Bayesian Inference for Latent Gaussian Models Governed by PDEs

Sonia Reilly, Georg Stadler (Courant Institute, New York University)

Goal

Characterize the posterior of a linear PDE-governed Bayesian inverse problem with Gaussian prior, marginalizing out hyperparameters.

Latent Gaussian Models (LGMs)

Hierarchical model with Gaussian prior:

$$\begin{aligned} \boldsymbol{\theta} &\sim \pi_{\text{hyp}}(\boldsymbol{\theta}) && \text{low-dim hyperparameter} \\ \boldsymbol{m}|\boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{\mu}_{\text{pr}}(\boldsymbol{\theta}), \boldsymbol{Q}_{\text{pr}}^{-1}(\boldsymbol{\theta})) && \text{high-dim latent variable} \\ \boldsymbol{y}|\boldsymbol{m},\boldsymbol{\theta} &\sim \pi_{\text{like}}(\boldsymbol{y}|\boldsymbol{m},\boldsymbol{\theta}) && \text{observed data} \end{aligned}$$

Linear Gaussian Bayesian inverse problem as LGM:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{m} + \boldsymbol{\varepsilon} \quad \text{with} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{Q}_{\varepsilon}(\boldsymbol{\theta})),$$

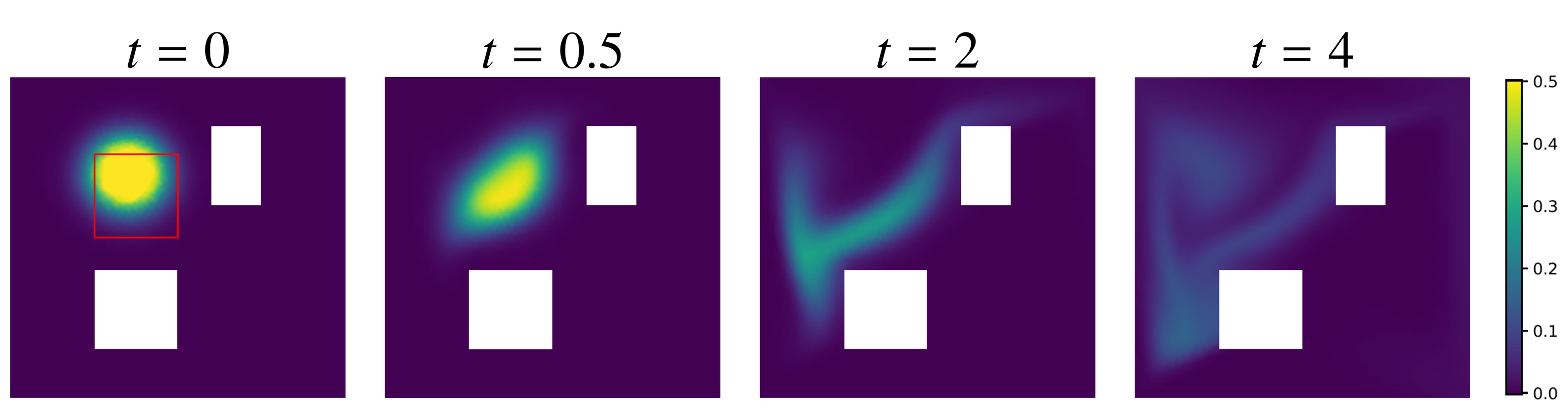
with $\boldsymbol{A}$ a discretization of a linear PDE. Want to characterize $\pi(\boldsymbol{m}|\boldsymbol{y})$.

Example: Initial Condition Inference

Infer distribution of initial condition of advection diffusion equation

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

from solution evaluated at points x_i and times $t_i > 0$.



$u(x,t)$ for various times t.

- Data: $y_i = u(x_i, t_i) + \varepsilon_i$
- Latent var.: initial condition $\boldsymbol{m}$
- Forward map: discretized PDE $\boldsymbol{A}$
- Prior: Matérn GRF with precision $(\delta I - \gamma\Delta)^2$
- Likelihood: $\mathcal{N}(\boldsymbol{A}\boldsymbol{m}, \lambda\boldsymbol{I})$
- Hyperparam. $\boldsymbol{\theta}$: $(\delta, \gamma, \lambda)$
- Hyperprior: independent gamma distributions
- QoI: q = avg. of $m(x)$ in red box

Sampling & Integration by Marginalization

Sampling: sample $\boldsymbol{m}^* \sim \pi(\boldsymbol{m}|\boldsymbol{y})$ by

1. sampling $\boldsymbol{\theta}^* \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$ (low-dimensional, so can use MCMC)
2. sampling $\boldsymbol{m}^* \sim \pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{\theta}^*, \boldsymbol{y})$ (Gaussian, posterior of a linear IP)

Integration: (low-dim quadrature)

$$\int f(\boldsymbol{m})\pi(\boldsymbol{m}|\boldsymbol{y})\,d\boldsymbol{m} = \int\left(\int f(\boldsymbol{m})\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{\theta},\boldsymbol{y})\,d\boldsymbol{m}\right)\pi(\boldsymbol{\theta}|\boldsymbol{y})\,d\boldsymbol{\theta}$$

Find max of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and quadrature points around max.

Fast Computation of Marginal: Low Rank Approx

Both require many evaluations of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ with different $\boldsymbol{\theta}$:

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{y}) &\propto \frac{\pi(\boldsymbol{m},\boldsymbol{\theta},\boldsymbol{y})}{\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{\theta},\boldsymbol{y})} = \frac{\pi_{\text{like}}(\boldsymbol{y}|\boldsymbol{m},\boldsymbol{\theta})\pi_{\text{pr}}(\boldsymbol{m}|\boldsymbol{\theta})\pi_{\text{hyp}}(\boldsymbol{\theta})}{\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{\theta},\boldsymbol{y})} \\ &\propto \left(\frac{|\boldsymbol{Q}_{\text{pr}}||\boldsymbol{Q}_{\varepsilon}|}{|\boldsymbol{Q}_{\text{post}}|}\right)^{1/2}\exp\left(-\frac{1}{2}\left[\|\boldsymbol{y}\|_{\boldsymbol{Q}_{\varepsilon}} + \|\boldsymbol{\mu}_{\text{pr}}\|_{\boldsymbol{Q}_{\text{pr}}} - \|\boldsymbol{\mu}_{\text{post}}\|_{\boldsymbol{Q}_{\text{post}}}\right]\right)\pi_{\text{hyp}}(\boldsymbol{\theta}) \end{aligned}$$

Each new $\boldsymbol{\theta}$ requires solving a linear BIP. Cost is dominated by forward and adjoint PDE solve in the prior-to-posterior precision update:

$$\boldsymbol{Q}_{\text{post}} = \boldsymbol{Q}_{\text{pr}} + \underbrace{\boldsymbol{A}^T\boldsymbol{Q}_{\varepsilon}\boldsymbol{A}}_{\text{update}}$$

Instead, precompute rank-r approximation to update. Requires $2r$ PDE solves up front, but avoids most PDE solves for each $\boldsymbol{\theta}$ (a few are needed in practice to improve accuracy of $\boldsymbol{\mu}_{\text{post}}$).

✗ Prior Preconditioned Approximation

Standard approach: use randomized eigensolver to find

$$\boldsymbol{Q}_{\text{pr}}^{-1/2}(\boldsymbol{\theta})\boldsymbol{A}^T\boldsymbol{Q}_{\varepsilon}\boldsymbol{A}\boldsymbol{Q}_{\text{pr}}^{-1/2}(\boldsymbol{\theta}) \approx \boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^T,$$

since $\boldsymbol{Q}_{\text{pr}}$ is smoothing, it lowers required r and we have a natural bound on truncation error. *But:* $\boldsymbol{Q}_{\text{pr}}$ depends on $\boldsymbol{\theta}$!

✓ Weakest Prior Preconditioned Approximation

Instead,

$$\boldsymbol{Q}_{\text{pr}}^{-1/2}(\boldsymbol{\theta}_0)\boldsymbol{A}^T\boldsymbol{Q}_{\varepsilon}\boldsymbol{A}\boldsymbol{Q}_{\text{pr}}^{-1/2}(\boldsymbol{\theta}_0) \approx \boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^T$$

for some fixed $\boldsymbol{\theta}_0$ and convert for each $\boldsymbol{\theta}$:

$$\boldsymbol{Q}_{\text{pr}}^{-1/2}(\boldsymbol{\theta})\boldsymbol{A}^T\boldsymbol{Q}_{\varepsilon}\boldsymbol{A}\boldsymbol{Q}_{\text{pr}}^{-1/2}(\boldsymbol{\theta}) \approx \boldsymbol{Q}_{\text{pr}}^{-1/2}(\boldsymbol{\theta})\boldsymbol{Q}_{\text{pr}}^{1/2}(\boldsymbol{\theta}_0)\boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^T\boldsymbol{Q}_{\text{pr}}^{1/2}(\boldsymbol{\theta}_0)\boldsymbol{Q}_{\text{pr}}^{-1/2}(\boldsymbol{\theta}).$$

Choose $\boldsymbol{\theta}_0$ to be the parameters of the least smoothing, "weakest" prior to avoid amplifying truncation error.

✓ Unpreconditioned Approximation

Alternatively,

$$\boldsymbol{A}^T\boldsymbol{Q}_{\varepsilon}\boldsymbol{A} \approx \boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^T.$$

Higher rank needed, and no principled way to choose rank, but simplest.

Computational Complexity

Main computational cost in evaluating $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ comes from forward/adjoint PDE solves and prior precision solves. The table below reports the complexity of N evaluations in terms of these operations.

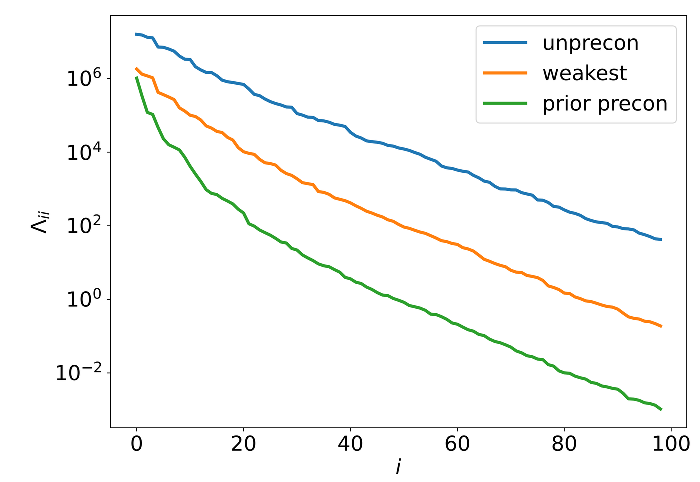
	prior preconditioned	weakest	unpreconditioned
$\boldsymbol{A}, \boldsymbol{A}^T$ applies	$2Nr_p + 2N\ell$	$2r_w + 2N\ell$	$2r_u + 2N\ell$
$\boldsymbol{Q}_{\text{prior}}$ solves	$N(r_p + \ell)$	$r_w + N(r_w + \ell)$	$N(r_u + \ell)$
rank	$r_p = 57$	$r_w = 66$	$r_u = 84$
time for $N = 100$	1167 s	167 s	182 s

Number of most expensive operations (first two rows) for the low-rank approximation methods above (columns); numerical rank used in each method (3rd row); and time for $N = 100$ evaluations in the initial-condition inference example (4th row); ℓ = num. of CG iterations to find $\boldsymbol{\mu}_{\text{post}}$.

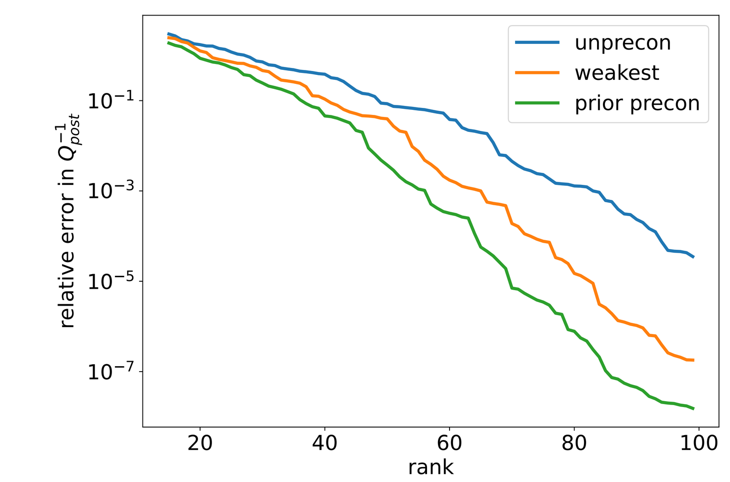
Numerical Results: Initial Condition Inference

Comparison of low rank approximations

- Fastest decay when preconditioned with $\boldsymbol{Q}_{\text{pr}}(\boldsymbol{\theta})$ (as expected)
- Preconditioning with weakest $\boldsymbol{Q}_{\text{pr}}(\boldsymbol{\theta}_0)$ improves over no precond
- Error plot allows for comparison of unpreconditioned method
- Using weakest $\boldsymbol{Q}_{\text{pr}}(\boldsymbol{\theta}_0)$ may be a good tradeoff between up-front and online computation



Eigenvalue spectrum of update, by type of preconditioning.



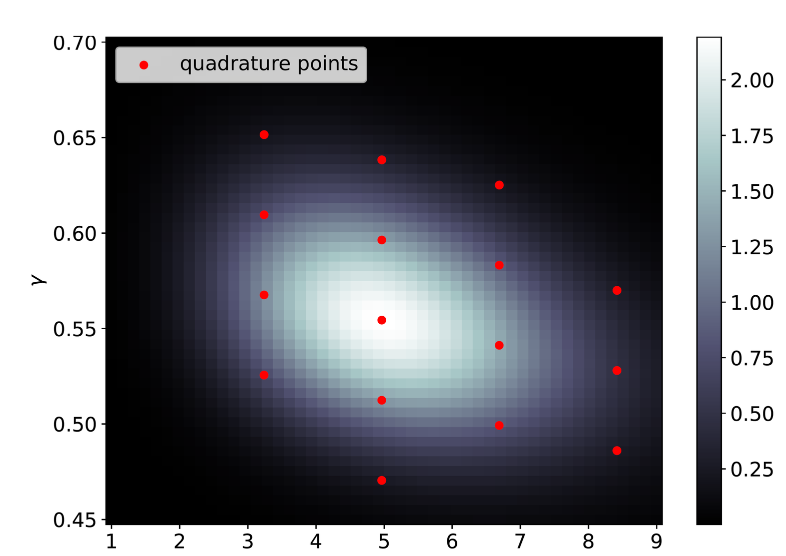
Error in norm of pointwise posterior variance as a function of rank.

Hyperparameter marginal $\pi(\boldsymbol{\theta}|\boldsymbol{y})$

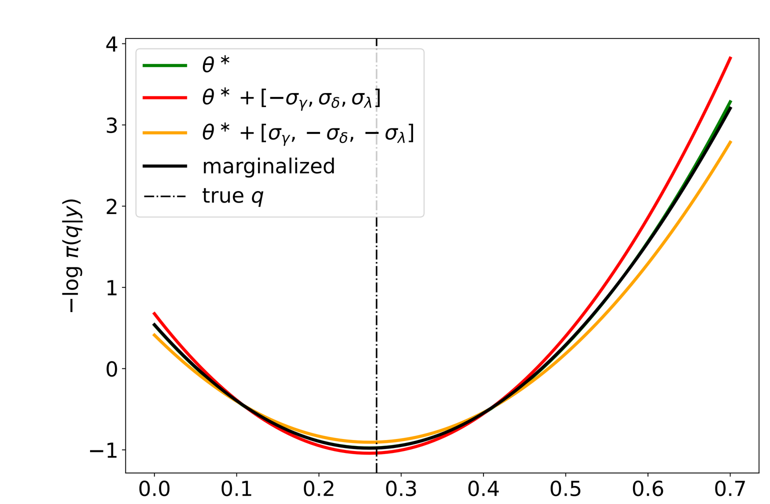
- Argmax γ^*, δ^* approximate smoothness and variance of true IC
- Unimodal, easy to sample and to integrate using quadrature

Quantity of Interest distribution

- With quadrature, can approximate $\pi(q|\boldsymbol{y})$ for any quantity of interest q that is a linear function of $\boldsymbol{m}$ (here, avg. of $m(x)$ in a region)
- In this case $\pi(q|\boldsymbol{y})$ is close to $\pi(q|\boldsymbol{y}, \boldsymbol{\theta}^*)$, but other choices of $\boldsymbol{\theta}$ change the variance significantly.



$\pi(\boldsymbol{\theta}|\boldsymbol{y})$ (only γ, δ shown) in gray scale, with tailored quadrature points in red.



$\pi(q|\boldsymbol{y})$ compared to $\pi(q|\boldsymbol{y}, \boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}^* = \arg\max \pi(\boldsymbol{\theta}|\boldsymbol{y})$ and other quadrature points.

Next Steps

- Use various discretizations of PDE to speed up MCMC
- Find a way to adapt low rank approximation when $\boldsymbol{A}$ depends on $\boldsymbol{\theta}$
- Extend to nonlinear PDEs using ideas from Integrated Nested Laplace Approximation

Acknowledgements

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of ASCR, DOE Computational Science Graduate Fellowship under Award Number DE-SC0022158.