Optimal design of experiments for inverse problems: OED for linearized Bayesian problems in high dimensions

Georg Stadler

Courant Institute, New York University
stadler@cims.nyu.edu

June 29, 2023

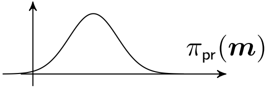
Bayes formula (finite dimensions)

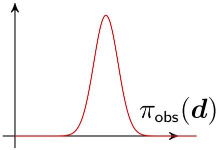
Given:

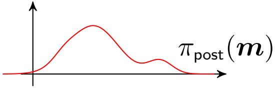
$$
\begin{aligned}
\pi_{\text{pr}}(\boldsymbol{m}) &: \text{ prior p.d.f. of model parameters } \boldsymbol{m} \\
\pi_{\text{obs}}(\boldsymbol{d}) &: \text{ prior p.d.f. of measurement error } \boldsymbol{d} \\
\pi_{\text{model}}(\boldsymbol{d}|\boldsymbol{m}) &: \text{ conditional p.d.f. combining } \boldsymbol{d} \text{ and } \boldsymbol{m} \text{ (model)}
\end{aligned}
$$

Then, the *posterior p.d.f. of the model parameters* is given by:

$$\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d}) \propto \pi_{\text{pr}}(\boldsymbol{m})\, \pi_{\text{like}}(\boldsymbol{d}|\boldsymbol{m})$$







model connecting $\boldsymbol{m}$ and $\boldsymbol{d}$

Outline

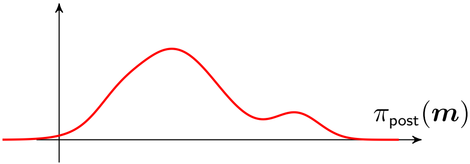
Posterior approximations

Linear/linearized problems

Approximation of the posterior distribution

Despite the explicit form of $\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d})$, its exploration is difficult due to:

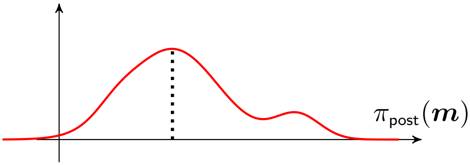
- the high/infinite dimension of $\boldsymbol{m}$
- the expensive PDE-based parameter-to-observable map $\boldsymbol{f}$



Approximation of the posterior distribution

Despite the explicit form of $\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d})$, its exploration is difficult due to:

- the high/infinite dimension of $\boldsymbol{m}$
- the expensive PDE-based parameter-to-observable map $\boldsymbol{f}$



Approximation I: MAP estimation

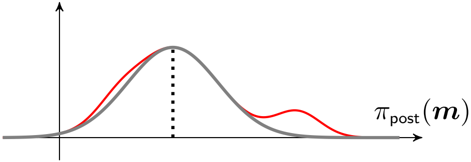
Find the maximum a posteriori (MAP) point.

- requires solution of PDE-constrained optimization problem $\sim$ deterministic inversion
- Computation of derivatives using adjoint methods: 2(+) PDE solves per gradient

Approximation of the posterior distribution

Despite the explicit form of $\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d})$, its exploration is difficult due to:

- the high/infinite dimension of $\boldsymbol{m}$
- the expensive PDE-based parameter-to-observable map $\boldsymbol{f}$



Approximation II: Gaussian around MAP point

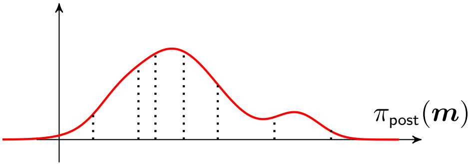
Use a Gaussian approximation around the MAP point based on second derivatives (Hessians) of $\mathcal{J}$

- requires the Hessian matrix which is not directly available for PDE-constrained problems... but the Hessian can be applied to vectors by solving 2(+) PDEs

Approximation of the posterior distribution

Despite the explicit form of $\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d})$, its exploration is difficult due to:

- the high/infinite dimension of $\boldsymbol{m}$
- the expensive PDE-based parameter-to-observable map $\boldsymbol{f}$



Approximation III: Sampling

Use sampling (Metropolis Hastings/Marcov chain Monte Carlo) to approximate statistics

- sampling in high dimensions is challenging, requires many evaluations of $\boldsymbol{f}$ (and good proposal distributions)
- Exploit low rank properties of update of distribution; "feels" the curse of the effective dimensionality rather than the discretization dimension

Outline

Posterior approximations

Linear/linearized problems

Gaussian distributions: Finite dimensions

Finite-dimensional $\mathcal{N}(\boldsymbol{m}_0, \boldsymbol{\Gamma})$, $\boldsymbol{m}_0 \in \mathbb{R}^n$, $\Gamma \in \mathbb{R}^{n\times n}$ spd. Interested in:

- Samples from this distribution: $\boldsymbol{s} = \boldsymbol{m}_0 + \boldsymbol{\Gamma}^{1/2}\boldsymbol{n}$, where n_i is iid.
- Diagonal of $\boldsymbol{\Gamma}$ contains variances (tomorrow we will use its trace as measure for information gain in OED)
- Need to apply $\boldsymbol{\Gamma}$ and its inverse (precision matrix) to vectors fast
- Gaussian density:

$$\pi(\boldsymbol{m}) \propto \exp\left(-\frac{1}{2} \parallel \boldsymbol{m} - \boldsymbol{m}_0 \parallel^2_{\boldsymbol{\Gamma}^{-1}}\right)$$

Gaussian distributions: Infinite dimensions

Let $\mathcal{D} \subset \mathbb{R}^d$, $d \in \mathbb{N}$ be a bounded domain.

Infinite-dimensions: $\mathcal{N}(\boldsymbol{m}_0, \mathcal{C})$, $\boldsymbol{m}_0 \in L_2(\mathcal{D})$, $\mathcal{C} : L_2(\mathcal{D}) \mapsto L_2(\mathcal{D})$ a trace-class operator.

- Defines a distribution of functions
- In infinite dimensions, no Lebesgue density exists. Typically use a Gaussian as reference density (e.g. in formulation of the Bayes theorem in infinite dimensions)
- Discretization possible but needs to be careful with inner products etc.
- I use specific Matern kernel covariance operators related to PDE operators $\rightarrow$ fast solvers:

$$\mathcal{C} = (-\alpha\Delta + \beta I)^{-\delta}, \qquad \alpha, \beta > 0$$

This operators is trace-class for $\delta = 1$ in 1D, and $\delta = 2$ in 2D and 3D.

Gaussian prior and noise

Assume additive Gaussian noise $\boldsymbol{e}$ in the measurements

$$\boldsymbol{d} = \boldsymbol{f}(\boldsymbol{m}) + \boldsymbol{e}, \qquad \boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{\text{noise}})$$

and Gaussian prior $\boldsymbol{\Gamma}_{\text{pr}}$ with mean $\boldsymbol{m}_0$, then the posterior density is:

$$\pi_{\text{post}}(\boldsymbol{m}) \propto \exp\left(-\frac{1}{2} \parallel \boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d} \parallel^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} - \frac{1}{2} \parallel \boldsymbol{m} - \boldsymbol{m}_{\text{pr}} \parallel^2_{\boldsymbol{\Gamma}^{-1}_{\text{pr}}}\right)$$

The "maximum a posteriori" point is

$$\begin{aligned} \boldsymbol{m}_{\text{MAP}} &\stackrel{\text{def}}{=} \arg \max_{\boldsymbol{m}} \pi_{\text{post}}(\boldsymbol{m}) \\ &= \arg \min_{\boldsymbol{m}} \frac{1}{2} \parallel \boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d} \parallel^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2} \parallel \boldsymbol{m} - \boldsymbol{m}_{\text{pr}} \parallel^2_{\boldsymbol{\Gamma}^{-1}_{\text{pr}}} \end{aligned}$$

$\Rightarrow$ deterministic inverse problem with appropriate weighted norms!

Gaussian prior and noise

- Solution of inverse problem in Bayesian approach is this distribution $\pi_{\text{post}}(\boldsymbol{m}) = \pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d})$.
- Can be written down explicitly, but not immediately useful to work with. Want: ability to draw samples, compute moments (mean, covariance,...)
- Characterization of distribution (e.g., through moments) gives a complete picture of the inverse problem
- Main reference if $\boldsymbol{m}$ is a function: A. Stuart, Acta Numerica '10
- Goal: making things computable for (discretized) parameter functions $\boldsymbol{m}$ and expensive-to-compute maps $\boldsymbol{f}$

Gaussian prior and noise and linear $\boldsymbol{f}$

Assume that

$$\boldsymbol{d} = \boldsymbol{F}\boldsymbol{m}$$

is linear. Then, the posterior p.d.f. is given by:

$$\pi_{\text{post}}(\boldsymbol{m}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{m}_{\text{MAP}})^T(\boldsymbol{F}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{F} + \boldsymbol{\Gamma}_{\text{pr}}^{-1})(\boldsymbol{m} - \boldsymbol{m}_{\text{MAP}})\right)$$

Thus, the posterior is also Gaussian, i.e.,

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

The covariance matrix is the inverse Hessian of the negative log of the posterior, i.e., the usual function from deterministic inversion:

$$\begin{aligned} \boldsymbol{\Gamma}_{\text{post}}^{-1} &= \boldsymbol{F}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{F} + \boldsymbol{\Gamma}_{\text{pr}}^{-1} \\ &= \nabla^2_{\boldsymbol{m}}(-\log \pi_{\text{post}}) \end{aligned}$$

Approximation of the covariance matrix

The covariance matrix is the inverse Hessian.
Idea: Never form $\boldsymbol{H}$ explicitly, but:

- use that $\boldsymbol{H}$ is the sum of the misfit Hessian, which often is a compact operator, and of the inverse prior, usually a differential operator:

$$\boldsymbol{H} = \boldsymbol{F}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \boldsymbol{F} + \boldsymbol{\Gamma}_{\text{pr}}^{-1}$$

- Use a low-rank approximation of the first term (requires fixed number of incremental forward/adjoint solutions)
- Use Sherman-Morrison-Woodbury formula for inversion

Low-rank approximation

$$\begin{aligned}\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{H}^{-1} &= \left(\boldsymbol{F}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{F} + \boldsymbol{\Gamma}_{\text{pr}}^{-1}\right)^{-1} \\ &= \boldsymbol{\Gamma}_{\text{pr}}^{1/2}\left(\boldsymbol{\Gamma}_{\text{pr}}^{1/2}\boldsymbol{F}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{F}\,\boldsymbol{\Gamma}_{\text{pr}}^{1/2} + \boldsymbol{I}\right)^{-1}\boldsymbol{\Gamma}_{\text{pr}}^{1/2} \\ &\overset{\text{low rank approx.}}{\approx} \boldsymbol{\Gamma}_{\text{pr}}^{1/2}\left(\boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^T + \boldsymbol{I}\right)^{-1}\boldsymbol{\Gamma}_{\text{pr}}^{1/2} \\ &\overset{SMW}{=} \boldsymbol{\Gamma}_{\text{pr}}^{1/2}\left[\boldsymbol{I} - \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T + \mathcal{O}\left(\sum_{i=r+1}^{n}\frac{\lambda_i}{\lambda_i+1}\right)\right]\boldsymbol{\Gamma}_{\text{pr}}^{1/2},\end{aligned}$$

where $\boldsymbol{V}_r, \boldsymbol{\Lambda}_r$ are the eigenvalues and eigenvectors of the prior-preconditioned Hessian, and $\boldsymbol{D}_r = \mathrm{diag}(\lambda_i/(\lambda_i+1))$.

Low-rank ideas: Liberty, et al. (2007); Biros & Chaillat (2011); Demanet et al. (2011); Halko, Martinsson, Tropp (2011);...

Low-rank-based posterior covariance

The posterior covariance is given by the prior minus the information gained by the data:

$$\boldsymbol{\Gamma}_{\text{post}} \approx \boldsymbol{\Gamma}_{\text{pr}} - \boldsymbol{\Gamma}_{\text{pr}}^{1/2} \boldsymbol{V}_r \boldsymbol{D}_r \boldsymbol{V}_r^T \boldsymbol{\Gamma}_{\text{pr}}^{1/2}$$

Full Hessian is never assembled. Complexity of Newton step, the sampling from the posterior p.d.f., computation of the variance etc. are scalable (i.e. the required number of PDE solves is independent of the discretization) if:

- the prior-preconditioned misfit Hessian is compact with mesh-independent spectrum (as for many inverse problems)
- low-rank algorithms (we use Lanczos or a randomized singular value decomposition) resolves the dominant spectrum in a fixed number of Hessian-vector products
- matrix-free Hessian-vector products using adjoint methods

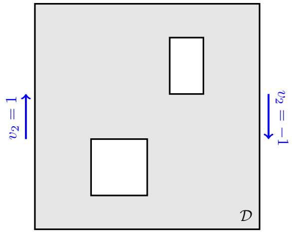
Time dependent advection-diffusion

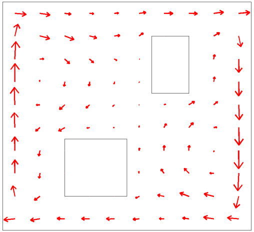
https://hippylib.github.io/tutorials_v3.0.0/4_AdvectionDiffusionBayesian/

- Forward problem: Given initial condition m, solve

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

Inverse problem: Recover unknown initial condition $m = m(x)$ from point observations at time instances. Optimal design will be *location of point sensors*.





- $m = m(x)$: *unknown* initial condition
- $\boldsymbol{v}$: Velocity field — solve steady-state Navier-Stokes