OPTIMAL LOW-RANK APPROXIMATIONS OF GOAL-ORIENTED BAYESIAN LINEAR INVERSE PROBLEMS

ALESSIO SPANTINI*, TIANGANG CUI*, LUIS TENORIO‡, AND YOUSSEF MARZOUK*

Copper Mountain 2016 **student competition**

1. Introduction. In the Bayesian approach to inversion, parameters are treated as random variables and endowed with a prior distribution that encodes one's knowledge of the parameters before data are collected. The distribution of the data conditioned on the parameters is specified through the likelihood model. Bayes' theorem combines prior and likelihood information to yield the posterior distribution, i.e., the distribution of the parameters conditioned on the data. The posterior distribution reflects our updated knowledge of the parameters once measurements are collected and gives the Bayesian solution to the inverse problem. Characterizing the posterior distribution is of primary interest in real-life engineering and science applications (e.g., computerized tomography, optical imaging, spatial statistics). For instance, we might want to compute posterior marginals, the posterior probability of some functionals of the parameters, or the probability of rare events under the posterior measure. In all these cases we need samples[1] from the posterior distribution. This task tends to be extremely challenging in large scale applications, especially when the parameters represent a finite-dimensional approximation to a distributed stochastic process like a permeability or a temperature field.

We begin by considering a finite-dimensional Bayesian Gaussian linear inverse problem of the form

$$\boldsymbol{Y} = G\,\boldsymbol{X} + \varepsilon \tag{1.1}$$

where $\boldsymbol{X} \in \mathbb{R}^n$ represents the inversion parameters, $\boldsymbol{Y} \in \mathbb{R}^d$ denotes the noisy observations, $G \in \mathbb{R}^{d\times n}$ is a linear forward operator, and $\varepsilon \sim \mathcal{N}(0, \Gamma_{\text{obs}})$ is zero-mean additive Gaussian noise, statistically independent of $\boldsymbol{X}$ and endowed with covariance $\Gamma_{\text{obs}} \succ 0$. We prescribe a Gaussian prior distribution on the parameters, $\boldsymbol{X} \sim \mathcal{N}(0, \Gamma_{\text{pr}})$, where we assume, without loss of generality, zero prior mean and $\Gamma_{\text{pr}} \succ 0$. One is usually concerned with the posterior distribution of the parameters, $\boldsymbol{X}|\boldsymbol{Y} \sim \mathcal{N}(\mu_{\text{pos}}(\boldsymbol{Y}), \Gamma_{\text{pos}})$,[2] with posterior mean and covariance matrix given by

$$\mu_{\text{pos}}(\boldsymbol{Y}) = \Gamma_{\text{pos}}\, G^\top \Gamma_{\text{obs}}^{-1}\, \boldsymbol{Y}, \qquad \Gamma_{\text{pos}} = (H + \Gamma_{\text{pr}}^{-1})^{-1}, \tag{1.2}$$

where $H \coloneqq G^\top \Gamma_{\text{obs}}^{-1} G$ is the Hessian of the negative log-likelihood. In this paper, however, we are not interested in the parameters $\boldsymbol{X}$ per se, but rather in a quantity of interest (QoI) $\boldsymbol{Z}$ that is a function of the parameters

$$\boldsymbol{Z} = \mathcal{O}\boldsymbol{X} \tag{1.3}$$

for some linear and, without loss of generality, full row-rank operator $\mathcal{O} \in \mathbb{R}^{p\times n}$ with $p < n$. Our interests are thus *goal-oriented*, as we wish only to characterize $\boldsymbol{Z}$ and not the inversion parameters $\boldsymbol{X}$. Including such ultimate goals in the inference formulation is an essential modeling step in virtually every application of Bayesian inverse problems. The hope underlying this additional step is to reduce the computational complexity of inference by making the ultimate goals explicit. Nevertheless, it is still not well understood how to leverage ultimate goals in order to yield more efficient Bayesian inference algorithms (see [18] for computationally efficient approaches to non-Bayesian regularization techniques in goal-oriented problems). The present paper will precisely address this issue, filling a gap in the existing literature.

The Bayesian solution to the goal-oriented inverse problem is the posterior distribution of the QoI, i.e., $\boldsymbol{Z}|\boldsymbol{Y}$. It is easy to see that $\boldsymbol{Z}|\boldsymbol{Y}$ is once again Gaussian with mean and covariance matrix given by

$$\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y}) = \mathcal{O}\,\mu_{\text{pos}}(\boldsymbol{Y}), \qquad \Gamma_{\boldsymbol{Z}|\boldsymbol{Y}} = \mathcal{O}\,\Gamma_{\text{pos}}\,\mathcal{O}^\top. \tag{1.4}$$

The goal of this paper is to characterize statistically optimal, computationally efficient, and structure-exploiting approximations of the statistics of $\boldsymbol{Z}|\boldsymbol{Y}$ whenever the use of direct formulas such as (1.4) is challenging or impractical (perhaps due to the high computational complexity or excessive storage requirements). We will approximate $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ as a low-rank negative update of the prior covariance of the QoI. Optimality will hold with

*Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, {spantini, tcui, ymarz}@mit.edu.

‡Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO 80401, USA, ltenorio@mines.edu.

[1] Alternatively, we can generate quadratures or use any other method to perform integration under the posterior measure.

[2] $\boldsymbol{X}|\boldsymbol{Y}$ refers to a random variable distributed according to the measure of $\boldsymbol{X}$ conditioned on $\boldsymbol{Y}$.

respect to the natural geodesic distance on the manifold of symmetric and positive definite matrices [12]. The posterior mean, $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$, will be approximated as a low-rank function of the data and optimality will follow from the minimization of the Bayes risk for squared-error loss weighted by $\Gamma^{-1}_{\boldsymbol{Z}|\boldsymbol{Y}}$. The essence of these approximations is the restriction of the inference process to directions in the parameter space that are informed by the data relative to the prior *and* that are relevant to the QoI, by finding the leading generalized eigenpairs of a suitable matrix pencil.

This paper is an extension of the work on goal-oriented inference originally presented in [18] in a number of different ways. First of all, we will introduce the notion of *optimal approximation*, rather than exact computation, for both the posterior covariance matrix and the posterior mean of the QoI. We will propose computationally efficient algorithms to determine these optimal approximations. The complexity of our algorithms will scale with the intrinsic dimensionality of the goal-oriented problem—which here reflects the dimension of the parameter subspace that is simultaneously relevant to the QoI and informed by the data, as noted above. In particular, the full posterior distribution of the parameters need not be computed at any stage of the algorithms. This is in stark contrast to [18]. Moreover, we introduce the possibility to handle high-dimensional QoIs such as those arising from the discretization of a distributed stochastic process. This class of problems is extremely relevant in applications (see, e.g., Section 3).

The remainder of the paper is organized as follows. In Section 2 we introduce the statistically optimal approximations of the posterior statistics of the QoI. In Section 3 we use an inverse problem in heat transfer to illustrate the theory. In Section 4 we offer some concluding remarks. Appendix A contains the proofs of the main results of this paper.

2. Theory. We first focus on the approximation of the posterior covariance of the QoI, $\boldsymbol{Z}$. The cost of computing $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ according to (1.4) is dominated by the solution of p linear systems with coefficient matrix[3] Γ_{pos}^{-1} in order to determine $\Gamma_{\text{pos}}\mathcal{O}^\top$. Moreover, the storage requirements for $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ scale as $O(p^2)$. If the dimension of the QoI is inherently low, e.g., $p = O(1)$, then the use of direct formulas like (1.4) can be remarkably efficient. For instance, if we are just interested in the average of $\boldsymbol{X}$, i.e., $\boldsymbol{Z} \coloneqq \frac{1}{n}\sum_i X_i$, then the QoI is only one-dimensional and computing the posterior covariance of the QoI amounts to solving essentially a single linear system. As the dimension of the QoI increases, however, direct formulas like (1.4) become increasingly impractical due to high computational and storage complexities; in many cases of interest the dimension of the QoI can even be arbitrarily large! Consider the following simple example. If $\boldsymbol{X}$ represents a finite-dimensional approximation of a spatially distributed stochastic process (e.g., a temperature field), then the QoI could be the restriction of this process to a domain of interest. In this case, the QoI must be a finite-dimensional approximation to a spatially distributed process and thus, it can be arbitrarily high-dimensional depending on the chosen level of discretization of the process (we will revisit this example in Section 3). Thus, there is a clear need for new inference algorithms that can efficiently tackle these challenging problems.

Even though direct formulas like (1.4) can be intractable owing to the high-dimensional QoI, essential features of large-scale Bayesian 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 are typically finite, scarce, indirect, corrupted by noise, and related to the inversion parameters by the action of a forward operator that filters out some information [24, 8]. As a result, data are usually informative, relative to the prior, only about a low-dimensional subspace of the parameter space. That is, the relevant difference between prior and posterior distribution is confined to a low-dimensional subspace. This source of low-dimensional structure is key to the development of efficient Bayesian inference algorithms [11, 8] and plays a crucial role also when dealing with goal-oriented problems. In [24] we studied optimal approximations of the posterior covariance of the parameters as a fixed negative definite low-rank update of the prior covariance matrix with respect to the Förstner–Moonen metric[4]: the geodesic distance on the manifold of symmetric and positive definite matrices [12]. In particular, we

[3] In large-scale inverse problems only the action of the precision matrix Γ_{pos}^{-1} on a vector is usually available and it is not reasonable to expect direct factorizations of Γ_{pos}^{-1} such as, for instance, a Cholesky decomposition. Thus, the solution of linear systems with coefficient matrix Γ_{pos}^{-1} is often times necessarily iterative (e.g., Lanczos iteration [17]).

[4] For a pair of symmetric and positive definite matrices A, B, the Förstner metric, $d_{\mathcal{F}}(A, B)$, is defined in terms of the generalized eigenvalues, (σ_i), of the matrix pencil (A, B) as $d^2_{\mathcal{F}}(A, B) = \sum_i \log^2(\sigma_i)$. The Förstner metric leverages the geometry of the manifold of positive definite matrices and it satisfies important invariance properties:

$$d_{\mathcal{F}}(A, B) = d_{\mathcal{F}}(A^{-1}, B^{-1}) \quad \text{and} \quad d_{\mathcal{F}}(A, B) = d_{\mathcal{F}}(MAM^\top, MBM^\top), \tag{2.1}$$

for any nonsingular matrix M, making it an ideal metric to compare covariance matrices [12]. Moreover, optimality of the covariance approximation in the Förstner metric leads to optimality of the approximation in distribution with respect to familiar measures of similarities between probability distributions like the Kullback–Leibler divergence and the Hellinger distance [24, 20]. Notice, in

focused on the approximation class

$$\mathcal{M}_r = \{\Gamma_{\text{pr}} - KK^\top \succ 0 : \text{rank}(K) \leq r\} \tag{2.2}$$

of positive definite matrices that can be written as a low-rank update of the prior covariance matrix in order to leverage the low-dimensional structure of the prior-to-posterior update[5]. The following theorem characterizes the optimal approximation of Γ_{pos} (see [24] for a proof).

THEOREM 2.1 (Optimal posterior covariance approximation). *Let (δ_i^2, w_i) be the generalized eigenvalues–eigenvector pairs of the matrix pencil $(H, \Gamma_{\text{pr}}^{-1})$ with the ordering $\delta_i^2 \geq \delta_{i+1}^2$ and $H := G^\top \Gamma_{\text{obs}}^{-1} G$ as in (1.2). Then, a minimizer, $\widehat{\Gamma}_{\text{pos}}$, of the Förstner metric between Γ_{pos} and an element of $\mathcal{M}_r$ is given by*

$$\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top, \qquad KK^\top = \sum_{i=1}^{r} \delta_i^2 (1+\delta_i^2)^{-1} \, w_i \, w_i^\top, \tag{2.3}$$

where the distance between Γ_{pos} and the optimal approximation is

$$d_{\mathcal{F}}(\Gamma_{\text{pos}}, \widehat{\Gamma}_{\text{pos}}) = \sqrt{\sum_{i>r} \log^2(1+\delta_i^2)}. \tag{2.4}$$

Theorem 2.1 tells us that the optimal way to update the prior covariance matrix to yield an approximation of Γ_{pos} is along the generalized eigenvectors of the matrix pencil $(H, \Gamma_{\text{pr}}^{-1})$[6]. These eigenvectors are the directions that are most informed by the data and are obtained from a precise balance between forward model, measurement noise and prior information. This update is typically low-rank for precisely the same reasons discussed above: the data are informative relative to the prior only about a low-dimensional subspace of the parameter space [5]. Notice that (2.3) is not only an optimal and structure-exploiting approximation of Γ_{pos}, but it is also computationally efficient as the generalized eigenpairs of $(H, \Gamma_{\text{pr}}^{-1})$ can be easily computed using a matrix-free algorithm like a Lanczos iteration (including its block version) [17, 9] or a randomized SVD [15]. This approximation of Γ_{pos}, originally introduced in [11] for computational convenience and justified by intuitive arguments, has been deployed successfully in a number of extremely large-scale applications in Bayesian inversion [4]. It is our starting point for the analysis of goal-oriented linear inverse problems.

The combination of Theorem 2.1 with the direct formulas (1.4) suggests a first approximation strategy for the posterior covariance of the QoI: we just replace Γ_{pos} in (1.4) with the optimal approximation described by Theorem 2.1

$$\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}} \approx \widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}} := \mathcal{O}\,\widehat{\Gamma}_{\text{pos}}\,\mathcal{O}^\top = \mathcal{O}\,\Gamma_{\text{pr}}\,\mathcal{O}^\top - \mathcal{O}\,KK^\top\,\mathcal{O}^\top, \tag{2.5}$$

where the low-rank update $KK^\top$ is given by (2.3). Approximation (2.5) is already a big computational improvement over the direct formulas (1.4): there is no need to compute p linear systems, we just need to compute the leading eigenpairs of $(H, \Gamma_{\text{pr}}^{-1})$ with a matrix-free algorithm. The rank of the update depends on the dimension of the subspace of the parameter space that is informed the most by the data and thus, whenever the parameters represent a finite-dimensional approximation of a distributed stochastic process, the dimension of this subspace is eventually independent of the chosen level of discretization of the process (at least below a critical level of resolution). This feature of the approximation is essential if we want to deal with truly large-scale inverse problems. Notice that $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ in (2.5) need not be formed explicitly but can be easily stored in terms of the prior covariance matrix, the goal-oriented operator, and the low-rank update $KK^\top$. Moreover, it follows from results in [24] that we can easily obtain an expression for a square root of $\widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}}$. This allows efficient sampling from $\mathcal{N}(\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y}), \widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}})$ in high-dimensional problems.

Despite these favorable computational properties, the approximation (2.5) is still not satisfactory as it does not account explicitly for the goal-oriented feature of the problem. $\widehat{\Gamma}_{\text{pos}}$ in (2.5) is the optimal approximation of the posterior covariance of the parameters but is by no means tailored to the QoI. The pencil $(H, \Gamma_{\text{pr}}^{-1})$ used to

particular, that the distance induced by the Frobenius norm does not satisfy any of the aforementioned properties.

[5] Many approximate inference algorithms, especially in the context of Kalman filtering, exploit the class (2.2) to deliver an approximation of Γ_{pos} (e.g., [2]). These algorithms, however, are suboptimal in the sense defined by the forthcoming Theorem 2.1 (see [24] for further details and numerical examples).

[6] The properties of the pencil $(H, \Gamma_{\text{pr}}^{-1})$ have been studied extensively in the literature on classical regularization techniques for linear inverse problems (e.g., [16, 10, 6]). These papers, however, do not adopt a statistical approach to inversion and thus have not considered the optimal approximation of the posterior covariance matrix.

compute the approximation $\widehat{\Gamma}_{\text{pos}}$ according to Theorem 2.1 does not contain the goal-oriented operator. That is, the directions, (w_i), that define the optimal prior-to-posterior update in (2.3) are certainly the most informed by the data relative to the prior but need not be relevant at all to the QoI. For instance, some of the (w_i) could lie in the nullspace of the goal-oriented operator. Computing such eigenvectors would be a clear waste of computational resources and should be avoided. Of course, as the rank of the optimal prior-to-posterior update increases, the corresponding approximation, $\widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}}$, will continue to improve until eventually $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}} = \widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}}$. However, in the worst case scenario, $\widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}}$ will be a good approximation of $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ only as we start computing eigenpairs of $(H, \Gamma_{\text{pr}}^{-1})$ associated with the smallest nonzero generalized eigenvalues. This is clearly unacceptable as the overall complexity of the approximation algorithm would not depend on the nature of the goal-oriented operator. Therefore, the approximation (2.5) cannot possibly satisfy any reasonable optimality statement in the spirit of Theorem 2.1 and thus it calls for a proper modification.

The form of $\widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}}$ in (2.5) clearly shows that the posterior covariance of the QoI can be written as a low-rank update of the prior on $\boldsymbol{Z}$ whose marginal distribution is Gaussian and simply given by $\mathcal{N}(0, \Gamma_{\boldsymbol{Z}})$ with $\Gamma_{\boldsymbol{Z}} \coloneqq \mathcal{O}\,\Gamma_{\text{pr}}\,\mathcal{O}^\top$. This is perfectly consistent with our intuition of the Bayesian update: the data will update the prior distribution on the QoI only along certain directions. Thus, a structure-exploiting and computationally efficient approximation class for $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ is given by the set of positive definite matrices that can be written as a fixed maximum rank negative definite update of $\Gamma_{\boldsymbol{Z}}$:

$$\mathcal{M}_r^{\boldsymbol{Z}} = \{\Gamma_{\boldsymbol{Z}} - KK^\top \succ 0 : \text{rank}(K) \leq r\}. \tag{2.6}$$

Notice that the definition of $\mathcal{M}_r^{\boldsymbol{Z}}$ is analogous to that of $\mathcal{M}_r$ in (2.2).

We are now in a position to introduce one of the main results of this paper. The following theorem defines the optimal approximation of $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ and its proof can be found in Appendix A.

THEOREM 2.2 (Optimal approximation of the posterior covariance of the QoI). *Let (λ_i, q_i) be the generalized eigenpairs of the pencil:*

$$(G\,\Gamma_{\text{pr}}\,\mathcal{O}^\top\,\Gamma_{\boldsymbol{Z}}^{-1}\,\mathcal{O}\,\Gamma_{\text{pr}}\,G^\top\,,\,\Gamma_{\boldsymbol{Y}}) \tag{2.7}$$

with the ordering $\lambda_i \geq \lambda_{i+1} > 0$ and normalization $q_i^\top\,G\,\Gamma_{\text{pr}}\,\mathcal{O}^\top\,\Gamma_{\boldsymbol{Z}}^{-1}\,\mathcal{O}\,\Gamma_{\text{pr}}\,G^\top\,q_i = 1$, where $\Gamma_{\boldsymbol{Y}} \coloneqq \Gamma_{\text{obs}} + G\,\Gamma_{\text{pr}}\,G^\top$ is the covariance matrix of the marginal distribution of $\boldsymbol{Y}$. Then, a minimizer, $\tilde{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}}$, of the Förstner metric between $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ and an element of $\mathcal{M}_r^{\boldsymbol{Z}}$ is given by:

$$\tilde{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}} = \Gamma_{\boldsymbol{Z}} - KK^\top, \quad KK^\top = \sum_{i=1}^{r} \lambda_i\,\widehat{q}_i\widehat{q}_i^\top, \quad \widehat{q}_i \coloneqq \mathcal{O}\,\Gamma_{\text{pr}}\,G^\top q_i, \tag{2.8}$$

where the corresponding minimum distance is:

$$d_{\mathcal{F}}^2(\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}, \tilde{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}}) = \sum_{i>r} \ln^2(1 - \lambda_i). \tag{2.9}$$

The optimal approximation in Theorem 2.2 yields the best possible accuracy for any given rank of the prior-to-posterior update and, most importantly, never requires the full posterior covariance of the parameters. (This should be contrasted with [18].) The directions (q_i) that define the optimal update are just the leading eigenvectors of the matrix pencil $(G\,\Gamma_{\text{pr}}\,\mathcal{O}^\top\,\Gamma_{\boldsymbol{Z}}^{-1}\,\mathcal{O}\,\Gamma_{\text{pr}}\,G^\top, \Gamma_{\boldsymbol{Y}})$ and stem from a careful balance of all the ingredients of the goal-oriented inverse problem: forward model, measurement noise, prior information, and ultimate goals. Incorporating ultimate goals reduces the intrinsic dimensionality of the inverse problem: the rank of the optimal update (2.8) can only be lower than that of the suboptimal approximation introduced in (2.5) for any fixed approximation error. The quality of the optimal approximation as a function of the rank of the update can be monitored using the formula for the minimum distance given in (2.9).

Finding the leading generalized eigenpairs of (2.7) requires the solution of a Hermitian generalized eigenvalue problem [3]. Unfortunately, it is not easy to reduce (2.7) to a standard eigenvalue problem[7] since that would require the action of a square root of the matrix $\Gamma_{\boldsymbol{Y}} \coloneqq \Gamma_{\text{obs}} + G\,\Gamma_{\text{pr}}\,G^\top$ or of its inverse. Nevertheless, there exist a plethora of matrix-free algorithms to deal with possibly large-scale generalized eigenvalue problems: generalized Lanczos iteration [3, Section 5.5], randomized SVD type methods [22], manifold optimization algorithms [1], the

[7] Notice that this is often possible in the non goal-oriented case when dealing with the pencil $(H, \Gamma_{\text{pr}}^{-1})$ as the action of a square root of Γ_{pr}^{-1}, or of its inverse, is available in many cases of interest (e.g., [19]).

trace minimization algorithm [23] and the inverse free preconditioned Krylov subspace method [13] just to name a few. These algorithms require the iterative solution of linear systems associated with $\Gamma_{\boldsymbol{Y}}$ (in some cases to low accuracy [23, 13]). However, applying $\Gamma_{\boldsymbol{Y}}$ to a vector entails the evaluation of the possibly expensive forward model. Thus, these algorithms can lead to more expensive computations than in the non goal-oriented case (for a fixed dimension of the desired eigenspace). Nevertheless, the optimal approximation in Theorem 2.2 guarantees the minimum prior-to-posterior rank update for each given accuracy of the approximation and for each possible configuration of the inverse problem.

Another important consequence of the optimal approximation of $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ with respect to the Förstner metric is optimality in distribution whenever we assume exact knowledge of the posterior mean of the QoI. It follows from [24, Lemma 2.2] that the minimizer of the Hellinger distance (or the Kullback–Leibler divergence) between the posterior distribution of the QoI, $\mathcal{N}(\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y}), \Gamma_{\boldsymbol{Z}|\boldsymbol{Y}})$, and the approximation $\mathcal{N}(\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y}), \Gamma)$ for a matrix $\Gamma \in \mathcal{M}_r^{\boldsymbol{Z}}$ is given by the optimal approximation (2.8) defined in Theorem 2.2.

We conclude this theory section with an analysis of the optimal approximation of the posterior mean of the QoI. The cost of computing

$$\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y}) \coloneqq \mathcal{O}\,\mu_{\text{pos}}(\boldsymbol{Y}) = \mathcal{O}\,\Gamma_{\text{pos}}\,G^\top \Gamma_{\text{obs}}^{-1}\,\boldsymbol{Y} \tag{2.10}$$

for a single realization of the data is usually dominated by the solution of a single linear system associated with Γ_{pos}^{-1}, to determine $\mu_{\text{pos}}(\boldsymbol{Y})$. This task can be efficiently tackled with state-of-the-art matrix-free iterative solvers for symmetric linear systems (e.g., [3]) even for million-dimensional parameter spaces [4]. However, if one is interested in the fast computation of $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$ for multiple realizations of the data that are not known a priori, e.g., in the context of online inference, then the situation is quite different [7]. Solving a linear system to compute $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$ each time a new measurement is available might just be infeasible in practical applications. If the dimension of the QoI is small, say $p = O(1)$, then there is an easy solution to this problem. One can just precompute the matrix $M \coloneqq \mathcal{O}\,\Gamma_{\text{pos}}\,G^\top \Gamma_{\text{obs}}^{-1}$ in an offline stage and then compute the posterior mean of the QoI as $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y}) = M\,\boldsymbol{Y}$ each time a new realization of the data becomes available. The computational efficiency of this procedure breaks down as the dimension of the QoI increases. For instance, this is the case when the QoI is a finite-dimensional approximation to a distributed stochastic process. In this case, the matrix M would be large and dense. Storing such a matrix could be quite inefficient. Moreover, performing a dense matrix-vector product to compute $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y}) = M\,\boldsymbol{Y}$ might be more expensive than solving a single linear system associated with Γ_{pos}^{-1}. Thus, our goal is to characterize computationally efficient and statistically optimal approximations of $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$. In particular, we seek an approximation of $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$ as a low-rank linear function of the data[8], i.e., $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y}) \approx A\,\boldsymbol{Y}$ for some low-rank matrix A. Thus, computing $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y}) \approx A\,\boldsymbol{Y}$ for each new realization of the data is fast and computationally efficient. We define optimality of the approximation with respect to the Bayes risk for squared-error loss weighted by the posterior precision matrix of the QoI, i.e.,

$$\mathcal{B}(A) \coloneqq \mathbb{E}[\,\|\,A\,\boldsymbol{Y} - \boldsymbol{Z}\,\|^2_{\Gamma^{-1}_{\boldsymbol{Z}|\boldsymbol{Y}}}\,], \tag{2.11}$$

where $\mathcal{B}(A)$ denotes the Bayes risk associated with the matrix A, $\|\,\boldsymbol{v}\,\|^2_{\Gamma^{-1}_{\boldsymbol{Z}|\boldsymbol{Y}}} \coloneqq \boldsymbol{v}^\top\,\Gamma^{-1}_{\boldsymbol{Z}|\boldsymbol{Y}}\,\boldsymbol{v}$ for all vectors $\boldsymbol{v}$ and where the expectation is taken over the joint distribution of $\boldsymbol{Z}$ and $\boldsymbol{Y}$. The weighted Frobenius norm in (2.11) penalizes errors in the approximation of $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$ more strongly in directions of lower posterior variance. The result is that the approximation of $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$ is more likely to fall within the bulk of the posterior density of the QoI. The following theorem characterizes the optimal approximation of $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$ and is proved in Appendix A.

THEOREM 2.3 (Optimal approximation of the posterior mean of the QoI). *Let $(\lambda_i, q_i, \widehat{q}_i)$ be defined as in Theorem 2.2, $\boldsymbol{X} \in \mathbb{R}^n$, and consider the minimization of the following Bayes risk over the set of low-rank matrices:*

$$\min \quad \mathbb{E}[\,\|\,A\,\boldsymbol{Y} - \boldsymbol{Z}\,\|^2_{\Gamma^{-1}_{\boldsymbol{Z}|\boldsymbol{Y}}}\,], \qquad \text{s.t.} \quad \text{rank}(A) \le r \tag{2.12}$$

Then a minimizer of (2.12) is given by:

$$A^* = \sum_{i=1}^{r} \lambda_i\,\widehat{q}_i\,q_i^\top, \tag{2.13}$$

[8] Under the assumption of zero prior mean, $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$ is just a linear function of $\boldsymbol{Y}$. There is no loss of generality in assuming zero prior mean.

where the minimum Bayes risk is:

$$\mathcal{B}(A^*) = \mathbb{E}[\,\|\,A^*\,\boldsymbol{Y} - \boldsymbol{Z}\,\|^2_{\Gamma^{-1}_{\boldsymbol{Z}|\boldsymbol{Y}}}\,] = \sum_{i>r} \frac{\lambda_i}{1-\lambda_i} + n\,. \tag{2.14}$$

Notice that the optimal approximation of $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$ shown in (2.13) can be computed from the optimal approximation of $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ introduced in Theorem 2.2 for free. In particular, both the optimal approximations of the posterior mean and covariance of the QoI are quite accurate whenever we include generalized eigenvalues $\lambda \ll 1$ in the corresponding form of the approximations (cf. minimum loss (2.9) and Bayes risk (2.14)).

3. Numerical examples: CPU cooling. We consider a goal-oriented inference problem in heat transfer. We study the cooling of a CPU by means of a heat sink. The goal is to infer the temperature field over the CPU from local noisy temperature measurements over the heat sink. Figure 3.1 shows the problem set up: the three layers of different materials correspond, respectively, to the CPU ($\mathcal{D}_1$), a thin silicon layer that connects the CPU to the heat sink ($\mathcal{D}_2$), and an aluminum fin ($\mathcal{D}_3$). We denote by $\mathcal{D}$ the union of these domains. Each $\mathcal{D}_i$ represents a two-dimensional cross section of the material of constant width W along the horizontal direction and height L_i. We assume that no heat transfer happens along the third dimension. This is a common engineering approximation. Each material has a constant density ρ_i, a constant specific heat c_i, and a constant thermal conductivity k_i shown in the table at the right of Figure 3.1. We want to describe the evolution over time $t \in (0, t_{\text{end}}]$ of the temperature fields $\Theta^{(i)} : \mathcal{D}_i \to \mathbb{R}$ for $i = 1, 2, 3$.

3.1. Forward, observational and prior models. The time evolution of the temperature field $\Theta^{(i)}$, in the interior of the domain $\mathcal{D}_i$, is described by a linear time dependent PDE of the form

$$\rho_i\, c_i\, \partial_t\, \Theta^{(i)} = \operatorname{div}(k_i \nabla \Theta^{(i)}) \qquad (i = 1, \dots, 3), \tag{3.1}$$

where ∂_t denotes partial integration with respect to time and where we assume no volumetric heat production and the Fourier's law for the heat flux [14]. Equations 3.1 should be complemented with appropriate boundary and initial conditions to have a well-posed problem. We use the independent variables s_1 and s_2 to denote, respectively, the horizontal and vertical directions and let $\boldsymbol{s} = (s_1, s_2)$. The point $\boldsymbol{s} = (0, 0)$ corresponds to the lower left corner of $\mathcal{D}$. At the lower interface of $\mathcal{D}_1$ we impose a space-time dependent heat flux: $k_1\, \partial_{\vec{n}}\, \Theta^{(1)} = q(\boldsymbol{s}, t)$ for $\boldsymbol{s} \in \mathcal{D}_{1,\text{bottom}}$, where $\vec{n}$ refers always to the outward pointing normal and q is a given scalar function nonconstant in $\boldsymbol{s}$. At the interface between domains $\mathcal{D}_i$ and $\mathcal{D}_{i+1}$ we assume heat transfer by conduction with no thermal contact resistance: $k_i\, \partial_{\vec{n}}\, \Theta^{(i)} = k_{i+1}\, \partial_{\vec{n}}\, \Theta^{(i+1)}$ and $\Theta^{(i)} = \Theta^{(i+1)}$ for $\boldsymbol{s} \in \text{interface}(\mathcal{D}_i, \mathcal{D}_{i+1})$ and $i = 1, 2$. At the top, left and right boundaries of $\mathcal{D}_3$, we assume heat transfer by convection with a fluid at constant temperature Θ_∞: $-k_3\, \partial_{\vec{n}}\, \Theta^{(3)} = h_c(\Theta^{(3)} - \Theta_\infty)$ for $\boldsymbol{s} \in \mathcal{D}_{3,\text{top}} \cup \mathcal{D}_{3,\text{left}} \cup \mathcal{D}_{3,\text{right}}$, where h_c is a constant convection coefficient. Finally, we impose adiabatic conditions (no heat exchange) on the left and right boundaries of $\mathcal{D}_1$ and $\mathcal{D}_2$: $\partial_{\vec{n}}\, \Theta^{(i)} = 0$ for $\boldsymbol{s} \in \mathcal{D}_{i,\text{left}} \cup \mathcal{D}_{i,\text{right}}$ and $i = 1, 2$. We do not specify here the initial conditions as they will be the subject of the forthcoming inference problem.

We consider a finite element spatial approximation of the weak form of (3.1) by means of linear elements on simplices [21]. We denote by $\boldsymbol{\Theta}_h(t) \in \mathbb{R}^n$ the collection of temperature values at the finite element nodes on $\mathcal{D}$ at time $t \in (0, t_{\text{end}})$. Notice that $\boldsymbol{\Theta}_h$ satisfies a system of ODEs of the form $M\, \partial_t \boldsymbol{\Theta}_h(t) + A\, \boldsymbol{\Theta}_h(t) = \boldsymbol{f}(t)$, with $t \in (0, t_{\text{end}})$, for a suitable mass matrix M, stiffness matrix A, known time dependent forcing term $\boldsymbol{f}$ and initial conditions $\boldsymbol{\Theta}_{0h} \coloneqq \boldsymbol{\Theta}_h(0)$.

The initial conditions $\boldsymbol{\Theta}_{0h}$ for $t = 0$ are unknown and must be estimated from local measurements of the temperature field Θ at few locations in space and time. The locations of the sensors $s^1, \dots, s^N$ are shown in Figure 3.1 (black dots). Observations are collected every Δt time units for $t \in (0, t_{\text{end}})$. The first observation happens at time $t = \Delta t$ and we assume that there are M of these. We denote measurements at time $t_i = i\Delta t$ as $\hat{\boldsymbol{Y}}_i = \big[\Theta(s^1, i\Delta t), \dots, \Theta(s^N, i\Delta t)\big]$. We can concatenate the observations into a unique vector $\hat{\boldsymbol{Y}} = (\hat{\boldsymbol{Y}}_1, \dots, \hat{\boldsymbol{Y}}_M)$ such that $\hat{\boldsymbol{Y}} \in \mathbb{R}^d$. The actual observed measurements are corrupted with additive Gaussian noise: $\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2_{\text{obs}}\, I)$ with I being the identity matrix. Notice that $\hat{\boldsymbol{Y}}$ is an affine function of $\boldsymbol{\Theta}_{0h}$. This relationship can be made linear by a suitable redefinition of the data vector. Thus, we are lead to a linear Gaussian inverse problem in standard form, $\boldsymbol{Y} = G\, \boldsymbol{\Theta}_{0h} + \epsilon$, where G defines the forward operator, $\boldsymbol{\Theta}_{0h} \mapsto \hat{\boldsymbol{Y}}$, that can be evaluated implicitly by solving a heat equation with no forcing term and initial conditions $\boldsymbol{\Theta}_{0h}$ for a time interval necessary to collect the corresponding observations $\hat{\boldsymbol{Y}}$.

We adopt a statistical approach to inversion. To define the zero mean Gaussian prior distribution[9] on $\boldsymbol{\Theta}_{0h}$, we model $\boldsymbol{\Theta}_{0h}$ as a discretized solution of a stochastic PDE of the form

$$\gamma\left(\kappa^2\mathcal{I}-\triangle\right)\Theta(\boldsymbol{s})=\mathcal{W}(\boldsymbol{s}),\qquad \boldsymbol{s}\in\mathcal{D}, \tag{3.2}$$

where $\mathcal{W}$ is a white noise process, κ is a positive scalar parameter, $\triangle$ is the Laplacian operator and $\mathcal{I}$ is the identity operator. In particular, we exploit the explicit link between Gaussian Markov random fields with the Matérn covariance function and solutions to stochastic PDEs as outlined in [19]. Notice, in particular, that the action of a square root of the prior covariance matrix on a vector is readily available as the solution of an elliptic PDE on $\mathcal{D}$ and thus, it is scalable to very large inverse problems [19].

3.2. Goal-oriented linear inverse problem. We now introduce the goal-oriented feature of the problem. We assume that we are only interested in the initial temperature distribution over the CPU ($\mathcal{D}_1$). Let $\boldsymbol{Z}$ be the restriction of $\boldsymbol{\Theta}_{0h}$ to the domain of interest $\mathcal{D}_1$. Clearly, there exists a linear map between $\boldsymbol{Z}$ and $\boldsymbol{\Theta}_{0h}$. That is, $\boldsymbol{Z}=\mathcal{O}\,\boldsymbol{\Theta}_{0h}$ for some goal-oriented linear operator $\mathcal{O}\in\mathbb{R}^{p\times n}$ with $p\ll n$. Thus, we have a linear Gaussian goal-oriented inverse problem as introduced in section 2 (we denote the parameters by $\boldsymbol{\Theta}_{0h}$):

$$\begin{cases}\boldsymbol{Y}=G\,\boldsymbol{\Theta}_{0h}+\epsilon\\ \boldsymbol{Z}=\mathcal{O}\,\boldsymbol{\Theta}_{0h}\end{cases} \tag{3.3}$$

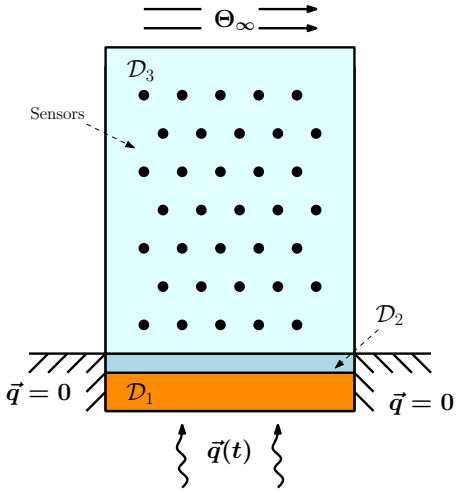
where both the marginal distribution of $\boldsymbol{\Theta}_{0h}$ and the likelihood $\boldsymbol{Y}|\boldsymbol{\Theta}_{0h}$ are specified. In particular, we choose a finite element discretization of the temperature field such that $\boldsymbol{\Theta}_{0h}\in\mathbb{R}^{2400}$ and $\boldsymbol{Z}\in\mathbb{R}^{370}$. Our goal is to characterize optimal approximations of the posterior statistics of the QoI, $\boldsymbol{Z}|\boldsymbol{Y}$, for a given set of observations (Figure 3.2 (*left*)). In this case, computing the posterior distribution of the QoI using direct formulas like (1.4) is infeasible as the QoI is a finite-dimensional approximation to a distributed stochastic process, $\Theta(0)|_{\mathcal{D}_1}$, and can be arbitrarily high-dimensional depending on the chosen level of discretization of the process. Thus, we need appropriate dimensionality reduction techniques in order to tackle this challenging inference task as explained in section 2.

The configuration of this problem highlights a crucial aspect of dimensionality reduction of goal-oriented inverse problems. Ideally we would position the measurement sensors on $\mathcal{D}_1$ since we are interested in inferring the temperature field on the CPU. However, due to clear geometrical constraints, we are forced to place our sensors on the heat sink ($\mathcal{D}_3$). As a result, observations are much more informative about the parameters in $\mathcal{D}_3$ rather than in $\mathcal{D}_1$. We see a hint of this by looking at Figure 3.2 (*right*). Figure 3.2 (*right*) shows the normalized difference between prior and posterior variance $(\mathbb{V}\mathrm{ar}(\boldsymbol{\Theta}_{0h})-\mathbb{V}\mathrm{ar}(\boldsymbol{\Theta}_{0h}|\boldsymbol{Y}))/\mathbb{V}\mathrm{ar}(\boldsymbol{\Theta}_{0h})$. In particular, notice that the prior variance is reduced the most in a neighborhood of the sensor locations in $\mathcal{D}_3$ and this makes intuitive sense as the collected data will be increasingly less informative as we move away from the sensors.

We first focus on the approximation of the posterior covariance of the QoI. If we use the suboptimal approximation introduced in (2.5), motivated by the optimality results presented in [24] for the non goal-oriented case, then we have to pay a considerable computational price as a result of the data being informative about directions in the parameter space that do not matter to the QoI. This is clear from the numerical results shown in Figure 3.3 (*left*). Notice that if we try to approximate the posterior covariance, Γ_{pos}, of $\boldsymbol{\Theta}_{0h}$ by its optimal approximation, $\widehat{\Gamma}_{\text{pos}}=\Gamma_{\text{pr}}-KK^\top$, introduced in [11, 24] and shown in Theorem 2.1, then the convergence of the approximation is rather slow (cf. blue dotted line in Figure 3.3 (*left*)). This is because there are many data informed directions in the parameter space (notice the multitude of sensors on the heat sink in Figure 3.1 (*left*)). If we use $\widehat{\Gamma}_{\text{pos}}$ to yield an approximation of the actual posterior covariance of interest, $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$, by means of the approximation $\widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}}=\mathcal{O}\,\widehat{\Gamma}_{\text{pos}}\,\mathcal{O}^\top$ as shown in (2.5), then the convergence of this approximation is still slow (cf. green solid line in Figure 3.3 (*left*)). This slow convergence can be easily justified. The optimal approximation, $\widehat{\Gamma}_{\text{pos}}=\Gamma_{\text{pr}}-KK^\top$, of Γ_{pos} will account first for those directions that are the most informed by the data. These directions correspond to modes with features near the locations of the sensors in $\mathcal{D}_3$. Thus, these modes will be little informative about the parameters in the region of interest ($\mathcal{D}_1$). This explains the slow convergence of the solid green line in Figure 3.3 (*left*). On the other hand, if we use the optimal approximation of $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ defined in Theorem 2.2, then the convergence of the approximation is remarkably fast (red solid line in Figure 3.3 (*left*)): we just need to update $\Gamma_{\boldsymbol{Z}}$ along a handful of directions, say 20, to achieve a satisfactory approximation of $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$.

[9] We assume, without loss of generality, zero mean of the parameters. In fact, if we are given a statistical model of the form $\boldsymbol{Y}=G\,\boldsymbol{\Theta}_{0h}+\epsilon$ where $\boldsymbol{\Theta}_{0h}\sim\mathcal{N}(\mu_{\text{pr}},\Gamma_{\text{pr}})$ has a nonzero prior mean, then we can trivially rewrite the statistical model as $\hat{\boldsymbol{Y}}:=\boldsymbol{Y}-G\mu_{\text{pr}}=G\,(\boldsymbol{\Theta}_{0h}-\mu_{\text{pr}})+\epsilon$ for a modified data vector $\hat{\boldsymbol{Y}}$ and infer, equivalently, a zero prior mean process $\boldsymbol{\Theta}_{0h}-\mu_{\text{pr}}\sim\mathcal{N}(0,\Gamma_{\text{pr}})$.

Notice also that the optimal approximation of the posterior mean of the QoI as a low-rank linear function of the data introduced in Theorem 2.3 converges quite fast as a function of the rank of the approximation (Figure 3.3 (*right*)). Once a low-rank approximation of the form (2.13) is available, then it is possible to compute a very good approximation of $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$, for each new realization of the data $\boldsymbol{Y}$, by just performing a low-rank (20 in this case) matrix-vector product as opposed to the solution of an expensive linear system.



Material	α_i at 20 °C	Domain
—	m/s^2	—
Copper	1.11×10^{-4}	$\mathcal{D}_1$
Silicon	8.8×10^{-5}	$\mathcal{D}_2$
Aluminium	8.42×10^{-5}	$\mathcal{D}_3$

Fig. 3.1: (*left*) CPU cooling problem. Inversion for the initial temperature field on $\mathcal{D}_1$ given noisy sparse temperature measurements in space and time on an aluminium heat sink ($\mathcal{D}_3$). The figure shows the problem configuration, the locations of the sensors (black dots), and the boundary conditions for the heat equation that describe the time evolution of the temperature field on the domain. $\mathcal{D} := \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$. (*right*) Material properties of the different layers $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$.

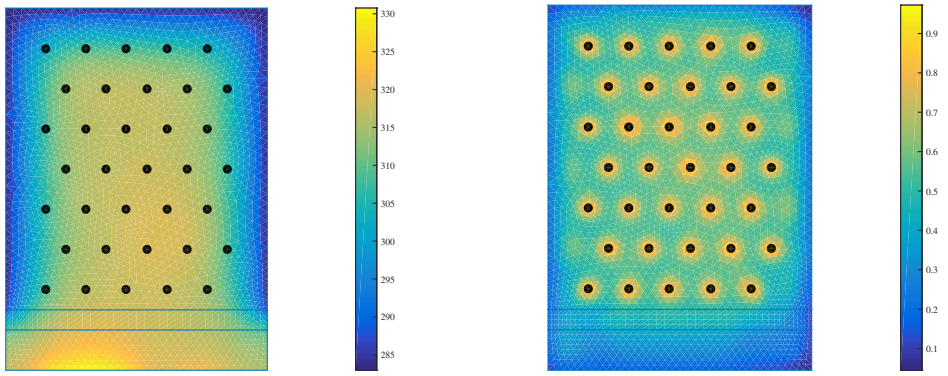


Fig. 3.2: (*left*) Initial temperature field used to generate synthetic data according to the observational set up described in the CPU cooling inverse problem. We remark that the initial temperature field used to generate synthetic data was not drawn from the marginal distribution of $\boldsymbol{\Theta}_{0h}$: it corresponds to a finer discretization of the continuous stochastic process $\Theta(0)$ compared to $\boldsymbol{\Theta}_{0h}$. (*right*) Normalized difference of prior to posterior variance of the parameters, i.e., $(\mathbb{V}\mathrm{ar}(\boldsymbol{\Theta}_{0h}) - \mathbb{V}\mathrm{ar}(\boldsymbol{\Theta}_{0h}|\boldsymbol{Y}))/\mathbb{V}\mathrm{ar}(\boldsymbol{\Theta}_{0h})$. Notice that the regions of greatest relative decrease of prior variance are localized in a neighborhood of the sensor locations (black dots).

4. Conclusions. In this paper we proposed statistically optimal and computationally efficient approximations of the posterior statistics of the QoI in a goal-oriented linear Gaussian inverse problem. The posterior covariance of the QoI is approximated as a low-rank negative update of the prior covariance of the QoI. Optimality holds with respect to the Förstner metric: the natural geodesic distance on the manifold of symmetric and positive definite matrices. The posterior mean of the QoI is approximated as a low-rank function of the data and optimality follows from the minimization of the Bayes risk for squared-error loss weighted by the posterior

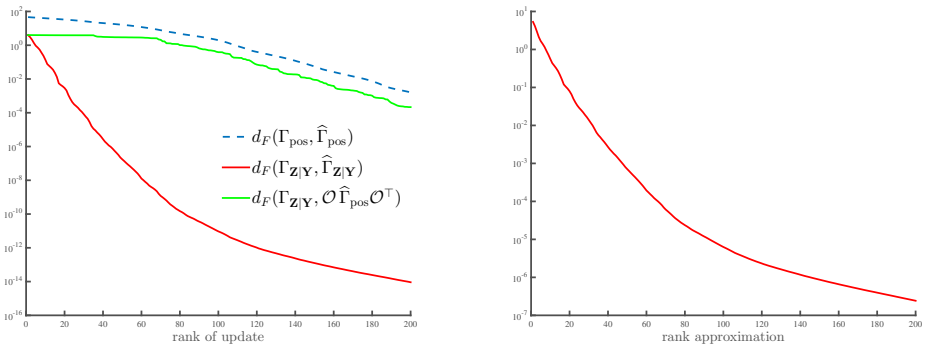


Fig. 3.3: (*left*) Convergence of the covariance approximations in the Förstner metric. The blue dotted line shows the Förstner distance between the covariance of $\boldsymbol{\Theta}_{0h}|\boldsymbol{Y}$, i.e., Γ_{pos}, and its optimal approximation introduced in [24], $\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top$, as a function of the rank of K (see Theorem 2.1). The red line shows the Förstner distance between the posterior covariance of the QoI, $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$, and its optimal approximation introduced in Theorem 2.2, $\widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}} = \Gamma_{\boldsymbol{Z}} - KK^\top$, as a function of the rank of K. Finally, the green line shows the Förstner distance between $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ and the suboptimal approximation (2.5) obtained as $\mathcal{O}\,\widehat{\Gamma}_{\text{pos}}\,\mathcal{O}^\top$ where $\widehat{\Gamma}_{\text{pos}}$ is the optimal approximation of Γ_{pos} introduced in [24]. (*right*) Error in the optimal approximation of the posterior mean of the QoI, $\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y})$. The error is measured as the square root of $\mathbb{E}[\,\|\,\mu_{\boldsymbol{Z}|\boldsymbol{Y}}(\boldsymbol{Y}) - A^*\,\boldsymbol{Y}\|^2_{\Gamma^{-1}_{\boldsymbol{Z}|\boldsymbol{Y}}}]$ and is a function of $\text{rank}(A^*)$.

precision matrix of the QoI. These optimal approximations avoid computations of the full posterior distribution of the parameters and focus only on directions in the parameter space that are informed by the data and that are relevant to the QoI. These directions are obtained as the leading generalized eigenvectors of a suitable matrix pencil and stem from a careful balance between all the ingredients of the goal oriented inverse problem: prior information, forward model, measurement noise and ultimate goals. Future work includes the extension of these optimality results to the nonlinear case.

Acknowledgments. This work was supported by the US Department of Energy, Office of Advanced Scientific Computing (ASCR), under grant numbers DE-SC0003908 and DE-SC0009297.

Appendix A. Technical results. The following lemma will be useful in proving theorems 2.2 and 2.3.

LEMMA A.1. *A linear Gaussian model consistent with* (1.4) *is given by:* $\boldsymbol{Y} = G\,\mathcal{O}_\dagger\,\boldsymbol{Z} + \delta$, *with* $\boldsymbol{Z} \sim \mathcal{N}(0, \Gamma_{\boldsymbol{Z}})$, $\mathcal{O}_\dagger := \Gamma_{\text{pr}}\mathcal{O}^\top\Gamma^{-1}_{\boldsymbol{Z}}$ *and* $\delta \sim \mathcal{N}(0, \Gamma_\delta)$ *is independent of* $\boldsymbol{Z}$ *with* $\Gamma_\delta := \Gamma_{\text{obs}} + G(\Gamma_{\text{pr}} - \Gamma_{\text{pr}}\mathcal{O}^\top\,\Gamma^{-1}_{\boldsymbol{Z}}\,\mathcal{O}\Gamma_{\text{pr}})G^\top$.

Proof. Consider the identity $\boldsymbol{Y} = G\,\boldsymbol{X} + \varepsilon = G\,\mathcal{O}_\dagger\,\mathcal{O}\,\boldsymbol{X} + G\,(I - \mathcal{O}_\dagger\,\mathcal{O})\,\boldsymbol{X} + \varepsilon = G\,\mathcal{O}_\dagger\,\boldsymbol{Z} + \delta$, where $\mathcal{O}_\dagger := \Gamma_{\text{pr}}\mathcal{O}^\top\Gamma^{-1}_{\boldsymbol{Z}}$ and $\delta := G\,(I - \mathcal{O}_\dagger\,\mathcal{O})\,\boldsymbol{X} + \varepsilon$. A simple computation shows that $\mathbb{E}[(I - \mathcal{O}_\dagger\,\mathcal{O})\boldsymbol{X}\,\boldsymbol{Z}^\top] = 0$. Hence, $(I - \mathcal{O}_\dagger\,\mathcal{O})\boldsymbol{X}$ and $\boldsymbol{Z}$ are uncorrelated, and, more importantly, independent since they are also jointly Gaussian. It follows that δ and $\boldsymbol{Z}$ are also independent since ε was independent of $\boldsymbol{X}$ and $\boldsymbol{Z} = \mathcal{O}\,\boldsymbol{X}$. In the hypothesis of zero prior mean, the mean of δ is also zero. Moreover, $\Gamma_\delta = \mathbb{V}\text{ar}[G\,(I - \mathcal{O}_\dagger\,\mathcal{O})\boldsymbol{X}] + \mathbb{V}\text{ar}[\varepsilon]$ since $\boldsymbol{X}$ and ε are independent. Simple algebra leads to the particular form of Γ_δ. ∎

Proof of Theorem 2.2. By applying [24, Theorem 2.3] to the linear Gaussian model defined in Lemma A.1, we know that a minimizer, $\widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}}$, of the Förstner metric between $\Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}$ and an element of $\mathcal{M}^{\boldsymbol{Z}}_r$ is given by: $\widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}} = \Gamma_{\boldsymbol{Z}} - \sum_{i=1}^r \eta_i^2(1+\eta_i^2)^{-1}\,\widehat{q}_i\widehat{q}_i^\top$, where $(\eta_i^2, \widehat{q}_i)$ are the generalized eigenvalue-eigenvector pairs of the pencil $(H_{\boldsymbol{Z}}, \Gamma^{-1}_{\boldsymbol{Z}})$, with the ordering $\eta_i^2 \geq \eta_{i+1}^2$, the normalization $\widehat{q}_i^\top\Gamma^{-1}_{\boldsymbol{Z}}\widehat{q}_i = 1$ and where $H_{\boldsymbol{Z}} := \mathcal{O}_\dagger^\top\,G^\top\,\Gamma^{-1}_\delta\,G\,\mathcal{O}_\dagger$ is the Hessian of the negative log–likelihood $\boldsymbol{Y}|\boldsymbol{Z} \sim \mathcal{N}(G\,\mathcal{O}_\dagger, \Gamma_\delta)$. Moreover, [24, Theorem 2.3] implies that the Förstner metric, at optimality, is given by: $d^2_{\mathcal{F}}(\widehat{\Gamma}_{\boldsymbol{Z}|\boldsymbol{Y}}, \Gamma_{\boldsymbol{Z}|\boldsymbol{Y}}) = \sum_{i>r}\ln^2(1+\eta_i^2)$ and that the minimizer is unique if the first r eigenvalues of the pencil $(H_{\boldsymbol{Z}}, \Gamma^{-1}_{\boldsymbol{Z}})$ are distinct. Now let (λ_i, q_i) be defined as in Theorem 2.2. A simple computation shows that $(\,\lambda_i(1-\lambda_i)^{-1}, q_i)$ are the generalized eigenvalue-eigenvector pairs of the pencil $(G\,\Gamma_{\text{pr}}\,\mathcal{O}^\top\,\Gamma^{-1}_{\boldsymbol{Z}}\,\mathcal{O}\,\Gamma_{\text{pr}}\,G^\top\,,\,\Gamma_\delta)$. Moreover, $(\,\lambda_i(1-\lambda_i)^{-1}\,, q_i)$ are also the generalized eigenpairs of the pencil $(H_{\boldsymbol{Z}}\,\mathcal{O}\,\Gamma_{\text{pr}}\,G^\top\,,\,\Gamma^{-1}_{\boldsymbol{Z}}\,\mathcal{O}\,\Gamma_{\text{pr}}\,G^\top)$. Then, it must be that $\eta_i^2 = \lambda_i(1-\lambda_i)^{-1}$ and $\widehat{q}_i = \alpha\,\mathcal{O}\,\Gamma_{\text{pr}}\,G^\top\,q_i$ for some real $\alpha > 0$ since $(\eta_i^2, \widehat{q}_i)$ are the generalized eigenpairs of $(H_{\boldsymbol{Z}}, \Gamma^{-1}_{\boldsymbol{Z}})$. Given the normalizations $\widehat{q}^\top\Gamma^{-1}_{\boldsymbol{Z}}\widehat{q} = 1$ and $q_i^\top(\,G\,\Gamma_{\text{pr}}\,\mathcal{O}^\top\,\Gamma^{-1}_{\boldsymbol{Z}}\,\mathcal{O}\,\Gamma_{\text{pr}}\,G^\top\,)q_i = 1$, it must be $\alpha = 1$. It is easy to see, using a counting argument, that the $(\widehat{q}_i)$ are indeed all the generalized eigenvectors of the pencil $(H_{\boldsymbol{Z}}, \Gamma^{-1}_{\boldsymbol{Z}})$ associated with positive eigenvalues. Simple

algebra then leads to (2.8) and (2.9). □

Proof of Theorem 2.3. By applying [24, Theorem 4.1] to the linear Gaussian model defined in Lemma A.1, we know that a minimizer of 2.12 is given by: $A^* = \sum_{i=1}^{r} \eta_i (1+\eta_i^2)^{-1} \widehat{q}_i \widehat{v}_i^\top$, where $(\eta_i^2, \widehat{q}_i)$ are generalized eigenvalue-eigenvector pairs of the pencil $(H_{\boldsymbol{Z}}, \Gamma_{\boldsymbol{Z}}^{-1})$ with normalization $\widehat{q}_i^\top \Gamma_{\boldsymbol{Z}}^{-1} \widehat{q}_i = 1$, whereas $(\widehat{v}_i)$ are generalized eigenvectors of the pencil $(G\,\mathcal{O}_\dagger\,\Gamma_{\boldsymbol{Z}}\,\mathcal{O}_\dagger^\top\,G^\top, \Gamma_\delta)$ with normalization $\widehat{v}_i^\top \Gamma_\delta\, \widehat{v}_i = 1$. Moreover, [24, Theorem 4.1] tells us that the Bayes risk associated with the minimizer A^* can be written as: $\mathbb{E}[\,\| A^* \boldsymbol{Y} - \boldsymbol{Z} \|^2_{\Gamma^{-1}_{\boldsymbol{Z}|\boldsymbol{Y}}}\,] = \sum_{i>r} \eta_i^2 + n$, where n is the dimension of the parameter space. The fact that the vectors $(\widehat{q}_i)$ can be written as $\widehat{q}_i = \mathcal{O}\,\Gamma_{\rm pr}\,G^\top\,q_i$ was proved in Theorem 2.2. Furthermore, in the proof of Theorem 2.2 we showed that: $\eta_i^2 = \lambda_i (1-\lambda_i)^{-1}$. Using the latter expression we can rewrite the minimizer as: $A^* = \sum_{i=1}^{r} \sqrt{\lambda_i\,(1-\lambda_i)}\, \widehat{q}_i \widehat{v}_i^\top$. If $(\widehat{v}_i)$ are generalized eigenvectors of the pencil $(G\,\mathcal{O}_\dagger\,\Gamma_{\boldsymbol{Z}}\,\mathcal{O}_\dagger^\top\,G^\top, \Gamma_\delta)$, then they must also be generalized eigenvectors of the pencil $(G\,\Gamma_{\rm pr}\,\mathcal{O}^\top\,\Gamma_{\boldsymbol{Z}}^{-1}\,\mathcal{O}\,\Gamma_{\rm pr}\,G^\top\,,\,\Gamma_{\boldsymbol{Y}})$. In particular, it has to be $\widehat{v}_i = \alpha\, q_i$ for some real $\alpha > 0$. Given the normalizations $q_i^\top\,G\,\Gamma_{\rm pr}\,\mathcal{O}^\top\,\Gamma_{\boldsymbol{Z}}^{-1}\,\mathcal{O}\,\Gamma_{\rm pr}\,G^\top\,q_i = 1$ and $\widehat{v}_i^\top \Gamma_\delta\, \widehat{v}_i = 1$, it must be: $\alpha = \lambda_i^{1/2}\,(1-\lambda_i)^{-1/2}$. Simple algebra then leads to (2.13). □

REFERENCES

[1] P. A. Absil, C. G. Baker, and K. A. Gallivan, *A truncated-CG style method for symmetric generalized eigenvalue problems*, Journal of computational and applied mathematics, 189 (2006), pp. 274–285.

[2] H. Auvinen, J. M. Bardsley, H. Haario, and T. Kauranne, *Large-scale Kalman filtering using the limited memory BFGS method*, Electronic Transactions on Numerical Analysis, 35 (2009), pp. 217–233.

[3] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, vol. 11, SIAM, 2000.

[4] T. Bui-Than, 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.

[5] 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), p. 055001.

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

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

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

[9] 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 Decision and Control including the 13th Symposium on Adaptive Processes, 1974 IEEE Conference on, IEEE, 1974, pp. 505–509.

[10] L. Dykes and L. Reichel, *Simplified GSVD computations for the solution of linear discrete ill-posed problems*, Journal of Computational and Applied Mathematics, 255 (2014), pp. 15–27.

[11] 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 Journal on Scientific Computing, 33 (2011), pp. 407–432.

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

[13] G. H. Golub and Q. Ye, *An inverse free preconditioned Krylov subspace method for symmetric generalized eigenvalue problems*, SIAM Journal on Scientific Computing, 24 (2002), pp. 312–334.

[14] G. Guglielmini and C. Pisoni, *Elementi di trasmissione del calore*, Veschi, 1990.

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

[16] P. C. Hansen, *Regularization, GSVD and truncated GSVD*, BIT Numerical Mathematics, 29 (1989), pp. 491–504.

[17] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, United States Governm. Press Office, 1950.

[18] C. Lieberman and K. Willcox, *Goal-oriented inference: Approach, linear theory, and application to advection diffusion*, SIAM Journal on Scientific Computing, 34 (2012), pp. A1880–A1904.

[19] F. Lindgren, H. Rue, and J. Lindström, *An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach*, Journal of the Royal Statistical Society: Series B, 73 (2011), pp. 423–498.

[20] L. Pardo, *Statistical Inference Based on Divergence Measures*, CRC Press, 2005.

[21] A. Quarteroni and A. Valli, *Numerical approximation of partial differential equations*, vol. 23, Springer Science & Business Media, 2008.

[22] A. K. Saibaba and P. K. Kitanidis, *Randomized square-root free algorithms for generalized hermitian eigenvalue problems*, arXiv preprint arXiv:1307.6885, (2013).

[23] A. H. Sameh and J. A. Wisniewski, *A trace minimization algorithm for the generalized eigenvalue problem*, SIAM Journal on Numerical Analysis, 19 (1982), pp. 1243–1259.

[24] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk, *Optimal low-rank approximations of Bayesian linear inverse problems*, SIAM Journal on Scientific Computing, 37 (2015), pp. A2451–A2487.