SAND2015-10328C

Optimization and Uncertainty Quantification

Habib N. Najm

hnnajm@sandia.gov

Sandia National Laboratories
Livermore, CA, USA

Optimization Program Planning Workshop
SAMSI, RTP, NC
Nov 16, 2015

Outline

1. Introduction
2. Optimization Utility in UQ
   - Inverse UQ – Bayesian methods
     - Statistical Inverse Problem
     - Optimal Experimental Design
   - Forward UQ with Polynomial Chaos Expansions
     - Optimal Sampling
     - Compressive Sensing
     - PC Extrema Estimation
3. UQ Utility in Optimization
   - Optimization under Uncertainty

Introduction

- Explore connections between Optimization and Uncertainty Quantification
- Optimization problems in UQ
  - Inverse UQ
    - Bayesian methods
    - Statistical inverse problem
    - Experimental design
  - Forward UQ
    - Polynomial Chaos (PC) methods
    - Sampling ... quadrature, response surface fitting
- UQ problems in optimization
  - Forward model surrogate construction
  - Estimation of moments/probabilities in optimization under uncertainty

Statistical Inverse Problems

- Estimation of model parameters/inputs given (noisy) data on model output observables
  - with quantified uncertainty in inferred parameters
- Conventional deterministic context:
  - Model $y_m = f(x; \lambda)$, data y
  - Least squares fitting, minimizing residual $\|f(x; \lambda) - y\|$
  - Regularization with suitable norms
  - End result is x_{BestFit}
- Statistical Bayesian context:
  - Use Bayes rule to infer parameter λ
    - Combine prior information with learning from data
  - Information on λ is in terms of a posterior density conditioned on the data $p(\lambda|y)$

Bayes formula for Parameter Inference

- Data Model (fit model + noise model): $y = f(\lambda) * g(\epsilon)$
- Bayes Formula:

$$p(\lambda, y) = p(\lambda|y)p(y) = p(y|\lambda)p(\lambda)$$

$$\underbrace{p(\lambda|y)}_{\text{Posterior}} = \frac{\overbrace{p(y|\lambda)}^{\text{Likelihood}}\ \overbrace{p(\lambda)}^{\text{Prior}}}{\underbrace{p(y)}_{\text{Evidence}}}$$

- Prior: knowledge of λ prior to data
- Likelihood: forward model and measurement noise
- Posterior: combines information from prior and data
- Evidence: normalizing constant for present context

Exploring the Posterior

- Given any sample λ, the un-normalized posterior probability can be easily computed

$$p(\lambda|y) \propto p(y|\lambda)p(\lambda)$$

- Explore posterior w/ Markov Chain Monte Carlo (MCMC)
  - Metropolis-Hastings algorithm:
    - Random walk with proposal PDF & rejection rules
  - Computationally intensive, $\mathcal{O}(10^5)$ samples
  - Each sample: evaluation of the forward model
    - Surrogate models
- Evaluate moments/marginals from the MCMC statistics

Optimization and MCMC

- MCMC needs to get enough good λ-samples to describe the posterior density well
- Frequently, the focus is on a given mode/peak
  – although generally multimodal
- In order to get good samples of a particular peak, the random walk needs to be directed to the vicinity of the peak as efficiently as possible
- The structure of the proposal distribution and the random walk algorithm are crucial
- Generally, this is about climbing the posterior density towards its peak at the Maximum A-Posteriori (MAP) parameter value, employing a random walk
- Gradient & Hessian information is useful in this regard

Optimal Experimental Design – Stochastic Optimization

- Setup:
  - Choose experimental design x
  - Collect data y to estimate parameters θ
- Challenge:
  - Choose an **optimal** design x^* that maximizes the **expected information gain** from the experiment
- Bayesian formulation:

$$
\begin{aligned}
p(\theta|y,x) &= p(y|\theta,x)p(\theta|x) \,/\, p(y|x) \\
D(y,x) &= D_{\mathrm{KL}}(p(\cdot|y,x) \| p(\cdot|x)) \equiv \int p(\theta|y,x) \ln \frac{p(\theta|y,x)}{p(\theta|x)} \mathrm{d}\theta \\
U(x) &= \mathbf{E}_{p(\cdot|x)}[D(y,x)] \equiv \int D(y,x)p(y|x)\mathrm{d}y \\
x^* &= \operatorname*{argmax}_{x \in \mathcal{D}} U(x)
\end{aligned}
$$

- A stochastic optimization problem
  - noisy random-sampling estimation of integrals for $U(x)$

Probabilistic Forward UQ & Polynomial Chaos Representation of Random Variables

With $y = f(x)$, x a random variable, estimate the RV y

- Can describe a RV in terms of its
  - density, moments, characteristic function, or
  - as a function on a probability space
- Constraining the analysis to RVs with finite variance
  - $\Rightarrow$ Represent RV as a spectral expansion in terms of orthogonal functions of standard RVs
    - Polynomial Chaos Expansion
- Enables the use of available functional analysis methods for forward UQ

Polynomial Chaos Expansion (PCE)

- Model uncertain quantities as random variables (RVs)
- Given a *germ* $\boldsymbol{\xi}(\omega) = \{\xi_1, \cdots, \xi_n\}$ – a set of *i.i.d.* RVs
  - where $p(\boldsymbol{\xi})$ is uniquely determined by its moments

Any RV in $L^2(\Omega, \mathfrak{S}(\boldsymbol{\xi}), P)$ can be written as a PCE:

$$u(\boldsymbol{x}, t, \omega) = f(\boldsymbol{x}, t, \boldsymbol{\xi}) \simeq \sum_{k=0}^{P} u_k(\boldsymbol{x}, t) \Psi_k(\boldsymbol{\xi}(\omega))$$

- $u_k(\boldsymbol{x}, t)$ are mode strengths
- $\Psi_k()$ are multivariate functions orthogonal w.r.t. $p(\boldsymbol{\xi})$

With dimension n and order p: $$P + 1 = \frac{(n+p)!}{n!p!}$$

Orthogonality

By construction, the functions $\Psi_k()$ are orthogonal with respect to the density of ξ

$$u_k(\boldsymbol{x},t) \quad = \quad \frac{\langle u\Psi_k\rangle}{\langle\Psi_k^2\rangle} = \frac{1}{\langle\Psi_k^2\rangle}\int u(\boldsymbol{x},t;\lambda(\boldsymbol{\xi}))\Psi_k(\boldsymbol{\xi})\; p_{\boldsymbol{\xi}}(\boldsymbol{\xi})\; d\boldsymbol{\xi}$$

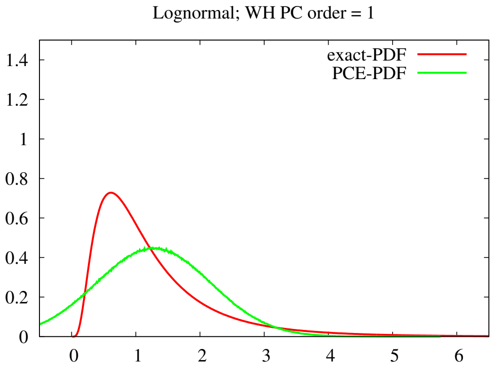
Examples:

- Hermite polynomials with Gaussian basis
- Legendre polynomials with Uniform basis, ...
- Global versus Local PC methods
  - Adaptive domain decomposition of the support of ξ

PC Illustration: WH PCE for a Lognormal RV

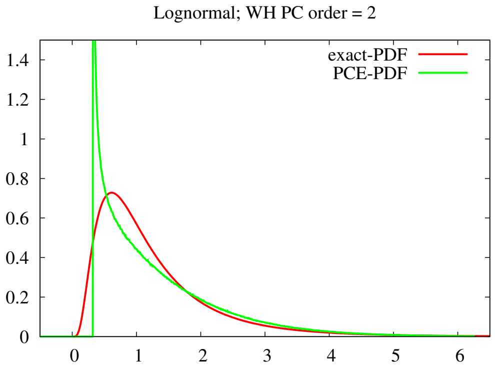
- Wiener-Hermite PCE constructed for a Lognormal RV
- PCE-sampled PDF superposed on true PDF
- Order = 1

$$u = \sum_{k=0}^{P} u_k \Psi_k(\xi)$$
$$= u_0 + u_1 \xi$$



PC Illustration: WH PCE for a Lognormal RV

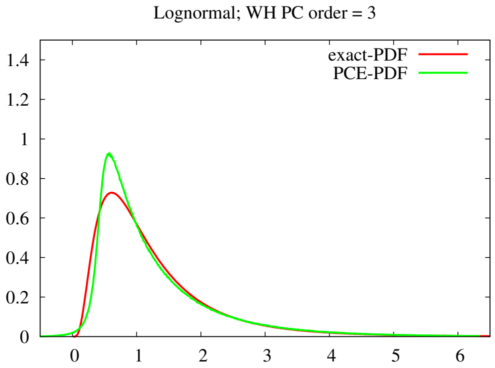
- Wiener-Hermite PCE constructed for a Lognormal RV
- PCE-sampled PDF superposed on true PDF
- Order = 2



$$
\begin{aligned}
u &= \sum_{k=0}^{P} u_k \Psi_k(\xi) \\
&= u_0 + u_1\xi + u_2(\xi^2 - 1)
\end{aligned}
$$

PC Illustration: WH PCE for a Lognormal RV

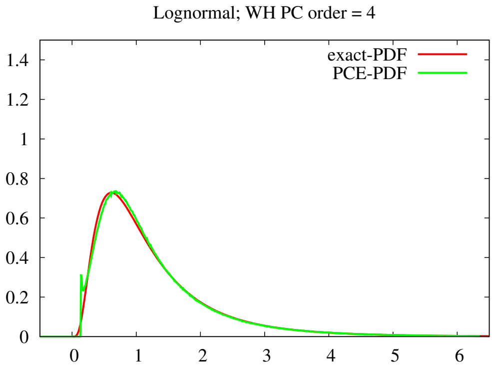
- Wiener-Hermite PCE constructed for a Lognormal RV
- PCE-sampled PDF superposed on true PDF
- Order = 3



$$u = \sum_{k=0}^{P} u_k \Psi_k(\xi) = u_0 + u_1\xi + u_2(\xi^2 - 1) + u_3(\xi^3 - 3\xi)$$

PC Illustration: WH PCE for a Lognormal RV

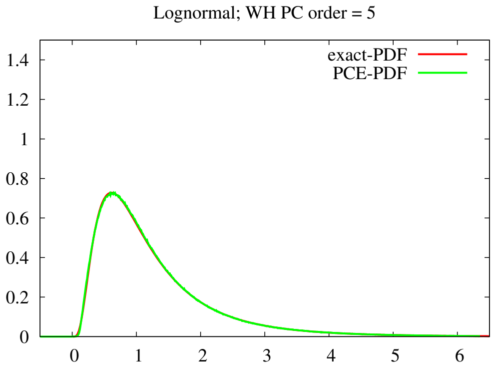
- Wiener-Hermite PCE constructed for a Lognormal RV
- PCE-sampled PDF superposed on true PDF
- Order = 4



$$
\begin{aligned}
u &= \sum_{k=0}^{P} u_k \Psi_k(\xi) \\
&= u_0 + u_1\xi + u_2(\xi^2 - 1) + u_3(\xi^3 - 3\xi) + u_4(\xi^4 - 6\xi^2 + 3)
\end{aligned}
$$

PC Illustration: WH PCE for a Lognormal RV

- Wiener-Hermite PCE constructed for a Lognormal RV
- PCE-sampled PDF superposed on true PDF
- Order = 5



$$
\begin{aligned}
u &= \sum_{k=0}^{P} u_k \Psi_k(\xi) \\
&= u_0 + u_1\xi + u_2(\xi^2 - 1) + u_3(\xi^3 - 3\xi) + u_4(\xi^4 - 6\xi^2 + 3) \\
&\quad + u_5(\xi^5 - 10\xi^3 + 15\xi)
\end{aligned}
$$

Essential Use of PC in UQ

Strategy:

- Represent model parameters/solution as random variables
- Construct PCEs for uncertain parameters
- Evaluate PCEs for model outputs

Advantages:

- Computational efficiency
- Utility
  - Moments: $\mathrm{E}(u) = u_0, \ \mathrm{var}(u) = \sum_{k=1}^{P} u_k^2 \langle \Psi_k^2 \rangle, \ \ldots$
  - Global Sensitivities – fractional variances, Sobol' indices
  - Surrogate for forward model

Requirement:

- RVs in L^2, *i.e.* with finite variance, on $(\Omega, \mathfrak{S}(\xi), P)$

Intrusive PC UQ: A direct *non-sampling* method

- Given model equations: $\mathcal{M}(u(\boldsymbol{x},t);\lambda) = 0$
- Express uncertain parameters/variables using PCEs

$$u = \sum_{k=0}^{P} u_k \Psi_k; \quad \lambda = \sum_{k=0}^{P} \lambda_k \Psi_k$$

- Substitute in model equations; apply Galerkin projection
- New set of equations: $\mathcal{G}(U(\boldsymbol{x},t),\Lambda) = 0$
  - with $U = [u_0,\ldots,u_P]^T$, $\Lambda = [\lambda_0,\ldots,\lambda_P]^T$
- Solving this <u>deterministic</u> system <u>once</u> provides the full specification of uncertain model ouputs

Non-intrusive PC UQ

- *Sampling*-based
- Relies on black-box utilization of the computational model
- Evaluate projection integrals *numerically*
- For any quantity of interest $\phi(\boldsymbol{x}, t; \lambda) = \sum_{k=0}^{P} \phi_k(\boldsymbol{x}, t)\Psi_k(\boldsymbol{\xi})$

$$\phi_k(\boldsymbol{x}, t) = \frac{1}{\langle \Psi_k^2 \rangle} \int \phi(\boldsymbol{x}, t; \lambda(\boldsymbol{\xi}))\, \Psi_k(\boldsymbol{\xi}) p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) d\boldsymbol{\xi}, \quad k = 0, \ldots, P$$

- Integrals can be evaluated using
  - A variety of (Quasi) Monte Carlo methods
    - Slow convergence; $\sim$ indep. of dimensionality
  - Quadrature/Sparse-Quadrature methods
    - Fast convergence; depends on dimensionality

Optimal Sparse Quadrature – forward UQ

Integration problem, with $x \in D \subset \mathbb{R}^N$:

$$I = \int_D f(x)\,\mathrm{d}x \approx \hat{I} = \sum_{j=1}^{M} w_j f(x_j)$$

- Optimization problem
  - Minimize number of (sparse) quadrature points, M
  - Optimize their locations and weights, $\{w_j, x_j\}_{j=1}^{M}$
  - For a requisite integration accuracy, $\|I - \hat{I}\| < \epsilon$
- Regular domains – hypercubes Sinsbeck & Nowak, IJUQ 2015
- Arbitrary domains Ryu & Boyd, Found. Comput. Math. 2015
  - By construction
  - $f(x)$ model failure – unrealistic conditions
  - $f(x)$ code failure – numerical stability / machine faults

Greedy Sampling Algorithms

- Find the optimal location for the next evaluation $f(x_k)$
  - given existing samples $x_j, j = 1, \ldots, k-1$
- Maximize expected reduction in error
  - given one additional sample/batch-of-samples

- Adaptive multilevel/hierarchical sparse quadrature
  - Selective evaluation of corner samples
- Non-isotropic sparse quadrature
  - Dimension-adaptive sampling

Interpolant/Regression Surrogates – Optimal Design

- PCE or other surrogate functions built via
  - Interpolation
  - Least-squares regression – noisy forward models
    - intrinsic noise
    - discretization errors
    - sample-averaging noise
    - sparse samples
- The optimal set of points – design
  - Minimize oscillations – particularly in Hi-D
  - Minimize cross-validation fit errors
- Recent work (Narayan *et al.*)
  - Leja sequences for optimal interpolation in Hi-D
  - Optimal random sampling of design points for weighted least-squares regression

PC coefficients via sparse regression

PCE:

$$y = f(x) \simeq \sum_{k=0}^{K-1} c_k \Psi_k(x)$$

with $x \in \mathbb{R}^n$, Ψ_k max order p, and $K = (p+n)!/p!/n!$

- N samples $(x_1, y_1), \ldots, (x_N, y_N)$
- Estimate K terms $c_0, \ldots, c_{K-1}$, s.t.

$$\min ||\boldsymbol{y} - \boldsymbol{Ac}||_2^2$$

where $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{c} \in \mathbb{R}^K$, $\boldsymbol{A}_{ik} = \Psi_k(x_i)$, $\boldsymbol{A} \in \mathbb{R}^{N\times K}$

With $N << K \Rightarrow$ under-determined

- Need some form of regularization

Regularization – Compressive Sensing (CS)

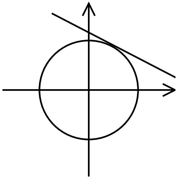
- ℓ_2-norm — Tikhonov regularization; Ridge regression:

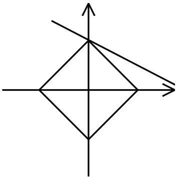
$$\min \; \{ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 + \|\boldsymbol{c}\|_2^2 \}$$

- ℓ_1-norm — Compressive Sensing; LASSO; basis pursuit

$$\begin{aligned} \min \; \{ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 + \|\boldsymbol{c}\|_1 \} & \\ \min \; \{ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 \} & \quad \text{subject to } \|\boldsymbol{c}\|_1 \leq \epsilon \\ \min \; \{ \|\boldsymbol{c}\|_1 \} & \quad \text{subject to } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 \leq \epsilon \end{aligned}$$

$\Rightarrow$ discovery of sparse signals





Bayesian Regression

- Bayes formula

$$p(\boldsymbol{c}|D) \propto p(D|\boldsymbol{c})\pi(\boldsymbol{c})$$

- Bayesian regression: prior as a regularizer, *e.g.*
    - Log Likelihood $\Leftrightarrow$ $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2$
    - Log Prior $\Leftrightarrow$ $\|\boldsymbol{c}\|_p^p$
- Laplace sparsity priors $\pi(c_k|\alpha) = \frac{1}{2\alpha} e^{-|c_k|/\alpha}$
- LASSO (Tibshirani 1996) ... formally:

$$\min \{\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 + \lambda \|\boldsymbol{c}\|_1\}$$

  Solution $\sim$ the posterior mode of $\boldsymbol{c}$ in the Bayesian model

$$y \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{c}, I_N), \qquad c_k \sim \frac{1}{2\alpha} e^{-|c_k|/\alpha}$$

- Bayesian LASSO (Park & Casella 2008)
- Bayesian compressive sensing (Ji 2008)

PC Extrema Estimation – Global Optimization

- Often it is important to establish PCE positivity

$$
\begin{aligned}
u_{\mathrm{PC}}(\boldsymbol{\xi}) &\equiv \sum_{k=1}^{P} u_k \Psi_k(\boldsymbol{\xi}) \\
u_{\min} &\equiv \min_{\boldsymbol{\xi}\in\Xi} u_{\mathrm{PC}}(\boldsymbol{\xi}) > 0
\end{aligned}
$$

- A global optimization problem
- Nonlinear
- High-dimensional

UQ Utility in Optimization

Deterministic optimization problems

– Model surrogates constructed using forward UQ

- A wide range of UQ methods for efficient surrogate construction in hi-D
- Surrogates can be built over deterministic spaces employing uniform RVs
- Readily available surrogate gradient/hessian information

Optimization under uncertainty

- Stochastic optimization
- Distributionally Robust optimization
- Robust optimization

Stochastic Optimization

e.g. stochastic objective

$$\begin{aligned} \underset{x \in \mathcal{X}}{\text{minimize}} \quad & \mathbf{E}_{p_\Theta}[u(x,\theta)] \\ \text{subject to} \quad & c(x) > 0 \end{aligned}$$

or penalize variability, and chance constraint

$$\begin{aligned} \underset{x \in \mathcal{X}}{\text{minimize}} \quad & \mathbf{E}_{p_\Theta}[u(x,\theta) - \gamma \mathrm{Var}(u(x,\theta))] \\ \text{subject to} \quad & \mathrm{P}[c(x,\theta) > 0] > 1 - \alpha \end{aligned}$$

or minimize conditional value at risk (CVaR)

$$\begin{aligned} \underset{x \in \mathcal{X}}{\text{minimize}} \quad & \mathbf{E}_{p_\Theta}[u(x,\theta) | u(x,\theta) > u_0] \\ \text{subject to} \quad & \mathrm{P}[u(x,\theta) < u_0] = 1 - \alpha \end{aligned}$$

Stochastic Optimization – SAA

$$\underset{x \in \mathcal{X}}{\text{minimize}} \quad \{ \ U(x) \equiv \mathbf{E}_{p_\Theta}[u(x,\theta)] \ \}$$

where $\Theta \sim p_\Theta(\theta)$, $\theta \in \mathbb{R}^n$, $\mathcal{X} \subset \mathbb{R}^N$

- Presumes knowledge of $p_\Theta(\cdot)$
- Typically relies on sample averaged approximation (SAA)

$$U(x) \approx \hat{U}(x) := \frac{1}{K} \sum_{k=1}^{K} u(x, \theta_k)$$

- Accurate Monte Carlo estimation requires large K
- $\hat{U}(x)$ is a noisy estimator of $U(x)$
  - Gradients of $\hat{U}(x)$ challenging to estimate

Stochastic Optimization with PCE

$$\theta_{PC}(\xi) = \sum_k \alpha_k \Psi_k(\xi), \quad u_{PC}(x,\xi) = \sum_k u_k(x)\Psi_k(\xi)$$

$$\hat{U}(x) = \mathbf{E}_{p_\xi}[u_{PC}(x,\xi)] = u_0(x)$$

where

$$u_0(x) = \int u\, p_\xi \mathrm{d}\xi = \sum_i w_i\, u(x, \theta(\xi_i))$$

is estimated using forward UQ methods

- perhaps intrusively, if $u(x,\theta)$ is relatively simple
- otherwise non-intrusive, *e.g.* sparse quadrature

Computational efficiency relative to Monte Carlo depends on

- the dimensionality of θ
- the θ-smoothness of $u(x,\theta)$

Gradients over the design space

- Estimation of $\frac{du_0}{dx}$ requires
  - A functional representation of $u_0(x)$ to be differentiated, or
  - A hi-resolution estimation of $u_{PC}(x_i, \xi),\ i = 1, \ldots, I_{mesh}$, or
  - A PCE for $\frac{du}{dx}(x, \xi)$, and hence gradients of the obj. func.

- Alternatively, the PCE can be built over (x, ξ) Eldred, IJUQ 2011

$$u_{PC}(x, \xi) = \sum_k u_k \Psi_k(x, \xi)$$

  - Functional representation of $u_{PC}(x, \xi)$ over x is built-in
  - Easy access to gradients/hessians over x
  - But a higher dimensional forward UQ problem

PCE in Power Grid Stochastic Optimization

Scenario Generation – Random Field (RF) Inputs

- Power-grid optimization involves uncertainties
  - in both loads and alternative energy sources
  - The largest uncertainties are in wind and solar generation
  - Being uncertain functions of time, these are RFs
- The Karhunen-Loeve expansion (KLE) provides an optimal representation of RFs, capturing both mean & covariance

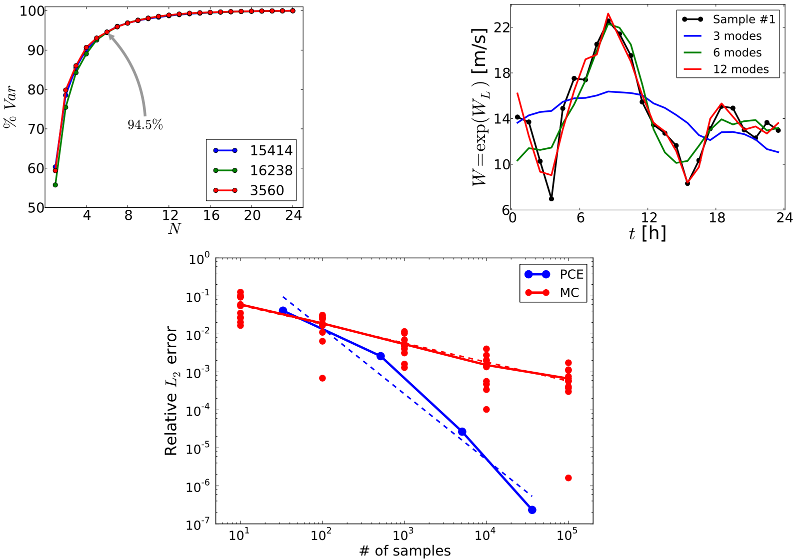
$$W(t,\omega) = \mu(t) + \sum_{i=1}^{\infty} \sqrt{\lambda_i}\eta_i(\omega)\phi_i(t)$$

- $\mu(t)$ is the mean of $W(t,\omega)$ at t
- λ_i and $\phi_i(t)$ are the eigenvalues and eigenfunctions of the covariance $C(t_1,t_2) = \langle [W(t_1,\omega) - \mu(t_1)][W(t_2,\omega) - \mu(t_2)] \rangle$
- The η_i are uncorrelated zero-mean unit-variance RVs

PCE in Power Grid Stochastic Optimization

- Consider the Economic Dispatch problem
  - Given a set of generators online
  - Find optimal expected power generation schedules over the next 24 hr
  - Feasibility and operational constraints
- IEEE 118 bus system – 54 generators, 64 loads
- 3 generators replaced by wind farms
- wind data from two sites in Wyoming and one in California
- KLE $\Rightarrow$ 16-dimensional forward UQ problem
- Minimum cost $Q(x, W(t,\omega)) \approx Q_{\mathrm{PC}}(x, \eta(\xi)) = \sum_k q_k(x)\Psi_k(\xi)$
  - Estimate PC coefficients using sparse quadrature
  - Expectation $q_0(x)$

Scenario Generation – Random Field (RF) Inputs



Distributionally Robust Optimization (DRO)

- Presumes imperfect knowledge of $p_\Theta(\cdot)$
- Consider $p_\Theta \in \mathcal{D}$,

$$\underset{x\in\mathcal{X}}{\text{minimize}} \quad \{ \max_{p_\Theta\in\mathcal{D}} \mathbf{E}_{p_\Theta}[u(x,\theta)] \}$$

- Implementation: define ambiguity set $\mathcal{D}$, *e.g.*
  - given presumed $\mathcal{S} \supset \text{supp}(p_\Theta)$ & moments of Θ: (μ_0, Σ_0)
    - Allow uncertainty in moments Delage & Ye, OR 2010
  - given max KL-divergence between p_Θ and a nominal p_0 Hu & Hong, 2013
- Utility of PC methods
  - Moment constraints accessible with PCE Eldred, IJUQ 2011
- Connections to "Optimal UQ" Owhadi, SIAM Review 2013
- Possible role for Maximum Entropy methods?

Robust optimization

- Set-based approach – support of uncertain parameter PDF
- Protect against worse case scenario in the set
- Learn set based on samples/historical-realizations of the uncertain-parameters as RVs

- Topics:
  - Role for PDF quantiles?
  - PCE extrema
  - Data-analysis/classification for establishing set boundaries
  - PDF tail behavior – Extreme Value Theory

The End