Received: 5 June 2019 | Revised: 17 April 2020 | Accepted: 31 May 2020

DOI: 10.1002/nla.2325

RESEARCH ARTICLE

WILEY

Efficient Krylov subspace methods for uncertainty quantification in large Bayesian linear inverse problems

Arvind K. Saibaba[1] | **Julianne Chung**[2] | **Katrina Petroske**[1]

[1]Department of Mathematics, North Carolina State University, Raleigh, North Carolina, USA

[2]Department of Mathematics, Computational Modeling Data Analytics Division, Academy of Integrated Science, Virginia Tech, Blacksburg, Virginia, USA

Correspondence
Arvind K. Saibaba, Department of Mathematics, North Carolina State University, Raleigh, NC.
Email: asaibab@ncsu.edu

Funding information
Division of Mathematical Sciences, Grant/Award Numbers: 1654175, 1720398, 1723005, 1845406

Summary

Uncertainty quantification for linear inverse problems remains a challenging task, especially for problems with a very large number of unknown parameters (e.g., dynamic inverse problems) and for problems where computation of the square root and inverse of the prior covariance matrix are not feasible. This work exploits Krylov subspace methods to develop and analyze new techniques for large-scale uncertainty quantification in inverse problems. In this work, we assume that generalized Golub-Kahan-based methods have been used to compute an estimate of the solution, and we describe efficient methods to explore the posterior distribution. In particular, we use the generalized Golub-Kahan bidiagonalization to derive an approximation of the posterior covariance matrix, and we provide theoretical results that quantify the accuracy of the approximate posterior covariance matrix and of the resulting posterior distribution. Then, we describe efficient methods that use the approximation to compute measures of uncertainty, including the Kullback-Liebler divergence. We present two methods that use the preconditioned Lanczos algorithm to efficiently generate samples from the posterior distribution. Numerical examples from dynamic photoacoustic tomography demonstrate the effectiveness of the described approaches.

KEYWORDS

generalized Golub-Kahan, preconditioned iterative methods, Bayesian inverse problems, uncertainty measures, Krylov subspace samplers

1 | INTRODUCTION

Inverse problems arise in various scientific applications, and a significant amount of effort has focused on developing efficient and robust methods to compute approximate solutions. However, as these numerical solutions are increasingly being used for data analysis and to aid in decision-making, there is a critical need to be able to obtain valuable uncertainty information (e.g., solution variances, samples, and credible intervals) to assess the reliability of computed solutions. Tools for inverse uncertainty quantification (UQ) often build upon the Bayesian framework from statistical inverse problems. Great overviews and introductions can be found in, for example, References 1-5.

Unfortunately, for very large inverse problems, UQ using the Bayesian approach is prohibitively expensive from a computational standpoint. This is partly because the posterior covariance matrices are so large that constructing, storing, and working with them directly are not computationally feasible. For these scenarios, hybrid Krylov subspace methods based on the generalized Golub-Kahan (genGK) bidiagonalization were proposed in Reference 6 to compute Tikhonov regularized solutions efficiently and to select a regularization parameter simultaneously and automatically. The next step is to go beyond computing reconstructions (e.g., maximum a posteriori [MAP] estimates) and to develop efficient methods for inverse UQ. In Reference 7 we used the genGK bidiagonalization to efficiently approximate the posterior covariance matrix and to estimate the posterior variance (diagonals of the posterior covariance matrix). In this paper, we extend the work in Reference 7 by providing theoretical analysis of the accuracy of the approximations to the posterior covariance and the posterior distribution. We then use the approximate posterior distribution to compute measures of uncertainty and develop preconditioned iterative solvers to efficiently sample from the posterior distribution by exploiting various tools from numerical linear algebra.

For concreteness, we consider linear inverse problems of the form

$$\mathbf{d} = \mathbf{A}\mathbf{s} + \delta, \tag{1}$$

where the goal is to reconstruct the desired parameters $\mathbf{s} \in \mathbb{R}^n$, given forward operator (or parameter-to-observable map) $\mathbf{A} \in \mathbb{R}^{m\times n}$ and the observed data $\mathbf{d} \in \mathbb{R}^m$. The inverse problem involves estimating the unknown parameters $\mathbf{s}$ from the data $\mathbf{d}$, and the recovery of these parameters is typically an ill-posed problem. We adopt a Bayesian approach where we assume that the measurement errors δ and the unknowns $\mathbf{s}$ are mutually independent Gaussian variables, that is, $\delta \sim \mathcal{N}(0, \mathbf{R})$ and $\mathbf{s} \sim \mathcal{N}(\boldsymbol{\mu}, \lambda^{-2}\mathbf{Q})$ where $\mathbf{R}$ and $\mathbf{Q}$ are symmetric positive definite matrices, $\boldsymbol{\mu} \in \mathbb{R}^n$, and λ is a scaling parameter that controls the prior precision and is also known as the regularization parameter. For the problems of interest, computing the inverse and square root of $\mathbf{R}$ are inexpensive (e.g., $\mathbf{R}$ is a diagonal matrix), but explicit computation of $\mathbf{Q}$ (or its inverse or square root) may not be possible. However, we assume that matrix-vector multiplications (matvecs) involving $\mathbf{A}$, $\mathbf{A}^\top$, and $\mathbf{Q}$ can be done efficiently (e.g., in $\mathcal{O}(n \log n)$ operations rather than $\mathcal{O}(n^2)$ operations for an $n \times n$ matrix).

Recall Bayes' theorem, which states that the posterior probability distribution function is given by

$$\pi_{\text{post}} = \pi(\mathbf{s}|\mathbf{d}) = \frac{\pi(\mathbf{d}|\mathbf{s})\pi(\mathbf{s})}{\pi(\mathbf{d})} .$$

Under our assumptions, the posterior distribution has the following probability density function

$$\pi_{\text{post}} \propto \exp\left(-\frac{1}{2}\|\mathbf{A}\mathbf{s} - \mathbf{d}\|^2_{\mathbf{R}^{-1}} - \frac{\lambda^2}{2}\|\mathbf{s} - \boldsymbol{\mu}\|^2_{\mathbf{Q}^{-1}}\right), \tag{2}$$

where $\|\mathbf{x}\|_{\mathbf{M}} = \sqrt{\mathbf{x}^\top\mathbf{M}\mathbf{x}}$ is a vector norm for any symmetric positive definite matrix $\mathbf{M}$, and $\propto$ means "proportional to." Thus, the posterior distribution is Gaussian, with corresponding measure $\rho_{\text{post}} = \mathcal{N}(\mathbf{s}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$, where the posterior covariance and mean are given as

$$\boldsymbol{\Gamma}_{\text{post}} \equiv (\lambda^2\mathbf{Q}^{-1} + \mathbf{A}^\top\mathbf{R}^{-1}\mathbf{A})^{-1} \quad \text{and} \quad \mathbf{s}_{\text{post}} = \boldsymbol{\Gamma}_{\text{post}}(\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{d} + \lambda^2\mathbf{Q}^{-1}\boldsymbol{\mu}), \tag{3}$$

respectively.[5] In the Bayesian framework, the solution to the inverse problem is the posterior distribution. However, for practical interpretation and data analysis, it is necessary to describe various characteristics of the posterior distribution.[2]

Typical inverse UQ approaches model the inverse of the prior covariance matrix (known as the precision matrix) as a discretized partial differential operator (e.g., Laplacian). This results in a sparse precision matrix that is relatively easy to factorize or solve linear systems with. In contrast, we model the prior covariance matrix entry-wise using covariance kernels (e.g., γ-exponential, or Matérn class), which allows the user the flexibility to incorporate a wide range of prior models, including nonseparable spatiotemporal kernels.[7] The main challenge is that the resulting prior covariance matrices are large and dense; explicitly forming and factorizing these matrices is prohibitively expensive. For such prior models, efficient matrix-free techniques (e.g., FFT embedding and $\mathcal{H}$-matrix approaches) can be used, for example, to compute matvecs with the prior covariance matrix $\mathbf{Q}$. In this paper, we describe new algorithms to perform inverse UQ for the scenario where obtaining the square root and inverse of $\mathbf{Q}$ are not computationally feasible, but matvecs with $\mathbf{Q}$ can be done efficiently. Specifically, we develop Krylov subspace algorithms that exploit the genGK bidiagonalization for approximating the posterior covariance matrix and for sampling from the posterior distribution.

1.1 | Overview of main contributions

In this paper, we use efficient Krylov subspace iterative methods, previously used for solving the weighted least squares problem, for inverse UQ. This motivates new algorithms and analyses, which is the central focus of this paper. The main contributions are as follows:

- We propose an approximation to the posterior covariance matrix using the genGK bidiagonalization that has an efficient representation (low-rank perturbation of the prior covariance matrix). This enables the computation of uncertainty measures involving the posterior distribution by storing bases for the Krylov subspaces during the computation of the MAP estimate and reusing the information contained in these subspaces for large-scale inverse UQ. We develop error bounds for monitoring the accuracy of the approximate posterior covariance matrix, based on the genGK iterates.
- We relate the error in the approximate posterior covariance matrix to the error in the approximate posterior distribution. We also show how to efficiently compute measures of uncertainty, such as the Kullback–Leibler (KL) divergence from the posterior to the prior distributions.
- We develop two different algorithms for generating samples from the posterior distribution using preconditioned Lanczos methods. The first algorithm uses the approximate posterior covariance matrix, whereas the second algorithm uses the true posterior covariance matrix but in different ways.

The idea of using low-rank perturbative approximations for the posterior covariance matrix previously appeared in References 8-11; however, these approaches rely on the ability to work with the square root or an appropriate factorization of **Q** (or its inverse). The authors in References 9,12 use randomized approaches to efficiently compute a low-rank approximation; in particular, the algorithm in Reference 12 does not require a factorization of **Q**. However, theoretical bounds suggest that randomized algorithms are effective when the singular values decay sufficiently rapidly. This assumption is valid for moderately or severely ill-posed inverse problems; however, for tomography-based applications, which we consider in this paper, the decay of the singular values is not sufficiently rapid, and therefore we pursue Krylov subspace methods. Previous work on Lanczos methods for sampling from Gaussian distributions can be found in, for example References 13-16, but these algorithms are meant for sampling from generic Gaussian distributions and do not exploit the structure of the posterior covariance matrix as we do.

The paper is organized as follows. In Section 2, we provide a brief overview of the genGK bidiagonalization and preconditioning of Krylov methods for sampling. Then, in Section 3, we use elements from the genGK bidiagonalization to approximate the posterior covariance matrix and provide theoretical bounds for the approximation. Not only are these bounds of interest for subsequent analysis and sampling, but they can also be used to determine a good stopping criterion for the iterative methods. In Section 4 we describe efficient Krylov subspace samplers for sampling from the posterior distribution. Numerical results for large inverse problems from image reconstruction are provided in Section 5, and conclusions and future work are provided in Section 6.

2 | BACKGROUND

In this section, we provide a brief background on two core topics that will be heavily used in the development of efficient methods to explore the posterior. In Section 2.1, we review an iterative hybrid method based on the genGK bidiagonlization that can be used to approximate the MAP estimate, which amounts to minimizing the negative log likelihood of the posterior probability distribution function, that is,

$$\mathbf{s}_{\text{post}} = \arg\min_{\mathbf{s}\in\mathbb{R}^n} - \log \pi_{\text{post}} = \arg\min_{\mathbf{s}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{As} - \mathbf{d}\|^2_{\mathbf{R}^{-1}} + \frac{\lambda^2}{2}\|\mathbf{s} - \boldsymbol{\mu}\|^2_{\mathbf{Q}^{-1}} . \tag{4}$$

Notice that with a change of variables, $\mathbf{s}_{\text{post}} = \boldsymbol{\mu} + \mathbf{Qx}$ where $\mathbf{x}$ is the solution to

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{AQx} - \mathbf{b}\|^2_{\mathbf{R}^{-1}} + \frac{\lambda^2}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}, \tag{5}$$

where $\mathbf{b} = \mathbf{d} - \mathbf{A}\boldsymbol{\mu}$. This change of variables is motivated by the fact that factorizing and/or inverting **Q** is infeasible in many applications. For more details on choices of prior covariance matrices **Q** for which this holds, we refer the reader to the discussion in our previous works (Reference 6, section 2.1) and (Reference 7, section 2.3). For readers familiar with hybrid Krylov iterative methods, Section 2.1 can be skipped. Then in Section 2.2, we review preconditioned Krylov subspace solvers for generating samples from normal distributions.

2.1 | Generalized hybrid iterative methods

Here, we provide an overview of the hybrid method based on the genGK bidiagonalization, but refer the interested reader to References 6,17 for more details.

The basic idea behind the generalized hybrid methods is first to generate a basis $\mathbf{V}_k$ for the Krylov subspace

$$\mathcal{S}_k \equiv \text{Span}\{\mathbf{V}_k\} = \mathcal{K}_k(\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{AQ}, \mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b}), \tag{6}$$

where $\mathcal{K}_k(\mathbf{M},\mathbf{g}) = \text{Span}\{\mathbf{g}, \mathbf{Mg}, \ldots, \mathbf{M}^{k-1}\mathbf{g}\}$, and second to solve (5) in this subspace. A basis for $\mathcal{S}_k$ can be generated using the genGK bidiagonalization process[*] summarized in Algorithm 1, where at the end of k steps, we have the matrices

$$\mathbf{U}_{k+1} \equiv [\mathbf{u}_1, \ldots, \mathbf{u}_{k+1}], \mathbf{V}_k \equiv [\mathbf{v}_1, \ldots, \mathbf{v}_k], \quad \text{and} \quad \mathbf{B}_k \equiv \begin{bmatrix} \alpha_1 & & \\ \beta_2 & \ddots & \\ & \ddots & \alpha_k \\ & & \beta_{k+1} \end{bmatrix}, \tag{7}$$

that in exact arithmetic satisfy

$$\mathbf{AQV}_k = \mathbf{U}_{k+1}\mathbf{B}_k, \quad \mathbf{A}^\top\mathbf{R}^{-1}\mathbf{U}_{k+1} = \mathbf{V}_k\mathbf{B}_k^\top + \alpha_{k+1}\mathbf{v}_{k+1}\mathbf{e}_{k+1}^\top, \tag{8}$$

and

$$\mathbf{U}_{k+1}^\top\mathbf{R}^{-1}\mathbf{U}_{k+1} = \mathbf{I}_{k+1}, \quad \mathbf{V}_k^\top\mathbf{QV}_k = \mathbf{I}_k. \tag{9}$$

Vector $\mathbf{e}_{k+1}$ corresponds to the $(k+1)$th standard unit vector.

Algorithm 1. genGK bidiagonlization

```
Result: $[\mathbf{U}_k, \mathbf{V}_k, \mathbf{B}_k] = \texttt{genGK}(\mathbf{A}, \mathbf{R}, \mathbf{Q}, \mathbf{b}, k)$
$\beta_1\mathbf{u}_1 = \mathbf{b}$, where $\beta_1 = ||\mathbf{b}||\mathbf{R}^{-1}$
$\alpha_1\mathbf{v}_1 = \mathbf{A}^\top\mathbf{R}^{-1}\mathbf{u}_1$, where $\alpha_1 = ||\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{u}_1||\mathbf{Q}$
for $i = 1, \ldots, k$ do
    $\beta_{i+1}\mathbf{u}_{i+1} = \mathbf{AQv}_i - \alpha_i\mathbf{u}_i$, where $\beta_{i+1} = ||\mathbf{AQv}_i - \alpha_i\mathbf{u}_i||\mathbf{R}^{-1}$
    $\alpha_{i+1}\mathbf{v}_{i+1} = \mathbf{A}^\top\mathbf{R}^{-1}\mathbf{u}_{i+1} - \beta_{i+1}\mathbf{v}_i$, where $\alpha_{i+1} = ||\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{u}_{i+1} - \beta_{i+1}\mathbf{v}_i||\mathbf{Q}$
end for
```

We seek an approximate solution to (5) of the form $\mathbf{x}_k = \mathbf{V}_k\mathbf{z}_k$, so that $\mathbf{x}_k \in \mathcal{S}_k$, where the coefficients $\mathbf{z}_k$ can be determined by solving the following problem,

$$\min_{\mathbf{x}_k\in\mathcal{S}_k} \frac{1}{2}||\mathbf{AQx}_k - \mathbf{b}||_{\mathbf{R}^{-1}}^2 + \frac{\lambda^2}{2}||\mathbf{x}_k||_{\mathbf{Q}}^2 \quad \Leftrightarrow \quad \min_{\mathbf{z}_k\in\mathbb{R}^k} \frac{1}{2}||\mathbf{B}_k\mathbf{z}_k - \beta_1\mathbf{e}_1||_2^2 + \frac{\lambda^2}{2}||\mathbf{z}_k||_2^2, \tag{10}$$

where the equivalency uses the relations in (8) and (9). For fixed λ, an approximate MAP estimate can be recovered by undoing the change of variables,

$$\mathbf{s}_k = \boldsymbol{\mu} + \mathbf{Qx}_k = \boldsymbol{\mu} + \mathbf{QV}_k(\mathbf{B}_k^\top\mathbf{B}_k + \lambda^2\mathbf{I})^{-1}\mathbf{B}_k^\top\beta_1\mathbf{e}_1 \ , \tag{11}$$

[*] genGK methods were first proposed by Benbow[18] for generalized least squares problems, and used in several applications, see, for example References 17,19,20. However, the specific form of the bidiagonalization was developed in Reference 6.

where now $\mathbf{s}_k \in \boldsymbol{\mu} + \mathbf{Q}\mathcal{S}_k$. If λ is not known a priori, a hybrid approach can be used where sophisticated SVD-based methods are applied to the right equation in (10). In this work, we use the hybrid implementation described in Reference 6 called genHyBR. The benefit of using this hybrid approach is that this algorithm automatically determines the number of iterations k and the regularization parameter λ_k at each iteration. In particular, the stopping condition is determined using a combination of approaches, including a maximum number of iterations, a generalized cross validation (GCV) function defined in terms of the iteration, and tolerances on the residual. The GCV function is given by

$$\hat{G}(k) = \frac{n\|(\mathbf{I} - \mathbf{A}\mathbf{A}_k^{\dagger})\mathbf{b}\|_{\mathbf{R}^{-1}}^2}{(\text{trace}(\mathbf{I} - \mathbf{R}^{-1}\mathbf{A}\mathbf{A}_k^{\dagger}))^2},$$

where $\mathbf{A}_k^{\dagger} = \mathbf{Q}\mathbf{V}_k(\mathbf{B}_k^{\top}\mathbf{B}_k + \lambda_k^2\mathbf{I})^{-1}\mathbf{B}_k^{\top}\mathbf{U}_{k+1}{}^{\top}$. Similar to the approach described in Reference 21, we select stopping iteration k that minimizes $\hat{G}(k)$ or when the absolute difference $|\hat{G}(k+1) - \hat{G}(k)|$ is smaller than some tolerance. Furthermore, the SVD of $\mathbf{B}_k$ can be used to simplify the expression for $\hat{G}(k)$. We remark that early termination of the genHyBR iterations can negatively affect the reconstruction and later approximations, but a later termination will not have a significant impact due to the inclusion of proper regularization. Thus, for subsequent UQ, a safer option is to perform a few extra iterations of the genHyBR method.

2.2 | Sampling from a Gaussian distribution

Let $\overline{\boldsymbol{\nu}} \in \mathbb{R}^n$ and let $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times n}$ be any symmetric positive definite matrix. Suppose the goal is to obtain samples from the Gaussian distribution $\mathcal{N}(\overline{\boldsymbol{\nu}}, \boldsymbol{\Gamma})$. Throughout this paper, let $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. If we have or are able to obtain a factorization of the form $\boldsymbol{\Gamma} = \mathbf{S}\mathbf{S}^{\top}$, then

$$\boldsymbol{\nu} = \overline{\boldsymbol{\nu}} + \mathbf{S}\boldsymbol{\epsilon},$$

is a sample from $\mathcal{N}(\overline{\boldsymbol{\nu}}, \boldsymbol{\Gamma})$, where it can be readily shown that $\mathbb{E}[\boldsymbol{\nu}] = \overline{\boldsymbol{\nu}}$ and

$$\text{Cov}(\boldsymbol{\nu}) = \mathbb{E}[(\boldsymbol{\nu} - \overline{\boldsymbol{\nu}})(\boldsymbol{\nu} - \overline{\boldsymbol{\nu}})^{\top}] = \mathbb{E}[\mathbf{S}\boldsymbol{\epsilon}\boldsymbol{\epsilon}\mathbf{S}^{\top}] = \boldsymbol{\Gamma}.$$

Note that any matrix $\mathbf{S}$ that satisfies $\mathbf{S}\mathbf{S}^{\top} = \boldsymbol{\Gamma}$ can be used to generate samples. We show how Krylov subspace solvers, in particular preconditioned versions, can be used to efficiently generate approximate samples from $\mathcal{N}(0, \boldsymbol{\Gamma})$ and $\mathcal{N}(0, \boldsymbol{\Gamma}^{-1})$. These approaches will be extended for sampling from the posterior in Section 4.

Given $\boldsymbol{\Gamma}$ and starting guess $\boldsymbol{\epsilon}$, after K steps of the symmetric Lanczos process, we have matrix $\mathbf{W}_K = [\mathbf{w}_1, \dots, \mathbf{w}_K] \in \mathbb{R}^{n\times K}$ that contains orthonormal columns and tridiagonal matrix

$$\mathbf{C}_K = \begin{bmatrix} \gamma_1 & \delta_2 & & & \\ \delta_2 & \gamma_2 & \delta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \delta_{K-1} & \gamma_{K-1} & \delta_K \\ & & & \delta_K & \gamma_K \end{bmatrix} \in \mathbb{R}^{K\times K}, \tag{12}$$

such that in exact arithmetic we have the following relation,

$$\boldsymbol{\Gamma}\mathbf{W}_K = \mathbf{W}_K\mathbf{C}_K + \delta_{K+1}\mathbf{w}_{K+1}\mathbf{e}_K^{\top}\,.$$

The Lanczos process is summarized in Algorithm 2. In practice, the vectors $\mathbf{W}_k$ tend to lose orthogonality in floating point arithmetic[16] and, therefore, we reorthogonalize the vectors to alleviate potential loss of orthogonality. Computed matrices $\mathbf{W}_K$ and $\mathbf{C}_K$ can then be used to obtain approximate draws from $\mathcal{N}(0, \boldsymbol{\Gamma})$ and $\mathcal{N}(0, \boldsymbol{\Gamma}^{-1})$ as

$$\boldsymbol{\xi}_K = \mathbf{W}_K\mathbf{C}_K^{1/2}\delta_1\mathbf{e}_1 \quad \text{and} \quad \boldsymbol{\zeta}_K = \mathbf{W}_K\mathbf{C}_K^{-1/2}\delta_1\mathbf{e}_1, \tag{13}$$

respectively. We remark that the Lanczos process, and in particular the computed matrices $\mathbf{W}_K$ and $\mathbf{C}_K$, depend on the realization $\boldsymbol{\epsilon}$. Although this can become costly if many samples are desired, the process is embarrassingly parallel. Furthermore, various methods (e.g., recycling techniques) can be exploited for handling multiple right-hand sides efficiently.

Algorithm 2. Lanczos tridiagonalization

```
Result: $[\mathbf{W}_K, \mathbf{C}_K] = \texttt{Lanczos}(\Gamma, \epsilon, K)$
$\delta_0 = 1, \mathbf{w}_0 = 0, \delta_1 = ||\epsilon||_2, \mathbf{w}_1 = \epsilon/\delta_1$
for $i = 1, \dots, K$ do
    $\gamma_i = \mathbf{w}_i^\top \Gamma \mathbf{w}_i$,
    $\mathbf{r} = \Gamma \mathbf{w}_i - \gamma_i \mathbf{w}_i - \delta_{i-1}\mathbf{w}_{i-1}$
    $\mathbf{w}_{i+1} = \mathbf{r}/\delta_i$, where $\delta_i = ||\mathbf{r}||_2$
end for
```

2.2.1 | Convergence

Let $\boldsymbol{\epsilon}$ be fixed for this discussion. The approximation improves as K increases, and we expect typical convergence behavior for the Lanczos process whereby convergence to extremal (i.e., largest and smallest) eigenvalues will be fast. The following result (Reference 15, theorem 3.3) sheds light onto the convergence of Krylov subspace methods for sampling. The error in the sample $\boldsymbol{\zeta}_K$ is given by

$$||\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\epsilon} - \boldsymbol{\zeta}_K||_2 \leq \sqrt{\lambda_{\min}(\boldsymbol{\Gamma})}||\mathbf{r}_K||_2,$$

where $\lambda_{\min}(\boldsymbol{\Gamma})$ is the smallest eigenvalue of $\boldsymbol{\Gamma}$. The term $\mathbf{r}_K = \boldsymbol{\epsilon} - \boldsymbol{\Gamma}\mathbf{x}_K$ is the residual vector at the K-th iteration of the conjugate gradient method and $\mathbf{x}_K = \mathbf{W}_K \mathbf{C}_K^{-1}\delta_1 \mathbf{e}_1$. The residual vector $||\mathbf{r}_K||_2$ can be bounded using standard techniques in Krylov subspace methods.[22] To use this as a stopping criterion, we note that $||\mathbf{r}_K||_2 = \delta_1 |\mathbf{e}_K^\top \mathbf{C}_K^{-1}\mathbf{e}_1|$ and by the Cauchy interlacing theorem [Reference 23, corollary III.1.3] $\lambda_{\min}(\boldsymbol{\Gamma}) \leq \lambda_{\min}(\mathbf{C}_K)$. Combining the two bounds we have

$$||\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\epsilon} - \boldsymbol{\zeta}_K||_2 \leq \sqrt{\lambda_{\min}(\mathbf{C}_K)}\delta_1 |\mathbf{e}_K^\top \mathbf{C}_K^{-1}\mathbf{e}_1|.$$

However, in numerical experiments we found that the bound was too pessimistic and instead adopted the approach in [Reference 16, section 4.1). Suppose we define the relative error norm as

$$e_K = \frac{||\boldsymbol{\zeta}_K - \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\epsilon}||_2}{||\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\epsilon}||_2}.$$

In practice, this quantity cannot be computed, therefore we use a different stopping criterion based on successive iterates (Reference 16, section 4.1) as

$$\tilde{e}_K = \frac{||\boldsymbol{\zeta}_K - \boldsymbol{\zeta}_{K+1}||_2}{||\boldsymbol{\zeta}_{K+1}||_2}.$$

The downside is that computing this error estimate is expensive since it costs $\mathcal{O}(nK^2)$ flops. However, this cost can be avoided by first writing

$$\boldsymbol{\zeta}_K = \mathbf{W}_K \hat{\boldsymbol{\zeta}}_K, \quad \text{where} \quad \hat{\boldsymbol{\zeta}}_K = \delta_1 \mathbf{C}_K^{-1/2}\mathbf{e}_1.$$

Since the columns of $\mathbf{W}_K$ are orthonormal, then

$$\tilde{e}_K = \frac{||\hat{\boldsymbol{\zeta}}'_K - \hat{\boldsymbol{\zeta}}_{K+1}||_2}{||\hat{\boldsymbol{\zeta}}_{K+1}||_2} \quad \hat{\boldsymbol{\zeta}}'_K \equiv \begin{bmatrix} \hat{\boldsymbol{\zeta}}_K \\ 0 \end{bmatrix}. \tag{14}$$

Therefore, $\tilde{e}_K$ can be computed in $\mathcal{O}(K^3)$ operations rather than $\mathcal{O}(nK^2)$ operations. A similar approach can be used to monitor the convergence of $\boldsymbol{\xi}_K$ to $\boldsymbol{\Gamma}^{1/2}\boldsymbol{\epsilon}$.

2.2.2 | Preconditioning

It is well known that an appropriate preconditioner can significantly accelerate convergence of Krylov subspace methods for solving linear systems. Assume that we have a preconditioner $\mathbf{G}$ which satisfies $\boldsymbol{\Gamma}^{-1} \approx \mathbf{G}^\top\mathbf{G}$. Then, the same preconditioner can be used to accelerate the convergence of Krylov subspace methods for generating samples, as we now show. Let

$$\mathbf{S} = \mathbf{G}^{-1}(\mathbf{G}\boldsymbol{\Gamma}\mathbf{G}^\top)^{1/2} \quad \text{and} \quad \mathbf{T} = \mathbf{G}^\top(\mathbf{G}\boldsymbol{\Gamma}\mathbf{G}^\top)^{-1/2} \ ,$$

then it is easy to see that

$$\boldsymbol{\Gamma} = \mathbf{G}^{-1}(\mathbf{G}\boldsymbol{\Gamma}\mathbf{G}^\top)\mathbf{G}^{-\top} = \mathbf{G}^{-1}(\mathbf{G}\boldsymbol{\Gamma}\mathbf{G}^\top)^{1/2}(\mathbf{G}\boldsymbol{\Gamma}\mathbf{G}^\top)^{1/2}\mathbf{G}^{-\top} = \mathbf{S}\mathbf{S}^\top, \tag{15}$$

and similarly $\boldsymbol{\Gamma}^{-1} = \mathbf{T}\mathbf{T}^\top$. The Lanczos process is then applied to $\mathbf{G}\boldsymbol{\Gamma}\mathbf{G}^\top$ and approximate samples from $\mathcal{N}(0, \boldsymbol{\Gamma})$ and $\mathcal{N}(0, \boldsymbol{\Gamma}^{-1})$ can be obtained by computing

$$\xi_K = \mathbf{G}^{-1}\mathbf{W}_K\mathbf{C}_K^{1/2}\delta_1\mathbf{e}_1 \qquad \zeta_K = \mathbf{G}^\top\mathbf{W}_K\mathbf{C}_K^{-1/2}\delta_1\mathbf{e}_1. \tag{16}$$

The preconditioned Lanczos sampling algorithm is summarized in Algorithm 3. If $\mathbf{G}$ is a good preconditioner, in the sense that $\boldsymbol{\Gamma}^{-1} \approx \mathbf{G}^\top\mathbf{G}$ (alternatively, $\mathbf{G}\boldsymbol{\Gamma}\mathbf{G}^\top \approx \mathbf{I}$), then the Krylov subspace method is expected to converge rapidly. The choice of preconditioner depends on the specific problem; we comment on the choice of preconditioners in the numerical experiments in Section 5.

Algorithm 3. Preconditioned Lanczos sampling

```
Result: $[\xi, \zeta]$ = LanczosSampling($\Gamma$, $\mathbf{G}$ $\epsilon$)
$\delta_0 = 1, \mathbf{w}_0 = 0, \delta_1 = ||\epsilon||_2, \mathbf{w}_1 = \epsilon/\delta_1$
Set $i = 0$
while not converged do
    Set $i \leftarrow i + 1$
    $\gamma_i = \mathbf{w}_i^\top \mathbf{G}\Gamma\mathbf{G}^\top\mathbf{w}_i$,
    $\mathbf{r} = \mathbf{G}\Gamma\mathbf{G}^\top\mathbf{w}_i - \gamma_i\mathbf{w}_i - \delta_{i-1}\mathbf{w}_{i-1}$
    $\mathbf{w}_{i+1} = \mathbf{r}/\delta_i$, where $\delta_i = \|\mathbf{r}\|_2$
end while
Let the number of iterations be $K$. Set $\hat{\xi}_K = \mathbf{C}_K^{1/2}\delta_1\mathbf{e}_1$, $\hat{\zeta}_K = \mathbf{C}_K^{-1/2}\delta_1\mathbf{e}_1$ (see (12))
Set $\xi = \mathbf{G}^{-1}\mathbf{W}_K\hat{\xi}_K$, $\zeta = \mathbf{G}^\top\mathbf{W}_K\hat{\zeta}_K$.
```

3 | APPROXIMATING THE POSTERIOR DISTRIBUTION USING THE GENGK BIDIAGONALIZATION

The basic goal of this section is to enable exploration of the posterior distribution for large-scale inverse problems by exploiting elements and relationships from the genGK bidiagonalization (cf, Equations (7)–(9)) to approximate the posterior covariance matrix $\boldsymbol{\Gamma}_{\text{post}}$.

Consider computing an approximate eigenvalue decomposition of $\mathbf{H} = \mathbf{A}^\top\mathbf{R}^{-1}\mathbf{A}$. We define the Ritz pairs $(\theta, \mathbf{y})$ obtained as the solution of the following eigenvalue problem,

$$(\mathbf{H}\mathbf{Q}\mathbf{V}_k\mathbf{y} - \theta\mathbf{V}_k\mathbf{y}) \perp_{\mathbf{Q}} \text{Span}\{\mathbf{V}_k\}.$$

Here the orthogonality condition $\perp_{\mathbf{Q}}$ is defined with respect to the weighted inner product $\langle\cdot,\cdot\rangle_{\mathbf{Q}}$. From (8) and (9), the Ritz pairs can be obtained by the solution of the eigenvalue problem

$$\mathbf{B}_k^\top\mathbf{B}_k\mathbf{y}_j = \theta_j\mathbf{y}_j \qquad j = 1, \ldots, k.$$

The Ritz pairs can be combined to express the eigenvalue decomposition in matrix form as,

$$\mathbf{B}_k^\top \mathbf{B}_k = \mathbf{Y}_k \boldsymbol{\Theta}_k \mathbf{Y}_k^\top.$$

The accuracy of the Ritz pairs can be quantified by the residual norm, defined as

$$\|\mathbf{r}_j\|_{\mathbf{Q}} \equiv \|\mathbf{HQV}_k \mathbf{y}_j - \theta_j \mathbf{V}_k \mathbf{y}_j\|_{\mathbf{Q}} = \alpha_{k+1}\beta_{k+1}|\mathbf{e}_k^\top \mathbf{y}_j| \qquad j = 1, \dots, k.$$

Furthermore, using arguments from (Reference 24, theorem 11.4.2) it can be shown that

$$\mathbf{T}_k \equiv \mathbf{B}_k^\top \mathbf{B}_k = \min_{\boldsymbol{\Delta} \in \mathbb{R}^{k\times k}} \|\mathbf{HQV}_k - \mathbf{V}_k \boldsymbol{\Delta}\|_{\mathbf{Q}},$$

is the best rank-k approximation over the subspace $\mathcal{S}_k \equiv \mathcal{K}_k(\mathbf{HQ}, \mathbf{A}^\top \mathbf{R}^{-1}\mathbf{b})$. Thus, the best rank-$k$ approximation of $\mathbf{H}$ over the space $\mathcal{S}_k$ is given by $\mathbf{H} \approx \mathbf{V}_k \mathbf{T}_k \mathbf{V}_k^\top$. Here we define the matrix $\|\cdot\|_{\mathbf{Q}}$ norm to be $\|\mathbf{M}\|_{\mathbf{Q}} = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Mx}\|_{\mathbf{Q}}$.

An approximation of this kind has been previously explored in References 8-10,12; however, the error estimates developed in the above references assume that the exact eigenpairs are available. If the Ritz pairs converge to the exact eigenpairs, then furthermore, the optimality result in (Reference 11, theorem 2.3) applies here as well.

For the rest of this paper, we use the following low-rank approximation of $\mathbf{H}$ which is constructed using the genGK bidiagonalization

$$\hat{\mathbf{H}} \equiv \mathbf{V}_k \mathbf{T}_k \mathbf{V}_k^\top. \tag{17}$$

Using this low-rank approximation, we can define the *approximate posterior distribution* $\hat{\pi}_{\text{post}}$, with the corresponding measure $\hat{\rho}_{\text{post}} = \mathcal{N}(\mathbf{s}_k, \hat{\boldsymbol{\Gamma}}_{\text{post}})$, which is a Gaussian distribution with covariance matrix

$$\hat{\boldsymbol{\Gamma}}_{\text{post}} \equiv (\lambda^2 \mathbf{Q}^{-1} + \hat{\mathbf{H}})^{-1}, \tag{18}$$

and mean $\mathbf{s}_k$ defined in (11). Using (18), we note that

$$\mathbf{s}_k = \boldsymbol{\mu} + \hat{\boldsymbol{\Gamma}}_{\text{post}} \mathbf{A}^\top \mathbf{R}^{-1} \mathbf{b}. \tag{19}$$

See Appendix A1 for the derivation.

3.1 | Posterior covariance approximation

First, we derive a way to monitor the accuracy of the low-rank approximation using the information available from the genGK bidiagonalization. This result is similar to (Reference 25, proposition 3.3).

Proposition 1. *Let* $\mathbf{H}_{\mathbf{Q}} = \mathbf{Q}^{1/2}\mathbf{H}\mathbf{Q}^{1/2}$ *and* $\hat{\mathbf{H}}_{\mathbf{Q}} = \mathbf{Q}^{1/2}\hat{\mathbf{H}}\mathbf{Q}^{1/2}$. *After k steps of Algorithm 1, the error in the low-rank approximation* $\hat{\mathbf{H}}$, *measured as*

$$\omega_k = \|\mathbf{H}_{\mathbf{Q}} - \hat{\mathbf{H}}_{\mathbf{Q}}\|_F, \tag{20}$$

satisfies the recurrence (for $k < n - 1$*)*

$$\omega_{k+1}^2 = \omega_k^2 - 2|\alpha_{k+1}\beta_{k+1}|^2 - |\alpha_{k+1}^2 + \beta_{k+2}^2|^2.$$

Proof. See Appendix A.2. ■

This proposition shows that, in exact arithmetic, the error in the low-rank approximation $\hat{\mathbf{H}}$ to $\mathbf{H}$ decreases monotonically as the iterations progress. Estimates for ω_k can be obtained in terms of the singular values of $\mathbf{R}^{-1/2}\mathbf{A}\mathbf{Q}^{1/2}$ following the approach in (Reference 25, theorem 3.2) and (Reference 26, theorem 2.7). However, we do not pursue them here.

Given the low-rank approximation, we can define the approximate posterior covariance $\hat{\mathbf{\Gamma}}_{\text{post}}$ in (18). The recurrence relation in Proposition 1 can be used to derive the following error estimates for $\mathbf{\Gamma}_{\text{post}}$.

Theorem 1. *The approximate posterior covariance matrix* $\hat{\mathbf{\Gamma}}_{\text{post}}$*, for* $k < n - 1$*, satisfies*

$$\|\mathbf{\Gamma}_{\text{post}} - \hat{\mathbf{\Gamma}}_{\text{post}}\|_F \leq \min \left\{ \omega_k \lambda^{-4} \|\mathbf{Q}\|_2, \frac{\omega_k \lambda^{-2} \|\mathbf{Q}\|_F}{\lambda^2 + \omega_k} \right\}.$$

Proof. See Appendix A.2. ■

The above theorem quantifies the error in the posterior covariance matrix in the Frobenius norm. However, the authors in Reference 11 argue that the Frobenius norm is not the appropriate metric to measure the distance between covariance matrices. Instead, they advocate the Förstner distance since it respects the geometry of the cone of positive definite covariance matrices. We take a different approach and consider metrics between the approximate and the true posterior distributions.

3.2 | Accuracy of posterior distribution

The KL divergence is a measure of "distance" between two different probability measures. The KL divergence is not a true metric on the set of probability measures, since it is not symmetric and does not satisfy the triangle inequality.[27] Despite these short-comings, the KL divergence is widely used since it has many favorable properties. Both the true and the approximate posterior measures are Gaussian, so the KL divergence from the approximate posterior measure to the posterior measure takes the form (using exercise 5.2 of Reference 27):

$$D_{\text{KL}}(\hat{\rho}_{\text{post}} \| \rho_{\text{post}}) = \frac{1}{2} \left[\text{trace}(\mathbf{\Gamma}_{\text{post}}^{-1} \hat{\mathbf{\Gamma}}_{\text{post}}) + \|\mathbf{s}_{\text{post}} - \mathbf{s}_k\|^2_{\mathbf{\Gamma}_{\text{post}}^{-1}} - n + \log \frac{\det \mathbf{\Gamma}_{\text{post}}}{\det \hat{\mathbf{\Gamma}}_{\text{post}}} \right].$$

We first present a result that can be used to monitor the accuracy of the trace of $\mathbf{H}_{\mathbf{Q}}$.

Proposition 2. *Let* $\theta_k = \text{trace}(\mathbf{H}_{\mathbf{Q}} - \hat{\mathbf{H}}_{\mathbf{Q}})$*. Then* θ_k *satisfies the recurrence relation*

$$\theta_{k+1} = \theta_k - (\alpha_{k+1}^2 + \beta_{k+2}^2) \qquad \text{for } \; k < n - 1.$$

Proof. See Appendix A.4. ■

Note that the Cauchy interlacing theorem implies that θ_k is non-negative; therefore, as with Proposition 1, this result implies that θ_k is monotonically decreasing.

Theorem 2. *At the end of k iterations, the KL divergence from the approximate to the true posterior measure satisfies*

$$0 \leq D_{\text{KL}}(\hat{\rho}_{\text{post}} \| \rho_{\text{post}}) \leq \frac{\lambda^{-2}}{2} \left[\theta_k + \frac{\omega_k^2}{\lambda^2 + \omega_k} \alpha_1^2 \beta_1^2 \right].$$

Proof. See Appendix A.4. ■

Both θ_k and ω_k are monotonically decreasing, implying that the the upper bound for the KL divergence from the approximate and the true posterior is getting smaller as the iterations progress. This theorem can be useful in providing bounds for the error using other metrics.

For example, consider the Hellinger metric and total variation (TV) distance denoted by $d_H(\rho_{\text{post}}, \hat{\rho}_{\text{post}})$ and $d_{\text{TV}}(\rho_{\text{post}}, \hat{\rho}_{\text{post}})$ respectively. Combining Pinsker's inequality (Reference 27, theorem 5.4) and Kraft's inequality (Reference 27, theorem 5.10), we have the following relationship

$$d_H^2(\rho_{\text{post}}, \hat{\rho}_{\text{post}}) \leq d_{\text{TV}}(\rho_{\text{post}}, \hat{\rho}_{\text{post}}) \leq \sqrt{2 D_{\text{KL}}(\hat{\rho}_{\text{post}} \| \rho_{\text{post}})}. \tag{21}$$

Thus, Theorem 2 can be used to find upper bounds for the Hellinger metric and the TV distance between the true and approximate posterior distributions. Furthermore, suppose $f : (\mathbb{R}^n, \|\cdot\|_{\mathbb{R}^n}) \to (\mathbb{R}^d, \|\cdot\|_{\mathbb{R}^d})$ is a function with finite second moments with respect to both measures, then by (Reference 27, proposition 5.12)

$$\|\mathbb{E}_{\rho_{\text{post}}}[f] - \mathbb{E}_{\hat{\rho}_{\text{post}}}[f]\|_{\mathbb{R}^n} \leq 2\sqrt{\mathbb{E}_{\rho_{\text{post}}}[\|f\|^2_{\mathbb{R}^d}] + \mathbb{E}_{\hat{\rho}_{\text{post}}}[\|f\|^2_{\mathbb{R}^d}]}\, d_{\mathcal{H}}(\rho_{\text{post}}, \hat{\rho}_{\text{post}}).$$

This implies that the error in the expectation of a function computed using the approximate posterior instead of the true posterior can be bounded by combining (21) and Theorem 2.

3.3 | Computation of information-theoretic metrics

In addition to providing a measure of distance from the approximate to the true posterior distributions, the KL divergence can also be used to measure the information gain between the prior and the posterior distributions. Similar to the derivation in Section 3.2 since both $\rho_{\text{prior}} = \mathcal{N}(\boldsymbol{\mu}, \lambda^{-2}\mathbf{Q})$ and ρ_{post} are Gaussian, the KL divergence takes the form

$$\begin{aligned} D_{\text{KL}}(\rho_{\text{post}}\|\rho_{\text{prior}}) &= \frac{1}{2}[\text{trace}(\lambda^2\mathbf{Q}^{-1}\boldsymbol{\Gamma}_{\text{post}}) + \lambda^2(\mathbf{s}_{\text{post}} - \boldsymbol{\mu})^\top\mathbf{Q}^{-1}(\mathbf{s}_{\text{post}} - \boldsymbol{\mu}) \\ &\quad - n - \log\det(\lambda^2\mathbf{Q}^{-1}\boldsymbol{\Gamma}_{\text{post}})] \\ &= \frac{1}{2}[\text{trace}(\mathbf{I} + \lambda^{-2}\mathbf{H}_{\mathbf{Q}})^{-1} + \lambda^2(\mathbf{s}_{\text{post}} - \boldsymbol{\mu})^\top\mathbf{Q}^{-1}(\mathbf{s}_{\text{post}} - \boldsymbol{\mu}) \\ &\quad - n + \log\det(\mathbf{I} + \lambda^{-2}\mathbf{H}_{\mathbf{Q}})]\ . \end{aligned}$$

Then, using the approximations generated by the genGK bidiagonlization, we consider the approximation

$$D_{\text{KL}} \equiv D_{\text{KL}}(\rho_{\text{post}}\|\rho_{\text{prior}}) \approx D_{\text{KL}}(\hat{\rho}_{\text{post}}\|\rho_{\text{prior}}) \equiv \hat{D}_{\text{KL}}.$$

Using the fact that $\log\det(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_{\mathbf{Q}}) = \log\det(\mathbf{I} + \lambda^{-2}\mathbf{T}_k)$,

$$\text{trace}(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_{\mathbf{Q}})^{-1} = n - \text{trace}(\mathbf{T}_k(\mathbf{T}_k + \lambda^2\mathbf{I})^{-1}),$$

and

$$\|\mathbf{s}_k - \boldsymbol{\mu}\|_{\mathbf{Q}^{-1}} = \|\mathbf{Q}\mathbf{V}_k(\mathbf{B}_k^\top\mathbf{B}_k + \lambda^2\mathbf{I})^{-1}\mathbf{B}_k^\top\beta_1\mathbf{e}_1\|_{\mathbf{Q}^{-1}} = \|\alpha_1\beta_1(\mathbf{T}_k + \lambda^2\mathbf{I})^{-1}\mathbf{e}_1\|_2^2\ ,$$

we get

$$\hat{D}_{\text{KL}} = \frac{1}{2}[-\text{trace}(\mathbf{T}_k(\mathbf{T}_k + \lambda^2\mathbf{I})^{-1}) + \lambda^2\|\mathbf{z}_k\|_2^2 + \log\det(\mathbf{I} + \lambda^{-2}\mathbf{T}_k)]$$

where $\mathbf{z}_k = \alpha_1\beta_1(\mathbf{T}_k + \lambda^2\mathbf{I})^{-1}\mathbf{e}_1$. Note that all of the terms only involve $k \times k$ tridiagonal matrices and, therefore, $\hat{D}_{\text{KL}}$ can be computed in $\mathcal{O}(k^3)$ operations once the genGK bidiagonalization has been computed.

The following result quantifies the accuracy of the estimator for the KL divergence from the posterior and the prior. Notice that the bound is similar to Theorem 2.

Theorem 3. *The error in the KL divergence for $k < n - 1$, in exact arithmetic, is given by*

$$|D_{\text{KL}} - \hat{D}_{\text{KL}}| \leq \lambda^{-2}\left[\theta_k + \frac{\lambda^2\omega_k}{\lambda^2 + \omega_k}\alpha_1^2\beta_1^2\right],$$

where ω_k and θ_k were defined in Propositions 1 and 2, respectively.

Proof. See Appendix A.4. ■

The D-optimal criterion for optimal experimental design is related to the expected KL divergence, where a precise connection was derived in (Reference 28, theorem 1). In D-optimal experimental design, the objective function is to maximize

$$\phi_D \equiv \log\det(\mathbf{I} + \lambda^{-2}\mathbf{H}_{\mathbf{Q}}).$$

Similar to the KL divergence, we can estimate the D-optimal criterion using the genGK bidiagonalization as

$$\hat{\phi}_D = \log\det(\mathbf{I} + \lambda^{-2}\mathbf{T}_k).$$

From the proof of Theorem 3, it can be readily seen that a bound for the error in approximating ϕ_D is given by

$$|\phi_D - \hat{\phi}_D| \leq \lambda^{-2}\theta_k.$$

4 | SAMPLING FROM THE POSTERIOR DISTRIBUTION

Since the posterior distribution is very high-dimensional, visualizing this distribution is challenging. A popular method is to generate samples from the posterior distribution (also sometimes known as conditional realizations), which provides a family of solutions and can be used for quantifying the reconstruction uncertainty. For instance, to compute the expected value of a quantity of interest $q(\cdot)$, defined as

$$\mathcal{Q} \equiv \mathbb{E}\,[q(\mathbf{s})|\mathbf{d}] = \int_{\mathbb{R}^n} q(\mathbf{s})\pi(\mathbf{s}|\mathbf{d})d\mathbf{s}.$$

Suppose, we have samples $\{\mathbf{s}^{(j)}\}_{j=1}^N$ then $\mathcal{Q}_N \equiv N^{-1}\sum_{j=1}^N q(\mathbf{s}^{(j)})$ is the *Monte Carlo estimate* of $\mathcal{Q}$. Furthermore, the Monte Carlo estimate converges to the expected value of the quantity of interest, that is, $\mathcal{Q}_N \to \mathcal{Q}$ as $N \to \infty$ almost surely, by the strong law of large numbers.

We now show how to draw samples from the posterior distribution $\mathcal{N}(\mathbf{s}_{\text{post}}, \mathbf{\Gamma}_{\text{post}})$. As described in Section 2.2, if $\epsilon \sim \mathcal{N}(0, I)$ and $\mathbf{S}\mathbf{S}^\top = \mathbf{\Gamma}_{\text{post}}$, then

$$\mathbf{s} = \mathbf{s}_{\text{post}} + \mathbf{S}\epsilon,$$

is a sample from $\mathcal{N}(\mathbf{s}_{\text{post}}, \mathbf{\Gamma}_{\text{post}})$. However, computing the posterior covariance matrix $\mathbf{\Gamma}_{\text{post}}$ and its factorization $\mathbf{S}$ are infeasible for reasons described before. We use preconditioned Krylov subspace methods to generate samples from the posterior distribution. A direct application of the approach in Section 2.2 to the posterior covariance matrix is expensive since it involves application of $\mathbf{Q}^{-1}$. To avoid this, we present several reformulations. The first approach we describe computes a low-rank approximation of $\mathbf{H}$ using the genGK approach and then uses this low-rank approximation to generate samples from the *approximate* posterior distribution. Any low-rank approximation can be used, provided it is sufficiently accurate. On the other hand, the second approach generates samples from the *exact* posterior distribution. Both methods use a preconditioner, albeit in different ways.

Before describing our proposed methods, we briefly review a few methods for sampling from high-dimensional Gaussian distributions. The idea of using Krylov subspace methods for sampling from Gaussian random processes seems to have originated from Reference 14. Variants of this idea have also been proposed in References 13,16 and have found applications in Bayesian inverse problems in References 15,29. The use of a low-rank surrogate of $\mathbf{H}_{\mathbf{Q}}$ has also been explored in References 9,10 and is similar to Method 1 (cf, Section 4.1) that we propose. Other approaches to sampling from the posterior distribution include randomize-then-optimize[30,31] and randomized MAP approach.[32] However, none of these methods can handle the case where $\mathbf{Q}^{-1}$ or $\mathbf{Q}^{-1/2}$ is not available.

4.1 | Method 1: Sampling from $\hat{\pi}_{\text{post}}$

Consider generating samples from $\hat{\pi}_{\text{post}}$, where $\hat{\mathbf{\Gamma}}_{\text{post}} = (\lambda^2\mathbf{Q}^{-1} + \hat{\mathbf{H}})^{-1}$ is the approximate posterior covariance matrix. Rewrite this as $\hat{\mathbf{\Gamma}}_{\text{post}} = \lambda^{-2}(\mathbf{Q}^{-1} + \lambda^{-2}\hat{\mathbf{H}})^{-1}$. Given a preconditioner $\mathbf{G}$, which we assume to be invertible, we can write

$$\mathbf{Q}^{-1} = \mathbf{G}^\top(\mathbf{G}\mathbf{Q}\mathbf{G}^\top)^{-1}\mathbf{G}.$$

Then, consider the factorization $\mathbf{Q}^{-1} = \mathbf{L}\mathbf{L}^\top$ where

$$\mathbf{L} \equiv \mathbf{G}^\top(\mathbf{G}\mathbf{Q}\mathbf{G}^\top)^{-1/2}. \tag{22}$$

An important point to note is that, while writing such an factorization, we do not propose to compute it explicitly. Instead, we access it in a matrix-free fashion using techniques from Algorithm 3.

Plugging the factorization of $\mathbf{Q}^{-1}$ into the expression for the approximate posterior covariance, we obtain

$$\hat{\boldsymbol{\Gamma}}_{\text{post}} = \lambda^{-2}(\mathbf{L}\mathbf{L}^\top + \lambda^{-2}\hat{\mathbf{H}})^{-1} = \lambda^{-2}\mathbf{L}^{-\top}(\mathbf{I} + \lambda^{-2}\mathbf{L}^{-1}\hat{\mathbf{H}}\mathbf{L}^{-1})^{-1}\mathbf{L}^{-1}.$$

We now have the following factorization of the approximate posterior covariance matrix

$$\hat{\boldsymbol{\Gamma}}_{\text{post}} = \hat{\mathbf{S}}\hat{\mathbf{S}}^\top \quad \text{where} \quad \hat{\mathbf{S}} \equiv \lambda^{-1}\mathbf{L}^{-\top}(\mathbf{I} + \lambda^{-2}\mathbf{L}^{-1}\hat{\mathbf{H}}\mathbf{L}^{-\top})^{-1/2}. \tag{23}$$

Algorithm 4. Low-rank representation $\mathbf{Z}\boldsymbol{\Theta}\mathbf{Z}^\top = \mathbf{Y}\mathbf{Y}^\top$

```
Result: [Z, Θ] = Lowrank(Y) for an arbitrary $\mathbf{Y} \in \mathbb{R}^{n\times k}$ with $k \le n$
Compute thin-QR factorization $\mathbf{QR} = \mathbf{Y}$
Compute eigenvalue decomposition $\mathbf{R}\mathbf{R}^\top = \mathbf{U}\Theta\mathbf{U}^\top$
Compute $\mathbf{Z} = \mathbf{QU}$
```

To efficiently compute matvecs with $\hat{\mathbf{S}}$, there are two stages: a precomputation stage and the sampling stage. In the precomputation stage, we first compute the low-rank representation

$$\lambda^{-2}\mathbf{L}^{-1}\hat{\mathbf{H}}\mathbf{L}^{-\top} = \lambda^{-2}\mathbf{L}^{-\top}\mathbf{V}_k\mathbf{T}_k\mathbf{V}_k^\top\mathbf{L}^{-1} = \mathbf{Z}_k\boldsymbol{\Theta}_k\mathbf{Z}_k^\top,$$

where $\mathbf{Z}_k$ has orthonormal columns and Θ_k is a diagonal matrix with nonnegative entries. Computing the low-rank representation is accomplished using Algorithm 4 with $\mathbf{Y}_k = \mathbf{L}^{-1}\mathbf{V}_k\mathbf{M}_k$ where $\mathbf{M}_k$ is the lower Cholesky factorization of $\lambda^{-2}\mathbf{T}_k = \lambda^{-2}\mathbf{B}_k^\top\mathbf{B}_k$. Now, we have

$$\hat{\mathbf{S}} \equiv \lambda^{-1}\mathbf{L}^{-\top}(\mathbf{I} + \lambda^{-2}\mathbf{L}^{-1}\hat{\mathbf{H}}\mathbf{L}^{-\top})^{-1/2} = \lambda^{-1}\mathbf{L}^{-\top}(\mathbf{I} - \mathbf{Z}_k\mathbf{D}_k\mathbf{Z}_k^\top).$$

In this step, we have used a variation of the Woodbury identity (Reference 33, equation (0.7.4.1))

$$(\mathbf{I} + \mathbf{Z}_k\boldsymbol{\Theta}_k\mathbf{Z}_k^\top)^{-1/2} = \mathbf{I} - \mathbf{Z}_k\mathbf{D}_k\mathbf{Z}_k^\top \qquad \mathbf{D}_k = \mathbf{I}_k \pm (\mathbf{I}_k + \boldsymbol{\Theta}_k)^{-1/2}.$$

In the sampling stage, given $\epsilon \sim \mathcal{N}(0, I)$, we can compute a sample from $\hat{\pi}_{\text{post}}$ as

$$\xi = \mathbf{s}_k + \lambda^{-1}\mathbf{L}^{-\top}(\epsilon - \mathbf{Z}_k\mathbf{D}_k\mathbf{Z}_k^\top\epsilon).$$

In summary, the procedure for computing samples $\xi^{(j)} \sim \mathcal{N}(0, \hat{\boldsymbol{\Gamma}}_{\text{post}})$ is provided in Algorithm 5. Computing matvecs with $\mathbf{L}$ (including its inverse and transpose) is done using the preconditioned Lanczos method described in subsection 2.2. Note that

$$\mathbf{L}^{-1} = (\mathbf{G}\mathbf{Q}\mathbf{G}^\top)^{1/2}\mathbf{G}^{-\top} \qquad \mathbf{L}^{-\top} = \mathbf{G}^{-1}(\mathbf{G}\mathbf{Q}\mathbf{G}^\top)^{1/2}.$$

The accuracy of the generated samples is discussed in Section 4.3.

Algorithm 5. Method 1: Generate N samples from $\hat{\pi}_{\text{post}}$

```
Result: $[\xi^{(1)}, \dots, \xi^{(N)}] = \texttt{Method1}(\mathbf{A}, \mathbf{R}, \mathbf{Q}, \mathbf{G}, \mathbf{b}, N)$
Use genHyBR to get $k$, $\mathbf{s}_k$, $\lambda$, $\mathbf{V}_k$, $\mathbf{B}_k$ (see Section 2.1)
Compute Cholesky factorization $\mathbf{M}_k\mathbf{M}_k^\top = \lambda^{-2}\mathbf{B}_k^\top\mathbf{B}_k$
{Stage 1: Precomputation stage}
for $j = 1, \dots, k$ do
    $\mathbf{z}^{(j)} = \mathbf{G}^{-\top}\mathbf{V}_k\mathbf{M}_k(:,j)$
    $[\xi^{(j)}, \sim] = \texttt{LanczosSampling}(\mathbf{G}\mathbf{Q}\mathbf{G}^\top, \mathbf{I}, \mathbf{z}^{(j)})$
    $\mathbf{Y}_k(:,j) = \xi^{(j)}$ {Computes $\mathbf{Y}_k(:,j) = \mathbf{L}^{-1}\mathbf{V}_k\mathbf{M}_k(:,j)$ }
end for
Compute $[\mathbf{Z}_k, \Theta_k] = \texttt{Lowrank}(\mathbf{Y}_k)$
Compute $\mathbf{D}_k = \mathbf{I}_k \pm (\mathbf{I}_k + \Theta_k)^{-1/2}$
{Stage 2: Sampling stage}
for $j = 1, \dots, N$ do
    Draw sample $\epsilon^{(j)} \sim \mathcal{N}(0, \mathbf{I})$. Compute $\mathbf{z}^{(j)} = \epsilon^{(j)} - \mathbf{Z}_k\mathbf{D}_k\mathbf{Z}_k^\top\epsilon^{(j)}$
    $[\xi^{(j)}, \sim] = \texttt{LanczosSampling}(\mathbf{Q}, \mathbf{G}, \mathbf{z}^{(j)})$ {Computes $\xi^{(j)} = \lambda^{-1}\mathbf{L}^{-\top}\epsilon^{(j)}$}
    Compute $\xi^{(j)} \leftarrow \mathbf{s}_k + \xi^{(j)}$
end for
```

4.2 | Method 2: Sampling from π_{post}

The second approach we describe generates samples from the exact posterior distribution π_{post}. First, we rewrite the posterior covariance matrix as

$$\mathbf{\Gamma}_{\text{post}} = (\lambda^2\mathbf{Q}^{-1} + \mathbf{H})^{-1} = \mathbf{Q}\mathbf{F}^{-1}\mathbf{Q} \qquad \mathbf{F} \equiv \lambda^2\mathbf{Q} + \mathbf{Q}\mathbf{H}\mathbf{Q}.$$

We define

$$\mathbf{S}_{\mathbf{F}} \equiv \mathbf{Q}\mathbf{F}^{-1/2}$$

such that $\mathbf{\Gamma}_{\text{post}} = \mathbf{S}_{\mathbf{F}}\mathbf{S}_{\mathbf{F}}^\top$. In this method, computing a factorization of $\mathbf{\Gamma}_{\text{post}}$ requires computing square roots with $\mathbf{F}$. Assume that we have a preconditioner $\mathbf{G}$ satisfying $\mathbf{G}\mathbf{G}^\top \approx \mathbf{F}^{-1}$. Armed with this preconditioner, we have the following factorization

$$\mathbf{\Gamma}_{\text{post}} = \mathbf{S}_{\mathbf{F}}\mathbf{S}_{\mathbf{F}}^\top \qquad \mathbf{S}_{\mathbf{F}} \equiv \mathbf{Q}\mathbf{G}^\top(\mathbf{G}\mathbf{F}\mathbf{G}^\top)^{-1/2}.$$

Note that this factorization of $\mathbf{\Gamma}_{\text{post}}$ is exact even though $\mathbf{G}\mathbf{G}^\top \approx \mathbf{F}^{-1}$; see (15). The application of the matrix $\mathbf{G}^\top(\mathbf{G}\mathbf{F}\mathbf{G}^\top)^{-1/2}$ to a randomly drawn vector can be accomplished by the preconditioned Lanczos approach described in Section 2.2. A matvec with $\mathbf{F}$ requires one matvec each with $\mathbf{A}$ and $\mathbf{A}^\top$ and two matvecs with $\mathbf{Q}$.

Algorithm 6. Method 2: Sampling from π_{post}

```
Result: $[\xi] = \texttt{Method2}(\mathbf{A}, \mathbf{R}, \mathbf{Q}, \mathbf{G}, \mathbf{s}_{\text{post}})$
Draw sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
Compute $[\sim, \zeta] = \texttt{LanczosSampling}(\mathbf{F}, \mathbf{G}, \epsilon)$ {Computes $\zeta = \mathbf{G}^\top(\mathbf{G}\mathbf{F}\mathbf{G}^\top)^{-1/2}\epsilon$}
Compute $\xi = \mathbf{s}_{\text{post}} + \mathbf{Q}\zeta$
```

As currently described, computing samples from $\mathbf{\Gamma}_{\text{post}}$ requires computing $\mathbf{s}_{\text{post}}$ and applying the matrix $\mathbf{A}$ and its adjoint $\mathbf{A}^\top$. However, this may be computationally expensive for several problems of interest. Here we use $\mathbf{s}_k$ as an approximation to $\mathbf{s}_{\text{post}}$. A variant of this method, not considered in this paper, follows by replacing the data-misfit part of the

TABLE 1 A summary of the main computational costs for Methods 1 and 2. The number of genHyBR iterations is denoted k, and the number of iterations in the preconditioned Lanczos sampling algorithm is denoted K. The columns labeled $\mathbf{A}$ and $\mathbf{A}^\top$ contain the number of matvecs with the forward and the adjoint operator respectively; the column labeled $\mathbf{Q}$ contains the number of matvecs with $\mathbf{Q}$, the column labeled $\mathbf{G}/\mathbf{G}^\top$ contains the number of matvecs with the preconditioner, and the column labeled $\mathbf{G}^{-1}/\mathbf{G}^{-\top}$ contains the number of solves involving the preconditioner

	Component	$\mathbf{A}$	$\mathbf{A}^\top$	$\mathbf{Q}$	$\mathbf{G}/\mathbf{G}^\top$	$\mathbf{G}^{-1}/\mathbf{G}^{-\top}$
Method 1	genHyBR	k	k	$2k$	–	–
	Precomputation	–	–	kK	$2kK$	K
	Sampling	–	–	$K+1$	$2K$	1
Method 2	Sampling	K	K	$2K+1$	$2K+1$	–

Hessian $\mathbf{H}$ by its low-rank approximation $\hat{\mathbf{H}}$, defined in (17). Define

$$\hat{\mathbf{F}} \equiv \lambda^2\mathbf{Q} + \mathbf{Q}\hat{\mathbf{H}}\mathbf{Q}.$$

Therefore, we compute the following factorization of the approximate posterior covariance

$$\hat{\boldsymbol{\Gamma}}_{\text{post}} = \hat{\mathbf{S}}_{\mathbf{F}}\hat{\mathbf{S}}_{\mathbf{F}}^\top \qquad \hat{\mathbf{S}}_{\mathbf{F}} \equiv \mathbf{Q}\mathbf{G}^\top(\mathbf{G}\hat{\mathbf{F}}\mathbf{G}^\top)^{-1/2}.$$

4.3 | Discussion

We now compare the two proposed methods for generating samples from the posterior. Method 1 only uses the forward operator $\mathbf{A}$ in the precomputation phase to generate the low-rank approximation and subsequently uses the low-rank approximation as a surrogate. This can be computationally advantageous if the forward operator is very expensive or if many samples are desired. On the other hand, if accuracy is important or only a few samples are needed, then Method 2 is recommended since it targets the exact posterior distribution. We summarize the computational costs of both methods in Table 1.

In Method 1, we generate samples from the approximate posterior distribution; the following result quantifies the error in the samples. Define $\mathbf{S} = \mathbf{Q}^{1/2}(\lambda^2\mathbf{I} + \mathbf{H}_{\mathbf{Q}})^{-1/2}$ such that $\boldsymbol{\Gamma}_{\text{post}} = \mathbf{S}\mathbf{S}^\top$ and let ϵ be a random draw from $\mathcal{N}(0, \mathbf{I})$, then

$$\mathbf{s} = \mathbf{s}_{\text{post}} + \mathbf{S}\epsilon \quad \text{and} \quad \hat{\mathbf{s}} = \mathbf{s}_k + \hat{\mathbf{S}}\epsilon$$

are samples from π_{post} and $\hat{\pi}_{\text{post}}$ respectively, where $\hat{\mathbf{S}}$ is defined in (23).

Theorem 4. *Let $\hat{\boldsymbol{\Gamma}}_{\text{post}}$ be the approximate posterior covariance matrix generated by running k steps of the genGK bidiagonalization algorithm and let ϵ be a fixed sample. The error in the sample $\hat{\mathbf{s}}$ satisfies*

$$\|\mathbf{s} - \hat{\mathbf{s}}\|_{\lambda^2\mathbf{Q}^{-1}} \leq \lambda^{-1}\left(\frac{\omega_k\alpha_1\beta_1}{\lambda^2 + \omega_k} + \sqrt{\frac{\lambda^2\omega_k}{\lambda^2 + \omega_k}}\|\epsilon\|_2\right).$$

Proof. See Appendix A.5. ■

Theorem 4 states that if ω_k is sufficiently small, then the accuracy of the samples is high. The samples, thus generated, can then be used *as is* in applications. Otherwise, they can be used as candidate draws from a proposal distribution $\hat{\pi}_{\text{post}}$. This proposal distribution can be used inside an independence sampler, similar to the approach in Reference 34.

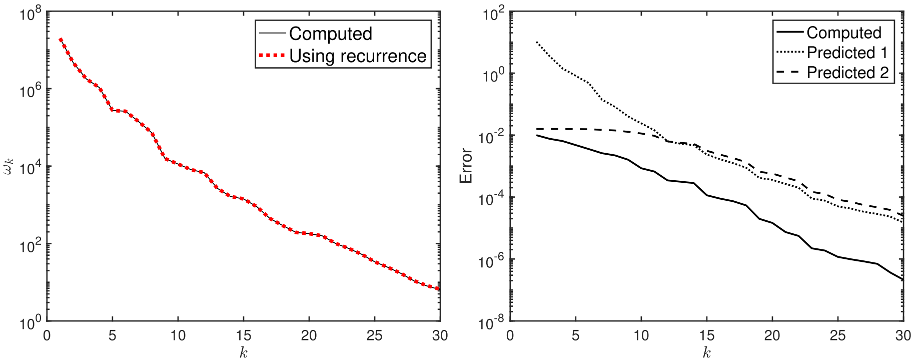

FIGURE 1 The left plot contains computed values of ω_k: the error between the true and the approximate prior-preconditioned Hessian for the data-misfit, as a function of the iteration k. The values for ω_k, as computed by the recurrence relationship presented in Proposition 1, are provided in the dotted line. The right plot contains the errors for the posterior covariance matrix $\|\boldsymbol{\Gamma}_{\text{post}} - \hat{\boldsymbol{\Gamma}}_{\text{post}}\|_F$ as a function of the iteration, along with two predicted bounds, whose minimum is given in Theorem 1

5 | NUMERICAL RESULTS

In Section 5.1, we investigate the accuracy of the low-rank approximation to $\mathbf{H}$ and the subsequent bounds that were derived in Section 3. Then, in Section 5.2, we describe our choice of preconditioners and demonstrate the efficiency of the preconditioned approaches proposed in Section 4 for generating samples from the posterior and approximate posterior. In the final experiment provided in Section 5.3, we demonstrate our methods on a very large dynamic tomography reconstruction problem. Our MATLAB code will be made freely available (upon acceptance of the article) at https://github.com/juliannechung/uq_krylov.

5.1 | Bounds for the posterior covariance matrix

For this example, we use the `heat` example from the Regularization Toolbox.[35] Matrix $\mathbf{A}$ is 256×256, and the observations are generated as (1), where $\boldsymbol{\delta}$ models the observational error. In the experiments, we take $\boldsymbol{\delta}$ to be 1% additive Gaussian white noise. We let $\mathbf{Q}$ be a 256×256 covariance matrix that was generated using an exponential kernel $\kappa(r) = \exp(-r/\ell)$ where r is the distance between two points and $\ell = 0.1$ is the correlation length. First, we use genHyBR to compute an approximate MAP estimate and simultaneously estimate a good regularization parameter. Using a weighted generalized cross validation (WGCV) method, the computed regularization parameter was $\lambda^2 \approx 5 \times 10^3$. The regularization parameter was then fixed for the remainder of the experiment.

Figure 1 shows the performance of the derived bounds. In the left plot, we track the accuracy of the prior-preconditioned data-misfit Hessian $\omega_k = \|\mathbf{H}_\mathbf{Q} - \hat{\mathbf{H}}_\mathbf{Q}\|_F$ as a function of the number of iterations. The error shows a decrease with increasing number of iterations k, and ω_k obtained by recursion is in close agreement with the actual error. This plot shows that, even in floating point arithmetic, the recursion relation for ω_k can be used to monitor the error of $\mathbf{H}_\mathbf{Q}$. In the right plot of Figure 1, we provide in the solid line the computed errors for the posterior covariance matrix $\|\boldsymbol{\Gamma}_{\text{post}} - \hat{\boldsymbol{\Gamma}}_{\text{post}}\|_F$, which decrease considerably with more iterations. To better illustrate the bound in Theorem 1, we provide predicted bounds $\omega_k \lambda^{-4} \|\mathbf{Q}\|_2$ (denoted "Predicted 1") and $\frac{\omega_k \lambda^{-2} \|\mathbf{Q}\|_F}{\lambda^2 + \omega_k}$ (denoted "Predicted 2"). While both bounds are qualitatively good, the first bound is slightly better at later iterations, whereas the second bound is more informative at earlier iterations. This can be attributed to the difference in the behavior of ω_k in the first bound versus $\omega_k/(\lambda^2 + \omega_k)$ in the second bound. The overall bound in Theorem 1 is obtained by taking the minimum value per iteration. These plots provide evidence that the low-rank approximation $\hat{\mathbf{H}}_\mathbf{Q}$ constructed using available components from the genGK bidiagonalization is quite accurate, and the bounds describing their behavior are informative.

5.2 | Sampling from the posterior

After describing the choice of preconditioners, we show the performance of these preconditioners within Lanczos approaches for sampling from the prior and the posterior.

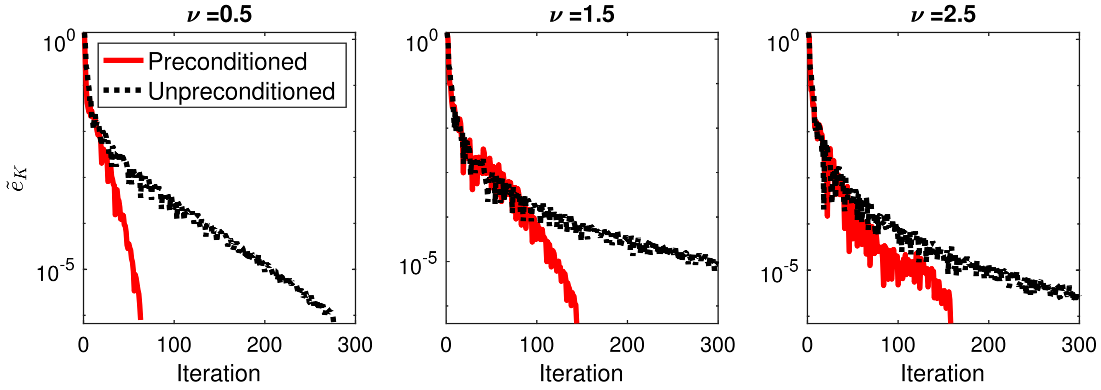

FIGURE 2 Relative differences $\tilde{\mathrm{e}}_K$ with and without preconditioning for sampling from $\mathcal{N}(0, \mathbf{Q})$. Preconditioners are based on the powers of the Laplacian $(-\boldsymbol{\Delta})^\gamma$. The plots correspond to various choices of ν in the Matérn covariance kernel and γ in the preconditioner. (left) $\nu = 1/2$ and $\mathbf{G}$ is the Cholesky factorization of $(-\boldsymbol{\Delta})$, (middle) $\nu = 3/2$ and $\gamma = 1$, and (right) $\nu = 5/2$ and $\gamma = 2$

5.2.1 | Preconditioners for Matérn covariance matrices

In this experiment, we investigate preconditioned Lanczos methods described in Section 2.2 for sampling from $\mathcal{N}(0, \mathbf{Q})$ where $\mathbf{Q}$ is defined by a Matérn covariance kernel. We pick three covariance matrices $\mathbf{Q}$ corresponding to Matérn parameters $\nu = 1/2,\ 3/2$, and $5/2$; this parameter controls the mean-squared differentiability of the underlying process. For a precise definition of the Matérn covariance function, see Reference 36, equation (1). The domain is set to $[0, 1]^2$, and we choose a 300×300 grid of evenly spaced points; thus, $\mathbf{Q}$ is a $90\,000 \times 90\,000$ matrix that is block-Toeplitz with Toeplitz blocks. Constructing such a matrix is never done explicitly; instead, circulant embedding and FFT-based techniques are used to efficiently perform matvecs. The correlation length ℓ is 0.25.

We use preconditioners of the form $\mathbf{G} = (-\boldsymbol{\Delta})^\gamma$ for parameters $\gamma \geq 1$, where $\boldsymbol{\Delta}$ is the Laplacian operator discretized using finite differences. These preconditioners are inspired by Reference 36 and exploit the fact that integral operators based on Matérn kernels have inverses that are fractional differential operators. We choose $\gamma = 1$ corresponding to $\nu = 3/2$ and $\gamma = 2$ corresponding to $\nu = 5/2$, respectively; for $\nu = 1/2$, we use $\mathbf{G}$ as the Cholesky factorization of $(-\boldsymbol{\Delta})$.

In Figure 2, we provide the relative differences (computed as $\tilde{\mathrm{e}}_K$ from (14)) per iteration of the preconditioned and unpreconditioned Lanczos approach. We use a fixed sample $\boldsymbol{\epsilon}$ for all the values of ν that we tested and remark that we observe similar results for other samples. It is readily seen that for $\nu = 1/2$ and $3/2$, including the preconditioner can dramatically speed up the convergence. Some improvement is seen for the case of $\nu = 5/2$, but the unpreconditioned solver does not converge within the maximum allotted number of iterations, which was set to 300. Also, the number of iterations that it takes to converge increases with increasing parameter ν; this is because the systems become more and more ill-conditioned, with increasing ν, for a fixed grid size. Finally, regarding the erratic convergence behavior, there is no guarantee that $\tilde{\mathrm{e}}_K$ decreases monotonically. In summary, we see that integral powers of the Laplacian operator can be good preconditioners for sampling from priors with Matérn covariance matrices. Next we investigate the use of these preconditioners for efficient sampling from the posterior.

5.2.2 | Sampling from the posterior distribution

In this experiment, we use the `PRspherical` test problem from the IRTools toolbox.[37,38] The true image $\mathbf{s}$ and forward model matrix $\mathbf{A}$ that models spherical means tomography are provided. We use the default settings provided by the toolbox; see Reference 37 for details. To simulate measurement error, we add 2% additive Gaussian noise.

For a grid size of 128×128 and for $\mathbf{Q}$ that represents a Matérn kernel with $\nu = 1/2$, we compute the MAP estimate using genHyBR and provide the reconstruction in the left panel of Figure 3. The relative reconstruction error in the 2-norm was 0.0168, and the regularization parameter determined using WGCV was $\lambda^2 \approx 19.48$. The regularization parameter was fixed for the remainder of this experiment. In Figure 3, we also show a random draw from the prior distribution $\mathcal{N}(0, \lambda^{-2}\mathbf{Q})$ in the middle panel and a random draw from the posterior distribution (computed using Method 2 in subsection 4.2) in the right panel. The same random vector $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ was used for both draws.

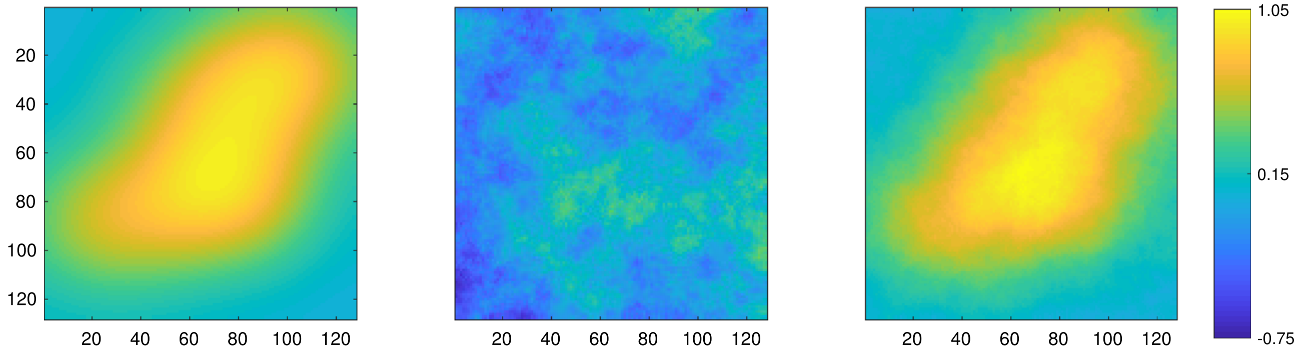

FIGURE 3 For the `PRspherical` problem, we provide the computed maximum a posteriori estimate (left), a random draw from the prior distribution (middle), and a random draw from the posterior distribution using Method 2 (right)

TABLE 2 For various examples of the `PRspherical` problem, we compare the performance of Methods 1 and 2 for sampling from the posterior. The notation n and m is used to denote the number of unknowns and measurements respectively. For Method 1, we provide the number of genHyBR iterations required to compute the MAP estimate (k), the total number of Lanczos iterations required for steps 4 to 8 of Algorithm 5 (Precomputation), and the average number of iterations required for steps 12 to 16 of Algorithm 5 (Sampling). For Method 2, we provide the number of iterations (averaged over 10 different runs) required for convergence in the preconditioned and unpreconditioned cases

		Method 1			**Method 2**	
n	***m***	***k***	**Precomputation**	**Sampling**	**Preconditioned**	**Unpreconditioned**
16×16	368	52	761	14.7	20.0	39.0
32×32	1,440	32	653	21.1	28.0	67.5
64×64	5,824	27	740	29.8	38.9	118.0
128×128	23,168	36	1386	42.5	53.7	206.3
256×256	92,672	63	3308	61.1	73	359.3

Next we demonstrate the performance of Method 1 in subsection 4.1 for sampling from the approximate posterior distribution $\hat{\pi}_{\text{post}}$ and the performance of Method 2 in Section 4.2 for sampling from the posterior. We vary the grid sizes from 16×16 to 256×256, and fix all other parameters ($\nu = 1/2$, 2% additive Gaussian noise) except the regularization parameter, which was determined for each problem using WGCV. The choice of ν defining the prior in this example was investigated in Reference 6, where there are comparisons to other choices of ν. The choice of preconditioners was described in Section 5.2.1.

For Method 1, we use the genHyBR method to obtain the MAP estimate, the regularization parameter λ^2, and the low-rank approximation $\hat{\mathbf{H}}_{\mathbf{Q}}$. In Table 2 we report the number of genHyBR iterations as k; see References 6,21 for details on stopping criteria. Then, we use Algorithm 5 to generate samples. Notice that steps 4 to 8 of Algorithm 5 require the application of $\mathbf{L}^{-\top}$ to the low-rank approximation; this is accomplished by using the approach described in Section 2.2, coupled with the choice of preconditioner described in Section 5.2.1. The number of total Lanczos iterations required is reported in the "Precomputation" column of Table 2. Then, for each sample, step 14 of Algorithm 5 requires the application of $\mathbf{L}^{-1}$, which is also done using a Lanczos iterative process; the number of iterations for this step, averaged over 10 samples, is listed under "Sampling" in Table 2.

For Method 2, we report the average number of iterations for the Lanczos solver to converge (i.e., achieving a residual tolerance of 10^{-6}) with and without a preconditioner in Table 2. We observe that the number of iterations required to achieve a desired tolerance increases with increasing problem size. This may be either because the matrices are becoming more ill-conditioned or because the number of measurements increases with increasing problem size, and the iterative solver has to work harder to process the additional "information content." We also notice that including the preconditioner cuts the number of iterations roughly in half. For the largest problem we consider here, the unpreconditioned iterative solver required over four times the number of iterations as the preconditioned solver. Since each iteration requires

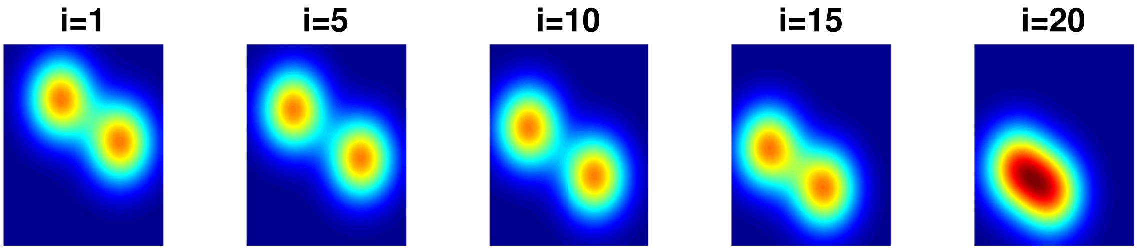

FIGURE 4 Five of the 20 true images for dynamic tomography example are provided

one matvec with **A** and one with $\mathbf{A}^\top$, each iteration can be quite expensive; the use of a preconditioner is beneficial in this case. Finally, another important observation is that although the preconditioners proposed in Section 5.2.1 were designed for the prior covariance matrix **Q**, here they were used for the matrix **F** instead; nevertheless, the results in Table 2 demonstrate that the preconditioners were similarly effective.

We make a few remarks about the results. First, the precomputation step to generate the low-rank approximation in Method 1 requires a considerable number of matvecs involving **Q** but far fewer involving **A**. Next, the number of iterations required for generating the samples in Method 1 is, on average, smaller than those reported for Method 2 for comparable problem size. The reason for this is that the preconditioner is designed for **Q** rather than **F**.

5.3 | Dynamic tomography example

In this experiment, we consider a dynamic tomography setup where the goal is to reconstruct a sequence of images from a sequence of projection datasets. Such scenarios are common in dynamic photoacoustic or dynamic electrical impedance tomography, where the underlying parameters change during the data acquisition process.[39-41] Reconstruction is particularly challenging for nonlinear or nonparametric deformations and often requires including a spatiotemporal prior.[7,42]

For this example, the true images were generated using two Gaussians moving in different directions in the image domain. We consider a sequence of 20 images (e.g., time points), where each image is 256 × 256. In Figure 4, we provide five of the true images.

We consider a linear problem of the form (1), where

$$\mathbf{s} = \begin{bmatrix} \mathbf{s}^{(1)} \\ \vdots \\ \mathbf{s}^{(20)} \end{bmatrix} \in \mathbb{R}^{20*256^2}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}^{(1)} & & \\ & \ddots & \\ & & \mathbf{A}^{(20)} \end{bmatrix}, \quad \text{and} \quad \mathbf{d} = \begin{bmatrix} \mathbf{d}^{(1)} \\ \vdots \\ \mathbf{d}^{(20)} \end{bmatrix}, \tag{24}$$

where $\mathbf{A}^{(i)} \in \mathbb{R}^{18*362\times 256^2}$ represents a spherical projection matrix corresponding to 18 equally spaced angles between i and $340+i$ for $i=1,\ldots,20$, and $\mathbf{d}^{(i)} \in \mathbb{R}^{18*362}$ contains projection data. To simulate measurement error we add 2% Gaussian noise.

For the spatiotemporal prior, we let $\mathbf{Q} = \mathbf{Q}_t \otimes \mathbf{Q}_s$, where $\mathbf{Q}_t \in \mathbb{R}^{20\times 20}$ and $\mathbf{Q}_s \in \mathbb{R}^{256^2\times 256^2}$ correspond to Matérn kernels with $\nu = 2.5, \ell = 0.1$ and $\nu = 0.5, \ell = 0.25$, respectively. First we use the genHyBR method to compute an approximation of the MAP estimate and to determine λ using WGCV. In Figure 5 we provide five of the images from the MAP reconstruction.

Since we can easily obtain a Cholesky factorization of $\mathbf{Q}_t^{-1} = \mathbf{G}_t^\top \mathbf{G}_t$, we define a preconditioner of the form $\mathbf{G} = \mathbf{G}_t \otimes \mathbf{G}_s$ where $\mathbf{G}_s$ is the Cholesky factorization of $(-\Delta)^\gamma$, with the exponent $\gamma = 1$ and $\otimes$ represents the Kronecker product. Then, we use the preconditioned sampling methods described in Section 4 to generate 10 samples from the prior, the approximate posterior (Method 1) and the posterior (Method 2). Note that each sample is a 256 × 256 × 20 volume. In Figure 6, we select one sample and provide five slices.

Next, we compare CPU timings (in seconds) and number of iterations for sampling, averaged over 10 samples, for both the preconditioned and unpreconditioned versions. In Table 3, we provide timings and iteration counts in parentheses for generating a sample from the prior, the approximate posterior using Method 1, and the posterior using Method 2.

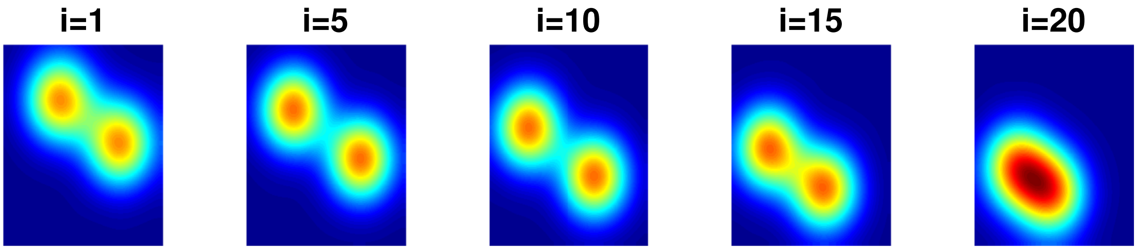

FIGURE 5 Slices of the maximum a posteriori estimate obtained using genHyBR for the dynamic tomography problem

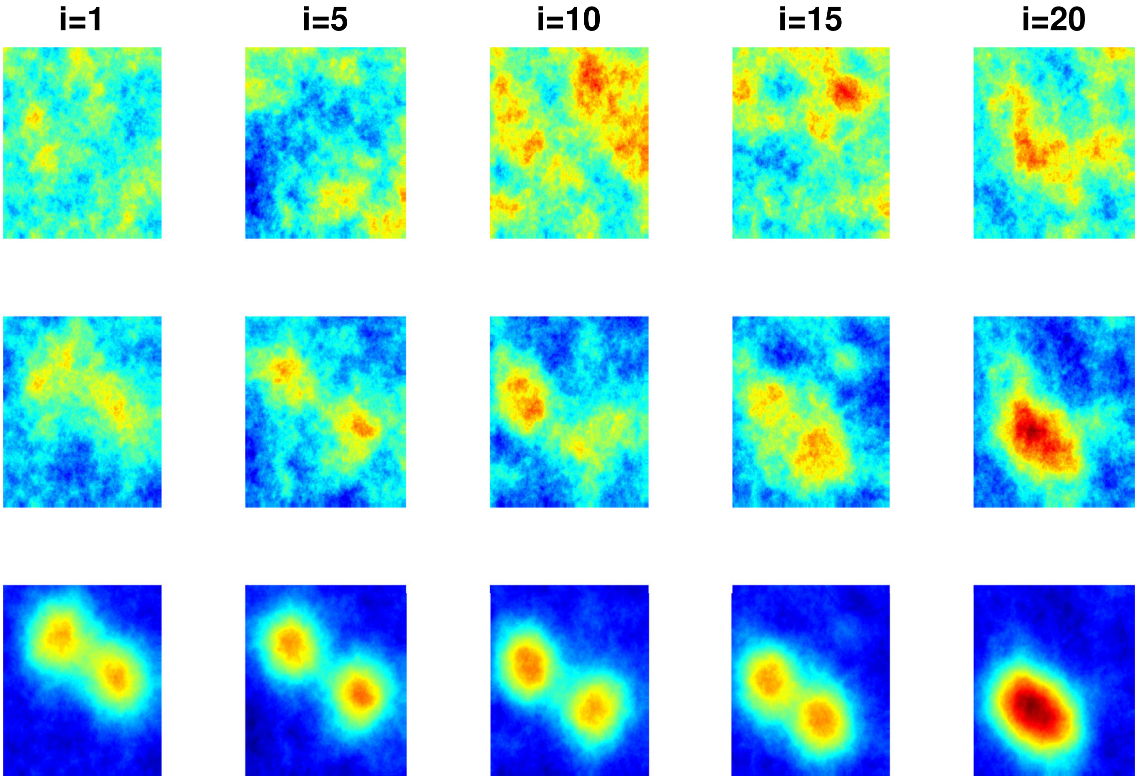

FIGURE 6 A random sample from the prior (first row), the approximate posterior using Method 1 (middle row), and the posterior using Method 2 (bottom row) for the dynamic tomography problem

For Method 1, we provide the total number of Lanczos iterations for precomputation followed by the average number of iterations for sampling (similar to Table 2). Again, we provide results for various problem dimensions. We remark that sampling from the approximate posterior, using Method 1, requires an upfront cost from precomputation, but if many samples are required, that cost can be amortized (cf, Table 1). On the other hand, if we need only a few, more accurate samples, then sampling from the true posterior, using Method 2, may be more efficient. We also observe that the use of a preconditioner significantly cuts the number of required iterations. Indeed, none of the unpreconditioned iterative solvers considered for this example converged within the maximum number of iterations taken to be 500.

6 | CONCLUSIONS

This paper addresses the challenging problem of providing an efficient representation for the posterior covariance matrix arising in high-dimensional inverse problems. To this end, we exploit Krylov subspace methods to derive an approximation to the posterior covariance matrix as a low-rank perturbation of the prior covariance matrix. The approximation is computed using information from the genGK bidiagonalization generated while computing the MAP estimate. As a

64 × 64 × 20	Preconditioned (iter)	Unpreconditioned (iter)
Prior sample	1.56 *s* (31.4)	297.54 s (500+)
Method 1	8.00 s (1379, 31)	1013.84 s (16143, 500+)
Method 2	15.55 s (104.5)	312.54 s (500+)
128 × 128 × 20	Preconditioned (iter)	Unpreconditioned (iter)
Prior sample	15.37 s (52.3)	1182.51 s (500+)
Method 1	58.02 s (1850, 48.9)	4984.79 s (18247, 500+)
Method 2	140.11 s (158.2)	1246.74 s (500+)
256 × 256 × 20	Preconditioned (iter)	Unpreconditioned (iter)
Prior sample	137.81 s (75)	5358.00 s (500+)
Method 1	460.29 s (2381, 75)	23147.48 s (18494, 500+)
Method 2	1255.31 s (238.4)	5563.96 s (500+)

TABLE 3 A comparison of CPU timings in seconds (s) and iteration counts for the dynamic tomography problem for both the preconditioned and unpreconditioned cases. Timings are averaged over 10 samples and average iteration counts are provided in parentheses, with iteration counts separated for Method 1 into total number of iterations for precomputation followed by average iteration counts for sampling.

result, we obtain an approximate representation at very little additional computational cost. Several results are presented to quantify the accuracy of this representation and of the resulting posterior distribution. We also show how to efficiently compute measures of uncertainty involving the posterior distribution. Then we present two methods that utilize a preconditioned Lanczos solver to efficiently generate samples from the posterior distribution. The first approach generates samples from an approximate posterior distribution, whereas the second approach generates samples from the exact posterior distribution. The approximate samples can be used *as is* or as candidate draws from a proposal distribution that closely approximates the exact posterior distribution.

There are several avenues for further research. The first important question is: Can we replace the bounds in the Frobenius norm by the spectral norm? The reason we employed the Frobenius norm is because of the recurrence relation in Proposition 1. Another issue worth exploring is if we can give bounds for the error in the low-rank approximation ω_k explicitly in terms of the eigenvalues of $\mathbf{H}_\mathbf{Q}$. This can be beneficial for deciding *a priori* the number of iterations required for an accurate low-rank approximation when the rate of decay of eigenvalues of $\mathbf{H}_\mathbf{Q}$ is known. Finally, we are interested in exploring the use of the approximate posterior distribution as a surrogate for the exact posterior distribution inside a Markov Chain Monte Carlo (MCMC) sampler. This is of particular interest for nonlinear problems where the posterior distribution is non-Gaussian. MCMC methods rely heavily on the availability of a good proposal distribution. One approach is to linearize the forward operator about the MAP estimate (the so-called Laplace's approximation) resulting in a Gaussian distribution with similar structure to π_{post}. This approximation to the true posterior distribution can be used as a proposal distribution, see for example References 43,44.

ACKNOWLEDGEMENTS

This work was partially supported by NSF DMS 1720398 (A.S. and K.P.), NSF DMS 1845406 (A.S.), NSF DMS 1654175 (J.C.), and NSF DMS 1723005 (J.C.). The authors would like to thank Silvia Gazzola, Per Christian Hansen, and James Nagy for generously sharing an advanced copy of their preprint[37] and code that we used in our numerical experiments.

ORCID

Arvind K. Saibaba https://orcid.org/0000-0002-8698-6100

REFERENCES

1. Biegler L, Biros G, Ghattas O, et al. Large-scale inverse problems and quantification of uncertainty. Vol 712. Hoboken, NJ: John Wiley & Sons, 2011.
2. Tenorio L. An introduction to data analysis and uncertainty quantification for inverse problems. Philadelphia, PA: SIAM, 2017.
3. Tenorio L, Andersson F, De Hoop M, Ma P. Data analysis tools for uncertainty quantification of inverse problems. Inverse Probl. 2011;27(4):045001.
4. Kaipio J, Somersalo E. Statistical and computational inverse problems. Vol 160. Berlin, Germany: Springer Science & Business Media, 2006.

5. Calvetti D, Somersalo E. An introduction to bayesian scientific computing: Ten lectures on subjective computing. Vol 2. New York, NY: Springer, 2007.
6. Chung J, Saibaba AK. Generalized hybrid iterative methods for large-scale Bayesian inverse problems. SIAM J Sci Comput. 2017;39(5):S24–S46.
7. Chung J, Saibaba AK, Brown M, Westman E. Efficient generalized Golub–Kahan based methods for dynamic inverse problems. Inverse Probl. 2018;34(2):024005.
8. Flath H, Wilcox L, Akçelik V, Hill J, van Bloemen-Waanders B, Ghattas O. Fast algorithms for Bayesian uncertainty quantification in large–scale linear inverse problems based on low-rank partial Hessian approximations. SIAM J Sci Comput. 2011;33(1):407–432.
9. Bui-Thanh T, Burstedde C, Ghattas O, Martin J, Stadler G, Wilcox LC. Extreme-scale UQ for Bayesian inverse problems governed by PDEs. Paper presented at: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis; 2012:3; Salt Lake City, UT, USA: IEEE, Computer Society Press.
10. Bui-Thanh T, Ghattas O, Martin J, Stadler G. A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion. SIAM J Sci Comput. 2013;35(6):A2494–A2523.
11. Spantini A, Solonen A, Cui T, Martin J, Tenorio L, Marzouk Y. Optimal low-rank approximations of Bayesian linear inverse problems. SIAM J Sci Comput. 2015;37(6):A2451–A2487.
12. Saibaba AK, Kitanidis PK. Fast computation of uncertainty quantification measures in the geostatistical approach to solve inverse problems. Adv Water Resour. 2015;82:124–138.
13. Parker A, Fox C. Sampling Gaussian distributions in Krylov spaces with conjugate gradients. SIAM J Sci Comput. 2012;34(3):B312–B334.
14. Schneider MK, Willsky AS. A Krylov subspace method for covariance approximation and simulation of random processes and fields. Multidim Syst Sign Process. 2003;14(4):295–318.
15. Simpson DP. Krylov Subspace Methods for Approximating Functions of Symmetric Positive Definite Matrices with Applications to Applied Statistics and Anomalous Diffusion [PhD thesis]. Queensland University of Technology; 2008.
16. Chow E, Saad Y. Preconditioned Krylov subspace methods for sampling multivariate Gaussian distributions. SIAM J Sci Comput. 2014;36(2):A588–A608.
17. Arioli M. Generalized Golub–Kahan bidiagonalization and stopping criteria. SIAM J Matrix Anal Appl. 2013;34(2):571–592.
18. Benbow SJ. Solving generalized least-squares problems with LSQR. SIAM J Matrix Anal Appl. 1999;21(1):166–177.
19. Arioli M, Orban D. Iterative methods for symmetric quasi-definite linear systems–Part I: Theory. Cahier du GERAD G-2013-32. Montréal, Canada: GERAD, Montréal, QC, 2013.
20. Orban D, Arioli M. Iterative solution of symmetric quasi-definite linear systems. Philadelphia, PA: SIAM, 2017.
21. Chung J, Nagy JG, O'Leary DP. A weighted GCV method for Lanczos hybrid regularization. Electron Trans Numer Anal. 2008;28:149–167.
22. Saad Y. Iterative methods for sparse linear systems. 2nd ed. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2003.
23. Bhatia R. Matrix analysis. Vol 169. Berlin, Germany: Springer Science & Business Media, 2013.
24. Parlett BN. The symmetric eigenvalue problem. Classics in Applied Mathematics, Vol 20, (xxiv+398). Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 1980. https://doi-org.prox.lib.ncsu.edu/10.1137/1.9781611971163.
25. Simon HD, Zha H. Low-rank matrix approximation using the Lanczos bidiagonalization process with applications. SIAM J Sci Comput. 2000;21(6):2257–2274.
26. Huang Y, Jia Z. Some results on the regularization of LSQR for large-scale discrete ill-posed problems. Sci Chin Math. 2017;60(4):701–718.
27. Sullivan TJ. Introduction to uncertainty quantification. Vol 63. New York, NY: Springer, 2015.
28. Alexanderian A, Saibaba AK. Efficient D-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems. SIAM J Sci Comput. 2018;40(5):A2956–A2985.
29. Gilavert C, Moussaoui S, Idier J. Efficient Gaussian sampling for solving large-scale inverse problems using MCMC. IEEE Trans Signal Process. 2015;63(1):70–80.
30. Bardsley JM, Solonen A, Haario H, Laine M. Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems. SIAM J Sci Comput. 2014;36(4):A1895–A1910.
31. Bardsley JM, Seppänen A, Solonen A, Haario H, Kaipio J. Randomize-then-optimize for sampling and uncertainty quantification in electrical impedance tomography. SIAM/ASA J Uncert Quantif. 2015;3(1):1136–1158.
32. Wang K, Bui-Thanh T, Ghattas O. A randomized maximum a posteriori method for posterior sampling of high dimensional nonlinear Bayesian inverse problems. SIAM J Sci Comput. 2018;40(1):A142–A171.
33. Horn RA, Johnson CR. Matrix analysis. Cambridge, MA: Cambridge University Press, 2012.
34. Brown DA, Saibaba A, Vallélian S. Low-rank independence samplers in hierarchical Bayesian inverse problems. SIAM/ASA J Uncert Quantif. 2018;6(3):1076–1100.
35. Hansen PC. Regularization tools: A MATLAB package for analysis and solution of discrete ill-posed problems. Numer Alg. 1994;6(1):1–35.
36. Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. J Royal Stat Soc Ser B (Stat Methodol). 2011;73(4):423–498.
37. Gazzola S, Hansen PC, Nagy JG. IR Tools: a MATLAB package of iterative regularization methods and large-scale test problems. Numer Alg. 2019;81(3):773–811.
38. Hansen PC, Jørgensen JS. AIR tools II: Algebraic iterative reconstruction methods, improved implementation. Numer Alg. 2018;79(1):107–137.
39. Wang K, Xia J, Li C, Wang LV, Anastasio MA. Fast spatiotemporal image reconstruction based on low-rank matrix estimation for dynamic photoacoustic computed tomography. J Biomed Opt. 2014;19(5):056007–056007.

40. Schmitt U, Louis AK, Wolters CH, Vauhkonen M. Efficient algorithms for the regularization of dynamic inverse problems: II. Applications. Inverse Probl. 2002;18(3):659.
41. Hahn BN. Efficient algorithms for linear dynamic inverse problems with known motion. Inverse Probl. 2014;30(3):035008.
42. Schmitt U, Louis AK. Efficient algorithms for the regularization of dynamic inverse problems: I. Theory. Inverse Probl. 2002;18(3):645.
43. Martin J, Wilcox LC, Burstedde C, Ghattas O. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. SIAM J Sci Comput. 2012;34(3):A1460–A1487.
44. Petra N, Martin J, Stadler G, Ghattas O. A computational framework for infinite-dimensional Bayesian inverse problems Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems. SIAM J Sci Comput. 2014;36(4):A1525–A1555.
45. Ouellette DV. Schur complements and statistics. Linear Algebra Appl. 1981;36:187–295.

How to cite this article: Saibaba AK, Chung J, Petroske K. Efficient Krylov subspace methods for uncertainty quantification in large Bayesian linear inverse problems. *Numer Linear Algebra Appl.* 2020;e2325. https://doi.org/10.1002/nla.2325

APPENDIX A. PROOFS

A.1 Derivation of (19)

First, we plug in $\hat{\boldsymbol{\Gamma}}_{\text{post}} = (\lambda^2\mathbf{Q}^{-1} + \hat{\mathbf{H}})^{-1}$ and rearrange to get

$$\begin{aligned}\hat{\boldsymbol{\Gamma}}_{\text{post}}\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b} &= (\lambda^2\mathbf{Q}^{-1} + \mathbf{V}_k\mathbf{T}_k\mathbf{V}_k^\top)^{-1}\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b}\\ &= (\lambda^2\mathbf{I} + \mathbf{Q}\mathbf{V}_k\mathbf{T}_k\mathbf{V}_k^\top)^{-1}\mathbf{Q}\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b}\ .\end{aligned}$$

Then, using the genGK relationships, we note that

$$\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b} = \mathbf{A}^\top\mathbf{R}^{-1}\mathbf{U}_{k+1}\beta_1\mathbf{e}_1 = \mathbf{V}_k\mathbf{B}_k^\top\beta_1\mathbf{e}_1\ .$$

Furthermore, using the Woodbury formula (Reference 33, equation (0.7.4.1)), we have

$$(\lambda^2\mathbf{I} + \mathbf{Q}\mathbf{V}_k\mathbf{T}_k\mathbf{V}_k^\top)^{-1} = \lambda^{-2}\mathbf{I} - \lambda^{-4}\mathbf{Q}\mathbf{V}_k(\mathbf{T}_k^{-1} + \lambda^{-2}\mathbf{I})^{-1}\mathbf{V}_k^\top\ .$$

Thus, we get

$$\begin{aligned}\hat{\boldsymbol{\Gamma}}_{\text{post}}\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b} &= (\lambda^{-2}\mathbf{I} - \lambda^{-4}\mathbf{Q}\mathbf{V}_k(\mathbf{T}_k^{-1} + \lambda^{-2}\mathbf{I})^{-1}\mathbf{V}_k^\top)\mathbf{Q}\mathbf{V}_k\mathbf{B}_k^\top\beta_1\mathbf{e}_1\\ &= \mathbf{Q}\mathbf{V}_k(\lambda^{-2}\mathbf{I} - \lambda^{-4}(\mathbf{T}_k^{-1} + \lambda^{-2}\mathbf{I})^{-1})\mathbf{B}_k^\top\beta_1\mathbf{e}_1\\ &= \mathbf{Q}\mathbf{V}_k(\mathbf{T}_k + \lambda^2\mathbf{I})^{-1}\mathbf{B}_k^\top\beta_1\mathbf{e}_1\ ,\end{aligned}$$

where the last equality uses the fact that $(\mathbf{T}_k^{-1} + \lambda^{-2}\mathbf{I})^{-1} = \lambda^2\mathbf{I} - \lambda^4(\mathbf{T}_k + \lambda^2\mathbf{I})^{-1}$. Since $\mathbf{T}_k = \mathbf{B}_k^\top\mathbf{B}_k$, we have the desired result.

A.2 Proofs for Section 3.1

Proof. First, we recognize that $\hat{\mathbf{H}} = \mathbf{V}_k\mathbf{T}_k\mathbf{V}_k^\top$, where $\mathbf{T}_k = \mathbf{V}_k^\top\mathbf{Q}\mathbf{H}\mathbf{Q}\mathbf{V}_k$ is a tridiagonal matrix of the form

$$\mathbf{T}_k = \begin{bmatrix} \mu_1 & \nu_2 & & & \\ \nu_2 & \mu_2 & \nu_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \nu_{k-1} & \mu_{k-1} & \nu_k \\ & & & \nu_k & \mu_k \end{bmatrix},$$

where $\mu_j = \alpha_j^2 + \beta_{j+1}^2$ and $\nu_j = \alpha_j\beta_j$ for $j = 1, \ldots, k$.

For simplicity denote $\hat{\mathbf{V}}_k = \mathbf{Q}^{1/2}\mathbf{V}_k$ and note that the columns of $\hat{\mathbf{V}}_k$ are orthonormal. Then write

$$\mathbf{H_Q} - \hat{\mathbf{H}}_\mathbf{Q} = (\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top)\mathbf{H_Q} + \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top\mathbf{H_Q}(\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top).$$

The observation that $(\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top)\mathbf{H_Q} \perp \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top\mathbf{H_Q}(\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top)$ with respect to the trace inner product, it is easy to show that

$$\omega_k^2 = \|(\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top)\mathbf{H_Q}\|_F^2 + \|\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top\mathbf{H_Q}(\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top)\|_F^2.$$

The second term is easy since using the genGK relationships and following (Reference 22, section 6.6.3), we have

$$\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top\mathbf{H_Q}(\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top) = \alpha_{k+1}\beta_{k+1}\hat{\mathbf{v}}_k\hat{\mathbf{v}}_{k+1}^\top,$$

and thus $\|\alpha_{k+1}\beta_{k+1}\hat{\mathbf{v}}_k\hat{\mathbf{v}}_{k+1}^\top\|_F^2 = |\alpha_{k+1}\beta_{k+1}|^2$. For the first term, we denote $\eta_k = \|(\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top)\mathbf{H_Q}\|_F$, so that

$$\omega_k^2 = \eta_k^2 + |\alpha_{k+1}\beta_{k+1}|^2. \tag{A1}$$

Then write $\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top = \mathbf{I} - \hat{\mathbf{V}}_{k+1}\hat{\mathbf{V}}_{k+1}^\top + \hat{\mathbf{v}}_{k+1}\hat{\mathbf{v}}_{k+1}^\top$ and again apply Pythagoras' theorem to get

$$\eta_k^2 = \eta_{k+1}^2 + \|\hat{\mathbf{v}}_{k+1}\hat{\mathbf{v}}_{k+1}^\top\mathbf{H_Q}\|_F^2.$$

From the genGK relations, it can be verified that

$$\begin{aligned}\mathbf{H_Q}\hat{\mathbf{v}}_{k+1}\hat{\mathbf{v}}_{k+1}^\top = {} & \alpha_{k+1}\beta_{k+1}\hat{\mathbf{v}}_k\hat{\mathbf{v}}_{k+1}^\top + (\alpha_{k+1}^2 + \beta_{k+2}^2)\hat{\mathbf{v}}_{k+1}\hat{\mathbf{v}}_{k+1}^\top \\ & + \alpha_{k+2}\beta_{k+2}\hat{\mathbf{v}}_{k+2}\hat{\mathbf{v}}_{k+1}^\top.\end{aligned} \tag{A2}$$

Since each term is mutually orthogonal, this implies

$$\eta_k^2 = \eta_{k+1}^2 + |\alpha_{k+1}\beta_{k+1}|^2 + |\alpha_{k+1}^2 + \beta_{k+2}^2|^2 + |\alpha_{k+2}\beta_{k+2}|^2.$$

Together with (A1), we get the desired recurrence. ■

Proof of Theorem 1. We now consider the error in the posterior covariance matrix. For the first bound, using

$$\mathbf{\Gamma}_{\text{post}} = \mathbf{Q}^{1/2}(\lambda^2\mathbf{I} + \mathbf{H_Q})^{-1}\mathbf{Q}^{1/2}\ ,$$

we have

$$\begin{aligned}\|\mathbf{\Gamma}_{\text{post}} - \hat{\mathbf{\Gamma}}_{\text{post}}\|_F &\leq \|\mathbf{Q}\|_2\|(\lambda^2\mathbf{I} + \mathbf{H_Q})^{-1} - (\lambda^2\mathbf{I} + \hat{\mathbf{H}}_\mathbf{Q})^{-1}\|_F \\ &= \lambda^{-2}\|\mathbf{Q}\|_2\|(\mathbf{I} + \lambda^{-2}\mathbf{H_Q})^{-1} - (\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}\|_F.\end{aligned}$$

With $f(x) = x/(1+x)$, it is verifiable that

$$(\mathbf{I} + \lambda^{-2}\mathbf{H_Q})^{-1} - (\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1} = f(\lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q}) - f(\lambda^{-2}\mathbf{H_Q}).$$

The function f is operator monotone (Reference 23, proposition V.1.6) and satisfies $f(0) = 0$. Since both $\lambda^{-2}\mathbf{H_Q}$ and $\lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q}$ are positive semi-definite, using (Reference 23, theorem X.1.3), we obtain

$$\|(\mathbf{I} + \lambda^{-2}\mathbf{H_Q})^{-1} - (\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}\|_F \leq \||\mathbf{E}|(\mathbf{I} + |\mathbf{E}|)^{-1}\|_F,$$

where we let $\mathbf{E} = \lambda^{-2}(\mathbf{H_Q} - \hat{\mathbf{H}}_\mathbf{Q})$, and $|\mathbf{E}| = (\mathbf{E}^*\mathbf{E})^{1/2}$. Note that both $|\mathbf{E}|$ and $\mathbf{E}$ have the same singular values, so $\||\mathbf{E}|\|_F = \|\mathbf{E}\|_F$. Since $|\mathbf{E}|$ is positive semi-definite, the singular values of $(\mathbf{I} + |\mathbf{E}|)^{-1}$ are at most 1. By submultiplicativity inequality and $\||\mathbf{E}|\|_F = \|\mathbf{E}\|_F$, we have

$$\|(\lambda^2\mathbf{I} + \mathbf{H_Q})^{-1} - (\lambda^2\mathbf{I} + \hat{\mathbf{H}}_\mathbf{Q})^{-1}\|_F \leq \|\lambda^{-2}(\mathbf{H_Q} - \hat{\mathbf{H}}_\mathbf{Q})\|_F = \lambda^{-2}\omega_k, \tag{A3}$$

and hence the desired result:

$$\|\mathbf{\Gamma}_{\text{post}} - \hat{\mathbf{\Gamma}}_{\text{post}}\|_F \leq \lambda^{-2}\|\mathbf{Q}\|_2\|\lambda^{-2}(\mathbf{H_Q} - \hat{\mathbf{H}}_\mathbf{Q})\|_F = \lambda^{-4}\omega_k\|\mathbf{Q}\|_2. \tag{A4}$$

For the second bound, we reverse the use of spectral and Frobenius norms

$$\|\mathbf{\Gamma}_{\text{post}} - \hat{\mathbf{\Gamma}}_{\text{post}}\|_F \leq \lambda^{-2}\|\mathbf{Q}\|_F\|(\mathbf{I} + \lambda^{-2}\mathbf{H_Q})^{-1} - (\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}\|_2.$$

Again, let $\mathbf{E} = \lambda^{-2}(\mathbf{H_Q} - \hat{\mathbf{H}}_\mathbf{Q})$, and use (Reference 23, theorem X.1.1) with $f(x) = x/(1+x)$, to obtain

$$\|(\mathbf{I} + \lambda^{-2}\mathbf{H_Q})^{-1} - (\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}\|_2 \leq \frac{\|\mathbf{E}\|_2}{1 + \|\mathbf{E}\|_2}.$$

It is readily verified that if $0 \leq a \leq b$, then $a(1+a)^{-1} \leq b(1+b)^{-1}$, and so

$$\|(\mathbf{I} + \lambda^{-2}\mathbf{H_Q})^{-1} - (\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}\|_2 \leq \frac{\|\mathbf{E}\|_2}{1 + \|\mathbf{E}\|_2} \leq \frac{\|\mathbf{E}\|_F}{1 + \|\mathbf{E}\|_F} = \frac{\omega_k}{\lambda^2 + \omega_k}. \tag{A5}$$

The recognition that $\|\mathbf{E}\|_F = \lambda^{-2}\omega_k$ completes the proof. ■

A.3 Lemma of independent interest

We will need the following lemma to prove Theorem 2 and 3. This may be of independent interest beyond this paper.

Lemma 1. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric positive semidefinite and let $\mathbf{P} \in \mathbb{R}^{n\times n}$ be an orthogonal projection matrix. Then the following results hold*

$$\begin{aligned} |\text{trace}(\mathbf{I} + \mathbf{A})^{-1} - \text{trace}(\mathbf{I} + \mathbf{PAP})^{-1}| &\leq \text{trace}(\mathbf{A} - \mathbf{PAP}), \\ \text{trace}[(\mathbf{I} + \mathbf{A})(\mathbf{I} + \mathbf{PAP})^{-1}] &\leq n + \text{trace}(\mathbf{A} - \mathbf{PAP}) \\ 0 \leq \log\det(\mathbf{I} + \mathbf{A}) - \log\det(\mathbf{I} + \mathbf{PAP}) &\leq \text{trace}(\mathbf{A} - \mathbf{PAP}). \end{aligned}$$

Proof. Let $\{\lambda_i\}_{i=1}^n$ and $\{\mu_i\}_{i=1}^n$ denote the eigenvalues of $\mathbf{A}$ and $\mathbf{PAP}$. Since both matrices are positive semidefinite, their eigenvalues are nonnegative. Since $\mathbf{P}$ is a projection matrix, its singular values are at most 1. The multiplicative singular value inequalities (Reference 23, problem III.6.2) say $\sigma_i(\mathbf{PA}^{1/2}) \leq \sigma_i(\mathbf{A}^{1/2})$, so $\lambda_i \geq \mu_i$ for $i = 1, \dots, n$, and therefore, $\text{trace}(\mathbf{A}) \geq \text{trace}(\mathbf{PAP})$. Then for the first inequality

$$\begin{aligned} |\text{trace}(\mathbf{I} + \mathbf{PAP})^{-1} - \text{trace}(\mathbf{I} + \mathbf{A})^{-1}| &= \left|\sum_{i=1}^n \frac{\lambda_i - \mu_i}{(1+\mu_i)(1+\lambda_i)}\right| \\ &\leq \left|\sum_{i=1}^n (\lambda_i - \mu_i)\right| = |\text{trace}(\mathbf{A} - \mathbf{PAP})|. \end{aligned}$$

The inequalities follow since λ_i, μ_i are nonnegative. The absolute value disappears since $\text{trace}(\mathbf{A}) \geq \text{trace}(\mathbf{PAP})$.

For the second inequality, write

$$(\mathbf{I} + \mathbf{A})(\mathbf{I} + \mathbf{PAP})^{-1} = \mathbf{A}(\mathbf{I} + \mathbf{PAP})^{-1} - \mathbf{PAP}(\mathbf{I} + \mathbf{PAP})^{-1} + \mathbf{I}.$$

Both $\mathbf{A}$ and $(\mathbf{I} + \mathbf{PAP})^{-1}$ are positive semidefinite (the second matrix is definite), so the trace of their product is nonnegative (Reference 33, exercise 7.2.26). Then a straightforward application of the von Neumann trace theorem (Reference

33, theorem 7.4.1.1) leads to

$$\mathrm{trace}(\mathbf{A}(\mathbf{I}+\mathbf{PAP})^{-1}) \leq \sum_{i=1}^{n} \frac{\lambda_i}{1+\mu_i}.$$

By utilizing its eigendecomposition, we see that $\mathrm{trace}[\mathbf{PAP}(\mathbf{I}+\mathbf{PAP})^{-1}] = \sum_{i=1}^{n} \frac{\mu_i}{1+\mu_i}$. Putting it together, we get

$$\begin{aligned}\mathrm{trace}[(\mathbf{I}+\mathbf{A})(\mathbf{I}+\mathbf{PAP})^{-1}] &\leq n + \sum_{i=1}^{n}\left(\frac{\lambda_i}{1+\mu_i} - \frac{\mu_i}{1+\mu_i}\right)\\ &\leq n + \sum_{i=1}^{n}\frac{\lambda_i-\mu_i}{1+\mu_i} \leq n + \sum_{i=1}^{n}(\lambda_i-\mu_i).\end{aligned}$$

Connecting the sum of the eigenvalues with the trace delivers the desired result.

For the third inequality, use Sylvester's determinant identity (Reference 45, corollary 2.1) to write

$$\log\det(\mathbf{I}+\mathbf{PAP}) = \log\det(\mathbf{I}+\mathbf{A}^{1/2}\mathbf{P}\mathbf{A}^{1/2}).$$

Denote $\mathbf{B} = \mathbf{A}^{1/2}\mathbf{P}\mathbf{A}^{1/2}$ and introduce the notation of Loewner partial ordering (Reference 33, section 7.7). Let $\mathbf{M},\mathbf{N} \in \mathbb{R}^{n\times n}$ be symmetric. Then, $\mathbf{M} \preccurlyeq \mathbf{N}$ means $\mathbf{N}-\mathbf{M}$ is positive semidefinite. Since $\mathbf{P} \preccurlyeq \mathbf{I}$, it follows that $\mathbf{B} \preccurlyeq \mathbf{A}$ (Reference 33, theorem 7.7.2). Then apply (Reference 28, lemma 9), to obtain

$$0 \leq \log\det(\mathbf{I}+\mathbf{A}) - \log\det(\mathbf{I}+\mathbf{B}) \leq \log\det(\mathbf{I}+\mathbf{A}-\mathbf{B}).$$

Finally, since $\log(1+x) \leq x$ for $x \geq 0$, $\log\det(\mathbf{I}+\mathbf{A}-\mathbf{B}) \leq \mathrm{trace}(\mathbf{A}-\mathbf{B})$. The proof is completed by observing that $\mathrm{trace}(\mathbf{B}) = \mathrm{trace}(\mathbf{PAP})$ by the cyclic property of trace. ■

A.4 Proofs of Sections 3.2 and 3.3

Proof. The linearity and cyclic property of trace estimator implies

$$\theta_k = \mathrm{trace}((\mathbf{I}-\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^{\top})\mathbf{H}_{\mathbf{Q}}).$$

As in the proof of Proposition 1, write $\mathbf{I}-\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^{\top} = \mathbf{I}-\hat{\mathbf{V}}_{k+1}\hat{\mathbf{V}}_{k+1}^{\top} + \hat{\mathbf{v}}_{k+1}\hat{\mathbf{v}}_{k+1}^{\top}$, so that

$$\theta_k = \theta_{k+1} + \mathrm{trace}(\hat{\mathbf{v}}_{k+1}\hat{\mathbf{v}}_{k+1}^{\top}\mathbf{H}_{\mathbf{Q}}).$$

The proof is finished if we apply the trace to the right-hand side of (A2). ■

Proof of Theorem 2. The lower bound follows from the property of the KL divergence and the fact that the distributions are not degenerate. The proof for the upper bound begins by providing an alternate expression for the error in the KL divergence.

$$D_{KL}(\hat{\pi}_{\mathrm{post}}\|\pi_{\mathrm{post}}) = \frac{1}{2}[\mathcal{E}_1+\mathcal{E}_2+\mathcal{E}_3],$$

where $\mathcal{E}_1 = \mathrm{trace}(\hat{\boldsymbol{\Gamma}}_{\mathrm{post}}\boldsymbol{\Gamma}_{\mathrm{post}}^{-1}) - n$,

$$\mathcal{E}_2 = \log\det(\boldsymbol{\Gamma}_{\mathrm{post}}) - \log\det(\hat{\boldsymbol{\Gamma}}_{\mathrm{post}}), \quad \text{and} \quad \mathcal{E}_3 = \|\mathbf{s}_{\mathrm{post}} - \mathbf{s}_k\|^2_{\boldsymbol{\Gamma}_{\mathrm{post}}^{-1}}.$$

We tackle each term individually. The second term $\mathcal{E}_2$ simplifies since

$$\log\det(\boldsymbol{\Gamma}_{\mathrm{post}}) - \log\det(\hat{\boldsymbol{\Gamma}}_{\mathrm{post}}) = \log\det(\mathbf{I}+\lambda^{-2}\hat{\mathbf{H}}_{\mathbf{Q}}) - \log\det(\mathbf{I}+\lambda^{-2}\mathbf{H}_{\mathbf{Q}}).$$

Let $\mathbf{M} = \lambda^{-2}(\mathbf{H_Q})$, then with $\mathbf{P} = \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top$ we have $\lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q} = \mathbf{PMP}$. Apply the third inequality in Lemma 1 to conclude $\mathcal{E}_2 \leq 0$. For the first term $\mathcal{E}_1$, apply the second part of Lemma 1 to obtain

$$\begin{aligned}\text{trace}(\hat{\boldsymbol{\Gamma}}_\text{post}\boldsymbol{\Gamma}_\text{post}^{-1}) =& \text{trace}[(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}(\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})] \\ \leq& n + \lambda^{-2}\text{trace}(\mathbf{H}_\mathbf{Q} - \hat{\mathbf{H}}_\mathbf{Q}).\end{aligned}$$

Therefore, $\mathcal{E}_1 \leq \lambda^{-2}\theta_k$. For the third term, notice that

$$\boldsymbol{\Gamma}_\text{post} - \hat{\boldsymbol{\Gamma}}_\text{post} = \lambda^{-2}\mathbf{Q}^{1/2}((\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})^{-1} - (\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1})\mathbf{Q}^{1/2}$$

and let $\mathbf{D} = (\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})^{-1} - (\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}$. Then

$$\|\mathbf{s}_\text{post} - \mathbf{s}_k\|^2_{\boldsymbol{\Gamma}^{-1}_\text{post}} = \hat{\mathbf{b}}^\top\mathbf{D}(\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})\mathbf{D}\hat{\mathbf{b}} \leq \|\mathbf{D}\hat{\mathbf{b}}\|_2\|(\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})\mathbf{D}\hat{\mathbf{b}}\|_2,$$

where $\hat{\mathbf{b}} = \mathbf{Q}^{1/2}\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b}$. The inequality is due to Cauchy–Schwartz. Using (A5), we can bound

$$\|\mathbf{D}\hat{\mathbf{b}}\|_2 \leq \frac{\omega_k\|\hat{\mathbf{b}}\|_2}{\lambda^2 + \omega_k}.$$

Next, with $\mathbf{E} = \lambda^{-2}(\mathbf{H}_\mathbf{Q} - \hat{\mathbf{H}}_\mathbf{Q})$, consider the simplification

$$(\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})\mathbf{D} = -\mathbf{E}(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1},$$

so that $\|(\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})\mathbf{D}\hat{\mathbf{b}}\|_2 \leq \lambda^{-2}\omega_k\|\hat{\mathbf{b}}\|_2$. Here, we have used submultiplicativity and the fact that singular values of $(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}$ are at most 1. From the genGK relations (see Algorithm 1)

$$\hat{\mathbf{b}} = \mathbf{Q}^{1/2}\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b} = \beta_1\mathbf{Q}^{1/2}\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{u}_1 = \beta_1\alpha_1\mathbf{Q}^{1/2}\mathbf{v}_1$$

which gives $\|\hat{\mathbf{b}}\|_2 = \alpha_1\beta_1\|\mathbf{Q}^{1/2}\mathbf{v}_1\|_2 = \alpha_1\beta_1$. Putting everything together, we see

$$\mathcal{E}_3 \leq \frac{\lambda^{-2}\omega_k^2\alpha_1^2\beta_1^2}{\lambda^2 + \omega_k}.$$

Gathering the bounds for $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ we have the desired result. ■

Proof of Theorem 3. The error in the KL-divergence satisfies

$$|D_\text{KL} - \hat{D}_\text{KL}| \leq \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3,$$

where

$$\begin{aligned}\mathcal{E}_1 &= \frac{1}{2}|\text{trace}(\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})^{-1} - \text{trace}(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}| \\ \mathcal{E}_2 &= \frac{1}{2}|\log\det(\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q}) - \log\det(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})| \\ \mathcal{E}_3 &= \frac{1}{2}\lambda^2|(\mathbf{s}_\text{post} - \boldsymbol{\mu})^\top\mathbf{Q}^{-1}(\mathbf{s}_\text{post} - \boldsymbol{\mu}) - (\mathbf{s}_k - \boldsymbol{\mu})^\top\mathbf{Q}^{-1}(\mathbf{s}_k - \boldsymbol{\mu})|.\end{aligned}$$

We tackle the first two terms together. As in the proof of Theorem 2, let $\mathbf{M} = \lambda^{-2}(\mathbf{H_Q})$, then with $\mathbf{P} = \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^\top$ we have $\lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q} = \mathbf{PMP}$. Apply the first and the third parts of Lemma 1 to obtain

$$\mathcal{E}_1 \leq \frac{\lambda^{-2}}{2}\text{trace}(\mathbf{H}_\mathbf{Q} - \hat{\mathbf{H}}_\mathbf{Q}) \qquad \mathcal{E}_2 \leq \frac{\lambda^{-2}}{2}\text{trace}(\mathbf{H}_\mathbf{Q} - \hat{\mathbf{H}}_\mathbf{Q}).$$

For the third term, let $\mathbf{s}_{\text{post}} = \mathbf{s}_k + \mathbf{e}$; add and subtract $(\mathbf{s}_k - \boldsymbol{\mu})\mathbf{Q}^{-1}(\mathbf{s}_{\text{post}} - \boldsymbol{\mu})$

$$\mathcal{E}_3 = \frac{1}{2}\lambda^2|\mathbf{e}^\top\mathbf{Q}^{-1}(\mathbf{s}_{\text{post}} - \boldsymbol{\mu}) + (\mathbf{s}_k - \boldsymbol{\mu})^\top\mathbf{Q}^{-1}\mathbf{e}|.$$

Notice that $\mathbf{e} = \mathbf{s}_{\text{post}} - \mathbf{s}_k = (\boldsymbol{\Gamma}_{\text{post}} - \hat{\boldsymbol{\Gamma}}_{\text{post}})\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b}$. Let

$$\hat{\mathbf{b}} \equiv \mathbf{Q}^{1/2}\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b} = \alpha_1\beta_1\mathbf{Q}^{1/2}\mathbf{v}_1,$$

and write

$$\mathbf{Q}^{-1/2}\mathbf{e} = ((\lambda^2\mathbf{I} + \mathbf{H}_\mathbf{Q})^{-1} - (\lambda^2\mathbf{I} + \hat{\mathbf{H}}_\mathbf{Q})^{-1})\hat{\mathbf{b}}.$$

So, the submultiplicative inequality and (A5) implies

$$\begin{aligned}\|\mathbf{Q}^{-1/2}\mathbf{e}\|_2 \leq& \lambda^{-2}\|(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1} - (\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})^{-1}\|_2\|\hat{\mathbf{b}}\|_2\\ \leq& \lambda^{-2}\frac{\omega_k}{\lambda^2 + \omega_k}\alpha_1\beta_1\ ,\end{aligned}$$

where we have used (A3). Next, applying the Cauchy–Schwartz inequality

$$|\mathbf{e}^\top\mathbf{Q}^{-1}(\mathbf{s}_{\text{post}} - \boldsymbol{\mu})| \leq \|\mathbf{Q}^{-1/2}\mathbf{e}\|_2\|\mathbf{Q}^{-1/2}(\mathbf{s}_{\text{post}} - \boldsymbol{\mu})\|_2.$$

Then, rewriting $\mathbf{s}_{\text{post}} = \boldsymbol{\mu} + \boldsymbol{\Gamma}_{\text{post}}\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b}$, we have

$$\|\mathbf{Q}^{-1/2}(\mathbf{s}_{\text{post}} - \boldsymbol{\mu})\|_2 = \|(\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})^{-1}\hat{\mathbf{b}}\|_2 \leq \|\hat{\mathbf{b}}\|_2 = \alpha_1\beta_1,$$

since the singular values of $(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}$ are less than 1. The other term is bounded in the same way. So, we have

$$\mathcal{E}_3 \leq \frac{\omega_k}{\lambda^2 + \omega_k}\alpha_1^2\beta_1^2.$$

Putting everything together along with $\mathcal{E}_1 + \mathcal{E}_2 \leq \lambda^{-2}\theta_k$ gives the desired result. ■

A.5 Proofs of Section 4

Proof of Theorem 4. By the triangle inequality, we have

$$\|\mathbf{s} - \hat{\mathbf{s}}\|_{\lambda^2\mathbf{Q}^{-1}} \leq \|\mathbf{s}_{\text{post}} - \mathbf{s}_k\|_{\lambda^2\mathbf{Q}^{-1}} + \|\mathbf{S}\epsilon - \hat{\mathbf{S}}\epsilon\|_{\lambda^2\mathbf{Q}^{-1}}.$$

Similar to previous proofs, we use $\mathbf{s}_{\text{post}} - \mathbf{s}_k = (\boldsymbol{\Gamma}_{\text{post}} - \hat{\boldsymbol{\Gamma}}_{\text{post}})\mathbf{A}^\top\mathbf{R}^{-1}\mathbf{b} = \lambda^{-2}\mathbf{Q}^{1/2}\mathbf{D}\hat{\mathbf{b}}$, where $\mathbf{D} = (\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})^{-1} - (\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1}$ to get

$$\|\mathbf{s}_{\text{post}} - \mathbf{s}_k\|^2_{\lambda^2\mathbf{Q}^{-1}} = \hat{\mathbf{b}}^\top\mathbf{D}\mathbf{Q}^{1/2}\lambda^{-2}(\lambda^2\mathbf{Q}^{-1})\lambda^{-2}\mathbf{Q}^{1/2}\mathbf{D}\hat{\mathbf{b}} = \lambda^{-2}\|\mathbf{D}\hat{\mathbf{b}}\|_2^2.$$

Thus,

$$\|\mathbf{s}_{\text{post}} - \mathbf{s}_k\|_{\lambda^2\mathbf{Q}^{-1}} = \lambda^{-1}\|\mathbf{D}\hat{\mathbf{b}}\|_2 \leq \lambda^{-1}\frac{\omega_k\alpha_1\beta_1}{\lambda^2 + \omega_k},$$

For the second term, we write

$$\lambda\mathbf{Q}^{-1/2}(\mathbf{S} - \hat{\mathbf{S}})\epsilon = [(\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})^{-1/2} - (\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1/2}]\epsilon.$$

Then, applying submultiplicativity

$$\|\mathbf{S}\epsilon - \hat{\mathbf{S}}\epsilon\|_{\lambda^2\mathbf{Q}^{-1}} = \|\lambda\mathbf{Q}^{-1/2}(\mathbf{S} - \hat{\mathbf{S}})\epsilon\|_2 \leq \|(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_\mathbf{Q})^{-1/2} - (\mathbf{I} + \lambda^{-2}\mathbf{H}_\mathbf{Q})^{-1/2}\|_2\|\epsilon\|_2.$$

When we apply theorem X.1.1 and (X.2) of Reference 23, we have

$$\|(\mathbf{I} + \lambda^{-2}\hat{\mathbf{H}}_{\mathbf{Q}})^{-1/2} - (\mathbf{I} + \lambda^{-2}\mathbf{H}_{\mathbf{Q}})^{-1/2}\|_2 \leq \|\mathbf{D}\|_2^{1/2}.$$

From (A5), $\|\mathbf{D}\|_2 \leq \omega_k/(\lambda^2 + \omega_k)$. Plugging this in gives the desired result. ■