BME695 Lecture Notes

April 2021

1 What is an algorithm?

- An algorithm is the list of instructions and rules that a computer needs to do to complete a task
- In essence, algorithms are simply a series of instructions that are followed, step by step, to do something useful or **solve a problem**.
  - consider a cake recipe. It is essentially an algorithm for making a cake!
- In computing, algorithms provide the computer with a successive guide to completing actions. They're comprised of a precise **list of instructions** that outline exactly how to complete a task.
- *Commonly used algorithms in your research?*

2 How do computer algorithms work?

- Computer algorithms work via input and output.
- They take the input and **apply each step** of the algorithm to that information to generate an output.
- You can easily visualise algorithms as a flowchart. The input leads to steps and questions that need handling in order. When each section of the flowchart is completed, the generated result is the output.

- *Talk about an algorithm you are familiar with?*
- Automation works by following set rules to complete tasks. Those rules form an algorithm. Think about *Automation Software*.

3 Noticeable Algorithm Types

- **Iterative algorithms**. Iteration uses looping statements such as for loop, while loop or do-while loop to repeat the same steps.
- *Can you give an example*?
- **Newton's method**. A root-finding algorithm, which produces successively better approximations to the roots (or zeroes) of a real-valued function.
  - In calculus, Newton's method is an iterative method for finding the roots of a differentiable function F, which are solutions to the equation $F(x) = 0$.
  - In optimization, Newton's method is applied to the derivative f' of a twice-differentiable function f to find the roots of the derivative (solutions to $f'(x) = 0$), also known as the stationary points of f.
  - These solutions may be minima, maxima, or saddle points.
  - $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$.
- evolutionary computing; local (or neighborhood) search.

- **Recursive algorithms**. A module (function) calls itself again and again until the *base condition or stopping condition* is satisfied.
- **divide-and-conquer algorithms**
  - recursively breaks down a problem into two or more sub-problems of the same or related type, until these become simple enough to be solved directly.
  - The solutions to the sub-problems are then combined to give a solution to the original problem.

- *Can you give an example*?

- **Binary Search**. Given an array A of n elements with values or records $A_0 \leq A_1 \leq A_2 \leq \ldots \leq A_{n-1}$, and target value T, the following subroutine uses binary search to find the index of T in A.
  - Set L to 0 and R to $n-1$.
  - If $L > R$, the search terminates as unsuccessful.
  - Set m (the position of the middle element) to the floor of $\frac{L+R}{2}$.
  - If $A_m < T$, set L to $m+1$ and go to step 2.
  - If $A_m > T$, set R to $m-1$ and go to step 2.
  - Now $A_m = T$, the search is done, return m.

- **Branch-and-bound for discrete and combinatorial optimization problems**.
  - The algorithm consists of a systematic enumeration of candidate solutions by means of state space search.
  - The set of candidate solutions is thought of as forming a *rooted tree* with the full set at the root.
  - The algorithm explores *branches* of this tree, which represent subsets of the solution set.
  - Before enumerating the candidate solutions of a branch, the branch is checked against upper and lower estimated *bounds* on the optimal solution, and is discarded if it cannot produce a better solution than the best one found so far by the algorithm.

4 Algorithm Correctness and Efficiency

- In theoretical computer science, **correctness of an algorithm** is asserted when it is said that the algorithm is correct with respect to a specification.

- Functional correctness refers to the input-output behavior of the algorithm (i.e., for each input it produces the expected output).

- *talk about search algorithm, optimization algorithms*

- In computer science, **efficiency of an algorithm** is a property of the algorithm which relates to the number of computational resources used by the algorithm.
- The efficiency can be measured based on the usage of different resources (time and space).
- Algorithmic efficiency can be thought of as analogous to engineering productivity for a repeating or continuous process.
- **best case, average case, worst case**.
- **big O notation**.
  - a mathematical notation that describes the limiting behavior of a function when the argument tends towards a particular value or infinity.
  - In computer science, big O notation is used to classify algorithms according to how their run time or space requirements grow as the input size grows.
- *look at the classic sorting algorithm comparison table!*

5 Iterative Algorithm Convergence and Convergence Speed

- **Convergent series**, the process of some functions and sequences approaching a limit under certain conditions.
- A series $\{a_n\}$ is **convergent** (or **converges**) if the sequence $(S_1, S_2, S_3, \dots)$ of its partial sums tends to a limit. That is, when adding one a_k after the other in the order given by the indices, one gets partial sums that become closer and closer to a given number. More precisely, a series converges, if there exists a number ℓ such that for every arbitrarily small positive number ε , there is a (sufficiently large) integer N such that for all $n \geq N$,

$$| S_n - \ell | < \varepsilon.$$

- **Convergence tests**. Ratio tests, for example. But not important in our context.

- In numerical analysis, the **rate of convergence** and **orde of convergence** of a convergent sequence are the quantities that represent how quickly the sequence approaches its limit. A sequence (x_n) that converges to $x*$ is said to have *order of convergence* $q \geq 1$ and *rate of convergence* μ if

$$\lim_{n \to \infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|^q} = \mu$$

- Let us take a look at **Netwon's method in optimization**.

6 Common Numerical Methods

- Numerical linear algebra – e.g., Solving systems of linear equations; Eigenvalue algorithms

- Interpolation and approximation – e.g., Polynomial interpolation, Spline interpolation, Trigonometric interpolation, Other interpolants, **approximation theory (for metamodeling, surrogate modeling)**

- Finding roots of nonlinear equations

- Optimization – e.g., Linear programming, Convex optimization, Nonlinear programming, Optimal control and infinite-dimensional optimization, **black-box optimization (for model calibration)**

- Numerical quadrature (integration)

- *Numerical methods for ordinary differential equations* (?)

- *Numerical methods for partial differential equations* (?)

- **Monte Carlo method (for sensitive analysis, uncertainty quantification, surrogate modeling)**

7 Very basic stuff on Differential Equations

- Newton's Second Law Motion, i.e., $F = ma$. Rewrite it. What about the acceleration a? $a = \frac{dv}{dt}$ or $a = \frac{d^2u}{dt^2}$.
- a few more examples

$$e.g., \alpha^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial x}; \frac{\partial^3 u}{\partial x^2 \partial t} = 1 + \frac{\partial u}{\partial y}$$

- the concept of **Order**
- ODE/PDE
- linear differential equations

$$a_n(t)y^{(n)}(t) + a_{n-1}(t)y^{n-1}(t) + \cdots + a_1(t)y'(t) + a_0(t) = g(t)$$

  - there are no products of the function, $y(t)$, and its derivatives
  - neither the function and its derivatives occur to any power other than the first power
  - neither the function and its derivatives are "inside" another function

- What is a solution to a differential equation?
  - Example: Show that $y(x) = x^{-\frac{3}{2}}$ s a solution to $4x^2y'' + 12xy' + 3y = 0$ for $x > 0$.
- initial condition(s) – a condition, or a set of conditions, on the solution
- initial value problem (IVP) – a differential equation along with an appropriate number of initial conditions
- interval of validity
- general solution
  - Example: $y(t) = \frac{3}{4} + \frac{c}{t^2}$ is the general solution to $2ty' + 4y = 3$.
- actual solution

  - Example: What is the actual solution to the following IVP? $2ty' + 4y = 3$, $y(1) = -4$.

- explicit/implicit solution
  - An **explicit solution** is any solution that is given in the form $y = y(t)$. In other words, the only place that y actually shows up is once on the left side and
  - Example: $y^2 = t^2 - 3$ is the actual implicit solution to $y' = \frac{t}{y}$, $y(2) = -1$.
  - An **implicit solution** is any solution that isn't in explicit form.
  - Example: Find an actual explicit solution to $y' = \frac{t}{y}$, $y(2) = -1$.
  - Note that it is possible to have either general implicit/explicit solutions and actual implicit/explicit solutions.

- To us, what is useful?
  - How to model physical situations, especially using first order differential equations
  - Characterize your model. Identify the right built-in routine or code your own.

8 First Order Differential Equations

The most general first order differential equation can be written as,

$$\frac{dy}{dx} = f(y, t)$$

- Note that there is no general formula for the solution to the above equation.

- Typically, people look at special cases and go further on these cases.
  - Linear Equations – linear first order differential equations, i.e., differential equations in the form $y' + p(t)y = g(t)$.
  - Separate Equations – $N(y)y' = M(x)$.
  - Exact Equations.

  - Bernoulli Differential Equations – $y' + p(t)y = y^n$.
  - Substitutions – $y' = F(\frac{y}{x})$ and $y' = G(ax + by)$.
- Modeling with first order differential equations. I will upload a doc.
- Additional Concepts
  - Equilibrium Solutions
  - Euler's Method

9 Euler's Method

- The vast majority of first order differential equations can't be solved.
- Let us start with a general first order IVP with respect to t

$$\frac{dy}{dt} = f(t, y), \; y(t_0) = y_0$$

  where $f(t, y)$ is a known function and the values in the initial condition are also known numbers.
- From the second theorem in the Intervals of Validity section we know that if f and f_y are continuous functions then there is a unique solution to the IVP in some interval surrounding $t = t_0$.
- In engineering, most times, everything is nice and continuous. Even if not, we can make a reasonable approximation on it.
- Let us do some derivations now
- Next take a look at a figure
- A bit of pseudo-code that you can use to write a routine for Euler's Method that uses a uniform step size, h.
  1. **define** $f(t, y)$
  2. **input** t_0 and y_0
  3. **input** step size, h and the number of steps, n

4. **for** j from 1 to n **do**
   (a) $m = f(t_0, y_0)$
   (b) $y_1 = y_0 + hm$
   (c) $t_1 = t_0 + h$
   (d) Print t_1 and y_1
   (e) $t_0 = t_1$
   (f) $y_0 = y_1$
5. **end**

10 More on Euler's Method

Review Chapter 5 of Burton and Faires! I have placed the e-book a doc.

- Let us look at Table 5.1 on Page 259 of BF
- Higher-order Taylor Methods
- Runge-Kutta Methods
  - Runge-Kutta Order Two (RK2)
  - pay attention to $T^{(2)}(t, y)$, while RK methods have the high-order local truncation error of the Taylor method while eliminating the need to compute and evaluate the derivaties of $f(t, y)$.
  - How to replace $T^{(2)}(t, y)$ is the key, leading to different methods
  - The difference-equation method, i.e., replacing $T^{(2)}(t, y)$ in Taylor's method of order 2 by $f(t+(h/2), y+(h/2)f(t, y))$ is a specific RK method known as the midpoint method.
    * Three commonly used methods: midpoint, modified midpoint, and Heun's method
    * Comparison between the three. Pay attention to Table 5.5 on Page 277 of BF.
- Runge-Kutta Order Four (RK4). Pay attention to Table 5.6 on Page 279 of BF.

11 Second Order Differential Equations

- Just like we did before, we will look at some special cases of second order differential equations that we can solve.
- However, we are going to have to be even more restrictive as to the kinds of differential equations that we'll look at. This will be required in order for us to actually be able to solve them.
- Homogenous, linear, second order differential equations,
  - $ay'' + by' + cy = 0$.
  - What is a famous example in physics? Spring-mass system
- some key concepts: real roots, complex roots,

12 Review of ODEs

- general definition, explicit ODEs; implicit ODEs
- further classification:
  - linear, i.e., F can be written as a linear combination of the derivatives of y
  - autonomous, i.e., a differential equation not depending on x
  - homogeneous, non-homogeneous (or inhomogenous)
  - nonlinear
  - high-order
- techniques to transform a higher-order equation into a system of first-order differential equations
- you can find the discussion on the transformation in most of dynamic systems theory books.
  - Ex. $ay'' + by' + cy = 0$.
  - covert initial conditions
- Extend the model to incorporate stochasticity: $y' = f(t, y) + \epsilon(t, y)$.

13 Systems of ODEs

A number of coupled differential equations from a system of equations. If $\mathbf{y}$ is a vector whose elements are functions: $\mathbf{y}(x) = [y_1(x), y_2(x), \ldots, y_m(x)]$, and $\mathbf{F}$ is a vector-valued function of $\mathbf{y}$ and its derivates, then

$$\mathbf{y}^{(n)} = \mathbf{F}(x, \mathbf{y}, \mathbf{y}', \mathbf{y}'', \ldots, \mathbf{y}^{(n-1)})$$

is an *explicit system of ordinary differential equations of order n and dimension m.*

- What about in column vector form?
- System of linear differential equations, $x' = Ax + g(t)$ where x is a vector. We say the system is homogeneous if $g(t) = 0$ and we say the system is nonhomogeneous if $g(t) \neq 0$.
- not necessarily linear.
- One of the most famous examples of a non-linear system of ODEs is the Navier-Stokes equations.
- The SIR model
- The analysis and computation can be very difficult.
- the implicit analog, $\mathbf{F}(x, \mathbf{y}, \mathbf{y}', \mathbf{y}'', \ldots, \mathbf{y}^{(n-1)}) = 0$
- convert one of implicit form to explicit form, if the Jacobian matrix $\frac{\partial \mathbf{F}(x, \mathbf{y}, \mathbf{y}')}{\partial \mathbf{y}}$ is non-singular
- RK method, review Ch 5.7 (Page 315) of BF
- convert one of implicit form to differential algebraic equations (DAEs); with a singular Jacobian

14 The SIR Model

- Compartment models simplify the mathematical modeling of infectious disease
- The population is assigned to compartments with labels, e.g., S (susceptible), I (infected), R (recovered).

15 Further Reading on ODEs/PDEs

- Chapter 11 of BF, Boundary-Value Problems of ODEs
- Chapter 12 of BF, Numerical Solutions to PDEs

16 Discrete-Time Markov Chains

- Difference equations/systems of difference equations.
  - Many problems in Probability give rise to *differential equations*
  - Difference equations related to differential equations as discrete mathematics relates to continuous mathematics
  - a standard approach to solving a nasty differential equation is to convert it to an approximate equivalent difference equation. actually the spirit of RK methods.
  - let us look at an n^{th} order linear difference equation
- The other way of modeling stochasticity, instantaneous random jumps at different (often uniformly distributed) time points.
- individualized state transition models, e.g., in disease modeling
- Let me now use the slides DTMC Modeling Lecture Notes 02042021.

17 Review of Warm-up assignment II

- Rate of convergence

  Definition. In numerical analysis, the order of convergence and the rate of convergence of a convergent sequence are quantities that represent how quickly the sequence approaches its limit. A sequence ($\{x_n\}$) that converges to x^* is said to have order of convergence $q \geq 1$ and rate of convergence μ if

$$\lim_{n \to \infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|^q} = \mu$$

- log-log plot, additional material

- Matlab coding, using functions from Matlab built-in functions. Can you find a built-in function for the Bisection method?
- Local truncation error. BF Book, p. 266 - 267, definition 5.11

  $$\omega_{i+1} = \omega_i + h\phi(t_i, \omega_i)$$

  has **local truncation error**

  $$\tau_{i+1}(h) = \frac{y_{i+1} - (y_i + h\phi(t_i, y_i))}{h} = \frac{y_{i+1} - y_i}{h} - \phi(t_i, y_i),$$

  for each $i = 0, 1, \ldots, N-1$.
- Euler's method has

  $$\tau_{i+1}(h) = \frac{h}{2} y''(\xi_i)$$

  for some ξ_i in (t_i, t_{i+1}). When $y''(t)$ is known to be bounded by a constant M on $[a, b]$, this implies

  $$|\tau_{i+1}(h)| \leq \frac{h}{2} M,$$

  so the local truncation error in Euler's method is O(h).
- Then what is the difference between high-order Taylor methods? Taylor method of n vs. Runge-Kutta method.

18 Review of Markov Transition Models

- Markovian assumption, memoryless property
- Markov chain

  Definition. A stochastic process $\{X_n\}$ is called a Markov chain if

  $$Pr\{X_{n+1} = j | X_0 = k_0, ..., X_{n-1} = k_{n-1}, X_n = i\} = Pr\{X_{n+1} = j | X_n = i\}$$

  for every $i, j, k_0, ..., k_{n-1}$ and for every n.
- state transition matrix and diagram. Use an small example, Gambler's Ruin Example.

- stationary distribution, $P\bar{\pi} = \bar{\pi}$.
- If $\bar{\pi}$ is the initial state distribution, then the chain has this distribution at every step. In other words, $\bar{\pi}$ is the equilibrium distribution.
- An absorbing state is a state that, once entered, cannot be left.
- An absorbing Markov chain is a Markov chain in which every state can reach an absorbing state.

19 Elements of Markov models

- The first element of Markov model are so-called **health states**, such as well, ill, dead, relief, no relief, headache, no headache
- People **transition** from one health state to another. For this reason, Markov models are sometimes called **transition models**.
- Each transition has a probability (**transition probability**)
- Transitions happen over a period of time (called a **cycle**)
- Each health state can have a cost and/or benefit associated to it (called **rewards** in Markov models; the equivalent of payoffs in decision model)

20 A HIV Example

- Two therapeutic strategies for HIV: zidovudine monotherapy and zidovudine in combination with lamivudine (for simplicity, "monotherapy" versus "combination" therapy)
- Four possible health states, some of them depending on CD4 counts:
  - State A: CD4 from 200 to 500 (best)
  - State B: CD4 less than 200 (not so good)
  - State C: AIDS (bad)
  - State D: Death (very bad)
- Cycle length is **one year**

21 HIV state transition diagram

Let us draw the transition diagram

- All paths lead to death and death leads to ... nowhere (aka absorbing state). Everybody starts in state A.
- Note that there is no way to get better in this example (still true for HIV?)
- Possible to remain in same state after a cycle

22 Transition probability matrix

- use white board
- where do probabilities come from?

23 Solving Markov models

- We now have **health states**, **transition probabilities**.
- "Solving" the Markov model means that we will simulate what would happen to a group (cohort) of people over time
  - That is why they are called **cohort models** sometimes
  - We also have **rewards** for each cycle
  - We will then calculate **expected costs**
  - *why need simulation?*
- The HIV model (contd.). We will simulate the transitions of 1,000 patients in each type of therapy over 20 years.

24 Markov model variants and extensions

- discrete-time Markov chain, DTMC
- ignore covariates and time variables

- individualized transition probability modeling, e.g.

$$p(s_2|s_1) = f(a_1, a_2, \ldots, a_n)$$

- renewal process, semi-Markov process
- a semi-Markov process is called CTMC where all the holding times are exponentially distributed.

25 Sensitivity Analysis

- It is the study of how the uncertainty in the output of a mathematical model or system (numerical or otherwise) can be divided or allocated to different sources of uncertainty in its inputs.
- A related practice is **uncertainty analysis**, which has a greater focus on **uncertainty quantification** and **propagation of uncertainty**.
- ideally, uncertainty and sensitivity analysis should be run in tandem

26 Meta-Modeling

- metamodeling or meta-modeling is the analysis, construction and development of the frames, rules, constraints, models and theories applicable and useful for modeling a predefined class of problem.
- A metamodel or surrogate model is a model of a model.
- metamodeling is the process of generating such metamodels.
- types of metamodels:
  - polynomial metamodel
  - piecewise polynomial (spline) metamodels
  - neural network
  - Kriging metamodels

27 Model Calibration

- Model calibration is the process of adjustment of the model parameters and forcing within the margins of the uncertainties (in model parameters and / or model forcing) to obtain a model representation of the processes of interest that satisfies pre-agreed criteria (Goodness-of-Fit or Cost Function).
- Think about model calibration with a DTMC or CTMD
- Let us talk about some use of the Nelder-Mead algorithm

28 Sensitivity Analysis

The process of recalculating outcomes under alternative assumptions to determine the impact of a variable under sensitivity analysis can be useful for a range of purposes, including:

- *** Testing the robustness of the results of a model or system in the presence of uncertainty.
- *** Searching for errors in the model (by encountering unexpected relationships between inputs and outputs).
- *** Finding regions in the space of input factors for which the model output is either maximum or minimum or meets some optimum criterion (optimization and Monte Carlo filtering).
- *** In case of calibrating models with large number of parameters, a primary sensitivity test can ease the calibration stage by focusing on the sensitive parameters. Not knowing the sensitivity of parameters can result in time being uselessly spent on non-sensitive ones.
- *** Increased understanding of the relationships between input and output variables in a system or model.
- *** Enhancing communication from modelers to decision makers (e.g. by making recommendations more credible, understandable, compelling or persuasive).

  - In a comment published in 2020 in the journal *Nature* 22 scholars take COVID-19 as the occasion for suggesting five ways to make models serve society better.
  - One of the five recommendations, under the heading of 'Mind the assumptions' is to 'perform global uncertainty and sensitivity analyses [...] allowing all that is uncertain — variables, mathematical relationships and boundary conditions — to vary simultaneously as runs of the model produce its range of predictions

- Uncertainty reduction, through the identification of model input that cause significant uncertainty in the output and should therefore be the focus of attention in order to increase robustness (perhaps by further research).

- Model simplification – fixing model input that has no effect on the output, or identifying and removing redundant parts of the model structure.

- To seek to identify important connections between observations, model inputs, and predictions or forecasts, leading to the development of better models.

29 Sensitivity Analysis (contd.)

- A mathematical model (for example in biology, climate change, economics or engineering) can be highly complex

- Its relationships between inputs and outputs may be poorly understood.

- The model can be viewed as a black box, i.e. the output is an "opaque" function of its inputs.

- A diagram to discuss. Can you explain this in your research context?

30 DE Models

Some areas of application

- Ecology – predator-prey
- Epidemiology – spread of infections
- Immunology – immune response models
- HIV Infection
- Physiology – human respiration system
- Neural Networks
- Cell Kinetics
- Chemical kinetics – The Oregonator
- Physics – Ring Cavity Lasers, two-body problem of electrodynamics

31 Various Formulations

- An Initial Value Problem (IVP) for ODEs

$$\begin{aligned} y'(t) &= f(t, y(t)) \\ y(t_0) &= y_0 \end{aligned}$$

- Retarded Delay Differential Equations (RDDEs)

$$\begin{aligned} y'(t) &= f(t, y(t), y(t-\sigma_1), \ldots, y(t-\sigma_\nu)), \text{ for } t_0 \le t \le t_F \\ y(t_0) &= \phi(t), \text{ for } t \le t_0 \end{aligned}$$

$\sigma_i = \sigma_i(t, y(t)) \ge 0$ delay (constant/time dependent/state dependent)
$\phi(t)$ *history function* (constant/time dependent)

- Neural Delay Differential Equations (NDDEs)

$$\begin{aligned} y'(t) &= f(t, y(t), y(t-\sigma_1), \ldots, y(t-\sigma_\nu), \\ &\qquad y'(t-\sigma_{\nu+1}), \ldots, y'(t-\sigma_{\nu+\omega})), \text{ for } t_0 \le t \le t_F \\ y(t_0) &= \phi(t), y'(t) = \phi'(t), \text{ for } t \le t_0 \end{aligned}$$

32 Numerical Simulation of DEs

- Classic theory of step by step integration for IVPs
  - Runge-Kutta (RK)
  - Linear Multistep (LM) – BF Book Ch. 5.6, 5.7.
- Continuous solutions using polynomial approximation
  - Continuous Runge-Kutta (CRK)
  - Linear multistep methods have natural approximating polynomials.
- DDEs: combining an "interpolation" method (for evaluating delayed solution values) with an ODE integration method (for solving the resulting ODE).

33 Sensitivity Analysis – Model Setup

- Parameterized Models
  - A parameterized IVP

$$\begin{aligned} y'(t;\mathbf{p}) &= f(t, y(t;\mathbf{p});\mathbf{p}) \\ y(t_0) &= y_0(\mathbf{p}) \end{aligned}$$

    For example, $\mathbf{p} = \mu$ in the Van der Pol oscillator

$$\begin{aligned} y_1'(t) &= y_2(t) \\ y_2'(t) &= \mu(1 - y_1^2)y_2 - y_1 \end{aligned}$$

- A simple parameterized DDE

$$\begin{aligned} y'(t;\mathbf{p}) &= f(t, y(t;\mathbf{p}), y(t-\sigma);\mathbf{p}), \text{ for } t_0(\mathbf{p}) \le t \\ y(t;\mathbf{p}) &= \phi(t;\mathbf{p}), \text{ for } t \le t_0(\mathbf{p}) \end{aligned}$$

34 Sensitivity Analysis – Definitions

- Forward Sensitivity Analysis
  - The (first order) solution sensitivity analysis with respect to the model parameter p_i is defined as a vector

$$s_i(t;\mathbf{p}) = \{\frac{\partial}{\partial p_i}\} y(t;\mathbf{p}),\ (i = 1, \ldots, \mathcal{L})$$

  - The second order solution sensitivity with respect to the model parameters p_i and p_j is defined as the vector

$$r_{ij}(t;\mathbf{p}) = \{\frac{\partial}{\partial p_j}\} s_i(t;\mathbf{p}) = \{\frac{\partial^2}{\partial p_j \partial p_i}\} y(t;\mathbf{p}),\ (i, j = 1, \ldots, \mathcal{L})$$

35 Sensitivity Analysis – Importance

Sensitivity information can be used to:

- Estimate which parameters are most influential in affecting the behavior of the simulation. Such information is crucial for
  - Experimental Design
  - Data Assimilation
  - Reduction of complex nonlinear models
- Study of Dynamic Systems: Periodic orbits, the Lyapunov exponents, chaos indicators, and bifurcation analysis are fundamental objects for the complete study of a dynamical system, and they require computation of the sensitivities with respect to the initial conditions of the problem.
- Evaluate optimization gradients and Jacobians in the setting of
  - Dynamic Optimization
  - Parameter Estimation

36 Numerical SA of IVPs

- Finite Difference Approach

$$\{\frac{\partial}{\partial p_i}\}y(t;\mathbf{p}) \approx \frac{y(t;\mathbf{p}+e_i\Delta p_i)-y(t;\mathbf{p})}{\Delta p_i}$$

  This approximation may not be good due to the rounding errors.

- Internal Differentiation

$$y'(t;\mathbf{p}) = f(t, y(t;\mathbf{p});\mathbf{p}),\ y(t_0) = y_0(\mathbf{p})$$

Differentiation + Chain Rule + Clairaut's Theorem

$$s_i' = \frac{\partial f}{\partial y}s_i + \frac{\partial f}{\partial p_i},\ s_i(t_0) = \frac{\partial y_0(\mathbf{p})}{\partial p_i},\ (i = 1,\ldots,\mathcal{L})$$

- Taylor Series method using extended rules of Automatic Differentiation (Barrio 2006).
  - second(or higher)-order sensitivities.
- Let us look at one example problem together, i.e., the Van der Pol oscillator
- Look for built-in Matlab routines for numerical SA of IVPs. One example that asks you to apply the Finite Difference Approach and replicate a set of results.

37 One-Way Sensitivity Analysis

- One at a time variation: the simplest way to approach SA is to *vary each factor one at a time* (OAT).
- The biggest advantage is that any changes (including model failure) observed can be ascribed to the change in that one factor.
- Any limitation?
- Tornado diagram
- n-way SA

38 Sampling-based SA – Overview

- based on Monte Carlo approaches
- The underlying idea:

$$\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_{nY}(\mathbf{x})]$$

  are functions of uncertain analysis inputs $\mathbf{x} = [x_1, x_2, \ldots, x_{nX}]$.
- In turn, uncertainty in $\mathbf{x}$ results in a corresponding uncertainty in $\mathbf{y}(\mathbf{x})$
- Two questions:
  - What is the uncertainty in $\mathbf{y}(\mathbf{x})$ given the uncertainty in $\mathbf{x}$? – uncertainty analysis
  - How important are the individual elements of $\mathbf{x}$ with respect to the uncertainty in $\mathbf{y}(\mathbf{x})$? – sensitivity analysis
  - In practice, the implementation of an uncertainty analysis and the implementation of a sensitivity analysis are very closely connected on both a conceptual and a computational level.

39 Key Topics

1. definition of distributions $D_1, D_2, \ldots, D_{nX}$ that characterize the epistemic uncertainty in the elements $x_1, x_2, \ldots, x_{nX}$ of $\mathbf{x}$.
   - epistemic: relating to knowledge or to the degree of its validation.
2. generation of a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{nS}$ from the $\mathbf{x}$'s in consistency with the distributions $D_1, D_2, \ldots, D_{nX}$.
3. propagation of the sample through the analysis to produce a mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$, from analysis inputs to analysis results.
4. presentation of uncertainty analysis results (i.e., approximations to the distributions of the elements of $\mathbf{y}$ constructed from the corresponding elements of $\mathbf{y}(\mathbf{x}_i)$, $i = 1, 2, \ldots, nS$).
5. determination of sensitivity analysis results (i.e., exploration of the mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$).

- We only consider probabilistic characterizations of uncertainty.
- alternative uncertainty representations are active areas of research ***
  - evidence theory
  - possibility theory
  - fuzzy set theory
  - interval analysis

40 Characterization of uncertainty

- Definition of the distributions $D_1, D_2, \ldots, D_{nX}$ that characterize the epistemic uncertainty in the elements $x_1, x_2, \ldots, x_{nX}$ is the most important part of a sampling-based uncertainty and sensitivity analysis.
- These distributions determine both the uncertainty in $\mathbf{y}$ and the sensitivity of the elements of $\mathbf{y}$ to the elements of $\mathbf{x}$.
- The distributions $D_1, D_2, \ldots, D_{nX}$ are typically defined through an expert review process, and their development can constitute a major analysis cost.

41 Generation of sample

- random sampling
- Latin hypercube sampling
- orthogonal sampling
- importance sampling

42 Latin hypercube sampling

- The range of each x_j is exhaustively divided into nS disjoint intervals of equal probability;
- One value x_{ij} is randomly selected from each interval;

- The nS values of x_1 are randomly paired without replacement with the nS value for x_2 to produce nS pairs;
- These pairs are then randomly combined without replacement with the nS values for x_3 to produce nS triples;
- This process is continued until a set of nS nX-tuples $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{i,nX}]$, $i = 1, 2, \ldots, nS$, is obtained.
- This set constitutes the Latin hypercube sample with nS data (prespecified).

Let us look at an illustration! Please see Figure 1.
Let us look at an example. Please see Table 1. *More in the homework.*

- Orthogonal sampling is basically an extension to Latin hypercube sampling that ensures that each subspace is evenly sampled.
- this can ensure that correlation between sampling dimensions is minimized.

Let us look at a comparative example! Please Figure 2.

43 Control of correlations

- An important aspect: correlated variables should have correlations close to their specified values, and uncorrelated variables should have correlations close to zero.
- In general, the imposition of complex correlation structures is not easy.
- Iman and Conover developed a broadly applicable procedure to impose rank correlations on sample values that
  1. is distribution free, i.e., does not depend on the assumed marginal distributions for the sampled variables,
  2. can impose complex correlation structures involving multiple variables,

3. works with both random and Latin hypercube sampling, and
4. preserves the intervals used in Latin hypercube sampling.

- Iman and Conover (1982) proposed a method to induce correlation among the variables by restricting the way variables are paired based on the *rank correlation* of some target values.
  - The method is based on the Cholesky decomposition of the correlation matrix.
  - The objective is to rearrange the input variables close to the target correlation matrix.
  - *Implementation and analysis in the homework!*
  - Additionally, Stein (1987) also proposed a method for sampling dependent variables based on the rank of a target multivariate distribution.

Let us revisit the example. Please see Table 1 again.

44 Importance sampling - Brief introduction

- What if large sample sizes are required to provide appropriate coverage of low probability but high consequence subsets of values of $\mathbf{x}$?
- Then importance sampling may be a more effective sampling procedure.
- Importance sampling is a variance reduction technique that can be used in the Monte Carlo method.
- Importance sampling complicates sensitivity analysis as the individual sampling elements do not have equal weight (i.e., likelihood of occurrence).

45 Propagation of the sample through the analysis

- Propagation of the sample through the analysis to produce the mapping $[\mathbf{x}_i, \mathbf{y}(\mathbf{x}_i)]$, $i = 1, 2, \ldots, nS$, from analysis inputs to analysis results

- often the most computationally demanding part of a sampling-based uncertainty and sensitivity analysis
- ranging from very simple for analyses that involve a single model to very complicated for large analyses that involve complex systems of linked models
- When a single model is under consideration, put a loop around the model
    1. supply the sampled input to the model
    2. run the model
    3. store model results for later analysis
- When more complex analyses with multiple models are involved, considerable sophistication may be required
    1. development of simplified models to approximate more complex models
    2. clustering of results at model interfaces
    3. reuse of model results through interpretation or linearity properties
    4. complex procedures for the storage and retrieval of analysis results

46 Presentation of uncertainty analysis results

- In general, straightforward.
- Presentation possibilities: means and standard deviations, desnity functions, CDFs, complementary cumulative distribution functions (CCDFs), and box plots
- The latter few techniques are more preferable.

47 Determination of sensitivity analysis results

- much more demanding than the last one
- Let us summarize the notation
- Scatterplots
  - A plot of the points $[x_{ij}, y_j]$ for $i = 1, 2, \ldots, nS$, i.e., a scatterplot of y vs. x_j
  - can reveal nonlinear or other unexpected relationships between analysis inputs and analysis results.
  - Scatterplots are a natural starting point in a complex analysis that can help in the development of a sensitivity analysis strategy using one or more additional techniques.
  - Often the examination of scatterplots is all that is needed to understand the relationships between the uncertainty in analysis inputs and the uncertainty in analysis results.
  - one-dimensional scatterplots? most analyses start with 2D scatterplots.
  - What about 3D? Sometimes good too. See Figure 3, together with a parameter table.
- Correlation
  - Correlation provides a measure of the strength of the linear relationship between x_j and y.
  - Pearson correlation coefficient (CC) $c(x_j, y)$ between x_j and y.
  - CC has a value between -1 and 1, with positive value indicating that x_j and y tend to increase and decrease together; with negative value indicating they tend to move in opposite directions.
  - 0: no linear relationship; 1: exact linear relationship between x_j and y.
  - CC is closely related to results obtained in a linear regression relating y to x_j. Specifically, $c(x_j, y)$ is equal to the standardized regression coefficient (SRC).

- Regression analysis
  - provide an algebraic representation of the relationships between y and one or more of the x_j's.
  - commonly, linear models
  - SRC: $b_j \hat{s}_j / \hat{s}$, providing a meaure of variable importance based on the effect on y relative to the standard deviation $\hat{s}$ of y.
  - When x_j''s are not independent, SRCs do not provide reliable indications of variable importance.

- Partial correlation characterizes the linear relationship between x_j and y after a correction that has been made for the linear effects on y of the remaining elements of $\mathbf{x}$

- Other techniques: rank transformations, statistical tests for patterns based on gridding, entropy tests for patterns based on gridding (*nonlinear and even nonmonotonic*); nonparametric regression; squared rank differences/rank correlation coefficient (SRD/RCC) test; 2D KS test, top down coefficient of concordance (TDCC), variance decomposition.

48 More on importance sampling

- The essence is to approximate a random quantity of interest for a given distribution even if we cannot directly sample from that particular distribution.

- See the white board for the setup of the task

- Watch a Youtube video for further introduction

- https://www.youtube.com/watch?v=V8f8ueBc9sYt=331s

- prior knowledge is needed!

49 Uncertainty Analysis

- Experimental Uncertainty Assessment, or Uncertainty Analysis in physical experiments, deals with assessing the uncertainty in measurement.

  - An experiment can be affected by errors due to instrumentation, methodology, and presence of confounding effects and so on.
  - Experimental uncertainty estimates are needed to assess the confidence in the results.
  - A related field is design of experiments.
  - I have uploaded a summary document on experimental uncertainty assessment methodology with an example provided.

- Interval finite element method (interval FEM) is a finite element method that uses interval parameters.
  - Interval FEM can be applied in situations where it is not possible to get reliable probabilistic characteristics of the structure.

- Propagation of Uncertainty (or propagation of error) is the effect of variables' uncertainties (or errors, more specifically random errors) on the uncertainty of a function based on them.
  - Forward uncertainty propagation (will get back to this again)

50 Uncertainty Quantification

- The science of quantitative characterization and reduction of uncertainties in both computational and real-world applications.
- It attempts to determine how likely certain outcomes are if some aspects of the system are not exactly known.

51 Sources of Uncertainty

Uncertainty can enter mathematical models and experimental measurements in various contexts.

To categorize the sources of uncertainty, one can consider:

- Parameter uncertainty. This comes from the model parameters that are inputs to the computer model (mathematical model) but whose exact values are unknown to experimentalists and cannot be controlled

in physical experiments, or whose values cannot be exactly inferred by statistical methods.

- Parametric variability. This comes from the variability of input variables of the model.

- Structural uncertainty. Also known as model inadequacy, model bias, or model discrepancy, this comes from the lack of knowledge of the underlying physics in the problem.

- Algorithmic uncertainty. Also known as numerical uncertainty, or discrete uncertainty. This type comes from numerical errors and numerical approximations per implementation of the computer model. For example, the *finite element method* or *finite difference method* may be used to approximate the solution of a *partial differential equation*, which introduces numerical errors.

- Experimental uncertainty. Also known as observation error, this comes from the variability of experimental measurements.

- Interpolation uncertainty. This comes from a lack of available data collected from computer model simulations and/or experimental measurements.

Sometimes uncertainty is classified into two categories, prominently seen in medical applications.

- Aleatoric uncertainty is also known as statistical uncertainty, and is representative of unknowns that differ each time we run the same experiment.

- Epistemic uncertainty is also known as systematic uncertainty. It is due to things one could in principle know but do not in practice. This may be because a measurement is not accurate, the model neglects certain effects, particular data has been delibrately hidden.

- In real life applications, both types of uncertainties are present.

- Uncertainty quantification intends to explicitly express both types of uncertainty separately.

- The quantification for the aleatoric uncertainty can be relatively straightforward.
- Traditional (frequentist) probability is the most basic form.
- Techniques such as the Monte Carlo method are frequently used.
- A probability distribution can be represented by its moments.
- In the Gaussian case, the mean and covariance suffice.

- To evaluate epistemic uncertainties, the efforts are made to understand the (lack of) knowledge of the system, process or mechanism.
- Representation of epistemic uncertainty can be based on methods such as probability bounds analysis, fuzzy logic or evidence/brief theories such as subjective logic.

- In mathematics, uncertainty is often characterized in terms of a probability distribution.
- In that perspective, epistemic uncertainty means not being certain what the relevant probability distribution is, and aleatoric uncertainty means not being certain what a random sample drawn from a probability distribution will be.

52 Uncertainty Quantification

Two types of problems: forward propagation and inverse assessment.

- Forward propagation of uncertainty. It is the process of quantifying uncertainties in system output(s) propagated from uncertain input(s).
- It focuses on the influence on the outputs from the *parametric variability* listed in the sources of uncertainty.
- There has been a proliferation of research on it and a majority of uncertainty analysis techniques were developed for it.
- The objectives:

  - To evaluate low-order moments of the outputs, i.e., mean and variance.
  - To evaluate the reliability of the outputs, especially useful in reliability engineering where outputs of a system are usually closely related to the performance of the system.
  - To assess the complete probability distribution of the outputs. This is useful in the scenario of utility optimization where the complete distribution is used to calculate the utility.

- Probabilistic approaches to uncertainty propagation, most rigorous approach to uncertainty analysis in engineering design
  - simulation-based methods: Monte Carlo simulations, importance sampling, and adaptive sampling
  - local expansion-based methods
  - functional expansion-based methods
  - numerical integration-based methods
- Non-probabilistic approaches, e.g., interval analysis, fuzzy theory

53 Inverse uncertainty quantification

- assess model uncertainty and parameter uncertainty (where the model parameters are calibrated simultaneously using test data).
- drawing increasing attention in the engineering design community, since uncertainty quantification of a model and the subsequent predictions of the true system response(s) are of great interest in designing robust systems.
- What is an inverse problem?
- Three scenarios in inverse uncertainty quantification.
  - Bias correction only

  - Parameter calibration only
  - Bias correction and parameter calibration
- Frequentist: regression and least squares problem
- Bayesian: GP modeling...

54 Interpolation and Polynomial Approximation – Introduction

Let us look at one example first!

What is *Interpolation*?

What is the potential need for approximation in your research?

One of the most useful and well-known classes of functions mapping the set of real number into itself is the class of *algebraic polynomials*

55 Algebraic Polynomials

Algebraic polynomials are the set of functions of the form

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

where n is a nonnegative integer and $a_0, \ldots, a_n$ are real constants.

- They can uniformly approximate continuous functions.
- Given any function, defined and continuous on a closed and bounded interval, there exists a polynomial that is as "close" to the given function as desired.
- Weierstress Approximation Theorem
- Taylor polynomials about $x_0 = 0$ for $f(x) = e^x$.
- Is higher-order Taylor polynomials always the best?

56 Taylor Polynomials

- The Taylor polynomials have the property that all the information used in the approximation is concentrated at the single point x_0.
- This limits the applicability to the situation where approximations are only needed at points close to x_0.
- It should be better we can include information at various points
- Thus, Taylor polynomials is NOT for approximation purposes but for derivation of numerical techniques (e.g., Taylor expansion based numerical methods for solving ODEs) and error estimation.

57 Interpolation and Lagrange Polynomial

- *Linear Interpolation.* Given points (x_0, y_0) and (x_1, y_1), determine a polynomial of degree 1.
- Generalize the concept of linear interpolation.
  - Consider the construction of a polynomial of degree at most n that passes through the $n + 1$ points,
    $$(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))$$
  - We need to construct, for each $k = 0, 1, \ldots, n$, a function $L_{n,k}(x)$ with the property that $L_{n,k}(x_i) = 0$ when $i \neq k$ and $L_{n,k}(x_k) = 1$.
  - To satisfy $L_{n,k}(x_i) = 0$ for each $i \neq k$ requires that the numerator of $L_{n,k}(x)$ contains the form
    $$(x - x_0)(x - x_1) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)$$
  - To satisfy $L_{n,k}(x_k) = 1$, the denominator of $L_{n,k}(x)$ must be equal to this term evaluated at $x = x_k$. Thus,
    $$L_{n,k}(x) = \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}.$$
  - The interpolating polynomial is easily described once the form of $L_{n,k}$ is known.
  - This polynomial is called the *nth* Lagrange interpolating polynomial.

58 Practical Difficulty

- A practical difficulty with Lagrange interpolation is that since the error term is difficult to apply, the degree of the polynomial needed for the desired accuracy is generally not known until computations are determined.
- The usual practice is to compute the results given from various polynomials until appropriate agreement is obtained.
- However, there is no "boosting" effect in computing a high-degree approximation once lower-degree approximations are available.

- iterated interpolation, i.e., generating successively higher-degree polynomial approximations at a specific point
- divided-difference methods, i.e., successively generate the polynomials themselves.
- Hermite interpolation, i.e., generalize both the Taylor polynomials and the Lagrange polynomials.

59 Cubic Spline Interpolation

- Previous techniques concerned the approximation of arbitrary functions on closed intervals by the use of polynomials.
- However, the oscillatory nature of high-degree polynomials and the property that a fluctuation over a small portion of the interval can induce large fluctuations over the entire range restricts their use.

- An alternative approach is to divide the interval into a collection of subintervals and construct a (generally) different approximation polynomial on each subinterval.
- Piecewise-polynomial approximation.

- the simplest form is piecewise-linear approximation
- the concept of cubic spline interpolant
- natural cubic spline, algorithm statement on page 146 in BF
- clamped cubic spline, algorithm statement on page 148 in BF

60 Parametric Curves

- What if a curve cannot be expressed as a function of one coordinate variable in terms of the other.
- To represent general curves, We use a parameter express both the x- and y-coordinate variables.
- This technique can be extended to represent general curves and surfaces in space.
- A straightforward parametric technique is to use a parameter t to construct a pair of approximate polynomials. See Page 157 in BF

61 Surrogate Modeling

- **Surrogate Modeling** is an engineering method used when an outcome of interest cannot be easily directly measured so a model of the outcome is used instead.
- Most engineering design problems require experiments and/or simulations to evaluate design objective and constraint functions as a function of design variables.
- However, a single experiment/simulation may take many minutes/hours/days for many real-world problems to complete.
- For stochastic simulation or web-lab experiments that require repeated measures, many replications are needed to obtain representative results and gain statistical confidence.
- Let us use a couple of design optimization problems

  - Calibration of state-transition model
  - Capacity optimization with QoS requirements
  - see a later section for the distinction with design space approximation problem

- One way to alleviate the above burden is by constructing approximation models, known as surrogate models, response surface models, metamodels, that mimics the behavior of the simulation model as closely as possible while being computationally cheap to evaluate. A data-driven, bottom-up approach.
  - what is as closely as possible?
  - how to intelligently choose data points?
  - does the closeness translate into accurate solution to the engineering problem?

- The exact, inner working of the simulation code is not assumed to be known (or even understood). Thus, we solely rely on the input-output pairs. Black-box modeling.
  - When only a single design variable, the process is known as curve fitting.
  - now we see the connection with polynomial interpolating curve (spline) fitting.

62 Goals and what we need to do

The scientific challenge of surrogate modeling is the generation of a surrogate that is as accurate as possible, using as few evaluations as possible.

The process comprises three major steps which may be interleaved iteratively:

- Sample selection (also known as sequential design, optimal experimental design or active learning)
- Construction of the surrogate model and optimizing the model parameters (known as bias-variance trade-off, see a later section; an explanation based off machine learning).

- Appraisal of the accuracy of the surrogate.

The accuracy of the surrogate depends on the number and location of samples (expensive experiments and simulations) in the design space.

63 Bias-variance dilemma

- **Bias-variance trade-off** is the property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters.
  - The *bias error* is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (i.e., underfitting).
  - The *variance* is an error from sensitivity to small fluctuations in the training set.
- The **bias-variance dilemma** is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set.
- interval validity (interpretation) vs. external validity (prediction)

64 Two types of applications

design optimization vs. design space approximation

- Note that both applications are based on surrogate modeling.
- In surrogate model based optimization, an initial surrogate is constructed using some of the available budgets of expensive experiments and/or simulations.

  The remaining experiments/simulations are run for designs which the surrogate model predicts may have promising performance.

- In design space approximation, one is not interested in finding the optimal parameters but rather in the global behavior of the system.

  Here the surrogate is tuned to mimic the underlying model as closely as needed over the complete design space.

  The underlying model is a probability distribution; sampling is the key, later we will talk about Markov chain Monte Carlo sampling

65 Surrogate model based design optimization

Search/update procedure:

1. initial sample selection (experiment/simulation)
2. construct surrogate model
3. run and update experiment/simulation at a new location(s) found by search and add to sample
4. iterate steps 2 to 4 until out of time or design "good enough"

Key steps:

- determine which factors really influence the outcome; **tool; screening designs like fractional factorial**; *sample selection/factor screening*
- approach optimum by repeated change of factor settings; **tools: Box/simplex or steepest ascent approach**, *optimum search/improvement*
- *determination of optimum* (find optimal settings of factor settings; tool: response surface designs + analysis of response surface using eigenvalues)

Now let us see a few illustrations.

Now let us see a practical and representative optimization scheme. At this point, we only see a small portion of the diagram. We will visit it again.

Now let us see an example, using regression models and emphasize the factor screening.

66 Regression models used in optimization

building regression models for factor screening

- "Far away" from the optimum a first-order model often suffices. For example,
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$
- "Near" the optimum often a quadratic (second order) model suffices. For example,
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$$
- Lack-of-fit metric must be applied to check whether these models are appropriate, since we cannot directly see whether we are near the optimum.

67 Search/Improvement Step

In order to efficiently move from current factor settings to factor setting that yield near-optimal values, 2 methods are available:

- Box/Simplex method
  - idea: form new full factorials in direction of largest increase in current full factorial
  - simple; no statistics needed for implementation
  - not efficient
- Steepest ascent/descent method
  - idea: use 1st order regression model from fractional factorial to obtain direction of largest increase ("steepest ascent")
  - perform single runs in direction of largest increase until increase stops
  - advanced
  - recommended since it is more efficient

68 Quadratic models

Near the optimum, lack-of-fit indicates curvature. Hence, we are now probably near the optimum.

In order to fit a quadratic model (suitable when we are near the optimum), we must vary the factors at 3 levels.

A 2^p-design with center points does not suffice, because then all quadratic factors are confounded.

A 3^p-design is possible, but not to be recommended:

- the number of runs grows fast
- uses more runs than necessary to fit a quadratic model.

69 Response surface designs

The following designs are widely used for fitting a quadratic model:

- Central Composite Design (uniform precision of effect estimates)
- Box-Behnken Design (almost uniform precision of effect estimates, but usually fewer runs required than for CCD)

The choice between these models is usually decided by the availability of these designs for a given number of runs and number of factors.

Note that there are other suitable designs (usually available in statistical software that supports DOE).

Finally, fit a quadratic regression model (previous page)

70 More on CCD Designs

- A Box-Wilson Central Composite Design, commonly called "a central composite design."
- The design contains an imbedded factorial or fractional factorial design with center points that is augmented with a group of 'star points' that allow estimation of curvature.
- see an illustration for the generation of a **Central Composite Design** for two factors.

- A central composite design always contains twice as many star points as there are factors in the design.
- The star points represent new extreme values (low and high) for each factor in the design.

See a pictorial representation of where the star points are placed for the 3 types of CCD designs.

- **Circumscribed – CCC**. CCC designs are the original form of the central composite design. The star points are at some distance alpha from the center based on the properties desired for the design and the number of factors in the design. The star points establish new extremes for the low and high settings for all factors.

  A CCC design has circular, spherical, or hyperspherical symmetry and require 5 levels for each factor. Augmenting an existing factorial design with star points can produce this design.

- **Inscribed – CCI**. For those situations in which the limits specified for factor settings are truly limits, the CCI design uses the factor settings as the star points and creates a factorial or fractional factorial design within those limits. In other words, a CCI design is a scaled down CCC design.

  A CCI design also requires 5 levels of each factor.

- **Face Centered – CCF**. In this design the star points are at the center of each face of the factorial space.

  This variety requires 3 levels of each factor. Augmenting an existing factorial design with appropriate star points can also produce this design.

- CCC explores the largest process space and CCI explores the smallest process space. Both CCC and CCI are rotatable designs, but the CCF is not. In CCC, the design points describe a circle circumscribed about the factorial square. For three factors, the CCC design points describe a sphere around the factorial cube.

What is blocking in DoE? In the statistical theory of the design of experiments, blocking is the arranging of experimental units in groups (blocks)

that are similar to one another. Blocking reduces unexplained variability. So in our setting, this is not a desirable thing!

Example. **Male and Female**: An experiment is designed to test a new drug on patients. There are two levels of the treatment, drug, and placebo, administered to male and female patients in a double blind trial. The sex of the patient is a blocking factor accounting for treatment variability between males and females.

71 More on Box-Behnken Designs

- The Box-Behnken design is an independent quadratic design in that it does not contain an embedded factorial or fractional factorial design.
- In this design the treatment combinations are at the midpoints of edges of the process space and at the center.
- These designs are rotatable (or near rotatable) and require 3 levels of each factor.
- The designs have limited capability for orthogonal blocking compared to the central composite designs.
- See a pictorial representation of a Box-Behnken Design for three factors.

72 When comes to optimization

Recap: optimization in one dimension

- Necessary condition for optimum: 1st derivative = 0
- not sufficient: "point of inflection"
- extra sufficient condition: 2nd derivative $\neq 0$

Determination of type of optimum

- Graphically: make contour plot (if 2 factors)

- Analytically:

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \hat{\beta}_{ii} x_i^2 + \sum_{i<j} \sum \hat{\beta}_{ij} x_i x_j$$

  matrix notation:

$$\hat{y} = \hat{\beta}_0 + x^T b + x^T B x$$

  B must be a symmetric matrix. write on board.

  stationary point (zero first-order derivatives):

$$\frac{\partial y}{\partial x} = b + 2Bx = 0 \implies x_0 = -\frac{1}{2} B^{-1} b$$

  characterization through eigenvalues of matrix B:

$$Bx = \lambda x$$

  - all eigenvalues positive: min
  - all eigenvalues negative: max
  - eigenvalues different signs: saddle point

73 ML based surrogate models

- Support Vector Regression (SVR) and Artificial Neural Networks (ANN)
- SVMs solve binary classification problems by formulating them as convex optimization problems.
- SVM generalization to SVR is accomplished by introducing an ϵ-insensitive region around the function, called the ϵ-tube.
- This tube reformulates the optimization problem to find the tube that best approximates the continuous-valued function, while balancing model complexity and prediction error.
- For more information, see "Support Vector Machine" in the supplemental materials folder.

- As for ANN, I will give you guys a task. Please find out from your self-learning the following?
  - what is ANN?
  - how to use ANN as surrogate models for measurements in optimization algorithms
  - what are the built-in functions in Matlab and/or Python?
- I will assign a task where you use SVR as surrogate models with the same set of prompts as above.
- look for a built-in function in Matlab/Python. Look for Youtube videos.

74 Radial basis function

- RBF is an effective algorithm for smoothing and interpolating the experimental data.
- The form of the approximate model is a basis function of the Euclidean distance between the sampled data point and the point to be predicted.
- RBF uses linear combinations of radial symmetric functions of the Euclidean distance to build approximate models.
- Mathematically, the model can be expressed as

$$\hat{y}(x) = \sum_{i=1}^{N} w_i \phi(||x - c_i||)$$

  where the approximated function $\hat{y}(x)$ is represented as a sum of N RBF ϕ, each associated with a difference center (measured location) c_i and weighted by an appropriate coefficient, w_i.
- RBF approximation is capable of producing good fits to arbitrary contours of both deterministic and stochastic response functions.
- The radial function $\phi(r)$ can take many forms
  - Linear Function: $\phi(r) = r$.

  - Cubic Function: $\phi(r) = r^3$.
  - Thin Plate Function: $\phi(r) = r^2 \log r$.
  - Multi-quadratic Function: $\phi(r) = \sqrt{1 + a^2 r^2}$.
  - Gaussian Function: $\phi(r) = e^{-ar^2}$.

- The weights w_i can be estimated using the matrix methods of linear least squares, because the approximating function is linear in the weights w_i.

- The model (sum) can be interpreted as a rather simple single-layer type of artificial neural network called a radial basis function network, with the radial basis functions taking on the role of the activation functions of the network.

- It can be shown that any continuous function on a compact interval can in principle be interpolated with arbitrary accuracy by a sum of the sum form, if a sufficiently large number of N of radial basis functions is used.

- I will assign a task where you use RBF with different forms of radial functions as surrogate models. With the following two machine learning models, you are expected to make a comparative investigation.

75 Spatial statistics with interpolation

The general formula is formed as a weighted sum of the data (i.e., linear combination). Given N data,

$$\hat{Y}(\mathbf{u}_0) = \sum_{i=1}^{N} \lambda_i Y(\mathbf{u}_i)$$

where:

- $Y(\mathbf{u}_i)$ = the measured value at the ith location
- λ_i = an unknown weight for the measured value at the ith location
- $\mathbf{u}_0$ = the prediction location

- N = the number of measured values
- See an illustration for the general idea. Slides no. 1, no. 2
- All interpolation methods (inverse distance squared, splines, radial basis functions, triagulation, etc.) weight the surrounding measured values (locations) to derive a prediction for an unmeasured location.
- Almost all interpolation methods assign weights according to functions that give a decreasing weight with increasing separation distance.
- IDW (Inverse Distance Weighted): estimates *cell values* by weighted averaging the values of *sample data points* in the *neighborhood of each processing cell*. The closer a point is to the center of the cell being estimated, the more influence, or weight, it has in the averaging process.
- In IDW, the weight λ_i, depends solely on the distance to the prediction location.
- The IDW, BRF, and spline interpolations are referred to as deterministic interpolation methods because they are directly based on the surrounding measured values or on specified mathematical formulas (through some arbitrary function) that determine the smoothness of the resulting surface.

76 Krigging metamodeling in simulation

- it is also known as a Gaussian process (GP) model, which is a mathematical interpolation method.
- it is named after the South African mining-engineer Krige (1919 – 2013). He solved the problem of interpolating results (outputs, responses) that were obtained at a limited number of locations for gold mining
- have seen widely applied in geostatistics (spatial statistics).
- The technique was formalized by the French mathematician Matheron (1930 – 2020).

- This basis of this formalization is the stationary GP.
- This stationarity means that the GP has a constant mean, a constant variance, and conavriances that depend only on the distance between "points" in a k-dimensional space.
  - obviously, in spatial statistics $k \leq 3$ (length, width, and height)
- This GP defines a multivariate normal (or Gaussian) distribution. That is, output X of our computer simulation $X \sim (\mu, P)$. Both μ and P need to be estimated.
- At present, GP is well applied in *machine learning*. So effective and efficient constrution of a GP has been intensively studied.
- The connection with our context: how to develop and apply GPs in experiments with computerized simulation models, also including how to determine sample size and how to generate samples, etc.
- So we call a GP a metamodel of the underlying simulation model.
- Many applications arose in the area of spatial statistics with increasing applications in biomedicine and population health.
- A few points before getting into details.
  - We can use LHS to select the sample inputs to be simulated.
  - Let us focus on deterministic simulation; for stochastic simulation, it requires adjusting the design and analysis.

77 Kriging method – Overview

- What about kriging? Optimal interpolation based on regression against observed z values of surrounding data points, weighted according to spatial covariance values.
- A probabilistic framework. Kriging can take into account the data configuration, the distance between data and target, the spatial correlations among the measured points, and potential external information.

- To use the spatial arrangement in the weights, the spatial autocorrelation must be quantified.
- Two kriging methods: ordinary kriging and universal kriging.
- In ordinary kriging (linear regression), the weight, λ_i, depends on a fitted model to the measured points, the distance to the prediction location, and the spatial relationships among the measured values around the prediction location.
- In universal kriging, there is the assumption on an overriding trend in the data (i.e., potential external information).
- One significant advantage of using a GP model is that Kriging also quantifies the uncertainty of its predictor, i.e., Kriging also gives the variance on the predictor, whereas many other metamodels (e.g., neural nets, splines) do not. See an illustration slide – Example: porosity of a geological medium. slide no. 3.

It is because of these two distinct tasks that it has been said that kriging uses the data twice: the first time to estimate the spatial autocorrelation of the data and the second to make the predictions.

78 General description of the kriging approach

All kriging estimators $\hat{Z}(\cdot)$ are some variant of the basic linear regression estimator.

$$\hat{Z}(\mathbf{u}) - \mu(\mathbf{u}) = \sum_{i=1}^{n(\mathbf{u}_0)} \lambda_i [Z(\mathbf{u}_i) - \mu(\mathbf{u}_i)],$$

where

- $\mathbf{u}$, $\mathbf{u}_i$: data point (location) whose response are needed for estimation, and observed data points, indexed by i, in the neighborhood.
- $n(\mathbf{u})$: number of data points in the local neighborhood used for estimation of $\hat{Z}(\mathbf{u})$.
- $\mu(\mathbf{u}), \mu(\mathbf{u}_i)$: expected values (means) of $Z(\mathbf{u})$ and $Z(\mathbf{u}_i)$.

- $\lambda_i(\mathbf{u})$: kriging weight assigned to response $Z(\mathbf{u})$ for estimation location $\mathbf{u}$.

$Z(\mathbf{u})$ is treated as a random field (Gaussian process) with a trend component, $\mu(\mathbf{u})$, and a residual component, $R(\mathbf{u}) = Z(\mathbf{u}) - \mu(\mathbf{u})$.

The goal is to determine weights, λ_i, that minimizes the variance of the estimator (estimation error variance)

$$\sigma_e^2(\mathbf{u}) = \text{Var}\{\hat{Z}(\mathbf{u}) - Z(\mathbf{u})\}$$

under the unbiasedness constraint $E\{\hat{Z}(\mathbf{u}) - Z(\mathbf{u})\} = 0$.

The random field (Gaussian process) $Z(\mathbf{u})$ is decomposed into residual and trend components, i.e., $Z(\mathbf{u}) = R(\mathbf{u}) + \mu(\mathbf{u})$, with the residual component treated as a Gaussian process with a stationary mean of 0 and a stationary covariance (a function of lag, $\mathbf{h}$, but not of position, $\mathbf{u}$):

$$E\{R(\mathbf{u})\} = 0$$

$$Cov\{R(\mathbf{u}), R(\mathbf{u}+\mathbf{h})\} = E\{R(\mathbf{u}) \cdot R(\mathbf{u}+\mathbf{h})\} = C_R(\mathbf{h}).$$

The residual covariance function is generally derived from the input semivariogram model. Thus the semivariogram we feed to a kriging program should represent the residual component of the variable.

Kriging estimates residual at $\mathbf{u}$ as weighted sum of residuals at surrounding data points.

Kriging weights, λ_i, are derived from covariance function or semivariogram, which should characterize residual component.

79 Simple Kriging

For simple kriging, we assume that the trend component is a constant and known mean, $\mu(\mathbf{u}) = \mu$. Thus, the estimate is automatically unbiased. So the optimization problem becomes unconstrained. That is, given

$$\hat{Y}(\mathbf{u}_0) = \sum_{i=1}^{N} \lambda_i Y(\mathbf{u}_i),$$

we want to minimize

$$\sigma_e^2(\mathbf{u}) = \text{Var}\{\hat{Y}(\mathbf{u}) - Y(\mathbf{u})\}.$$

Still quite mathematically involved.

Now let us just introduce some terminology along the procedure.

80 Variogram

- an important concept in geostatistics
- Fitting a model, or spatial modeling, is also known as structural analysis, or variography.
- In spatial modeling of the structure of the measured points, you begin with a graph of the empirical semivariogram (or just call variogram), computed with the following equation for all pairs of locations (i.e., $\mathbf{u}_i$ and $\mathbf{u}_j$) separated by distance (or lag) h:

$$\text{Semivariogram(distant\_h)} = 0.5 \times ((Y(\mathbf{u}_i) - Y(\mathbf{u}_j))^2$$

- The above formula involves calculating the difference squared between the values of the paired measured locations. See slide no. 4.
- Often, each pair of locations has a unique distance, and there are often many pairs of points. See slide no. 5 (variogram cloud)
- To plot all pairs may quickly become unmanageable. Thus, instead of plotting each pair, the pairs are grouped into lag bins.
- The empirical semivariogram is a graph of the averaged semivariogram values on the y-axis and the distance (or lag) on the x-axis. See slide no. 6.

81 Fitting a model to the empirical variogram

- The next step is to fit a model to the points forming the empirical variogram.
- Variogram modeling is a key step between spatial description and spatial prediction.
- It provides information on the spatial autocorrelation of datasets.
- However, it does not provide information for all possible directions and distances.

- For this reason, and to ensure that kriging predictions have positive kriging variances, it is necessary to fit a model—that is, a continuous function or curve—to the empirical variogram.
- Abstractly, this is similar to regression analysis, in which a continuous line or curve is fitted to the data points.
- Select a variogram, see slide no. 7.
  - Circular, Spherical, Exponential, Gaussian, Linear
  - The selected model influences the prediction of the unknown values, particularly when the shape of the curve near the origin differs significantly.
  - The steeper the curve near the origin, the more influence the closest neighbors will have on the prediction.
  - As a result, the output surface will be less smooth.
  - Each model is designed to fit different types of phenomena more accurately
  - See slide no. 8
  - Understanding a semivariogram—Range, sill, and nugget

82 Make predictions with kriging

- have uncovered the dependence or autocorrelation in your data
- have fitted a model for the spatial correlation function
- next, to construct the kriging model, which is the most computatonally challenging step.
- in general, μ, σ^2, and θ, hyperparameters transforming the degree of resembalance, all need to be estimated.
- Note that with the simple kriging model, μ is known, hyperparameters are set in a forward manner, i.e., the spatial correlation function. What is left is σ^2. We omit the remaining discussion.
- Now one can make predictions on the unknown locations

- Like IDW interpolation, kriging forms weights from surrounding measured values. Just in a more sophisticated manner.

83 Gradient-enhanced kriging

- further address the challenge where only a small number of evaluations of the expensive computer code can be used.
- the spirit is to add gradient information to the assumed GP model. See slide no. 9
- Adjoint solvers. Originally developed for optimization, adjoint solvers are now finding more and more use in uncertainty quantification.
- indirect GEK
- Direct GEK (through prior covariance matrix)
- Direct GEK (through observation matrix)
- Gradient-enhanced kriging for high-dimensional problems (Indirect method)

84 Simulation metamodeling with Bayesian networks

- As a an exploratory metamodelling tool for supporting simulation studies conducted with stochastic simulation models containing multiple inputs and outputs.
- Bayesian network metamodels combine simulation data with available expert knowledge into a non-parametric description of the joint probability distribution of discrete random variables representing simulation inputs and outputs.
- focus on stochastic, e.g., discrete event, simulation where systems with internal uncertainties and random factors.

So, we resort to simulation metamodeling:

- To avoid the repetition of the time-consuming steps in the simulation analyses, simulation metamodeling are used to represent the dependence between simulation inputs and outputs.
- The most commonly used metamodels are input-output mappings that project the values of the simulation inputs to the expected value of a simulation output. (**this is what we have learnt!**)
- tradition multile input and multiple output (MIMO) metamodels, such as regression models or neural networks, are based on constructing separate models.
  - expected values and variances of simulation outputs
  - percentiles of an output's distribution
- Meanwhile, simulation metamodeling and statistical input modeling dealing with input uncertainty are often considered as separate lines of analyses.
- The existing MIMO metamodels do not enable the analysis of the joint probability distribution of the simulation outputs without limiting assumptions about its functional form.
- this limitation can be overcome by using Bayesian networks (BNs)
  - as an exploratory metamodeling tool for supporting studies conducted with MIMO metamodeling models.

85 Bayesian Network (BN)

- A BN is a graphical and numerical model that represents a set of random variables and their dependencies.
- More specifically, it is a directed acyclic graph that consists of a set of chance nodes representing random variables and their interconnecting arcs.
- The arcs and the conditional probabilities of the nodes imply dependencies between the variables.

- As BNs can only deal with continuous variables in a limited manner, random variables are assumed to be discrete.

86 Exploratory use of BN metamodel

- In a BN metamodel, the random variables are associated with simulation inputs and outputs.
- The structure of the BN metamodel as well as the probabilities required are obtained by combining simulation data with expert knowledge.
- The exploratory use of a BN metamodel enables flexible and efficient analysis for studying the dependencies among the inputs and the outputs, input uncertainty, and inverse reasoning.

87 Monte Carlo Sampling – Brief Review

- Monte Carlo methods, or Monte Carlo experiments, are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.
- The underlying concept is to use randomness to solve problems that might be deterministic in principle.
- They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to use other approaches.
- In principle, Monte Carlo methods can be used to solve any problem having a probabilistic interpretation.
- By the law of large numbers, integrals described by the expected value of some random variable can be approximated by taking the empirical mean (a.k.a. the sample mean) of independent samples of the variable.
- Monte Carlo methods are mainly used in three problem classes, optimization, numerical integration, and *generating draws from a probability distribution.*
- This benefit of MC sampling is most pronounced when random samples are easy to draw.

- We could have given more discussion on MC. But we went with Importance Sampling directly. Direct sampling is hard!

88 Markov Chain Monte Carlo (MCMC)

- MCMC is a class of algorithms for sampling from a probability distribution.
- Markov chain of samples' equilibrium distribution is the target distribution.
- Each Markov step is sampling the target. distribution, more steps, more samples, more accurate.
- **Markov screening**: state $t+1$ is only dependent on state at t.

- The central idea is to design a judicious Markov chain model with a prescribed stationary probability distribution.
- The Markov chain property of MCMC is the idea that the random samples are generated by a special sequential process.
- Each random sample is used as a stepping stone to generate the next random sample.
- A special property of the chain is that, while each new sample depends on the one before it, new samples do not depend on any samples before the previous one (this is the "Markov" property).
- MCMC is particularly useful in Bayesian inference because of the focus on posterior distributions which are often difficult to work with via analytic examination.
- In rare event sampling, they are also used for generating samples that gradually populate the rare failure region.

89 Model Calibration and Validation

- The credibility of the use of computational models for real-world applications is a growing concern.
  - There is an increasing interest in improving and certifying the credibility of computational models.
- The presence of unknown input variables causes credibility concerns in computational model predictions.
- Model calibration can improve the credibility of a computational model by estimating the unknown input variables,
  - if sufficient identifiability exists for the unknown input variables.
- To check the credibility of a computational model, model validation determines the degree to which a model is an accurate representation of the real phenomenon.
- Model validation can be executed at the completion of model calibration to check the predictive robustness of the calibrated model.
- model validation not only assesses the accuracy of a computational model but also helps the process of improving the model based on the validation results.
- Let us revisit HW2
  - Cubic Interpreting Spline Modeling (HW2 Prob. 1)
  - Individualized State Transition Modeling (HW2 Prob. 3)

90 Statistical Approaches

- Deterministic approaches have been traditionally adopted.
- However, it is important to consider various uncertainties and know how to improve and certify the credibility of computational models.

- For this reason, statistical approaches have received significant attention
  - Deterministic approaches may incorrectly validate a computational model;
  - they may significantly degrade the predictive capability of the computational model.
- Conducting model calibration and validation in a statistical sense is not easy,
  - how to efficiently and accurately conduct uncertainty quantification?
  - how to reduce the degree of uncertainty in the epistemic variables? (***)
  - how to statistically check the validity of a model?
- Statistical approaches are beneficial because they attempt to enhance the model's predictive capability by thoroughly addressing uncertainty issues that arise in the experiments and computational models.
- Understand the nature of uncertainties is thus crucial to statistical model calibration and validation.

91 Sources of Uncertainty

- Three categories:
  - physical
  - modeling
  - statistical
- Physical uncertainty arises from the inherent variability in physical quantities.
- Modeling uncertainty comes from inadequate or erroneous physics models and their numerical implementation and solution.

- Statistical uncertainty arises from a lack of data associated with uncertainties. solution
- In principle, the existence of these uncertainties in engineering systems can be either recognized (accounted for), unrecognized (unaccounted for), or a combination of both.
- From the perspective of the uncertainty structure, three types of problems in statistical model calibration and validation:
  - forward
  - inverse
  - validation

92 Overview of a probabilistic description of system responses

An observed response can be expressed in a probabilistic form as

$$\mathbf{Y}_{\text{obs}}(\mathbf{X}, \epsilon) = \mathbf{Y}(\mathbf{X}) + \epsilon$$

- $\mathbf{Y}_{\text{obs}}$: all observed system responses from experiments
- $\mathbf{Y}$: true system responses
- $\mathbf{X}$: system variables subject to uncertainty
- ϵ: a measurement error, which can have both random and systematic components

Note that physical uncertainties ($\mathbf{X}$).

To compute uncertainty in predicted responses ($\mathbf{Y}_{\text{pre}}$), the uncertainty quantification is performed.

Nonetheless, computational models suffer from the effect of model form uncertainty (e), which embodies the errors from improper assumptions and discretization related numerical solution convergence errors.

Using the model form uncertainty, a predicted response model can be presented in a probabilistic form as

$$\mathbf{Y}_{\text{pre}}(\mathbf{X}, e) = \mathbf{Y}(\mathbf{X}) + e$$

where e represents the model form error and $\mathbf{Y}_{\text{pre}}$ represents all predicted system responses from simulations.

When characterizing physical and modeling uncertainties, statistical uncertainty arises when related data are insufficient.

- Aleatory uncertainty: natural randomness in a process.
- For discrete variables, the randomness is parameterized by the probability of each possible value.
- For continuous variables, the randomness is paratermized by the probability density function.
- Epistemic uncertainty: scientific uncertainty in the model of the process. It is due to limited data and knowledge.
- The epistemic uncertainty is characterized by alternative models.
- For discrete random variables, the epistemic uncertainty is modelled by alternative probability distributions
- For continuous random variables, the epistemic uncertainty is modelled by alternative probability density functions.
- In addition, there is epistemic uncertainty in parameters that are not random but have only a single correct (but unknown) value.
- see the supplemental material for a detailed discussion on the two types of uncertainty.

A certain level of data sufficiency can effectively eliminate any epistemic uncertainty ($\hat{\mathbf{X}}_{\text{e}}$) associated with the characterization of the aleatory uncertainty ($\mathbf{X}_{\text{a}}$).

Given the sources of uncertainty (i.e., physical, modeling, and statistical) in a computational model, uncertainty structures must be well understood to properly formulate statistical model calibration and validation problems.

93 Inverse Problem

$$\mathbf{Y}_{\text{pre}}(\hat{\mathbf{X}}_{\text{e}}, \mathbf{X}_{\text{a}}) = \mathbf{Y}_{\text{obs}}$$
$$[\hat{\mathbf{X}}_{\text{e}}, \mathbf{X}_{\text{a}}] = \mathbf{Y}_{\text{pre}}^{-1}(\mathbf{Y}_{\text{obs}})$$

Formulating the mathematical expression of the inverse problem is based on two assumptions.

- Experiments are presumed to be conducted with care, thereby the measurement error ϵ is negligible.
- $\mathbf{Y}_{\text{obs}}$ can be considered as a reference, and then the predicted system response ($\mathbf{Y}_{\text{pre}}$) with calibrated epistemic variables ($\hat{\mathbf{X}}_{\text{e}}$) should be equivalent to the observed system response ($\mathbf{Y}_{\text{obs}}$).

94 Contemporary issues in the inverse problem

The objective of the inverse problem is to statistically calibrate epistemic variables ($\mathbf{X}_{\text{e}}$) of a computational model. What are the issues?

- How to solve the implicit inverse problem?
- How to calibrate epistemic variables given a dearth of data?
- How to calibrate multiple epistemic variables?

First, the inverse problem can be solved by taking the inverse of the function ($\mathbf{Y}_{\text{pre}}$). However, for most computational models that emulate the behavior of engineered systems, it is infeasible to obtain a closed form (explicit form) of the inverse function ($\mathbf{Y}_{\text{pre}}^{-1}$).

Two approaches:

- optimization-based model calibration
- Bayesian-based model calibration

95 Formulation of optimization-based model calibration

The objective of optimization-based model calibration is to inversely estimate epistemic variables so that the prediction is consistent with the observation data.

Two ways of achieving this goal:

- maximize the agreement
- minimize the disagreement between the two probability distributions that were found from the computational prediction and experimental observation

So the mathematical formulation is
$\max_{\mathbf{X}_\mathrm{e}}$ or $\min_{\mathbf{X}_\mathrm{e}}$ $f(\mathbf{Y}_\mathrm{obs}, \mathbf{Y}_{pre}(\mathbf{X}_\mathrm{a}, \mathbf{X}_\mathrm{e}))$

- If the degree of uncertainty due to measurement error (σ) is negligible, the PDF of the experimental observation ($\mathbf{Y}_\mathrm{obs}$) is considered to be the true system response ($\mathbf{Y}$).
- For computational prediction, all recognized (accounted for) uncertainty sources are incorporated as the input ($\mathbf{X}$) to the computational model.
- In the presence of epistemic variables ($\mathbf{X}_\mathrm{e}$), the computational prediction ($\mathbf{Y}_\mathrm{pre}$) has an error.
- Then an objective function is formulated with the meaning of agreement or disagreement between true responses ($\mathbf{Y}_\mathrm{obs}$) and incorrect ones ($\mathbf{Y}_\mathrm{pre}$).
- By maximizing agreement or minimizing disagreement on the objective function, the epistemic variables (X_e) are calibrated.

96 Calibration metric: objective function for the optimization problem

- Establishing a relevant objective function (f) is the key for success of calibration using optimization techniques

- There are a substantial number of similarity or dissimilarity measures encountered in many different fields, such as pattern classification and clustering.
- Two issues:
  - a calibration metric should have is that the function should be globally convex or concave
  - It is also important to examine how each calibration metric deals with statistical uncertainty due to a lack of data
- In many cases, we focus on the latter and use least square error. **Loss function!**
- A bit more on whether the calibrated value is the ultimate answer.
  - not really especially when lack of observational data
  - However, an important element in solving an inverse problem is that calibrated results with *sufficient* data should converge.

97 Quantifying the disagreement

- minimize the weighted sum of the *spatial distance* between the statistical moments (e.g., mean, standard deviation, covariance) of observation data and that of prediction data.
- However, using only a few statistical parameters (e.g., moments) is difficult to fully describing the PDF in general cases
- In many practical settings, it can be erroneous to assume the system responses follow a particular PDF.
- Various categories of measures that are applicable to compare two PDFs.

98 The likelihood function – Briefly Mention

Quantify the agreemet/similarity

- The most common measure used to quantify the agreement between two probability distributions
- The log-likelihood is commonly incorporated; its probability is estimated in the exponential scale.
- To evaluate the likelihood function, an assumption needs to be made on the type of PDF (p_{pre}) for computational prediction ($\mathbf{Y}_{\text{pre}}$).
- However, it is advantageous that uncertainty characterization is not required for experimental observation ($\mathbf{Y}_{\text{obs}}$).
- Instead all frequentist information on discrete points are used.

99 Measures applicable to compare two PDFs

A probability distribution of any shape can be facilitated and directly compared, which results in a low computational cost for evaluation of the likelihood function.

1. the L_p Minkowski family, developed on the basis of Euclidean distance, e.g., Euclidean L_2 (straightforward)
2. the L_1 family; facilitates the absolute difference
3. the intersection family; incorporates the inner product
4. the squared-chord family; based on the sum of geometric means
5. the squared L_1 family, based on the squared Euclidean distance
6. the Shannon's entropy family, developed from the relative entropy, Kullback-Leibler
7. combinations; utilizes multiples ideas or measures

For more information, see the full text of the paper **Cha (2007)** I place in the supplemental materials folder.

100 Lingering Issues

- Measures based on Euclidean distance are straightforward.
- However, it can be computationally extremely challenging to perform nonlinear optimization on blackbox simulator.
- Meanwhile, a hindrance of this measure is the need to characterize the PDF ($p_{\text{obs}}(y_i)$) of the experiments.
- The process of characterizing the uncertainty in experimental data may lead to statistical uncertainty, especially with a dearth of data.
- An accurate calibration using the optimization-based model calibration highly depends on both the quantity and quality of the given experimental data.
- Also, if the assumption on the distribution type of unknown input variables ($\mathbf{X}_{\text{a}}$) is wrong, it may lead to inaccurate calibrated results.

101 Bayesian-based optimization calibration

- On crucial issue in the inverse problem is how to estimate the value of calibration variables with limited observation data.
- The major benefit of the Bayesian approach is its ability to incorporate prior information.
- Optimization-based model calibration is based on a frequentist approach, which exclusively relies on the sample data to estimate calibration variables.
  - For optimization-based model calibration, a small number of experimental data can lead to incorrect estimation of the calibration variables
- The Bayesian approach utilizes the prior information in conjunction with newly available data to obtain the posterior knowledge for calibration variables

  - Bayesian-based model calibration is capable of continuously updating the prior information with evolving experimental data to obtain the posterior information

- Using prior knowledge of uncertainties, the Bayesian approach has strength in calibration compared to frequentist probability.

- Later, will talk about prior knowledge/prior distribution; computational challenges.

102 Bayesian Inference

Bayesian inference updates the probability of unknown parameters (calibration variables) as more observational data become available. The prior information of the unknown parameters vector (θ) is given in the form of join PDF ($p_{\mathbf{\Theta}}(\theta)$), and the experimental data ($\mathbf{y}$) are given with variability.

The posterior distribution ($p_{\mathbf{\Theta}|\mathbf{Y}}(\theta|\mathbf{y})$) of the parameter ($\theta$) can then be expressed as

$$p_{\mathbf{\Theta}|\mathbf{Y}}(\theta|\mathbf{y}) \propto p_{\mathbf{Y}|\mathbf{\Theta}}(\mathbf{y}|\theta)p_{\mathbf{\Theta}}(\theta)$$

where $p_{\mathbf{Y}|\mathbf{\Theta}}(\mathbf{y}|\theta)$ is the likelihood function that elucidates the probability of observing data ($\mathbf{y}$), given parmeter (θ).

Now let us introduce/revisit Baye's rule!
Next look at the application of Bayesian updating.

- The Bayesian updating above can be applicable when the values of parameters (θ) are observable.

- In the case of model calibration, the epistemic variables ($\mathbf{X}_e$), which replace parameter vector (θ) above, of the model are not directly observable; however, the model responses ($\mathbf{Y}_{\text{obs}}$) are.

- The Bayesian calibration requires the relationship between the model and experimental observations.

- In statistical sense, the inverse problem estimates the statistical parameters of the calibration variables ($\Theta_{\text{X}_\text{e}}$).

For Bayesian-based model calibration, the Bayesian formulation can be expressed as

$$p_{\mathbf{\Theta}_{\mathrm{e}}|\mathbf{Y}}(\theta_{\mathrm{e}}|\mathbf{Y}=\mathbf{Y}_{\mathrm{obs}}) \propto p_{\mathbf{Y}|\mathbf{\Theta}_{\mathrm{e}}}(\mathbf{Y}=\mathbf{Y}_{\mathrm{obs}}|\theta_{\mathrm{e}})p_{\mathbf{\Theta}_{\mathrm{e}}}(\theta_{\mathrm{e}})$$

where the posterior distribution $p_{\mathbf{\Theta}_{\mathrm{X_e}}|\mathbf{Y}}(\theta_{\mathrm{X_e}}|\mathbf{Y}=\mathbf{Y}_{\mathrm{obs}})$ of statistical parameters (θ_{e}) of calibration variables ($\mathrm{X_e}$) is proportional to the likelihood $p_{\mathbf{Y}|\mathbf{\Theta}_{\mathrm{e}}}(\mathbf{Y}=\mathbf{Y}_{\mathrm{obs}}|\theta_{\mathrm{e}})$ times the prior distribution θ_{e}. In Bayesian statistics, a state of knowledge or belief of an unknown parameter is expressed in terms of PDF ($p(\cdot)$).

What does the above mean?

- If the value of an unknown parameter is well known, the distribution will have small uncertainty, and vice versa.
- Bayesian-based model calibration uses Bayesian inference to reduce epistemic uncertainty in calibration variables.
- As experimental data set ($\mathbf{Y}_{\mathrm{obs}}$) becomes larger, the variation in the posterior distribution becomes narrower, which means the degree of uncertainty decreases.

103 An extended model for Bayesian calibration

Include not only calibration variables but also a discrepancy function, which represents the effect of model-form error as well as numerical error.

In this framework, the relationship between the prediction model and observation can be represented by

$$\mathbf{Y}_{\mathrm{obs}}(\mathbf{X}) = \mathbf{Y}_{\mathrm{pre}}(\mathbf{X};\theta_{\mathrm{e}}) + \delta(\mathbf{X}) + \epsilon$$

where $\mathbf{X}$ is the vector of inputs; θ_{e} is a set of epistemic variables (unknown model parameters); $\delta(\mathbf{X})$ is the model error (discrepancy function), which is defined as the difference between model prediction and reality; and ϵ is the measurement error, which is usually assumed to be a Gaussian distribution ($\epsilon \sim N(0, \sigma_\epsilon^2)$).

- Using the experimental data to update the posterior distribution.

- MCMC methods are commonly used.
- surrogate methods have been attemped.
- Using this approach emphasize that the advantage of introducing the discrepancy function is that it can consider the possible existence of unrecognized (unaccounted for) uncertainty sources.
- This approach can deal with missing physics and other inaccuracies of the computer model or experimental error, while updating calibration variables.