BAYESCG AS AN UNCERTAINTY AWARE VERSION OF CG*

TIM W. REID†, ILSE C. F. IPSEN†, JON COCKAYNE‡, AND CHRIS J. OATES§

Abstract. The Bayesian Conjugate Gradient method (BayesCG) is a probabilistic generalization of the Conjugate Gradient method (CG) for solving linear systems with real symmetric positive definite coefficient matrices. Our CG-based implementation of BayesCG under a structure-exploiting prior distribution represents an 'uncertainty-aware' version of CG. Its output consists of CG iterates and posterior covariances that can be propagated to subsequent computations. The covariances have low-rank and are maintained in factored form. This allows easy generation of accurate samples to probe uncertainty in downstream computations. Numerical experiments confirm the effectiveness of the low-rank posterior covariances.

Key words. Symmetric positive semi-definite matrix, Krylov space method, Gaussian probability distribution, Bayesian inference, covariance matrix, mean, Moore-Penrose inverse, projectors in semi-definite inner products

AMS subject classifications. 65F10, 62F15, 65F50, 15A06, 15A10

1. Introduction.

The solution of linear systems

$$\mathbf{A}\mathbf{x}_* = \mathbf{b}, \tag{1.1}$$

with symmetric positive definite coefficient matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is an important problem in computational science and engineering. For large and sparse matrices $\mathbf{A}$, the preferred solver is the Conjugate Gradient method (CG) [26, 31]. This is a Krylov subspace method that, starting from a user-specified initial guess $\mathbf{x}_0$, produces iterates $\mathbf{x}_m$ that, the user hopes, ultimately converge to the solution $\mathbf{x}_*$. In practice, CG is terminated early, once the residual $\|\mathbf{b} - \mathbf{A}\mathbf{x}_m\|$ is sufficiently small in some norm. Early termination introduces a source of uncertainty since the solution $\mathbf{x}_*$ has not been exactly computed.

We seek to create an 'uncertainty aware' version of CG that models the uncertainty in our knowledge of $\mathbf{x}_*$ due to early termination. From the UQ perspective, this represents an instance of model discrepancy with epistemic uncertainties. Our motivation is to understand how the accuracy of the CG output $\mathbf{x}_m$ affects downstream computations in a *computational pipeline* [12, Section 5], [25], that is, sequences of computations where the output of one computation is the input to another [7, 23, 40, 43, 44]. Traditional normwise CG error estimates are inadequate, because subsequent computations may not be able to make effective use of them. In contrast, a probabilistic model of the uncertainty, in the form of a distribution, can be propagated so that downstream computations can sample from the distribution to probe the effect of uncertainty on their own computations.

This is the mission of *probabilistic numerics*[1]: Modelling the uncertainty in de-

*Submitted to the editors

Funding: The work was supported in part by NSF grant DMS-1745654 (TWR, ICFI), NSF grant DMS-1760374 and DOE grant DE-SC0022085 (ICFI), and the Lloyd's Register Foundation Programme on Data Centric Engineering at the Alan Turing Institute (CJO).

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA, (twreid@alumni.ncsu.edu, ipsen@ncsu.edu)

‡Department of Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK (jon.cockayne@soton.ac.uk)

§School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne NE1 7RU, UK (chris.oates@ncl.ac.uk)

[1]https://www.probabilistic-numerics.org/

terministic computations with a probabilistic treatment of the errors [25, 42]. The origins of probabilistic numerics can be traced back to Poincaré [42], while a rigorous modern perspective is established in [12]. Probabilistic numerical methods have been developed for Bayesian optimization [38], subsequently applied to hyperparameter optimization in machine learning [46]; numerical integration [4, 14, 29], sparse Cholesky decompositions [45], and solution of ordinary and partial differential equations [8, 34, 41, 52].

In the context of linear solvers, probabilistic solvers posit a *prior distribution* representing initial epistemic uncertainty about a quantity of interest, which can be the solution [1, 7, 9, 53] or the matrix inverse [1, 2, 24]. They then condition on the finite amount of information obtained during m iterations to produce a *posterior distribution* that reflects the reduced uncertainty [9, Section 1.2], [42]. The interpretation of CG as a probabilistic solver was pioneered in the context of optimization [24], followed by the development of the *Bayesian Conjugate Gradient method (BayesCG)* [9] as a general purpose solver in statistics. However, current versions of BayesCG have two drawbacks: they are computationally expensive; and their posterior distributions do not model the uncertainty accurately.

1.1. Contributions and outline. We propose an efficient uncertainty-aware CG implementation in the form of BayesCG (Algorithm 3.1), and establish its proper foundation within probabilistic numerics (sections 2 and 3).

We design a new *Krylov prior* distribution for BayesCG, which is motivated by the *Krylov subspace prior* [9, section 4.1], which is a *non-singular* structured prior based on Krylov spaces, whose posterior distributions are expensive and not always meaningful. In contrast, our new Krylov prior is generally singular, depends on quantities computed by CG, and produces low-rank posteriors that lend themselves to efficient sampling in downstream computations. We proceed in two steps.

1. Extension of BayesCG to singular prior covariances (section 2).
   We show that under reasonable assumptions, the theoretical and computational properties of BayesCG from [9] extend to prior covariances that are singular. This extension to singular priors paves the way for an efficient BayesCG implementation that produces meaningful posteriors. Auxiliary results and technical proofs are postponed to the end (Appendices A and B).
2. Introduction of the new Krylov prior and its properties (section 3).
   This singular prior covariance exploits structure and adapts to BayesCG, with posteriors whose means are identical to the corresponding CG iterates, and whose covariances describe a realistic level of uncertainty. The posterior covariances are maintained in factored form, and are therefore highly accurate and easy to approximate, as confirmed by numerical experiments (section 4).

1.2. Notation. Bold uppercase letters, like $\mathbf{A}$, represent matrices, with $\mathbf{I}$ denoting the identity. The Moore-Penrose inverse of $\mathbf{A}$ is $\mathbf{A}^\dagger$. Bold lowercase letters, like $\mathbf{x}_*$, represent vectors; italic lowercase letters, like α, scalars; and italic uppercase letters, like X_0, random variables. A multivariate Gaussian distribution with mean $\mathbf{x}$ and covariance $\mathbf{\Sigma}$ is denoted by $\mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$, and $X \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$ is a Gaussian random variable. We assume exact arithmetic throughout the theoretical sections 2 and 3.

2. Introduction to BayesCG with singular priors. We extend the applicability of BayesCG from definite to semi-definite prior covariances, and discuss the theory (section 2.1), recursive computation of posterior distributions (section 2.2), and choices for prior distributions (section 2.3).

2.1. Theoretical properties of BayesCG under singular priors. We derive expressions for the BayesCG posterior means and covariances under singular priors (Theorem 2.1), express the posteriors in terms of projectors (Theorem 2.4), and establish the optimality of the posterior means (Theorem 2.6). The proofs are analogous to earlier proofs for non-singular priors in [1, 9], and relegated to Appendix A and the supplement.

BayesCG computes posterior distributions $\mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$ by conditioning the prior $\mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$ on information from $m \leq n$ linearly independent search directions $\mathbf{S}_m$. Specifically, the posterior is the distribution of the random variable $X \sim \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$ conditioned on the random variable $Y = \mathbf{S}_m^T \mathbf{A} X$ taking the value $\mathbf{S}_m^T \mathbf{A} \mathbf{x}_*$. The conditioning relies on two properties of Gaussian distributions:
(i) *Stability:* linear transformations of Gaussians remain Gaussian [39, Section 1.2].
(ii) *Conjugacy:* posteriors from Gaussian priors conditioned under linear information remain Gaussian [51, Theorem 6.20].

We start with the extension of BayesCG to singular priors.

THEOREM 2.1 (Extension of [9, Proposition 1]). *Let $\mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$ be a prior with a symmetric positive semi-definite covariance $\mathbf{\Sigma}_0 \in \mathbb{R}^{n \times n}$. Let $m \leq \operatorname{rank}(\mathbf{\Sigma}_0)$, and let the matrix of search directions $\mathbf{S}_m \equiv \begin{bmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_m \end{bmatrix} \in \mathbb{R}^{n \times m}$ have linearly independent columns so that $\mathbf{\Lambda}_m \equiv \mathbf{S}_m^T \mathbf{A} \mathbf{\Sigma}_0 \mathbf{A} \mathbf{S}_m$ is non-singular. Then the BayesCG posterior $\mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$ has mean and covariance*

$$\mathbf{x}_m = \mathbf{x}_0 + \mathbf{\Sigma}_0 \mathbf{A} \mathbf{S}_m \mathbf{\Lambda}_m^{-1} \mathbf{S}_m^T (\mathbf{b} - \mathbf{A}\mathbf{x}_0) \tag{2.1}$$

$$\mathbf{\Sigma}_m = \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0 \mathbf{A} \mathbf{S}_m \mathbf{\Lambda}_m^{-1} \mathbf{S}_m^T \mathbf{A} \mathbf{\Sigma}_0. \tag{2.2}$$

Proof. See supplement. ∎

REMARK 2.2. *Theorem 2.1 requires the existence of search directions that produce a nonsingular $\mathbf{\Lambda}_m$, and the purpose this theorem is to derive an expression for how to compute the posterior distribution resulting from any valid set of search directions. Section 2.2 presents the recursive computation of search directions that make $\mathbf{\Lambda}_m$ non-singular, while the supplement presents an example of a a non-recursive construction.*

Next we derive explicit expressions for the posterior covariances in terms of orthogonal projectors onto range$(\mathbf{\Sigma}_0 \mathbf{A} \mathbf{S}_m)$. To this end we exploit the close relation between Gaussian conditioning and orthogonal projections [1, Section 3]; and generalize the notion of projector [48, page 111] to semi-definite inner products to allow for singular priors $\mathbf{\Sigma}_0$,

DEFINITION 2.3 ([28, section 0.6.1]). *Let $\mathbf{B} \in \mathbb{R}^{n \times n}$ be symmetric positive semi-definite, and $\mathbf{P} \in \mathbb{R}^{n \times n}$. If $\mathbf{P}^2 = \mathbf{P}$ and $(\mathbf{BP})^T = \mathbf{BP}$, then $\mathbf{P}$ is a* $\mathbf{B}$-orthogonal projector*, with $(\mathbf{I} - \mathbf{P})^T \mathbf{B} \mathbf{P} = \mathbf{0}$.*

Now we are ready to express the posterior distributions in Theorem 2.1 in terms of $\mathbf{\Sigma}_0^\dagger$-orthogonal projectors.

THEOREM 2.4 (Extension of [10, Proposition 3]). *Under the assumptions of Theorem 2.1*

$$\mathbf{P}_m \equiv \mathbf{\Sigma}_0 \mathbf{A} \mathbf{S}_m \mathbf{\Lambda}_m^{-1} \mathbf{S}_m^T \mathbf{A} \mathbf{\Sigma}_0 \mathbf{\Sigma}_0^\dagger \tag{2.3}$$

is a $\mathbf{\Sigma}_0^\dagger$-orthogonal projector onto $K_m \equiv \operatorname{range}(\mathbf{\Sigma}_0 \mathbf{A} \mathbf{S}_m)$.

If additionally $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\boldsymbol{\Sigma}_0)$*, then the posterior satisfies*

$$\mathbf{x}_m = (\mathbf{I} - \mathbf{P}_m)\mathbf{x}_0 + \mathbf{P}_m\mathbf{x}_*$$
$$\boldsymbol{\Sigma}_m = (\mathbf{I} - \mathbf{P}_m)\boldsymbol{\Sigma}_0, \qquad \mathbf{P}_m\boldsymbol{\Sigma}_m = \mathbf{0}.$$

Proof. See Appendix A. ∎

Theorem 2.4 expresses the posterior mean $\mathbf{x}_m$ as the sum of two projections: the projection of the solution $\mathbf{x}_*$ onto $\mathrm{range}(\mathbf{P}_m)$, and the projection of the prior mean $\mathbf{x}_0$ onto the complementary space $\mathrm{range}(\mathbf{P}_m)^\perp$. As for the posterior covariance $\boldsymbol{\Sigma}_m$, it is the projection of the prior covariance $\boldsymbol{\Sigma}_0$ onto the complementary space $\mathrm{range}(\mathbf{P}_m)^\perp$.

REMARK 2.5. *Theorem 2.4 implies that* $\mathbf{P}_m\mathbf{x}_m = \mathbf{P}_m\mathbf{x}_*$ *and* $\mathbf{P}_m\boldsymbol{\Sigma}_m\mathbf{P}_m^T = \mathbf{0}$*. As a consequence, if* $X \sim \mathcal{N}(\mathbf{x}_m, \boldsymbol{\Sigma}_m)$*, then the distribution of* $\mathbf{P}_m(X - \mathbf{x}_*)$ *is Gaussian with mean* $\mathbf{P}_m\mathbf{x}_m - \mathbf{P}_m\mathbf{x}_* = \mathbf{0}$ *and covariance* $\mathbf{P}_m\boldsymbol{\Sigma}_m\mathbf{P}_m^T = \mathbf{0}$*. Thus, within* $\mathrm{range}(\mathbf{P}_m)$*, there is no uncertainty in our knowledge of* $\mathbf{x}_*$ *We can interpret the posterior as a conjecture about the unknown location of* $\mathbf{x}_*$ *in the complementary subspace* $\mathrm{range}(\mathbf{P}_m)^\perp$*.*

Theorem 2.4 implies the following optimality for the posterior mean: It is the vector closest to the solution $\mathbf{x}_*$ in the affine space $\mathbf{x}_0 + K_m$, with K_m as in Theorem 2.1.

THEOREM 2.6 (Extension of [1, Proposition 4]). *Under all the assumptions of Theorem 2.4, the posterior mean satisfies*

$$\mathbf{x}_m = \operatorname*{arg\,min}_{\mathbf{x} \in \mathbf{x}_0 + K_m} (\mathbf{x}_* - \mathbf{x})^T \boldsymbol{\Sigma}_0^\dagger (\mathbf{x}_* - \mathbf{x}). \tag{2.4}$$

Additionally, $(\mathbf{x}_* - \mathbf{x}_m)^T \boldsymbol{\Sigma}_0^\dagger (\mathbf{x}_* - \mathbf{x}_m) = 0$ *if and only if* $\mathbf{x}_m = \mathbf{x}_*$*.*

Proof. See Appendix A. ∎

Theorems 2.1, 2.4, and 2.6 assume that the search directions are chosen so that $\boldsymbol{\Lambda}_m$ is non-singular. The additional assumption $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\boldsymbol{\Sigma}_0)$ in Theorems 2.4 and 2.6 guarantees this nonsingularity for the specific search directions computed by BayesCG, as will be shown in Theorem 2.11.

2.2. Recursive computation of BayesCG posteriors under singular priors. We extend the recursions for posterior distributions under nonsingular prior covariances in [9] to singular ones, and present three results for the efficient implementation of BayesCG: New recursions for the posterior covariances (Theorem 2.7) and the search directions (Theorem 2.8); and a proof that the search directions are well-defined (Theorem 2.11).

The residuals of the posterior means are defined as

$$\mathbf{r}_m \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_m, \qquad 0 \leq m. \tag{2.5}$$

THEOREM 2.7 (Extension of Proposition 6 in [9]). *Under the assumptions of Theorem 2.1 if, in addition, the search directions* $\mathbf{S}_m$ *are* $\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}$*-orthogonal, then the posterior means and covariances admit the recursions*

$$\mathbf{x}_j = \mathbf{x}_{j-1} + \frac{\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{s}_j \left(\mathbf{s}_j^T\mathbf{r}_{j-1}\right)}{\mathbf{s}_j^T\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{s}_j}, \qquad 1 \leq j \leq m, \tag{2.6}$$

and

$$\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_{j-1} - \frac{\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{s}_j \left(\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{s}_j\right)^T}{\mathbf{s}_j^T\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{s}_j}, \qquad 1 \leq j \leq m. \tag{2.7}$$

Proof. See Appendix A. ∎

The denominators $(\mathbf{\Lambda}_m)_{jj} = \mathbf{s}_j^T \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_j$ in (2.6) and (2.7) are non-zero because Theorem 2.1 assumes that $\mathbf{\Lambda}_m$ is non-singular.

Next is a Lanczos-like recurrence for the $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$-orthogonal search directions from [9, Proposition 7].

THEOREM 2.8 ([9, Proposition 7] and [11, Proof of Proposition 7, Proposition S4, and Section S2]). *If the search directions*

$$\mathbf{s}_1 = \mathbf{r}_0 \neq \mathbf{0}, \qquad \mathbf{s}_j = \mathbf{r}_{j-1} - \frac{\mathbf{r}_{j-1}^T\mathbf{r}_{j-1}}{\mathbf{r}_{j-2}^T\mathbf{r}_{j-2}}\mathbf{s}_{j-1}, \qquad 2 \leq j \leq m, \tag{2.8}$$

satisfy the assumptions of Theorem 2.1, then they are an $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$-orthogonal basis for the Krylov space

$$\mathcal{K}_m(\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}, \mathbf{r}_0) \equiv \operatorname{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{r}_0, \ldots, (\mathbf{A}\mathbf{\Sigma}_0\mathbf{A})^{m-1}\mathbf{r}_0\}, \tag{2.9}$$

while the residuals $\mathbf{r}_0, \ldots, \mathbf{r}_{m-1}$ are an orthogonal basis for $\mathcal{K}_m(\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}, \mathbf{r}_0)$.

The maximal number of search directions in (2.8) can be less than n, because they are a basis for the Krylov subspace $\mathcal{K}_m(\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}, \mathbf{r}_0)$ whose maximal dimension can be less than n.

DEFINITION 2.9 (Section 2 in [3], Definition 4.2.1 in [31]). *Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ be symmetric positive semi-definite and let $\mathbf{w} \in \mathbb{R}^n$ be a non-zero vector. The* grade of $\mathbf{w}$ with respect to $\mathbf{B}$, *or the* invariance index for $(\mathbf{B}, \mathbf{w})$ *is the maximal dimension $1 \leq \mathrm{K} \leq n$ of the Krylov space,*

$$\mathcal{K}_{\mathrm{K}}(\mathbf{B}, \mathbf{w}) = \mathcal{K}_{\mathrm{K}+i}(\mathbf{B}, \mathbf{w}), \qquad i \geq 1.$$

REMARK 2.10. *In Theorem 2.8, if* K *is the grade of $\mathbf{r}_0$ with respect to $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$, then $\mathbf{s}_{\mathrm{K}+1} = \mathbf{0}$, $\mathbf{r}_{\mathrm{K}} = \mathbf{0}$, while $\mathbf{s}_j \neq \mathbf{0}$ and $\mathbf{r}_{j-1} \neq \mathbf{0}$ for $1 \leq j \leq \mathrm{K}$. Additionally,* $\mathrm{K} \leq \operatorname{rank}(\mathbf{\Sigma}_0)$.

In the following theorem, we show that with the additional assumption that $\mathbf{x}_* - \mathbf{x}_0 \in \operatorname{range}(\mathbf{\Sigma}_0)$, the $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$-orthogonal search directions from Theorem 2.8 satisfy the assumptions of Theorem 2.1.

THEOREM 2.11. *Let $\mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$ be a prior with symmetric positive semi-definite $\mathbf{\Sigma}_0 \in \mathbb{R}^{n\times n}$,* K *the grade of $\mathbf{r}_0$ with respect to $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$, and $m \leq K$. If $\mathbf{x}_* - \mathbf{x}_0 \in \operatorname{range}(\mathbf{\Sigma}_0)$, then the search directions from Theorem 2.1 produce a nonsingular $\mathbf{\Lambda}_m$, and $\mathbf{S}_m$ is $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$-orthogonal.*

Proof. Recursive computation of the BayesCG posteriors requires the search directions $\mathbf{S}_m = \begin{bmatrix}\mathbf{s}_1 & \cdots & \mathbf{s}_m\end{bmatrix}$ to be $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$-orthogonal, so that $\mathbf{\Lambda}_m = \mathbf{S}_m^T\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_m$ is diagonal [9, Section 2.3]. Furthermore, if $\mathbf{s}_j \notin \ker(\mathbf{\Sigma}_0\mathbf{A})$, $1 \leq j \leq m$, then $\mathbf{\Lambda}_m$ has non-zero diagonal elements and is nonsingular.

In the following induction proof we show that the search directions are $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$-orthogonal and that $\mathbf{s}_i \notin \ker(\mathbf{\Sigma}_0\mathbf{A})$ and $\mathbf{s}_i \neq \mathbf{0}$, $1 \leq i \leq m$. Since $\mathbf{A}$ and $\mathbf{\Sigma}_0$ are symmetric, $\ker(\mathbf{\Sigma}_0\mathbf{A}) = \ker(\mathbf{\Sigma}_0^T\mathbf{A}^T) = \ker\left((\mathbf{A}\mathbf{\Sigma}_0)^T\right)$ is the orthogonal complement of $\operatorname{range}(\mathbf{A}\mathbf{\Sigma}_0)$ in $\mathbb{R}^n$. Therefore, we can show $\mathbf{s}_i \notin \ker(\mathbf{\Sigma}_0\mathbf{A})$ by showing $\mathbf{s}_i \in \operatorname{range}(\mathbf{A}\mathbf{\Sigma}_0)$ and $\mathbf{s}_i \neq \mathbf{0}$, $1 \leq i \leq m$.

By assumption $m \leq \mathrm{K}$, so Remark 2.10 implies $\mathbf{r}_i \neq \mathbf{0}$, $1 \leq i \leq m-1$.

Induction basis. The assumption $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\boldsymbol{\Sigma}_0)$ implies

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0 = \mathbf{A}(\mathbf{x}_* - \mathbf{x}_0) \in \mathrm{range}(\mathbf{A}\boldsymbol{\Sigma}_0).$$

Thus $\mathbf{s}_1 = \mathbf{r}_0 \in \mathrm{range}(\mathbf{A}\boldsymbol{\Sigma}_0)$, and $\mathbf{r}_0 \neq \mathbf{0}$ by assumption. Thus $\mathbf{s}_1 \neq \mathbf{0}$, $\mathbf{s}_1 \notin \ker(\boldsymbol{\Sigma}_0\mathbf{A})$, and $\boldsymbol{\Lambda}_1 = \mathbf{s}_1^T\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{s}_1 \neq 0$.

Induction hypothesis. Assume that $\mathbf{s}_i, \mathbf{r}_i \in \mathrm{range}(\mathbf{A}\boldsymbol{\Sigma}_0)$, $\mathbf{s}_i, \mathbf{r}_i \neq \mathbf{0}$, and $\boldsymbol{\Lambda}_i$ is nonsingular, $1 \leq i \leq m-1$. This, along with Theorem 2.8 implies that $\mathbf{s}_1, \ldots, \mathbf{s}_{m-1}$ are $\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}$-orthogonal so that $\boldsymbol{\Lambda}_{m-1}$ is a diagonal matrix.

Induction step. Applying the induction hypothesis $\mathbf{s}_{m-1}, \mathbf{r}_{m-1} \in \mathrm{range}(\mathbf{A}\boldsymbol{\Sigma}_0)$ to (2.8) gives

$$\mathbf{s}_m = \mathbf{r}_{m-1} - \frac{\mathbf{r}_{m-1}^T\mathbf{r}_{m-1}}{\mathbf{r}_{m-2}^T\mathbf{r}_{m-2}}\mathbf{s}_{m-1}. \tag{2.10}$$

Hence $\mathbf{s}_m \in \mathrm{range}(\mathbf{A}\boldsymbol{\Sigma}_0)$. Multiply (2.10) on the left by $\mathbf{r}_{m-1}^T$ and insert $\mathbf{s}_{m-1}^T\mathbf{r}_{m-1} = 0$ from Lemma B.1 into the last summand to get $\mathbf{r}_{m-1}^T\mathbf{s}_m = \mathbf{r}_{m-1}^T\mathbf{r}_{m-1}$, where $\mathbf{r}_{m-1} \neq 0$ implies $\mathbf{s}_m \neq 0$. Then $\mathbf{s}_m \in \mathrm{range}(\mathbf{A}\boldsymbol{\Sigma}_0)$ and $\mathbf{s}_m \neq \mathbf{0}$ imply $\mathbf{s}_m \notin \ker(\boldsymbol{\Sigma}_0\mathbf{A})$.

The induction hypothesis, Theorem 2.8, and (2.10) imply that the search directions $\mathbf{s}_1, \ldots, \mathbf{s}_m$ are non-zero and $\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}$-orthogonal. Thus $\boldsymbol{\Lambda}_m$ is nonsingular diagonal, which implies that $\mathbf{s}_i \notin \ker(\boldsymbol{\Sigma}_0\mathbf{A})$, $1 \leq i \leq m$; and with Lemma A.1 that $\mathbf{x}_* - \mathbf{x}_m \in \mathrm{range}(\boldsymbol{\Sigma}_0)$, thus $\mathbf{r}_m = \mathbf{A}(\mathbf{x}_* - \mathbf{x}_m) \in \mathrm{range}(\mathbf{A}\boldsymbol{\Sigma}_0)$. ∎

REMARK 2.12. *The assumption* $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\boldsymbol{\Sigma}_0)$ *in Theorem 2.11, which holds automatically if the prior covariance* $\boldsymbol{\Sigma}_0$ *is nonsingular, is required to guarantee the nonsingularity of the diagonal matrices* $\boldsymbol{\Lambda}_m$.

The statistical interpretation of the assumption $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\boldsymbol{\Sigma}_0)$ *is that the solution* $\mathbf{x}_*$ *must live in the support of the prior, that is, in the subspace of* $\mathbb{R}^n$ *where the probability density function of* $\mathcal{N}(\mathbf{x}_0, \boldsymbol{\Sigma}_0)$ *is nonzero.*

Theorems 2.7, 2.8, and 2.11 form the basis for the BayesCG Algorithm 2.1, which differs from the original BayesCG [9, Algorithm 1] only in the computation of the posterior covariances as a sequence of rank-1 downdates rather than just a single rank-m downdate at the end. Algorithm 2.1 is a Krylov space method; for nonsingular priors $\boldsymbol{\Sigma}_0$ this was established in [9, Section 3], while for singular priors this follows from (2.9) and Theorem 2.6. To show the similarity of BayesCG Algorithm 2.1 to CG, we present the most common implementation of CG in Algorithm 2.2; it is the original version due to Hestenes and Stiefel [26, Section 3].

The posterior means in Algorithm 2.1 are closely related to the CG iterates in Algorithm 2.2. In the special case $\boldsymbol{\Sigma}_0 = \mathbf{A}^{-1}$, the BayesCG posterior means are identical to the CG iterates [9, Section 2.3]. The relationship between CG and BayesCG is discussed further in [5, 9, 10, 11, 30], and the results are summarized in the supplement.

2.3. Choice of BayesCG prior distribution. The mean $\mathbf{x}_0$ in the prior $\mathcal{N}(\mathbf{x}_0, \boldsymbol{\Sigma}_0)$ corresponds to the initial guess in CG, while the covariance $\boldsymbol{\Sigma}_0$ can be any symmetric positive semi-definite matrix that satisfies $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\boldsymbol{\Sigma}_0)$. Nonsingular priors examined in [9, Section 4.1] include

- Inverse prior $\boldsymbol{\Sigma}_0 = \mathbf{A}^{-1}$: The posterior means in Algorithm 2.1 are equal to the CG iterates.
- Natural prior $\boldsymbol{\Sigma}_0 = \mathbf{A}^{-2}$: The posterior means in Algorithm 2.1 converge in a single iteration.

Algorithm 2.1 Bayesian Conjugate Gradient Method (BayesCG)

Input: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}_0 \in \mathbb{R}^n$
  spds $\mathbf{\Sigma}_0 \in \mathbb{R}^{n\times n}$ so that $\mathbf{x}_* - \mathbf{x}_0 \in \text{range}(\mathbf{\Sigma}_0)$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ ▷ define initial values
$\mathbf{s}_1 = \mathbf{r}_0$
$m = 0$
while not converged **do** ▷ iterate through BayesCG Recursions
  $m = m + 1$
  $\alpha_m = \left(\mathbf{r}_{m-1}^T \mathbf{r}_{m-1}\right) / \left(\mathbf{s}_m^T \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_m\right)$
  $\mathbf{x}_m = \mathbf{x}_{m-1} + \alpha_m \mathbf{\Sigma}_0 \mathbf{A}\mathbf{s}_m$
  $\mathbf{\Sigma}_m = \mathbf{\Sigma}_{m-1} - \mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_m \left(\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_m\right)^T / (\mathbf{s}_m^T \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_m)$
  $\mathbf{r}_m = \mathbf{r}_{m-1} - \alpha_m \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_m$
  $\beta_m = \left(\mathbf{r}_m^T \mathbf{r}_i\right) / \left(\mathbf{r}_{m-1}^T \mathbf{r}_{m-1}\right)$
  $\mathbf{s}_{m+1} = \mathbf{r}_m + \beta_m \mathbf{s}_m$
end while
Output: $\mathbf{x}_m$, $\mathbf{\Sigma}_m$

Algorithm 2.2 Conjugate Gradient Method (CG)

Input: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}_0 \in \mathbb{R}^n$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ ▷ define initial values
$\mathbf{v}_1 = \mathbf{r}_0$
$m = 0$
while not converged **do** ▷ iterate through CG Recursions
  $m = m + 1$
  $\gamma_m = (\mathbf{r}_{m-1}^T \mathbf{r}_{m-1}) / (\mathbf{v}_m^T \mathbf{A}\mathbf{v}_m)$
  $\mathbf{x}_m = \mathbf{x}_{m-1} + \gamma_m \mathbf{v}_m$
  $\mathbf{r}_m = \mathbf{r}_{m-1} - \gamma_m \mathbf{A}\mathbf{v}_m$
  $\delta_m = (\mathbf{r}_m^T \mathbf{r}_m) / (\mathbf{r}_{m-1}^T \mathbf{r}_{m-1})$
  $\mathbf{v}_{m+1} = \mathbf{r}_m + \delta_m \mathbf{v}_m$
end while
Output: $\mathbf{x}_m$

- Identity prior $\mathbf{\Sigma}_0 = \mathbf{I}$: The prior is easy to compute, but the posterior means in Algorithm 2.1 converge slowly.
- Preconditioner prior $\mathbf{\Sigma}_0 = \left(\mathbf{M}^T\mathbf{M}\right)^{-1}$ where $\mathbf{M} \approx \mathbf{A}$: This prior approximates the natural prior.
- Krylov subspace prior $\mathbf{\Sigma}_0$: This prior is defined in terms of a basis for the Krylov space $\mathcal{K}(\mathbf{A}, \mathbf{r}_0)$.

Figure 2.1 illustrates the convergence of posterior means and covariances from Algorithm 2.1 under the priors $\mathbf{\Sigma}_0 = \mathbf{A}^{-1}$ and $\mathbf{\Sigma}_0 = \mathbf{I}$. In both cases the posterior means converge faster than the posterior covariances, suggesting that the covariances are unreasonably pessimistic about the size of the error $\mathbf{x}_* - \mathbf{x}_m$. Section 3.3 presents a detailed discussion of the relation between the trace of the posterior covariance and the error $\mathbf{x}_* - \mathbf{x}_m$ in the posterior means.

The example below presents a prior of minimal rank that comprises a maximal amount of information.

EXAMPLE 2.13. *If* $\mathbf{x}_0 \neq \mathbf{x}_*$, *then* $\mathbf{\Sigma}_0 = (\mathbf{x}_* - \mathbf{x}_0)(\mathbf{x}_* - \mathbf{x}_0)^T$ *is is a rank-one*

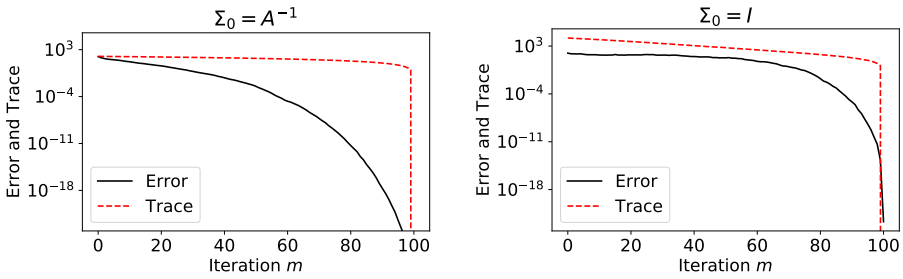

FIGURE 2.1. *Convergence of BayesCG Algorithm 2.1 applied to the linear system in section 4.2 under different priors: inverse prior (left panel) and identity prior (right panel). Convergence of the means is displayed as* $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$*, while convergence of the covariances is displayed as* $\mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma}_m)$.

covariance that satisfies $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\boldsymbol{\Sigma}_0)$. *All rank-one prior covariances for BayesCG are multiples of this prior.*

To see this, note that Theorem 2.11 and $\mathbf{A}^{-1}\mathbf{r}_0 = \mathbf{x}_* - \mathbf{x}_0$ *imply termination of Algorithm 2.1 under this prior in a single iteration,*

$$\mathbf{x}_1 = \mathbf{x}_0 + \frac{1}{\mathbf{r}_0^T\mathbf{A}\underbrace{\mathbf{A}^{-1}\mathbf{r}_0\mathbf{r}_0^T\mathbf{A}^{-1}}_{\boldsymbol{\Sigma}_0}\mathbf{A}\mathbf{r}_0}\underbrace{\mathbf{A}^{-1}\mathbf{r}_0\mathbf{r}_0^T\mathbf{A}^{-1}}_{\boldsymbol{\Sigma}_0}\mathbf{A}\mathbf{r}_0(\mathbf{r}_0^T\mathbf{r}_0) = \mathbf{x}_0 + \mathbf{x}_* - \mathbf{x}_0 = \mathbf{x}_*.$$

3. Prior distributions informed by Krylov subspaces. Motivated by the 'Krylov subspace prior' [9, section 4.1], we introduce a new 'Krylov prior' (section 3.1), derive expressions for the Krylov posteriors (section 3.2), ensure the Krylov posteriors accurately model uncertainty in $\mathbf{x}_*$ (section 3.3), and develop a practical Krylov posterior and an efficient implementation of BayesCG as a uncertainty-aware version of CG (section 3.4).

3.1. General Krylov prior. We introduce our new Krylov prior (Definition 3.1) and show that the BayesCG Krylov subspace under the Krylov prior is identical to the CG Krylov subspace (Lemma 3.2). This Krylov prior is impractical because its computation amounts to the direct solution of (1.1), however it is the foundation for the efficient low-rank approximations in section 3.4.

The new Krylov prior is defined in terms of the maximal CG Krylov subspace $\mathcal{K}_{\mathrm{K}}(\mathbf{A}, \mathbf{r}_0)$, where K is the grade of $\mathbf{r}_0$ with respect to $\mathbf{A}$ (Definition 2.9). The $\mathbf{A}$-orthonormal versions of the search directions $\mathbf{v}_m$ in Algorithm 2.2 are

$$\tilde{\mathbf{v}}_m \equiv \mathbf{v}_m / \sqrt{\mathbf{v}_m^T\mathbf{A}\mathbf{v}_m}, \qquad 1 \leq m \leq \mathrm{K}. \tag{3.1}$$

As columns of

$$\mathbf{V} \equiv \begin{bmatrix} \tilde{\mathbf{v}}_1 & \cdots & \tilde{\mathbf{v}}_{\mathrm{K}} \end{bmatrix} \in \mathbb{R}^{n \times \mathrm{K}} \qquad \text{with} \quad \mathbf{V}^T\mathbf{A}\mathbf{V} = \mathbf{I}_{\mathrm{K}} \tag{3.2}$$

they represent an $\mathbf{A}$-orthonormal basis for $\mathrm{range}(\mathbf{V}) = \mathcal{K}_{\mathrm{K}}(\mathbf{A}, \mathbf{r}_0)$ [26, Theorem 5.1].

DEFINITION 3.1. *The (general) Krylov prior is* $\mathcal{N}(\mathbf{x}_0, \boldsymbol{\Gamma}_0)$*, where the mean* $\mathbf{x}_0$ *is an initial guess for* $\mathbf{x}_*$*, and the covariance matrix is*

$$\boldsymbol{\Gamma}_0 \equiv \mathbf{V}\boldsymbol{\Phi}\mathbf{V}^T \in \mathbb{R}^{n \times n} \tag{3.3}$$

where $\mathbf{V}$ *is as defined in (3.2) and* $\mathbf{\Phi} \equiv \operatorname{diag}\begin{pmatrix}\phi_1 & \phi_2 & \cdots & \phi_{\mathrm{K}}\end{pmatrix} \in \mathbb{R}^{\mathrm{K}\times\mathrm{K}}$ *with* $\phi_i > 0$, $1 \leq i \leq \mathrm{K}$. *The Krylov prior is 'general' because the diagonal elements of* $\mathbf{\Phi}$ *are unspecified.*

The results in this section and in section 3.2 are valid for any choice of positive diagonal elements in $\mathbf{\Phi}$. A specific choice of diagonal elements is presented in section 3.3.

The Krylov prior covariance has $\operatorname{rank}(\mathbf{\Gamma}_0) = \mathrm{K}$ and is singular for $\mathrm{K} < n$, hence the need for singular priors in section 2. Fortunately, $\mathbf{\Gamma}_0$ is a well-defined BayesCG prior, because it satisfies the crucial condition in Theorem 2.11,

$$\mathbf{x}_* - \mathbf{x}_0 \in \mathcal{K}_{\mathrm{K}}(\mathbf{A}, \mathbf{r}_0) = \operatorname{range}(\mathbf{V}) = \operatorname{range}(\mathbf{\Gamma}_0).$$

Intuition. We give two different interpretations of the decomposition (3.3).

1. Hermitian eigenvalue problem $\mathbf{A}^{1/2}\mathbf{\Gamma}_0\mathbf{A}^{1/2} = \mathbf{W}\mathbf{\Phi}\mathbf{W}^T$, where $\mathbf{\Phi}$ contains the positive eigenvalues, and the eigenvector matrix $\mathbf{W} \equiv \mathbf{A}^{1/2}\mathbf{V}$ has orthonormal columns with $\mathbf{W}^T\mathbf{W} = \mathbf{I}_{\mathrm{K}}$.
2. Non-Hermitian eigenvalue problem $\mathbf{\Gamma}_0\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{\Phi}$ with eigenvalues and eigenvectors

$$\mathbf{\Gamma}_0\mathbf{A}\tilde{\mathbf{v}}_m = \phi_m\tilde{\mathbf{v}}_m, \qquad 1 \leq m \leq \mathrm{K}. \tag{3.4}$$

This is the property to be exploited in section 3.2.

We show that the BayesCG Krylov subspace under the Krylov prior is identical to the CG Krylov subspace.

LEMMA 3.2. *If* $\mathbf{\Gamma}_0$ *is the Krylov prior in Definition 3.1, then*

$$\mathcal{K}_m(\mathbf{A}, \mathbf{r}_0) = \mathcal{K}_m(\mathbf{A}\mathbf{\Gamma}_0\mathbf{A}, \mathbf{r}_0), \qquad 1 \leq m \leq \mathrm{K}.$$

Consequently, K *is also the grade of* $\mathbf{r}_0$ *with respect to* $\mathbf{A}\mathbf{\Gamma}_0\mathbf{A}$ *is* K.

Proof. An induction proof shows that the Krylov subspaces are the same for the first K dimensions. Then we prove that the grade of $\mathbf{r}_0$ with respect to $\mathbf{A}\mathbf{\Sigma}\mathbf{A}$ is K.

Induction basis. Since one-dimensional Krylov subspaces are independent of the matrix,

$$\mathcal{K}_1(\mathbf{A}, \mathbf{r}_0) = \operatorname{span}\{\mathbf{r}_0\} = \mathcal{K}_1(\mathbf{A}\mathbf{\Gamma}_0\mathbf{A}, \mathbf{r}_0).$$

Induction hypothesis. Assume that

$$\mathcal{K}_i(\mathbf{A}, \mathbf{r}_0) = \mathcal{K}_i(\mathbf{A}\mathbf{\Gamma}_0\mathbf{A}, \mathbf{r}_0), \qquad 1 \leq i \leq m-1.$$

With $\mathbf{V}_{1:m-1} = \begin{bmatrix}\tilde{\mathbf{v}}_1 & \tilde{\mathbf{v}}_2 & \cdots & \tilde{\mathbf{v}}_{m-1}\end{bmatrix}$ in (3.2) this implies

$$\operatorname{range}(\mathbf{V}_{1:m-1}) = \mathcal{K}_{m-1}(\mathbf{A}, \mathbf{r}_0) = \mathcal{K}_{m-1}(\mathbf{A}\mathbf{\Gamma}_0\mathbf{A}, \mathbf{r}_0). \tag{3.5}$$

Induction step. From (3.5) follow the expressions for the direct sums,

$$\mathcal{K}_m(\mathbf{A}, \mathbf{r}_0) = \operatorname{span}\{\mathbf{r}_0\} \oplus \operatorname{range}(\mathbf{A}\mathbf{V}_{1:m-1}) \tag{3.6}$$

$$\mathcal{K}_m(\mathbf{A}\mathbf{\Gamma}_0\mathbf{A}, \mathbf{r}_0) = \operatorname{span}\{\mathbf{r}_0\} \oplus \operatorname{range}(\mathbf{A}\mathbf{\Gamma}_0\mathbf{A}\mathbf{V}_{1:m-1}). \tag{3.7}$$

Then (3.4) and the non-singularity of $\mathbf{\Phi}$ imply

$$\operatorname{range}(\mathbf{A}\mathbf{\Gamma}_0\mathbf{A}\mathbf{V}_{1:m-1}) = \operatorname{range}(\mathbf{A}\mathbf{V}_{1:m-1}\mathbf{\Phi}_{1:m-1}) = \operatorname{range}(\mathbf{A}\mathbf{V}_{1:m-1}).$$

Combining this with (3.6) and (3.7) completes the induction,

$$\begin{aligned}\mathcal{K}_m(\mathbf{A}, \mathbf{r}_0) &= \text{span}\{\mathbf{r}_0\} \oplus \text{range}(\mathbf{A}\mathbf{V}_{1:m-1}) \\ &= \text{span}\{\mathbf{r}_0\} \oplus \text{range}(\mathbf{A}\boldsymbol{\Gamma}_0\mathbf{A}\mathbf{V}_{1:m-1}) = \mathcal{K}_m(\mathbf{A}\boldsymbol{\Gamma}_0\mathbf{A}, \mathbf{r}_0).\end{aligned}$$

Maximal Krylov space dimension. If K' is the grade of $\mathbf{r}_0$ with respect to $\mathbf{A}\boldsymbol{\Gamma}_0\mathbf{A}$, then the induction implies

$$\mathrm{K}' \geq \dim(\mathcal{K}_{\mathrm{K}}(\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}, \mathbf{r}_0)) = \dim(\mathcal{K}_{\mathrm{K}}(\mathbf{A}, \mathbf{r}_0)) = \mathrm{K}.$$

On the other hand, $\text{rank}(\mathbf{A}\boldsymbol{\Gamma}_0\mathbf{A}) = \mathrm{K}$ implies $\mathrm{K}' \leq \mathrm{K}$. Therefore $\mathrm{K}' = \mathrm{K}$. ∎

3.2. General Krylov posteriors. We show (Theorem 3.3) that under the Krylov prior, the BayesCG posteriors have means that are identical to the CG iterates, and covariances that can be factored as in Definition 3.1. This represents the foundation for an efficient implementation of BayesCG (Remark 3.4).

Define appropriate submatrices of $\mathbf{V}$ and $\boldsymbol{\Phi}$,

$$\mathbf{V}_{i:j} \equiv \begin{bmatrix}\tilde{\mathbf{v}}_i & \cdots & \tilde{\mathbf{v}}_j\end{bmatrix}, \qquad \boldsymbol{\Phi}_{i:j} \equiv \text{diag}\begin{pmatrix}\phi_i & \cdots & \phi_j\end{pmatrix}, \qquad 1 \leq i < j \leq \mathrm{K}. \tag{3.8}$$

In particular, $\mathbf{V} = \mathbf{V}_{1:K}$ and $\boldsymbol{\Phi} = \boldsymbol{\Phi}_{1:K}$.

THEOREM 3.3. *Let $\mathcal{N}(\mathbf{x}_0, \boldsymbol{\Gamma}_0)$ be the Krylov prior in Definition 3.1, and let $\mathcal{N}(\mathbf{x}_m, \boldsymbol{\Gamma}_m)$ be the posteriors from BayesCG Algorithm 2.1, $1 \leq m \leq \mathrm{K}$. Then the posterior means $\mathbf{x}_m$ are identical to the corresponding CG iterates in Algorithm 2.2, and the posterior covariances can be factored as*

$$\boldsymbol{\Gamma}_m = \mathbf{V}_{m+1:\mathrm{K}}\boldsymbol{\Phi}_{m+1:\mathrm{K}}(\mathbf{V}_{m+1:\mathrm{K}})^T, \qquad 1 \leq m < \mathrm{K}, \tag{3.9}$$

and $\boldsymbol{\Gamma}_m = \mathbf{0}$ for $m = \mathrm{K}$.

Proof. We first derive the equality of the posterior means, and then the factorizations of the covariances.

Posterior means. The idea is to show equality of the BayesCG posterior means under Krylov and inverse priors since, per the discussion in [9, Section 2.3] and section 2.3, BayesCG posterior means under the inverse prior are identical to CG iterates.

From Theorem 2.1, and the 'equivalence' of Algorithm 2.1 under $\boldsymbol{\Sigma}_0 = \mathbf{A}^{-1}$ and Algorithm 2.2 follows that the BayesCG posterior means under the inverse prior are equal to

$$\mathbf{x}_m = \mathbf{x}_0 + \mathbf{V}_{1:m}\mathbf{V}_{1:m}^T\mathbf{r}_0. \tag{3.10}$$

Similarly, Theorem 2.1 implies that the BayesCG posterior under the Krylov prior are equal to

$$\mathbf{x}_m = \mathbf{x}_0 + \boldsymbol{\Gamma}_0\mathbf{A}\widetilde{\mathbf{S}}_m(\widetilde{\mathbf{S}}_m^T\mathbf{A}\boldsymbol{\Gamma}_0\mathbf{A}\widetilde{\mathbf{S}}_m)^{-1}\widetilde{\mathbf{S}}_m^T\mathbf{r}_0, \tag{3.11}$$

where the columns of $\widetilde{\mathbf{S}}_m$ are the search directions from Algorithm 2.1 under the Krylov prior. To show the equality of (3.10) and (3.11), we need to relate $\widetilde{\mathbf{S}}_m$ and $\mathbf{V}_{1:m}$ and then include the Krylov prior $\boldsymbol{\Gamma}_0$.

With the submatrices defined as in (3.8) we conclude from (3.2) and Lemma 3.2 that

$$\text{range}(\widetilde{\mathbf{S}}_m) = \mathcal{K}_m(\mathbf{A}\boldsymbol{\Gamma}_0\mathbf{A}, \mathbf{r}_0) = \text{range}(\mathbf{V}_{1:m}),$$

where the columns of $\widetilde{\mathbf{S}}_m$ are $\mathbf{A\Gamma}_0\mathbf{A}$-orthogonal. To show that the columns of $\mathbf{V}_{1:m}$ are also $\mathbf{A\Gamma}_0\mathbf{A}$-orthogonal, exploit the fact that they are $\mathbf{A}$-orthonormal and apply Definition 3.1,

$$\mathbf{V}_{1:m}^T\mathbf{A\Gamma}_0\mathbf{A}\mathbf{V}_{1:m} = \mathbf{V}_{1:m}^T\mathbf{A}\mathbf{V}\mathbf{\Phi}\mathbf{V}^T\mathbf{A}\mathbf{V}_{1:m} = \mathbf{\Phi}_{1:m},$$

which is a diagonal matrix. We have established that the columns of $\widetilde{\mathbf{S}}_m$ and $\mathbf{V}_{1:m}$ are $\mathbf{A\Gamma}_0\mathbf{A}$-orthogonal, with respective leading columns being multiples of $\mathbf{r}_0$, thus are $\mathbf{A\Gamma}_0\mathbf{A}$-orthogonal bases of $\mathcal{K}_m(\mathbf{A\Gamma}_0\mathbf{A}, \mathbf{r}_0)$. Therefore the columns of $\mathbf{V}_{1:m}$ are multiples of the columns of $\widetilde{\mathbf{S}}_m$. That is

$$\widetilde{\mathbf{S}}_m = \mathbf{V}_{1:m}\mathbf{\Delta} \tag{3.12}$$

for some non-singular diagonal matrix $\mathbf{\Delta} \in \mathbb{R}^{m\times m}$. Substitute (3.12) into the third interpretation (3.4) of the Krylov prior,

$$\mathbf{\Gamma}_0\mathbf{A}\widetilde{\mathbf{S}}_m = \mathbf{\Gamma}_0\mathbf{A}\mathbf{V}_{1:m}\mathbf{\Delta} = \mathbf{V}_{1:m}\mathbf{\Phi}_{1:m}\mathbf{\Delta}$$

and this in turn into the second summand of (3.11). Then the non-singularity and diagonality of both $\mathbf{\Delta}$ and $\mathbf{\Phi}$ lead to the simplification

$$\mathbf{x}_m = \mathbf{x}_0 + \mathbf{V}_{1:m}\mathbf{\Phi}_{1:m}\mathbf{\Delta}(\mathbf{\Delta}\mathbf{\Phi}_{1:m}\mathbf{\Delta})^{-1}\mathbf{\Delta}\mathbf{V}_{1:m}^T\mathbf{r}_0 = \mathbf{x}_0 + \mathbf{V}_{1:m}\mathbf{V}_{1:m}^T\mathbf{r}_0, \tag{3.13}$$

which is (3.10).

Posterior covariances. Substituting (3.12) into Theorem 2.1 and simplifying as in (3.13) gives

$$\begin{aligned}\mathbf{\Gamma}_m &= \mathbf{\Gamma}_0 - \mathbf{\Gamma}_0\mathbf{A}\widetilde{\mathbf{S}}_m(\widetilde{\mathbf{S}}_m^T\mathbf{A\Gamma}\mathbf{A}\widetilde{\mathbf{S}}_m)^{-1}\widetilde{\mathbf{S}}^T\mathbf{A\Gamma}_0 \\ &= \mathbf{V}\mathbf{\Phi}\mathbf{V}^T - \mathbf{V}_{1:m}\mathbf{\Phi}_{1:m}\mathbf{V}_{1:m}^T = \mathbf{V}_{m+1:\mathrm{K}}\mathbf{\Phi}_{m+1:\mathrm{K}}\mathbf{V}_{m+1:\mathrm{K}}^T. \end{aligned}$$ ∎

REMARK 3.4. *Theorem 3.3 implies that the posteriors from BayesCG under the Krylov prior have means that can be computed with CG, and covariances can be maintained in factored form without any arithmetic operations. This is the key to the efficient implementation of BayesCG in section 3.4.*

3.3. Krylov posteriors that capture CG convergence. We present a Krylov prior with specific diagonal elements (section 3.3.1), discuss the calibration of BayesCG under this prior (section 3.3.2) and its relation to existing CG error estimation theory (section 3.3.3).

3.3.1. Specific Krylov prior. We choose a specific diagonal matrix $\mathbf{\Phi}$ for the Krylov prior (Definition 3.6), so that the Krylov posteriors accurately model the uncertainty in our knowledge of $\mathbf{x}_*$ due to the error $\mathbf{x}_* - \mathbf{x}_m$. We derive error estimates from samples of the posteriors (Lemma 3.5) and then relate them to CG errors (Theorem 3.7).

Let us start with a general posterior distribution $\mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$. If it indeed accurately modeled the uncertainty in $\mathbf{x}_*$ due to the approximation error $\mathbf{x}_* - \mathbf{x}$, then we would expect the difference between samples of $\mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$ and its posterior mean $\mathbf{x}$ to be close to the actual error,

$$\mathbb{E}\left[\|X - \mathbf{x}\|_{\mathbf{A}}^2\right] = \|\mathbf{x}_* - \mathbf{x}\|_{\mathbf{A}}^2 \qquad \text{where} \quad X \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma}). \tag{3.14}$$

The squared $\mathbf{A}$-norm error $\|X - \mathbf{x}\|_{\mathbf{A}}^2$ is a quadratic form, whose expected value has an explicit expression.

LEMMA 3.5. *If $X \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$ is a Gaussian random variable with mean $\mathbf{x} \in \mathbb{R}^n$ and symmetric positive semi-definite covariance $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$, then*

$$\mathbb{E}\left[\|X - \mathbf{x}\|_{\mathbf{A}}^2\right] = \operatorname{trace}(\mathbf{A}\mathbf{\Sigma}). \tag{3.15}$$

Proof. The proof relies on the expected value of a quadratic form in Appendix B. Set $Z \equiv X - \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and apply Lemma B.2 to $Z^T\mathbf{A}Z$,

$$\mathbb{E}\left[\|X - \mathbf{x}\|_{\mathbf{A}}^2\right] = \mathbb{E}\left[\|Z\|_{\mathbf{A}}^2\right] = \mathbb{E}\left[Z^T\mathbf{A}Z\right] = \operatorname{trace}(\mathbf{A}\mathbf{\Sigma}). \qquad \square$$

Thus, $\operatorname{trace}(\mathbf{A}\mathbf{\Sigma})$ has the potential to be an error indicator. We present a specific diagonal matrix for the Krylov prior $\mathbf{\Gamma}_0$ in Definition 3.1, so that its posterior covariances produce meaningful error estimates $\operatorname{trace}(\mathbf{A}\mathbf{\Gamma}_m)$.

DEFINITION 3.6. *The (specific) Krylov prior is $\mathcal{N}(\mathbf{x}_0, \mathbf{\Gamma}_0)$, where the mean $\mathbf{x}_0$ is an initial guess for $\mathbf{x}_*$, and the covariance matrix is*

$$\mathbf{\Gamma}_0 \equiv \mathbf{V}\mathbf{\Phi}\mathbf{V}^T \in \mathbb{R}^{n\times n} \tag{3.16}$$

where $\mathbf{V}$ is defined in (3.2) and $\mathbf{\Phi} \equiv \operatorname{diag}\begin{pmatrix}\phi_1 & \phi_2 & \cdots & \phi_{\mathrm{K}}\end{pmatrix} \in \mathbb{R}^{\mathrm{K}\times\mathrm{K}}$ has diagonal elements

$$\phi_i = \gamma_i\|\mathbf{r}_{i-1}\|_2^2, \qquad 1 \le i \le \mathrm{K},$$

where $\gamma_i = \mathbf{r}_{i-1}^T\mathbf{r}_{i-1}/\mathbf{v}_i^T\mathbf{A}\mathbf{v}_i$ are the step sizes in line 7 of CG Algorithm 2.2.

Now we show that the posterior covariances from BayesCG under the specific Krylov prior reproduce the CG error.

THEOREM 3.7. *Let $\mathcal{N}(\mathbf{x}_0, \mathbf{\Gamma}_0)$ be the Krylov prior in Definition 3.6, and $\mathcal{N}(\mathbf{x}_m, \mathbf{\Gamma}_m)$ be the posteriors from BayesCG Algorithm 2.1, $1 \le m \le \mathrm{K}$. Then*

$$\operatorname{trace}(\mathbf{A}\mathbf{\Gamma}_m) = \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2, \qquad 1 \le m \le \mathrm{K}.$$

Proof. Apply Lemma 3.5 to the specific Krylov prior in Definition 3.6. From the cyclic commutativity of the trace and $\mathbf{A}$-orthonormality of the columns of $\mathbf{V}$ follows

$$\begin{aligned}\operatorname{trace}(\mathbf{A}\mathbf{\Gamma}_m) &= \operatorname{trace}(\mathbf{A}\mathbf{V}_{m:\mathrm{K}}\mathbf{\Phi}_{m:\mathrm{K}}(\mathbf{V}_{m:\mathrm{K}})^T)\\ &= \operatorname{trace}((\mathbf{V}_{m:\mathrm{K}})^T\mathbf{A}\mathbf{V}_{m:\mathrm{K}}\mathbf{\Phi}_{m:\mathrm{K}}) = \operatorname{trace}(\mathbf{\Phi}_{m:\mathrm{K}}).\end{aligned} \tag{3.17}$$

The diagonal matrix $\mathbf{\Phi}$ for the specific Krylov prior in Definition 3.6 is chosen so that $\operatorname{trace}(\mathbf{\Phi}_{m:\mathrm{K}}) = \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$. Remember that the reduction in the squared $\mathbf{A}$-norm error from iteration m to $m + d$ of Algorithm 2.2 equals [26, Theorem 6:1] and [31, Theorem 5.6.1]

$$\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2 - \|\mathbf{x}_* - \mathbf{x}_{m+d}\|_{\mathbf{A}}^2 = \sum_{i=m+1}^{m+d} \gamma_i\|\mathbf{r}_{i-1}\|_2^2, \qquad 0 \le m < m + d \le \mathrm{K}. \tag{3.18}$$

Setting $d = \mathrm{K} - m$ gives $\mathbf{x}_{\mathrm{K}} = \mathbf{x}_*$ and

$$\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2 = \sum_{i=m+1}^{\mathrm{K}} \gamma_i\|\mathbf{r}_{i-1}\|_2^2, \qquad 0 \le m \le \mathrm{K}.$$

Combine this equality with (3.17) to conclude $\phi_i = \gamma_i\|\mathbf{r}_{i-1}\|_2^2$, $1 \le i \le \mathrm{K}$. $\square$

Thus, the specific Krylov posteriors have covariances that converge at the same speed as their means.

3.3.2. Calibration of BayesCG under the specific Krylov prior. A probabilistic numerical linear solver is considered *calibrated* if its posterior distribution accurately models the uncertainty in $\mathbf{x}_*$ due to the approximation error $\mathbf{x}_* - \mathbf{x}_m$. Calibration of general probabilistic methods is discussed in [6] and of linear solvers in [7]. We briefly discuss how Lemma 3.5 and Theorem 3.7 contribute to better calibration of BayesCG under the specific Krylov prior.

Previous probabilistic extensions of CG do not produce posteriors that accurately model the uncertainty in $\mathbf{x}_*$ [1, Section 6.4], [9, Section 6.1], [53, Section 3]. For instance, Figure 2.1 illustrates that BayesCG under the priors $\mathbf{\Sigma}_0 = \mathbf{A}^{-1}$ and $\mathbf{\Sigma}_0 = \mathbf{I}$ has errors $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$ that converge faster than $\mathrm{trace}(\mathbf{A}\mathbf{\Sigma}_m)$. Furthermore, according to Lemma 3.5, the estimators $\mathrm{trace}(\mathbf{A}\mathbf{\Sigma}_m)$ from posterior samples are inaccurate and do not reflect the true error $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$. In other words, the posteriors do not accurately model uncertainty in $\mathbf{x}_*$.

Our approach towards designing posteriors that accurately model the uncertainty in $\mathbf{x}_*$ relies a judicious choice of the diagonal matrix $\mathbf{\Phi}$ for the specific Krylov prior, so that sampling from the posteriors produces accurate error estimates. This can be viewed as a scaling of the posterior covariance that forces $\mathrm{trace}(\mathbf{\Phi}_{m:\mathrm{K}}) = \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$. Alternative approaches for improving posteriors via scaling of the posterior covariances include [9, Section 4.2], [13, Section 7], and [53, Section 3]

Empirical evidence demonstrating that BayesCG under the specific Krylov prior produces posterior samples with accurate error estimates suggests but does not guarantee that it accurately models the uncertainty in $\mathbf{x}_*$. A rigorous investigation of the calibration of BayesCG under the specific Krylov prior is the subject of a separate paper.

3.3.3. Relation to CG error estimation. The purpose of Lemma 3.5 is to motivate a choice of $\mathbf{\Phi}$ so that BayesCG under the specific Krylov prior accurately models the uncertainty in $\mathbf{x}_*$ due to the approximation error $\mathbf{x}_* - \mathbf{x}_m$.

Effective CG error estimation is a well researched area, with most effort focused on the absolute $\mathbf{A}$-norm error. One option [49] is to run d additional CG iterations and apply (3.18) to obtain the underestimate [49, Equation (4.9)],

$$\sum_{i=m+1}^{m+d} \gamma_i \|\mathbf{r}_{i-1}\|_2^2 \leq \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2. \tag{3.19}$$

The rationale is that the error after $m+d$ iterations has become negligible compared to the error after m iterations, especially in the case of fast convergence. The number of additional iterations d is usually called the 'delay' [37, Section 1], and larger values of d lead to more accurate error estimates.

The estimate (3.19) also coincides with the lower bound from Gaussian quadrature [49, Section 3]. Other lower and upper bounds for the $\mathbf{A}$-norm error based on quadrature formulas and tunable with a delay include [17, 18, 19, 35, 36, 37, 49, 50].

3.4. Practical specific Krylov posteriors. We define low rank approximations of specific Krylov posterior covariances (Definition 3.8), and present an efficient CG-based implementation of BayesCG (Algorithm 3.1). It approximates the Krylov posteriors from delay iterations, thereby avoiding explicit computation of the Krylov prior, and inherits the fast convergence of CG.

The following low-rank approximations are based on the factored form of the Krylov posteriors in Theorem 3.3 and make use of the submatrices defined in (3.8).

DEFINITION 3.8. *Let $\mathcal{N}(\mathbf{x}_0, \mathbf{\Gamma}_0)$ be the specific Krylov prior from Definition 3.6 with posteriors*

$$\mathbf{\Gamma}_m = \mathbf{V}_{m+1:\mathrm{K}} \mathbf{\Phi}_{m+1:\mathrm{K}} \left(\mathbf{V}_{m+1:\mathrm{K}}\right)^T, \qquad 1 \leq m < \mathrm{K}.$$

For $1 \leq d \leq \mathrm{K} - m$, extract the leading rank-d submatrices from $\mathbf{V}_{m+1:\mathrm{K}}$ and $\mathbf{\Phi}_{m+1:\mathrm{K}}$, and define the rank-$d$ approximate Krylov posteriors as $\mathcal{N}(\mathbf{x}_m, \widehat{\mathbf{\Gamma}}_m)$ with

$$\widehat{\mathbf{\Gamma}}_m \equiv \mathbf{V}_{m+1:m+d} \mathbf{\Phi}_{m+1:m+d} (\mathbf{V}_{m+1:m+d})^T. \tag{3.20}$$

REMARK 3.9. *We view (3.20) as approximations of the posteriors resulting from the full-rank prior. Instead, we could also view (3.20) as posteriors from rank-$(m+d)$ approximations of the prior $\mathcal{N}(\mathbf{x}_0, \widehat{\mathbf{\Gamma}}_0)$ with $\widehat{\mathbf{\Gamma}}_0 = \mathbf{V}_{1:m+d} \mathbf{\Phi}_{1:m+d} (\mathbf{V}_{1:m+d})^T$. This interpretation of (3.20) is discussed in the supplement. However, from a practical point of view, explicit computation of $\widehat{\mathbf{\Gamma}}_0$ is too expensive and it is not necessary.*

Following the same argument as Theorem 3.7, one can express the underestimate (3.19) for the CG error in terms of the posterior covariance,

$$\operatorname{trace}(\mathbf{A} \widehat{\mathbf{\Gamma}}_m) = \sum_{i=m+1}^{m+d} \gamma_i \|\mathbf{r}_{i-1}\|_2^2 \leq \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2.$$

If the posterior distribution accurately models the uncertainty in the solution, then we expect (3.14) to hold. This means the accuracy of the uncertainty from the approximate Krylov posterior is related to the accuracy of the underestimate (3.19).

Algorithm 3.1 represents an efficient computation of BayesCG under rank-d approximate Krylov posteriors, and consists of two loops[2]:

1. Run CG until convergence in iteration m and compute the posterior mean $\mathbf{x}_m$
2. Run d additional CG iterations and compute the factors $\mathbf{V}_{m+1:m+d}$ and $\mathbf{\Phi}_{m+1:m+d}$ of the rank-d approximate posterior $\widehat{\mathbf{\Gamma}}_m$.

Correctness. Theorem 3.3 asserts that posteriors of BayesCG under the Krylov prior have means that are identical to CG iterates, and covariances that can be maintained in factored form involving submatrices of $\mathbf{V}$ and $\mathbf{\Phi}$ from Definition 3.6. The rank d of $\widehat{\Gamma}_m$ has the same purpose as the 'delay' in CG error estimation: a small number of additional iterations to capture the error, and $\operatorname{trace}(\mathbf{A}\widehat{\mathbf{\Gamma}}_m) = \operatorname{trace}(\mathbf{\Phi}_{m+1:m+d})$ is equal to the error underestimate (3.19). As a termination criterion one can choose the usual residual norm, or a statistically motivated criterion.

Computational cost. Algorithm 3.1 performs fewer arithmetic operations than Algorithm 2.1. Specifically, Algorithm 3.1 runs $m+d$ iterations of Algorithm 2.2, and a total of $m+d$ matrix vector products with $\mathbf{A}$ and storage of at most $d+2$ vectors. This is less than Algorithm 2.1, which requires $2m$ matrix vector products with $\mathbf{A}$, m matrix vector products with $\mathbf{\Sigma}_0$, and storage of $m+2$ vectors.

In addition, Algorithm 2.1 requires reorthogonalization to ensure positive semi-definiteness of the posterior covariances [9, Section 6.1]. In contrast, Algorithm 3.1 maintains the Krylov posteriors in factored form, thus (i) ensuring symmetric positive semi-definiteness by design; and (ii) reducing the cost of sampling, because the factorizations $\mathbf{\Sigma}_m = \mathbf{F}_m \mathbf{F}_m^T$ are readily available without any computations. The

[2]The partition of Algorithm 3.1 into two loops is for the purpose expositional clarity. Alternatively, everything could have been merged into a single loop with a conditional.

Algorithm 3.1 BayesCG under rank-d approximations of specific Krylov posterior covariances

Inputs: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}_0 \in \mathbb{R}^n$, $d \geq 1$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ ▷ define initial values
$\mathbf{v}_1 = \mathbf{r}_0$
$m = 0$
while not converged **do** ▷ CG recursions for posterior means
  $m = m + 1$
  $\eta_m = \mathbf{v}_m^T \mathbf{A}\mathbf{v}_m$
  $\gamma_m = (\mathbf{r}_{m-1}^T \mathbf{r}_{m-1})/\eta_m$
  $\mathbf{x}_m = \mathbf{x}_{m-1} + \gamma_i \mathbf{v}_i$
  $\mathbf{r}_m = \mathbf{r}_{m-1} - \gamma_i \mathbf{A}\mathbf{v}_i$
  $\delta_m = (\mathbf{r}_m^T \mathbf{r}_m)/(\mathbf{r}_{m-1}^T \mathbf{r}_{m-1})$
  $\mathbf{v}_{m+1} = \mathbf{r}_m + \delta_m \mathbf{v}_m$
end while
$d = \min\{d, \mathrm{K} - m\}$ ▷ compute full rank posterior if $d > \mathrm{K} - m$
$\mathbf{V}_{m+1:m+d} = \mathbf{0}_{n\times d}$ ▷ define posterior factor matrices
$\mathbf{\Phi}_{m+1:m+d} = \mathbf{0}_{d\times d}$
for $j = m + 1 : m + d$ **do** ▷ d additional iterations for posterior covariance
  $\eta_j = \mathbf{v}_j^T \mathbf{A}\mathbf{v}_j$
  $\gamma_j = (\mathbf{r}_{j-1}^T \mathbf{r}_{j-1})/\eta_j$
  $\mathbf{V}_j = \mathbf{v}_j/\eta_j$ ▷ store column j of $\mathbf{V}$
  $\mathbf{\Phi}_j = \gamma_j \|\mathbf{r}_{j-1}\|_2^2$ ▷ store element j of $\mathbf{\Phi}$
  $\mathbf{r}_j = \mathbf{r}_{j-1} - \gamma_j \mathbf{A}\mathbf{v}_j$
  $\delta_j = (\mathbf{r}_j^T \mathbf{r}_j)/(\mathbf{r}_{j-1}^T \mathbf{r}_{j-1})$
  $\mathbf{v}_{j+1} = \mathbf{r}_j + \delta_j \mathbf{v}_j$
end for
Output: $\mathbf{x}_m$, $\mathbf{V}_{m+1:m+d}$, $\mathbf{\Phi}_{m+1:m+d}$

last point is important, since the posterior is propagated to subsequent computations which sample from it to probe the effect of the uncertainty in the linear solve. So far, analytical propagation of the posterior has proved elusive, and empirical propagation is our only option.

4. Numerical experiments. We present numerical experiments to compare (i) Algorithm 3.1 under full or rank-d approximations of specific Krylov posteriors with (ii) Algorithm 2.1 under the inverse prior. After describing the experimental set up (section 4.1), we apply the algorithms to two matrices: a matrix of small dimension (section 4.2), and one of larger dimension (section 4.3).

4.1. Set up of the numerical experiments. We describe the linear systems in the experiments, reorthogonalization in the algorithms, and sampling from the posterior distributions.[3]

Linear systems. We consider two types of symmetric positive-definite linear systems $\mathbf{A}\mathbf{x}_* = \mathbf{b}$: one with a dense matrix $\mathbf{A}$ of dimension $n = 100$, and the other with a sparse preconditioned matrix $\mathbf{A}$ of dimension $n = 11948$. We fix the solution $\mathbf{x}_*$, and compute the right hand side from $\mathbf{b} = \mathbf{A}\mathbf{x}_*$.

[3]The Python code used in the numerical experiments can be found at https://github.com/treid5/ProbNumCG_Supp

For $n = 100$, the matrix is $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T$ [22, Section 2], where $\mathbf{Q}$ is a random[4] orthogonal matrix with Haar distribution [47, Section 3], and $\mathbf{D}$ is a diagonal matrix with eigenvalues [20]

$$d_{ii} = (10^3)^{(i-1)/99}, \qquad 1 \leq i \leq 100. \tag{4.1}$$

The condition number is $\kappa(\mathbf{A}) = 10^3$, and the solution $\mathbf{x}_*$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$.

For $n = 11948$, the matrix $\mathbf{A} = \mathbf{L}^{-1}\mathbf{B}\mathbf{L}^{-T}$ is a sparse preconditioned matrix where $\mathbf{B}$ is `BCSSTK18` from the Harwell-Boeing collection [33], and $\mathbf{L}$ is the incomplete Cholesky factorization [21, Section 11.1] of the diagonally shifted matrix

$$\widetilde{\mathbf{B}} = \mathbf{B} + 9.0930 \cdot 10^8 \cdot \mathrm{diag}(\mathbf{B}) \qquad \text{with} \qquad \max_{1 \leq i \leq n} \left\{ -b_{ii} + \sum_{j \neq i} b_{ij} \right\} = 9.0930 \cdot 10^8.$$

The shift forces $\widetilde{\mathbf{B}}$ to be diagonally dominant. We compute the factorization of $\widetilde{\mathbf{B}}$ with a threshold drop tolerance 10^{-6} to make $\mathbf{L}$ diagonal. The condition number is $\kappa(\mathbf{A}) \approx 1.57 \cdot 10^6$, and the solution $\mathbf{x}_* = \mathbf{1}$ is the all ones vector.

Reorthogonalization. Since the posterior covariances in Algorithm 2.1 become indefinite when the search directions lose orthogonality, reorthogonalization of the search directions is recommended in every iteration, [9, Section 6.1] and [11, Section 4.1]. Following [22, Section 2], we reorthogonalize the residual vectors instead, as it has the additional advantage of better numerical stability in our experience. Reorthogonalization consists of classical Gram-Schmidt performed twice because it is efficient, easy to implement, and produces vectors orthogonal to almost machine precision [15, 16].

Sampling from the Gaussian distributions. We exploit the stability of Gaussians, see section 2.1, to sample from $\mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$ as follows. Let $\mathbf{\Sigma} = \mathbf{F}\mathbf{F}^T$ be a factorization of the covariance with $\mathbf{F} \in \mathbb{R}^{n \times d}$. Sample a standard Gaussian vector[5] $Z \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$; multiply it by $\mathbf{F}$; and add the mean to obtain $X \equiv \mathbf{x} + \mathbf{F}Z \sim \mathcal{N}(\mathbf{x}, \mathbf{F}\mathbf{F}^T)$.

By design, the rank-d approximate Krylov posteriors are maintained in factored form

$$\widehat{\mathbf{\Gamma}}_m = \mathbf{F}_m \mathbf{F}_m^T \qquad \text{where} \qquad \mathbf{F}_m \equiv \mathbf{V}_{m+1:m+d}\, \mathbf{\Phi}_{m+1:m+d}^{1/2} \in \mathbb{R}^{n \times d}.$$

For all other posteriors $\mathbf{\Sigma}_m$, we factor the matrix square root [27, Chapter 6] of the matrix absolute value [27, Chapter 8] of $\mathbf{\Sigma}_m$[6]. Factoring the absolute value of $\mathbf{\Sigma}_m$ enforces positive semi-definiteness of the posteriors which may be lost if BayesCG is implemented without reorthogonalization.

Convergence. We display convergence of the mean and covariance with $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$ and $\mathrm{trace}(\mathbf{A}\mathbf{\Sigma}_m)$. In addition, we sample from the posterior, $X \sim \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$ and compare the resulting estimate $\|X - \mathbf{x}_m\|_{\mathbf{A}}^2$ to the error $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$. If the samples X are accurate estimates, then the posterior distribution is likely to be a reliable indicator of the uncertainty in the solution $\mathbf{x}_*$.

[4] The exact random matrix can be reproduced with the python files in our code repository because we specified the random seed.

[5] Most scientific computing packages come with built in functions for sampling from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In Matlab and Julia the function is `randn` and in Python it is `numpy.random.randn`.

[6] The matrix absolute value of $\mathbf{B} \in \mathbb{R}^{n \times n}$ is $\mathrm{abs}(\mathbf{B}) = (\mathbf{B}^T\mathbf{B})^{1/2}$. If $\mathbf{B}$ is symmetric positive semi-definite, then $\mathrm{abs}(\mathbf{B}) = \mathbf{B}$. Otherwise, the square root of the absolute value is $(\mathrm{abs}(\mathbf{B}))^{1/2} = \mathbf{V}\mathbf{S}^{1/2}\mathbf{V}^T$, where $\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ is a SVD.

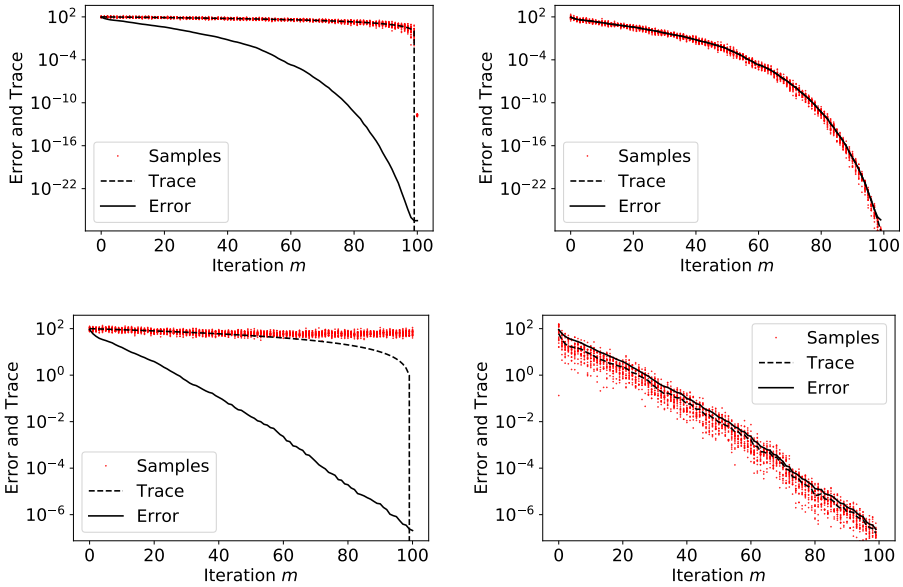

FIGURE 4.1. *Error estimates* $\|X - \mathbf{x}_m\|_{\mathbf{A}}^2$ *and* $\operatorname{trace}(A\boldsymbol{\Sigma}_m)$ *from samples* $X \sim \mathcal{N}(\mathbf{x}_m, \boldsymbol{\Sigma}_m)$, *for the matrix with small dimension* $n = 100$. *Top row: Algorithm 2.1 with reorthogonalization under the inverse prior (left panel), and Algorithm 3.1 under the full Krylov prior (right panel). Bottom row: Algorithm 2.1 without reorthogonalization under the inverse prior (left panel), and Algorithm 3.1 under the rank-5 approximate Krylov prior (right panel).*

4.2. Matrix with small dimension. We compare Algorithm 2.1 under the inverse prior, with Algorithm 3.1 under full or rank-5 approximate Krylov posteriors when applied to the matrix with small dimension $n = 100$.

Figure 4.1 illustrates that the posterior means converge at the same speed, regardless of reorthogonalization. However, without reorthogonalization, the convergence is slower.

Algorithm 2.1 under the inverse prior. The posterior covariances converge more slowly than the squared errors of the means. Without reorthogonalization, the posterior covariances are indefinite, and the error estimates from the posterior samples diverge from $\operatorname{trace}(\mathbf{A}\boldsymbol{\Sigma}_m)$ and violate Lemma 3.5. Thus, posteriors from BayesCG under the inverse prior are not reliable indicators of uncertainty.

Algorithm 3.1 under full or approximate Krylov priors. The quantity $\operatorname{trace}(\mathbf{A}\boldsymbol{\Sigma}_m)$ equals the error for full rank Krylov posteriors, while it underestimates the error for rank-5 approximate posteriors. Error estimates from samples of Krylov posteriors are significantly more accurate than those from the inverse posteriors. Thus, posteriors from BayesCG under (approximate) Krylov priors are more reliable indicators uncertainty.

4.3. Matrix with larger dimension. We compare Algorithm 3.1 under rank-1 and rank-50 approximate Krylov posteriors, when applied to the matrix with large dimension $n = 11948$.

Figure 4.2 illustrates that the traces of the posterior covariances underestimate the error. However, the trace of the rank-50 approximate Krylov covariance is more accurate, because CG error estimates (3.19) are more accurate for larger delays [49,

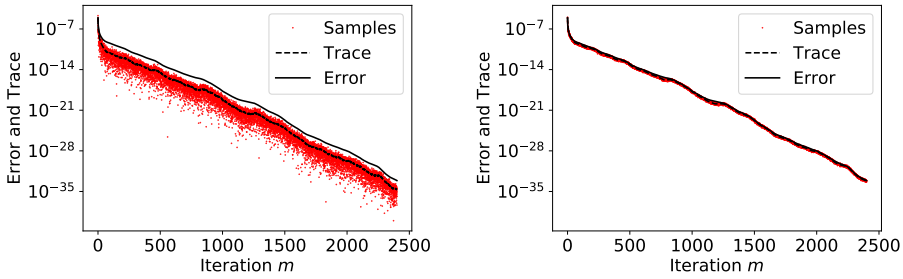

FIGURE 4.2. *Error estimates* $\|X - \mathbf{x}_m\|_{\mathbf{A}}^2$ *and* $\mathrm{trace}(\mathbf{A}\Sigma_m)$ *from samples* $X \sim \mathcal{N}(\mathbf{x}_m, \boldsymbol{\Sigma}_m)$*, for the matrix with large dimension* $n = 11948$*. Left: Algorithm 3.1 under rank-1 approximate Krylov posterior. Right: Algorithm 3.1 under rank-50 approximate Krylov posterior.*

Section 4]. As expected, error estimates from rank-50 posterior samples are more tightly concentrated around the true error than those of rank-1 posterior samples. Thus, BayesCG under higher rank approximate posteriors produces more reliable indicators of uncertainty.

5. Conclusion. BayesCG is our 'uncertainty-aware' version of CG, that is, a probabilistic numerical extension of CG that produces a probabilistic model of the uncertainty about our knowledge of the solution $\mathbf{x}_*$ due to early termination of CG. Under our Krylov prior, BayesCG produces iterates that are identical to those of CG (in exact arithmetic), thus converges at the same speed as CG; and its posterior distributions can be cheaply approximated. Samples from the Krylov posterior and its low rank approximations produce accurate error estimates, thus represent realistic indicators of the uncertainty about $\mathbf{x}_*$.

Future work. In a forthcoming paper, we focus on the statistical aspects of BayesCG under the Krylov prior. More specifically, we quantify the approximation error of low rank approximate Krylov posteriors and investigate the calibration of BayesCG under low-rank approximate Krylov posteriors.

In a separate paper, we assess the effect of CG accuracy in a computational pipeline in the form of a randomized algorithm for generalized singular value decomposition [44] with applications to hyper-differential sensitivity analysis [23].

Appendix A. Proofs of Theorems 2.4, 2.6 and 2.7.

Proof of Theorem 2.4. The proof is inspired by the proof of [10, Proposition 3] for nonsingular $\boldsymbol{\Sigma}_0$. For singular $\boldsymbol{\Sigma}_0$, we replace the inverse by the Moore-Penrose inverse which satisfies

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_0^{\dagger} \boldsymbol{\Sigma}_0. \tag{A.1}$$

The assumption $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\boldsymbol{\Sigma}_0)$ implies that there exists $\mathbf{y} \in \mathbb{R}^n$ so that

$$\mathbf{x}_* - \mathbf{x}_0 = \boldsymbol{\Sigma}_0 \mathbf{y} = \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_0^{\dagger} \boldsymbol{\Sigma}_0 \mathbf{y} = \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_0^{\dagger} (\mathbf{x}_* - \mathbf{x}_0). \tag{A.2}$$

The proof proceeds in four steps.

Range of $\mathbf{P}_m$*.* On the one hand (2.3) implies

$$\mathrm{range}(\mathbf{P}_m) = \mathrm{range}\left(\boldsymbol{\Sigma}_0 \mathbf{A} \mathbf{S}_m \boldsymbol{\Lambda}_m^{-1} \mathbf{S}_m^T \mathbf{A} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_0^{\dagger} \right) \subset \mathrm{range}\left(\boldsymbol{\Sigma}_0 \mathbf{A} \mathbf{S}_m \right).$$

On the other hand (2.3) and (A.1) imply

$$\mathbf{P}_m\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m = \boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m\boldsymbol{\Lambda}_m^{-1}\overbrace{\mathbf{S}_m^T\mathbf{A}\underbrace{\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\dagger\boldsymbol{\Sigma}_0}_{\boldsymbol{\Sigma}_0}\mathbf{A}\mathbf{S}_m}^{\boldsymbol{\Lambda}_m} = \boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m$$

so that $\mathrm{range}(\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m) \subset \mathrm{range}(\mathbf{P}_m)$.

Combining the two inclusions gives $\mathrm{range}(\mathbf{P}_m) = K_m \equiv \mathrm{range}(\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m)$.

$\mathbf{P}_m$ is a $\boldsymbol{\Sigma}_0^\dagger$-orthogonal projector. The above implies

$$\mathbf{P}_m^2 = \underbrace{\mathbf{P}_m\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m}_{\boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m}\boldsymbol{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{A}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\dagger = \boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m\boldsymbol{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{A}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\dagger = \mathbf{P}_m. \tag{A.3}$$

Thus $\mathbf{P}_m$ is a projector. The $\boldsymbol{\Sigma}_0^\dagger$-orthogonality of $\mathbf{P}_m$ follows from the symmetry of $\boldsymbol{\Sigma}_0^\dagger\mathbf{P}$.

Posterior mean. From (2.1), (A.2), and (2.3) follows

$$\begin{aligned}\mathbf{x}_m &= \mathbf{x}_0 + \boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m\boldsymbol{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{A}(\mathbf{x}_* - \mathbf{x}_0)\\ &= \mathbf{x}_0 + \boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m\boldsymbol{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{A}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\dagger(\mathbf{x}_* - \mathbf{x}_0) = (\mathbf{I} - \mathbf{P}_m)\mathbf{x}_0 + \mathbf{P}_m\mathbf{x}_*.\end{aligned}$$

Posterior covariance. From (2.2), (A.1) and (2.3) follows

$$\begin{aligned}\boldsymbol{\Sigma}_m &= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m\boldsymbol{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{A}\boldsymbol{\Sigma}_0\\ &= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m\boldsymbol{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{A}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^\dagger\boldsymbol{\Sigma}_0 = (\mathbf{I} - \mathbf{P}_m)\boldsymbol{\Sigma}_0.\end{aligned}$$

Multiply $\boldsymbol{\Sigma}_m$ on the left by $\mathbf{P}_m$ and apply (A.3) to obtain $\mathbf{P}_m\boldsymbol{\Sigma}_m = \mathbf{P}_m(\mathbf{I}-\mathbf{P}_m)\boldsymbol{\Sigma}_0 = \mathbf{0}$. ∎

The proof of Theorem 2.6 relies on the next three results related to semi-definite inner product spaces and orthogonal projectors in those spaces.

LEMMA A.1. *Under the assumptions of Theorem 2.1, if $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\boldsymbol{\Sigma}_0)$, then $\mathbf{x}_* - \mathbf{x}_m \in \mathrm{range}(\boldsymbol{\Sigma}_0)$, $1 \le m \le n$.*

Proof. Subtract from $\mathbf{x}_*$ both sides of the posterior mean (2.1),

$$\mathbf{x}_* - \mathbf{x}_m = (\mathbf{x}_* - \mathbf{x}_0) - \boldsymbol{\Sigma}_0\mathbf{A}\mathbf{S}_m\boldsymbol{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{A}(\mathbf{x}_* - \mathbf{x}_0), \qquad 1 \le m \le n.$$

The first summand $\mathbf{x}_* - \mathbf{x}_0$ is in $\mathrm{range}(\boldsymbol{\Sigma}_0)$ by assumption, and the second one by design, hence so is the sum. ∎

LEMMA A.2. *Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ be symmetric positive semi-definite. If $\mathbf{z} \in \mathrm{range}(\mathbf{B})$, then $\mathbf{z}^T\mathbf{B}\mathbf{z} = 0$ if and only if $\mathbf{z} = \mathbf{0}$.*

Proof. Since $\mathbf{B}$ is symmetric positive semi-definite, we can factor $\mathbf{F}\mathbf{F}^T = \mathbf{B}$, where $\mathbf{F}$ has full column rank. Let $\mathbf{w} = \mathbf{F}^T\mathbf{z}$. From $\mathbf{z} \in \mathrm{range}(\mathbf{B}) = \mathrm{range}(\mathbf{F})$, and $\mathrm{range}(\mathbf{F}) = \ker(\mathbf{F}^T)^\perp$ follows that $\mathbf{w} = \mathbf{F}^T\mathbf{z} = \mathbf{0}$ if and only if $\mathbf{z} = \mathbf{0}$. Therefore $\mathbf{w}^T\mathbf{w} = \mathbf{z}^T\mathbf{B}\mathbf{z} = 0$ if and only if $\mathbf{z} = \mathbf{0}$. ∎

LEMMA A.3. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a subspace, $\mathbf{B} \in \mathbb{R}^{n\times n}$ symmetric positive semi-definite, and $\mathbf{v} \in \mathbb{R}^n$. If $\mathbf{P}$ is a $\mathbf{B}$-orthogonal projector onto $\mathcal{X}$, then*

$$\arg\min_{\mathbf{x}\in\mathcal{X}}(\mathbf{v} - \mathbf{x})^T\mathbf{B}(\mathbf{v} - \mathbf{x}) = \{\mathbf{x} \in \mathcal{X} : (\mathbf{x} - \mathbf{P}\mathbf{v})^T\mathbf{B}(\mathbf{x} - \mathbf{P}\mathbf{v}) = 0\}.$$

If additionally $\mathcal{X} \subseteq \mathrm{range}(\mathbf{B})$*, then*

$$\arg\min_{\mathbf{x}\in\mathcal{X}} (\mathbf{v}-\mathbf{x})^T\mathbf{B}(\mathbf{v}-\mathbf{x}) = \mathbf{P}\mathbf{v}.$$

Proof. After proving the general case, we show that the minimizer is unique if $\mathcal{X} \subseteq \mathrm{range}(\mathbf{B})$.

General case. Abbreviate the induced semi-norm by $|\mathbf{z}|_{\mathbf{B}}^2 = \mathbf{z}^T\mathbf{B}\mathbf{z}$. Since $\mathbf{P}$ is a projector onto $\mathcal{X}$, we can write $\mathbf{x} = \mathbf{P}\mathbf{x}$ for $\mathbf{x} \in \mathcal{X}$. Add and subtract $\mathbf{P}\mathbf{v}$ inside the norm to obtain a Pythagoras-like theorem,

$$\begin{aligned}|\mathbf{v}-\mathbf{x}|_{\mathbf{B}}^2 &= |(\mathbf{I}-\mathbf{P})\mathbf{v}+\mathbf{P}(\mathbf{v}-\mathbf{x})|_{\mathbf{B}}^2 \\ &= |(\mathbf{I}-\mathbf{P})\mathbf{v}|_{\mathbf{B}}^2 + |\mathbf{P}(\mathbf{v}-\mathbf{x})|_{\mathbf{B}}^2 + 2\mathbf{v}^T\underbrace{(\mathbf{I}-\mathbf{P})^T\mathbf{B}\mathbf{P}}_{=\mathbf{0}}(\mathbf{v}-\mathbf{x}) \\ &= |(\mathbf{I}-\mathbf{P})\mathbf{v}|_{\mathbf{B}}^2 + |\mathbf{P}\mathbf{v}-\mathbf{x}|_{\mathbf{B}}^2.\end{aligned}$$

Since the first summand is independent of $\mathbf{x}$, the minimum is achieved if the second summand is zero.

Uniqueness. Since $\mathbf{P}$ is a projector onto $\mathcal{X}$, $\mathbf{P}\mathbf{v} \in \mathcal{X}$. From $\mathcal{X} \subseteq \mathrm{range}(\mathbf{B})$ follows $\mathbf{P}\mathbf{v} \in \mathrm{range}(\mathbf{B})$ and $\mathbf{x} \in \mathrm{range}(\mathbf{B})$. With Lemma A.2 this implies: $|\mathbf{P}\mathbf{v}-\mathbf{x}|_{\mathbf{B}}^2 = 0$ only if $\mathbf{P}\mathbf{v} = \mathbf{x}$. ∎

Proof of Theorem 2.6. This is similar to [1, Proof of Proposition 4]. Minimizing (2.4) over the affine space $\mathbf{x}_0 + K_m = \mathbf{x}_0 + \mathrm{range}(\mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_m)$ is equivalent to shifting by $\mathbf{x}_0$ and minimizing over K_m,

$$\min_{x\in\mathbf{x}_0+K_m} (\mathbf{x}_*-\mathbf{x})^T\mathbf{\Sigma}_0^\dagger(\mathbf{x}_*-\mathbf{x}) = \min_{\mathbf{x}\in K_m} ((\mathbf{x}_*-\mathbf{x}_0)-\mathbf{x})^T\mathbf{\Sigma}_0^\dagger((\mathbf{x}_*-\mathbf{x}_0)-\mathbf{x}).$$

Since $\mathbf{\Sigma}_0$ is symmetric, the $\mathbf{\Sigma}_0^\dagger$-orthogonal projector $\mathbf{P}_m$ from Theorem 2.4 satisfies $\mathrm{range}(\mathbf{P}_m) = K_m \subseteq \mathrm{range}(\mathbf{\Sigma}_0) = \mathrm{range}(\mathbf{\Sigma}_0^\dagger)$. Therefore, Lemma A.3 implies

$$\arg\min_{\mathbf{x}\in K_m} ((\mathbf{x}_*-\mathbf{x}_0)-\mathbf{x})^T\mathbf{\Sigma}_0^\dagger((\mathbf{x}_*-\mathbf{x}_0)-\mathbf{x}) = \mathbf{P}(\mathbf{x}_*-\mathbf{x}_0).$$

From Theorem 2.4 and $K_m = \mathrm{range}(\mathbf{P}_m)$ follows $\mathbf{x}_m - \mathbf{x}_0 = \mathbf{P}_m(\mathbf{x}_*-\mathbf{x}_0) \in K_m$. Thus $\mathbf{x}_m \in \mathbf{x}_0 + K_m$ is the minimizer.

The symmetry of $\mathbf{\Sigma}_m$ and Lemmas A.1 and A.2 imply that $(\mathbf{x}_*-\mathbf{x}_m)^T\mathbf{\Sigma}_0^\dagger(\mathbf{x}_*-\mathbf{x}_m) = 0$ only if $\mathbf{x}_m = \mathbf{x}_*$. ∎

Proof of Theorem 2.7. Recursion (2.6) was shown in [9, Proposition 6]. The following proof for (2.7) is analogous to [11, Proof of Proposition 6]. From (2.2) follows that the posterior covariance at iteration i amounts to a rank-i downdate of the prior,

$$\mathbf{\Sigma}_i = \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_i\mathbf{\Lambda}_i^{-1}\left(\mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_i\right)^T, \qquad 1 \le i \le m.$$

Here $\mathbf{\Lambda}_i$ is diagonal due to the $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$-orthogonality of the search directions, hence a rank-i downdate can be computed as a recursive sequence of i rank-1 downdates,

$$\mathbf{\Sigma}_i = \underbrace{\mathbf{\Sigma}_0 - \mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_{i-1}\mathbf{\Lambda}_{i-1}^{-1}(\mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_{i-1})^T}_{\mathbf{\Sigma}_{i-1}} - \frac{\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_i\left(\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_i\right)^T}{\mathbf{s}_i^T\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_i}.$$ ∎

Appendix B. Auxiliary results.

LEMMA B.1 (Lemma S3 in [11]). *Under the assumptions of Theorem 2.7,*

$$\mathbf{s}_j^T \mathbf{r}_i = 0, \qquad 1 \leq j \leq i \leq m.$$

LEMMA B.2 (Sections 3.2b.1–3.2b.3 in [32]). *Let $Z \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$ be a Gaussian random variable with mean $\mathbf{x} \in \mathbb{R}^n$ and covariance $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$, and let $\mathbf{B} \in \mathbb{R}^{n \times n}$ be symmetric positive definite. The mean and variance of $Z^T \mathbf{B} Z$ are*

$$\mathbb{E}[Z^T \mathbf{B} Z] = \operatorname{trace}(\mathbf{B}\mathbf{\Sigma}) + \mathbf{x}^T \mathbf{B} \mathbf{x},$$
$$\mathbb{V}[Z^T \mathbf{B} Z] = 2 \operatorname{trace}((\mathbf{B}\mathbf{\Sigma})^2) + 4\, \mathbf{x}^T \mathbf{B}\mathbf{\Sigma}\mathbf{B}\mathbf{x}.$$

Acknowledgments. We thank Eric Hallman, Joseph Hart, and the members of the NCSU Randomized Numerical Analysis RTG for helpful discussions. We are also most grateful to the reviewers for their recommendations that helped to ensure mathematical correctness and improve exposition.

REFERENCES

[1] S. BARTELS, J. COCKAYNE, I. C. F. IPSEN, AND P. HENNIG, *Probabilistic linear solvers: a unifying view*, Stat. Comput., 29 (2019), pp. 1249–1263, https://doi.org/10.1007/s11222-019-09897-7.

[2] S. BARTELS AND P. HENNIG, *Probabilistic approximate least-squares*, in Proc. 19th Int. Conf. Artificial Intelligence and Statistics, vol. 51 of Proc.Machine Learning Research, MLR Press, 2016, pp. 676—684.

[3] M. BERLJAFA AND S. GÜTTEL, *Generalized rational Krylov decompositions with an application to rational approximation*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 894–916, https://doi.org/10.1137/140998081.

[4] F.-X. BRIOL, C. J. OATES, M. GIROLAMI, M. A. OSBORNE, AND D. SEJDINOVIC, *Probabilistic Integration: A Role in Statistical Computation?*, Statist. Sci., 34 (2019), pp. 1 – 22, https://doi.org/10.1214/18-STS660.

[5] D. CALVETTI, *Contributed discussion for "A Bayesian conjugate gradient method"*, Bayesian Anal., 14 (2019), pp. 937–1012, https://doi.org/10.1214/19-BA1145.

[6] J. COCKAYNE, M. M. GRAHAM, C. J. OATES, AND T. J. SULLIVAN, *Testing whether a learning procedure is calibrated*, 2021, https://arxiv.org/abs/2012.12670. arXiv:2012.12670.

[7] J. COCKAYNE, I. C. F. IPSEN, C. J. OATES, AND T. W. REID, *Probabilistic iterative methods for linear systems*, J. Mach. Learn. Res., 22 (232) (2021), pp. 1–34.

[8] J. COCKAYNE, C. OATES, T. SULLIVAN, AND M. GIROLAMI, *Probabilistic numerical methods for PDE-constrained Bayesian inverse problems*, AIP Conference Proceedings, 1853 (2017), p. 060001, https://doi.org/10.1063/1.4985359.

[9] J. COCKAYNE, C. J. OATES, I. C. F. IPSEN, AND M. GIROLAMI, *A Bayesian conjugate gradient method (with discussion)*, Bayesian Anal., 14 (2019), pp. 937–1012, https://doi.org/10.1214/19-BA1145. Includes 6 discussions and a rejoinder from the authors.

[10] J. COCKAYNE, C. J. OATES, I. C. F. IPSEN, AND M. GIROLAMI, *Rejoinder for "A Bayesian conjugate gradient method"*, Bayesian Anal., 14 (2019), pp. 937–1012, https://doi.org/10.1214/19-BA1145.

[11] J. COCKAYNE, C. J. OATES, I. C. F. IPSEN, AND M. GIROLAMI, *Supplementary material for 'A Bayesian conjugate-gradient method'*, Bayesian Anal., (2019), https://doi.org/10.1214/19-BA1145SUPP.

[12] J. COCKAYNE, C. J. OATES, T. J. SULLIVAN, AND M. GIROLAMI, *Bayesian probabilistic numerical methods*, SIAM Rev., 61 (2019), pp. 756–789, https://doi.org/10.1137/17M1139357.

[13] V. FANASKOV, *Uncertainty calibration for probabilistic projection methods*, Stat. Comput., 31 (2021), pp. Paper No. 56, 17, https://doi.org/10.1007/s11222-021-10031-9, https://doi.org/10.1007/s11222-021-10031-9.

[14] A. GESSNER, O. KANJILAL, AND P. HENNIG, *Integrals over Gaussians under linear domain constraints*, in Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, S. Chiappa and R. Calandra, eds., vol. 108 of Proceedings of Machine Learning Research, 2020, pp. 2764–2774, http://proceedings.mlr.press/v108/gessner20a.html.

[15] L. Giraud, J. Langou, M. Rozložník, and J. van den Eshof, *Rounding error analysis of the classical Gram-Schmidt orthogonalization process*, Numer. Math., 101 (2005), pp. 87–100, https://doi.org/10.1007/s00211-005-0615-4.
[16] L. Giraud, J. Langou, and M. Rozloznik, *The loss of orthogonality in the Gram-Schmidt orthogonalization process*, Comput. Math. Appl., 50 (2005), pp. 1069–1075, https://doi.org/10.1016/j.camwa.2005.08.009.
[17] G. H. Golub and G. Meurant, *Matrices, moments and quadrature*, in Numerical analysis 1993 (Dundee, 1993), vol. 303 of Pitman Res. Notes Math. Ser., Longman Sci. Tech., Harlow, 1994, pp. 105–156.
[18] G. H. Golub and G. Meurant, *Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705, https://doi.org/10.1007/BF02510247.
[19] G. H. Golub and Z. Strakoš, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268, https://doi.org/10.1007/BF02142693.
[20] A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551, https://doi.org/10.1137/S0895479895284944.
[21] A. Greenbaum, *Iterative methods for solving linear systems*, vol. 17 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997, https://doi.org/10.1137/1.9781611970937.
[22] A. Greenbaum and Z. Strakoš, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 121–137, https://doi.org/10.1137/0613011.
[23] J. Hart, B. van Bloemen Waanders, and R. Herzog, *Hyperdifferential sensitivity analysis of uncertain parameters in PDE-constrained optimization*, Int. J. for Uncertain. Quantif., 10 (2020), pp. 225–248, https://doi.org/10.1615/Int.J.UncertaintyQuantification.2020032480.
[24] P. Hennig, *Probabilistic interpretation of linear solvers*, SIAM J. Optim., 25 (2015), pp. 234–260, https://doi.org/10.1137/140955501.
[25] P. Hennig, M. A. Osborne, and M. Girolami, *Probabilistic numerics and uncertainty in computations*, Proc. A., 471 (2015), pp. 20150142, 17.
[26] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436, https://doi.org/10.6028/jres.049.044.
[27] N. J. Higham, *Functions of matrices. Theory and computation*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008, https://doi.org/10.1137/1.9780898717778.
[28] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, second ed., 2013.
[29] T. Karvonen, C. J. Oates, and S. Sarkka, *A Bayes-Sard cubature method*, in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., vol. 31, Curran Associates, Inc., 2018, https://proceedings.neurips.cc/paper/2018/file/6775a0635c302542da2c32aa19d86be0-Paper.pdf.
[30] L. Li and E. X. Fang, *Invited discussion for "A Bayesian conjugate gradient method"*, Bayesian Anal., 14 (2019), pp. 937–1012, https://doi.org/10.1214/19-BA1145.
[31] J. Liesen and Z. Strakos, *Krylov Subspace Methods: Principles and Analysis*, Oxford University Press, 2013.
[32] A. M. Mathai and S. B. Provost, *Quadratic forms in random variables: Theory and applications*, Dekker, 1992.
[33] Matrix Market, *BCSSTK18: BCS Structural Engineering Matrices (linear equations) R.E. Ginna Nuclear Power Station*, https://math.nist.gov/MatrixMarket/data/Harwell-Boeing/bcsstruc2/bcsstk18.html.
[34] T. Matsuda and Y. Miyatake, *Estimation of ordinary differential equation models with discretization error quantification*, SIAM/ASA J. Uncertain. Quantif., 9 (2021), pp. 302–331, https://doi.org/10.1137/19M1278405.
[35] G. Meurant, *The computation of bounds for the norm of the error in the conjugate gradient algorithm*, Numer. Algorithms, 16 (1997), pp. 77–87 (1998), https://doi.org/10.1023/A:1019178811767. Sparse matrices in industry (Lille, 1997).
[36] G. Meurant and P. Tichý, *On computing quadrature-based bounds for the A-norm of the error in conjugate gradients*, Numer. Algorithms, 62 (2013), pp. 163–191, https://doi.org/10.1007/s11075-012-9591-9.
[37] G. Meurant and P. Tichý, *Approximating the extreme Ritz values and upper bounds for the A-norm of the error in CG*, Numer. Algorithms, 82 (2019), pp. 937–968, https://doi.org/

10.1007/s11075-018-0634-8.
[38] J. Močkus, *On Bayesian methods for seeking the extremum*, in Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974, G. I. Marchuk, ed., Berlin, Heidelberg, 1975, Springer Berlin Heidelberg, pp. 400–404.
[39] R. J. Muirhead, *Aspects of multivariate statistical theory*, John Wiley & Sons, Inc., New York, 1982. Wiley Series in Probability and Mathematical Statistics.
[40] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, second ed., 2006.
[41] C. J. Oates, J. Cockayne, R. G. Aykroyd, and M. Girolami, *Bayesian probabilistic numerical methods in time-dependent state estimation for industrial hydrocyclone equipment*, J. Amer. Statist. Assoc., 114 (2019), pp. 1518–1531, https://doi.org/10.1080/01621459.2019.1574583.
[42] C. J. Oates and T. J. Sullivan, *A modern retrospective on probabilistic numerics*, Stat. Comput., 29 (2019), pp. 1335–1351, https://doi.org/10.1007/s11222-019-09902-z.
[43] N. Petra, H. Zhu, G. Stadler, T. Hughes, and O. Ghattas, *An inexact Gauss-Newton method for inversion of basal sliding and rheology parameters in a nonlinear Stokes ice sheet model*, J. Glaciology, 58 (2012), p. 889–903, https://doi.org/10.3189/2012JoG11J182.
[44] A. K. Saibaba, J. Hart, and B. van Bloemen Waanders, *Randomized algorithms for generalized singular value decomposition with application to sensitivity analysis*, Numer. Linear Algebra Appl., (2021), p. e2364, https://doi.org/10.1002/nla.2364.
[45] F. Schäfer, T. J. Sullivan, and H. Owhadi, *Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity*, Multiscale Model. Simul., 19 (2021), pp. 688–730, https://doi.org/10.1137/19M129526X.
[46] J. Snoek, H. Larochelle, and R. P. Adams, *Practical Bayesian optimization of machine learning algorithms*, in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2, NIPS'12, Red Hook, NY, USA, 2012, Curran Associates Inc., p. 2951–2959.
[47] 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, https://doi.org/10.1137/0717034.
[48] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, 1990.
[49] Z. Strakoš and P. Tichý, *On error estimation in the conjugate gradient method and why it works in finite precision computations*, Electron. Trans. Numer. Anal., 13 (2002), pp. 56–80.
[50] Z. Strakoš and P. Tichý, *Error estimation in preconditioned conjugate gradients*, BIT, 45 (2005), pp. 789–817, https://doi.org/10.1007/s10543-005-0032-1.
[51] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559, https://doi.org/10.1017/S0962492910000061.
[52] F. Tronarp, H. Kersting, S. Särkkä, and P. Hennig, *Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective*, Stat. Comput., 29 (2019), pp. 1297–1315, https://doi.org/10.1007/s11222-019-09900-1.
[53] J. Wenger and P. Hennig, *Probabilistic linear solvers for machine learning*, 2020, https://arxiv.org/abs/2010.09691. arXiv:2010.09691.