Simulation-based, high-dimensional stochastic optimization

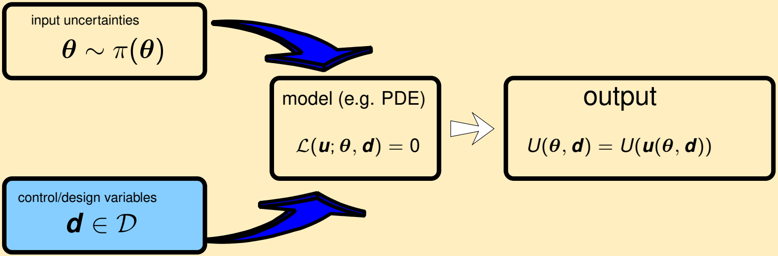
application in robust topology optimization under large material uncertainties

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

MascotNum Workshop on Computer Experiments and Meta-models for Uncertainty Quantification
ETH Zurich, April 24 2014

Motivation

Uncertainty quantification



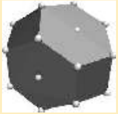
- uncertainties $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$, $n_\theta >> 1$
- design/control variables $\boldsymbol{d} \in \mathcal{D} \subset \mathbb{R}^{n_d}$, $n_d >> 1$
- Goal - Syochastic Optimization: Can we *efficiently* optimize w.r.t $\boldsymbol{d}$ and some output utility $U(\boldsymbol{\theta},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

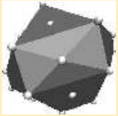
$$V(\boldsymbol{d}) = \int U(\boldsymbol{\theta}, \boldsymbol{d}) \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}$$

Motivation

Designing materials at the micro/atomistic level

$$V(\boldsymbol{d}) = \int U(\boldsymbol{\theta}) \frac{e^{-\beta W(\boldsymbol{\theta};\boldsymbol{d})}}{Z} \, d\boldsymbol{\theta}$$





- $V(\boldsymbol{d})$: macroscopic/thermodynamic property
- $\boldsymbol{d}$: design parameters (e.g. potential form, order of interactions)
- $W(\boldsymbol{\theta};\boldsymbol{d})$: interatomic potential
- $\boldsymbol{\theta}$: atomistic configuration

Stochastic topology optimization:

- Controlling statistics of the random material properties (Sternfels, PSK 2011).

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

- *Controlling geometry/spatial distribution of materials with random properties.*

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

Motivation

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

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

- Why is this interesting?
  1) Suppose $U(\theta, \boldsymbol{d}) = 1_{\mathcal{A}}(\theta, \boldsymbol{d})$ is the indicator function of some response event $\mathcal{A}$, e.g. failure, then:

$$min\ or\ max V(\boldsymbol{d}) \equiv \text{min or max the } \underline{\text{probability of failure}}$$

Motivation

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

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

- Why is this interesting?
  2) Suppose $U(\theta, \boldsymbol{d}) = \| \boldsymbol{u}(\theta, \boldsymbol{d}) - \boldsymbol{u}_{target} \|$ where $\boldsymbol{u}_{target}$ is a *desired* response, then:

$$min V(\boldsymbol{d}) \equiv \text{stochastic control}$$

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

Approach

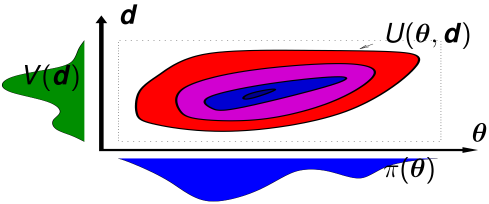
Optimize the *expected* utility $V(\boldsymbol{d})$:

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

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

Approach

We discuss two alternatives:

1. Adaptive Sequential Monte Carlo
2. Variational Bayes

Adaptive Sequential Monte Carlo

Sequential Monte Carlo:

A combination of Importance sampling and MCMC that provides a particulate approximation $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{d}^{(i)}), \boldsymbol{W}^{(i)}\}_{i=1}^{N}$ (Doucet 2001):

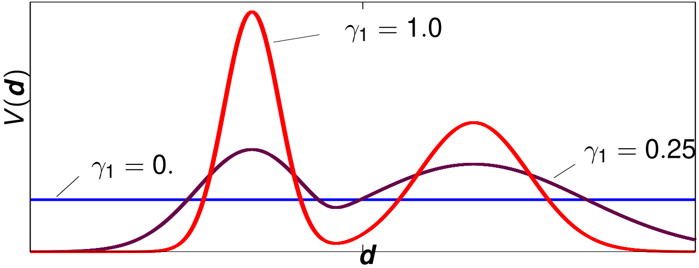
$$p(\boldsymbol{\theta}, \boldsymbol{d}) \propto U(\boldsymbol{\theta}, \boldsymbol{d})\pi(\boldsymbol{\theta}) \approx \sum_{i=1}^{N} \boldsymbol{W}^{(i)} \delta_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{\theta}) \delta_{\boldsymbol{d}^{(i)}}(\boldsymbol{d})$$

almost sure convergence of expectations of p-measurable functions

Adaptive Sequential Monte Carlo

We operate on a *sequence* of distributions (from simple to complicated) (Amzal et al 2003, Johansen et al 2006, Kück et al. 2006):

$$p_\gamma(\theta, \boldsymbol{d}) \propto U^\gamma(\theta, \boldsymbol{d})\pi(\theta), \quad \gamma \in [0, 1]$$



Adaptive Sequential Monte Carlo

We operate on a *sequence* of distributions (from simple to complicated):

$$p_\gamma(\boldsymbol{\theta}, \boldsymbol{d}) \propto U^\gamma(\boldsymbol{\theta}, \boldsymbol{d})\pi(\boldsymbol{\theta}), \quad \gamma \in [0, 1]$$

Adaptive SMC (PSK, *J. Comp. Phys.* 2009, Sternfels, PSK, *Int. J. Mult. Comp. Eng* 2010):

- If γ increases slowly, we do too many forward runs (cost)
- If γ increases too fast we loose accuracy (accuracy)

Adaptive SMC

- Generate initial particle population $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{d}^{(i)}), W^{(i)}\}_{i=1}^N$ from $\pi_{\gamma=0} \equiv p(\boldsymbol{\theta})$. Set $\gamma_{current} = 0$.
- Iterate until $\gamma_{current} = 1$.
  - **Reweight**: Find γ_{next} based on the relative reduction in the Effective Sample Size ESS :

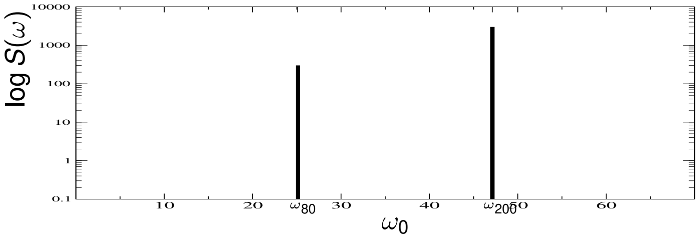
$$w^{(i)} = W^{(i)} \frac{\pi_{\gamma_{next}}(\boldsymbol{\theta}^{(i)}, \boldsymbol{d}^{(i)})}{p_{\gamma_{current}}(\boldsymbol{\theta}^{(i)}, \boldsymbol{d}^{(i)})}, \quad ESS = \frac{(\sum_{i=1}^N w^{(i)})^2}{\sum_{i=1}^N (w^{(i)})^2}$$

  - **Resample:** If ESS drops below a specified threshold (typically $N/2$) , then resample.
  - **Rejuvenate:** Move particles using a $p_{\gamma_{next}}$-invariant MCMC kernel:
    - We employed a Metropolis-adjusted Langevin (MALA) sampler which implies calculation of U as well as derivatives $\frac{\partial f}{\partial \boldsymbol{\theta}}$
    - These were calculated using *adjoint formulations*
  - Set $\gamma_{current} = \gamma_{next}$

Verification

$$\ddot{x}(t) + \omega_0^2 x(t) = f(t)$$

- uncertainties $\boldsymbol{\theta} \sim U(0, 2\pi)^{200}$: $f(t) = \sum_{k=1}^{n_\theta=200} \sqrt{2S(\omega_n)\Delta\omega_k} \cos(\omega_k t + \theta_k)$
- design variable $\boldsymbol{d} = \omega_0$
- utility $U(\boldsymbol{\theta}, \boldsymbol{d}) = e^{\frac{1}{T}\int_0^T x^2(t)dt}$

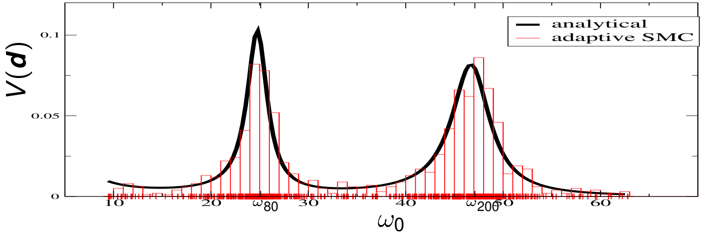


Verification

$$\ddot{x}(t) + \omega_0^2 x(t) = f(t)$$

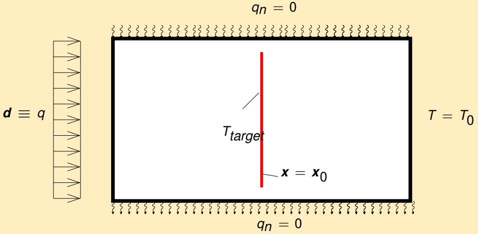
- uncertainties $\boldsymbol{\theta} \sim U(0, 2\pi)^{200}$: $f(t) = \sum_{k=1}^{n_\theta=200} \sqrt{2S(\omega_n)\Delta\omega_k} \cos(\omega_k t + \theta_k)$
- design variable $\boldsymbol{d} = \omega_0$
- utility $U(\boldsymbol{\theta}, \boldsymbol{d}) = e^{\frac{1}{T}\int_0^T x^2(t)dt}$

Sampling in $200 + 1 = 201$ *dimensions*

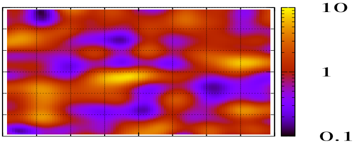


Heat diffusion in a random medium

$$\nabla \cdot (-\lambda(\boldsymbol{x}) \nabla T(\boldsymbol{x})) = 0$$



- uncertainties $\boldsymbol{\theta} \in \mathbb{R}^{1,000}$: $\lambda(\boldsymbol{x}) = h\left(\sum_{k=1}^{1,000} \theta_k \phi_k(\boldsymbol{x})\right)$

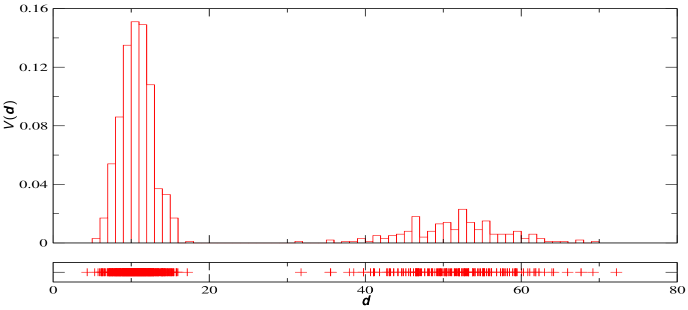


- control variable(s) $\boldsymbol{d}$: flux on the left

$\| T(\boldsymbol{x}_0; \boldsymbol{\theta}, \boldsymbol{d}) - T^{(1)}_{target} \|^2$ $\| T(\boldsymbol{x}_0; \boldsymbol{\theta}, \boldsymbol{d}) - T^{(2)}_{target} \|^2$

Controlling the input of random systems

Sampling in $1,000 + 1 = 1,001$ *dimensions*



Controlling the input of random systems

- What if we are really interested in the *global* maximum?
- State augmentation (Brooks et al. 1995):

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M, \boldsymbol{d}) \propto \prod_{m=1}^{M} U(\boldsymbol{\theta}_m, \boldsymbol{d}) \pi(\boldsymbol{\theta}_m)$$

- Note that the *marginal* w.r.t. the design variables $\boldsymbol{d}$ is:

$$\int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M, \boldsymbol{d}) d\boldsymbol{\theta}_{1:M} \propto V^M(\boldsymbol{d})$$

- The adaptive SMC scheme discussed can be readily adjusted

Controlling the input of random systems

State augmentation

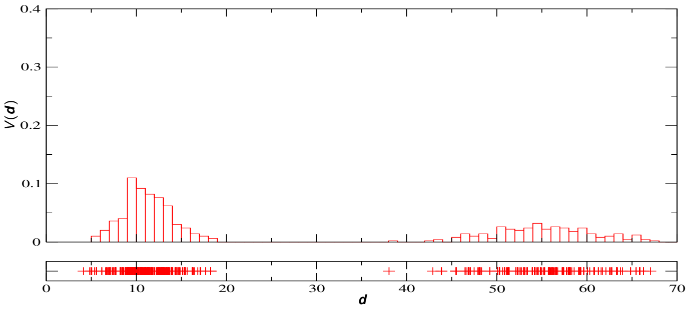


Figure: $M = 1$: Sampling in $1,001$ dimensions

Controlling the input of random systems

State augmentation

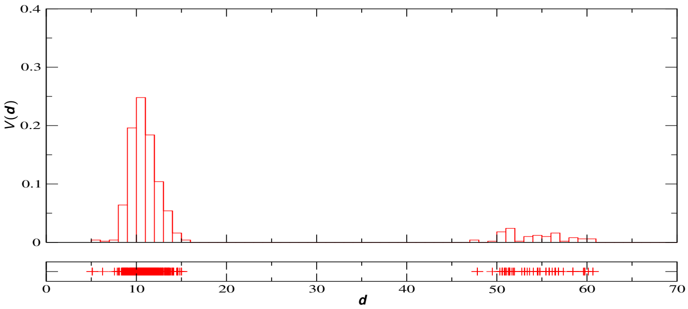


Figure: $M = 3$: Sampling in $3,001$ dimensions

Controlling the input of random systems

State augmentation

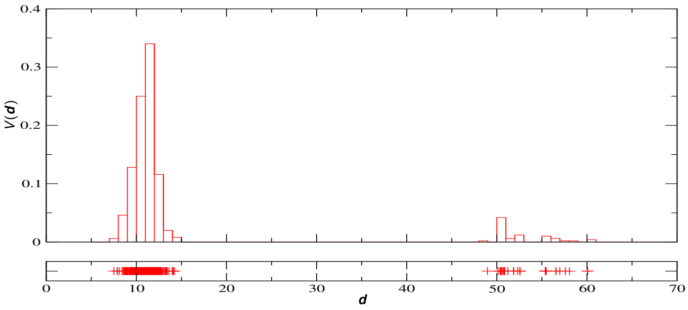


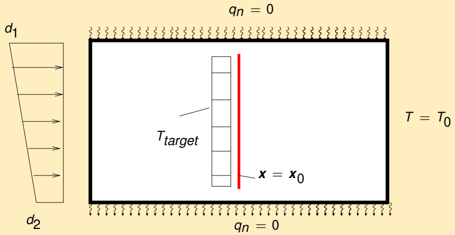
Figure: $M = 5$: Sampling in $5,001$ dimensions

Controlling the input of random systems

- What if we had more design variables $\boldsymbol{d}$?

Heat diffusion in a random medium

$$\nabla \cdot (-\lambda(\boldsymbol{x}) \nabla T(\boldsymbol{x})) = 0$$



Controlling the input of random systems

Two design variables

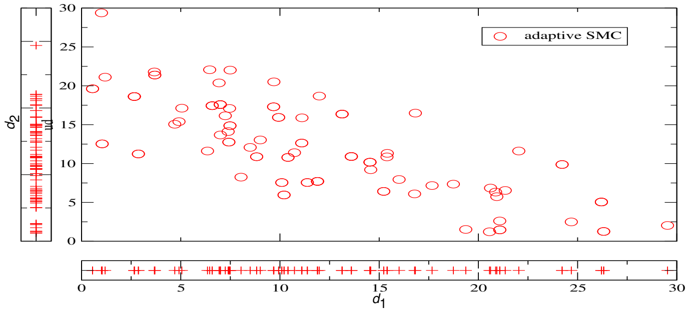


Figure: $M = 1$: Sampling in 1,002 dimensions

Controlling the input of random systems

Two design variables - State augmentation

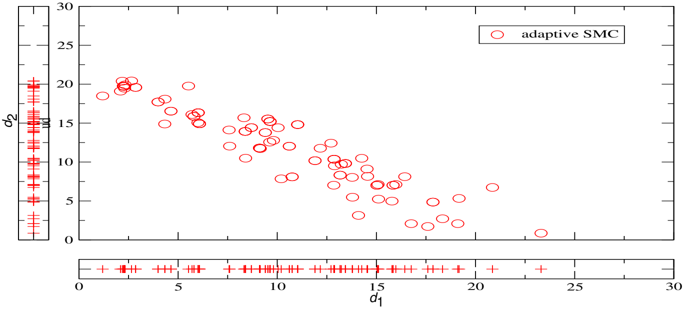
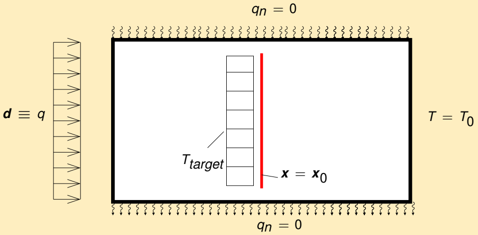


Figure: $M = 5$: Sampling in 5,002 dimensions

Approximate solvers for reducing cost

Heat diffusion in a random medium

$$\nabla \cdot (-\lambda(\boldsymbol{x}) \nabla T(\boldsymbol{x})) = 0$$



- utility $U(\theta, \boldsymbol{d}) = e^{-\frac{\| T(\boldsymbol{x}_0; \boldsymbol{\theta}, \boldsymbol{d}) - T_{target} \|^2}{2\sigma^2}}$
  $(T_{target} = 35)$

Approximate solvers for reducing cost

One design variable

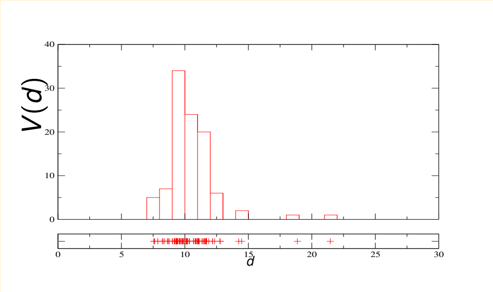


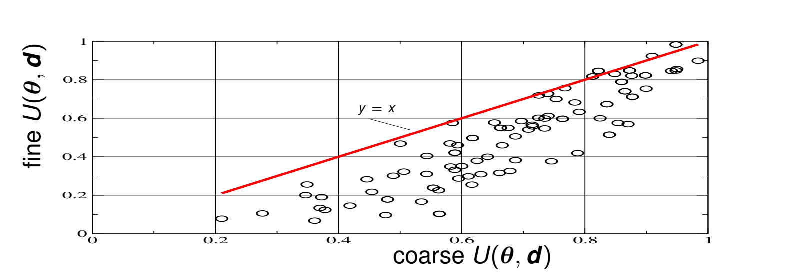
Figure: $M = 1$: Sampling in $1,001$ dimensions

- cost: $7,200$ calls to the forward model (particles $N = 100$, iterations 33)
- The simulation is *embarrassingly parallelizable* but still the cost is quite significant.

- Can we use *less-expensive* but *less-accurate* forward models?

Approximate solvers for reducing cost

Coarse (10 × 10) vs. Fine (200 × 200)



Approximate solvers for reducing cost

Adaptive SMC

- Sequence 1 (use the coarse model to drive you close to the solution):

$$p_{\gamma_1}(\boldsymbol{\theta}, \boldsymbol{d}) \propto U_{coarse}^{\gamma_1}(\boldsymbol{\theta}, \boldsymbol{d})\pi(\boldsymbol{\theta}), \quad \gamma_1 \in [0,1]$$

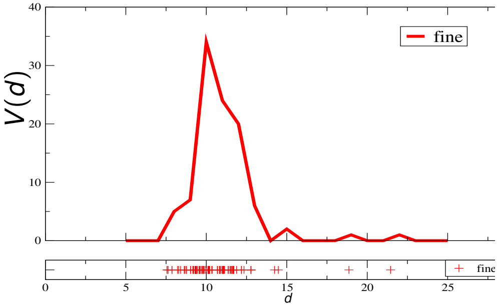
- Sequence 2 (correct for the discrepancies between coarse and fine models):

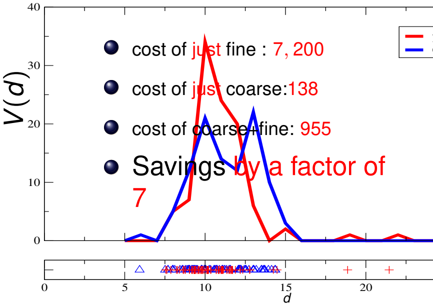
$$p_{\gamma_2}(\boldsymbol{\theta}, \boldsymbol{d}) \propto U_{coarse}^{1-\gamma_2}(\boldsymbol{\theta}, \boldsymbol{d}) U_{fine}^{\gamma_2}(\boldsymbol{\theta}, \boldsymbol{d})\pi(\boldsymbol{\theta}), \quad \gamma_2 \in [0,1]$$

- More levels can readily be added
- It suffices that the *coarse* model drives the sampling in the “right direction”. The less approximate it is the larger the savings.

Approximate solvers for reducing cost

One design variable - Sampling in 1, 001 dimensions



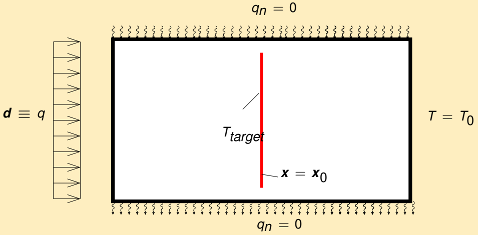


- cost of just fine : 7, 200
- cost of just coarse:138
- cost of coarse+fine: 955
- Savings by a factor of 7

Approximate solvers for reducing cost

Heat diffusion in a random medium

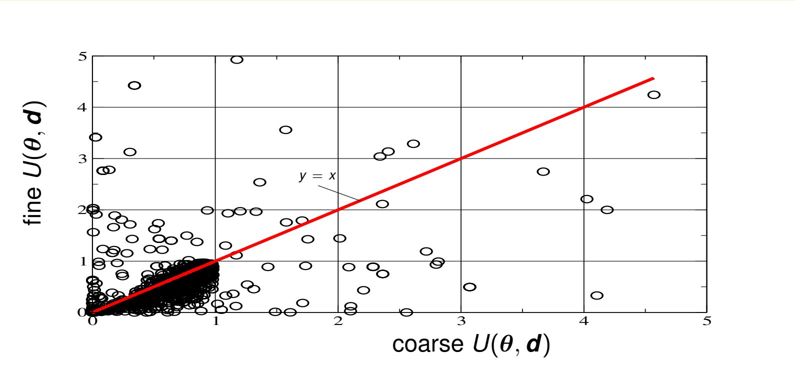
$$\nabla \cdot (-\lambda(\boldsymbol{x}) \nabla T(\boldsymbol{x})) = 0$$



- utility $U(\theta, \boldsymbol{d}) = e^{-\frac{\| T(\boldsymbol{x}_0; \boldsymbol{\theta}, \boldsymbol{d}) - T^{(1)}_{target} \|^2}{2\sigma^2}} + 6e^{-\frac{\| T(\boldsymbol{x}_0; \boldsymbol{\theta}, \boldsymbol{d}) - T^{(2)}_{target} \|^2}{2\sigma^2}}$
  $\left(T^{(1)}_{target} = 35,\ T^{(2)}_{target} = 70\right)$

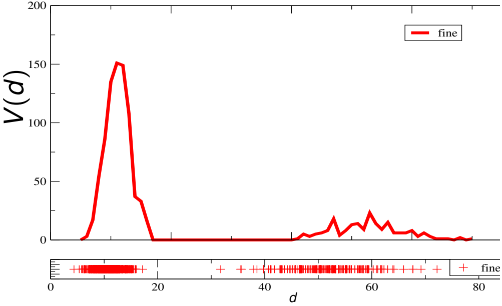
Approximate solvers for reducing cost

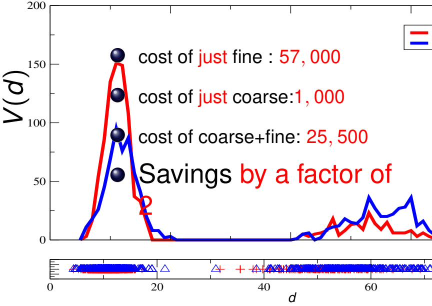
Coarse (10 × 10) vs. Fine (200 × 200)



Approximate solvers for reducing cost

One design variable - Sampling in $1,001$ dimensions





- cost of just fine : $57,000$
- cost of just coarse:$1,000$
- cost of coarse+fine: $25,500$
- Savings by a factor of 2

Deterministic topology optimization

Shape/topology optimization:

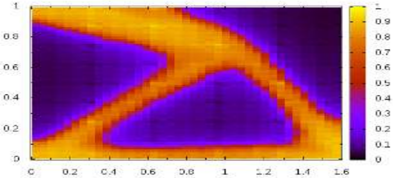
$$\min_{\boldsymbol{d}} \quad compliance(\boldsymbol{d}) = \boldsymbol{b}^T \boldsymbol{u}(\boldsymbol{d})$$

such that:

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

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

(a) domain



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

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

Stochastic topology optimization

Shape/topology optimization:

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

such that:

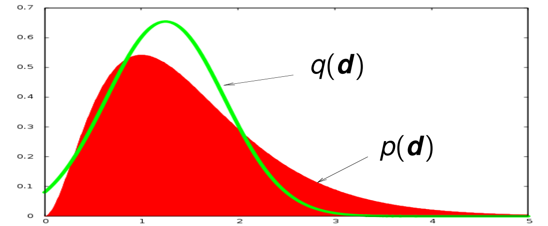
$\boldsymbol{K}(\boldsymbol{d}, \boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{d}, \boldsymbol{\theta}) = \boldsymbol{b}$ (governing equation)
$\int d(\boldsymbol{x})\, d\boldsymbol{x} = V_0,$ (volume fraction)
$d(\boldsymbol{x}) \in [0, 1]$
$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$

Variational Inference

Our goal is to infer:

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

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

- 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 :

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

  - Difficult term: $E_q(\log U(\boldsymbol{\theta}, \boldsymbol{d}))$
  - Easy/Tractable terms: $E_q(\log \pi(\boldsymbol{\theta}))$, $E_q(\log q)$

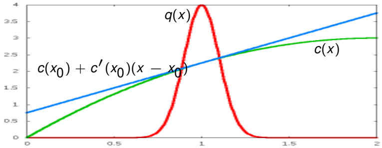
Variational Inference

- Assumption 1: Mean field approximation (Wainwright & Jordan, 2008):

$$q(\theta, \boldsymbol{d}) = q_1(\theta) q_2(\boldsymbol{d})$$

- Assumption 2: Family of approximating distributions $\boldsymbol{q} \in \mathcal{Q}$ are *multivariate Gaussians* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$.
- Assumption 3: Linearization - E.g. $U(\boldsymbol{\theta}, \boldsymbol{d}) = e^{-\frac{1}{2}|c(\boldsymbol{\theta}, \boldsymbol{d}) - c_{target}|^2}$:

$$\begin{aligned} c(\boldsymbol{\theta}, \boldsymbol{d}) \quad & \approx c(\boldsymbol{\theta}_0, \boldsymbol{d}_0) \\ & + \boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \boldsymbol{d}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ & + \boldsymbol{G}_{\boldsymbol{d}}(\boldsymbol{\theta}_0, \boldsymbol{d}_0)(\boldsymbol{d} - \boldsymbol{d}_0) \end{aligned}$$



where $\boldsymbol{G}_{\boldsymbol{\theta}} = \frac{\partial c}{\partial \boldsymbol{\theta}}$ and $\boldsymbol{G}_{\boldsymbol{d}} = \frac{\partial c}{\partial \boldsymbol{d}}$ available with minimal cost from adjoint-PDE.

Variational Inference

Algorithm:

$$\boxed{\mathcal{F}(q) = E_q(\log U(\theta, \boldsymbol{d})) + E_q(\log \pi(\boldsymbol{\theta})) - E_q(\log q)}$$

0. Initialize $q(\boldsymbol{\theta}) \equiv \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{S}_{\boldsymbol{\theta}})$ and $q(\boldsymbol{d}) \equiv \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{d}}, \boldsymbol{S}_{\boldsymbol{d}})$
1. Set $\theta_0 = \boldsymbol{\mu}_{\boldsymbol{\theta}}$, $\boldsymbol{d}_0 = \boldsymbol{\mu}_{\boldsymbol{d}}$ and linearize $c(\theta, \boldsymbol{d})$ around $(\theta_0, \boldsymbol{d}_0)$.
2. Fixed-point iterations for $q(\theta), q(\boldsymbol{d})$ [a]:

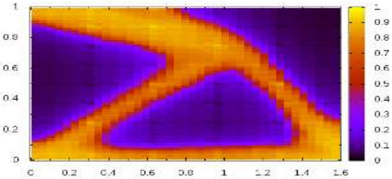
$$\begin{aligned}
\boldsymbol{S}_{\boldsymbol{d}}^{-1} &= \boldsymbol{G}_{\boldsymbol{d}}^T \boldsymbol{G}_{\boldsymbol{d}} \\
\boldsymbol{S}_{\boldsymbol{\theta}}^{-1} &= \boldsymbol{G}_{\boldsymbol{\theta}}^T \boldsymbol{G}_{\boldsymbol{\theta}} + \hat{\boldsymbol{S}}^{-1} \\
\boldsymbol{S}_{\boldsymbol{d}}^{-1} \boldsymbol{\mu}_{\boldsymbol{d}} &= \boldsymbol{G}_{\boldsymbol{d}}^T (c_0 - c_{target} - \boldsymbol{G}_{\boldsymbol{d}} \boldsymbol{d}_0) + \boldsymbol{G}_{\boldsymbol{\theta}} (\boldsymbol{\mu}_{\boldsymbol{\theta}} - \theta_0) \\
\boldsymbol{S}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}} &= \boldsymbol{G}_{\boldsymbol{\theta}}^T (c_0 - c_{target} - \boldsymbol{G}_{\boldsymbol{\theta}} \theta_0) + \boldsymbol{G}_{\boldsymbol{d}} (\boldsymbol{\mu}_{\boldsymbol{d}} - \boldsymbol{d}_0) + \hat{\boldsymbol{S}}^{-1} \hat{\boldsymbol{\mu}}
\end{aligned}$$

3. Goto 1. until convergence

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

Variational Inference

- What about high-dimensional ***d*** (or θ)?
  - high-dimensional Gaussian
  - quality of KL-divergence decays as measure of proximity
- What about any regularization?



Sparse Variational Inference

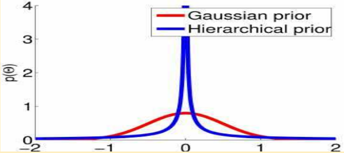
Sparse Bayesian Learning

$$\underbrace{\boldsymbol{d}}_{N\times 1} = \underbrace{\boldsymbol{W}}_{N\times n} \underbrace{\boldsymbol{y}}_{n\times 1}$$

where $\boldsymbol{W}$ contains basis/features/vocabulary

- Hierarchical heavy-tailed prior:

$$p(y_j|\tau_j) \equiv \mathcal{N}(0, \tau_j^{-1})$$
$$p(\tau_j) \equiv Gamma(\alpha, \beta), \quad j = 1, \ldots, n$$



- Automatic Relevance Determination priors (ARD, MacKay 1994)): $\tau_j \to \infty$ then $y_j \to 0$ (i.e. feature j is inactive)
- Closely related to LASSO (Tibshirani 1996), Compressive Sensing (Candés et al 2006, Donoho et al 2006)

Sparse Variational Inference

Variational Inference

$$\mathcal{F}(q, \boldsymbol{W}) = E_q\left(\log \frac{U(\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\tau})}\right) + E_q\left(\log p(\boldsymbol{y}|\boldsymbol{\tau})p(\boldsymbol{\tau})\right)$$

where $q(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\tau}) = q(\boldsymbol{\theta})q(\boldsymbol{y})q(\boldsymbol{\tau})$

Update equations for $q(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\tau})$:

$$q(\tau_j) \equiv Gamma(\alpha_j, \beta_j), \alpha_j = \alpha + \frac{1}{2},\ \beta_j = \beta + \frac{1}{2} E_{q(\boldsymbol{y})}(y_j^2)$$

$$\boldsymbol{S}_{\boldsymbol{y}}^{-1} = \boldsymbol{W}^T \boldsymbol{G}_{\boldsymbol{d}}^T \boldsymbol{G}_{\boldsymbol{d}} \boldsymbol{W} + E_{q(\boldsymbol{\tau})}(\boldsymbol{T}), \quad \boldsymbol{T} = diag(\tau_j)$$

$$\boldsymbol{S}_{\boldsymbol{\theta}}^{-1} = \boldsymbol{G}_{\boldsymbol{\theta}}^T \boldsymbol{G}_{\boldsymbol{\theta}} + \hat{\boldsymbol{S}}^{-1}$$

$$\boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{W}^T \boldsymbol{G}_{\boldsymbol{d}}^T (c_0 - c_{target} - \boldsymbol{G}_{\boldsymbol{d}} \boldsymbol{W} \boldsymbol{y}_0) + \boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

$$\boldsymbol{S}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{G}_{\boldsymbol{\theta}}^T (c_0 - c_{target} - \boldsymbol{G}_{\boldsymbol{\theta}} \boldsymbol{\theta}_0) + \boldsymbol{G}_{\boldsymbol{d}} \boldsymbol{W}(\boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{y}_0) + \hat{\boldsymbol{S}}^{-1} \hat{\boldsymbol{\mu}}$$

Sparse Variational Inference

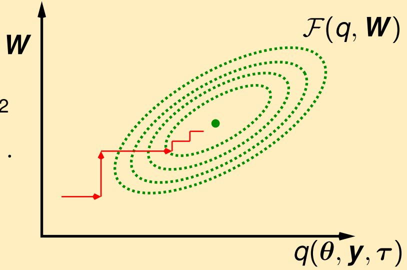
Sparse Bayesian Learning

$$\underbrace{\boldsymbol{d}}_{N\times 1} = \underbrace{\boldsymbol{W}}_{N\times n} \underbrace{\boldsymbol{y}}_{n\times 1}$$

Can we find a concise vocabulary $\boldsymbol{W}$ i.e. $n << N$?

- Sparse Coding (Olshausen & Field 1996, Lewicki & Sejnowski 2000)
- Given $q(\theta, \boldsymbol{y}, \boldsymbol{\tau})$, what is the best $\boldsymbol{W}$?

$$\mathcal{F}(q, \boldsymbol{W}) = \begin{array}{l} -\frac{1}{2}(c(\theta_0, \boldsymbol{W}\boldsymbol{y}_0) - c_{target})^2 \\ -\frac{1}{2}\boldsymbol{W}^T \boldsymbol{G}_{\boldsymbol{d}}^T \boldsymbol{G}_{\boldsymbol{d}} \boldsymbol{W} : \boldsymbol{S}_{\boldsymbol{y}} + \ldots \end{array}$$



Variational Inference

Algorithm:

$$\mathcal{F}(q, \boldsymbol{W}) = E_q\left(\log \frac{U(\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\tau})p(\boldsymbol{\tau})}{q(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\tau})}\right)$$

0. Initialize $\boldsymbol{W}$, $q(\boldsymbol{\theta}) \equiv \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{S}_{\boldsymbol{\theta}})$ and $q(\boldsymbol{y}) \equiv \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{y}}, \boldsymbol{S}_{\boldsymbol{y}})$, $q(\boldsymbol{\tau})$.
1. Set $\boldsymbol{\theta}_0 = \boldsymbol{\mu}_{\boldsymbol{\theta}}$, $\boldsymbol{d}_0 = \boldsymbol{W}\boldsymbol{\mu}_{\boldsymbol{y}}$ and linearize $c(\boldsymbol{\theta}, \boldsymbol{d})$ around $(\boldsymbol{\theta}_0, \boldsymbol{d}_0)$.
2. Fix $\boldsymbol{W}$, update $q(\boldsymbol{\theta})$, $q(\boldsymbol{y})$, $q(\boldsymbol{\tau})$ Cost: 1 forward call
3. Fix $q(\boldsymbol{\theta})$, $q(\boldsymbol{y})$, $q(\boldsymbol{\tau})$, update $\boldsymbol{W}$: Cost: 1 forward call

$$\boldsymbol{W} \leftarrow \boldsymbol{W} + \eta \frac{\partial \mathcal{F}}{\partial \boldsymbol{W}}$$
$$\text{such that } \sum_{i=1}^{N} W_{ij}^2 = 1, j = 1, \ldots, n$$

4. Goto 1. until convergence

Variational Inference - Constraints

Shape/topology optimization:

$$\min_{\boldsymbol{d}} \quad compliance(\boldsymbol{d}) = \boldsymbol{b}^T \boldsymbol{u}(\boldsymbol{d})$$

such that:

$$\boldsymbol{K}(\boldsymbol{d})\boldsymbol{u}(\boldsymbol{d}) = \boldsymbol{b} \quad \text{(governing equation)}$$

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

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

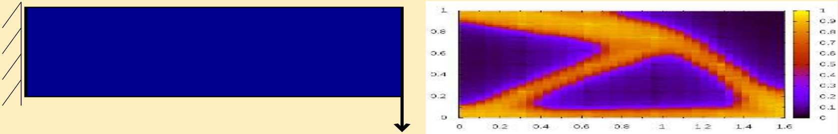
$$d(\boldsymbol{x}) = \begin{cases} 1, & material \\ 0, & void \end{cases}$$

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

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

Numerical Illustration

Deterministic topology optimization



(a) domain (b) *compliance*($\boldsymbol{d}$) ≈ 55

Figure: Deterministic topology optimization - $O(100)$ forward runs

Stochastic topology optimization

- $dim(\boldsymbol{d}) = 5120$ (design variables), $dim(\boldsymbol{\theta}) = 5120$ (random variables)
- $\log \boldsymbol{\theta} \sim N(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$
  - $C.O.V.[\theta_i] = 1$
  - $\boldsymbol{\Sigma}_\theta = Cov[\log\theta(\boldsymbol{x}_i), \log\theta(\boldsymbol{x}_j)] = e^{-|\boldsymbol{x}_i - \boldsymbol{x}_j|/l_0}$
  - $l_0 = 0.1$ (correlation length)
- Volume constraint: $\int d(\boldsymbol{x})\, d\boldsymbol{x} = 0.4$

Numerical Illustration

$$\underbrace{\boldsymbol{d}}_{5120\times 1} = \underbrace{\boldsymbol{W}}_{5120\times 100} \underbrace{\boldsymbol{y}}_{100\times 1}$$

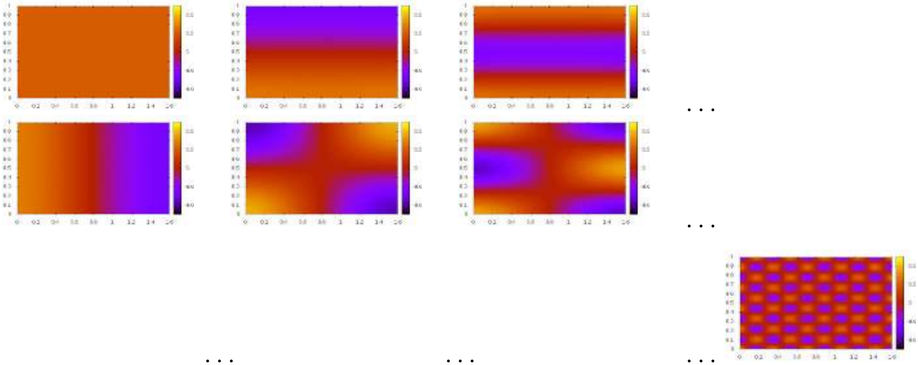


Figure: Initial $\boldsymbol{W}$ - DCT basis vectors

Numerical Illustration

$$\mathcal{F}(q, \boldsymbol{W}) = E_q \left(\log \frac{U(\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\tau})p(\boldsymbol{\tau})}{q(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\tau})} \right)$$

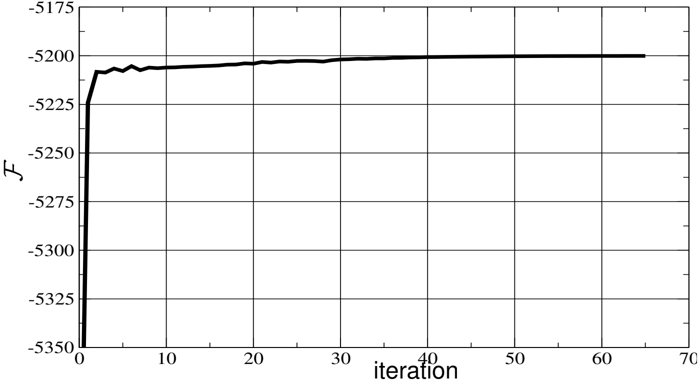


Figure: Evolution of Variational bound $\mathcal{F}(q, \boldsymbol{W})$

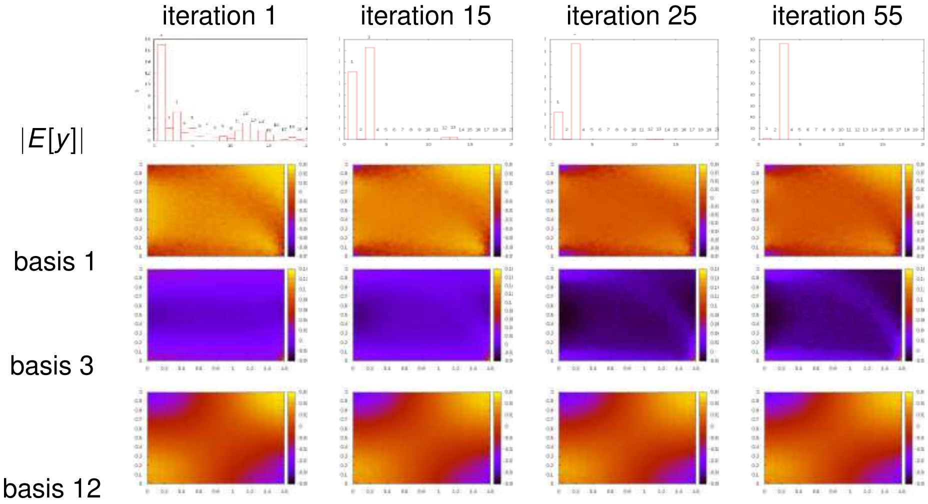
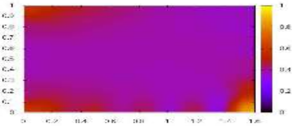
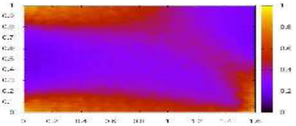


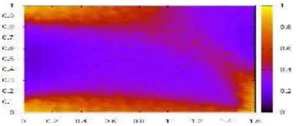
Table: Evolution of basis vectors in $\boldsymbol{W}$



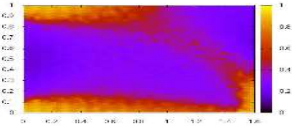
(a) iteration 1



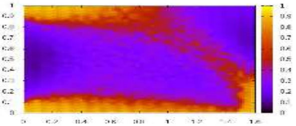
(b) iteration 5



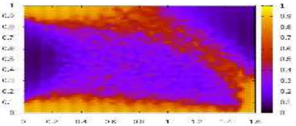
(c) iteration 10



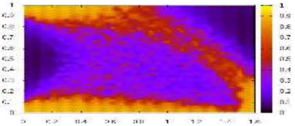
(d) iteration 15



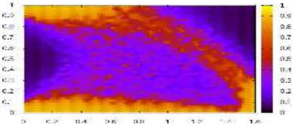
(e) iteration 25



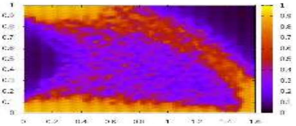
(f) iteration 35



(g) iteration 45



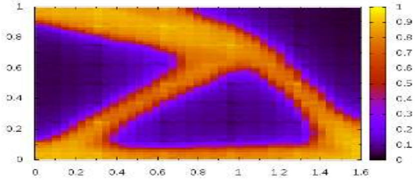
(h) iteration 55



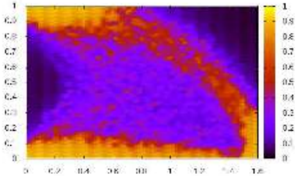
(i) iteration 400

Figure: Evolution of $\mu_{\boldsymbol{d}} = E_q(\boldsymbol{d})$

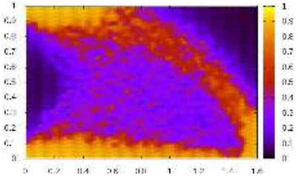
Numerical Illustration



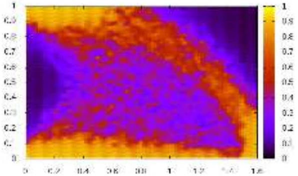
(a) deterministic



(b) mean-st.dev.*



(c) mean



(d) mean+st.dev.*

Figure: Deterministic vs. (Variational) Stochastic

Summary & Outlook

- Stochastic optimization poses significantly more challenges than uncertainty propagation when *thousands* of random and design variables are present.
- We advocate a probabilistic inference treatment
- Sequential Monte Carlo tools offer a general and (asymptotically) exact strategy
- Variational inference techniques offer more efficeint but approximate solutions
- Sparse Bayesian Learning can lead to significant dimensionality reduction and facilitate/expedite solution