Bayesian Multi-Fidelity Optimization under Uncertainty

Maximilian Koschade
maximilian.koschade@tum.de

Phaedon S. Koutsourelakis
p.s.koutsourelakis@tum.de

Continuum Mechanics Group
Department of Mechanical Engineering
Technical University of Munich, Germany

SIAM Computational Science and Engineering, Atlanta, 2017

Example: Material property as random field $\lambda(\boldsymbol{x})$

- $\boldsymbol{z}$: design variables (topology or shape)
- $\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$: stochastic influences, e.g.
  - material : discretized random field $\lambda(\boldsymbol{x})$
  - temperature / load : stochastic process
  - manufacturing tolerances : distributed around nominal value

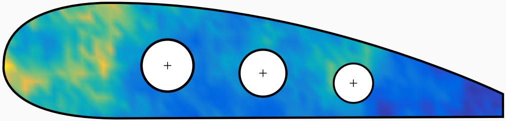


Figure 1: Cross section view of stiffening rib

Introducing uncertainty to optimization problems

In many engineering applications deterministic optimization is a simplification neglecting aleatory and epistemic uncertainty.

Optimization under Uncertainty - Objective Function

Maximize the expected utility

$$\boldsymbol{z}^* = \arg\max_{z} V(\boldsymbol{z}) = \arg\max_{z} \int U(\boldsymbol{z}, \boldsymbol{\theta})\, p_\theta(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

- $\boldsymbol{z}$: design variables
- $\boldsymbol{\theta} \sim p_\theta(\boldsymbol{\theta})$: stochastic influences on the system

Example: minimize probability of failure

$$U(\boldsymbol{z}, \boldsymbol{\theta}) = \mathbf{1}_{\mathcal{A}}(\boldsymbol{z}, \boldsymbol{\theta}) \text{ (where } \mathcal{A} \text{ the event of non-failure)}$$

Example: design goal $\boldsymbol{u}_{target}$

$$U(\boldsymbol{z}, \boldsymbol{\theta}) = \exp\left\{-\frac{1}{2}\tau_Q \left(\boldsymbol{u}(\boldsymbol{z}, \boldsymbol{\theta}) - \boldsymbol{u}_{target}\right)^2\right\}$$

τ_Q : penalty parameter enforcing the design goal

Probabilistic Formulation of Optimization under Uncertainty

Reformulation as Probabilistic Inference[1]

Solution is given by an auxiliary posterior distribution[2] $\boxed{\pi(\boldsymbol{z}, \boldsymbol{\theta})}$

$$V(\boldsymbol{z}) \propto \int \underbrace{\pi(\boldsymbol{z}, \boldsymbol{\theta})}_{\textit{posterior}} \mathrm{d}\boldsymbol{\theta}$$

$$\propto \int \boxed{\underbrace{U(\boldsymbol{z}, \boldsymbol{\theta})}_{\textit{likelihood}} \underbrace{p_{\theta}(\boldsymbol{\theta})}_{\text{prior}}} \mathrm{d}\boldsymbol{\theta}$$

since the *marginal* $\pi(\boldsymbol{z}) \propto V(\boldsymbol{z})$, given a flat prior $p_{\boldsymbol{z}}(\boldsymbol{z})$.

Conducive to consistent incorporation of epistemic uncertainty due to approximate, lower-fidelity solvers!

[1] Mueller (2005)

[2] This approach should NOT be confused with Bayesian optimization

Probabilistic Formulation of Optimization under Uncertainty

Reformulation as Probabilistic Inference[1]

Solution is given by an auxiliary posterior distribution[2] $\boxed{\pi(\boldsymbol{z}, \boldsymbol{\theta})}$

$$V(\boldsymbol{z}) \propto \int \underbrace{\pi(\boldsymbol{z}, \boldsymbol{\theta})}_{\textit{posterior}} \mathrm{d}\boldsymbol{\theta}$$

$$\propto \left[\int \boxed{\underbrace{U(\boldsymbol{z}, \boldsymbol{\theta})}_{\textit{likelihood}} \underbrace{p_\theta(\boldsymbol{\theta})}_{\text{prior}}} \mathrm{d}\boldsymbol{\theta}\right] \underbrace{p_{\boldsymbol{z}}(\boldsymbol{z})}_{\text{flat prior}}$$

since the *marginal* $\pi(\boldsymbol{z}) \propto V(\boldsymbol{z})$, given a flat prior $p_{\boldsymbol{z}}(\boldsymbol{z})$.

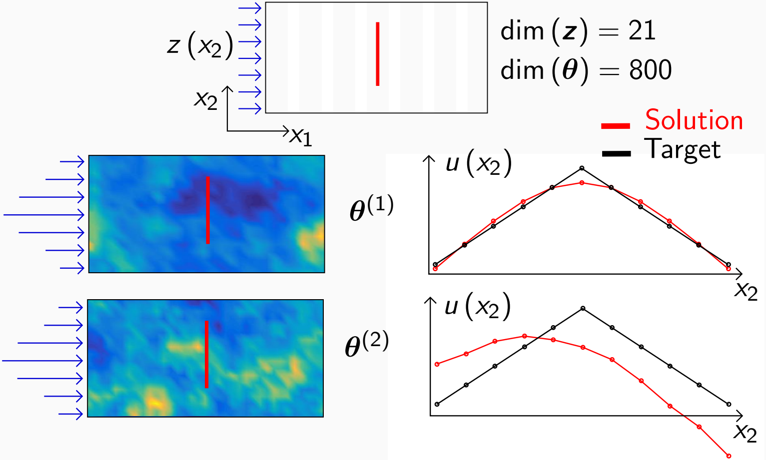
Conducive to consistent incorporation of epistemic uncertainty due to approximate, lower-fidelity solvers!

[1] Mueller (2005)

[2] This approach should NOT be confused with Bayesian optimization

Example: Stochastic Poisson Equation

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



$\boldsymbol{z}$: Control heat influx $\qquad$ $\boldsymbol{\theta}$: Log-Normal conductivity field

Solution via rank-1-perturbed Gaussian q^*

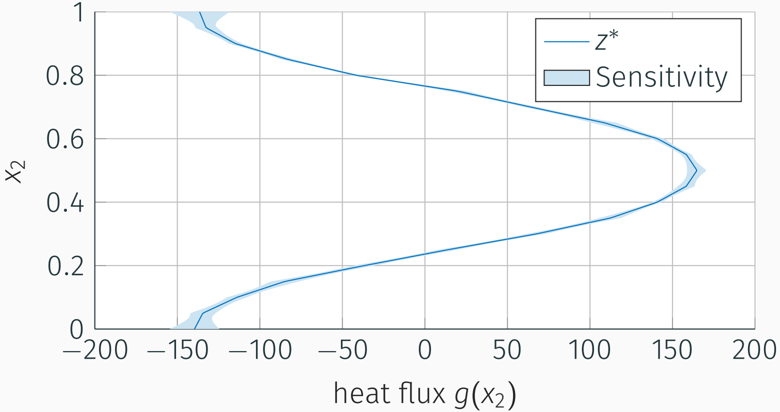
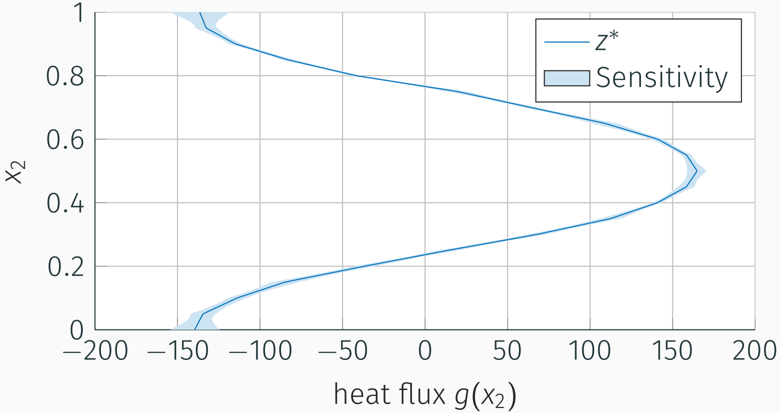


Figure 2: Black-box stochastic variational inference in dimension 821 ($\dim(\boldsymbol{\theta}) = 800, \dim(\boldsymbol{z}) = 21$)
(Hoffman et al., 2013; Ranganath et al., 2013)

Solution via rank-1-perturbed Gaussian q^*



Cost : $\mathcal{O}\left(10^3\right)$ forward evaluations

- **high dimension**
- **expensive numerical model**

$\Rightarrow$ **probabilistic inference can quickly become prohibitive.**

How can we address this issue?

Introduction of approximate solvers

If we denote $\boxed{a = \log U}$ and $\boldsymbol{y} = [\boldsymbol{z}, \boldsymbol{\theta}]^T$ we can rewrite $\pi(\boldsymbol{y})$

$$\begin{aligned}\pi_a(\boldsymbol{y}) \propto U(\boldsymbol{y})\, p_{\boldsymbol{y}}(\boldsymbol{y}) &= \exp(a(\boldsymbol{y}))\, p_{\boldsymbol{y}}(\boldsymbol{y}) \\ &= \int \exp(a)\, \boxed{\delta(a - \log U(\boldsymbol{y}))}\, p_{\boldsymbol{y}}(\boldsymbol{y})\, \mathrm{d}a \\ &= \int \exp(a)\, \boxed{p(a|\boldsymbol{y})}\, p_{\boldsymbol{y}}(\boldsymbol{y})\, \mathrm{d}a\end{aligned}$$

Approximate solvers = Epistemic uncertainty

- As long as $p(a|\boldsymbol{y})$ is a Dirac, we recover posterior perfectly
- Introduction of cheap, approximate solvers leads to dispersion of $p(a|\boldsymbol{y})$ and irrevocable loss of information regarding $\boldsymbol{y}$
- **We can consistently incorporate this epistemic uncertainty in the Bayesian framework**

Introduction of approximate solvers

If we denote $a = \log U$ and $\boldsymbol{y} = [\boldsymbol{z}, \boldsymbol{\theta}]^T$ we can rewrite $\pi(\boldsymbol{y})$

$$\begin{aligned}\pi_a(\boldsymbol{y}) \propto U(\boldsymbol{y})\, p_{\boldsymbol{y}}(\boldsymbol{y}) &= \exp(a(\boldsymbol{y}))\, p_{\boldsymbol{y}}(\boldsymbol{y}) \\ &= \int \exp(a)\, \delta(a - \log U(\boldsymbol{y}))\, p_{\boldsymbol{y}}(\boldsymbol{y})\, \mathrm{d}a \\ &= \int \exp(a)\, p(a|\boldsymbol{y})\, p_{\boldsymbol{y}}(\boldsymbol{y})\, \mathrm{d}a\end{aligned}$$

Regression Model

We may learn $p(a|\boldsymbol{y})$ from e.g. a Bayesian regression model or a Gaussian process $\mathcal{GP}$

$$a = \phi(\boldsymbol{y})^T \boldsymbol{w} + \epsilon$$

This approach is impractical for a high-dimensional probability space $\boldsymbol{y} = [\boldsymbol{z}, \theta]^T$!

Introduction of approximate solvers

If we denote $\boxed{a = \log U}$ and $\boldsymbol{y} = [\boldsymbol{z}, \boldsymbol{\theta}]^T$ we can rewrite $\pi(\boldsymbol{y})$

$$\begin{aligned}\pi_a(\boldsymbol{y}) &\propto U(\boldsymbol{y})\, p_{\boldsymbol{y}}(\boldsymbol{y}) = \exp(a(\boldsymbol{y}))\, p_{\boldsymbol{y}}(\boldsymbol{y}) \\ &= \int \exp(a)\, \boxed{\delta(a - \log U(\boldsymbol{y}))}\, p_{\boldsymbol{y}}(\boldsymbol{y})\, \mathrm{d}a \\ &= \int \exp(a)\, \boxed{p(a|\boldsymbol{y})}\, p_{\boldsymbol{y}}(\boldsymbol{y})\, \mathrm{d}a\end{aligned}$$

Suppose instead we introduce a low-fidelity log-likelihood A

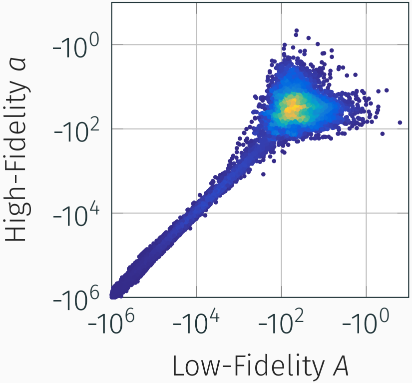
$$\begin{aligned}p(a|\boldsymbol{y}) &= \int p(a, A|\boldsymbol{y})\, \mathrm{d}A = \int p(a|A, \boldsymbol{y})\, p(A|\boldsymbol{y})\, \mathrm{d}A \\ &\approx \int \boxed{p(a|A)}\, \delta(A - \log U_{LowFi})\, \mathrm{d}A := p_A(a|\boldsymbol{y})\end{aligned}$$

$$\Rightarrow \pi_A(\boldsymbol{y}) \propto \int \exp(a)\, p_A(a|\boldsymbol{y})\, p_{\boldsymbol{y}}(\boldsymbol{y})\, \mathrm{d}a$$

Learning $p\left(a|y\right)$: Probabilistic multi-fidelity approach[3]

Introduce low-fidelity log-likelihood A

$$p_A\left(a|\boldsymbol{y}\right) \approx \int \boxed{p\left(a|A\right)} \, \delta\left(A - \log U_{\text{LowFi.}}\left(\boldsymbol{y}\right)\right) \mathrm{d}A$$



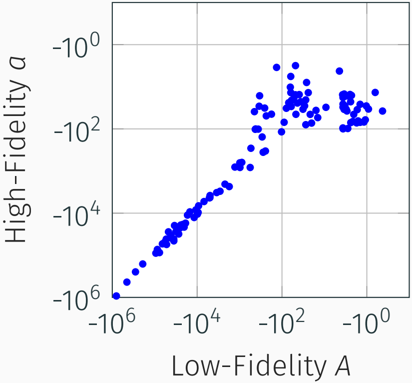
Pred. density $p_A\left(a|A\right)$

- belief of high-fidelity a given low-fidelity A
- learn from a limited set of forward solver evaluations $\mathcal{D}$
- $\mathcal{D} = \left\{a\left(\boldsymbol{y}_n\right), A\left(\boldsymbol{y}_n\right)\right\}_{n=1}^{N}$

Learning $p(a|y)$: Probabilistic multi-fidelity approach[3]

Introduce low-fidelity log-likelihood A

$$p_A(a|\boldsymbol{y}) \approx \int \boxed{p(a|A)}\, \delta(A - \log U_{\text{LowFi.}}(\boldsymbol{y}))\, \mathrm{d}A$$



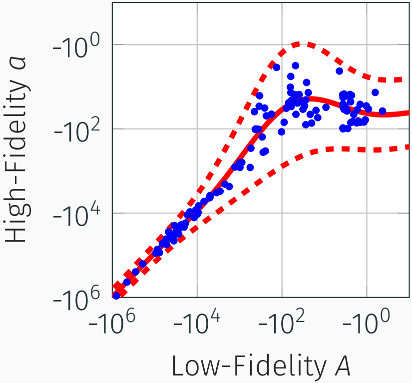
Pred. density $p_A(a|A, \mathcal{D})$

- belief of high-fidelity a given low-fidelity A
- learn from a limited set of forward solver evaluations $\mathcal{D}$
- $\mathcal{D} = \{a(\boldsymbol{y}_n), A(\boldsymbol{y}_n)\}_{n=1}^N$

Learning $p\left(a|y\right)$: Probabilistic multi-fidelity approach[3]

Introduce low-fidelity log-likelihood A

$$p_A\left(a|\boldsymbol{y}\right) \approx \int \boxed{p\left(a|A\right)}\ \delta\left(A - \log U_{\text{LowFi.}}\left(\boldsymbol{y}\right)\right) \mathrm{d}A$$



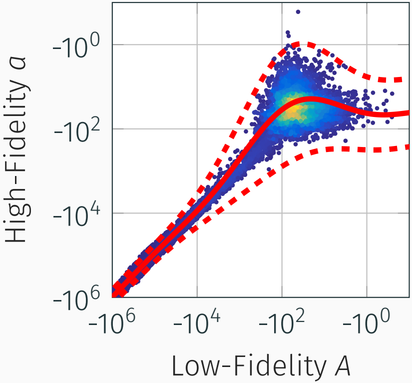
Learn $p_A\left(a|A, \mathcal{D}\right)$

- Learn predictive density
- Using e.g. variational relevance vector machine (VRVM) or Variational Heteroscedastic Gaussian Process (VHGP)
- $\mathcal{D} = \left\{a\left(\boldsymbol{y}_n\right), A\left(\boldsymbol{y}_n\right)\right\}_{n=1}^N$

Learning $p\left(a|y\right)$: Probabilistic multi-fidelity approach[3]

Introduce low-fidelity log-likelihood A

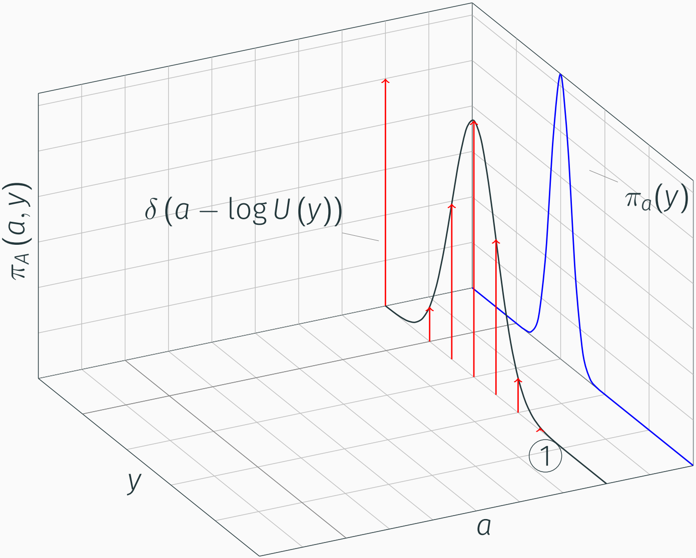
$$p_A\left(a|\boldsymbol{y},\mathcal{D}\right) \approx \int \boxed{p\left(a|A,\mathcal{D}\right)}\ \delta\left(A - \log U_{\text{LowFi.}}\left(\boldsymbol{y}\right)\right) \mathrm{d}A$$



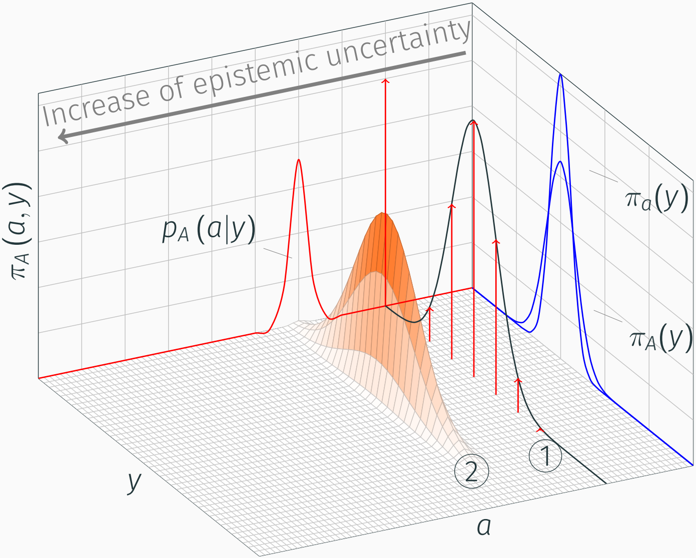
Pred. density $p_A\left(a|A,\mathcal{D}\right)$

- belief of high-fidelity a given low-fidelity A
- learn from a limited set of forward solver evaluations $\mathcal{D}$
- $\mathcal{D} = \left\{a\left(\boldsymbol{y}_n\right), A\left(\boldsymbol{y}_n\right)\right\}_{n=1}^{N}$

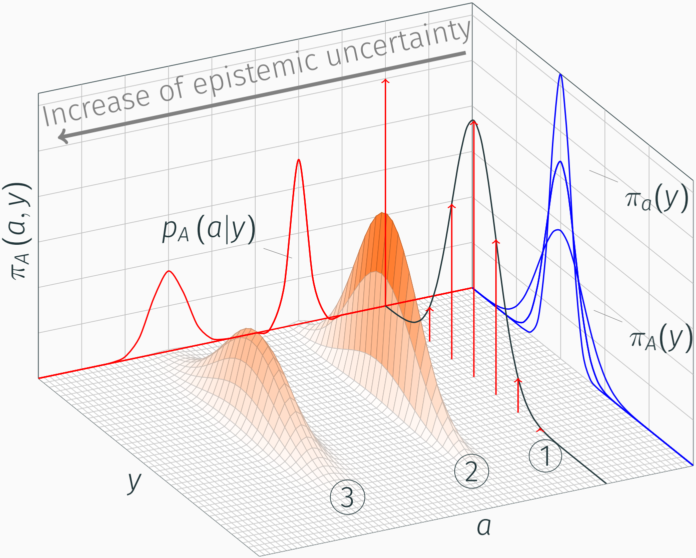
Extended Probability Space - Illustration



Extended Probability Space - Illustration



Extended Probability Space - Illustration



Multi-Fidelity posterior $\pi_A(\boldsymbol{y})$

Approximate $\pi_A(\boldsymbol{y})$

If predictive density $p(a|\boldsymbol{y})$ is given by a Gaussian $\mathcal{N}\left(a\middle|\mu(A(\boldsymbol{y})), \sigma^2(A(\boldsymbol{y}))\right)$, then we obtain

$$\log \pi_A(\boldsymbol{y}) = \ \mu(A(\boldsymbol{y})) \ + \ \frac{1}{2}\sigma^2(A(\boldsymbol{y})) \ + \log p_{\boldsymbol{y}}(\boldsymbol{y})$$

Place probability mass on $\boldsymbol{y}$ associated with

Multi-Fidelity posterior $\pi_A(\boldsymbol{y})$

Approximate $\pi_A(\boldsymbol{y})$

If predictive density $p(a|\boldsymbol{y})$ is given by a Gaussian $\mathcal{N}\left(a \middle| \mu(A(\boldsymbol{y})), \sigma^2(A(\boldsymbol{y}))\right)$, then we obtain

$$\log \pi_A(\boldsymbol{y}) = \underbrace{\boxed{\mu(A(\boldsymbol{y}))}}_{\textbf{A}} + \frac{1}{2}\sigma^2(A(\boldsymbol{y})) + \log p_{\boldsymbol{y}}(\boldsymbol{y})$$

Place probability mass on $\boldsymbol{y}$ associated with

(A): high predictive mean $\mu(\boldsymbol{y})$

Multi-Fidelity posterior $\pi_A(\boldsymbol{y})$

Approximate $\pi_A(\boldsymbol{y})$

If predictive density $p(a|\boldsymbol{y})$ is given by a Gaussian $\mathcal{N}\left(a \middle| \mu(A(\boldsymbol{y})), \sigma^2(A(\boldsymbol{y}))\right)$, then we obtain

$$\log \pi_A(\boldsymbol{y}) = \underbrace{\boxed{\mu(A(\boldsymbol{y}))}}_{\textbf{A}} + \frac{1}{2}\underbrace{\boxed{\sigma^2(A(\boldsymbol{y}))}}_{\textbf{B}} + \log p_{\boldsymbol{y}}(\boldsymbol{y})$$

Place probability mass on $\boldsymbol{y}$ associated with

(A): high predictive mean $\mu(\boldsymbol{y})$

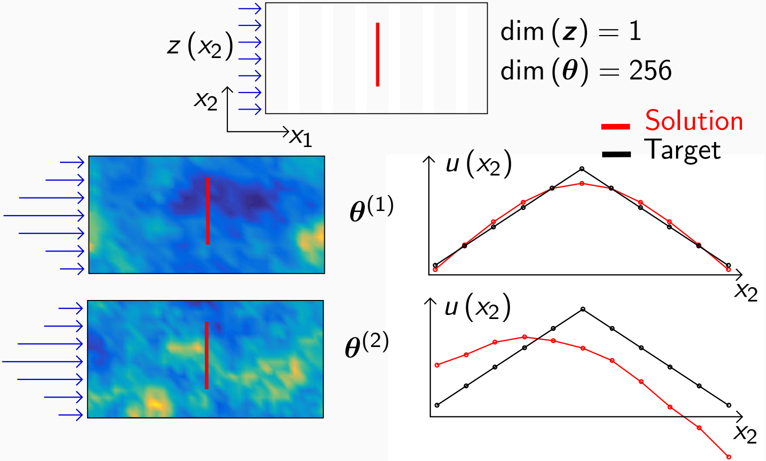
(B): large epistemic uncertainty $\sigma^2(\boldsymbol{y})$

Example: Stochastic Poisson Equation

$$\nabla \cdot \left(-\lambda\left(\boldsymbol{x}\right)\nabla u\left(\boldsymbol{x}\right)\right) = 0$$

$\dim\left(\boldsymbol{z}\right) = 1$

$\dim\left(\boldsymbol{\theta}\right) = 256$



$\boldsymbol{z}$: Control heat influx

$\boldsymbol{\theta}$: Log-Normal conductivity field

Effect of lower-fidelity solvers[4]

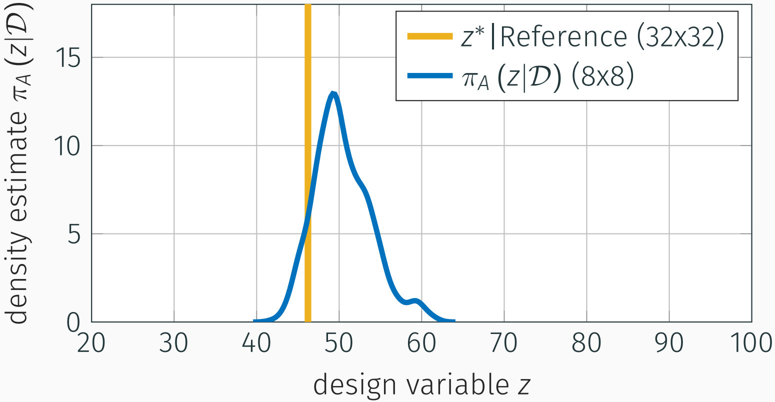


Figure 2: $\dim(\boldsymbol{\theta}) = 256$, speedup $\mathcal{S}_{4\times 4} \approx 2,000$, $N = 200$ training data samples, density estimate obtained using MALA

[4] here the low-fidelity solvers are simply coarser discretizations of the stochastic Poisson equation

Effect of lower-fidelity solvers[4]

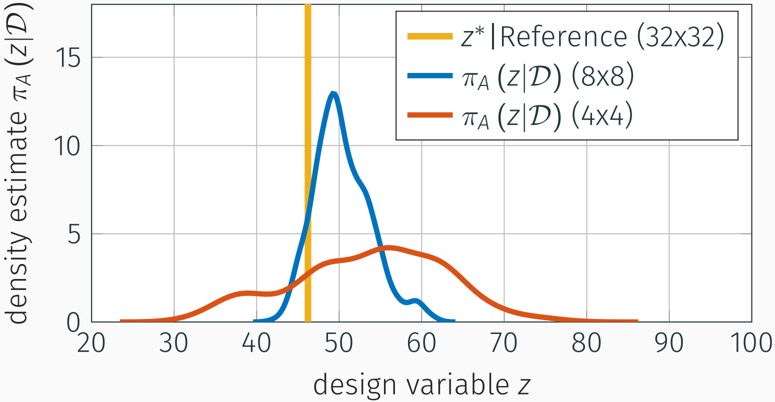


Figure 2: $\dim(\boldsymbol{\theta}) = 256$, speedup $\mathcal{S}_{4\times 4} \approx 2,000$, $N = 200$ training data samples, density estimate obtained using MALA

[4]here the low-fidelity solvers are simply coarser discretizations of the stochastic Poisson equation

Effect of training data $\mathcal{D}$

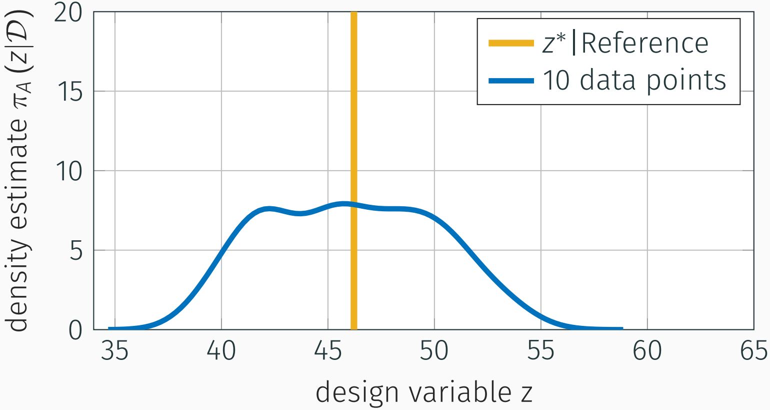


Figure 3: The data restricted likelihood becomes more confident due to the reduction of epistemic uncertainty by additional training samples. ($\dim(\boldsymbol{\theta}) = 256$, averaged for 100 random sub-samplings of the data)

Effect of training data $\mathcal{D}$

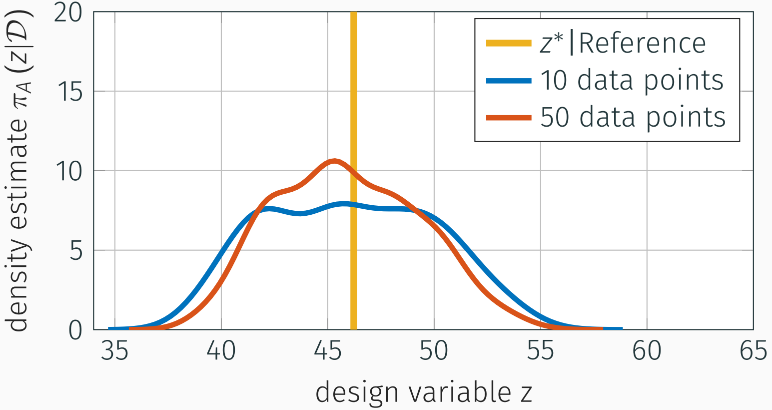


Figure 3: The data restricted likelihood becomes more confident due to the reduction of epistemic uncertainty by additional training samples. ($\dim(\boldsymbol{\theta}) = 256$, averaged for 100 random sub-samplings of the data)

Effect of training data $\mathcal{D}$

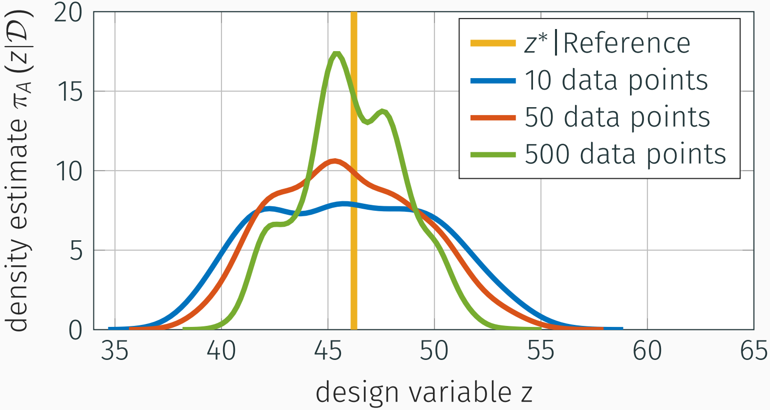


Figure 3: The data restricted likelihood becomes more confident due to the reduction of epistemic uncertainty by additional training samples. ($\dim(\boldsymbol{\theta}) = 256$, averaged for 100 random sub-samplings of the data)

Review

Summary:

- Optimization under uncertainty can be reformulated as Bayesian inference
- Allows consistent incorporation of epistemic uncertainty introduced by cheaper, approximate (probabilistic) models
- Inevitable loss of information, but me way obtain multi-fidelity posterior which contains the optimal design $\boldsymbol{z}^*$ (MAP)
- Approach applicable to any problem of Bayesian inference

Outlook:

- introduction of multiple predictors $A_{(p)}$
- adaptive enrichment of training data $\mathcal{D}$
- more flexible approach to learn $p(a|A)$

Review

Summary:

- Optimization under uncertainty can be reformulated as Bayesian inference
- Allows consistent incorporation of epistemic uncertainty introduced by cheaper, approximate (probabilistic) models
- Inevitable loss of information, but me way obtain multi-fidelity posterior which contains the optimal design z^* (MAP)
- Approach applicable to any problem of Bayesian inference

Outlook:

- introduction of multiple predictors $A_{(p)}$
- adaptive enrichment of training data $\mathcal{D}$
- more flexible approach to learn $p\,(a|A)$

Review

Summary:

- Optimization under uncertainty can be reformulated as Bayesian inference
- Allows consistent incorporation of epistemic uncertainty introduced by cheaper, approximate (probabilistic) models
- Inevitable loss of information, but me way obtain multi-fidelity posterior which contains the optimal design $\boldsymbol{z}^*$ (MAP)
- Approach applicable to any problem of Bayesian inference

Outlook:

- introduction of multiple predictors $A_{(p)}$
- adaptive enrichment of training data $\mathcal{D}$
- more flexible approach to learn $p\,(a|A)$

Review

Summary:

- Optimization under uncertainty can be reformulated as Bayesian inference
- Allows consistent incorporation of epistemic uncertainty introduced by cheaper, approximate (probabilistic) models
- Inevitable loss of information, but me way obtain multi-fidelity posterior which contains the optimal design $\boldsymbol{z}^*$ (MAP)
- Approach applicable to any problem of Bayesian inference

Outlook:

- introduction of multiple predictors $A_{(p)}$
- adaptive enrichment of training data $\mathcal{D}$
- more flexible approach to learn $p\,(a|A)$

References

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347.

Mueller, P. (2005). Simulation based optimal design. In Dey, D. and Rao, C., editors, *Bayesian Thinking Modeling and Computation*, volume 25 of *Handbook of Statistics*, pages 509 – 518. Elsevier.

Ng, L. W. and Willcox, K. E. (2014). Multifidelity approaches for optimization under uncertainty. *International Journal for Numerical Methods in Engineering*, 100(10):746–772.

Perdikaris, P., Venturi, D., Royset, J., and Karniadakis, G. (2015). Multi-fidelity modelling via recursive co-kriging and gaussian–markov random fields. In *Proc. R. Soc. A*, volume 471, page 20150018. The Royal Society.

Ranganath, R., Gerrish, S., and Blei, D. M. (2013). Black Box Variational Inference. *Aistats*, 33.

Addendum (1): generate Training Data $\mathcal{D} = \{a(\boldsymbol{y}_n), A(\boldsymbol{y}_n)\}_{n=1}^{N}$

Batch: Generate $\mathcal{D}$ such that equal numbers of $A(\boldsymbol{y}_n)$ fall in

$$\mathcal{M}_l = \left\{ \boldsymbol{y} \middle| \pi_c^{(l)} \leq \pi_c(\boldsymbol{y}) \leq \pi_c^{(l+1)} \right\}$$

- $l = 1, ..., L$
- $\pi_c^{(l+1)} = \text{const} \cdot \pi_c^{(l)}$
- π_c : posterior defined low-fidelity solver.

Adaptive Refinement: Use $\pi(\boldsymbol{y}|\mathcal{D})$ as acquisition function, corresponding to large predictive mean and epistemic uncertainty

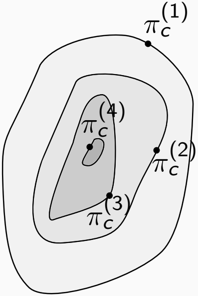


Figure 4: Iso-probability lines of the coarse posterior $\pi_c(\boldsymbol{y})$

Addendum (2): is it possible to exclude the optimal design z^* (MAP)?

This approach will, if executed correctly, never put zero probability mass on the MAP z^* or any other value deemed probable under the high fidelity posterior.

Potential Errors:

- Generated $\mathcal{D}$ does not sufficiently encapsulate $p\,(a|A)$
- Regression model is not flexible enough to learn $p\,(a|A, \mathcal{D})$ correctly
- Approximation of intractable posterior $\pi_A\,(\boldsymbol{y}|\mathcal{D})$ using e.g. VB, MCMC or SMC.