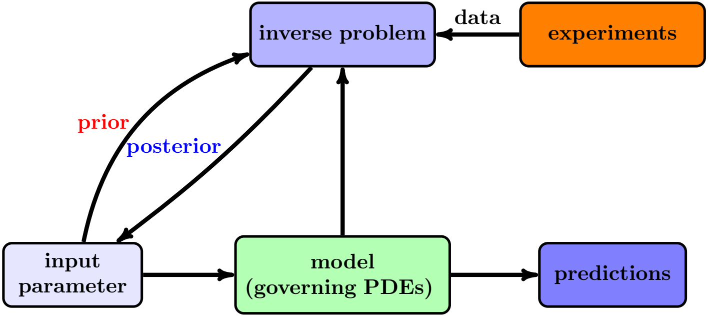
Optimal Experimental Design For Large-scale Bayesian Inverse Problems

Alen Alexanderian

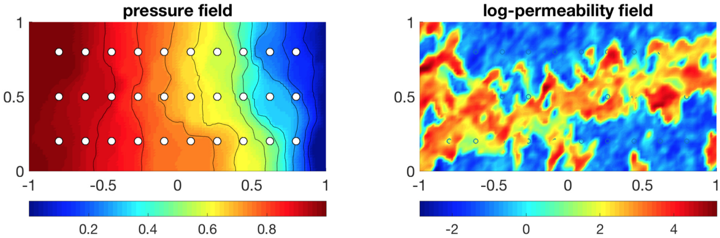
Department of Mathematics
North Carolina State University

May 25, 2021

Inverse problems governed by PDEs and design of experiments

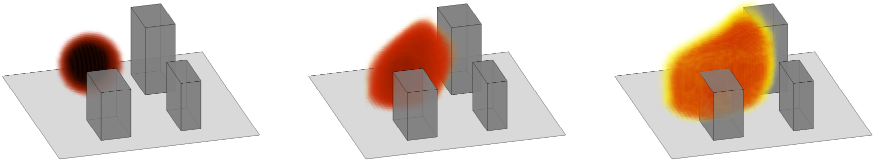


Example: permeability estimation in porous media

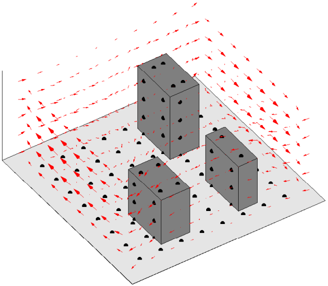


- **Governing PDE** (forward model): equations of subsurface flow
- **Unknown/uncertain parameter**: permeability field
- **Inverse problem**: Use a vector of sensor measurements of pressure to estimate the permeability field

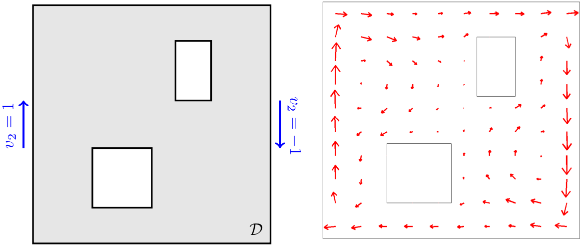
Example: contaminant source identification



- **Governing PDE**: advection-diffusion equation
- **Unknown/uncertain parameter**: initial concentration field
- **Inverse problem**: Use a vector $\boldsymbol{y}$ of point (sensor) measurements of concentration at final time to reconstruct the intial state



2D Model problem

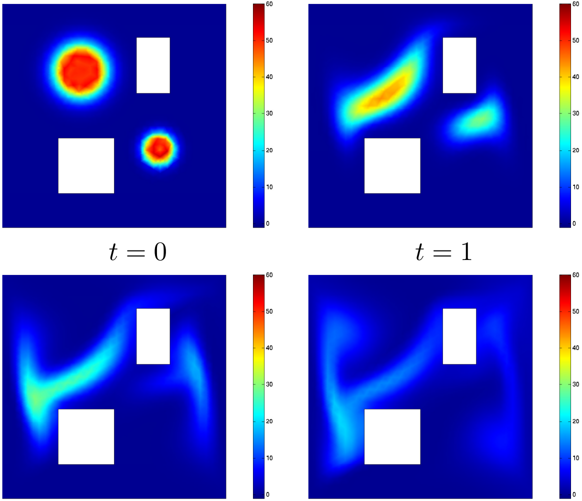


- Forward problem: time dependent advection-diffusion equation

$$
\begin{aligned}
u_t - \kappa \Delta u + \mathbf{v} \cdot \nabla u &= 0 \qquad && \text{in } \mathcal{D} \times [0, T] \\
u(0, \boldsymbol{x}) &= m && \text{in } \mathcal{D} \\
\kappa \nabla u \cdot \boldsymbol{n} &= 0 && \text{on } \partial\mathcal{D} \times [0, T]
\end{aligned}
$$

- m: *unknown* initial condition
- $\boldsymbol{v}$: velocity field

Solution of the forward problem



$t = 0$ $t = 1$

$t = 2$ $t = 3$

The inverse problem: reconstruct initial condition

The inverse problem of finding the unknown initial state based on sensor data

$$\min_m \frac{1}{2}\|\mathcal{B}u(m) - \boldsymbol{d}\|^2 + \frac{\alpha}{2}\langle \mathcal{A}m, m\rangle$$

where

$$\begin{aligned} u_t - \kappa\Delta u + \mathbf{v}\cdot\nabla u &= 0 && \text{in } \mathcal{D}\times[0,T] \\ u(0,\boldsymbol{x}) &= m && \text{in } \mathcal{D} \\ \kappa\nabla u\cdot\boldsymbol{n} &= 0 && \text{on } \partial\mathcal{D}\times[0,T] \end{aligned}$$

- $\mathcal{B}$: observation operator
- $\boldsymbol{d} = [\boldsymbol{d}_1^T \boldsymbol{d}_2^T \cdots \boldsymbol{d}_{n_t}^T]^T$, $\quad \boldsymbol{d}_i \in \mathbb{R}^{n_s}$, n_s = number of sensors
- u linear in m, $u = \mathcal{S}m \Longrightarrow$ linear parameter-to-observable map: $\mathcal{F} = \mathcal{B}\mathcal{S}$
- Can rewrite the optimization problem as

$$\min_m \mathcal{J}(m) := \frac{1}{2}\|\mathcal{F}m - \boldsymbol{d}\|^2 + \frac{\alpha}{2}\langle \mathcal{A}m, m\rangle$$

Solving the inverse problem

- Derivative of $\mathcal{J}$

$$
\begin{aligned}
D\mathcal{J}(m)(\tilde{m}) &= \frac{d}{d\varepsilon}\mathcal{J}(m + \varepsilon\tilde{m}) \,|_{\varepsilon=0} \\
&= \langle \mathcal{F}^*(\mathcal{F}m - \boldsymbol{d}) + \alpha\mathcal{A}m, \tilde{m} \rangle
\end{aligned}
$$

- Action of $\mathcal{F}^*$
  $\mathcal{F}^*\boldsymbol{y} = -p(\cdot, 0)$, where p is solution of the adjoint equation

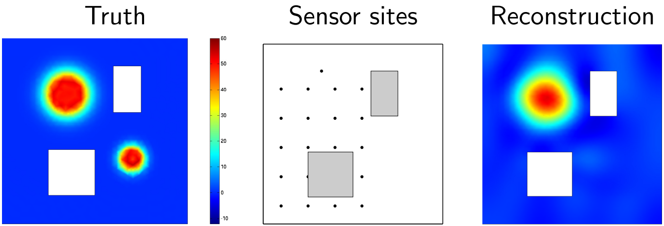
$$
\begin{aligned}
-p_t - \nabla \cdot (p\boldsymbol{v}) - \kappa\Delta p &= -\mathcal{B}^*\boldsymbol{y} \\
p(T) &= 0 \\
(\boldsymbol{v}p + \kappa\nabla p) \cdot \boldsymbol{n} &= 0
\end{aligned}
$$

- Optimality condition

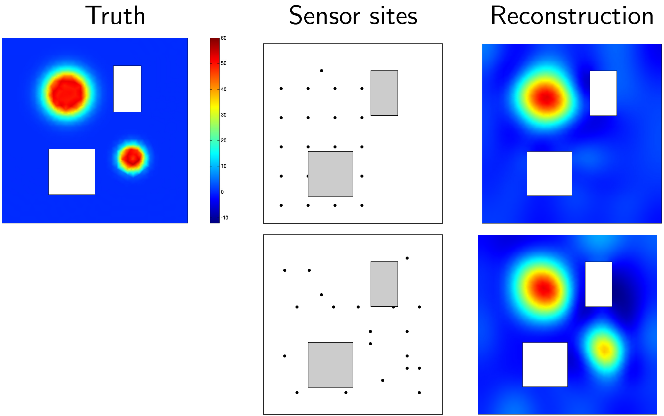
$$
(\mathcal{F}^*\mathcal{F} + \alpha\mathcal{A})m = \mathcal{F}^*\boldsymbol{d} \quad \stackrel{discretize}{\Longrightarrow} \quad (\mathbf{F}^*\mathbf{F} + \alpha\mathbf{A})\boldsymbol{m} = \mathbf{F}^*\boldsymbol{d}
$$

  Solve the linear system using an iterative method, e.g. conjugate gradient

Solving the inverse problem: numerical results



Solving the inverse problem: numerical results



Optimal sensor placement as an optimal design problem

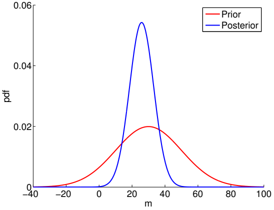
How to place sensors in an "optimal" way?

- Can formulate the optimal sensor placement problem as an optimal experimental design (OED) problem
- Can consider a statistical formulation of the inverse problem
- In addition to a reconstruction, we can also compute a statistical distribution of the parameters, conditioned on experimental data
- Find sensor locations so as to optimize the statistical quality of the reconstructed/inferred parameter
- In context of inverse problems a Bayesian formulation is natural

Bayesian inference: Bayes' formula

$$\pi_{\text{post}}(m|\boldsymbol{y}) \propto \pi_{\text{like}}(\boldsymbol{y}|m)\pi_{\text{prior}}(m)$$

$\pi_{\text{post}}(m\|\boldsymbol{y})$	posterior pdf of m
$\pi_{\text{like}}(\boldsymbol{y}\|m)$	pdf of $\boldsymbol{y}$ given m (data likelihood)
$\pi_{\text{prior}}(m)$	prior pdf of m
	pdf = probability density function



Bayes, T., An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, FRS Communicated by Mr. Price, in a Letter to John Canton, AMFRS. Philosophical Transactions, 1763.

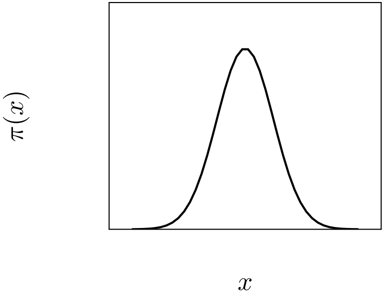
Laplace, P.S., Théorie analytique des probabilités. 1820.

Gaussian random variables

Consider a scalar valued random variable X

- X is Gaussian with mean a and variance σ^2 if it has pdf

$$\pi_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-a)^2\right)$$



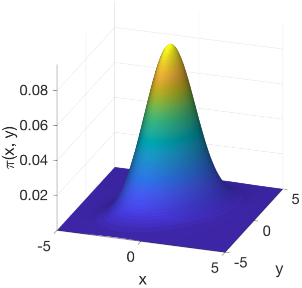
- Notation: $X \sim \mathcal{N}\left(a, \sigma^2\right)$

Gaussian random variables

Consider a vector valued random variable $\boldsymbol{X}$

- $\boldsymbol{X}$ is Gaussian with mean $\boldsymbol{a}$ and covariance matrix $\boldsymbol{\Sigma}$ if it has pdf

$$\pi_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{a})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{a})\right)$$



- Notation: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{a}, \boldsymbol{\Sigma})$

A useful property of Gaussian distribution

Gaussian random vector

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{bmatrix} \sim \mathcal{N}(\bar{\boldsymbol{z}}, \mathbf{C}) = \mathcal{N}\left(\begin{bmatrix} \bar{\boldsymbol{z}}_1 \\ \bar{\boldsymbol{z}}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \right)$$

Marginals of $\boldsymbol{Z}$

$$\boldsymbol{Z}_1 \sim \mathcal{N}(\bar{\boldsymbol{z}}_1, \mathbf{C}_{11}) \quad \text{and} \quad \boldsymbol{Z}_2 \sim \mathcal{N}(\bar{\boldsymbol{z}}_2, \mathbf{C}_{22})$$

Analogue of this result can be derived in infinite-dimensional setting as well

Bayesian inference: a simple but important example

- Let $m \sim \mathcal{N}\left(m_{\text{pr}}, \sigma_{\text{pr}}^2\right)$ with pdf π_{pr}
- Linear model

$$y = am + \eta$$

with $\eta \sim \mathcal{N}\left(0, \sigma_{\text{noise}}\right)$, independent of m

$$y|m \sim \mathcal{N}\left(am, \sigma_{\text{noise}}\right)$$

- Bayes' formula

$$\begin{aligned}
\pi(m|y) &\propto \pi(y|m)\pi_{\text{pr}}(m) \\
&\propto \exp\left(-\frac{1}{2\sigma_{\text{noise}}}(am-y)^2\right)\exp\left(-\frac{1}{2\sigma_{\text{pr}}^2}(m-m_{\text{pr}})^2\right) \\
&= \exp\left[-\frac{1}{2}\left(\sigma_{\text{noise}}^{-2}(am-y)^2 + \sigma_{\text{pr}}^{-2}(m-m_{\text{pr}})^2\right)\right].
\end{aligned}$$

High school algbra: completing the square

$$\begin{aligned}
&\sigma_{\text{noise}}^{-2}(am-y)^2+\sigma_{\text{pr}}^{-2}(m-m_{\text{pr}})^2\\
&\quad=\sigma_{\text{noise}}^{-2}(a^2m^2-2amy+y^2)+\sigma_{\text{pr}}^{-2}(m^2-2mm_{\text{pr}}+m_{\text{pr}}^2)\\
&\quad=(a^2\sigma_{\text{noise}}^{-2}+\sigma_{\text{pr}}^{-2})m^2-2a\sigma_{\text{noise}}^{-2}ym-2\sigma_{\text{pr}}^{-2}mm_{\text{pr}}+c_y\\
&\quad=\underbrace{(a^2\sigma_{\text{noise}}^{-2}+\sigma_{\text{pr}}^{-2})}_{\sigma_{\text{post}}^{-2}}m^2-2(a\sigma_{\text{noise}}^{-2}y+\sigma_{\text{pr}}^{-2}m_{\text{pr}})m+c_y\\
&\quad=\sigma_{\text{post}}^{-2}\Big(m^2-2\,\underbrace{\sigma_{\text{post}}^{2}(a\sigma_{\text{noise}}^{-2}y+\sigma_{\text{pr}}^{-2}m_{\text{pr}})}_{m_{\text{post}}}\,m\Big)+c_y\\
&\quad=\sigma_{\text{post}}^{-2}(m^2-2mm_{\text{post}}+m_{\text{post}}^2)+\tilde{c}_y=\sigma_{\text{post}}^{-2}(m-m_{\text{post}})^2+\tilde{c}_y
\end{aligned}$$

Back to Bayes' formula

$$\begin{aligned}
\pi(m|y)\propto\pi(y|m)\pi_{\text{pr}}(m)&\propto\exp\left[-\frac{1}{2}\left(\sigma_{\text{noise}}^{-2}(am-y)^2+\sigma_{\text{pr}}^{-2}(m-m_{\text{pr}})^2\right)\right]\\
&\propto\exp\left(-\frac{1}{2\sigma_{\text{post}}^2}(m-m_{\text{post}})^2\right)
\end{aligned}$$

posterior: $\boxed{m|y\sim\mathcal{N}\left(m_{\text{post}},\sigma_{\text{post}}^2\right)}$

Summary of linear case

Scalar case

- Gaussian prior: $m \sim \mathcal{N}\left(m_{\text{pr}}, \sigma_{\text{pr}}^2\right)$
- Linear model:

$$y = am + \eta$$

- Gaussian noise: $\eta \sim \mathcal{N}\left(0, \sigma_{\text{noise}}\right)$
- Posterior:

$$m|y \sim \mathcal{N}\left(m_{\text{post}}, \sigma_{\text{post}}^2\right)$$

with

$$\begin{aligned} \sigma_{\text{post}}^2 &= (a^2\sigma_{\text{noise}}^{-2} + \sigma_{\text{pr}}^{-2})^{-1} \\ m_{\text{post}} &= \sigma_{\text{post}}^2(a\sigma_{\text{noise}}^{-2} y + \sigma_{\text{pr}}^{-2} m_{\text{pr}}) \end{aligned}$$

Summary of linear case

Multivariate case

- Gaussian prior: $\boldsymbol{m} \sim \mathcal{N}(\boldsymbol{m}_{\mathsf{pr}}, \boldsymbol{\Gamma}_{\text{prior}})$
- Linear model:

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{m} + \boldsymbol{\eta}$$

- Gaussian noise: $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\text{noise}})$
- Posterior:

$$\boldsymbol{m}|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}_{\mathsf{post}}, \boldsymbol{\Gamma}_{\text{post}})$$

  with

$$\begin{aligned}
\boldsymbol{\Gamma}_{\text{post}} &= (\mathbf{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\mathbf{A} + \boldsymbol{\Gamma}_{\text{prior}}^{-1})^{-1} \\
\boldsymbol{m}_{\mathsf{post}} &= \boldsymbol{\Gamma}_{\text{post}}(\mathbf{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{y} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}\boldsymbol{m}_{\mathsf{pr}})
\end{aligned}$$

Gaussian linear case: remarks

Posterior pdf:

$$\pi_{\mathsf{post}}(\boldsymbol{m}|\boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{m}-\boldsymbol{m}_{\mathsf{post}})^T \boldsymbol{\Gamma}_{\text{post}}^{-1}(\boldsymbol{m}-\boldsymbol{m}_{\mathsf{post}})\right\}$$

with

$$\begin{aligned}\boldsymbol{\Gamma}_{\text{post}} &= (\mathbf{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\mathbf{A} + \boldsymbol{\Gamma}_{\text{prior}}^{-1})^{-1}\\ \boldsymbol{m}_{\mathsf{post}} &= \boldsymbol{\Gamma}_{\text{post}}(\mathbf{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{y} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}\boldsymbol{m}_{\mathsf{pr}})\end{aligned}$$

- Finding $\boldsymbol{m}_{\mathsf{post}}$: solve the linear system

  $$(\mathbf{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\mathbf{A} + \boldsymbol{\Gamma}_{\text{prior}}^{-1})\boldsymbol{m}_{\mathsf{post}} = \mathbf{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{y} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}\boldsymbol{m}_{\mathsf{pr}}$$

  SPD coefficient matrix; iterative methods for large-scale problems
- Note:

  $$\boldsymbol{m}_{\mathsf{post}} = \arg\max_{\boldsymbol{m}} \pi_{\mathsf{post}}(\boldsymbol{m}|\boldsymbol{y})$$

  $\boldsymbol{m}_{\mathsf{post}} \equiv$ the maximum a posteriori probability (MAP) estimator

Simple example: polynomial data fitting

- Fitting a line $y(t) = m_1 + m_2 t$ to measurement data y_i, $i = 1, \ldots, n$

$$\mathbf{A} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

- Linear model

$$\mathbf{A}\boldsymbol{m} + \boldsymbol{\eta} = \boldsymbol{y}$$

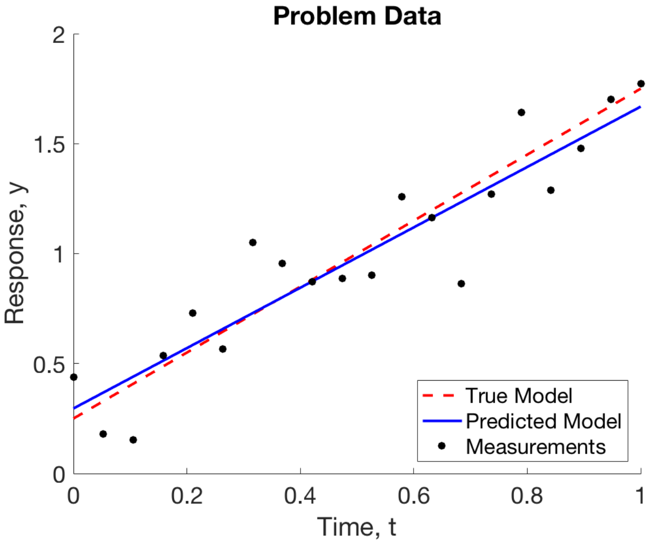
- Gaussian prior $\mathcal{N}(\boldsymbol{m}_{\text{pr}}, \boldsymbol{\Gamma}_{\text{prior}})$ and noise $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\text{noise}})$
- Sample problem setup $n = 20$, data generated from a true model $\boldsymbol{m} = (.25, 1.5)$ and we use

$$\boldsymbol{m}_{\text{pr}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \boldsymbol{\Gamma}_{\text{prior}} = \sigma_{\text{pr}}^2 \mathbf{I}_{2\times 2}, \quad \boldsymbol{\Gamma}_{\text{noise}} = \sigma_{\text{noise}} \mathbf{I}_{20\times 20}$$

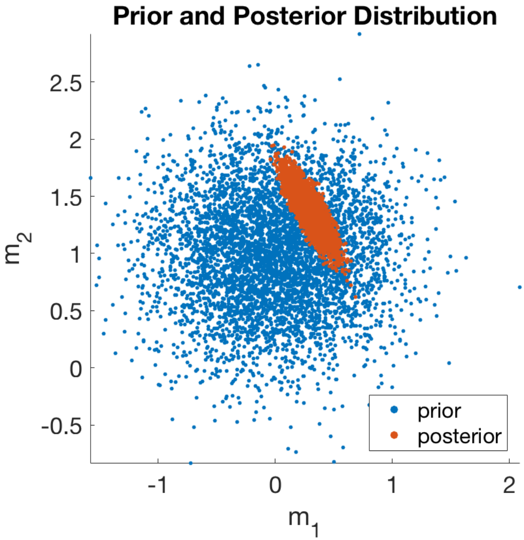
with

$$\sigma_{\text{pr}} = .5 \quad \sigma_{\text{noise}} = .2457$$

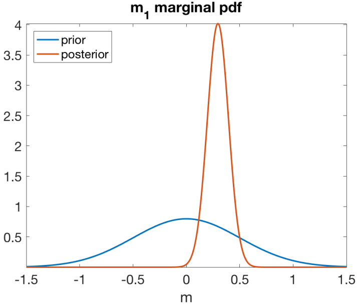
Simple example: polynomial data fitting (continued)

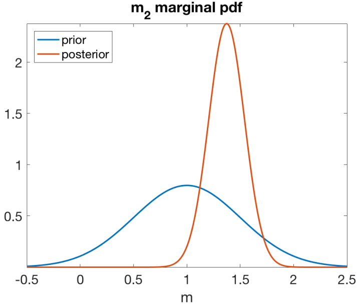


Simple example: polynomial data fitting (continued)



Simple example: polynomial data fitting (continued)





Bayesian approach to inverse problems

Physical/biological process

$$F(\boldsymbol{m}) \to \boldsymbol{y}$$

- F — physical process
- $\boldsymbol{m}$ — uncertain parameter; usually not directly observable
- $\boldsymbol{y}$ — results/observation (data)

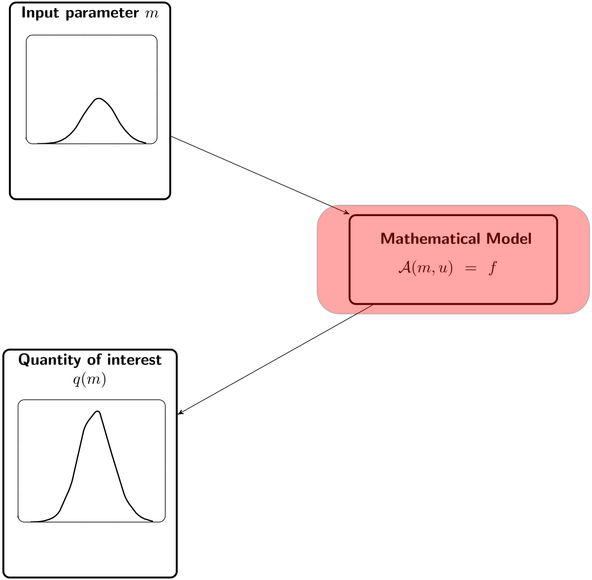
Mathematical model:

$$\boldsymbol{f}(\boldsymbol{m}) + \boldsymbol{\eta} = \boldsymbol{y}$$

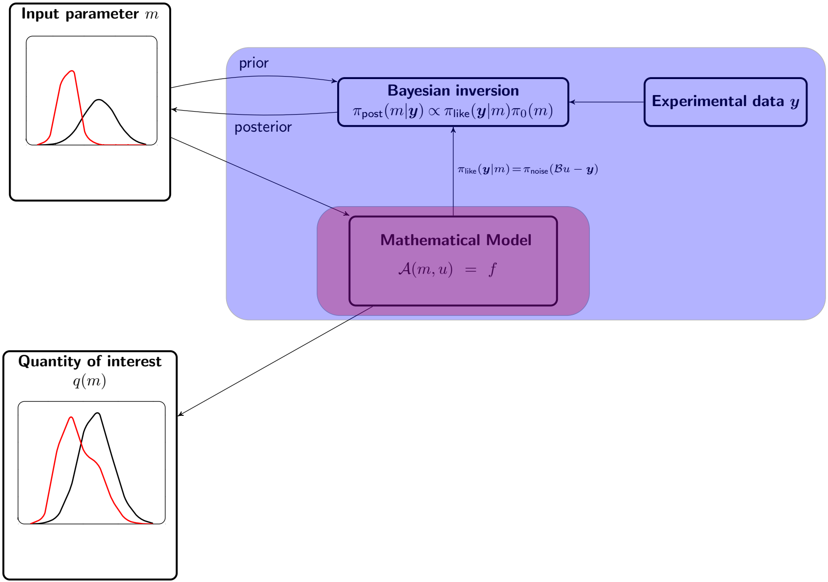
- $\boldsymbol{f}$ — forward mapping (mathematical model), aka parameter-to-observable map
- $\boldsymbol{\eta}$ — measurement or model errors

Goal: Combine model, data, and prior knowledge to estimate $\boldsymbol{m}$

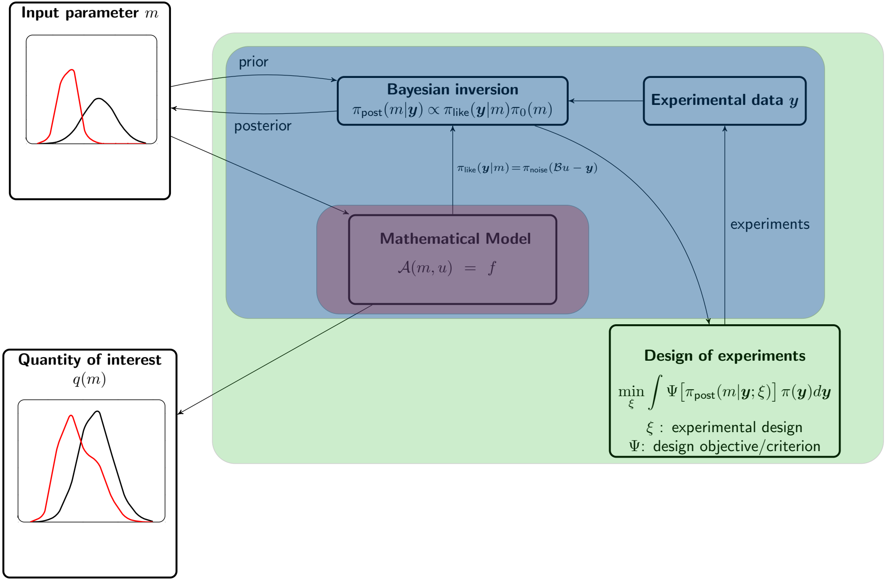
Modeling and decision making under uncertainty



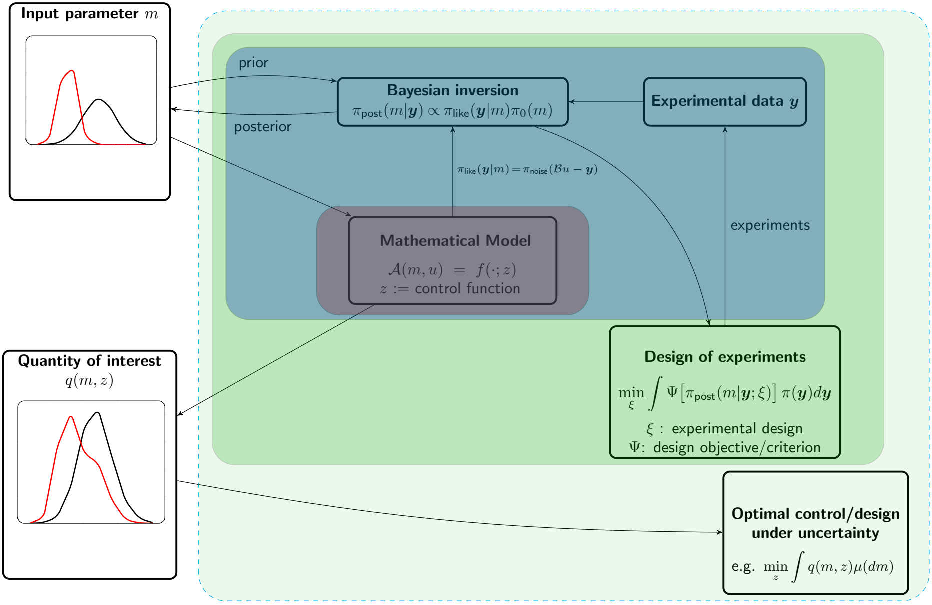
Modeling and decision making under uncertainty



Modeling and decision making under uncertainty



Modeling and decision making under uncertainty



Bayesian linear inverse problems

Assume linear parameter-to-observable map and additive Gaussian noise:

$$\boldsymbol{y} = \mathbf{F}\boldsymbol{m} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Gamma}_{\mathrm{noise}}\right)$$

Likelihood:

$$\pi_{\mathrm{like}}(\boldsymbol{y}|\boldsymbol{m}) \propto \exp\Big\{ -\frac{1}{2}(\mathbf{F}\boldsymbol{m} - \boldsymbol{y})^* \boldsymbol{\Gamma}_{\mathsf{noise}}^{-1} (\mathbf{F}\boldsymbol{m} - \boldsymbol{y}) \Big\}$$

Gaussian prior:

$$\pi_0(\boldsymbol{m}) \propto \exp(-\tfrac{1}{2}\boldsymbol{m}^T \boldsymbol{\Gamma}_{\mathrm{prior}}^{-1} \boldsymbol{m})$$

Bayesian linear inverse problems

For Bayesian linear inverse problem with Gaussian prior and noise, the posterior pdf is

$$\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{y}) \propto \exp\left\{-\tfrac{1}{2}(\boldsymbol{m}-\boldsymbol{m}_{\text{MAP}})^T(\mathbf{F}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\mathbf{F}+\boldsymbol{\Gamma}_{\text{prior}}^{-1})(\boldsymbol{m}-\boldsymbol{m}_{\text{MAP}})\right\}$$

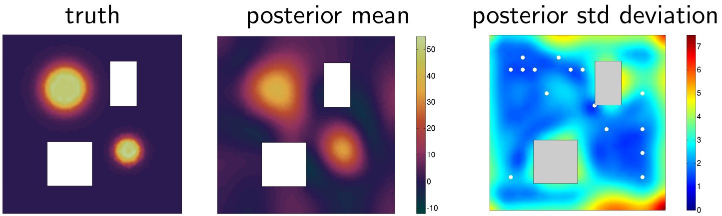
$$\Rightarrow \mu_{\text{post}} = \mathcal{N}\left(\boldsymbol{m}_{\text{MAP}}, \boldsymbol{\Gamma}_{\text{post}}\right)$$

$$\boldsymbol{\Gamma}_{\text{post}}^{-1} = \underbrace{\mathbf{F}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\mathbf{F}}_{\mathbf{H}_{\text{misfit}}} + \boldsymbol{\Gamma}_{\text{prior}}^{-1} \quad (= D^2_{\boldsymbol{m}}(-\log \pi_{\text{post}}))$$

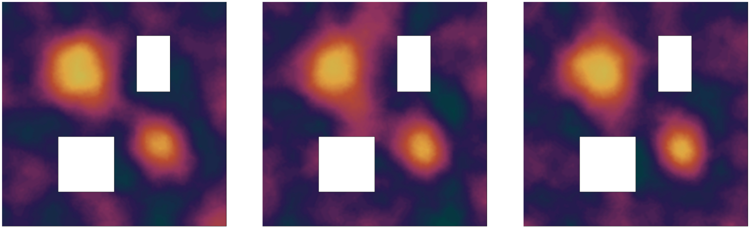
$$\boldsymbol{m}_{\text{MAP}} = \arg\min_{\boldsymbol{m}} \frac{1}{2}\|\mathbf{F}\boldsymbol{m}-\boldsymbol{d}\|^2_{\boldsymbol{\Gamma}_{\text{noise}}^{-1}} + \frac{1}{2}\left\langle \boldsymbol{\Gamma}_{\text{prior}}^{-1}\boldsymbol{m}, \boldsymbol{m}\right\rangle$$

Bayesian inversion of the initial condition for 2D advection-diffusion equation

- Posterior mean, and posterior variance

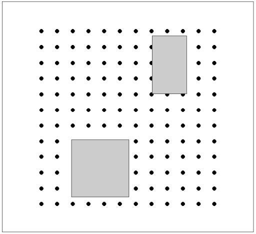


- Posterior samples



The optimal experimental design problem

A grid of candidate locations for observation points



- **Experimental design**: locations of observation points / sensors

$$\text{design} := \left\{ \begin{array}{c} \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_s} \\ w_1, \ldots, w_{N_s} \end{array} \right\}$$

- **Bayesian inversion**:

  data + likelihood, prior $\Longrightarrow$ posterior distribution of inversion parameter

- **Optimal experimental design (OED)**:

  Find sensor locations that result in minimized posterior uncertainty

Mathematical/computational challenges

- The inference problem is in an infinite-dimensional space
- Need to compute functionals of posterior covariance (inverse of Hessian, large, dense, expensive matvecs)
- With nonlinear inverse problems we are led to a bilevel optimization problem
- Optimal experimental design problem can have combinatorial complexity
- Conventional OED algorithms intractable for large-scale problem (due to high-dimensional parameters, expensive-to-evaluate PDE-based parameter-to-observable map, ...)

Key idea: understand and exploit problem structure

Optimal experimental design

- A-optimal design:

  Minimize "average variance" of parameter function m

- Covariance function: $c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathsf{Cov}\,\{m(\boldsymbol{x}_1), m(\boldsymbol{x}_2)\}$
- Covariance operator:

$$[\mathcal{C}_{\mathsf{post}} u](\boldsymbol{x}) = \int_{\mathcal{D}} c(\boldsymbol{x}_1, \boldsymbol{x}_2) u(\boldsymbol{x}_2)\, d\boldsymbol{x}_2$$

Optimal experimental design

- A-optimal design:

Minimize "average variance" of parameter function m

- Covariance function: $c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathsf{Cov}\left\{m(\boldsymbol{x}_1), m(\boldsymbol{x}_2)\right\}$
- Covariance operator:

$$[\mathcal{C}_{\mathsf{post}} u](\boldsymbol{x}) = \int_{\mathcal{D}} c(\boldsymbol{x}_1, \boldsymbol{x}_2) u(\boldsymbol{x}_2)\, d\boldsymbol{x}_2$$

- Variance of m at a given $\boldsymbol{x}$:

$$\mathsf{Var}\{m(\boldsymbol{x})\} = c(\boldsymbol{x}, \boldsymbol{x})$$

Optimal experimental design

- A-optimal design:

Minimize "average variance" of parameter function m

- Covariance function: $c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathsf{Cov}\left\{m(\boldsymbol{x}_1), m(\boldsymbol{x}_2)\right\}$
- Covariance operator:

$$[\mathcal{C}_{\mathsf{post}} u](\boldsymbol{x}) = \int_{\mathcal{D}} c(\boldsymbol{x}_1, \boldsymbol{x}_2) u(\boldsymbol{x}_2)\, d\boldsymbol{x}_2$$

- Variance of m at a given $\boldsymbol{x}$:

$$\mathsf{Var}\{m(\boldsymbol{x})\} = c(\boldsymbol{x}, \boldsymbol{x})$$

- Average variance:

$$\int_{\mathcal{D}} c(\boldsymbol{x}, \boldsymbol{x})\, d\boldsymbol{x}$$

Optimal experimental design

- A-optimal design:

  Minimize "average variance" of parameter function m

- Covariance function: $c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathsf{Cov}\left\{m(\boldsymbol{x}_1), m(\boldsymbol{x}_2)\right\}$
- Covariance operator:

$$[\mathcal{C}_{\mathsf{post}} u](\boldsymbol{x}) = \int_{\mathcal{D}} c(\boldsymbol{x}_1, \boldsymbol{x}_2) u(\boldsymbol{x}_2)\, d\boldsymbol{x}_2$$

- Variance of m at a given $\boldsymbol{x}$:

$$\mathsf{Var}\{m(\boldsymbol{x})\} = c(\boldsymbol{x}, \boldsymbol{x})$$

- Average variance:

$$\int_{\mathcal{D}} c(\boldsymbol{x}, \boldsymbol{x})\, d\boldsymbol{x} = \mathsf{tr}(\mathcal{C}_{\mathsf{post}})$$

Optimal experimental design

- A-optimal design:

Minimize "average variance" of parameter function m

- Covariance function: $c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathsf{Cov}\,\{m(\boldsymbol{x}_1), m(\boldsymbol{x}_2)\}$
- Covariance operator:

$$[\mathcal{C}_{\mathsf{post}} u](\boldsymbol{x}) = \int_{\mathcal{D}} c(\boldsymbol{x}_1, \boldsymbol{x}_2) u(\boldsymbol{x}_2)\, d\boldsymbol{x}_2$$

- Variance of m at a given $\boldsymbol{x}$:

$$\mathsf{Var}\{m(\boldsymbol{x})\} = c(\boldsymbol{x}, \boldsymbol{x})$$

- Average variance:

$$\int_{\mathcal{D}} c(\boldsymbol{x}, \boldsymbol{x})\, d\boldsymbol{x} = \mathsf{tr}(\mathcal{C}_{\mathsf{post}})$$

- Optimal design criterion:

Choose a "design" to minimize $\mathsf{tr}(\mathcal{C}_{\mathsf{post}})$

A-optimal experimental design with sparsity control

$$\begin{aligned} \underset{\boldsymbol{w} \in \mathbb{R}^{N_s}}{\text{minimize}} \quad & \mathrm{tr}\big[\boldsymbol{\Gamma}_{\text{post}}(\boldsymbol{w})\big] + \gamma P(\boldsymbol{w}) \\ \text{subject to} \quad & \mathbf{0} \le \boldsymbol{w} \le \mathbf{1} \end{aligned} \qquad (*)$$

- $\boldsymbol{\Gamma}_{\text{post}}(\boldsymbol{w}) = (\sigma_{\text{noise}}^{-2}\mathbf{F}^*\mathbf{W}\mathbf{F} + \boldsymbol{\Gamma}_{\text{prior}}^{-1})^{-1}$, $\mathbf{W}$: diagonal matrix with w_i on its diagonal
- $P(\boldsymbol{w})$: penalty term, $\gamma > 0$ (e.g., $P(\boldsymbol{w}) = \sum_j w_j$)
- Need trace of inverse Hessian and its derivative
- Need many applications of the forward operator $\mathbf{F}$ $\implies$ many PDE solves

Randomized trace estimation

- $\mathbf{A} \in \mathbb{R}^{n \times n}$ — symmetric
- Trace estimator:

$$\mathrm{tr}(\mathbf{A}) \approx \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \boldsymbol{z}_i^T \mathbf{A} \boldsymbol{z}_i, \quad \boldsymbol{z}_i \text{ — random vectors}$$

- Gaussian trace estimator: $\boldsymbol{z}_i$ independent draws from $\mathcal{N}(\mathbf{0}, \mathbf{I})$
- For $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$$\mathrm{E}\{\boldsymbol{z}^T \mathbf{A} \boldsymbol{z}\} = \mathrm{tr}(\mathbf{A}) \qquad \mathrm{Var}\{\boldsymbol{z}^T \mathbf{A} \boldsymbol{z}\} = 2 \|\mathbf{A}\|_F^2$$

Efficient means of approximating trace of posterior covariance

M. Hutchinson, A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines (1990).

H. Avron and S. Toledo, Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix (2011).

A-optimal design: the objective function

- Randomized trace estimator:

$$\mathsf{tr}\big[\boldsymbol{\Gamma}_{\mathrm{post}}(\boldsymbol{w})\big] \approx \frac{1}{N}\sum_{i=1}^{N} \langle \boldsymbol{z}_i, \boldsymbol{\Gamma}_{\mathrm{post}}(\boldsymbol{w})\boldsymbol{z}_i \rangle =: \phi(\boldsymbol{w})$$

- $\boldsymbol{z}_i$ random vectors (e.g. Gaussian)
- Note: $\boldsymbol{\Gamma}_{\mathrm{post}} = \mathbf{H}^{-1}$

$$\begin{aligned} \mathbf{H} &= \sigma_{\mathsf{noise}}^{-2}\mathbf{F}^*\mathbf{W}\mathbf{F} + \boldsymbol{\Gamma}_{\mathrm{prior}}^{-1} \\ &= \mathbf{H}_{\mathsf{misfit}}(\boldsymbol{w}) + \boldsymbol{\Gamma}_{\mathrm{prior}}^{-1} \end{aligned}$$

(for notational convenience, let $\sigma_{\mathsf{noise}} = 1$ from now on)

Application of inverse Hessian

- Inverse of the Hessian:

$$
\begin{aligned}
\mathbf{H}^{-1} &= \left(\mathbf{H}_{\mathsf{misfit}} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}\right)^{-1} \\
&= \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\Big(\underbrace{\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\mathbf{H}_{\mathsf{misfit}}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}}_{\tilde{\mathbf{H}}_{\mathsf{misfit}}} + \mathbf{I}\Big)^{-1}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}
\end{aligned}
$$

Application of inverse Hessian

- Inverse of the Hessian:

$$\begin{aligned}\mathbf{H}^{-1} &= \left(\mathbf{H}_{\mathsf{misfit}} + \mathbf{\Gamma}_{\text{prior}}^{-1}\right)^{-1} \\ &= \mathbf{\Gamma}_{\text{prior}}^{1/2}\Big(\underbrace{\mathbf{\Gamma}_{\text{prior}}^{1/2}\mathbf{H}_{\mathsf{misfit}}\mathbf{\Gamma}_{\text{prior}}^{1/2}}_{\tilde{\mathbf{H}}_{\mathsf{misfit}}} + \mathbf{I}\Big)^{-1}\mathbf{\Gamma}_{\text{prior}}^{1/2}\end{aligned}$$

- Low-rank approximation:

$$\tilde{\mathbf{H}}_{\mathsf{misfit}} \approx \sum_{i=1}^{r} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T$$

Application of inverse Hessian

- Inverse of the Hessian:

$$\mathbf{H}^{-1} = \left(\mathbf{H}_{\text{misfit}} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}\right)^{-1}$$
$$= \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\Big(\underbrace{\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\mathbf{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}}_{\tilde{\mathbf{H}}_{\text{misfit}}} + \mathbf{I}\Big)^{-1}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$$

- Low-rank approximation:

$$\tilde{\mathbf{H}}_{\text{misfit}} \approx \sum_{i=1}^{r} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T$$

- Efficient $\mathbf{H}^{-1}$ apply:

$$\mathbf{H}^{-1}\boldsymbol{q} \approx \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\left(\mathbf{V}_r\boldsymbol{\Lambda}_r\mathbf{V}_r^* + \mathbf{I}\right)^{-1}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{q}$$
$$= \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\left(\mathbf{I} - \mathbf{V}_r\mathbf{D}_r\mathbf{V}_r^*\right)\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{q} \qquad \text{(Sherman-Morrison-Woodbury)}$$

- $\mathbf{D} = \text{diag}\{\lambda_1/(1+\lambda_1), \ldots, \lambda_r/(1+\lambda_r)\}$

A-optimal design: the gradient

Objective function: $\phi(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} \langle \boldsymbol{z}_i, \boldsymbol{q}_i \rangle \qquad \boldsymbol{q}_i = \mathbf{H}^{-1}(\boldsymbol{w})\boldsymbol{z}_i$

A-optimal design: the gradient

Objective function: $\phi(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} \langle \boldsymbol{z}_i, \boldsymbol{q}_i \rangle \qquad \boldsymbol{q}_i = \mathbf{H}^{-1}(\boldsymbol{w})\boldsymbol{z}_i$

- Gradient:

$$\frac{\partial}{\partial w_j}\mathbf{H}^{-1}(\boldsymbol{w}) = -\mathbf{H}^{-1}(\boldsymbol{w})\partial_j \mathbf{H}(\boldsymbol{w})\mathbf{H}^{-1}(\boldsymbol{w}) = -\mathbf{H}^{-1}(\boldsymbol{w})\partial_j \mathbf{H}_{\mathsf{misfit}}(\boldsymbol{w})\mathbf{H}^{-1}(\boldsymbol{w})$$

$$\frac{\partial \phi}{\partial w_j} = -\frac{1}{N}\sum_{i=1}^{N} \langle \boldsymbol{q}_i, \partial_j \mathbf{H}_{\mathsf{misfit}} \boldsymbol{q}_i \rangle \qquad j = 1, \ldots, N_s$$

A-optimal design: the gradient

Objective function: $\phi(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\langle \boldsymbol{z}_i, \boldsymbol{q}_i\rangle \qquad \boldsymbol{q}_i = \mathbf{H}^{-1}(\boldsymbol{w})\boldsymbol{z}_i$

- Gradient:

$$\frac{\partial}{\partial w_j}\mathbf{H}^{-1}(\boldsymbol{w}) = -\mathbf{H}^{-1}(\boldsymbol{w})\partial_j\mathbf{H}(\boldsymbol{w})\mathbf{H}^{-1}(\boldsymbol{w}) = -\mathbf{H}^{-1}(\boldsymbol{w})\partial_j\mathbf{H}_{\mathsf{misfit}}(\boldsymbol{w})\mathbf{H}^{-1}(\boldsymbol{w})$$

$$\frac{\partial\phi}{\partial w_j} = -\frac{1}{N}\sum_{i=1}^{N}\langle \boldsymbol{q}_i, \partial_j\mathbf{H}_{\mathsf{misfit}}\boldsymbol{q}_i\rangle \qquad j = 1,\dots,N_s$$

- For each i, need one application of parameter-to-observable map $\mathbf{F}$:

$$\boldsymbol{q}_i \overset{\mathbf{F}}{\mapsto} (\boldsymbol{d}_1^T, \boldsymbol{d}_2^T, \dots \boldsymbol{d}_{N_\tau}^T)^T,$$

where $\boldsymbol{d}_s = (d_s^1, d_s^2, \dots, d_s^{N_s})^T$. Then,

$$\langle \boldsymbol{q}_i, \partial_j\mathbf{H}_{\mathsf{misfit}}\boldsymbol{q}_i\rangle = \sum_{s=1}^{N_\tau} d_s^j d_s^j.$$

The forward operator

- Need many forward solves in the optimization process
- Idea: $\mathbf{F}$ is low-rank (often)
- Note:

$$\mathbf{F} = \underbrace{\mathbf{B}}_{\text{observation operator}} \underbrace{\mathbf{S}}_{\text{solution operator}}$$

- Idea: compute a low-rank SVD surrogate for $\mathbf{F}$

$$\mathbf{F} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

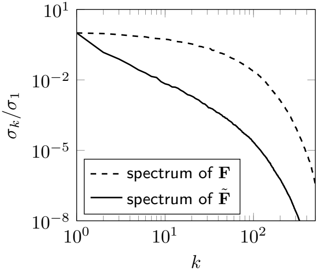
- Randomized SVD:
  - Independent matvecs (forward/adjoint) — can do in parallel
  - Simple but very robust algorithms
  - Backed by rigorous theory
  - Almost deterministic behavior

N. Halko, P.G. Martinsson, J.A. Tropp, Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. SIAM Review (2011).

Randomized SVD for forward operator

- Idea: $\mathbf{F}$ is low-rank (often); better idea: $\tilde{\mathbf{F}} = \mathbf{F}\mathbf{\Gamma}_{\text{prior}}^{1/2}$ is even more so ...

$$\tilde{\mathbf{F}} \approx \sum_{j=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$$



SVD surrogate for $\tilde{\mathbf{F}}$ $\Rightarrow$ no forward PDE solves in OED algorithm

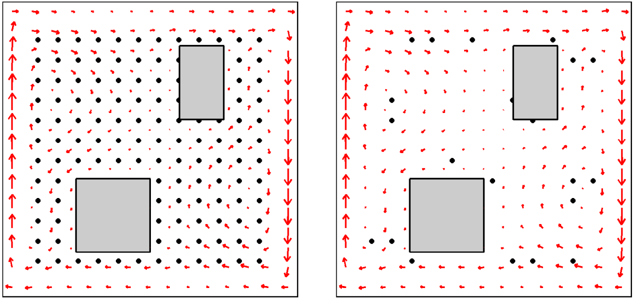
A-optimal design: computational cost

At the cost of an upfront SVD for $\mathbf{F}$:

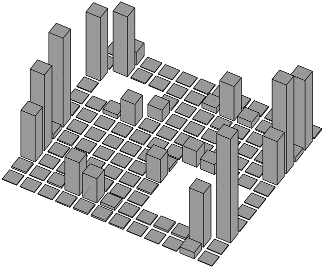
- No PDE solves in the optimization process
- Efficient computation of cost/derivatives
- Independent of temporal/spatial mesh

A-optimal design: numerical results

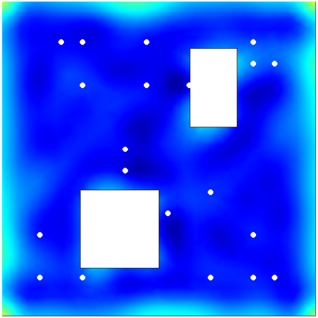
- Sensor allocation



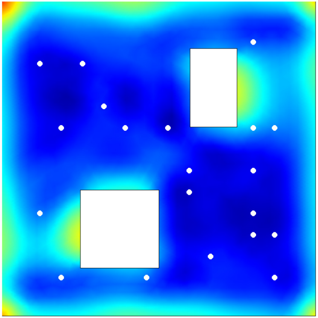
- Weight distribution



A-optimal design: the variance field

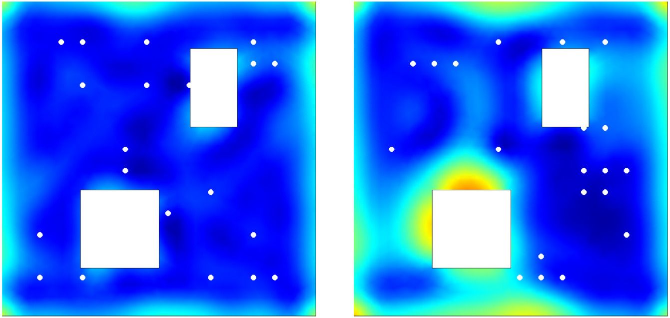


Optimal



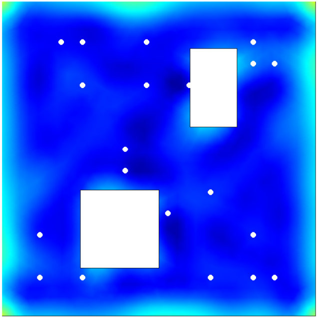
Sub-optimal

A-optimal design: the variance field

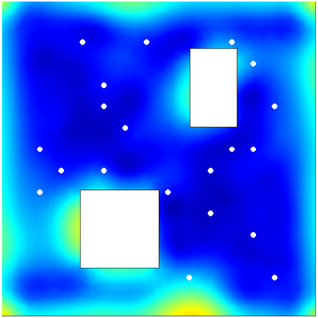


Optimal   Sub-optimal

A-optimal design: the variance field

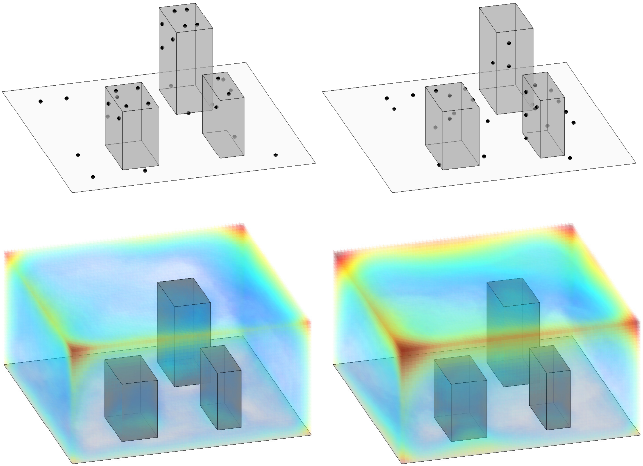


Optimal



Sub-optimal

OED for 3D model (parameter dim $\sim 10^4$)



A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas. A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ^0-sparsification. SISC. 2014.

Design of nonlinear inverse problems: a glimpse

- Nonlinear parameter-to-observable map
- Posterior not Gaussian in general
- No closed-form expression for measures of posterior uncertainty
- A straightforward idea: use linearized the model $\Longrightarrow$ locally optimum designs
- Bayesian optimal design: maximize expected information gain
- To illustrate the issues let's consider a simpler case, Bayes risk minimization, where we need to minimize

$$\Psi(\boldsymbol{w}) := \int_{\mathscr{H}} \int_{\mathbb{R}^d} \|m_{\mathsf{MAP}}(\boldsymbol{y}, \boldsymbol{w}) - m\|^2 \, \pi_{\text{like}}(\boldsymbol{y}|m) d\boldsymbol{y} \, \mu_{\mathsf{pr}}(dm). \tag{1}$$

$\Longrightarrow$ bilevel PDE constrained optimization

Summary and outlook

Summary:

- Bayesian approach to inverse problems
- A-Optimal sensor placement for PDE-based Bayesian linear inverse problems
- Scalable algorithms
- Efficient computation of OED objective/gradient (randomized methods in numerical linear algebra, low-rank approximations, ...)

Outlook:

- OED for nonlinear inverse problems governed by PDEs
- Sensitivity analysis of inverse problems with respect to problem data
- Goal oriented OED (OED for prediction)
- OED under model uncertainty
- Sequential design of experiments

Further reading

Books on OED:

- D. Ucinski. Optimal measurement methods for distributed parameter system identification. 2005.
- A. C. Atkinson and A. N. Donev. Optimum Experimental Designs. 1992.
- F. Pukelsheim. Optimal Design of Experiments. 1993.

Some papers

- E. Haber, L. Horesh, and L. Tenorio. Numerical methods for experimental design of large-scale linear ill-posed inverse problems. Inverse Problems, 2008.
- E. Haber, L. Horesh, and L. Tenorio. Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems. Inverse Problems, 2010.
- M. Chung and E. Haber. Experimental design for biological systems. SICON, 2012.
- X. Huan and Y. M. Marzouk. Simulation-based optimal Bayesian experimental design for nonlinear systems. JCP, 2013.
- Sunseri, Hart, van Bloemen Waanders, and Alexanderian. Hyper-Differential Sensitivity Analysis for Inverse Problems Constrained by Partial Differential Equations. Inverse Problems, 2020.
- A. Alexanderian, Optimal experimental design for infinite-dimensional Bayesian inverse problems governed by PDEs: a review, Inverse Problems, 2021.