MIT Open Access Articles

Optimal Low-rank Approximations
of Bayesian Linear Inverse Problems

OPTIMAL LOW-RANK APPROXIMATIONS OF BAYESIAN LINEAR INVERSE PROBLEMS[*]

ALESSIO SPANTINI[†], ANTTI SOLONEN[‡], TIANGANG CUI[†], JAMES MARTIN[§], LUIS TENORIO[¶], AND YOUSSEF MARZOUK[†]

Abstract. In the Bayesian approach to inverse problems, data are often informative, relative to the prior, only on a low-dimensional subspace of the parameter space. Significant computational savings can be achieved by using this subspace to characterize and approximate the posterior distribution of the parameters. We first investigate approximation of the posterior covariance matrix as a low-rank update of the prior covariance matrix. We prove optimality of a particular update, based on the leading eigendirections of the matrix pencil defined by the Hessian of the negative log-likelihood and the prior precision, for a broad class of loss functions. This class includes the Förstner metric for symmetric positive definite matrices, as well as the Kullback–Leibler divergence and the Hellinger distance between the associated distributions. We also propose two fast approximations of the posterior mean and prove their optimality with respect to a weighted Bayes risk under squared-error loss. These approximations are deployed in an offline-online manner, where a more costly but data-independent offline calculation is followed by fast online evaluations. As a result, these approximations are particularly useful when repeated posterior mean evaluations are required for multiple data sets. We demonstrate our theoretical results with several numerical examples, including high-dimensional X-ray tomography and an inverse heat conduction problem. In both of these examples, the intrinsic low-dimensional structure of the inference problem can be exploited while producing results that are essentially indistinguishable from solutions computed in the full space.

Key words. inverse problems, Bayesian inference, low-rank approximation, covariance approximation, Förstner–Moonen metric, posterior mean approximation, Bayes risk, optimality

AMS subject classifications. 15A29, 62F15, 68W25

DOI. 10.1137/140977308

1. Introduction. In the Bayesian approach to inverse problems, the parameters of interest are treated as random variables, endowed with a prior probability distribution that encodes information available before any data are observed. Observations are modeled by their joint probability distribution conditioned on the parameters of interest, which defines the likelihood function and incorporates the forward model and a stochastic description of measurement or model errors. The prior and likelihood then combine to yield a probability distribution for the parameters conditioned on the observations, i.e., the posterior distribution. While this formulation is quite general, essential features of inverse problems bring additional structure to the Bayesian update. The prior distribution often encodes some kind of smoothness or correlation

among the inversion parameters; observations typically are finite, few in number, and corrupted by noise; and the observations are indirect, related to the inversion parameters by the action of a forward operator that destroys some information. A key consequence of these features is that the data may be informative, relative to the prior, only on a *low-dimensional subspace* of the entire parameter space. Identifying and exploiting this subspace—to design approximations of the posterior distribution and related Bayes estimators—can lead to substantial computational savings.

In this paper we investigate approximation methods for finite-dimensional Bayesian linear inverse problems with Gaussian measurement and prior distributions. We characterize approximations of the posterior distribution that are structure-exploiting and that are *optimal* in a sense to be defined below. Since the posterior distribution is Gaussian, it is completely determined by its mean and covariance. We therefore focus on approximations of these posterior characteristics. Optimal approximations will reduce computation and storage requirements for high-dimensional inverse problems and will also enable fast computation of the posterior mean in a many-query setting.

We consider approximations of the posterior covariance matrix in the form of low-rank negative updates of the prior covariance matrix. This class of approximations exploits the structure of the prior-to-posterior update and also arises naturally in Kalman filtering techniques (e.g., [2, 3, 76]); the challenge is to find an *optimal* update within this class and to define in what sense it is optimal. We will argue that a suitable loss function with which to define optimality is the Förstner metric [29] for symmetric positive definite (SPD) matrices, and we will show that this metric generalizes to a broader class of loss functions that emphasize relative differences in covariance. We will derive the optimal low-rank update for this entire class of loss functions. In particular, we will show that the prior covariance matrix should be updated along the leading generalized eigenvectors of the pencil $(H, \Gamma_{\mathrm{pr}}^{-1})$ defined by the Hessian of the negative log-likelihood and the prior precision matrix. If we assume exact knowledge of the posterior mean, then our results extend to optimality statements between distributions (e.g., optimality in Kullback–Leibler (K-L) divergence and in Hellinger distance). The form of this low-rank update of the prior is not new [6, 9, 28, 60], but previous work has not shown whether—and if so, in exactly what sense—it yields optimal approximations of the posterior. A key contribution of this paper is to establish and explain such optimality.

Properties of the generalized eigenpairs of $(H, \Gamma_{\mathrm{pr}}^{-1})$ and related matrix pencils have been studied previously in the literature, especially in the context of classical regularization techniques for linear inverse problems[1] [81, 68, 38, 37, 25]. The joint action of the log-likelihood Hessian and the prior precision matrix have also been used in related regularization methods [13, 10, 12, 43]. However, these efforts have not been concerned with the posterior covariance matrix or with its optimal approximation, since this matrix is a property of the Bayesian approach to inversion.

One often justifies the assumption that the posterior mean is exactly known by arguing that it can easily be computed as the solution of a regularized least-squares problem [42, 69, 1, 62, 5]; indeed, evaluation of the posterior mean to machine precision is now feasible even for million-dimensional parameter spaces [6]. If, however, one needs multiple evaluations of the posterior mean for different realizations of the data

---

[1] In the framework of Tikhonov regularization [80], the regularized estimate coincides with the posterior mean of the Bayesian linear model we consider here, provided that the prior covariance matrix is chosen appropriately.

(e.g., in an online inference context), then solving a linear system to determine the posterior mean may not be the most efficient strategy. A second goal of our paper is to address this problem. We will propose two computationally efficient approximations of the posterior mean based on (i) evaluating a low-rank affine function of the data or (ii) using a low-rank update of the prior covariance matrix in the exact formula for the posterior mean. The optimal approximation in each case is defined as the minimizer of the Bayes risk for a squared-error loss weighted by the posterior precision matrix. We provide explicit formulas for these optimal approximations and show that they can be computed by exploiting the optimal posterior covariance approximation described above. Thus, given a new set of data, computing an optimal approximation of the posterior mean becomes a computationally trivial task.

Low-rank approximations of the posterior mean that minimize the Bayes risk for squared-error loss have been proposed in [17, 20, 19, 18, 16] for a general non-Gaussian case. Here, instead we develop analytical results for squared-error loss weighted by the posterior precision matrix. This choice of norm reflects the idea that approximation errors in directions of low posterior variance should be penalized more strongly than errors in high-variance directions, as we do not want the approximate posterior mean to fall outside the bulk of the posterior probability distribution. Remarkably, in this case, the optimal approximation only requires the leading eigenvectors and eigenvalues of a single eigenvalue problem. This is the same eigenvalue problem we solve to obtain an optimal approximation of the posterior covariance matrix, and thus we can efficiently obtain both approximations at the same time.

While the efficient solution of large-scale linear-Gaussian Bayesian inverse problems is of standalone interest [28], optimal approximations of Gaussian posteriors are also a building block for the solution of nonlinear Bayesian inverse problems. For example, the stochastic Newton Markov chain Monte Carlo (MCMC) method [60] uses Gaussian proposals derived from local linearizations of a nonlinear forward model; the parameters of each Gaussian proposal are computed using the optimal approximations analyzed in this paper. To tackle even larger nonlinear inverse problems, [6] uses a Laplace approximation of the posterior distribution wherein the Hessian at the mode of the log-posterior density is itself approximated using the present approach. Similarly, approximations of local Gaussians can facilitate the construction of a nonstationary Gaussian process whose mean directly approximates the posterior density [8]. Alternatively, [22] combines data-informed directions derived from local linearizations of the forward model—a direct extension of the posterior covariance approximations described in the present work—to create a global *data-informed subspace*. A computationally efficient approximation of the posterior distribution is then obtained by restricting MCMC to this subspace and treating complementary directions analytically. Moving from the finite- to the infinite-dimensional setting, the same global data-informed subspace is used to drive efficient dimension-independent posterior sampling for inverse problems in [21].

Earlier work on dimension reduction for Bayesian inverse problems used the Karhunen–Loève expansion of the prior distribution [61, 54] to describe the parameters of interest. To reduce dimension, this expansion is truncated; this step renders both the prior and posterior distributions singular—i.e., collapsed onto the prior mean—in the neglected directions. Avoiding large truncation errors then requires that the prior distribution impose significant smoothness on the parameters, so that the spectrum of the prior covariance kernel decays quickly. In practice, this requirement restricts the choice of priors. Moreover, this approach relies entirely on properties of the prior

distribution and does not incorporate the influence of the forward operator or the observational errors. Alternatively, [56] constructs a reduced basis for the parameter space via greedy model-constrained sampling, but this approach can also fail to capture posterior variability in directions uninformed by the data. Both of these earlier approaches seek reduction in the overall description of the parameters. This notion differs fundamentally from the dimension reduction technique advocated in this paper, where low-dimensional structure is sought in the *change* from prior to posterior.

The rest of this paper is organized as follows. In section 2 we introduce the posterior covariance approximation problem and derive the optimal prior-to-posterior update with respect to a broad class of loss functions. The structure of the optimal posterior covariance matrix approximation is examined in section 3. Several interpretations are given in this section, including an equivalent reformulation of the covariance approximation problem as an optimal *projection* of the likelihood function onto a lower-dimensional subspace. In section 4 we characterize optimal approximations of the posterior mean. In section 5 we provide several numerical examples. Section 6 offers concluding remarks. Appendix A collects proofs of many of the theorems stated throughout the paper, along with additional technical results.

2. Optimal approximation of the posterior covariance matrix. Consider the Bayesian linear model defined by a Gaussian likelihood and a Gaussian prior with a nonsingular covariance matrix $\Gamma_{\mathrm{pr}} \succ 0$ and, without loss of generality, zero mean:

$$(2.1) \qquad y \mid x \sim \mathcal{N}(Gx, \Gamma_{\mathrm{obs}}), \quad x \sim \mathcal{N}(0, \Gamma_{\mathrm{pr}}).$$

Here x represents the parameters to be inferred, G is the linear forward operator, and y are the observations, with $\Gamma_{\mathrm{obs}} \succ 0$. The statistical model (2.1) also follows from

$$y = Gx + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \Gamma_{\mathrm{obs}})$ is independent of x. It is easy to see that the posterior distribution is again Gaussian (see, e.g., [14]): $x \mid y \sim \mathcal{N}(\mu_{\mathrm{pos}}(y), \Gamma_{\mathrm{pos}})$, with mean and covariance matrix given by

$$(2.2) \qquad \mu_{\mathrm{pos}}(y) = \Gamma_{\mathrm{pos}}\, G^{\top} \Gamma_{\mathrm{obs}}^{-1}\, y \quad \text{and} \quad \Gamma_{\mathrm{pos}} = \left(H + \Gamma_{\mathrm{pr}}^{-1}\right)^{-1},$$

where

$$(2.3) \qquad H = G^{\top} \Gamma_{\mathrm{obs}}^{-1} G$$

is the Hessian of the negative log-likelihood (i.e., the Fisher information matrix). Since the posterior is Gaussian, the posterior mean coincides with the posterior mode: $\mu_{\mathrm{pos}}(y) = \arg\max_x \pi_{\mathrm{pos}}(x; y)$, where π_{pos} is the posterior density. Note that the posterior covariance matrix does not depend on the data.

2.1. Defining the approximation class. We will seek an approximation, $\widehat{\Gamma}_{\mathrm{pos}}$, of the posterior covariance matrix that is optimal in a class of matrices to be defined shortly. As we can see from (2.2), the posterior precision matrix $\Gamma_{\mathrm{pos}}^{-1}$ is a nonnegative update of the prior precision matrix $\Gamma_{\mathrm{pr}}^{-1}$: $\Gamma_{\mathrm{pos}}^{-1} = \Gamma_{\mathrm{pr}}^{-1} + ZZ^{\top}$, where $ZZ^{\top} = H$. Similarly, using Woodbury's identity we can write Γ_{pos} as a nonpositive update of Γ_{pr}: $\Gamma_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^{\top}$, where $KK^{\top} = \Gamma_{\mathrm{pr}} G^{\top} \Gamma_{\mathrm{y}}^{-1} G \Gamma_{\mathrm{pr}}$ and $\Gamma_{\mathrm{y}} = \Gamma_{\mathrm{obs}} + G \Gamma_{\mathrm{pr}} G^{\top}$ is the covariance matrix of the marginal distribution of y [47]. This update of Γ_{pr} is negative semidefinite because the data add information: the posterior variance in

any direction is always smaller than the corresponding prior variance. Moreover, the update is usually low rank for exactly the reasons described in the introduction: there are directions in the parameter space along which the data are not very informative, relative to the prior. For instance, H might have a quickly decaying spectrum [7, 75]. Note, however, that Γ_{pos} itself might *not* be low-rank. Low-rank structure, if any, lies in the update of Γ_{pr} that yields Γ_{pos}. Hence, a natural class of matrices for approximating Γ_{pos} is the set of negative semidefinite updates of Γ_{pr}, with a fixed maximum rank, that lead to positive definite matrices:

$$(2.4) \qquad \mathcal{M}_r = \left\{ \Gamma_{\text{pr}} - KK^\top \succ 0 : \operatorname{rank}(K) \leq r \right\}.$$

This class of approximations of the posterior covariance matrix takes advantage of the structure of the prior-to-posterior update.

2.2. Loss functions. Optimality statements regarding the approximation of a covariance matrix require an appropriate notion of distance between SPD matrices. We shall use the metric introduced by Förstner and Moonen [29], which is derived from a canonical invariant metric on the cone of real SPD matrices and is defined as follows: the Förstner distance, $d_{\mathcal{F}}(A, B)$, between a pair of SPD matrices, A and B, is given by

$$d_{\mathcal{F}}^2(A, B) = \operatorname{tr}\left[\ln^2(A^{-1/2} B A^{-1/2}) \right] = \sum_i \ln^2(\sigma_i),$$

where (σ_i) is the sequence of generalized eigenvalues of the pencil (A, B). The Förstner metric satisfies the following important invariance properties:

$$(2.5) \qquad d_{\mathcal{F}}(A, B) = d_{\mathcal{F}}(A^{-1}, B^{-1}) \quad \text{and} \quad d_{\mathcal{F}}(A, B) = d_{\mathcal{F}}(MAM^\top, MBM^\top)$$

for any nonsingular matrix M. Moreover, $d_{\mathcal{F}}$ treats under- and overapproximations similarly in the sense that $d_{\mathcal{F}}(\Gamma_{\text{pos}}, \alpha \widehat{\Gamma}_{\text{pos}}) \to \infty$ as $\alpha \to 0$ and as $\alpha \to \infty$.[2] Note that the metric induced by the Frobenius norm does not satisfy any of the aforementioned invariance properties. In addition, it penalizes under- and overestimation differently.

We will show that our posterior covariance matrix approximation is optimal not only in terms of the Förstner metric but also in terms of the following more general class, $\mathcal{L}$, of loss functions for SPD matrices.

DEFINITION 2.1 (loss functions). *The class $\mathcal{L}$ is defined as the collection of functions of the form*

$$(2.6) \qquad L(A, B) = \sum_{i=1}^{n} f(\sigma_i),$$

where A and B are SPD matrices, (σ_i) are the generalized eigenvalues of the pencil (A, B), and

$$(2.7) \qquad f \in \mathcal{U} = \{g \in \mathcal{C}^1(\mathbb{R}_+) : g'(x)(1 - x) < 0 \text{ for } x \neq 1, \text{ and } \lim_{x \to \infty} g(x) = \infty\}.$$

Elements of $\mathcal{U}$ are differentiable real-valued functions defined on the positive axis that decrease on $x < 1$, increase on $x > 1$, and tend to infinity as $x \to \infty$. The squared

---

[2]This behavior is shared by Stein's loss function, which has been proposed to assess estimates of a covariance matrix [45]. Stein's loss function is just the K-L distance between two Gaussian distributions with the same mean (see (A.10)), but it is not a metric for SPD matrices.

Förstner metric belongs to the class of loss functions defined by (2.6), whereas the distance induced by the Frobenius norm does not.

Lemma 2.2, whose proof can be found in Appendix A, justifies the importance of the class $\mathcal{L}$. In particular, it shows that optimality of the covariance matrix approximation with respect to any loss function in $\mathcal{L}$ leads to an optimal approximation of the posterior distribution using a Gaussian (with the same mean) in terms of other familiar criteria used to compare probability measures, such as the Hellinger distance and the K-L divergence [70]. More precisely, we have the following result.

LEMMA 2.2 (equivalence of approximations). *If $L \in \mathcal{L}$, then a matrix $\widehat{\Gamma}_{\mathrm{pos}} \in \mathcal{M}_r$ minimizes the Hellinger distance and the K-L divergence between $\mathcal{N}(\mu_{\mathrm{pos}}(y), \Gamma_{\mathrm{pos}})$ and the approximation $\mathcal{N}(\mu_{\mathrm{pos}}(y), \widehat{\Gamma}_{\mathrm{pos}})$ iff it minimizes $L(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}})$.*

Remark 1. We note that neither the Hellinger distance nor the K-L divergence between the distributions $\mathcal{N}(\mu_{\mathrm{pos}}(y), \Gamma_{\mathrm{pos}})$ and $\mathcal{N}(\mu_{\mathrm{pos}}(y), \widehat{\Gamma}_{\mathrm{pos}})$ depends on the data y. Optimality in distribution does not necessarily hold when the posterior means are different.

2.3. Optimality results. We are now in a position to state one of the main results of the paper. For a proof see Appendix A.

THEOREM 2.3 (optimal posterior covariance approximation). *Let $(\delta_i^2, \widehat{w}_i)$ be the generalized eigenvalue-eigenvector pairs of the pencil*

$$(2.8) \qquad\qquad (H, \Gamma_{\mathrm{pr}}^{-1}),$$

with the ordering $\delta_i^2 \geq \delta_{i+1}^2$, and $H = G^\top \Gamma_{\mathrm{obs}}^{-1} G$ as in (2.3). Let L be a loss function in the class $\mathcal{L}$ defined in (2.6). Then the following hold:

(i) *A minimizer, $\widehat{\Gamma}_{\mathrm{pos}}$, of the loss, L, between Γ_{pos} and an element of $\mathcal{M}_r$ is given by*

$$(2.9) \qquad \widehat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^\top, \quad KK^\top = \sum_{i=1}^{r} \delta_i^2 \left(1 + \delta_i^2\right)^{-1} \widehat{w}_i \widehat{w}_i^\top.$$

The corresponding minimum loss is given by

$$(2.10) \qquad L(\widehat{\Gamma}_{\mathrm{pos}}, \Gamma_{\mathrm{pos}}) = f(1)\, r + \sum_{i>r} f(1/(1 + \delta_i^2)).$$

(ii) *The minimizer (2.9) is unique if the first r eigenvalues of $(H, \Gamma_{\mathrm{pr}}^{-1})$ are different.*

Theorem 2.3 provides a way to compute the best approximation of Γ_{pos} by matrices in $\mathcal{M}_r$: it is just a matter of computing the eigenpairs corresponding to the decreasing sequence of eigenvalues of the pencil $(H, \Gamma_{\mathrm{pr}}^{-1})$ until a stopping criterion is satisfied. This criterion can be based on the minimum loss (2.10). Notice that the minimum loss is a function of the generalized eigenvalues $(\delta_i^2)_{i \geq r}$ that have not been computed. This is quite common in numerical linear algebra (e.g., error in the truncated SVD [26, 32]). However, since the eigenvalues (δ_i^2) are computed in a decreasing order, the minimum loss can be easily bounded.

The generalized eigenvectors, $\widehat{w}_i$, are orthogonal with respect to the inner product induced by the prior precision matrix, and they maximize the Rayleigh ratio,

$$\widehat{\mathcal{R}}(z) = \frac{z^\top H z}{z^\top \Gamma_{\mathrm{pr}}^{-1} z},$$

over subspaces of the form $\widehat{\mathcal{W}}_i = \mathrm{span}^{\perp}(\widehat{w}_j)_{j<i}$. Intuitively, the vectors $\widehat{w}_i$ associated with generalized eigenvalues greater than one correspond to directions in the parameter space (or subspaces thereof) where the curvature of the log-posterior density is constrained more by the log-likelihood than by the prior.

2.4. Computing eigenpairs of $(H, \Gamma_{\mathrm{pr}}^{-1})$. If a square root factorization of the prior covariance matrix $\Gamma_{\mathrm{pr}} = S_{\mathrm{pr}} S_{\mathrm{pr}}^{\top}$ is available, then the Hermitian generalized eigenvalue problem can be reduced to a standard one: find the eigenpairs, (δ_i^2, w_i), of $S_{\mathrm{pr}}^{\top} H S_{\mathrm{pr}}$, and transform the resulting eigenvectors according to $w_i \mapsto S_{\mathrm{pr}} w_i$ [4, section 5.2]. An analogous transformation is also possible when a square root factorization of $\Gamma_{\mathrm{pr}}^{-1}$ is available. Notice that only the actions of S_{pr} and $S_{\mathrm{pr}}^{\top}$ on a vector are required. For instance, evaluating the action of S_{pr} might involve the solution of an elliptic PDE [57]. There are numerous examples of priors for which a decomposition $\Gamma_{\mathrm{pr}} = S_{\mathrm{pr}} S_{\mathrm{pr}}^{\top}$ is readily available, e.g., [82, 24, 57, 84, 79]. Either direct methods or, more often, matrix-free algorithms (e.g., Lanczos iteration or its block version [50, 66, 23, 33]) can be used to solve the standard Hermitian eigenvalue problem [4, section 4]. Reference implementations of these algorithms are available in ARPACK [52]. We note that the Lanczos iteration comes with a rich literature on error analysis (e.g., [49, 65, 67, 46, 73, 32]). Alternatively, one can use randomized methods [35], which offer the advantage of parallelism (asynchronous computations) and robustness over standard Lanczos methods [6]. If a square root factorization of Γ_{pr} is not available, but it is possible to solve linear systems with $\Gamma_{\mathrm{pr}}^{-1}$, we can use a Lanczos method for generalized Hermitian eigenvalue problems [4, section 5.5], where a Krylov basis orthogonal with respect to the inner product induced by $\Gamma_{\mathrm{pr}}^{-1}$ is maintained. Again, ARPACK provides an efficient implementation of these solvers. When accurately solving linear systems with $\Gamma_{\mathrm{pr}}^{-1}$ is a difficult task, we refer the reader to alternative algorithms proposed in [74] and [34].

Remark 2. If a factorization $\Gamma_{\mathrm{pr}} = S_{\mathrm{pr}} S_{\mathrm{pr}}^{\top}$ is available, then it is straightforward to obtain an expression for a nonsymmetric square root of the optimal approximation of Γ_{pos} (2.9) as in [9],

$$(2.11) \qquad \widehat{S}_{\mathrm{pos}} = S_{\mathrm{pr}} \left(\sum_{i=1}^{r} \left[\left(1 + \delta_i^2\right)^{-1/2} - 1 \right] w_i w_i^{\top} + I \right),$$

such that $\widehat{\Gamma}_{\mathrm{pos}} = \widehat{S}_{\mathrm{pos}} \widehat{S}_{\mathrm{pos}}^{\top}$ and $w_i = S_{\mathrm{pr}}^{-1} \widehat{w}_i$. This expression can be used to efficiently sample from the approximate posterior distribution $\mathcal{N}(\mu_{\mathrm{pos}}(y), \widehat{\Gamma}_{\mathrm{pos}})$ (e.g., [28, 60]). Alternative techniques for sampling from high-dimensional Gaussian distributions can be found, for instance, in [71, 30].

3. Properties of the optimal covariance approximation. Now we discuss several implications of the optimal approximation of Γ_{pos} introduced in the previous section. We start by describing the relationship between this approximation and the directions of greatest relative reduction of prior variance. Then we interpret the covariance approximation as the result of projecting the likelihood function onto a "data-informed" subspace. Finally, we contrast the present approach with several other approximation strategies: using the Frobenius norm as a loss function for the covariance matrix approximation, or developing low-rank approximations based on prior or Hessian information alone. We conclude by drawing the connections with the BFGS Kalman filter update.

3.1. Interpretation of the eigendirections. Thanks to the particular structure of loss functions in $\mathcal{L}$, the problem of approximating Γ_{pos} is equivalent

to that of approximating Γ_{pos}^{-1}. Yet the form of the optimal approximation of Γ_{pos}^{-1} is important, as it explicitly describes the directions that control the ratio of posterior to prior variance. The following corollary to Theorem 2.3 characterizes these directions. The proof is in Appendix A.

COROLLARY 3.1 (optimal posterior precision approximation). *Let* $(\delta_i^2, \widehat{w}_i)$ *and* $L \in \mathcal{L}$ *be defined as in Theorem* 2.3. *Then the following hold:*

(i) *A minimizer of* $L(B, \Gamma_{\text{pos}}^{-1})$ *for*

$$(3.1) \qquad B \in \mathcal{M}_r^{-1} := \left\{ \Gamma_{\text{pr}}^{-1} + JJ^\top : \text{rank}(J) \leq r \right\}$$

is given by

$$(3.2) \qquad \widehat{\Gamma}_{\text{pos}}^{-1} = \Gamma_{\text{pr}}^{-1} + UU^\top, \quad UU^\top = \sum_{i=1}^r \delta_i^2 \widetilde{w}_i \widetilde{w}_i^\top, \quad \widetilde{w}_i = \Gamma_{\text{pr}}^{-1} \widehat{w}_i.$$

The minimizer (3.2) *is unique if the first* r *eigenvalues of* $(H, \Gamma_{\text{pr}}^{-1})$ *are different.*

(ii) *The optimal posterior precision matrix* (3.2) *is precisely the inverse of the optimal posterior covariance matrix* (2.9).

(iii) *The vectors* $\widetilde{w}_i$ *are generalized eigenvectors of the pencil* $(\Gamma_{\text{pos}}, \Gamma_{\text{pr}})$:

$$(3.3) \qquad \Gamma_{\text{pos}} \widetilde{w}_i = \frac{1}{1 + \delta_i^2} \Gamma_{\text{pr}} \widetilde{w}_i.$$

Note that the definition of the class $\mathcal{M}_r^{-1}$ is analogous to that of $\mathcal{M}_r$. Indeed, Lemma A.2 in Appendix A defines a bijection between these two classes.

The vectors $\widetilde{w}_i = \Gamma_{\text{pr}}^{-1} \widehat{w}_i$ are orthogonal with respect to the inner product defined by Γ_{pr}. By (3.3), we also know that $\widetilde{w}_i$ minimizes the generalized Rayleigh quotient,

$$(3.4) \qquad \mathcal{R}(z) = \frac{z^\top \Gamma_{\text{pos}} z}{z^\top \Gamma_{\text{pr}} z} = \frac{\mathbb{V}\text{ar}(z^\top x \mid y)}{\mathbb{V}\text{ar}(z^\top x)},$$

over subspaces of the form $\widetilde{\mathcal{W}}_i = \text{span}^\perp(\widetilde{w}_j)_{j<i}$. This Rayleigh quotient is precisely the *ratio of posterior to prior variance* along a particular direction, z, in the parameter space. The smallest values that $\mathcal{R}$ can take over the subspaces $\widetilde{\mathcal{W}}_i$ are exactly the smallest generalized eigenvalues of $(\Gamma_{\text{pos}}, \Gamma_{\text{pr}})$. In particular, the data are most informative along the first r eigenvectors $\widetilde{w}_i$ and, since

$$(3.5) \qquad \mathcal{R}(\widetilde{w}_i) = \frac{\mathbb{V}\text{ar}(\widetilde{w}_i^\top x \mid y)}{\mathbb{V}\text{ar}(\widetilde{w}_i^\top x)} = \frac{1}{1 + \delta_i^2},$$

the posterior variance is smaller than the prior variance by a factor of $(1 + \delta_i^2)^{-1}$. In the span of the other eigenvectors, $(\widetilde{w}_i)_{i>r}$, the data are not as informative. Hence, $(\widetilde{w}_i)$ are the directions along which the ratio of posterior to prior variance is minimized. Furthermore, a simple computation shows that these directions also maximize the relative difference between prior and posterior variance normalized by the prior variance. Indeed, if the directions $(\widetilde{w}_i)$ minimize (3.4), then they must also maximize $1 - \mathcal{R}(z)$, leading to

$$(3.6) \qquad 1 - \mathcal{R}(\widetilde{w}_i) = \frac{\mathbb{V}\text{ar}(\widetilde{w}_i^\top x) - \mathbb{V}\text{ar}(\widetilde{w}_i^\top x \mid y)}{\mathbb{V}\text{ar}(\widetilde{w}_i^\top x)} = \frac{\delta_i^2}{1 + \delta_i^2}.$$

3.2. Optimal projector. Since the data are most informative on a subspace of the parameter space, it should be possible to reduce the *effective dimension* of the inference problem in a manner that is consistent with the posterior approximation. This is essentially the content of the following corollary, which follows by a simple computation.

COROLLARY 3.2 (optimal projector). *Let* $\widehat{\Gamma}_{\mathrm{pos}}$ *and the vectors* $(\widehat{w}_i, \widetilde{w}_i)$ *be defined as in Theorems 2.3 and 3.1. Consider the reduced forward operator* $\widehat{G}_r = G \circ P_r$, *where* P_r *is the oblique projector (i.e.,* $P_r^2 = P_r$ *but* $P_t^\top \neq P_r$*):*

$$(3.7) \qquad P_r = \sum_{i=1}^{r} \widehat{w}_i \widetilde{w}_i^\top.$$

Then $\widehat{\Gamma}_{\mathrm{pos}}$ *is precisely the posterior covariance matrix corresponding to the Bayesian linear model:*

$$(3.8) \qquad y \mid x \sim \mathcal{N}(\widehat{G}_r\, x,\, \Gamma_{\mathrm{obs}}), \quad x \sim \mathcal{N}(0,\, \Gamma_{\mathrm{pr}}).$$

The projected Gaussian linear model (3.8) reveals the intrinsic dimensionality of the inference problem. The introduction of the optimal projector (3.7) is also useful in the context of dimensionality reduction for nonlinear inverse problems. In this case a particularly simple and effective approximation of the posterior density, $\pi_{\mathrm{pos}}(x|y)$, is of the form $\widehat{\pi}_{\mathrm{pos}}(x|y) \propto \pi(y; P_r\, x)\, \pi_{\mathrm{pr}}(x)$, where π_{pr} is the prior density and $\pi(y; P_r\, x)$ is the density corresponding to the likelihood function with parameters constrained by the projector. The range of the projector can be determined by combining locally optimal data-informed subspaces from high-density regions in the support of the posterior distribution. This approximation is the subject of a related paper [22].

Returning to the linear inverse problem, notice also that the posterior mean of the projected model (3.8) might be used as an efficient approximation of the exact posterior mean. We will show in section 4 that this posterior mean approximation in fact minimizes the Bayes risk for a weighted squared-error loss among all low-rank linear functions of the data.

3.3. Comparison with optimality in Frobenius norm. Thus far our optimality results for the approximation of Γ_{pos} have been restricted to the class of loss functions $\mathcal{L}$ given in Definition 2.1. However, it is also interesting to investigate optimality in the metric defined by the Frobenius norm. Given any two matrices A and B of the same size, the Frobenius distance between them is defined as $\|A - B\|$, where $\|\cdot\|$ is the Frobenius norm. Note that the Frobenius distance does not exploit the structure of the positive definite cone of symmetric matrices. The matrix $\widehat{\Gamma}_{\mathrm{pos}} \in \mathcal{M}_r$ that minimizes the Frobenius distance from the exact posterior covariance matrix is given by

$$(3.9) \qquad \widehat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^\top, \quad KK^\top = \sum_{i=1}^{r} \lambda_i\, u_i u_i^\top,$$

where (u_i) are the directions corresponding to the r largest eigenvalues of $\Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pos}}$. This result can be very different from the optimal approximation given in Theorem 2.3. In particular, the directions (u_i) are solutions of the eigenvalue problem

$$(3.10) \qquad \Gamma_{\mathrm{pr}}\, G^\top \Gamma_{\mathrm{y}}^{-1} G\, \Gamma_{\mathrm{pr}}\, u = \lambda u,$$

which maximize

$$(3.11) \qquad u^\top (\Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pos}})u = \mathbb{V}\mathrm{ar}(u^\top x) - \mathbb{V}\mathrm{ar}(u^\top x \mid y).$$

That is, while optimality in the Förstner metric identifies directions that maximize the *relative* difference between prior and posterior variance, the Frobenius distance favors directions that maximize only the absolute value of this difference. There are many reasons to prefer the former. For instance, data might be informative along directions of low prior variance (perhaps due to inadequacies in prior modeling); a covariance matrix approximation that is optimal in Frobenius distance may ignore updates in these directions entirely. Also, if parameters of interest (i.e., components of x) have differing units of measurement, relative variance reduction provides a unit-independent way of judging the quality of a posterior approximation; this notion follows naturally from the second invariance property of $d_{\mathcal{F}}$ in (2.5). From a computational perspective, solving the eigenvalue problem (3.10) is quite expensive compared to finding the generalized eigenpairs of the pencil $(H, \Gamma_{\mathrm{pr}}^{-1})$. Finally, optimality in the Frobenius distance for an approximation of Γ_{pos} does not yield an optimality statement for the corresponding approximation of the posterior distribution, as shown in Lemma 2.2 for loss functions in $\mathcal{L}$.

3.4. Suboptimal posterior covariance approximations.

3.4.1. Hessian-based and prior-based reduction schemes. The posterior approximation described by Theorem 2.3 uses both Hessian and prior information. It is instructive to consider approximations of the linear Bayesian inverse problem that rely only on one or the other. As we will illustrate numerically in section 5.1, these approximations can be viewed as natural limiting cases of our approach. They are also closely related to previous efforts in dimensionality reduction that propose only Hessian-based [55] or prior-based [61] reductions. In contrast with these previous efforts, here we will consider versions of Hessian- and prior-based reductions that do not discard prior information in the remaining directions. In other words, we will discuss posterior covariance approximations that remain in the form of (2.4)—i.e., updating the prior covariance only in r directions.

A Hessian-based reduction scheme updates Γ_{pr} in directions where the data have greatest influence in an absolute sense (i.e., not relative to the prior). This involves approximating the negative log-likelihood Hessian (2.3) with a low-rank decomposition as follows: let (s_i^2, v_i) be the eigenvalue-eigenvector pairs of H with the ordering $s_i^2 \geq s_{i+1}^2$. Then a best low-rank approximation of H in the Frobenius norm is given by

$$H \approx \sum_{i=1}^{r} s_i^2 \, v_i v_i^\top = V_r S_r V_r^\top,$$

where v_i is the ith column of V_r and $S_r = \mathrm{diag}\{s_1^2, \ldots, s_r^2\}$. Using Woodbury's identity we then obtain an approximation of Γ_{pos} as a low-rank negative semidefinite update of Γ_{pr}:

$$(3.12) \qquad \Gamma_{\mathrm{pos}} \approx \left(V_r S_r V_r^\top + \Gamma_{\mathrm{pr}}^{-1} \right)^{-1} = \Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pr}} V_r \left(S_r^{-1} + V_r^\top \Gamma_{\mathrm{pr}} V_r \right)^{-1} V_r^\top \Gamma_{\mathrm{pr}}.$$

This approximation of the posterior covariance matrix belongs to the class $\mathcal{M}_r$. Thus, Hessian-based reductions are in general suboptimal when compared to the optimal

approximations defined in Theorem 2.3. Note that an equivalent way to obtain (3.12) is to use a reduced forward operator of the form $\widehat{G} = G \circ V_r V_r^\top$, which is the composition of the original forward operator with a projector onto the leading eigenspace of H. In general, the projector $P_r = V_r V_r^\top$ is different from the optimal projector defined in Corollary 3.2 and is thus suboptimal.

To achieve prior-based reductions, on the other hand, we restrict the Bayesian inference problem to directions in the parameter space that explain most of the prior variance. More precisely, we look for a rank-r orthogonal projector, P_r, that minimizes the mean squared-error defined as

$$(3.13) \qquad \mathcal{E}\left(P_r\right) = \mathbb{E}\left(\|x - P_r x\|^2\right),$$

where the expectation is taken over the prior distribution (assumed to have zero mean) and $\|\cdot\|$ is the standard Euclidean norm [44]. Let (t_i^2, u_i) be the eigenvalue-eigenvector pairs of Γ_{pr} ordered as $t_i^2 \geq t_{i+1}^2$. Then a minimizer of (3.13) is given by a projector, P_r, onto the leading eigenspace of Γ_{pr} defined as $P_r = \sum_{i=1}^r u_i u_i^\top = U_r U_r^\top$, where u_i is the ith column of U_r. The actual approximation of the linear inverse problem consists of using the projected forward operator, $\widehat{G} = G \circ U_r U_r^\top$. By direct comparison with the optimal projector defined in Corollary 3.2, we see that prior-based reductions are suboptimal in general. Also in this case, the posterior covariance matrix with the projected Gaussian model can be written as a negative semidefinite update of Γ_{pr}:

$$\Gamma_{\mathrm{pos}} \approx \Gamma_{\mathrm{pr}} - U_r T_r [\,(U_r^\top H U_r)^{-1} + T_r\,]^{-1} T_r U_r^\top,$$

where $T_r = \mathrm{diag}\{t_1^2, \ldots, t_r^2\}$. The double matrix inversion makes this low-rank update computationally challenging to implement. It is also not optimal, as shown in Theorem 2.3.

To summarize, the Hessian and prior-based dimensionality reduction techniques are both suboptimal. These methods do not take into account the interactions between the dominant directions of H and those of Γ_{pr}, nor the relative importance of these quantities. Accounting for such interaction is a key feature of the optimal covariance approximation described in Theorem 2.3. Section 5.1 will illustrate conditions under which these interactions become essential.

3.4.2. Connections with the BFGS Kalman filter. The linear Bayesian inverse problem analyzed in this paper can be interpreted as the analysis step of a linear Bayesian filtering problem [27]. If the prior distribution corresponds to the forecast distribution at some time t, the posterior coincides with the so-called analysis distribution. In the linear case, with Gaussian process noise and observational errors, both of these distributions are Gaussian. The Kalman filter is a Bayesian solution to this filtering problem [48]. In [2] the authors propose a computationally feasible way to implement (and approximate) this solution in large-scale systems. The key observation is that when solving an SPD linear system of the form $Ax = b$ by means of BFGS or limited memory BFGS (L-BFGS [59]), one typically obtains an approximation of A^{-1} for free. This approximation can be written as a low-rank correction of an arbitrary positive definite initial approximation matrix A_0^{-1}. The matrix A_0^{-1} can be, for instance, the scaled identity. Notice that the approximation of A^{-1} given by L-BFGS is full rank and positive definite. This approximation is in principle convergent as the storage limit of L-BFGS increases [64]. An L-BFGS approximation of A is also possible [83].

There are many ways to exploit this property of the L-BFGS method. For example, in [2] the posterior covariance is written as a low-rank update of the prior

covariance matrix, $\Gamma_{\text{pos}} = \Gamma_{\text{pr}} - \Gamma_{\text{pr}}G^\top \Gamma_{\text{y}}^{-1} G\Gamma_{\text{pr}}$, where $\Gamma_{\text{y}} = \Gamma_{\text{obs}} + G\Gamma_{\text{pr}}G^\top$, and Γ_{y}^{-1} itself is approximated using the L-BFGS method. Since this approximation of Γ_{y} is full rank, however, this approach does not exploit potential low-dimensional structure of the inverse problem. Alternatively, one can obtain an L-BFGS approximation of Γ_{pos} when solving the linear system $\Gamma_{\text{pos}}^{-1} x = G^\top \Gamma_{\text{obs}}^{-1} y$ for the posterior mean $\mu_{\text{pos}}(y)$ [3]. If one uses the prior covariance matrix as an initial approximation matrix, A_0^{-1}, then the resulting L-BFGS approximation of Γ_{pos} can be written as a low-rank update of Γ_{pr}. This approximation format is similar to the one discussed in [28] and advocated in this paper. However, the approach of [3] (or its ensemble version [76]) does not correspond to any known optimal approximation of the posterior covariance matrix, nor does it lead to any optimality statement between the corresponding probability distributions. This is an important contrast with the present approach, which we will revisit numerically in section 5.1.

4. Optimal approximation of the posterior mean. In this section, we develop and characterize fast approximations of the posterior mean that can be used, for instance, to accelerate repeated inversion with multiple data sets. Note that we are not proposing alternatives to the efficient computation of the posterior mean for a single realization of the data. This task can already be accomplished with current state-of-the-art iterative solvers for regularized least-squares problems [42, 69, 1, 62, 5]. Instead, we are interested in constructing *statistically optimal approximations*[3] of the posterior mean as linear functions of the data. That is, we seek a matrix A, from an approximation class to be defined shortly, such that the posterior mean can be approximated as $\mu_{\text{pos}}(y) \approx A y$. We will investigate different approximation classes for A; in particular, we will only consider approximation classes for which applying A to a vector y is relatively inexpensive. Computing such a matrix A is more expensive than solving a single linear system associated with the posterior precision matrix to determine the posterior mean. However, once A is computed, it can be applied inexpensively to any realization of the data.[4] Our approach is therefore justified when the posterior mean must be evaluated for multiple instances of the data. This approach can thus be viewed as an offline-online strategy, where a more costly but data-independent offline calculation is followed by fast online evaluations. Moreover, we will show that these approximations can be obtained from an optimal approximation of the posterior covariance matrix (cf. Theorem 2.3) with minimal additional cost. Hence, if one is interested in *both* the posterior mean and covariance matrix (as is often the case in the Bayesian approach to inverse problems), then the approximation formulas we propose can be more efficient than standard approaches even for a single realization of the data.

4.1. Optimality results. For the Bayesian linear model defined in (2.1), the posterior mode is equal to the posterior mean, $\mu_{\text{pos}}(y) = \mathbb{E}(x|y)$, which is in turn the minimizer of the Bayes risk for squared-error loss [51, 58]. We first review this fact and establish some basic notation. Let S be an SPD matrix and let

$$L(\delta(y), x) = (x - \delta(y))^\top S (x - \delta(y)) = \|x - \delta(y)\|_S^2$$

be the loss incurred by the estimator $\delta(y)$ of x. The Bayes risk, $R(\delta(y), x)$, of $\delta(y)$ is defined as the average loss over the joint distribution of x and y [14, 51]: $R(\delta(y), x) = \mathbb{E}(L(\delta(y), x))$. Since

----

[3] We will precisely define this notion of optimality in section 4.1.
[4] In particular, applying A is much cheaper than solving a linear system.

$$(4.1) \qquad R(\delta(y), x) = \mathbb{E}\left(\| \delta(y) - \mu_{\mathrm{pos}}(y) \|_S^2\right) + \mathbb{E}\left(\|\mu_{\mathrm{pos}}(y) - x\|_S^2\right),$$

it follows that $\delta(y) = \mu_{\mathrm{pos}}(y)$ minimizes the Bayes risk over all estimators of x.

To study approximations of $\mu_{\mathrm{pos}}(y)$, we use the squared-error loss function defined by the Mahalanobis distance [15] induced by $\Gamma_{\mathrm{pos}}^{-1}$: $L(\delta(y), x) = \|\delta(y) - x\|_{\Gamma_{\mathrm{pos}}^{-1}}^2$. This loss function accounts for the geometry induced by the posterior measure on the parameter space, penalizing errors in the approximation of $\mu_{\mathrm{pos}}(y)$ more strongly in directions of lower posterior variance.

Under the assumption of zero prior mean, $\mu_{\mathrm{pos}}(y)$ is a linear function of the data. Hence we seek approximations of $\mu_{\mathrm{pos}}(y)$ of the form Ay, where A is a matrix in a class to be defined. Our goal is to obtain fast posterior mean approximations that can be applied repeatedly to multiple realizations of y. We consider two classes of approximation matrices:

$$(4.2)$$
$$\mathcal{A}_r := \{A : \mathrm{rank}(A) \le r\} \quad \text{and} \quad \widehat{\mathcal{A}}_r := \left\{A = (\Gamma_{\mathrm{pr}} - B)\, G^\top \Gamma_{\mathrm{obs}}^{-1} : \mathrm{rank}(B) \le r\right\}.$$

The class $\mathcal{A}_r$ consists of low-rank matrices; it is standard in the statistics literature [44]. The class $\widehat{\mathcal{A}}_r$, on the other hand, can be understood via comparison with (2.2); it simply replaces Γ_{pos} with a low-rank negative semidefinite update of Γ_{pr}. We shall henceforth use $\mathcal{A}$ to denote either of the two classes above.

Let $R_{\mathcal{A}}(Ay, x)$ be the Bayes risk of Ay subject to $A \in \mathcal{A}$. We may now restate our goal as follows: find a matrix, $A^* \in \mathcal{A}$, that minimizes the Bayes risk $R_{\mathcal{A}}(Ay, x)$. That is, find $A^* \in \mathcal{A}$ such that

$$(4.3) \qquad R_{\mathcal{A}}(A^*y, x) = \min_{A \in \mathcal{A}} \mathbb{E}(\|Ay - x\|_{\Gamma_{\mathrm{pos}}^{-1}}^2).$$

The following two theorems show that for either class of approximation matrices, $\mathcal{A}_r$ or $\widehat{\mathcal{A}}_r$, this problem admits a particularly simple analytical solution that exploits the structure of the optimal approximation of Γ_{pos}. The proofs of the theorems rely on a result developed independently by Sondermann [77] and Friedland and Torokhti [31] and are given in Appendix A. We also use the fact that $\mathbb{E}(\|\mu_{\mathrm{pos}}(y) - x\|_{\Gamma_{\mathrm{pos}}^{-1}}^2) = \ell$, where ℓ is the dimension of the parameter space.

THEOREM 4.1. *Let* $(\delta_i^2, \widehat{w}_i)$ *be defined as in Theorem 2.3 and let* $(\widehat{v}_i)$ *be generalized eigenvectors of the pencil* $(G\Gamma_{\mathrm{pr}}G^\top, \Gamma_{\mathrm{obs}})$ *associated with a nonincreasing sequence of eigenvalues, with the normalization* $\widehat{v}_i^\top \Gamma_{\mathrm{obs}} \widehat{v}_i = 1$. *Then the following hold:*

(i) *A solution of* (4.3) *for* $A \in \mathcal{A}_r$ *is given by*

$$(4.4) \qquad A^* = \sum_{i=1}^r \frac{\delta_i}{1 + \delta_i^2}\, \widehat{w}_i \widehat{v}_i^\top.$$

(ii) *The corresponding minimum Bayes risk over* $\mathcal{A}_r$ *is given by*

$$(4.5)$$
$$R_{\mathcal{A}_r}(A^*y, x) = \mathbb{E}\left(\|A^*y - \mu_{\mathrm{pos}}(y)\|_{\Gamma_{\mathrm{pos}}^{-1}}^2\right) + \mathbb{E}\left(\|\mu_{\mathrm{pos}}(y) - x\|_{\Gamma_{\mathrm{pos}}^{-1}}^2\right) = \sum_{i>r} \delta_i^2 + \ell.$$

Notice that the rank-r posterior mean approximation given by Theorem 4.1 coincides with the posterior mean of the projected linear Gaussian model defined in (3.8). Thus, applying this approximation to a new realization of the data requires only a *low-rank* matrix-vector product, a computationally trivial task. We define the quality

of a posterior mean approximation as the minimum Bayes risk (4.5). Notice, however, that for a given rank r of the approximation, (4.5) depends on the eigenvalues that have not yet been computed. Since (δ_i^2) are determined in order of decreasing magnitude, (4.5) can be easily bounded (cf. discussion after Theorem 2.3). The forthcoming minimum Bayes risk (4.8) can be bounded analogously.

Remark 3. Equation (4.4) can be interpreted as the truncated GSVD solution of a Tikhonov regularized linear inverse problem [38] (with unit regularization parameter). Hence, Theorem 4.1 also describes a Bayesian property of the (frequentist) truncated GSVD estimator.

Remark 4. If factorizations of the form $\Gamma_{\mathrm{pr}} = S_{\mathrm{pr}} S_{\mathrm{pr}}^\top$ and $\Gamma_{\mathrm{obs}} = S_{\mathrm{obs}} S_{\mathrm{obs}}^\top$ are readily available, then we can characterize the triplets $(\delta_i, \widehat{w}_i, \widehat{v}_i)$ from an SVD, $S_{\mathrm{obs}}^{-1} G S_{\mathrm{pr}} = \sum_{i \geq 1} \delta_i v_i w_i^\top$, of the matrix $S_{\mathrm{obs}}^{-1} G S_{\mathrm{pr}}$ with the transformations $\widehat{w}_i = S_{\mathrm{pr}} w_i$, $\widehat{v}_i = S_{\mathrm{obs}}^{-\top} v_i$ and the ordering $\delta_i \geq \delta_{i+1}$. In particular, the approximate posterior mean can be written as

$$(4.6) \qquad \mu_{\mathrm{pos}}^{(r)}(y) = S_{\mathrm{pr}} (S_{\mathrm{obs}}^{-1} G S_{\mathrm{pr}})_r^{\mathrm{Tikh}} S_{\mathrm{obs}}^{-1} y,$$

where $(S_{\mathrm{obs}}^{-1} G S_{\mathrm{pr}})_r^{\mathrm{Tikh}}$ is the best rank-r approximation to a Tikhonov regularized inverse.[5] That is, for any matrix A, $(A)_r$ is the best rank-r approximation of A (e.g., computed via SVD), whereas $(A)^{\mathrm{Tikh}} := (A^\top A + I)^{-1} A^\top$.

THEOREM 4.2. *Let $\widehat{\Gamma}_{\mathrm{pos}} \in \mathcal{M}_r$ be the optimal approximation of Γ_{pos} defined in Theorem 2.3. Then the following hold:*

(i) *A solution of* (4.3) *for $A \in \widehat{\mathcal{A}}_r$ is given by*

$$(4.7) \qquad\qquad\qquad \widehat{A}^* = \widehat{\Gamma}_{\mathrm{pos}} \, G^\top \Gamma_{\mathrm{obs}}^{-1}.$$

(ii) *The corresponding minimum Bayes risk over $\widehat{\mathcal{A}}_r$ is given by*

$$(4.8)$$
$$R_{\widehat{\mathcal{A}}_r}(\widehat{A}^* y, x) = \mathbb{E}\left(\left\| \widehat{A}^* y - \mu_{\mathrm{pos}}(y) \right\|_{\Gamma_{\mathrm{pos}}^{-1}}^2 \right) + \mathbb{E}\left(\| \mu_{\mathrm{pos}}(y) - x \|_{\Gamma_{\mathrm{pos}}^{-1}}^2 \right) = \sum_{i > r} \delta_i^6 + \ell.$$

Once the optimal approximation of Γ_{pos} described in Theorem 4.2 is computed, the cost of approximating $\mu_{\mathrm{pos}}(y)$ for a new realization of y is dominated by the adjoint and prior solves needed to apply $G^\top$ and Γ_{pr}, respectively. Combining the optimal approximations of $\mu_{\mathrm{pos}}(y)$ and Γ_{pos} given by Theorems 4.2 and 2.3, respectively, yields a complete approximation of the Gaussian posterior distribution. This is precisely the approximation adopted by the stochastic Newton MCMC method [60] to describe the Gaussian proposal distribution obtained from a local linearization of the forward operator of a nonlinear Bayesian inverse problem. Our results support the algorithmic choice of [60] with precise optimality statements.

It is worth noting that the two optimal Bayes risks, (4.5) and (4.8), depend on the parameter, r, that defines the dimension of the corresponding approximation classes $\mathcal{A}_r$ and $\widehat{\mathcal{A}}_r$. In the former case, r is the rank of the optimal matrix that defines the approximation. In the latter case, r is the rank of a negative update of Γ_{pr} that yields the posterior covariance matrix approximation. We shall thus refer to the estimator given by Theorem 4.1 as the low-rank approximation and to the estimator given by Theorem 4.2 as the low-rank *update* approximation. In both cases, we shall refer to r as the order of the approximation. A posterior mean approximation of order

---

[5]With unit regularization parameter and identity regularization operator [39].

r will be called *underresolved* if more than r generalized eigenvalues of the pencil $(H, \Gamma_{\mathrm{pr}}^{-1})$ are greater than one. If this is the case, then using the low-rank update approximation is not appropriate because the associated Bayes risk includes high-order powers of eigenvalues of $(H, \Gamma_{\mathrm{pr}}^{-1})$ that are greater than one. Thus, underresolved approximations tend to be more accurate when using the low-rank approximation. As we will show in section 5, this estimator is also less expensive to compute than its counterpart in Theorem 4.2. If, on the other hand, fewer than r eigenvalues of $(H, \Gamma_{\mathrm{pr}}^{-1})$ are greater than one, then the optimal low-rank *update* estimator will have better performance than the optimal low-rank estimator in the following sense:

$$0 < R_{\mathcal{A}_r}(A^* y, x) - R_{\widehat{\mathcal{A}}_r}(\widehat{A}^* y, x) = \sum_{i > r} \delta_i^2 \left(1 + \delta_i^2\right) \left(1 - \delta_i^2\right).$$

4.2. Connection with "priorconditioners". In this subsection we draw connections between the low-rank approximation of the posterior mean given in Theorem 4.1 and the regularized solution of a discrete ill-posed inverse problem, $y = Gx + \varepsilon$ (using the notation of this paper), as presented in [13, 10]. In [13, 10], the authors propose an early stopping regularization using iterative solvers preconditioned by prior statistical information on the parameter of interest, say, $x \sim \mathcal{N}(0, \Gamma_{\mathrm{pr}})$, and on the noise, say, $\varepsilon \sim \mathcal{N}(0, \Gamma_{\mathrm{obs}})$.[6] That is, if factorizations $\Gamma_{\mathrm{pr}} = S_{\mathrm{pr}} S_{\mathrm{pr}}^\top$ and $\Gamma_{\mathrm{obs}} = S_{\mathrm{obs}} S_{\mathrm{obs}}^\top$ are available, then [13] provides a solution, $x = S_{\mathrm{pr}} q$, to the inverse problem, where q comes from an early stopping regularization applied to the preconditioned linear system:

$$(4.9) \qquad S_{\mathrm{obs}}^{-1} G S_{\mathrm{pr}} q = S_{\mathrm{obs}}^{-1} y.$$

The iterative method of choice in this case is the conjugate gradient least squares (CGLS) algorithm [13, 36] (or GMRES for nonsymmetric square systems [11]) equipped with a proper stopping criterion (e.g., the discrepancy principle [47]). Although the approach of [13] is not exactly Bayesian, we can still use the optimality results of Theorem 4.1 to justify the observed good performance of this particular form of regularization. By a property of the CGLS algorithm, the rth iterate, $x^r = S_{\mathrm{pr}} q^r$, satisfies

$$(4.10) \qquad q^r = \operatorname*{argmin}_{q \in \mathcal{K}_r(\widehat{H}, \widehat{y})} \| S_{\mathrm{obs}}^{-1} y - S_{\mathrm{obs}}^{-1} G S_{\mathrm{pr}} q \|.$$

where $\mathcal{K}_r(\widehat{H}, \widehat{y})$ is the r-dimensional Krylov subspace associated with the matrix $\widehat{H} = S_{\mathrm{pr}}^\top H S_{\mathrm{pr}}$ and starting vector $\widehat{y} = S_{\mathrm{pr}}^\top G^\top \Gamma_{\mathrm{obs}}^{-1} y$. It was shown in [41] that the CGLS solution, at convergence, can be written as $x^* = S_{\mathrm{pr}} (S_{\mathrm{obs}}^{-1} G S_{\mathrm{pr}})^\dagger S_{\mathrm{obs}}^{-1} y$, where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudoinverse [63, 72]. To highlight the differences between the CGLS solution and (4.6), we assume that $\mathcal{K}_r(\widehat{H}, y) \approx \operatorname{ran}(W_r)$ for all r, where $W_r = [w_1 | \cdots | w_r]$, $\operatorname{ran}(A)$ denotes the range of a matrix A, and $\widehat{H} = \sum_i \delta_i^2 w_i w_i^\top$ is an SVD of $\widehat{H}$. Notice that the condition $\mathcal{K}_r(\widehat{H}, y) \approx \operatorname{ran}(W_r)$ is usually quite reasonable for moderate values of r. This practical observation is at the heart of the Lanczos iteration for symmetric eigenvalue problems [50]. With simple algebraic manipulations we conclude that

$$(4.11) \qquad x^r \approx S_{\mathrm{pr}} (S_{\mathrm{obs}}^{-1} G S_{\mathrm{pr}})_r^\dagger S_{\mathrm{obs}}^{-1} y.$$

___
[6]It suffices to consider a Gaussian approximation to the distribution of x and ε.

Recall from (4.6) that the optimal rank-r approximation of the posterior mean defined in Theorem 4.1 can be written as

$$(4.12) \qquad \mu_{\text{pos}}^{(r)}(y) = S_{\text{pr}}(S_{\text{obs}}^{-1}GS_{\text{pr}})_r^{\text{Tikh}} S_{\text{obs}}^{-1}y.$$

The only difference between (4.11) and (4.12) is the use of a Tikhonov-regularized inverse in (4.12) as opposed to a Moore–Penrose pseudoinverse. If $S_{\text{obs}}^{-1}GS_{\text{pr}} = \sum_{i\geq 1} \delta_i v_i w_i^\top$ is a reduced SVD of the matrix $S_{\text{obs}}^{-1}GS_{\text{pr}}$, then

$$(4.13) \qquad (S_{\text{obs}}^{-1}GS_{\text{pr}})_r^\dagger = \sum_{i\leq r} \frac{1}{\delta_i} w_i v_i^\top, \qquad (S_{\text{obs}}^{-1}GS_{\text{pr}})_r^{\text{Tikh}} = \sum_{i\leq r} \frac{\delta_i}{1+\delta_i^2} w_i v_i^\top.$$

These two matrices are *nearly* identical for values of r corresponding to δ_r^2 greater than one[7] (assuming the ordering $\delta_i^2 \geq \delta_{i+1}^2$). Beyond this regime, it might be convenient to stop the CGLS solver to obtain (4.11) (i.e., early stopping regularization). The similarity of these expressions is quite remarkable since (4.12) was derived as the minimizer of the optimization problem (4.3) with $\mathcal{A} = \mathcal{A}_r$. This informal argument may explain why *priorconditioners* perform so well in applications [12, 43]. Yet we remark that the goals of Theorem 4.1 and [13] are still quite different; [13] is concerned with preconditioning techniques for early stopping regularization of ill-posed inverse problems, whereas Theorem 4.1 is concerned with statistically optimal approximations of the posterior mean in the Bayesian framework.

ALGORITHM 1. Optimal *low-rank* approximation of the posterior mean.

INPUT: forward and adjoint models G, $G^\top$; prior and noise precisions Γ_{pr}^{-1}, Γ_{obs}^{-1}; approximation order $r \in \mathbb{N}$

OUTPUT: approximate posterior mean $\mu_{\text{pos}}^{(r)}(y)$

  1: Find the r leading generalized eigenvalue-eigenvector pairs $(\delta_i^2, \widehat{w}_i)$ of the pencil $(G^\top \Gamma_{\text{obs}}^{-1} G, \Gamma_{\text{pr}}^{-1})$
  2: Find the r leading generalized eigenvector pairs $(\widehat{v}_i)$ of the pencil $(G\Gamma_{\text{pr}}G^\top, \Gamma_{\text{obs}})$
  3: For each new realization of the data y, compute $\mu_{\text{pos}}^{(r)}(y) = \sum_{i=1}^r \delta_i(1+\delta_i^2)^{-1}\widehat{w}_i\widehat{v}_i^\top y$.

ALGORITHM 2. Optimal *low-rank update* approximation of the posterior mean.

INPUT: forward and adjoint models G, $G^\top$; prior and noise precisions Γ_{pr}^{-1}, Γ_{obs}^{-1}; approximation order $r \in \mathbb{N}$

OUTPUT: approximate posterior mean $\widehat{\mu}_{\text{pos}}^{(r)}(y)$

  1: Obtain $\widehat{\Gamma}_{\text{pos}}$ as described in Theorem 2.3.
  2: For each new realization of the data y, compute $\widehat{\mu}_{\text{pos}}^{(r)}(y) = \widehat{\Gamma}_{\text{pos}} G^\top \Gamma_{\text{obs}}^{-1} y$.

---

5. Numerical examples. Now we provide several numerical examples to illustrate the theory developed in the preceding sections. We start with a synthetic example to demonstrate various posterior covariance matrix approximations and continue with two more realistic linear inverse problems, where we also study posterior mean approximations.

---

[7]In section 5 we show that by the time we start including generalized eigenvalues $\delta_i^2 \approx 1$ in (4.4), the approximation of the posterior mean is usually already satisfactory. Intuitively, this means that all the directions in parameter space where the data are more informative than the prior have been considered.

5.1. Example 1: Hessian and prior with controlled spectra. We begin by investigating the approximation of Γ_{pos} as a negative semidefinite update of Γ_{pr}. We compare the optimal approximation obtained in Theorem 2.3 with the Hessian-, prior-, and BFGS-based reduction schemes discussed in section 3.4. The idea is to reveal differences between these approximations by exploring regimes where the data have differing impacts on the prior information. Since the directions defining the optimal update are the generalized eigenvectors of the pencil $(H, \Gamma_{\mathrm{pr}}^{-1})$, we shall also refer to this update as the *generalized* approximation.

To compare these approximation schemes, we start with a simple example with diagonal Hessian and prior covariance matrices: $G = I$, $\Gamma_{\mathrm{obs}} = \mathrm{diag}\{\sigma_i^2\}$, and $\Gamma_{\mathrm{pr}} = \mathrm{diag}\{\lambda_i^2\}$. Since the forward operator G is the identity, this problem can (loosely) be thought of as denoising a signal x. In this case, $H = \Gamma_{\mathrm{obs}}^{-1}$ and $\Gamma_{\mathrm{pos}} = \mathrm{diag}\{\lambda_i^2 \sigma_i^2/(\sigma_i^2 + \lambda_i^2)\}$. The ratios of posterior to prior variance in the canonical directions (e_i) are

$$\frac{\mathbb{V}\mathrm{ar}(e_i^\top x \mid y)}{\mathbb{V}\mathrm{ar}(e_i^\top x)} = \frac{1}{1 + \lambda_i^2/\sigma_i^2}.$$

We note that if the observation variances σ_i^2 are constant, $\sigma_i = \sigma$, then the directions of greatest variance reduction are those corresponding to the largest prior variance. Hence the prior distribution alone determines the most informed directions, and the prior-based reduction is as effective as the generalized one. On the other hand, if the prior variances λ_i^2 are constant, $\lambda_i = \lambda$, but the σ_i vary, then the directions of highest variance reduction are those corresponding to the smallest noise variance. This time the noise distribution alone determines the most important directions, and Hessian-based reduction is as effective as the generalized one. In the case of more general spectra, the important directions depend on the ratios λ_i^2/σ_i^2 and thus one has to use the information provided by both the noise and prior distributions. This is done naturally by the generalized reduction.

We now generalize this simple case by moving to full matrices H and Γ_{pr} with a variety of prescribed spectra. We assume that H and Γ_{pr} have SVDs of the form $H = U \Lambda U^\top$ and $\Gamma_{\mathrm{pr}} = V \widetilde{\Lambda} V^\top$, where $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $\widetilde{\Lambda} = \mathrm{diag}\{\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n\}$ with

$$\lambda_k = \lambda_0/k^\alpha + \tau \quad \text{and} \quad \widetilde{\lambda}_k = \widetilde{\lambda}_0/k^{\widetilde{\alpha}} + \widetilde{\tau}.$$

To consider many different cases, the orthogonal matrices U and V are randomly and independently generated uniformly over the orthogonal group [78], leading to different realizations of H and Γ_{pr}. In particular, U and V are computed with a QR decomposition of a square matrix of independent standard Gaussian entries using a Gram–Schmidt orthogonalization. (In this case, the standard Householder reflections cannot be used.)

Before discussing the results of the first experiment, we explain our implementation of BFGS-based reduction. We ran the BFGS optimizer with a dummy quadratic optimization target $\mathcal{J}(x) = \frac{1}{2} x^\top \Gamma_{\mathrm{pos}}^{-1} x$ and used Γ_{pr} as the initial approximation matrix for Γ_{pos}. Thus, the BFGS approximation of the posterior covariance matrix can be written as $\Gamma_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^\top$ for some rank-r matrix K. The rank-r update is constructed by running the BFGS optimizer for r steps from random initial conditions as shown in [3]. Note that in order to obtain results for sufficiently high-rank updates, we use BFGS rather than L-BFGS in our numerical examples. While [2, 3] in principle employ L-BFGS, the results in these papers use a number of optimization

steps roughly equal to the number of vectors stored in L-BFGS; our approach thus is comparable to [2, 3]. Nonetheless, some results for the highest-rank BFGS updates are not plotted in Figures 1 and 2, as the optimizer converged so close to the optimum that taking further steps resulted in numerical instabilities.

Figure 1 summarizes the results of the first experiment. The top row shows the prescribed spectra of H^{-1} (red) and Γ_{pr} (blue). The parameters describing the eigenvalues of Γ_{pr} are fixed to $\widetilde{\lambda}_0 = 1$, $\widetilde{\alpha} = 2$, and $\widetilde{\tau} = 10^{-6}$. The corresponding parameters for H are given by $\lambda_0 = 500$ and $\tau = 10^{-6}$ with $\alpha = 0.345$ (left), $\alpha = 0.690$ (middle), and $\alpha = 1.724$ (right). Thus, moving from the leftmost column to the rightmost column, the data become increasingly less informative. The second row in the figure shows the Förstner distance between Γ_{pos} and its approximation, $\widehat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^\top$, as a function of the rank of $KK^\top$ for 100 different realizations of H and Γ_{pr}. The third row shows, for each realization of $(H, \Gamma_{\mathrm{pr}})$ and for each fixed rank of $KK^\top$, the difference between the Förstner distance obtained with a prior-, Hessian-, or BFGS-based dimensionality reduction technique and the minimum distance obtained with the generalized approximation. All these differences are positive—a confirmation of Theorem 2.3. But Figure 1 also shows interesting patterns consistent with the observations made for the simple example above. When the spectrum of H is basically flat (left column), the directions along which the prior variance is reduced the most are likely to be those corresponding to the largest prior variances, and thus a prior-based reduction is almost as effective as the generalized one (as seen in the bottom two rows on the left). As we move to the third column, eigenvalues of H^{-1} increase more quickly. The data provide significant information only on a lower-dimensional subspace of the parameter space. In this case, it is crucial to combine this information with the directions in the parameter space along which the prior variance is the greatest. The generalized reduction technique successfully accomplishes this task, whereas the prior and Hessian reductions fail as they focus either on Γ_{pr} or H alone; the key is to combine the two. The BFGS update performs remarkably well across all three configurations of the Hessian spectrum, although it is clearly suboptimal compared to the generalized reduction.

In Figure 2 the situation is reversed and the results are symmetric to those of Figure 1. The spectrum of H (red) is now kept fixed with parameters $\lambda_0 = 500$, $\alpha = 1$, and $\tau = 10^{-9}$, while the spectrum of Γ_{pr} (blue) has parameters $\widetilde{\lambda}_0 = 1$ and $\widetilde{\tau} = 10^{-9}$ with decay rates increasing from left to right: $\widetilde{\alpha} = 0.552$ (left), $\widetilde{\alpha} = 1.103$ (middle), and $\widetilde{\alpha} = 2.759$ (right). In the first column, the spectrum of the prior is nearly flat. That is, the prior variance is almost equally spread along every direction in the parameter space. In this case, the eigenstructure of H determines the directions of greatest variance reduction, and the Hessian-based reduction is almost as effective as the generalized one. As we move toward the third column, the spectrum of Γ_{pr} decays more quickly. The prior variance is restricted to lower-dimensional subspaces of the parameter space. Mismatch between prior- and Hessian-dominated directions then leads to poor performance of both the prior- and Hessian-based reduction techniques. However, the generalized reduction performs well also in this more challenging case. The BFGS reduction is again empirically quite effective in most of the configurations that we consider. It is not always better than the prior- or Hessian-based techniques when the update rank is low, or when the prior spectrum decays slowly; for example, Hessian-based reduction is more accurate than BFGS across all ranks in the first column of Figure 2. But when either the prior covariance or the Hessian has quickly decaying spectra, the BFGS approach performs almost as well as the generalized
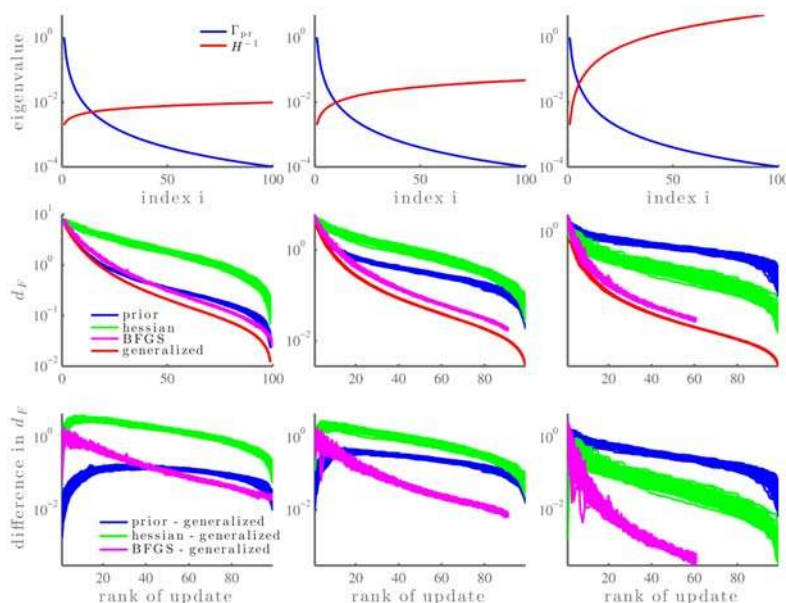
FIG. 1. *Top row: Eigenspectra of* Γ_{pr} *(blue) and* H^{-1} *(red) for three values for the decay rate of the eigenvalues of* H: $\alpha = 0.345$ *(left),* $\alpha = 0.690$ *(middle), and* $\alpha = 1.724$ *(right). Second row: Förstner distance between* Γ_{pos} *and its approximation versus the rank of the update for* 100 *realizations of* Γ_{pr} *and* H *using prior-based (blue), Hessian-based (green), BFGS-based (magenta), and optimal (red) updates. Bottom row: Differences of posterior covariance approximation error (measured with the Förstner metric) between the prior-based and optimal updates (blue), between the Hessian-based and optimal updates (green), and between the BFGS-based and optimal updates (magenta).*

reduction. Though this approach remains suboptimal, its approximation properties deserve further theoretical study.

5.2. Example 2: X-ray tomography. We consider a classical inverse problem of X-ray computed tomography (CT), where X-rays travel from sources to detectors through an object of interest. The intensities from multiple sources are measured at the detectors, and the goal is to reconstruct the density of the object. In this framework, we investigate the performance of the optimal mean and covariance matrix approximations presented in sections 2 and 4. This synthetic example is motivated by a real application: real-time X-ray imaging of logs that enter a saw mill for the purpose of automatic quality control. For instance, in the system commercialized by Bintec (www.bintec.fi), logs enter the X-ray system on a fast-moving conveyer belt and fast reconstructions are needed. The imaging setting (e.g., X-ray source and detector locations) and the priors are fixed; only the data changes from one log cross section to another. The basis for our posterior mean approximation can therefore be precomputed, and repeated inversions can be carried out quickly with direct matrix formulas.

We model the absorption of an X-ray along a line, ℓ_i, using Beer's law:

$$(5.1) \qquad I_d = I_s \exp\left(-\int_{\ell_i} x(s)ds \right),$$

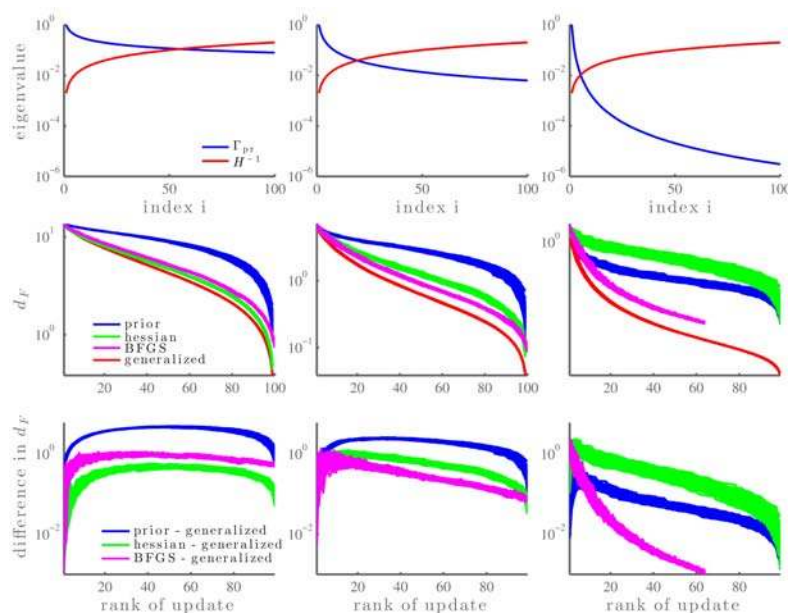where I_d and I_s are the intensities at the detector and at the source, respectively,

FIG. 2. *Analogous to Figure* 1 *but this time the spectrum of H is fixed, while that of* Γ_{pr} *has varying decay rates:* $\widetilde{\alpha} = 0.552$ *(left),* $\widetilde{\alpha} = 1.103$ *(middle), and* $\widetilde{\alpha} = 2.759$ *(right).*

and $x(s)$ is the density of the object at position s on the line ℓ_i. The computational domain is discretized into a grid and the density is assumed to be constant within each grid cell. The line integrals are approximated as

$$(5.2) \qquad \int_{\ell_i} x(s)ds \approx \sum_{j=1}^{\# \text{ of cells}} g_{ij} x_j,$$

where g_{ij} is the length of the intersection between line ℓ_i and cell j, and x_j is the unknown density in cell j. The vector of absorptions along m lines can then be approximated as

$$(5.3) \qquad I_d \approx I_s \exp\left(-Gx\right),$$

where I_d is the vector of m intensities at the detectors and $G = (g_{ij})$ is the $m \times n$ matrix of intersection lengths for each of the m lines. Even though the forward operator (5.3) is nonlinear, the inference problem can be recast in a linear fashion by taking the logarithm of both sides of (5.3). This leads to the following linear model for the inversion: $y = Gx + \epsilon$, where the measurement vector is $y = -\log(I_d/I_s)$ and the measurement errors are assumed to be independent and identically distributed Gaussian, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

The setup for the inference problem, borrowed from [40], is as follows. The rectangular domain is discretized with an $n \times n$ grid. The true object consists of three circular inclusions, each of uniform density, inside an annulus. Ten X-ray sources are positioned on one side of a circle, and each source sends a fan of 100 X-rays that are measured by detectors on the opposite side of the object. Here, the 10 sources are distributed evenly so that they form a total illumination angle of 90 degrees, resulting in a limited-angle CT problem. We use the exponential model
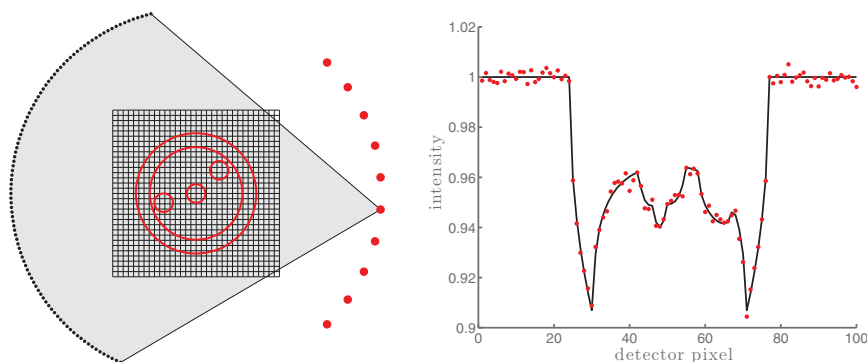
FIG. 3. *X-ray tomography problem. Left: Discretized domain, true object, sources (red dots), and detectors corresponding to one source (black dots). The fan transmitted by one source is illustrated in gray. The density of the object is* 0.006 *in the outer ring and* 0.004 *in the three inclusions; the background density is zero. Right: The true simulated intensity (black line) and noisy measurements (red dots) for one source.*

(5.1) to generate synthetic data in a discretization-independent fashion by computing the exact intersections between the rays and the circular inclusions in the domain. Gaussian noise with standard deviation $\sigma = 0.002$ is added to the simulated data. The imaging setup and data from one source are illustrated in Figure 3.

The unknown density is estimated on a 128×128 grid. Thus the discretized vector, x, has length 16,384, and direct computation of the posterior mean and the posterior covariance matrix, as well as generation of posterior samples, can be computationally nontrivial. To define the prior distribution, x is modeled as a discretized solution of a stochastic PDE of the form

$$(5.4) \qquad \gamma \left(\kappa^2 \mathcal{I} - \triangle \right) x(s) = \mathcal{W}(s), \qquad s \in \Omega,$$

where $\mathcal{W}$ is a white noise process, $\triangle$ is the Laplacian operator, and $\mathcal{I}$ is the identity operator. The solution of (5.4) is a Gaussian random field whose correlation length and variance are controlled by the free parameters κ and γ, respectively. A square root of the prior precision matrix of x (which is positive definite) can then be easily computed (see [57] for details). We use $\kappa = 10$ and $\gamma = \sqrt{800}$ in our simulations.

Our first task is to compute an optimal approximation of Γ_{pos} as a low-rank negative update of Γ_{pr} (cf. Theorem 2.3). Figure 4 (top row) shows the convergence of the approximate posterior variance as the rank of the update increases. The zero-rank update corresponds to Γ_{pr} (first column). For this formally 16,384-dimensional problem, a good approximation of the posterior variance is achieved with a rank 200 update; hence the data are informative only on a low-dimensional subspace. The quality of the covariance matrix approximation is also reflected in the structure of samples drawn from the approximate posterior distributions (bottom row). All five of these samples are drawn using the same random seed and the exact posterior mean, so that all the differences observed are due to the approximation of Γ_{pos}. Already with a rank 100 update, the small-scale features of the approximate posterior sample match those of the exact posterior sample. In applications, agreement in this "eyeball norm" is important. Of course, Theorem 2.3 also provides an exact formula for the error in the posterior covariance; this error is shown in the right panel of Figure 7 (blue curve).
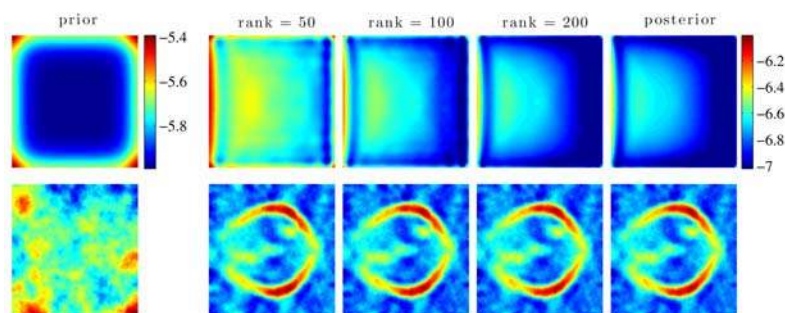
FIG. 4. *X-ray tomography problem. First column: Prior variance field, in log scale (top), and a sample drawn from the prior distribution (bottom). Second through last columns (left to right): Variance field, in log scale, of the approximate posterior as the rank of the update increases (top); samples from the corresponding approximate posterior distributions (bottom) assuming exact knowledge of the posterior mean.*

Our second task is to assess the performances of the two optimal posterior mean approximations given in section 4. We will use $\mu_{\mathrm{pos}}^{(r)}(y)$ to denote the low-rank approximation and $\widehat{\mu}_{\mathrm{pos}}^{(r)}(y)$ to denote the low-rank *update* approximation. Recall that both approximations are linear functions of the data y, given by $\mu_{\mathrm{pos}}^{(r)}(y) = A^* y$ with $A^* \in \mathcal{A}_r$ and $\widehat{\mu}_{\mathrm{pos}}^{(r)}(y) = \widehat{A}^* y$ with $\widehat{A}^* \in \widehat{\mathcal{A}}_r$, where the classes $\mathcal{A}_r$ and $\widehat{\mathcal{A}}_r$ are defined in (4.2). As in section 4, we shall use $\mathcal{A}$ to denote either of the two classes.

Figure 5 shows the normalized error $\|\mu(y) - \mu_{\mathrm{pos}}(y)\|_{\Gamma_{\mathrm{pos}}^{-1}} / \|\mu_{\mathrm{pos}}(y)\|_{\Gamma_{\mathrm{pos}}^{-1}}$ for different approximations $\mu(y)$ of the true posterior mean $\mu_{\mathrm{pos}}(y)$ and a fixed realization y of the data. The error is a function of the order r of the approximation class $\mathcal{A}$. Snapshots of $\mu(y)$ are shown along the two error curves. For reference, $\mu_{\mathrm{pos}}(y)$ is also shown at the top. We see that the errors decrease monotonically but that the low-rank approximation outperforms the low-rank update approximation for lower values of r. This is consistent with the discussion at the end of section 4; the crossing point of the error curves is also consistent with that analysis. In particular, we expect the low-rank update approximation to outperform the low-rank approximation only when the approximation starts to include generalized eigenvalues of the pencil $(H, \Gamma_{\mathrm{pr}}^{-1})$ that are less than one—i.e., once the approximations are no longer *underresolved*. This can be confirmed by comparing Figure 5 with the decay of the generalized eigenvalues of the pencil $(H, \Gamma_{\mathrm{pr}}^{-1})$ in the right panel of Figure 7 (blue curve).

On top of each snapshot in Figure 5, we show the *relative CPU time* required to compute the corresponding posterior mean approximation for each new realization of the data. The relative CPU time is defined as the time required to compute this approximation[8] divided by the time required to apply the posterior precision matrix to a vector. This latter operation is essential to computing the posterior mean via an iterative solver, such as a Krylov subspace method. These solvers are a standard choice for computing the posterior mean in large-scale inverse problems. Evaluating the ratio allows us to determine how many solver iterations could be performed with a computational cost roughly equal to that of approximating the posterior mean for a new realization of the data. Based on the reported times, a

---

[8]This timing does not include the computation of (4.4) or (4.7), which should be regarded as *offline* steps. Here we report the time necessary to apply the optimal linear function to any new realization of the data, i.e., the *online* cost.
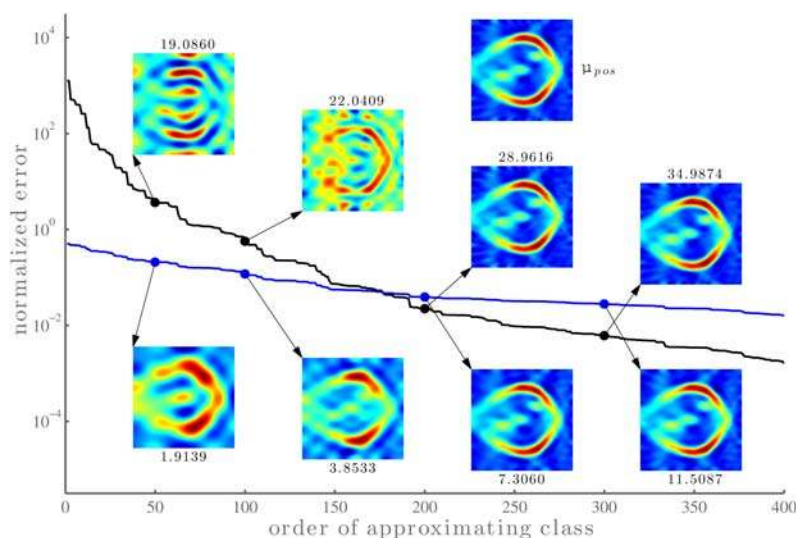
FIG. 5. *Limited-angle X-ray tomography: Comparison of the optimal posterior mean approximations, $\mu_{\mathrm{pos}}^{(r)}(y)$ (blue) and $\widehat{\mu}_{\mathrm{pos}}^{(r)}(y)$ (black) of $\mu_{\mathrm{pos}}(y)$ for a fixed realization of the data y, as a function of the order r of the approximating classes $\mathcal{A}_r$ and $\widehat{\mathcal{A}}_r$, respectively. The normalized error for an approximation $\mu(y)$ is defined as $\|\mu(y) - \mu_{\mathrm{pos}}(y)\|_{\Gamma_{\mathrm{pos}}^{-1}} / \|\mu_{\mathrm{pos}}(y)\|_{\Gamma_{\mathrm{pos}}^{-1}}$. The numbers above or below the snapshots indicate the relative CPU time of the corresponding mean approximation—i.e., the time required to compute the approximation divided by the time required to apply the posterior precision matrix to a vector.*

few observations can be made. First of all, as anticipated in section 4, computing $\mu_{\mathrm{pos}}^{(r)}(y)$ for any new realization of the data is faster than computing $\widehat{\mu}_{\mathrm{pos}}^{(r)}(y)$. Second, obtaining an accurate posterior mean approximation requires roughly $r = 200$, and the relative CPU times for this order of approximation are 7.3 for $\mu_{\mathrm{pos}}^{(r)}(y)$ and 29.0 for $\widehat{\mu}_{\mathrm{pos}}^{(r)}(y)$. These are roughly the number of iterations of an iterative solver that one could take for equivalent computational cost. That is, the speedup of the posterior mean approximation compared to an iterative solver is not particularly dramatic in this case, because the forward model A is simply a sparse matrix that is cheap to apply. This is different for the heat equation example discussed in section 5.3.

Note that the above computational time estimates exclude other costs associated with iterative solvers. For instance, preconditioners are often applied; these significantly decrease the number of iterations needed for the solvers to converge but, on the other hand, increase the cost per iteration. A popular approach for solving the posterior mean efficiently is to use the prior covariance as the preconditioner [6]. In the limited-angle tomography problem, including the application of this preconditioner in the reference CPU time would reduce the relative CPU time of our $r = 200$ approximations to 0.48 for $\mu_{\mathrm{pos}}^{(r)}(y)$ and 1.9 for $\widehat{\mu}_{\mathrm{pos}}^{(r)}(y)$. That is, the cost of computing our approximations is roughly equal to *one iteration* of a prior-preconditioned iterative solver. The large difference compared to the case without preconditioning is due to the fact that the cost of applying the prior here is computationally much higher than applying the forward model.

Figure 6 (left panel) shows unnormalized errors in the approximation of $\mu_{\mathrm{pos}}(y)$,

(5.5)
$$\|e(y)\|_{\Gamma_{\mathrm{pos}}^{-1}}^2 = \|\mu_{\mathrm{pos}}^{(r)}(y) - \mu_{\mathrm{pos}}(y)\|_{\Gamma_{\mathrm{pos}}^{-1}}^2 \quad \text{and} \quad \|\widehat{e}(y)\|_{\Gamma_{\mathrm{pos}}^{-1}}^2 = \|\widehat{\mu}_{\mathrm{pos}}^{(r)}(y) - \mu_{\mathrm{pos}}(y)\|_{\Gamma_{\mathrm{pos}}^{-1}}^2,$$
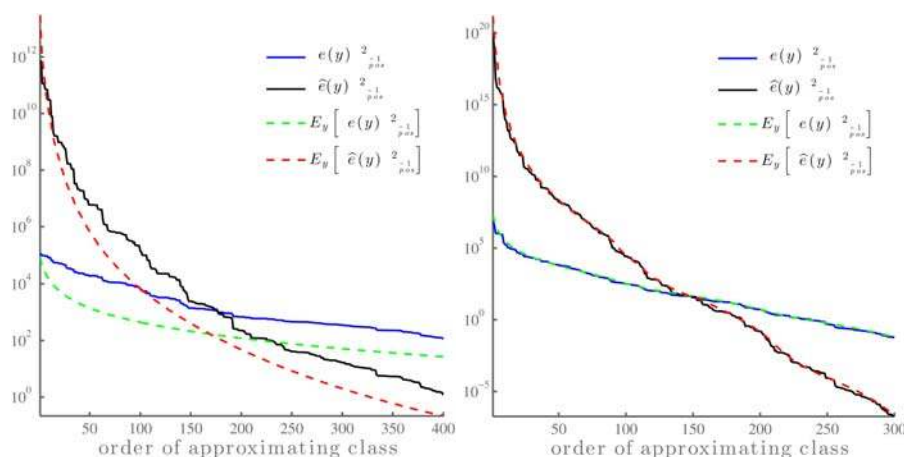
FIG. 6. *The errors* $\|e(y)\|^2_{\Gamma^{-1}_{\mathrm{pos}}}$ *(blue) and* $\|\widehat{e}(y)\|^2_{\Gamma^{-1}_{\mathrm{pos}}}$ *(black) defined by* (5.5)*, and their expected values in green and red, respectively, for Example* 2 *(left panel) and Example* 3 *(right panel).*

for the same realization of y used in Figure 5. In the same panel we also show the expected values of these errors over the prior predictive distribution of y, which is exactly the r-dependent component of the Bayes risk given in Theorems 4.1 and 4.2. Both sets of errors decay with increasing r and show a similar crossover between the two approximation classes. But the particular error $\|e(y)\|^2_{\Gamma^{-1}_{\mathrm{pos}}}$ departs consistently from its expectation; this is not unreasonable in general (the mean estimator has a nonzero variance), but the offset may be accentuated in this case because the data are generated from an image that is not drawn from the prior. (The right panel of Figure 6, which comes from Example 3, represents a contrasting case.)

By design, the posterior approximations described in this paper perform well when the data inform a low-dimensional subspace of the parameter space. To better understand this effect, we also consider a *full-angle* configuration of the tomography problem, wherein the sources and detectors are evenly spread around the entire unknown object. In this case, the data are more informative than in the limited-angle configuration. This can be seen in the decay rate of the generalized eigenvalues of the pencil $(H, \Gamma^{-1}_{\mathrm{pr}})$ in the center panel of Figure 7 (blue and red curves); eigenvalues for the full-angle configuration decay more slowly than for the limited-angle configuration. Thus, according to the optimal loss given in (2.10) (Theorem 2.3), the prior-to-posterior update in the full-angle case must be of greater rank than the update in the limited-angle case for any given approximation error. Also, good approximation of $\mu_{\mathrm{pos}}(y)$ in the full-angle case requires higher order of the approximation class $\mathcal{A}$, as is shown in Figure 8. But because the data are strongly informative, they allow an almost perfect reconstruction of the underlying truth image. The relative CPU times are similar to the limited angle case: roughly 8 for $\mu^{(r)}_{\mathrm{pos}}(y)$ and 14 for $\widehat{\mu}^{(r)}_{\mathrm{pos}}(y)$. If preconditioning with the prior covariance is included in the reference CPU time calculation, the relative CPU times drop to 1.5 for $\mu^{(r)}_{\mathrm{pos}}(y)$ and to 2.6 for $\widehat{\mu}^{(r)}_{\mathrm{pos}}(y)$. We remark that in realistic applications of X-ray tomography, the limited angle setup is extremely common as it is cheaper and more flexible (yielding smaller and lighter devices) than a full-angle configuration.
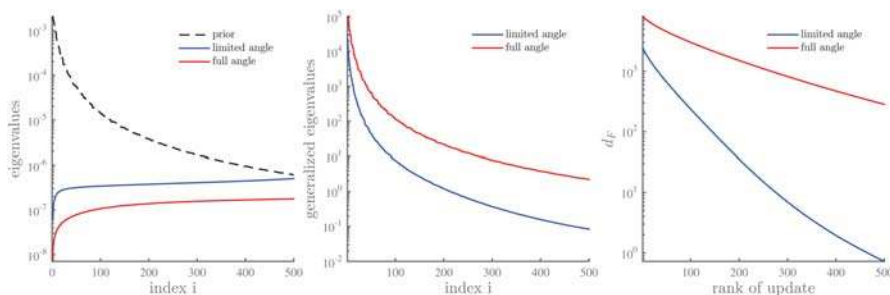
Fig. 7. *Left: Leading eigenvalues of Γ_{pr} and H^{-1} in the limited-angle and full-angle X-ray tomography problems. Center: Leading generalized eigenvalues of the pencil $(H, \Gamma_{\mathrm{pr}}^{-1})$ in the limited-angle (blue) and full-angle (red) cases. Right: $d_{\mathcal{F}}(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}})$ as a function of the rank of the update $KK^{\top}$, with $\widehat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^{\top}$, in the limited-angle (blue) and full-angle (red) cases.*
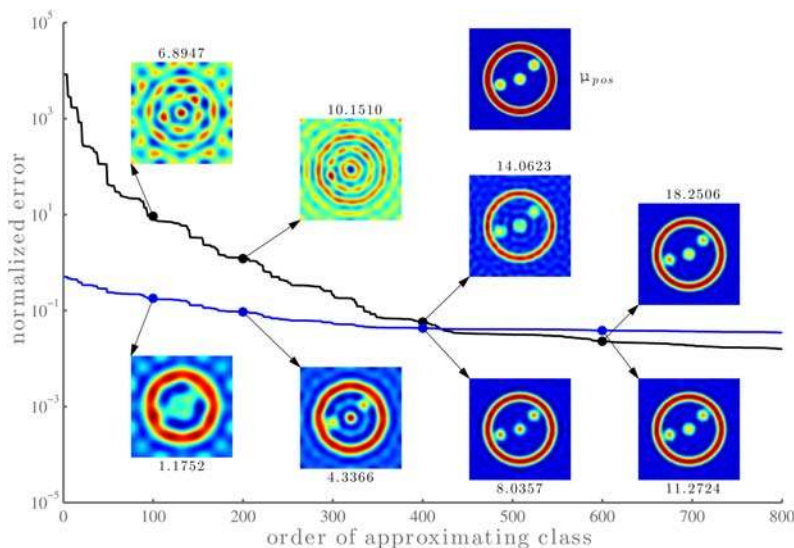

Fig. 8. *Same as Figure 5, but for full-angle X-ray tomography (sources and receivers spread uniformly around the entire object).*

5.3. Example 3: Heat equation. Our last example is the classic linear inverse problem of solving for the initial conditions of an inhomogeneous heat equation. Let $u(s,t)$ be the time-dependent state of the heat equation on $s = (s_1, s_2) \in \Omega = [0,1]^2$, $t \geq 0$, and let $\kappa(s)$ be the heat conductivity field. Given initial conditions, $u_0(s) = u(s,0)$, the state evolves in time according to the linear heat equation:

$$\frac{\partial u(s,t)}{\partial t} = -\nabla \cdot (\kappa(s)\nabla u(s,t)), \qquad s \in \Omega, \ t > 0,$$

$$(5.6) \qquad \kappa(s)\nabla u(s,t) \cdot n(s) = 0, \qquad s \in \partial\Omega, \ t > 0,$$

where $n(s)$ denotes the outward-pointing unit normal at $s \in \partial\Omega$. We place $n_s = 81$ sensors at the locations $s_1, \ldots, s_{n_s}$, uniformly spaced within the lower left quadrant of the spatial domain, as illustrated by the black dots in Figure 9. We use a finite-dimensional discretization of the parameter space based on the finite element method on a regular 100×100 grid, $\{s_i'\}$. Our goal is to infer the vector $x = (u_0(s_i'))$ of initial
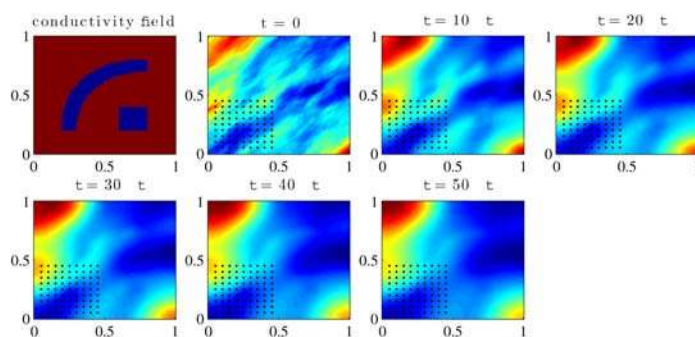
FIG. 9. *Heat equation (Example* 3*). Initial condition (top left) and several snapshots of the states at different times. Black dots indicate sensor locations.*

conditions on the grid. Thus, the dimension of the parameter space for the inference problem is $n = 10^4$. We use data measured at 50 discrete times $t = t_1, t_2, \ldots, t_{50}$, where $t_i = i\triangle t$, and $\triangle t = 2 \times 10^{-4}$. At each time t_i, pointwise observations of the state u are taken at these sensors, i.e.,

$$(5.7) \qquad d_i = \mathcal{C}u(s, t_i),$$

where $\mathcal{C}$ is the observation operator that maps the function $u(s, t_i)$ to $d = (u(s_1, t_i), \ldots, u(s_n, t_i))^\top$. The vector of observations is then $d = [d_1; d_2; \ldots; d_{50}]$. The noisy data vector is $y = d + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $\sigma = 10^{-2}$. Note that the data are a linear function of the initial conditions, perturbed by Gaussian noise. Thus the data can be written as

$$(5.8) \qquad y = Gx + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

where G is a linear map defined by the composition of the forward model (5.6) with the observation operator (5.7), both linear.

We generate synthetic data by evolving the initial conditions shown in Figure 9. This "true" value of the inversion parameters x is a discretized realization of a Gaussian process satisfying an SPDE of the same form used in the previous tomography example, but now with a nonstationary permeability field. In other words, the truth is a draw from the prior in this example (unlike in the previous example), and the prior Gaussian process satisfies the following SPDE:

$$(5.9) \qquad \gamma \left(\kappa^2 \mathcal{I} - \nabla \cdot \mathbf{c}(s)\nabla \right) x(s) = \mathcal{W}(s), \qquad s \in \Omega,$$

where $\mathbf{c}(s)$ is the space-dependent permeability tensor.

Figure 10 and the right panel in Figure 6 show our numerical results. They have the same interpretations as Figures 5 and 6 in the tomography example. The trends in the figures are consistent with those encountered in the previous example and confirm the good performance of the optimal low-rank approximation. Notice that in Figures 10 and 6 the approximation of the posterior mean appears to be nearly perfect (visually) once the error curves for the two approximations cross. This is somewhat expected from the theory since we know that the crossing point should occur when the approximations start to use eigenvalues of the pencil $(H, \Gamma_{\mathrm{pr}}^{-1})$ that are less than one—that is, once we have exhausted directions in the parameter space where the data are more constraining than the prior.
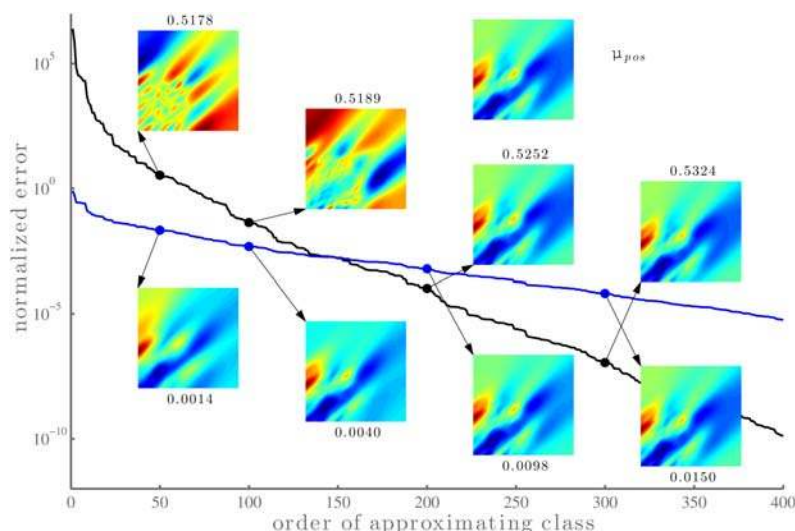
FIG. 10. *Same as Figure* 5, *but for Example* 3 *(initial condition inversion for the heat equation).*

Again, we report the relative CPU time for each posterior mean approximation above/below the corresponding snapshot in Figure 10. The results differ significantly from the tomography example. For instance, at order $r = 200$, which yields approximations that are visually indistinguishable from the true mean, the relative CPU times are 0.001 for $\mu_{\mathrm{pos}}^{(r)}(y)$ and 0.53 for $\widehat{\mu}_{\mathrm{pos}}^{(r)}(y)$. Therefore we can compute an accurate mean approximation for a new realization of the data much more quickly than taking one iteration of an iterative solver. Recall that, consistent with the setting described at the start of section 4, this is a comparison of *online* times, after the matrices (4.4) or (4.7) have been precomputed. The difference between this case and the tomography example of section 5.2 is due to the higher CPU cost of applying the forward and adjoint models for the heat equation—solving a time-dependent PDE versus applying a sparse matrix. Also, because the cost of applying the prior covariance is negligible compared to that of the forward and adjoint solves in this example, preconditioning the iterative solver with the prior would not strongly affect the reported relative CPU times, unlike the tomography example.

Figure 11 illustrates some important directions characterizing the heat equation inverse problem. The first two columns show the four leading eigenvectors of, respectively, Γ_{pr} and H. Notice that the support of the eigenvectors of H concentrates around the sensors. The third column shows the four leading directions $(\widehat{w}_i)$ defined in Theorem 2.3. These directions define the optimal prior-to-posterior covariance matrix update (cf. (2.9)). This update of Γ_{pr} is *necessary* to capture directions $(\widetilde{w}_i)$ of greatest relative difference between prior and posterior variance (cf. Corollary 3.1). The four leading directions $(\widetilde{w}_i)$ are shown in the fourth column. The support of these modes is again concentrated around the sensors, which intuitively makes sense as these are directions of greatest variance reduction.

6. Conclusions. This paper has presented and characterized optimal approximations of the Bayesian solution of linear inverse problems, with Gaussian prior and noise distributions defined on finite-dimensional spaces. In a typical large-scale inverse problem, observations may be informative—relative to the prior—only on a
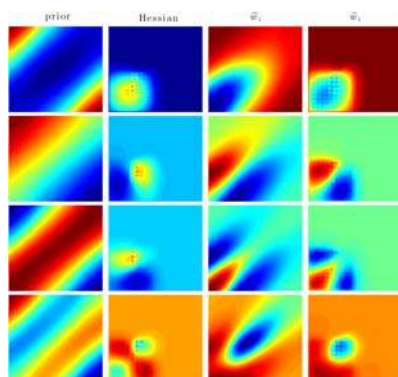
FIG. 11. *Heat equation (Example* 3*). First column: Four leading eigenvectors of* Γ_{pr}*. Second column: Four leading eigenvectors of* H*. Third column: Four leading directions* $(\widehat{w}_i)$ *(cf. (2.9)). Fourth column: Four leading directions* $(\widetilde{w}_i)$ *(cf. Corollary* 3.1*)*

low-dimensional subspace of the parameter space. Our approximations therefore identify and exploit low-dimensional structure in the *update* from prior to posterior.

We have developed two types of optimality results. In the first, the posterior covariance matrix is approximated as a low-rank negative semidefinite update of the prior covariance matrix. We describe an update of this form that is optimal with respect to a broad class of loss functions between covariance matrices, exemplified by the Förstner metric [29] for symmetric positive definite matrices. We argue that this is the appropriate class of loss functions with which to evaluate approximations of the posterior covariance matrix, and we show that optimality in such metrics identifies directions in parameter space along which the posterior variance is reduced the most, relative to the prior. Optimal low-rank updates are derived from a generalized eigendecomposition of the pencil defined by the negative log-likelihood Hessian and the prior precision matrix. These updates have been proposed in previous work [28], but our work complements these efforts by characterizing the optimality of the resulting approximations. Under the assumption of exact knowledge of the posterior mean, our results extend to optimality statements between the associated distributions (e.g., optimality in the Hellinger distance and in the K-L divergence). Second, we have developed fast approximations of the posterior mean that are useful when repeated evaluations thereof are required for multiple realizations of the data (e.g., in an online inference setting). These approximations are optimal in the sense that they minimize the Bayes risk for squared-error loss induced by the posterior precision matrix. The most computationally efficient of these approximations expresses the posterior mean as the product of a single low-rank matrix with the data. We have demonstrated the covariance and mean approximations numerically on a variety of inverse problems: synthetic problems constructed from random Hessian and prior covariance matrices; an X-ray tomography problem with different observation scenarios; and inversion for the initial condition of a heat equation, with localized observations and a nonstationary prior.

This work has several possible extensions of interest, some of which are already part of ongoing research. First, it is natural to generalize the present approach to infinite-dimensional parameter spaces endowed with Gaussian priors. This setting is essential to understanding and formalizing Bayesian inference over function spaces [9, 79]. Here, by analogy with the current results, one would expect the posterior covariance operator to be well approximated by a finite-rank negative perturbation of the

prior covariance operator. A further extension could allow the data to become infinite-dimensional as well. Another important task is to generalize the present methodology to inverse problems with nonlinear forward models. One approach for doing so is presented in [22]; other approaches are certainly possible. Yet another interesting research topic is the study of analogous approximation techniques for sequential inference. We note that the assimilation step in a linear (or linearized) data assimilation scheme can be already tackled within the framework presented here. But the nonstationary setting, where inference is interleaved with evolution of the state, introduces the possibility for even more tailored and structure-exploiting approximations.

Appendix A. Technical results. Here we collect the proofs and other technical results necessary to support the statements made in the previous sections.

We start with an auxiliary approximation result that plays an important role in our analysis. Given a semipositive definite diagonal matrix D, we seek an approximation of $D + I$ by a rank r perturbation of the identity, $UU^\top + I$, that minimizes a loss function from the class $\mathcal{L}$ defined in (2.6). The following lemma shows that the optimal solution $\widehat{U}\widehat{U}^\top$ is simply the best rank r approximation of the matrix D in the Frobenius norm.

LEMMA A.1 (approximation lemma). *Let* $D = \mathrm{diag}\{d_1^2, \ldots, d_n^2\}$, *with* $d_i^2 \geq d_{i+1}^2$, *and* $L \in \mathcal{L}$. *Define the functional* $\mathcal{J} : \mathbb{R}^{n \times r} \to \mathbb{R}$, *as* $\mathcal{J}(U) = L(UU^\top + I, D + I) = \sum_i f(\sigma_i)$, *where* (σ_i) *are the generalized eigenvalues of the pencil* $(UU^\top + I, D + I)$ *and* $f \in \mathcal{U}$. *Then the following hold:*
  (i) *There is a minimizer,* $\widehat{U}$, *of* $\mathcal{J}$ *such that*

$$\text{(A.1)} \qquad \widehat{U}\widehat{U}^\top = \sum_{i=1}^{r} d_i^2\, e_i e_i^\top,$$

      where (e_i) *are the columns of the identity matrix.*
 (ii) *If the first* r *eigenvalues of* D *are distinct, then any minimizer of* $\mathcal{J}$ *satisfies* (A.1).

Proof. The idea is to apply [53, Theorem 1.1] to the functional $\mathcal{J}$. To this end, we notice that $\mathcal{J}$ can be equivalently written as $\mathcal{J}(U) = F \circ \rho_n \circ g(U)$, where $F : \mathbb{R}_+^n \to \mathbb{R}$ is of the form $F(x) = \sum_{i=1}^{n} f(x_i)$; ρ_n denotes a function that maps an $n \times n$ SPD matrix A to its eigenvalues $\sigma = (\sigma_i)$ (i.e., $\rho_n(A) = \sigma$ and since F is a symmetric function, the order of the eigenvalues is irrelevant); and the mapping g is given by $g(U) = (D+I)^{-1/2}(UU^\top + I)(D+I)^{-1/2}$ for all $U \in \mathbb{R}^{n \times r}$. Since the function $F \circ \rho_n$ satisfies the hypotheses in [53, Theorem 1.1], $F \circ \rho_n$ is differentiable at the SPD matrix X iff F is differentiable at $\rho_n(X)$, in which case $(F \circ \rho_n)'(X) = Z S_\sigma Z^\top$, where

$$S_\sigma = \mathrm{diag}[\, F'(\rho_n(X))\,] = \mathrm{diag}\{f'(\sigma_1), \ldots, f'(\sigma_n)\},$$

and Z is an orthogonal matrix such that $X = Z\, \mathrm{diag}[\, \rho_n(X)\,] Z^\top$. Using the chain rule, we obtain

$$\frac{\partial \mathcal{J}(U)}{\partial U_{ij}} = \mathrm{tr}\left(Z S_\sigma Z^\top\, \frac{\partial g(U)}{\partial U_{ij}} \right),$$

which leads to the following gradient of $\mathcal{J}$ at U:

$$\mathcal{J}'(U) = 2(D+I)^{-1/2} Z S_\sigma (D+I)^{-1/2} Z^\top U = 2\, W S_\sigma W^\top U,$$

where the orthogonal matrix Z is such that the matrix $W = (D + I)^{-1/2}Z$ satisfies

$$(A.2) \qquad\qquad (UU^\top + I)W = (D + I)W\Upsilon_\sigma$$

with $\Upsilon_\sigma = \mathrm{diag}(\sigma)$. Now we show that the functional $\mathcal{J}$ is coercive. Let (U_k) be a sequence of matrices such that $\|U_k\|_F \to \infty$. Hence, $\sigma_{max}(g(U_k)) \to \infty$ and so does $\mathcal{J}$ since

$$\mathcal{J}(U_k) \geq f(\sigma_{max}(g(U_k))) + (n - 1)f(1)$$

and $f(x) \to \infty$ as $x \to \infty$. Thus, $\mathcal{J}$ is a differentiable coercive functional and has a global minimizer $\widehat{U}$ with zero gradient:

$$(A.3) \qquad\qquad \mathcal{J}'(\widehat{U}) = 2WS_\sigma W^\top \widehat{U} = 0.$$

However, since $f \in \mathcal{U}$, $f'(x) = 0$ iff $x = 1$. It follows that condition (A.3) is equivalent to

$$(A.4) \qquad\qquad (I - \Upsilon_\sigma)W^\top \widehat{U} = 0.$$

Equations (A.2) and (A.4) give $\widehat{U}\widehat{U}^\top - D = W^{-\top}\Upsilon^{-1}(\Upsilon - I)W^\top$, and right-multiplication by $\widehat{U}\widehat{U}^\top$ then yields

$$(A.5) \qquad\qquad D\,(\widehat{U}\widehat{U}^\top) = (\widehat{U}\widehat{U}^\top)^2.$$

In particular, if u is an eigenvector of $\widehat{U}\widehat{U}^\top$ with nonzero eigenvalue α, then u is an eigenvector of D, $Du = \alpha u$, and thus $\alpha = d_i^2 > 0$ for some i. Thus, any solution of (A.5) is such that

$$(A.6) \qquad\qquad \widehat{U}\widehat{U}^\top = \sum_{i=1}^{r_k} d_{k_i}^2 e_{k_i} e_{k_i}^\top$$

for some subsequence (k_ℓ) of $\{1, \ldots, n\}$ and rank $r_k \leq r$. Notice that any $\widehat{U}$ satisfying (A.5) is also a critical point according to (A.4). From (A.6) we also find that $g(\widehat{U})$ is a diagonal matrix,

$$g(\widehat{U}) = (D + I)^{-1}\left(\sum_{i=1}^{r_k} d_{k_i}^2 e_{k_i} e_{k_i}^\top + I\right).$$

The diagonal entries σ_i, which are the eigenvalues of $g(\widehat{U})$, are given by $\sigma_i = 1$ if $i = k_\ell$ for some $\ell \leq r_k$, or $\sigma_i = 1/(1 + d_i^2)$ otherwise. In either case, we have $0 < \sigma_i \leq 1$ and the monotonicity of f implies that $\mathcal{J}(\widehat{U})$ is minimized by the subsequence $k_1 = 1, \ldots, k_r = r$ and by the choice $r_k = r$. This proves (A.1). It is clear that if the first r eigenvalues of D are distinct, then any minimizer of $\mathcal{J}$ satisfies (A.1). $\quad\square$

Most of the objective functions we consider have the same structure as the loss function $\mathcal{J}$, hence the importance of Lemma A.1.

The next lemma shows that searching for a negative update of Γ_{pr} is equivalent to looking for a positive update of the prior precision matrix. In particular, the lemma provides a bijection between the two approximation classes, $\mathcal{M}_r$ and $\mathcal{M}_r^{-1}$, defined

by (2.4) and (3.1). In what follows, S_{pr} is any square root of the prior covariance matrix such that $\Gamma_{\mathrm{pr}} = S_{\mathrm{pr}} S_{\mathrm{pr}}^\top$.

LEMMA A.2 (prior updates). *For any negative semidefinite update of* Γ_{pr}, $\widehat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^\top$ *with* $\widehat{\Gamma}_{\mathrm{pos}} \succ 0$, *there is a matrix* U *(of the same rank as K) such that* $\widehat{\Gamma}_{\mathrm{pos}} = \left(\Gamma_{\mathrm{pr}}^{-1} + UU^\top\right)^{-1}$. *The converse is also true.*

Proof. Let $ZDZ^\top = S_{\mathrm{pr}}^{-1} KK^\top S_{\mathrm{pr}}^{-\top}$, $D = \mathrm{diag}\{d_i^2\}$, be a reduced SVD of $S_{\mathrm{pr}}^{-1} KK^\top S_{\mathrm{pr}}^{-\top}$. Since $\widehat{\Gamma}_{\mathrm{pos}} \succ 0$ by assumption, we must have $d_i^2 < 1$ for all i, and we may thus define $U = S_{\mathrm{pr}}^{-\top} ZD^{1/2}(I - D)^{-1/2}$. By Woodbury's identity,

$$\left(\Gamma_{\mathrm{pr}}^{-1} + UU^\top\right)^{-1} = \Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pr}} U \left(I + U^\top \Gamma_{\mathrm{pr}}^{-1} U\right)^{-1} U^\top \Gamma_{\mathrm{pr}} = \Gamma_{\mathrm{pr}} - KK^\top = \widehat{\Gamma}_{\mathrm{pos}}.$$

Conversely, given a matrix U, we use again Woodbury's identity to write $\widehat{\Gamma}_{\mathrm{pos}}$ as a negative semidefinite update of Γ_{pr}: $\widehat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^\top \succ 0$. ☐

Now we prove our main result on approximations of the posterior covariance matrix.

Proof of Theorem 2.3. Given a loss function $L \in \mathcal{L}$, our goal is to minimize

$$(A.7) \qquad L(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}}) = \sum_i f(\sigma_i)$$

over $K \in \mathbb{R}^{n \times r}$ subject to the constraint $\widehat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^\top \succ 0$, where (σ_i) are the generalized eigenvalues of the pencil $(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}})$ and f belongs to the class $\mathcal{U}$ defined by (2.7). We also write $\sigma_i(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}})$ to specify the pencil corresponding to the eigenvalues. By Lemma A.2, the optimization problem is equivalent to finding a matrix, $U \in \mathbb{R}^{n \times r}$, that minimizes (A.7) subject to $\widehat{\Gamma}_{\mathrm{pos}}^{-1} = \Gamma_{\mathrm{pr}}^{-1} + UU^\top$. Observe that (σ_i) are also the eigenvalues of the pencil $(\widehat{\Gamma}_{\mathrm{pos}}^{-1}, \Gamma_{\mathrm{pos}}^{-1})$.

Let $WDW^\top = S_{\mathrm{pr}}^\top H S_{\mathrm{pr}}$ with $D = \mathrm{diag}\{\delta_i^2\}$ be an SVD of $S_{\mathrm{pr}}^\top H S_{\mathrm{pr}}$. Then, by the invariance properties of the generalized eigenvalues we have

$$\sigma_i(\widehat{\Gamma}_{\mathrm{pos}}^{-1}, \Gamma_{\mathrm{pos}}^{-1}) = \sigma_i(W^\top S_{\mathrm{pr}}^\top \widehat{\Gamma}_{\mathrm{pos}}^{-1} S_{\mathrm{pr}} W, W^\top S_{\mathrm{pr}}^\top \Gamma_{\mathrm{pos}}^{-1} S_{\mathrm{pr}} W) = \sigma_i(ZZ^\top + I, D + I),$$

where $Z = W^\top S_{\mathrm{pr}}^\top U$. Therefore, our goal reduces to finding a matrix, $Z \in \mathbb{R}^{n \times r}$, that minimizes (A.7) with (σ_i) being the generalized eigenvalues of the pencil $(ZZ^\top + I, D + I)$. Applying Lemma A.1 leads to the simple solution $ZZ^\top = \sum_{i=1}^r \delta_i^2 e_i e_i^\top$, where (e_i) are the columns of the identity matrix. In particular, the solution is unique if the first r eigenvalues of $S_{\mathrm{pr}}^\top H S_{\mathrm{pr}}$ are distinct. The corresponding approximation $UU^\top$ is then

$$(A.8) \qquad UU^\top = S_{\mathrm{pr}}^{-\top} WZZ^\top W^\top S_{\mathrm{pr}}^{-1} = \sum_{i=1}^r \delta_i^2 \, \widetilde{w}_i \widetilde{w}_i^\top,$$

where $\widetilde{w}_i = S_{\mathrm{pr}}^{-\top} w_i$ and w_i is the ith column of W. Woodbury's identity gives the corresponding negative update of Γ_{pr} as

$$(A.9) \qquad \widehat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^\top, \quad KK^\top = \sum_{i=1}^r \delta_i^2 \left(1 + \delta_i^2\right)^{-1} \widehat{w_i} \widehat{w_i}^\top$$

with $\widehat{w_i} = S_{\mathrm{pr}} w_i$. Now, it suffices to note that the couples $(\delta_i^2, \widehat{w_i})$ defined here are precisely the generalized eigenpairs of the pencil $(H, \Gamma_{\mathrm{pr}}^{-1})$. At optimality, $\sigma_i = 1$ for $i \le r$ and $\sigma_i = (1 + \delta_i^2)^{-1}$ for $i > r$, proving (2.10). ☐

Before proving Lemma 2.2, we recall that the K-L divergence and the Hellinger distance between two multivariate Gaussians, $\nu_1 = \mathcal{N}(\mu, \Sigma_1)$ and $\nu_2 = \mathcal{N}(\mu, \Sigma_2)$, with the same mean and full rank covariance matrices are given, respectively, by [70]:

$$(A.10) \qquad D_{\mathrm{KL}}\left(\nu_1 \| \nu_2\right) = \frac{1}{2}\left[\mathrm{trace}\left(\Sigma_2^{-1}\Sigma_1\right) - \mathrm{rank}(\Sigma_1) - \ln\left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right)\right],$$

$$(A.11) \qquad d_{\mathrm{Hell}}\left(\nu_1, \nu_2\right) = \sqrt{1 - \frac{|\Sigma_1|^{1/4}\,|\Sigma_2|^{1/4}}{|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2|^{1/2}}}.$$

Proof of Lemma 2.2. By (A.10), the K-L divergence between the posterior $\nu_{\mathrm{pos}}(y)$ and the Gaussian approximation $\widehat{\nu}_{\mathrm{pos}}(y)$ can be written in terms of the generalized eigenvalues of the pencil $(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}})$ as

$$D_{\mathrm{KL}}\left(\nu_{\mathrm{pos}}(y) \| \widehat{\nu}_{\mathrm{pos}}(y)\right) = \sum_i \left(\sigma_i - \ln \sigma_i - 1\right)/2,$$

and since $f(x) = (x - \ln x - 1)/2$ belongs to $\mathcal{U}$, we see that the K-L divergence is a loss function in the class $\mathcal{L}$ defined by (2.6). Hence, Theorem 2.3 applies and the equivalence between the two approximations follows trivially. An analogous argument holds for the Hellinger distance. The squared Hellinger distance between $\nu_{\mathrm{pos}}(y)$ and $\widehat{\nu}_{\mathrm{pos}}(y)$ can be written in terms of the generalized eigenvalues, (σ_i), of the pencil $(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}})$, as

$$(A.12) \qquad d_{\mathrm{Hell}}\left(\nu_{\mathrm{pos}}(y), \widehat{\nu}_{\mathrm{pos}}(y)\right)^2 = 1 - 2^{\ell/2}\prod_i \sigma_i^{1/4}\left(1 + \sigma_i\right)^{-1/2}.$$

where ℓ is the dimension of the parameter space. Minimizing (A.12) is equivalent to maximizing $\prod_i \sigma_i^{1/4}\left(1 + \sigma_i\right)^{-1/2}$, which in turn is equivalent to minimizing the functional:

$$(A.13) \qquad L(\Gamma_{\mathrm{pos}}, \widehat{\Gamma}_{\mathrm{pos}}) = -\sum_i \ln(\sigma_i^{1/4}(1 + \sigma_i)^{-1/2}) = \sum_i \ln(2 + \sigma_i + 1/\sigma_i)/4.$$

Since $f(x) = \ln(2 + x + 1/x)/4$ belongs to $\mathcal{U}$, Theorem 2.3 can be applied once again. $\square$

Proof of Corollary 3.1. The proofs of parts (i) and (ii) were already given in the proof of Theorem 2.3. Part (iii) holds because

$$(1 + \delta_i^2)\Gamma_{\mathrm{pos}}\,\widetilde{w}_i = (1 + \delta_i^2)(H + \Gamma_{\mathrm{pr}}^{-1})^{-1}S_{\mathrm{pr}}^{-\top}w_i$$
$$= (1 + \delta_i^2)\,S_{\mathrm{pr}}(S_{\mathrm{pr}}^{\top}H\,S_{\mathrm{pr}} + I)^{-1}w_i = S_{\mathrm{pr}}\,w_i = \Gamma_{\mathrm{pr}}\,\widetilde{w}_i,$$

because w_i is an eigenvector of $(S_{\mathrm{pr}}^{\top}H\,S_{\mathrm{pr}} + I)^{-1}$ with eigenvalue $(1 + \delta_i^2)^{-1}$ as shown in the proof of Theorem 2.3. $\square$

Now we turn to optimality results for approximations of the posterior mean. In what follows, let S_{pr}, S_{obs}, S_{pos}, and S_{y} be the matrix square roots of, respectively, Γ_{pr}, Γ_{obs}, Γ_{pos}, and $\Gamma_{\mathrm{y}} := \Gamma_{\mathrm{obs}} + G\,\Gamma_{\mathrm{pr}}\,G^{\top}$ such that $\Gamma = S\,S^{\top}$ (i.e., possibly nonsymmetric square roots).

Equation (4.1) shows that to minimize $\mathbb{E}(\,\|Ay - x\|^2_{\Gamma^{-1}_{\mathrm{pos}}}\,)$ over $A \in \mathcal{A}$, we need only to minimize $\mathbb{E}(\,\|\,Ay - \mu_{\mathrm{pos}}(y)\,\|^2_{\Gamma^{-1}_{\mathrm{pos}}}\,)$. Furthermore, since $\mu_{\mathrm{pos}}(y) = \Gamma_{\mathrm{pos}}\,G^\top \Gamma^{-1}_{\mathrm{obs}}\,y$, it follows that

$$(A.14) \qquad \mathbb{E}(\,\|\,Ay - \mu_{\mathrm{pos}}(y)\,\|^2_{\Gamma^{-1}_{\mathrm{pos}}}\,) = \|\,S^{-1}_{\mathrm{pos}}\,(A - \Gamma_{\mathrm{pos}}\,G^\top \Gamma^{-1}_{\mathrm{obs}})\,S_{\mathrm{y}}\,\|^2_F.$$

We are therefore led to the following optimization problem:

$$(A.15) \qquad \min_{A \in \mathcal{A}} \|\,S^{-1}_{\mathrm{pos}}A\,S_{\mathrm{y}} - S^\top_{\mathrm{pos}}G^\top \Gamma^{-1}_{\mathrm{obs}}\,S_{\mathrm{y}}\,\|_F.$$

The following result shows that an SVD of the matrix $S_{\widehat{H}} := S^\top_{\mathrm{pr}}\,G^\top S^{-\top}_{\mathrm{obs}}$ can be used to obtain simple expressions for the square roots of Γ_{pos} and Γ_{y}.

LEMMA A.3 (square roots). *Let $WDV^\top$ be an SVD of $S_{\widehat{H}} = S^\top_{\mathrm{pr}}\,G^\top S^{-\top}_{\mathrm{obs}}$. Then*

$$(A.16) \qquad\qquad\qquad S_{\mathrm{pos}} = S_{\mathrm{pr}}\,W(\,I + DD^\top\,)^{-1/2}\,W^\top,$$

$$(A.17) \qquad\qquad\qquad S_{\mathrm{y}} = S_{\mathrm{obs}}\,V(\,I + D^\top D\,)^{1/2}\,V^\top$$

are square roots of Γ_{pos} and Γ_{y}.

Proof. We can rewrite $\Gamma_{\mathrm{pos}} = (\,G^\top \Gamma^{-1}_{\mathrm{obs}}G + \Gamma^{-1}_{\mathrm{pr}}\,)^{-1}$ as

$$\Gamma_{\mathrm{pos}} = S_{\mathrm{pr}}\,(\,S_{\widehat{H}}\,S^\top_{\widehat{H}} + I\,)^{-1}\,S^\top_{\mathrm{pr}} = S_{\mathrm{pr}}\,W(\,DD^\top + I\,)^{-1}W^\top S^\top_{\mathrm{pr}}$$
$$= [\,S_{\mathrm{pr}}\,W(\,DD^\top + I\,)^{-1/2}\,W^\top\,][\,S_{\mathrm{pr}}\,W(\,DD^\top + I\,)^{-1/2}\,W^\top\,]^\top,$$

which proves (A.16). The proof of (A.17) follows similarly using $S^\top_{\widehat{H}}\,S_{\widehat{H}} = S^{-1}_{\mathrm{obs}}G\Gamma_{\mathrm{pr}}\,G^\top \Gamma^{-\top}_{\mathrm{obs}}$. $\qquad\square$

In the next two proofs we use $(C)_r$ to denote a rank r approximation of the matrix C in the Frobenius norm.

Proof of Theorem 4.1. By [31, Theorem 2.1], an optimal $A \in \mathcal{A}_r$ is given by

$$(A.18) \qquad\qquad\qquad A = S_{\mathrm{pos}}\,\big(\,S^\top_{\mathrm{pos}}G^\top \Gamma^{-1}_{\mathrm{obs}}\,S_{\mathrm{y}}\,\big)_r\,S^{-1}_{\mathrm{y}}.$$

Now, we need some computations to show that (A.18) is equivalent to (4.4). Using (A.16) and (A.17) we find $S^\top_{\mathrm{pos}}G^\top \Gamma^{-1}_{\mathrm{obs}}\,S_{\mathrm{y}} = W(I + DD^\top)^{-1/2}D\,(I + D^\top D)^{1/2}\,V^\top$, and therefore $(\,S^\top_{\mathrm{pos}}G^\top \Gamma^{-1}_{\mathrm{obs}}\,S_{\mathrm{y}}\,)_r = \sum^r_{i=1}\delta_i w_i v^\top_i$, where w_i is the ith column of W, v_i is the ith column of V, and δ_i is the ith diagonal entry of D. Inserting this back into (A.18) yields $A = \sum_{i \leq r}\delta_i(1 + \delta^2_i)^{-1}S_{\mathrm{pr}}w_i v^\top_i S^{-1}_{\mathrm{obs}}$. Now it suffices to note that $\widehat{w}_i := S_{\mathrm{pr}}w_i$ is a generalized eigenvector of $(H, \Gamma^{-1}_{\mathrm{pr}})$, that $\widehat{v}_i := S^{-\top}_{\mathrm{obs}}v_i$ is a generalized eigenvector of $(G\Gamma_{\mathrm{pr}}G^\top, \Gamma_{\mathrm{obs}})$, and that (δ^2_i) are also eigenvalues of $(H, \Gamma^{-1}_{\mathrm{pr}})$. The minimum Bayes risk is a straightforward computation for the optimal estimator (4.4) using (A.14). $\qquad\square$

Proof of Theorem 4.2. Given $A \in \widehat{\mathcal{A}}_r$, we can restate (A.15) as the problem of finding a matrix B, of rank at most r, that minimizes

$$(A.19) \qquad \|\,S^{-1}_{\mathrm{pos}}(\Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pos}})\,G^\top \Gamma^{-1}_{\mathrm{obs}}\,S_{\mathrm{y}} - S^{-1}_{\mathrm{pos}}B\,(G^\top \Gamma^{-1}_{\mathrm{obs}}\,S_{\mathrm{y}})\,\|_F$$

such that $A = (\Gamma_{\mathrm{pr}} - B)\,G^\top \Gamma^{-1}_{\mathrm{obs}}$. By [31, Theorem 2.1], an optimal B is given by

$$(A.20) \qquad B = S_{\mathrm{pos}}(\,S^{-1}_{\mathrm{pos}}\,(\Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pos}})\,G^\top \Gamma^{-1}_{\mathrm{obs}}\,S_{\mathrm{y}}\,)_r(G^\top \Gamma^{-1}_{\mathrm{obs}}\,S_{\mathrm{y}})^\dagger,$$

where $\dagger$ denotes the pseudoinverse operator. A closer look at [31, Theorem 2.1] reveals that another minimizer of (A.19), itself not necessarily of minimum Frobenius norm, is given by

$$(A.21) \qquad B = S_{\mathrm{pos}}(\,S^{-1}_{\mathrm{pos}}\,(\Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pos}})\,G^\top \Gamma^{-1}_{\mathrm{obs}}\,S_{\mathrm{y}}\,)_r\,(S^\top_{\mathrm{pr}}G^\top \Gamma^{-1}_{\mathrm{obs}}\,S_{\mathrm{y}})^\dagger S^\top_{\mathrm{pr}}.$$

SPANTINI ET AL.

By Lemma A.3,

$$S_{\mathrm{pr}}^{\top}G^{\top}\Gamma_{\mathrm{obs}}^{-1}S_{\mathrm{y}} = W[\,D\left(I+D^{\top}D\right)^{1/2}\,]V^{\top},$$
$$S_{\mathrm{pos}}^{-1}\,\Gamma_{\mathrm{pr}}G^{\top}\Gamma_{\mathrm{obs}}^{-1}S_{\mathrm{y}} = W[\,(I+DD^{\top})^{1/2}D(I+D^{\top}D)^{1/2}\,]V^{\top},$$
$$S_{\mathrm{pos}}^{-1}\,\Gamma_{\mathrm{pos}}G^{\top}\Gamma_{\mathrm{obs}}^{-1}S_{\mathrm{y}} = W[\,(I+DD^{\top})^{-1/2}D(I+D^{\top}D)^{1/2}\,]V^{\top}$$

and therefore $(S_{\mathrm{pr}}^{\top}G^{\top}\Gamma_{\mathrm{obs}}^{-1}S_{\mathrm{y}})^{\dagger} = \sum_{i=1}^{q}\delta_{i}^{-1}\left(1+\delta_{i}^{2}\right)^{-1/2}v_{i}w_{i}^{\top}$ for $q = \mathrm{rank}(S_{\widehat{H}})$, whereas

$$(\,S_{\mathrm{pos}}^{-1}\,(\Gamma_{\mathrm{pr}}-\Gamma_{\mathrm{pos}})G^{\top}\Gamma_{\mathrm{obs}}^{-1}S_{\mathrm{y}}\,)_{r} = \sum_{i=1}^{r}\delta_{i}^{3}\,w_{i}v_{i}^{\top}.$$

Inserting these expressions back into (A.21), we obtain

$$B = S_{\mathrm{pr}}\left(\sum_{i=1}^{r}\frac{\delta_{i}^{2}}{1+\delta_{i}^{2}}\,w_{i}w_{i}^{\top}\right)S_{\mathrm{pr}}^{\top},$$

where w_i is the ith column of W, v_i is the ith column of V, and δ_i is the ith diagonal entry of D. Notice that $(\delta_i^2, \widehat{w}_i)$, with $\widehat{w}_i = S_{\mathrm{pr}}w_i$, are the generalized eigenpairs of $(H, \Gamma_{\mathrm{pr}}^{-1})$ (cf. proof of Theorem 2.3). Hence, by Theorem 2.3, we recognize the optimal approximation of Γ_{pos} as $\widehat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - B$. Plugging this expression back into (A.21) gives (4.7). The minimum Bayes risk in (ii) follows readily using the optimal estimator given by (4.7) in (A.14). □

REFERENCES

[1] V. AKÇELIK, G. BIROS, O. GHATTAS, J. HILL, D. KEYES, AND B. VAN BLOEMEN WAANDERS, *Parallel algorithms for PDE-constrained optimization*, Parallel Process. Sci. Comput., 20 (2006), p. 291.

[2] H. AUVINEN, J. M. BARDSLEY, H. HAARIO, AND T. KAURANNE, *Large-scale Kalman filtering using the limited memory BFGS method*, Electron. Trans. Numer. Anal., 35 (2009), pp. 217–233.

[3] H. AUVINEN, J. M. BARDSLEY, H. HAARIO, AND T. KAURANNE, *The variational Kalman filter and an efficient implementation using limited memory BFGS*, Internat. J. Numer. Methods Fluids, 64 (2010), pp. 314–335.

[4] Z. BAI, J. DEMMEL, J. DONGARRA, A. RUHE, AND H. VAN DER VORST, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, Software Environ. Tools 11, SIAM, Philadelphia, 2000.

[5] R. BARRETT, M. W. BERRY, T. F. CHAN, J. DEMMEL, J. DONATO, J. DONGARRA, V. EIJKHOUT, R. POZO, C. ROMINE, AND H. VAN DER VORST, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, 1994.

[6] T. BUI-THANH, C. BURSTEDDE, O. GHATTAS, J. MARTIN, G. STADLER, AND L. WILCOX, *Extreme-scale UQ for Bayesian inverse problems governed by PDEs*, in Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, IEEE Computer Society Press, 2012, p. 3.

[7] T. BUI-THANH AND O. GHATTAS, *Analysis of the Hessian for inverse scattering problems*: I. *Inverse shape scattering of acoustic waves*, Inverse Problems, 28 (2012), 055001.

[8] T. BUI-THANH, O. GHATTAS, AND D. HIGDON, *Adaptive Hessian-based nonstationary Gaussian process response surface method for probability density approximation with application to Bayesian solution of large-scale inverse problems*, SIAM J. Sci. Comput., 34 (2012), pp. A2837–A2871.

[9] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler, *A computational framework for infinite-dimensional Bayesian inverse problems part* I*: The linearized case, with application to global seismic inversion*, SIAM J. Sci. Comput., 35 (2013), pp. A2494–A2523.

[10] D. Calvetti, *Preconditioned iterative methods for linear discrete ill-posed problems from a Bayesian inversion perspective*, J. Comput. Appl. Math., 198 (2007), pp. 378–395.

[11] D. Calvetti, B. Lewis, and L. Reichel, *On the regularizing properties of the GMRES method*, Numer. Math., 91 (2002), pp. 605–625.

[12] D. Calvetti, D. McGivney, and E. Somersalo, *Left and right preconditioning for electrical impedance tomography with structural information*, Inverse Problems, 28 (2012), 055015.

[13] D. Calvetti and E. Somersalo, *Priorconditioners for linear systems*, Inverse Problems, 21 (2005), p. 1397.

[14] B. P. Carlin and T. A. Louis, *Bayesian Methods for Data Analysis*, 3rd ed., Chapman & Hall/CRC Press, Boca Raton, FL, 2009.

[15] R. A. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer-Verlag, Berlin, 1987.

[16] J. Chung and M. Chung, *Computing optimal low-rank matrix approximations for image processing*, in Proceedings of the Asilomar Conference on Signals, Systems and Computers, IEEE, New York, 2013, pp. 670–674.

[17] J. Chung and M. Chung, *An efficient approach for computing optimal low-rank regularized inverse matrices*, Inverse Problems, 30 (2014), 114009.

[18] J. Chung, M. Chung, and D. P. O'Leary, *Designing optimal spectral filters for inverse problems*, SIAM J. Sci. Comput., 33 (2011), pp. 3132–3152.

[19] J. Chung, M. Chung, and D. P. O'Leary, *Optimal filters from calibration data for image deconvolution with data acquisition error*, J. Math. Imaging Vis., 44 (2012), pp. 366–374.

[20] J. Chung, M. Chung, and D. P. O'Leary, *Optimal regularized low rank inverse approximation*, Linear Algebra Appl., 468 (2015), pp. 260–269.

[21] T. Cui, K. J. H. Law, and Y. Marzouk, *Dimension-independent likelihood-informed MCMC*, arXiv:1411.3688, 2014.

[22] T. Cui, J. Martin, Y. Marzouk, A. Solonen, and A. Spantini, *Likelihood-informed dimension reduction for nonlinear inverse problems*, Inverse Problems, 30 (2014), 114015.

[23] J. Cullum and W. Donath, *A block Lanczos algorithm for computing the q algebraically largest eigenvalues and a corresponding eigenspace of large, sparse, real symmetric matrices*, in Proceedings of the IEEE Conference on Decision and Control Including the 13th Symposium on Adaptive Processes, IEEE, New York, 1974, pp. 505–509.

[24] C. R. Dietrich and G. N. Newsam, *Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix*, SIAM J. Sci. Comput., 18 (1997), pp. 1088–1107.

[25] L. Dykes and L. Reichel, *Simplified gsvd computations for the solution of linear discrete ill-posed problems*, J. Comput. Appl. Math., 255 (2014), pp. 15–27.

[26] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[27] G. Evensen, *Data Assimilation*, Springer, Berlin, 2007.

[28] H. P. Flath, L. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas, *Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations*, SIAM J. Sci. Comput., 33 (2011), pp. 407–432.

[29] W. Förstner and B. Moonen, *A metric for covariance matrices*, in Geodesy: The Challenge of the 3rd Millennium, Springer, Berlin, 2003, pp. 299–309.

[30] C. Fox and A. Parker, *Convergence in variance of Chebyshev accelerated Gibbs samplers*, SIAM J. Sci. Comput., 36 (2014), pp. A124–A147.

[31] S. Friedland and A. Torokhti, *Generalized rank-constrained matrix approximations*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 656–659.

[32] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Vol. 3, Johns Hopkins University Press, Baltimore, 2012.

[33] G. H. Golub and R. Underwood, *The block Lanczos method for computing eigenvalues*, Math. Software, 3 (1977), pp. 361–377.

[34] G. H. Golub and Q. Ye, *An inverse free preconditioned krylov subspace method for symmetric generalized eigenvalue problems*, SIAM J. Sci. Comput., 24 (2002), pp. 312–334.

[35] N. Halko, P. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288.

[36] M. Hanke, *Conjugate Gradient Type Methods for Ill-Posed Problems*, Pitman Res. Notes in Math. Ser. 327, Longman, Harlow, UK, 1995.

[37] M. HANKE AND P. C. HANSEN, *Regularization methods for large-scale problems*, Surv. Math. Ind., 3 (1993), pp. 253–315.

[38] P. C. HANSEN, *Regularization, GSVD and truncated GSVD*, BIT, 29 (1989), pp. 491–504.

[39] P. C. HANSEN, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, Monogr. Math. Model. Comput. 4, SIAM, Philadelphia, 1998.

[40] J. HEIKKINEN, *Statistical Inversion Theory in X-Ray Tomography*, Master's thesis, Lappeenranta University of Technology, Finland, 2008.

[41] M. R. HESTENES, *Pseudoinversus and conjugate gradients*, Comm. ACM, 18 (1975), pp. 40–43.

[42] M. R. HESTENES AND E. STIEFEL, *Methods of Conjugate Gradients for Solving Linear Systems*, Vol. 49, National Bureau of Standards Washington, DC, 1952.

[43] L. HOMA, D. CALVETTI, A. HOOVER, AND E. SOMERSALO, *Bayesian preconditioned CGLS for source separation in MEG time series*, SIAM J. Sci. Comput., 35 (2013), pp. B778–B798.

[44] Y. HUA AND W. LIU, *Generalized Karhunen-Loève transform*, IEEE Signal Process. Lett., 5 (1998), pp. 141–142.

[45] W. JAMES AND C. STEIN, *Estimation with quadratic loss*, in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 1961, pp. 361–379.

[46] W. KAHAN AND B. N. PARLETT, *How Far Should You Go with the Lanczos Process*, No UCB/ERL-M78/48, Electronics Research Lab, University of California, Berkeley, 1978.

[47] J. KAIPIO AND E. SOMERSALO, *Statistical and Computational Inverse Problems*, Springer, Berlin, 2005.

[48] R. E. KALMAN, *A new approach to linear filtering and prediction problems*, J. Fluids Engrg., 82 (1960), pp. 35–45.

[49] S. KANIEL, *Estimates for some computational techniques in linear algebra*, Math. Comp., (1966), pp. 369–378.

[50] C. LANCZOS, *An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators*, U.S. Government Press Office, 1950.

[51] E. L. LEHMANN AND G. CASELLA, *Theory of Point Estimation*, Springer-Verlag, Berlin, 1998.

[52] R. B. LEHOUCQ, D. C. SORENSEN, AND C. YANG, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, Software Environ. Tools 6, SIAM, Philadelphia, 1998.

[53] A. S. LEWIS, *Derivatives of spectral functions*, Math. Oper. Res., 21 (1996), pp. 576–588.

[54] W. LI AND O. A. CIRPKA, *Efficient geostatistical inverse methods for structured and unstructured grids*, Water Resources Res., 42 (2006), W06402.

[55] C. LIEBERMAN, K. FIDKOWSKI, K. WILLCOX, AND B. VAN BLOEMEN WAANDERS, *Hessian-based model reduction: Large-scale inversion and prediction*, Internat. J. Numer. Methods Fluids, 71 (2013), pp. 135–150.

[56] C. LIEBERMAN, K. WILLCOX, AND O. GHATTAS, *Parameter and state model reduction for large-scale statistical inverse problems*, SIAM J. Sci. Comput., 32 (2010), pp. 2523–2542.

[57] F. LINDGREN, H. RUE, AND J. LINDSTRÖM, *An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach*, J. R. Stat. Soc. Ser. B, 73 (2011), pp. 423–498.

[58] D. V. LINDLEY AND A. F. M. SMITH, *Bayes estimates for the linear model*, J. R. Stat. Soc. Ser. B, 34 (1972), pp. 1–41.

[59] D. C. LIU AND J. NOCEDAL, *On the limited memory BFGS method for large scale optimization*, Math. Program., 45 (1989), pp. 503–528.

[60] J. MARTIN, L. WILCOX, C. BURSTEDDE, AND O. GHATTAS, *A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion*, SIAM J. Sci. Comput., 34 (2012), pp. A1460–A1487.

[61] Y. MARZOUK AND H. N. NAJM, *Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems*, J. Comput. Phys., 228 (2009), pp. 1862–1902.

[62] X. MENG, M. A. SAUNDERS, AND M. W. MAHONEY, *LSRN: A parallel iterative solver for strongly over-or underdetermined systems*, SIAM J. Sci. Comput., 36 (2014), pp. C95–C118.

[63] E. H. MOORE, *On the reciprocal of the general algebraic matrix*, Bull. Amer. Math. Soc., 26, pp. 394–395.

[64] J. NOCEDAL, *Updating quasi-Newton matrices with limited storage*, Math. Comput., 35 (1980), pp. 773–782.

[65] C. C. PAIGE, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*, Ph.D. thesis, University of London, 1971.

[66] C. C. PAIGE, *Computational variants of the Lanczos method for the eigenproblem*, IMA J. Appl. Math., 10 (1972), pp. 373–381.

[67] C. C. PAIGE, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, IMA J. Appl. Math., 18 (1976), pp. 341–349.

[68] C. C. PAIGE AND M. A. SAUNDERS, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.

[69] C. C. PAIGE AND M. A. SAUNDERS, *Algorithm 583: LSQR: Sparse linear equations and least squares problems*, ACM Trans. Math. Software, 8 (1982), pp. 195–209.

[70] L. PARDO, *Statistical Inference Based on Divergence Measures*, CRC Press, Boca Raton, FL, 2005.

[71] A. PARKER AND C. FOX, *Sampling Gaussian distributions in Krylov spaces with conjugate gradients*, SIAM J. Sci. Comput., 34 (2012), pp. B312–B334.

[72] R. PENROSE, *A generalized inverse for matrices*, Math. Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.

[73] Y. SAAD, *On the rates of convergence of the Lanczos and the block-Lanczos methods*, SIAM J. Numer. Anal., 17 (1980), pp. 687–706.

[74] A. H. SAMEH AND J. A. WISNIEWSKI, *A trace minimization algorithm for the generalized eigenvalue problem*, SIAM J. Numer. Anal., 19 (1982), pp. 1243–1259.

[75] C. SCHILLINGS AND C. SCHWAB, *Scaling Limits in Computational Bayesian Inversion*, Research report 2014-26, ETH-Zürich, 2014.

[76] A. SOLONEN, H. HAARIO, J. HAKKARAINEN, H. AUVINEN, I. AMOUR, AND T. KAURANNE, *Variational ensemble Kalman filtering using limited memory BFGS*, Electron. Trans. Numer. Anal., 39 (2012), pp. 271–285.

[77] D. SONDERMANN, *Best approximate solutions to matrix equations under rank restrictions*, Statist. Papers, 27 (1986), pp. 57–66.

[78] G. W. STEWART, *The efficient generation of random orthogonal matrices with an application to condition estimators*, SIAM J. Numer. Anal., 17 (1980), pp. 403–409.

[79] A. M. STUART, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[80] A. TIKHONOV, *Solution of incorrectly formulated problems and the regularization method*, Soviet Math. Dokl., 5 (1963), pp. 1035–1038.

[81] C. F. VAN LOAN, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.

[82] A. T. A. WOOD AND G. CHAN, *Simulation of stationary Gaussian processes in $[0,1]d$*, J. Comput. Graph. Statist., 3 (1994), pp. 409–432.

[83] S. J. WRIGHT AND J. NOCEDAL, *Numerical Optimization*, Vol. 2, Springer, New York, 1999.

[84] Y. YUE AND P. L. SPECKMAN, *Nonstationary spatial Gaussian Markov random fields*, J. Comput. Graphical Statist., 19 (2010), pp. 96–116.