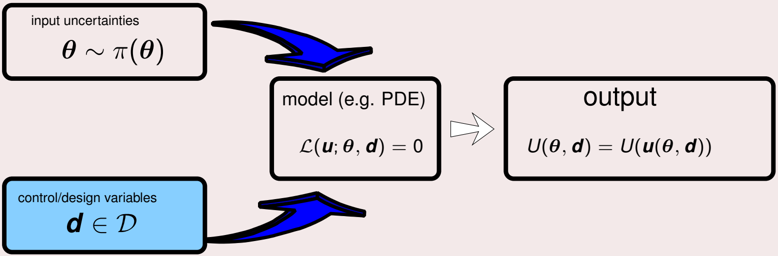
Variational Bayesian strategies for high-dimensional, stochastic design problems

Professur für Kontinuumsmechanik
p.s.koutsourelakis@tum.de

Big Data and Predictive Computational Modeling
IAS-TUMunich
May 21 2015

Motivation

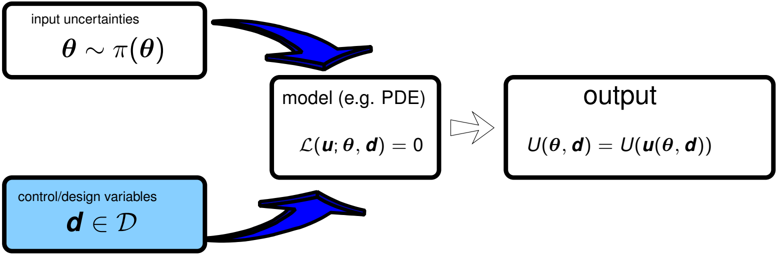
Uncertainty quantification



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

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

Motivation



Big Data Challenges

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

Motivation

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

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

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

Desiderata - The proposed scheme should be able to:

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

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

where:

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

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

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

Approach

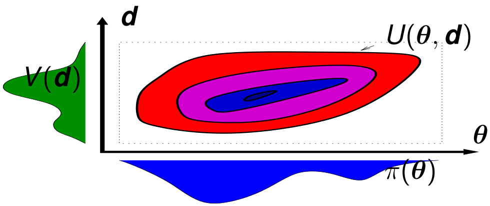
Optimize the *expected* utility $V(\boldsymbol{d})$:

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

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

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

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



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

Approach

the good:

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

the bad:

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

Approach

the good:

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

the bad:

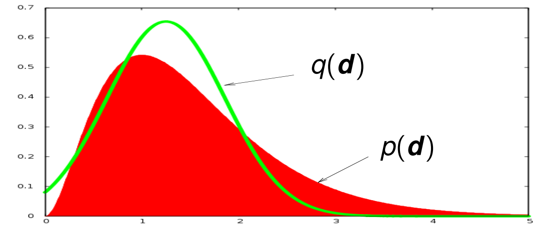
- we have to work on the joint space $\boldsymbol{\theta} \otimes \boldsymbol{d}$
- standard inference tools (e.g. plain vanilla Monte Carlo) can be very demanding in terms of forward runs.
- multiple local optima of $V(\boldsymbol{d})$

Variational Inference & Learning

Our goal is to infer:

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

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



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

Variational Inference & Learning

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

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

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

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

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

Variational Inference & Learning

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

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

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

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

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

Sparse Bayesian Learning

$$\underbrace{\boldsymbol{d}}_{N\times 1} = \boldsymbol{\mu}_d + \underbrace{\boldsymbol{W}}_{N\times n} \underbrace{\boldsymbol{y}}_{n\times 1} + \boldsymbol{\eta}_d$$

where:

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

Variational Inference & Learning

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

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

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

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

$$q(\boldsymbol{y}, \boldsymbol{\eta}_\theta) = \mathcal{N}(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{C}_{\theta\theta} & \boldsymbol{C}_{\theta y} \\ \boldsymbol{C}_{\theta y}^T & \boldsymbol{C}_{yy} \end{bmatrix}), \quad q(\boldsymbol{\eta}_d) = \mathcal{N}(\boldsymbol{0}, \sigma_d^2(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$$

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

Variational Inference & Learning

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

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

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

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

$$q(\boldsymbol{y}, \boldsymbol{\eta}_\theta) \equiv \mathcal{N}(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{C}_{\theta\theta} & \boldsymbol{C}_{\theta y} \\ \boldsymbol{C}_{\theta y}^T & \boldsymbol{C}_{yy} \end{bmatrix}), \quad q(\boldsymbol{\eta}_d) \equiv \mathcal{N}(\boldsymbol{0}, \sigma_d^2(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$$

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

Variational Inference & Learning

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

- Assumption 3: Model parameters $\boldsymbol{P} = \{\boldsymbol{\mu}_d, \boldsymbol{W}, \boldsymbol{\mu}_\theta, \sigma_d^2\}$
  - prior $p(\boldsymbol{\mu}_d)$ 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. $U(\boldsymbol{\theta}, \boldsymbol{d}) = e^{-\frac{1}{2\sigma^2}||\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{d})||^2}$:

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

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

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

- Assumption 3: Model parameters $\boldsymbol{P} = \{\boldsymbol{\mu}_d, \boldsymbol{W}, \boldsymbol{\mu}_\theta, \sigma_d^2\}$
  - prior $p(\boldsymbol{\mu}_d)$ 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. $U(\boldsymbol{\theta}, \boldsymbol{d}) = e^{-\frac{1}{2\sigma^2}||\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{d})||^2}$:

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

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

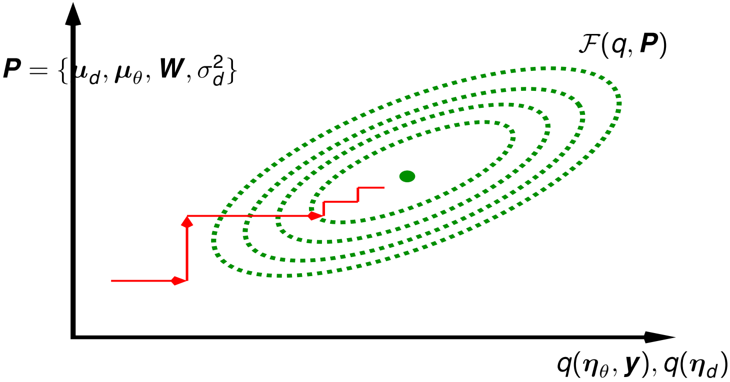


Figure : Variational Bayesian Expectation-Maximization (VB-EM, Beal & Ghahramani, 2003)

VB-EM Algorithm:

$$\mathcal{F}(\boldsymbol{P}, q) = E_q(\log U(\boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta, \boldsymbol{\mu}_d + \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\eta}_d)) + E_q(\log \pi(\boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta) p(\boldsymbol{y}) p(\boldsymbol{\eta}_d)) - E_q(\log q)$$

0. Initialize with $p(\boldsymbol{y}) \equiv \mathcal{N}(\boldsymbol{0}, \sigma_{y0}^2 \boldsymbol{I})$, $p(\boldsymbol{\eta}_d) \equiv \mathcal{N}(0, \sigma_{y0}^2(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$

1. Update $\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta$ (forward calls) [a]:

$$\max_{\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta} \mathcal{F}_\mu = -\frac{1}{2\sigma^2}|\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta)|^2 - \frac{1}{2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_0)^T \boldsymbol{S}_0^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_0) + \log p(\boldsymbol{\mu}_d)$$

2.1 Update $\boldsymbol{W}$ (No forward calls):

$$\max_{\boldsymbol{W}} \mathcal{F}_W = -\frac{1}{2\sigma^2} \boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W} : (\boldsymbol{C}_{yy} - \sigma_d^2 \boldsymbol{I}) + \frac{1}{\sigma^2} \boldsymbol{G}_\theta^T \boldsymbol{G}_d \boldsymbol{W} : \boldsymbol{C}_{\theta y}$$

2.2 Update $q(\boldsymbol{\eta}_\theta, \boldsymbol{y}) \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$, $q(\boldsymbol{\eta}_d) \equiv \mathcal{N}(\boldsymbol{0}, \sigma_d^2(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$ (No forward calls):

$$\boldsymbol{C}^{-1} = \begin{bmatrix} \boldsymbol{C}_{\theta\theta} & \boldsymbol{C}_{\theta y} \\ \boldsymbol{C}_{\theta y}^T & \boldsymbol{C}_{yy} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} \boldsymbol{G}_\theta^T \boldsymbol{G}_\theta + \boldsymbol{S}_0^{-1} & \frac{1}{\sigma^2} \boldsymbol{G}_\theta^T \boldsymbol{G}_d \\ sym. & \frac{1}{\sigma^2} \boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W} + \sigma_{y0}^{-2} \boldsymbol{I} \end{bmatrix}$$

$$\frac{1}{\sigma_d^2} = \frac{1}{\sigma_{y0}^2} + \frac{1}{(dim(\boldsymbol{d}) - dim(\boldsymbol{y}))} \frac{1}{\sigma^2} (tr(\boldsymbol{G}_d^T \boldsymbol{G}_d) - tr(\boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W}))$$

[a] Assuming $\pi(\theta) \equiv \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_0)$

VB-EM Algorithm:

$$\mathcal{F}(\boldsymbol{P}, q) = E_q(\log U(\boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta, \boldsymbol{\mu}_d + \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\eta}_d)) + E_q(\log \pi(\boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta) p(\boldsymbol{y}) p(\boldsymbol{\eta}_d)) - E_q(\log q)$$

0. Initialize with $p(\boldsymbol{y}) \equiv \mathcal{N}(\boldsymbol{0}, \sigma_{y0}^2 \boldsymbol{I})$, $p(\boldsymbol{\eta}_d) \equiv \mathcal{N}(0, \sigma_{y0}^2(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$
1. Update $\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta$ (forward calls) [a]:

$$\max_{\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta} \mathcal{F}_\mu = -\frac{1}{2\sigma^2} |\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta)|^2 - \frac{1}{2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_0)^T \boldsymbol{S}_0^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_0) + \log p(\boldsymbol{\mu}_d)$$

2.1 Update $\boldsymbol{W}$ (No forward calls):

$$\max_{\boldsymbol{W}} \mathcal{F}_W = -\frac{1}{2\sigma^2} \boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W} : (\boldsymbol{C}_{yy} - \sigma_d^2 \boldsymbol{I}) + \frac{1}{\sigma^2} \boldsymbol{G}_\theta^T \boldsymbol{G}_d \boldsymbol{W} : \boldsymbol{C}_{\theta y}$$

2.2 Update $q(\boldsymbol{\eta}_\theta, \boldsymbol{y}) \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$, $q(\boldsymbol{\eta}_d) \equiv \mathcal{N}(\boldsymbol{0}, \sigma_d^2(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$ (No forward calls):

$$\boldsymbol{C}^{-1} = \begin{bmatrix} \boldsymbol{C}_{\theta\theta} & \boldsymbol{C}_{\theta y} \\ \boldsymbol{C}_{\theta y}^T & \boldsymbol{C}_{yy} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} \boldsymbol{G}_\theta^T \boldsymbol{G}_\theta + \boldsymbol{S}_0^{-1} & \frac{1}{\sigma^2} \boldsymbol{G}_\theta^T \boldsymbol{G}_d \\ sym. & \frac{1}{\sigma^2} \boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W} + \sigma_{y0}^{-2} \boldsymbol{I} \end{bmatrix}$$

$$\frac{1}{\sigma_d^2} = \frac{1}{\sigma_{y0}^2} + \frac{1}{(dim(\boldsymbol{d}) - dim(\boldsymbol{y}))} \frac{1}{\sigma^2} (tr(\boldsymbol{G}_d^T \boldsymbol{G}_d) - tr(\boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W}))$$

[a] Assuming $\pi(\boldsymbol{\theta}) \equiv \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_0)$

VB-EM Algorithm:

$$\mathcal{F}(\boldsymbol{P}, q) = E_q(\log U(\boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta, \boldsymbol{\mu}_d + \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\eta}_d)) + E_q(\log \pi(\boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta) p(\boldsymbol{y}) p(\boldsymbol{\eta}_d)) - E_q(\log q)$$

0. Initialize with $p(\boldsymbol{y}) \equiv \mathcal{N}(\boldsymbol{0}, \sigma_{y0}^2 \boldsymbol{I})$, $p(\boldsymbol{\eta}_d) \equiv \mathcal{N}(0, \sigma_{y0}^2(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$
1. Update $\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta$ (forward calls) [a]:

$$\max_{\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta} \mathcal{F}_\mu = -\frac{1}{2\sigma^2} |\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta)|^2 - \frac{1}{2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_0)^T \boldsymbol{S}_0^{-1} (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_0) + \log p(\boldsymbol{\mu}_d)$$

2.1 Update $\boldsymbol{W}$ (No forward calls):

$$\max_{\boldsymbol{W}} \mathcal{F}_W = -\frac{1}{2\sigma^2} \boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W} : (\boldsymbol{C}_{yy} - \sigma_d^2 \boldsymbol{I}) + \frac{1}{\sigma^2} \boldsymbol{G}_\theta^T \boldsymbol{G}_d \boldsymbol{W} : \boldsymbol{C}_{\theta y}$$

2.2 Update $q(\boldsymbol{\eta}_\theta, \boldsymbol{y}) \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$, $q(\boldsymbol{\eta}_d) \equiv \mathcal{N}(\boldsymbol{0}, \sigma_d^2(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$ (No forward calls):

$$\boldsymbol{C}^{-1} = \begin{bmatrix} \boldsymbol{C}_{\theta\theta} & \boldsymbol{C}_{\theta y} \\ \boldsymbol{C}_{\theta y}^T & \boldsymbol{C}_{yy} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{\sigma^2}\boldsymbol{G}_\theta^T \boldsymbol{G}_\theta + \boldsymbol{S}_0^{-1} & \frac{1}{\sigma^2}\boldsymbol{G}_\theta^T \boldsymbol{G}_d \\ sym. & \frac{1}{\sigma^2}\boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W} + \sigma_{y0}^{-2}\boldsymbol{I} \end{bmatrix}$$

$$\frac{1}{\sigma_d^2} = \frac{1}{\sigma_{y0}^2} + \frac{1}{(dim(\boldsymbol{d}) - dim(\boldsymbol{y}))} \frac{1}{\sigma^2} (tr(\boldsymbol{G}_d^T \boldsymbol{G}_d) - tr(\boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W}))$$

[a] Assuming $\pi(\boldsymbol{\theta}) \equiv \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_0)$

VB-EM Algorithm:

$$\mathcal{F}(\boldsymbol{P}, q) = E_q(\log U(\boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta, \boldsymbol{\mu}_d + \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\eta}_d)) + E_q(\log \pi(\boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta) p(\boldsymbol{y}) p(\boldsymbol{\eta}_d)) - E_q(\log q)$$

0. Initialize with $p(\boldsymbol{y}) \equiv \mathcal{N}(\boldsymbol{0}, \sigma_{y0}^2 \boldsymbol{I})$, $p(\boldsymbol{\eta}_d) \equiv \mathcal{N}(0, \sigma_{y0}^2(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$
1. Update $\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta$ (forward calls) [a]:

$$\max_{\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta} \mathcal{F}_\mu = -\frac{1}{2\sigma^2} |\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta)|^2 - \frac{1}{2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_0)^T \boldsymbol{S}_0^{-1} (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_0) + \log p(\boldsymbol{\mu}_d)$$

2.1 Update $\boldsymbol{W}$ (No forward calls):

$$\max_{\boldsymbol{W}} \mathcal{F}_W = -\frac{1}{2\sigma^2} \boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W} : (\boldsymbol{C}_{yy} - \sigma_d^2 \boldsymbol{I}) + \frac{1}{\sigma^2} \boldsymbol{G}_\theta^T \boldsymbol{G}_d \boldsymbol{W} : \boldsymbol{C}_{\theta y}$$

2.2 Update $q(\boldsymbol{\eta}_\theta, \boldsymbol{y}) \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$, $q(\boldsymbol{\eta}_d) \equiv \mathcal{N}(\boldsymbol{0}, \sigma_d^2(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T))$ (No forward calls):

$$\boldsymbol{C}^{-1} = \begin{bmatrix} \boldsymbol{C}_{\theta\theta} & \boldsymbol{C}_{\theta y} \\ \boldsymbol{C}_{\theta y}^T & \boldsymbol{C}_{yy} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} \boldsymbol{G}_\theta^T \boldsymbol{G}_\theta + \boldsymbol{S}_0^{-1} & \frac{1}{\sigma^2} \boldsymbol{G}_\theta^T \boldsymbol{G}_d \\ sym. & \frac{1}{\sigma^2} \boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W} + \sigma_{y0}^{-2} \boldsymbol{I} \end{bmatrix}$$

$$\frac{1}{\sigma_d^2} = \frac{1}{\sigma_{y0}^2} + \frac{1}{(dim(\boldsymbol{d}) - dim(\boldsymbol{y}))} \frac{1}{\sigma^2} (tr(\boldsymbol{G}_d^T \boldsymbol{G}_d) - tr(\boldsymbol{W}^T \boldsymbol{G}_d^T \boldsymbol{G}_d \boldsymbol{W}))$$

[a] Assuming $\pi(\boldsymbol{\theta}) \equiv \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_0)$

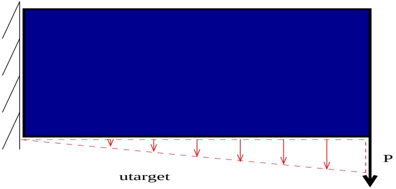
Shape/topology optimization:

$$\min_{\boldsymbol{d}} \quad |\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{d})|^2$$

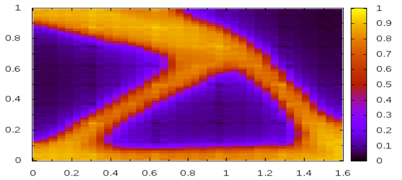
such that:

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

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



(a) domain



(b) $\textit{compliance}(\boldsymbol{d}) \approx 55$

Figure : Adjoint-based gradient optimization - $O(100)$ forward runs

Stochastic topology optimization

Shape/topology optimization:

$\boldsymbol{K}(\boldsymbol{d}, \boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{d}, \boldsymbol{\theta}) = \boldsymbol{b}$ (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: $\max_{\boldsymbol{d}} \int e^{-\frac{1}{2}|\boldsymbol{u}(\boldsymbol{d},\boldsymbol{\theta})-\boldsymbol{u}_0|^2}\, \pi(\boldsymbol{\theta})d\boldsymbol{\theta}$

such that:

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

Variational Inference - Constraints

Shape/topology optimization:

$$
\begin{array}{ll}
\min_{\boldsymbol{d}} & |\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{d})|^2 \\
\text{such that:} & \\
 & \boldsymbol{K}(\boldsymbol{d})\boldsymbol{u}(\boldsymbol{d}) = \boldsymbol{b} \quad \text{(governing equation)} \\
 & \int d(\boldsymbol{x})\, d\boldsymbol{x} = V_0, \quad \text{(volume fraction)} \\
 & d(\boldsymbol{x}) \in [0, 1]
\end{array}
\qquad
d(\boldsymbol{x}) = \begin{cases} 1, & \textit{material} \\ 0, & \textit{void} \end{cases}
$$

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

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

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

Numerical Illustration

Stochastic topology optimization

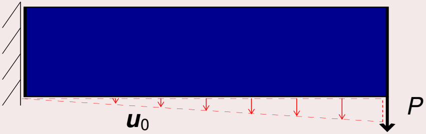
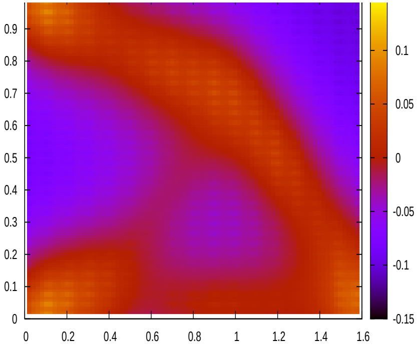


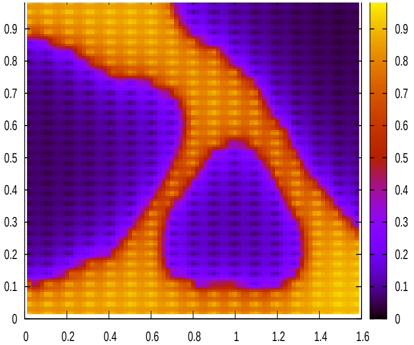
Figure : Problem Domain

- $dim(\boldsymbol{d}) = 2048$ (design variables), $dim(\boldsymbol{\theta}) = 2048$ (random variables)
- $\log \boldsymbol{\theta} \sim N(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$
    - $C.O.V.[\theta_i] = 0.25$
    - $\boldsymbol{\Sigma}_\theta = Cov[\log \theta(\boldsymbol{x}_i), \log \theta(\boldsymbol{x}_j)] = e^{-|\boldsymbol{x}_i - \boldsymbol{x}_j|/l_0}$
    - $l_0 = 0.1$ (correlation length)
    - target:
      $\boldsymbol{u}_0 = [-6, 25, -12.5, -18.75, -25., -31.25, -37.5, -43.75, -50]^T \times 10^{-3}$,
      $\sigma^2 = 5 \times 10^{-3}$.
- Volume constraint: $\int d(\boldsymbol{x})\, d\boldsymbol{x} = 0.4$

Numerical Illustration



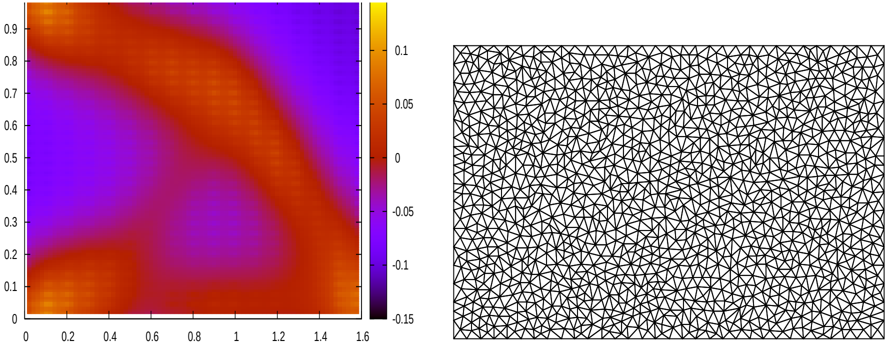
(a) $\boldsymbol{\mu}_\theta$



(b) $\boldsymbol{\mu}_d$ (Volume fraction=0.4)

Figure : Computational Cost: 46 forward runs (output and gradient computation)

Numerical Illustration

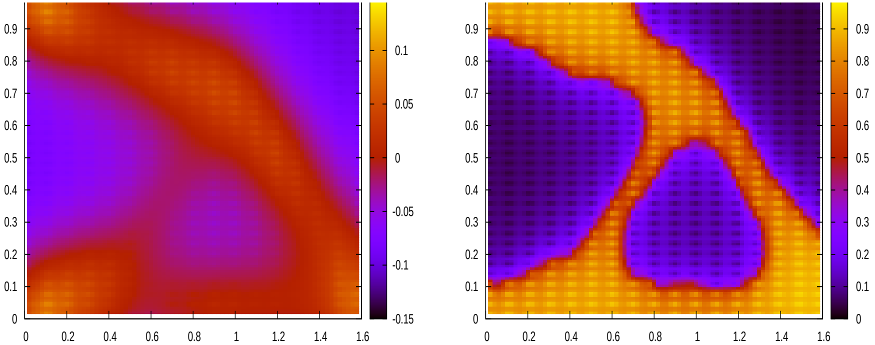


(a) $\boldsymbol{\mu}_\theta$

(b) $\boldsymbol{\mu}_d$ (Volume fraction=0.4)

Figure : Computational Cost: 46 forward runs (output and gradient computation)

Numerical Illustration



(a) $\boldsymbol{\mu}_\theta$

(b) $\boldsymbol{\mu}_d$ (Volume fraction=0.4)

Figure : Computational Cost: 46 forward runs (output and gradient computation)

Numerical Illustration

$$\mathcal{F}(\boldsymbol{P}, q) = -\frac{1}{2\sigma^2}|\boldsymbol{u}_0 - \boldsymbol{u}(\boldsymbol{\mu}_d, \boldsymbol{\mu}_\theta)|^2 - \frac{1}{2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_0)^T \boldsymbol{S}_0^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_0) + \frac{1}{2}\log|\boldsymbol{C}| + \frac{dim(\boldsymbol{d}) - dim(\boldsymbol{y})}{2}\log\sigma_d^2$$

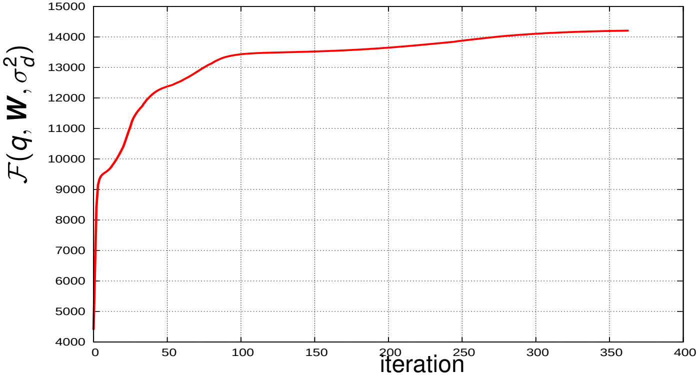


Figure : Evolution of VB lower-bound $\mathcal{F}(q, \boldsymbol{W}, \sigma_d^2)$ (No forward solves)

Numerical Illustration

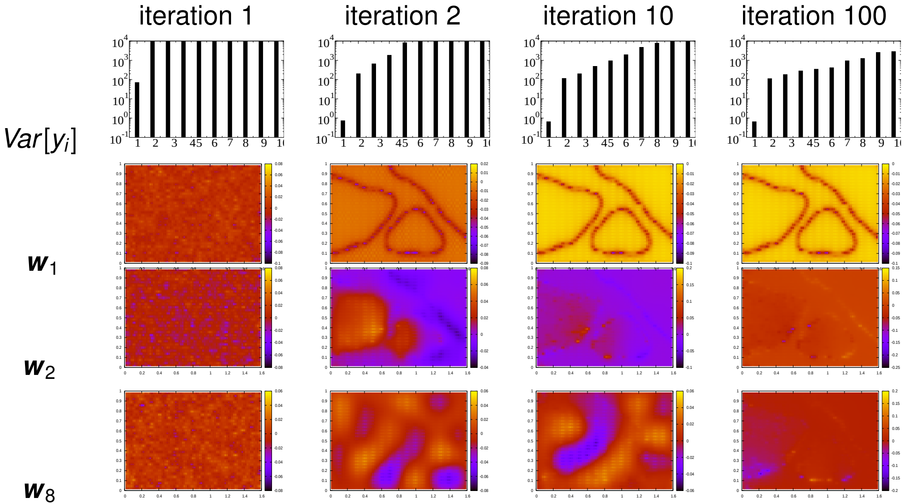
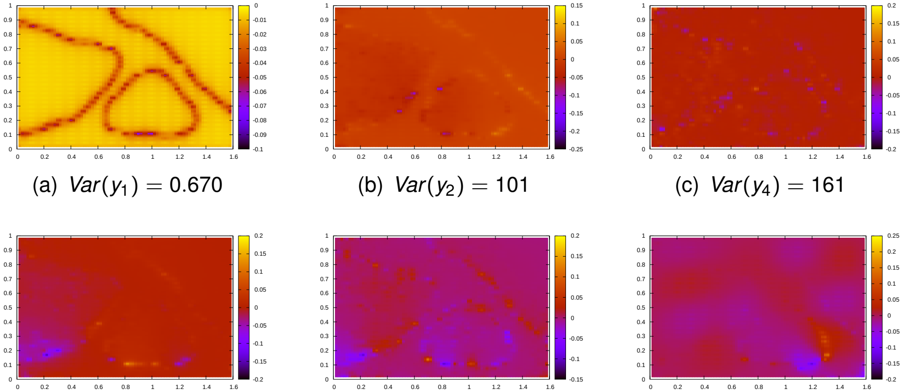


Table : Evolution of basis vectors in $\boldsymbol{W}$

Numerical Illustration

$$\underbrace{\boldsymbol{d}}_{2048\times 1} = \boldsymbol{\mu}_d + \underbrace{\boldsymbol{W}}_{2048\times 20} \underbrace{\boldsymbol{y}}_{20\times 1} + \boldsymbol{\eta}_d$$

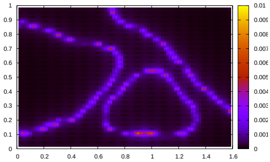


(a) $Var(y_1) = 0.670$
(b) $Var(y_2) = 101$
(c) $Var(y_4) = 161$
(d) $Var(y_8) = 305$
(e) $Var(y_{12}) = 2728$
(f) $Var(y_{14}) = 22925$

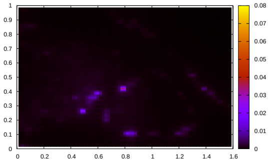
Figure : Learned dictionary of most *sensitive* directions $\boldsymbol{W}$

Numerical Illustration

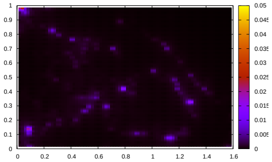
$$\underbrace{\boldsymbol{d}}_{2048\times 1} = \mu_d + \underbrace{\boldsymbol{W}}_{2048\times 20}\underbrace{\boldsymbol{y}}_{20\times 1}$$



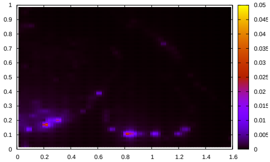
(a) $Var(y_1) = 0.670$



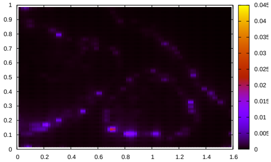
(b) $Var(y_2) = 101$



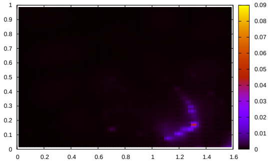
(c) $Var(y_4) = 161$



(d) $Var(y_8) = 305$



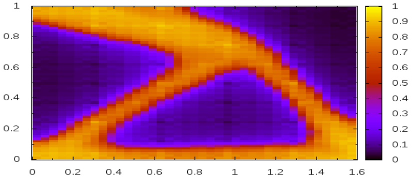
(e) $Var(y_{12}) = 2728$



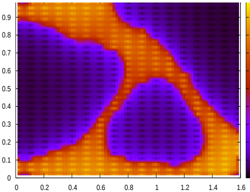
(f) $Var(y_{14}) = 22925$

Figure : Learned dictionary of most *sensitive* directions $\boldsymbol{W}$. Plotted $\{W_{i,j}^2\}_{i=1}^{2048}$, $j = 1 \div 20$

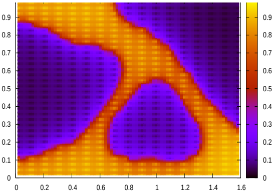
Numerical Illustration



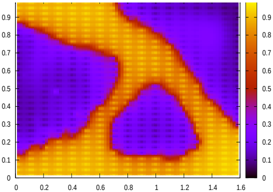
(a) deterministic



(b) mean-st.dev.*



(c) mean ($\boldsymbol{\mu}_d$)

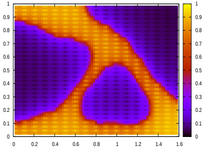


(d) mean+st.dev.*

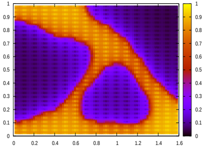
Figure : Deterministic vs. Stochastic (Variational Bayes)

$$\frac{V(\boldsymbol{d})}{V(\boldsymbol{d}^{opt})} = 0.95$$

Sample design 1



Sample design 2



Sample design 3

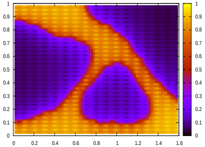


Table : Sample Design

Numerical Illustration

$$\frac{V(\boldsymbol{d})}{V(\boldsymbol{d}^{opt})} = 0.95 \qquad \frac{V(\boldsymbol{d})}{V(\boldsymbol{d}^{opt})} = 0.85$$

Sample design 1

Sample design 2

Sample design 3

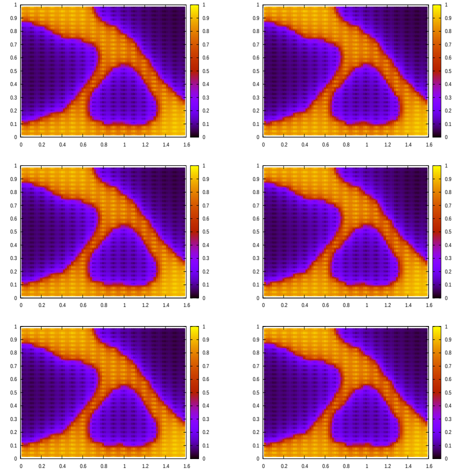


Table : Sample Design

$$\frac{V(\boldsymbol{d})}{V(\boldsymbol{d}^{opt})} = 0.95 \qquad \frac{V(\boldsymbol{d})}{V(\boldsymbol{d}^{opt})} = 0.85 \qquad \frac{V(\boldsymbol{d})}{V(\boldsymbol{d}^{opt})} = 0.75$$

Sample design 1

Sample design 2

Sample design 3

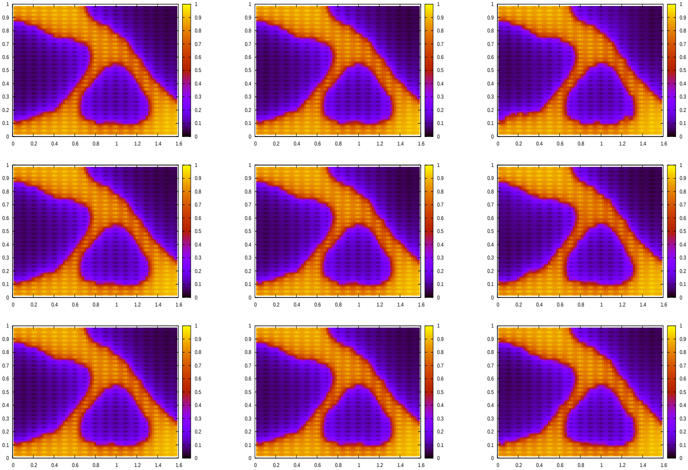


Table : Sample Design

Numerical Illustration

Convergence with reduced dimension $dim(\boldsymbol{y})$

$$err(dim(\boldsymbol{y})) = \frac{KL(q(\boldsymbol{\mu}_d + \boldsymbol{W}\boldsymbol{y}, \boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta)||p(\boldsymbol{\mu}_d + \boldsymbol{W}\boldsymbol{y}, \boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta))}{H(q(\boldsymbol{\mu}_d + \boldsymbol{W}\boldsymbol{y}, \boldsymbol{\mu}_\theta + \boldsymbol{\eta}_\theta))}$$

$dim(\boldsymbol{y})$	$err(dim(\boldsymbol{y}))$
5	5.1×10^{-3}
10	4.5×10^{-3}
15	2.9×10^{-3}
20	2.7×10^{-3}

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: MoG to capture non-Gaussian and multi-modal design objectives