Efficient sample average approximation techniques for hyperparameter estimation in Bayesian inverse problems

Julianne Chung, Malena Sabaté Landman, Scot M. Miller, Arvind K. Saibaba

Abstract

Inverse problems arise in many important applications, where the aim is to estimate some unknown inverse parameters from given observations. For large-scale problems where the number of unknowns can be large (e.g., due to the desire to reconstruct high-resolution images or dynamic image reconstructions) or for problems where observational datasets are huge, estimating the inverse parameters can be a computationally challenging task. Although there have been significant advancements in solving inverse problems, many of these approaches rely on a pre-determined, carefully-tuned set of hyperparameters (e.g., that define the noise and prior models) that must be estimated from the data. The need to estimate these hyperparameters further exacerbates the problem, often requiring repeated solves for many combinations of hyperparameters. In this work, we propose a sample average approximation (SAA) method that couples a Monte Carlo estimator with a preconditioned Lanczos method for the efficient estimation of hyperparameters in Bayesian inverse problems.

We are interested in linear inverse problems that involve recovering the parameters $\mathbf{s} \in \mathbb{R}^n$ from measurements $\mathbf{d} \in \mathbb{R}^m$, which have been corrupted by additive Gaussian measurement noise, $\boldsymbol{\eta} \in \mathbb{R}^m$, and takes the form

$$\mathbf{d} = \mathbf{A}\mathbf{s} + \boldsymbol{\eta}, \qquad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}(\boldsymbol{\theta}))$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ represents the forward map and $\boldsymbol{\theta} \in \mathbb{R}^K_+$, represents the (nonnegative) hyperparameters. In the hierarchical Bayes approach, we treat $\boldsymbol{\theta}$ as a random variable, which we endow with prior density $\pi_{\text{hyp}}(\boldsymbol{\theta})$. We assume that the noise covariance matrix $\mathbf{R} : \mathbb{R}^K_+ \to \mathbb{R}^{m\times m}$, where $\mathbf{R}(\cdot)$ is symmetric and positive definite (SPD), and has an inverse and square root that is computationally easy to obtain for any input (e.g., a diagonal matrix or a scalar times the identity). We assume that the prior distribution for the parameters $\mathbf{s}$ is also Gaussian of the form $\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{Q}(\boldsymbol{\theta}))$, where $\boldsymbol{\mu} : \mathbb{R}^K_+ \to \mathbb{R}^n$ and $\mathbf{Q} : \mathbb{R}^K_+ \to \mathbb{R}^{n\times n}$, where $\mathbf{Q}(\cdot)$ is assumed to be SPD.

With the above assumptions, we obtain the marginal posterior density,

$$\pi(\boldsymbol{\theta} \,|\, \mathbf{d}) \propto \pi_{\text{hyp}}(\boldsymbol{\theta}) \det(\boldsymbol{\Psi}(\boldsymbol{\theta}))^{-1/2} \exp\left(-\frac{1}{2}\|\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \mathbf{d}\|^2_{\boldsymbol{\Psi}^{-1}(\boldsymbol{\theta})}\right), \tag{1}$$

where $\boldsymbol{\Psi}(\boldsymbol{\theta}) = \mathbf{A}\mathbf{Q}(\boldsymbol{\theta})\mathbf{A}^\top + \mathbf{R}(\boldsymbol{\theta})$. One goal would be to draw samples (e.g., using Markov Chain Monte Carlo) from (1), and using the samples to quantify the uncertainty in the hyperparameters. However, this may be prohibitive for large-scale problems because evaluating the density function (or its logarithm) requires evaluating the determinant of and multiple solves with the matrix $\boldsymbol{\Psi}$ that depends on $\boldsymbol{\theta}$, which can be expensive. To compound matters, hundreds of samples are required to get accurate statistics, which can involve several hundred thousand density function evaluations.

Instead, we follow an empirical Bayes approach and focus on computing the maximum a posteriori (MAP) estimate, that is, the point estimate that maximizes the marginal posterior distribution or, equivalently, minimizes the negative log of the marginal posterior. That is, the problem of hyperparameter estimation becomes solving an optimization problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^K_+} \mathcal{F}(\boldsymbol{\theta}) \equiv -\log \pi_{\text{hyp}}(\boldsymbol{\theta}) + \frac{1}{2}\text{logdet}(\boldsymbol{\Psi}(\boldsymbol{\theta})) + \frac{1}{2}\|\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \mathbf{d}\|^2_{\boldsymbol{\Psi}(\boldsymbol{\theta})^{-1}}. \tag{2}$$

Notice that solving (2) is a computationally intensive task since it involves computing log determinants. To address this challenge, we consider an SAA method for computing the MAP estimate of the marginalized posterior distribution that combines a stochastic average approximation of the objective function and the preconditioned Lanczos method to compute efficient approximations of the function and gradient evaluations. The novel contributions of this work are as follows.

1. The method to estimate the objective function combines a Monte Carlo estimator for the log-determinant of the matrix with a preconditioned Lanczos approach to apply the matrix logarithm. We analyze the impact of the number of Monte Carlo samples and Lanczos iterations on the accuracy of the log-determinant estimator.

2. We use a novel preconditioner to accelerate the Lanczos iterations. The preconditioner is based on a parametric low-rank approximation of the prior covariance matrix, that is easy to update for new values of the hyperparameters. In particular, no access to the forward/adjoint solver is needed to update the preconditioner, and only a modest amount of precomputation is needed as a setup cost (independent of the optimization).

3. We also use a trace estimator to approximate the gradient that has two features: first, it works with a symmetric form of the argument inside the trace, and second, it is able to reuse Lanczos iterates from the objective function computations. Therefore, the gradient can be computed essentially for free (i.e., requiring no additional forward/adjoint applications).

Related works. The methods we describe here have some similarity to existing literature and share certain techniques in common. The problem of optimizing for hyperparameters is closely related to parameter estimation in Gaussian processes on maximum likelihood (we may think of it as setting the forward operator as the identity matrix). The literature on this topic is vast, but we mention a few key references that are relevant to our approach. In [3], the authors propose a matrix-free approach to estimate the hyperparameters and also use an SAA for optimization. In [2], the authors propose a reformulation of the problem that avoids computing the inversion of the (prior) covariance matrix. Approaches based on hierarchical matrices are considered in [8, 10, 1]. Preconditioned Lanczos methods for estimating the log-determinant and its gradient are considered in [6, 7]. However, the main difference is that the Gaussian process methods do not involve forward operators. This raises two issues: first, we have to account for the problem structure which is different from Gaussian processes, and second, we have to account for the computational cost of the forward operator (and its adjoint), which may be comparable or greater than the cost of the covariance matrices.

On the inverse problem side, there have been relatively few works on computing the hyperparameters by optimization. Several works (e.g., [4]) instead use sampling methods (e.g., Markov Chain Monte Carlo), but these methods are extremely expensive since they require several thousand evaluations of the likelihood to achieve accurate uncertainty estimates. In [9], we developed efficient methods for hyperparameter estimation based on low-rank approximations using the generalized Golub-Kahan iterative method. A brief review of other techniques is also given in the same paper.

References

[1] S. Ambikasaran, A. K. Saibaba, E. F. Darve, and P. K. Kitanidis. Fast algorithms for Bayesian inversion. In Computational Challenges in the Geosciences, pages 101–142. Springer, 2013.

[2] M. Anitescu, J. Chen, and M. L. Stein. An inversion-free estimating equations approach for Gaussian process models. Journal of Computational and Graphical Statistics, 26(1):98–107, 2017.

[3] M. Anitescu, J. Chen, and L. Wang. A matrix-free approach for solving the parametric Gaussian process maximum likelihood problem. SIAM Journal on Scientific Computing, 34(1):A240–A262, 2012.

[4] J. M. Bardsley. Computational uncertainty quantification for inverse problems, volume 19 of Computational Science & Engineering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2018.

[5] E. Chow and Y. Saad. Preconditioned Krylov subspace methods for sampling multivariate Gaussian distributions. *SIAM Journal on Scientific Computing*, 36(2):A588–A608, 2014.

[6] K. Dong, D. Eriksson, H. Nickisch, D. Bindel, and A. G. Wilson. Scalable log determinants for Gaussian process kernel learning. Advances in Neural Information Processing Systems, 30, 2017.

[7] J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. Advances in neural information processing systems, 31, 2018.

[8] C. J. Geoga, M. Anitescu, and M. L. Stein. Scalable Gaussian process computations using hierarchical matrices. Journal of Computational and Graphical Statistics, 29(2):227–237, 2020.

[9] K. A. Hall-Hooper, A. K. Saibaba, J. Chung, and S. M. Miller. Efficient iterative methods for hyperparameter estimation in large-scale linear inverse problems. arXiv preprint arXiv:2311.15827, 2023.

[10] V. Minden, A. Damle, K. L. Ho, and L. Ying. Fast spatial Gaussian process maximum likelihood estimation via skeletonization factorizations. Multiscale Modeling & Simulation, 15(4):1584–1611, 2017.