Amortized Low-Rank Approximation for Hyperparameter Marginalization in PDE-Governed Bayesian Inverse Problems

Sonia Reilly

Georg Stadler

Courant Institute, New York University

SIAM Annual Meeting, July 6, 2026

Hierarchical PDE-Governed Bayesian Inverse Problems

- Find parameter function m from measurements y of PDE solution:

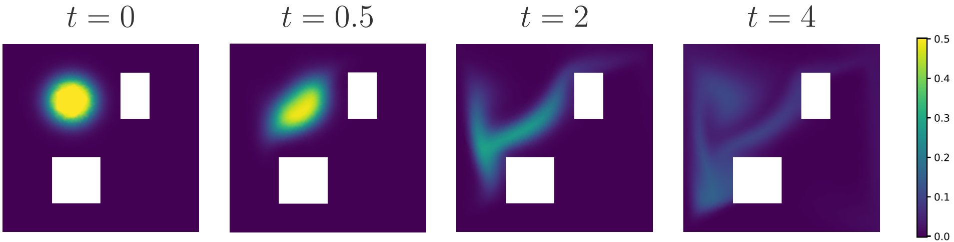
$$y = Am + \varepsilon, \quad A \text{ a linear PDE}$$

- Low-dimensional hyperparameter θ generalizes Gaussian prior and likelihood:

$$m \sim \mathcal{N}(\mu_{\text{pr}}(\theta), Q_{\text{pr}}^{-1}(\theta)), \quad \varepsilon \sim \mathcal{N}(0, Q_{\varepsilon}^{-1}(\theta))$$

- Goal: marginalize out hyperparameter to get $\pi(m|y)$

Example: Initial Condition Inference in Advection-Diffusion



Find initial concentration from measurements at later times.

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

$$
\begin{gathered}
m \sim \mathcal{N}(0, \delta^{-2}(I - \gamma\Delta)^{-2}) \text{ (Matérn)} \\
\varepsilon \sim \mathcal{N}(0, \sigma^2 I) \\
\theta = [\gamma, \delta, \sigma],\ \pi(\theta) \text{ uniform}
\end{gathered}
$$

Marginalization

$$\pi(m|y) = \int \underbrace{\pi(m|\theta, y)}_{\text{Gaussian}} \underbrace{\pi(\theta|y)}_{??} \, d\theta$$

- Empirical Bayes (EB): $\pi(m|y) \approx \pi(m|\theta^*, y)$, where $\theta^* = \arg\max \pi(\theta|y)$
  - Fastest approximation, Gaussian
- Quadrature around θ^*
  - Slightly slower (for low dim θ), mixture of Gaussians
- MCMC sampling
  - Typically most expensive, mixture of Gaussians

All require many $\pi(\theta|y)$ evaluations!

Marginalization

$$\pi(m|y) = \int \underbrace{\pi(m|\theta, y)}_{\text{Gaussian}} \underbrace{\pi(\theta|y)}_{??} \, d\theta$$

- Empirical Bayes (EB): $\pi(m|y) \approx \pi(m|\theta^*, y)$, where $\theta^* = \arg\max \pi(\theta|y)$
  - Fastest approximation, Gaussian
- **Quadrature around $\boldsymbol{\theta^*}$**
  - Slightly slower (for low dim θ), mixture of Gaussians
- MCMC sampling
  - Typically most expensive, mixture of Gaussians

All require many $\pi(\theta|y)$ evaluations!

Hyperparameter Marginal

$$\pi(m,\theta,y) = \pi(m|\theta,y)\pi(\theta|y)\pi(y) \implies \pi(\theta|y) \propto \frac{\pi(m,\theta,y)}{\pi(m|\theta,y)} = \frac{\pi(y|m,\theta)\pi(m|\theta)\pi(\theta)}{\pi(m|\theta,y)}$$

plug in Gaussian densities and cancel m:

$$\boxed{\pi(\theta|y) \propto \left(\frac{|Q_{\text{pr}}||Q_\varepsilon|}{|Q_{\text{post}}|}\right)^{1/2} \exp\left(-\frac{1}{2}\left[\|y\|^2_{Q_\varepsilon} + \|\mu_{\text{pr}}\|^2_{Q_{\text{pr}}} - \|\mu_{\text{post}}\|^2_{Q_{\text{post}}}\right]\right)\pi(\theta)}$$

$$\pi(m|\theta,y) = \mathcal{N}(\mu_{\text{post}}, Q_{\text{post}}^{-1})$$

$$\begin{aligned} Q_{\text{post}} &= Q_{\text{pr}} + A^\top Q_\varepsilon A \\ \mu_{\text{post}} &= Q_{\text{post}}^{-1}(Q_{\text{pr}}\mu_{\text{pr}} + A^\top Q_\varepsilon y) \end{aligned}$$

Computational Challenges

$$\pi(\theta|y) \propto \left(\frac{|Q_{\text{pr}}||Q_\varepsilon|}{|Q_{\text{post}}|}\right)^{1/2} \exp\left(-\frac{1}{2}\left[\|y\|^2_{Q_\varepsilon} + \|\mu_{\text{pr}}\|^2_{Q_{\text{pr}}} - \|\mu_{\text{post}}\|^2_{Q_{\text{post}}}\right]\right)\pi(\theta)$$

- No access to matrices $A, Q_{\text{pr}}, Q_{\text{post}}$! Can only apply to vectors.
- $A, A^\top$ very expensive to apply, e.g., 3D time-dependent PDE solve
- $\mu_{\text{post}} = Q^{-1}_{\text{post}}(Q_{\text{pr}}\mu_{\text{pr}} + A^\top Q_\varepsilon y)$ requires iterative applications of $A, A^\top$
- $|Q_{\text{pr}}|, |Q_{\text{post}}| \to \infty$ as discretization of m is refined

Standard Approach[2]: Low Rank Approximation

Data usually informs posterior in a low-rank subspace

$$Q_{\text{post}} = Q_{\text{pr}} + \underbrace{A^\top Q_\varepsilon A}_{\text{precision update}}$$

Prior-preconditioned (PP) update is even lower rank (in fact, optimal!)[1]

$$Q_{\text{pr}}^{-1/2} A^\top Q_\varepsilon A Q_{\text{pr}}^{-1/2} \approx V_r \Lambda_r V_r^\top$$

Final approximation:

$$Q_{\text{post}} \approx Q_{\text{post},r} := Q_{\text{pr}}^{1/2}(I + V_r \Lambda_r V_r^\top) Q_{\text{pr}}^{1/2}$$

[1] Spantini et al., 2015
[2] Bui-Thanh et al., 2013

Standard Approach: Pros and Cons

Posterior covariance and determinant ratio can be found explicitly:

$$Q_{\text{post},r}^{-1} = Q_{\text{pr}}^{-1/2}(I - V_r D_r V_r^{\top})Q_{\text{pr}}^{-1/2}, \quad D_r = \text{diag}\left(\frac{\lambda_i}{\lambda_i + 1}\right)$$

$$-\log \frac{|Q_{\text{pr}}|}{|Q_{\text{post},r}|} = \sum_{i=1}^{r} \log(1 + \lambda_i)$$

If θ in $Q_{\text{pr}}^{-1/2} A^{\top} Q_{\varepsilon} A Q_{\text{pr}}^{-1/2}$ is only multiplicative, approx can be precomputed.

But for non-multiplicative hyperparameters, e.g., $Q_{\text{pr}} = \delta^2(I - \gamma\Delta)^2$, approx needs to be recomputed for each θ.

Assumption: non-multiplicative only in Q_{pr}, not Q_{ε}. Multiplicative are 1, WLOG.

Weakest Preconditioning (WP)

Idea: offline, approximate update preconditioned with a different reference prior $\hat{Q}_{\text{pr}}$:

$$\hat{Q}_{\text{pr}}^{-1/2} A^\top Q_\varepsilon A \hat{Q}_{\text{pr}}^{-1/2} \approx \hat{V}_r \hat{\Lambda}_r \hat{V}_r^\top$$

Online, undo reference prior and add back current prior:

$$\begin{aligned} Q_{\text{post}} \approx \hat{Q}_{\text{post},r} &:= Q_{\text{pr}}^{1/2}(I + Q_{\text{pr}}^{-1/2} \hat{Q}_{\text{pr}}^{1/2} \hat{V}_r \hat{\Lambda}_r \hat{V}_r^\top \hat{Q}_{\text{pr}}^{1/2} Q_{\text{pr}}^{-1/2}) Q_{\text{pr}}^{1/2} \\ &= Q_{\text{pr}}^{1/2}(I + \bar{V}_r \bar{\Lambda}_r \bar{V}_r^\top) Q_{\text{pr}}^{1/2} \end{aligned}$$

"Weakest" because stronger $\hat{Q}_{\text{pr}}$ can over-precondition and underestimate rank

Truncation Error

Approximation introduces error in two terms:

$$-\log \pi(\theta|y) - (-\log \pi_{\text{approx}}(\theta|y)) = \underbrace{\frac{1}{2}\log\frac{|Q_{\text{post}}|}{|Q_{\text{post},r}|}}_{e_1} + \underbrace{\frac{1}{2}\|Q_{\text{pr}}\mu_{\text{pr}} + A^\top Q_\varepsilon y\|^2_{(Q^{-1}_{\text{post},r} - Q^{-1}_{\text{post}})}}_{e_2 = 0}$$

- e_1 can be bounded!
- e_2 is trickier, and often bypassed by solving $\mu_{\text{post}} = Q^{-1}_{\text{post}}(Q_{\text{pr}}\mu_{\text{pr}} + A^\top Q_\varepsilon y)$ exactly with CG preconditioned with $Q_{\text{post},r}$

Theoretical Results

Lemma (PP Error Bound)

$$e_1(r,\theta) \leq \frac{1}{2}\sum_{i=r+1}^{n} \lambda_i(\theta) \implies \textit{choose } r \textit{ s.t. } \lambda_{r+1}(\theta) \ll 1$$

Theorem (WP Error Bound)

If $\hat{Q}_{pr} \preceq Q_{pr}(\theta)$ *for all* θ, *then*

$$e_1(\hat{r},\theta) \leq \frac{1}{2}\left(\hat{r}\hat{\lambda}_{\hat{r}+1} + \sum_{\hat{r}+1}^{n} \hat{\lambda}_i\right) \implies \textit{choose } \hat{r} \textit{ s.t. } \hat{\lambda}_{\hat{r}+1} \ll 1$$

WP exchanges θ-dependent bound for higher, θ-independent bound.

Unpreconditioned Method (UP)

What if there's no reference prior that satisfies $\hat{Q}_{\text{pr}} \preceq Q_{\text{pr}}(\theta)$?
E.g. gamma hyperpriors, instead of uniform.

Alternative:

$$A^\top Q_\varepsilon A \approx \tilde{V}_r \tilde{\Lambda}_r \tilde{V}_r^\top$$

$$\begin{aligned} Q_{\text{post}} \approx \hat{Q}_{\text{post},r} &:= Q_{\text{pr}}^{1/2}(I + Q_{\text{pr}}^{-1/2}\tilde{V}_r \tilde{\Lambda}_r \tilde{V}_r^\top Q_{\text{pr}}^{-1/2})Q_{\text{pr}}^{1/2} \\ &= Q_{\text{pr}}^{1/2}(I + \bar{V}_r \bar{\Lambda}_r \bar{V}_r^\top)Q_{\text{pr}}^{1/2} \end{aligned}$$

WP error bound applies, with $\hat{Q}_{\text{pr}} = cI$ for some c (defined in $\dim(m) \to \infty$ limit).

No preconditioning $\implies$ slower eigenvalue decay $\implies$ larger error bound

Algorithm Summary

PP

For each θ:

- low rank approx
- evaluate det. ratio using λ_i
- CG iteration for μ_{post}

WP

Offline

- low rank approx

For each θ:

- replace preconditioning
- evaluate det. ratio using $\bar{\lambda}_i$
- CG iteration for μ_{post}

UP

Offline

- low rank approx

For each θ:

- add preconditioning
- evaluate det. ratio using $\bar{\lambda}_i$
- CG iteration for μ_{post}

Cost Comparison of N Evaluations of $\pi(\theta|y)$

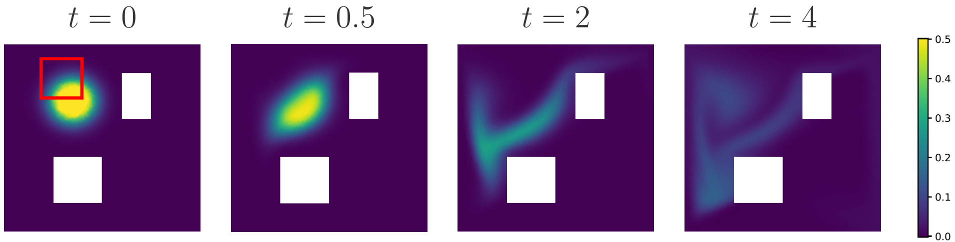
$A, A^\top$ applies dominate cost, followed by Q_{pr} applies/solves

- Low-rank approximation: $2r$ $A, A^\top$ applies, using randomized methods
- Each of ℓ CG iterations: 2 $A, A^\top$ applies

	PP	WP	UP
$\boldsymbol{A}, \boldsymbol{A}^\top$ applies	$2N\ell + 2Nr$	$2N\ell + 2\hat{r}$	$2N\ell + 2\bar{r}$
$\boldsymbol{Q}_{\text{pr}}$ solves	$N(r+\ell)$	$N(\hat{r}+\ell)+\hat{r}$	$N(\bar{r}+\ell)$
$\boldsymbol{Q}_{\text{pr}}$ applies	$N(3r+\ell)$	$N(4\hat{r}+\ell)+3\hat{r}$	$N(3\bar{r}+\ell)$

WP and UP avoid $O(Nr)$ PDE solves at the cost of additional cheaper prior computations

Back to Advection-Diffusion Example

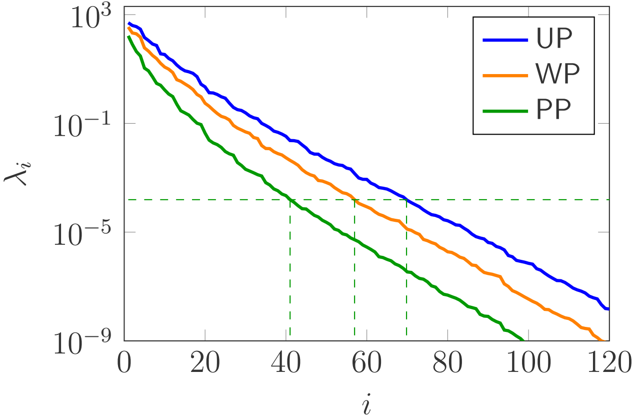


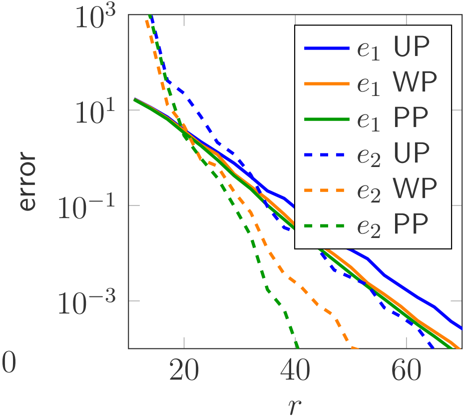
$$m \sim \mathcal{N}(0, \delta^{-2}(I - \gamma\Delta)^{-2}) \text{ (Matérn)}$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$
$$\theta = [\gamma, \delta, \sigma],\ \pi(\theta) \text{ uniform}$$

QoI: avg. init. concentration in red box, q

Find $\pi(q|y)$ from measurements along building edges at later times.

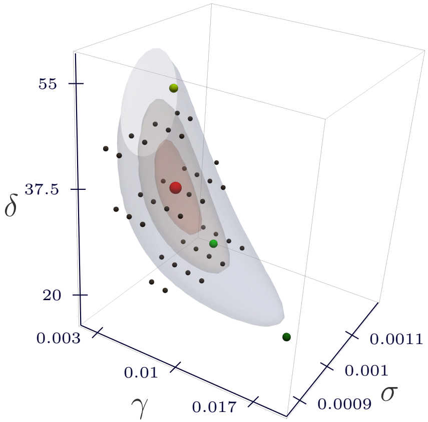
Rank and Error: PP vs WP vs UP

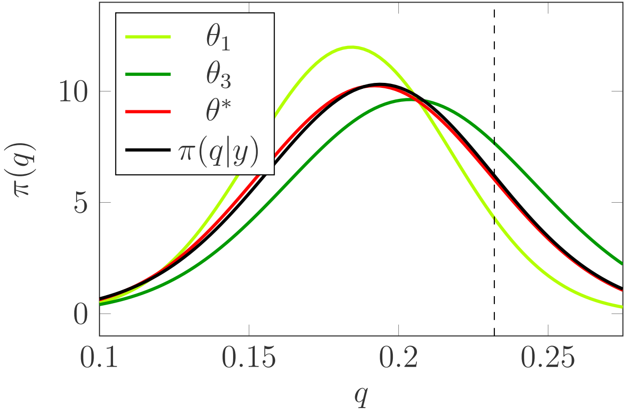




Stronger preconditioning $\implies$ lower rank, lower error

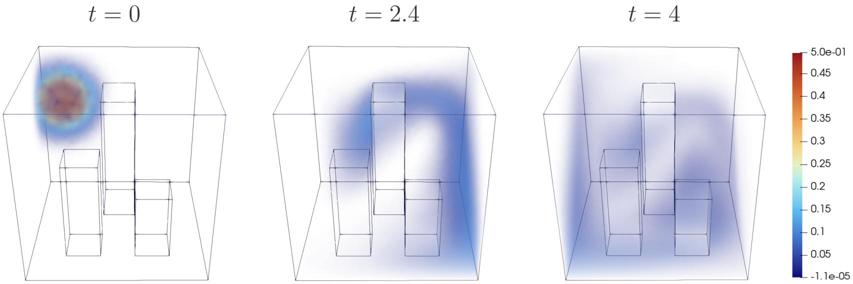
Marginalizing the QoI





Modest but visible improvement over Empirical Bayes

Timing Comparison, 100 Evaluations of $\pi(\theta|y)$



	PP	WP	UP
time, 2D	870 s	79 s	86 s
time, 3D	65.6 hrs	2.02 hrs	3.00 hrs

WP achieves speedup of $\sim 30\times$ on 3D problem!

Conclusions

- Marginalization is expensive, especially with nonlinear prior hyperparameters
- Low rank approximation can be adapted with a different preconditioner (WP) to amortize cost across $\pi(\theta|y)$ evaluations
- Future work: nonlinear or θ-dependent PDE A

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy, Computational Science Graduate Fellowship under Award Number DE-SC0022158. Registration and travel support for this presentation was provided by the Society for Industrial and Applied Mathematics.