# Bayesian Inverse Problems
# Oral Exam Notes

Sonia Reilly
NYU Courant

October 2023

These notes are based on Georg Stadler's and Andy Davis's Spring 2022 lecture notes for Bayesian Inverse Problems, as well as Alen Alexandarian's brief note on the KL expansion and Lindgren, Rue, and Lindström's paper on the link between Matérn kernels and SPDEs.

## 1 Bayes' Rule

- Fundamentally defined as
$$P(A|B) = \frac{P(A,B)}{P(B)}.$$
For pdfs, Bayes' rule takes the form
$$\pi(x|y) = \frac{\pi(x,y)}{\pi(y)} = \frac{\pi(x)\pi(y|x)}{\pi(y)},$$
which can be derived from the other form by letting $A = \{X \in dx\}$ and $B = \{Y \in dy\}$ and letting $\max(|dx|, |dy|) \to 0$.
- Simple consequence of the definition of conditional probability.
- This can also be written in the context of Bayesian inverse problems as
$$\pi_{\text{post}}(x) \propto \pi_{\text{prior}}(x)\pi_{\text{like}}(y|x),$$
where $x$ is the parameter, $y$ is the data, and the posterior is expressed as a product of the prior and the likelihood. The denominator $\pi(y)$ is called the model evidence, which can be used for Occam's razor model selection, but is often omitted when the prior only needs to be known up to a normalization factor.

## 2 The Linear Gaussian Bayesian Inverse Problem

- The simplest Bayesian inverse problem has forward model
$$Y = AX + \epsilon,$$
where the data $Y$ is a linear function of the parameter $X$ with some added noise $\epsilon$. We assume the noise is chosen from a Gaussian distribution $N(0, \Sigma_{\text{noise}})$ and the prior over the parameter is a Gaussian distribution $N(\mu, \Sigma_{\text{pr}})$.

- The posterior distribution in this case is also a multivariate Gaussian. This can be shown by writing the posterior distribution using Bayes' rule as

$$\pi_{\text{post}}(X) \propto \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma_{\text{pr}}^{-1}(X-\mu) - \frac{1}{2}(Y-AX)^T\Sigma_{\text{noise}}^{-1}(Y-AX)\right)$$

and completing the square.

# 3 Approximations to the Posterior

If the problem is nonlinear, the posterior may not be Gaussian, even if the prior and likelihood are Gaussian. Exploring the posterior can be difficult in high dimensions, so we often seek approximations to the posterior.

## 3.1 MAP estimate

- Simplest approximation: maximum a-posteriori (MAP) estimate:

$$x_{\text{MAP}} = \arg\max_x \pi_{\text{post}}(x) = \arg\min_x \frac{1}{2}||f(x)-y||^2_{\Gamma^{-1}_{\text{noise}}} + \frac{1}{2}||x-\mu||^2_{\Gamma^{-1}_{\text{pr}}}.$$

The second formulation comes from taking the negative log of the expression for the posterior using Bayes' rule, where the forward model is $y = f(x) + \epsilon$ and the prior and likelihood are Gaussian.

- This is the same as the deterministic inverse problem with a Tikhonov regularization term. So the deterministic inverse problem gives the MAP estimate of the Bayesian posterior distribution.

## 3.2 Laplace approximation

- The Laplace approximation is a Gaussian approximation to the posterior at the MAP point.
- Letting $h(x) = -\log(\pi(x))$, we can define $\pi(x) = \exp(-h(x))$. Then we take a second-order Taylor expansion of $h(x)$ around $x_{\text{MAP}}$,

$$\tilde{h}(x) = h(x_{\text{MAP}}) + \cancelto{0}{\nabla h(x_{\text{MAP}})}\cdot(x - x_{\text{MAP}}) + \frac{1}{2}(x-x_{\text{MAP}})^T\nabla^2 h(x_{\text{MAP}})(x-x_{\text{MAP}}).$$

The first-order term goes to 0 since the MAP point is a maximum. The Laplace approximation is given by

$$\pi_{\text{Laplace}}(x) = \exp(-\tilde{h}(x)).$$

- Therefore the Laplace approximation is the (multivariate) Gaussian with the MAP estimate as the mean and the local inverse Hessian of the negative log posterior, $(\nabla^2 h(x_{\text{MAP}}))^{-1}$, as the covariance.
- If the inverse problem is linear, $y = Ax + \epsilon$, then the Hessian is

$$\nabla^2 h(x_{\text{MAP}}) = A^T\Gamma^{-1}_{\text{noise}}A + \Gamma^{-1}_{\text{pr}}.$$

If the problem is nonlinear, we get the same Hessian where $A$ is now the Jacobian of $f(x)$ at the MAP point.

### 3.3 Low-rank approximation

- For very high dimension, computing or storing the Jacobian in order to find the posterior covariance of the Laplace approximation may be impossible. So instead of explicitly computing the Jacobian, we use the Sherman-Morrison-Woodbury low rank update formula to express the Laplace posterior covariance as an update on the prior covariance.

- We write

$$\begin{aligned}\Gamma_{\text{post}} &= \left(A^T\Gamma_{\text{noise}}^{-1}A + \Gamma_{\text{pr}}\right)^{-1} \\ &= \Gamma_{\text{pr}}^{1/2}\left(\Gamma_{\text{pr}}^{1/2}A^T\Gamma_{\text{noise}}^{-1}A\Gamma_{\text{pr}}^{1/2} + I\right)^{-1}\Gamma_{\text{pr}}^{1/2} \\ &\approx \Gamma_{\text{pr}}^{1/2}\left(V_r\Lambda_r V_r^T + I\right)^{-1}\Gamma_{\text{pr}}^{1/2} \\ &= \Gamma_{\text{pr}} - \Gamma_{\text{pr}}^{1/2}V_r D_r V_r^T\Gamma_{\text{pr}}^{1/2},\end{aligned}$$

  using the fact that the prior-preconditioned misfit Hessian $\Gamma_{\text{pr}}^{1/2}A^T\Gamma_{\text{noise}}^{-1}A\Gamma_{\text{pr}}^{1/2}$ is often close to low-rank, so it can be approximated by some $V_r\Lambda_r V_r^T$ with diagonal $\Lambda_r \in \mathbb{R}^{r\times r}$. We have also applied the Sherman-Morrison-Woodbury low rank update formula, which tells us that

$$(V_r\Lambda_r V_r^T + I)^{-1} = I - V_r D_r V_r^T,$$

  where $D_r = \text{diag}(\frac{\lambda_i}{\lambda_i+1})$ for $\Lambda_r = \text{diag}(\lambda_i)$.

- The low-rank approximation to the prior-preconditioned misfit Hessian can be found using only matvecs of $A$ (without constructing the full Jacobian) by a randomized SVD method.

## 4 Karhunen-Loève Expansions

- A mean-square continuous stochastic process $X_t$ with mean 0 has a basis $\{e_i\}$ such that

$$X_t(\omega) = \sum_{i=1}^{\infty}\sqrt{\lambda_i}\xi_i(\omega)e_i(t),$$

  where the $\xi_i$ are mean 0, variance 1, mutually uncorrelated random variables given by

$$\xi_i(\omega) = \frac{1}{\sqrt{\lambda_i}}\int_D X_t(\omega)e_i(t)dt.$$

- The basis also represents an eigenfunction decomposition of the covariance function,

$$C(s,t) = \sum_{i=1}^{\infty}\lambda_i e_i(s)e_i(t).$$

- To simulate realizations of $X_t$ we pick values of the $\xi_i$ and compute a truncated KL expansion. Approximations with more modes will be less smooth.

- The KL expansion for a Gaussian process with covariance $C(|s-t|) = \exp(-|s-t|)$ has iid normal coefficients $\xi(\omega) \sim N(0,1)$ and sines and cosines as eigenfunctions.

# 5 Matérn Kernels and Connection to PDEs

- For $x, y \in \mathbb{R}^d$, the Matérn covariance function is

$$k(x,y) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left( \frac{||x-y||}{\rho} \right)^{\nu} K_\nu \left( \frac{||x-y||}{\rho} \right),$$

  where $\sigma^2$ is the marginal variance, $\rho > 0$ is a characteristic length, and $K_\nu$ is the modified Bessel function of the second kind of order $\nu > 0$.

- For some common values of $\nu$, the covariance function reduces to

$$\begin{aligned}
\nu &= \frac{1}{2} \qquad k_{1/2}(r) \propto \exp\left(-\frac{r}{\ell}\right) \\
\nu &= \frac{3}{2} \qquad k_{3/2}(r) = \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right) \\
\nu &= \frac{5}{2} \qquad k_{5/2}(r) = \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}r}{\ell}\right),
\end{aligned}$$

  where $r = ||x-y||$ and $\ell = \sqrt{2\nu}\rho$.

- GRFs with the Matérn kernel are smoother for higher $\nu$, and as $\nu \to \infty$, the kernel approaches the Radial Basis Function (RBF) kernel, whose covariance function is an isotropic Gaussian.

- How do we simulate instances of a GRF with Matérn covariance? Could use KL expansion, but computing (and storing) the eigenfunctions may be too expensive.

- Instead, use the fact that a GRF $u$ in $\mathbb{R}^d$ with the Matérn covariance is a solution to the linear fractional SPDE

$$(\kappa^2 - \Delta)^{\alpha/2} u(\omega) = \mathcal{W}(\omega), \qquad \alpha = \nu + d/2,$$

  where $\mathcal{W}(\omega)$ is spatial Gaussian white noise with unit variance and $\kappa = \frac{1}{\rho}$. The marginal variance $\sigma^2$ of this solution can be computed as a function of $\rho, \nu$, and $d$. The Matérn fields are the only stationary solutions of this SPDE.

- The fractional Laplacian operator is defined using its Fourier transform,

$$(\mathcal{F}[(\kappa^2 - \Delta)^{\alpha/2}\phi])(\vec{k}) = (\kappa^2 + ||\vec{k}||^2)^{\alpha/2} (\mathcal{F}\phi)(\vec{k}).$$

- An instance of the GRF is given by $u = (\kappa^2 - \Delta)^{\alpha/2}\mathcal{W}(\omega)$ for a particular instance of white noise. This is analogous to sampling a distribution with covariance $C = R^T R$ by evaluating $R\xi$ with $\xi \sim N(0, I)$, so the covariance operator (whose discretization is the covariance matrix) is given by $(\kappa^2 - \Delta)^{-\alpha}$.

- In 1D, $\nu = 3/2 \implies (\kappa^2 I - \Delta)u = \mathcal{W}$, with covariance $(\kappa^2 I - \Delta)^{-2}$. Simulating $u$ requires one elliptic solve. Choosing $\nu = 1/2 \implies (\kappa^2 I - \Delta)^{1/2} u = \mathcal{W}$ with covariance $(\kappa^2 I - \Delta)^{-1}$ gives rougher GRFs.

- In 2D, $\nu = 1 \implies (\kappa^2 I - \Delta)u = \mathcal{W}$. We cannot, however, choose $\nu = 0 \implies (\kappa^2 - \Delta)^{1/2}$, because in 2D the covariance $(\kappa^2 - \Delta)^{-1}$ is not trace class (eigenvalues sum to $\infty$), so it is not a valid covariance.

- The SPDE interpretation of Matérn kernels allows us to avoid KL expansions and take advantage of PDE theory. For example, we can evaluate GRFs on irregular grids or triangulations, taking advantage of finite element methods.
- The SPDE interpretation also provides a simple way to implement anisotropic covariances, by replacing the Laplacian $\Delta$ with $\nabla \cdot (\theta(x)\nabla)$, for some anisotropic diffusion matrix $\theta(x)$.