Optimization under uncertainty of high-dimensional, sloppy models

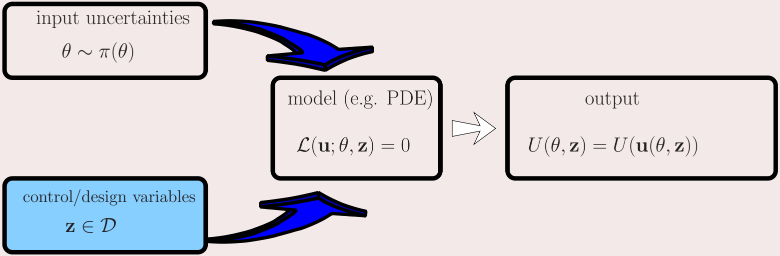
Professur für Kontinuums Mechanik

Continuum Mechanics Group
p.s.koutsourelakis@tum.de

SIAM - UQ 2016
April 6 2016

Motivation

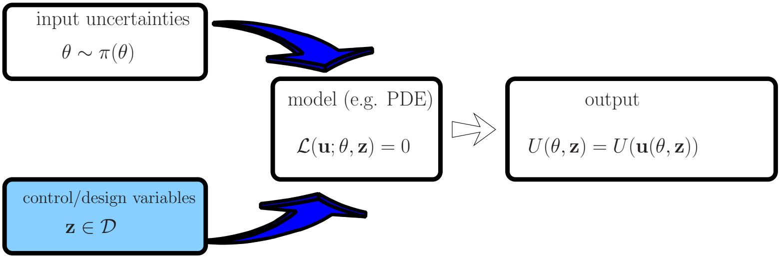
Uncertainty quantification



- uncertainties $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$, $n_\theta >> 1$
- design/control variables $\boldsymbol{z} \in \mathcal{D} \subset \mathbb{R}^{n_d}$, $n_d >> 1$
- Goal - Stochastic Optimization: Can we *efficiently* optimize w.r.t $\boldsymbol{z}$ and some output utility/gain $U(\boldsymbol{\theta}, \boldsymbol{z})$:

$$V(\boldsymbol{z}) = \int U(\boldsymbol{\theta}, \boldsymbol{z})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

Motivation



Big Data Challenges

- Solve model (e.g. PDE) to obtain: $u(\boldsymbol{\theta}, \boldsymbol{z}), \frac{\partial u}{\partial \boldsymbol{\theta}}, \frac{\partial u}{\partial \boldsymbol{z}}$
  - ✓ High-dimensional
  - ✓ Complex
  - ✓ Structured
  - × *Very Expensive*: The cost of the forward solves is a major factor in the overall efficiency

Motivation

Stochastic, model-based design/optimization: Find the design $\boldsymbol{z}$ that "on average" will respond the closest to the desired/target response $\boldsymbol{u}_0$

$$\max_{\boldsymbol{z}} \quad V(\boldsymbol{z}) = \int U(\theta, \boldsymbol{z}) \pi(\theta) \, d\theta$$

$$\text{where: } U(\theta, \boldsymbol{z}) = e^{-\frac{1}{2\sigma^2}||\boldsymbol{u}_0 - \boldsymbol{u}(\theta, \boldsymbol{z})||^2}$$

Desiderata - The proposed scheme should be able to:

- handle high-dimensional uncertainties θ (e.g $\mathcal{O}(dim(\theta)) = 1000$)
- handle high-dimensional design spaces $\boldsymbol{z}$ (e.g $\mathcal{O}(dim(\boldsymbol{z})) = 1000$)
- assess the sensitivity of the objective to design features (robustness)
- require the least possible evaluations of $\boldsymbol{u}(\theta, \boldsymbol{z})$ (and its derivatives)

Motivation

Deterministic optimization

- There is a wealth of techniques adapted to PDE-settings (e.g. adjoint formulations)
- Their direct transition to the stochastic setting is infeasible/impractical.

Stochastic Approximation (Robbins & Monro 1951)

- Perform gradient ascent i.e.:

$$\boldsymbol{z}^{(k+1)} = \boldsymbol{z}^{(k)} + \alpha_k \hat{\boldsymbol{J}}(\boldsymbol{z}^{(k)})$$

where:

- $\alpha_k > 0$, $\alpha_k \to 0$, $\sum_{k=0}^{\infty} \alpha_k = +\infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < +\infty$.
- $\hat{\boldsymbol{J}}(\boldsymbol{z}^{(k)}) = \text{unbiased estimator of } \frac{\partial V}{\partial \boldsymbol{z}} = \int \frac{\partial U(\boldsymbol{\theta}, \boldsymbol{z})}{\partial \boldsymbol{z}} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ (e.g. with Monte Carlo and a single $\boldsymbol{\theta}-$sample)

Motivation

Surrogate Models (e.g. gen. Pol. Chaos, Multi-dimensional Gaussian Processes): $\hat{\boldsymbol{u}}(\boldsymbol{z}, \theta) \approx \boldsymbol{u}(\boldsymbol{z}, \theta)$

- Not competitive when $dim(\theta), dim(\boldsymbol{z}) >> 1$
- Accuracy can also be poor in such settings.

Approach

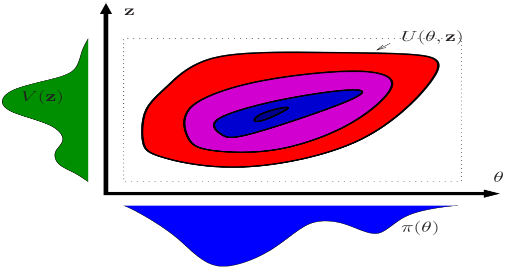
Optimize the *expected* utility/gain $V(\boldsymbol{z})$:

$$V(\boldsymbol{z}) = \int U(\boldsymbol{\theta}, \boldsymbol{z})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \quad U(\boldsymbol{\theta}, \boldsymbol{z}) = e^{-\frac{1}{2\sigma^2}||\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{z})||^2}$$

We adopt a *probabilistic inference* approach (*Müller 1999*) in the joint $\boldsymbol{\theta} \times \boldsymbol{z}$ space [a]:

$$p(\boldsymbol{\theta}, \boldsymbol{z}) \propto U(\boldsymbol{\theta}, \boldsymbol{z})\pi(\boldsymbol{\theta})$$

Note that the $\boldsymbol{z}$-coordinates of $(\boldsymbol{\theta}, \boldsymbol{z})$ samples from $p(\boldsymbol{\theta}, \boldsymbol{z})$ will concentrate on the maxima of V.



[a] $U(\boldsymbol{\theta}, \boldsymbol{z})$ is assumed positive or in general bounded from below

Approach

the good:

- uniform treatment as a probabilistic inference problem
- inferring the density $p(\boldsymbol{z})$ rather than a single-point estimate $\boldsymbol{z}^*$ can provide useful information about sensitivity of the solution

the bad:

- we have to work on the joint space $\theta \otimes \boldsymbol{z}$
- standard inference tools (e.g. plain vanilla Monte Carlo) can be very demanding in terms of forward runs.
- multiple local optima of $V(\boldsymbol{z})$

Approach

the good:

- uniform treatment as a probabilistic inference problem
- inferring the density $p(\boldsymbol{z})$ rather than a single-point estimate $\boldsymbol{z}^*$ can provide useful information about sensitivity of the solution

the bad:

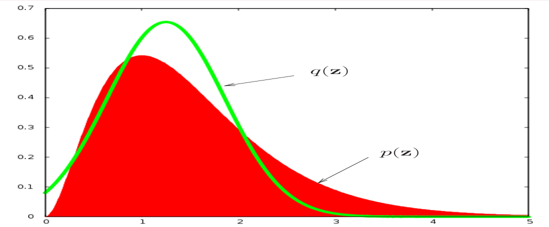
- we have to work on the joint space $\boldsymbol{\theta} \otimes \boldsymbol{z}$
- standard inference tools (e.g. plain vanilla Monte Carlo) can be very demanding in terms of forward runs.
- multiple local optima of $V(\boldsymbol{z})$

Variational Inference & Learning

Our goal is to infer:

$$p(\boldsymbol{\theta}, \boldsymbol{z}) \propto U(\boldsymbol{\theta}, \boldsymbol{z})\pi(\boldsymbol{\theta}) \rightarrow p(\boldsymbol{z}) \propto V(\boldsymbol{z}) = \int U(\boldsymbol{\theta}, \boldsymbol{z})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

Variational inference attempts to *approximate* $p(\boldsymbol{z})$ with a density $q^*(\boldsymbol{z})$ (belonging to an appropriate family of distributions $\mathcal{Q}$) such that [Bishop 2006]:



$$q^*(\boldsymbol{z}) = \arg\min_{q \in \mathcal{Q}} KL(q(\boldsymbol{z})||p(\boldsymbol{z})) = -\int q(\boldsymbol{z}) \log \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}\, d\boldsymbol{z}$$

Variational Inference & Learning

- In the joint space $\boldsymbol{\theta} \otimes \boldsymbol{z}$, we seek $q(\boldsymbol{\theta}, \boldsymbol{z})$ that minimizes the KL-divergence with the target joint density $p(\boldsymbol{\theta}, \boldsymbol{z}) = \frac{U(\boldsymbol{\theta},\boldsymbol{z})\pi(\boldsymbol{\theta})}{Z}$

$$
\begin{aligned}
KL(q(\boldsymbol{\theta}, \boldsymbol{z})||p(\boldsymbol{\theta}, \boldsymbol{z})) \quad &= -\int q(\boldsymbol{\theta}, \boldsymbol{z}) \log \frac{p(\boldsymbol{\theta},\boldsymbol{z})}{q(\boldsymbol{\theta},\boldsymbol{z})} \, d\boldsymbol{\theta} \, d\boldsymbol{z} \\
&= \log Z - \mathcal{F}(q)
\end{aligned}
$$

- Minimizing the Kullback-Leibler divergence is equivalent to maximizing :

$$
\begin{aligned}
\mathcal{F}(q) \quad &= E_q \left(\log \frac{U(\boldsymbol{\theta},\boldsymbol{z})\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta},\boldsymbol{z})} \right) \\
&= E_q(\log U(\boldsymbol{\theta}, \boldsymbol{z})) + E_q(\log \pi(\boldsymbol{\theta})) - E_q(\log q)
\end{aligned}
$$

  - Easy/Tractable terms: $E_q(\log \pi(\boldsymbol{\theta}))$, $E_q(\log q)$
  - Difficult term: $E_q(\log U(\boldsymbol{\theta}, \boldsymbol{z})) = -\frac{1}{2\sigma^2} E_q(||\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{z})||^2)$
  - What about high-dimensional $\boldsymbol{z}$ (or $\boldsymbol{\theta}$)?
  - What about any regularization/prior on $\boldsymbol{z}$?

Variational Inference & Learning

- In the joint space $\boldsymbol{\theta} \otimes \boldsymbol{z}$, we seek $q(\boldsymbol{\theta}, \boldsymbol{z})$ that minimizes the KL-divergence with the target joint density $p(\boldsymbol{\theta}, \boldsymbol{z}) = \frac{U(\boldsymbol{\theta}, \boldsymbol{z})\pi(\boldsymbol{\theta})}{Z}$

$$\begin{aligned} KL(q(\boldsymbol{\theta}, \boldsymbol{z})||p(\boldsymbol{\theta}, \boldsymbol{z})) &= -\int q(\boldsymbol{\theta}, \boldsymbol{z}) \log \frac{p(\boldsymbol{\theta}, \boldsymbol{z})}{q(\boldsymbol{\theta}, \boldsymbol{z})} \, d\boldsymbol{\theta} \, d\boldsymbol{z} \\ &= \log Z - \mathcal{F}(q) \end{aligned}$$

- Minimizing the Kullback-Leibler divergence is equivalent to maximizing :

$$\begin{aligned} \mathcal{F}(q) &= E_q \left(\log \frac{U(\boldsymbol{\theta}, \boldsymbol{z})\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta}, \boldsymbol{z})} \right) \\ &= E_q(\log U(\boldsymbol{\theta}, \boldsymbol{z})) + E_q(\log \pi(\boldsymbol{\theta})) - E_q(\log q) \end{aligned}$$

  - Easy/Tractable terms: $E_q(\log \pi(\boldsymbol{\theta}))$, $E_q(\log q)$
  - Difficult term: $E_q(\log U(\boldsymbol{\theta}, \boldsymbol{z})) = -\frac{1}{2\sigma^2} E_q(||\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{z})||^2)$
  - What about high-dimensional $\boldsymbol{z}$ (or $\boldsymbol{\theta}$)?
  - What about any regularization/prior on $\boldsymbol{z}$?

Sparse Bayesian Learning and "Sloppiness" [Brown & Sethna 2003]

$$\underbrace{\boldsymbol{z}}_{N\times 1} = \boldsymbol{\mu}_z + \underbrace{\boldsymbol{W}}_{N\times n}\underbrace{\boldsymbol{y}}_{n\times 1} + \eta_z$$

where:

- $\boldsymbol{W}$: set of reduced basis/features/vocabulary ($n << N$)
- $\boldsymbol{y}$: reduced-coordinates
- η_z: remaining "noise"

Sparse Bayesian Learning and "Sloppiness" [Brown & Sethna 2003]

$$\boldsymbol{z} = \boldsymbol{\mu}_z + \underbrace{\boldsymbol{W}}_{N \times n} \boldsymbol{y} + \boldsymbol{\eta}_z, \quad \boldsymbol{\theta} = \boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{\eta}_\theta$$

- Assumption 1: Latent variables $\boldsymbol{y}, \boldsymbol{\eta}_z, \boldsymbol{\eta}_\theta$

$$q(\boldsymbol{y}, \boldsymbol{\eta}_z, \boldsymbol{\eta}_\theta) = q(\boldsymbol{y}, \boldsymbol{\eta}_\theta) q(\boldsymbol{\eta}_z)$$

- Assumption 2: Family of approximating distributions $\boldsymbol{q} \in \mathcal{Q}$ are *multivariate Gaussians* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$.

$$q(\boldsymbol{y}, \boldsymbol{\eta}_\theta) \equiv \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{C}_{\theta\theta} & \boldsymbol{C}_{\theta y} \\ \boldsymbol{C}_{\theta y}^T & \boldsymbol{C}_{yy} \end{bmatrix}\right), \quad q(\boldsymbol{\eta}_z) \equiv \mathcal{N}(\boldsymbol{0}, \tau_z^{-1}(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$$

  - This is NOT PCA
  - Directions $\boldsymbol{y}$ have the lowest variance i.e. variations along them, cause (locally) smaller changes in $V(\boldsymbol{z})$.
  - *Implicit assumption: $dim(\boldsymbol{y}) << dim(\boldsymbol{z})$*

Sparse Bayesian Learning and "Sloppiness" [Brown & Sethna 2003]

$$\boldsymbol{z} = \boldsymbol{\mu}_z + \underbrace{\boldsymbol{W}}_{N\times n} \boldsymbol{y} + \boldsymbol{\eta}_z, \quad \boldsymbol{\theta} = \boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{\eta}_\theta$$

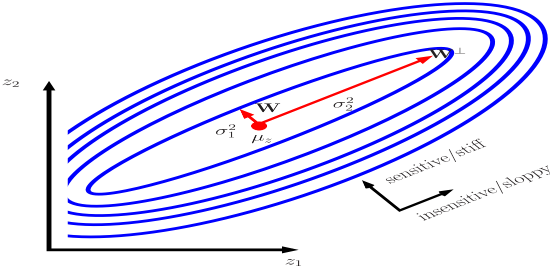
- Assumption 1: Latent variables $\boldsymbol{y}, \boldsymbol{\eta}_z, \boldsymbol{\eta}_\theta$

$$q(\boldsymbol{y}, \boldsymbol{\eta}_z, \boldsymbol{\eta}_\theta) = q(\boldsymbol{y}, \boldsymbol{\eta}_\theta) q(\boldsymbol{\eta}_z)$$

- Assumption 2: Family of approximating distributions $\boldsymbol{q} \in \mathcal{Q}$ are *multivariate Gaussians* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$.

$$q(\boldsymbol{y}, \boldsymbol{\eta}_\theta) \equiv \mathcal{N}(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{C}_{\theta\theta} & \boldsymbol{C}_{\theta y} \\ \boldsymbol{C}_{\theta y}^T & \boldsymbol{C}_{yy} \end{bmatrix}), \quad q(\boldsymbol{\eta}_z) \equiv \mathcal{N}(\boldsymbol{0}, \tau_z^{-1}(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$$

  - This is NOT PCA
  - Directions $\boldsymbol{y}$ have the lowest variance i.e. variations along them, cause (locally) smaller changes in $V(\boldsymbol{z})$.
  - *Implicit assumption:* $dim(\boldsymbol{y}) << dim(\boldsymbol{z})$



- Since $\boldsymbol{z} = \boldsymbol{\mu}_z + \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\eta}_z$:

$$V(\boldsymbol{z}) = \int U(\theta, \boldsymbol{z}) p(\theta) d\theta \approx q(\boldsymbol{z}) \propto e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_z)^T \boldsymbol{C}_{zz}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z)}$$

where:

$$\begin{aligned} \boldsymbol{C}_{zz} &= \boldsymbol{W}\boldsymbol{C}_{yy}\boldsymbol{W}^T + \tau_z^{-1}\boldsymbol{I} \\ &= \boldsymbol{W}\boldsymbol{U} \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \sigma_{d_y}^2 \end{bmatrix} \boldsymbol{U}^T \boldsymbol{W}^T + \tau_z^{-1}\boldsymbol{I} \end{aligned}$$

$$\boxed{\boldsymbol{z} = \boldsymbol{\mu}_z + \underbrace{\boldsymbol{W}}_{N\times n}\boldsymbol{y} + \boldsymbol{\eta}_z, \quad \boldsymbol{\theta} = \boldsymbol{\mu_\theta} + \eta_\theta}$$

- Assumption 3: Model parameters $\boldsymbol{R} = \{\boldsymbol{\mu}_z, \boldsymbol{W}, \boldsymbol{\mu}_\theta, \sigma_d^2\}$
  - prior $p(\boldsymbol{\mu}_z)$ for regularization (problem-dependent)
  - $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$, i.e. $p(\boldsymbol{W}) \equiv$ uniform on Stiefel manifold $V_n(\mathbb{R}^N)$
  - $\boldsymbol{\mu}_\theta$ from $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta)$
- Assumption 4: Linearization at $(\boldsymbol{\mu}_\theta, \boldsymbol{\mu}_d)$ - E.g. $U(\theta, \boldsymbol{z}) = e^{-\frac{1}{2\sigma^2}||u_0 - u(\theta,\boldsymbol{z})||^2}$:

$$\boldsymbol{u}(\theta, \boldsymbol{z}) \quad \approx \boldsymbol{u}(\boldsymbol{\mu}_\theta, \boldsymbol{\mu}_z) + \boldsymbol{G}_\theta \eta_\theta + \boldsymbol{G}_d(\boldsymbol{W}\boldsymbol{y} + \boldsymbol{\eta}_z)$$

where $\boldsymbol{G}_\theta = \frac{\partial u}{\partial \theta}|_{\boldsymbol{\mu}_z, \boldsymbol{\mu}_\theta}$ and $\boldsymbol{G}_z = \frac{\partial u}{\partial \boldsymbol{z}}|_{\boldsymbol{\mu}_z, \boldsymbol{\mu}_\theta}$ available with minimal cost from adjoint-PDE.

$$\boxed{\boldsymbol{z} = \boldsymbol{\mu}_z + \underbrace{\boldsymbol{W}}_{N\times n}\boldsymbol{y} + \boldsymbol{\eta}_z, \quad \boldsymbol{\theta} = \boldsymbol{\mu}_{\boldsymbol{\theta}} + \eta_\theta}$$

- <u>Assumption 3</u>: Model parameters $\boldsymbol{R} = \{\boldsymbol{\mu}_z, \boldsymbol{W}, \boldsymbol{\mu}_\theta, \sigma_d^2\}$
  - prior $p(\boldsymbol{\mu}_z)$ for regularization (problem-dependent)
  - $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$, i.e. $p(\boldsymbol{W}) \equiv$ uniform on Stiefel manifold $V_n(\mathbb{R}^N)$
  - $\boldsymbol{\mu}_\theta$ from $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta)$
- <u>Assumption 4</u>: Linearization at $(\boldsymbol{\mu}_\theta, \boldsymbol{\mu}_d)$ - E.g. $U(\boldsymbol{\theta}, \boldsymbol{z}) = e^{-\frac{1}{2\sigma^2}||\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{z})||^2}$:

$$\boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{z}) \quad \approx \boldsymbol{u}(\boldsymbol{\mu}_\theta, \boldsymbol{\mu}_z) + \boldsymbol{G}_\theta \boldsymbol{\eta}_\theta + \boldsymbol{G}_d(\boldsymbol{W}\boldsymbol{y} + \boldsymbol{\eta}_z)$$

where $\boldsymbol{G}_{\boldsymbol{\theta}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\mu}_z, \boldsymbol{\mu}_\theta}$ and $\boldsymbol{G}_{\boldsymbol{z}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{z}}|_{\boldsymbol{\mu}_z, \boldsymbol{\mu}_\theta}$ available with minimal cost from adjoint-PDE.

Variational Inference & Learning

$$
\begin{aligned}
\mathcal{F}(q(\boldsymbol{y},\boldsymbol{\eta}_z,\boldsymbol{\eta}_\theta),\boldsymbol{R}) = \quad & -\frac{\tau_Q}{2}\Big(|\boldsymbol{u}_{target}-\boldsymbol{u}(\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_z)|^2 && \text{(from } E_q[U]) \\
& +tr(\boldsymbol{G}_\theta^T\boldsymbol{G}_\theta\ \boldsymbol{C}_{\theta\theta}) + tr(\boldsymbol{W}^T\boldsymbol{G}_z^T\boldsymbol{G}_z\boldsymbol{W}\ \boldsymbol{C}_{yy}) \\
& +\tau_z^{-1}tr(\boldsymbol{G}_z^T\boldsymbol{G}_z\ (\boldsymbol{I}-\boldsymbol{W}\boldsymbol{W}^T)) \\
& +2tr(\boldsymbol{G}_\theta^T\boldsymbol{G}_z\boldsymbol{W}\ \boldsymbol{C}_{\theta y})\Big) \\
& -\frac{1}{2}(\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\theta 0})^T\boldsymbol{C}_{\theta 0}^{-1}(\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\theta 0}) && \text{(from } E_q[p_\theta]) \\
& -\frac{1}{2}tr(\boldsymbol{C}_{\theta 0}^{-1}\ \boldsymbol{C}_{\theta\theta}) \\
& -\frac{\tau_{y0}}{2}tr(\boldsymbol{C}_{yy}) && \text{(from } E_q[p(\boldsymbol{y})]) \\
& -\frac{d_z-d_y}{2}\frac{\tau_{z0}}{\tau_z} \qquad \text{(from } E_q[p(\boldsymbol{\eta}_z)]) \\
& +\frac{1}{2}\log\begin{vmatrix}\boldsymbol{C}_{\theta\theta} & \boldsymbol{C}_{\theta y}\\ \boldsymbol{C}_{\theta y}^T & \boldsymbol{C}_{yy}\end{vmatrix} && \text{(from } E_q[q(\boldsymbol{\eta}_\theta,\boldsymbol{y})]) \\
& +\frac{d_z-d_y}{2}\log\tau_z && \text{(from } E_q[q(\boldsymbol{\eta}_z)]) \\
& +\log p_{\mu_z}(\boldsymbol{\mu}_z)+\log p_W(\boldsymbol{W}).
\end{aligned}
$$

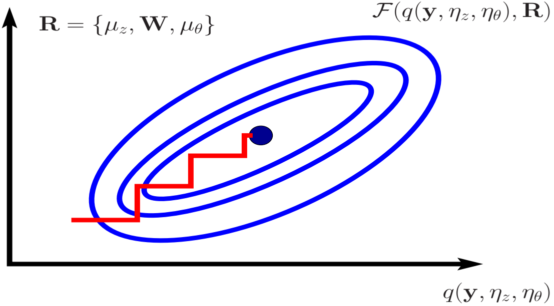


Figure : Variational Bayesian Expectation-Maximization (VB-EM, Beal & Ghahramani, 2003)

Iterate until convergence:

- **VB-Expectation**: Given the current $\boldsymbol{R}$ find the optimal q (i.e. the optimal $\boldsymbol{C}_{\theta\theta}, \boldsymbol{C}_{\theta y}, \boldsymbol{C}_{yy}, \tau_z$)
- **VB-Maximization**: Given the current q (i.e. $\boldsymbol{C}_{\theta\theta}, \boldsymbol{C}_{\theta y}, \boldsymbol{C}_{yy}, \tau_z$), find the optimal $\boldsymbol{R} = \{\boldsymbol{\mu}_z, \boldsymbol{W}, \boldsymbol{\mu}_\theta\}$.

Validation

$$\boxed{KL(q(\boldsymbol{y}, \boldsymbol{\eta}_z, \boldsymbol{\eta}_\theta) || p_{aux}(\boldsymbol{y}, \boldsymbol{\eta}_z, \boldsymbol{\eta}_\theta | \boldsymbol{R}))}$$

- Estimate with Importance Sampling
- To "normalize" w.r.t the dimension we use:

$$nKL = \frac{KL(q(\boldsymbol{y}, \boldsymbol{\eta}_z, \boldsymbol{\eta}_\theta) || p_{aux}(\boldsymbol{y}, \boldsymbol{\eta}_z, \boldsymbol{\eta}_\theta | \boldsymbol{R}))}{H(q)}.$$

where:

$$H(q) = \frac{d_\theta + d_y}{2} \log 2\pi + \frac{1}{2} \log \begin{vmatrix} \boldsymbol{C}_{\theta\theta}^{opt} & \boldsymbol{C}_{\theta y}^{opt} \\ sym. & \boldsymbol{C}_{yy}^{opt} \end{vmatrix} + \frac{d_z - d_y}{2} \log \frac{2\pi}{\tau_z^{opt}}.$$

Numerical Illustrations

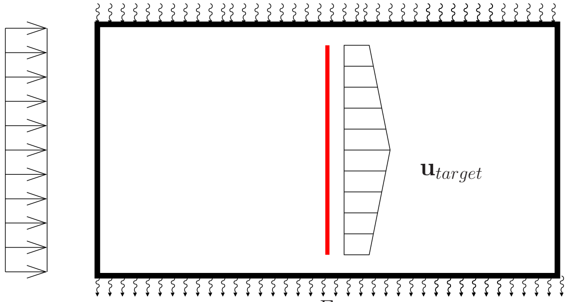
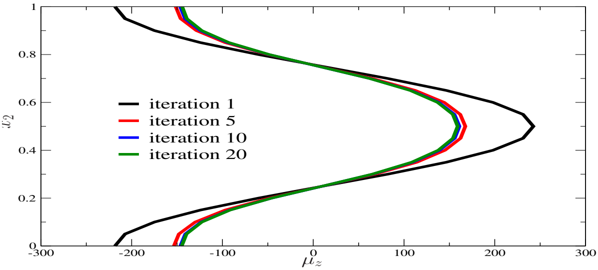


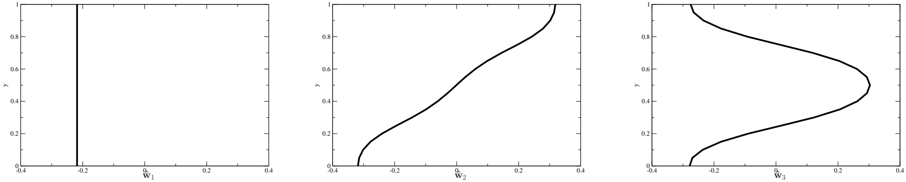
Figure : Random variables $d_\theta = 1600$, Number of design variables $d_z = 21$

- Forward model: $\nabla \cdot \big(-\lambda(\boldsymbol{x}) \nabla u(\boldsymbol{x}) \big) = 0$ in $[-1, 1] \times [0, 1]$
- Uncertainties (coef. of variation 0.5): $\lambda(\boldsymbol{x}) = e^{\lambda_g(\boldsymbol{x})}$,

$$\lambda_g(\boldsymbol{x}) \sim \mathcal{N}(\mu_g, C_g), \quad C_g(\Delta x_1, \Delta x_2) = \sigma_g^2 \exp\{-\frac{\sqrt{\Delta x_1^2 + \Delta x_2^2}}{x_0}\}, x_0 = \sqrt{0.1}$$

Numerical Illustrations





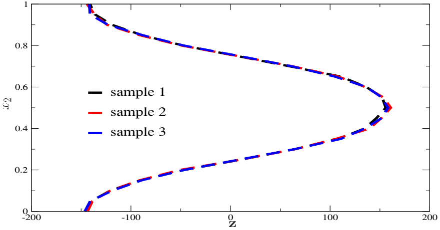
(a) $\sigma_1^2 = 4.0 \times 10^{-2}$

(b) $\sigma_2^2 = 1.5 \times 10^3$

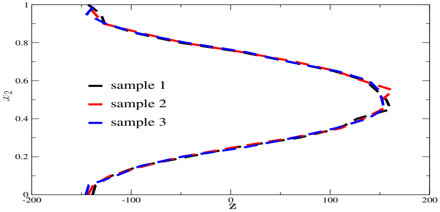
(c) $\sigma_3^2 = 6.6 \times 10^3$

Figure : First three most sensitive eigenvectors $\{\hat{\boldsymbol{w}}_j\}_{j=1}^3$ and associated variances σ_j^2.

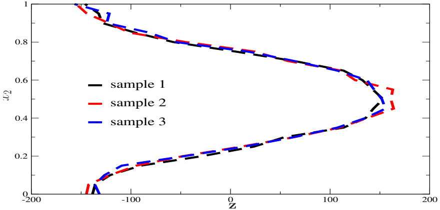
Numerical Illustration



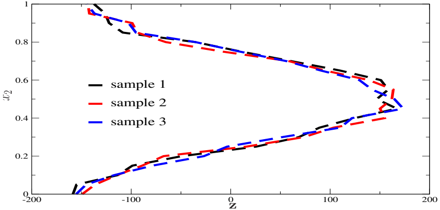
(a) $\frac{V(\boldsymbol{z})}{V(\boldsymbol{\mu}_z)} = 0.95$



(b) $\frac{V(\boldsymbol{z})}{V(\boldsymbol{\mu}_z)} = 0.75$



(c) $\frac{V(\boldsymbol{z})}{V(\boldsymbol{\mu}_z)} = 0.50$



(d) $\frac{V(\boldsymbol{z})}{V(\boldsymbol{\mu}_z)} = 0.25$

Figure : Alternative designs z at various levels of expected gain $\frac{V(\boldsymbol{z})}{V(\boldsymbol{\mu}_z)}$ as compared to the optimal $\boldsymbol{\mu}_z$

Validation

d_y	nKL
1	1.5×10^{-1}
2	1.2×10^{-1}
5	4.7×10^{-2}
10	2.5×10^{-2}
20	9.8×10^{-3}

Table : Normalized KL-divergence

Deterministic topology optimization

Shape/topology optimization:

$$\max_{\boldsymbol{z}} \quad e^{-|\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{z})|^2/2\sigma_0^2}$$

such that:

$$\boldsymbol{K}(\boldsymbol{z})\boldsymbol{u}(\boldsymbol{z}) = \boldsymbol{b} \quad \text{(governing equation)}$$
$$\int d(\boldsymbol{x})\, d\boldsymbol{x} = V_0, \quad \text{(volume fraction)}$$
$$d(\boldsymbol{x}) \in [0, 1]$$

$$d(\boldsymbol{x}) = \begin{cases} 1, & \textit{material} \\ 0, & \textit{void} \end{cases}$$

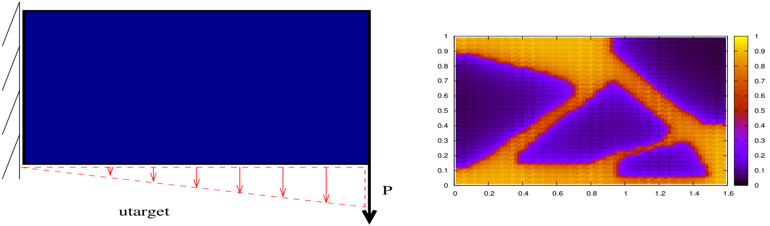


Figure : *Deterministic* adjoint-based gradient optimization - $O(100)$ forward runs

Stochastic topology optimization

Shape/topology optimization:

$\boldsymbol{K}(\boldsymbol{z},\boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{z},\boldsymbol{\theta})=\boldsymbol{b}$ (governing equation)
$\int d(\boldsymbol{x})\,d\boldsymbol{x}=V_0,$ (volume fraction)
$d(\boldsymbol{x})\in[0,1]$

$$d(\boldsymbol{x})=\begin{cases}1, & \textit{material}\\ 0, & \textit{void}\end{cases}$$

$\boldsymbol{\theta}\sim\pi(\boldsymbol{\theta}),$ (random material properties)

Stochastic topology optimization

Targeted design: $\max_{\boldsymbol{z}}\int e^{-\frac{1}{2}|\boldsymbol{u}(\boldsymbol{z},\boldsymbol{\theta})-\boldsymbol{u}_0|^2}\,\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$

such that:

$\boldsymbol{K}(\boldsymbol{z},\boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{z},\boldsymbol{\theta})=\boldsymbol{b}$ (governing equation)
$\int d(\boldsymbol{x})\,d\boldsymbol{x}=V_0,$ (volume fraction)
$d(\boldsymbol{x})\in[0,1]$
$\boldsymbol{\theta}\sim\pi(\boldsymbol{\theta})$

- Number of random variables $d_\theta=3536$
- Number of design variables $d_z=3536$

Variational Inference - Constraints

Shape/topology optimization:

$\max_{\boldsymbol{z}} \quad e^{-|\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{z})|^2/2\sigma_0^2}$

such that:

$\boldsymbol{K}(\boldsymbol{z})\boldsymbol{u}(\boldsymbol{z}) = \boldsymbol{b}$ (governing equation)

$\int d(\boldsymbol{x})\, d\boldsymbol{x} = V_0,$ (volume fraction)

$d(\boldsymbol{x}) \in [0, 1]$

$$d(\boldsymbol{x}) = \begin{cases} 1, & \textit{material} \\ 0, & \textit{void} \end{cases}$$

- Equality constraint $h(\boldsymbol{z}) = 0$: *probabilistic enforcement*

$$\text{Target density: } p(\boldsymbol{\theta}, \boldsymbol{z}) \propto U(\boldsymbol{\theta}, \boldsymbol{z})\pi(\boldsymbol{\theta})\, e^{-\frac{h(\boldsymbol{z})^2}{2\epsilon^2}}, \quad \epsilon \to 0$$

- $p(\boldsymbol{\mu}_z)$: penalize jumps with ARD prior
- Use logit to convert binary to real variables

Numerical Illustrations

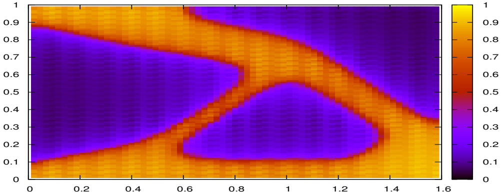
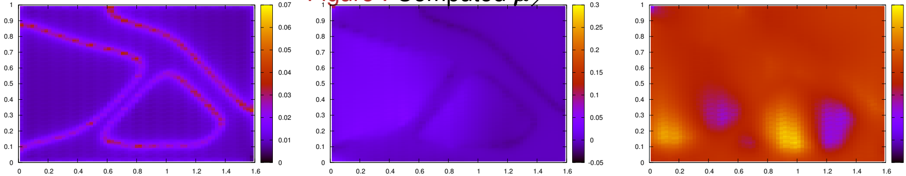


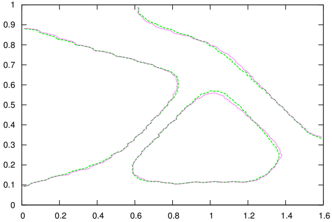
Figure : Computed $\boldsymbol{\mu}_z$



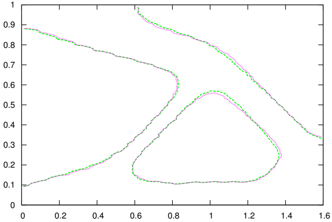
(a) $\sigma_1^2 = 7.31 \times 10^{-1}$ (b) $\sigma_2^2 = 1.25 \times 10^{2}$ (c) $\sigma_7^2 = 1.36 \times 10^{4}$

Figure : Sensitive eigenvectors $\hat{\boldsymbol{w}}_j$ and associated variances σ_j^2.

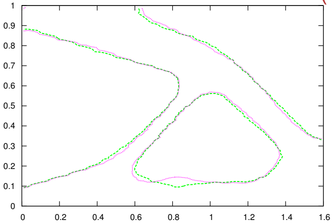
Numerical Illustrations



(a) $\frac{V(\mathbf{z})}{V(\boldsymbol{\mu}_z)} = 0.75$



(b) $\frac{V(\mathbf{z})}{V(\boldsymbol{\mu}_z)} = 0.50$



(c) $\frac{V(\mathbf{z})}{V(\boldsymbol{\mu}_z)} = 0.25$

Figure : Outline of alternative designs z at various levels of expected gain $\frac{V(\mathbf{z})}{V(\boldsymbol{\mu}_z)}$ as compared to the optimal $\boldsymbol{\mu}_z$ ($VF = 0.4$)

Numerical Illustration

Validation

	nKL	
d_y	$VolumeFraction = 0.4$	$VolumeFraction = 0.2$
5	1.5×10^{-2}	3.4×10^{-1}
10	8.7×10^{-3}	1.9×10^{-1}
15	3.9×10^{-3}	1.3×10^{-1}
20	6.0×10^{-4}	6.8×10^{-2}

Table : Normalized KL-divergence

Summary & Outlook

- Stochastic *optimization/design* poses significantly more challenges than *uncertainty propagation* when *thousands* of random and design variables are present.
- We advocate a probabilistic inference reformulation
- Variational Bayesian inference and learning techniques lead to efficient computation of approximate solutions
- Dictionary learning can lead to significant dimensionality reduction and identify most sensitive directions
- Extension 1: MoG to capture non-Gaussian and multi-modal design objectives
- Extension 2: Integration of *probabilistic surrogates* [Bilionis & Zabaras 2014]