Control Theory and Matrix Analysis

Adam Li,
Department of Applied Mathematics & Statistics,
Department of Biomedical Engineering,
Johns Hopkins University,
Baltimore, MD, 21218
ali39@jhu.edu / adam2392@gmail.com

April 25, 2021

Contents

1 Introduction

Control systems are important in many applied problems. The problem of robust control is determining a set of performance guarantees under parameter uncertainty of a system.

We begin by defining basic aspects of a control system. Namely, what is a control system made up of.

Definition 1.1. A control system is

Stability and performance are two fundamental problems in any design, analysis and implementation of a control system.

Definition 1.2. Stability means that in the absence of external excitation, all signals in the system decay to zero.

If a system's stability is preserved in the face of uncertainties, then that system has **robust stability**

Definition 1.3.

2 Matrix Analysis

3 Important Facts

3.1 Finite-Dimensional Vector Spaces

We begin by defining **finite-dimensional** vector spaces, which are fundamental for matrix analysis. A vector space, V over a field, F, is a set of objects called vectors that have the following properties:

i) closed under binary addition ii) associative: x iii) commutative iv) has the zero vector

A field is typically the real numbers, or complex numbers, and has the following properties:

1. Closure under addition: $\forall x, y \in F$, we have that $x + y \in F$
2. Closure under multiplication: $\forall x, y \in F$, we have that $xy \in F$

3.1.1 Subspaces

Sylvester inequality: $A \in M_{m,k}$

3.1.2 Sherman-Morrison-Woodbury formula

This formula deals with a nonsingular matrix $A \in M_n$, with a known inverse A^{-1}. Then any matrix:

$$B = A + XRY$$

with X (n-by-r matrix), R (r-by-r) nonsingular and Y (r-by-n), then if B, $R^{-1} + YA^{-1}X$ are both nonsingular, then we have a closed formula for the inverse of B:

$$B^{-1} = A^{-1} - A^{-1}X(R^{-1} + YA^{-1}X)^{-1}YA^{-1}$$

where if r « n, then the cost of computing R^{-1} is **significantly cheaper** then trying to compute the inverse of B, which is a n-by-n matrix.

Note on rank-1 perturbations If X = x, and Y = y^T, and R = [1], then we have that:

$$B = A + xy^T$$

is a rank-one perturbation of A. Now, if we want the inverse of B, we get that:

$$B^{-1} = A^{-1} - (1 + y^TA^{-1}x)^{-1})^{-1}A^{-1}xy^TA^{-1}$$

3.1.3 Schur Complements

4 Canonical Forms: Jordan and Weyr

When we are faced with two arbitrary matrices, A, B both n-by-n matrices, we would like to understand when and why they would be similar. We know that similarity equivalence condition says two matrices can be written as:

$$A = SBS^{-1}$$

We know that they share the same eigenvalues and also the same rank, but there are matrices that have the same eigenvalues and rank, but are **similar**. Take for example, a nilpotent matrix and construct a matrix that has also 0 eigenvalues, but is not nilpotent.

One then asks what are necessary and sufficient conditions for two matrices to be similar? Since we know that similarity is a transitive property, then perhaps we would seek a similar matrix to both A and B, such that if A and B can be broken down into this matrix, then A and B must be similar. If we take eigenvalue decomposition, then we are not always guaranteed existence of an eigenvalue decomposition, so this does not hold for **all square matrices**. If we take Schur's upper-triangular decomposition, we have a uniqueness problem. First off, upper-triangularization may have different off-diagonals, so upper-triangular matrices are too **large of a set** for our desired "canonical form".

4.1 Jordan Form

The first canonical form is a compromise between diagonal and upper-triangular matrices. They are block-jordan matrices. Note that **every square matrix** has a unique Jordan form up to permutations of its Jordan blocks. We list some of its important properties.

1. Linearly independent eigenvectors: number of Jordan blocks.
2. Geometric multiplicity of eigenvalues: number of Jordan blocks associated with that eigenvalue.
3. Algebraic multiplicity of eigenvalues: total size of all associated Jordan blocks with that eigenvalue.

4.1.1 Real Jordan Canonical Form

When we know that the matrix A is a real matrix, then we can make use of that fact to come up with another canonical form. Note that real matrices have complex eigenvalues $\lambda = a + bi$ come in complex conjugate paires, so $\lambda_2 = a - bi$ is also an eigenvalue of A.

5 Eigenvalues, Eigenvectors and Similarity

We begin with the notion of eigenvalues and eigenvectors.

Here we summarize some related results on rank-one perturbation of eigenvalues. The first theorem by Brauer states that we can move any eigenvalue arbitrarily without moving other eigenvalues by taking linear combinations of its corresponding eigenvector.

Theorem. *Suppose $A \in M_n$ is a n by n matrix that has eigenvalues, $\lambda, \lambda_2, ..., \lambda_n$ and let x be a non-zero vector such that $Ax = \lambda x$ (i.e x is an eigenvector of λ). Then for any $v \in \mathbb{C}^n$, the eigenvalues of $A + xv^*$ are $\lambda + v^*x, \lambda_2, ..., \lambda_n$.*

Proof. We take the non-normalized eigenvector x and WLOG, say $\eta = \frac{x}{||x||}$, so that η has unit norm. Next, we construct a unitary matrix U with η as its first column and orthogonal vectors in the next n-1 columns chosen via Gram-Schmidt orthogonalization. Then each vector is normalized giving us an orthonormal column matrix, U, which is unitary.

Next, we take a Schur decomposition of $A = UTU^*$, namely with U as just constructed, so that λ is on the top diagonal entry of T. The other eigenvalues $\lambda_2, ..., \lambda_n$ are along the rest of the diagonal of T.

Next, we consider U^*xv^*U, which equals:

$$\begin{pmatrix} \eta^*x \\ u_2^*x \\ \vdots \\ u_n^*x \end{pmatrix} v^*U$$

whcih then can be shown has its first row with v^*x as the leading entry and then zeros on the next n-1 rows. Thus when we take $U^*(A + xv^*)U$, then we get that the first entry of T has its eigenvalue $\lambda + v^*x$. □

Thus note, we can take any matrix and the corresponding eigenvector of an eigenvalue we are interested in and perturb it to any value, r. One just needs to select the vector $v \in F$ accordingly.

Next, we recognize that Gershgorin disc theorems are a general framework for obtaining an initial bound on eigenvalue locations purely from the column and row sums of any matrix. We state the main theorem:

Theorem. *Let $A \in M_n$ and let $R_i(A) = \sum_{j \neq i} |a_{ij}|$ for i = 1, ..., n. $R_i(A)$ is the ith row sum of A.*

*We denote the **deleted absolute row sums** of A and consider the following n Gershgorin discs:*

$$\{z \in \mathcal{C}^n : |z - a_{ii}| \leq R_i(A)\}$$

Then the n eigenvalues of A are located in the union of the n Gershgorin discs:

$$G(A) = \bigcup_{i=1}^{n} \{z \in \mathcal{C}^n : |z - a_{ii}| \leq R_i(A)\}$$

Furthermore if the union of k discs are disjoint from the remaining n - k discs, then the k discs have exactly k eigenvalues of A.

Note that the second assertion does not require the set $G_k(A)$ to be connected. If we did for some reason know that $G_k(A)$ was disconnected, then the theorem could be applied for each disjoint Gershgorin disc until we get the largest set of disjoint discs, which must each contain their respective number of eigenvalues.

6 Random Matrix Analysis

7 Important Facts

7.1 Finite-Dimensional Vector Spaces

We begin by defining **finite-dimensional** vector spaces, which are fundamental for matrix analysis. A vector space, V over a field, F, is a set of objects called vectors that have the following properties:

i) closed under binary addition ii) associative: x iii) commutative iv) has the zero vector

A field is typically the real numbers, or complex numbers, and has the following properties:

1. Closure under addition: $\forall x, y \in F$, we have that $x + y \in F$
2. Closure under multiplication: $\forall x, y \in F$, we have that $xy \in F$

8 Random Matrix Theory

Random matrix theory is the study of matrices that arise with some probability distribution. This has interesting intersections with concentration inequality theory and obtaining bounds on eigenvalues for statistical estimates.

9 Resolvent Operator

In its most general form, we can start off by defining the Cauchy-Stieljes transform for all $z \in \mathbb{C}_+ = \{z \in \mathbb{C} : \eta(z) > 0\}$:

$$g_\mu(z) = \int \frac{1}{\lambda - z} d\mu(\lambda)$$

where μ is a finite measure on $\mathbb{R}$.

If we consider a matrix, $A \in H_n(\mathbb{C})$ and $\psi \in \mathbb{C}^n$ is a vectro with unit l2-norm. The spectral theorem guarantees the existence of $(v_1, ..., v_n)$ set of orthonormal basis vectors that are **eigenvectors** of A, such that:

$$Av_i = \lambda_i v_i$$

We say that the **spectral measure with vector** ψ is the real probability measure defined by:

$$\mu_A^\psi = \sum_{k=1}^{n} |\langle v_k, \psi \rangle|^2 \delta_{\lambda_k}(A)$$

If $z \notin \sigma(A)$, then A - zI is invertible. We define the resolvent of A as the function: $R : \mathbb{C}_+ \mapsto M_n(\mathbb{C})$

10 System Identification

10.1 Proper Orthogonal Decomposition (or Principal Component Analysis)

The motivation of POD, or PCA is to obtain an **approximate** and low-dimensional description of a function. If we wish to approximate a function $z(x,t)$ as a finite sum of the variables:

$$z(x,t) = \sum_{i=1}^{M} a_k(t)\phi_k(x)$$

then we would hope that the approximation is consistent (i.e. approaches z as M goes to infinity, except possibly on sets of measure zero). We by convention call x the "space" variable and t the "time" variable. So $x \in X \subset \mathbb{R}^d$ and $t \in T \subset \mathbb{R}$.

The functions a_k, ϕ_k may take many forms or basis functions. For example, they might be the Fourier basis, Legendre polynomial basis, simple function basis, or step function basis. POD/PCA attempts to consider an **orthonormal** basis. Namely the functions have the following property:

$$\int_X \phi_{k_i}(x)\phi_{k_j}(x)dx = \begin{cases} 1 \text{ if } k_1 = k_2 \\ 0 \text{ else} \end{cases}$$

where the integrands are replaced with sums in the discrete case. We consider the finite-dimensional case for now, reserving the infinite-dimensional case for the Koopman theory.

We now consider taking n measurements over time of a system with m state variables. This can be arranged into a $n \times m$ data matrix A. We **consider only real-valued matrices**.

10.1.1 Singular Value Decomposition

From this data matrix, A, we assume it is in general some rectangular matrix (i.e. $n \neq m$). We consider the singular value decomposition (SVD) that always exists for any matrix:

$$A = U\Sigma V^*$$

where U is a $n \times n$ orthonormal left singular vector matrix and V is a $m \times m$ orthonormal right singular vector matrix, and Σ is an $n \times m$ matrix that is almost diagonal. The non-zero diagonal elements, σ_{ii} are the singular values of the matrix.

Connection with original approximation problem

Low-rank approximation - based on singular values We know that the number of non-zero singular values, σ_{ii} is the rank of the matrix, A. However, due to sampling noise, one can imagine many $0 < \sigma_{ii} < \epsilon$, for some $\epsilon > 0$. One can simply "truncate" the singular values at some threshold, and get a lower-rank approximation of the A matrix.

This useful for interpretation, noisy datasets, dimensionality reduction.

A note on the "method of snapshots" Consider the case where m » n. That is there are significantly more state variables, then there are samples. Computationally, one would seek a lower-dimensional representation without having to compute the full SVD. This can be done by first considering the matrix AA^T, which is now a n x n matrix. Now by left-multiplying A by U^T (and the orthonormality of the U matrix), we get the following:

$$U^T A = \Sigma V^T$$

which is simplified since Σ has the last m - n columns being zero (remember m > n), which means that the last m - n columns of V are irrelevant, and simply required to be orthogonal to the first n columns of V. Creating the next set of orthogonal vectors can be done via say Gram Schmidt algorithm. Of course, if the rank of A is less than n, then there would be even more columns that are "irrelevant".

This method of snapshots presents a computationally efficient way of computing the V right-singular vector matrix, without directly performing a SVD.

10.2 Dynamic Mode Decomposition

The standard DMD algorithm is formulated using the SVD. It concerns a linear dynamical system as follows:

$$z_{k+1} = Az_k$$

for some **unknown** matrix A. Alternatively, one can assume a discrete system:

$$z((k+1)\delta t) = Az(k\delta t)$$

where the sampling rate is δt fixed. Even when DMD is applied to a nonlinear system, it is **assumed that there exists a linear operator A** that approximates the nonlinear dynamics according to some norm. We assume $z \in \mathbb{R}^d$.

The standard algorithm

1. First, one arranges m samples into a data matrix and a one-time shifted version into another data matrix:

$$X = \begin{pmatrix} z_0 \cdots z_{m-1} \end{pmatrix}$$
$$Y = \begin{pmatrix} z_1 \cdots z_m \end{pmatrix}$$

So $X, Y \in M_{d,m}$.

2. Next one computes the SVD of X (you can also compute a "low-rank" approximation by truncating singular values).

$$X = U\Sigma V^*$$

If we truncate to obtain r non-zero singular values, then U is a $n \times r$ matrix, Σ is an r x r matrix and V is a m x r matrix.

3. Next, define an intermediate matrix by projecting the time-shifted data matrix onto the singular vector space of the X matrix.

$$\tilde{A} = U^* Y V \Sigma^{-1}$$

It's eigenvalue and eigenvectors are:

$$\tilde{A}w = \lambda w$$

4. Compute the DMD mode that corresponds to the eigenvalue λ

$$\hat{\psi} = Uw$$

We project the eigenvector of $\tilde{A}$ onto the column space of the left-singular vector matrix of X.

Exact DMD Exact DMD proposes a change on the computation of the DMD mode. It is instead defined as:

$$\psi = \frac{1}{\lambda} Y V \Sigma^{-1} w$$

The original definiton has the DMD modes as the column space of X, while this defintion has the DMD modes as the column space of Y. Rather we will see in this theorem that $\hat{\psi}$ in standard DMD are related to the ψ written in Exact DMD by an orthogonal projection matrix that projects onto the column space of X.

Theorem (DMD modes in standard DMD are projections onto X of the exact DMD modes)**.** *TBD*

10.2.1 Multi-Resolution DMD

10.2.2 Randomized DMD

10.3 Koopman Operator Theory - Orthogonal Basis Functions in Hilbert Spaces

We consider now a discrete-time system with the following dynamics:

$$x_{k+1} = f(x_k)$$

where $x \in M$, a finite-dimensional manifold. The Koopman operator $\mathcal{U}$ acts on scalar functions $g : M \to \mathbb{R}$, mapping the g function to a new function Ug given by:

$$Ug(x_k) = g(f(x_k)) = g(x_{k+1})$$

Note here that U operates linearly on functions g, even when f is a nonlinear function. U is a linear operator that operates on possibly infinite-dimensional space. We consider the eigendecomposition:

$$U\theta_j(x) = \lambda_j \theta_j(x)$$

for j = 0, 1, ... possibly infinite amount of eigenfunctions. We are interested in expressing functions of x using the eigenvectors and Koopman modes.

$$h(x) = \sum_{j=0}^{\infty} \theta_j(x)\hat{\psi}_j$$

where $\hat{psi}_j$ are Koopman modes that have growth/frequency depending on the eigenvalues λ_j

Note Koopman analagoy to the SVD-based DMD algorithm is only appropriate when X describes a sequential time-series, with linearly independent columns and distinct DMD eigenvalues.

10.4 The Kalman Filter - Optimal System Identification for Linear Systems with Gaussian Noise

We consider the system:

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + v(t) \\ y(t) &= Cx(t) + Du(t) + w(t) \end{aligned}$$

where $x \in \mathbb{R}^d$, $y \in \mathbb{R}^m$ and $v\tilde{N}(0, Q)$ iid normally distributed.

Why is noise normally distributed? This is just a sensible option, and also has deep connections to the L2 norm loss, which when solved is optimal for Gaussian systems.

10.5 Numerical Algorithms for State-Space Subspace Identification (N4SID)

11 Online Convex Optimization for Online System Identification

In the online convex optimization problem, we are interested in an algorithm that optimizes a cost function based on data seen so far. In addition, one must account for the fact that a decision as data is congested results in a regret, specifically a value of how poor the decision was in retrospect with the new data.

Definition 11.1. We define regret of algorithm, $\mathcal{A}$ after T iterations as:

$$regret_T(A) = \sup_{f_1,\ldots,f_T \subset \mathcal{F}} \{\sum_{t=1}^{T} f_t(x_t^A) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x)\}$$

Note that ideally, an algorithm performs with regret that is sublinear as a function of T implying that on average the algorithm performs as well as the best fixed strategy in hindsight.