Bayesian
Optimal Experimental Design (OED)

OED: Definitions

- **Given: model and the model parameters**
- **Prior PDF of model parameters: quantifies the uncertainty in the model parameters (describes the state of knowledge of the parameters) before an experiments takes place**
- **Prediction error: The discrepancy between the model predictions and measurements. Usually assumed Gaussian.**
- **Posterior PDF: quantifies the uncertainty in the model parameters given data from an experiment**
- **Design variables: These are the experimental variables to be optimized in the OED. May include sensor types, number of sensors, location of sensors, excitation characteristics (frequency content and amplitude), etc. Change in the values of design variables affect the quality of data and the information contained in the data. As a result, it affects the posterior PDF of the model parameters.**
- **OED Objective: Collect data from the system so that they are most informative for identifying the model parameters and predicting output QoI**

OED Objectives

- **Collect data from the system so that they are most informative for identifying the model parameters and/or predicting output QoI**
- **We will address the problem estimating the model parameters. Similar analysis will hold for the problem of predicting output QoI.**
- **OBJECTIVE: Select the values of the design variables such that the data are most informative for the model parameters**
- **The MUTUAL INFORMATION between two uncertain parameters measure the amount of information these parameters share. Equivalently, the mutual information measures the reduction of uncertainty in one variable after having observed the other. Large reduction of uncertainty corresponds to large mutual information.**
- **MUTUAL INFORMATION:**

$$I(\theta; y) = \int_{\Theta} \int_{Y} p(\theta, y) \log \frac{p(\theta, y)}{p(\theta)\, p(y)} \, d\theta \, dy$$

OED Objectives

$$I(\theta; y) = \int_\Theta \int_Y p(\theta, y) \log \frac{p(\theta, y)}{p(\theta)\, p(y)} \, d\theta \, dy$$

- **This can be seen as the relative entropy between the joint PDF and the independent approximation of it using the product of the marginals.**
- **It measures how much the two PDF differ, essentially measuring how much dependence (shared information) exist between them through their joint PDF.**
- **In experimental design, the mutual information is conditioned on the design variables. The resulting CONDITIONAL MUTUAL INFORMATION, conditioned on the values of the design variables, is used to formulate the OED problem.**
- **OBJECTIVE: Find the values of the design variables that maximize the information the data carries about the model parameters. Equivalent to maximizing conditional mutual information wrt design variables**

$$U(d) = I(\theta; y|d) = \int_\Theta \int_Y p(\theta, y|d) \log \frac{p(\theta, y|d)}{p(\theta|d)\, p(y|d)} \, d\theta \, dy$$

Mutual Information, Relative Entropy, Entropy

- **In information theory, these quantities have properties which relate them with each other. The mutual information can be written as relative entropy (Kullback-Leibler (KL) divergence) or the difference between two entropies.**
- **Lets derive the relations**
- **Expanding the joint PDF as**

$$p(\theta, y|d) = p(y|d)\, p(\theta|y, d)$$

and substituting into the conditional mutual information

$$U(d) = I(\theta; y|d) = \int_\Theta \int_Y p(\theta, y|d) \log \frac{p(\theta, y|d)}{p(\theta|d)\, p(y|d)}\, d\theta\, dy$$

one derives

Mutual Information, Relative Entropy, Entropy

$$U(d) = I(\theta; y|d) = \int_\Theta \int_Y p(y|d)\, p(\theta|y,d) \log \frac{p(y|d)\, p(\theta|y,d)}{p(\theta|d)\, p(y|d)}\, d\theta\, dy$$

$$= \int_Y p(y|d) \int_\Theta p(\theta|y,d) \log \frac{p(\theta|y,d)}{p(\theta)}\, d\theta\, dy$$

$$= E_{y|d}[D_{KL}(p(\theta|y,d)\,||\,p(\theta))]$$

- **The conditional mutual information is expressed as the expected KL-divergence (information gain) from the posterior PDF to the prior PDF, where the expectation is taken over all possible outcomes (data) of the particular experiment with design variables *d***
- **The posterior PDF of the model parameters depend on the data so that the expectation gives an average of the information gain about the model parameters based on all data that can result from experiment *d*, as described by the likelihood function.**
- **This is same as the expected Utility Function proposed by Lindley (1956) to carry out the OED problem based on a decision-theoretic framework.**

Expected KL-Divergence

- **The objective of the experiment is to learn as much as possible about the model parameters**
- **The Kullback-Leibler (KL) Divergence, which is a scalar measure of the "distance" between the posterior and prior, is used to represent the information gain due to data.**
- **The objective of the experiment is to maximize the KL-Divergence with respect to the design variables of the experiment.**
- **The posterior, and therefore the KL-Divergence, need the data in order to be computed. But data do not exist before the experiment.**
- **Therefore, the best thing we can do is to average the KL-Divergence over all the possible data, based on the likelihood and prior PDFs of the model.**

Sampling Estimate of the Utility Function

- **Apply Bayes rule and substitute the posterior PDF in**

$$U(d) = I(\theta; y|d) = \int_\Theta \int_Y p(y|d)\, p(\theta|y, d) \log \frac{p(y|d)\, p(\theta|y, d)}{p(\theta|d)\, p(y|d)}\, d\theta\, dy$$

- **One derives a form in terms of the likelihood and prior, that is suitable for approximation with Monte Carlo sampling**

$$\begin{aligned}
U(d) = I(\theta; y|d) &= \int_Y p(y|d) \int_\Theta p(\theta|y, d) \log \frac{p(\theta|y, d)}{p(\theta)}\, d\theta\, dy \\
&= \int_Y p(y|d) \int_\Theta \frac{p(y|\theta, d)\, p(\theta)}{p(y|d)} \log \frac{p(y|\theta, d)\, p(\theta)}{p(y|d)\, p(\theta)}\, d\theta\, dy \\
&= \int_Y \int_\Theta p(y|\theta, d)\, p(\theta) \log \frac{p(y|\theta, d)}{p(y|d)}\, d\theta\, dy \\
&= \int_Y \int_\Theta p(\theta, y|d) \left[\log p(y|\theta, d) - \log p(y|d)\right] d\theta\, dy
\end{aligned}$$

Sampling Estimate of the Utility Function

- **Apply Bayes rule and substitute the posterior PDF in**

$$U(d) = \int_Y \int_\Theta p(\theta, y|d)\,[\log p(y|\theta, d) - \log p(y|d)]\, d\theta\, dy$$

- **Samples**

$$\theta^i, y^i \mid d \sim p(\theta, y \mid d), \qquad i = 1, \boxed{?} N$$

from the joint PDF

$$p(\theta, y \mid d) = p(y \mid \theta, d)\, p(\theta \mid d)$$

can be used to approximate the integrals with a Monte Carlo sum

- **Assuming that** $p(\theta \mid d) = p(\theta)$
- **the model parameter samples are generated from the prior**

$$\theta^i \sim p(\theta), \qquad i = 1, \boxed{?} N$$

- **While the data samples are generated from the likelihood**

$$y^i \sim N(g(\theta^i, d), \Sigma_e(\theta^i, d))$$

Sampling Estimate of the Utility Function

- **Apply Bayes rule and substitute the posterior PDF in**

$$U(d) = \int_Y \int_\Theta p(\theta, y|d) \left[\log p(y|\theta, d) - \log p(y|d)\right] d\theta \, dy$$

- **The evidence is approximated with another Monte Carlo sum**

$$p(y|d) = \int_\Theta p(y, \theta|d) \, d\theta = \int_\Theta p(y|\theta, d) \, p(\theta) \, d\theta$$

- **The Monte Carlo estimator is**

$$U(d) \approx \frac{1}{N} \sum_{i=1}^{N} \left\{ \log p(y^i|\theta^i, d) - \log \left\{ \frac{1}{M} \sum_{j=1}^{M} p(y^i|\theta^{i,j}, d) \right\} \right\}$$

Sampling Estimate of the Utility Function

$$U(d) \approx \frac{1}{N}\sum_{i=1}^{N}\left\{\log p(y^i|\theta^i,d) - \log\left\{\frac{1}{M}\sum_{j=1}^{M} p(y^i|\theta^{i,j},d)\right\}\right\}$$

- **The method requires N samples for the outer integral and M samples for the inner integral.**
- **Thus, there is N+NM likelihood function evaluations involved for different values of the model parameters.**
- **Since each likelihood function evaluation involves one model run, there is a number of N+NM model runs.**
- **Note that we need to keep a sufficiently large number of outer and inner samples (N,M) in order to reduce the bias and the variance of the Monte Carlo estimate**
- **This creates an enormous computational burden, especially when a model run is computationally expensive to perform.**

Simplification of Sampling Estimate

- **The utility function can be written in the form**

$$U(d) = \int_Y \int_\Theta p(y|\theta,d)\, p(\theta) \log p(y|\theta,d)\, d\theta\, dy - \int_Y \int_\Theta p(y|\theta,d)\, p(\theta) \log p(y|d)\, d\theta\, dy$$

- **The first term can be written as**

$$\int_Y \int_\Theta p(y|\theta,d)\, p(\theta) \log p(y|\theta,d)\, d\theta\, dy$$

$$= \int_\Theta p(\theta) \left\{ \int_Y p(y|\theta,d)\, \log p(y|\theta,d)\, dy \right\} d\theta$$

$$= - \int_\Theta p(\theta) \left\{ - \int_Y p(y|\theta,d)\, \log p(y|\theta,d)\, dy \right\} d\theta$$

$$= - \int_\Theta p(\theta) [H(p(y|\theta,d))]\, d\theta$$

$$= -E_{p(\theta)}[H(p(y|\theta,d))]$$

Simplification of Sampling Estimate

- **which is the minus of the expectation of the likelihood entropy over the prior.**
- **The likelihood function depends on the prediction error equation**

$$y = g(\theta, d) + e$$

- **where the prediction error is almost always modeled as Gaussian with zero mean and given covariance matrix. So the PDF of the data is given by a normal distribution**

$$p(y|\theta, d) = N(g(\theta, d), \Sigma_e(\theta, d))$$

- **The information entropy of the Gaussian likelihood depends only of the covariance matrix, given by (assume covariance does not depend on model parameters**

$$H(p(y|\theta, d)) = \frac{1}{2}\log\{(2\pi e)^k |\Sigma_e(d)|\}$$

- **Substituting, the utility function becomes**

Simplification of Sampling Estimate

$$U(d) = -\frac{1}{2}\log\{(2\pi e)^k|\Sigma_e(d)|\} - \int_Y \int_\Theta p(y|\theta,d)\,p(\theta)\log p(y|d)\,d\theta\,dy$$

- **and the Monte Carlo estimate simplifies to**

$$U(d) \approx -\frac{1}{2}\log\{(2\pi e)^k|\Sigma_e(d)|\} - \frac{1}{N}\sum_{i=1}^{N}\log\left\{\frac{1}{M}\sum_{j=1}^{M} p(y^i|\theta^{i,j},d)\right\}$$

- **Choosing the M inner model parameter samples to be exactly the same as the outer model parameter samples, the MC sampling estimates takes the form**

$$U(d) \approx -\frac{1}{2}\log\{(2\pi e)^k|\Sigma_e(d)|\} - \frac{1}{N}\sum_{i=1}^{N}\log\left\{\frac{1}{N}\sum_{j=1}^{N} p(y^i|\theta^{j},d)\right\}$$

- **which constitutes an improved version since one less integral is approximated by Monte Carlo estimate.**

Optimization Problem

- **Find design variables that maximize the utility function or minimize minus the utility function**

$$U(d) \approx -\frac{1}{2}\log\{(2\pi e)^k|\Sigma_e(d)|\} - \frac{1}{N}\sum_{i=1}^{N}\log\left\{\frac{1}{M}\sum_{j=1}^{M}p(y^i|\theta^{i,j},d)\right\}$$

- **That is the optimal design variables are obtained from**

$$d_{opt} = \max_d U(d) = \min_d[-U(d)]$$

- **For each value of the design variables *d* the utility function is obtained using sampling. Resampling for each *d* makes the utility function to fluctuate as a function of *d* of . So an optimization algorithm is needed that is effective to such fluctuation.**

- **CMA-ES is an effective algorithm that optimizes stochastic (fluctuating) objective functions.**

Optimization Problem

- **However, resampling means that for each design value N+NM expensive model runs have to be performed.**
- **To reduce the excessive computational burden, the model parameter and the data samples are kept the same when computing the utility function for all values of the design variables.**
- **Keeping the samples the same also guarantees a smooth variation of the utility function wrt the design variables. The fluctuations due to the sampling estimate is no longer manifested in the values of the utility function. This makes the optimization problem easier to carry out with available techniques.**
- **In OED, a number of local/global optima are manifested.**
- **Gradient-based optimization methods are trapped to a local optimum.**
- **CMA-ES is very efficient technique, with much higher chances to converge to the global optima.**

Optimization Problem – Implementation Issues

$$U(d) \approx \frac{1}{N} \sum_{i=1}^{N} \left\{ \log p(y^i | \theta^i, d) - \log \left\{ \frac{1}{M} \sum_{j=1}^{M} p(y^i | \theta^{i,j}, d) \right\} \right\}$$

- **Assume that the domain of the physical problem where our model is defined is continuum and that the design variables are continuous variables.**
- **Let the design variables be associated with the location of sensors in the continuum domain. The problem now is stated as to find the sensor locations so that the data are most informative for inferring the model parameter**
- **To efficiently compute the utility function for different values of design variables one can proceed as follows.**
- **Generate N+NM model parameter samples in the space of model parameter, i.e.**

$$\theta^{i} \sim p(\theta), \qquad i = 1, \ldots N$$

$$\theta^{i,j} \sim p(\theta), \qquad j = 1, \ldots M$$

Optimization Problem – Implementation Issues

$$U(d) \approx \frac{1}{N}\sum_{i=1}^{N}\left\{\log p(y^i|\theta^i, d) - \log\left\{\frac{1}{M}\sum_{j=1}^{M} p(y^i|\theta^{i,j}, d)\right\}\right\}$$

- **For each model parameter sample compute the output QoI at all possible positions of the domain. Usually the domain is discretized and the QoI is computed at the grid points. The values of the QoI at points different from the grid points are computed by interpolation within an element and the values of the QoI at the nodes of the element**

- **Given a value of the design variable *d*, compute**

$$g(\theta^i, d) \quad \text{and} \quad g(\theta^{i,j}, d)$$

using interpolation.

- **Then generate data samples by sampling from the Gaussian PDF**

$$y^i \sim N(g(\theta^i, d), \Sigma_e(\theta^i, d))$$

- **Then compute for each data sample and model parameter sample**

$$p(y^i|\theta^i, d) \quad \text{and} \quad p(y^i|\theta^{i,j}, d)$$

Homework Problems

Problem 1:

- (a) Write the algorithm for optimal experimental design using N+NM model parameter samples.

- (b) Implement the algorithm for optimizing the velocity sensors for estimating the location and strength of a vortex causing the flow. Assume that the sensors are equally spaced on a horizontal line. The design variables are the location of the first sensor in the x-y space and the distance between sensors. Solve the problem for different number n of velocity sensors placed in the flow. The output QoI is the velocity field caused by the vortex.

Problem 2:

- Write the algorithm for optimal experimental design using N model parameter samples.

- Implement the algorithm using the same physical system described in Problem 1.