^{n_p} w_k^{(i)}} \quad (16.255)$$

After a resampling step is performed, one resets the normalized importance weights to

$$\tilde{w}_k^{(i)} = \frac{1}{n_p} \quad (16.256)$$

which can be used in the recursion for the importance weights.

### Alternative Importance Function Approximation as Gaussian

Instead of using the state transition PDF as the importance function, i.e.,

$$\pi(\vec{x}_k|\vec{x}_{0:k-1}, \vec{y}_{1:k}) = f_X(\vec{x}_k|\vec{x}_{k-1}) \quad (16.257)$$

an alternative for a particle filter is to update each prior particle according to a multivariate Gaussian approximation of the optimal importance function, i.e.,

$$\pi(\vec{x}_k|\vec{x}_{0:k-1}, \vec{y}_{1:k}) = f_N(\vec{x}_k; \hat{\vec{x}}_{k|k-1}, P_{k|k-1}) \quad (16.258)$$

Here, for each *single* particle,  $i = 1, \dots, n_p$

$$\vec{X}_{k|k-1}^{(i)} \sim N(\hat{\vec{x}}_{k|k-1}^{(i)}, P_{k|k-1}^{(i)}) \quad (16.259)$$

where the hyper-parameters of each multivariate Gaussian importance function approximation is updated via an EKF or one of the SPKFs. This alternative method is known as an **extended Kalman particle filter (EKPF)** or a **sigma-point particle filter (SPPF)**, respectively. This use of an additional filter for *each* particle directly improves the importance function approximation and improves the PF performance in many cases, albeit at significant computational and memory costs. The **extended Kalman/sigma-point particle filter (EK/SPPF)** uses  $n_p$  EKF/SPKFs to approximate the importance function for each particle as a multivariate Gaussian and can be outlined as follows.

First, the **EK/SPPF initialization step** samples the initial  $i = 1, \dots, n_p$  particles from an initial distribution for the state, importance weights, and state covariances as

$$\begin{aligned} \hat{\vec{x}}_{0|0}^{(i)} &\sim f_{\vec{X}_0}(\vec{x}_0) \\ w_0^{(i)} &= f_{\vec{X}_0}(\vec{x}_{0|0}^{(i)}) \\ \tilde{w}_0^{(i)} &= \frac{w_0^{(i)}}{\sum_{i=1}^{n_p} w_0^{(i)}} \\ \text{Assign } P_{0|0}^{(i)} \end{aligned} \tag{16.260}$$

Also, if implementing adaptive resampling, one must choose an effective number threshold ratio for resampling,  $\tau_{eff}$ .

Then, the recursive **EK/SPPF SIR step** performs the sequential importance resampling as

S Sample particles from the importance functions.

For  $i = 1, \dots, n_p$ ,

P Compute  $\vec{x}_{k|k-1}^{(i)}$  and  $P_{k|k-1}^{(i)}$  with the prediction step of the corresponding EKF/SPKF

C Compute  $\vec{x}_{k|k}^{(i)}$  and  $P_{k|k}^{(i)}$  with the correction step of the corresponding EKF/SPKF

Sample  $\vec{x}_k^{(i)} \sim \mathcal{N}(\vec{x}_{k|k}^{(i)}, P_{k|k}^{(i)})$  as the  $i^{\text{th}}$  importance function from the EKF/SPKF

I Compute the importance weights.

For  $i = 1, \dots, n_p$ , compute the relative importance weights by

$$w_k^{(i)} = \frac{f_{\vec{Y}_k|\vec{X}_k}(\vec{y}_k|\vec{x}_{k|k-1}^{(i)})f_{\vec{X}_k|\vec{X}_{k-1}}(\vec{x}_k^{(i)}|\vec{x}_{k-1|k-1}^{(i)})}{f_{\mathcal{N}}(\vec{x}_k^{(i)}; \vec{x}_{k|k}^{(i)}, P_{k|k}^{(i)})} \tilde{w}_{k-1}^{(i)} \tag{16.261}$$

For  $i = 1, \dots, n_p$ , compute the normalized importance weights by

$$\tilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{i=1}^{n_p} w_k^{(i)}} \tag{16.262}$$

R Resample  $\vec{x}_k^{(i)}$  with weights,  $\tilde{w}_k^{(i)}$ , if

$$\frac{1}{n_p \sum_{i=1}^{n_p} (\tilde{w}_k^{(i)})^2} < \tau_{eff} \quad (16.263)$$

Reset the relative importance weights as

$$\tilde{w}_k^{(i)} = \frac{1}{n_p} \quad (16.264)$$

Finally, the **EKF/SPPF state extraction step** extracts the chosen statistics on the estimated posterior distribution of  $\vec{x}_k$ . e.g., the state mean and covariance as

$$\begin{aligned} \hat{\vec{x}}_{k,PF} &= \sum_{i=1}^{n_p} \tilde{w}_k^{(i)} \vec{x}_k^{(i)} \\ \hat{P}_{k,PF} &= \sum_{i=1}^{n_p} \tilde{w}_k^{(i)} \left( \vec{x}_k^{(i)} - \hat{\vec{x}}_{k,PF} \right) \left( \vec{x}_k^{(i)} - \hat{\vec{x}}_{k,PF} \right)^T \end{aligned} \quad (16.265)$$

### Rao-Blackwellized Particle Filter

The Rao-Blackwellized particle filter provides a Bayesian framework to evaluate some of the filtering equations analytically and the others with sequential Monte Carlo sampling. This approach is preferred, not just for the computational savings, but the less variance that can be proven via the **Rao-Blackwell theorem**. The Rao-Blackwellized particle filter considers marginalized filtering of the state,  $\vec{x}_k$ , which is a conditionally linear-Gaussian, or an approximate linear-Gaussian HMM if one knows the **latent variables** in vectorized form,  $\vec{\phi}_k$ , i.e.,

$$\begin{aligned} f_{\vec{X}_k | \vec{X}_{k-1}, \vec{\Phi}_{k-1}}(\vec{x}_k | \vec{x}_{k-1}, \vec{\phi}_{k-1}) &= f_N \left( \vec{x}_k; f_{k-1}(\vec{x}_{k-1}, \vec{\phi}_k, \vec{u}_{k-1}), Q_{k-1}(\vec{\phi}_{k-1}) \right) \\ f_{\vec{Y}_k | \vec{X}_k, \vec{\Phi}_k}(\vec{y}_k | \vec{x}_k, \vec{\phi}_k) &= f_N \left( \vec{y}_k; h_k(\vec{x}_k, \vec{\phi}_k), R_k(\vec{\phi}_k) \right) \\ f_{\vec{\Phi}_k | \vec{\Phi}_{k-1}}(\vec{\phi}_k | \vec{\phi}_{k-1}) &= \text{Markov model} \end{aligned} \quad (16.266)$$

where  $\vec{y}_k$  is the corrective measurement and  $\vec{u}_{k-1}$  is the input or predictive measurement.

This corresponds to an approximation of the joint filtering distribution as

$$f_{\vec{X}_k, \vec{\Phi}_k | \vec{Y}_{1:k}}(\vec{x}_k, \vec{\phi}_k | \vec{y}_{1:k}) \approx \sum_{i=1}^{n_p} \tilde{w}_k^{(i)} \delta(\vec{\phi}_k - \vec{\phi}_k^{(i)}) f_N(\vec{x}_k; \hat{\vec{x}}_k^{(i)}, P_k^{(i)}) \quad (16.267)$$

As the joint posterior distribution at time step  $k$ , one has

$$f_{\vec{X}_{0:k}, \vec{\Phi}_{0:k} | \vec{Y}_{1:k}}(\vec{x}_{0:k}, \vec{\phi}_{0:k} | \vec{y}_{1:k}) = f_{\vec{X}_{0:k}, \vec{\Phi}_{0:k}, \vec{Y}_{1:k}}(\vec{x}_{0:k} | \vec{\phi}_{0:k}, \vec{y}_{1:k}) f_{\vec{\Phi}_{0:k} | \vec{Y}_{1:k}}(\vec{\phi}_{0:k} | \vec{y}_{1:k}) \quad (16.268)$$

where the first term on the right side can be recursively updated using a linear or nonlinear Kalman filter with a fixed trajectory  $\vec{\Phi}_{0:k} = \vec{\phi}_{0:k}$  while the second term on the right side can be recursively updated according

to Bayes rule, i.e.,

$$f_{\vec{\Phi}_{0:k}|\vec{Y}_{1:k}}(\vec{\phi}_{0:k}|\vec{y}_{1:k}) = \frac{f_{\vec{Y}_k|\vec{\Phi}_{0:k},\vec{Y}_{1:k}}(\vec{y}_k|\vec{\phi}_{0:k}, \vec{y}_{1:k-1}) f_{\vec{\Phi}_{0:k}|\vec{Y}_{1:k-1}}(\vec{\phi}_{0:k}|\vec{y}_{1:k-1})}{f_{\vec{Y}_k}(\vec{y}_k)} \quad (16.269)$$

which by the Markov property for  $\vec{\phi}_k$ , one has

$$f_{\vec{\Phi}_{0:k}|\vec{Y}_{1:k}}(\vec{\phi}_{0:k}|\vec{y}_{1:k}) = \frac{f_{\vec{Y}_k|\vec{\Phi}_{0:k},\vec{Y}_{1:k}}(\vec{y}_k|\vec{\phi}_{0:k}, \vec{y}_{1:k-1}) f_{\vec{\Phi}_k|\vec{\Phi}_{k-1}}(\vec{\phi}_k|\vec{\phi}_{k-1}) f_{\vec{\Phi}_{0:k-1}|\vec{Y}_{1:k-1}}(\vec{\phi}_{0:k-1}|\vec{y}_{1:k-1})}{f_{\vec{Y}_k}(\vec{y}_k)} \quad (16.270)$$

which can be updated via a particle filter recursion if one chooses an importance function at time step  $k$  with a recursion

$$\pi(\vec{\phi}_{0:k}|\vec{y}_{1:k}) = \pi(\vec{\phi}_k|\vec{\phi}_{0:k-1}, \vec{y}_{1:k})\pi(\vec{\phi}_{0:k-1}|\vec{y}_{1:k-1}) \quad (16.271)$$

and corresponding relative importance weight recursion for  $i = 1, \dots, n_p$  particles as

$$w_k^{(i)} = \frac{f_{\vec{Y}_k|\vec{\Phi}_{0:k},\vec{Y}_{1:k}}(\vec{y}_k|\vec{\phi}_{0:k}^{(i)}, \vec{y}_{1:k-1}) f_{\vec{\Phi}_k|\vec{\Phi}_{k-1}}(\vec{\phi}_k^{(i)}|\vec{\phi}_{k-1}^{(i)})}{\pi(\vec{\phi}_k^{(i)}|\vec{\phi}_{0:k-1}^{(i)}, \vec{y}_{1:k})} \tilde{w}_k^{(i)} \quad (16.272)$$

The **Rao-Blackwellized particle filter (RBPF)**, also known as the **marginalized particle filter (MPF)** can outlined as follows. First, the **RBPF initialization step** samples the initial  $i = 1, \dots, n_p$  particles from an initial distribution as the state estimates, latent variable vector, importance weights, and

$$\begin{aligned} \hat{x}_{0|0}^{(i)} &\sim f_{\vec{X}_0}(\vec{x}_0) \\ \vec{\phi}_{0|0}^{(i)} &\sim f_{\vec{\Phi}_0}(\vec{\phi}_0) \\ w_0^{(i)} &= f_{\vec{X}_0}(\vec{x}_{0|0}^{(i)}) \\ \tilde{w}_0^{(i)} &= \frac{w_0^{(i)}}{\sum_{i=1}^{n_p} w_0^{(i)}} \\ \text{Assign } P_{0|0}^{(i)} & \end{aligned} \quad (16.273)$$

Then, the recursive **RBPF KF-SIR step** performs the prediction, sampling, importance computation, correction, and resampling steps as

P Perform (nonlinear) Kalman filter prediction steps for  $i = 1, \dots, n_p$  prior state estimates,  $\vec{x}_{k|k-1}^{(i)}$ , and state covariances,  $P_{k|k-1}^{(i)}$  with the process model conditioned on  $\vec{\phi}_{k-1}^{(i)}$

S Sample  $n_p$  particles by

For  $i = 1, \dots, n_p$

$$\vec{\phi}_k^{(i)} \sim \pi(\vec{\phi}_k|\vec{\phi}_{0:k-1}^{(i)}, \vec{y}_{1:k}) \quad (16.274)$$

I Compute the importance weights:

For  $i = 1, \dots, n_p$

$$w_k^{(i)} = \frac{f_{\vec{Y}_k | \vec{\Phi}_{0:k}, \vec{Y}_{1:k}}(\vec{y}_k | \vec{\phi}_{0:k}^{(i)}, \vec{y}_{1:k-1}) f_{\vec{\Phi}_k | \vec{\Phi}_{k-1}}(\vec{\phi}_k^{(i)} | \vec{\phi}_{k-1}^{(i)})}{\pi(\vec{\phi}_k^{(i)} | \vec{\phi}_{0:k-1}^{(i)}, \vec{y}_{1:k})} \tilde{w}_k^{(i)} \quad (16.275)$$

- where  $f_{\vec{Y}_k | \vec{\Phi}_{0:k}, \vec{Y}_{1:k}}(\vec{y}_k | \vec{\phi}_{0:k}^{(i)}, \vec{y}_{1:k-1})$  is the marginal measurement likelihood based on the Gaussian approximation with with  $\hat{x}_{k|k-1}^{(i)}$  and  $P_{k|k-1}^{(i)}$ .

and normalize the importance weights by:

$$\tilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{i=1}^{n_p} w_k^{(i)}} \quad (16.276)$$

C Perform (nonlinear) Kalman filter correction steps for  $i = 1, \dots, n_p$  posterior state estimates,  $\vec{x}_{k|k}^{(i)}$ , and state covariances,  $P_{k|k}^{(i)}$  with the measurement model conditioned on  $\vec{\phi}_k^{(i)}$

R Resample  $\vec{\phi}_k^{(i)}$  with weights,  $\tilde{w}_k^{(i)}$  if

$$\frac{1}{n_p \sum_{i=1}^{n_p} (\tilde{w}_k^{(i)})^2} < \tau_{eff} \quad (16.277)$$

Reset the relative importance weights as

$$\tilde{w}_k^{(i)} = \frac{1}{n_p} \quad (16.278)$$

Finally, the **RBPF state extraction step** extracts the chosen statistics on the estimated joint posterior distribution of  $\vec{X}_k, \vec{\Phi}_k$ . e.g., the state mean, state covariance, latent variables mean, and latent variables covariance as

$$\begin{aligned} \hat{x}_{k,PF} &= \sum_{i=1}^{n_p} \tilde{w}_k^{(i)} \vec{x}_k^{(i)} \\ P_{k,PF} &= \sum_{i=1}^{n_p} \tilde{w}_k^{(i)} P_k^{(i)} \\ \hat{\phi}_{k,PF} &= \sum_{i=1}^{n_p} \tilde{w}_k^{(i)} \vec{\phi}_k^{(i)} \\ \hat{P}_{k,PF} &= \sum_{i=1}^{n_p} \tilde{w}_k^{(i)} \left( \vec{\phi}_k^{(i)} - \hat{\phi}_{k,PF} \right) \left( \vec{\phi}_k^{(i)} - \hat{\phi}_{k,PF} \right)^T \end{aligned} \quad (16.279)$$

## Particle Smoothers

While the same SIR method can provide the state smoothing distribution by sequentially updating the importance weights and resampling the entire state trajectories instead of simply the current states. However, such an approach typically results in a degenerate distribution for the  $k^{\text{th}}$  time step when  $k \ll N$ . Thus, two alternative particle smoothers are summarized here.

The **backward-simulation particle smoother (BS-PS)** updates the state filtering distributions approximated as a weighted set of particles,  $w_k^{(i)}$  and  $\vec{x}_k^{(i)}$  for  $i = 1, \dots, n_p$  and  $k = 1, \dots, N$ . The initialization step by resampling  $\vec{x}_N^{(i)}$  as  $\tilde{x}_N^{(i)}$ . Then, one recursively approximates the state smoothing distribution for  $k = N - 1, \dots, 0$  by first computing the relative importance weights as

$$w_{k|k+1}^{(i)} = f_{\vec{X}_{k+1}|\vec{X}_k}(\tilde{x}_{k+1}^{(i)} | \vec{x}_k^{(i)}) w_k^{(i)} \quad (16.280)$$

and the normalized importance weights as

$$\tilde{w}_{k|k+1}^{(i)} = \frac{w_{k|k+1}^{(i)}}{\sum_{i=1}^{n_p} w_{k|k+1}^{(i)}} \quad (16.281)$$

and resamples  $\tilde{x}_k^{(i)}$  with weights  $\tilde{w}_{k|k+1}^{(i)}$ .

The **reweighted particle smoother (RPS)** is implemented by recomputing the normalized importance weights for  $k = N - 1, \dots, 0$

$$\tilde{w}_{k|N}^{(i)} = \sum_{j=1}^{n_p} \tilde{w}_{k+1|N}^{(j)} \frac{\tilde{w}_k^{(i)} f_{\vec{X}_{k+1}|\vec{X}_k}(\vec{x}_{k+1}^{(j)} | \vec{x}_k^{(i)})}{\sum_{\ell=1}^{n_p} \tilde{w}_{k+1|N}^{(\ell)} f_{\vec{X}_{k+1}|\vec{X}_k}(\vec{x}_{k+1}^{(\ell)} | \vec{x}_k^{(\ell)})} \quad (16.282)$$

where one defines  $\tilde{w}_{N|N}^{(i)} = \tilde{w}_N^{(i)}$  for  $i = 1, \dots, n_p$

## References

For more information, please refer to the following

- Sarkka, S., “7.3 Sequential importance sampling,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 120-123
- Sarkka, S., “7.4 Sequential importance resampling,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 123-129
- Sarkka, S., “7.5 Rao-Blackwellized particle filter,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 129-132
- Sarkka, S., “11.1 SIR particle smoother,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 165-167
- Sarkka, S., “11.2 Backward-simulation particle smoother,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 165-167

- Sarkka, S., “11.3 SIR particle smoother,” in *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013, pp. 165-167
- Simon, D., “15.3 Implementation Issues,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 469-480
- Simon, D., “15.2 Particle filtering,” in *Optimal State Estimation*, 1st Ed., John Wiley & Sons, 2006, pp. 466-469

---

# Stochastic Motion Models

## 17.1 Uncoupled-Coordinate Stochastic Motion Models

This section will consider uncoupled-coordinate motion models where the *random* control input is typically taken to be the target acceleration which is the most natural choice from Newtonian dynamics as acceleration is directly related to the net force acting on the target. However, for some targets, particularly agile targets, it is more convenient to use the target jerk as the control input. Although an acceleration model can always be obtained from a jerk model by integration, one may choose a jerk model if the target motion model would be simpler. Regardless of this choice, one typically models the *unknown* control input as one of three random process models: a white noise process model, a Markov process model, and a semi-Markov Jump process model. In this section, one each dynamics model is written for a single coordinate state vector and noise vector which would be stacked for each uncoupled-coordinate in application, e.g. full two-dimensional or three-dimensional vehicle motion. These models apply to any particular vehicle tracking application.

Each model will be represented for both a continuous-time LTI state-space model, i.e.

$$\dot{\vec{x}}(t) = A\vec{x}(t) + B\vec{w}(t) \quad (17.1)$$

as well as a discrete-time LTI state-space model, i.e.

$$\vec{x}_{k+1} = F\vec{x}_k + G\vec{w}_k \quad (17.2)$$

obtained using discretization techniques with a time interval,  $\Delta t = t(k+1) - t(k)$ . It should be noted that the continuous-time model is more accurate than its discrete-time counterpart for most practical situations since a target moves continuously over time, but the discrete-time model is simpler to implement, though higher fidelity numerical integration can be done.

As an aside, as any continuous target trajectory can be approximated by an  $n^{\text{th}}$  polynomial to any arbitrary accuracy, the highest fidelity uncoupled-coordinate model would be a continuous-time  $n^{\text{th}}$ -order

**general polynomial stochastic motion model** for the target trajectory, i.e.

$$\vec{x}(t) = \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} = \begin{bmatrix} a_0 & a_1 & \cdots & a_n \\ b_0 & b_1 & \cdots & b_n \\ c_0 & c_1 & \cdots & c_n \end{bmatrix} \begin{bmatrix} 1 \\ t \\ t^2 \\ \vdots \\ t^n \end{bmatrix} + \begin{bmatrix} w_x(t) \\ w_y(t) \\ w_z(t) \end{bmatrix} \quad (17.3)$$

However, such models are usually good for data-fitting/smoothing, not the prediction/correction of target tracking. However, one can view most of the models here as special cases of this general polynomial model with *different* noise models for  $w(t)$ .

### White Noise Motion Models

The simplest model is the **white noise acceleration model** which uses a state vector as

$$\vec{x} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix} \quad (17.4)$$

and can be written in continuous-time as

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad (17.5)$$

and

$$B = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (17.6)$$

and the zero-mean white process noise,  $w(t)$ , models *random* acceleration with variance  $\sigma_w^2$ .

The white noise acceleration model is typically written in discrete-time with

$$F = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \quad (17.7)$$

and

$$G = \begin{bmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{bmatrix} \quad (17.8)$$

where the zero-mean white noise sequence  $w_k$  multiplied by the noise gain has a covariance

$$\text{Cov}(Gw_k) = Q = \sigma_w^2 \begin{bmatrix} \frac{1}{4}\Delta t^4 & \frac{1}{2}\Delta t^3 \\ \frac{1}{2}\Delta t^3 & \Delta t^2 \end{bmatrix} \quad (17.9)$$

It should be noted that if  $\sigma_w^2$  is relatively small, this is also known as the “nearly” **constant-velocity (CV) model**. The effectiveness of this model depends greatly on the determination of a suitable  $\sigma_w^2$  for the target which is often unknown. Thus, this parameter is often estimated using adaptive Kalman filter (AKF) techniques for the process noise covariance, e.g. based on the measurement residual statistics.

The second simplest model is the **white noise jerk model**, also known as the **Wiener process acceleration model**, which uses a state vector as

$$\vec{x}(t) = \begin{bmatrix} x \\ \dot{x} \\ \ddot{x} \end{bmatrix} \quad (17.10)$$

and can be written in continuous-time as

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad (17.11)$$

and

$$B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (17.12)$$

and the zero-mean white process noise,  $w(t)$ , models *random jerk* with auto-correlation  $\sigma_w^2 \delta(h)$ .

The white noise jerk model is typically written in discrete-time with

$$F = \begin{bmatrix} 1 & \Delta t & \frac{1}{2}\Delta t^2 \\ 0 & 1 & \Delta t \\ 0 & 0 & 0 \end{bmatrix} \quad (17.13)$$

and

$$G = I_{3 \times 3} \quad (17.14)$$

where the zero-mean white process noise  $w_k$  has covariance

$$\text{Cov}(Gw_k) = Q = \sigma_w^2 \begin{bmatrix} \frac{1}{20}\Delta t^5 & \frac{1}{8}\Delta t^4 & \frac{1}{6}\Delta t^3 \\ \frac{1}{8}\Delta t^4 & \frac{1}{3}\Delta t^3 & \frac{1}{2}\Delta t^2 \\ \frac{1}{6}\Delta t^3 & \frac{1}{2}\Delta t^2 & \Delta t \end{bmatrix} \quad (17.15)$$

It should be noted that if  $\sigma_w^2$  is relatively small, this is also known as the “nearly” **constant-acceleration (CA) model**.

An alternative to this representation is the **Wiener sequence acceleration model** which changes the discrete-time noise gain matrix to

$$G = \begin{bmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \\ 1 \end{bmatrix} \quad (17.16)$$

and the noise covariance to

$$\text{Cov}(Gw_k) = Q = \sigma_w^2 \begin{bmatrix} \frac{1}{4}\Delta t^4 & \frac{1}{2}\Delta t^3 & \frac{1}{2}\Delta t^2 \\ \frac{1}{2}\Delta t^3 & \frac{1}{2}\Delta t^2 & \Delta t \\ \frac{1}{2}\Delta t^2 & \Delta t & 1 \end{bmatrix} \quad (17.17)$$

## Markov Process Motion Models

The white noise models presented previously are simple in form, but crude approximations as actual maneuvering targets seldom have “nearly” constant velocities or accelerations that are uncoupled across coordinate directions. For example, the Wiener process model assumes the acceleration increment is independent across different time intervals which is hardly justifiable, except for its mathematical simplicity. Thus, a more realistic acceleration/jerk model would depend on *at least* its previous value. Thus, one typically considers the use of Markov process models whenever white noise models are not good enough.

The **Singer acceleration model** uses a zero-mean, first-order Markov process acceleration model for the acceleration with auto-correlation

$$R_a(\Delta t) = \sigma_w^2 \exp\left(\frac{|\Delta t|}{\tau_m}\right) \quad (17.18)$$

where  $\tau_m$  is **maneuver time constant** and  $\sigma^2$  is the instantaneous variance of the acceleration as a random variable.

Thus, a full model defines the state vector as

$$\vec{x} = \begin{bmatrix} x \\ \dot{x} \\ \ddot{x} \end{bmatrix} \quad (17.19)$$

and can be written in continuous-time as

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -\frac{1}{\tau_m} \end{bmatrix} \quad (17.20)$$

and

$$B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (17.21)$$

and the zero-mean white noise driving function,  $w(t)$ , models *random* acceleration with variance  $2\frac{\sigma_w^2}{\tau}$ .

The Singer acceleration model is typically written in discrete-time with

$$F = \begin{bmatrix} 1 & \Delta t & \tau_m^2 \left( -1 + \frac{\Delta t}{\tau_m} + \exp\left(-\frac{\Delta t}{\tau_m}\right) \right) \\ 0 & 1 & \tau_m \left[ 1 - \exp\left(-\frac{\Delta t}{\tau_m}\right) \right] \\ 0 & 0 & \exp\left(-\frac{\Delta t}{\tau_m}\right) \end{bmatrix} \quad (17.22)$$

and

$$G = I_{3 \times 3} \quad (17.23)$$

where the zero-mean white process noise  $w_k$  has covariance

$$\text{Cov}(Gw_k) = Q = 2\frac{\sigma_w^2}{\tau_m} \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{12} & Q_{22} & Q_{23} \\ Q_{13} & Q_{23} & Q_{33} \end{bmatrix} \quad (17.24)$$

where

$$Q_{11} = \frac{1}{2}\tau_m^5 \left( 1 - \exp\left(-2\frac{\Delta t}{\tau_m}\right) + 2\frac{\Delta t}{\tau_m} - 2\frac{\Delta t^2}{\tau_m^2} + \frac{2}{3}\frac{\Delta t^3}{\tau_m^3} - 4\frac{\Delta t}{\tau_m} \exp\left(-\frac{\Delta t}{\tau_m}\right) \right) \quad (17.25)$$

$$Q_{12} = \frac{1}{2}\tau_m^4 \left( \exp\left(-2\frac{\Delta t}{\tau_m}\right) + 1 - 2\exp\left(-\frac{\Delta t}{\tau_m}\right) + 2\frac{\Delta t}{\tau_m} \exp\left(-\frac{\Delta t}{\tau_m}\right) - 2\frac{\Delta t}{\tau_m} + \frac{\Delta t^2}{\tau_m^2} \right) \quad (17.26)$$

$$Q_{13} = \frac{1}{2}\tau_m^3 \left( 1 - \exp\left(-2\frac{\Delta t}{\tau_m}\right) - 2\frac{\Delta t}{\tau_m} \exp\left(-\frac{\Delta t}{\tau_m}\right) \right) \quad (17.27)$$

$$Q_{22} = \frac{1}{2}\tau_m^3 \left( 4 \exp\left(-\frac{\Delta t}{\tau_m}\right) - 3 - \exp\left(-2\frac{\Delta t}{\tau_m}\right) + 2\frac{\Delta t}{\tau_m} \right) \quad (17.28)$$

$$Q_{23} = \frac{1}{2}\tau_m^2 \left( \exp\left(-2\frac{\Delta t}{\tau_m}\right) + 1 - 2 \exp\left(-\frac{\Delta t}{\tau_m}\right) \right) \quad (17.29)$$

$$Q_{33} = \frac{1}{2}\tau_m \left( 1 - \exp\left(-2\frac{\Delta t}{\tau_m}\right) \right) \quad (17.30)$$

The effectiveness of the Singer acceleration model relies on an accurate determination of the parameters:  $\tau_m$  and  $\sigma^2$ . For aircraft,  $\tau \approx 60$  for a lazy turn and  $\tau \approx 10$  to 20 for an evasive maneuver. The parameter  $\sigma^2$  often is designed with a **ternary-uniform mixture distribution** defined as: the target may move without acceleration with probability  $P_0$ , the target may accelerate or decelerate at a maximum/minimum rates  $\pm a_{max}$  with equal probability  $P_{max}$ , and the target may accelerate or decelerate at a rate uniformly distributed over  $(-a_{max}, a_{max})$ . Then, one can show

$$\sigma^2 = \frac{a_{max}^2}{3}(1 + 4P_{max} - P_0) \quad (17.31)$$

It should also be noted that as  $\tau_m$  increases, relative to  $\Delta t$ , the Singer acceleration model reduces to the white noise jerk model and as the  $\tau_m$  decreases, relative to  $\Delta t$ , the Singer model acceleration model reduces to the white noise acceleration model. The Singer acceleration model is a standard model for target maneuvers as it was the first model that characterized the unknown target acceleration as a time-correlated stochastic process and served as the basis for the further development of target motion models.

In essence, the Singer acceleration model is a prior model as it does not use online information about the target maneuver. As a consequence, the Singer model is symmetric as the ternary-uniform mixture distribution is symmetric which is the main shortcoming of the Singer model, i.e. the target acceleration has *zero-mean* at any moment. Notably, this is a very reasonable *a priori* assumption, but other methods may use on-line information, e.g. the mean-adaptive acceleration model, also known as the “current” model. If the target acceleration is oscillatory, a zero-mean, second-order Markov process acceleration model will serve much better than the Singer acceleration model. Furthermore, it should be noted that a Singer jerk model and non-zero-mean Singer jerk model can also be derived in a similar fashion to the Singer acceleration models.

### Semi-Markov Jump Process Motion Models

In practice, many target maneuvers involve an acceleration with nonzero-mean, that may be reasonably assumed piecewise-constant. However, neither the time intervals over which the acceleration mean is piecewise-constant nor the corresponding constant levels of the nonzero mean are known to a tracker. One of the simplest models for such a piecewise-constant random process is a **semi-Markov jump process**. It differs from a Markov jump process only in that the time it stays in a mode, i.e. its **sojourn-time**, is random, i.e. it jumps at random times, while a Markov jump process has jumps at deterministic times. For example, consider the unknown input  $u(t)$  to be the non-zero mean of the acceleration with *possible* mean quantized

into  $n$  known levels,  $\bar{a}_1, \dots, \bar{a}_n$ . Then, the sequence  $u(t_0), \dots, u(t_k), \dots$  among these level is a semi-Markov process with known transition probabilities

$$\Pr\{u(t_k) = \bar{a}_j | u(t_{k-1}) = \bar{a}_i\} \quad \text{for } i, j = 1, \dots, n \quad (17.32)$$

and the sojourn-time PMFs are

$$P_{ij}(t) = \Pr(\tau_{ij} \leq t) \quad (17.33)$$

where  $\tau_{ij} = t_k - t_{k-1}$  is the sojourn-time in mode  $\bar{a}_i$  before it jumps to mode  $\bar{a}_j$ .

Then, one such model is the **Markov jump-mean acceleration model** which models the target acceleration as the Singer acceleration  $(\tilde{a})(t)$ , with *non-zero mean* that is a semi-Markov process, i.e.

$$\ddot{x}(t) = -\beta \dot{x}(t) + u(t) + (\tilde{a})(t) \quad (17.34)$$

where  $\beta$  is a drag coefficient. Thus, this model defines the state vector as

$$\vec{x} = \begin{bmatrix} x \\ \dot{x} \\ \ddot{x} \end{bmatrix} \quad (17.35)$$

and can be written as a continuous-time state-space model as

$$\dot{\vec{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -\beta & 1 \\ 0 & 0 & -\frac{1}{\tau_m} \end{bmatrix} \vec{x} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u(t) + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} w(t) \quad (17.36)$$

where the zero-mean white noise driving function,  $\vec{w}(t)$ , models *random* acceleration with variance  $2\frac{\sigma_w^2}{\tau}$ . The unknown acceleration mean is typically estimated by a weighted sum of the quantization levels, i.e.

$$\hat{u}(t) = \sum_{i=1}^n \bar{a}_i \Pr\{u(t) = \bar{a}_i | y(s), s \leq t\} \quad (17.37)$$

where  $\vec{y}(s)$ ,  $s \leq t$  denotes all measurements up to time  $s$ . Three important issues associated with this model are the design of the input modes, the transition probabilities, and the sojourn-time.

## 17.2 Turning Stochastic Motion Models

Coupled-coordinate models are based on the target kinematics, in contrast to those of uncoupled-coordinate motion based on random processes due to the fact random processes are more natural for modeling time correlation than kinematics for modeling spatial correlation. Coordinate-coupled target models are highly dependent on the choice of the state components, e.g. to describe two-dimensional motion, the state vector is at least four-dimensional. The choice of the state components and kinematic model is a non-trivial with consideration given to the expected target trajectory with random perturbation, the accuracy of the unavoidable approximations in the model, and the sensor coordinate system. This section will focus on kinematic-based target models for maneuvering flight vehicles, in particular, **turning motion models** which can be applied to both ground, surface, and air vehicles.

As an aside, in some applications, the decision for using a non-maneuvering or maneuvering motion model can be made *a priori* or the problem can be posed to the target tracking system as it operates in real-time. This decision may be performed by an **adaptive tracker** which estimates the unknown model parameters, often noise, alongside the state. Secondly, the system may use a **maneuver detector** which attempts to determine the **onset time** of any maneuvers, e.g. using  $\chi^2$ -squared tests or GLRTs. Alternatively, the system may also use set-based methods known as **multi-model target tracking (MM-TT)** where one uses the full model set of all considered maneuvers for the vehicle(s) to be tracked in the target tracking algorithm.

## Two-Dimensional Coordinated-Turn Models

For two-dimensional motion, the **standard curvilinear motion model** can be written from kinematics as

$$\begin{aligned}\dot{x}(t) &= v(t) \cos \psi(t) \\ \dot{y}(t) &= v(t) \sin \psi(t) \\ \dot{v}(t) &= a_t(t) \\ \dot{\psi}(t) &= \frac{a_n(t)}{v(t)}\end{aligned}\tag{17.38}$$

where  $(x, y)$  are the target position in Cartesian coordinates,  $v$  is the speed,  $\psi$  is the heading angle,  $a_t$  is the target tangential acceleration, also known as the “along-track” acceleration, and  $a_n$  is the target normal acceleration, also known as the “cross-track” acceleration. These two accelerations are then treated as the unknown inputs to the target motion model.

One important variant of this motion model is the **bicycle motion model** which models  $a_n(t)$  using a single steering wheel at the front of the vehicle

$$a_n^2 = \frac{v^2(t)}{R_c} = \frac{v^2(t)}{L} \tan \gamma\tag{17.39}$$

where  $R_c$  is the radius of curvature of a turn,  $L$  is the length from the front wheel to the rear wheel, also known as the **wheel base**, and  $\gamma_s$  is the **steering angle**. This results in the dynamics

$$\begin{aligned}\dot{x}(t) &= v(t) \cos \psi(t) \\ \dot{y}(t) &= v(t) \sin \psi(t) \\ \dot{v}(t) &= a_t(t) \\ \dot{\psi}(t) &= \frac{v(t)}{L} \tan \gamma_s\end{aligned}\tag{17.40}$$

where  $a_t$  and  $\gamma_s$  are the unknown inputs to the target motion model.

Three special cases exist for this general curvilinear motion. If  $a_t = a_n = 0$ , then one has rectilinear, constant velocity (CV) motion. If  $a_t \neq 0$ ,  $a_n = 0$ , then one has rectilinear, acceleration motion, e.g. if  $a_t = \text{constant}$  then constant acceleration (CA) motion. For both cases, previous uncoupled-coordinate motion models apply. However, if  $a_t = 0$  and  $a_n \neq 0$ , then one has a **coordinated-turn (CT) motion models**, i.e. constant-speed, constant-turn-rate motion and can be simplified to two simpler models specified in terms of the **turn-rate**,  $\omega = \dot{\phi}$ , which may be known or unknown.

The **coordinated-turn with known turn-rate (CT-KTR) model** assumes  $\omega$  is known, uses a four-dimensional state vector as

$$\vec{x} = \begin{bmatrix} x \\ y \\ \dot{x} \\ \dot{y} \end{bmatrix} \quad (17.41)$$

and can be written in as a continuous-time LTI state-space model, i.e.

$$\dot{\vec{x}} = A_{ct}(\omega) \vec{x} + \vec{w}(t) \quad (17.42)$$

where

$$A_{ct}(\omega) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \omega & 0 \\ 0 & 0 & 0 & -\omega \end{bmatrix} \quad (17.43)$$

The CT-KTR model can be discretized as

$$\vec{x}_{k+1} = F_{ct}(\omega) \vec{x}_k + \vec{w}_k \quad (17.44)$$

$$F_{ct} = \begin{bmatrix} 1 & 0 & \frac{\sin(\omega\Delta t)}{\omega} & -\frac{1-\cos(\omega\Delta t)}{\omega} \\ 0 & 1 & \frac{1-\cos(\omega\Delta t)}{\omega} & \frac{\sin(\omega\Delta t)}{\omega} \\ 0 & 0 & \cos(\omega\Delta t) & -\sin(\omega\Delta t) \\ 0 & 0 & \sin(\omega\Delta t) & \cos(\omega\Delta t) \end{bmatrix} \quad (17.45)$$

or approximated for  $\omega\Delta t \approx 0$  by

$$F_{ct} \approx \begin{bmatrix} 1 & 0 & \Delta t & -\frac{1}{2}\omega^2\Delta t^2 \\ 0 & 1 & \frac{1}{2}\omega^2\Delta t^2 & \Delta t \\ 0 & 0 & 1 - \frac{1}{2}\omega^2\Delta t^2 & -\omega\Delta t \\ 0 & 0 & \omega\Delta t & 1 - \frac{1}{2}\omega^2\Delta t^2 \end{bmatrix} \quad (17.46)$$

and  $\vec{w}(t)$  is zero-mean AWGN used to perturb the model from the ideal CT motion by some amount. This is only a decent model if the constant turn-rate is known approximately *a priori*. However, this model is often utilized with adaptive approaches to learn  $\omega$  and with multiple-model methods using Markov or semi-Markov sequences for different discrete values of  $\omega$ .

The **coordinated-turn with unknown turn-rate (CT-UTR) models** assume that  $\omega$  is driven by a stochastic process, typically either a Wiener process

$$\begin{aligned} \dot{\omega}(t) &= w_\omega(t) \\ \omega_{k+1} &= \omega_k + w_{\omega,k} \end{aligned} \quad (17.47)$$

or a first-order Markov process

$$\begin{aligned} \dot{\omega}(t) &= -\frac{1}{\tau_\omega} \omega(t) + w_\omega(t) \\ \omega_{k+1} &= \exp\left(-\frac{\Delta t}{\tau_\omega}\right) \omega_k + w_{\omega,k} \end{aligned} \quad (17.48)$$

where  $\tau_\omega$  is the correlation time constant for the turn-rate and  $\vec{w}(t)$  is again some zero-mean AWGN used to perturb the model from the ideal CT motion by some amount.

Thus, the five-dimensional state vector is

$$\vec{x} = \begin{bmatrix} x \\ y \\ \dot{x} \\ \dot{y} \\ \omega \end{bmatrix} \quad (17.49)$$

and the noise vector is

$$\vec{w} = \begin{bmatrix} w_{\dot{x}} \\ w_{\dot{y}} \\ w_\omega \end{bmatrix} \quad (17.50)$$

The CT-UTR model can be written as a continuous-time LPV state-space model, i.e.

$$\dot{\vec{x}} = f(\vec{x}(t), \vec{w}(t)) = \begin{bmatrix} A_{ct}(\omega) & 0 \\ 0_{1 \times 4} & -\alpha \end{bmatrix} \vec{x} + L \vec{w}(t) \quad (17.51)$$

where  $\alpha = 0$  or  $\frac{1}{\tau_\omega}$ , the noise gain matrix is

$$L = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (17.52)$$

which can be discretized as

$$\vec{x}_{k+1} = f(\vec{x}_k, \vec{w}_k) = \begin{bmatrix} F_{ct}(\omega) & 0 \\ 0_{1 \times 4} & \beta \end{bmatrix} \vec{x}_k + \begin{bmatrix} \frac{1}{2}\Delta t^2 & 0 & 0 \\ 0 & \frac{1}{2}\Delta t^2 & 0 \\ \Delta t & 0 & 0 \\ 0 & \Delta t & 0 \\ 0 & 0 & 1 \end{bmatrix} \vec{w}_k \quad (17.53)$$

where  $\beta = 1$  or  $\exp\left(-\frac{\Delta t}{\tau_\omega}\right)$ . As this is a nonlinear equation in  $\omega$ , one must use a nonlinear filter with the CT-UTR model, e.g. an EKF. In this case, the Jacobians of the discrete-time model are notably

$$\frac{\partial f(\vec{x}, \vec{w})}{\partial \vec{x}} = \begin{bmatrix} 1 & 0 & \frac{\sin(\omega\Delta t)}{\omega} & -\frac{1-\cos(\omega\Delta t)}{\omega} & \left(\frac{\omega\Delta t \cos(\omega\Delta t)-\sin(\omega\Delta t)}{\omega^2}\right)\dot{x} - \left(\frac{\omega\Delta t \sin(\omega\Delta t)-1+\cos(\omega\Delta t)}{\omega^2}\right)\dot{y} \\ 0 & 1 & \frac{1-\cos(\omega\Delta t)}{\omega} & \frac{\sin(\omega\Delta t)}{\omega} & \left(\frac{\omega\Delta t \sin(\omega\Delta t)-1+\cos(\omega\Delta t)}{\omega^2}\right)\dot{x} + \left(\frac{\omega\Delta t \cos(\omega\Delta t)-\sin(\omega\Delta t)}{\omega^2}\right)\dot{y} \\ 0 & 0 & \cos(\omega\Delta t) & -\sin(\omega\Delta t) & -\Delta t \sin(\omega\Delta t)\dot{x} - \Delta t \cos(\omega\Delta t)\dot{y} \\ 0 & 0 & \sin(\omega\Delta t) & \cos(\omega\Delta t) & \Delta t \cos(\omega\Delta t)\dot{x} - \Delta t \sin(\omega\Delta t)\dot{y} \\ 0 & 0 & 0 & 0 & \beta \end{bmatrix} \quad (17.54)$$

and

$$\frac{\partial f(\vec{x}, \vec{w})}{\partial \vec{w}} = L_k = \begin{bmatrix} \frac{1}{2}\Delta t^2 & 0 & 0 \\ 0 & \frac{1}{2}\Delta t^2 & 0 \\ \Delta t & 0 & 0 \\ 0 & \Delta t & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (17.55)$$

### Three-Dimensional Turn Models

The basic equation for three-dimensional turn motion, relates the angular velocity,  $\vec{\omega}$ , to the vehicle velocity and acceleration vectors as

$$\vec{\omega} = \frac{\dot{\vec{x}} \times \ddot{\vec{x}}}{\dot{\vec{x}}^T \dot{\vec{x}}} \quad (17.56)$$

where for a three-dimensional fixed-center constant-angular-velocity (CAV) model, one has

$$\ddot{\vec{x}} = \vec{\omega} \times \dot{\vec{x}} \quad (17.57)$$

Thus, defining the derivative of acceleration as

$$\ddot{\vec{x}} = -\|\vec{\omega}\|_2^2 \dot{\vec{x}} = \frac{\|\dot{\vec{x}} \times \ddot{\vec{x}}\|}{\dot{\vec{x}}^T \dot{\vec{x}}} \dot{\vec{x}} \quad (17.58)$$

Thus, the three-dimensional **fixed-center constant-turn-rate (FC-CTR) model** is a second-order Markov process

$$\ddot{\vec{x}} = -\omega^2 \dot{\vec{x}} + \vec{w} \quad (17.59)$$

where  $\omega = \|\vec{\omega}\|_2$  and  $\vec{w}$  is white noise with variance  $\sigma_\omega^2 I_{3 \times 3}$ .

The FC-CTR has *for each coordinate* a state vector as

$$\vec{x} = \begin{bmatrix} x \\ \dot{x} \\ \ddot{x} \end{bmatrix} \quad (17.60)$$

and can be written as a continuous-time LTI state-space model as

$$\dot{\vec{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -\omega^2 & 0 \end{bmatrix} \vec{x} + \vec{w}(t) \quad (17.61)$$

which can be discretized as

$$\vec{x}_{k+1} = \begin{bmatrix} 1 & \frac{\sin(\omega\Delta t)}{\omega} & \frac{1-\cos(\omega\Delta t)}{\omega^2} \\ 0 & \cos(\omega\Delta t) & \frac{\sin(\omega\Delta t)}{\omega} \\ 0 & -\omega \sin(\omega\Delta t) & \cos(\omega\Delta t) \end{bmatrix} \vec{x}_k + \begin{bmatrix} \frac{\omega\Delta t - \sin(\omega\Delta t)}{\omega^3} \\ \frac{1-\cos(\omega\Delta t)}{\omega^2} \\ \frac{\sin(\omega\Delta t)}{\omega} \end{bmatrix} w_k \quad (17.62)$$

where  $\text{Cov}(w_k) = \sigma_\omega^2 I$ . Thus, the coordinates are coupled only through the common  $\omega$ , which must be estimated in this form. Note that for a target with constant speed, the velocity and acceleration vectors are orthogonal, i.e.  $\ddot{\vec{x}}^T \dot{\vec{x}} = 0$ , then

$$\omega = \frac{\|\ddot{\vec{x}}\|_2}{\|\dot{\vec{x}}\|_2} \quad (17.63)$$

which may be imposed as a constraint on the potentially unknown  $\vec{\omega}$ .

Alternatively, the **constant-angular-velocity (CAV) model** uses a state vector

$$\vec{x} = \begin{bmatrix} \vec{x} \\ \dot{\vec{x}} \\ \vec{\omega} \end{bmatrix} \quad (17.64)$$

and can be written as a continuous-time nonlinear state-space model as

$$\dot{\vec{x}}(t) = \begin{bmatrix} 0 & I_{3 \times 3} & 0 \\ 0 & [\vec{\omega}]^\times & 0 \\ 0 & 0 & 0 \end{bmatrix} \vec{x}(t) + \vec{w}(t) \quad (17.65)$$

where the angular velocity can be estimated as part of the state vector. Note that if  $\vec{\omega} = [0 \ 0 \ \omega]^T$ , then this model reduces to the CT-KTR model.

The **nearly constant-turn-rate model (N-CTR) model** assumes the acceleration is a second-order Gauss-Markov process with state-dependent coefficients, i.e.

$$\ddot{\vec{x}} = -2\alpha \dot{\vec{x}} - (\alpha^2 + \omega^2) \vec{x} + \vec{w} \quad (17.66)$$

where  $\omega = \|\vec{\omega}\|_2$  is the turn-rate,

$$\alpha = -\frac{\dot{\vec{x}}^T \ddot{\vec{x}}}{\dot{\vec{x}}^T \dot{\vec{x}}} \quad (17.67)$$

is a normalized target drag term and

$$\vec{w} = \vec{\omega}_B \times \vec{x} + \vec{x}_B \quad (17.68)$$

reflects the effect of the forces and moments applied to the target, which are unknown, and taken as AWGN with variance  $\sigma_\omega^2 I$ .

The N-CTR has *for each coordinate* a state vector as

$$\vec{x} = \begin{bmatrix} x \\ \dot{x} \\ \ddot{x} \end{bmatrix} \quad (17.69)$$

and can be written as a continuous-time LTI state-space model as

$$\dot{\vec{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -(\alpha^2 + \omega^2) & -2\alpha \end{bmatrix} \vec{x} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} w(t) \quad (17.70)$$

which can be discretized as

$$\vec{x}_{k+1} = F \vec{x}_k + G w_k \quad (17.71)$$

where

$$F = \begin{bmatrix} 1 & \frac{2\alpha\omega - \exp(-\alpha\Delta t)(2\alpha\omega \cos(\omega\Delta t) + (\alpha^2 - \omega^2)\sin(\omega\Delta t))}{\omega(\alpha^2 + \omega^2)} & \frac{\omega - \exp(-\alpha\Delta t)(\omega \cos(\omega\Delta t) + \alpha \sin(\omega\Delta t))}{\omega(\alpha^2 + \omega^2)} \\ 0 & \frac{\exp(-\alpha\Delta t)(\omega \cos(\omega\Delta t) + \alpha \sin(\omega\Delta t))}{\omega(\alpha^2 + \omega^2)} & \frac{\exp(-\alpha\Delta t) \sin(\omega\Delta t)}{\omega(\alpha^2 + \omega^2)} \\ 0 & -\frac{(\alpha^2 + \omega^2)\exp(-\alpha\Delta t)\sin(\omega\Delta t)}{\omega} & \frac{\exp(-\alpha\Delta t)(\omega \cos(\omega\Delta t) - \alpha \sin(\omega\Delta t))}{\omega} \end{bmatrix} \quad (17.72)$$

$$G = \begin{bmatrix} \frac{\omega(-2\alpha + (\alpha^2 + \omega^2)\Delta t) + \exp(-\alpha\Delta t)(2\alpha\omega \cos(\omega\Delta t) + (\alpha^2 - \omega^2)\sin(\omega\Delta t))}{\omega(\alpha^2 + \omega^2)^2} \\ \frac{\omega - \exp(-\alpha\Delta t)(\omega \cos(\omega\Delta t) + \alpha \sin(\omega\Delta t))}{\omega(\alpha^2 + \omega^2)} \\ \frac{\exp(-\alpha\Delta t) \sin(\omega\Delta t)}{\omega} \end{bmatrix} \quad (17.73)$$

and  $\text{Cov}(w_k) = \sigma_w^2$ . Thus, the coordinates are coupled only through the common  $\omega$  and  $\alpha$ .

Lastly, the **non-constant-speed coordinated-turn (NCS-CT) model** assumes the target acceleration is

$$\ddot{\vec{x}} = \vec{\omega} \times (\ddot{\vec{x}} - \vec{g}) + C_{N \leftarrow B}(\dot{\vec{x}}, \ddot{\vec{x}}) \vec{u} \quad (17.74)$$

where  $\vec{g}$  is the gravity vector and the direction cosine matrix is formed as

$$C_{N \leftarrow B}(\dot{\vec{x}}, \ddot{\vec{x}}) = \begin{bmatrix} \frac{\dot{\vec{x}}}{\|\dot{\vec{x}}\|_2} & \frac{(\ddot{\vec{x}} - \vec{g}) \times \dot{\vec{x}}}{\|(\ddot{\vec{x}} - \vec{g}) \times \dot{\vec{x}}\|_2} & \frac{\dot{\vec{x}}}{\|\dot{\vec{x}}\|_2} \times \frac{(\ddot{\vec{x}} - \vec{g}) \times \dot{\vec{x}}}{\|(\ddot{\vec{x}} - \vec{g}) \times \dot{\vec{x}}\|_2} \end{bmatrix} \quad (17.75)$$

and  $\vec{u}$  is a vector-valued zero-mean, first-order Gauss-Markov process that models random perturbations in the acceleration with respect to the body frame. The time-correlated perturbations accounts for the unmodeled motions and is modeled in the body frame to be more precise. This, however, necessitates a frame transformation which introduces a dependency of the model on velocity and acceleration estimates.

This creates a highly nonlinear model and may lead to large dynamic model errors. The continuous-time state-space form of this model is

$$\dot{\vec{x}} = \begin{bmatrix} \dot{\vec{x}} \\ \ddot{\vec{x}} \\ \frac{\dot{\vec{x}} \times \ddot{\vec{x}}}{\dot{\vec{x}}^T \dot{\vec{x}}} \times (\ddot{\vec{x}} - \vec{g}) + C_{N \leftarrow B}(\dot{\vec{x}}, \ddot{\vec{x}}) \vec{u} \\ -\frac{1}{\tau_m} \vec{u} \end{bmatrix} + \begin{bmatrix} 0_{3 \times 3} \\ 0_{3 \times 3} \\ 0_{3 \times 3} \\ I_{3 \times 3} \end{bmatrix} \vec{w}_B \quad (17.76)$$

which uses a state vector

$$\vec{x} = \begin{bmatrix} \vec{x} \\ \dot{\vec{x}} \\ \ddot{\vec{x}} \\ \vec{u} \end{bmatrix} \quad (17.77)$$

## 17.3 Orbital and Ballistic Stochastic Motion Models

**Ballistic flight** is defined as a flight path that is largely predetermined by the performance characteristics specific for type of flight vehicle which includes launch vehicles, ballistic missiles, decoys, debris, satellites,

and projectiles, though some advanced ballistic missiles can undergo small maneuvers usually for re-targeting. Regardless, tracking the motion of *unknown ballistic objects*, also known as **ballistic targets (BT)**, can be challenging due to the variety of uncertainties, most of which result from the target type uncertainty, the principal uncertainty in **ballistic target tracking**. The general BT flight path can be divided into three phases: boost, coast, and reentry. The **boost phase** is the endo-atmospheric, powered flight, which begins at launch and continues until thrust cutoff or burnout. The **coast phase** is the exo-atmospheric, free-motion flight which begins at thrust cutoff or burnout and continues until the atmosphere is reached again and falls under general orbital mechanics. The **reentry phase** is the endo-atmospheric, free-motion flight, which begins when the atmospheric drag becomes considerable and continues until landing or impact.

Though a general motion model based on random processes can be used for the entire flight path of a BT, those models are typically crude as opposed to taking advantage of the fact that ballistic flight can be primarily modeled by three forces: gravity, thrust, and aerodynamic, i.e. drag and possibly lift. Furthermore, not all of these primary forces significantly characterize every flight phase of BTs. The boost phase is characterized by a large thrust, which may be subject to abrupt, jump-wise changes, as well as drag and gravity. After the boost phase, drag is no longer present and thrust vanishes or drops to a very low level. Thus, the coast phase motion is characterized essentially by gravity only though small re-targeting maneuvers are possible. The reentry phase motion is characterized by a rapid drag-induced deceleration with possible lateral accelerations. In summary, the significant forces in the different phases are gravity in coast; aerodynamic and gravity in reentry; and thrust, gravity, and aerodynamic in boost. This section will discuss the relevant **ballistic flight motion models** in order from the least number of significant forces to greatest.

Lastly, this section uses  $\vec{a}$  for acceleration,  $\vec{v}$  for velocity, and  $\vec{p}$  for position where the reference frame for the coordinates are chosen as either the Earth-Centered, Inertial (ECI) or the local East-North-Up (ENU) reference frames. Thus, the kinematic portion of the state vector has the form

$$\vec{x} = \begin{bmatrix} \vec{p} \\ \vec{v} \end{bmatrix} \quad (17.78)$$

and its time derivative as

$$\dot{\vec{x}} = \begin{bmatrix} \vec{v} \\ \vec{a} \end{bmatrix} \quad (17.79)$$

and the primary consideration is how to model the acceleration  $\vec{a}$  appropriately.

## Acceleration Models

The total acceleration of a BT in an inertial frame,  $\vec{a}$ , can be decomposed as

$$\vec{a} = \vec{a}_G + \vec{a}_T + \vec{a}_A = \vec{a}_G + \vec{a}_T + \vec{a}_D + \vec{a}_L \quad (17.80)$$

where  $\vec{a}_G$  is the acceleration due to gravity,  $\vec{a}_T$  is the acceleration due to thrust,  $\vec{a}_A$  is the acceleration due to aerodynamic forces,  $\vec{a}_D$  is the acceleration due to drag, and  $\vec{a}_L$  is the acceleration due to lift. In an Earth-fixed frame, the total acceleration will also include those induced by the Earth's angular velocity,  $\vec{\omega}_E$ , i.e. the Coriolis and centrifugal acceleration

$$\vec{a}_C = 2 [\vec{\omega}_E]_{\times} \vec{v} + [\vec{\omega}_E]_{\times}^2 \vec{p} \quad (17.81)$$

where  $\vec{v}$  is the target velocity and  $[\bullet]_x$  is the skew-symmetric matrix. For ballistic target tracking, one typically uses a range of force models which are summarized as follows. Note that one would need to *subtract* the Earth rotation term from the following for an Earth-fixed frame.

For gravity, one may use a flat-Earth, spherical-Earth, or ellipsoidal-Earth model. A **flat-Earth gravity model** assumes a flat, non-rotating Earth and the gravity acting on the target is constant. In East-North-Up (ENU) coordinates, one has

$$\vec{a}_{G,f-E} = \begin{bmatrix} 0 \\ 0 \\ -g_0 \end{bmatrix} \quad (17.82)$$

where  $g_0$  is the **standard gravitational acceleration**, i.e.  $9.80665 \text{ m-s}^{-2}$ .

A **spherical-Earth gravity model** assumes that Earth and the target are point masses at their centers of gravity, no other body's gravity affects either, and the target has negligible mass. Then, one can use Newton's inverse-square gravitational law as

$$\vec{a}_{G,s-E}(\vec{r}) = -\frac{-\mu_E}{\|\vec{r}\|_2^3} \vec{r} = -\frac{-\mu_E}{\|\vec{r}\|_2^2} \vec{u}_r \quad (17.83)$$

where  $\mu_E$  is the Earth's gravitational constant, i.e.  $398600.440 \text{ km}^3\text{-s}^{-2}$ ,  $\vec{r}$  is the vector between Earth's and the target's center of gravity, and  $\vec{u}_r$  is its unit vector.

An **ellipsoidal-Earth gravity model** accounts for the Earth's oblate spheroid shape by including the **second-order gravitational harmonic**,  $J_2$ , of the Earth's gravitational field, i.e.  $1.75553 \times 10^{10} \text{ km}^5\text{-s}^{-2}$  from the Joint Gravity Model (JGM-3). This can be shown to be the following difference.

$$J_2 = \frac{I_{zz,E} - I_{xx,E}}{m_E R_e^2} \quad (17.84)$$

where  $I_{zz,E}$  is the polar moment of inertia of the ellipsoidal Earth,  $I_{xx,E}$  is the equatorial moment of inertia of the ellipsoidal Earth,  $m_E$  is the mass of Earth, and  $R_e$  is the equatorial radius of Earth. Then, accounting for this second-order correction, one has

$$\vec{a}_{G,e-E}(\vec{r}) = -\frac{-\mu_E}{\|\vec{r}\|_2^2} \left[ \vec{u}_r + \frac{3}{2} J_2 \left( \frac{R_e}{\|\vec{r}\|_2} \right)^2 \left( (1 - 5 (\vec{u}_r^T \vec{u}_z)^2) \vec{u}_r + 2 \vec{u}_r^T \vec{u}_z \vec{u}_z \right) \right] \quad (17.85)$$

where  $R_e$  is the equatorial radius of Earth and  $\vec{u}_z$  is the unit vector along the  $z$ -axis of an Earth-centered reference frame.

For the aerodynamic forces, one typically models the acceleration due to the drag force as

$$\vec{a}_D = -\frac{1}{2m} S C_D \rho(h) \|\vec{v}_r\|_2 \vec{v}_r = -\frac{1}{2m} S C_D \rho(h) \|\vec{v}\|_2^2 \vec{u}_v \quad (17.86)$$

where  $m$  is the target mass,  $S$  is the cross-sectional surface area to the velocity,  $\vec{v}$  is the velocity relative to the atmosphere in a non-inertial frame,  $C_D$  is the drag coefficient which generally depends on  $S$  and  $\vec{v}$ , but is sometimes assumed constant, and  $\rho(h)$  is the air density and depends on the target altitude,  $h$ , and is typically approximated analytically as

$$\rho(h) \approx \rho_0 \exp\left(-\frac{h}{h_0}\right) \quad (17.87)$$

where  $\rho_0 = 1.225 \text{ kg-m}^{-3}$  and  $h_0 = 10,400 \text{ m}$  are often used, though higher fidelity models may also be used. Here,  $\vec{u}_v$  is the unit vector in the direction of the velocity.

In ballistic target models, one often alternatively uses the **drag parameter** as

$$\alpha_D = \frac{SC_D}{m} \quad (17.88)$$

or the **ballistic coefficient (BC)** as

$$\beta_D = \frac{m}{SC_D} \quad (17.89)$$

Thus, the acceleration due to the drag force can be written as

$$\vec{a}_D = -\frac{1}{2}\alpha_D\rho_0 \exp\left(-\frac{h}{h_0}\right) \|\vec{v}\|_2^2 \vec{u}_v \quad (17.90)$$

For ballistic targets, lift occurs when the target velocity vector and thrust vectors are not aligned, i.e. the target is at some angle of attack, typically for the purposes of maneuvering during reentry or boosting. Furthermore, the lift during a maneuver can also be modeled as

$$\vec{a}_L = -\frac{1}{2}\rho(h)\|\vec{v}\|_2^2 \vec{\alpha}_L \vec{u}_L \quad (17.91)$$

where modeling the two coordinates of  $\vec{\alpha}_L$  and  $\vec{u}_L$  depend on the reference frame used for the maneuvering target. This will also add a lift-induced drag to the overall target drag. Lastly, when thrust is expected during the boost phase, one cannot observe both the drag and the thrust separately. Thus, one typically models the combined thrust and drag as the **net acceleration**,  $\vec{a}_{T-D}$ ,

$$\vec{a} = \vec{a}_G + \vec{a}_T + \vec{a}_D = \vec{a}_G + \vec{a}_{T-D} \quad (17.92)$$

This net acceleration may be taken as constant or time-varying in which case, it is modeled as an unknown stochastic process and included in the estimation procedure using basic assumptions on its form.

## Orbital Motion Models

In **orbital motion models**, also known as **coast motion model**, one typically only considers gravity as the significant force. Thus, one has for the total acceleration

$$\dot{\vec{x}} = \begin{bmatrix} \vec{v} \\ \vec{a}_G \end{bmatrix} \quad (17.93)$$

and one need only decide between the spherical-Earth or ellipsoidal-Earth models and the ECI or ENU reference frames. This provides four options for motion models. For the spherical-Earth and ECI model, one has

$$\vec{a} = \vec{a}_{G,s-E}(\vec{p}) = -\frac{\mu_E}{\|\vec{p}\|_2^3} \vec{p} \quad (17.94)$$

and the motion is governed by Kepler's equation of motion.

Though this motion is highly nonlinear, one can predict the state forward-in-time using the following algorithm. Given,  $\vec{x}(t_0) = [\vec{p}_0 \ \vec{v}_0]^T$ , the predicted state at time  $t = t_0 + \Delta t$  is given by

$$\begin{aligned}\vec{p} &= \left(1 - \frac{1 - \lambda^2 C}{\|\vec{p}\|_0}\right) \vec{p}_0 + \left(\Delta t - \frac{\lambda^3 S}{\sqrt{\mu_E}}\right) \vec{v}_0 \\ \vec{v} &= \left(\frac{\sqrt{\mu_E} \lambda (\zeta S - 1)}{\|\vec{p}_0\|_2 \|\vec{p}\|_2}\right) \vec{p}_0 + \left(1 - \frac{\lambda^2 C}{\|\vec{p}\|_0}\right) \vec{v}_0\end{aligned}\quad (17.95)$$

where  $C$ ,  $S$ , and  $\zeta$  are known functions of  $\lambda$ , and the parameter  $\lambda$  is defined through  $\dot{\lambda} = \sqrt{\mu_E}/\|\vec{p}\|_2$  and is typically computed using a Newton iteration scheme for the **time-of-flight (TOF) equation** which can be defined for elliptical, parabolic, and hyperbolic flight paths.

For elliptical orbits, i.e.  $a = 2/\|\vec{p}_0\|_2 - \|\vec{v}_0\|_2^2/\mu_E > 0$ , one can initialize the iteration on  $\lambda$  as

$$\begin{aligned}\alpha &= \frac{1 - a \|\vec{p}_0\|_2}{\sqrt{\mu_E}} \\ \beta &= \frac{\vec{p}_0^T \vec{v}_0}{\mu_E} \\ \gamma &= \frac{\|\vec{p}_0\|}{\sqrt{\mu_E}} \\ \lambda_0 &= \frac{a \Delta t}{\sqrt{\mu_E}}\end{aligned}\quad (17.96)$$

and iterate  $\lambda_i$  for  $i = 1, 2, \dots$  until  $|\Delta t - \tau| < \epsilon$ :

$$\begin{aligned}\zeta &= a \lambda_i^2 \\ C &= \frac{1 - \cos \sqrt{\zeta}}{\zeta} \\ S &= \frac{\sqrt{\zeta} - \sin \sqrt{\zeta}}{\zeta \sqrt{\zeta}} \\ \tau &= \alpha \lambda_i^3 S + \beta \lambda_i^2 C + \gamma \lambda_i \\ \frac{d\tau}{d\lambda} &= \alpha \lambda_i^2 C + \beta \lambda_i (1 - \zeta S) + \gamma \\ \lambda_{i+1} &= \lambda_i + \left[ \frac{d\tau}{d\lambda} \right]^{-1} (\Delta t - \tau)\end{aligned}\quad (17.97)$$

For the ellipsoidal-Earth and ECI model, one has

$$\vec{a} = \vec{a}_{G,e-E}(\vec{p}) \quad (17.98)$$

where

$$\vec{u}_p = \frac{\vec{p}}{\|\vec{p}\|} \quad (17.99)$$

and

$$\vec{u}_z = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (17.100)$$

Thus, one has

$$\vec{a} = -\frac{\mu_E}{\|\vec{p}\|_2^3} \begin{bmatrix} 1 + \frac{3J_2R_e^2}{2\|\vec{p}\|_2^2} \left( 1 - 5 \left( \frac{z}{\|\vec{p}\|_2} \right)^2 \right) \\ 1 + \frac{3J_2R_e^2}{2\|\vec{p}\|_2^2} \left( 1 - 5 \left( \frac{z}{\|\vec{p}\|_2} \right)^2 \right) \\ 1 + \frac{3J_2R_e^2}{2\|\vec{p}\|_2^2} \left( 3 - 5 \left( \frac{z}{\|\vec{p}\|_2} \right)^2 \right) \end{bmatrix} \vec{p} \quad (17.101)$$

For the spherical-Earth gravity model and ENU reference frame, one has

$$\vec{a} = \vec{a}_{G,s-E}(\vec{r}) - 2\vec{\omega}_E \Phi_0 \vec{v} - \omega_E^2 \Phi_0^2 \vec{x} \quad (17.102)$$

where

$$\Phi_0 = \begin{bmatrix} 0 & -\sin \phi_0 & \cos \phi_0 \\ \sin \phi_0 & 0 & 0 \\ -\cos \phi_0 & 0 & 0 \end{bmatrix} \quad (17.103)$$

and

$$\vec{r} = \begin{bmatrix} x \\ y \\ z + \|\vec{r}\|_0 \end{bmatrix} \quad (17.104)$$

where  $\phi$  is the latitude of the reference point of the ENU frame  $\vec{r}_0$  is the vector from the Earth center to the origin of the ENU reference frame, i.e. the sum of the Earth mean radius and origin altitude.

Likewise, for the ellipsoidal-Earth gravity model and ENU reference frame, one has

$$\vec{a} = \vec{a}_{G,e-E}(\vec{r}) - 2[\vec{\omega}_E]_{\times} \vec{v} - [\vec{\omega}_E]_{\times}^2 \vec{p} \quad (17.105)$$

For these coast models, one may employ the EKF as the *de facto* standard for tracking. However, the linearization error needed for covariance time update may not be sufficiently accurate for relatively long time intervals. However, for high sampling rates, i.e.  $\Delta t$  small, one may use a simple piecewise-constant acceleration model across time steps,  $k \rightarrow k+1$ , i.e.

$$\vec{a}(\hat{\vec{x}}_{k|k}) = \hat{\vec{a}}_{k|k} = \begin{bmatrix} \hat{\dot{x}}_{k|k} \\ \hat{\dot{y}}_{k|k} \\ \hat{\dot{z}}_{k|k} \end{bmatrix} \quad (17.106)$$

then, for the discrete-time state equation, one has for the  $x$ -coordinate

$$\begin{bmatrix} x_{k+1} \\ \dot{x}_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_k \\ \dot{x}_k \end{bmatrix} + \begin{bmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{bmatrix} (\hat{\dot{x}}_{k|k} + w_x) \quad (17.107)$$

with  $w_x$  as some small zero-mean white-noise with variance  $q_x^2$  and likewise for the  $y$ - and  $z$ -coordinates. In practice, the values of  $q_x$ ,  $q_y$  and  $q_z$  are design parameters.

## Reentry Motion Models

In reentry motion models, one typically considers gravity and aerodynamics as the significant forces. Thus, one has for the total acceleration

$$\dot{\vec{x}} = \begin{bmatrix} \vec{v} \\ \vec{a}_G + \vec{a}_A \end{bmatrix} \quad (17.108)$$

However, one may consider two types of targets: non-maneuvering and maneuvering where non-maneuvering will only have zero-lift drag while maneuvering may also have lift and lift-induced drag.

For non-maneuvering reentry targets, if one has a *known* drag parameter,  $\alpha_D$ , the total acceleration written in the ENU reference frame simply becomes

$$\vec{a} = \vec{a}_G - 2 [\vec{\omega}_E]_{\times} \vec{v} - [\vec{\omega}_E]_{\times}^2 \vec{p} - \frac{1}{2} \alpha_D \rho_0 \exp\left(-\frac{h}{h_0}\right) \|\vec{v}\|_2 \vec{v} \quad (17.109)$$

and the reentry motion model can be written with a suitable gravity model as shown previously for short time intervals. However, if one has an *unknown* drag parameter or it is time-varying or too complex to model adequately, one may instead estimate it alongside the position and velocity. This is typically done either assuming a “nearly” constant drag parameter using a Wiener process model, i.e.

$$\dot{\alpha}_D(t) = w_{\alpha}(t) \quad (17.110)$$

or a first-order Markov process, i.e.

$$\dot{\alpha}_D(t) = -\frac{1}{\tau_{\alpha}} \alpha_D(t) + w_{\alpha}(t) \quad (17.111)$$

where  $w_{\alpha}(t)$  is zero-mean white noise with variance  $\sigma_{\alpha}^2$  as a design parameter.

Alternatively, one may use a nominal value  $\bar{\alpha}$  and model this as an exponential, i.e.

$$\alpha_D = \bar{\alpha}_D e^{\delta} \quad (17.112)$$

or

$$\delta = \ln\left(\frac{\alpha_D}{\bar{\alpha}_D}\right) \quad (17.113)$$

where  $\delta$  can be taken as a Wiener process, i.e.

$$\dot{\delta}(t) = w_{\delta}(t) \quad (17.114)$$

or a first-order Markov process, i.e.

$$\dot{\delta}(t) = -\frac{1}{\tau_{\delta}} \delta(t) + w_{\delta}(t) \quad (17.115)$$

where  $w_{\delta}(t)$  is zero-mean white noise with variance  $\sigma_{\delta}^2$  as a design parameter.

For maneuvering reentry targets, one must account for possible lift and lift-induced drag. As such, one must choose the two coordinates for which one can decompose the lift. A common reference frame is the velocity-turn-climb (VTC) reference frame. Then, the acceleration due to aerodynamic forces becomes

$$\vec{a}_A = \frac{1}{2} \rho_0 \exp\left(-\frac{h}{h_0}\right) \|\vec{v}\|_2^2 \begin{bmatrix} \vec{u}_v & \vec{u}_t & \vec{u}_c \end{bmatrix} \vec{a}_A \quad (17.116)$$

where the **aerodynamic parameter vector**,  $\vec{\alpha}_A$  can be written as

$$\vec{\alpha}_A = \begin{bmatrix} -(1 + \lambda_0)\alpha_D \\ \alpha_t \\ \alpha_c \end{bmatrix} \quad (17.117)$$

and  $\lambda_0$  is the ratio of the lift to the **critical lift**, i.e. the lift at the maximum lift-to-drag ratio.

Here,  $\alpha_t > 0$  corresponds to a left turning maneuver,  $\alpha_t < 0$  corresponds to a right turning maneuver,  $\alpha_c > 0$  corresponds to a climbing maneuver, and  $\alpha_c < 0$  corresponds to a diving maneuver. It should also be pointed out that  $\vec{\alpha}_A$  corresponds to four generally unknown parameters which will not generally be observable in an estimator as one cannot distinguish between  $\lambda_0$  and  $\alpha_D$ . Thus, one typically uses

$$\vec{\alpha}_A = \begin{bmatrix} -\tilde{\alpha}_D \\ \alpha_t \\ \alpha_c \end{bmatrix} \quad (17.118)$$

For the total acceleration, one must transform this to the ENU reference frame through a rotation matrix,  $C_{ENU \leftarrow VTC}$ , i.e.

$$C_{ENU \leftarrow VTC} = \begin{bmatrix} \cos \epsilon \cos \psi & -\cos \psi & \sin \epsilon \cos \psi \\ \cos \epsilon \sin \psi & \sin \psi & \sin \epsilon \sin \psi \\ -\sin \epsilon & 0 & \cos \epsilon \end{bmatrix} \quad (17.119)$$

where the heading angle is

$$\psi = \tan^{-1} \frac{\dot{y}}{\dot{x}} \quad (17.120)$$

and the negative elevation angle is

$$\epsilon = \tan^{-1} \frac{-\dot{z}}{\sqrt{\dot{x}^2 + \dot{y}^2}} \quad (17.121)$$

For the motion model, one must choose either the ENU or ECI frames for the gravitational force where the former will also require the Coriolis and centrifugal accelerations. As before, one either assumes either a scalar Wiener process or Markov process for each drag parameter in the *unknown*  $\vec{\alpha}_A$ .

## Boost Motion Models

For boost motion models, one typically considers gravity and the net acceleration, i.e. thrust minus drag, as the significant forces. Thus, one has for the total acceleration

$$\dot{\vec{x}} = \begin{bmatrix} \vec{v} \\ \vec{a}_G + \vec{a}_{T-D} \end{bmatrix} \quad (17.122)$$

Here, the most general motion model uses one of the gravity models and either a Wiener process or Markov process for the net acceleration,  $\vec{a}_{T-D}$ .

However, to take advantage of the kinematics of a typical boost phase, one may use a **gravity turn (GT) motion model** which is characterized by the thrust being parallel to the Earth-relative velocity vector. In the

ECEF reference frame, the GT constraint implies that non-gravitational acceleration are related through a **acceleration-to-speed ratio (ASR)**,  $\lambda_{ASR}$ , i.e.

$$\vec{a} = \lambda_{ASR} \vec{v} \quad (17.123)$$

Thus, a simple GT motion model can be written in state-space form in continuous-time as

$$\begin{bmatrix} \vec{v} \\ \vec{a} \end{bmatrix} = \begin{bmatrix} 0_{3 \times 3} & I_{3 \times 3} \\ 0_{3 \times 3} & \lambda_{ASR} I_{3 \times 3} \end{bmatrix} \begin{bmatrix} \vec{p} \\ \vec{v} \end{bmatrix} \quad (17.124)$$

and in discrete-time as

$$\begin{bmatrix} \vec{p}_{k+1} \\ \vec{v}_{k+1} \end{bmatrix} = \begin{bmatrix} I_{3 \times 3} & \frac{1}{\lambda_{ASR,k}} (\exp(\lambda_{ASR,k} \Delta t) - 1) I_{3 \times 3} \\ 0_{3 \times 3} & \exp(\lambda_{ASR} \Delta t) I_{3 \times 3} \end{bmatrix} \begin{bmatrix} \vec{p}_k \\ \vec{v}_k \end{bmatrix} \quad (17.125)$$

which is noticeably very similar to the CT models where  $\lambda_{ASR}$  may be known *a priori*, adaptively estimated, or modeled using a Wiener or Markov process. In addition, this GT model must still be gravity-compensated in application, typically as a piecewise-constant model.

For rocket boost motion models where the mass rate has significant effects on the target dynamics, one may use a **rocket gravity turn (GT) motion model**

$$\vec{v}_r = \vec{v} - \vec{\omega} \times \vec{p} \quad (17.126)$$

or

$$\vec{a}_{T-D} = \alpha_a \frac{\vec{v}_r}{\|\vec{v}_r\|} \quad (17.127)$$

where  $\alpha_a = \|\vec{a}_{T-D}\|$ . Thus, the total acceleration is

$$\vec{a} = \alpha_a \frac{\vec{v}_r}{\|\vec{v}_r\|} + \vec{a}_G \quad (17.128)$$

and since the drag is in the opposite direction of the velocity, one has

$$\alpha_a = \frac{T - D}{m} \quad (17.129)$$

which explicitly models GT for rockets as opposed to general GT models using the  $\lambda_{ASR}$  parameter.

It is often assumed that the magnitude of the non-gravitational net force,  $F(t) = T(t) - D(t) = F(t_0)$  is constant where  $t_0$  is an arbitrary initial time. Then, for  $t \geq t_0$ ,

$$\alpha_a(t) = \frac{F(t)}{m(t)} = \frac{F(t_0)}{m(t_0) - (t - t_0)\delta} = \frac{\alpha_a(t_0)}{1 - \beta_a(t_0)(t - t_0)} \quad (17.130)$$

and

$$\beta_a(t) = \frac{\delta}{m(t)} = \frac{\delta}{m(t_0) - (t - t_0)\delta} = \frac{\beta_a(t_0)}{1 - \beta_a(t_0)(t - t_0)} \quad (17.131)$$

where  $\beta_a = -\dot{m}/m$ . From differentiation, one has

$$\dot{\alpha}_a = \alpha_a \beta_a \quad (17.132)$$

and

$$\dot{\beta}_a = \beta_a^2 \quad (17.133)$$

Then, the eight-state GT model for the state vector

$$\vec{x} = \begin{bmatrix} \vec{p} \\ \vec{v} \\ \alpha_a \\ \beta_a \end{bmatrix} \quad (17.134)$$

and the state rate is given by

$$\dot{\vec{x}} = \begin{bmatrix} \vec{v} \\ \alpha_a \frac{\vec{v}_r}{\|\vec{v}_r\|} + \vec{a}_G \\ \alpha_a \beta_a \\ \beta_a^2 \end{bmatrix} \quad (17.135)$$

which is a nonlinear model and can be difficult to use in filtering.

A generalization of this rocket GT model to the general boost case in VTC coordinates allows one to include the possible lateral forces during boost as velocity acceleration,  $\alpha_v$ , turning acceleration,  $\alpha_t$ , and climbing acceleration,  $\alpha_c$ , as the net acceleration

$$\vec{a}_{T-D} = [\vec{u}_v \quad \vec{u}_t \quad \vec{u}_c] \begin{bmatrix} \alpha_v \\ \alpha_t \\ \alpha_c \end{bmatrix} \quad (17.136)$$

where

$$\alpha_v = \frac{T_v - D}{m} \quad (17.137)$$

$$\alpha_t = \frac{T_t + L_t}{m} \quad (17.138)$$

and

$$\alpha_c = \frac{T_c + L_c}{m} \quad (17.139)$$

where  $\vec{T} = [T_v \ T_t \ T_c]^T$  is the thrust force,  $D$  is the magnitude of the drag force,  $L_t$  and  $L_c$  are the turning and climbing lift forces.

Then, the total acceleration,  $\vec{a}$ , in ENU coordinates can be written as

$$\vec{a} = \vec{a}_G + C_{ENU \leftarrow VTC} \begin{bmatrix} \alpha_v \\ \alpha_t \\ \alpha_c \end{bmatrix} \quad (17.140)$$

where

$$C_{ENU \leftarrow VTC} = [\vec{u}_v \quad \vec{u}_t \quad \vec{u}_c] \quad (17.141)$$

Then, under the same constant net force in the moving VTC reference frame, one has a ten-state model for the state vector

$$\vec{x} = \begin{bmatrix} \vec{p} \\ \vec{v} \\ \alpha_v \\ \alpha_t \\ \alpha_c \\ \beta_a \end{bmatrix} \quad (17.142)$$

with state equation

$$\dot{\vec{x}} = \begin{bmatrix} \vec{v} \\ \vec{a}_G + C_{ENU \leftarrow VTC} \begin{bmatrix} \alpha_v \\ \alpha_t \\ \alpha_c \end{bmatrix} \\ \beta_a \begin{bmatrix} \alpha_v \\ \alpha_t \\ \alpha_c \end{bmatrix} \\ \beta_a^2 \end{bmatrix} \quad (17.143)$$

where it should be noted that one should add zero-mean white noise to this model.

### Preliminary Orbit Determination

Tracking orbital objects often requires several stages to form a suitable motion model. **Preliminary orbit determination (POD)** utilizes the minimal number of pseudomeasurements, i.e., six, to estimate the six orbital elements of an orbiting body, e.g., planet, moon, asteroid, satellite, whereas **orbit estimation** utilizes large sets of data to determine the orbital elements typically as a batch process, e.g., least-squares, or a refinement of the estimated orbital elements from a POD algorithm in a nonlinear Bayesian filter, e.g., extended Kalman filter.

Gauss's method

Herrick-Gibbs method

Angles-Only method

### References

For more information, please refer to the following

- Curtis, H. D., “5 Preliminary Orbit Determination,” *Orbital Mechanics for Engineering Students*, 1st ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2005, pp. 448-459
- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “1.6 Geodesy, Coordinate Systems, Gravity,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 23-34

## **Part IV**

# **Aerospace Vehicle Sensors and Perception Systems**

---

# Sensor and Data Systems

## 18.1 Inertial Sensors

**Inertial sensors** are accelerometers and gyroscopes which are described in terms of the sensing concepts in this section.

### Accelerometer Design

An **accelerometer** is a sensor that directly measures the acceleration of a body in its own instantaneous rest frame. This provides a measurement of

Conceptually, an accelerometer is a **proof mass**, also known as a **seismic mass**, on a spring. For planetary-based accelerometers, the rest frame is the Earth so the gravity force is not included. When the accelerometer experiences an acceleration, the proof mass is deflected to the point that the spring accelerates the mass at the same speed as the casing. The deflection of the proof mass from its neutral position or the compression of the spring provides an acceleration measurement. However, the mass-spring system is also damped so that additional oscillations of the system do not affect the measurement output. Thus, all accelerometers have a frequency response and are designed to provide the required sensitivity and maximum expected acceleration. There are many different designs for converting this mass deflection or spring compression into a measurable electric signal including piezoelectric, piezoresistive, capacitive, and thermal.

**Piezoelectric accelerometers** use single crystals, e.g., quartz, or piezoceramics, e.g., lead-zirconate-titanate. They are constructed with two piezoelectric transducers. The unit consists of a hollow tube that is sealed by a piezoelectric transducer on each end. The transducers are oppositely polarized and are selected to have a specific series capacitance. The tube is then partially filled with a heavy liquid and the accelerometer is excited. While excited the total output voltage is continuously measured and the volume of the heavy liquid is adjusted until the desired output voltage is obtained. Finally the outputs of the individual transducers are measured, the residual voltage difference is tabulated, and the dominate transducer is identified. They are unmatched in high frequency measurements, low packaged weight, and resistance to high temperatures.

**Piezoresistive accelerometers** use a semiconductor layer that is attached to a handle wafer by a thick oxide layer. The semiconductor layer is then patterned to the accelerometer's geometry. This semiconductor layer has one or more apertures so that the underlying mass will have the corresponding apertures. Next the semiconductor layer is used as a mask to etch out a cavity in the underlying thick oxide. A mass in the cavity is supported in cantilever fashion by the piezoresistant arms of the semiconductor layer. Directly below the accelerometer's geometry is a flex cavity that allows the mass in the cavity to flex or move in direction that is orthogonal to the surface of the accelerometer. These resist shock better and are able to measure very high accelerations. Integrating piezoresistors in the springs to detect spring deformation, and thus deflection, is a good alternative, although a few more process steps are needed during the fabrication sequence.

**Capacitive accelerometer** use the properties of an opposed plate capacitor for which the distance between the plates varies proportionally to applied acceleration, thus altering capacitance. This variable is used in a circuit to ultimately deliver a voltage signal that is proportional to acceleration. Capacitive accelerometers are capable of measuring constant as well as slow transient and periodic acceleration. AC capacitive-acceleration sensors fundamentally contain at least two components; the primary is a 'stationary' plate (i.e., connected to the housing) and the secondary plate is attached to the inertial mass, which is free to move inside the housing. These plates form a capacitor whose value is a function of a distance between the plates. The sensing material is either a flat plate of nickel or electronic chip supported above the substrate surface by two torsion bars attached to a central pedestal. This design is simple, reliable, and inexpensive. They measure low frequencies well.

**Thermal accelerometer**, also known as **convective accelerometers**, contain a small heater in a very small dome. This heats the air or other fluid inside the dome. The thermal bubble acts as the proof mass. An accompanying temperature sensor, e.g., a thermistor; or thermopile, in the dome measures the temperature in one location of the dome. This measures the location of the heated bubble within the dome. When the dome is accelerated, the colder, higher density fluid pushes the heated bubble. The measured temperature changes. The temperature measurement is interpreted as acceleration. The fluid provides the damping. Gravity acting on the fluid provides the spring. Since the proof mass is very lightweight gas, and not held by a beam or lever, thermal accelerometers can survive high shocks. Another variation uses a wire to both heat the gas and detect the change in temperature. The change of temperature changes the resistance of the wire. A two dimensional accelerometer can be economically constructed with one dome, one bubble and two measurement devices.

## Gyroscope Design

A **rate gyroscope**, referred to as a **gyroscope** in this work, is a sensor that directly measures the rotational rate about an axis. Conceptually, it is a spinning wheel or disc in which the spin axis is free to assume any orientation by itself. When rotating, the orientation of this axis is unaffected by tilting or rotation of the mounting, according to the conservation of angular momentum. There are two broad design categories for gyroscopes: gyrostabilized and strapdown. **Gyrostabilized gyroscope** measure the angles of gimbal bearings or fluid bearings. **Strapdown** measure either optical effects or vibrating structures effects to infer the rotational rate.

A **gimbal** is a ring that is suspended so it can rotate about an axis and nested gimbals with bearings nominally at right angles to each other. Due to the suspended nature of these platforms, the sensors will keep the same orientation relative to the inertial frame while the housing rotates. In addition, two gyroscopes are

typically placed on each ring to cancel **gyroscopic precession**, i.e. the tendency of a gyroscope to twist at right angles to an input torque, thereby changing its axis of rotation. By placing a sensor at the bearings of the gimbals, one can measure the vehicle's attitude by the three sequential rotations of the nested gimbals, i.e. the inertial-to-body-frame Euler angles.

The three major disadvantages to gyrostabilized IMUs are the degradation of the many expensive precision mechanical parts, the mechanical parts' potential to jam, and the fatal phenomenon called gimbal lock. **Gimbal lock** is the loss of one degree of freedom that occurs when the axes of two of the three gimbal rings are driven into a parallel configuration. Because of the parallel orientation of two of the gimbals' axes there is no gimbal available to accommodate rotation about the orthogonal direction to the two axes, thereby "locking" the gyroscope into only measuring rotation in the resulting two-dimensional space.

To see this, consider the direction cosine matrix relationship to the Euler angles

$$C_{G \leftarrow F} = \begin{bmatrix} \cos \theta_y \cos \theta_z & \cos \theta_y \sin \theta_z & -\sin \theta_y \\ \sin \theta_x \sin \theta_y \cos \theta_z - \cos \theta_x \sin \theta_z & \sin \theta_x \sin \theta_y \sin \theta_z + \cos \theta_x \cos \theta_z & \sin \theta_x \cos \theta_y \\ \cos \theta_x \sin \theta_y \cos \theta_z + \sin \theta_x \sin \theta_z & \cos \theta_x \sin \theta_y \sin \theta_z - \sin \theta_x \cos \theta_z & \cos \theta_x \cos \theta_y \end{bmatrix} \quad (18.1)$$

If  $\theta_y = \pi/2$ , one has

$$C_{G \leftarrow F} = \begin{bmatrix} 0 & 0 & -1 \\ \sin \theta_x \cos \theta_z - \cos \theta_x \sin \theta_z & \sin \theta_x \sin \theta_z + \cos \theta_x \cos \theta_z & 0 \\ \cos \theta_x \cos \theta_z + \sin \theta_x \sin \theta_z & \cos \theta_x \sin \theta_z - \sin \theta_x \cos \theta_z & 0 \end{bmatrix} \quad (18.2)$$

which by trigonometric summation formulas can be rewritten as

$$C_{G \leftarrow F} = \begin{bmatrix} 0 & 0 & -1 \\ \sin(\theta_x - \theta_z) & \cos(\theta_x - \theta_z) & 0 \\ \cos(\theta_x - \theta_z) & -\sin(\theta_x - \theta_z) & 0 \end{bmatrix} \quad (18.3)$$

Similarly, if  $\theta_y = -\pi/2$ , one has

$$C_{G \leftarrow F} = \begin{bmatrix} 0 & 0 & 1 \\ -\sin \theta_x \cos \theta_z - \cos \theta_x \sin \theta_z & -\sin \theta_x \sin \theta_z + \cos \theta_x \cos \theta_z & 0 \\ -\cos \theta_x \cos \theta_z + \sin \theta_x \sin \theta_z & -\cos \theta_x \sin \theta_z - \sin \theta_x \cos \theta_z & 0 \end{bmatrix} \quad (18.4)$$

which by the trigonometric summation formulas can be rewritten as

$$C_{G \leftarrow F} = \begin{bmatrix} 0 & 0 & 1 \\ -\sin(\theta_x + \theta_z) & \cos(\theta_x + \theta_z) & 0 \\ -\cos(\theta_x + \theta_z) & -\sin(\theta_x + \theta_z) & 0 \end{bmatrix} \quad (18.5)$$

Thus, at the critical angles,  $\theta_y = \pm\pi/2$ , there is a loss in gimbal bearing angles to distinguish difference between  $\theta_x$  and  $\theta_z$  gimbal angles which reduces gyroscope to 2-D, i.e.  $\theta_y$  and  $\theta_x + \theta_z$  or  $\theta_x - \theta_z$ . In practice, mechanical gimbals encounter difficulties *near* these two critical angles where small rotations of the platform would require large motions of the surrounding gimbals which makes them resist motion due to their inertia, bearing friction, and fluid resistance.

**Optical gyroscopes** use the **Sagnac effect** of light interferometry to sense the angular motion. Optical gyroscopes are either ring laser or fiber optic gyroscopes. A **ring laser gyroscope (RLG)** uses a single

closed-path with a laser beam while a **fiber optic gyroscope (FOG)** uses multiple coils of the closed-path for a fiber optic cable of known length. A FOG is less sensitive than an RLG to shock and vibration and offers excellent thermal stability, but can be susceptible to magnetic interference on the fiber optics.

To explain the Sagnac effect, consider a light beam split into two beams that are constrained to follow the same closed path but in opposite directions where one can perform interferometer on the interacting beams when they return back to source. Then, gyroscopic motion can be determined when at rest, the light beams take the same amount of time to traverse the closed path in either direction while when rotating, one beam of light has a longer path to travel than the other and takes longer. This difference in time results in a phase difference in the interference pattern in the interferometer due to Sagnac effect where the relative phases or positions of the interference fringes that are shifted by tracking loop offset are proportional to the rotational rate of closed path, e.g. out of or into the page as shown above. Thus, three different loops would be required for a three-axis gyroscope.

Vibrating structure gyroscopes use the **Coriolis effect** to sense angular motion. Thus, these gyroscopes are also known as **Coriolis vibratory gyroscope (CVG)**. There are five general types of CVGs.

- Cylindrical Resonator Gyroscope (CRG)
- Piezoelectric gyroscopes
- Tuning fork gyroscope
- Wine-glass resonator
- Vibrating Wheel Gyroscope

For each of these types of CVGs, there are versions that can be made into microscopic devices incorporating both electronic and moving parts known as **micro-electromechanical systems (MEMS)**.

## Inertial Measurement Units

**Inertial measurement units (IMU)** are multi-sensor electronic devices which measure the **specific force**, i.e., the non-gravitational force per mass, and the angular velocity. For the three-dimensional vector quantities, an IMU uses three orthogonal accelerometers and three orthogonal gyroscopes. An IMU also typically has at least a simple onboard processor to use as a digital interface, long-term memory for unit conversion, and a temperature sensor for calibration modeling. It is also very common for IMUs to include magnetometers which are covered in the next section as they are not strictly inertial sensors. Notably, there are many different IMU technologies for the underlying IMU design with vastly different performance grades.

Modern mechanical IMUs are often small **micro-electro-mechanical systems (MEMS)**. MEMS range in size from  $0.004$  to  $1.0 \text{ mm}^2$  and are made up of components between  $0.00001$  to  $0.01 \text{ mm}^2$ , although components arranged in large arrays can be more than  $1000 \text{ mm}^2$ . They usually consist of an integrated circuit chip for simple data processing and microsensors, i.e. components that interact with the surrounding environment. Because of the large surface area to volume ratio of MEMS, forces produced by ambient electromagnetism, e.g., electrostatic charges and magnetic moments, and fluid dynamics, e.g., surface tension and viscosity, are more important design considerations than with larger scale mechanical devices. MEMS are manufactured by **lithographic construction**, i.e., the transfer of a pattern into a photosensitive material by selective exposure to a radiation source, e.g., light. The use of MEMS IMUs have allowed

IMUs to be made at the chip level and provide inertial information in a wide variety of applications beyond navigation.

Depending on the targeted applications, the choice of IMU should consider the needs regarding stability, repeatability, and environment sensitivity, mainly thermal and mechanical environments, on both short and long time scales. However, as the raw sensor performance is repeatable over time, IMUs are typically compensated in real-time to enhance their performance based on both the sensor data and IMU models. The model complexity is chosen according to the needed performance for the specific type of application considered and the model parameters are typically assessed through dedicated calibration tests using multi-axis turntables and climatic chambers. They can either be computed for each individual product or generic for the whole product line. Calibration will typically improve a sensor's raw performance by a factor of  $10\times$  to  $10,000\times$ . Due to these production and calibration factors, the price and performance of IMUs vary widely. Thus, it is important to understand the different IMU performance grades in order to use the right sensor for an application. These grades correspond to IMU errors resulting from imperfect mechanical mounting of the measurement axes, scaling nonlinearities, parasitic error induced by solicitation along an axis orthogonal to sensor axis, thermal gradients, linear accelerations for gyroscopes, and non-repeatable errors.

There are roughly four grades of IMUs, namely

- **Navigation grade**, also known as **marine grade**:
  - $\sim >\$50,000$
  - $\sim 0.5$  m after 1 minute of integration
- **Tactical grade**:
  - $\sim \$5,000\text{-}50,000$
  - $\sim 5$  m after 1 minute of integration
- **Industrial grade**:
  - $\sim \$100\text{-}1,000$
  - $\sim 50$  m after 1 minute of integration
- **Consumer grade**, also known as **automotive grade**:
  - $\sim \$10$ s
  - $\sim 500$  m after 1 minute of integration

Taking into account the major errors for aerospace attitude determination or navigation systems, a three-axis accelerometer measurement model is generally specified as

$$\vec{y}_{a,B} = \begin{bmatrix} 1 + s_{a,x} & \sigma_{a,xy} & \sigma_{a,xz} \\ 0 & 1 + s_{a,y} & \sigma_{a,yz} \\ 0 & 0 & 1 + s_{a,z} \end{bmatrix} (\vec{f}_B + \bar{b}_{a,B} + \vec{b}_{a,B} + \vec{\epsilon}_{a,B}) \quad (18.6)$$

where  $\vec{f}_B = \vec{a}_{B,B/I} - \vec{g}_B$  is the true specific force in the body-fixed frame,  $\vec{a}_{B,B/I}$  is the true total acceleration in the body-fixed frame,  $\vec{g}_B$  is the force of gravity in the body-fixed frame which requires a

rotation from the navigation, LVLH, or inertial frame,  $s_{a,z}$ ,  $s_{a,y}$ , and  $s_{a,z}$  are the linear accelerometer scale factor errors in each axis,  $\sigma_{a,xy}$ ,  $\sigma_{a,xz}$ , and  $\sigma_{a,yz}$  are the accelerometer non-orthogonality errors in each axis,  $\bar{b}_{a,B}$  is the accelerometer repeatability bias, also known as the turn-on bias stability, in the body-fixed frame which is constant during operation, but can fluctuate from one turn-on to the next over the lifetime of the accelerometer,  $\vec{b}_{a,B}$  is the accelerometer random bias drift, also known as the bias instability or the in-run bias stability, in the body-fixed frame, and  $\vec{\epsilon}_{a,B}$  is the accelerometer random noise in the body-fixed frame.

Likewise, a three-axis gyroscope measurement model can be generally specified as

$$\vec{y}_{g,B} = \begin{bmatrix} 1 + s_{g,x} & \sigma_{g,xy} & \sigma_{g,xz} \\ 0 & 1 + s_{g,y} & \sigma_{g,yz} \\ 0 & 0 & 1 + s_{g,z} \end{bmatrix} \left( \vec{\omega}_{B,I \leftarrow B} + \bar{b}_{g,B} + \vec{b}_{g,B} + \vec{\epsilon}_{g,B} \right) \quad (18.7)$$

where  $\vec{\omega}_{B,I \leftarrow B}$  is the angular velocity of the body-fixed frame relative to the inertial frame in body-fixed frame coordinates,  $s_{g,z}$ ,  $s_{g,y}$ , and  $s_{g,z}$  are the linear gyroscope scale factor errors in each axis,  $\sigma_{g,xy}$ ,  $\sigma_{g,xz}$ , and  $\sigma_{g,yz}$  are the gyroscope non-orthogonality errors in each axis,  $\bar{b}_{g,B}$  is the gyroscope repeatability bias, also known as the turn-on bias stability, in the body-fixed frame which is constant during operation, but can fluctuate from one turn-on to the next over the lifetime of the gyroscope,  $\vec{b}_{g,B}$  is the gyroscope random bias drift, also known as the bias instability or the in-run bias stability, in the body-fixed frame, and  $\vec{\epsilon}_{g,B}$  is the gyroscope random noise in the body-fixed frame.

Notably, the scale factor and non-orthogonality errors are often significantly reduced by calibration procedures as to be eliminated from the general three-axis measurement models and may also be tabulated with respect to the ambient temperature to further reduce them. The state augmentation of scale factor and non-orthogonality errors into an attitude determination or navigation filter may be worthwhile based on the specific application as it adds additional computational load. However, the random bias drift and noise terms for each measurement axis of an accelerometer or gyroscope forms the **IMU stochastic error** vector,  $\vec{e}_{a,B} = \vec{b}_{a,B} + \vec{\epsilon}_{a,B}$  and  $\vec{e}_{g,B} = \vec{b}_{g,B} + \vec{\epsilon}_{g,B}$ . For aerospace perception systems, these stochastic errors are typically specified for each component by a truncated series of their power spectral densities (PSD) terms dependent on angular frequency as

$$S_{e_\bullet}(\omega) = N^2 + \frac{\mathcal{B}^2}{\omega} + \frac{K^2}{\omega^2} + \frac{R^2}{\omega^3} \quad (18.8)$$

where  $\bullet$  represents a particular component, i.e.,  $a, B, x; a, B, y; a, B, z; g, B, x; g, B, y;$  or  $g, B, z$ ,  $N$  is the white noise coefficient,  $B$  is the pink noise coefficient,  $K$  is the brown or red noise coefficient, and  $R$  is the pink ramp coefficient. Notably, this equation is often described with cyclic frequency,  $f = \omega/(2\pi)$ .

As the high-frequency pink ramp effects are easily filtered out, one typically decomposes the stochastic error into three independent terms as

$$e_\bullet = \vec{e}_{\bullet,N} + \vec{e}_{\bullet,B} + \vec{e}_{\bullet,K} \quad (18.9)$$

where  $\vec{e}_{\bullet,N}$  is known as the **velocity random walk (VRW)** for accelerometers and the **angle random walk (ARW)** for gyroscopes,  $\vec{e}_{\bullet,B}$  is known as the bias instability, also known as the in-run bias stability or flicker noise, and  $\vec{e}_{\bullet,K}$  is known as the **acceleration random walk** for accelerometers and the **rate random walk** for gyroscopes. With respect to each of these terms, one can exactly model the white term of the stochastic error as

$$e_{\bullet,N} = \epsilon_{\bullet,N} \quad (18.10)$$

where  $\epsilon_N$  is white random noise with PSD  $S_N = N^2$  and variance  $\sigma_N^2 = S_N$ . Likewise, one can exactly model the brown term of the stochastic error as

$$\begin{aligned}\dot{b}_{\bullet,K} &= \epsilon_{\bullet,K} \\ e_{\bullet,K} &= b_{\bullet,K}\end{aligned}\tag{18.11}$$

where  $\epsilon_K$  is white random noise with PSD  $S_K = K^2$  and autocorrelation  $\sigma_K^2 = S_K\delta$ . Notably, the autocorrelation of  $e_{\bullet,K}$  is given by  $\sigma_{e_{\bullet,K}}^2 = \sigma_K^2|t|$  which is unbounded in time.

However, it is impossible to exactly model the bias instability with a finite-dimensional stochastic state-space model as it is an odd-power of frequency, so an approximation must be made for stochastic state-space models which is typically a first-order stationary Gauss-Markov model, i.e., an Ornstein-Uhlenbeck process or discrete-time AR(1) process, or higher-order AR( $n$ ) models. For example, a first-order stationary Gauss-Markov model approximation for the bias instability is given by

$$\dot{b}_{\bullet,B} = -\frac{1}{\tau_{\bullet,B}}b_{\bullet} + \epsilon_{\bullet,B}e_{\bullet,B} = b_{\bullet,B}\tag{18.12}$$

where  $\tau_{\bullet,B} > 0$  is the correlation time of the process and  $\epsilon_{\bullet,B}$  is white random noise with PSD  $S_B = \frac{2\ln(2)}{0.4365^2\pi\tau_B}\mathcal{B}^2$  and variance  $\sigma_B^2 = S_B$ . Notably, the autocorrelation of  $e_{\bullet,B}$  is given by  $\sigma_{e_{\bullet,B}}^2 = \frac{1}{2}\tau_{\bullet,B}\sigma_B^2 \exp(-|t|/\tau_{\bullet,B})$  which reaches a bounded steady-state value of  $\bar{\sigma}_{e_{\bullet,B}}^2 = \frac{1}{2}\tau_{\bullet,B}\sigma_B^2$ .

With these three models, one can obtain the following stochastic state-space model for the stochastic error

$$\begin{aligned}\begin{bmatrix}\dot{b}_{\bullet,B} \\ \dot{b}_{\bullet,K}\end{bmatrix} &= \begin{bmatrix}-\frac{1}{\tau_{\bullet,B}} & 0 \\ 0 & 0\end{bmatrix} \begin{bmatrix}b_{\bullet,B} \\ b_{\bullet,K}\end{bmatrix} + \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix} \begin{bmatrix}\epsilon_{\bullet,B} \\ \epsilon_{\bullet,K}\end{bmatrix} \\ e_{\bullet} &= [1 \quad 1] \begin{bmatrix}b_{\bullet,B} \\ b_{\bullet,K}\end{bmatrix} + \epsilon_{\bullet,N}\end{aligned}\tag{18.13}$$

where the covariance of  $\epsilon_{\bullet,N}$  is  $S_N$  and the covariance matrix of  $[\epsilon_{\bullet,B} \ \epsilon_{\bullet,K}]^T$  is

$$\begin{bmatrix}S_B & 0 \\ 0 & S_K\end{bmatrix}\tag{18.14}$$

This continuous-time model can be discretized at  $\Delta t \ll \tau_{\bullet,B}$  as

$$\begin{aligned}\begin{bmatrix}b_{\bullet,B,k} \\ b_{\bullet,K,k}\end{bmatrix} &= \begin{bmatrix}\exp\left(-\frac{\Delta t}{\tau_{\bullet,B}}\right) & 0 \\ 0 & 1\end{bmatrix} \begin{bmatrix}b_{\bullet,B,k-1} \\ b_{\bullet,K,k-1}\end{bmatrix} + \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix} \begin{bmatrix}\epsilon_{\bullet,B,k-1} \\ \epsilon_{\bullet,K,k-1}\end{bmatrix} \\ e_{\bullet,k} &= [1 \quad 1] \begin{bmatrix}b_{\bullet,B,k} \\ b_{\bullet,K,k}\end{bmatrix} + \epsilon_{\bullet,N,k}\end{aligned}\tag{18.15}$$

with the covariance of  $\epsilon_{\bullet,N,k}$  is  $S_N/\Delta t$  and the covariance matrix of  $[\epsilon_{\bullet,B,k-1} \ \epsilon_{\bullet,K,k-1}]^T$  is

$$\begin{bmatrix}S_B\Delta t & 0 \\ 0 & S_K\Delta t\end{bmatrix}\tag{18.16}$$

Noticeably with respect to Bayesian state estimation, it is not possible to separately estimate  $\bar{b}_{\bullet}$ ,  $b_{\bullet,B,k}$ , and  $b_{\bullet,K,k}$  through measuring  $e_{\bullet,k}$  which is a sum of their contributions. In this case, one typically models

the bias and noise for discrete-time estimation as either a random walk, i.e., using the  $N$  and  $K$  contributions as

$$\begin{aligned} b_{\bullet,k} &= \epsilon_{\bullet,k-1} \\ e_{\bullet,k} &= b_{\bullet,k} + \epsilon_{\bullet,e,k} \end{aligned} \quad (18.17)$$

or a Gauss-Markov process, i.e., using the  $N$  and  $\mathcal{B}$  contributions as

$$\begin{aligned} b_{\bullet,k} &= \exp\left(-\frac{\Delta t}{\tau_{\bullet,\mathcal{B}}}\right) b_{\bullet,k-1} + \epsilon_{\bullet,k-1} \\ e_{\bullet,k} &= b_{\bullet,k} + \epsilon_{\bullet,e,k} \end{aligned} \quad (18.18)$$

where  $\epsilon_{\bullet,k-1}$  and  $\epsilon_{\bullet,e,k}$  are white noise with inflated variances from the ideal to account for the initial uncertainty due to the unknown  $\bar{b}_{\bullet}$ , but also for the unmodeled linear and nonlinear scale factor errors, non-orthogonality, temperature variation, vibration-induced noise, acceleration sensitivity, etc. Typical values to use to inflate these variances from data range from  $2 - 10\times$  depending on your IMU grade. With the random walk model, the bias state variance grows unbounded and linearly with time when the bias is unobservable which is not true of the bounded physical sensor bias. When the bias becomes observable through an aiding measurement and vehicle motion, a large growth of bias variance may cause the estimator gain to be unreasonably large. With the Gauss–Markov model, the bias state variance is bounded even when the bias is unobservable. However, the bias state estimate will tend to zero during time update steps which may be relevant depending on the time duration of unobservability of the bias state and the Gauss–Markov model correlation time.

## References

For more information, please refer to the following

- J. A. Farrell, F. O. Silva, F. Rahman and J. Wendel, “Inertial Measurement Unit Error Modeling Tutorial: Inertial Navigation System State Estimation with Real-Time Sensor Calibration,” in IEEE Control Systems Magazine, vol. 42, no. 6, pp. 40-66, Dec. 2022
- S Gleason and D Gebre-Egziabher, “GNSS Applications and Methods,” Artech House, 2012
- IEEE Std. 647-2006 IEEE Standard Specification Format Guide and Test Procedure for Single-Axis Laser Gyros

## 18.2 Electromagnetic Sensors

### Magnetometers

A **magnetometer** is a sensor for measuring the strength and/or the direction of magnetic fields and can be based on a variety of techniques. The simplest is the use of a **magnetic compass** which is a permanently magnetized needle that is mounted so that it can pivot in the horizontal plane. With no interference, e.g., gravity, the needle will align itself exactly along the *local* magnetic field vector. Magnetic compasses typically have a response time less than  $1 \mu\text{s}$  and often can be sampled up to 1000 Hz. Many magnetic

compasses provide accurate measurements within  $1^\circ$  of the magnetic field direction, for which the underlying compass must reliably resolve  $0.1^\circ$  variations.

Excepting local variations in magnetic materials, a magnetic compass nominally points in the direction of **magnetic north**, i.e., the direction of the *horizontal* component of the geomagnetic field relative to the Earth's ellipsoid at a given location. This direction is only approximately **geographic north**, also known as **true north**, i.e., the point about which the Earth rotates, and near the poles becomes a poor approximation. This error between the direction of local magnetic north and geographic north is called the **magnetic declination**, also known as the **magnetic variation**. The geomagnetic field suffers from localized variations which can also change over time. These variations are especially noticeable at high or low latitudes where one can measure a significantly large magnetic variation. Furthermore, the strength along these field lines vary based on location, thus, calibration is necessary for useful measurements. Measurements of the geomagnetic field are publicly available, e.g., the World Magnetic Map (WMM) and the International Geomagnetic Reference Field (IGRF).

A **three-axis magnetometer (TAM)** provides measurements of the Earth's magnetic *vector* coordinates and is often compensated for any dip errors with an additional sensor that estimates the pitch and roll angles. This algorithm is called **tilt compensation**. However, a TAM is still susceptible to errors. In particular, the Biot-Savart Law states that ferromagnetic parts, and/or magnetic effects induced by electrical circuits will impact the TAM measurement as well potentially created external magnetic disturbances. The presence of any additional constant magnetic fields will form distortions that either bias the Earth's magnetic field, i.e., a **hard-iron distortion**, or distort the Earth's magnetic field relative to the TAM's attitude, i.e., a **soft-iron distortion**. Additionally, axes misalignment of the individual magnetometers may shift or rotate the axis of measurement. Non-magnetic effects may further locally distort the TAM measurement, e.g., rotating the TAM too fast while collecting measurements, roughly for angular speeds greater than  $\approx 150^\circ/\text{s}$ . However, careful calibration of the TAM can greatly reduce these errors.

Thus, taking into account the major errors for aerospace attitude determination or navigation systems, a suitable three-axis magnetometer measurement model can be written as

$$\vec{y}_{m,B} = \begin{bmatrix} 1 + s_{m,x} & \sigma_{m,xy} & \sigma_{m,xz} \\ 0 & 1 + s_{m,y} & \sigma_{m,yz} \\ 0 & 0 & 1 + s_{m,z} \end{bmatrix} \left( C_{s-i}^{-1} (\vec{h}_B + \vec{d}_{m,B}) + \vec{b}_{B,h-i} + \vec{\epsilon}_{m,B} \right) \quad (18.19)$$

where  $\vec{y}_{m,B}$  is the magnetic field measurement in the magnetometer's body-fixed frame,  $\vec{h}_B$  is the Earth's magnetic field vector in the body-fixed frame,  $s_{a,z}$ ,  $s_{a,y}$ , and  $s_{a,z}$  are the linear magnetometer scale factor errors in each axis,  $\sigma_{m,xy}$ ,  $\sigma_{m,xz}$ , and  $\sigma_{m,yz}$  are the magnetometer non-orthogonality errors in each axis,  $\vec{d}_{m,B}$  is the external magnetic disturbance,  $\vec{b}_{B,h-i}$  is the hard-iron distortion bias,  $C_{s-i}$  is the soft-iron distortion rotation matrix, and  $\vec{\epsilon}_{m,B}$  is the random error. Notably, in calibration, one can often remove or greatly reduce the scale factor and non-orthogonality errors which results in the measurement model

$$\vec{y}_{m,B} = C_{s-i}^{-1} (\vec{h}_B + \vec{d}_{m,B}) + \vec{b}_{B,h-i} + \vec{\epsilon}_{m,B} \quad (18.20)$$

which also allows one to perform an additional hard-soft-iron (HSI) calibration to form the HSI-calibrated three-axis magnetometer measurement model as

$$\vec{y}_{m-HSI,B} = C_{s-i} (\vec{y}_{m,B} - \vec{b}_{B,h-i}) = \vec{h}_B + \vec{d}_{m,B} + \vec{\epsilon}_{m-HSI,B} \quad (18.21)$$

where

$$\vec{\epsilon}_{m-HSI,B} = C_{s-i} \vec{\epsilon}_{m,B} \quad (18.22)$$

## Radar Sensors

Energy wave sensors use electromagnetic (EM) or acoustic signals to form measurements and these can be **active**, i.e., the energy is generated by the sensor, or **passive**, i.e. the energy is generated by the environment. Active energy wave sensing uses active sonar, radar, and lidar while passive energy wave sensing uses passive sonar and optical cameras. These energy wave sensors can produce one-, two-, or three-dimensional signals depending on the type and number of sensors, where two- or three-dimensional signals are also known as **images**.

**Radio detection and ranging (radar)** is a remote sensing technique that uses active radio wave transmission and reception to derive pseudorange, pseudoazimuth, pseudoelevation, and pseudorange rate measurements. Radar use antennas to both transmit and receive the radio signals. When a single antenna is used for both or separate transmitter and receiver antennas are used in close proximity, one has a **mono-static radar**. When separate transmitter and receiver antennas are used with spatial separation, one has a **bi-static radar**. When three or more transmitters and receivers are used with spatial separation, one has a **multi-static radar**.

When radio waves contact a target, also known as **illumination**, they are usually reflected or scattered in many directions, although some of them will be absorbed and penetrate into the target. Radar signals are reflected especially well by materials of considerable electrical conductivity—such as most metals, seawater, and wet ground. If electromagnetic waves traveling through one material meet another material, having a different dielectric constant from the first, the waves will reflect or scatter from the boundary between the materials. This means that a solid object or a significant change in the atomic density between the object and what is surrounding it will usually scatter radio waves from its surface. This is particularly true for electrically conductive materials such as metal and carbon fiber, making radar well-suited to the tracking of flight vehicles, ground vehicles, and surface vessels.

Radar waves scatter in a variety of ways depending on the wavelength of the radio wave and the shape of the target. If the wavelength is much shorter than the target's size, the wave will bounce off in a way similar to the way light is reflected by a mirror. If the wavelength is much longer than the size of the target, the target may not be visible because of poor reflection. The most reflective targets for short wavelengths have 90° angles between the reflective surfaces. A **corner reflector** consists of three flat surfaces meeting like the inside corner of a box. The structure will reflect waves entering its opening directly back to the source. They are commonly used as radar reflectors to make otherwise difficult-to-detect targets easier to detect. Conversely, targets intended to avoid detection will not have inside corners or surfaces and edges perpendicular to likely detection directions, e.g., stealth aircraft. The extent to which a target reflects or scatters radio waves is called its **radar cross section**,  $\sigma$ .

The **radar range equation** provides an equation for the power returning to the receiving antenna,  $P_r$ , as a function of the transmitter power,  $P_t$ , as

$$P_r = \frac{G_t A_r \sigma F^4}{(4\pi)^2 R_t^2 R_r^2} P_t \quad (18.23)$$

where  $G_t$  is the gain of the transmit antenna,  $A_r$  is the effective aperture of the receiving antenna,  $F$  is the pattern propagation factor which accounts for multipath and shadowing,  $R_t$  is the distance from the

transmitter to the target, and  $R_r$  is the distance from the target to the receiver. Notably, one can also

$$A_r = \frac{G_r \lambda^2}{4\pi} \quad (18.24)$$

where  $\lambda$  is the transmitted wavelength and  $G_r$  is the gain of the receiving antenna. Also, if the transmitter and receiver are the same antenna, then  $R_t = R_r = R$  and

$$P_r = \frac{G_t A_r \sigma F^4}{(4\pi)^2 R^4} P_t \quad (18.25)$$

which shows the distant target received power is relatively very small.

The **radar Doppler shift**,  $f_D$ , is related to the transmit frequency,  $f_t$  as

$$f_D = 2 \frac{\dot{\rho}}{c} f_t \quad (18.26)$$

where  $\dot{\rho}$  is the range rate between the target and radar. Of note, when the transmit frequency is pulsed at a frequency of  $f_r$ , the resulting frequency spectrum will contain harmonic frequencies above and below  $f_t$  with a distance of  $f_r$ . Thus, the Doppler shift is unambiguous if

$$|f_D| < \frac{1}{2} f_r \quad (18.27)$$

where  $f_r/2$  is known as the **Nyquist frequency** since the returned frequency cannot be distinguished from the shifting of a harmonic frequency above or below. When substituting with  $f_D$ , one has

$$|\dot{\rho}| < \frac{c f_r}{4 f_t} \quad (18.28)$$

**Shot noise** is produced by electrons in transit across a discontinuity, which occurs in all detectors. Shot noise is the dominant source in most receivers. There will also be **flicker noise** caused by electron transit through amplification devices, which is reduced using **beat frequency** amplification of mixed signals. Another reason for beat frequency processing is that for fixed fractional bandwidth, the instantaneous bandwidth increases linearly in frequency which improves range resolution. The one notable exception to beat frequency radar systems is **ultra-wideband (UWB) radar**. Noise is also generated by external sources, most importantly the natural thermal radiation of the background surrounding the target of interest. In modern radar systems, the internal noise is typically about equal to or lower than the external noise.

The **thermal noise** is given by  $k_B T B$ , where  $T$  is the temperature,  $B$  is the bandwidth after filtering, and  $k_B$  is Boltzmann's constant. Filtering allows the entire energy received from an object to be compressed into a single range, Doppler, elevation, or azimuth bin. Mathematically, this infers that one could obtain perfect, error-free, detection if one could compress the energy into an infinitesimal time slice. However, while time is arbitrarily divisible, current is not because the quantum of electrical energy is an electron, and so the best that one can do is filter the energy into a single electron. Furthermore, since this single electron would be moving at a certain temperature, i.e., the Planck spectrum, this noise source would still remain in the best possible case.

One technique to obtain a range measurement using radar is based on pulsed radar through the **time-of-flight (TOF)** measurement, i.e., by transmitting a short pulse of radio signal and measuring the time it takes

for the reflection to be received. The range to a target that reflects the radar signal is proportional to the time difference between the time of transmission (TOT) and the time of arrival (TOA) given by

$$r = \frac{c(TOF)}{2} = \frac{c(TOA - TOT)}{2} \quad (18.29)$$

where  $r$  is the relative range and  $c$  is the speed of light through the medium. Through the use of a **duplexer**, a mono-static radar switches between transmitting and receiving a pulse at a predetermined rate, i.e., the **pulse repetition frequency (PRF)**. This pulse process imposes a minimum and maximum measurable range for the radar. Smaller PRFs, i.e., longer times between pulses, are also needed to increase the maximum range. Short pulses are needed to decrease the minimum range, but with less total energy, any returns are harder to detect and reduce the effective range of the radar. This could be offset by using more pulses, but this would only serve to further decrease the maximum range. These two effects infer good short range and good long range measurements are not feasible in basic radar design and long-range radars tend to use long pulses with smaller PRFs and short-range radars use smaller pulses with higher PRFs. However, many modern radars can dynamically change their PRF, thereby dynamically changing their effective range for specific applications, e.g., object tracking. Other modern advanced radars fire can two pulses simultaneously, one for short range and one for long range, which are then processed separately.

Another technique to obtain a range measurement using radar is based on frequency modulation. In these systems, the frequency of the transmitted signal is changed over time. Since the signal takes a finite time to travel to and from the target, the received signal is a different frequency than what the transmitter is broadcasting at the time the reflected signal arrives back at the radar. By comparing the frequency of the two signals the difference can be easily measured. A further advantage is that the radar can operate effectively at relatively low frequencies. This technique can be used in **frequency modulated, continuous wave (FMCW) radar** where a carrier signal is frequency modulated in a predictable way, typically varying up and down with a sine wave or sawtooth pattern. This signal is sent out from one antenna and received on another and continuously compared using a simple beat frequency modulator that produces a frequency tone from the returned signal and a portion of the transmitted signal. The modulation index riding on the receive signal is proportional to the time delay between the radar and the reflector. The frequency shift becomes greater with greater time delay. The frequency shift is directly proportional to the distance traveled.

The two techniques outlined above both have their disadvantages. The TOF radar technique has an inherent tradeoff in that the accuracy of the range measurement is inversely related to the length of the pulse, while the energy, and effective range, is directly related. Increasing power for longer range while maintaining accuracy demands extremely high peak power in the pulsed signal. The FMCW technique spreads this energy out in time and thus requires much lower peak power compared to pulse techniques, but requires some method of allowing the sent and received signals to operate at the same time, often demanding two separate antennas. Thus, modern radar combine these two techniques using **pulse compression** which start with a longer pulse that is also frequency modulated. Spreading the broadcast energy out in time means lower peak energies can be used, but upon reception, the signal is sent into a system that delays different frequencies by different times. The resulting output is a much shorter pulse that is suitable for accurate distance measurement, while also compressing the received energy into a much higher energy peak and thus reducing the **signal-to-noise ratio (SNR)**.

To measure the bearing angles to a target, several receivers measure the relative arrival time to each which provides a bearing angle or with an array, one can measure the relative amplitude in the beams formed

through a process called **beam-forming**. Advanced radars often have multiple beams to provide all-round cover while simple ones only cover a narrow arc, although the beam may be rotated by mechanical scanning. The target signal together with noise is then passed through various forms of signal processing, which for simple sonars may be just energy measurement, and one uses classic detection algorithms, e.g., hypothesis testing, to decides whether the output is the required signal or noise. Further processing may be done to classify the target.

**Aperture radar** use radar arrays to illuminate an entire region, known as an **aperture**, whose output is a two- or three-dimensional image. Another technique is to use a single radar in motion as a **synthetic aperture radar (SAR)**. To create a SAR image, successive pulses of radio waves are transmitted to illuminate a target scene, and the echo of each pulse is received and recorded. The pulses are transmitted and the echoes received using a single beam-forming antenna with varying wavelengths. As the SAR moves, the antenna location relative to the target changes with time. Signal processing of the successive recorded radar echoes allows the combining of the recordings from these multiple antenna positions. This process forms the synthetic aperture and allows the creation of higher-resolution two- and three-dimensional images than would otherwise be possible with a given physical antenna.

## Optical Sensors

Optical sensors can be categorized into **light detection and ranging (lidar)**, i.e., active optical sensors, and **electro-optical and infrared (EO/IR)** sensors, i.e., passive optical sensors. These sensors detect the EM spectrum from infrared to visible to ultraviolet light and convert these EM signals to electric signals. These sensors can be used to produce one-, two-, or three-dimensional signals.

### Lidar

A **camera** is an EO/IR sensor that generates images and is specified by the wavelength band or bands used. An **infrared camera**, **near-infrared camera**, **visual camera**, **multispectral camera** which uses three to ten wide bands of infrared light, and **hyperspectral camera** which uses hundreds of narrow bands of optical light. Two types of cameras are commonly employed for aerospace vehicles, **monocular cameras** which use a single lens to form a two-dimensional image and **stereoscopic cameras** which use two cameras to capture two two-dimensional images which can be processed using **stereoscopic vision** techniques to form a three-dimensional image, e.g., stereo matching or machine learning. If the processed three-dimensional image does provide any wavelength information about the received signal, e.g., hue or shade, then the image becomes a **point cloud** as with lidar.

When utilizing cameras, one must consider the frame transformations from field-of-view (FOV) of the camera to the two-dimensional image frame.

### Stereo matching

To form consistent feature points from camera images, also known as **keypoints**, one must use algorithms from **computer vision**, also known as **machine vision**, a field which encompasses the techniques and algorithms for processing images into other information including useful measurements for navigation and perception.

difference of Gaussians (DoG)

scale-invariant feature transform (SIFT) speeded-up robust features (SURF)

An alternative to DoG features from accelerated segment test (FAST) is a corner detection method. binary robust independent elementary features (BRIEF) oriented FAST and rotated BRIEF (ORB)

(KAZE) and advanced KAZE (A-KAZE)

In recent years, **convolutional neural networks (CNN)** and **vision transformers** neural networks have performed very well in perception systems. These data-driven algorithms use truth databases to form object detection and measurements given the image information. From two-dimensional images, one can form bearing measurements as part of the CNN output. For three-dimensional images and point clouds, one can also obtain range measurements sometimes referred to as depth measurements.

## 18.3 Radionavigation Systems

### Navigation Satellite Systems

**Navigation satellite systems (NSS)**, also known as **NavSat**, are a specific type of radionavigation/radiopositioning system which transmit timed radio signals from artificial satellites orbiting Earth, known as a **satellite constellation**. In the future, one could also imagine using a satellite constellation around other planetary bodies which would assist in navigation on the surface of that body. As with most space technologies, these types of systems are notably expensive to launch as satellites are not cheap, especially those that must provide necessary information for navigation, i.e. positions of the broadcasting satellites, the timing of transmission, and the error models of significant system parameters. Thus, only governments have built NSS. There are three general types of NSS: global, regional, and augmentation.

There are currently four **global navigation satellite systems (GNSS)** constellations operating. The first GNSS installed was the United States' **Global Positioning System (GPS)** which is composed of 31-33 satellites stationed in 6 orbital planes. GPS became globally available in 1994. The second installed constellation of GNSS was Russia's **GLObal NAVigation Satellite System (GLONASS)** which is composed of 24-26 satellites stationed in 3 orbital planes. GLONASS was fully operational in 1995 over Russia, but reached a period of decline soon thereafter, but became globally available in 2016. The third constellation was the European Union's **galileo positioning system (GALILEO)** composed of 26 satellites in 3 orbital planes and became globally available in 2020. The fourth constellation was China's **BeiDou navigation satellite System (BDS)**, which is composed of 35-40+ satellites in 7 orbital planes and became globally available in 2020.

There are currently two **regional navigation satellite systems (RNSS)** constellations operating: India's **NAVigation with Indian Constellation (NAVIC)** with 7 satellites which became available in 2019 and Japan's **Quasi-Zenith Satellite System (QZSS)** composed of 4 satellites in elliptical GSO which became fully operational in 2018 with plans in 2023 to become independent of GPS with 7 satellites as an RNSS.

There are currently four **satellite-based augmentation systems (SBAS)** currently operating: the United States's **Wide Area Augmentation System (WAAS)**, the European Union's **European Geostationary Navigation Overlay Service (EGNOS)**, India's **GPS-aided GEO augmented navigation (GAGAN)**, and Japan's **Multi-functional Satellite Augmentation System (MSAS)**. SBAS only exist for enhancing the capabilities and performance of GPS by broadcasting additional corrections through GPS wavelengths to end users. These corrections are computed by ground station networks which use an uplink to the communication satellite constellation to broadcast corrections to users over a large region of the Earth's surface. **Ground-Based Augmentation Systems (GBAS)** which also exist for GPS, but use ground stations to broadcast computed corrections to GPS users within a smaller region than SBAS, but with potentially better corrections.

The history of NSS is relatively short in the history of navigation, but in recent times has become a ubiquitous technology in society. Below is an abbreviated timeline of its history.

- 1973: GPS proposed and funded by U.S. Department of Defense (DoD) as navigation technology for the military
- 1976: GLONASS proposed and funded by USSR
- 1978: GPS fully operational
- 1980: GPS civilian use allowed for signals with much poorer accuracy than military signals
- 1994: GPS globally available
- 1995: GLONASS fully operational at 24 satellites
- 1999: US military selectively denies access to India during Kargil War
- 2001: GLONASS at 6 operational satellites
- 2002: European Union (EU) and European Space Agency (ESA) agree to fund Galileo GNSS
- 2003: China joins the Galileo project
- 2003: Federal Aviation Administration (FAA) commissions WAAS
- 2006: China opts instead to upgrade then-regional BDS instead of joining Galileo GNSS
- 2006: WAAS operational
- 2008: GLONASS at 12 satellites
- 2009: ESA commissions EGNOS
- 2012: EGNOS operational
- 2016: GLONASS globally available at 24 satellites
- 2019: Galileo fully operational
- 2020: BDS globally available
- 2020: Galileo globally available

NSS are complex systems made up of three segments. The first is the **space segment** which contains the artificial satellites in the constellations, also known as its **space vehicles (SV)**. The second is the **ground segment** which consists of the master control station (MCS) and monitor stations of the space vehicles. Lastly, there is the **user segment** which is made up of millions of users using a wide variety of GNSS navigation devices which at the most basic level must have the following components: an antenna, a receiver-processor, a highly stable clock, and relevant PNT algorithms.

In general, geocentric satellites can have a variety of orbits, but one way to characterize them is by orbital altitude, of which there are four types. **Low Earth orbits (LEO)** have altitudes below 2,000 km. **Medium Earth orbits (MEO)** have altitudes between 2,000 and 35,786 km. **Geosynchronous orbits (GSO)** are a unique orbit altitude at 35,786 km as since the satellites are orbit at the same rate as the average rotation of the Earth, i.e. one rotation per day. Furthermore, a specific GSO is the **geostationary orbit (GEO)** where the satellite remains stationary overhead of a point on the Earth's equator, thus these are highly useful for communication satellites to service particular locations on Earth. Finally, **high earth orbits (HEO)** have altitudes above 35,786 km. For GNSS satellites, orbits near 20,200 km, i.e. MEO, are prescribed. This results in an orbital period of 12 hours and allows satellites to travel straight above the same location twice per day. Notably some GNSS orbits are also at GSO or GEO as well as the RNSS and SBAS orbits. The orbits and number of satellites in the constellation are designed to provide enough measurements for pseudorange multilateration.

The control segment of a GNSS is made up of a network of monitor stations with a centralized **master control system (MCS)** which for GPS is located at Schriever Space Force Base (SFB) in Colorado. The MCS has many duties to perform as part of the overseer of a GNSS including commanding the space vehicles to perform any orbital manuevers, estimating the navigational state of the satellites, generating the timing and navigational messages in the GNSS signal(s), monitoring the system integrity, and updating the processor programs of the GNSS satellite digital systems if necessary. In order to obtain information about the status of the satellites for the purposes of command and control, the MCS collects data from **monitor stations** around the globe. GPS has six monitor stations. These stations track and collect the GNSS signals and local atmospheric data and feed this information directly to the MCS for analysis. In addition, to communicate with the GNSS satellites, the MCS uses ground antennas stationed around the globe to uplink with the satellites, i.e. send commands. GPS has four ground antennas. As an operational system, there are several factors that must be continuously monitored. The first is that any of the various subsystems of the satellites will have different effects on the observable behavior of the satellites. The control segment monitors the satellites for a variety of faults or errors which make using GNSS unsafe for users. Secondly, the satellite clocks are not perfect and therefore the control segment must monitor and update the clocks to provide accurate timing for radionavigation. Another major consideration is the perturbations in the satellites' orbits caused by gravitational effects and space weather. The MCs estimates the navigational state of each satellite and can command corrective maneuvers if required to keep the constellation within operational limits. It also can command orbital maneuvers to de-service satellites when they are no longer operational.

The original users of GPS and GNSS were the military who desired a global solution to navigation for command and control of their assets. This is still a major part of the GNSS systems which contain encrypted military channels that enable high precision GNSS navigation. However, civilian navigation is still possible for pedestrians and a variety of vehicles including automobiles, aircraft, cyclists, ships, and even spacecraft in LEO. Some other uses for GNSS include mapping the Earth (e.g. Geographic Information Systems (GIS), archaeology, and surveying), synchronized timing (e.g. cell networks), mobile satellite communications (provides antenna pointing information in the global ECEF frame), weather prediction (infers atmospheric conditions from GNSS radio signals), and much more.

However, GNSS still has some shortcomings and challenges for navigational use in all circumstances, the primary being a requirement for a clear Line-of-Sight (LOS) to the satellites for the signal to be measurable by GNSS antennas. This causes users to lose the signal indoors and in canopies or canyons, either natural or urban, or which the latter is a large area of need for reliable navigation in the near future. Another challenge

in GNSS is that atmospheric transmission introduces significant error which has been a major subject of research in 2000-2010s. Third, is the possibility of jamming the GNSS signal due to its low power being susceptible to any elevated noise which can be intentional or unintentional. Lastly, there is the threat of **spoofing**, i.e. creating a false GPS signal which replicates the expected external radio signal. This can make receiver estimate completely a wrong trajectory. This last subject matter is of critical importance in the modern age of connected systems and thus there is a lot of current research into **Resilient PNT** to develop algorithms and systems to overcome system attacks.

The International GNSS Service (IGS) publishes standards for communicating GNSS information. The primary format is the **receiver independent exchange (RINEX)** which has a standard for storing the observation data, the navigation message, and the meteorological data. The RINEX format was adopted by the GNSS community to catalog raw GPS data for post-processing by any potential user. This allows more potentially more accurate positioning to be done with information better than or unknown to the original receiver, e.g. better atmospheric or orbital models than at the actual time of measurement. This format can be processed by a multitude of applications by using standard information parameters. This format has evolved over time with the development of GPS technologies with the most recent being RINEX 3. It should also be noted that often RINEX files are compressed as ASCII-based CompactRINEX or CRINEX using the Hatanaka compression scheme. As such, there are open-source parsers available, e.g. for python the `georinex 1.14.1` package is available.

The observation data contains the observations or measurements from all GNSS frequencies and all satellites. This includes the pseudoranges, Doppler shift and the carrier phase which will be discussed later in the course. The navigation message contains the ephemeris information from the MCS broadcasted model while the IGS precise orbit solutions are published in the *sp3* format. The IGS also publishes other supplementary materials for GNSS data processing. This includes IGS monitoring station position and velocity solutions which are published in the Station Independent Exchange (SINEX) format, the Earth rotation parameter (erp) files, the ionospheric exchange (IONEX) format for TEC grid products, and the Zenith path delay products computed at each IGS monitoring station in the format Tropo SINEX.

A second important piece of GNSS are the individual time standards used by each GNSS constellation. GPS uses **GPS Time (GPST)** as a continuous time scale and theoretically accurate to about 14 ns. However, most receivers lose accuracy in the interpretation of the signals and are only accurate to 100 ns. The GPST differs from TAI and UTC, but remains a constant 19 seconds behind TAI as it does not implement leap seconds. Periodic corrections are performed to the on-board clocks in the satellites to keep them synchronized with ground clocks. The GPS navigation message includes the difference between GPST and UTC. GPST 0 occurred at 12:00AM on January 6, 1980. As of July 2015, GPST is 17 seconds ahead of UTC because of the leap second added to UTC on June 30, 2015. GLONASS uses **GLONASS Time (GLONASST)** as generated by the GLONASS Central Synchroniser and is typically accurate to within 1,000 ns. Unlike GPS, the GLONASS time scale implements leap seconds, like UTC. Galileo uses **Galileo System Time (GST)** as a continuous time scale which is generated on the ground at the Galileo Control Centre in Fucino, Italy, by the Precise Timing Facility, based on averages of different atomic clocks and maintained by the Galileo Central Segment and synchronized with TAI with a nominal offset below 50 ns. According to the European GNSS Agency Galileo is accurate to about 30 ns. The Galileo navigation message includes the differences between GST, UTC and GPST. BeiDou uses **BeiDou Time (BDT)** as a continuous time scale starting at 1 January 2006 at 0:00:00 UTC and is synchronized with UTC within 100 ns. GPS satellites have at least two onboard caesium and as many as two rubidium atomic clocks. Each Galileo satellite has two passive hydrogen maser

and two rubidium atomic clocks for onboard timing. The vast majority of GNSS receivers use quartz clocks, typically symbolized by *OCXO*. Quartz clocks are used because they are relatively inexpensive, compact, use low amounts of power, and possess a long life, but are not at the atomic clock standard used by the GNSS SV. However, more expensive receivers may allow an external atomic timing source, e.g. a caesium or rubidium clock. GNSS systems use the **WGS84 parameters** which defines the following constants for Earth.

## GPS Signal Design

GPS satellites, a.k.a. **space vehicles (SV)** in GPS documentation broadcast the GPS signals as **right hand circular polarized (RHCP)** signals, i.e. the electromagnetic wave is polarized in a direction that oscillates in a continuous rotating fashion with a rotation defined by the right hand rule. In particular, the electrical field vectors of GPS signals have constant magnitude, but the electric field traces out a helix in direction of propagation.

The GPS carrier signals are broadcast on multiple frequencies. There are five GPS frequencies all within the **L-band** of the radio spectrum, i.e. 1 to 2 gigahertz (GHz), three of which are used as navigation frequencies. The L1 GPS carrier is broadcast at 1575.42 MHz and carries the L1C or *Civilian*, also known as the *Course/Acquisition (C/A)*, the encrypted *Precision* or P(Y) code used by the U.S. DoD, and a newer encrypted DoD *Military-code (M-code)*. The L2 GPS carrier is broadcast at 1227.60 MHz, and carries the L2C or *Civilian*, the second piece of the *Precision* P(Y), and the second piece of the *Military-code (M-code)*. The L5 GPS carrier is broadcast at 1176.45 MHz and is used for safety-of-life (SOL) information. It should be noted that the L3 & L4 carriers do not provide information for navigation, but instead are used for nuclear warhead enforcement and atmospheric research, respectfully.

GPS satellites in GPS documentation transmit modulated radio signals use multi-frequency carrier signals encoded with digital code through **binary code modulation (BCM)** at a much slower frequency than the carrier. These signals are broadcast in a cone of approximately  $28^\circ$ , which at a MEO of 20,200 km, this is enough to cover one-side of the Earth's geoid. BCM encodes 0's and 1's on the carrier signal through a variety of techniques including **amplitude modulation (AM)** for analog signals or **amplitude-shift keying (ASK)** for digital signals, **frequency modulation (FM)** for analog signals or **frequency-shift keying (FSK)** for digital signals, and **phase modulation (PM)** for analog signals or **phase-shift keying (PSK)** for digital signals. Of these, PSK has become a very popular form of radio communication technology including wireless local area networks (WLAN), radio frequency identification (RFID), Bluetooth, and GNSS. In addition, another form of modulation developed in particular for GNSS is **binary offset carrier (BOC) modulation**, also known as **split spectrum modulation**.

GPS phase modulation uses several techniques in order to provide a *spread spectrum* technique on the carrier in order to achieve better signal-to-noise ratio (SNR), provide more accurate ranging, less radio frequency interference (RFI), and increased security. However, this technique also weakens the signal tremendously, so much so that sensing the GPS signal is often compared metaphorically to identifying a 25-watt light bulb from 10,000 miles away.

As GPS signals are broadcast from multiple satellites at the same frequencies, GPS uses **code division multiple access (CDMA)** to track each signal. CDMA allows multiple operating SVs with unique IDs (SV ID) to be tracked where each SV ID broadcast is a unique pseudorandom noise (PRN) number, a predictable signal, but statistically resembles noise where each PRN does not correlate well with each other allowing for correlation tracking methods to be used to distinguish SVs. Lastly, the GPS signals each carry a navigation

message modulated on the PRN and carrier which is much slower than both. For GPS, the civilian codes are modulated using PSK (PSK) for the L1 C/A and L2C which modulates the code between 2 phases separated by  $180^\circ$  while BOC is used for the L1C signal to allow its simultaneous transmission at L1 with low interference with BPSK which has a centered peak around the carrier frequency.

As GPS depends on accurate timing of the signal transmissions and uses precise signal processing to overcome its low signal power, all components of GPS digital systems are multiples of a standard rate known as the **fundamental clock rate**, denoted by  $F_0$ , and equal to 10.23 MHz. This is set by caesium and rubidium clocks onboard the SVs. Of note, the three navigation L-band carriers are set at  $154F_0$  for L1,  $120F_0$  for L2, and  $115F_0$  for L5. However, the BCM signals are much slower than the clock rate, e.g. the L1 C/A PRN oscillates at 0.1,  $F_0 = 1.023$  Mbps, and the L1 C/A navigation message oscillates at 50 bps. SVs experience relativistic effects which cause SV clocks to appear to run  $38 \mu\text{s}/\text{day}$  faster rate than Earth clocks. As SVs travel at roughly 3.874 km/s, this produces a special relativistic effect on the SV clocks as appearing to run slower than Earth clocks at roughly  $-7 \mu\text{s}/\text{day}$ . In addition, the weaker gravity generates a general relativistic effect which causes the SV clocks to appear to run faster than earth clocks at about  $45 \mu\text{s}/\text{day}$ . Together, these create an overall . Thus, to produce  $F_0 = 10.23$  MHz in space, the SV clocks are set to 10.22999999543 MHz.

## GPS Signal Processing

The design of the GNSS signal greatly reduces the power of signal which causes the GNSS signals to be less than the natural receiver noise. This is notably the opposite of directional communication satellites which use large dish antennas to focus the signal on both transmit and receive. However, GPS signals are made observable through specialized signal processing. GNSS receiver-processors sense the radio signal using an antenna and process the complex radio signal to extract the relevant PNT information. The antenna serves to convert the radio waves to an electrical current and must have high sensitivity, i.e. gain, for which different types of antennas have been used for different applications and include patch, dipole, quadrifilar, and helix. Due to the wide broadcast beamwidth, these antennas are designed to be omni-directional. Once the signal has been converted to a current, the GPS receiver-processor can process the signal by either radio circuitry, i.e. a **hardware-defined radio (HDR)**, or digital signal processing, i.e. a **software-defined radio (SDR)**. GPS receiver-processors can have a wide variety of sizes and components, but require at least a pre-amplifier, a bandpass filter, a phase-lock loop (PLL), carrier signal demodulation, and a delay-lock loop (DLL).

As the SVs orbit the Earth, the observed carrier frequencies will increase and decrease by up to 5-10 kHz due to the Doppler effect. This change, known as the **Doppler shift** can be predicted as the SV pass across the sky. However, not only do the SVs move, but the GPS receiver may also move which will cause an additional Doppler shift which the receiver must detect and track. This Doppler shift captures the relative line-of-sight (LOS) motion between each SV and the receiver and can be used as the **GPS range rate measurement** in addition to the signal processing of the GPS signal. In order to acquire and track the Doppler shift, the receiver-processor replicates the original frequency carrier and differences it with the observed signal from the antenna. This new differenced signal oscillates with an **intermediate frequency (IF)**, also known as the **beat frequency**. The beat frequency error is then detected and tracked by a PLL in the receiver-processor which allows the carrier signal to be tracked accurately and forms the **GPS carrier phase measurement**. Finally, the receiver-processor demodulates the carrier signal from the encoded signal before passing it to the next tracking stage.

Next, the receiver-processor needs to acquire and track the encoded signals *from each SV*. To do so, the receiver-processor generates the unique PRNs for each SV ID. Recall that the each PRN does not correlate well with other PRNs, i.e. *orthogonal* signals, thereby allowing the receiver-processor to distinguish the different SV signals. For example, the L1C PRN codes consist of 1023 chips and are generated by an *exclusive or* combination of two 10-stage linear feedback shift registers (LFSR), i.e. bit-streams determined by linear function of certain stages, i.e. “taps”. The uniqueness of each L1C PRN is performed by a unique integer delay on second LFSR. Similar to the Doppler shift, the receiver-processor can acquire each PRN by generating replica PRNs and by shifting the PRNs bit-by-bit until one finds the “best fit” from all satellite PRNs, a process also known as **PRN auto-correlation** where matching the signals in time will theoretically create a signal correlation of 1 though this will be slightly lower in practice due to processing errors. Receiver-processors typically use a Fast Fourier Transform (FFT) to perform an efficient and accurate PRN auto-correlation.

After acquiring the PRN signal, a DLL tracks the PRN code shift over time to provide the **GPS pseudorange measurement** as the measured time shift that provides the maximum correlation. This is also known as the **code phase measurement**. In addition, this tracking allows the receiver-processor to read the much slower navigation message that still remains on the signal. The DLL is typically a digital proportional-integral (PI) controller that compares the phase of its last output with the input clock which it integrates and feeds back as the control to all delay elements. Furthermore, to acquire and track multiple satellites simultaneously, receiver-processors run multiple tracking channels in parallel. In particular, each single frequency from a single satellite has its own dedicated channel which notably requires hundreds of channels in multi-frequency, multi-constellation GNSS receiver-processors. This paradigm allows the receivers to maintain accuracy when moving, anti-jamming capability, and shortens the time to first fix on the PRNs. Thus, each channel has two modes of operation: *acquisition* which performs the PRN auto-correlation calculation and *tracking* which uses the DLL to track the slightly varying time shift to obtain the code phase measurement.

## GPS Navigation Message

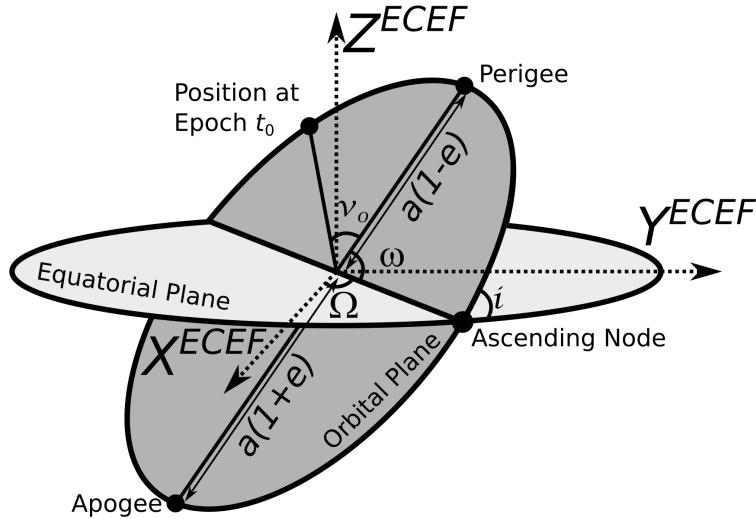
There are three specific formats for the GPS navigation message encoded on the civilian signals. However, each format of the GPS navigation message contains at least the following:

- **Almanac**
  - Information for every SV
  - Atmosphere model
  - Relate GPS time to UTC
  - Updated every 6 days
- **Broadcast Ephemeris**
  - Binary message in SV signal
  - Updated every 2 hours by MCS

The L1 C/A Legacy Navigation (LNAV) message which supports 32 SVs and has five sub-frames. The first frame contains information on the satellite clock and GPS Time. The second and third contains broadcast ephemeris for the particular satellite, and the fourth and fifth contain the almanac which provides coarse orbital elements for *all* SVs, a global ionospheric model, and GPS time relationship to UTC. The entire almanac takes 25 frames to complete (e.g. 12.5 minutes), but is used by receivers to acquire a set of viewable SVs which can be used for replicate PRNs for receivers with a low number of channels. However, once acquired, the specific SV broadcast ephemeris is used for positioning. The L1/2C: Civilian Navigation (CNAV/CNAV-2) which supports 63 SVs and has 12-second 300-bit messages which are analogous to LNAV sub-frames, but with additional information. In particular, this format contains a GPS-to-GNSS time offset which facilitates better interoperability with other GNSS, extra bandwidth which enables a differential correction which can correct the L1 navigation clock, and an alert flag which can be set by the MCS if the SV data cannot be trusted. This alert flag is a much more rapid notification than the LNAV message which is necessary for safety-of-life applications using GPS. Lastly, the GPS time format is also limited within each navigation message format. The LNAV expresses GPST as the *week* and *time of week (TOW)* which has a resolution of 1.5 seconds. The CNAV/CNAV-2 expresses GPST as the *interval time of week (ITOW)* and *time of interval (TOI)* to increase the bit length of GPS time. For LNAV, GPST rolls back to 0 every 1,024 GPS weeks (19.6 years) while CNAV/CNAV-2 GPST rolls back to 0 every 8192 weeks (157.0 years).

### GPS Satellite Positions from Ephemeris

The most crucial factor in GNSS pseudorange multilateration is determining the position of the transmitting artificial satellites. Thus, this subsection will cover an algorithm for converting the **GPS ephemeris**, i.e. a quantitative description of the positions, and possibly velocity, of satellites over time, to SV ECEF positions based on their elliptical orbits



where the labeled GPS ephemeris parameters are the six **Keplerian elements** of the orbit, i.e.

- the **eccentricity**,  $0 < e < 1$ , which is a measure of the elongation of the elliptical orbit compared to a circular orbit ( $e = 0$ ),
- the **semimajor axis**,  $a$ , which is the distance between the perigee and apogee, i.e. the closest and further points from the focus corresponding to the Earth's center of gravity,
- the **inclination** angle,  $i$ , which is the angle between the equatorial and orbital planes at the ascending node where the space vehicle moves from beneath the equatorial plane to above,
- the **longitude of ascending node**,  $\Omega$ , which is the angle relative to the prime meridian of Earth where the SV intersects the equatorial plane,
- the **argument of perigee**,  $\omega$ , which is the angle in the orbital plane from the ascending node to perigee,
- the **true anomaly at epoch**,  $v_0$

To compute an SV's ECEF position using the Keplerian elements, one must derive the true anomaly,  $v$ , from the eccentric anomaly,  $E$ , and the current mean anomaly,  $M$ . To do so, one must use  $M$  to solve for the eccentric anomaly  $E$  using **Kepler's equation**

$$M = E - e \sin E \quad (18.30)$$

which has no analytical solution and must be solved numerically. Next, one can compute  $v$  using

$$v = \text{atan2} \left( \frac{\sqrt{1 - e^2} \sin E}{1 - e \cos E}, \frac{\cos E - e}{1 - e \cos E} \right) \quad (18.31)$$

Then, one can convert the orbital angles to the position along the elliptical orbit and, finally, compute the ECEF position by a frame transformation using  $(i, \Omega, \omega)$  as Euler angles for elliptical orbit.

From the broadcasted navigation message, one can parse out the following information:

- $t_0$ : Reference time of ephemeris
- Keplerian Elements
  - $\sqrt{a}$ : square root of semimajor axis
  - $e$ : eccentricity
  - $i_0$ : inclination angle (at time  $t_0$ )
  - $\Omega_0$ : longitude of ascending node (at weekly epoch)
  - $\omega$ : argument of perigee (at time  $t_0$ )
  - $M_0$ : mean anomaly (at time  $t_0$ )
- Rates of change of select Keplerian elements
  - IDOT:  $di/dt$
  - OMEGADOT:  $\dot{\Omega}$

- $\Delta n$ : mean motion correction
- Amplitude of sinusoidal corrections
  - $C_{uc}/C_{us}$ : cosine/sine correction to argument of latitude
  - $C_{rc}/C_{rs}$ : cosine/sine correction to orbital radius
  - $C_{ic}/C_{is}$ : cosine/sine correction to inclination angle

As part of the algorithm, one must also use the following GPS-defined constants

- The WGS84 value of Earth's gravitational parameter

$$\mu = 3.986005 \times 10^{14} \text{ m}^3/\text{s}^2 \quad (18.32)$$

- GPS value for speed of light

$$c = 2.99792458 \times 10^8 \text{ m/s} \quad (18.33)$$

- WGS84 value of Earth's rotation rate

$$\dot{\Omega}_e = 7.2921151467 \times 10^{-5} \text{ rad/s} \quad (18.34)$$

- GPS value for  $\pi$

$$\pi = 3.1415926535898 \quad (18.35)$$

Alternatively, one can also compute the precise satellite positions from the orbital products of the IGS which use the IGS08 reference frame. This frame was first defined on January 1, 2005 and, similar to the WGS84, the IGS08 uses the Geodetic Reference System (GRS) 1980 ellipsoid with the Greenwich prime meridian and an origin at geocenter. It is derived from a subset of the 232 stable IGS station coordinates at epoch 2005.0 and is practically coincident with ITRF2008 and WGS84, all of which agree at the centimeter level which is also the uncertainty magnitude for each frame solution.

## 18.4 Air Data Systems

**Air data systems (ADS)** are multi-sensor perception systems that provide measurements of a flight vehicle's surrounding air mass, collectively known as the **air data**, typically quantified as the airspeed, altitude, rate-of-climb, and angles of attack and sideslip where the subset of airspeed, angle of attack, and angle of sideslip is known as the **air data triplet**. ADS are safety-critical for the majority of airplane operations as air data measurements are often used in the gain scheduling of the guidance and control laws. **Synthetic air data system (SADS)** provide only air data triplet, typically using the wind triangle, and can be model-based or model-free. SADS are primarily used for risk monitoring of a real air data system.

For the altitude and airspeed, most ADS use the principle of an air pressure gradient, i.e., pressure differences, as measured from a pitot tube and a static port/source, separately, or together as a **pitot-static system**.



The pitot tube measures the **pitot pressure**,  $P_p$ , also known as the **ram pressure**, defined as the air pressure created by the aircraft motion or the air “ramming” into the tube. Under ideal conditions, the pitot pressure is equal to the **total pressure**,  $P_t$ , also known as the **stagnation pressure**. The pitot tube is most often mounted on front of the wing or the nose, facing forward, where its opening is exposed to the relative wind and is less impacted by the aircraft’s structure. The static port measures the **static pressure**,  $P_s$ . Under ideal conditions, the static pressure is equal to the **atmospheric pressure**,  $P_a$ , which defines the nominal pressure of the surrounding air mass with no objects moving inside. The static port is often a flush-mounted hole on the fuselage of an aircraft where it can access the air flow in a relatively undisturbed area. This static port can also be used directly in a barometric altimeter using an ISA pressure model as function of altitude. The difference between the total and atmospheric pressures is called the **dynamic pressure**,  $Q$ , i.e.

$$Q = P_t - P_a = \frac{1}{2} \rho v_\infty^2 \approx P_p - P_s \quad (18.36)$$

which allows a direct computation of the airspeed *if* the air density is known. Typically, the air density is estimated with an ISA model with temperature, altitude, and static pressure.

Some pitot-static systems may also directly measure the rate-of-climb by intentionally releasing air from a static source port to a trailing air port. By measuring the rate of discharge through this leak relative to a calibrated nominal value, the rate-of-climb can be inferred. To measure the angles of attack and sideslip, an ADS use differential pressure ports, wind vanes, and/or null-seeking pressure tubes. In many ADS the differential pressure ports are typically placed on a single pitot tube and commonly use a five-hole or seven-hole geometry. With this tube, one simple method about straight-and-level flight is to use the center port for the airspeed, top-bottom port difference for angle of attack, and the left-right port difference for the angle of sideslip. However, more accurate ADS use calibration tables or a polynomial fit to calibration tests using all ports together to obtain the air data triplet. It should be noted that often the **flank angle**,  $\beta_f$ , is measured by the ADS instead of the sideslip angle, i.e.

$$\beta_f = \tan^{-1} \frac{v}{u} \quad (18.37)$$

which is approximately equal to the linearized sideslip angle, i.e.  $\Delta\beta_f = \Delta\beta$ .

**System faults** are failures of the ADS itself. Pitot-static ADS faults can occur from a blockage of the air passage due to foreign objects. A **blocked pitot tube** is caused by objects in the atmosphere entering and obstructing the tube, e.g. ice, water, or insects. Thus, the FAA recommends pitot tubes be checked for obstructions prior to any flight and to prevent icing, all pitot tubes for instrument-certified flight are equipped with a heating element. This blockage will only affect the ADS airspeed measurement as a blocked pitot tube will measure an increase or decrease in airspeed when the aircraft climbs or descends, even

though actual airspeed is constant due to the pitot pressure remaining constant while the static pressure is decreasing/increasing. However, a **blocked static port** typically only occurs due to icing. A blocked static port will cause the ADS to measure a constant altitude at which the static source became blocked, the measured rate-of-climb will be a constant value of zero, and the airspeed be measured as less or more than its true value as aircraft climbs or descends. This fault is much more dangerous than a blocked pitot tube because it affects *all* pitot-static differential pressures.

**Inherent errors** are the result of environmental factors and fall into several categories, each affecting different instruments. A **density error** is caused by variations of pressure and temperature in the atmosphere and affects the airspeed and altitude measurements. A **compressibility error** can arise because the ram air pressure will cause the air to compress inside the pitot tube. At standard sea level pressure altitude, this can be calibrated, but at higher altitudes, the compression may cause the sensor to measure greater than equivalent airspeed if a correction is not tabulated. These compressibility errors become significant at altitudes above 3,000 m and at airspeeds greater than 100 m/s. A **reversal error** is caused by a false static pressure measurement. This false measurement is typically caused by abnormally large changes in an aircraft's pitch. A large change in pitch will cause a momentary sensing of movement in the opposite direction. Reversal errors primarily affect the altitude and rate-of-climb measurements.

A **position error** is caused by the aircraft's static pressure being different from the air pressure of the ambient air mass far away from the aircraft due to the air flowing past the static source at a speed different from the aircraft's true airspeed. Position errors depend on several factors including airspeed, angle of attack, aircraft weight, acceleration, aircraft configuration, and in the case of rotorcraft, downwash. There are two categories of position errors: **fixed errors** and **variable errors**. Fixed errors are defined as specific to a particular model of aircraft, while variable errors are caused by external factors such as deformed panels obstructing the flow of air, or particular situations which may overstress the aircraft. A **lag error** is caused by the fact that any changes in the static or dynamic pressure outside the aircraft require a finite amount of time to affect the air in the tubing and the associated pressure gauges. Naturally, this type of error depends on the length and diameter of the tubing as well as the volume inside the gauges. Lag error is only significant around the time when the airspeed or altitude are changing rapidly and is not a concern for straight and level flight.

## 18.5 Clocks and Timing Systems

Timing systems provide a time output using sensors known as **clocks**, also known as **timepieces** or **chronometers**, which are studied as part of **horology**. Throughout history, different devices have been used as clocks including sundials, hourglasses, and water clocks. However, modern clocks use harmonic oscillators, also known as a **resonator**, which is a physical object that vibrates or oscillates at a particular *known* frequency. These include pendulums, tuning forks, quartz crystals, i.e. quartz clocks, or atomic vibrations from electrons emitting microwaves, i.e., **atomic clocks**. In most aerospace vehicle control and perception systems, either quartz clocks or atomic clocks are used.

The piezoelectric properties of crystalline quartz allows crystal oscillators to be used in **quartz clocks**. The National Bureau of Standards based the time standard of the United States on quartz clocks from late 1929 until the 1960s when it changed to atomic clocks. However, their inherent accuracy and low cost of production resulted in the subsequent proliferation of quartz clocks. Currently, **atomic clocks** are the most accurate clocks in existence. They are considerably more accurate than quartz clocks as they can be accurate

to within a few seconds over trillions of years. In the 1930s the development of magnetic resonance created practical methods for developing a prototype ammonia maser device in 1949 which was initially less accurate than existing quartz clocks. The first accurate atomic clock based on a certain transition of the caesium-133 atom was built in 1955. As of 2013, the most accurate atomic clocks are ytterbium clocks.

The development of atomic clocks has led to many scientific and technological advances, e.g., navigation satellite systems, long-baseline interferometry in radioastronomy, and the Internet, all of which critically depend on frequency and time standards. In these applications, atomic clocks are installed at broadcasting stations with radio transmitters which transmit a very precise carrier frequency used as a time signal. Many governments operate these transmitters for time-keeping purposes. A **radio clock** is a clock that synchronizes itself by means of these radio time signals received by a radio receiver. Consumer-grade receivers solely rely on the amplitude-modulated time signals and use narrow-band receivers, 10 Hz bandwidth, with small ferrite loopstick antennas and circuits with digital signal processing delays. Consumer-grade radio clocks are expected to determine the beginning of a second with an uncertainty of  $\pm 0.1$  second. Instrument-grade time receivers provide higher accuracy and incur a transit delay of approximately 1 ms for every 300 km of distance from the radio transmitter.

## Time Standards

A **time standard** is a specification for measuring time, either the rate at which time passes or specific points in time or both. Historically, time standards were often based on the Earth's rotational period as it was assumed that the Earth's daily rotational rate was constant which is false. These were replaced for astronomical use from 1952 onwards by an ephemeris time standard based on the Earth's orbital period and on the motion of the Moon. However, the invention of atomic clocks has led to the replacement of astronomical time standards by newer time standards based on atomic time.

**Ephemeris time (ET)** is a former standard astronomical time scale adopted in 1952 and superseded during the 1970s. This time scale was proposed to overcome the disadvantages of irregularly fluctuating mean solar time and to define a uniform time based on Newtonian theory. ET was a first application of the concept of a dynamical time scale, in which the time and time scale are defined implicitly, inferred from the observed position of an astronomical object via the dynamical theory of its motion.

**Terrestrial time (TT)** is a modern astronomical time standard defined by the International Astronomical Union and defines a coordinate time scale at the Earth's surface primarily for time-measurements of astronomical observations made from the Earth's surface. As TT shares the original purpose for which ET was designed, to be free of the irregularities in the rotation of Earth, it has succeeded the use of ET. It is a theoretical ideal, and clocks can only approximate it. The unit of TT is the **SI second**, the definition of which is based currently on the caesium atomic clock, though TT is not itself defined by atomic clocks.

**International Atomic Time (TAI)** is the primary international time standard from which other time standards are calculated. TAI is kept by the International Bureau of Weights and Measures (BIPM) and is based on the combined input of many atomic clocks around the world, each corrected for environmental and relativistic effects, both gravitational and speed. Of note, because of the historical difference between TAI and ET when TT was introduced, TT is approximately 32.184 seconds ahead of TAI.

**Universal Time (UT)** is the mean solar time at  $0^\circ$  longitude. However, as precise measurements of the sun are difficult, UT is computed from determining the positions of distant quasars using very long baseline interferometry (VLBI), satellite laser ranging (SLR), and global navigation satellite systems (GNSS). These

observations allow the determination of a measure of the **Earth rotation angle (ERA)**, i.e. Earth's angle with respect to the ICRF, neglecting some small adjustments. The principal form of UT, UT1 is required to follow the following relationship.

$$ERA = 2\pi(0.7790572732640 + 1.00273781191135448(JD - 2451545.0)) \text{ radians} \quad (18.38)$$

where *JD* is the **Julian date (JD)**, i.e. the Julian day number plus the fraction of a day since the preceding noon in UT1 where the **Julian day number (JDN)** is the continuous count of days since the beginning of the Julian period, i.e. noon Universal Time on January 1, 4713 BC, by the Julian calendar and November 24, 4714 BC, by the Gregorian calendar.

**Coordinated Universal Time (UTC)** is an atomic time scale designed to approximate UT. UTC differs from TAI by an integral number of seconds and is kept within 0.9 second of UT1 by the introduction of one-second steps to UTC, i.e. the “**leap second**.” To date these leap-seconds and the difference, TAI-UTC, have always been positive.

**Standard time**, also known as **civil time** in a region deviates a fixed, round amount, usually a whole number of hours, from UTC. The offset is chosen such that a new day starts approximately while the sun is crossing the nadir meridian. It may also change twice a year by a round amount, usually one hour, e.g. **daylight saving time (DST)**.

Timing systems within the context of information systems often use their own time standard which are typically related in some defined way to UTC. One common standard widely used in operating systems and file formats is **posix time**, also known as **unix time**, defined as the number of seconds that have elapsed since 00:00:00 UTC on January 1, 1970, excluding leap seconds. Posix time is nonlinear as a UTC leap second has the same Unix time as the second before it or after it, so that every day is treated as if it contains exactly 86400 seconds with no seconds added to or subtracted from the day as a result of positive or negative leap seconds.

### Clock Stability and Allan Variance

Quartz and atomic oscillators of clocks have a phase noise consisting of white noise and **flicker frequency noise** which is a divergent noise. Thus, all clocks exhibit **clock drift**. Furthermore, as this noise is divergent, one cannot compute a standard deviation. Thus, early efforts in analyzing the stability included both theoretical analysis and practical measurements. However, as various methods of measurements did not agree with each other, the key aspect of repeatability of a measurement could not be achieved, limiting the possibility to compare clock sources and make meaningful specifications to require from suppliers. Essentially all forms of scientific and commercial uses were then limited to dedicated measurements, which hopefully would capture the need for that application.

To address these problems, David Allan introduced the ***M*-sample variance** and, indirectly, the **2-sample variance**, also known as the **Allan variance**. The *M*-sample variance is defined as

$$\sigma_y^2(M, T, \tau) = \frac{1}{M-1} \left( \sum_{k=0}^{M-1} \bar{y}_k^2 - \frac{1}{M} \left( \sum_{k=0}^{M-1} \bar{y}_k \right)^2 \right) \quad (18.39)$$

where *M* is the number of frequency samples used in the variance, *T* is the time between adjacent start events,  $\tau$  is the observation period from start to stop events, and  $\bar{y}_k$  is the *k*<sup>th</sup> **average fractional frequency**

defined

$$\bar{y}_k = \frac{x(kT + \tau) - x(kT)}{\tau} \quad (18.40)$$

where  $x(t)$  is the clock reading (in seconds) measured at  $t$ . It should be noted that the  $M$ -sample allows the use of **dead-time** by letting  $T$  be different from  $\tau$ . Allan proved all  $M$ -sample variances were comparable and did not converge for large  $M$ , thus the Allan variance has become the preferred variance. The Allan variance is defined as

$$\sigma_y^2(\tau) = \mathbb{E} [\sigma_y^2(2, \tau, \tau)] \quad (18.41)$$

$$\sigma_y^2(\tau) = \frac{1}{2} \mathbb{E} [(\bar{y}_{k+1} - \bar{y}_k)^2] \quad (18.42)$$

or

$$\sigma_y^2(\tau) = \frac{1}{2\tau^2} \mathbb{E} [(x(k\tau + 2\tau) - 2x(k\tau + \tau) + x(k\tau))^2] \quad (18.43)$$

Lastly, the **Allan deviation** is given by

$$\sigma_y(\tau) = \sqrt{\sigma_y^2(\tau)} \quad (18.44)$$

The Allan variance as one half of the time average of the squares of the differences between successive samples of the frequency deviation over the sampling period which is a free parameter. Therefore, it is a function of the distribution being measured and is displayed as a graph rather than a single number. While the Allan variance does not completely allow all types of noise to be distinguished, it provides a means to meaningfully separate many noise-forms for time-series of phase or frequency measurements between two or more oscillators. However, the Allan deviation is the preferred graphical measure as it gives the relative amplitude stability. A low Allan deviation indicates a clock has good stability over the measured period. An Allan deviation of  $1.3 \times 10^{-9}$  at an observation period of 1 s, i.e.  $\tau = 1$  s, should be interpreted as there being an instability in frequency between two observations 1 second apart with a relative **root mean square (RMS)** value of  $1.3 \times 10^{-9}$ . For a 10 MHz clock, this would be equivalent to 13 mHz RMS movement. Ytterbium clocks are stable to within  $2 \times 10^{-18}$  in terms of Allan deviation. Though these definitions are given in terms of expectation, in practice, one must use estimate the Allan variance for which various statistical estimators have been proposed.

## Clock Networks and Synchronization

A **clock network** is a set of synchronized clocks designed to always show exactly the same time by communicating with each other, e.g. **wireless sensor networks (WSN)** in multi-sensor data fusion. Each clock,  $C_i(t)$ , should be configured such that

$$C_i(t) = t \quad (18.45)$$

where  $t$  is the reference time. However, even when initially set accurately, real clocks will differ after some amount of time due to imperfections in the clock oscillator, i.e. clock drift. This clock drift can be modeled as

$$C_i(t) = \omega t + \theta \quad (18.46)$$

where  $\omega$  is the **clock skew**, i.e. frequency difference, and  $\theta$  is the **clock offset**, i.e. phase difference. Over time, these clock parameters vary due to external effects, e.g. temperature, atmospheric pressure, voltage changes, and hardware aging. Thus, the network has to perform periodic clock synchronization routines

to adjust the estimated clock skew and offset and typically uses one of three schemes to perform clock synchronization: sender-receiver, receiver-receiver, or one-way sender.

A one-way sender strategy uses a master clock to broadcast its timing information, e.g. the **time of transmission (TOT)**, to slave clocks which also record the arrival times of the broadcast message, i.e., the **time of arrival (TOA)**. Traditionally, one models the master clock as the reference time, i.e.

$$C_M(t) = t \quad (18.47)$$

and a slave clock as having a relative clock skew and offset error terms

$$C_S(t) = \delta\omega t + \delta\theta \quad (18.48)$$

where the primary problem for each slave is to track the reference time  $t$  from the master clock given the TOT packets from the master. In modern applications, one typically uses digital clocks which use discrete clock ticks, i.e. an integer count of the periodic clock signal with some known period or **timebase**.

In this case, assume the master clock transmits a TOT packet with timebase  $T_M$ . This is received by the slave clock at time step  $k$  as  $T_{S,k}$ . With this model one can define the measurement model as

$$T_{S,k} \approx \delta\omega T_M + \delta\theta + \epsilon_k \quad (18.49)$$

where  $\epsilon_k$  is the random variable delay in the transmission. Thus, clock tracking by each slave is implemented with the state dynamics taking three inputs, the slave clock TOA and the master clock's TOT with its timebase. **Digital phase-locked loops (DPLL)** and Kalman filters have been widely used for clock synchronization.

Consider the following estimation of a phase and frequency offset for the master clock signal, i.e.

$$s_M(t) = A \cos \phi_M = A \cos(\omega_M t + \theta_M) \quad (18.50)$$

which is assumed to have some known values for  $\omega_M$  and  $\theta_M$ . A phase detector replicates this model and compares this to the slave clock signal, i.e.

$$s_S(t) = A \cos \phi_S = A \cos(\omega_S t + \theta_S) \quad (18.51)$$

which can be used to form the **loop phase error**,  $\delta\phi$  at time step  $k$  as

$$\delta\phi_k = \phi_M - \phi_S = \delta\omega t + \delta\theta \quad (18.52)$$

In tracking mode, a DPLL  $D(z)$  uses this to output a corrected sequence  $y_k$

$$y_k = D(z)\delta\phi_k \quad (18.53)$$

where a proportional-integral DPLL

$$D(z) = K_p + K_i(1 - z^{-1})^{-1} \quad (18.54)$$

results in

$$y_k = K_p \delta\phi_k + K_i \sum_{j=0}^k \delta\phi_j \quad (18.55)$$

Then, the corrected sequence,  $y_k$ , is used to control the next period of the sample,  $T_{k+1}$ , through the adjustment

$$T_{k+1} = T_M - y_k \quad (18.56)$$

where one should adjust  $T_k$  until the loop is in the locked state with clock period  $T_k$  that exactly matches  $t_{k+1} - t_k$ . Furthermore, the reference sampling time for each packet can be obtained through

$$\hat{t}_{k+1} = \hat{t}_k + T_M - K_p \delta\phi_k - K_i \sum_{j=0}^k \delta\phi_j \quad (18.57)$$

---

# Positioning Systems

## 19.1 Introduction to Positioning Systems

Vehicle perception systems that consider the problem of estimating the position of a vehicle relative to a predefined reference frame are said to be performing **positioning**. One potential reference frame could simply be chosen as the origin of some movement and the vehicle's position is then computed relative to this beginning by estimating the change in movement. Positioning using only the integration of the vehicle's movement or velocity is known as **dead reckoning (DR)** which is fully presented in the following chapter on navigation. **Navigation** is the estimation of a vehicle's position, velocity, *and* attitude, which together are also known as the vehicle's **navigation state**. It should be noted that the primary drawback of DR results from the compounding of velocity estimation errors through integration. Thus, the positioning error in DR increases unboundedly over time. Lastly, it should be noted that the vehicle's position *and* attitude are also known as the vehicle's **pose** whose combined estimation may also be considered positioning.

Importantly, positioning and navigation typically require one to be able to also sense and estimate any potential changes in the sources of measurements as well as any measurement signal processing computations. Thus, any positioning or navigation system will also require accurate time measurements provided by a clock as part of a timing system. Thus, due to their interrelationship, these three quantities are often considered as the single engineering discipline of **positioning, navigation, and timing (PNT)** in vehicle information systems.

Related to reference frames in positioning is the potential use of a **map** which defines the positions of collective features within a reference frame. In this case, one is said to be performing **localization** if one describes the **location** of vehicle relative to these collective objects. This type of localization may be qualitative or quantitative, but a position is always defined as the *coordinates* within a defined reference frame.

## Positioning Measurements and Sensors

The four fundamental positioning measurement types are **bearing**, i.e. the direction from the measurement source to the vehicle, **range**, i.e., the distance from the measurement source to the vehicle, **differential range**, i.e., the relative distance from two measurement sources to the vehicle, and **range rate**, i.e., the instantaneous change in the distance from the measurement source to the vehicle. These measurement types allow one to form corresponding **lines-of-position (LOP)** relating the measurement source(s) and the vehicle. When using only bearing LOPs, one can perform positioning using **multiangulation** techniques and when using only range or differential range LOPs, one can perform positioning using **multilateration** techniques. Using bearings and ranges LOPs, one can perform positioning using **multiangulateration** techniques.

Measurement sources can generally be considered as either intentionally-designed **beacons** or identifiable environmental **features**. In beacon-based positioning, a vehicle uses sensors to detect, identify, and process the *known* signals of the beacons which have known characteristics convenient for positioning algorithms. Common beacons include radio beacons and visible light beacons, e.g. lighthouses. In feature-based positioning, a vehicle uses sensors to detect, identify, and process the signals from environmental features in order to perform positioning within one's environment. Types of environmental features that are commonly used for positioning are topographical structures, magnetic fields, celestial bodies, atmospheric and ocean pressure, and exploitable radio signals also known as **signals of opportunity (SOP)**. Pre-modern positioning primarily used known topographical structures using eyesight, the Earth's magnetic field using compasses, and the observable movement of celestial bodies using eyesight.

Of note, until the advent of electronic inertial sensors, which can be used to perform accurate dead reckoning, navigators used the geographic position of celestial bodies to perform open ocean positioning. The **geographic position (GP)** of a celestial body is the point on the Earth's surface from which the body is currently directly overhead. However the computation of the GP requires celestial trigonometry and is also known as **sight reduction**. Thus by careful observation and calculation of the movement of the celestial body, celestial maps were created and used to be able to determine a celestial body's current GP. However, sufficient determination for reliable navigation was only made possible by the invention of accurate clocks with sufficient stability during long sea voyages. Furthermore as most celestial body GP maps were tabulated to GMT, it was standard to set one's clock to **Greenwich Mean Time (GMT)** where the prime meridian, i.e. 0° longitude, passes through Greenwich, England. In modern times, GMT has become the time standard UTC. Using this GP information, the **sextant** allowed one to measure the angle between the horizon in the direction of the GP and its corresponding celestial object in the sky. This angle can be shown to be directly proportional to the range of the navigator to the GP and this angle-to-distance conversion forms the origin of the distance unit of the nautical mile as equal to one minute of latitude.

With modern electronic sensors, positioning primarily uses energy wave signals, e.g. radio, light, acoustic. One uses **antennas** to sense radio signals, **photosensors**, also known as **photodetectors**, to sense ultraviolet (UV), visible, or near infrared (NIR) light signals, and **sonar** to sense transmitted and received acoustic signals. These sensors convert the sensed wave signal into an electrical signal for use in a signal processing algorithm before use in a positioning algorithm. This signal processing can use several characteristics of these received signals to derive an LOP. One may use the observed **angle of arrival (AOA)** of the received signal to derive a bearing measurement. One may also use the observed **time of flight (TOF)** of a transmitted *and* received signal to derive a range measurement, e.g. **radar**.

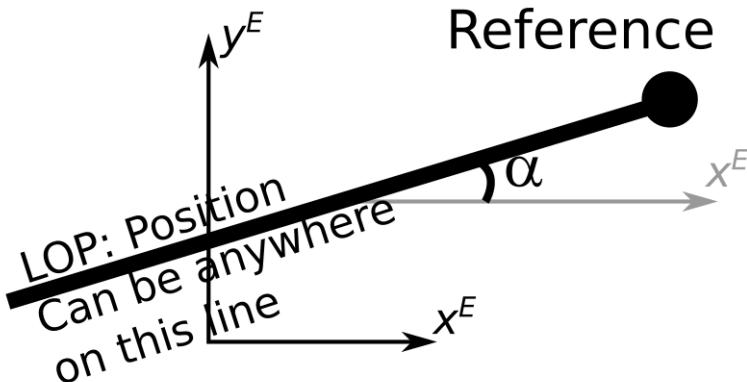
However, if one is simply receiving a *known* univariate signal, one may use the observed **time of arrival (TOA)** of the received signal to derive a **pseudorange** measurement, i.e. "range" based on the timing of

transmitted signals introduces additional timing errors in the measurement that must be estimated. One may also use the observed **time difference of arrival (TDOA)** of two received wave signals to derive a **differential pseudorange**. Another alternative is to use the observed **phase difference** for additional timing information for ranging. For range rate measurements using *known* univariate signals, one typically use Doppler effect of energy waves. The **Doppler effect** describes the increase/decrease in the frequency of an energy wave as the source and observer move toward/away from each other.

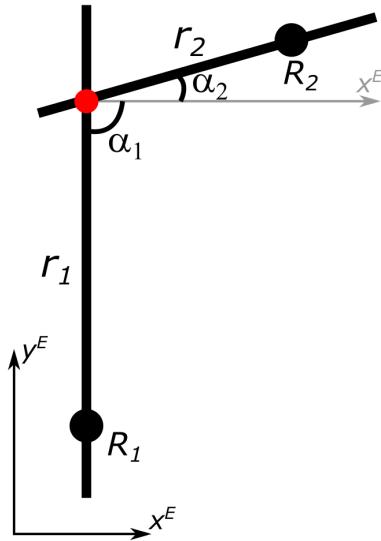
For the signal processing of two- or three-dimensional signals, one can produce **images** or **point clouds**, when often uses additional signal processing algorithms known as **image processing** or **computer vision** algorithms to derive bearing, range, and/or range rate measurements. In addition to these energy wave signals, one may also incorporate other sensors to aid in the positioning. One can use pressure gauges to sense air or water pressure, magnetometers to sense the Earth's magnetic field, and/or radiometers, e.g. WiFi signals, which allow one to localize one's position based on maps of the pressure, magnetic field and/or passive radio signals. One may also use inertial sensors and/or mechanical linkages which allow one to form proprioceptive sensors for estimating changes in position motion, i.e. movement.

### Bearing Triangulation

A bearing LOP from a measurement source in two-dimensions,  $\alpha$ , can be visualized as a straight line as shown in the following figure.



These LOPs provide an estimate of position using the intersection of straight lines shown here for two sources as

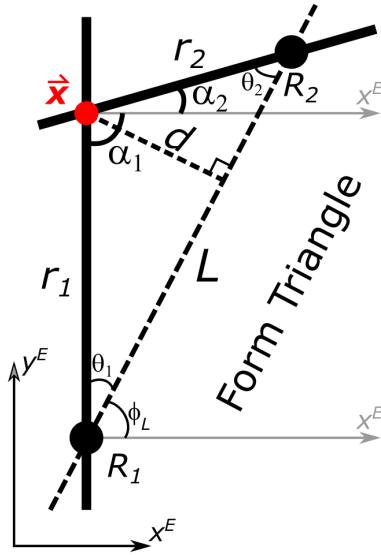


With two bearing measurements and a known **baseline** between the reference sources  $R_1$  and  $R_2$ , one can determine the position from the resulting triangle. Thus, this positioning is known as **triangulation**. For more than two sources of bearing measurements, one can perform positioning using **multiangulation** which will increase the accuracy of the position estimate due to any uncertainty in the measurements, i.e., noise. Note that for long distances on Earth, multiangulation may require spherical trigonometry for calculating the LOPs.

To solve the two-dimensional triangulation problem geometrically, assume that one has a two-dimensional unknown position represented as a vector

$$\vec{x} = [x \ y]^T \quad (19.1)$$

and that one knows the reference positions of the source of the bearing angles,  $\vec{x}_{R1}$  and  $\vec{x}_{R2}$ . Furthermore, assume that one has knowledge of the absolute reference frame, the direction of  $x^E$  from which one has measured  $\alpha_1$  and  $\alpha_2$ . Then, one can form the following triangle.



Defining, the relative angle of baseline to  $x^E$  as

$$\phi_L = \tan^{-1} \frac{y_{R2} - y_{R1}}{x_{R2} - x_{R1}} \quad (19.2)$$

by trigonometry, one has as base angles of the triangle

$$\theta_1 = 180^\circ + \alpha_1 - \phi_L \quad (19.3)$$

as  $\alpha_1 < 0^\circ$

$$\theta_2 = \phi_L - \alpha_2 \quad (19.4)$$

and a baseline length as

$$L = \|\vec{x}_{R1} - \vec{x}_{R2}\|_2 \quad (19.5)$$

one can write that

$$L = \frac{d}{\tan \theta_1} + \frac{d}{\tan \theta_2} \quad (19.6)$$

where  $d$  is the only unknown. This equation can be rearranged as

$$L = d \frac{\cos \theta_1}{\sin \theta_1} + d \frac{\cos \theta_2}{\sin \theta_2} \quad (19.7)$$

$$L = d \frac{\sin \theta_1 \cos \theta_2 + \sin \theta_2 \cos \theta_1}{\sin \theta_1 \sin \theta_2} \quad (19.8)$$

$$L = d \frac{\sin(\theta_1 + \theta_2)}{\sin \theta_1 \sin \theta_2} \quad (19.9)$$

$$d = L \frac{\sin \theta_1 \sin \theta_2}{\sin(\theta_1 + \theta_2)} \quad (19.10)$$

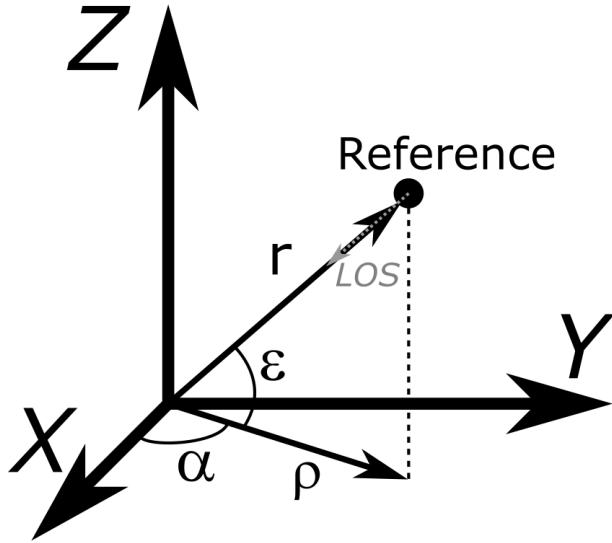
Now that one has  $d$ , one can find the position coordinates as

$$\vec{x} = \vec{x}_{R_1} + \begin{bmatrix} r_1 \sin \alpha_1 \\ r_1 \cos \alpha_1 \end{bmatrix} \quad (19.11)$$

$$\vec{x} = \vec{x}_{R_1} + \begin{bmatrix} \frac{d \sin \alpha_1}{\sin \theta_1} \\ \frac{d \cos \alpha_1}{\sin \theta_1} \end{bmatrix} \quad (19.12)$$

though one could also reference from  $\vec{x}_{R_2}$ .

For three-dimensional positioning, the bearing LOPs can be completely defined by two angles known as **azimuth**,  $\alpha \in [0, 2\pi]$ , and **elevation**,  $\epsilon \in [0, \pi]$ , as shown in the following figure for one measurement source,



Here,  $\rho$  is called the **ground range** and  $r$  is called the **slant range**.

To solve the three-dimensional multiangulation problem for an unknown position  $\vec{x} = [x \ y \ z]^T$  and a reference located at  $\vec{x}_R = [x_R \ y_R \ z_R]^T$ , one has can form the **line-of-sight (LOS)** unit vector as

$$\frac{\vec{x} - \vec{x}_R}{\|\vec{x} - \vec{x}_R\|_2} = \begin{bmatrix} \cos \alpha \cos \epsilon \\ \sin \alpha \cos \epsilon \\ \sin \epsilon \end{bmatrix} \quad (19.13)$$

or using trigonometry, one has

$$\begin{bmatrix} \tan \alpha \\ \tan \epsilon \end{bmatrix} = \begin{bmatrix} \frac{x - x_R}{y - y_R} \\ \frac{y - y_R}{z - z_R} \end{bmatrix} \quad (19.14)$$

$$\begin{bmatrix} \alpha \\ \epsilon \end{bmatrix} = \begin{bmatrix} \tan^{-1} \frac{x - x_R}{y - y_R} \\ \tan^{-1} \frac{y - y_R}{z - z_R} \end{bmatrix} \quad (19.15)$$

For this problem there are three unknown coordinates and two measurements for each reference. Thus, for two references, one has an over-constrained problem. As a solution, one could ignore one of the four measurements and solve this geometrically similarly to the two-dimensional case. However, as one typically must consider measurement noise, one can use a nonlinear estimation technique, e.g., nonlinear least-squares, to perform the positioning with the model chosen with a parameter vector as

$$\vec{\beta} = \vec{x} \quad (19.16)$$

a measurement vector as

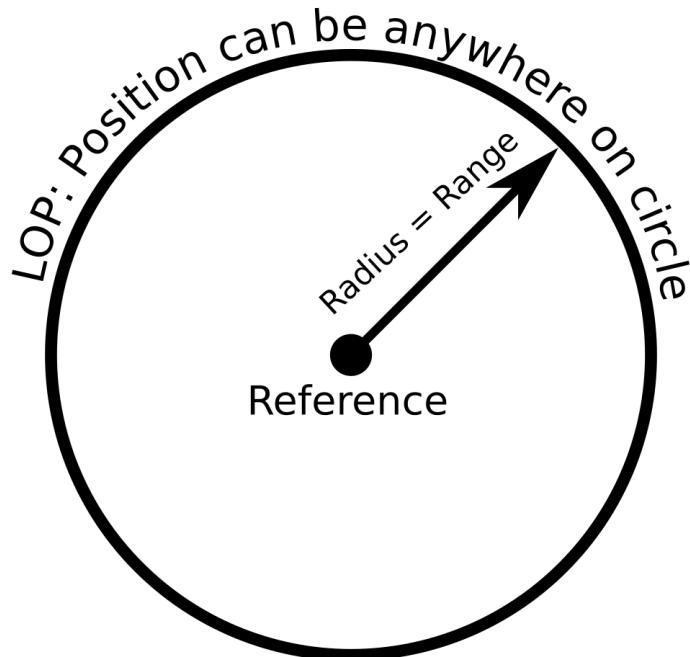
$$\vec{y} = \begin{bmatrix} \alpha_1 \\ \epsilon_1 \\ \alpha_2 \\ \epsilon_2 \end{bmatrix} \quad (19.17)$$

and a nonlinear measurement model as

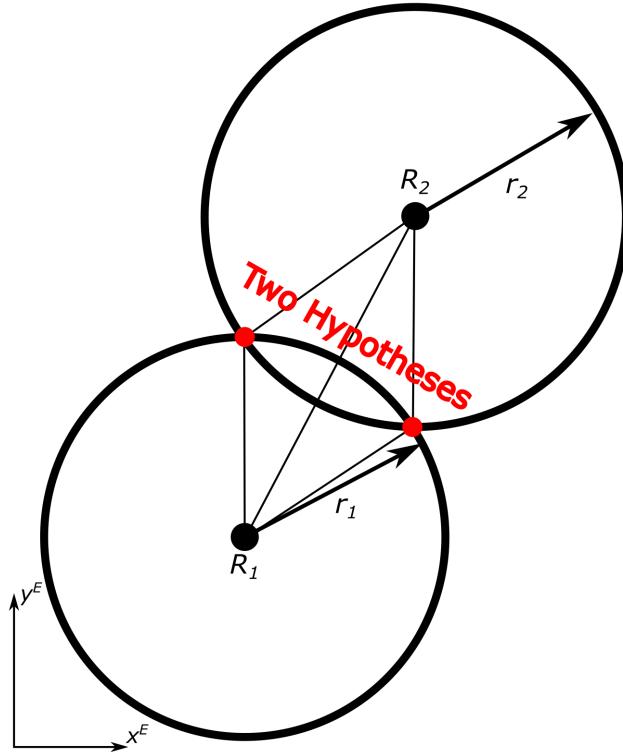
$$\vec{f}([\vec{x}_{R1} \vec{x}_{R2}]^T, \vec{x}) = \begin{bmatrix} \tan^{-1} \frac{x - x_{R1}}{y - y_{R1}} \\ \tan^{-1} \frac{x - x_{R1}}{\sqrt{(x - x_{R1})^2 + (y - y_{R1})^2}} \\ \tan^{-1} \frac{x - x_{R2}}{y - y_{R2}} \\ \tan^{-1} \frac{x - x_{R2}}{\sqrt{(x - x_{R2})^2 + (y - y_{R2})^2}} \end{bmatrix} \quad (19.18)$$

### Range Trilateration

A range LOP from a measurement source in two-dimensions,  $r$ , can be visualized as a circle as shown in the following figure

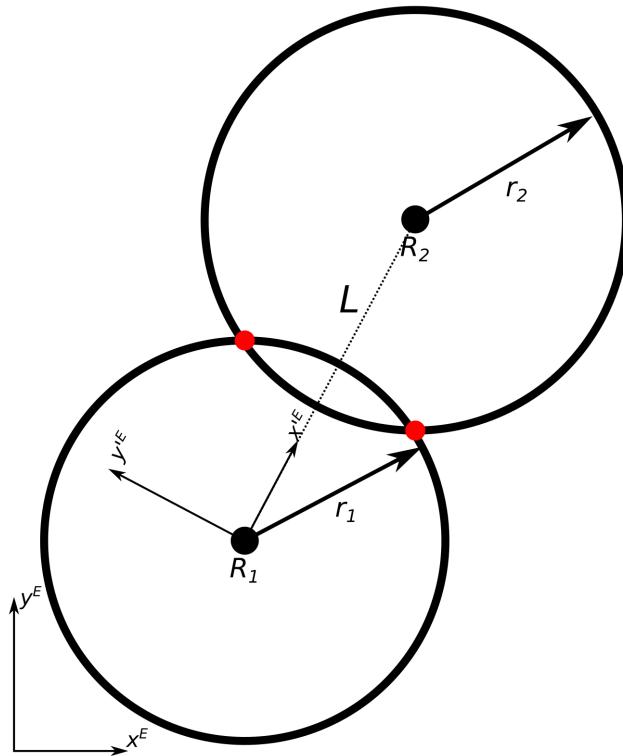


These LOPs provide an estimate of position using the intersection of circles shown here for two sources as



With an initial guess, one can select the more likely hypothesis of the two possibilities. For each hypothesis, one knows the lengths of the sides of the associated triangle. Thus, this positioning is known as **trilateration**.

To solve the two-dimensional multilateration problem geometrically, one can form a “new”  $x - y$  frame with the origin at  $R_1$  and the  $x'$ -axis pointing at  $R_2$ , i.e.



Next, by noting the baseline length is now

$$L = \|\vec{x}_{R1} - \vec{x}_{R2}\|_2 \quad (19.19)$$

one can write the range measurements in terms of the unknown coordinates as

$$r_1^2 = x'^2 + y'^2 \quad (19.20)$$

$$r_2^2 = (L - x')^2 + y'^2 \quad (19.21)$$

which can be solved simultaneously for the unknown position coordinates in the new frame as

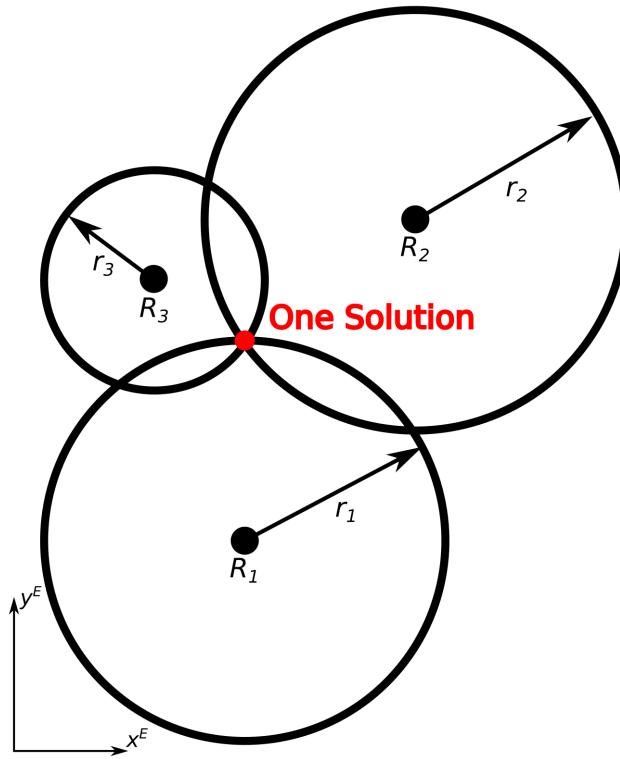
$$x' = \frac{r_1^2 - r_2^2 + L^2}{2L} \quad (19.22)$$

and

$$y' = \pm\sqrt{r_1^2 - x'^2} \quad (19.23)$$

and by transforming back to the original frame,  $x$  and  $y$  can be obtained.

With three or more references, one can obtain improved accuracy due to measurement noise which generalizes trilateration to multilateration positioning which is shown in the following figure



Lastly, it is important to note that in three-dimensions, range LOPs become spheres for which an intersection point requires at least three references for two position hypotheses from which one must infer which is correct.

A related measurement source is the range-rate,  $\dot{r}$ , which can be modeled as

$$\dot{r} = \frac{d}{dt} \|\vec{r}\| = \frac{\dot{\vec{r}}^T \vec{r}}{\|\vec{r}\|} \quad (19.24)$$

which can be visualized as the projection of the relative velocity onto the LOS vector. However, the inclusion of velocity requires the addition of source and/or vehicle movement which extends the positioning problem to a navigation problem.

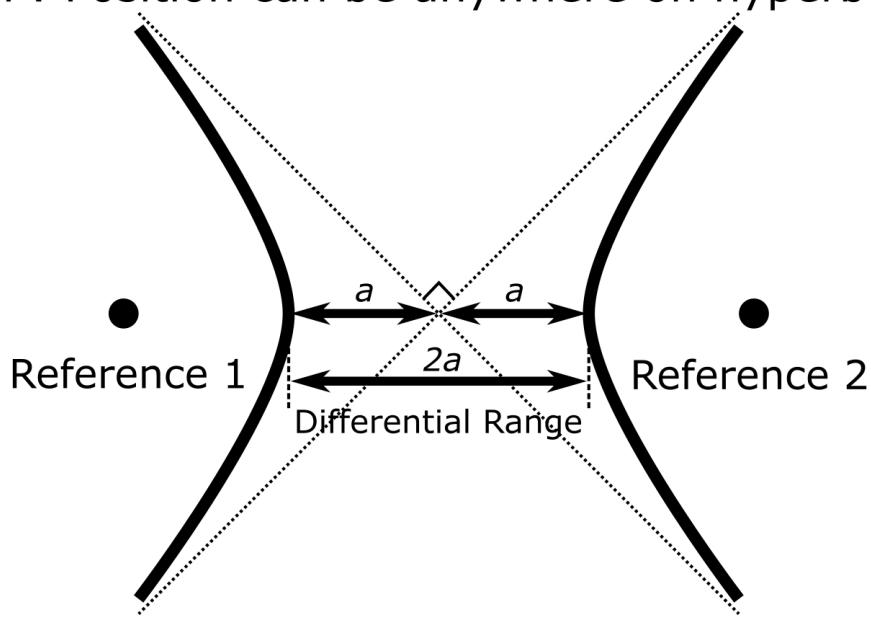
### Differential Range Multilateration

A differential range LOP from two measurement sources in two-dimensions can be modeled as

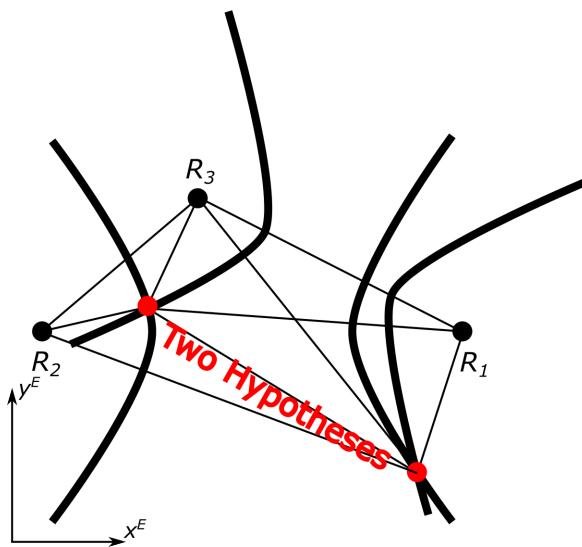
$$\|\vec{r}_1\|_2 - \|\vec{r}_2\|_2 = 2a \quad (19.25)$$

which can be visualized as a rectangular hyperbola as shown in the following figure

LOP: Position can be anywhere on hyperbola



These LOPs provide an estimate of position using the intersection of the rectangular hyperbolas shown here for three sources as



and with an initial guess, one can select the more likely hypothesis of the two possibilities. For each hypothesis, one knows the length of one side or the combined length of two sides of the two associated triangles in the diagram. Thus, this positioning method is also known as **multilateration**, albeit, the

trigonometric relationships are more difficult than true-range trilateration and are left for the reader.

### Range-Rate Positioning

The range-rate is related to the relative velocity of the vehicle,  $\dot{\vec{x}}_V$  and a measurement source,  $\dot{\vec{x}}_S$ , projected along its relative line-of-sight (LOS) vector, i.e.,

$$\frac{d}{dt} \|\vec{r}\| = \frac{d}{dt} \|\vec{x}_V - \vec{x}_S\|_2 = (\dot{\vec{x}}_V - \dot{\vec{x}}_S)^T \frac{(\vec{x}_V - \vec{x}_S)}{\|\vec{x}_V - \vec{x}_S\|_2} \quad (19.26)$$

where  $\vec{x}_V$  is the receiver position and  $\vec{x}_S$  is the source's position. Thus, to use range-rates for positioning, one must either know or estimate the velocity of the vehicle before positioning or estimate the velocity simultaneously with the position. The former requires two or more measurement sources in two-dimensions and three or more measurement sources in three-dimensions. The latter would require four or more measurement sources in two-dimensions and six or more measurement sources in three-dimensions. Notably, for known velocities, a difference in range-rate from two measurements source in two dimensions can be modeled as

$$\frac{d}{dt} \|\vec{r}_1\| - \frac{d}{dt} \|\vec{r}_2\| = 2a \quad (19.27)$$

which also corresponds to a rectangular hyperbola and with an initial guess, one can select the more likely hypothesis of the two possibilities.

### References

For more information, please refer to the following

- S. Gleason and D. Gebre-Egziabher, “GNSS Applications and Methods,” Artech House, 2012
- C. Jekeli, “Inertial Navigation Systems with Geodetic Applications,” de Gruyter, 2001

## 19.2 Pseudorange Positioning Systems

### Un-Differenced Pseudorange Multilateration

In true range multilateration, one directly measures the Euclidean distance, i.e. the vector  $L_2$ -norm, between the  $i^{\text{th}}$  transmitter and the signal receiver, i.e.

$$r_i = \|\vec{x} - \vec{x}_i\|_2 \quad (19.28)$$

which in three-dimensional space, i.e.  $\vec{x} = [x \ y \ z]^T$  and  $\vec{x}_i = [x_i \ y_i \ z_i]^T$ , is written by component as

$$r_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} \quad (19.29)$$

In pseudorange multilateration, one measures the change in time of transmission (TOT) of the  $i^{\text{th}}$  signal,  $TOT_i$ , to the time of arrival (TOA),  $TOA_i$ . Thus, by multiplying difference between TOA and TOT by the assumed constant signal speed,  $c$ , one can form the **pseudorange**,  $\rho_i$ , for the  $i^{\text{th}}$  transmitter as

$$\rho_i = c (TOA_i - TOT_i) \quad (19.30)$$

However, this rough approximation to the true range is corrupted by errors, both known and unknown. Thus, one typically relates the pseudorange to the true range by the **(un-differenced) pseudorange equation** defined as

$$\rho_i = r_i + c(\delta t - \delta t_i) + b_{p,i} + b_{c,i} + \epsilon \quad (19.31)$$

where  $\delta t$  accounts for any clock error in  $TOA_i$ ,  $\delta t_i$  accounts for any clock error in  $TOT_i$ ,  $b_{p,i}$  accounts for any distance error bias due to the assumed transmitter position,  $b_{c,i}$  accounts for any distance error bias due to the assumed signal speed. Lastly,  $\epsilon$  accounts for any additional unknown random errors, e.g. electrical noise. In pseudorange multilateration applications, one must typically refine the initially calculated  $\rho_i$  based on *transmitter-dependent* error models,  $\delta t_i$ ,  $b_{p,i}$ , and  $b_{c,i}$ . Furthermore,  $\delta t$  is a receiver clock error that changes with time and must be estimated along with the receiver position coordinates in pseudorange multilateration. Removing the modeled error biases, one has the **simplified pseudorange equation** defined as

$$\rho_i = r_i - c\delta t + \epsilon \quad (19.32)$$

Substituting for  $r_i$ , one has

$$\rho_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} - c\delta t + \epsilon \quad (19.33)$$

where  $x$ ,  $y$ ,  $z$ , and  $\delta t$  are the four unknown parameters. Thus, in pseudorange multilateration, one has for the parameter vector

$$\vec{\beta} = \begin{bmatrix} x \\ y \\ z \\ c\delta t \end{bmatrix} \quad (19.34)$$

which implies that pseudorange multilateration requires four or more transmitters to be able to obtain a position estimate as opposed to three or more for true range multilateration. The simplified pseudorange equation allows one to form a nonlinear observation equation with  $m$  transmitters as

$$\vec{y} = \begin{bmatrix} \rho_1 \\ \vdots \\ \rho_m \end{bmatrix} = \mathbf{f}(\vec{x}, \vec{\beta}) = \begin{bmatrix} \sqrt{(\hat{x} - x_1)^2 + (\hat{y} - y_1)^2 + (\hat{z} - z_1)^2} + c\hat{\delta}t \\ \vdots \\ \sqrt{(\hat{x} - x_m)^2 + (\hat{y} - y_m)^2 + (\hat{z} - z_m)^2} + c\hat{\delta}t \end{bmatrix} \quad (19.35)$$

To solve this problem using NLS with GNA, assume one has an initial estimate,  $\hat{\vec{\beta}}_0$ . Then, one can iteratively improve  $\hat{\vec{\beta}}$  by  $\vec{\Delta}_k$  for  $k = 0, 1, \dots$ , i.e.

$$\hat{\vec{\beta}}_{k+1} = \hat{\vec{\beta}}_k + \vec{\Delta}_k \quad (19.36)$$

where

$$\vec{\Delta}_k = (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T (\vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_k)) \quad (19.37)$$

where the  $i^{\text{th}}$  row of  $m \times 4 \mathbf{J}$  can be written as

$$\mathbf{J}_i = \begin{bmatrix} \frac{1}{2} [(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2]^{-1/2} 2(x - x_i) \\ \frac{1}{2} [(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2]^{-1/2} 2(y - y_i) \\ \frac{1}{2} [(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2]^{-1/2} 2(z - z_i) \\ 1 \end{bmatrix}_{\vec{x}=\hat{\vec{x}}_k}^T \quad (19.38)$$

or simplifying, one has

$$\begin{aligned}\mathbf{J}_i &= \left[ \frac{x-x_i}{r_i} \quad \frac{y-y_i}{r_i} \quad \frac{z-z_i}{r_i} \quad 1 \right]_{\vec{x}=\hat{\vec{x}}_k} \\ &= \left[ \frac{(\vec{x}-\vec{x}_i)^T}{\|\vec{x}-\vec{x}_i\|_2} \quad 1 \right]_{\vec{x}=\hat{\vec{x}}_k}\end{aligned}\quad (19.39)$$

where  $\frac{\vec{x}-\vec{x}_i}{\|\vec{x}-\vec{x}_i\|_2}$  is notably the **line-of-sight (LOS)** unit vector from the transmitter to the receiver.

Note that if one assumes  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ , where the covariance is approximately given by

$$\text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{J}^T \mathbf{J})^{-1} = \sigma^2 \begin{bmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} & c\sigma_{xt} \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yz} & c\sigma_{yt} \\ \sigma_{xz} & \sigma_{yz} & \sigma_z^2 & c\sigma_{zt} \\ c\sigma_{xt} & c\sigma_{yt} & c\sigma_{zt} & c^2\sigma_t^2 \end{bmatrix} \quad (19.40)$$

where  $\sigma^2$  would come from a statistical analysis of the pseudorange errors. In this case, as the matrix terms of  $(\mathbf{J}^T \mathbf{J})^{-1}$  term are completely dependent on the current LOS geometry of the transmitters to the receiver,  $\mathbf{J}$  is also known as the **geometry matrix**. It should be noted that this dependence implies the number of iterations to convergence depends on relative accuracy of LOS vectors.

One also analyze the relative precision of the pseudorange estimate due to the geometry of transmitter positions based on the diagonal terms of  $(\mathbf{J}^T \mathbf{J})^{-1}$ . These are called the **dilution of precision (DOP)** and for an NED reference frame for  $x - y - z$ , one can define the following DOPs:

- Vertical DOP (VDOP):  $\sigma_z$
- Horizontal DOP (HDOP):  $\sqrt{\sigma_x^2 + \sigma_y^2}$
- Position DOP (PDOP):  $\sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}$
- Geometric DOP (GDOP):  $\sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2 + c^2\sigma_t^2}$
- Time DOP (TDOP):  $\sigma_t$

### Differential Pseudorange Multilateration

Often one desires to form differential pseudoranges to obtain more accurate measurements for the positioning estimation. One common technique is to set the  $j^{\text{th}}$  transmitter as the **reference transmitter** with pseudorange equation

$$\rho_j = r_j + c(\delta t - \delta t_j) + b_{p,j} + b_{c,j} + \epsilon_j \quad (19.41)$$

Then  $\forall i \neq j$ , one can form the **between-transmitter single-difference (SD) pseudorange equation** with  $\rho_j$  as

$$\begin{aligned}\rho_i - \rho_j &= r_i + c\delta t - c\delta t_i + b_{p,i} + b_{c,i} + \epsilon_i \\ &\quad - r_j - c\delta t + c\delta t_j - b_{p,j} - b_{c,j} - \epsilon_j\end{aligned}\quad (19.42)$$

which can be rewritten succinctly by denoting the single-difference by the  $\nabla_j$ , one has

$$\nabla_j \rho_i = \nabla_j r_i - c \nabla_j \delta t_i + \nabla_j b_{p,i} + \nabla_j b_{c,i} + \nabla_j \epsilon_i \quad (19.43)$$

which explicitly eliminates the receiver clock error which could be very noisy relative to the other errors in some systems.

Thus, one has for the parameter vector

$$\vec{\beta} = \vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad (19.44)$$

In addition, note that the between-receiver SD noise term for each single-difference is

$$\nabla_j \epsilon_i = \epsilon_i - \epsilon_j \quad (19.45)$$

where the  $j$ -dependent noise that will be common to all measurements, resulting in correlated measurements. If one assumes  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ , the single-differencing with the reference transmitter produces correlated measurements with covariance matrix

$$\text{Cov}(\nabla_j \epsilon_i) = \sigma^2 \begin{bmatrix} 2 & 1 & \cdots & 1 & 1 \\ 1 & 2 & \cdots & 1 & 1 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 & 1 \\ 1 & 1 & \cdots & 1 & 2 \end{bmatrix} \quad (19.46)$$

Furthermore, note that  $\nabla_j \rho_i$  can be related to the time difference of arrival (TDOA) between signals  $i$  and  $j$ ,  $TDOA_{ij}$ , by noting

$$\rho_i - \rho_j = c (TOA_i - TOT_i) - c (TOA_j - TOT_j) \quad (19.47)$$

Then, assuming  $TOT_i - TOT_j \approx 0$ , one can form the difference

$$\nabla_j \rho_i = c (TOA_i - TOA_j) = c TDOA_{ij} \quad (19.48)$$

Thus, this technique is useful if one can measure  $TDOA_{ij}$  much more accurately than  $TOA_i$  and  $TOA_j$  individually, and one can setup the network of transmitters to transmit at the same time or know the differences with high precision, i.e. low  $\nabla_j \delta t_i$ . This technique is also known as **hyperbolic positioning** as one will obtain hyperbolic LOPs based on the differential ranges,  $r_{ij}$ .

The between-transmitter pseudorange equation allows one to form the nonlinear observation equation with  $m + 1$  transmitters as

$$\vec{y} = \begin{bmatrix} \nabla_j \rho_1 \\ \vdots \\ \nabla_j \rho_m \end{bmatrix} = \mathbf{f}(\vec{x}, \hat{\vec{\beta}}) = \begin{bmatrix} \|\hat{\vec{x}} - \vec{x}_1\|_2 - \|\hat{\vec{x}} - \vec{x}_j\|_2 \\ \vdots \\ \|\hat{\vec{x}} - \vec{x}_m\|_2 - \|\hat{\vec{x}} - \vec{x}_j\|_2 \end{bmatrix} \quad (19.49)$$

where  $j = m + 1$  denotes the reference transmitter and  $\vec{x}_i$  denotes the position of the  $i^{\text{th}}$  transmitter.

To solve this problem using NLS with GNA, assume one has an initial estimate,  $\hat{\beta}_0$ . Then, one can iteratively improve  $\hat{\beta}$  by  $\vec{\Delta}_k$  for  $k = 0, 1, \dots$ , i.e.

$$\hat{\beta}_{k+1} = \hat{\beta}_k + \vec{\Delta}_k \quad (19.50)$$

where

$$\vec{\Delta}_k = \left( \mathbf{J}^T \mathbf{J} \right)^{-1} \mathbf{J}^T \left( \vec{\mathbf{y}} - \mathbf{f}(\vec{\mathbf{x}}, \hat{\beta}_k) \right) \quad (19.51)$$

where the  $i^{\text{th}}$  row of the  $m \times 3$   $\mathbf{J}$  can be written as

$$\mathbf{J}_i = \left[ \frac{(\vec{\mathbf{x}} - \vec{\mathbf{x}}_i)^T}{\|\vec{\mathbf{x}} - \vec{\mathbf{x}}_i\|_2} - \frac{(\vec{\mathbf{x}} - \vec{\mathbf{x}}_j)^T}{\|\vec{\mathbf{x}} - \vec{\mathbf{x}}_j\|_2} \right]_{\vec{\mathbf{x}} = \hat{\mathbf{x}}_k} \quad (19.52)$$

Note that the single-difference with the reference transmitter produces correlated measurements and the estimator covariance is no longer well approximated by  $\sigma^2 (\mathbf{J}^T \mathbf{J})^{-1}$ .

Another technique for differential pseudorange multilateration considers the use of two or more receivers where one receiver is the **rover** whose position is to be estimated and may change over time, i.e. **kinematic**, while the other is the **base** whose position is *known* and typically is stationary, i.e. **static**. Thus, the position of the rover can be estimated by alternatively estimating the position of the rover relative to the base, also known as the **baseline**. Here, one can define the **rover pseudorange equation**, with dependent terms denoted by subscript  $R$ , and the  $i^{\text{th}}$  transmitter as

$$\rho_{i,R} = r_{i,R} + c(\delta t_R - \delta t_i) + b_{p,i} + b_{c,i,R} + \epsilon_{i,R} \quad (19.53)$$

Similarly, one can define the **base pseudorange equation**, with dependent terms denoted by subscript  $B$ , and the  $i^{\text{th}}$  transmitter as

$$\rho_{i,B} = r_{i,B} + c(\delta t_B - \delta t_i) + b_{p,i} + b_{c,i,B} + \epsilon_{i,B} \quad (19.54)$$

Then, one can form the **between-receiver single-difference (SD) pseudorange equation** between the rover and base as

$$\begin{aligned} \rho_{i,R} - \rho_{i,B} &= r_{i,R} + c(\delta t_R - \delta t_i) + b_{p,i} + b_{c,i,R} + \epsilon_{i,R} \\ &\quad - r_{i,B} - c(\delta t_B - \delta t_i) - b_{p,i} - b_{c,i,B} - \epsilon_{i,B} \end{aligned} \quad (19.55)$$

$$\begin{aligned} \rho_{i,R} - \rho_{i,B} &= r_{i,R} + c\delta t_R + b_{c,i,R} + \epsilon_{i,R} \\ &\quad - r_{i,B} - c\delta t_B - b_{c,i,B} - \epsilon_{i,B} \end{aligned} \quad (19.56)$$

which eliminates the transmitter clock and position errors exactly.

This equation can be rewritten succinctly by denoting the single difference between  $R$  and  $B$  terms with  $\Delta_{RB}$ , i.e.

$$\Delta_{RB}\rho_i = \Delta_{RB}r_{i,RB} + c\Delta_{RB}\delta t + \Delta_{RB}b_{c,i} + \epsilon_i \quad (19.57)$$

Furthermore, assuming the rover and base are sufficiently close relative to geographical variation in the signal speed errors, one can reduce this equation to

$$\Delta_{RB}\rho_i = r_i + c\Delta_{RB}\delta t + \Delta_{RB}\epsilon_i \quad (19.58)$$

Thus, with between-receiver SD, one need not rely on accurate models for the known sources of error in the pseudoranges. However, the signal speed error increases as the baseline increases.

The parameter vector here is the rover baseline and the relative clock error between the rover and the base,  $c\delta t_{RB}$ , i.e.

$$\vec{\beta} = \begin{bmatrix} x_R - x_B \\ y_R - y_B \\ z_R - z_B \\ c\Delta_{RB}\delta t \end{bmatrix} \quad (19.59)$$

If one assumes a static base receiver, then the base position does not change and one can choose the reference frame such that  $\vec{x}_B = [0 \ 0 \ 0]^T$  for relative positioning to the rover position  $\vec{x}_R$ . This results in the parameter vector

$$\vec{\beta} = \begin{bmatrix} x_R \\ y_R \\ z_R \\ c\Delta_{RB}\delta t \end{bmatrix} \quad (19.60)$$

The between-receiver single-differenced pseudorange equation allows one to form the nonlinear observation equation with  $m$  transmitters as

$$\vec{y} = \begin{bmatrix} \Delta_{RB}\rho_1 \\ \vdots \\ \Delta_{RB}\rho_m \end{bmatrix} = \mathbf{f}(\vec{x}, \hat{\vec{\beta}}) = \begin{bmatrix} \|\hat{\vec{x}}_R - \vec{x}_1\|_2 - \|\vec{x}_B - \vec{x}_1\|_2 + c\hat{\Delta}_{RB}\delta t \\ \vdots \\ \|\hat{\vec{x}}_R - \vec{x}_m\|_2 - \|\vec{x}_B - \vec{x}_m\|_2 + c\hat{\Delta}_{RB}\delta t \end{bmatrix} \quad (19.61)$$

If one assumes a static base receiver, then one can rewrite the nonlinear observation equation as

$$\mathbf{f}(\vec{x}, \hat{\vec{\beta}}) = \begin{bmatrix} \|\hat{\vec{x}}_R - \vec{x}_1\|_2 - \|\vec{x}_1\|_2 + c\hat{\Delta}_{RB}\delta t \\ \vdots \\ \|\hat{\vec{x}}_R - \vec{x}_m\|_2 - \|\vec{x}_m\|_2 + c\hat{\Delta}_{RB}\delta t \end{bmatrix} \quad (19.62)$$

To solve this static base problem using NLS with GNA, assume one has an initial estimate,  $\hat{\vec{\beta}}_0$ . Then, one can iteratively improve  $\hat{\vec{\beta}}_k$  by  $\vec{\Delta}_k$  for  $k = 0, 1, \dots$ , i.e.

$$\hat{\vec{\beta}}_{k+1} = \hat{\vec{\beta}}_k + \vec{\Delta}_k \quad (19.63)$$

where

$$\vec{\Delta}_k = (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T (\vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_k)) \quad (19.64)$$

where the  $i^{\text{th}}$  row of the  $m \times 4$   $\mathbf{J}$  can be written as

$$\mathbf{J}_i = \begin{bmatrix} (\vec{x}_R - \vec{x}_i)^T & 1 \end{bmatrix}_{\vec{x}_R = \hat{x}_{R,k}} \quad (19.65)$$

Lastly, another technique for differential pseudorange multilateration considers the between-transmitter differences for between-receiver differences which forms the **double-differenced (DD) pseudorange equation** as

$$\Delta_{RB}\rho_i - \Delta_{RB}\rho_j = \Delta_{RB}r_{i,RB} + c\Delta_{RB}\delta t + \Delta_{RB}\epsilon_i - \Delta_{RB}r_j - c\Delta_{RB}\delta t - \Delta_{RB}\epsilon_j \quad (19.66)$$

$$\Delta_{RB}\rho_i - \Delta_{RB}\rho_j = \Delta_{RB}r_i - \Delta_{RB}r_j + \epsilon_{i,RB} - \epsilon_{j,RB} \quad (19.67)$$

which can be rewritten succinctly as

$$\nabla_j \Delta_{RB}\rho_i = \nabla_j \Delta_{RB}r_i + \nabla_j \Delta_{RB}\epsilon_i \quad (19.68)$$

which eliminates the receiver clock error and leaves the double-differenced true range plus the double-differenced unmodeled errors. This technique can also reduce the contribution of the signal speed errors.

Here, one has for the parameter vector, the relative position of the rover to the base, i.e.

$$\vec{\beta} = \Delta_{RB} \vec{x} = \begin{bmatrix} x_R \\ y_R \\ z_R \end{bmatrix} - \begin{bmatrix} x_B \\ y_B \\ z_B \end{bmatrix} \quad (19.69)$$

where  $[x_B \ y_B \ z_B]^T$  is assumed known. If one assumes a static base receiver, then the base position does not change and one can choose the reference frame such that  $\vec{x}_B = [0 \ 0 \ 0]^T$  for relative positioning to the rover position  $\vec{x}_R$ . This results in the parameter vector

$$\vec{\beta} = \vec{x}_R = \begin{bmatrix} x_R \\ y_R \\ z_R \end{bmatrix} \quad (19.70)$$

In addition, note that the double-differencing noise term for each double-difference is

$$\nabla_j \Delta_{RB}\epsilon_i = \epsilon_{i,R} - \epsilon_{i,B} - \epsilon_{j,R} + \epsilon_{j,B} \quad (19.71)$$

where the two  $j$ -dependent noise will be common to all measurements, results in correlated measurements. If one assumes  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ , the double-differencing produces correlated measurements with covariance

$$\text{Cov}(\nabla_j \Delta_{RB}\epsilon_i) = \sigma^2 \begin{bmatrix} 4 & 2 & \cdots & 2 & 2 \\ 2 & 4 & \cdots & 2 & 2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 2 & 2 & \cdots & 4 & 2 \\ 2 & 2 & \cdots & 2 & 4 \end{bmatrix} \quad (19.72)$$

The double-differenced pseudorange equation allows one to form the nonlinear observation equation with  $m + 1$  transmitters as

$$\vec{y} = \begin{bmatrix} \nabla_j \Delta_{RB}\rho_1 \\ \vdots \\ \nabla_j \Delta_{RB}\rho_m \end{bmatrix} = \mathbf{f}(\vec{x}, \hat{\vec{\beta}}) = \begin{bmatrix} (\|\hat{\vec{x}}_R - \vec{x}_1\|_2 - \|\vec{x}_B - \vec{x}_1\|_2) - (\|\hat{\vec{x}}_R - \vec{x}_j\|_2 - \|\vec{x}_B - \vec{x}_j\|_2) \\ \vdots \\ (\|\hat{\vec{x}}_R - \vec{x}_m\|_2 - \|\vec{x}_B - \vec{x}_m\|_2) - (\|\hat{\vec{x}}_R - \vec{x}_j\|_2 - \|\vec{x}_B - \vec{x}_j\|_2) \end{bmatrix} \quad (19.73)$$

If one again assumes a static base receiver, then one can rewrite the nonlinear observation equation as

$$\vec{y} = \begin{bmatrix} \nabla_j \Delta_{RB}\rho_1 \\ \vdots \\ \nabla_j \Delta_{RB}\rho_m \end{bmatrix} = \mathbf{f}(\vec{x}, \hat{\vec{\beta}}) = \begin{bmatrix} (\|\vec{x}_R - \vec{x}_1\|_2 - \|\vec{x}_1\|_2) - (\|\vec{x}_R - \vec{x}_j\|_2 - \|\vec{x}_j\|_2) \\ \vdots \\ (\|\vec{x}_R - \vec{x}_m\|_2 - \|\vec{x}_m\|_2) - (\|\vec{x}_R - \vec{x}_j\|_2 - \|\vec{x}_j\|_2) \end{bmatrix} \quad (19.74)$$

To solve this static base problem using NLS with GNA, assume one has an initial estimate,  $\hat{\beta}_0$ . Then, one can iteratively improve  $\hat{\beta}_k$  by  $\vec{\Delta}_k$  for  $k = 0, 1, \dots$ , i.e.

$$\hat{\beta}_{k+1} = \hat{\beta}_k + \vec{\Delta}_k \quad (19.75)$$

where

$$\vec{\Delta}_k = \left( \mathbf{J}^T \mathbf{J} \right)^{-1} \mathbf{J}^T \left( \vec{\mathbf{y}} - \mathbf{f}(\vec{\mathbf{x}}, \hat{\beta}_k) \right) \quad (19.76)$$

where the  $i^{\text{th}}$  row of the  $m \times 3$   $\mathbf{J}$  can be written as

$$\mathbf{J}_i = \left[ \frac{(\vec{\mathbf{x}}_R - \vec{\mathbf{x}}_i)^T}{\|\vec{\mathbf{x}}_R - \vec{\mathbf{x}}_i\|_2} - \frac{(\vec{\mathbf{x}}_R - \vec{\mathbf{x}}_j)^T}{\|\vec{\mathbf{x}}_R - \vec{\mathbf{x}}_j\|_2} \right]_{\vec{\mathbf{x}}_R = \hat{\vec{\mathbf{x}}}_{R,k}} \quad (19.77)$$

### Pseudorange-Rate and Doppler Positioning

The true range rate,  $\dot{r}_i$ , for the  $i^{\text{th}}$  transmitter is related to the velocity of the receiver,  $\vec{\mathbf{x}} = [\dot{x} \ \dot{y} \ \dot{z}]^T$ , and the velocity of the  $i^{\text{th}}$  transmitter,  $\vec{\mathbf{x}}_i = [\dot{x}_i \ \dot{y}_i \ \dot{z}_i]^T$ , through the relative line-of-sight (LOS) vector, i.e.

$$\dot{r}_i = (\vec{\mathbf{x}} - \vec{\mathbf{x}}_i)^T \frac{(\vec{\mathbf{x}} - \vec{\mathbf{x}}_i)}{\|\vec{\mathbf{x}} - \vec{\mathbf{x}}_i\|_2} = \frac{(\vec{\mathbf{x}} - \vec{\mathbf{x}}_i)^T}{\|\vec{\mathbf{x}} - \vec{\mathbf{x}}_i\|_2} (\vec{\mathbf{x}} - \vec{\mathbf{x}}_i) \quad (19.78)$$

or

$$\dot{r}_i = \frac{(x - x_i)(\dot{x} - \dot{x}_i) + (y - y_i)(\dot{y} - \dot{y}_i) + (z - z_i)(\dot{z} - \dot{z}_i)}{\sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2}} \quad (19.79)$$

where  $\vec{\mathbf{x}} = [x \ y \ z]^T$  is the receiver position and  $\vec{\mathbf{x}}_i = [x_i \ y_i \ z_i]^T$  is the  $i^{\text{th}}$  transmitter's position.

The **pseudorange-rate equation**, one can relate the **pseudorange-rate**,  $\dot{\rho}_i$ , to the true range rate,  $\dot{r}_i$ , between the receiver and the  $i^{\text{th}}$  transmitter as

$$\dot{\rho}_i = \dot{r}_i + c\dot{t} + \epsilon \quad (19.80)$$

where  $\dot{t}$  is the **receiver clock drift** and accounts for the timing drift in the rate measurement. Lastly,  $\epsilon$  accounts for any additional unknown random errors, e.g. electrical noise.

A energy wave receiver typically measures the **Doppler shift** at the receiver,  $\Delta f_i$ , of a *known* signal from the  $i^{\text{th}}$  transmitter with *known* frequency of transmission,  $f$ . Thus, one can form the **pseudorange-rate measurement**,  $\dot{\rho}_i$ , for the  $i^{\text{th}}$  transmitter by

$$\dot{\rho}_i = c \frac{\Delta f_i}{f} \quad (19.81)$$

Thus, by substitution, one has the **Doppler positioning equation** which relates the Doppler shift,  $\Delta f_i$

$$\Delta f_i = \frac{f}{c} \frac{(\vec{\mathbf{x}} - \vec{\mathbf{x}}_i)^T}{\|\vec{\mathbf{x}} - \vec{\mathbf{x}}_i\|_2} (\vec{\mathbf{x}} - \vec{\mathbf{x}}_i) + f\dot{t} + \epsilon \quad (19.82)$$

which depends on seven unknowns, receiver position, velocity, and clock drift. One may also include bias errors due to the assumed transmitter positions and the signal transmission speed and frequency.

Thus, in pseudorange-rate positioning, one has for the parameter vector

$$\vec{\beta} = \begin{bmatrix} \vec{x} \\ \dot{\vec{x}} \\ f\dot{t} \end{bmatrix} \quad (19.83)$$

which implies that **Doppler positioning** requires seven or more transmitters to be able to obtain a position *and* velocity estimate. As such, Doppler positioning is more often used for *static* receivers, so only four transmitters required the Doppler range equation allows one to form a nonlinear observation equation with  $m$  transmitters as

$$\vec{y} = \begin{bmatrix} \Delta f_1 \\ \vdots \\ \Delta f_m \end{bmatrix} = \mathbf{f}(\vec{x}, \hat{\vec{\beta}}) = \begin{bmatrix} \frac{f}{c} \frac{(\vec{x} - \vec{x}_1)^T}{\|\vec{x} - \vec{x}_1\|_2} (\vec{x} - \vec{x}_1) + f\dot{t} \\ \vdots \\ \frac{f}{c} \frac{(\vec{x} - \vec{x}_m)^T}{\|\vec{x} - \vec{x}_m\|_2} (\vec{x} - \vec{x}_m) + f\dot{t} \end{bmatrix} \quad (19.84)$$

To solve this problem using NLS with GNA, assume one has an initial estimate,  $\hat{\vec{\beta}}_0$ . Then, one can iteratively improve  $\hat{\vec{\beta}}$  by  $\vec{\Delta}_k$  for  $k = 0, 1, \dots$  as

$$\hat{\vec{\beta}}_{k+1} = \hat{\vec{\beta}}_k + \vec{\Delta}_k \quad (19.85)$$

where

$$\vec{\Delta}_k = \left( \mathbf{J}^T \mathbf{J} \right)^{-1} \mathbf{J}^T \left( \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_k) \right) \quad (19.86)$$

where the  $i^{\text{th}}$  row of  $m \times 7 \mathbf{J}$  can be written as

$$\mathbf{J}_i = \left[ \begin{array}{c} \frac{f}{c} \left( \frac{(\vec{x} - \vec{x}_i)}{\|\vec{x} - \vec{x}_i\|_2} - (\vec{x} - \vec{x}_i) (\vec{x} - \vec{x}_i)^T \frac{(\vec{x} - \vec{x}_i)}{\|\vec{x} - \vec{x}_i\|_2^3} \right)^T \\ \frac{f}{c} \frac{(\vec{x} - \vec{x}_i)^T}{\|\vec{x} - \vec{x}_i\|_2} \\ 1 \end{array} \right]_{\vec{x} = \hat{\vec{x}}_k, \dot{\vec{x}} = \hat{\vec{x}}_k}^T \quad (19.87)$$

## References

For more information, please refer to the following

- S. Gleason and D. Gebre-Egziabher, “GNSS Applications and Methods,” Artech House, 2012
- C. Jekeli, “Inertial Navigation Systems with Geodetic Applications,” de Gruyter, 2001

## 19.3 Phase-Based Positioning Systems

### Phase-Based Ranging

Another technique for positioning is to accurately measure the phase of a transmitted wave signal at a *constant* frequency with wavelength,  $\lambda$ . Assuming a receiver can track the phase of the wave signal in time,  $\phi$ , as **fractional cycles**. Then, if the transmitter has created a very constant wavelength, a replicated wave signal

in the receiver signal processing can be roughly matched in time to when the signal was transmitted, thus the fractional phase cycles correspond to a *distance* from the beginning of the current cycle defined as the wavelength multiplied by the fractional cycle, i.e.  $\lambda\phi$ . However, the portion of the signal that extends from the beginning of the currently measured cycle to the signal source,  $B$ , is unknown and, hence, is called the **phase ambiguity**.

Thus, the **phase-based range equation** for the  $i^{\text{th}}$  transmitter is defined as

$$\lambda\phi_i + B_i = r_i + c(\delta t - \delta t_i) + K + b_{p,i} + b_{c,i} + \epsilon_i \quad (19.88)$$

where  $k$  is the receiver **phase delay**, or **phase ambiguity** in its measurement and

$$B_i = K_i + \lambda N_i \quad (19.89)$$

where  $K_i$  is the transmitter **phase delay**, or *phase ambiguity*, and  $N_i$  is the **integer ambiguity**, i.e. the integer number of wavelengths from the transmitter to the beginning of the fractional cycles distance,  $\lambda\phi$ . It should be noted that  $K$  and  $K_i$  are also known as the **fractional ambiguities** as they can be any real number as opposed to  $N_i$ .

Furthermore, substituting for  $r_i$  and  $B_i$ , then by rearranging one has

$$\lambda\phi_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} + c(\delta t - \delta t_i) + (K - K_i) - \lambda N_i + b_{p,i} + b_{c,i} + \epsilon_i \quad (19.90)$$

where it should be noted that rotating transmitters and receivers can also cause an additional additive **phase wind-up** error term,  $\lambda\omega$ , due to any angular velocity interacting with circularly polarized wave signals. In phase-based positioning, one may be able to refine the initially calculated  $\lambda\phi_i$  based on highly accurate transmitter-dependent errors,  $\delta t_i$ ,  $K_i$ ,  $b_{p,i}$ , and  $b_{c,i}$ , one has

$$\lambda\phi_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} + c\delta t + K - \lambda N_i + \epsilon_i \quad (19.91)$$

where  $x$ ,  $y$ ,  $z$ ,  $\delta t$ ,  $K$ , and  $N_i$  are the six unknown parameters.

Thus, in un-differenced phase-based positioning, one has for the parameter vector

$$\vec{\beta} = \begin{bmatrix} x \\ y \\ z \\ c\delta t \\ K \\ N_i \end{bmatrix} \quad (19.92)$$

where each transmitter will have an *unknown* integer ambiguity that must be estimated. This additional consideration of each  $N_i$  leads to additional estimation methods in phase-based positioning, namely estimating across multiple time steps as well as **ambiguity fixing**. The other potential difficulty in phase-based positioning is obtaining accurate enough models for the transmitter-dependent errors to be eliminated or reduced to a sufficient extent for estimating all  $N_i$ . Thus, the primary obstacles for un-differenced phase-based positioning are precise models for the transmitter position errors, clock errors, phase delays, and transmission speed errors which are typically only available in post-processing of phase data especially as the fractional ambiguities are not constant, but drift over time. Thus, differential phase-based positioning is typically used for real-time positioning and is discussed in this section.

## Differential Phase-Based Positioning

Due to  $N_i$  appearing in the parameter vector, using a between-transmitter phase measurement will only increase the number of unknowns in phase-based positioning. Thus, for single-differencing one typically uses two or more receivers in a **rover** and **base** configuration with dependent terms distinguished by subscripts  $R$  and  $B$ , respectively. Thus, for the  $i^{\text{th}}$  transmitter, one has

$$\lambda\phi_{i,R} = r_{i,R} + c(\delta t_R - \delta t_i) + (K_R - K_i) - \lambda N_{i,R} + b_{p,i} + b_{c,i} + \epsilon \quad (19.93)$$

and

$$\lambda\phi_{i,B} = r_{i,B} + c(\delta t_B - \delta t_i) + (K_B - K_i) - \lambda N_{i,B} + b_{p,i} + b_{c,i} + \epsilon \quad (19.94)$$

Then, one can define the **single-differenced (SD) phase-based range equation** for each  $i^{\text{th}}$  transmitter as

$$\lambda\Delta_{RB}\phi_i = \Delta_{RB}r_i + c\Delta_{RB}\delta t + \Delta_{RB}K - \lambda\Delta_{RB}N_i + \Delta_{RB}\epsilon_i \quad (19.95)$$

where the term  $\Delta_{RB}$  represents the difference of the corresponding rover term with the base term.

Thus, the SD operation eliminates the transmitter-dependent errors, i.e.  $\delta t_i$ ,  $K_i$ , and  $b_{p,i}$ , as well as  $b_{c,i}$  assuming a sufficiently small baseline. However, the differential clock and phase errors remain. Thus, one often sets the  $j = m + 1$  as the reference transmitter, then  $\forall i \neq j$ , one can form the **double-difference (DD) phase-based range equation** for each  $i^{\text{th}}$  transmitter as

$$\lambda\nabla_j\Delta_{RB}\phi_i = \nabla_j\Delta_{RB}r_i - \lambda\nabla_j\Delta_{RB}N_i + \nabla_j\Delta_{RB}\epsilon_i \quad (19.96)$$

Thus, in double-differenced phase-based positioning, one has for the parameter vector

$$\vec{\beta} = \begin{bmatrix} x_R - x_B \\ y_R - y_B \\ z_R - z_B \\ \nabla_j\Delta_{RB}N_1 \\ \vdots \\ \nabla_j\Delta_{RB}N_m \end{bmatrix} \quad (19.97)$$

If one assumes a static base receiver, then the base position does not change and one can choose the reference frame such that  $\vec{x}_B = [0 \ 0 \ 0]^T$  for relative positioning to the rover position  $\vec{x}_R$ . This results in the parameter vector

$$\vec{\beta} = \begin{bmatrix} x_R \\ y_R \\ z_R \\ \nabla_j\Delta_{RB}N_1 \\ \vdots \\ \nabla_j\Delta_{RB}N_m \end{bmatrix} \quad (19.98)$$

one can see the DD phase measurements provide  $m$  equations and  $m + 3$  unknowns, i.e. the position of rover  $\vec{x}_R$  and the  $m$  integer ambiguities,  $\nabla_j\Delta_{RB}N_i$ , which does not allow a determinable solution at a *single* time step.

However, if the receiver can track the additional cycles from  $i^{\text{th}}$  transmitter,  $C_i$ , across multiple time steps. Then, as  $\nabla_j \Delta_{RB} N_i$  are *constant* unknowns with respect to time, if four or more *of the same* transmitters are sampled at each subsequent time step, the number of unknowns will decrease since the *new* rover position has three unknown coordinates to consider. This concept requires a continuous lock on the signal in order to keep constant the initial integer number of unknown cycles,  $N$ . Thus, the rover and base receivers can alternatively measure the phase as a **fractional cycle count**,  $\Phi_i$ . This concept is called **cycle counting**. With this alternative measurement, one has at time step  $k$  for the phase of the signal from the  $i^{\text{th}}$  transmitter

$$\Phi_{i,k} = \lambda \phi_{i,k} \quad (19.99)$$

and at time step  $k + n$  for the phase of the signal from the  $i^{\text{th}}$  transmitter

$$\Phi_{i,k+n} = \lambda \phi_{i,k+n} = \lambda \phi_{i,k} + C_{i,n} \quad (19.100)$$

where  $C_{i,n}$  is the fractional cycle count from  $k$  to  $k + n$ . Then, one can form the alternate **double-differenced cycle count range equation** for each  $i^{\text{th}}$  transmitter at time step  $k$  as

$$\nabla_j \Delta_{RB} \Phi_{i,k} = \nabla_j \Delta_{RB} r_{i,k} + \nabla_j \Delta_{RB} \lambda N_i + \nabla_j \Delta_{RB} \epsilon_{i,k} \quad (19.101)$$

where the sign of  $N$  has been changed without loss of generality due to the arbitrary start of the cycle counting.

Defining the following vectors for the  $i = 1, \dots, m$  transmitters at time step  $k$  as

$$\vec{\Phi}_k = [\Phi_1 \ \dots \ \Phi_m]^T \quad (19.102)$$

$$\vec{r}_k = [r_1 \ \dots \ r_m]^T \quad (19.103)$$

$$\vec{N} = [N_1 \ \dots \ N_m]^T \quad (19.104)$$

$$\vec{\epsilon}_k = [\epsilon_1 \ \dots \ \epsilon_m]^T \quad (19.105)$$

one has the vector equation

$$\nabla_j \Delta_{RB} \vec{\Phi}_k = \nabla_j \Delta_{RB} \vec{r}_k + \lambda \nabla_j \Delta_{RB} \vec{N} + \nabla_j \Delta_{RB} \vec{\epsilon}_k \quad (19.106)$$

Then, for  $k = 1, \dots, n$ , one can form the stacked nonlinear observation equation as

$$\vec{y} = \begin{bmatrix} \nabla_j \Delta_{RB} \vec{\Phi}_1 \\ \vdots \\ \nabla_j \Delta_{RB} \vec{\Phi}_n \end{bmatrix} = \mathbf{f}(\vec{x}, \hat{\beta}) = \begin{bmatrix} \left[ \left( \|\vec{x}_{R,1} - \vec{x}_1\|_2 - \|\vec{x}_B - \vec{x}_1\|_2 \right) - \left( \|\vec{x}_{R,1} - \vec{x}_j\|_2 - \|\vec{x}_B - \vec{x}_j\|_2 \right) \right] \\ \vdots \\ \left[ \left( \|\vec{x}_{R,m} - \vec{x}_m\|_2 - \|\vec{x}_B - \vec{x}_m\|_2 \right) - \left( \|\vec{x}_{R,1} - \vec{x}_j\|_2 - \|\vec{x}_B - \vec{x}_j\|_2 \right) \right] \\ \vdots \\ \left[ \left( \|\vec{x}_{R,n} - \vec{x}_m\|_2 - \|\vec{x}_B - \vec{x}_m\|_2 \right) - \left( \|\vec{x}_{R,n} - \vec{x}_j\|_2 - \|\vec{x}_B - \vec{x}_j\|_2 \right) \right] \\ \vdots \\ \left[ \left( \|\vec{x}_{R,n} - \vec{x}_j\|_2 - \|\vec{x}_B - \vec{x}_j\|_2 \right) - \left( \|\vec{x}_{R,n} - \vec{x}_j\|_2 - \|\vec{x}_B - \vec{x}_j\|_2 \right) \right] \\ + \lambda \begin{bmatrix} I_{(m) \times (m)} \\ \vdots \\ I_{(m) \times (m)} \end{bmatrix} [\nabla_j \Delta_{RB} \vec{N}] \end{bmatrix} \quad (19.107)$$

If one again assumes a static base receiver, then one can rewrite the nonlinear observation equation as

$$\mathbf{f}(\vec{\mathbf{x}}, \hat{\vec{\beta}}) = \begin{bmatrix} \left( \|\vec{x}_{R,1} - \vec{x}_1\|_2 - \|\vec{x}_1\|_2 \right) - \left( \|\vec{x}_{R,1} - \vec{x}_j\|_2 - \|\vec{x}_j\|_2 \right) \\ \vdots \\ \left( \|\vec{x}_{R,m} - \vec{x}_m\|_2 - \|\vec{x}_m\|_2 \right) - \left( \|\vec{x}_{R,1} - \vec{x}_j\|_2 - \|\vec{x}_j\|_2 \right) \\ \vdots \\ \left( \|\vec{x}_{R,n} - \vec{x}_1\|_2 - \|\vec{x}_1\|_2 \right) - \left( \|\vec{x}_{R,n} - \vec{x}_j\|_2 - \|\vec{x}_j\|_2 \right) \\ \vdots \\ \left( \|\vec{x}_{R,n} - \vec{x}_m\|_2 - \|\vec{x}_m\|_2 \right) - \left( \|\vec{x}_{R,n} - \vec{x}_j\|_2 - \|\vec{x}_j\|_2 \right) \end{bmatrix} + \lambda \begin{bmatrix} I_{(m) \times (m)} \\ \vdots \\ I_{(m) \times (m)} \end{bmatrix} [\nabla_j \Delta_{RB} \vec{N}] \quad (19.108)$$

where the entire parameter vector to be estimated is

$$\vec{\beta} = \begin{bmatrix} \vec{x}_{R,1} \\ \vdots \\ \vec{x}_{R,n} \\ \nabla_j \Delta_{RB} \vec{N} \end{bmatrix} \quad (19.109)$$

where  $\vec{x}_{R,1}, \dots, \vec{x}_{R,n}$  are the rover positions from  $k = 1, \dots, n$ , and each transmitter position,  $\vec{x}_1, \dots, \vec{x}_m, \vec{x}_j$ , may also change with  $k$ , though not explicitly written above.

To solve this static base problem using NLS with GNA, assume one has an initial estimate,  $\hat{\vec{\beta}}_0$ . Then, one can iteratively improve  $\hat{\vec{\beta}}_p$  by  $\Delta_p$  for  $p = 1, \dots$

$$\vec{\Delta}_p = \left( \mathbf{J}^T \mathbf{J} \right)^{-1} \mathbf{J}^T \left( \vec{\mathbf{y}} - \mathbf{f}(\vec{\mathbf{x}}, \hat{\vec{\beta}}_p) \right) \quad (19.110)$$

where

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & 0 & \cdots & \cdots & 0 & \lambda I_{m \times m} \\ 0 & \ddots & \ddots & \ddots & 0 & \lambda I_{m \times m} \\ 0 & \ddots & \mathbf{J}_k & \ddots & 0 & \lambda I_{m \times m} \\ 0 & \ddots & \ddots & \ddots & 0 & \lambda I_{m \times m} \\ 0 & \cdots & \cdots & 0 & \mathbf{J}_n & \lambda I_{m \times m} \end{bmatrix} \quad (19.111)$$

and the  $i^{\text{th}}$  row of each  $m \times 3$   $\mathbf{J}_k$  matrix can be written as

$$\mathbf{J}_{k,i} = \left[ \frac{\vec{x}_{R,k} - \vec{x}_i}{\|\vec{x}_{R,k} - \vec{x}_i\|_2} - \frac{\vec{x}_{R,k} - \vec{x}_j}{\|\vec{x}_{R,k} - \vec{x}_j\|_2} \right]^T \Big|_{\vec{x}_{R,k} = \hat{\vec{x}}_{R,k,p}} \quad (19.112)$$

which provides the estimated rover positions at each time step  $k$  as  $\hat{\vec{x}}_{R,k}$ , and the estimated  $\nabla_j \Delta_{RB} \hat{N}_i$  as “real” numbers, a solution known as the **float solution** due to the computer science term for decimal

numbers in algorithms. This is an imprecise estimate due to the fact that  $\nabla_j \Delta_{RB} \hat{N}_i$  should be constrained to be integers.

### Integer Ambiguity Fixing

To obtain the most accurate positioning solution beyond the NLS estimate, regardless whether using undifferenced or double-differenced phase measurements, one must “fix” the integer ambiguities as integers, i.e. obtain a **fixed solution** for which there are three primary methods with the simplest as rounding the float solution to the nearest integers, i.e. the **rounding** method.

A second method is to select  $p$  “candidate sets” of integers “near” the float solution, i.e.  $\vec{N}_s \in \mathbb{S}$  with  $s = 1, \dots, p$ . Then, compute the positioning solution for each candidate set as

$$\hat{\vec{x}}_{1:n, \vec{N}_s} = [\hat{\vec{x}}_1, \dots, \hat{\vec{x}}_n] \text{ assuming } \nabla_j \Delta_{RB} \vec{N} = \vec{N}_s \quad (19.113)$$

for  $s = 1, \dots, p$ . Finally, the optimal fixed solution is chosen as the candidate set which minimizes the smallest post-fit residual between the measurements and the expected measurement based on the estimated position for that candidate set, i.e.

$$\hat{N} = \min_{\vec{N}_s \in \mathbb{S}} \left\| \vec{y} - \mathbf{f} \left( \vec{x}, [\hat{\vec{x}}_{1:n, \vec{N}_s} \ \vec{N}_s]^T \right) \right\|_2 \quad (19.114)$$

It should be noted that this method requires possibly computing the positioning solution for a large number of candidate sets which can be computationally expensive if the number of measurements and time steps is large. Thus, this method is typically called the **brute force** method if care is not taken when selecting the candidate sets. An optimal way to choose a small set of integer candidates is to use the **Least-squares AMbiguity Decorrelation Adjustment (LAMBDA)** method which identifies the most likely candidates using the geometry matrix and the correlation between the measurements to find the most likely candidate sets from a statistical perspective. Implementations of the LAMBDA method are available for free at [www.tudelft.nl](http://www.tudelft.nl).

A third method uses an additionally available pseudorange measurement from all transmitters, i.e.  $\vec{\rho}_k$  for  $k = 1, \dots, n$ , uses the integers which minimize the residual between phase-based measurement and the *expected* phase measurements based on the pseudorange solution. This is called the **geometry-free** method as its measurement model does not use the LOS vectors, i.e. the geometry matrix, for  $\hat{\vec{N}}$ . For the double-differenced phases and *expected* phases based on the double-differenced pseudoranges, i.e.

$$\nabla_j \Delta_{RB} \hat{N} = \text{round} \left( \text{mean} \left( \begin{bmatrix} \nabla_j \Delta_{RB} \vec{\Phi}_1 \\ \vdots \\ \nabla_j \Delta_{RB} \vec{\phi}_n \end{bmatrix} - \frac{1}{\lambda} \begin{bmatrix} \nabla_j \Delta_{RB} \vec{\rho}_1 \\ \vdots \\ \nabla_j \Delta_{RB} \vec{\rho}_n \end{bmatrix} \right) \right) \quad (19.115)$$

### Cycle Slip and Triple-Differencing

A **cycle slip** is a discontinuity in the receiver’s signal processor digital phase-lock loop (DPLL) due to a loss in tracking a transmitter’s signal. Cycle slip can be caused by several different factors including power loss, a failure of the receiver software, a malfunctioning clock oscillator, and most commonly, obstructions to the LOS to the transmitter such that the signal is lost. However, once the transmitter reappears, the reacquisition

of the signal occurs and the DPLL can continue to count the cycles. The precision of the phase measurement requires an accurate counting of the cycles that have occurred to maintain a fixed integer ambiguity required for phase-based positioning.

However, in many applications one may desire to fix cycle slips by estimating the missing number using the other transmitter measurements to estimate the discrepancy. This type of analysis is often done by first using an additional **between-epochs difference** to form the **triple-difference cycle count range equation**, also known as the **transmitter-receiver-time triple-difference**. This technique takes the double-differenced phase-based range equation at different time steps  $k$ , i.e.

$$\lambda \nabla_j \Delta_{RB} \vec{\phi}_k = \nabla_j \Delta_{RB} \vec{r}_k - \lambda \nabla_j \Delta_{RB} \vec{N} + \nabla_j \Delta_{RB} \vec{\epsilon}_k \quad (19.116)$$

and computes the difference with one of the time steps, e.g. differencing  $k = 1, \dots, n$  with  $k = 0$ , to form

$$\Delta_0 \nabla_j \Delta_{RB} \vec{\phi}_k = \nabla_j \Delta_{RB} \vec{\phi}_k - \nabla_j \Delta_{RB} \vec{\phi}_0 \quad (19.117)$$

$$\lambda \Delta_0 \nabla_j \Delta_{RB} \vec{\phi}_k = \Delta_0 \nabla_j \Delta_{RB} \vec{r}_k + \Delta_0 \nabla_j \Delta_{RB} \vec{\epsilon}_k \quad (19.118)$$

However, as before, though this eliminates a common known error, it increases the effect of the noise  $\epsilon$ . Thus, triple-differencing is not as useful as double-differenced phase-based solutions for positioning. However, because it is less sensitive to cycle slips, it allows for corrections to be made through residual analysis of the component double-differences of the triple-differenced solution.

## References

For more information, please refer to the following

- S. Gleason and D. Gebre-Egziabher, “GNSS Applications and Methods,” Artech House, 2012

## 19.4 Satellite-Based Positioning Systems

### GNSS Code Phase Positioning

GNSS uses modulated radio signals made up of a central carrier frequency and encodes digital code through binary code modulation (BCM) onto the carrier signal. The original design and standard method of using GNSS is reading and interpreting this digital code to form pseudorange measurements for positioning, a technique also known as the **code phase positioning** for GNSS. Using its own clock, GNSS receiver measures the time of arrival at the receiver,  $TOA_i$ , of the GPS code phase signal from the  $i^{\text{th}}$  satellite which has an encoded time of transmission,  $TOT_i$ . Thus, one can form the GNSS pseudorange measurement by

$$\rho_i = c (TOA_i - TOT_i) \quad (19.119)$$

where the GNSS timing is based off the fundamental clock rate,  $F_0$ . For the **GNSS pseudorange equation**, one can relate the pseudorange  $\rho_i$  to the true range  $r_i$  between the receiver and the  $i^{\text{th}}$  satellite as

$$\rho_i = r_i + c (\delta t - \delta t_i) + b_{E,i} + b_{I,i} + b_{T,i} + \epsilon_{M-P} + \epsilon \quad (19.120)$$

where  $\epsilon$  is the random unknown error sources, and  $c\delta t$ ,  $c\delta t_i$ ,  $b_{E,i}$ ,  $b_{I,i}$ ,  $b_{T,i}$ , and  $\epsilon_{M-P}$  are the known error sources to  $\rho_i$ , collectively known as the **User Equivalent Range Error (UERE)** in GNSS applications.

The individual terms of the UERE and the source of their error can be summarized as follows. The unknown  $\delta t$  results from the receiver clock error due to an imperfect oscillator and must be estimated alongside the position estimate. The control segments and IGS monitor the **satellite clock error**,  $\delta t_i$ , as well as the **satellite ephemeris error**,  $b_{E,i}$ , and provide models to remove this error. Both of these terms can be improved through optimal smoothing *after* real-time observations of the GNSS signals. The signal speed transmission errors are primarily due to the **ionospheric delay**,  $b_{I,i}$ , and the **tropospheric delay**,  $b_{T,i}$ , depend on the atmospheric conditions and are collectively known as the **atmospheric delay**. Weather stations around the world provide regional information about the measured atmospheric delay, but typically these model improvements are made *after* real-time observations. Lastly, multi-path errors,  $\epsilon_{M-P}$ , occur when the GNSS signal strongly reflects off a nearby object before reaching the receiver, thereby negating the assumption of straight flight from the transmitter to the receiver.

### GNSS Carrier Phase Positioning

An alternative to the code phase positioning is **carrier phase positioning**. The primary reason that using the GNSS carrier phase in a phase-based positioning algorithm is the fact that it offers increased precision of range due to the fact that its wavelength is much shorter than the modulated encoded signal. A convenient approximation for the maximum resolution of the encoded and carrier signals is roughly one percent of the signal wavelength, also known as the **one percent rule**. For reference, the C/A code phase has a frequency  $f$  of 1.023 MHz which corresponds to a wavelength  $\lambda$  of 293 m by  $\lambda = \frac{c}{f}$  where  $c \approx 2.998 \times 10^8$  m/s. Then, by the one percent rule, this corresponds to a resolution of approximately  $\pm 3$  m. In practice, C/A positioning offers around  $\pm 20$  m of accuracy. Conversely, the L1 and L2 carrier signals have frequencies of 1575.42 and 1227.60 MHz which translates to 19 and 24 cm wavelengths. This effectively puts the one percent rule resolution of the carrier phase measurements as  $\pm 2$  mm.

Similar to the code phase, the **GNSS carrier phase range equation**, one can relate the carrier phase fractional cycle count,  $\Phi_i$ , to the true range,  $r_i$ , between the receiver and the  $i^{\text{th}}$  satellite as

$$\Phi_i = \lambda\phi_i = r_i + c(\delta t - \delta t_i) + (K - K_i) + \lambda N_i + b_{E,i} - b_{I,i} + b_{T,i} + \lambda\omega + \epsilon_{M-P} + \epsilon \quad (19.121)$$

where  $\epsilon$  is the random unknown error sources, and  $c\delta t$ ,  $c\delta t_i$ ,  $K$ ,  $K_i$ ,  $b_{E,i}$ ,  $b_{I,i}$ ,  $b_{T,i}$ , and  $\epsilon_{M-P}$  are the carrier phase **user equivalent range error (UERE)**. Note that the ionosphere error is negative for the carrier phase measurement.

Also, the additional  $\lambda\omega$  is known as the **phase wind-up** due to the circular polarization of the GNSS signal. This wind-up can often be estimated using the **geometry-free combination** for the carrier phase as

$$\Phi_{G-F} = \Phi_1 - \Phi_2 \quad (19.122)$$

where  $\Phi_1$  and  $\Phi_2$  are GNSS cycle counts on different frequencies. It should also be noted that one can form the **wide-lane combination** for the carrier phase as

$$\Phi_{W-L} = \frac{f_1\Phi_1 - f_2\Phi_2}{f_1 - f_2} \quad (19.123)$$

where  $\Phi_1$ , and  $\Phi_2$  are GNSS cycle counts on frequencies,  $f_1$  and  $f_2$ , respectively, which allows the integer ambiguities to be increased up to four times larger for the wavelength distance which makes the ambiguity fixing easier to estimate.

GNSS un-differenced carrier phase positioning is known as **precise point positioning (PPP)**. In order to remove as much error as possible, PPP requires error models for the precise transmitter orbits and clocks, the atmospheric effects, the antenna biases and orientation, and solid Earth tides with optional ocean loading and pole tides. Normally, the IGS data products for these errors are used due to their high accuracy. PPP also uses dual-frequency GNSS to remove the ionospheric errors. The PPP algorithm uses both carrier phase and pseudorange measurements across time in a nonlinear Kalman filter to estimate the receiver position and the receiver clock bias as well as the  $m$  transmitter combined fractional and integer phase ambiguities, and the wet tropospheric delay. However, a large convergence time is required for PPP filters to resolve the ambiguity fix, typically on the order of tens of minutes. Thus, PPP is primarily used as a post-processed positioning technique as an alternative to GNSS double-differenced carrier phase positioning as PPP allows a ubiquitous solution as no baseline limitations apply for the positioning.

GNSS double-differenced carrier phase positioning is known as **real-time kinematic (RTK)** and **post-processed kinematic (PPK)** depending on the time when the positioning is performed. The simplest form of RTK uses a base station with a radio data link to broadcast the raw observational data, and/or other correctional data, at about a 1 Hz rate while including its ECEF location every 10-30 seconds which allows rovers to compute a double-differenced NLS solution. An open protocol for general-purpose RTK is the **Radio Technical Commission for Maritime Services (RTCM)** standard. It should be noted that as RTK requires a data link to transmit the information to the rovers which it must first calculate, there will typically be a latency to the RTK communication. Thus, RTK receivers must send a range rate correction along with their data at the time of their observations for the rovers to apply the observations correctly. This radio data link also requires that a communication frequency is available that is not being used by other RTK users in the area.

Alternatively, **real-time networks (RTN)** use an internet connection to supply the corrections and typically received through the standard Networked Transport of RTCM via Internet Protocol (NTRIP) message using a cell phone modem. With this configuration, the corrections can be supplied by a network of stationary GNSS receivers, e.g. **continuous operating reference stations (CORS)**, which compute the GNSS distance-dependent biases, i.e. ephemeris, ionospheric, and tropospheric, over a wide area and compute corrections for users within the network area. One common method is the use of a **virtual reference station (VRS)** for which the rover receives the optimal corrections. Then, if a rover moves significantly away from the VRS, e.g. 10 km, a new VRS is provided to the rover. However, VRS requires feedback to the RTN to provide an approximate position. This is typically handled as a **National Marine Electronics Association (NMEA)** standard string format, of which there are seven. It should also be noted that CORS networks can also be used for PPK as well as they provide their raw GNSS observables for public use.

## GNSS-Based Velocity Estimation

GNSS signal processing of the carrier phase measurement provides the Doppler shift of the carrier frequency for velocity information. For GNSS, it is typical to estimate position and velocity with the Doppler shift and code/carrier phase measurements as these are already available to GNSS receivers. The first and simplest method for adding velocity estimation to the GNSS receiver-processor is to first estimate the GNSS receiver

position,  $\hat{\vec{x}}$ , and clock offset,  $\hat{f}\dot{t}$ . Then, one can estimate the receiver velocity and clock drift as the second parameter vector, i.e.,

$$\vec{\beta} = \begin{bmatrix} \dot{\vec{x}} \\ f\dot{t} \end{bmatrix} \quad (19.124)$$

Then, by rearranging the GNSS pseudorange rate equation

$$\Delta f_i + \frac{f}{c} \frac{(\vec{x} - \vec{x}_i)^T}{\sqrt{(\hat{x} - x_i)^2 + (\hat{y} - y_i)^2 + (\hat{z} - z_i)^2}} \dot{\vec{x}}_i = \frac{f}{c} \frac{(\vec{x} - \vec{x}_i)^T}{\sqrt{(\hat{x} - x_i)^2 + (\hat{y} - y_i)^2 + (\hat{z} - z_i)^2}} \dot{\vec{x}} + f\dot{t} + \epsilon \quad (19.125)$$

which allows one to form the linear observation equation with  $m \geq 4$  satellites

$$\vec{y} = \begin{bmatrix} \Delta f_1 + \frac{f}{c} \frac{(\hat{\vec{x}} - \vec{x}_1)^T}{\|\hat{\vec{x}} - \vec{x}_1\|} \dot{\vec{x}}_1 \\ \vdots \\ \Delta f_m + \frac{f}{c} \frac{(\hat{\vec{x}} - \vec{x}_m)^T}{\|\hat{\vec{x}} - \vec{x}_m\|} \dot{\vec{x}}_m \end{bmatrix} = X\vec{\beta} = \begin{bmatrix} \frac{f}{c} \frac{(\hat{\vec{x}} - \vec{x}_1)^T}{\|\hat{\vec{x}} - \vec{x}_1\|} & 1 \\ \vdots & \vdots \\ \frac{f}{c} \frac{(\hat{\vec{x}} - \vec{x}_m)^T}{\|\hat{\vec{x}} - \vec{x}_m\|} & 1 \end{bmatrix} \begin{bmatrix} \dot{\vec{x}} \\ f\dot{t} \end{bmatrix} \quad (19.126)$$

where the optimal estimate is given by the OLS solution

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y} \quad (19.127)$$

The second and more accurate method would be to simultaneously estimate the velocity and position as a single parameter vector, i.e.

$$\vec{\beta} = \begin{bmatrix} \vec{x} \\ \dot{\vec{x}} \\ c\dot{t} \\ f\dot{t} \end{bmatrix} \quad (19.128)$$

Then, the combined pseudorange and pseudorange rate equation allows one to form a nonlinear observation equation with  $m$  satellites as

$$\vec{y} = \begin{bmatrix} \rho_1 \\ \vdots \\ \rho_m \\ \Delta f_1 \\ \vdots \\ \Delta f_m \end{bmatrix} = \mathbf{f}(\vec{x}, \hat{\vec{\beta}}) = \begin{bmatrix} \|\vec{x} - \vec{x}_1\|_2 + c\delta t \\ \vdots \\ \|\vec{x} - \vec{x}_m\|_2 + c\delta t \\ \frac{(\vec{x} - \vec{x}_1)^T}{\|\vec{x} - \vec{x}_1\|_2} (\dot{\vec{x}} - \dot{\vec{x}}_1) + f\dot{t} \\ \vdots \\ \frac{f}{c} \frac{(\vec{x} - \vec{x}_m)^T}{\|\vec{x} - \vec{x}_m\|_2} (\dot{\vec{x}} - \dot{\vec{x}}_m) + f\dot{t} \end{bmatrix} \quad (19.129)$$

To solve this problem using NLS with GNA, assume one has an initial estimate,  $\hat{\vec{\beta}}_0$ . Then, one can iteratively improve  $\hat{\vec{\beta}}$  by  $\vec{\Delta}_k$  for  $k = 0, 1, \dots$  as

$$\hat{\vec{\beta}}_{k+1} = \hat{\vec{\beta}}_k + \vec{\Delta}_k \quad (19.130)$$

where

$$\vec{\Delta}_k = \left( \mathbf{J}^T \mathbf{J} \right)^{-1} \mathbf{J}^T \left( \vec{y} - \mathbf{f}(\vec{x}, \hat{\vec{\beta}}_k) \right) \quad (19.131)$$

where the  $i^{\text{th}}$  row of  $2m \times 7 \mathbf{J}$  can be written for  $i = 1, \dots, m$  as

$$\mathbf{J}_i = \begin{bmatrix} \frac{(\vec{x} - \vec{x}_i)^T}{\|\vec{x} - \vec{x}_i\|_2} \\ \vec{0}_{3 \times 1} \\ 1 \\ 0 \end{bmatrix}^T \quad (19.132)$$

$\vec{x} = \hat{\vec{x}}_k$

and for  $i = m + 1, \dots, 2m$  as

$$\mathbf{J}_i = \begin{bmatrix} \frac{f}{c} \left( \frac{(\vec{x} - \vec{x}_i)^T}{\|\vec{x} - \vec{x}_i\|_2} - (\vec{x} - \vec{x}_1)^T (\vec{x} - \vec{x}_i) \frac{(\vec{x} - \vec{x}_1)^T}{\|\vec{x} - \vec{x}_1\|_2^3} \right) \\ \frac{f}{c} \frac{(\vec{x} - \vec{x}_i)^T}{\|\vec{x} - \vec{x}_i\|_2} \\ 0 \\ 1 \end{bmatrix}^T \quad (19.133)$$

$\vec{x} = \hat{\vec{x}}_k, \vec{x} = \hat{\vec{x}}_k$

### User Equivalent Range Error Models

The satellite clock error models include a clock offset, a clock drift model, as well as general and special relativity, i.e.

$$\delta t_i = a_0 + a_1(t - t_0) + a_2(t - t_0)^2 + \Delta_{\text{rel}} \quad (19.134)$$

where  $a_0$  is the **satellite clock offset**,  $a_1$  is the **satellite clock drift**,  $a_2$  is the **satellite clock drift rate**, and  $\Delta_{\text{rel}}$  is the relativistic error term. This is typically modeled as

$$\Delta_{\text{rel}} = -2 \frac{\vec{r}_s \vec{v}_s}{c^2} = -2 \frac{\sqrt{\mu a}}{c^2} e \sin(E) \quad (19.135)$$

where  $\vec{r}_s$  is the radial position of the satellite,  $\vec{v}_s$  is the velocity of the satellite,  $c$  is the speed of light in a vacuum,  $\mu$  is the geocentric gravitational constant,  $a$  is the semimajor axis,  $e$  is the eccentricity (about 0.02), and  $E$  is the eccentric anomaly. This effect can cause a pseudorange error of about 14 m. In addition, there is an additional effect due to the motion of the receiver on the Earth's surface as the earth rotates during transmission time. This is known as the *Sagnac effect* and can result in around 133 nanoseconds of error at its maximum.

As part of GNSS, the control segments monitor all satellite systems and broadcasted signals, thus making it capable of monitoring and controlling the satellite clocks which it guides to be in line with the GNSS time. Similar to the receiver clocks, the satellite clock are subject to destabilizing effects, e.g. temperature, satellite acceleration, radiation. However, the control segment must keep the satellite clocks within strict standards of GNSS time and publishes the current satellite clock error models through the broadcast message onboard the satellites. Though the control segments can monitor the clocks closely, they do not constantly tweak the satellite clocks since doing so would cause them to deteriorate much more rapidly. As such, clock failure is the most common failure mode currently. Thus, as these models are not completely accurate, but the remainder of the error can be treated as part of the other noise/uncertainty in the pseudorange,  $\epsilon$ .

Other forces and moments than Newtonian gravity act on the satellites due to the non-uniformity of the Earth's gravity, the Sun's gravity, the Moon's gravity, and solar radiation pressure. However, each of these

error sources are difficult to directly model the orbital errors analytically. Thus, the best model of these disturbing forces and moments is by real-time estimation of the current satellite motion. These estimates are then broadcast by control segment to the satellites for use in the navigation message. The real-time parameters are updated every 1-4 hours, but are relatively accurate. However, due to this real-time modeling, one can also post-process the GNSS data to form more optimized orbits for the satellites for post-processed positioning. Three publicly available orbital error products are published by the International GNSS Service (IGS) and are named by GPS Week number and Day of Week. These are the **ultra-rapid orbital errors** (iguWWWD.sp3) which has a 6-hour latency release and uses a constrained optimization technique that does not allow any net rotation or translation, the **rapid** (igrWWWD.sp3) which has a 13-hour latency release and uses a constrained optimization technique that also does not allow any net rotation or translation, and the **final** (igsWWWD.sp3) which has a 12 to 14 day latency release and uses a minimally constrained optimization that only does not allow any net rotation.

The electromagnetic properties of the GNSS signal changes as it passes through the Earth's **atmosphere** which can be modeled as four regimes. Namely, the ionosphere which spans roughly 50-1000 km, the mesosphere which spans roughly 40-60 km, stratosphere which spans roughly 15-40 km, troposphere which spans roughly 0-15 km. Through refraction and diffraction, the atmosphere alters the apparent speed and direction of the GNSS signal which causes an apparent delay in the signal transit. This atmospheric delay is primarily caused by the physics of the ionosphere and troposphere whose error contributions to the GNSS range equations can be modeled as follows.

When gas molecules are ionized by the Sun's ultraviolet radiation, free electrons are released into the atmosphere. These free electrons gather into the ionosphere, but their number and dispersion varies across time and geography and affect the electron density in ionosphere. This phenomenon can be quantified by the **total electron count (TEC)** whose TEC units (TECU) are multiples of  $10^{16}$  free electrons in a vertical ionospheric column with a cross-sectional area of  $1 \text{ m}^2$ , i.e.  $1\text{TECU} = 10^{16} e^- / \text{m}^2$ . The TEC depends on the time of day primarily through the day/night cycle, 11-year solar cycle, magnetic activity of the Earth, and geographic location. The highest TEC values and variations generally occur in a band of  $60^\circ$  of latitude around magnetic equator, but can also be harsh in the polar regions.

As the free electrons in the ionosphere dispersively slows the signal, there are frequency and path of transmission dependencies for the ionospheric delay. With these considerations, it has been shown that the first order ionospheric delay is proportional to the **slant total electron count (STEC)**, i.e. the total count along the length of transmission distance,  $l$ , i.e.

$$STEC = \int N_e dl \quad (19.136)$$

where  $N_e$  is the electron density. This effects can be described by the heuristic equation

$$b_I = \frac{40.3 \times 10^{16}}{f^2} STEC \quad (19.137)$$

where  $f$  is the frequency in Hz, and the  $STEC$  is in TECUs. However, as some ionospheric delay may remain, one often uses a **mask angle** for elevations of  $15^\circ$ , i.e. satellites with elevations below  $15^\circ$  are not used in an un-differenced positioning solution.

Alternatively, for GNSS broadcast at multiple frequencies, e.g. GPS, one can use a multi-frequency receiver to form an **ionosphere-free combination** using two frequencies for the ionosphere-free pseudorange,

$\rho_{I-F}$ , as

$$\rho_{I-F} = \frac{f_1^2 \rho_1 - f_2^2 \rho_2}{f_1^2 - f_2^2} \quad (19.138)$$

or the ionosphere-free cycle count,  $\Phi_{I-F}$ , as

$$\Phi_{I-F} = \frac{f_1^2 \Phi_1 - f_2^2 \Phi_2}{f_1^2 - f_2^2} \quad (19.139)$$

where  $\rho_1$ ,  $\rho_2$ ,  $\Phi_1$ , and  $\Phi_2$  are GNSS pseudoranges and cycle counts on frequencies,  $f_1$  and  $f_2$ , respectively.

The tropospheric delay consists of two components which cause refraction of the GNSS signal: the **hydrostatic delay**, also known as the **dry delay**, and the **wet delay**. The hydrostatic delay causes 80-90% of the total tropospheric delay and is closely correlated to the atmospheric pressure, thus easily estimated. The wet delay is due to the highly variable water vapor distribution which is only roughly correlated to temperature and humidity. Thus, some high-precision GNSS applications may use water vapor radiometers and radiosondes to better model this effect. Modeling the troposphere delay can be up to 95% effective, but unlike the ionospheric effect it is nondispersive, i.e. it affects all frequencies the same, and cannot be removed using a dual-frequency receiver. However, it is more consistent than the ionosphere delay. It is also known that the index of refraction decreases with elevation angle  $\epsilon$ . At MSL a 90° elevation only produces about 2.5 m of error in the pseudorange, while at 75° elevation, this would be roughly 9 m of error, and at 10° elevation this would be 20 m of error. Thus, using a mask angle can also be useful for removing a significant portion of the tropospheric delay.

Therefore, one typically uses a model of the form:

$$b_T = m_h(\epsilon)ZHD + m_d(\epsilon)ZWD \quad (19.140)$$

where  $m_h(\epsilon)$  and  $m_d(\epsilon)$  are the Niell mapping functions based on  $\epsilon$ ,  $ZHD$  is the zenith hydrostatic delay, and  $ZWD$  is the zenith wet delay.  $ZHD$  and  $ZWD$  related to the air total pressure,  $p$  (hPa), the temperature,  $T$  (K), and the water vapor pressure,  $e$  (hPa). Two common models for these are

$$ZHD = \frac{0.0022767p}{1 - 0.00266 \cos 2\phi - 0.00028h_0} \quad (19.141)$$

where  $\phi$  is the ellipsoidal latitude,  $h_0$  is the surface height above ellipsoid (m),  $p$  is the total surface pressure (hPa), and

$$ZWD = 10^{-3} \int_{h_0}^{\infty} \left( 22.1 \frac{e}{T} + 3.754 \frac{e}{T^2} \right) \quad (19.142)$$

Multi-path occurs when part of the GNSS signal broadcast from the satellite reaches the receiver antenna after one or more reflections which can be caused by objects such as the ground, buildings, cliffs, and trees. These reflections are still sensed by the antenna and interfere with direct signal by skewing the correlation peak during the replica code matching done for acquiring the PRN code and obtaining a time difference measurement. This skewing can be up to 1.5 code chip lengths. However, once a replica PRN code is correlated with the incoming signal, other multi-path signals outside the expected chip length are easily rejected by modern receivers. However, smaller multi-path delays less than the expected chip length are harder to filter out and can result in typically 1-3 meters of additional error to the pseudorange measurement.

One method for further reducing multi-path is to use the fact that for an odd number of reflections, the multi-path GNSS signals become **left hand circular polarized** which allow advanced receivers to identify and remove. Multi-path effects can also practically be lessened by using a mask angle for elevations of 15°, elevating the antenna, and using a plate to block signals from below.

## References

For more information, please refer to the following

- S. Gleason and D. Gebre-Egziabher, “GNSS Applications and Methods,” Artech House, 2012

## 19.5 Map-Based Positioning Systems

### Altimeters

**Altimeters** are sensors provide a pseudo-altitude measurement through different sensing phenomenology. As altitude is one of the three components of the geodetic position coordinates along with latitude and longitude, altimeters are almost always used in conjunction with other positioning systems to improve the estimation of the altitude component via multi-sensor fusion algorithms. There are two general types of altimeter sensing phenomenology, passive signal reception and active signal transmission and reception. For aerospace vehicles, the primary passive altimeter is the barometric altimeter and the two types of active altimeters use radar or lidar.

**Barometric altimeters**, also known as **pressure altimeters** measure the ambient pressure and use a pressure map to invert this to a pseudo-altitude, typically using the International Standard Atmosphere (ISA) model which linearly relates absolute temperature of the air,  $T$ , with altitude,  $h$ , at different atmospheric layers. For the pressure,  $P$ , the ISA and barometric altimeters generally use the hydrostatic balance and ideal gas law to compute the pseudo-altitude as

$$h = \frac{RT}{g} \log \left( \frac{P_0}{P_h} \right) \quad (19.143)$$

where  $R$  is the specific gas constant for dry air ( $287.058 \text{ J kg}^{-1} \text{ K}^{-1}$ ),  $T$  is the absolute temperature of the air,  $g$  is the acceleration due to gravity,  $P_h$  is the pressure at altitude  $h$ , and  $P_0$  is the reference pressure at MSL.

Differences in temperature from the ISA model will also cause errors in the pseudo-altitude as well as small errors due to assumptions about the acceleration due to gravity in the altimeter due to its variance with latitude and altitude. Barometric altimeters can exhibit errors of hundreds of feet due to sudden changes in air pressure, e.g., weather fronts, without any actual change in altitude. Modern aircraft use a barometric altimeter known as a **sensitive altimeter** which allows one to adjust the reference pressure at MSL which is supplied by local airports since MSL reference atmospheric pressure at a given location varies over time with temperature and the movement of pressure systems in the atmosphere.

As opposed to barometric altimeters, **Radar altimeters** and **lidar altimeters** provide a relative pseudo-altitude, i.e., the **height above ground level (HAGL)** for narrow-beam or point measurements or **height above average terrain (HAAT)** for wide-beam measurements, through pseudorange measurements beneath the aircraft corrected by the pitch and roll estimates of the aircraft. Radar and lidar altimeters were developed as a more reliable and accurate sensor for aircraft than barometric altimeters which are almost unusable in

heavy fog or rain which hinder the ability of aircraft to accurately fly close to the terrain, e.g., landing. Thus, in modern aircraft, radar altimeters are standard sensors for automated landing systems and terrain-avoidance systems. In satellite systems, radar altimeters are also used in remote sensing of the Earth's surface.

Pulsed radar and lidar altimeters measure the time-of-flight (TOF) for a radio or light signal to travel to and reflect back from the ground which allows them to provide, i.e.,

$$h = c \text{TOF} = c (TOA_i - TOT_i) \quad (19.144)$$

which results in a pseudo-altimeter measurement as

$$h = \frac{c \text{TOF}}{2} \quad (19.145)$$

FMCW radar and lidar provide the pseudo-altitude as

$$h = \frac{c \Delta f_h}{2 \frac{df}{dt}} = \frac{c \frac{\Delta f_{re} + \Delta f_{fe}}{2}}{2 \frac{df}{dt}} \quad (19.146)$$

where  $df/dt$  is the frequency shift of the FMCW signal per unit time,  $f_h$  is the Doppler-adjusted frequency difference of the received signal to the transmitted signal which is formed by two different frequency differences,  $\Delta f_{re}$ , the frequency difference at the rising edge, and  $\Delta f_{fe}$ , the frequency difference at the falling edge. Notably, FMCW radar and lidar also provide the Doppler frequency as

$$f_d = \frac{\Delta f_{re} - \Delta f_{fe}}{2} \quad (19.147)$$

Frequency modulated continuous-wave (FMCW) radar and lidar are more commonly used than pulsed radar and lidar since the frequency shift can be more accurately tracked through Doppler tracking than the direct time difference.

### Terrain Map-Based Positioning

For an absolute altitude estimate, altimeters require a **terrain map** in which the positioning system must localize relative to the terrain. Often, this is in the form of a **digital elevation map (DEM)** which discretizes the continuously-valued terrain elevation into digital information of the elevation.

### Magnetic Map-Based Positioning

**Magnetic-based navigation (MBN)**, also known as **magnetic navigation**, uses magnetometer to sense the Earth's crustal **magnetic anomalies**, which are unique “fingerprints” of the Earth's crust, that can be collected to form a **magnetic map** for which one can perform positioning, e.g., the World Magnetic Model (WMM). Because this

### Visual Feature Map-Based Positioning

**Vision-based navigation (VBN)**, also known as **visual-based navigation** or **image-based navigation (IBN)**, uses an imaging sensor, i.e., cameras, lidar, and/or radar, to form the pseudo-measurements for the navigation

system from images of the environment. These environmental images can be either two-dimensional, e.g., from a monocular camera, or three-dimensional, e.g., from a stereo camera, red-green-blue-depth (RGB-D) camera, lidar, or synthetic aperture radar (SAR). These images are processed by the navigation system through one of two general methods. The first method is to process the full image intensities into odometry measurements. The second method is to identify, extract, and potentially track points of interest from the images known as **feature points**, also known as **keypoints** or **landmarks**, which may then be used for odometry measurement or for feature-relative pseudorange/pseudobearing measurements, i.e., localization measurements.

Feature-based VBN can be classified into two broad classes where the features in the environment are either **mapped**, i.e., the feature points' positions are known, or unmapped, i.e., the feature points' positions are unknown. For VBN with unmapped features, one may perform **simultaneous localization and mapping (SLAM)** which jointly estimates the navigation state of the vehicle and the feature points' positions throughout the process or perform navigation state-only estimation which utilizes only images as visual odometry measurements. Importantly, a fundamental challenge in feature-based VBN is the identification and association of extracted feature points to those of the feature map which is either known or being estimated. This challenge is also fundamental to the more general object tracking systems which are discussed in the subsequent chapter.

## References

For more information, please refer to the following

- S. Gleason and D. Gebre-Egziabher, “GNSS Applications and Methods,” Artech House, 2012

---

# Attitude Determination Systems

## 20.1 Introduction to Attitude Determination Systems

This section introduces some simple methods to determine attitude, i.e., roll, pitch, and yaw, from single sensor measurements either through direct measurements or through integration of rate measurements.

### Direct Roll and Pitch Determination

For roll  $\phi$  and pitch  $\theta$  Euler angle determination of the LVLH-to-body-fixed frame rotation, one uses infrared, optical, or inertial sensors for aerospace vehicles to obtain a **nadir** bearing measurement,  $\vec{b}_{LorN}$ , defined as straight down in the LVLH or NED navigation frame as

$$\vec{b}_L = \begin{bmatrix} 0 \\ 0 \\ \|\vec{b}_B\|_2 \end{bmatrix} \quad (20.1)$$

which can be related to the DCM by

$$\vec{b}_B = \begin{bmatrix} b_{B,1} \\ b_{B,2} \\ b_{B,3} \end{bmatrix} = C_{B \leftarrow L} \vec{b}_L \quad (20.2)$$

Thus, one has

$$\begin{bmatrix} b_{B,1} \\ b_{B,2} \\ b_{B,3} \end{bmatrix} = \|\vec{b}_B\|_2 \begin{bmatrix} -\sin \theta \\ \sin \phi \cos \theta \\ \cos \phi \cos \theta \end{bmatrix} \quad (20.3)$$

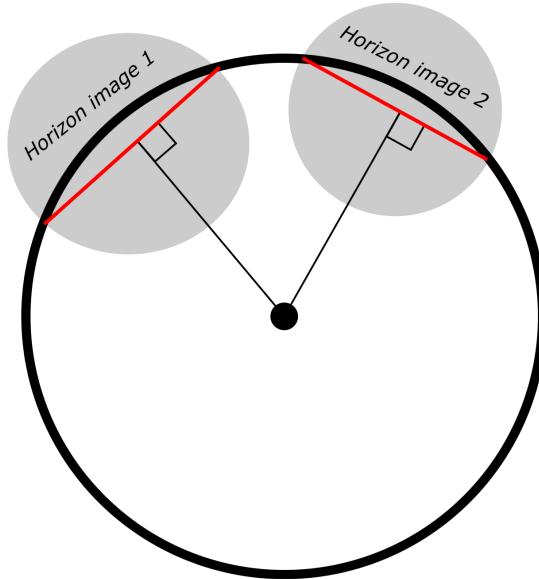
which implies one can estimate the pitch as

$$\hat{\theta} = \sin^{-1} \left( -\frac{b_{B,1}}{\|\vec{b}_B\|_2} \right) \quad (20.4)$$

and the roll as

$$\hat{\phi} = \tan^{-1} \left( \frac{b_{B,2}}{b_{B,3}} \right) \quad (20.5)$$

For spacecraft, infrared (IR) sensors can be used to sense the direction to the warmest source in the area of operation for satellites. If the source is the Earth's atmosphere, one has an **Earth sensor**. From the IR image, the centroid of the **Earth disk**, i.e. the 2D projection of the Earth on the image, must be rotated from the camera image frame to the body-fixed frame to construct the nadir vector measurement to the center of the Earth, i.e. an Earth sensor provides  $\vec{b}_B$  directly. Notably, at lower altitude orbits where the Earth disk is not completely contained within the sensor's field-of-view (FOV), the infrared imaging sensor is often referred to as an **Earth horizon sensor** from which the centroid must be estimated using the curvature of the horizon line, e.g., the intersection of the perpendiculars from two chords as shown.



For aircraft, an **accelerometer** is a sensor that directly measures the acceleration of a body in its own instantaneous rest frame. Conceptually, an accelerometer is a **proof mass**, also known as a **seismic mass**, on a spring. For planetary-based accelerometers, the rest frame is the Earth so the gravity force is not included. When the accelerometer experiences an acceleration, the proof mass is deflected to the point that the spring accelerates the mass at the same speed as the casing. The deflection of the proof mass from its neutral position or the compression of the spring provides an acceleration measurement. However, the mass-spring system is also damped so that additional oscillations of the system do not affect the measurement output. Thus, all accelerometers have a frequency response and are designed to provide the required sensitivity and maximum expected acceleration. There are many different designs for converting this mass deflection or spring compression into a measurable electric signal including piezoelectric, piezoresistive, capacitive, and thermal.

A three-axis accelerometer nominally measures the specific force experienced by the vehicle in the body-fixed frame, i.e.,

$$\vec{f}_B = \vec{a}_B - \vec{g} \quad (20.6)$$

where  $\vec{a}_B$  is the total acceleration and  $\vec{g}$  is the gravitational acceleration where nominally, the gravity acceleration vector points in the nadir direction, i.e., straight down

$$\vec{g}_L = \begin{bmatrix} 0 \\ 0 \\ \|g_L\|_2 \end{bmatrix} \quad (20.7)$$

Thus, for an accelerometer, one has

$$\vec{b}_B = \vec{a}_B - \vec{f}_B \quad (20.8)$$

However, this requires one is able to independently calculate the acceleration of the vehicle or allow it as an unknown and time-varying “bias” error in the measurement. Thus, the quality of these estimates depends on the acceleration experienced by the vehicle which can be quite large. Thus, accelerometer measurements for pitch and roll determination are typically low-pass filtered to remove any high frequency accelerations.

### Direct Yaw Determination

For a direct yaw solution for aerospace vehicles operating around Earth, one can use a magnetometer. To compute the yaw angle,  $\psi$ , one can use a TAM to nominally measure the local magnetic field vector as

$$\vec{h}_B = \begin{bmatrix} h_{B,1} \\ h_{B,2} \\ h_{B,3} \end{bmatrix} \quad (20.9)$$

Furthermore, if the model-based approach can ut

If the which can be projected into the horizontal plane through **tilt compensation (TC)** as

$$\vec{h}_{TC} = \begin{bmatrix} h_{TC,1} \\ h_{TC,2} \\ h_{TC,3} \end{bmatrix} = \begin{bmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \phi & -\sin \phi \\ 0 & \sin \phi & \cos \phi \end{bmatrix} \vec{h}_B \quad (20.10)$$

$$\vec{h}_{TC} = \begin{bmatrix} h_{TC,1} \\ h_{TC,2} \\ h_{TC,3} \end{bmatrix} = \begin{bmatrix} h_{B,1} \cos \theta + h_{B,2} \sin \phi \sin \theta - h_{B,3} \cos \phi \sin \theta \\ h_{B,2} \cos \phi - h_{B,3} \sin \phi \\ -h_{B,1} \sin \theta + h_{B,2} \sin \phi \cos \theta + h_{B,3} \cos \phi \cos \theta \end{bmatrix} \vec{h}_B \quad (20.11)$$

which requires the current roll and pitch estimates to compute the projection. Then, the angle between the leveled measurements and magnetic north can be used to estimate the yaw angle, i.e.,

$$\hat{\psi} = \tan^{-1} \left( \frac{-h_{TC,2}}{h_{TC,1}} \right) + \eta \quad (20.12)$$

$$\hat{\psi} = \tan^{-1} \left( \frac{h_{B,3} \sin \theta - h_{B,2} \cos \theta}{h_{B,1} \cos \theta + h_{B,2} \sin \phi \sin \theta + h_{B,3} \cos \phi \sin \theta} \right) + \eta \quad (20.13)$$

where  $\eta$  is the magnetic declination between true north and magnetic north which must be obtained from a magnetic field map based on estimated position. Notably, if this is unknown, a magnetometer will provide an estimate of the “magnetic” yaw angle.

## Attitude Determination from Angular Velocity

The Euler angles can be related to the angular velocity of the body-to-navigation frame  $\vec{\omega}_{B,N \leftarrow B}$  which can be described as the difference between the inertial frames, i.e.

$$\vec{\omega}_{B,N \leftarrow B} = \vec{\omega}_{B,I \leftarrow B} - \vec{\omega}_{B,I \leftarrow N} \quad (20.14)$$

where  $\vec{\omega}_{B,I \leftarrow B}$  can be nominally measured using a three-axis rate gyroscope as

$$\vec{\omega}_{B,I \leftarrow B} = \begin{bmatrix} p_g \\ q_g \\ r_g \end{bmatrix} \quad (20.15)$$

and  $\vec{\omega}_{B,I \leftarrow N}$  combines the transport rate and Earth rotation rate terms of the angular velocity and can be computed as a function of the current latitude  $\ell$  and longitude  $\lambda$  and their rates, i.e.,

$$\vec{\omega}_{B,I \leftarrow N} = C_{B \rightarrow N} \begin{bmatrix} (\dot{\ell} + \omega_{I \leftarrow E}) \cos \lambda \\ -\dot{\lambda} \\ -(\dot{\ell} + \omega_{I \leftarrow E}) \sin \lambda \end{bmatrix} \quad (20.16)$$

where  $\omega_{I \leftarrow E}$  is the Earth rotation rate, often taken as the WGS84 value  $7.292115 \times 10^{-5}$  rad/s. For automotive/consumer grade IMUs, one can typically neglect  $\vec{\omega}_{B,I \leftarrow N}$  due to the much higher bias errors.

By defining the Euler angle vector as

$$\vec{\psi}_{N \leftarrow B} = \begin{bmatrix} \phi \\ \theta \\ \psi \end{bmatrix} \quad (20.17)$$

Then, the Euler angle time derivatives can be related to the gyroscope measurements as

$$\dot{\vec{\psi}}_{N \leftarrow B} = \begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \frac{1}{\cos \theta} \begin{bmatrix} \cos \theta & \sin \theta \sin \phi & \sin \theta \cos \phi \\ 0 & \cos \theta \cos \phi & -\sin \phi \cos \theta \\ 0 & \sin \theta & \cos \phi \end{bmatrix} \begin{bmatrix} p_g \\ q_g \\ r_g \end{bmatrix} \quad (20.18)$$

which must be numerically integrated to determine the Euler angles. With proper initialization, this setup can be within several degrees of accuracy for tens of seconds without aiding even for low grade gyroscopes. However, the integration of any error in the angular velocity measurements will cause unbounded growth of the Euler angle estimate error which requires “aiding” from additional attitude sensors to bound, e.g., aiding with an accelerometer and magnetometer produces an attitude and heading reference system (AHRS) which is discussed in the next section.

## References

For more information, please refer to the following

- H. Mokhtarzadeh. “Appendix D Attitude Heading Reference System,” in *Correlated-Data Fusion and Cooperative Aiding in GNSS-Stressed or Denied Environments*, Ph.D. dissertation, Department of Aerospace Engineering & Mechanics, University of Minnesota, 2014, pp. 151-172

## 20.2 Attitude and Heading Reference Systems

The most common type of attitude determination system which employs multi-sensor information fusion uses magnetometers, gyroscopes, and accelerometers to estimate the attitude and heading referenced to the Earth's surface, i.e. the local navigation frame. This type of system is known as an **attitude and heading reference system (AHRS)**, also known as a **magnetic, angular rate, and gravity (MARG) system**. Furthermore, one can form an **Air Data AHRS (ADAHRS)** which uses the additional air data measurements to improve the models in an AHRS, e.g., the airspeed, angle of attack, and angle of sideslip for additional error modeling. With proper error modeling and tuning, one can derive an (AD)AHRS using low grade IMUs, magnetometers, and optional air data sensors.

AHRS typically use **complementary filtering**, **error-state filtering**, or **equivariant filtering** as the multi-sensor data fusion algorithm. The complementary filter can be considered a state-to-state fusion algorithm with constant "Kalman" gains while the error-state and equivariant filters are sensor-to-sensor fusion algorithms which are chosen due to the 3-DOF attitude dynamics inherent in AHRS. This section presents a complementary filter- and an error-state-based AHRS with the Euler angle representation. In this case, it should be pointed out that care must be taken when using Euler angles due to the ambiguities due to gimbal lock and angular wrapping, i.e.,  $\theta = \theta + n2\pi$  for  $n = 1, 2, 3, \dots$ . Notably, if the local magnetic declination is unknown, an AHRS will provide an estimate of the **magnetic heading**, i.e., the yaw angle relative to magnetic north, not true north.

### Complementary Filter-Based AHRS

Recall that one can perform direct attitude determination of the Euler angles by integrating the gyroscope angular velocity measurement  $\hat{\omega}_{B,B \leftarrow I}$ , e.g., a forwards Euler integration provides

$$\hat{\psi}_{B \leftarrow N,g} = \begin{bmatrix} \hat{\phi}_g \\ \hat{\theta}_g \\ \hat{\psi}_g \end{bmatrix} = \hat{\psi}_{B \leftarrow N,g} + \Delta t \begin{bmatrix} 1 & \sin \phi \tan \theta & \cos \phi \tan \theta \\ 0 & \cos \phi & -\sin \phi \\ 0 & \sin \phi \sec \theta & \cos \phi \sec \theta \end{bmatrix} \left( \hat{\omega}_{B,B \leftarrow I} - \vec{\omega}_{B,N \leftarrow I} \right) \quad (20.19)$$

where  $\vec{\omega}_{B,N \leftarrow I}$  combines the transport rate and Earth rotation rate terms of the angular velocity which can be compensated using the current latitude and altitude if known. Notably, an additional design aspect for most complementary filters involves high-pass filtering of the gyroscope Euler angle estimates to bound the low-frequency bias drift which is not modeled nor estimated within the complementary filter due to the loosely-coupled fusion, e.g., a first-order Butterworth low-pass filter would produce

$$\begin{bmatrix} \hat{\phi}_{g,h-p} \\ \hat{\theta}_{g,h-p} \\ \hat{\psi}_{g,h-p} \end{bmatrix} = \begin{bmatrix} \frac{\omega_g s}{s+\omega_g} \hat{\phi}_g \\ \frac{\omega_g s}{s+\omega_g} \hat{\theta}_g \\ \frac{\omega_g s}{s+\omega_g} \hat{\psi}_g \end{bmatrix} \quad (20.20)$$

Furthermore, recall one can perform direct attitude determination of the Euler angles using a combination of an accelerometer and a magnetometer as

$$\hat{\psi}_{B \leftarrow N,a+m} = \begin{bmatrix} \hat{\phi}_a \\ \hat{\theta}_a \\ \hat{\psi}_m \end{bmatrix} = \begin{bmatrix} \tan^{-1} \left( \frac{b_{B,2}}{b_{B,3}} \right) \\ \sin^{-1} \left( -\frac{b_{B,1}}{\| \vec{b}_B \|_2} \right) \\ \tan^{-1} \left( \frac{h_{B,3} \sin \hat{\theta}_a - h_{B,2} \cos \hat{\theta}_a}{h_{B,1} \cos \hat{\theta}_a + h_{B,2} \sin \hat{\phi}_a \sin \hat{\theta}_a + h_{B,3} \cos \hat{\phi}_a \sin \hat{\theta}_a} \right) + \eta \end{bmatrix} \quad (20.21)$$

which notably couples the magnetometer with the accelerometer measurements for the tilt compensation. Notably, an additional design aspect for most complementary filters involves low-pass filtering of the accelerometer Euler angle estimates to increase the SNR and remove any high-frequency non-gravity accelerations and low-pass filtering of the magnetometer Euler angle estimates to increase the SNR and remove any high-frequency magnetic disturbances, e.g., a first-order Butterworth low-pass filter would produce

$$\begin{bmatrix} \hat{\phi}_{a,l-p} \\ \hat{\theta}_{a,l-p} \\ \hat{\psi}_{m,l-p} \end{bmatrix} = \begin{bmatrix} \frac{\omega_a}{s+\omega_a} \hat{\phi}_a \\ \frac{\omega_a}{s+\omega_a} \hat{\theta}_a \\ \frac{\omega_m}{s+\omega_m} \hat{\psi}_m \end{bmatrix} \quad (20.22)$$

For the complementary filter-based AHRS, these two measurement vectors are fused using a weighted average for each component, i.e.,

$$\begin{bmatrix} \hat{\phi}_f \\ \hat{\theta}_f \\ \hat{\psi}_f \end{bmatrix} = \begin{bmatrix} (1 - K_\phi) \hat{\phi}_g \\ (1 - K_\theta) \hat{\theta}_g \\ (1 - K_\psi) \hat{\psi}_g \end{bmatrix} + \begin{bmatrix} K_\phi \hat{\phi}_a \\ K_\theta \hat{\theta}_a \\ K_\psi \hat{\psi}_m \end{bmatrix} \quad (20.23)$$

or

$$\begin{bmatrix} \hat{\phi}_f \\ \hat{\theta}_f \\ \hat{\psi}_f \end{bmatrix} = \begin{bmatrix} \hat{\phi}_g \\ \hat{\theta}_g \\ \hat{\psi}_g \end{bmatrix} + \begin{bmatrix} K_\phi (\hat{\phi}_a - \hat{\phi}_g) \\ K_\theta (\hat{\theta}_a - \hat{\theta}_g) \\ K_\psi (\hat{\psi}_m - \hat{\psi}_g) \end{bmatrix} \quad (20.24)$$

where  $K_\phi \in [0, 1]$ ,  $K_\theta \in [0, 1]$ ,  $K_\psi \in [0, 1]$  are constant gains chosen via testing or it can be approximated via the univariate steady-state Kalman filter equations if the variances of the gyroscope, accelerometer, and magnetometer estimates are known and relatively constant. Notably, to simultaneously low-pass filter the accelerometer and magnetometer measurements with  $G(s)$  and high-pass filter the gyroscope measurements with first-order Butterworth filters with  $\omega = \omega_g = \omega_a = \omega_m$ , one has

$$\begin{bmatrix} \hat{\phi}_f \\ \hat{\theta}_f \\ \hat{\psi}_f \end{bmatrix} = \begin{bmatrix} \hat{\phi}_g \\ \hat{\theta}_g \\ \hat{\psi}_g \end{bmatrix} + \frac{\omega}{s + \omega} \begin{bmatrix} K_\phi (\hat{\phi}_a - \hat{\phi}_g) \\ K_\theta (\hat{\theta}_a - \hat{\theta}_g) \\ K_\psi (\hat{\psi}_m - \hat{\psi}_g) \end{bmatrix} \quad (20.25)$$

which can also be shown to be equivalent to the steady-state Kalman filter equations by incorporating the filter into the prediction step.

## ES-EKF-Based AHRS

An ES-EKF-based AHRS uses a prediction step for the gyroscope nominal-state and the gyroscope error-state covariance, a correction step for the error-state estimate and covariance using the accelerometer and magnetometer measurements, and a nominal-state update step for correcting the nominal-state estimate with the error-state estimate and resetting the error-state. In navigation, the *de facto* standard algorithm for data fusion is the error-state EKF although the UKF and the PF have been used in some cases. This section will describe the error-state Extended Kalman filter (ES-EKF) implementation in an AHRS using the Euler angle for attitude and a composition based on the linearized DCM error.

For the AHRS ES-EKF presented here, the estimated nominal-state contains the three nominal Euler angles  $\bar{\psi}_{B \leftarrow N} = [\bar{\phi} \ \bar{\theta} \ \bar{\psi}]^T$  and the three nominal gyroscope biases  $\bar{b}_g$ , i.e.,

$$\bar{x} = \begin{bmatrix} \bar{\psi}_{B \leftarrow N} \\ \bar{b}_g \end{bmatrix} \quad (20.26)$$

The error-state vector contains the three Euler angle errors  $\delta\vec{\psi}_{B \leftarrow N} = [\delta\phi \ \delta\theta \ \delta\psi]^T$  and three gyroscope bias errors  $\delta\vec{b}_g$ , i.e.,

$$\delta\vec{x} = \begin{bmatrix} \delta\vec{\psi}_{B \leftarrow N} \\ \delta\vec{b}_g \end{bmatrix} \quad (20.27)$$

where  $\bar{N}$  is the NED navigation frame due to the nominal Euler angle rotation. Notably, the true gyroscope biases composition function is simply given as additive gyroscope biases to the nominal gyroscope biases, i.e.,

$$\vec{b}_g = \bar{b}_g \oplus \delta\vec{b}_g = \bar{b}_g + \delta\vec{b}_g \quad (20.28)$$

For the composition function of the Euler angles,  $\vec{\psi}_{B \leftarrow N}$ , consider the composition of the true DCM,  $C_{B \leftarrow N}$ , the nominal DCM,  $C_{B \leftarrow \bar{N}}$ , and the error DCM,  $C_{\bar{N} \leftarrow N}$ , as functions of the true, nominal, and error 3–2–1 Euler angles given by

$$C_{B \leftarrow N}(\phi, \theta, \psi) = C_{B \leftarrow \bar{N}}(\bar{\phi}, \bar{\theta}, \bar{\psi})C_{\bar{N} \leftarrow N}(\delta\phi, \delta\theta, \delta\psi) \quad (20.29)$$

where the Euler angle errors correspond to some small axis offset between the estimated navigation frame and the true navigation frame. For these small Euler angle errors, one has the linear approximation

$$C_{\bar{N} \leftarrow N}(\delta\phi, \delta\theta, \delta\psi) \approx I_{3 \times 3} - [\delta\vec{\psi}_{B \leftarrow N}] \times \quad (20.30)$$

Thus, the true Euler angle composition function  $\vec{\psi}_{B \leftarrow N} = \bar{\psi}_{B \leftarrow N} \oplus \delta\vec{\psi}_{B \leftarrow N}$  can be obtained via substitution and extraction from the DCM relationships of the components of the following equation

$$C_{B \leftarrow N}(\vec{\psi}_{B \leftarrow N}) = C_{B \leftarrow \bar{N}}(\bar{\psi}_{B \leftarrow N}) \left( I_{3 \times 3} - [\delta\vec{\psi}_{B \leftarrow N}] \times \right) \quad (20.31)$$

or, alternatively, its transpose, i.e.,

$$C_{N \leftarrow B}(\vec{\psi}_{B \leftarrow N}) = \left( I_{3 \times 3} + [\delta\vec{\psi}_{B \leftarrow N}] \times \right) C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) \quad (20.32)$$

which will be used here. Notably, the DCMs in the second equation are not a 3–2–1 sequential rotation of the Euler angles, but a 1–2–3 sequential rotation of the negative Euler angles.

For the AHRS ES-EKF prediction step at time step  $k$ , the nominal-state estimate is modeled in continuous-time as

$$\hat{\vec{x}}_k = \begin{bmatrix} \hat{\vec{\psi}}_{B \leftarrow N} \\ \hat{\vec{b}}_g \end{bmatrix} \quad (20.33)$$

where the Euler angles have continuous-time dynamics

$$\dot{\hat{\vec{\psi}}}_{B \leftarrow N} = \begin{bmatrix} 1 & \sin \phi \tan \theta & \cos \phi \tan \theta \\ 0 & \cos \phi & -\sin \phi \\ 0 & \sin \phi \sec \theta & \cos \phi \sec \theta \end{bmatrix} \vec{\omega}_{B, B \leftarrow N} \quad (20.34)$$

where  $\vec{\omega}_{B, B \leftarrow N}$  is the angular velocity. This equation is a nonlinear function of the nominal-state and must be numerically integrated to obtain the Euler angles at time step  $k$ ,  $\hat{\vec{\psi}}_{B \leftarrow I, k}$ , based on the updated discrete-time gyroscope measurements of the inertial angular velocity,  $\hat{\vec{\omega}}_{B, B \leftarrow I, k}$ , which are corrupted by the gyroscope bias, i.e.,

$$\vec{\omega}_{B, B \leftarrow N, k} = \hat{\vec{\omega}}_{B, B \leftarrow I, k} - \vec{\omega}_{B, N \leftarrow I, k} - \hat{\vec{b}}_{g, k} \quad (20.35)$$

For low-grade IMUs, a forwards Euler integration from the previous time step  $k - 1$  to the current time step  $k$  is typically used for the update, i.e., for the Euler angles

$$\hat{\psi}_{B \leftarrow N, k|k-1} = \hat{\psi}_{B \leftarrow N, k-1|k-1} + \Delta t \begin{bmatrix} 1 & \sin \phi \tan \theta & \cos \phi \tan \theta \\ 0 & \cos \phi & -\sin \phi \\ 0 & \sin \phi \sec \theta & \cos \phi \sec \theta \end{bmatrix} \left( \hat{\omega}_{B, B \leftarrow I, k-1} - \vec{\omega}_{B, N \leftarrow I, k-1} - \hat{b}_{g, k-1} \right) \quad (20.36)$$

where  $\vec{\omega}_{B, B \leftarrow I}$  is dependent on the latitude and longitude of the vehicle which may not be known. For higher-grade IMUs, the RK(4) method is typically used to obtain intermediate points for the attitude state and is typically the DCM or rotation quaternion. The gyroscope biases are typically modeled as an Ornstein-Ohlenbeck or a Wiener process. A first-order Gauss-Markov process model provides

$$\dot{\bar{b}}_g = -\frac{1}{\tau_g} I_{3 \times 3} \bar{b}_g \quad (20.37)$$

which provides the nominal-state update as

$$\hat{b}_{g, k|k-1} = \exp \left( -\frac{\Delta t}{\tau_g} \right) I_{3 \times 3} \hat{b}_{g, k-1|k-1} \quad (20.38)$$

A Wiener process model provides

$$\dot{\bar{b}}_g = [0]_{3 \times 3} \quad (20.39)$$

which provides the nominal-state update as

$$\hat{b}_{g, k|k-1} = \hat{b}_{g, k-1|k-1} \quad (20.40)$$

For the error-state covariance update in the prediction step, one requires the linearized continuous-time Euler angle error dynamics equation which can be derived from the derivative of the composition function of the true, nominal, and error Euler angles, i.e.,

$$\dot{C}_{N \leftarrow B}(\vec{\psi}_{B \leftarrow N}) = \left( I_{3 \times 3} + [\delta \vec{\psi}_{B \leftarrow N}]_{\times} \right) \dot{C}_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) + [\dot{\delta \vec{\psi}}_{B \leftarrow N}]_{\times} C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) \quad (20.41)$$

By definition of the derivative of the DCM, one has

$$\begin{aligned} C_{N \leftarrow B}(\vec{\psi}_{N \leftarrow B}) [\vec{\omega}_{B, B \leftarrow N}]_{\times} &= \left( I_{3 \times 3} + [\delta \vec{\psi}_{B \leftarrow N}]_{\times} \right) C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) [\vec{\omega}_{B, B \leftarrow \bar{N}}]_{\times} \\ &\quad + [\dot{\delta \vec{\psi}}_{B \leftarrow N}]_{\times} C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) \end{aligned} \quad (20.42)$$

Rearranging and by definition, one has

$$\begin{aligned} [\delta \dot{\vec{\psi}}_{B \leftarrow N}]_{\times} C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) &= \left( I_{3 \times 3} + [\delta \vec{\psi}_{B \leftarrow N}]_{\times} \right) C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) [\vec{\omega}_{B, B \leftarrow N}]_{\times} \\ &\quad - \left( I + [\delta \vec{\psi}_{B \leftarrow N}]_{\times} \right) C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) [\vec{\omega}_{B, B \leftarrow \bar{N}}]_{\times} \end{aligned} \quad (20.43)$$

$$[\dot{\delta \vec{\psi}}_{B \leftarrow N}]_{\times} = \left( I_{3 \times 3} + [\delta \vec{\psi}_{B \leftarrow N}]_{\times} \right) C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) \left( [\vec{\omega}_{B, B \leftarrow N}]_{\times} - [\vec{\omega}_{B, B \leftarrow \bar{N}}]_{\times} \right) C_{B \leftarrow \bar{N}}(\bar{\psi}_{B \leftarrow N}) \quad (20.44)$$

Retaining the linear error terms, one has the approximation

$$[\delta \dot{\vec{\psi}}_{B \leftarrow N}]_{\times} \approx C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) \left( [\vec{\omega}_{B, B \leftarrow N}]_{\times} - [\vec{\omega}_{B, B \leftarrow \bar{N}}]_{\times} \right) C_{B \leftarrow \bar{N}}(\bar{\psi}_{B \leftarrow N}) \quad (20.45)$$

which can be written in equivalent vector form as

$$\delta \dot{\vec{\psi}}_{B \leftarrow N} \approx C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) (\vec{\omega}_{B, B \leftarrow N} - \vec{\omega}_{B, B \leftarrow \bar{N}}) \quad (20.46)$$

Splitting up the angular velocities into the components, one has

$$\delta \dot{\vec{\psi}}_{B \leftarrow N} \approx C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) \left( \vec{\omega}_{B, B \leftarrow I} - \vec{\omega}_{B, N \leftarrow I} - (\hat{\vec{\omega}}_{B, B \leftarrow I} - \vec{\omega}_{B, \bar{N} \leftarrow I}) \right) \quad (20.47)$$

where  $\hat{\vec{\omega}}_{B, B \leftarrow I}$  is the gyroscope angular velocity measurement. Then, defining

$$\delta \vec{\omega}_{B, B \leftarrow I} = \hat{\vec{\omega}}_{B, B \leftarrow I} - \vec{\omega}_{B, B \leftarrow I} \quad (20.48)$$

and

$$\delta \vec{\omega}_{B, N \leftarrow I} = \vec{\omega}_{B, \bar{N} \leftarrow I} - \vec{\omega}_{B, N \leftarrow I} \quad (20.49)$$

one has

$$\delta \dot{\vec{\psi}}_{B \leftarrow N} \approx C_{\bar{N} \leftarrow B}(\bar{\psi}_{B \leftarrow N}) (\delta \vec{\omega}_{B, N \leftarrow I} - \delta \vec{\omega}_{B, B \leftarrow I}) \quad (20.50)$$

Noting that one can rearrange the composition function as

$$C_{\bar{N} \leftarrow B}(\vec{\psi}_{B \leftarrow N}) = C_{N \leftarrow B}(\psi_{B \leftarrow N}) + [\delta \vec{\psi}_{B \leftarrow N}]_{\times} C_{\bar{N} \leftarrow B}(\vec{\psi}_{B \leftarrow N}) \quad (20.51)$$

one has

$$\delta \dot{\vec{\psi}}_{B \leftarrow N} \approx \left( C_{N \leftarrow B}(\psi_{B \leftarrow N}) + [\delta \vec{\psi}_{B \leftarrow N}]_{\times} C_{\bar{N} \leftarrow B}(\vec{\psi}_{B \leftarrow N}) \right) (\delta \vec{\omega}_{B, N \leftarrow I} - \delta \vec{\omega}_{B, B \leftarrow I}) \quad (20.52)$$

and discarding higher-order terms, one has the linear approximation

$$\delta \dot{\vec{\psi}}_{B \leftarrow N} \approx [\delta \vec{\psi}_{B \leftarrow N}]_{\times} \vec{\omega}_{N, N \leftarrow I} + \delta \vec{\omega}_{N, N \leftarrow I} - C_{N \leftarrow B}(\psi_{B \leftarrow N}) \delta \vec{\omega}_{B, B \leftarrow I} \quad (20.53)$$

or, using the identity  $[\vec{a}]_{\times} \vec{b} = -[\vec{b}]_{\times} \vec{a}$ , one has

$$\delta \dot{\vec{\psi}}_{B \leftarrow N} \approx -[\delta \vec{\omega}_{N, N \leftarrow I}]_{\times} \delta \vec{\psi}_{B \leftarrow N} + \delta \vec{\omega}_{N, N \leftarrow I} - C_{N \leftarrow B}(\psi_{B \leftarrow N}) \delta \vec{\omega}_{B, B \leftarrow I} \quad (20.54)$$

where one must drop the two terms with  $\delta \vec{\omega}_{N, N \leftarrow I}$  if the latitude and altitude of the vehicle is unknown. Assuming the axes errors in the gyroscope have been removed so that only the stochastic errors remain, one has

$$\delta \vec{\omega}_{B, B \leftarrow I} = \vec{b}_g + \vec{w}_g \quad (20.55)$$

For an Ornstein-Uhlenbeck process model for the gyroscope biases, one has

$$\dot{\vec{b}}_g = -\frac{1}{\tau_g} I_{3 \times 3} \vec{b}_g + \vec{\mu}_g \quad (20.56)$$

which can be stacked with the gyroscope white noise as the process noise vector

$$\vec{\mathbf{w}} = \begin{bmatrix} \vec{w}_g \\ \vec{\mu}_g \end{bmatrix} \quad (20.57)$$

with PSD

$$S = \begin{bmatrix} \sigma_{w_g}^2 I_{3 \times 3} & [0]_{3 \times 3} \\ [0]_{3 \times 3} & \sigma_{\mu_g}^2 I_{3 \times 3} \end{bmatrix} = \begin{bmatrix} \sigma_{w_g}^2 I_{3 \times 3} & [0]_{3 \times 3} \\ [0]_{3 \times 3} & \frac{2\sigma_{b_g}^2}{\tau_g} I_{3 \times 3} \end{bmatrix} \quad (20.58)$$

For a Wiener process model for the gyroscope biases, one has

$$\dot{\vec{b}}_g = \vec{\mu}_g \quad (20.59)$$

which can be stacked with the gyroscope white noise as the process noise vector

$$\vec{\mathbf{w}} = \begin{bmatrix} \vec{w}_g \\ \vec{\mu}_g \end{bmatrix} \quad (20.60)$$

with PSD

$$S = \begin{bmatrix} \sigma_{w_g}^2 I_{3 \times 3} & [0]_{3 \times 3} \\ [0]_{3 \times 3} & \sigma_{\mu_g}^2 I_{3 \times 3} \end{bmatrix} \quad (20.61)$$

Taken together, the linearized error-state dynamics equation is

$$\delta \dot{\vec{x}} = A \delta \vec{x} + L \vec{\mathbf{w}} \quad (20.62)$$

$$\begin{bmatrix} \delta \dot{\vec{\psi}}_{B \leftarrow N} \\ \dot{\vec{b}}_g \end{bmatrix} = A \begin{bmatrix} \delta \vec{\psi}_{B \leftarrow N} \\ \vec{b}_g \end{bmatrix} + L \begin{bmatrix} \vec{w}_g \\ \vec{\mu}_g \end{bmatrix} \quad (20.63)$$

where, for an Ornstein-Uhlenbeck process model for the bias, one has

$$A = \begin{bmatrix} [0]_{3 \times 3} & -C_{N \leftarrow B}(\psi_{B \leftarrow N}) \\ [0]_{3 \times 3} & -\frac{1}{\tau_g} I_{3 \times 3} \end{bmatrix} \quad (20.64)$$

or, for a Wiener process model for the bias, one has

$$A = \begin{bmatrix} [0]_{3 \times 3} & -C_{N \leftarrow B}(\psi_{B \leftarrow N}) \\ [0]_{3 \times 3} & [0]_{3 \times 3} \end{bmatrix} \quad (20.65)$$

and, for either process, one has

$$L = \begin{bmatrix} -C_{N \leftarrow B}(\psi_{B \leftarrow N}) & [0]_{3 \times 3} \\ [0]_{3 \times 3} & I_{3 \times 3} \end{bmatrix} \quad (20.66)$$

For the AHRS ES-EKF correction step at time step  $k$ , the innovation  $\delta \vec{y}_k = \vec{y}_k - h(\hat{\vec{x}}_{k|k-1})$  can be formed as

$$\delta \vec{y}_k = \begin{bmatrix} \delta \vec{g}_B \\ \delta \vec{h}_B \end{bmatrix} = \begin{bmatrix} \hat{\vec{g}}_{B,k} \\ \hat{\vec{h}}_{B,k} \end{bmatrix} - \begin{bmatrix} C_{B \leftarrow \bar{N}}(\hat{\psi}_{k|k-1}) \vec{g}_N \\ C_{B \leftarrow \bar{N}}(\hat{\psi}_{k|k-1}) \vec{h}_N \end{bmatrix} \quad (20.67)$$

where  $C_{B \leftarrow N}(\hat{\psi}_{k|k-1})$  is the DCM obtained from the Euler angle nominal-state,  $\hat{h}_{B,k}$  is the magnetic field measurement from magnetometer, and  $\hat{g}_{B,k}$  is the estimated gravity vector. Alternatively, one can multiply this equation by the  $C_{N \leftarrow B}^T(\hat{\psi}_{k|k-1})$  to obtain the equivalent innovation in NED navigation frame as

$$\delta \vec{y}_k = \begin{bmatrix} \delta \vec{g}_N \\ \delta \vec{h}_N \end{bmatrix} = \begin{bmatrix} C_{\bar{N} \leftarrow B}(\hat{\psi}_{k|k-1}) \hat{g}_{B,k} \\ C_{\bar{N} \leftarrow B}(\hat{\psi}_{k|k-1}) \hat{h}_{B,k} \end{bmatrix} - \begin{bmatrix} \vec{g}_N \\ \vec{h}_N \end{bmatrix} \quad (20.68)$$

which will be developed here.

Notably, an AHRS often models the  $\hat{g}_B$  as the opposite of the accelerometer's specific force measurement under the assumption of no other accelerations as discussed in previous sections. However, due to the coupling between the gyroscope and accelerometer, one can also estimate a slowly-varying acceleration with the accelerometer bias evolving according to an Ornstein-Uhlenbeck process which is often not a useful model for aerospace vehicle attitude determination. An alternative option for aircraft is to use an ADAHRS which provides an airspeed measurement,  $v_\infty$ , from which one can estimate the total acceleration with the Coriolis acceleration, which is typically much larger than the translational acceleration, as

$$\vec{a}_B = \dot{\vec{v}}_B + [\vec{\omega}_{I \leftarrow B, B}] \times \vec{v}_B \quad (20.69)$$

$$\vec{a}_B \approx [\vec{\omega}_{I \leftarrow B, B}] \times \vec{v}_B \quad (20.70)$$

$$\vec{a}_B \approx [\vec{\omega}_{I \leftarrow B, B}] \times \vec{v}_B \quad (20.71)$$

$$\vec{a}_B \approx \begin{bmatrix} 0 \\ v_\infty r_g \\ -v_\infty q_g \end{bmatrix} \quad (20.72)$$

where  $v_\infty$  is the airspeed and  $q_g$  and  $r_g$  are the gyroscope measurements of the angular velocity about the  $y$  body-fixed frame axis and  $z$  body-fixed frame axis, respectively. Notably, this correlates the gyroscope angular velocity measurements used in the nominal-state estimation with the airspeed-based error-state estimation. In this case, one may also be able to better model and estimate the accelerometer bias.

Here, the error-state covariance update requires the linearization of the innovation model with respect to the error-state. The accelerometer-based gravity model can be rewritten given by the errors in  $C_{\bar{N} \leftarrow B}(\hat{\psi}_{k|k-1})$  and  $\hat{g}_{B,k}$ , i.e.,

$$\delta \vec{g}_N = \left( I_{3 \times 3} - [\delta \vec{\psi}_{B \leftarrow N}] \right) C_{N \leftarrow B} (\vec{g}_B - \delta \vec{a}_{B,I \leftarrow B} - \delta \vec{f}_B - \vec{v}_f) - \vec{g}_N \quad (20.73)$$

or, multiplying out the  $\vec{g}_B$  term and combining the errors into a single lumped error term  $\vec{v}_a = \delta \vec{a}_{B,I \leftarrow B} + \delta \vec{f}_B + \vec{v}_f$  with covariance  $R_a$ , one has

$$\delta \vec{g}_N = -[\delta \vec{\psi}_{B \leftarrow N}] \times \vec{g}_N + \left( I_{3 \times 3} - [\delta \vec{\psi}_{B \leftarrow N}] \right) C_{N \leftarrow B} \vec{v}_a \quad (20.74)$$

Next, dropping the multiplicative error term, and using the identity  $[\vec{a}] \times \vec{b} = -[\vec{b}] \times \vec{a}$ , one has the accelerometer linearized measurement equation

$$\delta \vec{g}_N = [\vec{g}_N] \times \delta \vec{\psi}_{N \leftarrow B} + C_{N \leftarrow B} \vec{v}_a \quad (20.75)$$

or, for a gravity model only in the Down direction, one has

$$\delta \vec{g}_N = \begin{bmatrix} 0 & -\|g_N\| & 0 \\ \|g_N\| & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \delta \vec{\psi}_{N \leftarrow B} + C_{N \leftarrow B} \vec{v}_a \quad (20.76)$$

which demonstrates only roll and pitch are observable by the accelerometer as expected which means one may alternatively form the magnetometer measurement equation as

$$\delta \vec{g}_{N,1:2} = \begin{bmatrix} 0 & -\|g_N\| \\ \|g_N\| & 0 \end{bmatrix} \begin{bmatrix} \delta \phi \\ \delta \theta \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} C_{N \leftarrow B} \vec{v}_a \quad (20.77)$$

Notably, the acceleration due to gravity magnitude will vary with latitude and altitude, so the position estimate may be necessary for the attitude determination.

The linear error in the magnetometer model is given by

$$\delta \vec{h}_N = \left( I_{3 \times 3} - [\delta \vec{\psi}_{N \leftarrow B}]^\times \right) C_{N \leftarrow B} (\vec{h}_B + \vec{v}_h) - \vec{h}_N \quad (20.78)$$

which similarly provides

$$\delta \vec{h}_N = [\vec{h}_N]_\times \delta \vec{\psi}_{N \leftarrow B} + C_{N \leftarrow B} \vec{v}_h \quad (20.79)$$

$$\delta \vec{h}_N = \begin{bmatrix} 0 & -h_{N,3} & h_{N,2} \\ h_{N,3} & 0 & -h_{N,1} \\ -h_{N,2} & h_{N,1} & 0 \end{bmatrix} \delta \vec{\psi}_{N \leftarrow B} + C_{N \leftarrow B} \vec{v}_h \quad (20.80)$$

where  $\vec{v}_h$  has covariance  $R_h$  and accounts for the combined magnetometer measurement noise, magnetic field disturbances, and residual errors in the sensor calibration. If strong magnetic field disturbances are expected, then one may choose to restrict the aiding to only the necessary error-state, i.e., the yaw error-state, as

$$\delta \vec{h}_N = \begin{bmatrix} h_{N,2} \\ -h_{N,1} \\ 0 \end{bmatrix} \delta \psi + C_{N \leftarrow B} \vec{v}_h \quad (20.81)$$

Furthermore, as there is only a weak dependence on  $h_{N,2}$ , one may alternatively form the magnetometer measurement equation as

$$\delta h_{N,2} = -h_{N,1} \delta \psi + [0 \ 1 \ 0] C_{N \leftarrow B} \vec{v}_h \quad (20.82)$$

which is identical to forming a “leveled” direct yaw measurement model. Notably, this yaw angle error will be with respect to magnetic north, so the position estimate may be used to form the heading reference to true north.

The combined AHRS ES-EKF measurement model for the innovation  $\delta \vec{y}_k$  is

$$\delta \vec{y}_k = \begin{bmatrix} \delta \vec{g}_N \\ \delta \vec{h}_N \end{bmatrix} = \begin{bmatrix} C_{N \leftarrow B} (\hat{\vec{\psi}}_{k|k-1}) \hat{\vec{g}}_{B,k} \\ C_{N \leftarrow B} (\hat{\vec{\psi}}_{k|k-1}) \hat{\vec{h}}_{B,k} \end{bmatrix} - \begin{bmatrix} \vec{g}_N \\ \vec{h}_N \end{bmatrix} \quad (20.83)$$

The combined full linearized error-state measurement equation is

$$\delta \vec{y}_k = H_k \delta \vec{x}_k + M_k \vec{v}_k \quad (20.84)$$

with

$$H_k = \begin{bmatrix} 0 & -\|g_N\| & 0 & 0 & 0 & 0 \\ \|g_N\| & 0 & 0 & 0 & 0 & 0 \\ 0 & -h_{N,3} & h_{N,2} & 0 & 0 & 0 \\ h_{N,3} & 0 & -h_{N,1} & 0 & 0 & 0 \\ -h_{N,2} & h_{N,1} & 0 & 0 & 0 & 0 \end{bmatrix} \quad (20.85)$$

and

$$M_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} C_{N \leftarrow B,k}(\hat{\psi}_{k|k-1}) & [0]_{3 \times 3} \\ [0]_{3 \times 3} & C_{N \leftarrow B,k}(\hat{\psi}_{k|k-1}) \end{bmatrix} \quad (20.86)$$

or the combined alternative linearized error-state measurement equation is

$$\delta \vec{y}_k = \begin{bmatrix} \delta \vec{g}_{N,1:2} \\ \delta h_{N,2} \end{bmatrix} = H_k \delta \vec{x}_k + M_k \vec{v}_k \quad (20.87)$$

with

$$H_k = \begin{bmatrix} 0 & -\|g_N\| & 0 & 0 & 0 & 0 \\ \|g_N\| & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -h_{N,1} & 0 & 0 & 0 \end{bmatrix} \quad (20.88)$$

and

$$M_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} C_{N \leftarrow B,k}(\hat{\psi}_{k|k-1}) & [0]_{3 \times 3} \\ [0]_{3 \times 3} & C_{N \leftarrow B,k}(\hat{\psi}_{k|k-1}) \end{bmatrix} \quad (20.89)$$

Regardless, the stacked measurement noise vector is

$$\vec{v}_k = \begin{bmatrix} \vec{v}_{a,k} \\ \vec{v}_{h,k} \end{bmatrix} \quad (20.90)$$

with stacked covariance matrix

$$\mathbf{R} = \begin{bmatrix} R_a & [0]_{3 \times 3} \\ [0]_{3 \times 3} & R_h \end{bmatrix} \quad (20.91)$$

Lastly, for the ES-EKF nominal-state update step, the gyroscope nominal-state bias estimates are simply updated by their composition function

$$\hat{\vec{b}}_{g,k|k} = \hat{\vec{b}}_{g,k|k-1} + \delta \hat{\vec{b}}_{g,k} \quad (20.92)$$

For the Euler angle attitude update, one must use the corresponding DCM composition function, i.e.,

$$C_{N \leftarrow B}(\hat{\psi}_{B \leftarrow N,k|k}) = (I_{3 \times 3} + [\delta \hat{\psi}_{B \leftarrow N,k}] \times) C_{N \leftarrow B}(\hat{\psi}_{B \leftarrow N,k|k-1}) \quad (20.93)$$

where the updated Euler angles,  $\hat{\psi}_{B \leftarrow N,k|k}$  can be extracted from the elements of the  $3 \times 3$   $C_{N \leftarrow B}(\hat{\psi}_{B \leftarrow N,k|k})$  as

$$\hat{\psi}_{B \leftarrow N,k|k} = \begin{bmatrix} \tan^{-1} \frac{C_{32}}{C_{33}} \\ -\sin^{-1} C_{31} \\ \tan^{-1} \frac{C_{21}}{C_{11}} \end{bmatrix} \quad (20.94)$$

where  $C_{ij}$  is the element in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column of  $C_{N \leftarrow B}(\vec{\psi}_{B \leftarrow N, k|k})$ . Notably, the inverse tangent functions should be the four-quadrant inverse tangent based on the signs of the numerator and denominator.

## References

For more information, please refer to the following

- H. Mokhtarzadeh. “Appendix D Attitude Heading Reference System,” in *Correlated-Data Fusion and Cooperative Aiding in GNSS-Stressed or Denied Environments*, Ph.D. dissertation, Department of Aerospace Engineering & Mechanics, 2014, pp. 151-172

## 20.3 Multi-Source Direct Attitude Determination Systems

A **feature tracker** is a sensor that measures the features in the environment which can include the bearing, range, or position of passive or static objects and the strength or direction of a vector field, e.g., celestial bodies, edges, gravitational fields, or magnetic fields. Feature trackers can be a passive radio, infrared, or optical, or multi-spectral sensor or an active radar, sonar, or lidar.

### Attitude Determination from Arbitrary Vector Measurements

Using feature trackers provides the ability to compute a full attitude solution provided that two vectors can be measured in the body-fixed frame,  $\vec{u}_B$  and  $\vec{v}_B$ , and are known in the relative reference frame for attitude, e.g., the LVLH frame,  $\vec{u}_L$  and  $\vec{v}_L$ . From these vectors, one can form the following two pairs of orthogonal vector quantities to  $\vec{u}_B$  and  $\vec{u}_L$  as

$$\vec{r}_B = \frac{[\vec{u}_B] \times \vec{v}_B}{\|[\vec{u}_B] \times \vec{v}_B\|_2} \quad \vec{r}_L = \frac{[\vec{u}_L] \times \vec{v}_L}{\|[\vec{u}_L] \times \vec{v}_L\|_2} \quad (20.95)$$

$$\vec{s}_B = [\vec{u}_B] \times \vec{r}_B \quad \vec{s}_L = [\vec{u}_L] \times \vec{r}_L \quad (20.96)$$

Then, one can form the body matrix as

$$M_B = [\vec{u}_B \quad \vec{r}_B \quad \vec{s}_B] \quad (20.97)$$

and the reference matrix as

$$M_L = [\vec{u}_L \quad \vec{r}_L \quad \vec{s}_L] \quad (20.98)$$

which provides the relationship

$$M_B = C_{B \leftarrow L} M_L \quad (20.99)$$

which can be inverted to form the DCM estimate

$$\hat{C}_{B \leftarrow L} = M_B M_L^{-1} \quad (20.100)$$

which has covariance matrix

$$P_{\hat{C}} = \sigma_{\vec{u}}^2 I_{3 \times 3} + \frac{1}{\|\vec{u}_B \times \vec{v}_B\|_2^2} \left( \left( \sigma_{\vec{v}}^2 - \sigma_{\vec{u}}^2 \right) \vec{u}_B \vec{u}_B^T + \sigma_{\vec{u}}^2 (\vec{u}_B \cdot \vec{v}_B) \left( \vec{u}_B \vec{v}_B^T + \vec{v}_B \vec{u}_B^T \right) \right) \quad (20.101)$$

where  $\sigma_{\vec{u}}^2$  is the variance of the  $\vec{u}$  measurement and  $\sigma_{\vec{v}}^2$  is the variance of the  $\vec{v}$  measurement.

When more than two vectors quantities are present, then one can use an optimal parameter estimation algorithm to estimate the direction cosine matrix (DCM) or the quaternions. In this case, the weighted least-squares formulation of the attitude determination problem is known as **Wahba's problem** which states the following minimization for the DCM

$$\hat{C}_{B \leftarrow L} = \underset{C_{B \leftarrow L}}{\operatorname{argmin}} \frac{1}{2} \sum_{k=1}^N w_k \| \vec{b}_{B,k} - C_{B \leftarrow L} \vec{b}_{L,k} \|_2^2 \quad \text{for } N \geq 2 \quad (20.102)$$

where  $w_k$  are the individual weights of the  $N$  vector measurements,  $\vec{b}_{B,k}$  is the vector measurement for the  $k^{\text{th}}$  feature in the body-fixed frame, and  $\vec{b}_{L,k}$  is the reference vector for the  $k^{\text{th}}$  feature in the LVLH frame. This can also be rewritten in a more convenient form using the definition

$$B = \sum_{k=1}^N a_k \vec{b}_{B,k} \vec{b}_{L,k}^T \quad (20.103)$$

which provides

$$\hat{C}_{B \leftarrow L} = \underset{C_{B \leftarrow L}}{\operatorname{argmin}} \sum_{k=1}^N a_k - \operatorname{Tr} \left( C_{B \leftarrow L} B^T \right) \quad (20.104)$$

or

$$\hat{C}_{B \leftarrow L} = \underset{C_{B \leftarrow L}}{\operatorname{argmax}} \operatorname{Tr} \left( C_{B \leftarrow L} B^T \right) \quad (20.105)$$

The SVD solution method is to determine the singular value decomposition (SVD) of the non-square matrix  $B$ , i.e.

$$B = U \Sigma V^T \quad (20.106)$$

where  $U$  is an  $3 \times 3$  orthogonal matrix,  $V$  is an  $3 \times 3$  orthogonal matrix, and  $\Sigma$  is a diagonal  $3 \times 3$  matrix with *non-negative* real numbers on the diagonal, i.e.

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} \quad (20.107)$$

The diagonal entries of  $\Sigma$  are known as the **singular values** of  $B$  and are ordered in size with  $\sigma_1 \geq \sigma_2 \geq \sigma_3$ . The singular values are also the square root of the eigenvalues of  $B^T B$  or  $BB^T$  and are defined for all matrices, even non-square.

Then, the estimate of the DCM can be computed as

$$\hat{C}_{B \leftarrow L} = U \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(U) \det(V) \end{bmatrix} \quad (20.108)$$

with covariance

$$P_{\hat{C}} = U \begin{bmatrix} (\sigma_2 + \det(U) \det(V) \sigma_3)^{-1} & 0 & 0 \\ 0 & (\det(U) \det(V) \sigma_3 + \sigma_1)^{-1} & 0 \\ 0 & 0 & (\sigma_1 + \sigma_2)^{-1} \end{bmatrix} U^T \quad (20.109)$$

Other solution methods to Wahba's problem use quaternions methods to compute the optimal, e.g. the  $q$  and QUaternion ESTimator (QUEST) methods. The following two subsections discuss how one can use only one vector measurement from a single feature tracker to obtain a direct partial attitude determination.

Notably, for spacecraft, the bearing vector to many different features can be tracked including the Earth, the sun, the moon, and the stars. A **star tracker** is a sensor which uses optical photocells or cameras to sense the magnitude of brightness and spectral type to identify and measure the relative bearing vectors to specific identifiable stars, known as **navigational stars**. Thus, if one has a **star map** defined for the bearing vectors to identifiable stars in a known reference frame, e.g., the ICRF, ECI, or ECEF frame, then one can use these bearing vectors in the body-fixed frame to solve for the attitude of the spacecraft in the body-fixed frame relative to the LVLH frame by first transforming the reference bearing vectors to the LVLH or navigation frame using the estimated position of the vehicle first which couples the positioning and attitude determination problems.

### Attitude Determination from Phase Measurements

Another type of environmental feature that can be used for attitude determination is the phase measurement of a constant frequency signal, e.g., the carrier phase of GNSS, at multiple antennas. Phase-based attitude estimation is an extension of differential phase positioning with the parameter to estimate as the baseline vector but with an additional constraint for a constant baseline, or nearly constant due to elastic motion of the vehicle. This very precise relative position estimate between a pair of antennas allows one to transform this relative position into angular measurements of the baseline vector. Thus, two baselines composed of three non-collinear antennas can completely define the attitude, e.g. the Euler angles.

For one baseline, recall that the double-differenced phase range equation

$$\nabla_j \Delta_{RB} \Phi_i = \Delta_{RB} r_i + \lambda \nabla_j \Delta_{RB} N_i + \Delta_{RB} \epsilon_i \quad (20.110)$$

which can be shown to be related to the constant **baseline vector**, i.e.

$$\vec{x}_{RB} = \begin{bmatrix} x_{RB} \\ y_{RB} \\ z_{RB} \end{bmatrix} = \begin{bmatrix} x_R - x_B \\ y_R - y_B \\ z_R - z_B \end{bmatrix} \quad (20.111)$$

as

$$\nabla_j \Delta_{RB} \Phi_i = \frac{(\vec{x}_i - \vec{x})^T}{\|\vec{x} - \vec{x}_i\|_2} \vec{x}_{RB} + \lambda \nabla_j \Delta_{RB} N_i + \nabla_j \Delta_{RB} \epsilon_i \quad (20.112)$$

Thus, in phase-based attitude estimation, one has the parameter vector for  $m + 1$  transmitters

$$\vec{\beta} = \begin{bmatrix} x_{RB} \\ y_{RB} \\ z_{RB} \\ \nabla_j \Delta_{RB} N_1 \\ \vdots \\ \nabla_j \Delta_{RB} N_m \end{bmatrix} \quad (20.113)$$

which has  $m + 3$  unknowns with  $m$  DD measurements. Thus, one requires the use of multiple time steps in order to first resolve the integer ambiguities,  $\nabla_j \Delta_{RB} N_i$ . However, the known constraint of the baseline

allows the fixed solution to resolve the integer ambiguities quicker from the float, e.g. **baseline-constrained LAMBDA method**. Typically, one uses a short-baseline to use the same LOS vectors for all three receivers in the NLS algorithm and reduce the number of candidate sets of integers.

However, once one has resolved the integer ambiguities, one can form the OLS solution for the baseline vector estimate as

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\nabla_j \Delta_{RB} \Phi_i - \lambda \nabla_j \Delta_{RB} N_i) \quad (20.114)$$

where

$$\mathbf{X} = \begin{bmatrix} \frac{(\vec{x}_1 - \vec{x})^T}{\|\vec{x} - \vec{x}_1\|_2} \\ \vdots \\ \frac{(\vec{x}_m - \vec{x})^T}{\|\vec{x} - \vec{x}_m\|_2} \end{bmatrix} \quad (20.115)$$

Furthermore, for two baselines with one rover denoted by the subscript  $R1$  and the second rover denoted by the subscript  $R2$ , one can write

$$[\hat{x}_{N,R1B} \quad \hat{x}_{N,R2B} \quad \hat{x}_{N,R1B} \times \hat{x}_{N,R2B}] = C_{N \leftarrow B} [\vec{x}_{B,R1B} \quad \vec{x}_{B,R2B} \quad \vec{x}_{B,R1B} \times \vec{x}_{B,R2B}] \quad (20.116)$$

Thus, one can form the DCM estimate from

$$\hat{C}_{N \leftarrow B} = [\hat{x}_{N,R1B} \quad \hat{x}_{N,R2B} \quad \hat{x}_{N,R1B} \times \hat{x}_{N,R2B}] [\vec{x}_{B,R1B} \quad \vec{x}_{B,R2B} \quad \vec{x}_{B,R1B} \times \vec{x}_{B,R2B}]^{-1} \quad (20.117)$$

and the Euler angle estimates can be computed as

$$\begin{bmatrix} \hat{\phi} \\ \hat{\theta} \\ \hat{\psi} \end{bmatrix} = \begin{bmatrix} \tan^{-1} \frac{c_{32}}{c_{33}} \\ -\sin^{-1} \frac{c_{31}}{c_{33}} \\ \tan^{-1} \frac{c_{21}}{c_{11}} \end{bmatrix} \quad (20.118)$$

where  $c_{ij}$  is the element in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column of  $\hat{C}_{N \leftarrow B}$ .

## References

For more information, please refer to the following

•

- Stevens, B. L., Lewis, F. L., and Johnson, E. N., “7 Digital Control,” *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed., Vol. 1, Wiley-Blackwell, New York, 2015, pp. 584-620

---

# Navigation Systems

## 21.1 Introduction to Navigation Systems

**Navigation systems** are computing systems that provide estimates of the position of a vehicle relative to some reference frame while *in motion*. For Earth-based navigation, one typically uses the LLA coordinates of the ECEF reference frame to navigate from one point on the Earth's surface to another. For space-based navigation, one typically uses a planet-, moon-, or sun-centered inertial reference frame to navigate in orbit. Thus, positioning systems are a type of navigation system, but navigation systems also relate the velocity of the vehicle over time to infer the position of a moving vehicle. Thus, positioning systems can easily be expanded to navigation systems, by using a vehicle-specific motion model, known as a **vehicle dynamics model**, or a vehicle-agnostic motion model as a kinematic predictive measurement and the positioning sensor outputs as corrective measurements to perform navigation using a navigation state filter. A **dead reckoning navigation system** performs navigation only using a velocity measurement. However, most navigation systems use multiple types of measurements which result in navigation state filter designs to fuse these measurements, e.g., one can use a state filter to estimate a GNSS receiver's position, velocity, clock offset, and clock drift, as a function of time which forms the navigation state

$$\vec{x} = \begin{bmatrix} \vec{x} \\ \dot{\vec{x}} \\ c\delta t \\ f\dot{\delta t} \end{bmatrix} \quad (21.1)$$

However, most navigation systems are multi-sensor systems fusing measurements from dead reckoning sensor(s) with measurements from a positioning sensor (s) in a navigation filter. This section will discuss the general form of these types of systems while later sections will discuss a specific dead reckoning approach known as inertial navigation, aiding inertial navigation with different specific sensors/systems, and the problem of navigating when the map of the environment is unknown.

## Dead Reckoning Navigation

Some dead reckoning navigation systems use the vehicle's distance traveled directly in the position update. A sensor that measures distance traveled is known as an **odometer**. While odometers for ground vehicles can be based on wheel rotations, odometers for aerospace vehicles can be developed based on 2D or 3D imagers from electro-optical/infrared sensors which estimate the change in the vehicle's position from one image frame to the next using auto-correlation techniques between subsequent images. **Visual odometry (VO)**, also known as **electro-optical/infrared (EO-IR) odometry** or **camera odometry**, is the process of using image sequences to estimate the motion of the imager as odometry measurements to be used for dead reckoning positioning. These sequences can be of different lengths depending on the number of images used to form the odometry measurements. VO can use either features extracted from the images which requires feature tracking algorithms, a type of object tracking algorithms, or the entire image intensity values. In either case, the pseudomeasurement from the imager is either **optical flow**, the two-dimensional motion of the field of points from a two-dimensional imaging sensor, e.g., a single EO/IR camera, or **scene flow**, also known specifically as **radar flow** or **lidar flow**, the three-dimensional motion of the field of points from a three-dimensional imaging sensor, e.g., a radar, lidar, or stereo EO/IR camera.

Lastly, it should be noted that most odometers, velocimeters, and accelerometers are **strapdown sensors**, i.e., the sensors are attached to the structure of the vehicle and their measurements are referenced to the body-fixed frame of the vehicle. In this case, the navigation system will also require attitude determination to transform the velocity vector to the navigation frame for the dead reckoning computation. Thus, many navigation systems use a combination of attitude determination and dead reckoning equations which implicitly provides estimates of the position, velocity, and attitude (PVA) of the vehicle.

For a odometry-based dead reckoning system, the LLA position in the ECEF frame can be predicted by the model

$$\vec{p}_{E,k} = \begin{bmatrix} \dot{\ell}_k \\ \dot{\lambda}_k \\ \dot{h}_k \end{bmatrix} = \begin{bmatrix} \dot{\ell}_{k-1} \\ \dot{\lambda}_{k-1} \\ \dot{h}_{k-1} \end{bmatrix} + \begin{bmatrix} \frac{1}{R_N(\ell_{k-1})+h_{k-1}} & 0 & 0 \\ 0 & \frac{1}{(R_E(\ell_{k-1})+h_{k-1}) \cos \ell_{k-1}} & 0 \\ 0 & 0 & -1 \end{bmatrix} \vec{u}_{N,k} \quad (21.2)$$

where  $\vec{u}_{N,k}$  is the change in position from  $k - 1$  to  $k$  as measured in the navigation frame which may need to be computed from a strapdown odometer as

$$\vec{u}_{N,k} = C_{N \leftarrow B,k} \vec{u}_{B,k} \quad (21.3)$$

Notably in this case, an approximation of the velocity at time step  $k$ ,  $\vec{v}_{N,k}$ , is given simply by

$$\vec{v}_{N,k} \approx \vec{u}_{N,k} \Delta t \quad (21.4)$$

where  $\Delta t$  is the sampling time interval of the odometer.

Some dead reckoning navigation systems use velocity directly in the position update. A sensor that measures velocity is known as a **velocimeter**. Electro-optical velocimeters use Doppler frequency measurements from electromagnetic signals to infer relative range rate from features in the environment. In this case, one requires the pre-processing of multiple **pseudorange rates** using a least-squares estimator for the velocity measurement. Another approach for aircraft is to use the airspeed vector from an air data system (ADS) as a rough approximation for the inertial velocity, i.e. as no-wind and flat-Earth approximations, which may or may not include subtracting a provided wind speed from the provided airspeed.

For a velocimetry-based dead reckoning system, the LLA position in the ECEF frame is related to the velocity kinematically through the ordinary differential equation

$$\dot{\vec{p}}_E = \begin{bmatrix} \dot{\ell} \\ \dot{\lambda} \\ \dot{h} \end{bmatrix} = \begin{bmatrix} \frac{1}{R_N+h} & 0 & 0 \\ 0 & \frac{1}{(R_E+h) \cos \ell} & 0 \\ 0 & 0 & -1 \end{bmatrix} \vec{v}_N \quad (21.5)$$

where  $\vec{v}_N$  is the instantaneous velocity of the vehicle in the navigation frame which may need to be computed from a strapdown velocimeter as

$$\vec{v}_N = C_{N \leftarrow B} \vec{v}_B \quad (21.6)$$

This model can be integrated using an Euler integration to obtain a position prediction model as

$$\vec{p}_{E,k} = \begin{bmatrix} \ell_k \\ \lambda_k \\ h_k \end{bmatrix} = \begin{bmatrix} \ell_{k-1} \\ \lambda_{k-1} \\ h_{k-1} \end{bmatrix} + \Delta t \begin{bmatrix} \frac{1}{R_N(\ell_{k-1})+h_{k-1}} & 0 & 0 \\ 0 & \frac{1}{(R_E(\ell_{k-1})+h_{k-1}) \cos \ell_{k-1}} & 0 \\ 0 & 0 & -1 \end{bmatrix} \vec{v}_{N,k} \quad (21.7)$$

where  $\Delta t$  is the sampling time interval of the velocimeter. For an accelerometry-based dead reckoning system, the velocity is first updated as

$$\vec{v}_{N,k} = \vec{v}_{N,k-1} + \Delta t \vec{a}_{N,k-1} \quad (21.8)$$

where  $\vec{a}_{N,k-1}$  is the instantaneous velocity of the vehicle in the navigation frame and  $\Delta t$  is the sampling time interval of the accelerometer.

## 21.2 Inertial Navigation Systems

An **inertial navigation system (INS)** is a multi-sensor navigation system that uses an inertial measurement unit (IMU) and a model-based data fusion algorithm to continuously calculate the position, velocity, and attitude of a vehicle, i.e. the vehicle's **navigation state**. As INS do not require external references for computing a vehicle's position, un-aided INS perform dead reckoning (DR). An INS uses the IMU measurements of angular rate and specific force at every time step. Then, after initialization or the storage of the previous navigation state, one performs following three recursive updates:

1. Attitude update based on integrating the three-axis angular rate measurement;
2. Transformation of the specific force from body-fixed frame coordinates to navigation frame coordinates;
3. Gravitational force model calculation based on estimated position;
4. Velocity update based on integrating the acceleration; and
5. Position update based on integrating the velocity.

In INS design, there exists an inherent tradeoff among accuracy, processing efficiency, and model complexity. Before the 2000s, accuracy of the INS alone as well as processing efficiency were extremely important, thus, there exists many highly complex models for the INS equations. However, since the 2000s, most INS are aided with other sensors which allows a reduction in the complexity, e.g. GNSS coupling.

The discussion of all possible representations is beyond the scope of this textbook and a simple INS implementation for automotive/consumer grade IMUs is presented in this section. However, first, some notation must be introduced. In vehicle reference frames, recall the following frame subscripts

- $I$ : Earth-Centered Inertial (ECI)
- $E$ : Earth-Centered Earth Fixed (ECEF)
- $N$ : navigation frame
- $B$ : body-fixed frame

For an INS, the **navigation state** is composed of the vehicle's position, velocity, and attitude (PVA). The vehicle's position is typically expressed in geodetic coordinates, i.e., latitude ( $\ell$ ), longitude ( $\lambda$ ), and altitude ( $h$ ), grouped as the position vector given by

$$\vec{p}_E = \begin{bmatrix} \ell \\ \lambda \\ h \end{bmatrix} \quad (21.9)$$

The vehicle's velocity vector is defined in the North-East-Down (NED) navigation frame,  $\vec{v}_N$ . The vehicle's attitude can be defined in three different ways. The first is the body-to-navigation frame transformation sequence of 3 – 2 – 1 Euler angles, i.e., yaw ( $\psi$ ), pitch ( $\theta$ ), and roll ( $\phi$ ), grouped as the attitude vector

$$\vec{\psi}_{N \leftarrow B} = \begin{bmatrix} \phi \\ \theta \\ \psi \end{bmatrix} \quad (21.10)$$

The second is the body-to-navigation frame DCM,  $C_{N \leftarrow B}$ . The third is the body-to-navigation frame rotation quaternion,  $\vec{q}_{N \leftarrow B}$ . However, when using the DCM or rotation quaternion, the INS update equations require constraints relating the independent and dependent parameters, i.e.,  $C_{N \leftarrow B}$  must be positive definite and  $\vec{q}_{N \leftarrow B}$  must be a unit quaternion.

When initializing an INS, the initial position estimate is assumed to be at or near a known location with some uncertainty or is computed using a positioning algorithm with another sensor. For the initial velocity estimate, if the vehicle is static, the inertial velocity is approximately the velocity of the Earth's surface at the initial position. Alternatively, if the vehicle is in motion, the initial velocity must be supplied by another sensor. For the initial attitude, if the vehicle is static and then the initial pitch and roll can be computed from a leveling procedure and, if navigation grade IMUs are being used, the initial heading or yaw can be computed from a gyrocompassing procedure. **Leveling** assumes the IMU is static where the accelerometers will only measure the gravity vector which is known to point in the down direction of the navigation frame whose direction in the body frame can be related to the pitch and roll. **Gyrocompassing** assumes the IMU is static where gyroscopes will only measure the Earth's rotation. However, the gyroscope errors must be adequately small with little drift in order to sense the Earth's rotation rate for a significant period of time. Alternatively, the initialization of the Euler angles may come from another sensor, e.g., a magnetometer which are typically included within IMUs for this and other reasons.

Naturally, the INS update equations are given by the continuous-time rigid-body kinematics for position, velocity, and attitude (PVA). The **continuous-time INS attitude update** is given by

$$\dot{\vec{\psi}}_{N \leftarrow B} = \begin{bmatrix} 1 & \sin \phi \tan \theta & \cos \phi \tan \theta \\ 0 & \cos \phi & -\sin \phi \\ 0 & \sin \phi \sec \theta & \cos \phi \sec \theta \end{bmatrix} (\vec{\omega}_{B,I \leftarrow B} - \vec{\omega}_{B,I \leftarrow N}) = A (\vec{\omega}_{B,I \leftarrow B} - \vec{\omega}_{B,I \leftarrow N}) \quad (21.11)$$

The **continuous-time INS velocity update** is given by

$$\dot{\vec{v}}_N = (C_{N \leftarrow B} \vec{f}_B + \vec{g}_N) - [2\vec{\omega}_{N,I \leftarrow E} + \vec{\omega}_{N,E \leftarrow N}] \times \vec{v}_N \quad (21.12)$$

where  $\vec{f}_B$  is the specific force measured by the three-axis accelerometer in the body-fixed frame and  $\vec{g}_N$  is the gravity vector in the navigation frame. Notably, if the accelerometer is not at the center of mass of the vehicle or its axes are not aligned with the body-fixed frame corresponding to the vehicle's attitude vector, then the measured specific force,  $\vec{f}_S$ , may need to be translated and rotated from the sensor frame, subscript  $S$ , i.e.,

$$\vec{f}_B = C_{B \leftarrow S} \vec{f}_S - [\vec{x}_{B,S/B}] \times \vec{\omega}_{B,I \leftarrow B} \quad (21.13)$$

where  $C_{B \leftarrow S}$  is the DCM from the sensor frame to the body frame and  $\vec{x}_{B,S/B}$  is the relative position of the sensor frame origin with respect to the center of mass expressed in body frame coordinates.

The **continuous-time INS position update** is given by

$$\dot{\vec{p}}_E = \begin{bmatrix} \dot{\ell} \\ \dot{\lambda} \\ \dot{h} \end{bmatrix} = \begin{bmatrix} \frac{1}{N_\ell + h} & 0 & 0 \\ 0 & \frac{1}{(M_\ell + h) \cos \ell} & 0 \\ 0 & 0 & -1 \end{bmatrix} \vec{v}_N = T(\vec{p}_E) \vec{v}_N \quad (21.14)$$

where  $N_\ell$  is the **east-west radius of curvature**, also known as the **prime vertical radius of curvature**, of the reference ellipsoid, i.e.

$$N_\ell = \frac{a(1-e^2)}{(1-e^2 \sin^2 \ell)^{1.5}} \quad (21.15)$$

and  $M_\ell$  is **north-south radius of curvature**, also known as the **meridian radius of curvature**, of the reference ellipsoid, i.e.

$$M_\ell = \frac{a}{\sqrt{1-e^2 \sin^2 \ell}} \quad (21.16)$$

where  $R_e = a$  is the equatorial radius, also known as the Earth's semi-major axis, of the reference ellipsoid and  $e$  is the eccentricity of the reference ellipsoid. Typically, one uses the WGS84 which provides the following values

$$R_e = 6,378,137 \text{ m} \quad (21.17)$$

and

$$e = \sqrt{f(2-f)} \quad (21.18)$$

where  $f$  is the WGS84 flattening, i.e.

$$f = \frac{1}{298.257223563} \quad (21.19)$$

Notably, the attitude can alternatively be updated via the DCM directly as

$$\dot{C}_{N \leftarrow B} = [\vec{\omega}_{N,N \leftarrow B}] \times C_{N \leftarrow B} = C_{N \leftarrow B} [\vec{\omega}_{B,N \leftarrow B}] \times \quad (21.20)$$

or via the rotation quaternion directly as

$$\dot{q}_{N \leftarrow B} = \frac{1}{2} \vec{q} \otimes \vec{\omega}_{B,N \leftarrow B} \quad (21.21)$$

From the IMU, the rate gyroscopes measure the angular velocity of the body-to-inertial frame, i.e.,  $\vec{\omega}_{B,I \leftarrow B}$  which requires the following correction

$$\vec{\omega}_{B,N \leftarrow B} = \vec{\omega}_{B,I \leftarrow B} - \vec{\omega}_{B,I \leftarrow N} \quad (21.22)$$

where  $\vec{\omega}_{B,I \leftarrow N}$  can be computed by

$$\vec{\omega}_{B,I \leftarrow N} = C_{B \leftarrow N} (\vec{\omega}_{N,I \leftarrow N}) \quad (21.23)$$

$$\vec{\omega}_{B,I \leftarrow N} = C_{B \leftarrow N} (\vec{\omega}_{N,I \leftarrow E} + \vec{\omega}_{N,E \leftarrow N}) \quad (21.24)$$

where  $\vec{\omega}_{N,I \leftarrow E}$  is the **Earth rotation rate** expressed in the navigation frame and can be computed using the formula

$$\vec{\omega}_{N,I \leftarrow E} = 7.292115 \times 10^{-5} \begin{bmatrix} \cos \ell \\ 0 \\ -\sin \ell \end{bmatrix} \text{ rad/s} \quad (21.25)$$

and  $\vec{\omega}_{N,E \leftarrow N}$  is the **transport rate** expressed in the navigation frame and can be computed using the formula

$$\vec{\omega}_{N,E \leftarrow N} = \begin{bmatrix} \frac{v_{N,y}}{M_\ell + h} \\ -\frac{v_{N,x}}{N_\ell + h} \\ -\frac{v_{N,y} \tan \ell}{M_\ell + h} \end{bmatrix} \quad (21.26)$$

However, when using consumer grade gyroscopes,  $\vec{\omega}_{B,I \leftarrow N}$  is much smaller than the gyroscope errors and it is reasonable to assume that  $\vec{\omega}_{B,N \leftarrow B} \approx \vec{\omega}_{B,I \leftarrow B}$ , i.e., one is navigating on a non-rotating, flat-Earth.

## INS Integration

In addition, the differential equation must be integrated to update the Euler angles from the previous time step(s) to the current. For automotive or consumer grade gyroscopes, a forwards Euler integration from the previous time step  $k - 1$  to the current time step  $k$  is typically used for the update, i.e.,

$$\vec{\psi}_{N \leftarrow B,k} = \vec{\psi}_{N \leftarrow B,k-1} + \Delta t A(\vec{\psi}_{N \leftarrow B,k-1}) \vec{\omega}_{B,N \leftarrow B} \quad (21.27)$$

where  $\Delta t$  is the time increment between steps  $k - 1$  and  $k$ . Note that higher order Runge-Kutta integrations are also used, but require state estimates from multiple past time steps to be saved in memory.

## INS Velocity Update

The NED velocity in the navigation frame,  $\vec{v}_N$  is related to the specific force,  $\vec{f}_B$ , normal gravity  $\vec{g}_N$ , and the Coriolis forces through the following differential equation

$$\dot{\vec{v}}_N = (\vec{f}_N + \vec{g}_N) - [2\vec{\omega}_{N,I \leftarrow E} + \vec{\omega}_{N,E \leftarrow N}] \times \vec{v}_N \quad (21.28)$$

From the IMU, the accelerometers measure the specific force in the body frame,  $\vec{f}_B$ , which must be translated to the navigation frame, i.e.

$$\vec{f}_N = C_{N \leftarrow B} \vec{f}_B \quad (21.29)$$

and the local gravity vector is a function of latitude and altitude.

For industrial & consumer grade INS, the following gravity is typically adequate:

$$\vec{g}_N = \begin{bmatrix} 0 \\ 0 \\ g(\ell, h) \end{bmatrix} \quad (21.30)$$

where the gravity can be given by the **Somigliana gravity model**

$$g(\ell, h) = g_0(\ell) \left( 1 - (3.157042870579883 \times 10^{-7} - 2.102689650440234 \times 10^{-9} \sin^2 \ell)h + (7.374516772941995 \times 10^{-14})h^2 \right) \quad (21.31)$$

where

$$g_0(\ell) = 9.7803253359 \frac{1 + 0.001931853 \sin^2 \ell}{\sqrt{1 - 0.00669438 \sin^2 \ell}} \quad (21.32)$$

For consumer grade IMUs, the specific force, gravity, and Coriolis forces can be assumed to be constant across the time step  $\Delta t$ . Thus, a forwards Euler integration from the previous time step  $k - 1$  to the current time step  $k$  is typically used for the update, i.e.,

$$\vec{v}_{N,k} = \vec{v}_{N,k-1} + \Delta t \dot{\vec{v}}_{N,k-1} \quad (21.33)$$

It should be noted that in this integration, the transformation of the specific force is assumed to be integrated over  $\Delta t$  with a constant rotation matrix  $C_{N \leftarrow B,k}$  which is an approximation that can be improved with better integration such as averaging the values for  $C_{N \leftarrow B,k}$  from the previous time step(s) and the current time step. In addition, it is important to note that often the Coriolis forces can even be dropped for low grade IMUs due to their effect being much less than the process noise. This implies that the navigation is being done on a flat, non-rotating Earth.

## INS Position Update

The LLA position in the ECEF frame is related to the velocity through the differential equation

$$\dot{\vec{p}}_E = \begin{bmatrix} \dot{\ell} \\ \dot{\lambda} \\ \dot{h} \end{bmatrix} = \begin{bmatrix} \frac{1}{R_N+h} & 0 & 0 \\ 0 & \frac{1}{(R_E+h)\cos \ell} & 0 \\ 0 & 0 & -1 \end{bmatrix} \vec{v}_N \quad (21.34)$$

For consumer-grade and most industrial-grade IMUs, a forwards Euler integration, from the previous time step  $k - 1$  to the current time step  $k$  is typically used for the update, i.e.,

$$\vec{p}_{E,k} = \begin{bmatrix} \ell_k \\ \lambda_k \\ h_k \end{bmatrix} = \begin{bmatrix} \ell_{k-1} \\ \lambda_{k-1} \\ h_{k-1} \end{bmatrix} + \Delta t \begin{bmatrix} \frac{1}{R_N(\ell_{k-1})+h_{k-1}} & 0 & 0 \\ 0 & \frac{1}{(R_E(\ell_{k-1})+h_{k-1})\cos \ell_{k-1}} & 0 \\ 0 & 0 & -1 \end{bmatrix} \vec{v}_{N,k} = T(\vec{p}_{E,k-1}) \vec{v}_N \quad (21.35)$$

which assumes a constant integration period over  $\Delta t$ . For tactical-grade and navigation-grade IMUs, an RK(4) method may be preferred although if there are noticeable divergences in  $\Delta t$  from its nominal time interval, a combination of Euler integrations for noticeably off-nominal  $\Delta t$ 's and RK(4) for nominal  $\Delta t$ 's may be used. It is also important to note that care should be taken in the precision of the values of  $\ell$  and  $\lambda$  as 1 m of north-south displacement is approximately equivalent to  $1.6 \times 10^{-7}$  radians.

## INS Errors and Performance

Another consideration in the INS equations is the propagation of error, not only in the equation models, but also in the angular velocity and specific force measurements. By performing a perturbation analysis of this model using a first-order Taylor series expansion, i.e., linearization, one can identify how the errors, represented by  $\delta$ 's, propagate through the INS update equations, e.g.

$$\begin{aligned}\dot{\delta\vec{\psi}}_{N \leftarrow B} &\approx -[\vec{\omega}_{N,I \leftarrow N}]_x \delta\vec{\psi}_{N \leftarrow B} + \delta\vec{\omega}_{N,I \leftarrow N} - C_{N \leftarrow B} \delta\vec{\omega}_{B,I \leftarrow B} \\ \delta\vec{v}_N &= [C_{N \leftarrow B} \vec{f}_B]_x \delta\vec{\psi}_{N \leftarrow B} + C_{N \leftarrow B} \delta\vec{f}_B \\ &\quad - [2\vec{\omega}_{N,I \leftarrow E} + \vec{\omega}_{N,E \leftarrow N}]_x \delta\vec{v}_N \\ &\quad - [2\delta\vec{\omega}_{N,I \leftarrow E} + \delta\vec{\omega}_{N,E \leftarrow N}]_x \vec{v}_N + \delta\vec{g}_N \\ \dot{\delta\vec{p}}_E &= T' \delta\vec{p}_N + T \delta\vec{v}_N\end{aligned}\tag{21.36}$$

where the matrix  $T'$  relates the position errors to their time derivatives and  $\dot{\delta\vec{\psi}}_{N \leftarrow B}$  is not the Euler angle errors, but the attitude errors resolved about the navigation frame. The relation of these errors to Euler angle errors will be discussed later. It should be noted that  $\delta\vec{g}_N$  varies as a function of location, thus it is dependent on position errors which, in turn, lead to velocity errors. In particular, errors in altitude are significant as they lead to the well-known **vertical channel instability** problem. Thus, one typically uses some exteroceptive altimeter to form an aided INS.

Thus, INS performance is ultimately determined by the accelerometer and gyroscope measurement errors in the IMU. For these inertial sensors, simple errors models in the body frame can be chosen as

$$\delta\vec{f}_B = \vec{f}_B - (I_{3 \times 3} + M_a) \bar{f}_B + \vec{b}_a + \vec{w}_a\tag{21.37}$$

$$\delta\vec{\omega}_{B,I \leftarrow B} = \vec{\omega}_{B,I \leftarrow B} - (I_{3 \times 3} + M_g) \bar{\omega}_{B,I \leftarrow B} + \vec{b}_g + \vec{w}_g\tag{21.38}$$

where the  $a$  subscript represents the accelerometer parameters and  $g$  represents the gyroscope parameters. Here,  $\bar{f}_B$  and  $\bar{\omega}_{B,I \leftarrow B}$  are the *true* specific force and angular velocity,  $M_a$  and  $M_g$  account for scale factor and axes misalignment errors,  $\vec{b}_a$  and  $\vec{b}_g$  are the bias vectors, and  $\vec{w}_a$  and  $\vec{w}_g$  are uncorrelated additive measurement noise.

Each of the error terms can be modeled in the INS algorithms as constant, time-varying, or a combination of both whose magnitude and variability of these errors depends on the IMU grade. These errors limit the performance of an INS. In particular, the accelerometer errors are integrated twice in the INS equations and lead to position errors that grow as a function of  $t^2$ . Likewise, gyroscope errors are integrated thrice and lead to position errors that grow as a function of  $t^3$ . In order to mitigate the unbounded INS error growth, updates from exteroceptive sensor(s) are required which can be used to estimate these accelerometer and gyroscope biases forming an aided INS.

## References

For more information, please refer to the following

- S. Gleason and D. Gebre-Egziabher, “GNSS Applications and Methods,” Artech House, 2012
- C. Jekeli, “Inertial Navigation Systems with Geodetic Applications,” de Gruyter, 2001

## 21.3 Aided Inertial Navigation Systems

A specific class of VBNs with IMU fusion are called **vision-aided inertial navigation systems (VINS)**, also known as **vision-inertial navigation systems** or **visual-inertial navigation systems**. This fusion can either use the SLAM approach which substitutes the INS equations for the stochastic motion models of the general SLAM algorithm or uses the visual odometry approach, known specifically as **visual-inertial odometry (VIO)**.

A specific class of TBNs with IMU fusion are called **terrain-based navigation (TBN)**, also known as **terrain-referenced navigation (TRN)**, **terrain-relative navigation (TRN)**, **terrain-contour navigation (TCN)**, **terrain-aided navigation (TAN)**, and **terrain-matching navigation (TMN)**. This fusion substitutes the INS equations for the stochastic motion models of the general TBN algorithm.

GNSS and INS are complementary navigation systems utilizing three strengths of the INS and one strength of GNSS. First, the absence of an external signals for INS makes it more reliable and secure than GNSS. Second, the INS estimation rate is only limited by the processing power of the onboard computer which can be as fast as 100 Hz or more while most GNSS receivers only update from 1-20 Hz. Third, GNSS provides measurements for inferring position and velocity, but not attitude, while INS must estimate the attitude. However, one should note that multiple GNSS receivers with a known baseline can be used to estimate attitude. Fourth, INS output errors are time-correlated and unbounded (when using low-cost inertial sensors) while a GNSS receiver has bounded errors on its position and velocity estimates.

Thus, integration of GNSS and INS provides estimates to arrest the effect of the INS errors through the GNSS measurements, while the INS provides high bandwidth attitude, position, and velocity for vehicle guidance and control. There are many different integration schemes for the data fusion of GNSS and INS, but in general there are three different types: loose, tight, and deep. Though some crossover and ambiguity may exist between the three, these distinctions will assist in explaining the different concepts in GNSS/INS integration.

### Types of Integration for Aiding Inertial Navigation Systems

A typical **loose integration**, also known as **position-domain integration** or loosely-coupled, operate the INS and VBN/TBN/GNSS as independent navigation systems and the navigation solution from each is blended using an estimator to form a third navigation solution. The algorithms for these systems are state-to-state fusion algorithms. The primary variant to this scheme is the inclusion or exclusion of the dashed-line as a feedback signal in the blending software leading to either a closed-loop or open-loop configuration. The inclusion of the feedback signal is necessary for low-grade INS, but for navigation grade, this is not necessary, but does cause the INS to be dependent on the VBN/TBN. The primary strength to this approach is the independence of solutions, thus faults in either do not affect the other. Another advantage is the simple implementation, though time synchronization can still be a challenge. When a consumer grade INS is used, then the INS simply supplies attitude information and increased bandwidth to the navigation system.

A typical **tight integration**, also known as **range-domain integration** or tightly-coupled, blends the sensor pseudomeasurements together instead of the navigation solution. The algorithms for these systems are sensor-to-sensor fusion algorithms with the sensors' pseudomeasurements. The primary variant to this scheme for GNSS is the inclusion or exclusion of **tracking loop aiding** which feeds the blending filter's position and/or velocity estimate back to the GNSS receiver to enhance its performance. This is often necessary for aircraft in high-dynamic maneuvers to be able to continuously track the GNSS signals.

However, this integration adds complexity to the GNSS tracking loops and creates a dependency between the systems which does not allow for separate faults to be isolated. The primary advantage is that tight integration provides a more accurate and robust solution than loose integration. The enhanced accuracy is, in part, due to the fact that the GNSS measurements exhibit less temporal correlation than the INS navigation solution. It is robust because it can use the GNSS measurements when there are less than four satellites, though this will only be possible for a limited amount of time due to GNSS receiver clock drift.

A diagram of a typical **deep GNSS/INS integration** a.k.a. deeply-coupled, also known as **tracking-domain integration**. The algorithms for these systems are also sensor-to-sensor fusion algorithms with the sensors' raw measurements. The primary variant to this scheme is coherent or noncoherent. This scheme is an optimal fusion of the INS and GNSS measurements as it utilizes the IMU measurements in a **vector delay lock loop (VDLL)**, enhancing the robustness of the GNSS receiver to interference and jamming even better than tight integration. However, this design introduces enhanced complexity to integrating the GNSS and INS with the necessity of access to the receiver hardware or a software-defined receiver (SDR) for the signal processing.

As the 3-DOF attitude dynamics in navigation are highly nonlinear integrations of Euler angles, DCMs, or quaternions, one often uses **error-state filtering** as the multi-sensor data fusion algorithm. For GNSS/INS integration, error-state filtering uses a time update step for the INS and measurement update step for estimating the error-state of the INS using the GNSS measurements. Then, the additional state update step corrects the nominal-state INS estimate with the GNSS-based error-state estimate. In navigation, the *de facto* standard algorithm for data fusion is the error-state EKF although the UKF and the PF have been used in some cases. This section will describe the three update steps for and ES-EKF implementation of a loose GNSS/INS integration system.

### ES-EKF for Loose GNSS/INS Integration

For the GNSS/INS integration, the error-state dynamics model can vary greatly based on IMU grade, the GNSS update accuracy and frequency, and the overall vehicle mission. For simplicity, the ES-EKF here will only use the accelerometer and gyroscope biases which assumes one has sufficiently calibrated any axes misalignment and scale factor errors. Thus, the loose GNSS/INS integration nominal-state estimated by the INS is

$$\bar{x} = [\bar{p}^T \quad \bar{v}_N^T \quad \bar{\psi}_{N \leftarrow B}^T \quad \bar{b}_a^T \quad \bar{b}_g^T]^T \quad (21.39)$$

where  $\bar{p} = [\bar{\ell} \quad \bar{\lambda} \quad \bar{h}]^T$  is the LLA position vector,  $\bar{v}_N$  is the NED velocity vector,  $\bar{\psi}_{N \leftarrow B} = [\bar{\phi} \quad \bar{\theta} \quad \bar{\psi}]^T$  is the Euler angle vector,  $\bar{b}_a$  is the accelerometer bias vector, and  $\bar{b}_g$  is the gyroscope bias vector. Likewise, the loose GNSS/INS integration error-state vector is

$$\delta \vec{x} = [\delta \vec{p}^T \quad \delta \vec{v}_N^T \quad \delta \vec{\psi}_{N \leftarrow B}^T \quad \delta \vec{b}_a^T \quad \delta \vec{b}_g^T]^T \quad (21.40)$$

where  $\delta \vec{p} = [\delta x \quad \delta y \quad \delta z]^T$  is the NED position error-state,  $\delta \vec{v}_N$  is the NED velocity error-state,  $\delta \vec{\psi}_{N \leftarrow B}$  is the Euler angle error-state,  $\delta \vec{b}_a$  is the accelerometer bias error-state, and  $\delta \vec{b}_g$  is the gyroscope bias error. Notably, in a tight GNSS/INS integration utilizing pseudorange and pseudorange rate, one would need to include the clock offset and clock drift error-states,  $\delta t$  and  $\delta \dot{t}$ .

For the ES-EKF state update step, one must update the nominal-state INS estimates. For the LLA position update, one has

$$\hat{\vec{p}} = \begin{bmatrix} \hat{\ell} \\ \hat{\lambda} \\ \hat{h} \end{bmatrix} = \begin{bmatrix} \hat{\ell} \\ \hat{\lambda} \\ \hat{h} \end{bmatrix} + \begin{bmatrix} \frac{\delta\hat{x}}{R_N(\hat{\ell})+\hat{h}} \\ \frac{\delta\hat{y}}{(R_E(\hat{\ell})+\hat{h})\cos\hat{\ell}} \\ -\delta\hat{z} \end{bmatrix} \quad (21.41)$$

where  $R_N(\hat{\ell})$  is the **north-south radius of curvature** and  $R_E(\hat{\ell})$  is **east-west radius of curvature** of the reference ellipsoid. For the NED velocity update, one has

$$\hat{\vec{v}} = \hat{\vec{v}} + \delta\hat{\vec{v}} \quad (21.42)$$

However, for low-grade IMUs, often one relies only on GNSS-based estimates of  $\hat{h}$  and  $\hat{z}$  updates, due to the vertical channel instability problem.

For the accelerometer bias update, one has

$$\hat{\vec{b}}_a = \hat{\vec{b}}_a + \delta\hat{\vec{b}}_a \quad (21.43)$$

and for the gyroscope bias update, one has

$$\hat{\vec{b}}_g = \hat{\vec{b}}_g + \delta\hat{\vec{b}}_g \quad (21.44)$$

It should be noted that the accelerometer and gyroscope bias estimates are added to the raw INS measurements before use in the INS equations.

For the Euler angle attitude update, one must first correct the DCM using the error-state estimate. Using a first-order approximation for the DCM error, one has

$$C_{N \leftarrow B} = \left( I_{3 \times 3} - \left[ \delta\hat{\vec{\psi}}_{N \leftarrow B} \right]_{\times} \right) C_{N \leftarrow B} \quad (21.45)$$

Then, the updated Euler angles can be extracted from  $C_{N \leftarrow B}$  by

$$\hat{\vec{\psi}}_{N \leftarrow B} = \begin{bmatrix} \hat{\phi} \\ \hat{\theta} \\ \hat{\psi} \end{bmatrix} = \begin{bmatrix} \tan^{-1} \frac{C_{32}}{C_{33}} \\ -\sin^{-1} C_{31} \\ \tan^{-1} \frac{C_{21}}{C_{11}} \end{bmatrix} \quad (21.46)$$

where  $C_{ij}$  is the element in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column of  $C_{N \leftarrow B}$ .

For the ES-EKF time update step, the accelerometer's specific force measurement *error* can be modeled as

$$\delta\vec{f}_B = \delta\vec{b}_a + \vec{w}_a \quad (21.47)$$

and the gyroscope's angular rate measurement *error* can be modeled as

$$\delta\vec{\omega}_{B,I \leftarrow B} = \delta\vec{b}_g + \vec{w}_g \quad (21.48)$$

where  $\vec{w}_a$  and  $\vec{w}_g$  are additive zero-mean IMU noise processes with uncorrelated power spectral densities (PSD) of  $\sigma_{w_a}^2$  and  $\sigma_{w_g}^2$ , respectively. Notably, these IMU noise processes don't affect the biases, but do affect the Euler angle, velocity, and position estimates due to the INS integrations of the IMU measurements.

For low-grade IMUs, one typically assumes these biases evolve in time according to first-order Gauss-Markov processes, i.e.,

$$\dot{\vec{b}}_a = -\frac{1}{\tau_a} \delta \vec{b}_a + I_{3 \times 3} \vec{\mu}_a \quad (21.49)$$

and

$$\dot{\vec{b}}_g = -\frac{1}{\tau_g} \delta \vec{b}_g + I_{3 \times 3} \vec{\mu}_g \quad (21.50)$$

where  $\mu_a$  and  $\mu_g$  are the zero-mean correlated Markov bias errors with PSDs as

$$\sigma_{\mu_a}^2 = \frac{2\sigma_{b_a}^2}{\tau_a} \quad (21.51)$$

and

$$\sigma_{\mu_g}^2 = \frac{2\sigma_{b_g}^2}{\tau_g} \quad (21.52)$$

Thus, the stacked IMU noise vector contains the IMU output noise and first-order Markov noise as

$$\vec{w} = \begin{bmatrix} \vec{w}_a \\ \vec{w}_g \\ \vec{\mu}_a \\ \vec{\mu}_g \end{bmatrix} \quad (21.53)$$

with PSD

$$S_\omega = \begin{bmatrix} \sigma_{w_a}^2 I_{3 \times 3} & 0 & 0 & 0 \\ 0 & \sigma_{w_g}^2 I_{3 \times 3} & 0 & 0 \\ 0 & 0 & \sigma_{\mu_a}^2 I_{3 \times 3} & 0 \\ 0 & 0 & 0 & \sigma_{\mu_g}^2 I_{3 \times 3} \end{bmatrix} \quad (21.54)$$

Taken together, the linearized error-state dynamics equation is given by

$$\dot{\vec{x}}(t) = A(t) \delta \vec{x}(t) + L(t) \vec{w}(t) \quad (21.55)$$

where the time-varying state matrix,  $A$ , is approximated from the linearized INS error equations and bias models as

$$A(t) = \begin{bmatrix} -[\vec{\omega}_{N,E \leftarrow N}]_\times & I_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} \\ \frac{\|\vec{g}\|_2}{a} \text{diag}(-1, -1, 2) & -[2\vec{\omega}_{N,I \leftarrow E} + \vec{\omega}_{N,E \leftarrow N}]_\times & [\vec{C}_{N \leftarrow B} \vec{f}_B]_\times & \vec{C}_{N \leftarrow B} & 0_{3 \times 3} \\ 0_{3 \times 3} & 0_{3 \times 3} & -[\vec{\omega}_{N,I \leftarrow N}]_\times & 0_{3 \times 3} & -\vec{C}_{N \leftarrow B} \\ 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & -\frac{1}{\tau_a} I_{3 \times 3} & 0_{3 \times 3} \\ 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & -\frac{1}{\tau_g} I_{3 \times 3} \end{bmatrix} \quad (21.56)$$

where  $\|\vec{g}_N\|$  is the local gravity magnitude and depends on latitude and altitude and  $a$  is the semi-major axis of the Earth, a.k.a. equatorial radius. Furthermore, the time-varying process noise gain matrix  $L$  is

approximated from the INS noise models transformed to the correct frame as

$$L(t) = \begin{bmatrix} 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} \\ C_{N \leftarrow B} & 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} \\ 0_{3 \times 3} & -C_{N \leftarrow B} & 0_{3 \times 3} & 0_{3 \times 3} \\ 0_{3 \times 3} & 0_{3 \times 3} & I_{3 \times 3} & 0_{3 \times 3} \\ 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & I_{3 \times 3} \end{bmatrix} \quad (21.57)$$

For the ES-EKF measurement update step, one must compute the innovation between the expected measurement based on the INS navigation state and the actual GNSS measurements, i.e.

$$\delta \vec{y}_k = \vec{y}_k - \hat{\vec{y}}_k \quad (21.58)$$

For the loose GNSS/INS integration ES-EKF, the GNSS measurements are the position vector,  $\vec{p}$ , and velocity,  $\vec{v}$ , which are typically in the LLA frame and the NED frame, respectively. However, as the position error is in NED as that typically is how the measurement error covariance is defined, one is required to transform the nominal-state and GNSS position measurements to the *same* NED navigation frame with the INS nominal-state as the  $[0 \ 0 \ 0]^T$  reference point of the NED frame. Then, the expected INS measurement model can relate the GNSS measurement to the nominal-state estimate in NED coordinates as

$$\hat{\vec{y}}_k = H_k \hat{x}_k \quad (21.59)$$

and the linearized GNSS measurement model is

$$\vec{y}_k = H_k \vec{x}_k \quad (21.60)$$

where

$$H_k = [I_{6 \times 6} \ 0_{6 \times 9}] \quad (21.61)$$

The GNSS measurement noise covariance is typically modeled as

$$R_k = \begin{bmatrix} P_{p,k} & 0 \\ 0 & P_{v,k} \end{bmatrix} \quad (21.62)$$

where  $P_{p,k}$  and  $P_{v,k}$  are the estimated position and velocity covariance matrix from the GNSS state estimator, respectively. However, an important consideration is that GNSS noise is truly auto-correlated, thus violating the Markov process assumptions for the this Bayes filtering method. Thus, in theory, one should model and estimate this auto-correlation in the filter. However, in practice, one inflates the covariance matrix to compensate for this. This is sub-optimal, but with low-grade INS, one does not necessarily desire to drastically improve the GNSS positioning, but to simply directly estimate attitude and have a higher bandwidth for the vehicle state estimate provided by the overall navigation system.

## GNSS/INS Integration Implementation Considerations

When implementing a GNSS/INS integrated system, an important consideration is the spatial offset between the IMU and the GNSS antenna which should be compensated for when fusing the information from the two sensors together since they are referring to two different points of the vehicle. This spatial offset is typically

called the **lever arm**, parameterized in the body frame, and denoted by  $\vec{l}_B$ . The position and velocity of the GNSS and INS in the navigation frame can be related by the translation equations

$$\vec{p}_{N,GNSS} = \vec{p}_{N,INS} + C_{N \leftarrow B} \vec{l}_B \quad (21.63)$$

and

$$\vec{v}_{N,GNSS} = \vec{v}_{N,INS} + C_{N \leftarrow B} \left( [\vec{\omega}_{B,I \leftarrow B}]_x + [\vec{\omega}_{B,I \leftarrow N}]_x \right) \vec{l}_B \quad (21.64)$$

where the accuracy of the translation depends on the accuracy of the lever arm as well as the gyroscope accuracy for the velocity. In a loose GNSS/INS integration, one typically translates the GNSS position and velocity to the IMU before the error measurement update step, while for a tight GNSS/INS integration the INS position and velocity are typically translated to the GNSS antenna before fusing the error measurement update step.

A second related aspect of fusing two different sensors is **time synchronization** which must be handled by the dedicated hardware or data collection systems whose design is beyond the scope of this textbook. However, one should be aware that a timing error  $\delta t$  will affect the velocity and position according to Newtonian physics in the presence of an acceleration, e.g.

$$\delta v = a\delta t \quad (21.65)$$

$$\delta p = \frac{1}{2}a\delta t^2 \quad (21.66)$$

Thus, a timing error of 1 ms and an acceleration of 10 m/s<sup>2</sup>, the position and velocity error would be  $\ll 1$  mm and 1 cm/s, respectively, which shows that the position error is negligible, though the velocity error may be problematic for higher precision navigation systems. Lastly, when implementing the covariance updates in an EKF, one should force the covariance to be symmetric.

### Multi-State Constraint Kalman Filter

The **multi-state constraint Kalman filter (MSCKF)** is an EKF-based algorithm for an vision-aided inertial navigation system. The MSCKF uses a measurement model that considers the geometric constraints for features tracked through multiple camera frames without requiring the feature position in the state vector of the EKF as in a full SLAM approach, but still is linearly optimal for the navigation state estimate.

### References

For more information, please refer to the following

- S. Gleason and D. Gebre-Egziabher, “GNSS Applications and Methods,” Artech House, 2012
- C. Jekeli, “Inertial Navigation Systems with Geodetic Applications,” de Gruyter, 2001
- A. I. Mourikis and S. I. Roumeliotis, “A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation,” in *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, Rome, Italy, 2007

## 21.4 Simultaneous Localization and Mapping Systems

**EKF-SLAM**

**FastSLAM**

**References**

For more information, please refer to the following

- M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte and M. Csorba, “A solution to the simultaneous localization and map building (SLAM) problem,” in *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 229-241, June 2001
- M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, “FastSLAM: A factored solution to the simultaneous localization and mapping problem,” in *Proceedings of the AAAI National Conference on Artificial Intelligence*, Edmonton, Canada, 2002

---

# Object Tracking Systems

## 22.1 Introduction to Object Tracking Systems

An **object tracking system** is similar to a navigation system, in that one is estimating an object's state, but also includes detection and identification of the object(s). Specifically, **navigation** is the self-state estimation of the vehicle's position, velocity, attitude, i.e. the **navigation state**, while **object tracking** is the other-state estimation of an object's position and potentially its velocity and/or attitude, i.e., the **object state**. Thus, tracking systems fundamentally use similar sensors and signal/image processing as navigation to form **pseudomeasurements**, i.e., **pseudoranges**, **pseudobearings**, and **pseudorange rates**. For aerospace perception systems, these objects may be obstacles, features, or **targets** relating to the vehicle's mission. Thus, object tracking systems are also known as **target tracking systems** when the information is used by the planning and/or guidance systems of the aerospace vehicle with regards to the trajectory planning and execution. Another name for these systems is **surveillance system** which detects and tracks any "objects of interest" in their surveillance volume and may be considered static, ground-based, airborne, or spaceborne surveillance. Object tracking can be divided into static and dynamic objects as well as passive dynamic or **maneuvering objects**, i.e., actively controlled objects. In some instances, static and passive dynamic objects are also known as **features** and tracking these features is known as **feature tracking**.

### Object Tracking Challenges

There are two fundamental challenges in object tracking that drive the different approaches in tracking system design. The first challenge is that the tracking system must first detect the object(s) to be tracked. Thus, the object detection algorithms will exhibit missed detections and false alarms, known as **clutter** in object tracking, also known as the **measurement origin uncertainty**. Furthermore, one may not know the number of objects to track which may dynamically increase, known as **object birth** and **object spawning** in object tracking, or decrease, also known as **object death** in object tracking. This challenge must be accounted for in object tracking algorithms and lead to the different approaches. This introductory chapter of object tracking

will consider object tracking methods when the number of objects is known and constant. The following advanced chapter will consider object tracking methods for an unknown and time-varying number of objects.

**Clutter** refers to extracted measurements from the signals/images that are returned from “uninteresting objects” to the object tracking system. Such objects include natural objects such as ground, sea, precipitation, sand storms, birds, atmospheric turbulence, ionosphere reflections, meteor trails, and buildings. Clutter is detected and neutralized in several ways. Clutter tends to appear static between images; on subsequent scans, objects of interest will appear to move, and all stationary returns can be eliminated. Clutter can also often be reduced by incorporating a ground map of the sensor’s surroundings and eliminating all objects which appear to originate below ground or above a certain height, e.g., time gating. For radar, clutter may originate from multi-path echoes from valid objects caused by ground reflection, atmospheric ducting, or ionospheric reflection/refraction. This clutter type is especially bothersome since it appears to move and behave like other normal point objects of interest.

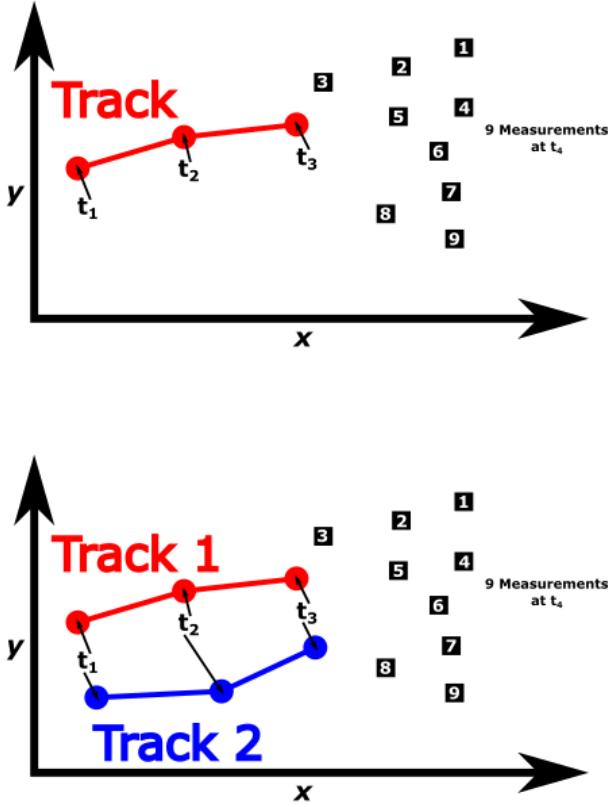
An extension of this challenge are the use of **superpositional sensors** where each object can contribute to any number of measurements, also known as **split measurements**; each measurement is potentially affected by multiple objects in an additive fashion, also known as **merged measurements**; and measurements are not independent. Other advanced topics beyond the scope of this textbook as current areas of active research include **multi-sensor**, **multi-object tracking** including distributed multi-object tracking (MOT) with **sensor networks** of heterogeneous and geographically dispersed nodes with sensing, communication, and processing capabilities as well as **multi-tracker fusion** where one desires to fuse the estimated objects’ states from multiple tracking systems.

The second challenge is that the surveillance system can only use remote sensing to estimate the object state. Thus, the control inputs of controlled objects and/or inertial sensors cannot be used as part of the time update in the traditional Bayes filtering approach used in navigation, also known as the **object motion uncertainty**. Thus, object tracking depends on general motion models for the objects one desires to track. Thus, this chapter of the textbook will discuss motion models for **point object tracking**, i.e., the object state contains dynamic characteristics and not spatial characteristics and only single detections per object are possible. In general, point object motion models can be distinguished between the amount of coupling assumed between the coordinates of the vehicle’s motion which will be discussed in this introductory chapter. This is opposed to **extended object tracking** where the object state contains spatial characteristics or **group object tracking** where multiple objects move according to some common group dynamics where the objects within a single group can be *resolvable* or *unresolvable* by the sensor measurements. Unresolvable group and extended objects are considered to be the same type of tracking problem.

## Measurement-Origin Uncertainty

The optimal state estimator is the Bayes filter which for linear, Gaussian stochastic systems is the Kalman filter. However, as any object detection procedure will exhibit absent and cluttered measurements of the  $N$  object(s), object tracking must account for object-originated measurements and clutter-originated measurements, i.e., the **measurement-origin uncertainty problem**, also known as the **data association problem**, when performing the **measurement update step** or **correction step** within the Bayes filter context.

This measurement origin uncertainty can be demonstrated using the following graphics. Here, at some time  $k = t_4$ , one is tracking  $N = 1$  or  $2$  objects and receives  $M = 9$  measurements, i.e., **sensor data**, which contain a combination of clutter and possibly the object-originated measurements.



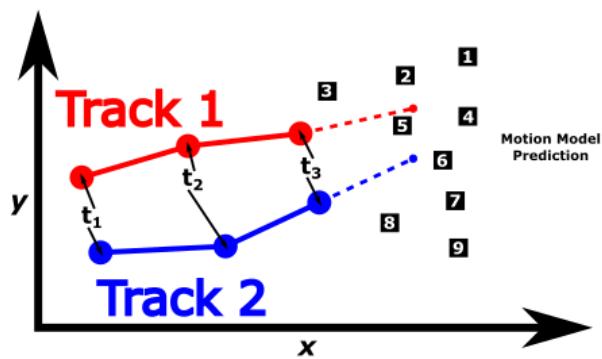
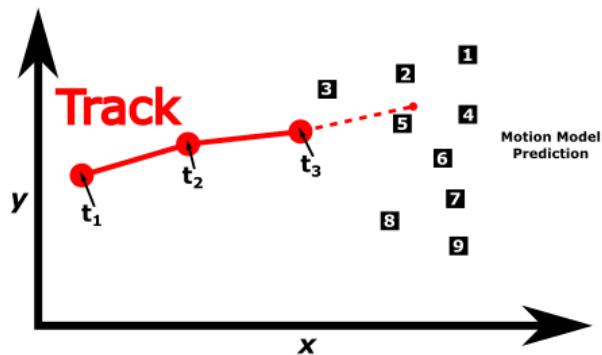
In the Bayes filter context, one can perform the prediction step of the object state estimator, i.e. obtain some  $\hat{\vec{x}}_{k|k-1}$  by using a chosen process equation

$$\hat{\vec{x}}_{k|k-1} = f \left( \hat{\vec{x}}_{k-1|k-1}, \hat{\vec{w}}_{k|k-1} \right) \quad (22.1)$$

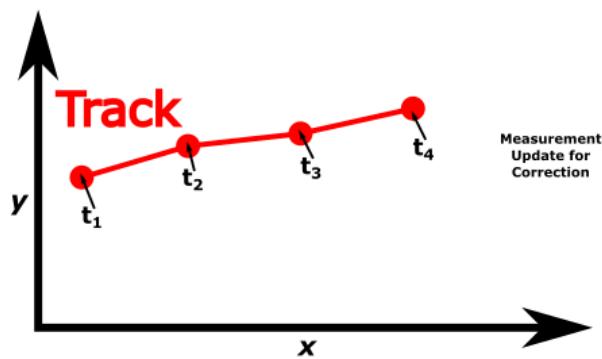
which could be any of the object motion models discussed in previous sections. This will notably also correspond to some predicted measurement,  $\hat{\vec{y}}_{k|k-1}$ , through the measurement equation, i.e.

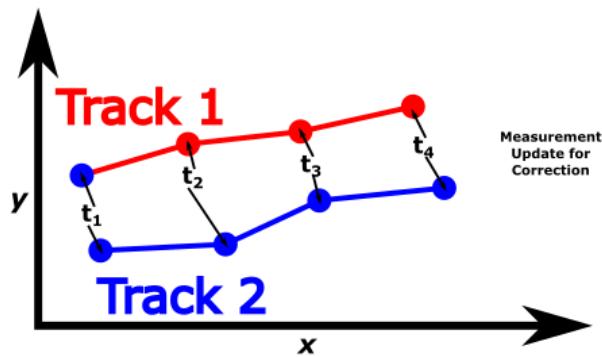
$$\hat{\vec{y}}_{k|k-1} = h \left( \hat{\vec{x}}_{k|k-1}, \hat{\vec{v}}_k \right) \quad (22.2)$$

which could be any of the pseudomeasurements for object tracking. This step can be represented by the small circle in the following figure.



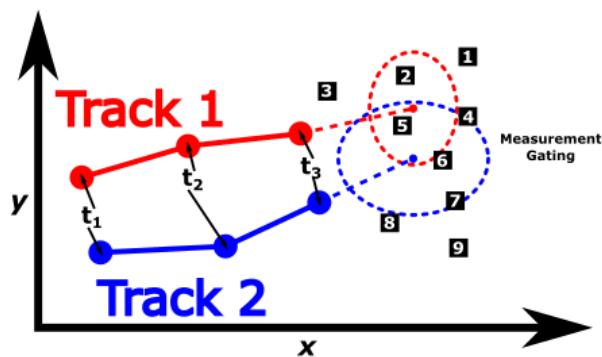
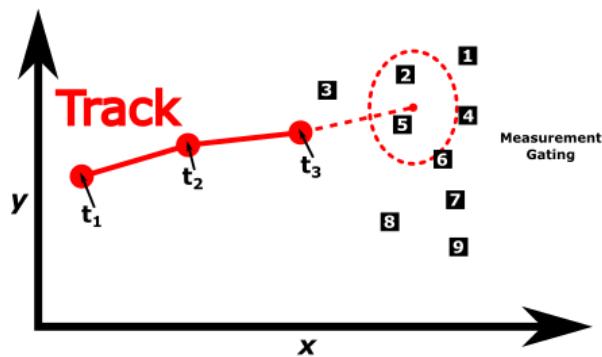
At this point, one must perform some **track-to-measurement association**, also known as **track-to-sensor data association** in order to generate a Bayes-like **correction step**, also known as the **measurement update**, on the predicted object state, i.e. obtain some  $\hat{x}_{k|k}$  regardless of the choice of data association method. This can be represented by the big circle in the following figure.





### Measurement Gating

Theoretically, to account for all possible measurements as object-originated, one should consider all possible associations between the predicted track and the sensor data, i.e. the **candidate measurement set**. However, if significant clutter is present, one often performs a **gating step** which reduces the candidate measurement set to contain only those within some partition of the measurement space,  $\mathbb{R}^{n_y}$  before performing the data association. This partitioning procedure typically depends on using some statistical characteristics relating the measurements to the predicted measurement. This concept is represented in the following figures.



where the gate is the ellipse centered on the predicted state estimate. The inside of the ellipse is the partition of measurements which should be considered as possibly object-originating, i.e. some  $M'$  subset of the  $M$  measurements  $\vec{y}_k \in \mathbb{R}^{n_y}$ .

When designing the gate, one typically specifies the **gate coverage**,  $1 - \alpha$ , defined as the probability of including a object-originated measurement, i.e.

$$\Pr(\vec{y}_k \in \mathcal{G}) \geq 1 - \alpha \quad (22.3)$$

Then, assuming that the innovation is Gaussian distributed, i.e.  $\vec{y}_k - \hat{\vec{y}}_{k|k-1} \sim \mathcal{N}(0, S)$ , two common gates used are rectangular gates and ellipsoidal gates.

**Rectangular gates (RG)** partition measurements inside a hyper-rectangular region, i.e.

$$\mathcal{G}_R = \left\{ \vec{y}_k \in \mathbb{R}^{n_y} : |\vec{y}_k - \hat{\vec{y}}_{k|k-1}|_j \leq F_N^{-1} \left( 1 - \frac{\alpha}{n_y} \right) S_{jj}^{0.5}, \quad j = 1, \dots, n_y \right\} \quad (22.4)$$

where  $|\bullet|_j$  is the  $j^{\text{th}}$  element of  $|\bullet|$ ,  $F_N$  is the standard normal distribution and  $F_N^{-1} \left( 1 - \frac{\alpha}{n_y} \right) S_{jj}^{0.5}$  is the **rectangular gating threshold**. Then, the gate coverage can be shown to be bounded by noting that

$$\Pr(\vec{y}_k \in \mathcal{G}_R) = \Pr \left( \bigcup_{j=1}^{n_y} \left( |\vec{y}_k - \hat{\vec{y}}_{k|k-1}|_j \leq F_N^{-1} \left( 1 - \frac{\alpha}{n_y} \right) S_{jj}^{0.5} \right) \right) \quad (22.5)$$

which by Bonferroni's inequality, one has

$$\Pr(\vec{y}_k \in \mathcal{G}_R) \geq 1 - \sum_{j=1}^{n_y} \Pr \left( |\vec{y}_k - \hat{\vec{y}}_{k|k-1}|_j > F_N^{-1} \left( 1 - \frac{\alpha}{n_y} \right) S_{jj}^{0.5} \right) \quad (22.6)$$

and simplifying, one has

$$\Pr(\vec{y}_k \in \mathcal{G}_R) \geq 1 - \frac{n_y \alpha}{n_y} \quad (22.7)$$

or

$$\Pr(\vec{y}_k \in \mathcal{G}_R) \geq 1 - \alpha \quad (22.8)$$

**Ellipsoidal gates (EG)** ignore any measurements outside the ellipsoidal region, i.e.

$$\mathcal{G}_E = \left\{ \vec{y}_k \in \mathbb{R}^{n_y} : (\vec{y}_k - \hat{\vec{y}}_{k|k-1})^T S^{-1} (\vec{y}_k - \hat{\vec{y}}_{k|k-1}) \leq F_{\chi(n_y)}^{-1} (1 - \alpha) \right\} \quad (22.9)$$

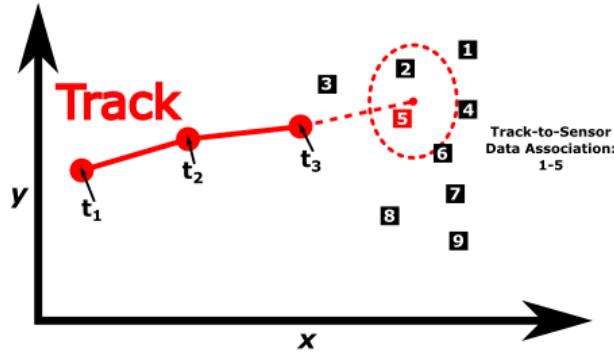
where  $F_{\chi(n_y)}$  is the centralized  $\chi^2$ -distribution with  $n_y$  degrees-of-freedom and  $F_{\chi(n_y)}^{-1} (1 - \alpha)$  is the **ellipsoidal gating threshold**. Furthermore, by the eigendecomposition of  $S$  as  $V^T \Lambda V$  and defining  $\vec{y}_k - \hat{\vec{y}}_{k|k-1}$  as  $\vec{r}_k$  and  $V \vec{r}_k$  as  $\tilde{r}_k$ , one has

$$\vec{r}_k^T V^T \Lambda^{-1} V \vec{r}_k = \tilde{r}_k^T \Lambda^{-1} \tilde{r}_k = \sum_{i=1}^{n_y} \frac{\text{num}}{\text{den}} \sim \chi^2(n_y) \quad (22.10)$$

which proves the gate coverage is exactly  $1 - \alpha$ . Thus, for a given  $\alpha$  and a truly Gaussian distribution for the innovation, the ellipsoidal gate is optimal in the minimal **gate volume** sense and thus is used most commonly. However, in practical tracking systems, one typically enlarges this ellipsoidal gate based on possible object maneuvers that may not be captured by the motion model prediction, also known as a **two-stage gating step**. Other types of gates are also possible if additional data and/or simulations are available for the tracking system design.

### Single-Object Data Association Methods

For single-object data association, the simplest method would be to use only the **nearest-neighbor (NN) association** which selects a *single* measurement with the smallest “distance” to the predicted measurement, i.e.  $\hat{y}_{k|k-1}$ . This concept is represented in the following figure.



Here, a statistical distance measure is typically used, either with the mean distance, i.e. the Euclidean distance

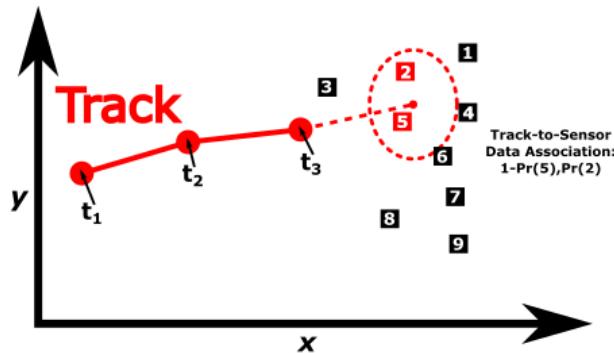
$$\vec{y}_k = \underset{\vec{y}_k \in \mathbb{R}^{n_y}}{\operatorname{argmin}} \|\vec{y}_k - \hat{y}_{k|k-1}\|_2^2 \quad (22.11)$$

or the covariance-weighted distance, i.e. the **Mahalanobis distance**

$$\vec{y}_k = \underset{\vec{y}_k \in \mathbb{R}^{n_y}}{\operatorname{argmin}} \left( \vec{y}_k - \hat{y}_{k|k-1} \right)^T S^{-1} \left( \vec{y}_k - \hat{y}_{k|k-1} \right) \quad (22.12)$$

In addition, one could compute additional “affinity scores” for each measurement based on expected characteristics, e.g. size, shape, appearance. Regardless, NN association allows one to use a standard Bayes filter correction step with the selected measurement.

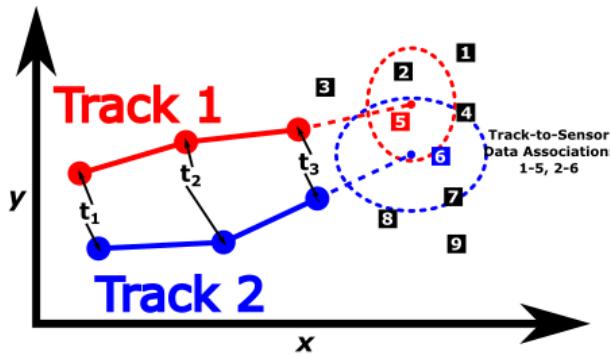
A second data association method would be to use the **all-neighbors (AN) association**, also known as the **probabilistic data association (PDA)**, which uses *all* measurements within the gate, but weights each probabilistically in the correction step of the resulting filter. This concept is represented in the following figure as



However, to weigh each measurement probabilistically, one must know have a statistical model for the probability of missed detections and clutter, as well as the process and measurement models.

### Multi-Object Data Association Methods

For multi-object tracking, the simplest method is **global nearest-neighbor (GNN) association** which assigns a *single* measurement to *each* predicted measurement, i.e. associates a single set of  $\hat{y}_{i,k|k-1}$  for  $i = 1, \dots, N$  to a *single*  $\vec{y}_{j,k|k-1}$  for  $j = 1, \dots, M'$ , based on the smallest sum of “distances”. This concept is represented in the following figure.

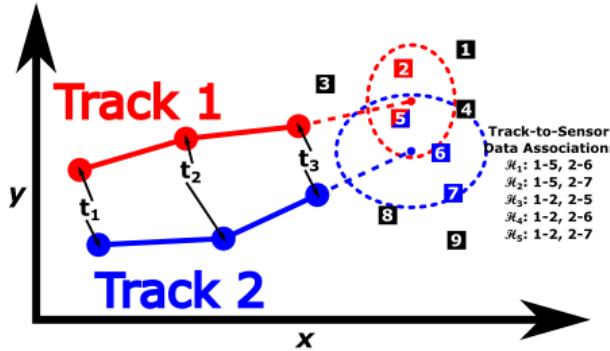


This **assignment problem** seeks to minimize the overall distance, i.e. cost, of each assignment of the  $N$  tracks to the gated  $M'$  measurements which can be efficiently solved in polynomial time using the **Hungarian algorithm** which is outlined as follows for  $N < M'$ .

1. Form the  $M' \times M'$  cost matrix:
  - $ij$  element for  $i < N$ : distance between  $i^{\text{th}}$  track and  $j^{\text{th}}$  measurement
  - Pad with small random numbers for  $M' - N \times M'$  sub-matrix
2. Subtract the minimal cost from each row
  - If one can form a permutation matrix out of 0 elements in  $N \times M'$  submatrix, stop
3. Subtract the minimal cost from each column
  - If one can form a permutation matrix out of 0 elements in  $N \times M'$  submatrix, stop
4. Draw as few row or column lines as possible to cover all 0's
5. From the elements remaining, subtract lowest value from all elements not struck and add to elements at the intersection of row/column lines
  - If one can form a permutation matrix out of 0 elements in  $N \times M'$  submatrix, stop
  - If not go back to step four until finished

A second, closely related, data association method would be to use the  $k$  **nearest-neighbors ( $k$ -NN) association**, which uses  $k$  combinations of measurement to *each* predicted measurement, i.e. associates  $k$

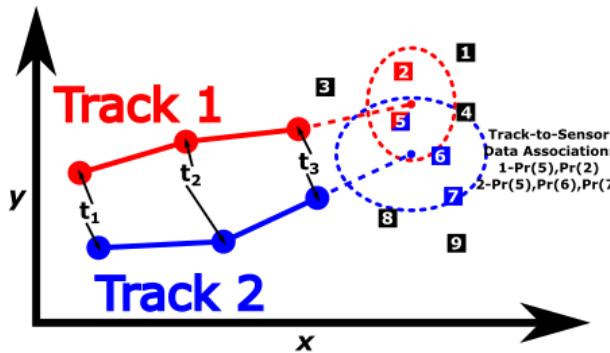
sets of  $\hat{y}_{i,k|k-1}$  for  $i = 1, \dots, N$  to the set of  $\vec{y}_{j,k|k-1}$  for  $j = 1, \dots, M'$ , based on the  $k$  smallest sums of “distances”. This concept is represented in the following figure.



This  **$k$ -best assignment problem** seeks to minimize the overall distance, i.e. cost, of each assignment of the  $N$  tracks to the  $M'$  measurements which can be efficiently solved in polynomial time using the **Murty's algorithm** which is outlined as follows for  $N < M$ .

- Start with best assignment through Hungarian algorithm
- Start methodically “tweaking” it by toggling associations in and out of assignment
- Maintain sorted list of best assignments so far
- During each iterative “sweep,” toggle associations in next best assignment
- $k$ -best assignments found in decreasing order, one per sweep

A third data association method would be to use the **global all-neighbors (GAN) data association**, also known as the **joint probabilistic data association (JPDA)**, which uses *all* measurements within the gate of *each* predicted measurement, but weights each probabilistically in the correction step of the resulting filter. This concept is represented in the following figure as



However, to weigh each measurement probabilistically, one must know have a statistical model for the probability of missed detections, clutter, process and measurement models, but also for the missed identification probabilities.

### Optimal Multi-Object Bayes Filter

By analogy with the single-object filtering problem, the multi-object Bayes filtering problem should use two models, one for prediction of the multi-target state and one for the correction of the multi-target state using the measurement set. However, the data association problem must be overcome with additional probabilities to account for the combinatorics of data association within the Bayes filter framework. The **information set** available at the current time step  $k$  is the cumulative set of measurements through the current time step. Furthermore, an **information state** is a function of the available information set that summarizes the past of the system in a probabilistic sense, i.e., the information state can be substituted into the PDF of any present or future state-related statistic conditioned on the past data.

In the context of known- $N$  multi-object tracking, consider the following **discrete-time stochastic state-space system with measurement-origin uncertainty**

$$\begin{aligned}\vec{\mathbf{x}}_k &= \mathbf{f}(\vec{\mathbf{x}}_{k-1}, \vec{\mathbf{w}}_{k-1}) \\ \vec{\mathbf{y}}_k &= \mathbf{h}(\vec{\mathbf{x}}_k, [DA]_k, \vec{\mathbf{v}}_k)\end{aligned}\quad (22.13)$$

where  $\vec{\mathbf{x}} \in \mathbb{R}^{n_x N}$  is the stacked state vector of all known  $N$  objects with dimension  $n_x$ ,  $\vec{\mathbf{w}}_k \in \mathbb{R}^{n_w N}$  is the *known* process noise,  $\vec{\mathbf{y}} \in \mathbb{R}^{n_y M_k}$  is the stacked measurement vector of all  $M_k$  measurements of dimension  $n_y$ ,  $[DA]_k$  is the data association event at  $k$  that specifies which measurement sub-vector originated from which object sub-vector *and* which measurements originated from the *random* clutter, and  $\vec{\mathbf{v}}_k \in \mathbb{R}^{n_v M_k}$  is the *known* measurement noise consisting of the error in the object-originated measurement and the clutter-originated measurements. The process and measurement noise have known PDFs and the *random* number of clutter-originated measurements has a *known* PMF. The initial state vector is assumed to have a known PDF and to be independent of the noises. For this stochastic state-space system, the information set is  $\vec{\mathbf{y}}_{1:k}$  and the information state is  $f_{\vec{\mathbf{X}}|\vec{\mathbf{Y}}}(\vec{\mathbf{x}}_{1:k}|\vec{\mathbf{y}}_{1:k})$ . However, if the process and object-originating measurement noise are white and mutually independent and the object detection and clutter is white, i.e.,  $\vec{\mathbf{x}}_k$  must be a partially observed Markov process, the information state can be reduced to the posterior conditional PDF of the state at the current time step  $k$ , i.e.,  $f_{\vec{\mathbf{X}}|\vec{\mathbf{Y}}}(\vec{\mathbf{x}}_k|\vec{\mathbf{y}}_{1:k})$ , and is sufficient to predict the PDF of every future state.

Thus, for the discrete-time stochastic state-space system with measurement-origin uncertainty, one can use an optimal Bayes filter for this information state provided one can account for the additional data association probabilities. To do so, consider the use of the law of total probability on the data association probabilities to obtain the information state, i.e.,

$$f_{\vec{\mathbf{X}}|\vec{\mathbf{Y}}}(\vec{\mathbf{x}}_k|\vec{\mathbf{y}}_{1:k}) = \sum_{i=1}^{n_{DA,k}} f_{\vec{\mathbf{X}}|\vec{\mathbf{Y}},[DA]}(\vec{\mathbf{x}}_k|\vec{\mathbf{y}}_k, [DA]_{k,i}) \Pr([DA]_{k,i}) \quad (22.14)$$

Then, using Bayes' theorem and the Chapman-Kolmogorov equation, the known- $N$  multi-object Bayes filter

consists of the recursion of

$$\begin{aligned} f_{\vec{\mathbf{x}}|\vec{\mathbf{y}}}(\vec{\mathbf{x}}_k|\vec{\mathbf{y}}_{1:k}) &= \frac{1}{c} \sum_{i=1}^{n_{DA,k}} f_{\vec{\mathbf{x}}|\vec{\mathbf{y}}, [\text{DA}]}(\vec{\mathbf{y}}_k|\vec{\mathbf{x}}_k, [\text{DA}]_{k,i}) \\ &\quad \times \int f_{\vec{\mathbf{x}}|\vec{\mathbf{x}}}(\vec{\mathbf{x}}_k|\vec{\mathbf{x}}_{k-1}) f_{\vec{\mathbf{x}}|\vec{\mathbf{y}}}(\vec{\mathbf{x}}_{k-1}|\vec{\mathbf{y}}_{1:k-1}) d\vec{\mathbf{x}}_{k-1} \\ &\quad \times \Pr([\text{DA}]_{k,i}) \end{aligned} \quad (22.15)$$

where  $c$  is the normalization constant. This recursion shows that the posterior conditional PDF for the optimal known- $N$  multi-object Bayes filter is a probability-weighted sum of PDFs, i.e., a **PDF mixture**, where the mixture weights are given by the probability of the data association events at the current time,  $[\text{DA}]_{k,i}$ ,  $i = 1, \dots, n_{DA,k}$ , where  $n_{DA,k}$  is the number of mutually exclusive and exhaustive data association events at time step  $k$ . Furthermore, it should be pointed out that the number of terms of the mixture at time  $k$  is given by the product

$$n_{DA,1:k} = \prod_{i=1}^k n_{DA,i} \quad (22.16)$$

which amounts to an exponential increase in the number of data association events in time. Secondly, it shows that for the optimal known- $N$  multi-object Bayes filter, if the exact posterior conditional PDF at  $k-1$  is available, then, only the most recent data association event is needed at each time. However, some practical decisions must be made about the form of the state-space system and the clutter process in order to derive tractable equations. For example, if one assumes linear process and measurement models, Gaussian distributions for the process and measurement noise, and uniformly and independently distributed clutter, then the exact posterior conditional PDFs are Gaussian mixtures (GMs) with an exponentially growing number of terms. Thus, approximations must be made with regards to this Gaussian mixture.

Notably, the optimal known- $N$  multi-object Bayes filter in the sense of providing the MMSE can also be obtained using the smoothing property of the expectation operation for the conditional mean of the state. This is obtained by averaging over all possible data association events, i.e.

$$\begin{aligned} \hat{\vec{\mathbf{x}}}_k^{MMSE} &= \mathbb{E}[\vec{\mathbf{x}}_k|\vec{\mathbf{y}}_{1:k}] \\ &= \mathbb{E}[\mathbb{E}[\vec{\mathbf{x}}_k|[\text{DA}]_{1:k}, \vec{\mathbf{y}}_{1:k}]|\vec{\mathbf{y}}_{1:k}] \\ &= \sum_{[\text{DA}]_{1:k,i} \in [\mathcal{DA}]_{1:k}} \mathbb{E}[\vec{\mathbf{x}}_k|[\text{DA}]_{1:k,i}, \vec{\mathbf{y}}_{1:k}] \Pr([\text{DA}]_{1:k,i}|\vec{\mathbf{y}}_{1:k}) \\ &= \sum_{i=1}^{n_{DA,1:k}} \hat{\vec{\mathbf{x}}}_{k,i} \Pr([\text{DA}]_{1:k,i}|\vec{\mathbf{y}}_{1:k}) \end{aligned} \quad (22.17)$$

where  $[\text{DA}]_{1:k,i}$  is a single set of data association events from time step 1 to  $k$ , with a corresponding state estimate at time step  $k$ ,  $\hat{\vec{\mathbf{x}}}_{k,i}$ , and  $[\mathcal{DA}]_{1:k}$  is the set of  $n_{DA,1:k}$  mutually exclusive and exhaustive data association events across time steps 1, ...,  $k$ .

For the general unknown- $N$ , consider a multi-target state  $\mathbf{x}_{k-1}$  at time step  $k-1$ , each  $\vec{\mathbf{x}}_{k-1} \in \mathbf{x}_{k-1}$  either continues to exist at time step  $k$  with **survival probability**,  $p_{s,k}(\vec{\mathbf{x}}_{k-1})$ , or dies with probability,  $1 - p_{s,k}(\vec{\mathbf{x}}_{k-1})$ . Furthermore, conditional on the existence at time step  $k$ , the transition PDF from state  $\vec{\mathbf{x}}_{k-1}$  to  $\vec{\mathbf{x}}_k$  is given by  $f_{k|k-1}(\vec{\mathbf{x}}_k|\vec{\mathbf{x}}_{k-1})$ . Thus, for a given state,  $\vec{\mathbf{x}}_{k-1} \in \mathbf{x}_{k-1}$  at time step  $k-1$ , its behavior at

the  $k$  can be modeled as the Bernoulli **survival RFS**  $\mathbf{S}_{k|k-1}(\vec{x}_{k-1})$  that either takes on  $\{\vec{x}_k\}$  when the target survives or  $\emptyset$  when the target dies. Furthermore, a new target at time step  $k$  can appear either by spawning from another target at time  $k-1$ , represented by the **spawn RFS**,  $\mathbf{G}_{k|k-1}(\vec{x}_{k-1})$ , or a birth independent of any existing target, represented by the **birth RFS**,  $\mathbf{B}_k$ .

Thus, given a multi-target state  $\mathbf{x}_{k-1}$  at time  $k-1$ , the multi-target state  $\mathbf{x}_k$  at time step  $k$  is given by the union of the surviving targets, the spawned targets, and the birthed targets, i.e. the **multi-target process equation**

$$\mathbf{X}_k = \left[ \bigcup_{\vec{x}_k \in \mathbf{x}_{k-1}} \mathbf{S}_{k|k-1}(\vec{x}_k) \right] \cup \mathbf{B}_k \cup \left[ \bigcup_{\vec{x}_k \in \mathbf{x}_{k-1}} \mathbf{G}_{k|k-1}(\vec{x}_k) \right] \quad (22.18)$$

which can be related to a Markov **multi-target transition density**

$$f_{k|k-1}(\mathbf{x}_k | \mathbf{x}_{k-1}) \quad (22.19)$$

which captures the underlying models of target motion, survival, spawns and births.

For the measurement model, let  $\mathcal{Y} = \mathbb{R}^{n_y}$  denote the single-measurement space whose elements are vectors of the form  $\vec{y} = [y_1 \dots y_{n_y}]^T$  where  $y_1, \dots, y_{n_y}$  in the set of real numbers, e.g., pseudorange, pseudobearing, pseudorange rate. Thus, the measurement set of a multi-target system is a finite set of measurement vectors,  $\mathbf{y} = \{\vec{y}_1, \dots, \vec{y}_M\}$  where  $M$  is the random number of measurements and  $\vec{y}_1, \dots, \vec{y}_M$  are the individual measurements. The multi-target measurement space is the class of finite subsets of  $\mathcal{Y}$ , i.e.  $\mathcal{F}(\mathcal{Y})$ . This allows one to define the random measurement set as RFS,  $\mathbf{Y}$ . The multi-target measurement  $\mathbf{y}_k$  at time step  $k$  is given by the union of the detected target-originating measurements, represented by the Bernoulli **detection RFS**,  $\mathbf{D}_k(\mathbf{x}_k)$ , and the clutter-originating measurements, represented by the **clutter RFS**,  $\mathbf{C}_k$ , i.e. the **multi-target measurement equation**

$$\mathbf{Y}_k = \left[ \bigcup_{\vec{x}_k \in \mathbf{x}_k} \mathbf{D}_k(\mathbf{x}_k) \right] \cup \mathbf{C}_k \quad (22.20)$$

which can be related to a **multi-target likelihood function**

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_k | \mathbf{x}_k) \quad (22.21)$$

It should be noted that each of the RFSs here are assumed independent of each other as in other Bayes formulations. Lastly, the actual forms of  $\mathbf{G}_{k|k-1}$ ,  $\mathbf{B}$ , and  $\mathbf{C}_k$  are problem dependent.

Thus, all information about the multi-target state history through time step  $k$  is encapsulated in the **multi-target posterior density**,  $f_{\mathbf{X}_{0:k}}(\mathbf{x} | \mathbf{y}_{1:k})$  which can be recursively computed from an initial *a priori* density, via the multi-target Bayes recursion, i.e.

$$f_{\mathbf{X}_{0:k}}(\mathbf{x} | \mathbf{y}_{1:k}) \propto f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_k | \mathbf{x}_k) f_{k|k-1}(\mathbf{x}_k | \mathbf{x}_{k-1}) f_{\mathbf{X}_{0:k-1}}(\mathbf{x}_{0:k-1} | \mathbf{y}_{1:k}) \quad (22.22)$$

where individual target trajectories can be accommodated by incorporating a label in each target's state vector. For online multi-target tracking, the **multi-target filtering density** can be computed using the Bayes prediction and correction equations, i.e.

$$f_{\mathbf{X}_{k|k-1}}(\mathbf{x}_k | \mathbf{y}_{1:k-1}) = \int f_{k|k-1}(\mathbf{x}_k | \mathbf{x}_{k-1}) f_{\mathbf{X}}(\mathbf{x}_{k-1} | \mathbf{y}_{1:k-1}) \delta \mathbf{x}_{k-1}$$

$$f_{\mathbf{X}}(\mathbf{x}_k | \mathbf{y}_{1:k}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_k | \mathbf{x}_k) f_{\mathbf{X}_{k|k-1}}(\mathbf{x}_k | \mathbf{y}_{1:k-1})}{\int f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_k | \mathbf{x}_k) f_{\mathbf{X}_{k|k-1}}(\mathbf{x}_k | \mathbf{y}_{1:k-1}) \delta \mathbf{x}_k} \quad (22.23)$$

which are the RFS-defined Chapman-Kolmogorov equation and Bayes rule. However, these general equations do not provide an analytical form, particle filters based on the multi-target Bayes recursions have been formulated; albeit, typically only for a few number of target tracks. Four other analytical approximating filters have been developed under this framework, namely, the PHD, the cardinalized PHD, and the generalized labeled multi-Bernoulli filters. Each of these will be discussed in this textbook.

In estimation, one must use RFS models and convert into their respective density functions. That is, one requires a systematic procedure for constructing the true multi-target measurement density  $f_{Y|X}(y|x)$  that faithfully describes the multi-sensor/multi-target measurement model, i.e. the likelihood that a multi-sensor measurement finite set,  $y$  will be collected if targets with finite set  $x$  are present. Likewise, one requires a general, systematic procedure for constructing the true multi-target Markov density  $f_{k|k-1}(x_k|x_{k-1})$  which faithfully describes the multi-target motion model, i.e. the likelihood that the targets will have finite set  $X_k$  at time step  $k$  if they had finite set  $X_{k-1}$  at time step  $k-1$ . However, for the construction of true multi-target measurement and Markov PDFs, one requires the inverse operation of set integrals, i.e. **set derivatives**, which, in turn requires an understanding of belief mass functions. These set integrals, set derivatives, BMFs, PGFLs, and PGFL derivatives are all used to prove the optimality of filters used in RFS-based tracking where it should be pointed out that the set and PGFL derivatives can be computed using rules similar to those of calculus, e.g. chain, product, sum, constant, power rules.

Thus, given the measurement and motion models previously, one can first construct their belief mass functions as

$$\beta_k(\mathcal{S}|x) = \Pr(y_k \subseteq \mathcal{S}) = \int_{\mathcal{S}} f_{Y|X}(y_k|x_k) \delta y_k \quad (22.24)$$

and

$$\beta_{k|k-1}(\mathcal{S}|x) = \Pr(x_{k|k-1} \subseteq \mathcal{S}) = \int_{\mathcal{S}} f_{k|k-1}(x_k|x_{k-1}) \delta x_k \quad (22.25)$$

From these definitions, one can show

$$f_{Y|X}(y_k|x_k) = \frac{\delta \beta_k}{\delta y}(\emptyset|x) \quad (22.26)$$

and

$$f_{k|k-1}(x_k|x_{k-1}) = \frac{\delta^N \beta_{k|k-1}}{\delta x_k}(\emptyset|x_{k-1}) \quad (22.27)$$

where for  $F(x)$  of finite set  $X$ , its **set derivatives** are defined as

$$\frac{\delta F}{\delta \vec{x}}(\mathcal{S}) = \lim_{V(\mathcal{E}_x) \rightarrow 0} \frac{F(\mathcal{S} \cup \mathcal{E}_x) - F(\mathcal{S})}{V(\mathcal{E}_x)} \quad (22.28)$$

and

$$\frac{\delta \beta_{k|k-1}}{\delta x_k}(\mathcal{S}) = \frac{\delta^N \beta}{\delta \vec{x}_N \cdots \delta \vec{x}_1}(\mathcal{S}) = \frac{\delta}{\delta \vec{x}_N} \frac{\delta^{N-1} \beta}{\delta \vec{x}_{N-1} \cdots \delta \vec{x}_1}(\mathcal{S}) \quad (22.29)$$

where  $\mathcal{E}_x$  is a small neighborhood of  $\vec{x}$  and  $V(\mathcal{S})$  is the hyper-volume of set  $\mathcal{S}$ . Likewise, the multi-target posterior PDFs are

$$f_{X|Y}(\mathbf{x}_k|\mathbf{y}_{1:k}) = \frac{\delta \beta_{k|k}}{\delta \mathbf{x}}(\emptyset|\mathbf{y}_{1:k}) \quad (22.30)$$

as constructed by their belief mass functions

$$\beta_{k|k}(\mathcal{S}|\mathbf{y}_{1:k}) = \Pr(\mathbf{x}_{k|k} \subseteq \mathcal{S}) = \int_{\mathcal{S}} f_{\mathbf{X}}(\mathbf{x}_k|\mathbf{y}_{1:k}) \delta \mathbf{x} \quad (22.31)$$

These general models allow for non-homogeneous, non-Poisson clutter, and state-dependent detection probabilities in a principled way, depending on the optimal Bayes filter or approximations to the optimal which often use these set derivatives to prove their approximations.

## Trajectory Set Theory

### References

For more information, please refer to the following

- ...

## 22.2 Probabilistic Data Association Filtering

For the known- $N$  multi-target Bayes filter, two practical sub-optimal estimators based on MMSE optimality and approximate the posterior conditional PDF are the probabilistic data association filter (PDAF) and the joint probabilistic data association filter (JPDAF) which approximate the general posterior conditional PDF by a Gaussian distribution using similar linearization approximations as the EKF. One can also improve the PDAF and JPDAF by using a Gaussian mixture approximation while *limiting* the number of components in the mixture, e.g., multi-hypothesis tracking considered in the following section. Alternatively, one may also use particles to approximate the posterior conditional PDFs which would also exponentially grow in number. In either case, a key design parameter is how to limit this growing number of terms in the PDF mixture or the number of particles.

### Probabilistic Data Association Filter

The **probabilistic data association filter (PDAF)** presented here consists of four steps: prediction, gating, data association, and correction, and has seven key assumptions. First, it assumes that only one object is present, i.e.  $N = 1$ , with state  $\vec{x} \in \mathbb{R}^{n_x}$ , process model

$$\vec{x}_k = f(\vec{x}_{k-1}, \vec{w}_{k-1}) \quad (22.32)$$

and a object-originated measurement,  $\vec{y}_k \in \mathbb{R}^{n_y}$

$$\vec{y}_k = h(\vec{x}_k, \vec{v}_k) \quad (22.33)$$

where  $\vec{w}_{k-1}$  and  $\vec{v}_k$  are zero-mean AWGN with covariances  $Q_{k-1}$  and  $R_k$ , respectively.

Second, the past information through time  $k - 1$  about the object is summarized approximately by a sufficient statistic in the form of the Gaussian **a posteriori** PDF

$$f_{\vec{X}|\vec{Y}}(\vec{x}_k|\vec{y}_{1:k}) = f_N(\hat{\vec{x}}_{k|k}, P_{k|k}) \quad (22.34)$$

Third, it assumes that the track has been initialized, typically through some detection algorithm in the signal processing. Fourth, the object detections occur independently over time with known probability  $P_D$ . Fourth, an ellipsoidal gate is set up at each time step centered on the predicted measurement to select the candidate measurement set for data association to the object. Fifth, if the object was detected and the corresponding measurement fell in the gate, then, at most one of the gated measurements can be object-originated. Sixth, the remaining measurements are assumed to be due to clutter are independent and identically distributed with uniform spatial distribution with the number of clutter points distributed as a **spatial Poisson process** with **clutter spatial density**,  $\lambda_C = \frac{\Lambda_C}{V_\phi}$ , where  $\Lambda_C$  is the **Poisson clutter rate** and  $V_\phi$  is the field-of-view or surveillance volume, or a **diffuse prior**, i.e., all observed measurements equally likely.

The **PDAF prediction step** at time step  $k$  computes the prior information state mean as

$$\hat{\vec{x}}_{k|k-1} = f(\hat{\vec{x}}_{k-1|k-1}, 0) \quad (22.35)$$

with predicted measurement

$$\hat{\vec{y}}_{k|k-1} = h(\hat{\vec{x}}_{k|k-1}, 0) \quad (22.36)$$

Then, one can approximate the prior information state covariance as

$$P_{k|k-1} = F_{k-1} P_{k-1|k-1} F_{k-1}^T + L_{k-1} Q_{k-1} L_{k-1}^T \quad (22.37)$$

where

$$F_{k-1} = \frac{\partial f(\vec{x}, \vec{w})}{\partial \vec{x}} \Big|_{\vec{x}=\hat{\vec{x}}_{k-1|k-1}, \vec{w}=0} \quad (22.38)$$

and

$$L_{k-1} = \frac{\partial f(\vec{x}, \vec{w})}{\partial \vec{w}} \Big|_{\vec{x}=\hat{\vec{x}}_{k-1|k-1}, \vec{w}=0} \quad (22.39)$$

and the innovation covariance as

$$S_k = H_k P_{k|k-1} H_k^T + M_k R_k M_k^T \quad (22.40)$$

where

$$H_k = \frac{\partial h(\vec{x}, \vec{v})}{\partial \vec{x}} \Big|_{\vec{x}=\hat{\vec{x}}_{k|k-1}, \vec{v}=0} \quad (22.41)$$

and

$$M_{k-1} = \frac{\partial h(\vec{x}, \vec{v})}{\partial \vec{v}} \Big|_{\vec{x}=\hat{\vec{x}}_{k|k-1}, \vec{v}=0} \quad (22.42)$$

The **PDAF gating step** uses an ellipsoidal gate as

$$\mathcal{G}_E = \left\{ \vec{y}_k \in \mathbb{R}^{n_y} : \left( \vec{y}_k - \hat{\vec{y}}_{k|k-1} \right)^T S_k^{-1} \left( \vec{y}_k - \hat{\vec{y}}_{k|k-1} \right) \leq F_{\chi^2(n_y)}^{-1} (1 - \alpha) \right\} \quad (22.43)$$

where  $F_{\chi^2(n_y)}^{-1} (1 - \alpha)$  is the **ellipsoidal gating threshold**, and  $P_G = 1 - \alpha$  is the **gate probability**. This gating selects  $M'_k \leq M_k$  measurements as the gated measurement subset,  $\{ \vec{y}_{k,i} \}_{i=1}^{M'_k}$ . It can be shown that the **gate volume**,  $V_{\mathcal{G},k}$ , is given by

$$V_{\mathcal{G},k} = c_{n_y} \left( F_{\chi^2(n_y)}^{-1} (1 - \alpha) \right)^{n_y/2} |S_k|^{0.5} \quad (22.44)$$

and  $c_{n_y}$  is the volume of the  $n_y$ -dimensional hypersphere, e.g.  $c_1 = 1$ ,  $c_2 = \pi$ ,  $c_3 = 4\pi/3$ .

The **PDAF data association step** computes the data association event probabilities using the measurement likelihoods, the clutter spatial density, and the probabilities of detection and gating, i.e.

$$\Pr([DA]_{k,i} | \vec{y}_{1:k}) = \begin{cases} \frac{\mathcal{L}_{k,i}}{1 - P_D P_G + \sum_{j=1}^{M'_k} \mathcal{L}_{k,j}}, & i = 1, \dots, M'_k \\ \frac{1 - P_D P_G}{1 - P_D P_G + \sum_{j=1}^{M'_k} \mathcal{L}_{k,j}}, & i = 0 \end{cases} \quad (22.45)$$

where  $1 - P_D P_G$  is the probability of an object-originating measurement being excluded from the candidate measurement set and the object-originating measurement likelihoods are assumed Gaussian with some probability of detection within the

$$\mathcal{L}_{k,i} = P_D \frac{f_N(\vec{y}_{k,i}; \hat{y}_{k|k-1}, S_k)}{\rho} \quad (22.46)$$

where  $\rho$  is the spatial density of the clutter within the gate. If one is using the spatial Poisson process clutter model with spatial density  $\lambda_C$

$$\rho = \lambda_C \quad (22.47)$$

or if one is using a diffuse prior

$$C = \frac{M'_k}{V_{G,k}} \quad (22.48)$$

which results from the assumption that either  $M'_k$  or  $M'_k - 1$  measurements are clutter-originating.

The **PDAF correction step** forms the filter gain as

$$K_k = P_{k|k-1} H_k^T S_k^{-1} \quad (22.49)$$

the combined probability-weighted averaged innovation is

$$\tilde{r}_k = \sum_{i=1}^{M'_k} \Pr([DA]_{k,i} | \vec{y}_{1:k}) (\vec{y}_{k,i} - \hat{y}_{k|k-1}) \quad (22.50)$$

the posterior information state mean is computed as

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k \tilde{r}_k \quad (22.51)$$

and the posterior information state covariance is computed as

$$\begin{aligned} P_{k|k} = & (1 - \Pr([DA]_{k,0} | \vec{y}_{1:k})) (P_{k|k-1} - K_k S_k K_k^T) + \Pr([DA]_{k,0} | \vec{y}_{1:k}) P_{k|k-1} \\ & + K_k \left( \sum_{i=1}^{M'_k} \Pr([DA]_{k,i} | \vec{y}_{1:k}) (\vec{y}_{k,i} - \hat{y}_{k|k-1}) (\vec{y}_{k,i} - \hat{y}_{k|k-1})^T - \tilde{r}_k \tilde{r}_k^T \right) K_k^T \end{aligned} \quad (22.52)$$

where the first term is the covariance decrease due to a correct measurement update, the second term is the covariance unchanging due to no measurement being correct, and the third term is covariance increase due to the uncertainty in which of the  $M'_k$  measurements is correct.

The **integrated probabilistic data association filter (IPDAF)** provides an augmentation of the PDAF where the probability of the target existence can also be modeled and combined into the filter.

### Joint Probabilistic Data Association Filter

The **joint probabilistic data association filter (JPDAF)** extends the PDAF for  $N > 1$  objects by considering the conditional probabilities of the following joint data association events

$$[DA]_k = \bigcap_{j=1}^M [DA]_{k,j t_j} \quad (22.53)$$

where  $[DA]_{k,j t_j}$  is the data association event at time step  $k$  that measurement  $j$  originated from object  $t$  where  $t_j$  is the index of the object to which measurement  $j$  is associated in the data association event under consideration. The JPDAF has the following five key assumptions. First,  $N$  is known. Second, the measurements from a object can fall in the gate of another object over several time steps. Third, the past information is summarized by approximate conditional means and covariances for each object which allows the data association probabilities to be computed only for the latest set of measurements. Fourth, the states are assumed to be Gaussian distributed. Fifth, each object has a process and measurement model which do not have to be identical.

Just as for the PDAF, the **JPDAF data association step**, has two possible clutter models. The first is the Poisson PMF with clutter rate  $\Lambda_C$  for the number of clutter,  $\phi$ , i.e.

$$p_\phi(\phi) = e^{-\Lambda_C} \frac{(\Lambda_C)^\phi}{\phi!} = e^{-\lambda_C V_\phi} \frac{(\lambda_C V_\phi)^\phi}{\phi!} \quad (22.54)$$

Then, the joint data association probabilities can be modeled as

$$\Pr([DA]_k | \vec{y}_{1:k}) = \frac{1}{c} \prod_j \left( \lambda_C^{-1} f_{t_j}(\vec{y}_{k,j}) \right) \prod_t (P_{D,t})^{\delta_t} (1 - P_D)^{1 - \delta_t} \quad (22.55)$$

where

$$f_{t_j}(\vec{y}_{k,j}) = f_N \left( \vec{y}_{k,j}; \hat{\vec{y}}_{k|k-1, t_j}, S_{k, t_j} \right) \quad (22.56)$$

$\hat{\vec{y}}_{k|k-1}^{t_j}$  is the predicted measurement for object  $t_j$  with associated innovation covariance  $S_k^{t_j}$ ,  $P_{D,t}$  is the detection probability of object  $j$ ,  $c$  is the normalization constant,  $\tau_j$  is the object detection indicator function, i.e.

$$\tau_j = \begin{cases} 1, & \text{if } t_j > 0 \\ 0, & \text{if } t_j = 0 \end{cases} \quad (22.57)$$

which indicates if measurement  $j$  is associated with *any* object, and  $\delta_t$  is the measurement data association indicator function, i.e.

$$\delta_t = \begin{cases} 1, & \text{if } t_j = t \text{ for some } j \\ 0, & \text{if } t_j \neq t \text{ for all } j \end{cases} \quad (22.58)$$

which indicates if *any* measurement is associated with object  $t$ .

The second is the diffuse prior i.e.

$$p_\phi(\phi) = \epsilon \text{ for all } \phi \quad (22.59)$$

Then, the joint data association probabilities can be modeled as

$$\Pr([DA]_k | \vec{y}_{1:k}) = \frac{1}{c} \phi! \prod_j \left( V_\phi f_{t_j}(\vec{y}_{k,j}) \right) \prod_t (P_{D,t})^{\delta_t} (1 - P_{D,t})^{1-\delta_t} \quad (22.60)$$

where  $V_\phi$  is the volume of the region in which measurements not associated with a object are uniformly distributed,  $\phi$  is the number of clutter in event  $[DA]$ , and  $c$  is the normalization constant. Here the term  $\frac{\phi!}{V_\phi} \phi$  substitutes for  $\lambda_C^\phi$ .

If one can assume the object states conditioned on the past measurements are mutually independent, then one requires the marginal data association probabilities which are obtained from the joint probabilities by summing over all joint data association events in which the marginal data association event occurs, i.e.

$$\Pr([DA]_{k,t} | \vec{y}_{1:k}) = \sum_{[DA]: DA_{jt} \in [DA]} \Pr([DA]_k | \vec{y}_{1:k}) \quad (22.61)$$

which allows one to decouple the JPDAF correction step for each object,  $t = 1, \dots, N$ , and are exactly the same as the PDAF.

However, as multiple objects may share measurements in their gates across several time steps, the estimation errors become statistically dependent. This can be taken into account by calculating state cross-covariances by using the  $N$  objects as a single stacked state vector in the correction step of the JPDAF. This additional consideration in the JPDAF results in the **joint probabilistic data association coupled filter (JPDACF)**.

For the **JPDACF data association step**, the conditional probability for a joint data association event is modeled as

$$\Pr([DA]_k | \vec{y}_{1:k}) = \frac{1}{c} \frac{\phi!}{M'_k!} p_C(\phi) V_\phi^{-\phi} f_{t_{j1}, t_{j2}, \dots}(\vec{y}_{k,j}, j : \tau_j = 1) \prod_t (P_{D,t})^{\delta_t} (1 - P_{D,t})^{1-\delta_t} \quad (22.62)$$

where  $f_{t_{j1}, t_{j2}, \dots}$  is the joint PDF of the measurements of the objects under consideration,  $t_{j1}$  is the object to which  $\vec{y}_{k,j1}$  is associated in data association event,  $[DA]_k$ , and  $c$  is the normalization constant. These joint probabilities are used directly in the coupled filter.

Then, for the **JPDACF correction step**, the stacked prior state vector is

$$\hat{\mathbf{x}}_{k|k-1} = \begin{bmatrix} \hat{x}_{k|k-1,1} \\ \vdots \\ \hat{x}_{k|k-1,N} \end{bmatrix} \quad (22.63)$$

The stacked predicted measurement is

$$\hat{\mathbf{y}}_{k|k-1} = \begin{bmatrix} h_1(\hat{x}_{k|k-1,1}, 0) \\ \vdots \\ h_N(\hat{x}_{k|k-1,N}, 0) \end{bmatrix} \quad (22.64)$$

The stacked measurement vector is

$$\vec{\mathbf{y}}_{k,[DA]} = \begin{bmatrix} \vec{y}_{k,j1([DA])} \\ \vdots \\ \vec{y}_{k,jN([DA])} \end{bmatrix} \quad (22.65)$$

where  $j_t([DA])$  is the index of the measurement associated with object  $t$  in data association event  $[DA]_k$ . The stacked combined innovation is

$$\tilde{\mathbf{r}}_k = \sum_{[DA]_k} \Pr([DA]_k | \vec{\mathbf{y}}_{1:k}) \left( \vec{\mathbf{y}}_{k,[DA]} - \hat{\vec{\mathbf{y}}}_{k|k-1} \right) \quad (22.66)$$

The stacked measurement matrix is

$$\mathbf{H}_k = \begin{bmatrix} H_{k,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & H_{k,N} \end{bmatrix} \quad (22.67)$$

The stacked measurement noise covariance is

$$\mathbf{R}_k = \begin{bmatrix} R_{k,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & R_{k,N} \end{bmatrix} \quad (22.68)$$

The stacked prior covariance matrix is

$$\mathbf{P}_{k|k-1} = \begin{bmatrix} P_{k|k-1,11} & \cdots & P_{k|k-1,1N} \\ \vdots & \ddots & \vdots \\ P_{k|k-1,1N} & \cdots & P_{k|k-1,NN} \end{bmatrix} \quad (22.69)$$

where the cross-covariance matrices are initially set to zero. The stacked innovation covariance matrix is

$$\mathbf{S}_k = \mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^T + \mathbf{R}_k \quad (22.70)$$

The stacked filter gain is

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{H}_k^T \mathbf{S}_k^{-1} \quad (22.71)$$

The stacked posterior information state mean is

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \tilde{\mathbf{r}}_k \quad (22.72)$$

The stacked posterior information state covariance is

$$\begin{aligned} P_{k|k} &= (1 - \Pr([DA]_{k,0} | \vec{\mathbf{y}}_{1:k})) (\mathbf{P}_{k|k-1} - \mathbf{K}_k \mathbf{S}_k \mathbf{K}_k^T) + \Pr([DA]_{k,0} | \vec{\mathbf{y}}_{1:k}) \mathbf{P}_{k|k-1} \\ &\quad + \mathbf{K}_k \left( \sum_{[DA]_k} \Pr([DA]_k | \vec{\mathbf{y}}_{1:k}) \left( \vec{\mathbf{y}}_{k,[DA]} - \hat{\vec{\mathbf{y}}}_{k|k-1} \right) \left( \vec{\mathbf{y}}_{k,,[DA]} - \hat{\vec{\mathbf{y}}}_{k|k-1} \right)^T - \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k^T \right) \mathbf{K}_k^T \end{aligned} \quad (22.73)$$

The **joint integrated probabilistic data association filter (JIPDAF)** provides an augmentation of the JPDAF where the probability of any target's existence can also modeled and combined into the filter.

## References

For more information, please refer to the following

- Bar-Shalom, Y., Daum, F., and Huang, J., “The Probabilistic Data Association Filter,” *IEEE Control Systems Magazine*, Vol. 29, Issue 6, 2009, pp. 82-100
- Musicki, D., and Evans, R. J., “Joint integrated probabilistic data association: JIPDA,” in *IEEE Transactions on Aerospace and Electronic Systems*, Volume 40, Issue 3, 2004, pp. 1093-1099

## 22.3 Multi-Hypothesis Tracking

**Multi-hypothesis tracking (MHT)** is a deferred decision logic in which alternative data association hypotheses are formed whenever track-to-measurement association conflict situations occur. Then, instead of choosing the single best association hypothesis, i.e. nearest-neighbor data association, or combining all probable association hypotheses, i.e. all-neighbor data association with gating, the different association hypotheses are propagated into the future in anticipation that subsequent data will resolve the uncertainty. Thus, one seeks to find a **global hypothesis** which is a collection of compatible tracks over some number of time steps representing the estimated sequence of target states where **compatible tracks** are defined as not having any measurements in common. This also allows for the possibility of new tracks originating in surveillance volume which is not directly applicable to the JPDAF technique.

As an example, consider that two target tracks, i.e.  $t_1$  and  $t_2$ , with predicted states, i.e.  $\hat{x}_1$  and  $\hat{x}_2$ , have been resolved before three measurements are obtained, i.e.  $\vec{y}_1$ ,  $\vec{y}_2$ , and  $\vec{y}_3$ , where  $\vec{y}_1$  and  $\vec{y}_2$  are within the gate of  $t_1$  and  $\vec{y}_2$  and  $\vec{y}_3$  are with the gate of  $t_2$ . Thus, the global hypotheses for this single time step can be written as

1.  $\mathcal{H}_1: \{t_1 - \emptyset, t_2 - \emptyset, \emptyset - \vec{y}_1, \emptyset - \vec{y}_2, \emptyset - \vec{y}_3\}$
2.  $\mathcal{H}_2: \{t_1 - \emptyset, t_2 - \vec{y}_2, \emptyset - \vec{y}_1, \emptyset - \vec{y}_3\}$
3.  $\mathcal{H}_3: \{t_1 - \emptyset, t_2 - \vec{y}_3, \emptyset - \vec{y}_1, \emptyset - \vec{y}_2\}$
4.  $\mathcal{H}_4: \{t_1 - \vec{y}_1, t_2 - \emptyset, \emptyset - \vec{y}_2, \emptyset - \vec{y}_3\}$
5.  $\mathcal{H}_5: \{t_1 - \vec{y}_1, t_2 - \vec{y}_2, \emptyset - \vec{y}_3\}$
6.  $\mathcal{H}_6: \{t_1 - \vec{y}_1, t_2 - \vec{y}_3, \emptyset - \vec{y}_2\}$
7.  $\mathcal{H}_7: \{t_1 - \vec{y}_2, t_2 - \emptyset, \emptyset - \vec{y}_1, \emptyset - \vec{y}_3\}$
8.  $\mathcal{H}_8: \{t_1 - \vec{y}_2, t_2 - \vec{y}_3, \emptyset - \vec{y}_1\}$

where  $t_i - \emptyset$  refers to a missed detection event and  $\emptyset - \vec{y}_j$  could refer to a clutter source or a potential new target event. Thus, there are 8 feasible hypotheses, and 5 feasible tracks, i.e. 2 *a priori* tracks and 3 potential “new” tracks.

**Reid’s algorithm** forms the basis of **hypothesis-oriented MHT (HOMHT)** which redefines the possible tracks from one time-step to the next as

$$t_i = t_i - \emptyset \quad \forall i = 1, \dots, N \quad (22.74)$$

for all  $l = 1, \dots, MN$  possible data associations of  $N$  targets and  $M$  measurements

$$t_{N+l} = t_i - \vec{y}_j \quad \forall i = 1, \dots, N, j = 1, \dots, M \quad (22.75)$$

and new track hypotheses as

$$t_{N+M:N+j} = \emptyset - \vec{y}_j \quad \forall j = 1, \dots, M \quad (22.76)$$

Thus, from one time step to another, the HOMHT maintains a number  $K$  of global hypotheses across time steps and expands these multi-step global hypotheses by considering all possible track-to-measurement associations for the current time step satisfying the track compatibility requirement. However, one can see the potential combinatorial explosion in the number of global hypotheses across multiple time steps, although many of these hypotheses may be very improbable. Thus, one common method is to use Murty's algorithm to find  $K$  for all multi-step hypotheses.

However, there can be a high cost to computationally evaluating a large number of global hypotheses in HOMHT which has led to the widespread use of the alternative **track-oriented MHT (TOMHT)** which instead uses track hypothesis trees that satisfy the track compatibility requirement instead of an exhaustive list of possible global hypotheses as the methods to prune the number of track hypothesis trees are more computationally tractable. However, TOMHT does not directly provide global hypothesis probabilities as in HOMHT, but instead uses multi-dimensional assignment (MDA) using Lagrangian relaxation to implement the hypothesis output.

## Track Scoring

The HOMHT yields the probability of a single cumulative joint event, i.e. a global hypothesis up to time step  $k$ , as

$$\begin{aligned} \Pr([DA]_{1:k}^l | \{\vec{y}_{1:k}\}) &= \frac{1}{c} \frac{\phi! \nu!}{M_k!} p_\phi(\phi) p_\nu(\nu) V^{-\phi-\nu} \prod_{j=1}^{M_k} [f_{t_j}(\vec{y}_{k,j})]^{\tau_j} \\ &\quad \prod_t [(P_{D_t,k})^{\delta_t} (1 - P_{D_t,k})^{1-\delta_t}] \\ &\Pr([DA]_{1:k-1}^s | \{\vec{y}_{1:k-1}\}) \end{aligned} \quad (22.77)$$

where  $[DA]_{1:k}^l$  is the joint data association hypothesis  $l$  through the current time step  $k$ ,  $\{\vec{y}_{1:k}\}$  is the cumulative set of measurements through the current time step  $k$ ,  $c$  is the normalization constant,  $\phi$  is the number of measurements at  $k$  deemed as false alarms in the hypothesis under consideration,  $\nu$  is the number of measurements at  $k$  deemed as new targets in the hypothesis under consideration,  $M_k$  is the number of measurements at  $k$ ,  $p_C$  is the PMF of the number of clutter in the current measurement set,  $p_\nu$  is the PMF of the number of new targets in the current measurement set,  $V$  is the surveillance volume,  $f_{t_j}$  is the innovation PDF for the predicted measurement of track  $t_j$  to which measurement  $\vec{y}_{k,j}$  is assigned in the current hypothesis,  $\tau_j$  is the indicator function for the assignment of measurement  $\vec{y}_{k,j}$  to a track in the hypothesis under consideration,  $P_{D_t,k}$  is the probability of detection for target  $t$  at time step  $k$ ,  $\delta_t$  is the indicator function for the assignment of a measurement to track  $t$  in the hypothesis under consideration, and  $[DA]_{1:k}^s$  is the parent joint data association hypothesis of  $[DA]_{1:k}^l$ .

Here, the whiteness of the innovation sequence of a track leads to the multiplication of the parent hypothesis probability by the PDF of the latest innovation. The factorial terms follow from a combinatorial analysis, starting with the *a priori* probabilities, that provides the correct *a posteriori* probabilities. Typically,

one assumes a Poisson PMF with clutter spatial density  $\lambda_C$ , i.e.

$$p_\phi(\phi) = e^{-\lambda_C V} \frac{(\lambda_C V_\phi)^\phi}{\phi!} \quad (22.78)$$

and, similarly, for the number of new targets, with spatial density,  $\lambda_\nu$ , i.e.

$$p_\nu(\nu) = e^{-\lambda_\nu V} \frac{(\lambda_\nu V_\phi)^\nu}{\nu!} \quad (22.79)$$

where  $V$  is the surveillance volume and  $\lambda_C/\lambda_\nu$  are the expected number of clutter/new targets of the measurement space per time step *per unit volume*.

Using these PMFs, one has

$$\begin{aligned} \Pr([DA]_{1:k}^l | \{\vec{y}_{1:k}\}) &= \frac{1}{c} \frac{e^{-\lambda_C V} e^{-\lambda_\nu V}}{M_k!} \lambda_C^\phi \lambda_\nu^\nu \prod_{j=1}^{M_k} [f_{t_j}(\vec{y}_{k,j})]^{\tau_j} \\ &\quad \prod_t [(P_{D_t,k})^{\delta_t} (1 - P_{D_t,k})^{1-\delta_t}] \\ &\quad \Pr([DA]_{1:k-1}^s | \{\vec{y}_{1:k-1}\}) \end{aligned} \quad (22.80)$$

The marginal likelihood for track-to-measurement association which multiplies the *a priori* probability from the parent track for measurement  $\vec{y}_{k,j}$  at time  $k$  having originated from track  $t$  is

$$\mathcal{L}_{k,tj}() = f_t(\vec{y}_{k,j}) P_{D_t,k} \quad (22.81)$$

and if it has no measurement, i.e. assigned the dummy measurement at index 0, one has

$$\mathcal{L}_{k,t0} = 1 - P_{D_t,k} \quad (22.82)$$

The marginal likelihood function for a clutter measurement is thus

$$\mathcal{L}_{k,\phi j} = \lambda_C \quad (22.83)$$

and the marginal likelihood function for a new target is

$$\mathcal{L}_{k,\nu j} = \lambda_\nu \quad (22.84)$$

which can both be interpreted as uniform distributions in the “average volume occupied by one such measurement.”

Thus, assuming an equal prior versus this track having no continuation, i.e.  $\vec{y}_{k,j}$  is either from clutter or a new target, the probability of the event  $[DA]_{k,tj}$  that measurement  $\vec{y}_{k,j}$  is the continuation of track  $t$  is

$$\begin{aligned} \Pr([DA]_{k,tj} | \vec{y}_k) &= \frac{f_t(\vec{y}_{k,j}) P_{D_t,k}}{f_t(\vec{y}_{k,j}) P_{D_t,k} + (\lambda_C + \lambda_\nu)[1 - P_{D_t,k}]} \\ &= \frac{\frac{f_t(\vec{y}_{k,j}) P_{D_t,k}}{\lambda_C + \lambda_\nu}}{\frac{f_t(\vec{y}_{k,j}) P_{D_t,k}}{\lambda_C + \lambda_\nu} + 1 - P_{D_t,k}} \\ &= \frac{\Lambda_{k,tj}}{\Lambda_{k,tj} + 1 - P_{D_t,k}} \end{aligned} \quad (22.85)$$

where  $\Lambda_{k,tj}$  is the likelihood ratio (LR) of track  $t$  continuing with measurement  $\vec{y}_{k,j}$  versus being “extraneous,” i.e.

$$\Lambda_{k,tj} = \lambda_{ex}^{-1} f_t(\vec{y}_{k,j}) P_{D_{t,k}} \quad (22.86)$$

where

$$\lambda_{ex} = \lambda_C + \lambda_v \quad (22.87)$$

Regardless of the choice of MHT, the hypothesis evaluation process relies on the ability to compute the **log-likelihood ratio (LLR)**, also known as the **track score**, for each individual track  $t$ -to-measurement  $j$  association (T2MA) hypothesis,  $\ell_{k,tj}$ , at each time step  $k$ , relative to the probability of being extraneous, i.e. either clutter or a new target, as

$$\ell_{k,tj} = \ln \Lambda_{k,tj} = \ln \frac{P_{t,j,k}}{P_{ex}} \quad (22.88)$$

Furthermore, if the innovations in a track are white, then the joint likelihood is the product of the marginal likelihoods, and the track score can be accumulated from  $l = 1, \dots, k$  as

$$\ell_{1:k,tj}() = \sum_{l=1}^k \ell_{l,tj(t,l)} \quad (22.89)$$

where  $\ell_{l,tj(t,l)}$  is the track score of associating to target  $t$  at time  $l$  measurement  $j(t,l)$ . Thus, for updating the information of the different data association hypotheses, one need only update the track score for each track.

To initialize the track scores for a new track, one typically sets the mean as the pseudomeasurement mapped to the target state domain with covariance corresponding to the mapped measurement noise covariance. However, if the state is not directly observable, then one typically initializes the state to the average and the standard deviation to half the maximum expected velocity. Lastly, one can note that the probability of the T2MA can be written as

$$P_{t,j,1:k} = \frac{\exp(\ell_{tj}(1:k))}{1 + \exp(\ell_{tj}(1:k))} \quad (22.90)$$

which can be used to determine thresholds to confirm/prune hypotheses with very high/low track scores.

## Multi-Dimensional Assignment

With this track score in mind, TOMHT only hypothesizes that measurements come from old tracks or are clutter, i.e.  $\lambda_v = 0$ . Then, any hypothesized clutter measurements are initiated as new tracks. This “two-way” measurement origin approach, as opposed allows TOMHT to use multi-dimensional assignment (MDA) which seeks to find the most likely set of tracks by maximizing the sum of all track scores while also maintaining the constraint that no tracks in the hypothesis share measurements, i.e. MDA seeks the best global hypothesis. To describe this assignment problem at time step  $k$ , consider one has  $N$  lists of hypotheses from the previous  $N$  time steps, where time step  $k - N$  has confirmed tracks to that point as the root nodes of each T2MA tree, followed by the most recent  $N - 1$  lists. Each item from each list should only be assigned once into a complete  $N$ -tuple consisting of one item from each of the  $N$  lists. However, to have a complete  $N$ -tuple, each list also must have a “dummy” element which stands for “clutter” or “missed detection” and has no constraints on how many times it can be assigned. For  $N > 2$ , MDA is an NP-hard problem and Lagrangian relaxation with approximate linear programming provides a near-optimal solution. Then, after

finding the best global hypothesis, one typically implements  $N$ -scan pruning which eliminates any branches of tracks not included in the solution starting from  $k - N$ , thus, confirming the tracks from step  $k - N$  to  $k + 1 - N$  for the next window. One typically also prunes any tracks with too low of track scores.

The basic principle of the Lagrangian relaxation approach is to replace constraints, e.g. a measurement can be used by at most a single track, by Lagrange multipliers in the objective function, e.g. the sum of track scores, used in the maximization. This method involves the proper choice of Lagrange multipliers so that the solution formed from maximizing the objective function approaches the best feasible solution, in which each measurement is used by at most a single track. This optimization is complex and requires sophisticated mathematics, one can summarize the basic principles as follows. Two solutions to the hypothesis formation problem are obtained with cost defined to be the negative of score. The first solution, defined to be the relaxed or dual solution, may not satisfy the constraints, i.e. a measurement should be used once and only once. However, Lagrange multipliers are introduced into this solution and are chosen so that constraint violations are, effectively, given high costs. Thus, the number of constraint violations should be reduced over time with successive iterations of the method. A second solution, denoted the recovered or primal solution, is obtained from the dual solution by enforcing the constraints.

For example, one method for obtaining this solution starts with the assignment of the first two scans of data that was obtained by the dual solution. Then, it adds measurements from the later scans by solving an assignment matrix, that enforces the constraints, on each later scan. Thus, a feasible, but likely sub-optimal, solution is obtained. The costs of the dual solution,  $q(\underline{u})$ , where  $\underline{u}$  represents the Lagrangian multipliers, and the primal solution,  $v(\bar{z})$ , represent bounds on the cost,  $v(z)$ , of the true, but unknown, solution, i.e.

$$q(\underline{u}) \leq v(z) \leq v(\bar{z}) \quad (22.91)$$

where  $z$  and  $\bar{z}$  are the set of binary variables that define which tracks are included in the true and the primal solutions, respectively. Successive iterations are performed by using updated Lagrange multipliers in an attempt to increase  $q(\underline{u})$  and decrease  $v(\bar{z})$  and a stopping rule is defined so that the feasible primary solution is accepted when  $q(\underline{u})$  and  $v(\bar{z})$  are “close enough,” or the max iterations are reached.

## 22.4 Probability Hypothesis Density Filtering

The probability hypothesis density (PHD) filter is an approximation developed to overcome the computational intractability of the multi-target Bayes filter by recursively updating the first-order statistical moment of the multi-target state, i.e. the **probability hypothesis density**, also known as the **intensity**, instead of the entire posterior multi-target density. This formulation takes advantage of the special properties of Poisson RFS. This strategy is similar to the constant-gain Kalman filter which recursively updates the first moment, i.e. the mean, of the single-target state.

Based on the facts for Poisson RFSs, the PHD filter relies on three assumptions.

- A. Each target evolves and generates measurements independently of one another.
- B. The clutter RFS,  $\mathbf{C}_k$ , is Poisson with PHD  $\kappa_k(\vec{y})$  and is independent of the target-originated measurements.
- C. The prior multi-target RFS governed by  $f_{\vec{X}_k | \vec{X}_{k-1}}(\mathbf{x}|\mathbf{z})$  is Poisson.

It can be shown that assumption C. is completely satisfied if the birth RFS,  $\mathbf{B}_k$ , is Poisson with PHD  $\beta_k(\vec{x})$  and there is no spawn RFS, though one can still include  $\mathbf{G}_k(\vec{x}|\vec{z})$  with PHD  $\gamma_{k|k-1}(\vec{x}|\vec{z})$  and approximate the posterior density as Poisson.

Next, let  $v_k(\vec{x})$  denote the PHD associated with the multi-target posterior density,  $f_{\mathbf{X}}(\mathbf{x}_k|\mathbf{y}_{1:k})$ , and let  $v_{k|k-1}(\vec{x})$  denote the PHD associated with the multi-target prior density,  $f_{\mathbf{X}_k|\mathbf{X}_{k-1}}(\mathbf{x}_k|\mathbf{y}_{1:k-1})$ . Then, with assumptions A.-C., the prior and posterior PHDs can be updated with the **probability hypothesis density filter** using the recursive prediction and correction steps

$$\begin{aligned} v_{k|k-1}(\vec{x}) &= \int p_{S,k}(\vec{z}) f_{\vec{X}_k|\vec{X}_{k-1}}(\vec{x}|\vec{z}) v_{k-1}(\vec{z}) d\vec{z} \\ &\quad + \int \gamma_{k|k-1}(\vec{x}|\vec{z}) v_{k-1}(\vec{z}) d\vec{z} + \beta_k(\vec{x}) \\ v_k(\vec{x}) &= (1 - p_{D,k}) v_{k|k-1}(\vec{x}) \\ &\quad + \sum_{\vec{y} \in \mathbf{y}_k} \frac{p_{D,k}(\vec{x}) f_{\vec{Y}_k|\vec{X}_k}(\vec{y}|\vec{x}) v_{k|k-1}(\vec{x})}{\kappa_k(\vec{y}) + \int p_{D,k}(\vec{z}) f_{\vec{Y}_k|\vec{X}_k}(\vec{y}|\vec{z}) v_{k|k-1}(\vec{z}) d\vec{z}} \end{aligned} \quad (22.92)$$

where  $\vec{z}$  is a dummy integration vector.

It is important to note that the PHD filter completely avoids the combinatorial computations arising from the unknown track-to-measurement. Furthermore, since the posterior PHD is a function on the single-target state space,  $X$ , the PHD filter requires much less computational power than the multi-target Bayes filter. However, the PHD filter does not admit a closed-form solution, in general, and **particle-PHD** approximations suffer from the “curse of dimensionality” of the number of targets and particles required to sufficiently cover the single-target space. Notably, these also require some form of clustering algorithm in order to extract meaningful tracks. However, a closed-form solution exists for linear, Gaussian models for individual targets as will be shown in the following subsection.

### Gaussian Mixture-Probability Hypothesis Density Filter

A closed-form to the PHD filter can be obtained by assuming the following

- D. Each target and sensor follows a linear-Gaussian dynamical model, i.e.

$$\begin{aligned} f_{\vec{X}_k|\vec{X}_{k-1}}(\vec{x}|\vec{z}) &= f_N(\vec{x}; F_{k-1}\vec{z}, Q_{k-1}) \\ f_{\vec{Y}_k|\vec{X}_k}(\vec{y}|\vec{z}) &= f_N(\vec{y}; H_k\vec{z}, R_k) \end{aligned} \quad (22.93)$$

- E. The survival and detection probabilities are state independent, i.e.

$$\begin{aligned} p_{S,k}(\vec{x}) &= p_{S,k} \\ p_{D,k}(\vec{x}) &= p_{D,k} \end{aligned} \quad (22.94)$$

- F. The birth and spawn RFS have Gaussian mixture (GM) PHDs, i.e.

$$\begin{aligned}\beta_k(\vec{x}) &= \sum_{j=1}^{J_{\beta,k}} w_{\beta,k}^{(j)} f_N(\vec{x}; \vec{m}_{\beta,k}^{(j)}, P_{\beta,k}^{(j)}) \\ \gamma_{k|k-1}(\vec{x}|\vec{z}) &= \sum_{j=1}^{J_{\beta,k}} w_{\gamma,k}^{(j)} f_N(\vec{x}; F_{\gamma,k-1}^{(j)} \vec{z} + \vec{d}_{\gamma,k}^{(j)}, Q_{\gamma,k-1}^{(j)})\end{aligned}\quad (22.95)$$

where  $J_{\beta,k}$  is the number of terms in the birth GM-PHD,  $w_{\beta,k}^{(j)}$  is the weight of the  $j^{\text{th}}$  term in the birth GM-PHD,  $\vec{m}_{\beta,k}^{(j)}$  is the mean of the  $j^{\text{th}}$  term in the birth GM-PHD,  $P_{\beta,k}^{(j)}$  is the covariance of the  $j^{\text{th}}$  term in the birth GM-PHD,  $J_{\gamma,k}$  is the number of terms in the spawn GM-PHD,  $w_{\gamma,k}^{(j)}$  is the weight of the  $j^{\text{th}}$  term in the spawn GM-PHD,  $F_{\gamma,k-1}^{(j)} \vec{z} + \vec{d}_{\gamma,k}^{(j)}$  is the mean of the  $j^{\text{th}}$  term in the spawn GM-PHD, and  $Q_{\gamma,k-1}^{(j)}$  is the covariance of the  $j^{\text{th}}$  term in the spawn GM-PHD.

In this model, it should be noted that the mean of the birth PHD,  $\vec{m}_{\beta,k}$  corresponds to the highest local concentrations of expected number of spontaneous births, whereas the mean of the spawn PHD,  $F_{\gamma,k-1}^{(j)} \vec{z} + \vec{d}_{\gamma,k}^{(j)}$  corresponds to the highest local concentration of the expected number of generated spawns which depends on the previous posterior state. For example,  $\vec{m}_{\beta,k}$  could be a static, known, airport location while  $\vec{x}_{k-1}$  could be an aircraft carrier which is also being tracked, but can spawn fighter planes. For both,  $w_{\bullet,k}^{(j)}$  corresponds to the expected number of targets originating from any particular Gaussian term, and  $P_{\beta,k}^{(j)}$  and  $Q_{\gamma,k-1}^{(j)}$  corresponds to the spread of the PHDs in the vicinity of the peaks.

However, it should be noted that one can derive closed-form expressions for exponential mixtures for  $p_{S,k}(\vec{x})$  and  $p_{D,k}(\vec{x})$ , i.e.

$$\begin{aligned}p_{S,k}(\vec{x}) &= w_{S,k}^{(0)} + \sum_{j=1}^{J_{S,k}} w_{S,k}^{(j)} f_N(\vec{x}; \vec{m}_{S,k}^{(j)}, P_{S,k}^{(j)}) \\ p_{D,k}(\vec{x}) &= w_{D,k}^{(0)} + \sum_{j=1}^{J_{D,k}} w_{D,k}^{(j)} f_N(\vec{x}; \vec{m}_{D,k}^{(j)}, P_{D,k}^{(j)})\end{aligned}\quad (22.96)$$

Furthermore, the use of Gaussian mixture PHDs allows one to approximate any other form of birth and spawn PHDs to any desired accuracy, albeit at a higher memory and computational cost due to the number of mixtures required to do so.

With assumptions A.-F. and an initial multi-target state as a Gaussian mixture, then, the **Gaussian mixture - probability hypothesis density (GM-PHD) filter** can be constructed as recursive prediction and correction steps. For the prediction step, the prior PHD at time step  $k$  will be a Gaussian mixture

$$\nu_{k|k-1}(\vec{x}) = \nu_{S,k|k-1}(\vec{x}) + \nu_{\beta,k}(\vec{x}) + \nu_{\gamma,k|k-1}(\vec{x}) \quad (22.97)$$

where the posterior PHD at time step  $k - 1$  is a Gaussian mixture

$$\nu_{k-1}(\vec{x}) = \sum_{i=1}^{J_{k-1}} w_{k-1}^{(i)} f_N(\vec{x}; \vec{m}_{k-1}^{(i)}, P_{k-1}^{(i)}) \quad (22.98)$$

and the individual PHDs are Gaussian mixtures

$$v_{S,k|k-1}(\vec{x}) = p_{S,k} \sum_{i=1}^{J_{k-1}} w_{k-1}^{(i)} f_N \left( \vec{x}; F_{k-1} \vec{m}_{k-1}^{(i)}, F_{k-1} P_{k-1}^{(i)} F_{k-1}^T + Q_{k-1} \right) \quad (22.99)$$

$$v_{\beta,k}(\vec{x}) = \sum_{j=1}^{J_{\beta,k}} w_{\beta,k}^{(j)} f_N \left( \vec{x}; \vec{m}_{\beta,k}^{(j)}, P_{\beta,k}^{(j)} \right) \quad (22.100)$$

$$v_{\gamma,k|k-1}(\vec{x}) = \sum_{i=1}^{J_{k-1}} \sum_{j=1}^{J_{\gamma,k}} w_{k-1}^{(i)} w_{\gamma,k}^{(j)} f_N \left( \vec{x}; F_{\gamma,k-1}^{(j)} \vec{m}_{k-1}^{(i)} + \vec{d}_{\gamma,k}^{(j)}, F_{\gamma,k-1}^{(j)} P_{k-1}^{(i)} \left( F_{\gamma,k-1}^{(j)} \right)^T + Q_{\gamma,k-1}^{(j)} \right) \quad (22.101)$$

Likewise, for the correction step, the posterior multi-target PHD at time step  $k$  will be a Gaussian mixture

$$v_k(\vec{x}) = (1 - p_{D,k}) v_{k|k-1}(\vec{x}) + \sum_{\vec{y} \in \mathbf{y}_k} \sum_{i=1}^{J_{k|k-1}} w_k^{(i)}(\vec{y}) f_N \left( \vec{x}; \vec{m}_{k|k}^{(i)}(\vec{y}), P_{k|k}^{(i)}(\vec{y}) \right) \quad (22.102)$$

where the prior multi-target state PHD at time step  $k$  is a Gaussian mixture

$$v_{k|k-1}(\vec{x}) = \sum_{i=1}^{J_{k|k-1}} w_{k|k-1}^{(i)} f_N(\vec{x}; \vec{m}_{k|k-1}^{(i)}, P_{k|k-1}^{(i)}) \quad (22.103)$$

and the Gaussian mixture terms for each measurement,  $\vec{y}$ , are given by

$$w_k^{(i)}(\vec{y}) = \frac{\mathcal{L}_k^{(i)}(\vec{y})}{\kappa_k(\vec{y}) + \sum_{j=1}^{J_{k|k-1}} \mathcal{L}_k^{(j)}(\vec{y})} \quad (22.104)$$

$$\mathcal{L}_k^{(i)}(\vec{y}) = p_{D,k} w_{k|k-1}^{(i)} f_N(\vec{y}; H_k \vec{m}_{k|k-1}^{(i)}, S_k^{(i)}) \quad (22.105)$$

$$S_k^{(i)} = H_k P_{k|k-1}^{(i)} H_k^T + R_k \quad (22.106)$$

$$K_k^{(i)} = P_{k|k-1}^{(i)} H_k^T \left( S_k^{(i)} \right)^{-1} \quad (22.107)$$

$$\vec{m}_{k|k}^{(i)}(\vec{y}) = \vec{m}_{k|k-1}^{(i)} + K_k^{(i)} (\vec{y} - H_k \vec{m}_{k|k-1}^{(i)}) \quad (22.108)$$

and

$$P_{k|k}^{(i)} = P_{k|k-1}^{(i)} - K_k^{(i)} S_k^{(i)} \left( K_k^{(i)} \right)^T \quad (22.109)$$

where it should be pointed out that the mean and covariance recursions for the surviving and detected targets are given by the classic Kalman filter equations. Thus, for nonlinear motion and measurement models, one could easily implement an extended or unscented Kalman filter-like versions of the GM-PHD filter through either Jacobian linearization or statistical linearization.

Furthermore, based on the properties of Poisson PHDs, the prior cardinality estimate can be shown to be

$$\hat{N}_{k|k-1} = \hat{N}_{k-1} \left( p_{S,k} + \sum_{j=1}^{J_{\gamma,k}} w_{\gamma,k}^{(j)} \right) + \sum_{j=1}^{J_{\beta,k}} w_{\beta,k}^{(j)} \quad (22.110)$$

and the posterior cardinality estimate can be shown to be

$$\hat{N}_k = \hat{N}_{k|k-1} (1 - p_{D,k}) + \sum_{\vec{y} \in \mathbf{y}_k} \sum_{i=1}^{J_{k|k-1}} w_k^{(i)}(\vec{y}) \quad (22.111)$$

Thus, the prior multi-target PHD consists of the sum of the survival, birth, and spawn GMs while the posterior multi-target PHD consists of the sum of each of these GMs possibly being missed detected as well as detected *for each measurement*. Thus, the number of terms at time step  $k$ ,  $J_k$ , is

$$J_k = (J_{k-1}(1 + J_{\gamma,k}) + J_{\beta,k})(1 + |\mathbf{y}_k|) \quad (22.112)$$

which increases without bound. This combinatorial nature of the GM-PHD requires that practical implementations of a GM-PHD filter to implement an additional **GM-PHD manage step** made up of pruning very unlikely terms based on the individual weight thresholds, merging similar terms based on the Mahalanobis distance, and capping the maximum number of terms in the posterior multi-target PHD based on the individual weight thresholds.

After pruning and before capping, the merging step evaluates the set of mixture terms with indices  $I$ , as

$$j = \underset{i \in I}{\operatorname{argmax}} w_k^{(i)} \quad (22.113)$$

and computes a subset of indices,  $L$ , for Gaussian terms  $(\vec{w}_k^{(i)}, \vec{m}_k^{(i)}, \vec{P}_k^{(i)})$  that are statistically close to this term based on thresholding the Mahalanobis distance, i.e.

$$L = \{i \in I : (\vec{m}_k^{(i)} - \vec{m}_k^{(j)})^T \left( P_k^{(i)} \right)^{-1} (\vec{m}_k^{(i)} - \vec{m}_k^{(j)}) \leq \tau\} \quad (22.114)$$

for some chosen  $\tau$ . This allows one to form the new merged Gaussian term  $(\bar{w}_k^{(\ell)}, \bar{m}_k^{(\ell)}, \bar{P}_k^{(\ell)})$ , with index  $\ell$  as

$$\bar{w}_k^{(\ell)} = \sum_{i \in L} w_k^{(i)} \quad (22.115)$$

$$\bar{m}_k^{(\ell)} = \frac{1}{\bar{w}_k^{(\ell)}} \sum_{i \in L} w_k^{(i)} \vec{m}_k^{(i)} \quad (22.116)$$

and

$$\bar{P}_k^{(\ell)} = \frac{1}{\bar{w}_k^{(\ell)}} \sum_{i \in L} w_k^{(i)} \left( P_k^{(i)} + (\bar{m}_k^{(\ell)} - \vec{m}_k^{(i)}) (\bar{m}_k^{(\ell)} - \vec{m}_k^{(i)})^T \right) \quad (22.117)$$

and reiterates over terms with indices in  $I$  with all  $L$  removed.

Furthermore, one must also form a multi-target state estimate using an **GM-PHD extraction step**. For the particle-PHD filter, one typically performs clustering of the particles representing the PHD based on the estimated number of targets or close to that number. In the GM-PHD, one can easily extract the expected number of targets *from each term* in the mixture. Thus, if there is significant separation for each term, a condition easily satisfied due to the merging step in the manage step, then, the mean of each term with a statistically significant weight, e.g.  $> 0.5$ , should be output as a likely target location. In addition, as a single Gaussian can represent any number of targets in its weight, one typically rounds the weight and reports that many copies of the mean.

## CPHD Filter

The PHD filter updates the cardinality information indirectly through the PHD of the posterior multi-target density which can be integrated over the entire single-target state space to obtain the mean of the cardinality distribution. This effectively approximates the cardinality distribution by a Poisson distribution. However, as the mean and variance of a Poisson distribution are equal, when a high number of targets are present, the PHD filter estimates the cardinality with a correspondingly high variance. In practice, this approximation manifests itself in drastically fluctuating estimates of the number of targets. To address this problem, Mahler relaxed the first-order assumption on the number of targets and derived a generalization of the PHD recursion known as the cardinalized PHD (CPHD) recursion, which jointly propagates the PHD and an arbitrary cardinality distribution through the use of IID cluster RFSs. Likewise, a closed-form solution for the CPHD recursion can be written for linear, Gaussian motion and measurement models which uses a Gaussian mixture formulation called the GM-CPHD.

This section will adopt the following notation.  $C_j^\ell$  is the **binomial coefficient**

$$C_j^\ell = \frac{\ell!}{j!(\ell-j)!} \quad (22.118)$$

$\mathcal{P}_j^N$  is the **permutation coefficient**

$$\mathcal{P}_j^N = \frac{N!}{(N-j)!} \quad (22.119)$$

$\langle \cdot, \cdot \rangle$  denotes the **inner product** between two real-valued functions,  $\alpha(\vec{x})$  and  $\beta(\vec{x})$ , as

$$\langle \alpha, \beta \rangle = \int \alpha(\vec{x})\beta(\vec{x})d\vec{x} \quad (22.120)$$

or real-valued sequences,  $\alpha(\ell)$  and  $\beta(\ell)$ , as

$$\langle \alpha, \beta \rangle = \sum_{\ell=0}^{\infty} \alpha(\ell)\beta(\ell) \quad (22.121)$$

The CPHD filter relies on four assumptions.

- A. Each target evolves and generates measurements independently of one another.
- B. The clutter RFS,  $\mathbf{C}_k$ , is IID cluster and is independent of the target-originated measurements.
- C. The birth RFS,  $\mathbf{B}_k$ , and surviving RFS,  $\mathbf{S}_k(\vec{x})$  are IID cluster and is independent of each other.
- D. The prior and posterior multi-target RFS's are IID cluster.

For the sake of simplicity of presentation, this version will also assume (E.) there is no spawn RFS, though the CPHD has been derived for Bernoulli, Poisson, and zero-inflated Poisson RFS spawn models.

Next, let  $v_k(\vec{x})$  denote the PHD and  $\varrho_k(N)$  denote the cardinality distribution associated with the multi-target posterior density,  $f_{\mathbf{X}_k}(\mathbf{x}_k|\mathbf{y}_{1:k})$ . Also, let  $v_{k|k-1}(\vec{x})$  denote the PHD and  $\varrho_{k|k-1}(N)$  denote the cardinality distribution associated with the multi-target prior density,  $f_{\mathbf{X}_k|\mathbf{X}_{k-1}}(\mathbf{x}_k|\mathbf{y}_{1:k-1})$ . Then, with assumptions A.-E., the prior and posterior PHDs and cardinality distributions can be updated with the

**cardinalized probability hypothesis density (CPHD) filter** using the following recursive prediction and correction steps.

The prediction step of the CPHD filter is given by

$$\begin{aligned} v_{k|k-1}(\vec{x}) &= \int p_{S,k}(\vec{z}) f_{\vec{X}_k|\vec{X}_{k-1}}(\vec{x}|\vec{z}) v_{k-1}(\vec{z}) d\vec{z} + \beta_k(\vec{x}) \\ \varrho_{k|k-1}(N) &= \sum_{j=0}^N \varrho_{B,k}(N-j) \sum_{\ell=j}^{\infty} C_j^\ell \frac{\langle p_{S,k}, v_{k-1} \rangle^j \langle 1 - p_{S,k}, v_{k-1} \rangle^{\ell-j}}{\langle 1, v_{k-1} \rangle^\ell} \varrho_{k-1}(\ell) \end{aligned} \quad (22.122)$$

where  $p_{S,k}(\vec{z})$  is the probability of target survival at time  $k$  given previous state  $\vec{z}$ ,  $\vec{z}$  is a dummy integration vector,  $f_{\vec{X}_k|\vec{X}_{k-1}}(\vec{x}|\vec{z})$  is the single-target transition density at time  $k$  given previous state  $\vec{z}$ , and  $\beta_k(\cdot)$  and  $\varrho_{B,k}(\cdot)$  are the birth PHD and cardinality distribution at time  $k$ , respectively. Notably, the prior PHD prediction is the same as for the PHD filter and uncoupled from the prior cardinality. The CPHD prior cardinality distribution is simply a convolution of the birth and survival cardinalities as the prior cardinality random variable is the sum of the birth and surviving targets.

The correction step of the CPHD filter is given by

$$\begin{aligned} v_k(\vec{x}) &= \frac{\langle \Upsilon_k^1[v_{k|k-1}, \mathbf{y}_k], \varrho_{k|k-1} \rangle}{\langle \Upsilon_k^0[v_{k|k-1}, \mathbf{y}_k], \varrho_{k|k-1} \rangle} (1 - p_{D,k}(\vec{x})) v_{k|k-1}(\vec{x}) \\ &\quad + \sum_{\vec{y} \in \mathbf{y}_k} \frac{\langle \Upsilon_k^1[v_{k|k-1}, \mathbf{y}_k \setminus \{\vec{y}\}], \varrho_{k|k-1} \rangle}{\langle \Upsilon_k^0[v_{k|k-1}, \mathbf{y}_k], \varrho_{k|k-1} \rangle} (\Psi_{k,\vec{y}}(\vec{x}) v_{k|k-1}(\vec{x})) \\ \varrho_k(N) &= \frac{\Upsilon_k^0[v_{k|k-1}, \mathbf{y}_k](N) \varrho_{k|k-1}(N)}{\langle \Upsilon_k^0[v_{k|k-1}, \mathbf{y}_k], \varrho_{k|k-1} \rangle} \end{aligned} \quad (22.123)$$

where

$$\Upsilon_k^u[v, \mathbf{y}](N) = \sum_{j=0}^{\min(|\mathbf{y}|, N)} (|\mathbf{y}| - j) \varrho_{C,k}(|\mathbf{y}| - j) \mathcal{P}_{j+u}^N \frac{\langle 1 - p_{D,k}, v \rangle^{N-(j+u)}}{\langle 1, v \rangle^N} e_j(\Xi_k(v, \mathbf{y})) \quad (22.124)$$

$$\Xi_k(v, \mathbf{y}) = \{ \langle v, \Psi_{k,\vec{y}} \rangle : \vec{y} \in \mathbf{y} \} \quad (22.125)$$

$$\Psi_{k,\vec{y}}(\vec{x}) = \frac{\langle 1, \kappa_k \rangle}{\kappa_k(\vec{y})} f_{\vec{Y}_k|\vec{X}_k}(\vec{y}|\vec{x}) p_{D,k}(\vec{x}) \quad (22.126)$$

$$e_j(\Xi) = \sum_{\mathbf{z} \subseteq \Xi, |\mathbf{z}|=j} \left( \prod_{\vec{z} \in \mathbf{z}} \vec{z} \right) \quad (22.127)$$

is the **elementary symmetric function** of order  $j$  with  $e_0(\Xi) = 1$  by convention,  $\mathbf{y}_k$  is the measurement set at time  $k$ ,  $p_{D,k}(\vec{x})$  is the probability of target detection at time step  $k$ ,  $f_{\vec{Y}_k|\vec{X}_k}(\vec{y}|\vec{x})$  is the single-target measurement likelihood at time step  $k$  given state  $\vec{x}$ , and  $\kappa_k(\cdot)$  and  $\varrho_{C,k}$  are the clutter PHD and cardinality distribution at time  $k$ , respectively.

The CPHD cardinality and PHD corrections are coupled; however, the posterior PHD is similar to the PHD filter as both have one missed detection term and  $|\mathbf{y}|$  detection terms. The cardinality correction is a Bayes

update incorporating the clutter cardinality, the measurement set, the prior PHD, and the prior cardinality into  $\Upsilon_k^0[v_{k|k-1}, \mathbf{y}_k](N)$ , i.e. the likelihood of the multi-target measurement  $\mathbf{y}$  where  $\langle \Upsilon_k^0[v_{k|k-1}, \mathbf{y}_k], \varrho_{k|k-1} \rangle$  is the normalizing constant. The particle-CPHD filter can be used to approximate this recursion for arbitrary cardinality distributions and PHDs. However, just like the particle-PHD filter, the state extraction involves clustering to partition the particle population into a given number of clusters, e.g. the estimated number of targets, but adds additional uncertainty in the target tracking system due to the potential clustering error. Furthermore, if the cardinality distribution is infinite tailed, the recursive updates of the entire posterior cardinality distribution is generally not possible as this would involve storing an infinite number of terms for the infinite sums. Thus, in practice, the cardinality distributions are truncated at and approximated with a finite number of terms which must be significantly greater than the maximum number of targets possible in the surveillance volume which must be chosen *a priori* for the CPHD filter.

Notably, the CPHD filter requires a solver of the elementary symmetric functions which can be done efficiently with the following techniques. Evaluating the elementary symmetric functions directly from the definition is clearly intractable. Using a basic result from combinatorics theory known as the Newton–Girard formulas, the elementary symmetric function  $e_j(\cdot)$  can be computed using the following procedure. Let  $\lambda_1, \dots, \lambda_M$  be distinct roots of the polynomial  $a_M x^M + a_{M-1} x^{M-1} + \dots + a_1 x + a_0$ . Then,  $e_j(\cdot)$  for orders  $j = 0, \dots, M$  is given by  $e_j(\lambda_1, \dots, \lambda_M) = (-1)^j a_{M-j} / a_M$ . Thus, the values  $e_j(\mathbf{y})$  over the finite set,  $\mathbf{y}$ , can be evaluated by expanding out the polynomial with roots given by the elements of  $\mathbf{y}$ , which can be implemented using an appropriate recursion or convolution. For finite set  $\mathbf{y}$ , this calculation of  $e_j(\mathbf{y})$  requires  $|\mathbf{y}|^2$  operations. In the CPHD recursion, each correction step requires the calculation of  $|\mathbf{y}| + 1$  elementary symmetric functions, i.e. one for  $\mathbf{y}$  and one for each set  $\{\mathbf{y} \setminus \vec{y}\}$  where  $\vec{y} \in \mathbf{y}$ . Thus, the CPHD recursion has a complexity of  $\mathcal{O}(|\mathbf{y}|^3)$ . However, in practice, the number of measurements considered in  $\mathbf{y}$  can be reduced by classical gating techniques.

### Gaussian Mixture - CPHD Filter

A closed-form to the CPHD filter can be obtained by assuming the following

F. Each target and sensor follows a linear-Gaussian dynamical model, i.e.

$$\begin{aligned} f_{\vec{X}_k | \vec{X}_{k-1}}(\vec{x} | \vec{z}) &= f_N(\vec{x}; F_{k-1} \vec{z}, Q_{k-1}) \\ f_{\vec{Y}_k | \vec{X}_k}(\vec{y} | \vec{z}) &= f_N(\vec{y}; H_k \vec{z}, R_k) \end{aligned} \tag{22.128}$$

G. The survival and detection probabilities are state independent, i.e.

$$\begin{aligned} p_{S,k}(\vec{x}) &= p_{S,k} \\ p_{D,k}(\vec{x}) &= p_{D,k} \end{aligned} \tag{22.129}$$

H. The PHD of the birth RFS is a GM, i.e.

$$\beta_k(\vec{x}) = \sum_{j=1}^{J_{\beta,k}} w_{\beta,k}^{(j)} f_N(\vec{x}; \vec{m}_{\beta,k}^{(j)}, P_{\beta,k}^{(j)}) \quad (22.130)$$

With assumptions A.-H. and an initial multi-target state as a Gaussian mixture, then, the **Gaussian mixture - cardinalized probability hypothesis density (GM-CPHD) filter** can be constructed as recursive prediction and correction steps. For the GM-CPHD prediction step, the prior PHD and cardinality distributions at time step  $k$  are given by

$$\begin{aligned} v_{k|k-1}(\vec{x}) &= v_{S,k|k-1}(\vec{x}) + v_{\beta,k}(\vec{x}) \\ \varrho_{k|k-1}(N) &= \sum_{j=0}^N \varrho_{B,k}(N-j) \sum_{\ell=j}^{\infty} C_j^\ell \varrho_{k-1}(\ell) p_{S,k}^j (1-p_{S,k})^{\ell-j} \end{aligned} \quad (22.131)$$

where the posterior PHD at time step  $k-1$  is a Gaussian mixture

$$v_{k-1}(\vec{x}) = \sum_{i=1}^{J_{k-1}} w_{k-1}^{(i)} f_N(\vec{x}; \vec{m}_{k-1}^{(i)}, P_{k-1}^{(i)}) \quad (22.132)$$

the posterior cardinality distribution at time step  $k-1$  is  $\varrho_{k-1}$ , and the individual PHDs are Gaussian mixtures

$$v_{S,k|k-1}(\vec{x}) = p_{S,k} \sum_{i=1}^{J_{k-1}} w_{k-1}^{(i)} f_N\left(\vec{x}; F_{k-1} \vec{m}_{k-1}^{(i)}, F_{k-1} P_{k-1}^{(i)} F_{k-1}^T + Q_{k-1}\right) \quad (22.133)$$

$$v_{\beta,k}(\vec{x}) = \sum_{j=1}^{J_{\beta,k}} w_{\beta,k}^{(j)} f_N\left(\vec{x}; \vec{m}_{\beta,k}^{(j)}, P_{\beta,k}^{(j)}\right) \quad (22.134)$$

Likewise, for the GM-CPHD correction step, the posterior PHD and cardinality distributions at time step  $k$  are given by

$$\begin{aligned} v_k(\vec{x}) &= \frac{\Omega_k^1[\vec{w}_{k|k-1}, \mathbf{y}_k](N) \varrho_{k|k-1}(N)}{\langle \Omega_k^0[\vec{w}_{k|k-1}, \mathbf{y}_k], \varrho_{k|k-1} \rangle} (1-p_{D,k}) v_{k|k-1}(\vec{x}) + \sum_{\vec{y} \in \mathbf{y}_k} \sum_{i=1}^{J_{k|k-1}} w_k^{(i)}(\vec{y}) f_N\left(\vec{x}; \vec{m}_{k|k}^{(i)}(\vec{y}), P_{k|k}^{(i)}(\vec{y})\right) \\ \varrho_k(N) &= \frac{\Omega_k^0[\vec{w}_{k|k-1}, \mathbf{y}_k](N) \varrho_{k|k-1}(N)}{\langle \Omega_k^0[\vec{w}_{k|k-1}, \mathbf{y}_k], \varrho_{k|k-1} \rangle} \end{aligned} \quad (22.135)$$

where the prior multi-target state PHD at time step  $k$  is a Gaussian mixture

$$v_{k|k-1}(\vec{x}) = \sum_{i=1}^{J_{k|k-1}} w_{k|k-1}^{(i)} f_N(\vec{x}; \vec{m}_{k|k-1}^{(i)}, P_{k|k-1}^{(i)}) \quad (22.136)$$

$$\Omega_k^u[\vec{w}, \mathbf{y}_k](N) = \sum_{j=0}^{\min(|\mathbf{y}|, N)} (|\mathbf{y}| - j) \varrho_{C,k}(|\mathbf{y}| - j) \mathcal{P}_{j+u}^N \frac{(1-p_{D,k})^{N-(j+u)}}{\langle 1, w \rangle^{j+u}} e_j(\Lambda_k(w, \mathbf{y})) \quad (22.137)$$

$$\Lambda_k(w, \mathbf{y}) = \left\{ \frac{\langle 1, \kappa_k \rangle}{\kappa_k(\vec{y})} p_{D,k} \vec{w}_{k|k-1}^T \vec{\mathcal{L}}_k(\vec{y}) : \vec{y} \in \mathbf{y} \right\} \quad (22.138)$$

$$\vec{w}_{k|k-1} = \left[ w_{k|k-1}^{(1)}, \dots, w_{k|k-1}^{(J_{k|k-1})} \right]^T \quad (22.139)$$

$$\mathcal{L}_k^{(i)}(\vec{y}) = f_N(\vec{x}_k, H_k \vec{m}_{k|k-1}^{(i)}, S_k^{(i)}) \quad (22.140)$$

$$\vec{\mathcal{L}}_k(\vec{y}) = \left[ \mathcal{L}_k^{(1)}(\vec{y}), \dots, \mathcal{L}_k^{(J_{k|k-1})}(\vec{y}) \right]^T \quad (22.141)$$

where the Gaussian mixture terms for each measurement,  $\vec{y}$ , are given by

$$w_k^{(i)}(\vec{y}) = p_{D,k} w_{k|k-1}^{(i)} \mathcal{L}_k^{(i)}(\vec{y}) \frac{\langle \Omega_k^1[\vec{w}_{k|k-1}, \mathbf{y}_k \setminus \{\vec{y}\}], \varrho_{k|k-1} \rangle \langle 1, \kappa_k \rangle}{\langle \Omega_k^0[w_{k|k-1}, \mathbf{y}_k], \varrho_{k|k-1} \rangle \kappa_k(\vec{y})} \quad (22.142)$$

$$S_k^{(i)} = H_k P_{k|k-1}^{(i)} H_k^T + R_k \quad (22.143)$$

$$K_k^{(i)} = P_{k|k-1}^{(i)} H_k^T \left( S_k^{(i)} \right)^{-1} \quad (22.144)$$

$$\vec{m}_{k|k}^{(i)}(\vec{y}) = \vec{m}_{k|k-1}^{(i)} + K_k^{(i)}(\vec{y} - H_k \vec{m}_{k|k-1}^{(i)}) \quad (22.145)$$

and

$$P_{k|k}^{(i)} = P_{k|k-1}^{(i)} - K_k^{(i)} S_k^{(i)} \left( K_k^{(i)} \right)^T \quad (22.146)$$

where it should be pointed out that the mean and covariance recursions for the surviving and detected targets are given by the classic Kalman filter equations. Thus, for nonlinear motion and measurement models, one could easily implement an extended or unscented Kalman filter-like versions of the GM-CPHD filter through either Jacobian linearization or statistical linearization. Also, similar to GM-PHD filter, the number of Gaussian components required to represent the posterior PHD increases without bound which can be mitigated using the same **manage step** as described for the GM-PHD.

For the **GM-CPHD extraction step**, the number of targets can be estimated using, for example, the expected posterior (EAP) or maximum a posteriori (MAP) estimators. However, the EAP estimator is likely to fluctuate and be unreliable as clutter and missed detections may induce minor modes in the posterior cardinality distribution, away from the target-induced primary mode which would effect the expected value. As such, the MAP estimator is typically more reliable since it ignores minor modes and locks directly onto the target-induced primary mode, thereby usually preferred over the EAP estimator.

### Constant-Cardinality CPHD Filter

A special case of the CPHD filter can be derived if the number of targets,  $N$ , is known *a priori* and constant, i.e. the cardinality distribution at any time step  $k$  is a Dirac delta function at  $N$ ,  $\varrho_{k|k-1} = \varrho_k = \delta_N(\cdot)$ . As there are no births, the birth PHD is  $\beta_k(\vec{x}) = 0 \forall k$  and no deaths, the survival probability is  $p_{S,k} = 1 \forall k$ .

Thus, the prediction and correction steps for the CPHD filter can be simplified as

$$\begin{aligned} v_{k|k-1}(\vec{x}) &= \int f_{\vec{X}_k|\vec{X}_{k-1}}(\vec{x}|\vec{z})v_{k-1}(\vec{z})d\vec{z} \\ v_k(\vec{x}) &= \frac{\Upsilon_k^1[v_{k|k-1}, \mathbf{y}_k](N)}{\Upsilon_k^0[v_{k|k-1}, \mathbf{y}_k](N)}(1 - p_{D,k}(\vec{x}))v_{k|k-1}(\vec{x}) \\ &\quad + \sum_{\vec{y} \in \mathbf{y}_k} \frac{\Upsilon_k^1[v_{k|k-1}, \mathbf{y}_k \setminus \{\vec{y}\}](N)}{\Upsilon_k^0[v_{k|k-1}, \mathbf{y}_k](N)} \Psi_{k,\vec{y}}(\vec{x})v_{k|k-1}(\vec{x}) \end{aligned} \quad (22.147)$$

Furthermore, with the assumptions F.-H., the GM-CPHD can be simplified as

$$\begin{aligned} v_{k|k-1}(\vec{x}) &= \sum_{i=1}^{J_{k-1}} w_{k-1}^{(i)} f_N \left( \vec{x}; F_{k-1} \vec{m}_{k-1}^{(i)}, F_{k-1} P_{k-1}^{(i)} F_{k-1}^T + Q_{k-1} \right) \\ v_k(\vec{x}) &= \frac{\Omega_k^1[v_{k|k-1}, \mathbf{y}_k](N)}{\Omega_k^0[v_{k|k-1}, \mathbf{y}_k](N)}(1 - p_{D,k}(\vec{x}))v_{k|k-1}(\vec{x}) \\ &\quad + \sum_{\vec{y} \in \mathbf{y}_k} \sum_{i=1}^{J_{k-1}} w_k^{(i)}(\vec{y}) f_N \left( \vec{x}; \vec{m}_{k|k}^{(i)}(\vec{y}), P_{k|k}^{(i)}(\vec{y}) \right) \end{aligned} \quad (22.148)$$

where the weights are now defined as

$$w_k^{(i)}(\vec{y}) = p_{D,k} w_{k|k-1}^{(i)} \mathcal{L}_k^{(i)}(\vec{y}) \frac{\Omega_k^1[\vec{w}_{k|k-1}, \mathbf{y}_k \setminus \{\vec{y}\}](N) \langle 1, \kappa_k \rangle}{\Omega_k^0[w_{k|k-1}, \mathbf{y}_k](N) \kappa_k(\vec{y})} \quad (22.149)$$

and the other terms have the same definitions as the standard GM-CPHD. It has been shown that this CC-GM-CPHD filter has better performance than the JPDAF for resolving target crossings with lower computational cost.

## References

For more information, please refer to the following

- ...

## 22.5 Bernoulli Filtering

**Single-Target Bernoulli Filter**

**Conjugate Prior of Multi-Target Bayes Filter**

**Cardinality-Balanced Multi-Target Multi-Bernoulli Filter**

**Track-Oriented Marginalized Multi-Bernoulli Filter**

**Measurement-Oriented Marginalized Multi-Bernoulli Filter**

**Poisson Multi-Bernoulli Mixture Filter**

**Poisson Multi-Bernoulli Filter**

**Generalized Labeled Multi-Bernoulli Filter**

Thus, to incorporate tracks into the multi-target Bayes filter, one can identify targets by an ordered pair of integers  $\ell = (k, i)$  where  $k$  is the **time of birth** and  $i \times \mathbb{N}$  is the unique index to distinguish objects born at the same time. In this way, the label space for targets born at time  $k$  is denoted  $\mathbb{L}_k = \{k\} \times \mathbb{N}$  and has a single-target state  $\vec{x} \in \mathbb{X} \times \mathbb{L}_k$ . Thus, the label space for targets at time  $k$ , including those born prior to  $k$ , denoted as  $\mathbb{L}_{0:k}$ , is given recursively by the disjoint union,  $\mathbb{L}_{0:k} = \mathbb{L}_{0:k-1} \cup \mathbb{L}_k$  and the multi-target state is a finite subset of  $\mathbb{X} \times \mathbb{L}_{0:k}$ .

Thus, for  $N(k)$  targets and  $M(k)$  measurements, the (labeled) multi-target state at time  $k$  is given by

$$\underline{\vec{x}} = \{\vec{x}_{k,1}, \dots, \vec{x}_{k,N(k)}\} \quad (22.150)$$

and the multi-target measurement at time  $k$  is given by

$$\mathbf{y}_k = \{\vec{y}_{k,1}, \dots, \vec{y}_{k,M}\} \quad (22.151)$$

and the optimal multi-target Bayes filter is given by the recursive prediction and correction steps as

$$\begin{aligned} f_{\underline{\vec{x}}|k-1}(\underline{\vec{x}}_k | \mathbf{y}_{1:k-1}) &= \int f_{k|k-1}(\underline{\vec{x}}_k | \underline{\vec{x}}_{k-1}) f_{\underline{\vec{x}}}(\underline{\vec{x}}_{k-1} | \mathbf{y}_{1:k-1}) \delta \underline{\vec{x}}_{k-1} \\ f_{\underline{\vec{x}}}(\underline{\vec{x}}_k | \mathbf{y}_{1:k}) &= \frac{f_{\mathbf{Y}|\underline{\vec{x}}}(\mathbf{y}_k | \underline{\vec{x}}_k) f_{\underline{\vec{x}}|k-1}(\underline{\vec{x}}_k | \mathbf{y}_{1:k-1})}{\int f_{\mathbf{Y}|\underline{\vec{x}}}(\mathbf{y}_k | \underline{\vec{x}}_k) f_{\underline{\vec{x}}|k-1}(\underline{\vec{x}}_k | \mathbf{y}_{1:k-1}) \delta \underline{\vec{x}}_k} \end{aligned} \quad (22.152)$$

where  $f_{k|k-1}(\underline{\vec{x}}_k | \underline{\vec{x}}_{k-1})$  is the transition density,  $f_{\mathbf{Y}|\underline{\vec{x}}}(\mathbf{y}_k | \underline{\vec{x}}_k)$  is the multi-target likelihood function, and the set integrals are defined as previously, but now also over all labels.

Furthermore, as part of this recursion, one requires the definition of an **association map** at time  $k$  as the function  $\vartheta : \mathbb{L}_{0:k} \rightarrow \{0, 1, \dots, M(k)\}$  such that  $\vartheta(\ell) = \vartheta(\ell') > 0$  implies  $\ell = \ell'$ . Such a function can be regarded as an assignment of labels to measurements, with undetected labels assigned to 0. The set of all such association maps is denoted as  $\Theta_k$ . Thus, the subset of association maps with domain  $L$  is denoted by  $\Theta_k(L)$ .

Then, with the assumption that the initial multi-target density is in the GLMB RFS, then the *a priori* and *a posteriori* densities are also in the GLMB RFS. Thus, a closed-form solution to the multi-target Bayes filter

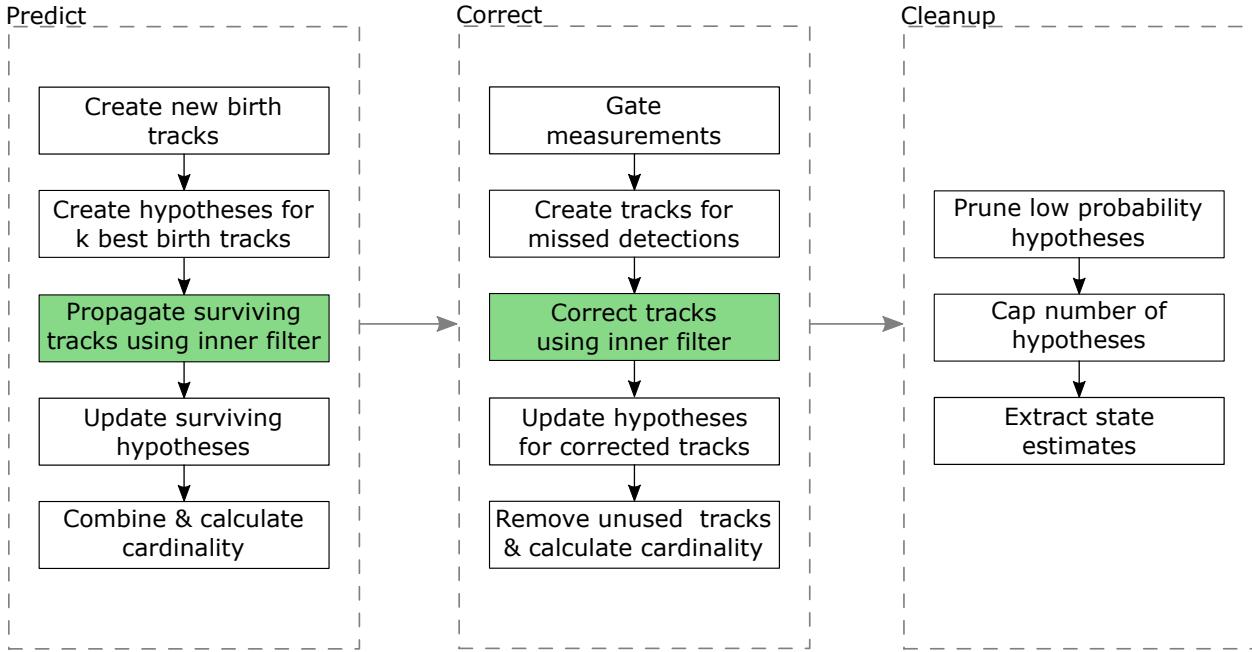
can be realized with a GLMB formulation. However, for numerical implementation, an alternative form of the GLMB known as the  **$\delta$ -generalized labeled multi-Bernoulli (GLMB) RFS**, i.e.

$$f_{\underline{\mathbf{x}}}(\underline{\mathbf{x}}) = \Delta(\underline{\mathbf{x}}) \sum_{(I, \xi) \in \mathcal{F}(\mathbb{L}) \times \Xi} w^{(I, \xi)} \delta_I(\mathcal{L}(\underline{\mathbf{x}})) [f_{\underline{\mathbf{x}}}^{(\xi)}]^\underline{\mathbf{x}} \quad (22.153)$$

which allows for numerical implementation of the labeling scheme presented above, by noting  $\Xi$  is the space of association map histories,  $\Theta_{0:k} = \Theta_0 \times \dots \times \Theta_k$ .

Here, each  $I \in \mathcal{F}(\mathbb{L}_{0:k})$  represents a set of tracks at time step  $k$ , and each  $\xi = (\vartheta_0, \dots, \vartheta_k) \in \Theta_{0:k}$  represents a history of association maps up to time step  $k$ , which also contains the history of target labels incorporating target births and deaths. The pair  $(I, \xi) \in \mathcal{F}(\mathbb{L}_{0:k}) \times \Theta_{0:k}$  is a **hypothesis** and its associated weight  $w^{(I, \xi)}$  is the hypothesis probability. The density  $f_{\underline{\mathbf{x}}}^{(\xi)}(\cdot, \ell)$  represents the kinematic state of track  $\ell$  for association map history  $\xi$ . Thus, one typical state extraction step for GLMB-based filters is to determine the MAP cardinality estimate  $\hat{N}$ , determine the MAP label set among those with cardinality  $\hat{N}$ , then extract the expected values of the states from the kinematic densities.

The GLMB filter can be outlined as follows using the bookkeeping outer filter for the hypotheses probabilities and cardinality and the inner filter for the kinematic filtering of each hypothesis.



In this fashion, the  **$\delta$ -GLMB** filter can be implemented with cubic complexity in the number of measurements. This multi-target tracker also enjoys a number of nice analytical properties, e.g. it is a necessary and sufficient statistic of a GLMB and minimizes the  $L_1$ -distance between a GLMB through its truncation, which both can be computed in closed-form. In addition, the simpler **labeled multi-Bernoulli (LMB) filter** was developed which propagated only one term in the label set significantly reducing the runtime of the GLMB filter equations. Lastly, the **joint generalized labeled multi-Bernoulli (JGLMB) filter** was developed in

2016 which combines the prediction and correction step steps of the GLMB into a single recursion with only one truncation step necessary. This JGLMB filter has sped up the LRFS-based filter considerably and is considered the state-of-the-art in multi-target tracking.

### Labeled Multi-Bernoulli Filter

#### References

For more information, please refer to the following

- A. F. Garcia-Fernandez, J. L. Williams, K. Granstrom and L. Svensson, “Poisson Multi-Bernoulli Mixture Filter: Direct Derivation and Implementation,” in *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1883-1901, Aug. 2018
- B. -T. Vo, B. -N. Vo and A. Cantoni, “The Cardinality Balanced Multi-Target Multi-Bernoulli Filter and Its Implementations,” in *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 409-423, 2009
- J. L. Williams, “Marginal Multi-Bernoulli Filters: RFS Derivation of MHT, JIPDA, and Association-Based MeMBER,” *IEEE Transactions on Aerospace and Electronic Systems*, Volume: 51, Issue: 3, 2015, pp. 1664-1687
- S. Reuter, B. -T. Vo, B. -N. Vo and K. Dietmayer, “The Labeled Multi-Bernoulli Filter,” in *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3246-3260, 2014
- B. -N. Vo, B. -T. Vo and D. Phung, “Labeled Random Finite Sets and the Bayes Multi-Target Tracking Filter,” in *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6554-6567, 2014

## 22.6 Extended Object Tracking

### Single Extended Object Tracking

### Multiple Extended Object Tracking

### Gaussian-Inverse-Wishart Filter

### Gamma-Gaussian-Inverse-Wishart Filter

---

# Monitoring Systems

## 23.1 Fuel and Battery Monitoring Systems

### Fuel Monitoring Systems

A **fuel monitoring system** estimates the fuel level of a vehicle. A typical monitoring system uses two sensors, a **fuel flow meter** and a **fuel tank level** as a simple system modeled as

$$\begin{aligned} \begin{bmatrix} f_k \\ b_k \end{bmatrix} &= \begin{bmatrix} 1 & A_{line}\Delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} f_{k-1} \\ b_{k-1} \end{bmatrix} + \begin{bmatrix} -A_{line}\Delta t \\ 0 \end{bmatrix} u_{k-1} + \vec{w} \\ y_k &= [A_{tank}^{-1} \quad 0] \begin{bmatrix} f_k \\ b_k \end{bmatrix} + v \end{aligned} \tag{23.1}$$

where

- $f_k$  is the fuel remaining in  $\text{cm}^3$ ,
- $b_k$  is the flow meter bias in  $\text{cm}$ ,
- $A_{line}$  is the cross-sectional area of the fuel line in  $\text{cm}^2$ ,
- $\Delta t$  is the sampling rate in  $\text{s}$ ,
- $u_{k-1}$  is the fuel flow meter measurement at time step  $k - 1$  in  $\text{cm/s}$ ,
- $\vec{w}$  is the zero-mean fuel flow meter measurement error in  $[\text{cm}^3, \text{cm}]^T$  with covariance  $Q$ ,
- $y_k$  is the fuel tank level measurement in  $\text{cm}$ ,
- $A_{tank}$  is the cross-sectional area of the fuel tank in  $\text{cm}^2$ , and
- $v$  is the fuel tank level measurement error in  $\text{cm}$  with variance  $R$ .

Note that in this model, the bias drift is modeled as a Gaussian random walk, but could be modeled as an *AR(1)* process or a jump process in some similar sensors.

## Battery Monitoring Systems

A model-based **battery monitoring system (BMS)** typically estimates three quantities, the **state-of-charge (SoC)** of the battery, the **state-of-health (SoH)**, and the **remaining useful life (RUL)**, using the voltage and current measurements at the nodes of the battery with three cascaded nonlinear Bayesian filters.

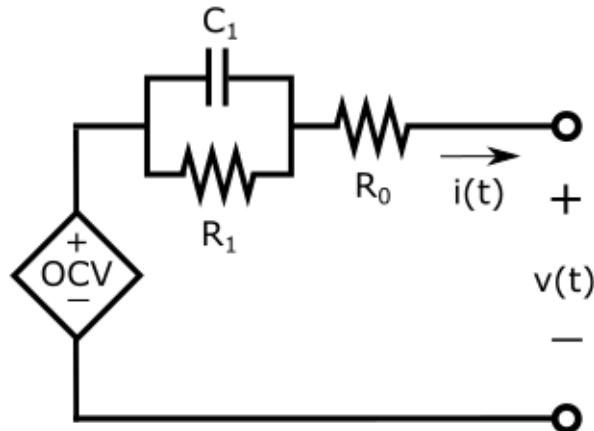
The SoC is defined as the percentage of current  $i$  remaining in the battery relative to the capacity,  $Q$ , i.e., as the

$$\dot{SoC} = -\frac{i(t)}{Q} \quad (23.2)$$

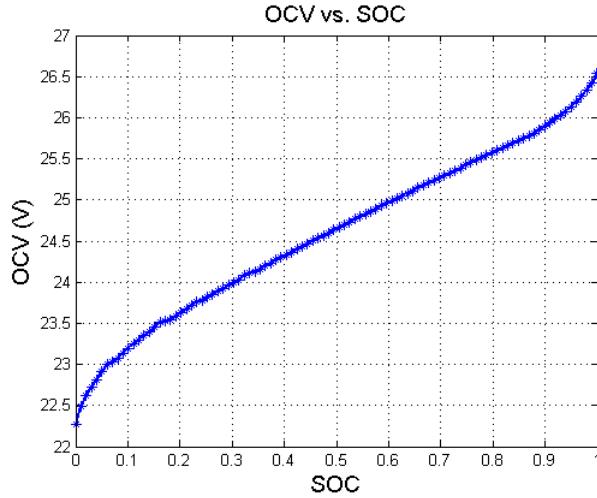
where one can use numerical integration to estimate the *SoC* if one can measure the current. Note that here  $i(t)$  should be positive for a discharging battery. Note that the total capacity of the battery can be considered as all of the possible current usable in the battery, i.e.

$$Q = \int_0^{\infty} i(t) dt \quad (23.3)$$

To relate the voltage and current measurements to the *SoC* and  $Q$ , consider the following equivalent circuit diagram of a battery.



where  $R_0$  is the internal resistance,  $\tau_1 = R_1 C_1$  is one dynamic element of the battery (there may be multiple of these),  $v(t)$  is the voltage measurement,  $i(t)$  is the current measurement, and **OCV** is the **open circuit voltage (OCV)** of the battery. Empirically, it is known that the OCV is a nonlinear function of the SoC as shown in the following figure.



which remains relatively constant over the lifetime of the battery.

With these two sensors of current and voltage, the *SoC* estimation and one *RC* element allows one to form the following state-space model

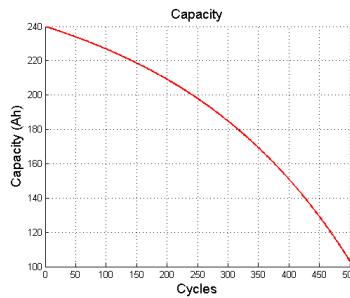
$$\begin{bmatrix} SoC_k \\ i_{1,k} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \exp^{-\frac{-\Delta t}{R_1 C_1}} \end{bmatrix} \begin{bmatrix} SoC_{k-1} \\ i_{1,k-1} \end{bmatrix} + \begin{bmatrix} \frac{-\Delta t}{Q} \\ 1 - \exp^{-\frac{-\Delta t}{R_1 C_1}} \end{bmatrix} i(k-1) \quad (23.4)$$

$$v(k) = OCV(SoC_k) - R_0 i(k) - R_1 i_1(k)$$

The *SoH* is the current value of the capacity  $Q(t)$  relative to the initial capacity,  $Q_0$ , i.e.,

$$SoH = \frac{Q(t)}{Q_0} \quad (23.5)$$

which will slowly decrease over the lifetime of the battery as shown in the following figure.



which is typically modeled as an exponential decay with the number of cycles,  $\ell$ , as a double exponential model

$$Q(\ell) = ae^{b\ell} + ce^{d\ell} \quad (23.6)$$

However, some batteries exhibit “burn-in” growth stage, a plateau stage, and then a decay stage which requires additional detection algorithms in estimation.

However, as  $Q(\ell)$  can change  $< 1\%$  during a single discharge cycle of the battery, estimating  $Q(\ell)$  continuously with  $k$  during a discharge cycle is unnecessary and can actually lead to volatility of the estimate. Thus, cascading *SoH* estimation after *SoC* estimation improves performance. To setup a cascaded Bayes filter, step propagates the *SoC* at the beginning of a discharge cycle,  $SoC_{start}$ , through the *entire*  $\ell^{\text{th}}$  discharge cycle with  $N_\ell$  time steps with current capacity estimate and current measurements, i.e.

$$\hat{SoC}_{proj} = SoC_{start} - \frac{\sum_{k=0}^{N_\ell-1} i_k}{SoHQ_0} \quad (23.7)$$

Then, one can correct the  $Q$  estimate with the observed *SoC* estimate at the end of the cycle,  $SoC_{end}$ . This provides the stochastic state-space model with a nonlinear measurement model as

$$\begin{aligned} SoH_\ell &= SoH_{\ell-1} + \vec{w}_{\ell-1} \\ SoC_\ell &= \frac{\sum_{k=0}^{N_\ell-1} i_k}{SoH_\ell Q_0} \end{aligned} \quad (23.8)$$

Notably, this method relies heavily on accurate *SoC* estimates across the entire discharge cycles.

The RUL is defined as the number of discharge cycles remaining before one reaches a *SoH* that is unusable for the system employing the battery, e.g., 50-80%. By recursively estimating the model parameters of the exponential decay model,  $\vec{x} = [a \ b \ c \ d]^T$ , the RUL can be calculated by numerically solving the equation

$$SoH_{min} = \hat{a}e^{\hat{b}RUL} + \hat{c}e^{\hat{d}RUL} \quad (23.9)$$

## 23.2 Fault Monitoring Systems

In the operation of multi-sensor systems, one or more sensors in the system may malfunction from its intended purpose, i.e. a sensor may experience a **fault**. In many designs, information systems often include **fault monitoring systems** to detect and isolate potential faulted sensor(s) from the multi-sensor data fusion algorithm, also known as **fault detection and isolation (FDI)** methods, which are either model-based or signal processing-based. Model-based FDI typically uses **stochastic state-space fault** models, i.e. for exteroceptive sensors as

$$\vec{y}_k = h_k (\vec{x}_k, \vec{v}_k, \vec{b}_k) \quad (23.10)$$

and proprioceptive sensors as

$$\vec{x}_k = f_{k-1} (\vec{x}_{k-1}, \vec{u}_{k-1}, \vec{w}_{k-1}, \vec{b}_{k-1}) \quad (23.11)$$

where  $\vec{b}$  is the **fault vector**, also known as the **bias vector**.

The fault vector may be constant or time-varying and deterministic or stochastic. Thus, one typically uses detection theory to formulate a detector and threshold based on the fault model(s) for  $\vec{b}$  and the type of data fusion algorithm.

Here, if one detects any faulted measurements and identifies the sensor associated with the faulted measurement, then one should only use the subset of fault-free measurements, thereby excluding or isolating

any faulted measurements from the fusion algorithm, a process known as **fault detection and exclusion (FDE)** or **fault detection and isolation (FDI)**. A common method for model-based FDE/FDI is called **solution separation** or  **$N$ -model redundancy** whereby one evaluates the data fusion solution for the full set of measurements and cycles through all subsets of measurements based on all fault models which notably may have single- or multiple-fault modes. Then, a detector,  $D$ , is formed from the solution based on each subset and the solution based on the full set which is then compared to some threshold,  $\tau$ . If  $D > \tau$ , then that fault-mode is declared and those faulted measurements are removed from the final data fusion output. However, these types of multi-sensor systems require

For constant, single-faults, the solution separation can be shown for OLS parameter estimation as follows. Consider  $n_y$  measurements linearly related to the parameter, i.e.

$$\vec{y} = X\vec{\beta} + \vec{v} + \vec{b} \quad (23.12)$$

where  $n_y > n_\beta$ ,  $v_i \sim \mathcal{N}(0, \sigma_v^2)$ ,  $b_i = 0$  when the  $i^{\text{th}}$  sensor measurement  $y_i$  is fault-free and  $b_i = \mu_i$  when the  $i^{\text{th}}$  sensor measurement  $y_i$  is faulted. Recall the OLS solution is

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y} \quad (23.13)$$

and the **measurement residual** is a Gaussian distributed random vector given by

$$\vec{r} = \vec{y} - \hat{\vec{y}} = \vec{y} - X\hat{\vec{\beta}} \quad (23.14)$$

where one can show

$$\vec{r} = (I - X(X^T X)^{-1} X^T)(\vec{v} + \vec{b}) \quad (23.15)$$

Then, one can define the detector as

$$D = \frac{1}{\sigma_v^2} \vec{r}^T \vec{r} \quad (23.16)$$

$$D = \frac{1}{\sigma_v^2} \left( r_1^2 + \dots + r_{n_y}^2 \right) \quad (23.17)$$

For fault-free measurements, i.e.

$$\mathbb{E}[\vec{r}] = \vec{0} \quad (23.18)$$

it can be shown that  $D \sim \chi^2(n_y - n_\beta)$ , i.e. a centralized chi-squared with  $n_y - n_\beta$  degrees-of-freedom. For single-fault measurements, i.e.

$$\mathbb{E}[\vec{r}] = [0 \ \dots \ \mu_i \ \dots \ 0]^T \quad (23.19)$$

it can be shown that  $D \sim \chi^2(n_y - n_\beta, \frac{\mu_i^2}{\sigma_v^2})$ , i.e. a noncentralized chi-squared with  $n_y - n_\beta$  degrees-of-freedom and a  $\frac{\mu_i^2}{\sigma_v^2}$  non-centrality parameter. Thus, for a given expected fault magnitude,  $\mu_i$  and  $P_{FA}$ , one can formulate a threshold,  $\tau$ , and probability of missed detection of the fault,  $P_{MD}$ , based on the  $\chi^2$  distribution models.

## Error Overbounding

Unfortunately, one does not always have accurate sensor error models used in model-based multi-sensor fusion, i.e. the probabilistic models may be difficult to exactly quantify. This leads to the concept of **overbounding** which models the probability of an error as always greater than or equal to the true probability. Mathematically, if the error is defined as a random negative or positive “distance” away from some error-free value  $\tilde{x}$ , one can model this error as the random variable  $X$  and  $B(x)$  is an **overbound** of  $X$  about  $\tilde{x}$  if

$$\begin{aligned} B(x) &\geq \Pr(X \leq x) \quad \forall x < \tilde{x} \\ B(x) &\geq \Pr(X > x) \quad \forall x > \tilde{x}. \end{aligned} \tag{23.20}$$

i.e. the overbound provides upper bounds for the probabilities of observing any negative error, i.e.  $x < \tilde{x}$ , and any positive error, i.e.  $x > \tilde{x}$ .

This separation of positive and negative error bounding is useful for the fundamental definition of overbounding since  $X$  may contain asymmetry about  $\tilde{x}$ . Furthermore, this overbound definition can be rewritten in terms of the *unknown* CDF of  $X$  i.e.

$$\begin{aligned} B(x) &\geq F_X(x) \quad \forall x < \tilde{x} \\ B(x) &\leq F_X(x) \quad \forall x > \tilde{x}. \end{aligned} \tag{23.21}$$

It should be noted that this formulation puts no requirement on the actual value of  $B(x)$  at  $x = \tilde{x}$ , which may be relevant for mixed random variables, but has no effect on continuous random variables. This section defines three types of univariate error overbounds for multi-sensor information systems. These cases are shown graphically along with the depiction of a general overbound. If  $F(x) = 0 \quad \forall x \leq \tilde{x}$ , then  $B(x)$  is the **one-sided overbound** of  $X$  if

$$B(x) \leq F_X(x) \quad \forall x > \tilde{x}. \tag{23.22}$$

If  $F(\tilde{x}) = 0.5$  where  $\tilde{x}$  = median of  $x$ , then  $B(x)$  is the **two-sided overbound** of  $X$  if

$$\begin{aligned} B(x) &\geq F_X(x) \quad \forall F_X(x) < 0.5 \\ B(x) &\leq F_X(x) \quad \forall F_X(x) > 0.5. \end{aligned} \tag{23.23}$$

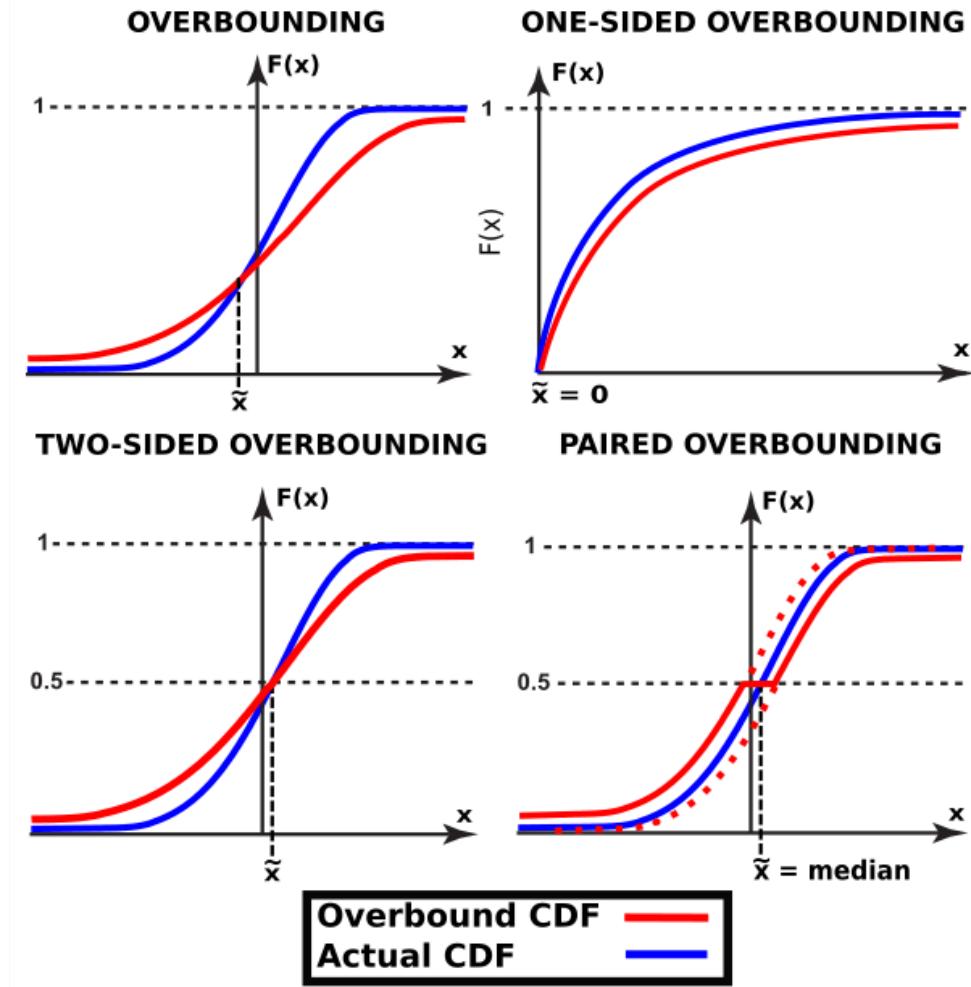
From this definition of the two-sided overbound, a common construction called paired overbounding. This approach can also be considered as a natural extension of one-sided overbounding to the two-sided case, where two separate models are used, one for each side of  $F(x) = 0.5$ . Mathematically,  $B(x)$  is called the **paired overbound** of  $X$  if

$$B(x) = \begin{cases} B_L(x) & \forall x : B_L(x) < 0.5 \\ 0.5 & \text{otherwise} \\ B_R(x) & \forall x : B_R(x) > 0.5 \end{cases} \tag{23.24}$$

where  $B_L(x)$  and  $B_R(x)$  satisfy

$$\begin{aligned} B_L(x) &\leq F_X(x) \quad \forall x \\ B_R(x) &\geq F_X(x) \quad \forall x. \end{aligned} \tag{23.25}$$

Although the mathematical formulations for the preceding overbounding definitions hold in principle, the fact that the true random variable remains *unknown* requires one estimate the overbound model of the



error. Thus, one typically uses error data to form the **empirical distribution function (EDF)** of  $X$ ,  $\hat{F}_X(x)$ , given by

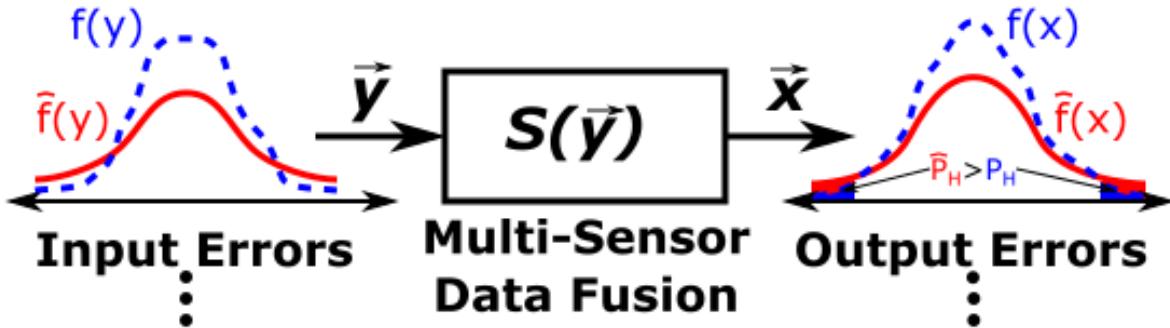
$$\hat{F}_X(x) = \begin{cases} 0, & \forall x < x_{(1)} \\ i/n, & \forall x_{(i)} \leq x < x_{(i+1)} \quad i = 1, \dots, n-1 \\ 1, & \forall x \geq x_{(n)} \end{cases} \quad (23.26)$$

and  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  is the ordered magnitudes of the samples  $x_1, \dots, x_n$ .

In addition, one typically desires to be able to compute overbounds for the output errors of multi-sensor information systems. This can be done by mapping the input error overbounds through the data fusion algorithm. When the algorithm is linear, the mapping can be computed by convolving the input error distributions. It has been proven that the convolution of one- or two-sided overbounds produce an overbound on the convolution of the true error distributions if they are unimodal, symmetric, and zero-mean. Furthermore, it has been shown that paired overbounds can remove these limitations by convolving each pair of overbounds separately. This convolution approach to output overbounding has also led to the Gaussian

distribution becoming by far the most popular overbound model primarily because convolutions of Gaussian random variables remain Gaussian random variables. However, if the overbounded distribution exhibits heavy-tailedness, i.e. it decays slower than a Gaussian distribution, then the input and output errors cannot be truly overbounded by a Gaussian overbound model.

The following figure displays the concept of applying the overbounding method for multi-sensor information systems by using conservative overbounds for calculating the probabilities of hazardously misleading information,  $P_{HMI}$ . Such a system with a linear mapping can be mathematically represented as



$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = S \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = S \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} \sum_{j=1:m} S_{1,j} y_j \\ \vdots \\ \sum_{j=1:m} S_{n,j} y_j \end{bmatrix} \quad (23.27)$$

where  $[y_1 \dots y_m]^T$  and  $[x_1 \dots x_n]^T$  are the input and output random vectors of lengths  $n$  and  $m$ , respectively, and  $S$  is the  $n \times m$  linear transformation matrix with  $S_{i,j}$  representing the element of the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column. Next, since the original inputs are first scaled, the scaled inputs can be computed by the formula for a function  $X = S(Y)$  of random variable,  $Y$ :

$$f_X(x) = f_Y(S^{-1}(y)) \left| \frac{dS^{-1}(y)}{dy} \right| \quad (23.28)$$

where if  $S(Y) = SY$ , then

$$f_X(y) = \frac{1}{|S|} f_Y \left( \frac{y}{S} \right) \quad (23.29)$$

Then, if the inputs are independent, the output elements  $x_1, \dots, x_i, \dots, x_n$  can be written as successive  $m - 1$  convolutions of the PDFs of inputs  $y_1, \dots, y_j, \dots, y_m$  where  $f_k(x)$  denotes the PDF for the  $k^{\text{th}}$  convolution which for  $1 < k \leq m - 1$  is computed by

$$f_k(u) = \int_{-\infty}^{\infty} \frac{1}{|S_{i,m-k}|} f_{y_{m-k}} \left( \frac{t_k}{S_{i,m-k}} \right) f_{k-1}(u - t_k) dt_k \quad (23.30)$$

and for  $k = 1$  is given as

$$f_1(u) = \int_{-\infty}^{\infty} \frac{1}{|S_{i,m-1}|} f_{y_{m-1}}\left(\frac{t_1}{S_{i,m-1}}\right) \frac{1}{|S_{i,m}|} f_{y_m}\left(\frac{u - t_1}{S_{i,m}}\right) dt_1 \quad (23.31)$$

where  $f_{m-1}(\cdot) = f_{x_i}(\cdot)$  is the PDF of output  $x_i$  and  $f_{y_j}(\cdot)$  is the PDF of input  $y_j$ .

Now suppose that  $y_1, \dots, y_m$  are left overbounded by CDFs  $B_{L,y_1}(y), \dots, B_{L,y_m}(y)$  and right overbounded by CDFs  $B_{R,y_1}(y), \dots, B_{R,y_m}(y)$ , respectively, then performing the same successive convolutions for the overbounding PDFs and integrating the resulting PDFs to CDFs produces left overbounds,  $B_{L,x_1}(x), \dots, B_{L,x_n}(x)$  and right overbounds  $B_{R,x_1}(x), \dots, B_{R,x_n}(x)$  for  $x_1, \dots, x_n$ , respectively. When the overbounds used for the inputs are Gaussians, then equations 23.30 and 23.31 do not need to be computed explicitly, but only the means and variances must be computed using the algebraic formulas that follow. This is the primary reason why Gaussian overbounds are simpler to handle for real-time integrity computations. The mean formula is

$$\mu_{x_i} = \sum_{j=1:m} S_{i,j} \mu_{y_j} \quad (23.32)$$

and the variance formula is

$$\sigma_{x_i}^2 = \sum_{j=1:m} S_{i,j}^2 \sigma_{y_j}^2 \quad (23.33)$$

where  $\mu_{x_i}$  and  $\mu_{y_j}$  are the means of the distributions of  $x_i$  and  $y_j$ , respectively, and  $\sigma_{x_i}^2$  and  $\sigma_{y_j}^2$  are the variances of the distributions of  $x_i$  and  $y_j$ , respectively. However, for input overbounds that are not Gaussian, then equations 23.30 and 23.31 must be computed numerically.

### 23.3 Integrity Monitoring Systems

Multi-sensor systems are often used in safety-critical operations for which it is imperative one has quantified the fault-free and faulted performance using risk analysis, namely, integrity risk and continuity risk. In this context, a **integrity monitoring system**, also known as a **integrity monitor**, assesses the safety-critical outputs for an potential **hazardously misleading information (HMI)** which is defined for specific types of safety-critical operations. **Integrity risk**, also known as the **probability of hazardously misleading information (HMI)**,  $P_{HMI}$ , is probability of a hazardously large output error occurring *without* the integrity monitor issuing an alert. As the risk of a large error can never practically be completely eliminated, it is necessary to accept a finite, albeit very small, risk of HMI. Integrity risk is divided into two sub-categories, *faulted* and *fault-free*. Faulted integrity risk is the probability that the integrity monitor will fail to issue an alert when it should due to a system fault.

Thus, one can model the integrity risk,  $P_{HMI}$ , as

$$P_{HMI} = P_{HMI|0} + P_{HMI|F}P_F + P_{MI|D}P_D \quad (23.34)$$

where  $P_{HMI|0}$  is the **probability of hazardously misleading information (HMI)** with a fault-free system,  $P_{HMI|F}$  is the **probability of hazardously misleading information (HMI)** with a faulted system,  $P_F$  is the probability of a fault,  $P_D$  is the **probability of fault detection**, and  $P_{MI|D}$  is the **probability of a**

**mis-identification** of the fault which leads to HMI *given* a system fault was detected. Furthermore, the first two terms can be grouped together to obtain

$$P_{MD} = P_{HMI|0} + P_{HMI|D}P_D \quad (23.35)$$

where  $P_{MD}$  is the **probability of missed detection** of HMI.

**Continuity risk** is the probability that an integrity monitor alert interrupts a safety-critical operation reliant on the information system. Integrity monitor alerts due to detected failures cause the failed measurements to be excluded from use, an action that may in turn reduce the information accuracy to a level at which the current operation is no longer safe to continue. In addition, integrity monitors may trigger false alerts. Thus, one can model the continuity risk,  $R_C$ , as

$$R_C = P_{FA} + P_{NI}P_D \quad (23.36)$$

where  $P_{FA}$  is the **probability of false alarm** of a HMI alert when HMI has not actually occurred and  $P_{NI}$  is the **probability of a non-isolable fault** for which the integrity monitor declares an HMI alert as it could not isolate the fault *given* one has detected a fault which could result in HMI. Continuity requirements in effect dictate integrity monitor thresholds and determine integrity monitor sensitivity. They also set maximum allowable fault probabilities for the underlying sensors.

As shown in the equations above, integrity monitors use FDI methods to determine the two terms within the integrity and continuity risks. In this context, an integrity monitor compares the safety-critical outputs to pre-defined **alert limits (ALs)**, i.e. the maximum allowable operational error beyond which the integrity monitor should issue a HMI alert. In this case, the “true” integrity risk,  $P_{HMI}$ , can be restated as the probability that, at any moment, the true output error exceeds the AL. However, as one does not typically know the exact output error statistics, e.g. the measurement noise variance  $\sigma_v^2$ , integrity monitors compute **protection levels (PLs)** using **error overbounds** of the system output. Overbounds model the probabilities of exceedance at all error levels as smaller than or equal to the true integrity risk of the system’s operation. Thus, protection levels are used to ensure with a very high probability that the provided information will not result in a hazardous situation with a very high confidence level. With these protection levels and alert limits, an integrity monitor uses a detector  $D$  to detect a system fault based on some threshold,  $\tau$ , and the integrity monitor can bound the integrity risk as

$$\Pr(PL > AL|\mathcal{H}_0) + \Pr(PL > AL \cup D < \tau|\mathcal{H}_F) \leq P_{HMI} \quad (23.37)$$

where  $\mathcal{H}_0$  is the fault-free system hypothesis and  $\mathcal{H}_F$  is the faulted system hypothesis.

As an example of an integrity monitoring system, consider one is given an alert limit,  $AL$ , for the parameter estimate,  $\vec{\beta}$  of a multi-sensor information system. Then, a single-fault **solution separation method** for integrity monitor can be outlined as follows.

1. Check measurement dimension  $n_y$  exceeds the parameter dimension  $n_\beta$
2. Determine suitable detection threshold,  $\tau$
3. Compute parameter estimate using all  $n_y$  measurements
4. FDI for  $k = 1, \dots, n_y$  measurements:
  - Compute parameter estimate with declared “fault-free” and unevaluated measurements excluding  $k$

- Compute fault detector,  $D_\alpha$ , with declared “fault-free” and unevaluated measurements excluding  $k$
  - Declare measurement  $k$  “fault-free” if  $D_\alpha < \tau$
5. Compute protection levels based on “fault-free” measurements and error overbounds:  $PL$
  6. Declare *system available* if  $PL \leq AL$

For many applications, these types of requirements on any potentially hazardously misleading information must often be computed by the multi-sensor systems during operation. Thus, a large part of *safety-critical* multi-sensor systems is the characterization of statistics of the estimate errors, i.e. the uncertainty quantification of the integrity risk. The performance of an integrity monitor is typically denoted by the **availability**, i.e. the fraction of the time that integrity and continuity requirements can be satisfied by the information system. To assess whether performance requirements are met, it is necessary to conduct an available test which confirms that integrity, and potentially continuity and accuracy as well, is met for both fault-free and faulted conditions.

# **Part V**

# **Appendices**

**A**

---

# Fundamental Mathematical Concepts

## A.1 Set Theory

A mathematical **set** is a collection of points denoted by a boldface capital letter or curly brackets, e.g.

$$\mathbf{S} = \{\dots, \omega, \dots\} \quad (\text{A.1})$$

where  $\omega$  is an **element**, also known as a **member**, of  $\mathbf{S}$  and is denoted by  $\omega \in \mathbf{S}$ . Of special note is the blackboard bold notation for the sets of the **natural numbers**  $0, 1, 2, 3, \dots$ , also known as the **counting numbers**, denoted by  $\mathbb{N}$ , the **integers**  $\dots, -3, -2, -1, 0, 1, 2, 3, \dots$  denoted by  $\mathbb{Z}$ , the **rational numbers** denoted by  $\mathbb{Q}$ , the **real numbers** denoted by  $\mathbb{R}$ , the real numbers with  $\infty$  denoted by  $\mathbb{R}_+$ , and the **complex numbers** denoted by  $\mathbb{C}$ .

The **cardinality** of set  $\mathbf{S}$ , denoted by  $|\mathbf{S}| \geq 0$ , is the number of elements in a set  $\mathbf{S}$ . Cardinality may be finite or infinite. The **empty set**, denoted by  $\emptyset$ , has cardinality zero and is also known as the **null set**. A set  $\mathbf{S}$  is a **finite set** if its cardinality is finite. A nonempty set is a **countably infinite set** if the elements  $\omega \in \mathbf{S}$  have a one-to-one correspondence with the natural numbers,  $\mathbb{N}$ , i.e., one can construct some infinite sequence,  $\omega_0, \omega_1, \omega_2, \dots$  where  $\omega_i$  are distinct and indexed by  $\mathbb{N}$ . A nonempty set  $\mathbf{S}$  is **countable** if  $\mathbf{S}$  is a finite set or a countably infinite set.

Let  $\mathbf{A}$  be another set of elements. If every point in  $\mathbf{A}$  also belongs contained within  $\mathbf{S}$ , then  $\mathbf{A}$  is a **subset** of  $\mathbf{S}$  denoted by  $\mathbf{A} \subseteq \mathbf{S}$ . If  $\mathbf{S} \subseteq \mathbf{A}$  also, then  $\mathbf{S} = \mathbf{A}$ . If  $\mathbf{S} \not\subseteq \mathbf{A}$ , then  $\mathbf{A} \subset \mathbf{S}$  is a **proper subset** as  $\mathbf{A} = \mathbf{S}$  is not possible.

For  $\mathbf{A} \subset \mathbf{S}$  and  $\omega \in \mathbf{S}, \omega \notin \mathbf{A}$  denotes that  $\omega$  is not in  $\mathbf{A}$ . The **complement** of  $\mathbf{A}$ , denoted  $\mathbf{A}^c$ , is defined as

$$\mathbf{A}^c = \{\omega \in \mathbf{S} : \omega \notin \mathbf{A}\} \quad (\text{A.2})$$

Note that  $\mathbf{S}^c = \emptyset$  and for any  $\mathbf{A} \subset \mathbf{S}, \emptyset \in \mathbf{A}$ .

A **function**, denoted by  $f : \mathbf{X} \rightarrow \mathbf{Y}$ , consists of a set  $\mathbf{X}$  of admissible inputs called the **domain**, a set  $\mathbf{Y}$  called the **co-domain**, and a **mapping**  $f$  that associates a value  $f(x) \in \mathbf{Y}$  to each  $x \in \mathbf{X}$ . This is typically

stated as “ $f$  maps  $\mathbf{X}$  into  $\mathbf{Y}$ .” The **range** is the set of all possible values of  $f(x)$ , i.e.,

$$\{f(x) : x \in \mathbf{X}\} \subseteq \mathbf{Y} \quad (\text{A.3})$$

where the range may not be equal to the co-domain. A function is **onto** if for every  $y \in \mathbf{Y}$ , the equation  $f(x) = y$  has *at least* one solution, i.e.,  $\{f(x) : x \in X\} = \mathbf{Y}$ . A function is **one-to-one** if for every  $y \in \mathbf{Y}$ , the equation  $f(x) = y$  has *at most* one solution. A function is **invertible** if for every  $y \in \mathbf{Y}$ , the equation  $f(x) = y$  has one *unique* solution, i.e., it is onto and one-to-one. Two functions are the same if and only if they have the same domain, co-domain, and mapping. Two mappings from the same domain and co-domain are the same if and only if they associate the same value to each  $x \in \mathbf{X}$ .

The **intersection** of two subsets  $\mathbf{A} \subseteq \mathbf{S}$  and  $\mathbf{B} \subseteq \mathbf{S}$  is

$$\mathbf{A} \cap \mathbf{B} = \{\omega \in \mathbf{S} : \omega \in \mathbf{A} \text{ and } \omega \in \mathbf{B}\} \quad (\text{A.4})$$

which can be generalized to collections of subsets,  $\mathbf{A}_1, \mathbf{A}_2, \dots \subseteq \mathbf{S}$ , as

$$\bigcap_i \mathbf{A}_i = \{\omega \in \mathbf{S} : \omega \in \mathbf{A}_i \forall i\} \quad (\text{A.5})$$

Note that if  $\mathbf{A} \subseteq \mathbf{B}$ , then  $\mathbf{A} \cap \mathbf{B} = \mathbf{A}$ . Two subsets  $\mathbf{A} \subseteq \mathbf{S}$  and  $\mathbf{B} \subseteq \mathbf{S}$  are **disjoint** or **mutually exclusive** if

$$\mathbf{A} \cap \mathbf{B} = \emptyset \quad (\text{A.6})$$

and the infinite collection of subsets  $\mathbf{A}_1, \mathbf{A}_2, \dots \subseteq \mathbf{S}$  are **pairwise disjoint** or **mutually exclusive** if

$$\mathbf{A}_i \cap \mathbf{A}_j = \emptyset \quad \forall i \neq j \quad (\text{A.7})$$

The **union** of two subsets  $\mathbf{A} \subseteq \mathbf{S}$  and  $\mathbf{B} \subseteq \mathbf{S}$  is

$$\mathbf{A} \cup \mathbf{B} = \{\omega \in \mathbf{S} : \omega \in \mathbf{A} \text{ or } \omega \in \mathbf{B}\} \quad (\text{A.8})$$

with “or” inclusive which can be generalized to collections of subsets,  $\mathbf{A}_1, \mathbf{A}_2, \dots \subseteq \mathbf{S}$ , as

$$\bigcup_i \mathbf{A}_i = \{\omega \in \mathbf{S} : \omega \in \mathbf{A}_i \text{ for some } i\} \quad (\text{A.9})$$

A **partition** is a collection of pairwise disjoint, nonempty sets,  $\mathbf{A}_1, \mathbf{A}_2, \dots \subseteq \mathbf{S}$ , with

$$\bigcup_i \mathbf{A}_i = \mathbf{S} \quad (\text{A.10})$$

The **set difference** to two subsets  $\mathbf{A} \subseteq \mathbf{S}$  and  $\mathbf{B} \subseteq \mathbf{S}$  is

$$\mathbf{B} \setminus \mathbf{A} = \mathbf{B} \cap \mathbf{A}^c \quad (\text{A.11})$$

The **commutative laws** for two subsets  $\mathbf{A} \subseteq \mathbf{S}$  and  $\mathbf{B} \subseteq \mathbf{S}$  are

$$\mathbf{A} \cap \mathbf{B} = \mathbf{B} \cap \mathbf{A} \quad \text{and} \quad \mathbf{A} \cup \mathbf{B} = \mathbf{B} \cup \mathbf{A} \quad (\text{A.12})$$

The **associative laws** for three subsets  $\mathbf{A} \subseteq \mathbf{S}$ ,  $\mathbf{B} \subseteq \mathbf{S}$ , and  $\mathbf{C} \subseteq \mathbf{S}$  are

$$\mathbf{A} \cap (\mathbf{B} \cap \mathbf{C}) = (\mathbf{A} \cap \mathbf{B}) \cap \mathbf{C} \quad \text{and} \quad \mathbf{A} \cup (\mathbf{B} \cup \mathbf{C}) = \mathbf{A} \cup (\mathbf{B} \cup \mathbf{C}) \quad (\text{A.13})$$

The **distributive laws** for three subsets  $\mathbf{A} \subseteq \mathbf{S}$ ,  $\mathbf{B} \subseteq \mathbf{S}$ , and  $\mathbf{C} \subseteq \mathbf{S}$  are

$$\mathbf{A} \cap (\mathbf{B} \cup \mathbf{C}) = (\mathbf{A} \cap \mathbf{B}) \cup (\mathbf{A} \cap \mathbf{C}) \quad \text{and} \quad \mathbf{A} \cup (\mathbf{B} \cap \mathbf{C}) = (\mathbf{A} \cup \mathbf{B}) \cap (\mathbf{A} \cup \mathbf{C}) \quad (\text{A.14})$$

which can be generalized to infinite collections of subsets,  $\mathbf{A}, \mathbf{B}_1, \mathbf{B}_2, \dots \subseteq \mathbf{S}$ , as

$$\mathbf{A} \cap \left( \bigcup_{i=1}^{\infty} \mathbf{B}_i \right) = \bigcup_{i=1}^{\infty} (\mathbf{A} \cap \mathbf{B}_i) \quad \text{and} \quad \mathbf{A} \cup \left( \bigcap_{i=1}^{\infty} \mathbf{B}_i \right) = \bigcap_{i=1}^{\infty} (\mathbf{A} \cup \mathbf{B}_i) \quad (\text{A.15})$$

**De Morgan's laws** for three subsets  $\mathbf{A} \subseteq \mathbf{S}$ ,  $\mathbf{B} \subseteq \mathbf{S}$ , and  $\mathbf{C} \subseteq \mathbf{S}$  are

$$(\mathbf{A} \cap \mathbf{B})^c = \mathbf{A}^c \cup \mathbf{B}^c \quad \text{and} \quad (\mathbf{A} \cup \mathbf{B})^c = \mathbf{A}^c \cap \mathbf{B}^c \quad (\text{A.16})$$

which can be generalized to infinite collections of subsets,  $\mathbf{A}_1, \mathbf{A}_2, \dots \subseteq \mathbf{S}$ , as

$$\left( \bigcap_{i=1}^{\infty} \mathbf{A}_i \right)^c = \bigcup_{i=1}^{\infty} \mathbf{A}_i^c \quad \text{and} \quad \left( \bigcup_{i=1}^{\infty} \mathbf{A}_i \right)^c = \bigcap_{i=1}^{\infty} \mathbf{A}_i^c \quad (\text{A.17})$$

## References

For more information, please refer to the following

- Gubner, J. A., “1.2 Review of set notation,” in *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, 2006

## A.2 Linear Algebra

### Vector and Matrix Definitions

This textbook denotes **scalars** by unbolded, block lowercase letters, e.g.  $x$ .

This textbook denotes **vectors** with **dimension**  $n_x$  by unbolded block lowercase letters with an arrow above, e.g.,  $\vec{x}$ , which takes on values from some set  $\mathbf{Y}$ , denoted as  $\vec{x} \in \mathbf{Y}^{n_x}$ , and is represented in a bracketed column as

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_{n_x} \end{bmatrix} \quad (\text{A.18})$$

where  $x_i$  denotes the scalar **element** of  $\vec{x}$  in the  $i^{\text{th}}$  row.  $\mathbf{Y}$  is a **vector space** defined as a set of vectors whose members can be added and scaled and still belong to the vector space. A **real vector** with dimension  $n_x$  takes on  $n_x$  real number values and is denoted as  $\vec{x} \in \mathbb{R}^{n_x}$ . A **complex vector** with dimension  $n_x$  takes on

$n_x$  complex number values and is denoted as  $\vec{x} \in \mathbb{C}^{n_x}$ . Note that this notation exclusively uses vectors as **column vectors** where a **row vector** is denoted by the **transpose**  $\vec{x}^T \in \mathbb{R}^{n_x}$ , i.e.,

$$\vec{x}^T = [x_1 \quad \cdots \quad x_{n_x}] \quad (\text{A.19})$$

This textbook denotes **matrices** by unbolded, block capital letters, e.g.,  $A$ , which takes on values from some set  $\mathbf{Y}$  with a **row dimension**  $m$  and a **column dimension**,  $n$ , is denoted as  $A \in \mathbf{Y}^{m \times n}$ , and is represented in a bracketed table as

$$A = \begin{bmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{bmatrix} \quad (\text{A.20})$$

where  $A_{i,j}$  denotes the scalar **element** of  $A$  in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column. The **diagonal**, also known as the **main diagonal****principal diagonal****primary diagonal****leading diagonal****major diagonal**, is the list of entries  $A_{i,j}$  where  $i = j$ . The **off-diagonal** is the list of entries  $A_{i,j}$  where  $i \neq j$ . A **real matrix** with  $m$  rows and  $n$  columns takes on  $mn$  real number values and is denoted as  $A \in \mathbb{R}^{m \times n}$ . A **complex matrix** with  $m$  rows and  $n$  columns takes on  $mn$  complex values and is denoted as  $A \in \mathbb{C}^{m \times n}$ .

The **zero matrix** of dimensions  $m \times n$  is defined as

$$[0]_{m \times n} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \quad (\text{A.21})$$

and the related **zero vector** is defined as

$$\vec{0} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \quad (\text{A.22})$$

$A$  is a **square matrix** if it has an equal number of rows and columns, i.e.,  $m = n$ .

The **identity matrix** is a square matrix of dimension  $n$  is defined as

$$I_n = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \vdots & 1 \end{bmatrix} \quad (\text{A.23})$$

The **matrix transpose** of  $A$  is  $A^T = B$  which assigns  $B_{i,j} = A_{j,i}$ .

The **conjugate matrix transpose** or **Hermitian transpose** of  $A$  is  $A^* = A^H = B$  which assigns  $B_{i,j} = A_{j,i}^*$ .

The **matrix inverse** of  $A$  is  $A^{-1}$  and is the solution to  $A^{-1}A = I$  which may or may not exist.

A useful **matrix inversion lemma** for the inverse of the sum of two matrices is

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}C)^{-1}DA^{-1} = A^{-1}B(C^{-1} + DA^{-1}C)^{-1}(BC)^{-1} \quad (\text{A.24})$$

$A$  is an **orthogonal matrix** if  $A^{-1} = A^T$ .

$A$  is an **unitary matrix** if  $A^{-1} = A^*$ .

$A$  is a **symmetric matrix** if  $A = A^T$ . This class of matrices is denoted as  $A \in \mathbb{S}^n$ .

$A$  is a **skew-symmetric matrix** if  $A = -A^T$  which notably requires  $A$  to be square and its diagonal to contain only zero elements.

$A$  is a **Hermitian matrix** if  $A = A^*$ . This class of matrices is denoted as  $A \in \mathbb{H}^n$ .

$A$  is a **diagonal matrix** if all off-diagonal entries are zero, i.e., for  $i \neq j$ ,  $A_{i,j} = 0$ .

$A$  is an **upper triangular matrix** if for  $i > j$ ,  $A_{i,j} = 0$ .

$A$  is a **lower triangular matrix** if for  $i < j$ ,  $A_{i,j} = 0$ .

The **rank** of a matrix  $M$ , denoted by  $\text{rank}(M)$ , is a measure of the “non-degenerateness” of the system of linear equations encoded by  $M$ . More formally, this can be expressed mathematically as the maximal number of linearly independent rows/columns of  $M$  or as the **dimension of the vector space** spanned by the rows/columns of  $M$ . Vector spaces are the formal subject of linear algebra and are well characterized by its dimension, or the number of linearly independent “directions” in the space. Thus, the rank of a matrix can describe the dimension of the column space, i.e. linearly independent columns, of  $M$  as the **column rank** and the dimension of the row space, i.e. linearly independent rows, of  $M$  as the **row rank**. A fundamental theorem of linear algebra is that column rank is equal to row rank. Lastly,  $M$  is said to have **full rank** if its rank equals the lesser of the number of rows and columns, i.e.  $\text{rank}(M) = \min(m, n)$ . Conversely,  $M$  is said to be **rank deficient** if it does not have full rank.

$A$  is **injective** if  $A^*A$  is invertible, i.e.,  $A$  has linearly independent columns.

$A$  is **surjective** if  $AA^*$  is invertible, i.e.,  $A$  has linearly independent rows.

The matrix **pseudoinverse**, also known as the **Moore-Penrose inverse** of  $A$  is  $A^+$  and must satisfy the following four Moore-Penrose conditions:

$$\begin{aligned} AA^+A &= A \\ A^+AA^+ &= A^+ \\ (AA^+)^* &= AA^+ \\ (A^+A)^* &= A^+A \end{aligned} \tag{A.25}$$

If  $A$  is injective, then the pseudoinverse can be computed as

$$A^+ = (A^*A)^{-1}A^* \tag{A.26}$$

If  $A$  is surjective, then the pseudoinverse can be computed as

$$A^+ = A^*(AA^*)^{-1} \tag{A.27}$$

The vector **dot product**, also known as the **vector inner product**, is defined as

$$\vec{v} \cdot \vec{w} = \vec{w} \cdot \vec{v} = \|\vec{v}\| \|\vec{w}\| \cos \theta = \vec{v}^T \vec{w} \tag{A.28}$$

The projection of  $\vec{v}$  onto  $\vec{w}$  is defined as

$$\text{proj}_{\vec{w}} \vec{v} = \vec{v}^T \frac{\vec{w}}{\|\vec{w}\|} \tag{A.29}$$

The vector **cross product** is defined as

$$\vec{v} \times \vec{w} = \|\vec{v}\| \|\vec{w}\| \sin \theta \vec{n}_{\vec{v}, \vec{w}} = [\vec{v}]_{\times} \vec{w} = \begin{bmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} v_2 w_3 - v_3 w_2 \\ v_3 w_1 - v_1 w_3 \\ v_1 w_2 - v_2 w_1 \end{bmatrix} \tag{A.30}$$

where  $\theta$  is the angle between vectors  $\vec{v} = [v_1 \ v_2 \ v_3]^T$  and  $\vec{w} = [w_1 \ w_2 \ w_3]^T$ ,  $\vec{n}_{\vec{v}, \vec{w}}$  is the unit vector normal to plane defined by  $\vec{v}$  and  $\vec{w}$ , and  $[\bullet]_{\times}$  denotes the skew-symmetric **cross product matrix** which operates on the vector  $\bullet$ .

Matrix addition of  $A$  and  $B$  with same dimensions  $m \times n$ :

$$C = A + B : C_{i,j} = A_{i,j} + B_{i,j} \quad 1 \leq i \leq m, 1 \leq j \leq n \quad (\text{A.31})$$

Matrix multiplication of  $A$  with dimensions  $m \times n$  by scalar  $b$ :

$$C = bA : C_{i,j} = bA_{i,j} \quad 1 \leq i \leq m, 1 \leq j \leq n \quad (\text{A.32})$$

Matrix multiplication of  $A$  with dimensions  $m \times n$  by matrix  $B$  with dimensions  $n \times p$ :

$$C = AB : C_{i,j} = \sum_{k=1}^n A_{i,k}B_{k,j} \quad 1 \leq i \leq m, 1 \leq j \leq p \quad (\text{A.33})$$

where  $C$  has dimensions  $m \times p$ . Notably, matrix multiplication satisfies associativity,  $(AB)C = A(BC)$ , right distributivity,  $(A+B)C = AC + BC$ , and left distributivity,  $C(A+B) = CA + CB$ , but not associativity,  $AB \neq BA$ . Furthermore, one has

$$C^T = (AB)^T = B^T A^T \quad (\text{A.34})$$

From this definition, one can specify matrix multiplication of  $A$  with dimensions  $m \times n$  by (column) vector  $\vec{x}$  with dimension  $n$ :

$$\vec{y} = A\vec{x} : y_i = \sum_{k=1}^n A_{i,k}x_k \quad 1 \leq i \leq m \quad (\text{A.35})$$

where  $\vec{y}$  has a dimension of  $m$ . Matrix multiplication by (row) vector  $\vec{x}^T$  with dimension  $m$ :

$$\vec{y}^T = \vec{x}^T A : y_j = \sum_{k=1}^m x_k A_{k,j} \quad 1 \leq j \leq n \quad (\text{A.36})$$

where  $\vec{y}$  has a dimension of  $n$ .

In this context of column vector multiplication,  $A$  is known as a **linear transformation**, also known as a **linear map**, and is represented as  $f : X \rightarrow Y$  where  $X$  is the domain and  $Y$  is the codomain of the function  $f$ . The **null space**, also known as the **kernel**, of  $A$  is the part of the domain  $X$  that is mapped to the zero vector  $\vec{0}$  of the codomain  $Y$  and its dimension is known as the **nullity**. The **rank-nullity theorem** states that the rank plus the nullity of a matrix is equal to its number of columns.

## Vector and Matrix Norms

The **vector norm**,  $\|\vec{x}\|$ , of any vector  $\vec{x} \in \mathbb{R}^{n_x}$  is a real valued function from  $\mathbb{R}^{n_x} \rightarrow \mathbb{R}$  with the following properties

1.  $\|\vec{x}\| \geq 0$ ;
2.  $\|\vec{x}\| = 0$  if and only if  $\vec{x}$  is the zero vector in  $\mathbb{R}^{n_x}$ ;
3. for any  $\lambda \in \mathbb{R}$ ,  $\|\lambda \vec{x}\| = |\lambda| \|\vec{x}\|$ ; and

4. for any  $\vec{y} \in \mathbb{R}^{n_x}$ , the **triangle inequality**

$$\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\| \quad (\text{A.37})$$

holds.

An important class of vector norms are the  **$L_p$ -norms**, also known as  **$p$ -norms**, defined as

$$\|\vec{x}\|_p = \left( \sum_{i=1}^{n_x} |x_i|^p \right)^{\frac{1}{p}}, \quad 1 \leq p \leq \infty \quad (\text{A.38})$$

Of particular interest are the  **$L_1$ -norm**, also known as the **taxicab norm**,

$$\|\vec{x}\|_1 = \sum_{i=1}^{n_x} |x_i| \quad (\text{A.39})$$

the  **$L_2$ -norm**, also known as the **Euclidean norm**,

$$\|\vec{x}\|_2 = \left( \sum_{i=1}^{n_x} x_i^2 \right)^{\frac{1}{2}} = \sqrt{\vec{x}^T \vec{x}} = \sqrt{\begin{bmatrix} x_1 & \cdots & x_{n_x} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_{n_x} \end{bmatrix}} \quad (\text{A.40})$$

and the  **$L_\infty$ -norm**, also known as the **vector maximum norm**,

$$\|\vec{x}\|_\infty = \max |x_i| \quad (\text{A.41})$$

As an aside, another popular norm is the **weighted  $L_2$ -norm**,

$$\|\vec{x}\|_W = \sqrt{\vec{x}^T W \vec{x}} \quad (\text{A.42})$$

where  $W$  is known as the **weight matrix**.

The **matrix norm**,  $\|A\|$ , of the matrix  $A \in \mathbb{R}^{m \times n}$  is a real valued function from  $\mathbb{R}^{m \times n} \rightarrow \mathbb{R}$  with the following properties

1.  $\|A\| \geq 0$ ;
2.  $\|A\| = 0$  if and only if  $A$  is the zero matrix in  $\mathbb{R}^{m \times n}$ ;
3. for any  $\lambda \in \mathbb{R}$ ,  $\|\lambda A\| = |\lambda| \|A\|$ ; and
4. for any  $B \in \mathbb{R}^{m \times n}$ , the **triangle inequality**

$$\|A + B\| \leq \|A\| + \|B\| \quad (\text{A.43})$$

holds.

An important class of matrix norms are **induced matrix norms** which are defined for a matrix  $A$  and some specified norm  $\|\vec{x}\|$  as

$$\|A\| = \sup_{\vec{x} \neq 0} \frac{\|A\vec{x}\|}{\|\vec{x}\|} \quad (\text{A.44})$$

where sup stands for the **supremum**, also known as the **least upper bound** of the specified set. Typically one uses  $L^p$ -norms for induced matrix norms which can be written as

$$\|A\|_p = \sup_{\vec{x} \neq 0} \frac{\|A\vec{x}\|_p}{\|\vec{x}\|_p} \quad (\text{A.45})$$

Another important class of matrix norms are **entry-wise matrix norms** which treats a matrix  $A$  as a vector of size  $m \times n$  and uses a vector norm. One such norm is the  $L_{p,q}$ -norms defined as

$$\|A\|_{p,q} = \left( \sum_{j=1}^n \left( \sum_{i=1}^m |A_{i,j}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} \quad (\text{A.46})$$

where  $A_{i,j}$  denotes the element of  $A$  in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column. The  $L_{2,2}$ -norm is also known as the **Frobenius norm** and can be defined as

$$\|A\|_{2,2} = \|A\|_F = \left( \sum_{i=1}^m \sum_{j=1}^n |A_{i,j}|^2 \right)^{\frac{1}{2}} \quad (\text{A.47})$$

In addition the  $L_{\infty,\infty}$ -norm is also known as the **matrix maximum norm** and can be defined as

$$\|A\|_{\max} = \max_{i,j} |A_{i,j}| \quad (\text{A.48})$$

## Matrix Decompositions

**Eigenvalue decomposition** of the state matrix. A **eigenvalue**,  $\lambda$ , with associated **(left) eigenvector**,  $\vec{v}$ , of a square matrix  $A$  is defined as a solution to the **eigenvalue problem** defined as “given some  $A$ , for what values of  $\lambda$  and  $\vec{v}$  does the following equation hold:”

$$A\vec{v} = \lambda\vec{v} \quad (\text{A.49})$$

To solve this problem, one can rewrite the equation above as

$$\lambda\vec{v} - A\vec{v} = 0 \quad (\text{A.50})$$

$$[\lambda I - A]\vec{v} = 0 \quad (\text{A.51})$$

which for nontrivial solutions, i.e.  $\vec{v} \neq 0$ , one must solve for

$$\det(\lambda I - A) = 0 \quad (\text{A.52})$$

to obtain the eigenvalues of  $A$ , i.e.  $\lambda(A)$ . Then, by substituting back into  $[\lambda I - A]\vec{v} = 0$ , one can also obtain the associated (left) eigenvectors.

In general,  $\lambda$  and  $\vec{v}$  can be complex-valued even if  $A$  is real-valued. If the real parts of all  $\lambda(A)$  are  $< 0$   $A$  is known as a **stable matrix** or **Hurwitz matrix**. For symmetric matrices, one can define the following:

- $A \in \mathbb{S}^n$  is **positive definite**, denoted by  $A > 0$ , if the real parts of all  $\lambda(A)$  are  $> 0$ ;
- $A \in \mathbb{S}^n$  is **positive semi-definite**, denoted by  $A \geq 0$ , if the real parts of all  $\lambda(A)$  are  $\geq 0$ ;
- $A \in \mathbb{S}^n$  is **negative definite**, denoted by  $A < 0$ , if the real parts of all  $\lambda(A)$  are  $< 0$ ;
- $A \in \mathbb{S}^n$  is **negative semi-definite**, denoted by  $A \leq 0$ , if the real parts of all  $\lambda(A)$  are  $\leq 0$ ; and
- $A \in \mathbb{S}^n$  is **indefinite** otherwise.

Furthermore, a square matrix  $A$  with dimensions  $n \times n$  is **diagonalizable**, if an eigenvalue decomposition can be performed as

$$A = V\Lambda V^{-1} \quad (\text{A.53})$$

where the  $n$  corresponding *non-repeated* eigenvalues of  $A$  makeup a diagonal  $\Lambda$  as

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \quad (\text{A.54})$$

the  $n$  linearly independent right eigenvectors of  $A$  makeup  $V$  as

$$V = [\vec{v}_1 \quad \cdots \quad \vec{v}_n] \quad (\text{A.55})$$

and the  $n$  linearly independent left eigenvectors of  $A$  makeup  $V^{-1}$  as

$$V^{-1} = \begin{bmatrix} \vec{\mu}_1^T \\ \vdots \\ \vec{\mu}_n^T \end{bmatrix} \quad (\text{A.56})$$

However, if there are *repeated* eigenvalues of  $A$ , then  $A$  may not be diagonalizable. Then, a **Jordan matrix** can be used to form  $\Lambda$  in a similar fashion to an eigenvalue decomposition where a Jordan matrix,  $J$ , is defined as

$$J = \begin{bmatrix} J_1 & 0 & \cdots & 0 \\ 0 & J_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_k \end{bmatrix} \quad (\text{A.57})$$

where the 0's are zero-valued matrices and the  $k$  **Jordan blocks**,  $J_k$ , are specified by dimension  $r$  and eigenvalue  $\lambda_r$ , i.e.

$$J_k(r, \lambda_r) = \begin{bmatrix} \lambda_r & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_r & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda_r & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_r & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_r \end{bmatrix} \quad (\text{A.58})$$

For diagonalizable matrices, the Jordan matrix is purely diagonal since each constituent Jordan block is  $1 \times 1$ . The Jordan matrix is useful in forming the **Jordan Canonical Form (JCF)** of LTI state-space systems through the substitution

$$\vec{x} = V \vec{z} \quad (\text{A.59})$$

which allows the continuous-time LTI state-space system to be rewritten as

$$\begin{aligned} \dot{V} \vec{z}(t) &= AV \vec{z}(t) + B \vec{u}(t) \\ \vec{y}(t) &= CV \vec{z}(t) + D \vec{u}(t) \end{aligned} \quad (\text{A.60})$$

$$\begin{aligned} \dot{\vec{z}}(t) &= V^{-1} AV \vec{z}(t) + V^{-1} B \vec{u}(t) \\ \vec{y}(t) &= CV \vec{z}(t) + D \vec{u}(t) \end{aligned} \quad (\text{A.61})$$

$$\begin{aligned} \dot{\vec{z}}(t) &= \Lambda \vec{z}(t) + \bar{B} \vec{u}(t) \\ \vec{y}(t) &= \bar{C} \vec{z}(t) + D \vec{u}(t) \end{aligned} \quad (\text{A.62})$$

where  $\Lambda$  is a Jordan matrix (i.e. diagonal or nearly diagonal), and  $\bar{B}$  and  $\bar{C}$  are new input and output matrices, respectively. To obtain  $V$ , one must solve for the **generalized eigenvectors** for each Jordan block,  $J_k(r, \lambda_r)$ , which solve

$$\begin{aligned} (A - \lambda_r I) \vec{v}_1 &= 0 \\ (A - \lambda_r I) \vec{v}_2 &= \vec{v}_1 \\ &\vdots \\ (A - \lambda_r I) \vec{v}_r &= \vec{v}_{r-1} \end{aligned} \quad (\text{A.63})$$

When considering the full LTI system behavior from input to output using the transfer function matrix, one often uses the **singular value decomposition (SVD)** defined for any  $A \in \mathbb{C}^{m \times n}$  as

$$A = U \Sigma V^{-1} \quad (\text{A.64})$$

where  $U$  is an  $m \times m$  orthogonal matrix,  $\Sigma$  is a diagonal  $m \times n$  matrix with *non-negative* real numbers on the diagonal, i.e.

$$\Sigma = \begin{cases} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \dots \\ 0 & 0 \end{bmatrix} & m < n \\ \begin{bmatrix} \Sigma_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} & m > n \\ \Sigma_1 & m = n \end{cases} \quad (\text{A.65})$$

$$\Sigma_1 = \text{diag}(\sigma_1, \dots, \sigma_k) \quad (\text{A.66})$$

and  $V$  is an  $n \times n$  orthogonal matrix. The diagonal entries of  $\Sigma$  are known as the **singular values** of  $A$  and are ordered in size with  $\tilde{\sigma} = \sigma_1 \geq \dots \geq \sigma_k = \underline{\sigma}$  with  $k = \min(m, n)$  singular values because of its non-square nature. In addition, one can say that  $U$  and  $V$  are unitary,  $U^{-1} = U^*$  and  $V^{-1} = V^*$ . Thus the SVD could also be written as

$$A = U \Sigma V^* \quad (\text{A.67})$$

which is much simpler to compute.

The singular values  $\sigma_i$  for  $A$  are the non-negative real numbers for which there exists unit-length vectors  $\vec{u}_i$  and  $\vec{v}_i$  such that

$$A \vec{v}_i = \sigma_i \vec{u}_i \quad (\text{A.68})$$

and

$$\vec{u}_i^T A = \sigma_i \vec{v}_i^T \quad \text{or} \quad A^T \vec{u}_i = \sigma_i \vec{v}_i \quad (\text{A.69})$$

where  $\vec{v}_i$  is the right-singular vector for  $\sigma_i$  and all together make up the columns of  $V$ , i.e.

$$V = [\vec{v}_1 \quad \cdots \quad \vec{v}_n] \quad (\text{A.70})$$

and  $\vec{u}_i$  is the left-singular vector for  $\sigma_i$  and all together make up the columns of  $U$ , i.e.

$$U = [\vec{u}_1 \quad \cdots \quad \vec{u}_m] \quad (\text{A.71})$$

The singular value problem is similar to the eigenvalue problem except one has two related problems due to the non-square nature of  $A$ . The  $k$  nonzero singular values of  $A$ ,  $\sigma_i(M)$ , can be shown to be related to the eigenvalue decomposition by

$$\sigma_i(M) = \sqrt{\lambda_i(A^*A)} = \sqrt{\lambda_i(AA^*)} > 0 \quad (\text{A.72})$$

and

$$\begin{aligned} A^*A \vec{v}_i &= \sigma_i^2 \vec{v}_i \\ AA^* \vec{u}_i &= \sigma_i^2 \vec{u}_i \end{aligned} \quad (\text{A.73})$$

Thus, all  $\sigma_i^2$  are eigenvalues of  $AA^*$  and  $A^*A$ , all  $\vec{v}_i$  are eigenvectors of  $A^*A$ , and all  $\vec{u}_i$  are eigenvectors of  $MM^*$ .

The **maximum singular value** of  $A$  can be shown to be

$$\bar{\sigma}(A) = \sup_{\vec{x} \neq 0} \frac{\|A\vec{x}\|_2}{\|\vec{x}\|_2} = \|A\|_2 \quad (\text{A.74})$$

while the **minimum singular value** of  $A$  can be shown to be

$$\underline{\sigma}(A) = \inf_{\vec{x} \neq 0} \frac{\|A\vec{x}\|_2}{\|\vec{x}\|_2} \quad (\text{A.75})$$

The  $\bar{\sigma}(A)$ , i.e.  $\|A\|_2$ , represents how “big”  $A$  is or how large the “gain” of  $A$  is. The  $\underline{\sigma}(A)$  represents how nearly singular  $A$  is. The **condition number** for  $A$  is defined as

$$\kappa(A) = \frac{\bar{\sigma}(A)}{\underline{\sigma}(A)} \quad (\text{A.76})$$

and can be used to determine how invertible  $A$  is.

Lastly, there are other matrix decompositions that may be used in computational algorithms for linear state-space systems. One such decomposition is the **QR decomposition** for any square matrix  $A$  defined as

$$A = QR \quad (\text{A.77})$$

where  $Q$  is orthogonal and  $R$  is upper triangular. A second decomposition is the **Cholesky decomposition** for Hermitian positive definite matrices defined as

$$A = LL^* \quad (\text{A.78})$$

where  $L$  is an upper triangular matrix with real and positive entries along its main diagonal. Note that if  $A$  is real-valued, then this reduces to  $A = LL^T$ .

### Linear Matrix Equalities and Inequalities

A **linear matrix equality (LME)** in the variable  $X \in \mathcal{X}$  is an equality of the form

$$F(X) = Q \quad (\text{A.79})$$

and a **linear matrix inequality (LMI)** in the variable  $X \in \mathcal{X}$  is an inequality of the form

$$F(X) \leq Q \quad (\text{A.80})$$

where  $\mathcal{X}$  is a real-valued vector space,  $F$  is a *linear* mapping from  $\mathcal{X} \rightarrow \mathbb{H}^n$ , and  $Q \in \mathbb{H}^n$  where  $\mathbb{H}^n$  is the space of  $n \times n$  Hermitian matrices. Furthermore, a **strict linear matrix inequality (LMI)** in  $X \in \mathcal{X}$  is an inequality of the form

$$F(X) < Q \quad (\text{A.81})$$

An alternative form for LMEs and LMIs can be formulated using vectors by defining  $\vec{v}_1, \dots, \vec{v}_m$  as the basis for the vector space,  $\mathcal{X}$ . Next, for any  $X \in \mathcal{X}$ , one has that there exists scalars  $x_1, \dots, x_m$  such that

$$X = x_1 \vec{v}_1 + \dots + x_m \vec{v}_m \quad (\text{A.82})$$

Then, by the linearity of  $F(X)$ , one has for LMEs

$$x_1 F(\vec{v}_1) + \dots + x_m F(\vec{v}_m) = Q \quad (\text{A.83})$$

and for LMIs

$$x_1 F(\vec{v}_1) + \dots + x_m F(\vec{v}_m) \leq Q \quad (\text{A.84})$$

where the variables become the set of scalars:  $x_1, \dots, x_m$ . However, this coordinate form is not typically how one encounters LMIs in control theory. Thus, to check whether an inequality is an LMI, one can simply check whether the three conditions above are satisfied.

While many matrix inequalities may not look like LMIs at first glance, consider the following matrix inequality

$$A^* X A - X = Q \quad (\text{A.85})$$

and

$$A^* X A - X < Q \quad (\text{A.86})$$

with  $A \in \mathbb{R}^{n \times n}$ ,  $X \in \mathbb{S}^n$ , and  $Q \in \mathbb{H}^n$  where  $\mathbb{S}^n$  is the space of  $n \times n$  symmetric matrices. Defining  $F(X) = A^* X A - X$ , one can clearly see this is a linear mapping in the space of symmetric matrices as  $F(X)\mathbb{S}^n$ . Next, consider

$$A^* X A - B Y + Y^* B^* = Q \quad (\text{A.87})$$

and

$$A^*XA - BY + Y^*B^* < Q \quad (\text{A.88})$$

with  $A \in \mathbb{C}^{n \times n}$ ,  $B \in \mathbb{C}^{n \times m}$ ,  $Q \in \mathbb{H}^n$ ,  $X \in \mathbb{S}^n$ , and  $Y \in \mathbb{R}^{m \times n}$ . Defining  $Z = (X, Y)$  and  $F(Z) = A^*XA - BY + Y^*B^*$ , one can see this is LME/LMI in  $X$  and  $Y$  as  $F(Z) : \mathbb{S}^n \times \mathbb{R}^{m \times n} \rightarrow \mathbb{H}^n$  is a linear mapping. Furthermore, it should be noted that in some problems LMEs/LMIs are written over the space  $\mathbb{S}^n$  instead of  $\mathbb{H}^n$ .

A useful fact for LMEs/LMIs is that a list of LMEs/LMIs can always be converted to a *single* LMI through block diagonal matrices, e.g. two LMIs:

$$F_1(X_1) < Q_1 \quad (\text{A.89})$$

and

$$F_2(X_2) < Q_2 \quad (\text{A.90})$$

are equivalent to one LMI

$$F(X) = \begin{bmatrix} F_1(X_1) & 0 \\ 0 & F_2(X_2) \end{bmatrix} < \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix} = Q \quad (\text{A.91})$$

Another useful fact for deriving LMIs is to use the **Schur Complement Lemma** which states for matrices:  $A \in \mathbb{H}^n$ ,  $B \in \mathbb{H}^n$ , and  $C$ , the matrix inequality

$$\begin{bmatrix} A & C \\ C^* & B \end{bmatrix} < 0 \quad (\text{A.92})$$

is satisfied if and only if

$$\begin{bmatrix} A - CB^{-1}C^* & 0 \\ 0 & B \end{bmatrix} < 0 \quad (\text{A.93})$$

also is satisfied.

To prove this lemma, one can left and right multiply this second inequality by the matrix

$$M = \begin{bmatrix} I & CB^{-1} \\ 0 & I \end{bmatrix} \quad (\text{A.94})$$

and its Hermitian transpose to obtain

$$\begin{bmatrix} A & C \\ C^* & B \end{bmatrix} = \begin{bmatrix} I & CB^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A - CB^{-1}C^* & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} I & 0 \\ B^{-1}C^* & I \end{bmatrix} < 0 \quad (\text{A.95})$$

where the right side uses the fact that if a matrix  $Q < 0$ , then by definition  $MQM^* < 0$ .

## References

For more information, please refer to the following

- Henderson, H. V., “On Deriving the Inverse of a Sum of Matrices,” in *Society for Industrial and Applied Mathematics*, Vol. 23, No. 1, 1981, pp. 53-60

## A.3 Numerical Methods

### Numerical Integration

Throughout this textbook, consider the following vector ordinary differential equation

$$\dot{\vec{x}} = f(t, \vec{x}) \quad (\text{A.96})$$

which is used throughout this textbook to represent system dynamics equations. **Numerical integration** computes a discrete solution for  $\vec{x}$  across  $t_k = k\Delta t_k$  for  $k = 0, 1, \dots$  as a difference equation, i.e.,

$$\vec{x}_{k+1} = \vec{x}_k + \int_{k\Delta t_k}^{(k+1)\Delta t_k} f(\tau, \vec{x}(\tau)) d\tau \quad (\text{A.97})$$

where  $k$  is the time step and  $\Delta t_k = t_{k+1} - t_k$  is the time interval from time step  $k$  to time step  $k + 1$ . One of the most common approaches for model-based design of aerospace systems use Runge-Kutta methods which can be divided into fixed-step, i.e.,  $\Delta t_k = \Delta t$  is constant with  $k$ , and variable-step, i.e.,  $\Delta t_k$  is allowed to vary with  $k$  to adapt to the local numerical integration error.

The general **fixed-step Runge-Kutta (RK) method** uses a weighted average of the derivative function at intermediate stages between  $t_{k+1}$  and  $t_k$  to approximate the integral of the shown above, i.e.,

$$\vec{x}_{k+1} = \vec{x}_k + \Delta t \sum_{i=1}^N b_i \tilde{f}_i \quad (\text{A.98})$$

where  $N$  is the number of fixed-step stages,  $b_i$  are the averaging weights, and  $\tilde{f}_i$  is the evaluation of the derivative function at the  $i^{\text{th}}$  stage, i.e.,

$$\tilde{f}_i = f(\tilde{t}_i, \tilde{\vec{x}}_i) = f(t_k + c_i \Delta t, \vec{x}_{k-1} + \Delta t \sum_{j=1}^i a_{ij} \tilde{f}_j) \quad i = 1, \dots, N \quad (\text{A.99})$$

where  $c_i$  is the  $i^{\text{th}}$  intermediate time node and  $a_{ij}$  is coupling coefficient between the  $i^{\text{th}}$  and  $j^{\text{th}}$  stages. Thus, an RK method is specified by the values of the Runge-Kutta matrix,  $[a]$  where  $a_{ij}$  is the element in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column, the weight vector,  $\vec{b}$ , and the node vector,  $\vec{c}$ . A RK method is called **explicit** if  $a_{ij} = 0$  for  $i \leq j$  with  $c_1 = 0$  and implicit otherwise. Notably, this makes the term for  $j = i$  in the summation for  $\tilde{\vec{x}}_i$  always be zero and is often dropped in its definition. There are different ways to form a RK method to maintain a certain total accumulated error of some order  $p$ , often using the rule

$$\sum_{j=1}^{i-1} a_{ij} = c_i \quad \text{for } i = 2, \dots, N \quad (\text{A.100})$$

with  $c_1 = 0$  and to maintain consistency with the Taylor Series expansion, i.e.,

$$\sum_{i=1}^N b_i = 1 \quad (\text{A.101})$$

A RK method of order  $p$  is denoted as an  $\text{RK}_p$  method where an  $\text{RK}_p$  method is as accurate in computing  $\vec{x}_{k+1}$  as the  $p^{\text{th}}$ -order Taylor series expansion of  $\vec{x}_{k+1}$ .  $\text{RK}_p$  methods only require the first-order derivative function,  $f(t, \vec{x})$ , and not higher-order derivatives through order  $p$  as for the Taylor series expansion. Furthermore, it can be shown that the number of stages  $N$  equals the order of the RK method  $p$  if  $p < 5$ . Notably, variable-step RK methods typically combine adjacent-order fixed-step RK methods to automatically adjust  $\Delta t$  to estimate and bound the accumulated error, e.g., RK4(5) embeds RK4 into RK5.

The fixed-step **single-stage explicit Runge-Kutta (RK1) method**, also known as the **forwards Euler integration**, method is given by

$$\vec{x}_{k+1} = \vec{x}_k + \Delta t f(t_n, \vec{x}_k) \quad (\text{A.102})$$

where  $\tilde{f}_1 = f(t_n, \vec{x}_{k-1})$  which implies

$$\begin{aligned} \vec{a} &= \text{undefined} \\ \vec{b} &= [1] \\ \vec{c} &= [0] \end{aligned} \quad (\text{A.103})$$

The family of explicit fixed-step two-stage Runge-Kutta (RK2) methods is parameterized by  $\alpha > 0$  and given by

$$\vec{x}_{k+1} = \vec{x}_k + \Delta t \left[ \left(1 - \frac{1}{2\alpha}\right) \tilde{f}_1 + \frac{1}{2\alpha} \tilde{f}_2 \right] \quad (\text{A.104})$$

where  $\tilde{f}_1 = f(t_k, \vec{x}_k)$  and  $\tilde{f}_2 = f(t_k + \alpha\Delta t, \vec{x}_k + \alpha\Delta t f(t_k, \vec{x}_k))$  which implies

$$\begin{aligned} \vec{a} &= \begin{bmatrix} 0 & 0 \\ \alpha & 0 \end{bmatrix} \\ \vec{b} &= \begin{bmatrix} 1 - \frac{1}{2\alpha} \\ \frac{1}{2\alpha} \end{bmatrix} \\ \vec{c} &= \begin{bmatrix} 0 \\ \alpha \end{bmatrix} \end{aligned} \quad (\text{A.105})$$

where  $\alpha = 0.5$  is the **midpoint method**,  $\alpha = 1$  is **Heun's method**, and  $\alpha = 2/3$  is **Ralston's method**.

The most common explicit fixed-step four-stage Runge-Kutta method is the **classic RK4 method** given by

$$\vec{x}_{k+1} = \vec{x}_k + \frac{\Delta t}{6} f(t_n, \vec{x}_k) \quad (\text{A.106})$$

where

$$\tilde{f}_1 = f(t_k, \vec{x}_k) \quad (\text{A.107})$$

$$\tilde{f}_2 = f(t_k + 0.5\Delta t, \vec{x}_k + 0.5\Delta t k_1) \quad (\text{A.108})$$

$$\tilde{f}_3 = f(t_k + 0.5\Delta t, \vec{x}_k + 0.5\Delta t k_2) \quad (\text{A.109})$$

$$\tilde{f}_4 = f(t_k, \vec{x}_k + \Delta t k_3) \quad (\text{A.110})$$

which implies

$$[a] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (\text{A.111})$$

$c_2 = c_3 = 0.5$ ,  $b_1 = b_4 = 1/6$ , and  $b_2 = b_3 = 1/3$ .

Variable-step Runge-Kutta methods typically

## References

For more information, please refer to the following

- Curtis, H. D., “1.8 Numerical integration,” *Orbital Mechanics for Engineering Students*, 4th ed., Vol. 1, Elsevier Butterworth-Heinemann, Massachusetts, 2021, pp. 34-49

## A.4 Complex Analysis and Stability Criterion

### Nyquist Plot

The **Nyquist plot** is a common tool used to understand the stability and robustness of a feedback control system. Similar to the Bode plot, the Nyquist plot can be used to analyze the frequency response of a transfer function, i.e.  $G(s)$  with  $s = j\omega$ . However, opposed to the Bode plot, the Nyquist plot visualizes the real and imaginary parts of  $G(j\omega)$  as a single curve with the real part on the horizontal axis and the imaginary on the vertical axis. It should also be noted the convention for the Nyquist plot is to plot over  $-\infty < \omega < \infty$  as opposed to  $\omega \geq 0$  for the Bode plot. This convention results in a reflected curve about the real axis for  $\omega < 0$  with respect to  $\omega > 0$  as a transfer function value  $G(j\omega) = \alpha + j\beta$  simply provides the complex conjugate for  $\omega < 0$  with respect to  $\omega > 0$ .

As an example of a Nyquist plot, consider the standard form of a first order LTI system, i.e.

$$\dot{y} + a_0 y = b_0 u \quad (\text{A.112})$$

and assume that  $a_0 \neq 0$ . The transfer function for this system is  $G(s)$

$$G(s) = \frac{b_0}{s + a_0} \quad (\text{A.113})$$

and has a pole at  $s = -a_0$ .

The Nyquist plot of the frequency response requires computing  $G(j\omega)$ , i.e.

$$G(j\omega) = \frac{b_0}{j\omega + a_0} \quad (\text{A.114})$$

$$G(j\omega) = \frac{b_0}{j\omega + a_0} \frac{a_0 - j\omega}{a_0 - j\omega} \quad (\text{A.115})$$

$$G(j\omega) = \frac{a_0 b_0}{a_0^2 + \omega^2} - j \frac{\omega b_0}{a_0^2 + \omega^2} \quad (\text{A.116})$$

where

$$\operatorname{Re}\{G(j\omega)\} = \frac{a_0 b_0}{a_0^2 + \omega^2} \quad (\text{A.117})$$

and

$$\operatorname{Im}\{G(j\omega)\} = \frac{-\omega b_0}{a_0^2 + \omega^2} \quad (\text{A.118})$$

Next, noting that

$$\operatorname{Re}\{G(j\omega)\}^2 = \frac{a_0^2 b_0^2}{(a_0^2 + \omega^2)^2} \quad (\text{A.119})$$

$$\operatorname{Im}\{G(j\omega)\}^2 = \frac{\omega^2 b_0^2}{(a_0^2 + \omega^2)^2} \quad (\text{A.120})$$

and adding, one has

$$\operatorname{Re}\{G(j\omega)\}^2 + \operatorname{Im}\{G(j\omega)\}^2 = \frac{a_0^2 b_0^2}{(a_0^2 + \omega^2)^2} + \frac{\omega^2 b_0^2}{(a_0^2 + \omega^2)^2} \quad (\text{A.121})$$

or

$$\operatorname{Re}\{G(j\omega)\}^2 + \operatorname{Im}\{G(j\omega)\}^2 = \frac{b_0^2}{(a_0^2 + \omega^2)} \quad (\text{A.122})$$

Then rearranging, one has

$$\operatorname{Re}\{G(j\omega)\}^2 - \frac{b_0^2}{(a_0^2 + \omega^2)} + \operatorname{Im}\{G(j\omega)\}^2 = 0 \quad (\text{A.123})$$

Then, adding  $\left(\frac{b_0}{2a_0}\right)^2$  to both sides, one has

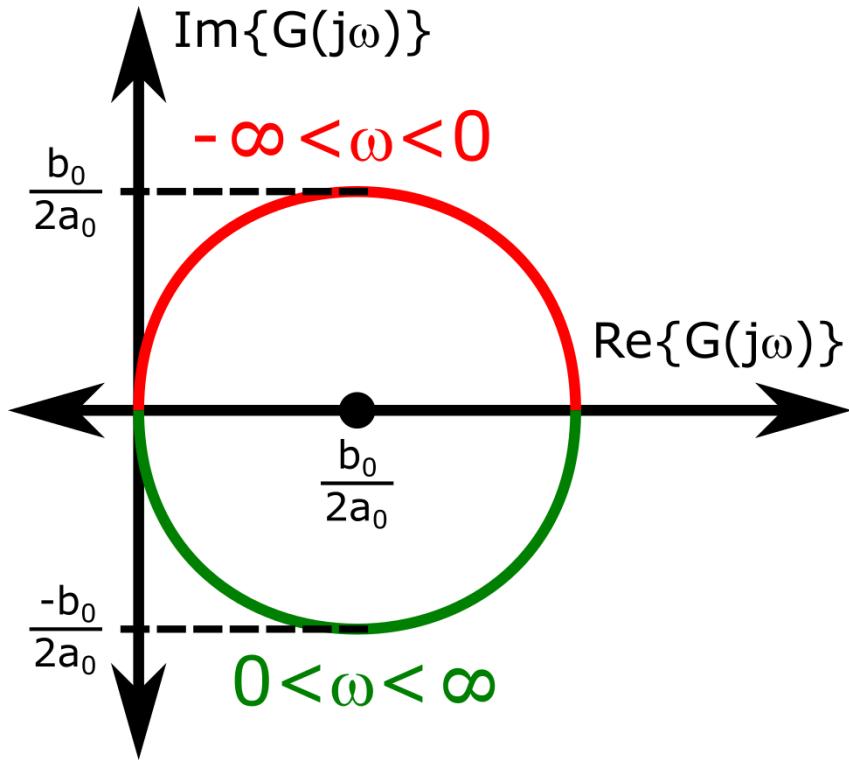
$$\operatorname{Re}\{G(j\omega)\}^2 - \frac{b_0^2}{(a_0^2 + \omega^2)} + \left(\frac{b_0}{2a_0}\right)^2 + \operatorname{Im}\{G(j\omega)\}^2 = \left(\frac{b_0}{2a_0}\right)^2 \quad (\text{A.124})$$

and recalling Equation A.119, one has

$$\left(\operatorname{Re}\{G(j\omega)\}^2 - \frac{b_0^2}{2a_0}\right)^2 + \operatorname{Im}\{G(j\omega)\}^2 = \left(\frac{b_0}{2a_0}\right)^2 \quad (\text{A.125})$$

which is the equation of a circle of radius  $\frac{b_0}{2a_0}$  centered at  $\left(\frac{b_0}{2a_0}, 0\right)$ .

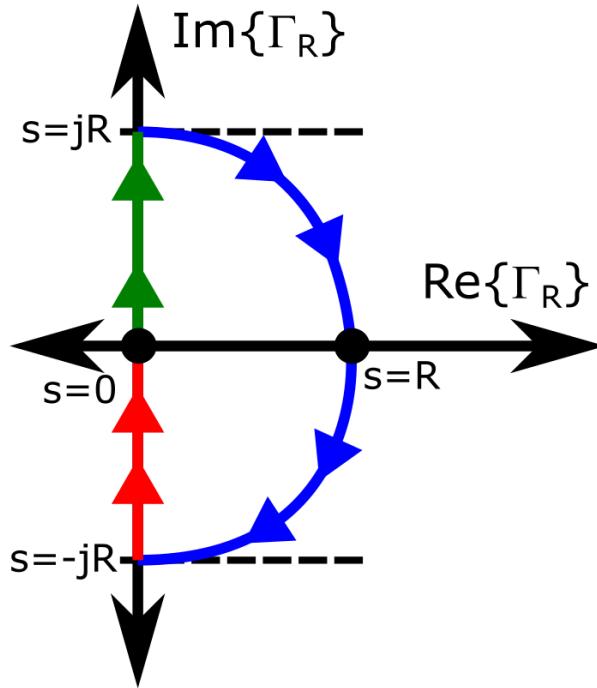
For plotting the variation with  $\omega$ , consider three particular points. At  $\omega = 0$ ,  $G(0) = \frac{b_0}{a_0}$ . As  $\omega \rightarrow \infty$ ,  $G(j\omega) \rightarrow -j\frac{b_0}{\omega}$ . At  $\omega = a_0$ :  $G(ja_0) = (1 - j)\frac{b_0}{2a_0}$ . Using these results, one can draw the Nyquist plot as



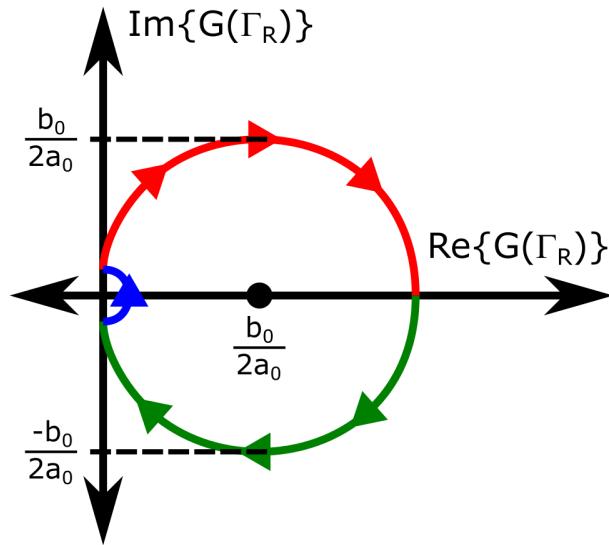
### Cauchy's Argument Principle

The Nyquist plot of the open-loop transfer function  $L(s)$  can be used to state a theorem concerning the stability of a feedback control system. However, to derive this analysis tool, one requires a result from complex analysis, Cauchy's argument principle, which will be summarized as follows. To begin, let  $G(s)$  be a transfer function of a system and let  $\Gamma$  be a simple, closed curve in the complex plane where a “simple” curve is one that does not intersect itself and a closed curve is completely connected. Here, the notation  $G(\Gamma)$  is the curve obtained by mapping each complex number  $s \in \Gamma$  to another complex number  $G(s)$ . In general,  $G(\Gamma)$  will be closed but need not be simple.

For example, consider the curve  $\Gamma_R$  shown as



which results in the mapping to  $G(\Gamma_R)$  for a first order system as



and it can be shown that  $G(\Gamma_R)$  converges to the Nyquist plot of  $G(s)$  as  $R \rightarrow \infty$ .

Next, define  $N_p$  and  $N_z$  as the number of poles and zeros of  $G(s)$  inside a general curve  $\Gamma$ , respectively.

Then, **Cauchy's argument principle** states that if  $\Gamma$  does not pass through any poles or zeros of  $G(s)$ , then the closed curve  $G(\Gamma)$  encircles the origin  $N_z - N_p$  times. Furthermore, if  $N_z - N_p > 0$ , the  $G(\Gamma)$  encircles the origin clockwise and if  $N_z - N_p < 0$ , then  $G(\Gamma)$  encircles the origin counter-clockwise.

Finally, it should also be noted that if  $\Gamma$  passes through a pole of  $G(s)$ , then  $G(\Gamma)$  diverges to  $\infty$  and if  $\Gamma$  passes through a zero of  $G(s)$ , then  $G(\Gamma)$  intersects the origin. In either case, Cauchy's argument principle would not apply. The proof of this principle can be found elsewhere, e.g. in textbooks on complex analysis, and is left for the reader.

## Nyquist Stability Criterion

To develop the stability criteria based on the Nyquist plot of the open-loop transfer function  $L(j\omega)$ , first note that  $s = -1$  is a critical point for the stability of the open-loop transfer function  $L(s)$ . Recall from previous analysis that the stability of the feedback control system could be assessed through the zeros of  $1 + G(s)K(s)$ , i.e.  $1 + L(s)$  as this expression appeared in the denominator of all four fundamental transfer functions. Thus, if the frequency response of the open-loop transfer function, i.e.  $L(j\omega)$ , passes through  $j\omega = -1$  on the Nyquist plot for some  $\omega = \omega_0$ , then the feedback control system will be unstable, as any signal that contains frequency  $\omega_0$  would cause one or more of the fundamental transfer functions to have infinite gain, e.g. as  $S(s) = \frac{E(s)}{Y_c(s)} = \frac{1}{1+L(s)}$ , then  $|S(j\omega_0)| = \infty$ .

Furthermore, recall the stability conditions that if no pole-zero cancellations exist between  $G(s)$  and  $K(s)$ , then the feedback control system is stable if and only if  $1 + L(s)$  has no zeros in the RHP. Thus, one can infer that the critical  $s = -1$  point also plays a role in the Nyquist theorem on the feedback control system stability.

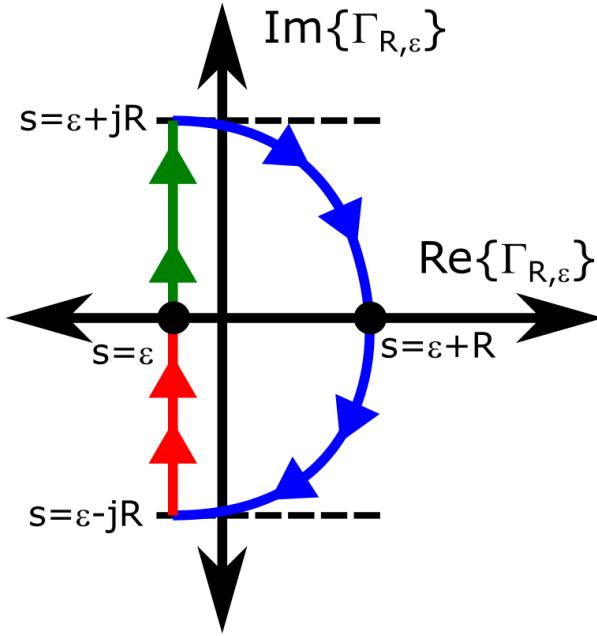
Next, define  $P_{c-l}(\epsilon)$  as the number of poles of the closed-loop feedback control system with real part greater than or equal to  $\epsilon$ ,  $P_{o-l}(\epsilon)$  as the number of poles of the open-loop transfer function  $L(s)$  with real part greater than or equal to  $\epsilon$ , and  $N_{ccw}(\epsilon)$  as the number of times the Nyquist curve of  $L(\epsilon + j\omega)$  encircles the critical  $s = -1$  point in the counterclockwise direction.

Then, the **Nyquist stability criterion** states that if  $L(s)$  has no poles with real part equal to  $\epsilon$ , then  $P_{c-l}(\epsilon) = P_{o-l}(\epsilon) - N_{ccw}(\epsilon)$ . Thus,  $P_{c-l}(\epsilon) = 0$  if and only if  $N_{ccw}(\epsilon) = P_{o-l}(\epsilon)$ .

A notable special case of this theorem, e.g. **simplified Nyquist stability criterion**, states that if  $\epsilon = 0$ , i.e.  $L(s)$  has no poles on the imaginary axis, then  $P_{c-l}$  and  $P_{o-l}$  are the number of RHP poles for the closed- and open-loops, respectively (including the imaginary axis) and  $N_{ccw}$  is the number of times the Nyquist plot of  $L(j\omega)$ , i.e. the frequency response of  $L(s)$ , encircles  $s = -1$  in the counterclockwise direction. In this case, the Nyquist theorem states that  $P_{c-l} = P_{o-l} - N_{ccw}$  and the closed-loop is stable if and only if  $N_{ccw} = P_{o-l}$ . Note that this is the case in MATLAB for assessing the Nyquist plot of  $L(j\omega)$ . Thus, if  $L(j\omega)$  has a pole on the imaginary axis, e.g.  $K(s)$  has an integral component  $\frac{1}{s}$ , then care must be taken in the stability analysis.

However, in either case, the Nyquist plot of the open-loop transfer function  $L(j\omega)$  provides a visual method for determining the stability of the closed-loop feedback control system, as well as a measure of "how much" stability there is for the system, i.e. its robustness, which will be discussed in the subsequent section.

To prove the general Nyquist theorem, one requires defining a "perturbed" simple, closed curve  $\Gamma_{R,\epsilon}$  which is a shifted  $\Gamma_R$  curve by some amount  $\epsilon$  along the real axis as shown below for some  $\epsilon < 0$ .



Note that as  $\epsilon \rightarrow 0$ ,  $\Gamma_{R,\epsilon} \rightarrow \Gamma_R$ . Next, defining the transfer function  $H(s) = 1 + L(s)$  and  $N_p$  and  $N_z$  as the number of poles and zeros of  $H(s)$  inside the simple, closed curve  $\Gamma_{R,\epsilon}$ , then Cauchy's argument principle states that  $H(\Gamma_{R,\epsilon})$  has  $N_p - N_z$  counterclockwise encirclements of the origin. Then, note the following three observations.

1. As  $L(s) = H(s) - 1$ ,  $L(\Gamma_{R,\epsilon})$  is the curve  $H(\Gamma_{R,\epsilon})$  shifted to the left by one unit. Thus,  $H(\Gamma_{R,\epsilon})$  encircling the origin is equivalent to  $L(\Gamma_{R,\epsilon})$  encircling  $s = -1$ . Furthermore, as  $R \rightarrow \infty$  and  $\epsilon \rightarrow 0$ ,  $L(\Gamma_{R,\epsilon})$  converges to the Nyquist plot of  $L(s)$ . Thus,  $N_{ccw} = N_p - N_z$ .
2.  $\Gamma_{R,\epsilon}$  contains the entire RHP as  $R \rightarrow \infty$  and  $\epsilon \rightarrow 0$ . Thus, if  $R$  is sufficiently large and  $\epsilon$  is sufficiently small, then the RHP zeros of  $H(s) = 1 + L(s)$  are precisely the closed-loop RHP poles, i.e.  $N_z = P_{c-l}$ .
3. The RHP poles of  $H(s)$  are precisely the RHP poles of the open-loop transfer function  $L(s)$ , i.e.  $N_p = P_{o-l}$ . This follows from the assumption that  $L(s)$  contains no pole-zero cancellations, and thus,  $|H(s_0)| = \infty$  for some  $s_0$  in the RHP if and only if  $|L(s_0)| = \infty$ .

Combining these three observations, one has that  $N_{ccw} = P_{o-l} - P_{c-l}$  or rearranging  $P_{c-l} = P_{o-l} - N_{ccw}$  which is what was needed to be proved.

## Multivariate Nyquist Criterion

### Circle Criterion

### References

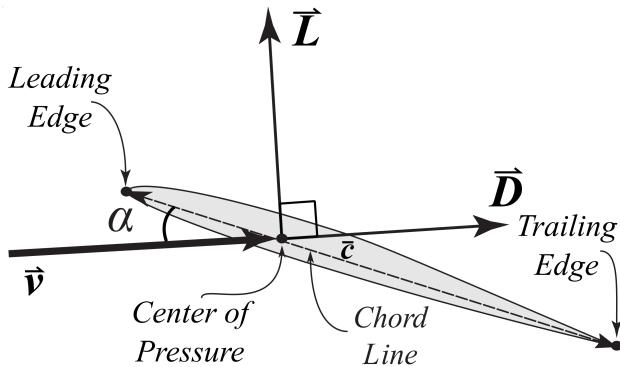
---

# Airplane Stability and Control Derivative Models

## B.1 Finite-Wing Theory and Component Build-Up Model

### Finite-Wing Theory

To develop the basic analytical models for modeling the aerodynamic forces and moments of fixed-wing aircraft, one can use **finite-wing theory** for each lifting surface on a aircraft while also accounting for some interactive aerodynamics between these surfaces and other structures. Finite-wing theory uses the simpler airfoil theory for airflow over any two-dimensional cross-section of a lifting surface, also known as an **airfoil** and resolves the two-dimensional pressure distribution of the moving air over the lifting surface into two perpendicular force contributions as shown in the following free body diagram.



where  $c$  is the **aerodynamic chord** from the leading edge to trailing edge,  $\vec{v}_\infty$  is the **free-stream velocity**,  $v_\infty$  is the **airspeed**,  $\alpha$  is the **airfoil angle of attack**,  $L$  is the **airfoil lift force** defined as perpendicular to the free-stream velocity and  $D$  is the **airfoil drag force** defined as parallel to the free-stream velocity. It should be noted that the **center of pressure** defines the location about which the pitching moment due to

the pressure distribution is currently zero, thus producing no pitching moment. However, as this location will generally change as a function of angle of attack, one often uses the **aerodynamic center** of the airfoil about which the pitching moment will not change with angle of attack, allowing one to resolve the pressure distribution as the lift, drag, and **airfoil pitching moment**,  $M$ , at any location.

Considering a finite-wing as a distribution of airfoils at a single angle of attack and defining the **dynamic pressure** for the free-stream velocity,  $Q_\infty$ , as

$$Q_\infty = \frac{1}{2} \rho v_\infty^2 \quad (\text{B.1})$$

where  $\rho$  is the **air density** which depends on altitude,  $h$ , one can model the lift, drag, and wing pitching moment for the entire wing,  $L_w$ ,  $D_w$ , and  $M_w$ , respectively, as

$$L_w = Q_\infty S_w C_{L,w} \quad (\text{B.2})$$

$$D_w = Q_\infty S_w C_{D,w} \quad (\text{B.3})$$

and

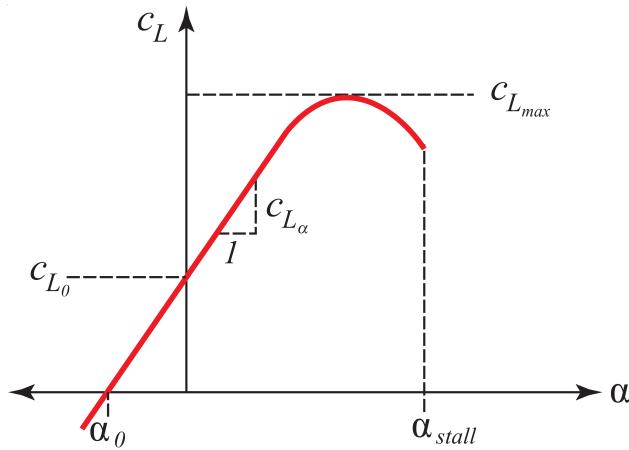
$$M_w = Q_\infty S_w \bar{c}_w C_{M,w} \quad (\text{B.4})$$

where  $S_w$  is the surface area of the wing,  $\bar{c}_w$  is the **mean aerodynamic chord** of the wing,  $C_{L,w}$  is the **wing lift coefficient**,  $C_{D,w}$  is the **wing drag coefficient**, and  $C_{M,w}$  is the **wing pitching moment coefficient**. Furthermore, due to their finite nature, wings may also have a **rolling moment**,  $L_{roll,w}$ , defined as

$$L_{roll,w} = Q_\infty S_w b_w C_{L_{roll,w}} \quad (\text{B.5})$$

where  $b_w$  is the **wing span** of the wing and  $C_{L_{roll,w}}$  is the **wing rolling moment coefficient**. It should be noted that often the “*roll*” subscript is dropped which noticeably overloads the notation for the letter,  $L$ , and must be inferred by context whether one is referring to the lift or the rolling moment.

The lift coefficient for a finite-wing generally depends on the wing angle of attack as depicted in the following plot.



At low angles of attack, the relationship is linear, i.e.

$$C_{L,w} = C_{L_\alpha,w}(\alpha_w - \alpha_{0,w}) \quad (\text{B.6})$$

where  $C_{L_\alpha,w} = \frac{dC_{L,w}}{d\alpha}$  is the **wing lift-curve slope** and  $\alpha_{0,w} \leq 0$  is the **wing zero-lift angle of attack**. This can also be written as

$$C_{L,w} = C_{L_\alpha,w}\alpha_w + C_{L_{0,w}} \quad (\text{B.7})$$

where  $C_{L_{0,w}}$  is the lift coefficient at  $\alpha = 0$  and equals  $C_{L_{0,w}} = -C_{L_\alpha,w}\alpha_{0,w}$ . If  $C_{L_{0,w}} = \alpha_{0,w} = 0$ , then one has a **symmetric wing**, otherwise one has a **cambered wing**. Note that at high angles of attack, a maximum lift coefficient is reached,  $C_{L_{max}}$ , after which there is a slight decrease in the lift coefficient and eventually a large decrease. With this large decrease in lift there is also a large increase in drag, a phenomenon known as **stall** which occurs at some  $\alpha_{stall}$ . As the purpose of wings are generally to generate lift with a small amount of drag, fixed-wing aircraft nominally operate at angles of attack well below stall and will be assumed in the modeling of this textbook. Stall aerodynamics are beyond the scope of this textbook.

The drag coefficient for a finite-wing flying can generally be separated into three terms as

$$C_{D,w} = C_{D_{0,w}} + C_{D_{i,w}} + C_{D_{w,w}} \quad (\text{B.8})$$

where  $C_{D_{0,w}}$  is the **wing parasitic drag coefficient**, also known as the **wing profile drag coefficient**, due to air pressure and skin friction,  $C_{D_{i,w}}$  is the **wing induced drag coefficient** due to the production of lift, and  $C_{D_{w,w}}$  is the **wing wave drag coefficient** due to generation of shockwaves at supersonic flight. As  $C_{D_{0,w}}$  is constant with respect to the flight conditions, it is also known as the **zero-lift drag coefficient**.

The induced drag coefficient is typically modeled as a quadratic function

$$C_{D_{i,w}} = \frac{C_{L,w}^2}{\pi e_0 A R_w} \quad (\text{B.9})$$

where  $A R_w$  is the **wing aspect ratio** defined as

$$A R_w = \frac{b_w^2}{S_w} \quad (\text{B.10})$$

and  $0 < e_0 < 1$  is the **wing Oswald efficiency** which is typically between 0.7 – 0.85 for subsonic fixed-wing aircraft with moderate aspect ratio and sweep angles. At supersonic speeds,  $e_0$  can range from 0.3 – 0.5. For flat plate airfoils, the wave drag coefficient can be modeled as

$$C_{D_{w,w}} = \frac{4\alpha_w}{\sqrt{M_\infty^2 - 1}} \quad (\text{B.11})$$

where  $M_\infty$  is the free-stream Mach number, i.e.

$$M_\infty = \frac{v_\infty}{c_s} \quad (\text{B.12})$$

where  $c_s$  is the **speed of sound** of the air mass which varies with the square root of the thermodynamic temperature.

For a wing with constant airfoil shape at subsonic speeds, one can use **lifting-line theory** which predicts that the lift-curve slope for a finite-wing,  $C_{L\alpha,w}$ , can be related to the section **airfoil's lift-curve slope**,  $c_{L\alpha}$ , for a wide range of aspect ratios as

$$C_{L\alpha,w} = \frac{2\pi AR_w}{2 + \sqrt{\frac{4\pi^2 AR_w^2 (1-M_\infty^2)}{c_{L\alpha}^2} \left(1 + \frac{\tan^2 \Lambda_{c/2}}{1-M_\infty^2}\right) + 4}} \text{ rad} \quad (\text{B.13})$$

where  $\Lambda_{c/2}$  is the **sweep angle** of the half chord line. In addition, lifting-line theory predicts that the induced drag coefficient can be modeled as

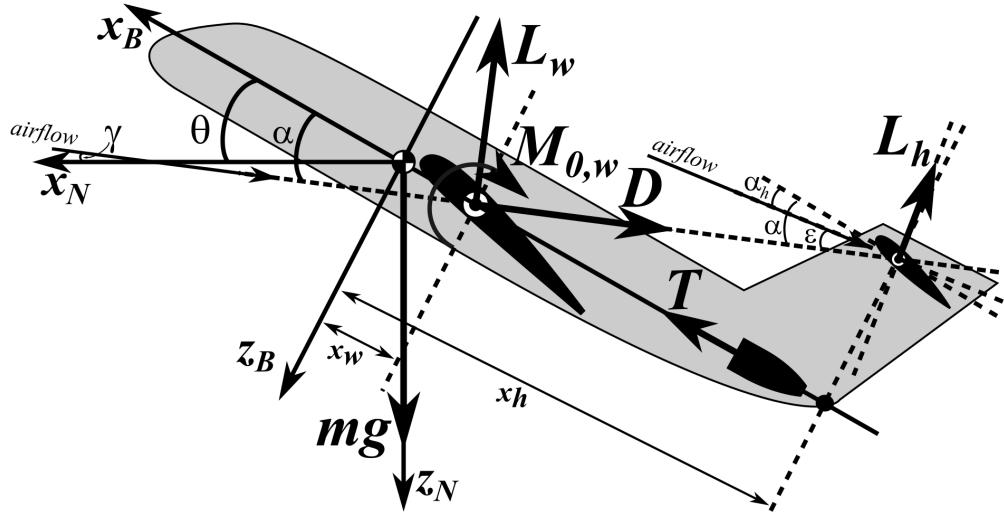
$$C_{D_i,w} = \frac{C_{L,w}^2}{\pi AR_w e_w} \quad (\text{B.14})$$

where  $e_w$  is the wing's **span efficiency** where if  $e_w = 1$ , the wing is defined as an **elliptical**. Typically  $e_0 < e_w$ , but it is often a good approximation without additional data. However, it should be noted that most aircraft use multiple types of airfoils at different points of a lifting surface, a technique called **aerodynamic twisting** which makes an analytical solution more difficult to obtain than lifting-line theory.

### Component Build-Up Model

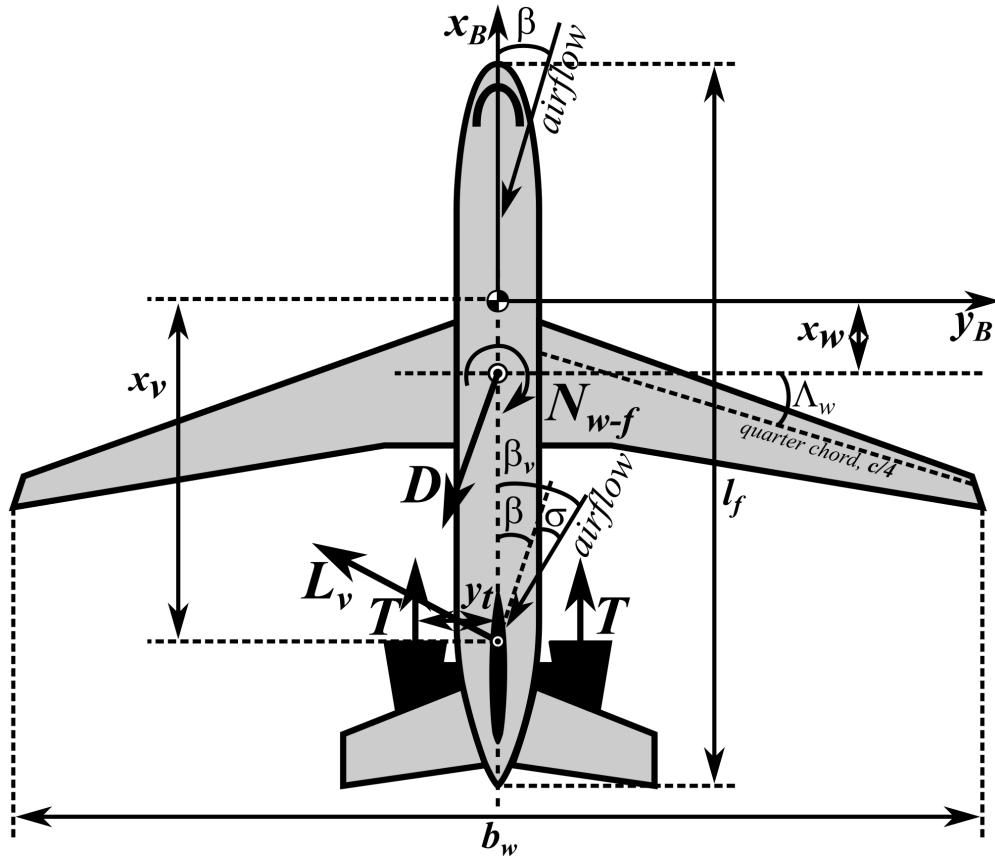
The **component build-up model** uses finite-wing theory for each primary lifting surface of a fixed-wing aircraft to build up the stability and control derivatives. For each surfaces, one must define the aerodynamic center about which lift, drag, pitching moment, and rolling moment can be resolved for that surface as prescribed via finite-wing theory. For conventional fixed-wing aircraft, there are four primary lifting surfaces to consider in aerodynamic coefficient modeling, namely the wing aerodynamic coefficients denoted with subscript  $w$ , the horizontal tail aerodynamic coefficients denoted with subscript  $h$ , the vertical tail aerodynamic coefficients denoted with subscript  $v$ , and the fuselage aerodynamic coefficients denoted with subscript  $f$ .

From the view of the  $x_B - z_B$ -plane, also known as the **longitudinal plane**, the simplified FBD with the  $y_B$ -axis drawn into the page can be drawn as



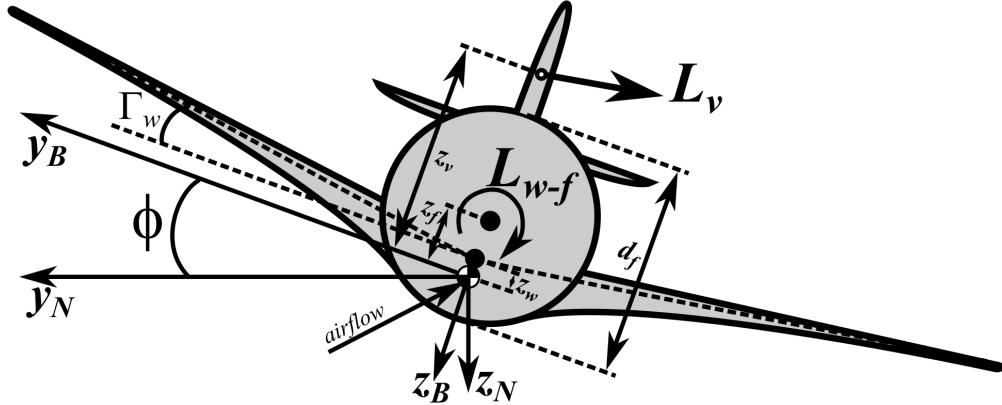
where the airplane's induced drag is dominated by the wing and the parasitic drag has been resolved at the wing (i.e. this replaces  $D_f$ ,  $D_w$ ,  $D_h$ , and  $D_v$  by  $D$  at wing aerodynamic center), the drag and thrust contribution to the pitching moment on the airplane has been neglected, the horizontal tail airfoil is symmetric (i.e.  $M_{0,h} = 0$ ), and  $\epsilon$  is the **downwash angle** which results from the airflow interacting with the wings and engines upstream of the horizontal tail. Note that the aerodynamic centers of each lifting surface in this plane are labeled by their  $x$  and  $z$  coordinates in the body frame. It should also be noted that the lift from the fuselage is often quite small in this plane.

From the view of the  $x_B - y_B$ -plane, also known as the **directional plane**, the simplified FBD with the  $z_B$ -axis drawn into the page can be drawn as



where the wing and fuselage effects can be combined into two terms,  $L_{w-f}$  and  $N_{w-f}$ , the vertical tail airfoil is symmetric (i.e.  $N_{0,v} = 0$ ), and  $\sigma$  is the **sidewash angle** which results from the airflow interacting with the wings and engines upstream of the vertical tail. It should be noted that each thrust vector has a  $y_B$  component which cancel each other with respect to the yawing moment.

From the view of the  $y_B - z_B$ -plane, also known as the **lateral plane**, the simplified FBD with the  $x_B$ -axis drawn out of the page can be drawn as



It should be noted that the aerodynamic centers of the fuselage and wing are additionally represented by  $z_B$  coordinates in the body frame which was neglected in the longitudinal frame.

Furthermore, if one assumes the small angle approximation for  $\alpha$  and  $\beta$ ,  $\alpha - \epsilon$ , and  $\beta + \sigma$ , then, one obtains the simplified wind-to-body aerodynamic and propulsive force and moment equations as

$$\begin{aligned}
 mX &= (L_w + L_h)\alpha - D + T \\
 mY &= -D\beta - L_v \\
 mZ &= (L_w + L_h) - D\alpha \\
 I_{xx}L &= L_{w-f} - z_v L_v \\
 I_{yy}M &= M_{0,w} + x_w L_w + x_h L_h \\
 I_{zz}N &= N_{w-f} - x_v L_v
 \end{aligned} \tag{B.15}$$

where  $\alpha$  and  $\beta$  are in radians and  $x_w$ ,  $x_h$ , and  $x_v$  are the body frame coordinates of the aerodynamic centers, i.e. negative values in the FBD. Note that  $L$  and  $L_{w-f}$  are moment terms and not lift terms here. By substituting for  $(L_w + L_h)$ ,  $D$ ,  $T$ ,  $L_w$ ,  $L_h$ ,  $L_v$ , and  $N_v$  by their aerodynamic coefficients, one has

$$\begin{aligned}
 mX &= Q_\infty S_w C_L \alpha - Q_\infty S_w C_D + Q_\infty S_w C_T \\
 mY &= -Q_\infty S_w C_D \beta - Q_v S_v C_{L,v} \\
 mZ &= -Q_\infty S_w C_L - Q_\infty S_w C_D \alpha \\
 I_{xx}L &= Q_\infty S_w b_w C_{l,w-f} - z_v Q_v S_v C_{L,v} \\
 I_{yy}M &= Q_\infty S_w \bar{c}_w C_{m_0,w} + x_w Q_\infty S_w C_{L,w} + x_h Q_h S_h C_{L,h} \\
 I_{zz}N &= Q_\infty S_w b_w C_{n,w-f} - x_v Q_v S_v C_{L,v}
 \end{aligned} \tag{B.16}$$

where the lowercase  $l$ ,  $m$ , and  $n$  are used for the body frame moment coefficient subscripts as  $C_L$  is used for the overall airplane lift coefficient, in this case  $L_w + L_h$ . These equations can be further simplified into

non-dimensional coefficient equations by the definitions

$$\begin{aligned} X &= \frac{Q_\infty S_w}{m} C_X \\ Y &= \frac{Q_\infty S_w}{m} C_Y \\ Z &= \frac{Q_\infty S_w}{m} C_Z \\ L &= \frac{Q_\infty S_w b_w}{I_{xx}} C_l \\ M &= \frac{Q_\infty S_w \bar{c}_w}{I_{yy}} C_m \\ N &= \frac{Q_\infty S_w b_w}{I_{zz}} C_n \end{aligned} \quad (\text{B.17})$$

which provides the following approximation for the force and moment coefficients as

$$\begin{aligned} C_X &= C_L \alpha - C_D + C_T \\ C_Y &= -C_D \beta - \frac{Q_v S_v}{Q_\infty S_w} C_{L,v} \\ C_Z &= -C_L - C_D \alpha \\ C_l &= C_{l,w-f} - \frac{Q_v S_v z_v}{Q_\infty S_w b_w} C_{L,v} \\ C_m &= C_{m_0,w} + \frac{x_w}{\bar{c}_w} C_{L,w} - \frac{Q_h S_h x_h}{Q_\infty S_w \bar{c}_w} C_{L,h} \\ C_n &= C_{n,w-f} - \frac{Q_v S_v x_v}{Q_\infty S_w b_w} C_{L,v} \end{aligned} \quad (\text{B.18})$$

Note that these simplified equations will also be used in this section to derive analytical models in the rest of this appendix for studying airplane dynamics about steady-flight conditions.

## B.2 Longitudinal Stability and Control Derivatives

This section will derive models for the primary longitudinal stability and control derivatives used in the Jacobian linearization of the aerodynamic and propulsive normalized forces and moments about trimmed steady flight, i.e.

$$\begin{aligned} X &= \bar{X} + X_u \Delta u + X_\alpha \Delta \alpha + X_{\delta_T} \Delta \delta_T \\ Z &= \bar{Z} + Z_u \Delta u + Z_\alpha \Delta \alpha + Z_q \Delta q + Z_{\dot{\alpha}} \Delta \dot{\alpha} + Z_{\delta_e} \Delta \delta_e + Z_{\delta_T} \Delta \delta_T \\ M &= \bar{M} + M_u \Delta u + M_\alpha \Delta \alpha + M_q \Delta q + M_{\dot{\alpha}} \Delta \dot{\alpha} + M_{\delta_e} \Delta \delta_e + M_{\delta_T} \Delta \delta_T \end{aligned} \quad (\text{B.19})$$

where the coefficients of the linear terms are called the **longitudinal stability and control derivatives**. Alternatively, these equations can be written in terms of nondimensional coefficients as

$$\begin{aligned} C_X &= \bar{C}_X + C_{X_u} \Delta u + C_{X_\alpha} \Delta \alpha + C_{X_{\delta_T}} \Delta \delta_T \\ C_Z &= \bar{C}_Z + C_{Z_u} \Delta u + C_{Z_\alpha} \Delta \alpha + C_{Z_q} \Delta q + C_{Z_\dot{\alpha}} \Delta \dot{\alpha} + C_{Z_{\delta_e}} \Delta \delta_e + C_{Z_{\delta_T}} \Delta \delta_T \\ C_m &= \bar{C}_m + C_{m_u} \Delta u + C_{m_\alpha} \Delta \alpha + C_{m_q} \Delta q + C_{m_{\dot{\alpha}}} \Delta \dot{\alpha} + C_{m_{\delta_e}} \Delta \delta_e + C_{m_{\delta_T}} \Delta \delta_T \end{aligned} \quad (\text{B.20})$$

where the coefficients of the linear terms are called the **longitudinal stability and control coefficients**.

It should be noted that a more comprehensive collection of aerodynamic stability and control prediction techniques can be found in the *USAF Stability and Control DATCOM (Data Compendium)* which was compiled between 1960 and 1978 by the McDonnell Douglas Corporation in conjunction with the Flight Dynamics Laboratory at Wright-Patterson Air Force Base.

### Longitudinal Static Stability Coefficient

For the  $M$ -moment, the coefficient is

$$C_m = C_{m_0,w} + \frac{x_w}{\bar{c}_w} C_{L,w} - \frac{Q_h S_h x_h}{Q_\infty S_w \bar{c}_w} C_{L,h} \quad (\text{B.21})$$

Next, modeling the lift coefficients for small angles of attack as

$$C_{L,w} = C_{L_0,w} + C_{L_{\alpha,w}} \alpha_w \quad (\text{B.22})$$

and

$$C_{L,h} = C_{L_{\alpha,h}} \alpha_h \quad (\text{B.23})$$

where it has been assumed that the vertical tail has a symmetric airfoil, thus providing no net force at  $\alpha_h = 0^\circ$ . Then, one has

$$C_m = C_{m_0,w} + \frac{x_w}{\bar{c}_w} (C_{L_0,w} + C_{L_{\alpha,w}} \alpha) - \frac{Q_h S_h x_h}{Q_\infty S_w \bar{c}_w} C_{L_{\alpha,h}} \alpha_h \quad (\text{B.24})$$

Then, by the previous FBD, one also has

$$\alpha_h = \alpha - \epsilon \quad (\text{B.25})$$

where  $\epsilon$  is typically approximated by a linear function

$$\epsilon = \frac{d\epsilon}{d\alpha} (\alpha + \alpha_{0,w}) = \frac{d\epsilon}{d\alpha} \alpha + \epsilon_0 \quad (\text{B.26})$$

where for an elliptical lift distribution

$$\epsilon_0 = \frac{2C_{L,w}}{\pi AR_w} \quad (\text{B.27})$$

and

$$\frac{d\epsilon}{d\alpha} = \frac{2C_{L_{\alpha,w}}}{\pi AR_w} \quad (\text{B.28})$$

though these are typically estimated using test data and/or CFD. Next, substituting for  $\alpha_h$  and  $\epsilon$ , one has

$$\begin{aligned} C_m = & C_{m_0,w} + \frac{x_w}{\bar{c}_w} (C_{L_0,w} + C_{L_\alpha,w}\alpha) \\ & + \frac{Q_h S_h x_h}{Q_\infty S_w \bar{c}_w} C_{L_\alpha,h} \left( \alpha - \left( \epsilon_0 + \frac{d\epsilon}{d\alpha} \alpha \right) \right) \end{aligned} \quad (\text{B.29})$$

and finally grouping constant and  $\alpha$ -varying terms provides

$$C_m = \left[ C_{m_0,w} + \frac{x_w}{\bar{c}_w} (C_{L_0,w} - \frac{Q_h S_h x_h}{Q_\infty S_w \bar{c}_w} C_{L_\alpha,h} \epsilon_0) \right] + \left[ \frac{x_w}{\bar{c}_w} C_{L_\alpha,w} + \frac{Q_h S_h x_h}{Q_\infty S_w \bar{c}_w} C_{L_\alpha,h} \left( 1 - \frac{d\epsilon}{d\alpha} \right) \right] \alpha \quad (\text{B.30})$$

and defining

$$\eta_h = \frac{Q_h}{Q_\infty} \quad (\text{B.31})$$

as the **horizontal tail efficiency** and

$$V_h = \frac{-x_h S_h}{\bar{c}_w S_w} \quad (\text{B.32})$$

as the **horizontal tail volume ratio**, one has

$$C_m = C_{m_0} + C_{m_\alpha} \alpha \quad (\text{B.33})$$

where

$$C_{m_0} = C_{m_0,w} + \frac{x_w}{\bar{c}_w} (C_{L_0,w} + \eta_h V_h C_{L_\alpha,h} \epsilon_0) \quad (\text{B.34})$$

and

$$C_{m_\alpha} = \frac{x_w}{\bar{c}_w} C_{L_\alpha,w} - \eta_h V_h C_{L_\alpha,h} \left( 1 - \frac{d\epsilon}{d\alpha} \right) \quad (\text{B.35})$$

## *u* Stability Derivatives and Coefficients

For  $X$ , consider the dominant drag and thrust forces for steady flight, i.e.

$$mX = -D + T \quad (\text{B.36})$$

or

$$mX = -Q_\infty S_w C_D + T \quad (\text{B.37})$$

and taking the derivative with respect to  $u$

$$mX_u = -Q_\infty S_w \left( \frac{2\bar{C}_D}{u} + C_{D_u} \right) + T_u \quad (\text{B.38})$$

and defining the coefficient

$$X_u = \frac{Q_\infty S_w}{m\bar{u}} C_{X_u} \quad (\text{B.39})$$

one has

$$C_{X_u} = - \left( 2\bar{C}_D + C_{D_u} \right) + C_{T_u} \quad (\text{B.40})$$

where  $C_{T_u} \approx 0$  for jet engines,  $C_{T_u} \approx -\bar{C}_D$  for piston engines, and  $C_{D_u}$  typically depends on the trim Mach number  $\bar{\mathcal{M}}$ , i.e.

$$C_{D_u} = C_{D_M} \bar{\mathcal{M}} \quad (\text{B.41})$$

where the Mach number is defined as

$$\bar{\mathcal{M}} = \frac{\bar{v}_\infty}{\bar{v}_s} \quad (\text{B.42})$$

where  $\bar{v}_s$  is the speed of sound at trim, and varies with altitude. Also, note that for low speed flight

$$C_{D_u} \approx 0 \quad (\text{B.43})$$

For  $Z$ , consider the dominant force

$$mZ = -L \quad (\text{B.44})$$

or by lifting-line theory

$$mZ = -Q_\infty S_w C_L \quad (\text{B.45})$$

and taking the derivative with respect to  $u$ , one has

$$mZ_u = -Q_\infty S_w \left( \frac{2\bar{C}_L}{u} + C_{L_u} \right) \quad (\text{B.46})$$

Defining the coefficient as

$$Z_u = \frac{Q_\infty S_w}{m\bar{u}} C_{Z_u} \quad (\text{B.47})$$

one has

$$C_{Z_u} = - (2\bar{C}_L + \bar{u} C_{L_u}) \quad (\text{B.48})$$

The Prantl-Glaudent formula for the lift coefficient of compressible flow given the incompressible lift coefficient,  $C_{L,M=0}$ , states

$$C_L = \frac{1}{\sqrt{1 - \bar{\mathcal{M}}^2}} C_{L,M=0} \quad (\text{B.49})$$

one can show

$$\bar{u} C_{L_u} = \left( \frac{\bar{\mathcal{M}}^2}{1 - \bar{\mathcal{M}}^2} \right) \bar{C}_L \quad (\text{B.50})$$

Then,

$$C_{Z_u} = - \left( 2 + \frac{\bar{\mathcal{M}}^2}{1 - \bar{\mathcal{M}}^2} \right) \bar{C}_L \quad (\text{B.51})$$

For  $M$ , the stability coefficient is defined as

$$M_u = \frac{Q_\infty S_w \bar{c}_w}{I_{yy} \bar{u}} C_{m_u} \quad (\text{B.52})$$

for which the stability coefficient can also be rewritten as

$$C_{m_u} = C_{m_M} \bar{\mathcal{M}} \quad (\text{B.53})$$

and for low speed flight

$$C_{m_u} \approx 0 \quad (\text{B.54})$$

### $\alpha$ Stability Derivatives and Coefficients

For  $X$ , consider the coefficient form where one has

$$C_X = C_L\alpha - C_D + C_T \quad (\text{B.55})$$

thus, the stability coefficient with respect to  $\alpha$  is

$$C_{X_\alpha} = \bar{C}_L - C_{D_\alpha} \quad (\text{B.56})$$

where, by lifting-line theory, one has

$$C_{D_\alpha} = \frac{2\bar{C}_L C_{L_\alpha}}{\pi A R_w e_w} \quad (\text{B.57})$$

where the stability derivative is defined as

$$X_\alpha = \frac{Q S_w}{m} C_{X_\alpha} \quad (\text{B.58})$$

For  $Z$ , consider

$$C_Z = -(C_L + C_D\alpha) \quad (\text{B.59})$$

$$C_Z = -(C_{L_\alpha}\alpha + C_{L_0} + C_D\alpha) \quad (\text{B.60})$$

$$C_Z = -(C_{L_\alpha} + C_D)\alpha - C_{L_0} \quad (\text{B.61})$$

thus, the stability coefficient with respect to  $\alpha$  is

$$C_{Z_\alpha} = -(C_{L_\alpha} + \bar{C}_D) \quad (\text{B.62})$$

where the stability derivative is defined as

$$Z_\alpha = \frac{Q_\infty S_w}{m} C_{Z_\alpha} \quad (\text{B.63})$$

For  $M$ , the stability derivative is defined as

$$M_\alpha = \frac{Q_\infty S_w \bar{c}_w}{I_{yy}} C_{m_\alpha} \quad (\text{B.64})$$

where the stability coefficient was derived previously for static stability as

$$C_{m_\alpha} = \frac{x_w}{\bar{c}_w} C_{L_{\alpha,w}} - \eta_h V_h C_{L_{\alpha,h}} \left( 1 - \frac{d\epsilon}{d\alpha} \right) \quad (\text{B.65})$$

It should be noted that sometimes a  $C_{m_{\alpha,f}}$  term is included in this derivative to account for fuselage effects.

### *q* Stability Derivatives and Coefficients

For  $Z$ , the primary change is due to the change of the lift at the horizontal tail. Thus, consider modeling the change in tail lift as

$$\Delta L_h = Q_h S_h C_{L_\alpha, h} \Delta \alpha_h \quad (\text{B.66})$$

where due to a rotation rate of  $q$  provides by small angle approximation

$$\Delta \alpha_h = -\frac{x_h q}{\bar{u}} \quad (\text{B.67})$$

Thus,

$$\Delta L_h = -Q_h S_h C_{L_\alpha, h} \frac{x_h q}{\bar{u}} \quad (\text{B.68})$$

or in terms of the stability derivative where  $L_h$  acts opposite the  $z_B$  axis

$$m \Delta Z = Q_h S_h C_{L_\alpha, h} \frac{x_h q}{\bar{u}} \quad (\text{B.69})$$

and taking the derivative with respect to  $q$

$$Z_q = Q_h S_h C_{L_\alpha, h} \frac{x_h}{m \bar{u}} \quad (\text{B.70})$$

By definition of the stability derivative as

$$Z_q = \frac{Q_\infty S_w \bar{c}_w}{2m \bar{u}} C_{Z_q} \quad (\text{B.71})$$

the stability coefficient is

$$C_{Z_q} = \frac{2Q_h x_h S_h}{Q_\infty S_w \bar{c}_w} C_{L_\alpha, h} \quad (\text{B.72})$$

Then, defining

$$V_h = \frac{-x_h S_h}{S_w \bar{c}_w} \quad (\text{B.73})$$

as the **horizontal tail volume ratio** and

$$\eta_h = \frac{Q_h}{Q_\infty} \quad (\text{B.74})$$

as the **horizontal tail efficiency**, one has

$$C_{Z_q} = -2\eta_h V_h C_{L_\alpha, h} \quad (\text{B.75})$$

Note that  $x_h$  is a negative number here. It should also be noted that typical values for  $\eta_h$  are 0.8-1.2 where  $\eta_h$  can be  $> 1$  because of slipstream or engine stream effects, i.e. the interaction of the airflow with the wing or engine before encountering the horizontal tail.

For  $M$ , consider the moment due to  $Z_q$  as modeled by the horizontal tail, i.e.

$$I_{yy} M_q = -x_h m Z_q \quad (\text{B.76})$$

By definition of the stability derivative as

$$M_q = \frac{Q_\infty S_w \bar{c}_w^2}{2I_{yy}\bar{u}} C_{m_q} \quad (\text{B.77})$$

and substituting for  $Z_q$  from the previous derivation, the stability coefficient is

$$C_{m_q} = -\frac{x_h}{\bar{c}_w} C_{Z_q} \quad (\text{B.78})$$

Note that  $Z_q$  was neglected in the linearized EOMs in the previous lecture due to its small magnitude relative to  $\bar{u}$ , though it appears indirectly in  $M_q$ .

Lastly, note that a common practice with these  $q$  derivatives is to increase these models by 10% to approximately account for the wing and fuselage as well.

### $\dot{\alpha}$ Stability Derivatives and Coefficients

For  $Z$ , consider the change in the angle of attack primarily changing the circulation around the wing. This most directly affects the downwash at the tail which occurs at a lag time approximated by

$$\Delta t = \frac{-x_t}{\bar{u}} \quad (\text{B.79})$$

Furthermore, the change in downwash can be related to the change in the horizontal tail angle of attack as

$$\Delta\alpha_h = \frac{d\epsilon}{dt} \Delta t \quad (\text{B.80})$$

Then, by substitution

$$\Delta\alpha_h = \frac{d\epsilon}{dt} \frac{-x_h}{\bar{u}} \quad (\text{B.81})$$

$$\Delta\alpha_h = \frac{d\epsilon}{d\alpha} \frac{d\alpha}{dt} \frac{-x_h}{\bar{u}} \quad (\text{B.82})$$

$$\Delta\alpha_h = \frac{d\epsilon}{d\alpha} \dot{\alpha} \frac{-x_h}{\bar{u}} \quad (\text{B.83})$$

and relating this to the change in the lift, one has

$$\Delta L_h = Q_h S_h C_{L_{\alpha},h} \Delta\alpha_h \quad (\text{B.84})$$

or in terms of  $Z$  where  $L_h$  acts opposite the  $z_B$  axis

$$m\Delta Z = -Q_h S_h C_{L_{\alpha},h} \frac{d\epsilon}{d\alpha} \dot{\alpha} \frac{-x_h}{\bar{u}} \quad (\text{B.85})$$

and taking the derivative with respect to  $\dot{\alpha}$

$$Z_{\dot{\alpha}} = -Q_h S_h \frac{-x_h}{m\bar{u}} C_{L_{\alpha},h} \frac{d\epsilon}{d\alpha} \quad (\text{B.86})$$

By definition of the stability derivative as

$$Z_{\dot{\alpha}} = \frac{Q_{\infty} S_w \bar{c}_w}{2m\bar{u}} C_{Z_{\dot{\alpha}}} \quad (\text{B.87})$$

the stability coefficient is

$$C_{Z_{\dot{\alpha}}} = -2V_h \eta_h C_{L_{\alpha,h}} \frac{d\epsilon}{d\alpha} \quad (\text{B.88})$$

For  $M$ , consider the moment due to  $Z_{\dot{\alpha}}$  as modeled by the horizontal tail, i.e.

$$I_{yy} M_{\dot{\alpha}} = -x_h m Z_{\dot{\alpha}} \quad (\text{B.89})$$

By definition of the stability derivative as

$$M_{\dot{\alpha}} = \frac{Q_{\infty} S_w \bar{c}_w^2}{2I_{yy}\bar{u}} C_{m_{\dot{\alpha}}} \quad (\text{B.90})$$

and substituting for  $Z_{\dot{\alpha}}$  from the previous derivation, the stability coefficient is

$$C_{m_{\dot{\alpha}}} = -\frac{x_h}{\bar{c}_w} C_{Z_{\dot{\alpha}}} \quad (\text{B.91})$$

Note that  $Z_{\dot{\alpha}}$  was neglected in the linearized EOMs in the previous lecture due to its small magnitude relative to  $\bar{u}$ , although it appears indirectly in  $M_{\dot{\alpha}}$

### $\delta_T$ and $\delta_e$ Control Derivatives and Coefficients

The longitudinal control inputs are the elevator angle,  $\delta_e$ , and throttle input  $\delta_T$ . The throttle coefficients and derivatives are determined solely by the engine type and can vary with the placement of the engines, e.g. under the wings. However, their modeling is beyond the scope of this course, but, in general the thrust can affect  $C_X$ ,  $C_Z$ , and  $C_m$ . This implies the following definitions for the control derivatives

$$X_{\delta_T} = \frac{Q_{\infty} S_w}{m} C_{X_{\delta_T}} \quad (\text{B.92})$$

$$Z_{\delta_T} = \frac{Q_{\infty} S_w}{m} C_{Z_{\delta_T}} \quad (\text{B.93})$$

$$M_{\delta_T} = \frac{Q_{\infty} S_w \bar{c}_w}{I_{yy}} C_{m_{\delta_T}} \quad (\text{B.94})$$

where the coefficients will depend on the design of the propulsion system.

Recalling for airplanes, one typically models the propulsive force as

$$\vec{F}_{p,B} = \begin{bmatrix} T \cos \theta_T \\ 0 \\ T \sin \theta_T \end{bmatrix} \quad (\text{B.95})$$

where  $T$  is the **thrust force** and  $\theta_T$  is a potential offset angle with respect to the  $x_B$ -axis of the body-fixed frame. Moreover, a propulsive moment may also be present nominally about the  $y_B$ -axis as

$$\vec{M}_{p,B} = \begin{bmatrix} 0 \\ T(z_T \cos \theta_T - x_T \sin \theta_T) \\ 0 \end{bmatrix} \quad (\text{B.96})$$

where  $(x_T, z_T)$  denotes the location of the thrust force in the  $x_B - z_B$  plane. It should be noted that often  $\theta_T = 0^\circ$  or approximately and  $T$  is generally a function of the airspeed, altitude, and throttle setting. And modeling the thrust force as a **thrust coefficient**, i.e.

$$T = Q_\infty S_w C_T \quad (\text{B.97})$$

one has

$$C_{X_{\delta_T}} = \cos \theta_T C_T \quad (\text{B.98})$$

$$C_{Z_{\delta_T}} = \sin \theta_T C_T \quad (\text{B.99})$$

$$C_{m_{\delta_T}} = (z_T \cos \theta_T - x_T \sin \theta_T) C_T \quad (\text{B.100})$$

The elevator angle input affects the aerodynamics of the horizontal tail which will affect the lift and induced drag of the tail, the second of which is relatively small compared to the total airplane drag and will be neglected in this course's models. Thus, the elevator control derivatives are defined as

$$Z_{\delta_e} = \frac{Q_\infty S_w}{m} C_{Z_{\delta_e}} \quad (\text{B.101})$$

$$M_{\delta_e} = \frac{Q_\infty S_w \bar{c}_w}{I_{yy}} C_{m_{\delta_e}} \quad (\text{B.102})$$

where the coefficients can be modeled using the tail lift effects as

$$C_{Z_{\delta_e}} = -C_{L_{\delta_e}} = -\eta_h \frac{S_h}{S_w} C_{L_{\delta_e}, h} \quad (\text{B.103})$$

$$C_{m_{\delta_e}} = -\frac{x_h}{\bar{c}_w} C_{Z_{\delta_e}} \quad (\text{B.104})$$

$C_{m_{\delta_e}}$  is also known as the **elevator control power** while the derivative  $C_{L_{\delta_e}}$  is the **elevator effectiveness** and can be rewritten as

$$C_{L_{\delta_e}} = \eta_h \frac{S_h}{S_w} \frac{\partial C_{L,h}}{\partial \delta_e} = \eta_h \frac{S_h}{S_w} \frac{\partial C_{L,h}}{\partial \alpha_h} \frac{\partial \alpha_h}{\partial \delta_e} = \eta_h \frac{S_h}{S_w} C_{L_{\alpha,h}} \tau_e \quad (\text{B.105})$$

where the  $\tau$  **empirical parameter** models the relationship between the change in the lifting surface's angle of attack and the control surface deflection angle. Intuitively,  $\tau$  depends on the ratio of the control surface area to the lifting surface area, e.g. if  $S_e$  is the surface area of the elevator, then  $\tau_e$  is related to  $S_e/S_h$ . This term will also be used for the rudder and aileron control coefficients. The following table provides discrete values for  $\tau$ .

$S_{control}/S_{lifting}$	0.0	0.05	0.10	0.15	0.20	0.25	0.30	0.35
$\tau$	0.0	0.16	0.26	0.34	0.41	0.47	0.52	0.56
$S_{control}/S_{lifting}$	0.40	0.45	0.50	0.55	0.60	0.65	0.70	
$\tau$	0.60	0.64	0.68	0.72	0.75	0.78	0.80	

For intermediate values one typically uses linear interpolation.

### B.3 Lateral-Directional Stability and Control Derivatives

This section will derive models for the lateral stability and control derivatives used in the Jacobian linearization of the normalized aerodynamic forces and moments about trimmed steady flight, i.e.

$$\begin{aligned} Y &= \bar{Y} + Y_\beta \Delta\beta + Y_p \Delta p + Y_r \Delta r + Y_{\delta_r} \Delta\delta_r \\ L &= \bar{L} + L_\beta \Delta\beta + L_p \Delta p + L_r \Delta r + L_{\delta_a} \Delta\delta_a + L_{\delta_r} \Delta\delta_r \\ N &= \bar{N} + N_\beta \Delta\beta + N_p \Delta p + N_r \Delta r + N_{\delta_a} \Delta\delta_a + N_{\delta_r} \Delta\delta_r \end{aligned} \quad (\text{B.106})$$

where the coefficients of the linear terms are called the **lateral-directional stability and control derivatives**. Alternatively, these equations can be written in terms of nondimensional coefficients as

$$\begin{aligned} C_Y &= \bar{C}_Y + C_{Y_\beta} \Delta\beta + C_{Y_p} \Delta p + C_{Y_r} \Delta r + C_{Y_{\delta_r}} \Delta\delta_r \\ C_l &= \bar{C}_l + C_{l_\beta} \Delta\beta + C_{l_p} \Delta p + C_{l_r} \Delta r + C_{l_{\delta_a}} \Delta\delta_a + C_{l_{\delta_r}} \Delta\delta_r \\ C_n &= \bar{C}_n + C_{n_\beta} \Delta\beta + C_{n_p} \Delta p + C_{n_r} \Delta r + C_{n_{\delta_a}} \Delta\delta_a + C_{n_{\delta_r}} \Delta\delta_r \end{aligned} \quad (\text{B.107})$$

where the coefficients of the linear terms are called the **lateral-directional stability and control coefficients**.

It should be noted that a more comprehensive collection of aerodynamic stability and control prediction techniques can be found in the *USAF Stability and Control DATCOM (Data Compendium)* which was compiled between 1960 and 1978 by the McDonnell Douglas Corporation in conjunction with the Flight Dynamics Laboratory at Wright-Patterson Air Force Base.

Lastly, it should also be noted that in the following models  $S_v$  is the vertical tail area and includes the submerged area to fuselage centerline.

#### Lateral-Directional Static Stability Coefficients

For  $L$  and  $N$  moments, the coefficients are

$$C_l = C_{l,w-f} - \frac{Q_v S_v z_v}{Q_\infty S_w b_w} C_{L,v} \quad (\text{B.108})$$

and

$$C_n = C_{n,w-f} - \frac{Q_v S_v x_v}{Q_\infty S_w b_w} C_{L,v} \quad (\text{B.109})$$

Next, modeling the moment and lift coefficients for small angles of attack as

$$C_{l,w-f} = C_{l_\beta,w-f} \beta \quad (\text{B.110})$$

$$C_{n,w-f} = C_{n_\beta,w-f}\beta \quad (\text{B.111})$$

and

$$C_{L,v} = C_{L_\alpha,v}\alpha_v \quad (\text{B.112})$$

where these surfaces are symmetric with respect to  $\beta$  and  $\alpha_v$ , respectively, thus providing no net moment or force at  $0^\circ$ . Then, one has

$$C_l = C_{l_\beta,w-f}\beta - \frac{Q_v S_v z_v}{Q_\infty S_w b_w} C_{L_\alpha,v} \alpha_v \quad (\text{B.113})$$

and

$$C_n = C_{n_\beta,w-f}\beta - \frac{Q_v S_v x_v}{Q_\infty S_w b_w} C_{L_\alpha,v} \alpha_v \quad (\text{B.114})$$

Then, inspecting the FBD above, one can see that

$$\alpha_v = \beta + \sigma \quad (\text{B.115})$$

where  $\sigma$  can be approximated by a linear function

$$\sigma = \frac{d\sigma}{d\beta} \beta \quad (\text{B.116})$$

which is typically estimated using test data and/or CFD. Next, substituting for  $\alpha_v$  and  $\sigma$ , one has

$$C_l = C_{l_\beta,w-f}\beta - \frac{Q_v S_v z_v}{Q_\infty S_w b_w} C_{L_\alpha,v} \left( \beta + \frac{d\sigma}{d\beta} \beta \right) \quad (\text{B.117})$$

and

$$C_n = C_{n_\beta,w-f}\beta - \frac{Q_v S_v x_v}{Q_\infty S_w b_w} C_{L_\alpha,v} \left( \beta + \frac{d\sigma}{d\beta} \beta \right) \quad (\text{B.118})$$

Finally, defining

$$\eta_v = \frac{Q_v}{Q_\infty} \quad (\text{B.119})$$

as the **vertical tail efficiency** and

$$V_v = \frac{-x_v S_v}{b_w S_w} \quad (\text{B.120})$$

as the **vertical tail volume ratio**, one has

$$C_l = C_{l_\beta}\beta \quad (\text{B.121})$$

$$C_n = C_{n_\beta}\beta \quad (\text{B.122})$$

where

$$C_{l_\beta} = C_{l_\beta,w-f} - \eta_v \frac{S_v z_v}{S_w b_w} C_{L_\alpha,v} \left( 1 + \frac{d\sigma}{d\beta} \right) \quad (\text{B.123})$$

and

$$C_{n_\beta} = C_{n_\beta,w-f} + \eta_v V_v C_{L_\alpha,v} \left( 1 + \frac{d\sigma}{d\beta} \right) \quad (\text{B.124})$$

## $\beta$ Stability Derivatives and Coefficients

For  $Y$ , consider the dominant force from the vertical tail

$$mY = -L_v \quad (\text{B.125})$$

and in coefficient form

$$mY = -Q_v S_v C_{L,v} \quad (\text{B.126})$$

or

$$mY = -Q_v S_v C_{L_{\alpha},v} (\beta + \sigma) \quad (\text{B.127})$$

and taking the derivative with respect to  $\beta$  provides

$$mY_{\beta} = -Q_v S_v C_{L_{\alpha},v} \left( 1 + \frac{d\sigma}{d\beta} \right) \quad (\text{B.128})$$

By definition of the stability derivative as

$$Y_{\beta} = \frac{Q_{\infty} S_w}{m} C_{Y_{\beta}} \quad (\text{B.129})$$

and defining

$$\eta_v = \frac{Q_v}{Q_{\infty}} \quad (\text{B.130})$$

as the **vertical tail efficiency**, the stability coefficient is

$$C_{Y_{\beta}} = -\eta_v \frac{S_v}{S_w} C_{L_{\alpha},v} \left( 1 + \frac{d\sigma}{d\beta} \right) \quad (\text{B.131})$$

The following empirical equation can be used in the previous equation

$$\eta_v \left( 1 + \frac{d\sigma}{d\beta} \right) = 0.724 + 3.06 \frac{S_v/S_w}{1 + \cos \Lambda_w} + 0.4 \frac{z_w - z_f}{d_f} + 0.009 AR_w \quad (\text{B.132})$$

where  $\Lambda_w$  is the sweep angle of the wing (measured at quarter chord),  $z_w$  is the  $z_B$  coordinate of the aerodynamic center of the wing,  $z_f$  is the  $z_B$  coordinate of the fuselage centerline,  $d_f$  is the maximum fuselage depth, and  $AR_w$  is the wing aspect ratio.

For  $L$ , the primary factors are the dihedral angle and the wing taper. The stability derivative is defined as

$$L_{\beta} = \frac{Q_{\infty} S_w b_w}{I_{xx}} C_{l_{\beta}} \quad (\text{B.133})$$

and the stability coefficient can be simply modeled by

$$C_{l_{\beta}} = \frac{\partial C_{l_{\beta}}}{\partial \Gamma_w} \Gamma_w \quad (\text{B.134})$$

where  $\Gamma_w$  is the wing dihedral angle,  $\frac{\partial C_{l_{\beta}}}{\partial \Gamma_w}$  is related to the aspect ratio and taper ratio of the root chord and tip chord ( $\approx -0.4$  to  $-0.9$  /rad<sup>2</sup>).

For  $N$ , the stability derivative is defined as

$$N_\beta = \frac{Q_\infty S_w b_w}{I_{zz}} C_{n_\beta} \quad (\text{B.135})$$

where the stability coefficient was derived previously for static stability as

$$C_{n_\beta} = C_{n_\beta, w-f} + \eta_v V_v C_{L_{\alpha_v}} \left( 1 + \frac{d\sigma}{d\beta} \right) \quad (\text{B.136})$$

where

$$V_v = \frac{-x_v S_v}{b_w S_w} \quad (\text{B.137})$$

is the **vertical tail volume ratio**.

### *p* Stability Derivatives and Coefficients

For  $Y$ , there will be a resulting side force only if there is wing sweep as the change in the relative airflow will only then have a component in the  $y_B$  direction. The stability derivative is defined as

$$Y_p = \frac{Q_\infty S_w b_w}{2m\bar{u}} C_{Y_p} \quad (\text{B.138})$$

It can be shown that the stability coefficient is

$$C_{Y_p} = \frac{AR_w + \cos \Lambda_w}{AR_w + 4 \cos \Lambda_w} \tan \Lambda_w \bar{C}_L \quad (\text{B.139})$$

For  $L$ , the primary factor is the change in lift distribution over the wing. Thus, consider modeling the change in tail lift over a cross-section of the wing of width  $dy$  and chord length  $c_w(y)$  as

$$\Delta(\text{Lift}) = Q_\infty(c_w(y)dy)C_{l_{\alpha,w}}\Delta\alpha \quad (\text{B.140})$$

where  $C_{l_{\alpha,w}}$  is the cross-sectional lift coefficient slope. A rotation rate of  $p$  provides by the small angle approximation

$$\Delta\alpha = \frac{yp}{\bar{u}} \quad (\text{B.141})$$

Thus, for the moment arm  $y$  for the contribution to the normalized moment  $L$  of the sectional mass

$$my^2\Delta L = -Q_\infty C_{l_{\alpha,w}} \frac{p}{\bar{u}} c_w(y) y^2 dy \quad (\text{B.142})$$

which can be integrated over half the span of the wing twice for the entire roll moment  $I_{xx}L$  due to the moment being in the opposite sense for the other side of the wing as

$$I_{xx}L = -2 \int_0^{b_w/2} Q_\infty C_{l_{\alpha,w}} \frac{p}{\bar{u}} c_w(y) y^2 dy \quad (\text{B.143})$$

Then, assuming that the cross-sectional lift coefficient slope is well approximated by the wing lift coefficient slope, one has

$$I_{xx}L = \frac{-2Q_\infty C_{L\alpha,w} p}{\bar{u}} \int_0^{b_w/2} c_w(y) y^2 dy \quad (\text{B.144})$$

and taking the derivative with respect to  $p$

$$I_{xx}L_p = \frac{-2Q_\infty C_{L\alpha,w}}{\bar{u}} \int_0^{b_w/2} c_w(y) y^2 dy \quad (\text{B.145})$$

By definition of the stability derivative as

$$L_p = \frac{Q_\infty S_w b_w^2}{2I_{xx}\bar{u}} C_{l_p} \quad (\text{B.146})$$

The stability coefficient is

$$C_{l_p} = -\frac{4C_{L\alpha,w}}{S_w b_w^2} \int_0^{b_w/2} c_w(y) y^2 dy \quad (\text{B.147})$$

where  $c_w(y)$  is the wing chord as a function of the lateral coordinate. For a tapered wing this chord function can be written as

$$c_w(y) = c_{r,w} \left[ 1 + \left( \frac{\frac{c_{t,w}}{c_{r,w}} - 1}{\frac{b_w}{2}} \right) y \right] \quad (\text{B.148})$$

where  $c_{r,w}$  and  $c_{t,w}$  are the root and tip chords for the wing which results in a stability coefficient

$$C_{l_p} = -\left( \frac{1+3\lambda_w}{1+\lambda_w} \right) \frac{C_{L\alpha,w}}{12} \quad (\text{B.149})$$

where  $\lambda_w$  is the taper ratio of the wing, i.e.

$$\lambda_w = \frac{c_{t,w}}{c_{r,w}} \quad (\text{B.150})$$

For  $N$ , the stability derivative is defined as

$$N_p = \frac{Q_\infty S_w b_w^2}{2I_{zz}\bar{u}} C_{n_p} \quad (\text{B.151})$$

It can be shown that the stability coefficient is

$$C_{n_p} = -\frac{\bar{C}_L}{8} \quad (\text{B.152})$$

which is due to conservation of angular momentum.

### *r* Stability Derivatives and Coefficients

For  $Y$ , consider the dominant force from the vertical tail side force and its moment arm due to the resulting sideslip, i.e.

$$Y_r = \frac{x_v}{\bar{u}} Y_{\beta,v} \quad (\text{B.153})$$

$$Y_r = \frac{x_v Q_\infty S_w}{m \bar{u}} C_{Y_{\beta,v}} \quad (\text{B.154})$$

By definition of the stability derivative as

$$Y_r = \frac{Q_\infty S_w b_w}{2m \bar{u}} C_{Y_r} \quad (\text{B.155})$$

the stability coefficient is

$$C_{Y_r} = 2 \frac{x_v}{b_w} C_{Y_{\beta,v}} \quad (\text{B.156})$$

For  $L$ , the stability derivative is defined as

$$L_r = \frac{Q_\infty S_w b_w^2}{2 I_{xx} \bar{u}} C_{l_r} \quad (\text{B.157})$$

It can be shown that the stability coefficient is

$$C_{l_r} = \frac{\bar{C}_L}{4} - 2 \frac{x_v z_v}{b_w^2} C_{Y_{\beta,v}} \quad (\text{B.158})$$

For  $N$ , consider the dominant force from the vertical tail side force and its moment arm, i.e.

$$I_{zz} \Delta N = -Q_v S_v C_{L_{\alpha,v}} x_v \Delta \beta \quad (\text{B.159})$$

where due to a rotation rate of  $r$  provides by small angle approximation

$$\Delta \beta = \frac{x_v r}{\bar{u}} \quad (\text{B.160})$$

Thus,

$$I_{zz} \Delta N = -Q_v S_v C_{L_{\alpha,v}} \frac{x_v^2 r}{\bar{u}} \quad (\text{B.161})$$

and taking the derivative with respect to  $r$

$$I_{zz} N_r = -Q_v S_v C_{L_{\alpha,v}} \frac{x_v^2}{\bar{u}} \quad (\text{B.162})$$

By definition of the stability derivative as

$$N_r = \frac{Q_\infty S_w b_w^2}{2 I_{zz} \bar{u}} C_{n_r} \quad (\text{B.163})$$

the stability coefficient is

$$C_{n_r} = 2 \eta_v V_v \frac{x_v}{b_w} C_{L_{\alpha,v}} \quad (\text{B.164})$$

Note that  $C_{Y_{\beta,v}} = C_{Y_{\beta}}$  for the tail-only model and  $x_v$  is a negative number in this appendix.

### $\delta_a$ and $\delta_r$ Control Derivatives and Coefficients

The lateral-directional control derivatives are defined as

$$Y_{\delta_r} = \frac{Q_\infty S_w}{m} C_{Y_{\delta_r}} \quad (\text{B.165})$$

$$L_{\delta_a} = \frac{Q_\infty S_w b_w}{I_{xx}} C_{l_{\delta_a}} \quad (\text{B.166})$$

$$L_{\delta_r} = \frac{Q_\infty S_w b_w}{I_{xx}} C_{l_{\delta_r}} \quad (\text{B.167})$$

$$N_{\delta_a} = \frac{Q_\infty S_w b_w}{I_{zz}} C_{n_{\delta_a}} \quad (\text{B.168})$$

$$N_{\delta_r} = \frac{Q_\infty S_w b_w}{I_{zz}} C_{n_{\delta_r}} \quad (\text{B.169})$$

For low speed flight, the following models can be used for the control coefficients.

The  $\delta_a$  control coefficient for  $L$  can be modeled as

$$C_{l_{\delta_a}} = \frac{2C_{L_{\alpha,w}}\tau_a}{S_w b_w} \int_{y_{a,i}}^{y_{a,o}} c_w(y) y dy \quad (\text{B.170})$$

where  $C_{l_{\delta_a}}$  is the **aileron control power**,  $C_{L_{\alpha,w}}\tau_a$  is the **aileron effectiveness**,  $y_{a,i}$  is the inner aileron  $y_B$  coordinate, and  $y_{a,o}$  is the outer aileron  $y_B$  coordinate. For a tapered wing, this becomes

$$C_{l_{\delta_a}} = \frac{2C_{L_{\alpha,w}}\tau_a}{S_w b_w} \int_{y_{a,i}}^{y_{a,o}} c_{r,w} \left[ 1 + \left( \frac{c_{t,w}/c_{r,w} - 1}{b_w/2} \right) y \right] y dy \quad (\text{B.171})$$

$$C_{l_{\delta_a}} = \frac{2C_{L_{\alpha,w}}\tau_a c_{r,w}}{S_w b_w} \left[ \frac{y^2}{2} + \left( \frac{c_{t,w}/c_{r,w} - 1}{b_w/2} \right) \frac{y^3}{3} \right]_{y_{a,i}}^{y_{a,o}} \quad (\text{B.172})$$

The  $\delta_a$  control coefficient for  $N$  can be modeled as

$$C_{n_{\delta_a}} = 2K \bar{C}_L C_{l_{\delta_a}} \quad (\text{B.173})$$

where  $K$  is an aileron empirical factor ( $\approx -0.3$  to  $-0.1$ ) that decreases with aspect ratio, increases slightly with aileron placement further out on the wing, and increases with taper ratio; and

The  $\delta_r$  control coefficient for  $Y$  can be modeled as

$$C_{Y_{\delta_r}} = \frac{S_v}{S_w} \tau_r C_{L_{\alpha,v}} \quad (\text{B.174})$$

The  $\delta_r$  control coefficient for  $L$  can be modeled as

$$C_{l_{\delta_r}} = \frac{S_v}{S_w} \tau_r \frac{z_v}{b_w} C_{L_{\alpha,v}} \quad (\text{B.175})$$

The  $\delta_r$  control coefficient for  $N$  can be modeled as

$$C_{n_{\delta_r}} = -\eta_v V_v C_{L_{\alpha,v}} \tau_r \quad (\text{B.176})$$

where  $C_{n_{\delta_r}}$  is the **rudder control power** and  $C_{L_{\alpha,v}}\tau_r$  is the **rudder effectiveness**.

# Index

- $J_2$   
WGS, 260
- $L_2$ -norm, 803  
weighted, 803
- $L_\infty$ -norm, 803
- $L_p$ -norms, 803
- $L_{2,2}$ -norm, 804
- $L_{\infty,\infty}$ -norm, 804
- $M$ -sample variance, 677
- $N$ -model redundancy, 789
- $\alpha - \beta$  filter, 173
- $\chi^2$ -test, 528
- $\delta$ -generalized labeled multi-Bernoulli random finite set, 783
- $\gamma$ -iteration, 162
- $\mathcal{H}_\infty$  loop-shaping  
mixed-sensitivity, 164
- $\mathcal{L}_1$ -norm  
signal, 57
- $\mathcal{L}_2$ -norm  
signal, 57
- $\mathcal{L}_p$ -norm  
signal, 57
- $\mathcal{L}_{p,q}$ -induced norm  
system, 57
- $\sigma$ -additivity, 446
- $\sigma$ -plot, 56
- $\tau$  empirical parameter, 833  
table, 833
- $k$ -nearest-neighbors association, 755
- $p$ -norm, 803  
signal, 57
- $t$ -test, 528
- a posteriori* PDF, 510
- a priori* PDF, 511
- 2-sample variance, 677
- A-tail, 333
- acceleration  
centrifugal, 224  
Coriolis, 224  
Euler, 224  
fictitious, 224  
inertial, point-mass, 222
- acceleration due to gravity  
standard, 262
- acceleration model  
constant-acceleration, 630  
constant-velocity, 629  
Singer, 631  
white noise, 629
- Wiener process, 629
- Wiener sequence, 630
- acceleration-to-speed ratio, 647

- accelerometer, 651, 717
- accelerometer stochastic error, 656
- acceptance probability, 619
- acceptance-rejection sampling, 467
- active
  - energy sensing, 660
- active nutation control
  - thruster-based, 426
  - wheel-based, 428
- active rotation matrix, 213
- actuation
  - system, 319
- actuator, 74
- adaptive control, 320
  - gain-scheduled, 195
  - NDI-based, 205
- adaptive resampling, 621
- addition law of probability, 447
- adjoint
  - classical matrix, 26
- adjoint equation, 139
- adjoint vector, 138
- adjugate, 26
- aerodynamic chord
  - airfoil, 818
- aerodynamic parameter vector, 646
- aerodynamic twisting, 821
- aerostatic equation, 296
- aerostatic pressure, 297
- affine transformation, 463
- aileron, 330
- aileron-rudder interconnect, 387
- air data, 320, 673
  - triplet, 320, 673
- air data system, 320
- air data systems, 673
- air density, 819
- airfoil, 818
- airplane
  - anatomy, 328
- airplane trim, 338
- airspeed, 253, 297, 818
- alert limit, 794
- algebraic Riccati equation
  - continuous-time, 142
  - discrete-time, 180
- algebraic Riccati inequality, 152
- all-neighbors association, 754
- Allan deviation, 678
- Allan variance, 677
- alternate conditional sampling, 522
- alternate hypothesis, 525
- altimeter
  - barometric, 713
  - lidar, 713
  - pressure, 713
  - radar, 713
  - sensitive, 713
- altimeters, 713
- altitude hold, 340
- ambiguity fixing, 701
  - brute force, 705
  - geometry-free, 705
  - LAMBDA, 705
  - rounding, 705
- amplitude
  - doublet, 33
  - step, 32
- analog system, 13
- analysis plot
  - flight envelope, 355
  - v-n, 359
- angle
  - sweep, 821
- angle of attack, 256
  - airfoil, 818
- angular acceleration
  - orbital, 407
  - three-dimensional, 216
- angular velocity
  - orbital, 407
  - three-dimensional, 216
  - two-dimensional, 215
- aperture, 663
- aperture radar, 663
- argument of periapsis, 405, 406

- argument of perigee, 405
- armature constant, 310, 428
- armature resistance, 309, 428
- artificial satellites, 401, 438
- ascent phase, 442
- aspect ratio
  - wing, 820
- assist maneuver, 402
- association map, 782
- associative laws
  - set, 799
- assumed density filters, 579
- asymptotic slope, 44
- asymptotically stable, 20
- atmospheric delay
  - code phase, 707
- atmosphere, 711
- atmospheric pressure, 674
  - static, 297
- atomic clocks, 675
- attitude and heading reference system, 720
  - air data, 720
- attitude dynamics
  - linearized gravity-gradient, 421
- attitude vector, 215
- augmented proportional navigation guidance, 324
- augmented rendezvous guidance, 325
- augmented sigma-points
  - prediction, 602
- augmented signma-points
  - correction, 603, 604
- augmented state, 135, 541
- augmented state covariance
  - FI-SPKS/SPRTSS process, 604
  - SPKF process, 603
- augmented state estimate
  - FI-SPKS/SPRTSS process, 604
  - SPKF process, 603
- auto-correlation
  - PRN, 670
- auto-correlation function, 472
  - residual, 503
- auto-covariance function, 472
- auto-regressive moving-average process, 477
  - vector, 477
- auto-regressive process, 476
- auto-throttle, 389
- automotive grade IMU, 655
- autonomous multi-model, 569
- autopilot, 11
- availability
  - information system, 795
- available bandwidth, 111
- avionics, 328
- axiomatic probability theory, 446
- axis
  - lateral, 252
  - local-horizontal, 251, 252
  - local-vertical, 251, 252
  - longitudinal, 252
- axis-angle, 215
- axisymmetric, 415
- aymptotic stability, 192
- azimuth, 686
- backward-simulation particle smoother, 626
- backwards recursive Bayes estimator, 539
- ballistic body, 442
- ballistic coefficient, 642
- ballistic flight
  - boost phase, 640
  - coast phase, 640
  - reentry phase, 640
- ballistic flight motion model, 640
- ballistic missiles, 440
- ballistic objects, 640
- ballistic space flight, 401
- ballistic targets, 640
- ballistic vehicles, 442
- ballistics, 442
- band-pass filter control stage, 101
- band-stop filter control stage, 101
- bang-bang property, 145
- bang-off-bang property, 147
- bank angle, 256
- barometric altimeter, 713

- base
  - receiver, 696, 702
- base state, 478
- baseline, 684, 696
- baseline vector, 731
- basic rotation matrices, 217
- battery monitoring system, 786
- Bayes cost, 533
- Bayes criterion, 533
- Bayes estimator, 511
  - parameter, 511
  - state, 536
- Bayes filter, 538
- Bayes parameter estimator, 511
- Bayes predictor, 537
- Bayes risk, 511, 533
- Bayes risk test, 533
- Bayes smoother
  - fixed-interval, 539
- Bayes state estimator, 536
  - median, 536
- Bayes' law
  - two events, 448
- Bayes' rule, 511
  - continuous random variables, 462
  - discrete random variables, 461
  - two events, 448
- Bayes' theorem
  - two events, 448
- Bayesian least-squares parameter estimator, 514
- Bayesian linear regression problem, 515
- Bayesian regression problem, 514
- beacon, 682
- beam-forming, 663
- bearing, 682
- beat frequency, 661, 669
- belief mass function
  - RFS, 485
- Bellman equation, 177
- Bernoulli distribution, 455
- Bernoulli random finite set, 488
- Bernoulli's equation, 296
- best linear unbiased estimator, 495
- between-epochs difference, 706
- bi-static radar, 660
- bias vector, 788
- bicycle motion model, 634
- binary hypothesis test, 526
- binomial coefficient, 776
- biplane, 328
- birth
  - object, 748
- block diagram, 6
- blocked pitot tube, 674
- blocked static port, 675
- Bode gain-phase formula, 89
- Bode plot, 42
  - asymptotic slope, 44
  - DC gain, 43
- bodies, 9
- body
  - ballistic, 442
- body-fixed frame, 252
  - mean-axis, 283
  - principal, 413
- boost
  - flight phase, 335, 439
- boost phase, 442, 640
- bootstrap particle filter, 618
- Borel set, 447
- boundary condition, 14
- bounded real lemma, 152
  - generalized, 198
- bounded set, 194
- Brownian motion, 476
- burn
  - maneuver, 402
- cambered wing, 820
- camera, 663
  - hyperspectral, 663
  - infrared, 663
  - monocular, 663
  - multispectral, 663
  - near-infrared, 663
  - stereoscopic, 663

- visual, 663
- canard, 329
- candidate set, 491
  - measurement, 752
- cardinality, 484, 797
  - classification, 471
- cardinality distribution, 484
- cardinalized probability hypothesis density filter, 777
  - Gaussian mixture, 779
- carrier phase measurement
  - GPS, 669
- carrier phase positioning, 707
- carrier phase range equation
  - GNSS, 707
- cascade control, 317
- cascade interconnection, 72
- Cauchy problem, 190
- Cauchy's argument principle, 816
- Cauchy-Peano Theorem, 190
- Cayley-Hamilton definition, 123
- center of mass, 222
- center of pressure
  - airfoil, 818
- central divided difference, 606
- central limit theorem, 457
  - generalized, 457
- central moment
  - random variable, 454
- Chapman-Kolmogorov equation
  - Bayesian state prediction, 538
  - Markov chain, 474
  - Markov jump process, 475
- characteristic equation
  - ODE, 26
  - SISO feedback control system, 81
    - state-space, 26
- characteristic exponent, 457
- characteristic function, 451
  - random vectors, 460
- characteristic polynomial
  - ODE, 26
  - SISO feedback control system, 81
- state-space, 26
- Cholesky decomposition, 808
- chronometers, 675
- civil time, 677
- classic Euler angles, 217
- classical adjoint of matrix, 26
- classical control theory, 78
- classical mechanics, 9
- classical optimal control problem, 138, 143
- classification problem, 526
- climb axis, 253
- clock
  - radio, 676
- clock drift, 677
  - receiver, 699
  - satellite, 710
- clock drift rate
  - satellite, 710
- clock error
  - satellite, 707
- clock network, 678
- clock offset, 678
  - satellite, 710
- clock skew, 678
- clocks, 675
  - atomic, 675
  - quartz, 675
- Clohessy-Wiltshire equations, 409
- closed set, 194
- closed-loop observer, 121, 173
- closed-loop transfer function, 79
- clutter, 748, 749
  - clutter rate
    - Poisson, 762
  - clutter spatial density, 762
- co-domain
  - function, 797
- coast, 404
  - flight phase, 439
  - spacecraft, 402
- coast phase, 640
- cockpit
  - airplane, 328

- helicopter, 396
- code division multiple access, 668
- code phase measurement
  - GPS, 670
- code phase positioning, 706
- coefficient of determination, 501
- colored-measurement-noise Kalman filter, 562
- column dimension, 800
- column rank, 801
- combination
  - ionosphere-free, 711
- commutative laws
  - set, 798
- compact set, 194
- complement
  - set, 797
- complementary cumulative distribution function,
  - 451
- complementary filter, 542
- complementary filtering, 720
- complementary sensitivity
  - transfer function, 80
- complex matrix, 800
- complex numbers, 797
- complex parametric uncertainty, 126
- complex vector, 799
- component build-up model, 366, 821
- composite hypothesis test, 526
- compressible flow, 296
- computer vision, 663, 683
- conditional PDF, 462
- conditional PMF, 461
- conditional probability, 447
- cone, 150
- cone program, 150
- confidence ellipse
  - Gaussian, 469
- confidence interval
  - $100(1-\alpha)$ , 503
- confidence level, 525
- confidence region
  - multivariate Gaussian, 468
  - univariate Gaussian, 456
- coning maneuver, 433
- conjugate prior, 512
- consistency
  - estimator, 492
- constant-mass approximation, 228
- constellation
  - satellite, 664
- constrained eigenvectors, 238
- consumer grade IMU, 655
- continuous-time stochastic state-space model,
  - 477
- continuously operating reference stations, 708
- control
  - automatic, 11
  - closed-loop, 75
  - feedback, 75
  - feedforward, 74
  - manual, 11
  - open-loop, 74
  - semi-automatic, 11
  - system, 319
- control design
  - performance requirements, 88
  - stability requirements, 88
- control horizon, 187
- control input
  - nonlinear, 204
  - virtual LTI, 202
- control moment gyro, 314
- control power
  - aileron, 840
  - elevator, 833
  - rudder, 840
- control stage
  - band-pass filter, 101
  - band-stop filter, 101
  - derivative, 97
  - high-pass filter, 100
  - integral, 96
  - lag, 102
  - lag-lead, 103
  - lead, 103
  - low-frequency boost, 101

- low-pass filter, 98
- notch filter, 101
- proportional, 95
- control theory
  - classical, 78
- control-affine dynamical system, 201
- controllability
  - mode, 30
  - output, 125
  - state, 123
  - under constraints, 125
- controllability Gramian
  - discrete-time, 174
  - LTI systems, 124
  - LTV systems, 195
- controllability grammian, 59
- controllability matrix, 124
- controllability test
  - Popov-Belevitch-Hautus, 124
- Controllable Canonical Form, 16
- controlled variable, 201
- controller, 8
- conventional tail, 329
- convex set, 150
- convext set, 151
- convolution integral, 35
- convolutional neural network, 664
- Cooper-Harper Rating Scale, 372
- cooperating multi-model, 569
  - MAP base state estimator, 572
  - MAP mode sequence estimator, 572
  - MMSE base state estimator, 570
  - MMSE modal state estimator, 571
- coordinated flight, 354
- coordinated steady-flight condition, 338
- coordinated universal time, 677
- coordinated-turn motion model, 634
- copula, 467
- Coriolis effect, 654
- corner reflector, 660
- correction step, 749, 751
  - Bayes filter, 539
  - extended Kalman filter, 584, 587
- information filter, 551
- innovation-saturated extended Kalman filter, 590
- iterative extended Kalman filter, 588
- Kalman filter, 546
- second-order extended Kalman filter, 593
- sigma-point Kalman filter, 603
- statistically linearized (Kalman) filter, 597
- correlated-noise Kalman filter, 559
  - continuous-time, 567
- correlated-noise Kalman-Bucy filter, 567
- correlation
  - random process, 474
- correlation coefficient, 465
- correlation matrix, 464
- cost function, 176
- cost functional, 137
- cost matrix, 140
  - cross-, 140
  - endpoint, 140
  - input, 140
  - state, 140
  - terminal, 140
- cost-to-come function, 184
- cost-to-go
  - continuous-time, 140
- cost-to-go function, 177
- costate equation, 139
- costate vector, 138
- countable, 797
- countably infinite set, 797
- counting numbers, 797
- counting process, 475
- covariance
  - sample, 513
- covariance intersection, 560
  - algorithm, 560
- covariance intersection Kalman filter, 560
- covariance matrix, 465
- covariance stationarity, 473
- critical lift, 646
- cross product matrix, 802
- cross-correlation

- function, 472
- cross-covariance
  - function, 472
  - noise, 556
- cross-covariance function
  - continuous-time noise, 565
- cross-covariance matrix, 465, 466
- crossover frequency separation, 318
- cruciform tail, 331
- cruise missiles, 335
- cumulative distribution function, 450
  - joint, 460
- cycle count
  - fractional, 703
- cycle counting, 703
- damped least-squares, 506
- data association
  - $k$ -nearest-neighbors, 755
  - all-neighbors, 754
  - global all-neighbor, 756
  - global nearest-neighbor, 755
  - nearest-neighbor, 754
- data association problem, 749
- DC gain, 43
- De Morgan's laws, 799
- dead reckoning, 681
  - navigation system, 733
- death
  - object, 748
- decade
  - frequency, 44
- decibel table, 43
- decision altitude, 341
- decision height, 341
- decision theory, 136, 491, 535
- declination
  - magnetic, 659
- decomposition
  - eigenvalue, 27
- deconvolution, 35
- decouples, 28, 69
- deep GNSS/INS integration, 742
- definiteness
  - matrix, 804
- deformable-body model, 9
- degeneracy problem, 616
- delay
  - dry, 712
  - hydrostatic, 712
  - wet, 712
- delay atmospheric
  - code phase, 707
- delay ionospheric
  - code phase, 707
- delay lock loop
  - vector, 742
- delay margin, 84
- delay tropospheric
  - code phase, 707
- delayed gratification, 506
- derivative control stage, 97
- derivative gain, 105
- descent direction, 492
- descent phase, 442
- design matrix, 496
- detectability, 125
- detection
  - fault probability, 793
- detection theory
  - signal, 526
- detection threshold, 526
- detector, 526
  - maneuver, 634
- deviation of the normal, 250
- diagonal
  - matrix, 800
- diagonal matrix, 801
- diagonalizable, 27, 805
- difference
  - set, 798
- difference equations, 65
- differential game theory, 136
- differential pseudorange, 683
- differential range, 682
- differentiator

- pure, 44
- diffuse prior, 762
- digital system, 13
- dilution of precision, 694
- dimension
  - row, 800
  - vector, 799
  - vector space, 801
- Dirac delta, 34
- direct method
  - Lyapunov, 192
- direction cosine matrix
  - three-dimensional, 216
  - two-dimensional, 213
- directional plane, 822
- discrete Fourier transform, 67
- discrete-time dynamics equation, 64
- discrete-time Fourier transform, 66
- discrete-time Gaussian random walk, 479
- discrete-time Kalman filter
  - one-step posterior, 547
  - one-step prior, 547
- discrete-time output equation, 64
- discrete-time random walk, 479
- discrete-time stochastic state-space model, 477
- discrete-time stochastic state-space model with measurement-origin uncertainty, 757
- discrete-time transfer function, 66
- discretization, 175
- discretized LTI stochastic state-space model, 481
- discretized measurement noise covariance, 481
- discretized process noise covariance, 480
- discrimination problem, 526
- discriminator, 526
- disjoint set, 798
- disk margin, 85
- distance
  - Mahalanobis, 466
- distance measuring equipment, 342
- distances
  - statistical, 466
- distinct label indicator, 484
- distributive laws
- set, 799
- divergence
  - Kullback-Leibler, 466
  - statistical, 466
- domain
  - function, 797
- Doppler effect, 683
- Doppler positioning, 700
- Doppler positioning equation, 699
- Doppler shift
  - GPS, 669
- dot product, 801
- double-differenced
  - cycle count range, 703
  - phase-based range, 702
  - pseudorange, 697
- doublet amplitude, 33
- doublet length, 33
- downwash angle, 822
  - equation, 826
- drag coefficient, 299
  - wing, 819
  - wing, induced, 820
  - wing, parasitic, 820
  - wing, profile, 820
  - wing, wave, 820
  - wing, zero-lift, 820
- drag force, 299
  - airfoil, 818
  - fixed-wing aircraft, 335
- drag parameter, 642
- dry delay, 712
- Dryden gust model, 305
- dual estimation, 541
- Dubins path OCP
  - 2D, 148
- Dubins path problem, 148
- Dubins vehicle, 148
- duplexer, 662
- dutch-roll mode, 369
- Dvoretzky-Kiefer-Wolfowitz inequality
  - multivariate, 461
  - univariate, 452

- dynamic inversion
  - LTI controller, 202
  - nonlinear controller, 205
- dynamic inversion control, 201
- dynamic pressure, 674
  - free-stream, 819
  - incompressible, 297
- dynamic programming, 177
- dynamic programming equation, 177
- dynamic uncertainty, 127
  - LTI, 127
- dynamic-controller feedback
  - discrete-time, 172
- dynamical system
  - control-affine, 201
  - linear-in-control, 201, 207
  - random, 477
  - stochastic, 477
- dynamical systems theory, 9
- dynamics
  - equation, 13
- dynamics equation
  - discrete-time state-space, 64
  - state-space, 15
- Earth angular rate
  - WGS, 248
- Earth disk, 717
- Earth Gravitational Models, 260
- Earth gravitational parameter
  - WGS, 248
- Earth horizon sensor, 717
- Earth radius
  - mean, 273
- Earth rotation angle, 677
- Earth rotation rate, 738
- Earth sensor, 717
- Earth's gravitational parameter
  - WGS, 260
- Earth's mean radius, 247, 262
- Earth's radius
  - mean, 247, 262
- Earth-centered, inertial, 247
- eccentricity, 405
  - WGS, 248
- eccentricity vector, 404
- effective navigation ratio, 323
- effective number threshold, 621
- effectiveness
  - aileron, 840
  - elevator, 833
  - rudder, 840
- efficient estimator, 495
- eigenfunctions, 230
- eigenvalue, 804
- eigenvalue decomposition, 27, 804
- eigenvalue equation
  - right, 27
- eigenvalue placement, 123
- eigenvalue problem, 804
- eigenvector
  - left, 804
- elastic stability and control derivative, 380
- elastic-body model, 9
- electro-optical sensor, 663
- electromotive force constant, 309, 428
- element
  - matrix, 800
  - set, 797
  - vector, 799
- elementary symmetric function, 777
- elevation, 686
- elevation angle, 258
- elevator, 329
- elevons, 329
- ellipsoidal coordinates, 248
- ellipsoidal gates, 753
- Ellipsoidal Gravity Formula
  - WGS 84, 261
- elliptical wing, 821
- empennage, 329
- empirical distribution function, 452, 791
  - multivariate, 461
- empty set, 797
- endpoint cost, 138
- engine

- airplane, 329
- rocket, 439
- entry-wise matrix norm, 804
- Epanechnikov kernel, 620
- ephemeris, 671
- ephemeris error
  - satellite, 707
- epoch, 404
- equation of motion, 13
  - Newton, inertial frame, 224
  - Newton, rotating frame, 225
- equations of motion
  - Newton-Euler, 227
  - Newton-Euler, rotating frame, 229, 260
- rigid aircraft, 348
- rigid aircraft alternative, 349
- rotation equation, 227
- satellite attitude, 414
- six degrees-of-freedom, 229
- translation equation, 227
- equator, 247, 248
- equatorial radius
  - WGS, 248
- equilibrium
  - flight conditions
    - rigid-body, 349
    - point-mass flight conditions, 337
- equilibrium linearization, 195
- equilibrium point, 20, 64
  - time-varying systems, 191
- equivariant filtering, 720
- error
  - estimator, 494
- error dynamics
  - LTI dynamic inversion, 202
  - nonlinear dynamic inversion, 204
- error transfer function, 79
- error-state correction step, 610
- error-state extended Kalman filter, 611
- error-state filter
  - EKF prediction step, 611
  - ES-EKF correction step, 612
  - prediction step, 610
- error-state filtering, 720
- error-state Kalman filtering, 610
- error-state measurement update step, 610
- error-state sigma-point Kalman filter, 613
- error-state time update step, 610
- error-state vector, 610
- estimator, 8
  - consistency, 492
  - efficient, 495
  - error, 494
  - linear, 495
  - maximum likelihood, 491
  - mean, 494
  - method of moments, 493
  - variance, 494
- estimator gain matrix, 495
- Euclidean norm, 803
- Euler angle
  - argument of periapsis, 406
  - attack, 256
  - bank, 256
  - ECEF-to-navigation, 254
  - ECI-to-ECEF, 254
  - elevation, 258
  - flight-path, 256
  - geodetic latitude, 254
  - heading, 256, 258
  - inclination, 406
  - inertial-to-body, 255, 406
  - longitude, 254
  - longitude of the ascending node, 406
  - lvlh-to-body, 255
  - navigation-to-body, 255
  - navigation-to-VTC, 258
  - navigation-to-wind, 256
  - nutation, 255
  - pitch, 255
  - precession, 255
  - roll, 255
  - sideslip, 256
  - spin, 255
  - wind-to-body, 256
  - yaw, 255

- Euler angle ambiguity, 218
- Euler angle rates
  - $3 - 1 - 3$ , 219
  - $3 - 2 - 1$ , 219
- Euler angles, 217
  - classic, 217
  - proper, 217
  - Tait-Bryan, 217
  - yaw-pitch-roll, 217
- Euler axis, 215
- Euler integration
  - forward, 811
- Euler symmetric parameters, 220
- Euler vector, 215
- Euler's formula, 28
- event, 445
- event space, 446
- events
  - elementary, 447
- exceedance, 451
- excess kurtosis, 456
- exit velocity, 277
- expectation
  - random process, 472
- expectation function, 472
- expectation-maximization algorithm, 519
- expected average power
  - random process, 473
- explained sum of squares, 502
- exponential function
  - matrix, 22
  - univariate, 21
- extended Kalman Filter
  - hybrid-time, 586
- extended Kalman filter
  - continuous-time, 586
  - error-state, 611
  - iterative, 587
  - multiplicative, 612
  - second-order, 591
- extended Kalman particle filter, 622
- extended Kalman smoother
  - fixed-interval, 593
- extended Kalman-Bucy filter, 586
- extended object tracking, 749
- extended proportional navigation guidance, 326
- extended Rauch-Tung-Striebel smoother, 593
- exteroceptive sensor, 482
- extreme value theory, 456
- F-test, 529
- fair, 446
- false alarm, 526
  - probability, 794
- fast Fourier transform, 67
- fault, 788
- fault detection
  - exclusion, 789
  - isolation, 789
- fault detection and isolation, 788
- fault vector, 788
- feasibility problem, 151
- feature, 682, 748
- feature map, 715
- feature points, 715
- feature tracker, 729
- feature tracking, 748
- feedback
  - discrete-time dynamic-controller, 172
- feedback control system
  - classical, 78
- feedback interconnection, 73
- feedback linearization control
  - input-output, 201
- feedback linearization loop
  - LTI, 202, 204
- feedthrough matrix, 15, 64
- filter
  - SISO LTI, 76
- filter bank
  - elementary, 568
- filtered derivative, 107
- filtering
  - equivariant, 720
  - error-state, 720, 742
- final orbital errors, 711

- final value, 32
- final value theorem, 18
- finite burn, 402
- finite mixture distribution, 452
- finite set, 797
- finite-wing theory, 818
- fins, 439
- first point of Aries, 247
- first-order approximation
  - function, 23
- Fisher information matrix, 492
- fixed solution, 705
- fixed-gain controller, 171
- fixed-interval Fourier-Hermite Kalman smoother, 599
- fixed-interval Kalman filter, 593, 598
- fixed-interval Kalman smoother, 548
- fixed-interval quadrature Kalman smoother, 609
- fixed-interval sigma-point Kalman smoothers, 602
- fixed-interval state smoothing posterior PDF, 537
- fixed-lag Kalman smoother, 553
- fixed-lag state smoothing posterior PDF, 537
- flank angle, 674
- flap, 334
- flattening
  - WGS, 248
- flicker frequency noise, 677
- flicker noise, 661
- flight condition
  - engine-out, 352
- flight dynamics and control, 11
- flight envelope plot, 355
- flight phase
  - ascent, 439, 442
  - ballistic, 439
  - boost, 442
  - descent, 439, 442
  - impact, 439
  - launch, 442
  - lift-off, 439
  - reentry, 439
  - terminal, 439
- terminal phase, 442
- flight profile, 335
- flight-path angle, 256
- float solution, 704
- fly-by maneuver, 402
- flying qualities, 370
- flywheel, 311
- folded cumulative distribution function, 451
- force
  - aerodynamic, 262
  - gravitational, 260
  - propulsive, 262
  - radiation pressure, 262
  - thrust, 336, 833
- force of light, 262
- forces
  - fictitious, 224
- forward value function, 184
- forwards Euler integration, 811
- Fourier inversion theorem, 41
- Fourier series, 41
- Fourier transform, 41
- Fourier-Hermite Kalman filter, 599
- Fourier-Hermite Rauch-Tung-Striebel smoother, 599
- fractional cycles, 700
- free air correction
  - gravity, 261
- free response, 31
  - LTI MIMO continuous-time, 55
- free-stream airflow, 252
- free-stream velocity
  - airfoil, 818
- frequency
  - angular, 41
  - beat, 669
  - intermediate, 669
  - periodic, 41
- frequency indexes, 503
- frequency modulated continuous wave radar, 662
- frequency resolution, 67
- frequency response, 41
  - gain, 40

- magnitude, 40
- phase, 41
- frequency separation
  - crossover, 318
- frequentist probability theory, 445
- Frobenius norm, 804
- Froude number, 299
- fuel monitoring system, 785
- full rank, 801
- function, 797
  - co-domain, 797
  - domain, 797
  - invertible, 798
  - mapping, 797
  - one-to-one, 798
  - range, 798
- functional, 137
- fundamental clock rate
  - GPS, 669
- fundamental transfer functions, 79
- fuselage, 439
  - airplane, 328
  - helicopter, 396
- fuselage frame, 350
- gain
  - frequency response, 40
- gain margin, 81
- gain scheduling, 320
- gain-scheduled adaptive control, 195
  - divide-and-conquer, 197
- Gamma function
  - multivariate, 470
- gate probability, 762
- gate volume, 753, 762
- gating step, 752
  - two-stage, 753
- gating threshold
  - ellipsoidal, 753, 762
  - rectangular, 753
- Gauss-Hermite cubature rule, 608
- Gauss-Hermite quadrature rule, 608
- Gauss-Markov process, 476
- Gauss-Markov theorem, 499
- Gauss-Newton algorithm, 505
- Gaussian distribution, 455
  - multivariate, 468
- Gaussian filtering, 579
- Gaussian process, 476
- Gaussian random walk
  - discrete-time, 479
- Gaussian sum filter, 567
- Gaussian-mixture
  - posterior approximation, 518
- generalized calculus of variations, 138
- generalized control input, 118
  - discrete-time, 170
- generalized covariance intersection
  - algorithm, 560
- generalized disturbance, 118
  - discrete-time, 170
- generalized eigenvalue problem, 237
- generalized eigenvectors, 806
- generalized error, 118
  - discrete-time, 170
- generalized labeled multi-Bernoulli random finite set family, 489
- generalized least-squares
  - estimator, 504
- generalized least-squares problem, 499
- generalized likelihood-ratio, 530
- generalized likelihood-ratio test, 531
- generalized LTI feedback controller, 118, 171
- generalized mass matrix, 240
- generalized matched filter, 533
- generalized plant, 118
- generating variate
  - scale mixture, 453
- geocentric orbits, 405
- geocentric rates
  - navigation frame, 267
- geodesy, 247
- geodetic frame, 251
- geodetic latitude, 254
- geodetic rates
  - navigation frame, 267

- geographic north, 659
- geographic position
  - celestial body, 682
- geoid, 247
- geometry matrix, 694
- geometry-free, 705
- geometry-free combination, 707
- geopotential, 296
- geostationary orbit, 666
- geosynchronous orbit, 666
- Gibbs sampler, 522
- Gilvenko-Cantelli theorem, 452
- gimbal, 652
- gimbal lock, 653
- glide
  - flight phase, 440
- glide-slope, 342
- glide-slope guidance system, 344
- global all-neighbor data association, 756
- global hypothesis
  - MHT, 767
- global nearest-neighbor association, 755
- global positive definite, 193
- global stability, 192
- global uniform asymptotic stability, 192
- globally asymptotically stable, 20
- globally negative definite, 193
- globally negative semi-definite, 193
- globally positive semi-definite, 193
- globally stable, 20
- GNSS
  - carrier phase range equation, 707
  - pseudorange equation, 706
- GNSS/INS integration
  - deep, 742
- goodness-of-fit test, 526
- Gramian
  - controllability, 124
  - observability, 126
- gravitational harmonic
  - second-order, 641
- gravity
  - Somigliana, 739
- gravity assist maneuver, 402
- gravity drag, 441
- gravity model
  - ellipsoidal-Earth, 641
  - flat-Earth, 641
  - spherical-Earth, 641
- gravity turn, 441
- gravity turn motion model, 646
  - rocket, 647
- gravity vector, 261
- gravity-gradient moment
  - linearized, 421
- gravity-gradient stabilization
  - regions, 423
- great circle, 339
- Greenwich mean time, 682
- ground range, 686
- ground segment, 665
- groundspeed, 253
- guidance
  - system, 319
- guidance law
  - augmented proportional navigation, 324
  - augmented rendezvous, 325
  - extended proportional navigation, 326
  - line-following, 319
  - proportional navigation, 323
  - rendezvous, 323
- guidance, navigation, and control, 318
- gust, 305
- gust model
  - Dryden, 305
  - von Kármán, 305
- gyrocompassing, 736
- gyroscope, 652
  - gyrostabilized, 652
  - strapdown, 652
- gyroscope stochastic error, 656
- gyroscopic precession, 653
- gyrostabilized gyroscope, 652
- gyrostat, 429
- gyrotorquer, 314
- H-tail, 331

- Hamilton-Jacobi-Bellman equation, 139
- Hamiltonian, 138
- handling qualities, 370
- hard-iron distortion, 659
- hardware-defined radio, 669
- hardware-in-the-loop, 8
- hazardously misleading information, 793
  - probability, 793
- heading angle, 256, 258
- heading hold, 340
- heavy-tailed, 613
- height above average terrain, 713
- height above ground level, 713
- helicopter
  - anatomy, 396
- heliocentric frame, 246
- Hermitian matrix, 801
- Hermitian transpose, 800
- hidden Markov model assumption, 536
- hidden Markov models, 479
- high Earth orbit, 666
- high-pass filter control stage, 100
- higher-order terms, 23
- higher-order unscented transform, 605
- Hill equations, 409
- Hill frame, 251
- hold control
  - altitude, 340
  - heading, 340
  - speed, 339
- holding time, 475
- homogeneous coordinates, 220
- homogeneous solution
  - ODE, 21
- horizontal tail
  - efficiency, 827
  - volume ratio, 827
- horology, 675
- Hungarian algorithm, 755
- Hurwitz matrix, 804
- hybrid process, 478
- hybrid-state estimation, 568
- hybrid-time feedback control system, 175
- hybrid-time stochastic state-space model, 477
- hydrostatic delay, 712
- hyper-parameters
  - conjugate prior, 512
- hyperspectral camera, 663
- hypotheses, 525
- hypothesis
  - alternate, 525
  - null, 525
- hypothesis test
  - M*-ary, 526
  - $\chi^2$ -test, 528
  - t*-test, 528
  - Bayes risk, 533
  - binary, 526
  - classification, 526
  - composite, 526
  - discrimination problem, 526
  - F-test, 529
  - generalized likelihood-ratio, 531
  - null, 526
  - simple, 526
  - Z-test, 527
- hypothesis testing
  - probability of error, 533
- hypothesis testing problem, 525
- hypothesis-oriented multi-hypothesis tracking, 767
- ideal gas equation, 299
- identity matrix, 16, 800
- IID cluster random finite set, 487
- illumination, 660
- image, 660
- image processing, 683
- image-based navigation, 714
- images, 683
- impact pressure
  - incompressible, 297
- importance function, 614
- importance sampling Monte Carlo, 523
- importance weights
  - normalized, 615

- impulse function, 34
- impulsive maneuver, 402
- IMU grade
  - automotive, 655
  - consumer, 655
  - industrial, 655
  - marine, 655
  - navigation, 655
  - tactical, 655
- IMU stochastic error, 656
- inclination, 405, 406
- inclusion-exclusion principle, 447
- incompressible dynamic pressure, 297
- incompressible flow, 296
- incompressible impact pressure, 297
- increment, 471
- incremental nonlinear dynamic inversion controller, 205
- indefinite matrix, 805
- independence
  - events, 448
  - pairwise, 465
  - random process, 472
- independent and identically distributed, 492
- independent increments, 473
- index set, 471
- indicator, 452
- indirect method
  - Lyapunov, 192
- induced matrix norm, 804
- induced norm
  - system, 57
- industrial grade IMU, 655
- inertia, 222
  - matrix, 227
  - moments, 227
  - products, 227
  - tensor, 227
- inertia matrix
  - aircraft, 347
  - stability frame, 351
- inertia tensor
  - satellite, 413
- inertia wheel, 311
- inertial measurement units, 654
- inertial sensors, 651
- infinite dimensional problem, 230
- information filter, 550
- information gain, 466
- information matrix, 550
- information set, 757
- information state, 757
- infrared camera, 663
- infrared sensor, 663
- inherent errors
  - ADS, 675
- initial condition, 14
- initial condition response, 31
- initial value problem, 31, 190
- initialization step
  - BPF, 618
  - EKPF, 622
  - extended Kalman filter, 583, 586
  - innovation-saturated extended Kalman filter, 590
  - iterative extended Kalman filter, 588
  - Kalman filter, 546
  - RBPF, 624
  - second-order extended Kalman filter, 592
  - sigma-point Kalman filter, 602
  - SPPF, 622
  - statistically linearized (Kalman) filter, 596
- injection maneuver, 402
- injective
  - matrix, 801
- inner product, 776
  - vector, 801
- inner-outer loop control, 317
- innovation covariance, 546
- innovation process, 546
- innovation-saturation extended Kalman filter, 589
- input
  - doublet, 33
  - impulse, 33
  - signal, 14
  - step, 32

- vector, 15, 64
- input matrix
  - continuous-time, 15
  - discrete-time, 64
- input-output feedback linearization control, 201
- INS attitude update
  - continuous-time, 736
- INS position update
  - continuous-time, 737
- INS velocity update
  - continuous-time, 737
- insertion maneuver, 402
- instability
  - LTI system, 29, 69
  - modal continuous-time, 28
  - modal discrete-time, 69
- instant cost, 138
- instrument guidance system, 342
- instrument landing systems, 342
- integer ambiguity, 701
- integral control stage, 96
- integral quadratic constraint, 127
- integration
  - loose, 741
  - tight, 741
- integration-points, 520
- integrator
  - pure, 45
- integrity monitor, 793
- integrity monitoring system, 793
- intensity, 771
- intensity function, 486
- interacting multi-model
  - filter, 572
  - Kalman filter, 572
- intercept guidance problem, 320
- intercept problem, 322
- interior point method, 151
- intermediate frequency, 669
- internal model principle, 132
- intersection
  - set, 798
- inverse
  - matrix, 800
  - inverse distribution function, 451
    - generalized, 451
  - inverse transform sampling, 467
  - inverted V-tail, 332
  - invertible
    - function, 798
  - ionosphere-free combination, 711
  - ionospheric delay
    - code phase, 707
  - ISMC estimator, 524
  - iterative extended Kalman filter, 587
  - iterative least-squares
    - estimator, 505
  - J2000 epoch, 247
  - Jacobian, 24
  - jerk model
    - white noise, 629
  - jet-damping effect, 280
  - joint CDF
    - random process, 471
  - joint cumulative distribution function, 460
  - joint estimation, 541
  - joint PDF
    - random process, 471
  - joint PMF
    - random sequence, 471
  - joint probabilistic data association filter, 764
    - coupled, 765
  - joint probability density function, 460
  - joint probability mass function, 460
  - joint state, 541
  - Jordan blocks, 805
  - Jordan canonical form, 28, 806
    - discrete-time, 69
  - Jordan Chains, 55
  - Joseph form, 546
  - Julian date, 677
  - Julian day number, 677
  - jump-linear system, 478
  - Kalman filter
    - colored-measurement-noise, 562

- continuous-time, 565
- correlated-noise, 559
- covariance intersection, 560
- multi-state constraint, 746
- nonlinear, 579
- quadrature, 609
- sigma-point, 602
- square-root, 551
- Kalman gain
  - continuous-time, 575
  - discrete-time, 576
  - hybrid-time, 577
- Kalman smoother
  - fixed-interval, 548
  - fixed-interval Fourier-Hermite, 599
  - fixed-interval quadrature, 609
  - fixed-interval sigma-point, 602
  - fixed-lag, 553
- Kalman-Bucy filter, 565
- Kalman-Yacubovich-Popov lemma, 152
- Kepler's equation, 672
- Keplerian elements, 404, 671
- kernel, 802
  - Epanechnikov, 620
- kernel PDF, 620
- keying
  - amplitude-shift, 668
  - frequency-shift, 668
  - phase-shift, 668
- keypoints, 715
- kinematic, 696
- kinematics
  - point-mass, 222
- kinetics
  - point-mass, 222
- Kolmogorov-Smirnov statistic, 452
- Krasovskii-LaSalle theorem, 194
- Kullback-Leibler divergence, 466
- kurtosis, 454
  - excess, 456
- L-band, 668
- L-moment, 493
- labeled IID cluster random finite set, 488
- labeled multi-Bernoulli random finite set, 489
- labeled Poisson random finite set, 488
- labeled random finite set, 485
  - $\delta$ -generalized multi-Bernoulli, 783
  - generalized multi-Bernoulli family, 489
  - IID cluster, 488
  - multi-Bernoulli, 489
  - Poisson, 488
- lag control stage, 102
- lag-lead control stage, 103
- Lagrangian term, 138
- LAMBDA method
  - baseline-constrained, 732
- landmarks, 715
- Laplace transform, 17
  - inverse, 17
- latent variables
  - Rao-Blackwellized particle filter, 623
- lateral plane, 823
- latitude, longitude, altitude
  - geocentric, 247
  - geodetic, 248
- launch
  - flight phase, 439
- launch phase, 442
- law of gravitation, 260
- law of total probability
  - continuous random variables, 461, 462
  - events, 448
- lead angle, 321
- lead control stage, 103
- leading diagonal
  - matrix, 800
- leap second, 677
- learning rate, 492
- least upper bound, 804
- least-squares
  - constrained regression problem, 507
  - damped, 506
  - iterative, 505
- least-squares estimator, 499
  - nonlinear, 505

- least-squares problem
  - generalized, 499
  - ordinary, 499
  - weighted, 499
- least-squares regression problem, 499
  - linear, 499
- left half plane, 29, 69
- level, 352
- level flight, 338
- level set, 193
- leveling, 736
- Levenburg-Marquardt algorithm, 506
  - modified, 506
- lever arm
  - GNSS/INS, 746
- LFT inverse
  - lower, 63
  - upper, 63
- lidar, 663
  - lidar altimeter, 713
  - lidar flow, 734
- Lie derivative, 204
- lift coefficient, 299
  - wing, 819
- lift force, 299
  - airfoil, 818
    - fixed-wing aircraft, 335
- lift-curve slope
  - airfoil, 821
  - wing, 820
- lifting-line theory, 821
- likelihood, 448
- likelihood function
  - multi-target, 759
- likelihood-ratio, 529
  - generalized, 530
- likelihood-ratio test, 530
  - generalized, 531
- limiter, 341
- line-of-sight, 686, 694
- line-of-sight angle, 321
- linear fractional transformation, 60
  - lower, 60
  - upper, 60
- linear least-squares estimator, 499
- linear least-squares regression problem, 499
- linear matrix equality, 808
- linear matrix inequality, 808
  - strict, 808
- linear objective problem, 151
- linear parameter estimator, 495
- linear time-invariant, 14
- linear velocity
  - apparent, 223, 224
- linear, parameter-varying system, 198
- linear-in-control dynamical system, 201, 207
- linear-quadratic estimator, 155
  - continuous-time, 565, 575
  - discrete-time, 576, 577
- linear-quadratic regulator, 140
  - extended, 185
  - iterative, 184
- linear-quadratic-Gaussian OCP, 155
- linearization
  - cosine, 25
  - error, 23
  - Jacobian, 24
  - sine, 25
  - theorem, 23
  - univariate, 23
- linearized dynamics
  - gravity-gradient attitude, 421
  - relative orbital, 409
- linearized LTI state-space model
  - discrete-time, 65
- linearized rigid airplane dynamics
  - lateral-directional, 364
  - longitudinal, 364
- lines-of-position, 682
- Lipschitz condition, 190
- lithographic construction, 654
- local tangent plane (LTP), 248
- local-vertical, local-horizontal frame
  - nadir-pointing, 251
- local-vertical, local-horizontal, normal frame
  - nadir-pointing, 251

- localization, 681
- localizer, 342
- localizer guidance system, 346
- localizer-type direction aid, 342
- locally negative definite, 193
- locally negative semi-definite, 193
- locally positive definite, 192
- locally positive semi-definite, 193
- location, 681
- log-likelihood ratio
  - target tracking, 770
- long-period mode, 369
- longitude angle, 254
- longitude of ascending node, 405
- longitude of the ascending node, 406
- longitudinal plane, 821
- loop bandwidth, 90
- loop phase error, 679
- loop-shaping
  - control stages, 88, 94
  - mixed-sensitivity  $\mathcal{H}_\infty$ , 164
  - SISO, 88
- loop-shaping design, 95
- loop-shifting, 119
- loose integration, 741
- low Earth orbit, 666
- low-frequency boost
  - control stage, 101
- low-pass filter control stage, 98
- low-thrust maneuver, 402
- lower triangular matrix, 801
- LQR-smoothing, 186
- LTI dynamic inversion controller, 202
- LTI MIMO continuous-time stability, 55
- LTI MIMO free response
  - continuous-time, 55
- LTI stochastic state-space model
  - discretized, 481
- LTI system
  - normality, 145
- Luenberger observer equation
  - continuous-time, 575
  - discrete-time, 576
- hybrid-time, 577
- Luenberger observer matrix, 121, 173
- lumped-mass model, 230
- Lyapunov function, 193
- Lyapunov function candidate, 193
- Lyapunov stability of equilibrium point
  - time-varying system, 192
- Lyapunov's direct method, 192
- Lyapunov's indirect method, 192
- Mach number, 298
- machine vision, 663
- magnetic compass, 658
- magnetic declination, 659
- magnetic heading, 720
- magnetic map, 714
- magnetic navigation, 714
- magnetic north, 659
- magnetic torquer, 315
- magnetic variation, 659
- magnetic-based navigation, 714
- magnetometer, 658
  - three-axis, 659
- magnetorquer, 315
- magnitude
  - frequency response, 40
- Mahalanobis distance, 754
- main diagonal
  - matrix, 800
- major diagonal
  - matrix, 800
- maneuver
  - time constant, 631
- maneuvering objects, 748
- map
  - feature, 715
  - localization, 681
- mapping
  - function, 797
- marginal likelihood, 511
- marginal parameter posterior PDF, 540
- marginal PDF, 461
- marginal stability

- LTI system, 29, 69
- modal continuous-time, 28
- modal discrete-time, 69
- marginal state posterior PDF, 540
- marginalized particle filter, 624
- marine grade IMU, 655
- marker beacons, 342
- Markov chain, 474
- Markov chain Monte Carlo, 522
- Markov jump system, 478
- Markov jump-linear system, 479
- Markov jump-mean acceleration model, 633
- Markov process, 474
- Markov property, 474
  - $n^{\text{th}}$ -order, 477
- Marquardt parameter, 506
- mask angle, 711
- mass matrix
  - generalized, 240
- master control station, 666
- matched filter, 532
  - generalized, 533
- matched uncertainties, 201
- mathematical optimization, 136
- mathematical programming, 136
- matrix
  - correlation, 464
  - covariance, 465
  - cross-covariance, 465, 466
  - diagonal, 801
  - element, 800
  - Hermitian, 801
  - lower triangular, 801
  - norm, 803
  - notation, 800
  - orthogonal, 800
  - skew-symmetric, 801
  - square, 800
  - symmetric, 801
  - unitary, 800
  - upper triangular, 801
  - variance-covariance, 465
- matrix decomposition
- Cholesky, 808
- eigenvalue, 804
- QR, 807
- singular value, 806
- matrix exponential
  - Cayley-Hamilton, 123
- matrix exponential function, 22
- matrix inverse, 800
- matrix inversion lemma, 515, 800
- matrix Lyapunov equation, 59
- matrix rank, 801
- matrix transpose, 800
  - conjugate, 800
  - Hermitian, 800
- maximum *a posteriori* estimator, 511
- maximum likelihood estimator, 491
- maximum modulus theorem, 116
- maximum norm
  - matrix, 804
  - vector, 803
- maximum overshoot, 33
  - percentage, 33
- maximum posterior
  - Bayes state estimator, 536
- maximum singular value, 807
- Mayer term, 138
- mean, 454
  - estimator, 494
  - function, 472
  - random process, 472
  - random vector, 464
  - sample, 512
- mean motion, 409
- mean radius of Earth, 273
- mean-axis constraints, 283
- measure, 447
- measurement
  - candidate set, 752
- measurement equation, 478
  - multi-target, 759
- measurement error, 496
- measurement noise, 478
- measurement noise covariance

- discretized, 481
- measurement origin uncertainty, 748
- measurement update step, 749, 751
  - error-state, 610
- measurement-origin uncertainty problem, 749
- median, 454
  - Bayes estimator, 512
  - Bayes state estimator, 536
- medium Earth orbit, 666
- member
  - set, 797
- memoryless process, 474
- method of moments estimator, 493
- metric, 466
- Metropolis-Hastings algorithm, 619
- MHA-MCMC parameter estimator, 523
- micro-electro-mechanical systems, 654
- micro-electromechanical systems, 654
- mid-course
  - flight phase, 439
- minimal realization, 18
- minimum  $\mathbb{P}_E$  test, 534
- minimum mean square error, 494
  - Bayes state estimator, 536
- minimum mean-square error
  - Bayes parameter estimator, 511
- minimum phase system, 90
- minimum singular value, 807
- minimum variance unbiased estimator, 494
- minimum-energy optimal control problem, 143
- minimum-fuel optimal control problem, 146
- minimum-time optimal control problem, 144
- minimum-time/fuel optimal control problem, 147
- mis-identified fault
  - probability, 794
- missed detection, 526
  - hazardously misleading information, 794
- missile
  - ballistic, 440
  - cruise, 335
- missions, 370
- mixed-sensitivity  $\mathcal{H}_\infty$  loop-shaping, 164
- mixture
- PDFs, 758
- modal analysis, 55
- modal coordinates, 230
- modal state, 478
- mode, 55, 454
  - system, 27
- mode approximation
  - phugoid, 369
  - roll, 370
  - short-period, 369
  - spiral, 370
- mode shapes, 230
- mode-finding algorithm, 517
- model
  - deformable-body, 9
  - elastic-body, 9
  - point particle, 9
  - point-mass, 9
  - rigid-body, 9
- model evidence, 511
- model predictive control, 137, 187
- model set, 568
- model-based design, 6
- modulation
  - binary code, 668
  - binary offset carrier, 668
  - phase, 668
  - split spectrum, 668
- molecular-scale temperature, 300
- moment
  - aerodynamic, 262, 271
  - propulsive, 262, 271
  - radiation pressure, 262, 271
  - random variable, 453
- moment-generating function, 451
  - random vectors, 460
- moments of inertia
  - principal, 413
- momentum exchange devices, 311
- momentum wheel, 311
- momentum-biased satellites, 434
- monitor stations, 666
- monitoring system

- battery, 786
- fuel, 785
- integrity, 793
- mono-static radar, 660
- monocular camera, 663
- Monte Carlo
  - importance sampling, 523
  - Markov chain, 522
  - methods, 521
  - rejection sampling, 521
- Moore-Penrose inverse, 500, 801
- motion model
  - ballistic flight, 640
  - bicycle, 634
  - coordinated-turn, 634
  - standard curvilinear, 634
  - turning, 633
- motor moment constant, 310
- moving-average process, 477
- multi-Bernoulli distribution, 455
- multi-Bernoulli mixture random finite set, 488
- multi-Bernoulli random finite set, 488
- multi-hypothesis tracking, 767
  - hypothesis-oriented, 767
  - track-oriented, 768
- multi-modal, 613
- multi-model
  - autonomous, 569
  - cooperating, 569
  - target tracking, 634
  - variable-structure, 569
- multi-model filtering, 568
  - interacting, 572
- multi-sensor information fusion, 541
- multi-sensor system, 541
- multi-state constraint Kalman filter, 746
- multi-static radar, 660
- multi-target filtering density, 759
- multi-target measurement equation, 759
- multi-target posterior density, 759
- multi-target process equation, 759
- multi-target transition density, 759
- multi-tracker fusion, 749
- multiangulateration, 682
- multiangulation, 682
- multilateration, 682, 691
- multiplicative extended Kalman filter, 612
- multispectral camera, 663
- Murty's algorithm, 756
- mutual orthogonality
  - random process, 474
- mutually exclusive, 447, 798
- mutually independence
  - random processes, 474
- mutually independent, 449
- nadir, 716
- National Marine Electronics Association, 708
- natural numbers, 797
- Navigation, 681
- navigation, 319, 748
  - image-based, 714
  - magnetic, 714
  - magnetic-based, 714
  - system, 319
  - terrain-aided, 741
  - terrain-based, 741
  - terrain-contour, 741
  - terrain-matching, 741
  - terrain-referenced, 741
  - terrain-relative, 741
  - vision-based, 714
  - visual-based, 714
- navigation frame, 251
- navigation grade IMU, 655
- navigation state, 681, 735, 748
  - INS, 736
- navigation system
  - dead reckoning, 733
- navigation systems, 733
- navigational stars, 731
- NavSat, 664
- near-infrared camera, 663
- nearest-neighbor association, 754
- negative definite matrix, 805
- negative semi-definite matrix, 805

- nested-loop control, 317
- net acceleration, 642
- net angular impulse, 225
- net force impulse, 224
- network
  - clock, 678
- Newton equation of motion
  - inertial, 224
  - rotating, 225
- Newton's second law, 224
- Newton-Raphson algorithm, 517
- Neyman-Pearson lemma, 530
- no-wind approximation, 348
- no-wind condition, 297
- noise, 535
  - flicker, 661
  - measurement, 478
  - process, 478
  - shot, 661
- noise cross-covariance, 556
- noise cross-covariance function
  - continuous-time, 565
- nominal-state update step, 611
- nominal-state vector, 610
- non-informative prior, 511
- non-isolable
  - fault probability, 794
- nonlinear dynamic inversion controller, 205
- nonlinear Kalman filtering, 579
- nonlinear least-squares estimator, 505
- nonlinear least-squares regression problem , 505
- norm
  - $L_2$ , 803
  - $L_\infty$ , 803
  - $L_p$ , 803
  - $L_{2,2}$ , 804
  - $L_{\infty,\infty}$ , 804
  - $p$ , 803
  - entry-wise matrix, 804
  - Euclidean, 803
  - Frobenius, 804
  - induced matrix, 804
  - matrix, 803
- matrix maximum, 804
- taxicab, 803
- vector, 802
- vector maximum, 803
- normal distribution, 455
- normal load factor, 353
- normal LTI system, 145
- normal modes, 239
- normalized Butterworth polynomial, 98
- normalized importance weights, 615
- nose
  - airplane, 328
  - ballistic vehicle, 439
  - helicopter, 396
- notch filter control stage, 101
- null hypothesis, 525
- null hypothesis test, 526
- null set, 797
- null space, 802
- nullity, 802
- numerical integration, 520, 810
  - fixed-step Runge-Kutta, 810
  - Heun's method, 811
  - midpoint method, 811
  - Ralston's method, 811
  - RK(4), 811
  - Runge-Kutta (1), 811
- nutation angle, 255
- nutation control
  - active thruster-based, 426
  - active wheel-based, 428
  - passive wheel-based, 427
  - wheel-based, 427
- nutation damper, 419
- nutation frequency, 426
- Nyquist frequency, 67
- Nyquist plot, 41, 812
- Nyquist stability criterion
  - simplified SISO, 816
  - SISO, 816
- Oberth maneuver, 402
- object

- birth, 748
- death, 748
- spawning, 748
- object motion uncertainty, 749
- object state, 748
- object tracking, 748
  - extended, 749
  - group, 749
  - point, 749
- object tracking system, 748
- objective function, 176
- objective functional, 137
- objects, 9
- oblate body, 417
- oblate spheriod, 248
- obliquity of ecliptic, 246
- observability, 125
  - mode, 30
  - state, discrete-time, 174
- observability Gramian, 126
  - LTV systems, 195
- observability grammian, 59
- observability matrix, 125, 175
- observability test
  - Popov-Belevitch-Hautus, 125
- observation equation, 478
- observation error, 496
- observation matrix, 496
- observer feedback control, 121, 172
- observer feedback control system, 121, 173
- OCP
  - linear-quadratic, 140
- ODE solution
  - homogeneous, 21
  - particular, 21
- odometer, 734
- odometry, 734
- off-diagonal
  - matrix, 800
- one-to-one
  - function, 798
- onset time, 634
- onto
- function, 798
- open circuit voltage, 786
- open-loop observer, 121, 173
- open-loop transfer function, 78
- operator, 11
- optical flow, 734
- optical gyroscopes, 653
- optimal control
  - function, 137
  - problem, 136
  - sequence, 176
- optimal control problem
  - fundamental stochastic, 574
  - linear-quadratic-Gaussian, 574
  - minimum-energy, 143
  - minimum-fuel, 146
  - minimum-time, 144
    - minimum-time/fuel, 147
- optimal control theory, 136
- optimal estimation
  - parameter, 262, 363
- optimal estimator
  - parameter, 491
- optimal guidance law
  - intercept, 325
- optimal information fusion
  - general, 542
  - linear-Gaussian, 542
- optimal tradeoff curve, 158
- optimality equation, 177
- orbit
  - geostationary orbit, 666
  - geosynchronous, 666
  - high Earth, 666
  - low Earth, 666
  - medium Earth, 666
- orbit determination
  - orbit estimation, 649
  - preliminary, 649
- orbit estimation, 649
- orbital angular momentum, 404
- orbital elements, 404
- orbital mean motion, 407

- orbital mechanics, 403
- orbital space flight, 401
- orbital vehicles, 401, 438
- order statistic, 493
- ordinary differential equation, 14
  - linear, 15
  - linear time-invariant, 16
  - order, 14
  - proper, 14
  - time-invariant, 15
  - unforced, 14
- ordinary least-squares
  - estimator, 500
- ordinary least-squares problem, 499
- orientation vector, 215
- orthogonal matrix, 800
- orthogonality
  - random processes, 474
- Oswald efficiency
  - wing, 820
- Oswald's efficiency
  - effective, 335
- outcome space, 446
- outcomes
  - frequentist, 445
- outer tracking loop
  - LTI, 202, 204
- output
  - signal, 14
  - vector, 15, 64
- output controllability, 125
- output equation, 15, 64
- output feedback control, 120, 171
- output matrix, 15
  - discrete-time, 64
- overbound, 790
  - one-sided, 790
  - two-sided, 790
- overbounding, 790
  - error, 794
- parallel interconnection, 72
- parameter estimation, 491
  - problem, 491
  - parameter estimator, 491
    - BLS, 516
    - ISMC, 524
    - least-squares, 514
    - MHA-MCMC, 523
    - optimal, 491
    - RSMC, 522
    - VB, 520
  - parameter vector, 198
  - parametric uncertainty, 126
    - complex, 126
    - real, 126
  - Parseval's theorem, 159
  - partial-fraction decomposition, 18
  - particle depletion, 619
  - particle dynamics, 222
  - particle filter, 614
    - extended Kalman, 622
    - extended/sigma-point, 622
    - marginalized, 624
    - Rao-Blackwellized, 624
    - sigma-point, 622
  - particle smoother
    - backward-simulation, 626
    - reweighted, 626
  - particular solution
    - ODE, 21
  - partition
    - set, 798
  - passive
    - energy sensing, 660
  - passive nutation control
    - wheel-based, 427
  - passive rotation matrix
    - two-dimensional, 213
  - passive rotation quaternion, 220
  - PDAF
    - prediction step, 762
  - peak time, 33
  - perception, 482
  - perception systems, 482
  - perfect-gas equation, 299

- performance index, 137
- performance requirements
  - control design, 88
- perifocal frame, 405
- period
  - function, 40
- permutation coefficient, 776
- perturbation
  - continuous-time vector, 24
  - discrete-time vector, 64
  - form, 23
  - scalar, 23
- phase
  - frequency response, 41
- phase ambiguity
  - receiver, 701
  - transmitter, 701
- phase delay
  - receiver, 701
  - transmitter, 701
- phase margin, 83
- phase wind-up, 701, 707
- phase-based range equation, 701
- phases of flight
  - aircraft, 334, 398
- phasing manuever, 402
- phugoid mode, 369
- PI control
  - with rate feedback, 108
- PID
  - interacting form, 106
  - parallel form, 105
  - proportional gain, 105
  - series form, 106
  - standard form, 106
- PID control
  - integral gain, 105
- PIDF control
  - filtered series form, 107
  - filtered standard form, 107
  - parallel form, 107
- pilot, 11
- pitch angle, 255
- pitch damper, 369
- pitching moment
  - airfoil, 819
- pitching moment coefficient
  - wing, 819
- pitchover maneuver, 441
- pitot pressure, 674
- pitot tube
  - blocked, 674
- pitot-static system, 674
- plan, 319
- planning, 319
  - mission, 319
  - path, 319
  - system, 319
  - trajectory, 319
- planning, control, and perception, 7
  - aerospace, 7
- plant, 8
- plant modeling, 8
- point cloud, 663
- point clouds, 683
- point object tracking, 749
- point particle, 9
- point-mass acceleration
  - inertial, 222
- point-mass dynamics, 222
- point-mass model, 9
- point-mass velocity
  - inertial, 222
- Pointryagin's principle, 139
- Poisson distribution, 455
- Poisson process, 475
  - spatial, 762
- Poisson random finite set, 487
- polar form, 40
- polar radius
  - WGS, 248
- pole placement, 123
- pole-zero cancellation, 18
- pole-zero plot, 30
- poles
  - system, 27

- polynomial-matrix model, 20, 366
- Popov-Belevitch-Hautus controllability test, 124
- Popov-Belevitch-Hautus observability test, 125
- pose, 681
- position-domain integration, 741
- positioning
  - carrier phase, 707
  - code phase, 706
  - Doppler, 700
- positioning, navigation, and timing, 681
- positive definite matrix, 805
- positive semi-definite matrix, 805
- posix time, 677
- possibility space, 446
- posterior PDF, 510
  - state filtering PDF, 537
  - state prediction, 537
  - state smoothing, 537
- posterior probability, 448
- power
  - detector, 526
- power spectral density, 473
- powered explicit guidance, 442
- PPP
  - random finite set, 487
- precession
  - prograde, 417
  - retrograde, 417
- precession angle, 255
- precision approach, 341
- precision matrix, 465
- prediction, 535
- prediction horizon, 187
- prediction step
  - Bayes filter, 538
  - extended Kalman filter, 583, 587
  - fixed-lag Kalman smoother, 556
  - information filter, 551
  - innovation-saturated extended Kalman filter, 590
  - iterative extended Kalman filter, 588
  - Kalman filter, 546
  - PDAF, 762
- second-order extended Kalman filter, 592
- sigma-point Kalman filter, 602
- square-root Kalman filter, 551
- statistically linearized (Kalman) filter, 596
- preliminary orbit determination, 649
- pressure
  - aerostatic, 297
  - atmospheric, 674
  - atmospheric static, 297
  - dynamic, 674
  - incompressible dynamic, 297
  - incompressible impact, 297
  - pitot, 674
  - ram, 674
  - stagnation, 297, 674
  - static, 674
  - total, 297, 674
- pressure altimeter, 713
- primary diagonal
  - matrix, 800
- prime meridian, 247, 248
- principal diagonal
  - matrix, 800
- principal moments of inertia, 413
- principle of optimality, 140, 176
- principle of superposition, 13, 15
- prior editing, 619
- prior PDF, 511
- prior probability, 448
- probabilistic data association filter, 761
  - joint, 764
  - joint coupled, 765
- probability
  - survival, 758
- probability density function, 450
  - joint, 460
- probability distribution
  - $K$ , 458
  - $\Gamma$ , 457
  - (student's)  $t$ , 457
  - Bernoulli, 455
  - Cauchy, 456, 457
  - central  $\chi^2$ , 458

- dboule exponential, 458
- exponential, 457
- finite mixture, 452
- Gaussian, 455
- Gaussian scale mixture, 469
- Gaussian-inverse-Wishart, 470
- Inverse-Wishart, 470
- Laplace, 458
- Levy, 457
- Levy  $\alpha$ -stable, 456
- multi-Bernoulli, 455
- noncentral  $\chi^2$ , 458
- normal, 455
- Pareto, 456
- Poisson, 455
- scale mixture, 453
- scaled inverse- $\chi^2$ , 458
- standard Cauchy, 456, 457
- symmetric  $\alpha$ -stable, 457
- uniform, 467
- Wishart, 470
- probability generating functional (PGFL), 486
- probability hypothesis density, 486, 771
  - particle, 772
- probability hypothesis density filter, 772
  - cardinalized, 777
  - Gaussian mixture, 773
  - Gaussian mixture-cardinalized, 779
- probability mass function, 450
  - joint, 460
- probability of error, 533
- probability of exceedance, 306
- probability space, 447
- probability theory
  - axiomatic, 446
  - frequentist, 445
- probability weighted moments, 493
- probability-generating function
  - discrete random variable, 450
  - random vectors, 460
- problem
  - hypothesis testing, 525
  - optimal control, 136
- parameter estimation, 491
- process equation, 478
  - multi-target, 759
- process noise, 478
- process noise covariance
  - discretized, 480
- prograde precession, 417
- projectile, 442
- prolate body, 417
- proof mass, 651, 717
- propeller, 329
- proper Euler angles, 217
- proper subset, 797
- proportional control stage, 95
- proportional navigation guidance, 323
- proportional-integral-derivative control, 105
- proposal distribution, 467
- proposal functions, 521
- proposal PDF, 521
  - PF, 614
- proprioceptive sensor, 482
- protection level, 794
- pseudo-random number generation, 467
- pseudobearings, 748
- pseudoinverse, 175, 203, 500, 801
  - left, 500
- pseudomeasurements, 748
- pseudometric
  - statistical, 466
- pseudorange, 682
  - differential, 683
  - measurement, 692
- pseudorange equation
  - base, 696
  - GNSS, 706
  - rover, 696
  - simplified, 693
  - un-differenced, 693
- pseudorange measurement
  - GPS, 670
- pseudorange rate, 734, 748
- pseudorange-rate, 699
  - measurement, 699

- pseudorange-rate equation, 699
- pseudoranges, 748
- pulse compression, 662
- pulse repetition frequency, 662
- pursuit-evasion guidance problem, 320
- Q-factor, 101
- QR decomposition, 807
- quadratic programming, 507
- quadrature, 520
- quadrature Kalman filter, 609
- quadrature Rauch-Tung-Striebel smoother, 609
- quantile function, 451
- quartz clocks, 675
- quasimetric
  - statistical, 466
- quaternion
  - passive rotation, 220
- radar, 682
  - bi-static, 660
  - duplexer, 662
  - frequency modulated continuous wave, 662
  - illumination, 660
  - mono-static, 660
  - multi-static, 660
  - ultra-wideband, 661
- radar altimeter, 713
- radar cross section, 660
- radar Doppler shift, 661
- radar flow, 734
- radar range equation, 660
- radially unbounded, 193
- radio clock, 676
- radio technical commission for maritime services, 708
- radius of curvature
  - east-west, 737, 743
  - instantaneous, 338
  - meridian, 266, 737
  - north-south, 266, 737, 743
  - prime vertical, 737
- ram pressure, 674
- random dynamical system, 477
- random finite set, 484
  - Bernoulli, 488
  - birth, 759
  - clutter, 759
  - detection, 759
  - IID cluster, 487
  - intensity function, 486
  - labeled, 485
  - multi-Bernoulli, 488
  - multi-Bernoulli mixture, 488
  - Poisson, 487
  - PPP, 487
  - spawn, 759
  - survival, 759
- random finite sets
  - BMF, 485
- random gusts, 305
- random number generation, 467
- random process, 471
  - auto-regressive, 476
  - auto-regressive moving-average, 477
  - Brownian motion, 476
  - colored noise, 473
  - counting, 475
  - expected average power, 473
  - moving-average, 477
  - Poisson, 475
  - standard Wiener, 476
  - strictly white noise, 473
  - white noise, 473
  - white noise Gaussian process, 476
  - Wiener, 476
- random sequence, 471
- random variable, 449
  - central moment, 454
  - continuous, 449
  - discrete, 449
  - kurotsis, 454
  - mean, 454
  - median, 454
  - mixed, 453
  - mode, 454
  - moment, 453

- skewness, 454
- standard deviation, 454
- standardized moment, 454
- variance, 454
- random variate, 467
- random variate generation, 467
- random vector, 459
- random walk
  - discrete-time, 479
- range, 682
  - differential, 682
  - function, 798
  - ground, 686
  - slant, 686
- range rate, 682
- range rate measurement
  - GPS, 669
- range-domain integration, 741
- rank
  - column, 801
  - deficient, 801
  - full, 801
  - matrix, 801
  - row, 801
- rank-nullity theorem, 802
- Rao-Blackwell theorem, 623
- Rao-Blackwellized particle filter, 624
- rapid orbital errors, 711
- rate feedback control, 108
- rate gyroscope, 652
- rational numbers, 797
- Rauch-Tung-Striebel smoother, 548
  - Fourier-Hermite, 599
  - quadrature, 609
  - sigma-point, 602
- reachability
  - state, 123
- reaction wheels, 311
- real matrix, 800
- real numbers, 797
- real parametric uncertainty, 126
- real vector, 799
- real-time networks, 708
- realization, 449
  - random process, 471
  - vector, 459
- receding horizon control, 187
- receding-horizon control, 137
- receiver clock drift, 699
- receiver-operating characteristic, 530
- rectangular gates, 753
- recursive Bayes estimator, 538
- recursive least-squares, 515
- reduced-order model, 30
- reentry flight, 640
- reference area, 299
- reference ellipsoid, 248
- reference frame
  - Earth-centered, Earth-fixed, 247
  - east-north-up, 251
  - international celestial reference frame, 246
  - international terrestrial reference frame, 247
  - north-east-down, 251
  - rotation, 212
  - topocentric-horizon, 248
  - translation, 212
  - velocity-turn-climb, 253
- reference frames, 212
  - Earth-centered, 246
- reference model, 205
- reference transmitter, 694
- region of attraction, 193
- region of convergence, 66
- regression error, 496
- regression matrix, 496
- regression model, 498
- regression problem, 498
  - Bayesian linear, 515
  - nonlinear least-squares, 505
- regression sum of squares, 502
- regression weights, 600
- regressor, 498
- regulator, 140
  - controller, 132
- Reid's algorithm, 767
- rejection sampling, 467

- rejection sampling Monte Carlo, 521
- relative entropy, 466
- relative frequency
  - probability, 445
- relative orbital dynamics, 408
  - linearized, 409
- remote sensing, 482
- rendezvous and proximity operation, 407
- rendezvous guidance, 323
- rendezvous guidance problem, 320
- rendezvous maneuver
  - spacecraft, 402
- rendezvous problem, 322
- rendezvous, proximity operations, docking, and undocking, 403
- requirements
  - design, 8
- resampling
  - adaptive, 621
  - low-variance, 617
  - multinomial, 616
  - regularized, 619
  - remainder, 617
  - residual, 617
  - stratified, 617
  - systematic, 617
- residual, 498
- residual sum of squares, 501
- residue, 18
- resolvent, 19
- resonant
  - peak, 52
- resonant valley, 51
- resonator, 675
- restricted-step methods, 507
- retrograde precession, 417
- reweighted particle smoother, 626
- Reynolds number, 299
- RFS
  - cardinality distribution, 484
  - density, 484
  - probability generating functional (PGFL), 486
- RFS moment density, 486
  - $N^{\text{th}}$ -order, 486
  - first-order, 486
- Riccati matrix
  - continuous-time, 141
  - discrete-time, 180
- right ascension of ascending node, 405
- right hand circular polarized, 668
- rigid-body model, 9
- rise time, 32
- RMS gust intensities, 306
- robust stability, 129
- robust stability margin, 129
- robustness
  - SISO feedback control system, 126
- rocket-powered, 439
- roll angle, 255
- roll damper, 370
- roll mode, 369
- roll-spiral mode, 370
- rolling moment
  - wing, 819
- rolling moment coefficient
  - wing, 819
- root locus, 104, 385
- root mean square, 678
- rotation angle
  - Euler axis, 215
- rotation vector
  - passive, 215
- rotor mast, 397
- roughening, 619
- rover
  - receiver, 696, 702
- row rank, 801
- RSMC estimator, 522
- rudder, 329
  - ballistic vehicle, 439
- ruddervator, 334
- Runge-Kutta
  - fixed-step, 810
- running cost, 138
- runway visual reference, 341

- Sagnac effect, 653
- GNSS, 710
- sample, 471
- sample covariance, 513
- sample function, 471
- sample impoverishment, 619
- sample mean, 512
- sample path, 471
- sample point, 447
- sample space, 446
- sample trajectory, 471
- sampling
  - Markov chain Monte Carlo, 619
  - pseudo-random number, 467
  - random number, 467
- sampling frequency, 67
- sampling importance resampling, 616
- sampling interval, 67
- satellite constellation, 664
- satellites, 401, 438
- saturation bound, 589
- saturation function, 589
- scalar
  - notation, 799
- scale mixture random variable, 453
- scale parameter
  - scale mixture, 453
- scene flow, 734
- Schur Complement Lemma, 809
- score
  - likelihood, 492
- search vector, 505
- second-order extended Kalman filter, 591
- seismic mass, 651, 717
- semi-lactus rectum, 406
- semi-major axis, 405
- semi-Markov jump process, 632
- semidefinite programming, 153
- semidefinite programs, 150
- semimetric
  - statistical, 466
- sensitive altimeter, 713
- sensitivity
  - transfer function, 79
- sensitivity tradeoff, 111
- sensor, 75, 482
  - electro-optical, 663
  - exteroceptive, 482
  - infrared, 663
  - proprioceptive, 482
  - remote, 482
  - system, 319
- sensor data, 749
- sensor error
  - corrective, 482
  - predictive, 483
- sensor networks, 749
- separation principle, 122, 173
  - LQG, 574
  - stochastic control, 578
- sequential importance resampling, 616
- sequential importance sampling with resampling, 616
- sequential Kalman filter
  - discrete-time, 550
- serial interconnection, 72
- servomechanism state-space model, 133
- set
  - associative laws, 799
  - commutative laws, 798
  - complement, 797
  - countably infinite, 797
  - De Morgan's laws, 799
  - disjoint, 798
  - distributive laws, 799
  - empty, 797
  - finite, 797
  - function, 797
  - intersection, 798
  - mathematical, 797
  - partition, 798
  - union, 798
- set derivatives, 760
- set difference, 798
- set element, 797
- set integral, 484

- settling time, 33
- sextant, 682
- Shannons sampling theorem, 67
- short-period mode, 369
- shortest-path problem
  - Euclidean, 148
- shot noise, 661
- side force
  - fixed-wing aircraft, 335
- sideslip angle, 256
- sidewash angle, 823
  - equation, 835
- sight reduction, 682
- sigma plot, 56
- sigma-point Kalman filter
  - error-state, 613
- sigma-point Kalman filters, 602
- sigma-point particle filter, 622
- sigma-point Rauch-Tung-Striebel smoothers, 602
- sigma-points, 521, 599
- signal, 13, 535
  - analog, 13
  - digital, 13
  - discrete-time, 13
  - periodic, 40
- signal detection theory, 526
- signal-to-noise ratio, 662
- signals of opportunity, 682
- significance level, 525
- simple hypothesis test, 526
- simulation
  - system, 8
- simultaneous localization and mapping, 715
- single-degree LTI feedback control system, 164
- single-difference pseudorange
  - between-receiver, 696
  - between-transmitter, 694
- single-differenced phase-based range equation, 702
- single-spin stabilized, 419
- singular value, 730, 806
  - maximum, 807
  - minimum, 807
- singular value decomposition, 806
- singular value plot, 56
- SIR step
  - BPF, 618
  - EKPF, 622
  - RBPF, 624
  - SPPF, 622
- SISO feedback control system
  - characteristic equation, 81
  - characteristic polynomial, 81
  - stability, 81
- skew-symmetric matrix, 801
- skewness, 454
- skip
  - flight phase, 440
- Sklar's theorem, 467
- slant range, 686
- slat, 334
- small angle approximation, 25
- small gain theorem, 129
- small-disturbance theory, 23
- small-perturbation theory, 23
- Smirnov transform, 467
- smoothing step
  - Bayes smoother, 540
  - extended Rauch-Tung-Striebel smoother, 594
  - fixed-interval extended Kalman smoother, 594
  - fixed-interval Kalman smoother, 549
  - fixed-interval sigma-point Kalman smoother, 604
  - fixed-interval statistically linearized Kalman smoother, 598
  - fixed-lag Kalman smoother, 556
  - Rauch-Tung-Striebel smoother, 549
  - sigma-point Rauch-Tung-Striebel smoother, 604
  - statistically linearized Rauch-Tung-Striebel smoother, 598
- soft-iron distortion, 659
- software system
  - validation & verification, 8
- software system analysis, 8

- software system design, 8
- software system synthesis, 8
- software-defined radio, 669
- software-in-the-loop, 8
- sojourn time, 475
- sojourn-time, 632
- solution separation, 789
- solution separation method, 794
- Somigliana gravity model, 261
- sonar, 682
- space cone, 415
- space flight
  - orbital, 401
  - sub-orbital, 401
- space segment, 665
- space vehicles, 665
  - GPS, 668
- spacecraft
  - chaser, 407
  - target, 407
- span
  - wing, 819
- span efficiency
  - wing, 821
- sparsity, 510
- spatial Poisson process, 762
- spawning
  - object, 748
- specific force, 654
- spectral radius, 131
- speed hold, 339
- speed of sound, 820
- spherical coordinates, 247
- spherical-Earth model, 273
- spin angle, 255
- spin stabilization, 426
- spiral mode, 369
- split measurements, 749
- spoiler, 334
- spoofing, 667
- square matrix, 800
- square-root correction step
  - Kalman filter, 552
- square-root Kalman filter, 551
- stabilizer, 329
- stability
  - asymptotic, 20
  - global, 20
  - global asymptotic, 20
  - lateral-directional, 370
  - longitudinal, 369
  - LTI MIMO continuous-time, 55
  - LTI system, 29, 69
  - Lyapunov, 20
  - modal continuous-time, 28
  - modal discrete-time, 69
  - SISO feedback control system, 81
- stability and control coefficients
  - longitudinal, 826
- stability and control derivative
  - elastic, 380
- stability and control derivatives, 363
  - lateral-directional, 834
  - longitudinal, 826
- stability augmentation system, 76, 108
- stability augmentation systems
  - airplane, 385
- stability frame, 257, 350
- stability index, 457
- stability margin
  - (time) delay, 84
  - disk, 85
  - gain, 81
  - phase, 83
  - SISO, 126
- stabilizability, 124
- stabilizer
  - horizontal, 329
  - vertical, 329
- stable matrix, 804
- stagnation pressure, 297, 674
- stall, 820
- standard acceleration due to gravity, 262
- standard curvilinear motion model, 634
- standard deviation, 454
- standard gravitational acceleration, 641

- standard gravitational parameter, 260
- standard normal distribution, 455
- standard score, 527
- standard time, 677
- standard Wiener process, 476
- standardized moment
  - random variable, 454
- star map, 731
- star tracker, 731
- state
  - base, 478
  - Markov process, 474
  - modal, 478
  - vector, 15, 64
- state controllability, 123
- state controllability matrix
  - discrete-time, 68, 174
- state equation, 15, 478
  - discrete-time, 64
- state error
  - continuous-time, 122, 173
- state estimate, 121, 172
- state estimation, 535
  - vehicle (self), 319
- state extraction step
  - BPF, 618
  - EKPF, 623
  - RBPF, 625
  - SPPF, 623
- state feedback control, 120, 172
- state filter, 535
- state filtering, 535
- state filtering PDF
  - posterior, 539, 540
  - prior, 538, 540
- state filtering posterior PDF, 537, 539, 540
- state matrix
  - continuous-time, 15
  - discrete-time, 64
- state observability
  - discrete-time, 174
- state posterior PDF
  - marginal, 540
- state prediction PDF
  - posterior, 538
- state prediction posterior PDF, 537, 538
- state predictor, 535
- state prior PDF, 536
- state reachability, 123
- state smoother, 535
- state smoothing, 535
- state smoothing PDF
  - posterior, 540
- state smoothing posterior PDF, 537, 540
- state space
  - probability, 449
  - random process, 471
- state transition
  - Markov process, 474
- state transition matrix
  - discrete-state Markov process, 474
- state transition PDF, 536
- state transition probabilities, 474
- state transition probability
  - one-step, 474
- state vector
  - modal, 27, 68
- state-of-health, 786
- state-space
  - continuous-time, 14
  - continuous-time linear, 15
  - continuous-time LTI, 16
  - continuous-time, time-invariant, 16
  - discrete-time, 64
  - discrete-time linear, 64
  - discrete-time LTI, 64
  - discrete-time, time-invariant, 64
- state-space model
  - stochastic, 477
  - stochastic continuous-time, 477
  - stochastic discrete-time, 477
  - stochastic discrete-time with
    - measurement-origin uncertainty, 757
  - stochastic hybrid-time, 477
- state-transition matrix
  - discrete-time, 68

- LTI, 22
- LTV, 194
- static, 696
- static port
  - blocked, 675
- static pressure, 674
  - atmospheric, 297
- static-controller feedback, 171
- static-controller feedback control, 120
- stationarity
  - covariance, 473
  - strict-sense, 473
  - wide-sense, 473, 474
- stationary distribution
  - Markov chain, 475
- statistic, 453
  - KS, 452
- statistical comparison, 525
- statistical inference, 491
- statistical linear regression, 599
- statistical linearization, 595
- statistical regularity, 445
- statistically linearized (Kalman) filter, 596
- statistically linearized Kalman smoother
  - fixed-interval, 598
- statistically linearized Rauch-Tung-Striebel smoother, 598
- steady wind, 304
- steady-flight condition
  - point-mass, 337
- steady-flight conditions
  - rigid-body, 349
- steady-flight equations
  - lateral-directional straight, 352
  - longitudinal straight, 352
  - point-mass alternative, 338
  - rigid-body, 349
- steady-state
  - condition, 32
  - gain, 32
  - input, 32
  - output, 32
- steady-state Kalman filter, 565
- steady-state sinusoidal response, 40
- steepest gradient, 506
- steering angle, 634
- step amplitude, 32
- step response, 32
  - unit, 32
- stereo matching, 663
- stereoscopic camera, 663
- stereoscopic vision, 663
- Stirling's formula
  - second-order , 606
- stochastic decoupling, 521
- stochastic dynamical system, 477
- stochastic error
  - accelerometer, 656
  - gyroscope, 656
  - IMU, 656
- stochastic motion model
  - general polynomial, 629
- stochastic OCP
  - fundamental, 574
- stochastic process, 471
- stochastic state-space
  - fault model, 788
- stochastic state-space model, 477
  - continuous-time, 477
  - discrete-time, 477
  - discrete-time with measurement-origin uncertainty, 757
  - discretized LTI, 481
  - hybrid-time, 477
  - linear, 478
- stochastic universal sampler, 617
- straight flight, 338
- straight-line approximation, 43
- strapdown gyroscope, 652
- strapdown sensor, 734
- strict-sense stationarity, 473
- structured singular value, 130
- structural-mode control, 390
  - active, 390
  - passive, 390
- sub-orbital space flight, 401

- subset, 797
- proper, 797
- sum of squares
  - explained, 502
  - regression, 502
  - residual, 501
  - total, 501
- superpositional sensors, 749
- support
  - probability, 448
- supremum, 804
- surjective
  - matrix, 801
- surveillance system, 748
- survival probability, 758
- swashplate, 397
- sweep angle, 821
- symmetric congruence transformation, 163
- symmetric matrix, 801
- symmetric wing, 820
- synthetic aperture radar, 663
- system
  - autonomous, 14
  - deterministic, 14
  - dynamical, 13
  - general, 13
  - linear, 13
  - linear, parameter-varying, 198
  - mimo, 14
  - miso, 14
  - mode, 27
  - nonlinear, 13
  - response, 13
  - simo, 14
  - siso, 14
  - static, 13
  - stochastic, 14
  - time-invariant, 13
  - time-varying, 14
- system faults, 674
- system identification
  - aircraft, 262
  - airplane, 363
- system response
  - free, 31
  - frequency, 41
  - impulse, 34
  - initial condition, 31
  - pole effects, 37
  - step, 32
  - unit step, 32
  - zero effects, 38
  - zero-input, 31
  - zero-state, 32
- system type, 132
- T-tail, 330
- tactical grade IMU, 655
- tail
  - A-, 333
  - conventional, 329
  - cruciform, 331
  - H-, 331
  - horizontal, 329
  - inverted V-, 332
  - T-, 330
  - twin, 331
  - V-, 331
  - vertical, 329
  - X-, 332
  - Y-, 333
- tail distribution, 451
- tail index, 457
- taileron, 329
- tailless, 329
- Tait-Bryan Euler angles, 217
- tandem wings, 328
- target function, 521
- target tracking
  - ballistic, 640
  - compatible tracks, 767
- target tracking system, 748
- targets, 748
- taxicab norm, 803
- Taylor series
  - multivariate, 24

- univariate, 22
- temperative lapse rate, 300
- terminal
  - flight phase, 335
- terminal cost, 138
- terminal phase, 442
- ternary-uniform mixture distribution, 632
- terrain map, 714
- terrain-aided navigation, 741
- terrain-based navigation, 741
- terrain-contour navigation, 741
- terrain-matching navigation, 741
- terrain-referenced navigation, 741
- terrain-relative navigation, 741
- test statistic, 525
- thermal noise, 661
- third-order spherical cubature rule (TOSCR), 607
- three-axis magnetometer, 659
- three-body problem, 407
- threshold
  - detection, 526
- thrust, 262, 271
- thrust coefficient, 833
- thrust vectoring, 262, 271
- thruster
  - pair, 316
  - single, 315
- thruster-based active nutation control, 426
- tight integration, 741
- tilt compensation, 659, 718
- time
  - civil, 677
  - coordinated universal, 677
  - daylight saving, 677
  - posix, 677
  - standard, 677
  - unix, 677
- time constant
  - maneuver, 631
  - modal, 29
- time difference of arrival, 683
- time horizon
  - continuous-time, 137
- discrete-time, 176
- time of arrival, 679, 682
- time of birth, 782
- time of flight, 682
- time of transmission, 679
- time sample, 471
- time shift operator, 66
- time step, 64
- time synchronization
  - GNSS/INS, 746
- time update step
  - error-state, 610
- time-homogeneous, 474
- time-invariant
  - continuous-time state-space, 16
  - discrete-time state-space, 64
  - ODE, 15
- time-of-flight, 661
- time-of-flight equation, 643
- time-varying dynamical system, 190
- timebase, 679
- timepieces, 675
- torque-free motion, 414
- torquer, 314
- total electron count, 711
- total pressure, 297, 674
- total sampling time, 67
- total sum of squares, 501
- track score, 770
- track-oriented multi-hypothesis tracking, 768
- track-to-measurement association, 751
- track-to-sensor data association, 751
- tracking loop aiding
  - GNSS, 741
- tracking-domain integration, 742
- transfer function, 17
  - discrete-time, 66
  - fundamental, 79
  - matrix, 19
  - pole, 18
  - standard form, 18
  - zero, 18
- transfer function matrix

- standard, 19
- transient response, 40
- transition probabilities, 475
- transition rate
  - Markov jump process, 475
- transmitter-receiver-time triple-difference, 706
- transport rate, 265, 738
- transpose, 800
  - matrix, 800
- trial step vector, 507
- triangle inequality, 803
- tricopter, 398
- trilateration, 688
- trim point, 20
- trim state, 20
- triple-difference
  - cycle count range equation, 706
- tropospheric delay
  - code phase, 707
- true anomaly, 405
- true north, 659
- true north pole, 246
- true-state vector, 610
- trust-region, 507
- trust-region methods, 507
- Tschauner-Hempel equations, 411
- turn axis, 253
- turn compensation, 390
- turn-rate, 634
- turning motion, 633
- twin tail, 331
- two-body approximation, 404
- type I error, 526
- type II error, 526
- ultra-rapid orbital errors, 711
- unbiased, 494
- uncertainty
  - measurement origin, 748
  - measurement-origin problem, 749
  - object motion, 749
- uncertainty set, 126
- uncorrelation, 465
- random process, 474
- uniform asymptotic stability, 192
- uniform distribution, 467
- uniform stability, 192
- uniformly most powerful test, 530
- union
  - set, 798
- unitary matrix, 800
- universal time, 676
- unix time, 677
- unmatched uncertainties, 201
- unscented transform, 605
  - higher-order, 605
- unsteady winds, 305
- unstructured uncertainty set, 128
- upper triangular matrix, 801
- user equivalent range error, 707
- user segment, 665
- v-n plot, 359
- V-tail, 331
- validation and verification
  - software system, 8
- value function, 177
  - continuous-time, 140
  - forward, 184
- variable-structure multi-model, 569
- variance, 454
  - $M$ -sample, 677
  - 2-sample, 677
  - Allan, 677
  - estimator, 494
- variance-covariance, 465
- variation
  - cost functional, 138
- variational Bayes, 519
- VB parameter estimator, 520
- vector
  - column, 800
  - element, 799
  - norm, 802
  - notation, 799
  - row, 800

- vector auto-regressive process, 479
- vector space, 799
  - dimension, 801
- vehicle, 10
  - aerospace, 10
  - ballistic, 442
  - flight, 10
- vehicle dynamics and control, 11
- vehicle dynamics model, 733
- vehicle-centered, inertial frame, 251
- velocimeter, 734
- velocity
  - inertial point-mass, 222
- velocity axis, 253
- velocity-based linearization, 196
- velocity-turn-climb frame, 253
- vernal point, 247
- vertical channel instability, 740
- vertical tail
  - efficiency, 835
  - volume ratio, 835
- virtual control input
  - LTI, 202
  - nonlinear, 204
- virtual reference station, 708
- virtual work, 243
- vision transformer, 664
- vision-aided inertial navigation system, 741
- vision-based navigation, 714
- vision-inertial navigation system, 741
- visual camera, 663
- visual-based navigation, 714
- visual-inertial navigation system, 741
- visual-inertial odometry, 741
- volume ratio
  - horizontal tail, 827
  - vertical tail, 835
- von Kármán gust model, 305
- Wahba's problem, 730
- waterbed effect, 111
- waypoint, 319
- weak white noise, 473
- weight, 260
- weight matrix, 140, 803
  - cross-, 140
  - endpoint, 140
  - input, 140
  - state, 140
  - terminal, 140
- weighted least-squares problem, 499
- wet delay, 712
- WGS parameters, 248, 668
- wheel base, 634
- wheel-based active nutation control, 428
- wheel-based passive nutation control, 427
- wide-lane combination, 707
- wide-sense stationarity, 473, 474
- Wiener filter, 565
- Wiener process, 476
- Wilks test, 530
- wind frame, 252
- wind shear, 304
- wind speed, 253
- wind triangle, 253
- wind-up
  - phase, 707
- wing span, 819
- wings, 328
- wings-level flight, 354
- wobble angle, 417
- X-tail, 332
- Y-tail, 333
- yaw angle, 255
- yaw damper, 370
- yaw-pitch-roll Euler angles, 217
- Z-test, 527
- zero
  - left half plane, 38
  - right half plane, 38
- zero dynamics
  - LTI, 202
  - nonlinear, 205
- zero matrix, 16

- zero-effort-miss, 323  
zero-input response, 31  
zero-lift angle of attack  
wing, 820
- zero-lift turn, 441  
zero-state response, 32  
Ziegler-Nichols PID tuning method, 108