Noname manuscript No.
(will be inserted by the editor)

Statistical Properties of the Probabilistic Numeric Linear Solver BayesCG

**Tim W. Reid, Ilse C. F. Ipsen,
Jon Cockayne, and Chris J. Oates**

July 23, 2022

Abstract We analyse the calibration of BayesCG under the Krylov prior, a probabilistic numeric extension of the Conjugate Gradient (CG) method for solving systems of linear equations with symmetric positive definite coefficient matrix. Calibration refers to the statistical quality of the posterior covariances produced by a solver. Since BayesCG is not calibrated in the strict existing notion, we propose instead two test statistics that are necessary but not sufficient for calibration: the Z-statistic and the new S-statistic. We show analytically and experimentally that under low-rank approximate Krylov posteriors,

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).

Tim W. Reid
North Carolina State University
Department of Mathematics
Raleigh, NC 27695-8205, US
twreid@alumni.ncsu.edu

Ilse C. F. Ipsen
North Carolina State University
Department of Mathematics
Raleigh, NC 27695-8205, US
ipsen@ncsu.edu

Jon Cockayne
University of Southampton
Department of Mathematical Sciences
Southampton, SO17 1BJ, UK
jon.cockayne@soton.ac.uk

Chris J. Oates
Newcastle University
School of Mathematics and Statistics
Newcastle-upon-Tyne, NE1 7RU, UK
chris.oates@ncl.ac.uk

BayesCG exhibits desirable properties of a calibrated solver, is only slightly optimistic, and is computationally competitive with CG.

1 Introduction

We present a rigorous analysis of the probabilistic numeric solver BayesCG under the Krylov prior [8,32] for solving systems of linear equations

$$\mathbf{A}\mathbf{x}_* = \mathbf{b}, \tag{1}$$

with symmetric positive definite coefficient matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$.

Probabilistic numerics. This area [10,18,29] seeks to quantify the uncertainty due to limited computational resources, and to propagate these uncertainties through computational pipelines—sequences of computations where the output of one computation is the input for the next. At the core of many computational pipelines are iterative linear solvers [7,16,28,31,35], whose computational resources are limited by the impracticality of running the solver to completion. The premature termination generates uncertainty in the computed solution.

Probabilistic numeric linear solvers. Probabilistic numeric extensions of Krylov space and stationary iterative methods [2,7,8,11,17,32,39] model the 'epistemic uncertainty' in a quantity of interest, which can be the matrix inverse $\mathbf{A}^{-1}$ [17,2,39] or the solution $\mathbf{x}_*$ [8,2,7,11]. Our quantity of interest is the solution $\mathbf{x}_*$, and the 'epistemic uncertainty' is the uncertainty in the user's knowledge of the true value of $\mathbf{x}_*$.

The probabilistic solver takes as input a *prior distribution* which models the initial uncertainty in $\mathbf{x}_*$ and then computes *posterior distributions* which model the uncertainty remaining after each iteration. Figure 1 depicts a prior and posterior distribution for the solution $\mathbf{x}_*$ of 2-dimensional linear system.

Calibration. An important criterion of probabilistic solvers is the statistical quality of their posterior distributions. A solver is 'calibrated' if its posterior distributions accurately model the uncertainty in $\mathbf{x}_*$ [8, Section 6.1]. Probabilistic Krylov solvers are not always calibrated because their posterior distributions tend to be *pessimistic.* This means, the posteriors imply that the error is larger than it actually is [2, Section 6.4], [8, Section 6.1]. Previous efforts for improving calibration have focused on scaling the posterior covariances [8, Section 4.2], [11, Section 7], [39, Section 3].

BayesCG. We analyze the calibration of BayesCG under the Krylov prior [8, 32]. BayesCG was introduced in [8] as a probabilistic numeric extension of the Conjugate Gradient (CG) method [19] for solving (1). The Krylov prior proposed in [32] makes BayesCG competitive with CG. The numerical properties of BayesCG under the Krylov prior are analysed in [32], while here we analyse its statistical properties.

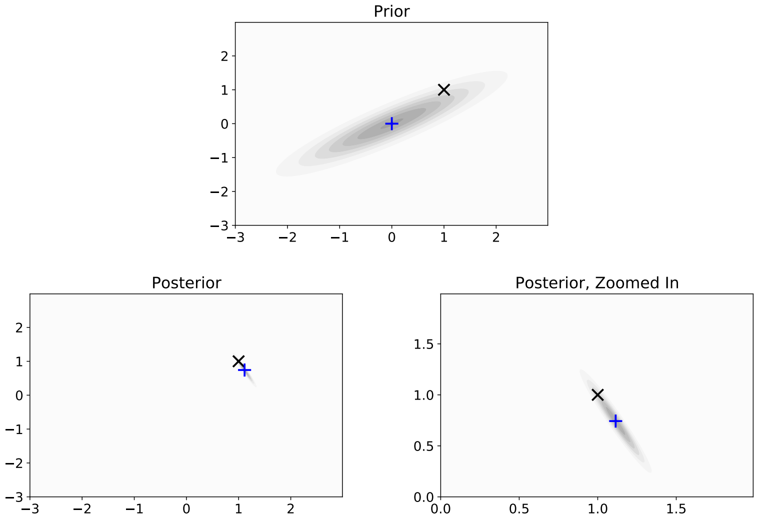

Figure 1 Prior and posterior distributions for a linear system (1) with $n = 2$. Top plot: prior distribution. Bottom plots: posterior distributions, where the bottom right is a zoomed in version of the bottom left. The gray shaded contours represent the areas in which the distributions are concentrated, the symbol '$\times$' represents the solution, and the symbol '$+$' the mean of the prior or posterior.

1.1 Contributions and Overview

Overall Conclusion. BayesCG under the Krylov prior is not calibrated in the strict sense, but has the desirable properties of a calibrated solver. Under the efficient approximate Krylov posteriors, BayesCG is only slightly optimistic and competitive with CG.

Background (Section 2). We present a short review of BayesCG, and the Krylov prior and posteriors.

Approximate Krylov posteriors (Section 3). We define the **A**-Wasserstein distance (Definition 11, Theorem 12); determine the error between Krylov posteriors and their low-rank approximations in the **A**-Wasserstein distance (Theorem 13); and present a statistical interpretation of a Krylov prior as an empirical Bayesian procedure (Theorem 15, Remark 16).

Calibration (Section 4). We review the strict notion of calibration for probabilistic solvers (Definition 17, Lemma 18), and show that it does not apply to BayesCG under the Krylov prior (Remark 22).

We relax the strict notion above and propose as an alternative form of assessment two test statistics that are necessary but not sufficient for calibration: the Z-statistic (Theorem 25) and the new S-statistic (Theorem 31, Definition 33). We present implementations for both (Algorithms 3 and 4);

and apply a Kolmogorov-Smirnov statistic (Definition 27) for evaluating the quality of samples from the Z-statistic.

The Z-statistic is inconclusive about the calibration of BayesCG under the Krylov prior (Theorem 29), while the S-statistic indicates that it is not calibrated (Section 4.3.4).

Numerical Experiments (Section 5). We create a calibrated but slowly converging version of BayesCG which has random search directions, and use it as a baseline for comparison with two BayesCG versions that both replicate CG: BayesCG under the inverse and the Krylov priors.

We assess calibration with the Z- and S-statistics for BayesCG with random search directions (Algorithms 5 and 6); BayesCG under the inverse prior (Algorithms 1 and 7); and BayesCG under the Krylov prior with full posteriors (Algorithm 8) and approximate posteriors (Algorithm 9).

Both, Z- and S statistics indicate that BayesCG with random search directions is indeed a calibrated solver, while BayesCG under the inverse prior is pessimistic.

The S-statistic indicates that BayesCG under full Krylov posteriors mimics a calibrated solver, and that BayesCG under rank-50 approximate posteriors does as well, although not as much, since it is slightly optimistic.

1.2 Notation

Matrices are represented in bold uppercase, such as $\mathbf{A}$; vectors in bold lowercase, such as $\mathbf{b}$; and scalars in lowercase, such as m.

The $m \times m$ identity matrix is $\mathbf{I}_m$, or just $\mathbf{I}$ if the dimension is clear. The Moore-Penrose inverse of a matrix $\mathbf{A}$ is $\mathbf{A}^\dagger$, and the matrix square root is $\mathbf{A}^{1/2}$ [20, Chapter 6].

Probability distributions are represented in lowercase Greek, such as μ_m; and random variables in uppercase, such as X. The random variable X having distribution μ is represented by $X \sim \mu$, and its expectation by $\mathbb{E}[X]$.

The Gaussian distribution with mean $\mathbf{x} \in \mathbb{R}^n$ and covariance $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ is denoted by $\mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$, and the chi-squared distribution with f degrees of freedom by χ_f^2.

2 Review of Existing Work

We review BayesCG (Section 2.1), the ideal Krylov prior (Section 2.2), and practical approximations for Krylov posteriors (Section 2.3). All statements in this section hold in exact arithmetic.

2.1 BayesCG

We review the computation of posterior distributions for BayesCG under general priors (Theorem 1), and present a pseudo code for BayesCG (Algorithm 1).

Given an initial guess $\mathbf{x}_0$, BayesCG [8] solves symmetric positive definite linear systems (1) by computing iterates $\mathbf{x}_m$ that converge to the solution $\mathbf{x}_*$. In addition, BayesCG computes probability distributions that quantify the uncertainty about the solution at each iteration m. Specifically, for a user-specified Gaussian prior $\mu_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$, BayesCG computes posterior distributions $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$, by conditioning a random variable $X \sim \mu_0$ on information from m search directions $\mathbf{S}_m$.

Theorem 1 ([8, Proposition 1], [32, Theorem 2.1]) *Let $\mathbf{A}\mathbf{x}_* = \mathbf{b}$ be a linear system where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric positive definite. Let $\mu_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$ be a prior with symmetric positive semi-definite covariance $\mathbf{\Sigma}_0 \in \mathbb{R}^{n\times n}$, and initial residual $\mathbf{r}_0 \equiv \mathbf{b}_0 - \mathbf{A}\mathbf{x}_0$.*

Pick $m \leq n$ so that $\mathbf{S}_m \equiv \begin{bmatrix}\mathbf{s}_1 \, \mathbf{s}_2 \cdots \mathbf{s}_m\end{bmatrix} \in \mathbb{R}^{n\times m}$ has $\mathrm{rank}(\mathbf{S}_m) = m$ and $\mathbf{\Lambda}_m \equiv \mathbf{S}_m^T\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{S}$ is non-singular. Then, the BayesCG posterior $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$ has mean and covariance

$$\mathbf{x}_m = \mathbf{x}_0 + \mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_m\mathbf{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{r}_0 \tag{2}$$

$$\mathbf{\Sigma}_m = \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_m\mathbf{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{A}\mathbf{\Sigma}_0. \tag{3}$$

Algorithm 1 represents the iterative computation of the posteriors from [8, Propositions 6 and 7], [32, Theorem 2.7]. To illustrate the resemblance of BayesCG and the Conjugate Gradient method, we present the most common implementation of CG in Algorithm 2.

BayesCG (Algorithm 1) computes specific search directions $\mathbf{S}_m$ with two additional properties:

1. They are $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$-orthogonal, which means that $\mathbf{\Lambda}_m = \mathbf{S}_m^T\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_m$ is diagonal [8, Section 2.3], thus easy to invert.
2. They form a basis for the Krylov space [9, Proposition S4]

$$\mathrm{range}(\mathbf{S}_m) = \mathcal{K}_m(\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}, \mathbf{r}_0) \equiv \mathrm{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{r}_0, \ldots, (\mathbf{A}\mathbf{\Sigma}_0\mathbf{A})^{m-1}\mathbf{r}_0\}.$$

Remark 2 The additional requirement $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\mathbf{\Sigma}_0)$ in Algorithm 1 ensures the nonsingularity of $\mathbf{\Lambda}_m$ as required by Theorem 1, even for singular prior covariance matrices $\mathbf{\Sigma}_0$ [32, Theorem 2.7].

2.2 The ideal Krylov Prior

After defining the Krylov space of maximal dimension (Definition 3), we review the ideal but impractical Krylov prior (Definition 4), and discuss its construction (Lemma 6) and properties (Theorem 7).

Algorithm 1 BayesCG [32, Algorithm 2.1]

```
Input: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, prior $\mu_0 = \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$    ▷ with $\mathbf{x}_* - \mathbf{x}_0 \in \text{range}(\mathbf{\Sigma}_0)$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$    ▷ Initial residual
$\mathbf{s}_1 = \mathbf{r}_0$    ▷ Initial search direction
$m = 0$    ▷ Initial iteration count
while not converged do
    $m = m + 1$    ▷ Increment iteration count
    $\alpha_m = \left(\mathbf{r}_{m-1}^T \mathbf{r}_{m-1}\right) / \left(\mathbf{s}_m^T \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_m\right)$
    $\mathbf{x}_m = \mathbf{x}_{m-1} + \alpha_m \mathbf{\Sigma}_0 \mathbf{A}\mathbf{s}_m$    ▷ Next posterior mean
    $\mathbf{\Sigma}_m = \mathbf{\Sigma}_{m-1} - \mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_m \left(\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_m\right)^T / (\mathbf{s}_m^T \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_m)$    ▷ Next posterior covariance
    $\mathbf{r}_m = \mathbf{r}_{m-1} - \alpha_m \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{s}_m$    ▷ Next residual
    $\beta_m = \left(\mathbf{r}_m^T \mathbf{r}_m\right) / \left(\mathbf{r}_{m-1}^T \mathbf{r}_{m-1}\right)$
    $\mathbf{s}_{m+1} = \mathbf{r}_m + \beta_m \mathbf{s}_m$    ▷ Next $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$-orthogonal search direction
end while
Output: $\mu_m = \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$    ▷ Final posterior
```

Algorithm 2 Conjugate Gradient Method (CG) [19, Section 3]

```
Input: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}_0 \in \mathbb{R}^n$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$    ▷ Initial residual
$\mathbf{w}_1 = \mathbf{r}_0$    ▷ Initial search direction
$m = 0$    ▷ Initial iteration count
while Not converged do
    $m = m + 1$    ▷ Increment iteration count
    $\gamma_m = (\mathbf{r}_{m-1}^T \mathbf{r}_{m-1}) / (\mathbf{w}_m^T \mathbf{A}\mathbf{w}_m)$    ▷ Next step size
    $\mathbf{x}_m = \mathbf{x}_{m-1} + \gamma_m \mathbf{w}_m$    ▷ Next iterate
    $\mathbf{r}_m = \mathbf{r}_{m-1} - \gamma_m \mathbf{A}\mathbf{w}_m$    ▷ Next residual
    $\delta_m = (\mathbf{r}_m^T \mathbf{r}_m) / (\mathbf{r}_{m-1}^T \mathbf{r}_{m-1})$
    $\mathbf{w}_{m+1} = \mathbf{r}_m + \delta_m \mathbf{w}_m$    ▷ Next search direction
end while
Output: $\mathbf{x}_m$    ▷ Final approximation for $\mathbf{x}_*$
```

Definition 3 The Krylov space of *maximal dimension* for Algorithm 2 is

$$\mathcal{K}_g(\mathbf{A}, \mathbf{r}_0) \equiv \text{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^{g-1}\mathbf{r}_0\}.$$

Here $g \leq n$ represents the *grade* of $\mathbf{r}_0$ with respect to $\mathbf{A} \in \mathbb{R}^{n\times n}$ [25, Definition 4.2.1], or the *invariance index* for $(\mathbf{A}, \mathbf{r}_0)$ [4, Section 2], which is the minimum value where

$$\mathcal{K}_g(\mathbf{A}, \mathbf{r}_0) = \mathcal{K}_{g+i}(\mathbf{A}, \mathbf{r}_0), \qquad i \geq 1.$$

The Krylov prior is a Gaussian distribution whose covariance is constructed from a basis for the maximal dimensional CG Krylov space.

Definition 4 ([32, Definition 3.1]) The *ideal Krylov prior* for $\mathbf{A}\mathbf{x}_* = \mathbf{b}$ is $\eta_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Gamma}_0)$ with symmetric positive semi-definite covariance

$$\mathbf{\Gamma}_0 \equiv \mathbf{V}\mathbf{\Phi}\mathbf{V}^T \in \mathbb{R}^{n\times n}. \tag{4}$$

The columns of $\mathbf{V} \equiv \begin{bmatrix}\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_g\end{bmatrix} \in \mathbb{R}^{n\times g}$ are an $\mathbf{A}$-orthonormal basis for $\mathcal{K}_g(\mathbf{A}, \mathbf{r}_0)$, which means that

$$\mathbf{V}^T\mathbf{A}\mathbf{V} = \mathbf{I}_g \quad \text{and} \quad \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_i\} = \mathcal{K}_i(\mathbf{A}, \mathbf{r}_0), \quad 1 \leq i \leq g.$$

The diagonal matrix $\mathbf{\Phi} \equiv \operatorname{diag}\begin{pmatrix}\phi_1 & \cdots & \phi_g\end{pmatrix} \in \mathbb{R}^{g\times g}$ has diagonal elements

$$\phi_i = (\mathbf{v}_i^T \mathbf{r}_0)^2, \qquad 1 \leq i \leq g. \tag{5}$$

Remark 5 The Krylov prior covariance satisfies the requirement of Algorithm 1 that $\mathbf{x}_* - \mathbf{x}_0 \in \operatorname{range}(\mathbf{\Gamma}_0)$. This follows from [25, Section 5.6],

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

If the maximal Krylov space $\mathcal{K}_g(\mathbf{A}, \mathbf{r}_0)$ has dimension $g < n$, then $\mathbf{\Gamma}_0$ is singular.

Lemma 6 ([32, Remark SM2.1]) *The Krylov prior $\mathbf{\Gamma}_0$ can be constructed from quantities computed by CG (Algorithm 2),*

$$\mathbf{v}_i \equiv \mathbf{w}_i/(\mathbf{w}_i^T \mathbf{A} \mathbf{w}_i), \quad \textit{and} \quad \phi_i \equiv \gamma_i \|\mathbf{r}_{i-1}\|_2^2, \qquad 1 \leq i \leq g.$$

The posterior distributions from BayesCG under the Krylov prior depend on submatrices of $\mathbf{V}$ and $\mathbf{\Phi}$,

$$\begin{aligned} \mathbf{V}_{i:j} &\equiv \begin{bmatrix}\mathbf{v}_i & \cdots & \mathbf{v}_j\end{bmatrix} \\ \mathbf{\Phi}_{i:j} &\equiv \operatorname{diag}\begin{pmatrix}\phi_i & \cdots & \phi_j\end{pmatrix}, \qquad 1 \leq i \leq j \leq g, \end{aligned} \tag{6}$$

where $\mathbf{V}_{1:g} = \mathbf{V}$, $\mathbf{\Phi}_{1:g} = \mathbf{\Phi}$, and $\mathbf{V}_{j+1:j} = \mathbf{\Phi}_{j+1:j} = \mathbf{0}$, $1 \leq j \leq n$.

Under suitable assumptions, BayesCG (Algorithm 1) produces the same iterates as CG (Algorithm 2).

Theorem 7 ([32, Theorem 3.3]) *Let $\mathbf{x}_0$ be the starting vector for CG (Algorithm 2). Then BayesCG (Algorithm 1) under the Krylov prior $\eta_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Gamma}_0)$ produces Krylov posteriors $\eta_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Gamma}_m)$ whose mean vectors*

$$\mathbf{x}_m = \mathbf{x}_0 + \mathbf{V}_{1:m} \mathbf{V}_{1:m}^T \mathbf{r}_0, \qquad 1 \leq m \leq g,$$

are identical to the iterates in CG (Algorithm 2), and whose covariance matrices

$$\mathbf{\Gamma}_m = \mathbf{V}_{m+1:g} \mathbf{\Phi}_{m+1:g} \mathbf{V}_{m+1:g}^T, \qquad 1 \leq m < g, \tag{7}$$

satisfy

$$\operatorname{trace}(\mathbf{A}\mathbf{\Gamma}_m) = \operatorname{trace}(\mathbf{\Phi}_{m+1:g}) = \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2. \tag{8}$$

Explicit construction of the ideal Krylov prior, followed by explicit computation of the Krylov posteriors in Algorithm 1 is impractical, because it is more expensive than solving the linear system (1) in the first place. That is the reason for introducing practical, approximate Krylov posteriors.

2.3 Practical Krylov posteriors

We dispense with the explicit computation of the Krylov prior, and instead compute a low-rank approximation of the final posterior (Definition 8) by running d additional iterations. The corresponding CG-based implementation of BayesCG under approximate Krylov posteriors is relegated to Algorithm 9 in Appendix B.

Definition 8 ([32, Definition 3.4]) Given the Krylov prior $\eta_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Gamma}_0)$ with posteriors $\eta_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Gamma}_m)$, pick some $d \geq 1$. The rank-d approximation of η_m is a Gaussian distribution $\widehat{\eta}_m \equiv \mathcal{N}(\mathbf{x}_m, \widehat{\mathbf{\Gamma}}_m)$ with the same mean $\mathbf{x}_m$ as η_m, and a rank-d covariance

$$\widehat{\mathbf{\Gamma}}_m \equiv \mathbf{V}_{m+1:m+d} \mathbf{\Phi}_{m+1:m+d} \mathbf{V}_{m+1:m+d}^T, \qquad 1 \leq m < g - d,$$

that consists of the leading d columns of $\mathbf{V}_{m+1:g}$.

In contrast to the full Krylov posteriors, which reproduce the error as in (8), approximate Krylov posteriors underestimate the error [32, Section 3.4],

$$\operatorname{trace}(\mathbf{A}\widehat{\mathbf{\Gamma}}_m) = \operatorname{trace}(\mathbf{\Phi}_{m+1:m+d}) = \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2 - \|\mathbf{x}_* - \mathbf{x}_{m+d}\|_{\mathbf{A}}^2, \qquad (9)$$

where $\|\mathbf{x}_* - \mathbf{x}_{m+d}\|_{\mathbf{A}}^2$ is the error after $m+d$ iterations of CG. The error underestimate $\operatorname{trace}(\mathbf{A}\widehat{\mathbf{\Gamma}}_m)$ is equal to [36, Equation(4.9)], and it is more accurate when convergence is fast. Fast convergence makes $\operatorname{trace}(\mathbf{A}\widehat{\mathbf{\Gamma}}_m)$ a more accurate estimate because fast convergence implies that $\|\mathbf{x}_* - \mathbf{x}_{m+d}\|_{\mathbf{A}}^2 \ll \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$, and this, along with (9), implies that $\operatorname{trace}(\mathbf{A}\widehat{\mathbf{\Gamma}}_m) \approx \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$ [36, Section 4].

3 Approximate Krylov Posteriors

We determine the error in approximate Krylov posteriors (Section 3.1), and interpret the Krylov prior as an empirical Bayesian method (Section 3.2).

3.1 Error in Approximate Krylov Posteriors

We review the p-Wasserstein distance (Definition 9), extend the 2-Wasserstein distance to the $\mathbf{A}$-Wasserstein distance weighted by a symmetric positive definite matrix $\mathbf{A}$ (Theorem 12), and derive the $\mathbf{A}$-Wasserstein distance between approximate and full Krylov posteriors (Theorem 13).

The p-Wasserstein distance is a metric on the set of probability distributions.

Definition 9 ([24, Definition 2.1], [38, Definition 6.1]) The p-Wasserstein distance between probability distributions μ and ν on $\mathbb{R}^n$ is

$$W_p(\mu, \nu) \equiv \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|M - N\|_2^p \, d\pi(M, N) \right)^{1/p}, \quad p \geq 1, \qquad (10)$$

where $\Pi(\mu, \nu)$ is the set of couplings between μ and ν, that is, the set of probability distributions on $\mathbb{R}^n \times \mathbb{R}^n$ that have μ and ν as marginal distributions.

In the special case $p = 2$, the 2-Wasserstein or *Fréchet distance* between two Gaussian distributions admits an explicit expression.

Lemma 10 ([12, Theorem 2.1]) *The 2-Wasserstein distance between Gaussian distributions $\mu \equiv \mathcal{N}(\mathbf{x}_\mu, \mathbf{\Sigma}_\mu)$ and $\nu \equiv \mathcal{N}(\mathbf{x}_\nu, \mathbf{\Sigma}_\nu)$ on $\mathbb{R}^n$ is*

$$(W_2(\mu, \nu))^2 = \|\mathbf{x}_\mu - \mathbf{x}_\nu\|_2^2 + \operatorname{trace}\left(\mathbf{\Sigma}_\mu + \mathbf{\Sigma}_\nu - 2\left(\mathbf{\Sigma}_\mu^{1/2}\mathbf{\Sigma}_\nu\mathbf{\Sigma}_\mu^{1/2}\right)^{1/2}\right).$$

We generalize the 2-Wasserstein distance to the $\mathbf{A}$-Wasserstein distance weighted by a symmetric positive definite matrix $\mathbf{A}$.

Definition 11 The two-norm of $\mathbf{x} \in \mathbb{R}^n$ weighted by a symmetric positive definite $\mathbf{A} \in \mathbb{R}^{n\times n}$ is

$$\|\mathbf{x}\|_\mathbf{A} \equiv \|\mathbf{A}^{1/2}\mathbf{x}\|_2. \tag{11}$$

The $\mathbf{A}$-Wasserstein distance between Gaussian distributions $\mu \equiv \mathcal{N}(\mathbf{x}_\mu, \mathbf{\Sigma}_\mu)$ and $\nu \equiv \mathcal{N}(\mathbf{x}_\nu, \mathbf{\Sigma}_\nu)$ on $\mathbb{R}^n$ is

$$W_\mathbf{A}(\mu, \nu) \equiv \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|M - N\|_\mathbf{A}^2 \, d\pi(M, N)\right)^{1/2}, \tag{12}$$

where $\Pi(\mu, \nu)$ is the set of couplings between μ and ν.

We derive an explicit expression for the $\mathbf{A}$-Wasserstein distance analogous to the one for the 2-Wasserstein distance in Lemma 10.

Theorem 12 *For symmetric positive definite $\mathbf{A} \in \mathbb{R}^{n\times n}$, the $\mathbf{A}$-Wasserstein distance between Gaussian distributions $\mu \equiv \mathcal{N}(\mathbf{x}_\mu, \mathbf{\Sigma}_\mu)$ and $\nu \equiv \mathcal{N}(\mathbf{x}_\nu, \mathbf{\Sigma}_\nu)$ on $\mathbb{R}^n$ is*

$$\begin{aligned}(W_\mathbf{A}(\mu, \nu))^2 = \|\mathbf{x}_\mu - \mathbf{x}_\nu\|_\mathbf{A}^2 &+ \operatorname{trace}(\widetilde{\mathbf{\Sigma}}_\mu) + \operatorname{trace}(\widetilde{\mathbf{\Sigma}}_\nu) \\ &- 2\operatorname{trace}\left((\widetilde{\mathbf{\Sigma}}_\mu^{1/2}\widetilde{\mathbf{\Sigma}}_\nu\widetilde{\mathbf{\Sigma}}_\mu^{1/2})^{1/2}\right),\end{aligned} \tag{13}$$

where $\widetilde{\mathbf{\Sigma}}_\mu \equiv \mathbf{A}^{1/2}\mathbf{\Sigma}_\mu\mathbf{A}^{1/2}$ and $\widetilde{\mathbf{\Sigma}}_\nu \equiv \mathbf{A}^{1/2}\mathbf{\Sigma}_\nu\mathbf{A}^{1/2}$.

Proof First express the $\mathbf{A}$-Wasserstein distance as a 2-Wasserstein distance, by substituting (11) into (12),

$$(W_\mathbf{A}(\mu, \nu))^2 = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{A}^{1/2}M - \mathbf{A}^{1/2}N\|_2^2 \, d\pi(M, N). \tag{14}$$

Lemma 35 in Appendix A implies that $\mathbf{A}^{1/2}M$ and $\mathbf{A}^{1/2}N$ are again Gaussian random variables with respective means and covariances

$$\tilde{\mu} \equiv \mathcal{N}(\mathbf{A}^{1/2}\mathbf{x}_\mu, \underbrace{\mathbf{A}^{1/2}\mathbf{\Sigma}_\mu\mathbf{A}^{1/2}}_{\widetilde{\mathbf{\Sigma}}_\mu}), \qquad \tilde{\nu} \equiv \mathcal{N}(\mathbf{A}^{1/2}\mathbf{x}_\nu, \underbrace{\mathbf{A}^{1/2}\mathbf{\Sigma}_\nu\mathbf{A}^{1/2}}_{\widetilde{\mathbf{\Sigma}}_\nu}).$$

Thus (14) is equal to the 2-Wasserstein distance

$$(W_{\mathbf{A}}(\mu,\nu))^2 = \inf_{\pi\in\Pi(\tilde{\mu},\tilde{\nu})} \int_{\mathbb{R}^n\times\mathbb{R}^n} \|\widetilde{M}-\widetilde{N}\|_2^2 \, d\pi(\widetilde{M},\widetilde{N}) = (W_2(\tilde{\mu},\tilde{\nu}))^2 \,. \qquad (15)$$

At last, apply Lemma 10 and the linearity of the trace. □

We are ready to derive the **A**-Wasserstein distance between approximate and full Krylov posteriors.

Theorem 13 *Let $\eta_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Gamma}_m)$ be a Krylov posterior from Theorem 7, and for some $d \geq 1$ let $\widehat{\eta}_m \equiv \mathcal{N}(\mathbf{x}_m, \widehat{\mathbf{\Gamma}}_m)$ be a rank-d approximation from Definition 8. The **A**-Wasserstein distance between η_m and $\widehat{\eta}_m$ is*

$$W_{\mathbf{A}}(\eta_m, \widehat{\eta}_m) = \left(\sum_{i=m+d+1}^{g} \phi_i \right)^{1/2}. \qquad (16)$$

Proof We factor the covariances into square factors, to obtain an eigenvalue decomposition for the congruence transformations of the covariances in (13).

Expand the column dimension of $\mathbf{V}_{m+1:g}$ from $g-m$ to n by adding an **A**-orthogonal complement $\mathbf{V}_m^{\perp} \in \mathbb{R}^{n\times(n-g+m)}$ to create an **A**-orthogonal matrix

$$\widetilde{\mathbf{V}} \equiv \begin{bmatrix}\mathbf{V}_{m+1:g} & \mathbf{V}_m^{\perp}\end{bmatrix} \in \mathbb{R}^{n\times n}$$

with $\widetilde{\mathbf{V}}^T\mathbf{A}\widetilde{\mathbf{V}} = \mathbf{I}_n$. Analogously expand the dimension of the diagonal matrices by padding with trailing zeros,

$$\begin{aligned} \widetilde{\mathbf{\Phi}}_{m+1:g} &\equiv \operatorname{diag}\begin{pmatrix}\phi_{m+1} & \cdots & \phi_g & \mathbf{0}_{1\times(n-g+m)}\end{pmatrix} \in \mathbb{R}^{n\times n}, \\ \widetilde{\mathbf{\Phi}}_{m+1:m+d} &\equiv \operatorname{diag}\begin{pmatrix}\phi_{m+1} & \cdots & \phi_{m+d} & \mathbf{0}_{1\times(n-d)}\end{pmatrix} \in \mathbb{R}^{n\times n}. \end{aligned}$$

Factor the covariances in terms of the above square matrices,

$$\mathbf{\Gamma}_m = \widetilde{\mathbf{V}}\widetilde{\mathbf{\Phi}}_{m+1:g}\widetilde{\mathbf{V}}^T \quad \text{and} \quad \widehat{\mathbf{\Gamma}}_m = \widetilde{\mathbf{V}}\widetilde{\mathbf{\Phi}}_{m+1:m+d}\widetilde{\mathbf{V}}^T.$$

Substitute the factorizations into (13), and compute the **A**-Wasserstein distance between η_m and $\widehat{\eta}_m$ as

$$(W_{\mathbf{A}}(\eta_m, \widehat{\eta}_m))^2 = \operatorname{trace}(\mathbf{G}) + \operatorname{trace}(\mathbf{J}) - 2\operatorname{trace}\left((\mathbf{G}^{1/2}\,\mathbf{J}\,\mathbf{G}^{1/2})^{1/2}\right), \qquad (17)$$

where the congruence transformations of $\mathbf{\Gamma}_m$ and $\widehat{\mathbf{\Gamma}}_m$ are again Hermitian,

$$\begin{aligned} \mathbf{G} &\equiv \mathbf{A}^{1/2}\,\underbrace{\widetilde{\mathbf{V}}\widetilde{\mathbf{\Phi}}_{m+1:g}\widetilde{\mathbf{V}}^T}_{\mathbf{\Gamma}_m}\,\mathbf{A}^{1/2} = \mathbf{U}\widetilde{\mathbf{\Phi}}_{m+1:g}\mathbf{U}^T, \qquad \mathbf{U} \equiv \mathbf{A}^{1/2}\widetilde{\mathbf{V}} \\ \mathbf{J} &\equiv \mathbf{A}^{1/2}\,\underbrace{\widetilde{\mathbf{V}}\widetilde{\mathbf{\Phi}}_{m+1:m+d}\widetilde{\mathbf{V}}^T}_{\widehat{\mathbf{\Gamma}}_m}\,\mathbf{A}^{1/2} = \mathbf{U}\widetilde{\mathbf{\Phi}}_{m+1:d}\mathbf{U}^T. \end{aligned}$$

Lemma 37 implies that $\mathbf{U}$ is an orthogonal matrix, so that the second factorizations of $\mathbf{G}$ and $\mathbf{J}$ represent eigenvalue decompositions. Commutativity of the trace implies

$$\begin{aligned}\operatorname{trace}(\mathbf{G}) &= \operatorname{trace}(\widetilde{\mathbf{\Phi}}_{m+1:g}) = \sum_{i=m+1}^{g} \phi_i\\ \operatorname{trace}(\mathbf{J}) &= \operatorname{trace}(\widetilde{\mathbf{\Phi}}_{m+1:m+d}) = \sum_{i=m+1}^{m+d} \phi_i.\end{aligned}$$

Since $\mathbf{G}$ and $\mathbf{J}$ have the same eigenvector matrix, they commute, and so do diagonal matrices,

$$\begin{aligned}\mathbf{G}^{1/2}\mathbf{J}\mathbf{G}^{1/2} &= \mathbf{U}\widetilde{\mathbf{\Phi}}_{m+1:g}\widetilde{\mathbf{\Phi}}_{m+1:m+d}\mathbf{U}^T\\ &= \mathbf{U}\operatorname{diag}\left(\phi_{m+1}^2 \cdots \phi_{m+d}^2 \ \mathbf{0}_{1\times(n-d)}\right)\mathbf{U}^T\end{aligned}$$

where the last equality follows from the fact that $\widetilde{\mathbf{\Phi}}_{m+1:g}$ and $\widetilde{\mathbf{\Phi}}_{m+1:m+d}$ share the leading d diagonal elements. Thus

$$\operatorname{trace}\left((\mathbf{G}^{1/2}\,\mathbf{J}\,\mathbf{G}^{1/2})^{1/2}\right) = \sum_{i=m+1}^{m+d} \phi_i.$$

Substituting the above expressions into (17) gives

$$\left(W_{\mathbf{A}}(\eta_m, \widehat{\eta}_m)\right)^2 = \sum_{i=m+1}^{g} \phi_i + \sum_{i=m+1}^{m+d} \phi_i - 2\sum_{i=m+1}^{m+d} \phi_i = \sum_{i=m+d+1}^{g} \phi_i.$$

□

Theorem 13 implies that the $\mathbf{A}$-Wasserstein distance between approximate and full Krylov posteriors is the sum of the CG steps sizes skipped by the approximate posterior, and this, as seen in (9) and [36, Equation (4.4)], is equal to the distance between the error estimate $\operatorname{trace}(\mathbf{A}\widehat{\mathbf{\Gamma}}_m)$ and the true error $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$. As a consequence, the approximation error decreases as the convergence of the posterior mean accelerates, or the rank d of the approximation increases.

Remark 14 The distance in Theorem 13 is a special case of the 2-Wasserstein distance between two distributions whose covariance matrices commute [24, Corollary 2.4].

To see this, consider the $\mathbf{A}$-Wasserstein distance between η_m and $\widehat{\eta}_m$ from Theorem 13, and the 2-Wasserstein distance between $\nu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{A}^{1/2}\mathbf{\Gamma}\mathbf{A}^{1/2})$ and $\widehat{\nu}_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{A}^{1/2}\widehat{\mathbf{\Gamma}}\mathbf{A}^{1/2})$. Then (15) implies that the $\mathbf{A}$-Wasserstein distance is equal to the 2-Wassterstein distance of a congruence transformation,

$$W_{\mathbf{A}}(\eta_m, \widehat{\eta}_m) = W_2(\nu_m, \widehat{\nu}_m).$$

The covariance matrices $\mathbf{A}^{1/2}\mathbf{\Gamma}_m\mathbf{A}^{1/2}$ and $\mathbf{A}^{1/2}\widehat{\mathbf{\Gamma}}_m\mathbf{A}^{1/2}$ associated with the 2-Wasserstein distance commute because they are both diagonalized by the same orthogonal matrix $\mathbf{A}^{1/2}\widetilde{\mathbf{V}}$.

3.2 Probabilistic Interpretation of the Krylov Prior

We interpret the Krylov prior as an 'empirical Bayesian procedure' (Theorem 15), and elucidate the connection between the random variables and the deterministic solution (Remark 16).

An *empirical Bayesian procedure* estimates the prior from data [3, Section 4.5]. Our 'data' are the pairs of normalized search directions $\mathbf{v}_i$ and step sizes ϕ_i, $1 \leq i \leq m+d$, from $m+d$ iterations of CG. In contrast, the usual data for BayesCG are the inner products $\mathbf{v}_i^T\mathbf{b}$, $1 \leq i \leq m$. However, if we augment the usual data with the search directions, which is natural due to their dependence on $\mathbf{x}_*$, then ϕ_i is just a function of the data.

From these data we construct a prior in an empirical Bayesian fashion, starting with a random variable

$$X = \mathbf{x}_0 + \sum_{i=1}^{m+d} \sqrt{\phi_i}\mathbf{v}_i Q_i \in \mathbb{R}^n,$$

where $Q_i \sim \mathcal{N}(0,1)$ are independent and identically distributed scalar Gaussian random variables, $1 \leq i \leq m+d$. Due to the independence of the Q_i, the above sum is the matrix vector product

$$X = \mathbf{x}_0 + \mathbf{V}_{1:m+d}\,\mathbf{\Phi}_{1:m+d}^{1/2}\,Q \tag{18}$$

where $Q \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{m+d})$ is a vector-valued Gaussian random variable.

The distribution of X is the *empirical prior*, while the distribution of X conditioned on the random variable $Y \equiv \mathbf{V}_{1:m}^T\mathbf{A}X$ taking the value $\mathbf{V}_{1:m}^T\mathbf{b}$ is the *empirical posterior*. We relate these distributions to the Krylov prior.

Theorem 15 *Under the assumptions of Theorem 7, the random variable X in (18) is distributed according to the empirical prior*

$$\mathcal{N}\left(\mathbf{x}_0, \mathbf{V}_{1:m+d}\mathbf{\Phi}_{1:m+d}\mathbf{V}_{1:m+d}^T\right),$$

which is the rank-$(m+d)$ approximation of the Krylov prior $\mathbf{\Gamma}_0$. The variable X conditioned on $Y \equiv \mathbf{V}_{1:m}^T\mathbf{A}X$ taking the value $\mathbf{V}_{1:m}^T\mathbf{b}$ is distributed according to the empirical posterior

$$\mathcal{N}\left(\mathbf{x}_m, \mathbf{V}_{m+1:m+d}\mathbf{\Phi}_{m+1:m+d}\mathbf{V}_{m+1:m+d}^T\right) = \mathcal{N}\left(\mathbf{x}_m, \widehat{\mathbf{\Gamma}}_m\right),$$

which, in turn, is the rank-d approximation of the Krylov posterior.

Proof As in the proof of Theorem 1 in [9, Proof of Proposition 1], we exploit the stability and conjugacy of Gaussian distributions in Lemmas 35 and 36 in Appendix A.

Prior. Lemma 35 implies that X in (18) is a Gaussian random variable with mean and covariance

$$X \sim \mathcal{N}\left(\mathbf{x}_0, \mathbf{V}_{1:m+d}\mathbf{\Phi}_{1:m+d}\mathbf{V}_{1:m+d}^T\right). \tag{19}$$

Thus, the approximate Krylov prior is an empirical Bayesian prior.

Posterior. From (19) follows that X and $Y \equiv \mathbf{V}_{1:m}^T \mathbf{A} X$ have the joint distribution

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_0 \\ \mathbb{E}[Y] \end{bmatrix}, \begin{bmatrix} \mathbf{V}_{1:m+d}\mathbf{\Phi}_{1:m+d}\mathbf{V}_{1:m+d}^T & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y)^T & \mathrm{Cov}(Y,Y) \end{bmatrix} \right) \tag{20}$$

and that $\mathbb{E}[Y] = \mathbf{V}_{1:m}^T \mathbf{A}\mathbf{x}_0$. This, together with the linearity of the expectation and the $\mathbf{A}$-orthonormality of $\mathbf{V}$ implies

$$\begin{aligned} \mathrm{Cov}(Y,Y) &= \mathbb{E}\left[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^T\right] \\ &= \mathbf{V}_{1:m}^T \mathbf{A}\ \mathbb{E}\left[(X - \mathbf{x}_0)(X - \mathbf{x}_0)^T\right] \mathbf{A}\mathbf{V}_{1:m} \\ &= \mathbf{V}_{1:m}^T \mathbf{A}\ \left(\mathbf{V}_{1:m+d}\mathbf{\Phi}_{1:m+d}\mathbf{V}_{1:m+d}^T\right)\ \mathbf{A}\mathbf{V}_{1:m} \\ &= \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \end{bmatrix} \mathbf{\Phi}_{1:m+d} \begin{bmatrix} \mathbf{I}_m \\ \mathbf{0} \end{bmatrix} = \mathbf{\Phi}_{1:m}. \end{aligned}$$

Analogously,

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathbb{E}[(X - \mathbf{x}_0)(Y - \mathbb{E}[Y])^T] = \mathbb{E}[(X - \mathbf{x}_0)(Y - \mathbf{V}_{1:m}^T \mathbf{A}\mathbf{x}_0)^T] \\ &= \mathbb{E}[(X - \mathbf{x}_0)(X - \mathbf{x}_0)^T]\mathbf{A}\mathbf{V}_{1:m} = \mathbf{V}_{1:m+d}\mathbf{\Phi}_{1:m+d}\mathbf{V}_{1:m+d}^T \mathbf{A}\mathbf{V}_{1:m} \\ &= \mathbf{V}_{1:m+d}\mathbf{\Phi}_{1:m+d} \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \end{bmatrix} = \mathbf{V}_{1:m}\mathbf{\Phi}_{1:m}. \end{aligned}$$

From [37, Theorem 6.20] follows the expression for the posterior mean,

$$\begin{aligned} \mathbf{x}_m &= \mathbf{x}_0 + \mathrm{Cov}(X,Y)\,\mathrm{Cov}(Y,Y)^{-1}\left(\mathbf{V}_{1:m}^T \mathbf{b} - \mathbf{V}_{1:m}^T \mathbf{A}\mathbf{x}_0\right) \\ &= \mathbf{x}_0 + \mathbf{V}_{1:m}\mathbf{\Phi}_{1:m}\mathbf{\Phi}_{1:m}^{-1}\mathbf{V}_{1:m}^T \mathbf{r}_0 = \mathbf{x}_0 + \mathbf{V}_{1:m}\mathbf{V}_{1:m}^T \mathbf{r}_0, \end{aligned}$$

and for the posterior covariance

$$\widehat{\mathbf{\Gamma}}_m = \mathbf{V}_{1:m+d}\mathbf{\Phi}_{1:m+d}\mathbf{V}_{1:m+d}^T - \mathrm{Cov}(X,Y)\,\mathrm{Cov}(Y,Y)^{-1}\,\mathrm{Cov}(X,Y)^T,$$

where

$$\begin{aligned} \mathrm{Cov}(X,Y)\,\mathrm{Cov}(Y,Y)^{-1}\,\mathrm{Cov}(X,Y)^T &= \mathbf{V}_{1:m}\mathbf{\Phi}_{1:m}\mathbf{\Phi}_{1:m}^{-1}\mathbf{\Phi}_{1:m}\mathbf{V}_{1:m}^T \\ &= \mathbf{V}_{1:m}\mathbf{\Phi}_{1:m}\mathbf{V}_{1:m}^T. \end{aligned}$$

Substituting this into $\widehat{\mathbf{\Gamma}}_m$ gives the expression for the posterior covariance

$$\begin{aligned} \widehat{\mathbf{\Gamma}}_m &= \mathbf{V}_{1:m+d}\mathbf{\Phi}_{1:m+d}\mathbf{V}_{1:m+d}^T - \mathbf{V}_{1:m}\mathbf{\Phi}_{1:m}\mathbf{V}_{1:m}^T \\ &= \mathbf{V}_{m+1:m+d}\mathbf{\Phi}_{m+1:m+d}\mathbf{V}_{m+1:m+d}^T. \end{aligned}$$

Thus, the posterior mean $\mathbf{x}_m$ is equal to the one in Theorem 7, and the posterior covariance $\widehat{\mathbf{\Gamma}}_m$ is equal to the rank-d approximate Krylov posterior in Definition 8. □

Remark 16 The random variable X in Theorem 15 is a surrogate for the unknown solution $\mathbf{x}_*$. The solution $\mathbf{x}_*$ is a deterministic quantity, but prior to solving the linear system (1), we are uncertain of $\mathbf{x}_*$, and the prior models this uncertainty.

During the course of the BayesCG iterations, we acquire information about $\mathbf{x}_*$, and the posterior distributions μ_m, $1 \le m \le n$ incorporate our increasing knowledge and, consequently, our diminishing uncertainty.

4 Calibration of BayesCG Under the Krylov Prior

We review the notion of calibration for probabilistic solvers, and show that this notion does not apply to BayesCG under the Krylov prior (Section 4.1). Then we relax this notion and analyze BayesCG with two test statistics that are necessary but not sufficient for calibration: the Z-statistic (Section 4.2) and the S-statistic (Section 4.3).

4.1 Calibration

We review the definition of calibration for probabilistic linear solvers (Definition 17, Lemma 18), discuss the difference between certain random variables (Remark 19), present two illustrations (Examples 20 and 21), and explain why this notion of calibration does not apply to BayesCG under the Krylov prior (Remark 22).

Informally, a probabilistic numerical solver is calibrated if its posterior distributions accurately model the uncertainty in the solution [6,7].

Definition 17 ([7, Definition 6]) Let $\mathbf{A}X_* = B$ be a class of linear systems where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric positive definite, and the random right hand sides $B \in \mathbb{R}^n$ are defined by random solutions $X_* \sim \mu_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$.

Assume that a probabilistic linear solver under the prior μ_0 and applied to a system $\mathbf{A}X_* = B$ computes posteriors $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$, $1 \le m \le n$. Let $\operatorname{rank}(\mathbf{\Sigma}_m) = p_m$, and let $\mathbf{\Sigma}_m$ have an orthogonal eigenvector matrix $\mathbf{U} = \begin{bmatrix}\mathbf{U}_m & \mathbf{U}_m^\perp\end{bmatrix} \in \mathbb{R}^{n\times n}$ where $\mathbf{U}_m \in \mathbb{R}^{n\times p_m}$ and $\mathbf{U}_m^\perp \in \mathbb{R}^{n\times(n-p_m)}$ satisfy

$$\operatorname{range}(\mathbf{U}_m) = \operatorname{range}(\mathbf{\Sigma}_m), \qquad \operatorname{range}(\mathbf{U}_m^\perp) = \ker(\mathbf{\Sigma}_m).$$

The probabilistic solver is *calibrated* if all posterior covariances $\mathbf{\Sigma}_m$ are independent of B and satisfy

$$\begin{aligned}(\mathbf{U}_m^T\mathbf{\Sigma}_m\mathbf{U}_m)^{-1/2}\mathbf{U}_m^T(X_* - \mathbf{x}_m) &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p_m}),\\ (\mathbf{U}_m^\perp)^T(X_* - \mathbf{x}_m) &= \mathbf{0}, \qquad 1 \le m \le n.\end{aligned} \tag{21}$$

Alternatively, one can think of a probabilistic linear solver as calibrated if and only if the solutions X_* are distributed according to the posteriors.

Lemma 18 *Under the conditions of Definition 17, a probabilistic linear solver is calibrated, if and only if*

$$X_* \sim \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m), \qquad 1 \le m \le n.$$

Proof Let $\mathbf{\Sigma}_m = \mathbf{U}\mathbf{D}\mathbf{U}^T$ be an eigendecomposition where the eigenvalue matrix $\mathbf{D} = \operatorname{diag}\begin{pmatrix}\mathbf{D}_m & \mathbf{0}\end{pmatrix}$ is commensurately partitioned with $\mathbf{U}$ in Definition 17. Multiply the first equation of (21) on the left by $\mathbf{D}_m = \mathbf{U}_m^T\mathbf{\Sigma}_m\mathbf{U}_m$,

$$\mathbf{U}_m^T(X_* - \mathbf{x}_m) \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_m),$$

combine the result with the second equation in (21),

$$\mathbf{U}^T(X_* - \mathbf{x}_m) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{D}\right).$$

and multiply by $\mathbf{U}$ on the left,

$$(X_* - \mathbf{x}_m) \sim \mathcal{N}(\mathbf{0}, \mathbf{U}\mathbf{D}\mathbf{U}^T), \qquad 1 \leq m \leq n.$$

At last, substitute $\mathbf{\Sigma}_m = \mathbf{U}\mathbf{D}\mathbf{U}^T$ and subtract $\mathbf{x}_m$. □

Since the covariance matrix $\mathbf{\Sigma}_m$ is singular, its probability density function is zero on the subspace of $\mathbb{R}^n$ where the solver has eliminated the uncertainty about X_*. From (21) follows that $X_* = \mathbf{x}_m$ in $\ker(\mathbf{\Sigma}_m)$. Hence, this subspace must be $\ker(\mathbf{\Sigma}_m)$, and any remaining uncertainty about X_* lies in $\mathrm{range}(\mathbf{\Sigma}_m)$.

Remark 19 We discuss the difference between the random variable X_* in Definition 17 and the random variable X in Theorem 15.

In the context of calibration, the random variable $X_* \sim \mu_0$ represents the set of *all possible* solutions that are accurately modeled by the prior μ_0. If the solver is calibrated, then Lemma 18 shows that $X_* \sim \mu_m$. Thus, solutions accurately modeled by the prior μ_0 are also accurately modeled by all posteriors μ_m.

By contrast, in the context of a deterministic linear system $\mathbf{A}\mathbf{x}_* = \mathbf{b}$, the random variable X represents a surrogate for the *particular* solution $\mathbf{x}_*$ and can be viewed as an abbreviation for $X \mid X_* = \mathbf{x}_*$. The prior μ_0 models the uncertainty in the user's initial knowledge of $\mathbf{x}_*$, and the posteriors μ_m model the uncertainty remaining after m iterations of the solver.

The following two examples illustrate Definition 17.

Example 20 Suppose there are three people: Alice, Bob, and Carol.

1. Alice samples $\mathbf{x}_*$ from the prior μ_0 and computes the matrix vector product $\mathbf{b} = \mathbf{A}\mathbf{x}_*$.
2. Bob receives μ_0, $\mathbf{b}$, and $\mathbf{A}$ from Alice. He estimates $\mathbf{x}_*$ by solving the linear system with a probabilistic solver under the prior μ_0, and then samples $\mathbf{y}$ from a posterior μ_m.
3. Carol receives μ_m, $\mathbf{x}_*$ and $\mathbf{y}$, but she is not told which vector is $\mathbf{x}_*$ and which is $\mathbf{y}$. Carol then attempts to determine which one of $\mathbf{x}_*$ or $\mathbf{y}$ is the sample from μ_m. If Carol cannot distinguish between $\mathbf{x}_*$ and $\mathbf{y}$, then the solver is calibrated.

Example 21 This is the visual equivalent of Example 20, where Carol receives the images in Figure 2 of three different probabilistic solvers, but without any identification of the solutions and posterior samples.

- Top plot. This solver is calibrated because the solutions look indistinguishable from the samples of the posterior distribution.

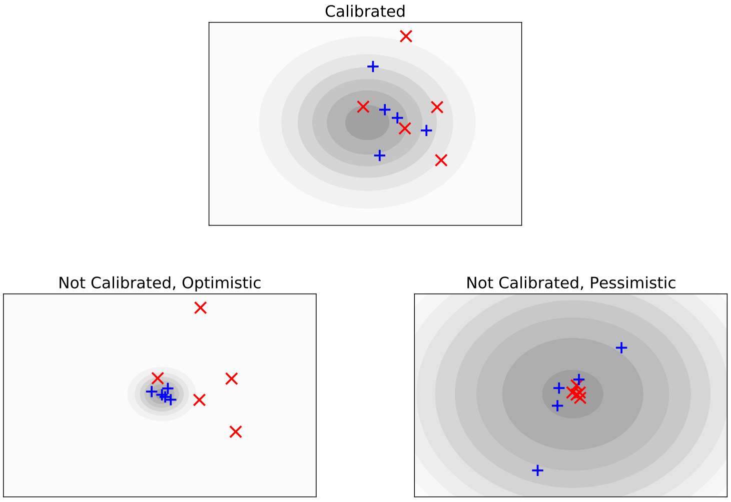

Figure 2 Posterior distributions and solutions from three different probabilistic solvers: calibrated (top), optimistic (bottom left), and pessimistic (bottom right). The gray contours represent the posterior distributions, the red symbols "×" the solutions, and the blue symbols "+" samples from the posterior distributions.

- Bottom left plot. This solver is not calibrated because the solutions are unlikely to be samples from the posterior distribution.
  The solver is *optimistic* because the posterior distribution is concentrated in an area of $\mathbb{R}^n$ that is too small to cover the solutions.
- Bottom right plot. The solver is not calibrated. Although the solutions could plausibly be sampled from the posterior, they are concentrated too close to the center of the distribution.
  The solver is *pessimistic* because the area covered by the posterior distribution is much larger than the area containing the solutions.

Remark 22 The posterior means and covariances from a probabilistic solver can depend on the solution $\mathbf{x}_*$, as is the case for BayesCG. If a solver is applied to a random linear system in Definition 17 and if the posterior means and covariances depend on the solution X_*, then the posterior means and covariances are also random variables.

Definition 17 prevents the posterior covariances from being random variables by forcing them to be independent of the random right hand side B. Although this is a realistic constraint for the stationary iterative solvers in [32], it does not apply to BayesCG under the Krylov prior, because Krylov posterior covariances depend non-linearly on the right-hand side. In Sections 4.2 and 4.3, we present a remedy for BayesCG in the form of test statistics that are motivated by Definition 17 and Lemma 18.

4.2 The Z-statistic

We assess BayesCG under the Krylov prior with an existing test statistic, the Z-statistic, which is a necessary condition for calibration and can be viewed as a weaker normwise version of criterion (21). We review the Z-statistic (Section 4.2.1), and apply it to BayesCG under the Krylov prior (Section 4.2.2).

4.2.1 Review of the Z-statistic

We review the Z-statistic (Definition 23), and the chi-square distribution (Definition 24), which links the Z-statistic to calibration (Theorem 25). Then we discuss how to generate samples of the Z-statistic (Algorithm 3), how to use them for the assessment of calibration (Remark 26), and then present the Kolmogorov-Smirnov statistic as a computationally inexpensive estimate for the difference between two general distributions (Definition 27).

The Z-statistic was introduced in [8, Section 6.1] as a means to assess the calibration of BayesCG, and has subsequently been applied to other probabilistic linear solvers [2, Section 6.4], [11, Section 9].

Definition 23 ([8, Section 6.1]) Let $\mathbf{A}X_* = B$ be a class of linear systems where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric positive definite, and $X_* \sim \mu_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$. Let $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$, $1 \leq m \leq n$, be the posterior distributions from a probabilistic solver under the prior μ_0 applied to $\mathbf{A}X_* = B$. The Z-statistic is

$$\mathrm{Z}_m(X_*) \equiv (X_* - \mathbf{x}_m)^T \mathbf{\Sigma}_m^\dagger (X_* - \mathbf{x}_m), \qquad 1 \leq m \leq n. \tag{22}$$

The chi-squared distribution below furnishes the link from Z-statistic to calibration.

Definition 24 ([33, Definition 2.2]) If $X_1, \ldots, X_f \in \mathcal{N}(0,1)$ are independent random normal variables, then $\sum_{j=1}^{f} X_j^2$ is distributed according to the chi-squared distribution χ_f^2 with f degrees of freedom and mean f.
In other words, if $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_f)$, then $X^T X \sim \chi_f^2$ and $\mathbb{E}[X^T X] = f$.

We show that the Z-statistic is a necessary condition for calibration. That is: If a probabilistic solver is calibrated, then the Z-statistic is distributed according to a chi-squared distribution.

Theorem 25 ([9, Proposition 1]) *Let $\mathbf{A}X_* = B$ be a class of linear systems where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric positive definite, and $X_* \sim \mu_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$. Assume that a probabilistic solver under the prior μ_0 applied to $\mathbf{A}X_* = B$ computes the posteriors $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$ with $\operatorname{rank}(\mathbf{\Sigma}_m) = p_m$, $1 \leq m \leq n$.*

If the solver is calibrated, then

$$\mathrm{Z}_m(X_*) \sim \chi_{p_m}^2, \qquad 1 \leq m \leq n.$$

Proof Write $\mathrm{Z}_m(X_*) = M_m^T M_m$, where $M_m \equiv (\mathbf{\Sigma}_m^\dagger)^{1/2}(X_* - \mathbf{x}_m)$. Lemma 18 implies that a calibrated solver produces posteriors with

$$(X_* - \mathbf{x}_m) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_m), \qquad 1 \leq m \leq n.$$

With the eigenvector matrix $\mathbf{U}_m \in \mathbb{R}^{n\times p_m}$ as in Definition 17, Lemma 35 in Appendix A implies

$$M_m \sim \mathcal{N}(\mathbf{0}, \mathbf{U}_m\mathbf{U}_m^T), \qquad 1 \leq m \leq n.$$

Since the covariance of M_m is an orthogonal projector, Lemma 41 implies $\mathrm{Z}_m(X_*) = (M_m^T M_m) \sim \chi^2_{p_m}$. □

Theorem 25 implies that BayesCG is calibrated if the Z-statistic is distributed according to a chi-squared distribution with $p_m = \mathrm{rank}(\mathbf{\Sigma}_0) - m$ degrees of freedom. For the Krylov prior specifically, $p_m = g - m$.

Generating samples from the Z-statistic and assessing calibration. For a user-specified probabilistic linear `solver` and a symmetric positive definite matrix $\mathbf{A}$, Algorithm 3 samples N_{test} solutions $\mathbf{x}_*$ from the prior distribution μ_0, defines the systems $\mathbf{b} \equiv \mathbf{A}\mathbf{x}_*$, runs m iterations of the `solver` on $\mathbf{b} \equiv \mathbf{A}\mathbf{x}_*$, and computes $\mathrm{Z}_m(\mathbf{x}_*)$ in (22).

The application of the Moore-Penrose inverse in Line 6 can be implemented by computing the minimal norm solution $\mathbf{q} = \mathbf{\Sigma}_m^\dagger(\mathbf{x}_* - \mathbf{x}_m)$ of the least squares problem

$$\min_{\mathbf{u}\in\mathbb{R}^n} \|(\mathbf{x}_* - \mathbf{x}_m) - \mathbf{\Sigma}_m\mathbf{u}\|_2, \tag{23}$$

followed by the inner product $z_i = (\mathbf{x}_* - \mathbf{x}_m)^T\mathbf{q}$.

Algorithm 3 Sampling from the Z-statistic

```
1: Input: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mu_0 = \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$, solver, $m$, $N_{\text{test}}$
2: for $i = 1 : N_{\text{test}}$ do
3:     Sample $\mathbf{x}_*$ from prior distribution $\mu_0$          ▷ Sample solution vector
4:     $\mathbf{b} = \mathbf{A}\mathbf{x}_*$          ▷ Define test problem
5:     $[\mathbf{x}_m, \mathbf{\Sigma}_m] =$ solver$(\mathbf{A}, \mathbf{b}, \mu_0, m)$          ▷ Compute posterior $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$
6:     $z_i = (\mathbf{x}_* - \mathbf{x}_m)^T\mathbf{\Sigma}_m^\dagger(\mathbf{x}_* - \mathbf{x}_m)$          ▷ Compute $Z$-statistic sample
7: end for
8: Output: $Z$-statistic samples $z_i$, $1 \leq i \leq N_{test}$.
```

Remark 26 We assess calibration of the `solver` by comparing the Z-statistic samples z_i from Algorithm 3 to the chi-squared distribution $\chi^2_{p_m}$ with $p_m \equiv \mathrm{rank}(\mathbf{\Sigma}_0) - m$ degrees of freedom, based on the following criteria from [8, Section 6.1].

Calibrated: If $z_i \sim \chi^2_{p_m}$, then $\mathbf{x}_* \sim \mu_m$ and the solutions $\mathbf{x}_*$ are distributed according to the posteriors μ_m.

Pessimistic: If the z_i are concentrated around smaller values than $\chi^2_{p_m}$, then the solutions $\mathbf{x}_*$ occupy a smaller area of $\mathbb{R}^n$ than predicted by μ_m.

Optimistic: If the z_i are concentrated around larger values than $\chi^2_{p_m}$, then the solutions cover a larger area of $\mathbb{R}^n$ than predicted by μ_m.

In [8, Section 6.1] and [2, Section 6.4], the Z-statistic samples and the predicted chi-squared distribution are compared visually. In Section 5, we make an additional quantitative comparison with the Kolmogorov-Smirnov test to estimate the difference between two probability distributions.

Definition 27 ([23, Section 3.4.1]) Given two distributions μ and ν on $\mathbb{R}^n$ with cumulative distribution functions F_μ and F_ν, the Kolmogorov-Smirnov statistic is

$$KS(\mu,\nu) = \sup_{x\in\mathbb{R}} |F_\mu(x) - F_\nu(x)|,$$

where $0 \leq KS(\mu,\nu) \leq 1$.

If $KS(\mu,\nu) = 0$, then μ and ν have the same cumulative distribution functions, $F_\mu = F_\nu$. If $KS(\mu,\nu) = 1$, then μ and ν do not overlap. In general, the lower $KS(\mu,\nu)$, the closer μ and ν are to each other.

In contrast to the Wasserstein distance in Definition 9, the Kolmogorov-Smirnov statistic can be easier to estimate—especially if the distributions are not Gaussian—but it is not a metric. Consequently, if μ and ν do not overlap, then $KS(\mu,\nu) = 1$ regardless of how far μ and ν are apart, while the Wasserstein metric still gives information about the distance between μ and ν.

4.2.2 Z-Statistic for BayesCG under the Krylov prior

We apply the Z-statistic to BayesCG under the Krylov prior. We start with an expression for the Moore-Penrose inverse of the Krylov posterior covariances (Lemma 28). Then we show that the Z-statistic for the full Krylov posteriors has the same *mean* as the corresponding chi-squared distribution (Theorem 29), but its *distribution* is different. Therefore the Z-statistic is inconclusive about the calibration of BayesCG under the Krylov prior (Remark 30).

Lemma 28 *In Definition 8, abbreviate* $\widehat{\mathbf{V}} \equiv \mathbf{V}_{m+1:m+d}$ *and* $\widehat{\mathbf{\Phi}} \equiv \mathbf{\Phi}_{m+1:m+d}$. *The rank-d approximate Krylov posterior covariances have the Moore-Penrose inverse*

$$\widehat{\mathbf{\Gamma}}_m^\dagger = \left(\widehat{\mathbf{V}}\widehat{\mathbf{\Phi}}\widehat{\mathbf{V}}^T\right)^\dagger = \widehat{\mathbf{V}}(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})^{-1}\widehat{\mathbf{\Phi}}^{-1}(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})^{-1}\widehat{\mathbf{V}}^T, \qquad 1 \leq m \leq g-d.$$

Proof We exploit the fact that all factors of $\widehat{\mathbf{\Gamma}}_m$ have full column rank.

The factors $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{V}}^T$ have full column and row rank, respectively, because $\mathbf{V}$ has $\mathbf{A}$-orthonormal columns. Additionally, the diagonal matrix $\widehat{\mathbf{\Phi}}$ is nonsingular. Then Lemma 39 in Appendix A implies that the Moore-Penrose inverses can be expressed in terms of the matrices proper,

$$\widehat{\mathbf{V}}^\dagger = (\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})^{-1}\widehat{\mathbf{V}}^T, \qquad (\widehat{\mathbf{V}}^T)^\dagger = \widehat{\mathbf{V}}(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})^{-1}, \tag{24}$$

and

$$(\widehat{\mathbf{\Phi}}\widehat{\mathbf{V}}^T)^\dagger = (\widehat{\mathbf{V}}^T)^\dagger\widehat{\mathbf{\Phi}}^{-1} = \widehat{\mathbf{V}}(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})^{-1}\widehat{\mathbf{\Phi}}^{-1}. \tag{25}$$

Since $\widehat{\mathbf{\Phi}}\widehat{\mathbf{V}}^T$ also has full row rank, apply Lemma 39 to $\widehat{\mathbf{\Gamma}}_m$,

$$\widehat{\mathbf{\Gamma}}_m^\dagger = (\widehat{\mathbf{\Phi}}\widehat{\mathbf{V}}^T)^\dagger \widehat{\mathbf{V}}^\dagger,$$

and substitute (24) and (25) into the above expression. □

We apply the Z-statistic to the full Krylov posteriors, and show that Z-statistic samples reproduce the dimension of the unexplored Krylov space.

Theorem 29 *Under the assumptions of Theorem 7, let BayesCG under the Krylov prior $\eta_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Gamma}_0)$ produce full Krylov posteriors $\eta_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Gamma}_m)$. Then the Z-statistic is equal to*

$$\mathrm{Z}_m(\mathbf{x}_*) = (\mathbf{x}_* - \mathbf{x}_m)^T \mathbf{\Gamma}_m^\dagger (\mathbf{x}_* - \mathbf{x}_m) = g - m, \qquad 1 \leq m \leq g.$$

Proof Express the error $\mathbf{x}_0 - \mathbf{x}_m$ in terms of $\widehat{\mathbf{V}} \equiv \mathbf{V}_{m+1:m+d}$ by inserting

$$\mathbf{x}_* = \mathbf{x}_0 + \mathbf{V}_{1:g}\mathbf{V}_{1:g}^T \mathbf{r}_0, \qquad \mathbf{x}_m = \mathbf{x}_0 + \mathbf{V}_{1:m}\mathbf{V}_{1:m}^T \mathbf{r}_0, \quad 1 \leq m \leq g, \qquad (26)$$

from Theorem 7 into

$$\mathbf{x}_* - \mathbf{x}_m = \mathbf{V}_{m+1:g}\mathbf{V}_{m+1:g}^T \mathbf{r}_0 = \widehat{\mathbf{V}}\widehat{\mathbf{V}}^T \mathbf{r}_0.$$

This expression is identical to [25, Equation (5.6.5)], which relates the CG error to the search directions and step sizes of the remaining iterations.

With Lemma 28, this implies for the Z-statistic in Theorem 25

$$\begin{aligned}
\mathrm{Z}_m(\mathbf{x}_*) &= (\mathbf{x}_* - \mathbf{x}_m)^T \mathbf{\Gamma}_m^\dagger \mathbf{x}_* - \mathbf{x}_m) \\
&= \underbrace{\mathbf{r}_0^T \widehat{\mathbf{V}}\widehat{\mathbf{V}}^T}_{(\mathbf{x}_* - \mathbf{x}_m)^T} \ \underbrace{\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})^{-1}\widehat{\mathbf{\Phi}}^{-1}(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})^{-1}\widehat{\mathbf{V}}^T}_{\mathbf{\Gamma}_m^\dagger} \ \underbrace{\widehat{\mathbf{V}}\widehat{\mathbf{V}}^T \mathbf{r}_0}_{(\mathbf{x}_* - \mathbf{x}_m)} \\
&= \mathbf{r}_0^T \widehat{\mathbf{V}} \ \underbrace{(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})^{-1}}_{\mathbf{I}} \ \widehat{\mathbf{\Phi}}^{-1} \ \underbrace{(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})^{-1}(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}})}_{\mathbf{I}} \ \widehat{\mathbf{V}}^T \mathbf{r}_0 \\
&= \mathbf{r}_0^T \widehat{\mathbf{V}} \widehat{\mathbf{\Phi}}^{-1} \widehat{\mathbf{V}}^T \mathbf{r}_0.
\end{aligned}$$

In other words,

$$\begin{aligned}
\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{\Gamma}_m^\dagger}^2 &= \left(\mathbf{V}_{m+1:g}^T \mathbf{r}_0\right)^T \mathbf{\Phi}_{m+1:m+d}^{-1} \left(\mathbf{V}_{m+1:g}^T \mathbf{r}_0\right) \\
&= \sum_{j=m+1}^{g} \phi_j^{-1} (\mathbf{v}_j^T \mathbf{r}_0)^2 = g - m, \qquad 0 \leq m < g,
\end{aligned}$$

where the last inequality follows from $\phi_j = (\mathbf{v}_j^T \mathbf{r}_0)^2$ in Definition 4. □

Remark 30 The Z-statistic is inconclusive about the calibration of BayesCG under the Krylov prior.

Theorem 29 shows that the Z-statistic is distributed according to a Dirac distribution at $g - m$. Thus, the Z-statistic has the same mean as the chi-squared distribution χ_{g-m}^2, which suggests that BayesCG under the Krylov prior is neither optimistic or pessimistic. However, although the means are the same, the distributions are not. Hence, Theorem 29 does not imply that BayesCG under the Krylov prior is calibrated.

4.3 The S-statistic

We introduce a new test statistic for assessing the calibration of probabilistic solvers, the S-statistic. After discussing the relation between calibration and error estimation (Section 4.3.1), we define the S-statistic (Section 4.3.2), compare the S-statistic to the Z-statistic (Section 4.3.3), and then apply the S-statistic to BayesCG under the Krylov prior (Section 4.3.4).

4.3.1 Calibration and Error Estimation

We establish a relation between the error of the posterior means (approximations to the solution) and the trace of posterior covariances (Theorem 31).

Theorem 31 *Let* $\mathbf{A}X_* = B$ *be a class of linear systems where* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is symmetric positive definite and* $X_* \sim \mu_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$. *Let* $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$, $1 \leq m \leq n$ *be the posterior distributions from a probabilistic solver under the prior* μ_0 *applied to* $\mathbf{A}X_* = B$.

If the solver is calibrated, then

$$\mathbb{E}[\|X_* - \mathbf{x}_m\|_{\mathbf{A}}^2] = \mathrm{trace}(\mathbf{A}\mathbf{\Sigma}_m), \qquad 1 \leq m \leq n. \tag{27}$$

Proof For a calibrated solver Lemma 18 implies that $X_* \sim \mu_m$. Then apply Lemma 42 in Appendix A to the error $\|X_* - \mathbf{x}_m\|_{\mathbf{A}}^2$. □

For a calibrated solver, Theorem 31 implies that the equality $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2 = \mathrm{trace}(\mathbf{A}\mathbf{\Sigma}_m)$ holds *on average.* This means the trace can overestimate the error for some solutions, while for others, it can underestimate the error.

We explain how Theorem 31 relates the errors of a calibrated solver to the area in which its posteriors are concentrated.

Remark 32 The trace of a posterior covariance matrix quantifies the spread of its probability distribution—because the trace is the sum of the eigenvalues, which in the case of a covariance are the variances of the principal components [22, Section 12.2.1].

In analogy to viewing the $\mathbf{A}$-norm as the 2-norm weighted by $\mathbf{A}$, we can view $\mathrm{trace}(\mathbf{A}\mathbf{\Sigma}_m)$ as the trace of $\mathbf{\Sigma}_m$ weighted by $\mathbf{A}$. Theorem 31 shows that the $\mathbf{A}$-norm errors of a calibrated solver are equal to the weighted sum of the principal component variances from the posterior. Thus, the posterior means $\mathbf{x}_m$ and the areas in which the posteriors are concentrated both converge to the solution at the same speed, provided the solver is calibrated.

4.3.2 Definition of the S-statistic

We introduce the S-statistic (Definition 33), present an algorithm for generating samples from the S-statistic (Algorithm 4), and discuss their use for assessing calibration of solvers (Remark 34).

The S-statistic represents a necessary condition for calibration, as established in Theorem 31.

Definition 33 Let $\mathbf{A}X_* = B$ be a class of linear systems where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric positive definite, and $X_* \sim \mu_0 \equiv \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$. Let $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$, $1 \leq m \leq n$, be the posterior distributions from a probabilistic solver under the prior μ_0 applied to $\mathbf{A}X_* = B$. The S-statistic is

$$\mathrm{S}_m(X_*) \equiv \|X_* - \mathbf{x}_m\|_{\mathbf{A}}^2. \tag{28}$$

If the solver is calibrated then Theorem 31 implies

$$\mathbb{E}[\mathrm{S}(X_*)] = \mathrm{trace}(\mathbf{A}\mathbf{\Sigma}_m). \tag{29}$$

Generating samples from the S-statistic and assessing calibration. For a user specified probabilistic linear `solver` and a symmetric positive definite matrix $\mathbf{A}$, Algorithm 4 samples N_{test} solutions $\mathbf{x}_*$ from the prior distribution μ_0, defines the linear systems $\mathbf{b} = \mathbf{A}\mathbf{x}_*$, runs m iterations of the solver on the system, and computes $\mathrm{S}_m(\mathbf{x}_*)$ and $\mathrm{trace}(\mathbf{A}\mathbf{\Sigma}_m)$ from (28).

As with the Z-statistic, Algorithm 4 requires a separate reference μ_{ref} when sampling solutions $\mathbf{x}_*$ for BayesCG under the Krylov prior.

Algorithm 4 Sampling from the S-statistic

```
Input: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mu_0 = \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$, solver, $m$, $N_{\text{test}}$
for $i = 1 : N_{\text{test}}$ do
    Sample $\mathbf{x}_*$ from prior distribution $\mu_0$          ▷ Sample solution vector
    $\mathbf{b} = \mathbf{A}\mathbf{x}_*$          ▷ Define test problem
    $[\mathbf{x}_m, \mathbf{\Sigma}_m] = \texttt{solver}(\mathbf{A}, \mathbf{b}, \mu_0, m)$          ▷ Compute posterior $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$
    $s_i = \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$          ▷ Compute $S$-statistic for test problem
    $t_i = \mathrm{trace}(\mathbf{A}\mathbf{\Sigma}_m)$          ▷ Compute trace for test problem
end for
$h = (1/N_{test}) \sum_{i=1}^{N_{\text{test}}} s_i$          ▷ Compute empirical mean of $S$-statistic samples
Output: $S$-statistic samples $s_i$ and traces $t_i$, $1 \leq i \leq N_{\text{test}}$; $S$-statistic mean $h$
```

Remark 34 We assess calibration of the solver by comparing the S-statistic samples s_i from Algorithm 4 to the traces t_i, $1 \leq i \leq N_{test}$. The following criteria are based on Theorem 31 and Remark 32.

Calibrated: If the solver is calibrated, the traces t_i should all be equal to the empirical mean h of the S-statistic samples s_i.

Pessimistic: If the s_i are concentrated around smaller values than the t_i, then the solutions $\mathbf{x}_*$ occupy a smaller area of $\mathbb{R}^n$ than predicted by the posteriors μ_m.

Optimistic: If the s_i are concentrated around larger values than the t_i, then the solutions $\mathbf{x}_*$ occupy a larger area of $\mathbb{R}^n$ than predicted by μ_m.

We can also compare the empirical means of the s_i and t_i, because a calibrated solver should produce s_i and t_i with the same mean. Note that a comparison via the Kolmogorov-Smirnov statistic is not appropriate because the empirical distributions of s_i and t_i are generally different.

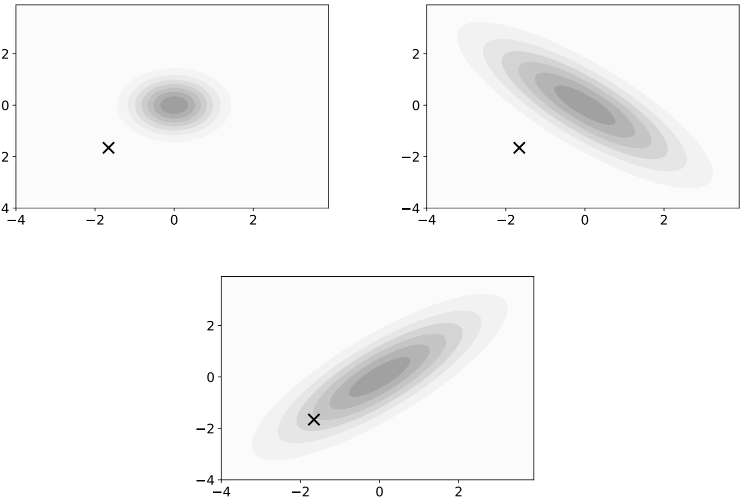

Figure 3 Assessment of calibration from Z-statistic and S-statistic. The contour plots represent the posterior distributions, and the symbol '$\times$' represents the solution.
Top left: Both statistics decide that the solver is not calibrated. Top right: The S-statistic decides that the solver is calibrated, while the Z-statistic does not. Bottom: Both statistics decide that the solver is calibrated.

4.3.3 Comparison of the Z- and S-statistics

Both, Z- and S-statistic represent necessary conditions for calibration ((27) and (29)); and both measure the norm of the error $X_* - \mathbf{x}_m$: The Z-statistic in the $\mathbf{\Sigma}_m^\dagger$-pseudo norm (Definition 23), and the S-statistic in the $\mathbf{A}$-norm (Definition 33). Deeper down, though, the Z-statistic projects errors onto a single dimension (Theorem 25), while the S-statistic relates errors to the areas in which the posterior distributions are concentrated.

Due to its focus on the area of the posteriors, the S-statistic can give a *false positive* for calibration. This occurs when the solution is not in the area of posterior concentration but the size of the posteriors is consistent with the errors. The Z-statistic is less likely to encounter this problem, as illustrated in Figure 3.

The Z-statistic is better at assessing calibration, while the S statistic produces accurate error estimates, which default to the traditional $\mathbf{A}$-norm estimates. The S-statistic is also faster to compute because it does not require the solution of a least squares problem.

4.3.4 S-statistic for BayesCG under the Krylov prior

We show that BayesCG under the Krylov prior is not calibrated, but that it is has similar performance to a calibrated solver under full posteriors and is optimistic under approximate posteriors.

Calibration of BayesCG under full Krylov posteriors. Theorem 7 implies that the S-statistic for any solution $\mathbf{x}_*$ is equal to

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

Thus, the S-statistic indicates that the size of Krylov posteriors is consistent with the errors, which is a desirable property of calibrated solvers. However, BayesCG under the Krylov prior is not a calibrated solver because the traces of posterior covariances from calibrated solvers are distributed around the *average* error instead of always being equal to the error.

Calibration of BayesCG under approximate posteriors. From (9) follows that $\mathrm{trace}(\mathbf{A}\widehat{\mathbf{\Gamma}}_m)$ is concentrated around smaller values than the S-statistic; and the underestimate of the trace is equal to the Wasserstein distance between full and approximate Krylov posteriors in Theorem 13. This underestimate points to the optimism of BayesCG under approximate Krylov posteriors. This optimism is expected because approximate posteriors model the uncertainty about $\mathbf{x}_*$ in a lower dimensional space than full posteriors.

5 Numerical experiments

We present numerical assessments of BayesCG calibration via the Z- and S-statistics.

After describing the setup of the numerical experiments (Section 5.1), we assess the calibration of three implementations of BayesCG: (i) BayesCG with random search directions (Section 5.2)—a solver known to be calibrated—so as to establish a baseline for comparisons with other versions of BayesCG; (ii) BayesCG under the inverse prior (Section 5.3); and (iii) BayesCG under the Krylov prior (Section 5.4). We choose the inverse prior and the Krylov priors because under each of these priors, the posterior mean from BayesCG coincides with the solution from CG.

Conclusions from all the experiments. Both, Z- and S statistics indicate that BayesCG with random search directions is indeed a calibrated solver, and that BayesCG under the inverse prior is pessimistic.

The S-statistic indicates that BayesCG under full Krylov posteriors mimics a calibrated solver, and that BayesCG under rank-50 approximate posteriors does as well, but not as much since it is slightly optimistic.

However, among all versions, BayesCG under approximate Krylov posteriors is the only one that is computationally practical and that is competitive with CG.

5.1 Experimental Setup

We describe the matrix $\mathbf{A}$ in the linear system (Section 5.1.1); the setup of the Z- and S-statistic experiments (Section 5.1.2); and the three BayesCG implementations (Section 5.1.3).

5.1.1 The matrix $\mathbf{A}$ *in the linear system (1)*

The symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ of dimension $n = 1806$ is a preconditoned version of the matrix BCSSTK14 from the Harwell-Boeing collection in [1]. Specifically,

$$\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}, \qquad \text{where} \quad \mathbf{D} \equiv \operatorname{diag}\begin{pmatrix}\mathbf{B}_{11} & \cdots & \mathbf{B}_{nn}\end{pmatrix}$$

and $\mathbf{B}$ is BCSSTK14. Calibration is assessed at iterations $m = 10, 100, 300$.

5.1.2 Z-statistic and S-statistic

The Z-statistic and S-statistic experiments are implemented as described in Algorithms 3 and 4, respectively. The calibration criteria for the Z-statistic are given in Remark 26, and for the S-statistic in Remark 34.

We sample from Gaussian distributions by exploiting their stability. According to Lemma 35 in Appendix A, if $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{F}\mathbf{F}^T = \mathbf{\Sigma}$ is a factorization of the covariance, then

$$\mathbf{F}Z + \mathbf{z} = X \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma}).$$

Samples $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are generated with `randn`$(n, 1)$ in Matlab, and with `numpy.random.randn`$(n, 1)$ in NumPy.

Z-statistic experiments. We quantify the distance between the Z-statistic samples and the chi-squared distribution by applying the Kolmogorov-Smirnov statistic (Definition 27) to the empirical cumulative distribution function of the Z-statistic samples and the analytical cumulative distribution function of the chi-squared distribution.

The degree of freedom in the chi-squared distribution is chosen as the median numerical rank of the posterior covariances. Note that the numerical rank of $\mathbf{\Sigma}_m$ can differ from

$$\operatorname{rank}(\mathbf{\Sigma}_m) = \operatorname{rank}(\mathbf{\Sigma}_0) - m,$$

and choosing the median rank gives an integer equal to the rank of at least one posterior covariance.

In compliance with the Matlab function `rank` and the NumPy function `numpy.linalg.rank`, we compute the numerical rank of $\mathbf{\Sigma}_m$ as

$$\operatorname{rank}(\mathbf{\Sigma}_m) = \text{cardinality}\{\sigma_i \mid \sigma_i > n\varepsilon\|\mathbf{\Sigma}_m\|_2\}, \tag{30}$$

where ε is machine epsilon and σ_i, $1 \leq i \leq n$, are the singular values of $\mathbf{\Sigma}_m$ [15, Section 5.4.1].

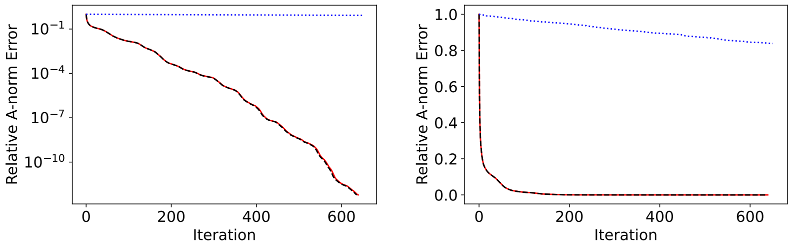

Figure 4 Relative error $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2/\|\mathbf{x}_*\|_{\mathbf{A}}^2$ for BayesCG under the inverse prior (solid line) and Krylov prior (dashed line), and BayesCG with random search directions (dotted line). Vertical axis has a logarithmic scale (left plot) and a linear scale (right plot).

5.1.3 Three BayesCG implementations

We use three versions of BayesCG: BayesCG with random search directions, BayesCG under the inverse prior, and BayesCG under the Krylov prior.

BayesCG with random search directions. The implementation in Algorithm 6 in Appendix B.2 computes posterior covariances that do not depend on the solution $\mathbf{x}_*$. This, in turn, requires search directions that do not depend on $\mathbf{x}_*$ [7, Section 1.1] which is achieved by starting with a random search direction $\mathbf{s}_1 = \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ instead of the initial residual $\mathbf{r}_0 \equiv \mathbf{b}_0 - \mathbf{A}\mathbf{x}_0$. The prior is $\mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$.

By design, this version of BayesCG is calibrated. However, it is also impractical due to its slow convergence, see Figure 4, and an accurate solution is available only after n iterations. The random initial search direction $\mathbf{s}_1$ leads to uninformative m-dimensional subspaces, so that the solver has to explore all of $\mathbb{R}^n$ before finding the solution.

BayesCG under the inverse prior $\mu_0 \equiv \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$. The implementation in Algorithm 7 in Appendix B.3 is a modified version of Algorithm 1 for general priors that maintains the posterior covariances in factored form.

BayesCG under the Krylov prior. For full posteriors, the modified Lanczos solver Algorithm 5 in Appendix B.4 computes the full prior, which is then followed by the direct computation of the posteriors from the prior in Algorithm 6.

For approximate posteriors, Algorithm 9 in Appendix B.4 computes rank-d covariances at the same computational cost as $m + d$ iterations of CG.

In Z- and S-statistic experiments, solutions $\mathbf{x}_*$ are usually sampled from the prior distribution. We cannot sample solutions from the Krylov prior because it differs from solution to solution. Instead we sample solutions from the reference distribution $\mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$. This is a reasonable choice because the posterior means in BayesCG under the inverse and Krylov priors coincide with the CG iterates [8, Section 3].

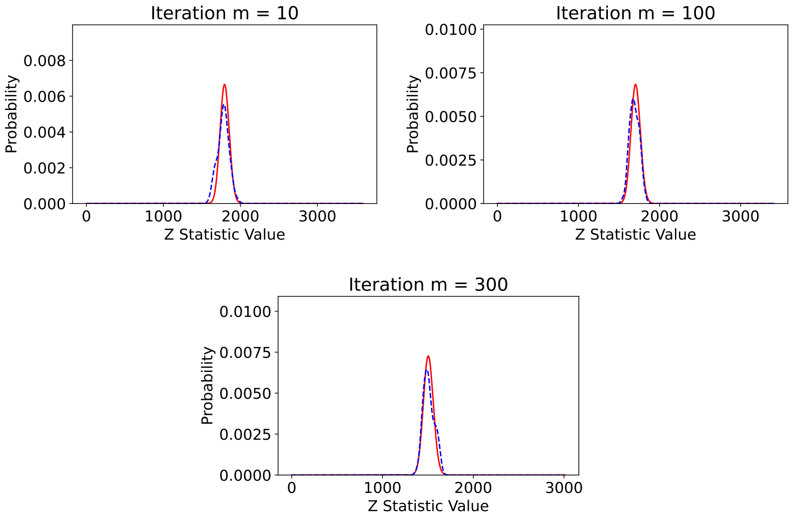

Figure 5 Z-statistic samples for BayesCG with random search directions after $m = 10, 100, 300$ iterations. The solid curve represents the chi-squared distribution and the dashed curve the Z-statistic samples.

Iteration	Z-stat mean	χ^2 mean	K-S statistic
10.0	1.79×10^3	1.8×10^3	0.139
100.0	1.69×10^3	1.71×10^3	0.161
300.0	1.5×10^3	1.51×10^3	9.65×10^{-2}

Table 1 This table corresponds to Figure 5. For BayesCG with random search directions, it shows the Z-statistic sample means; the chi-squared distribution means; and the Kolmogorov-Smirnov statistic between the Z-statistic samples and the chi-squared distribution.

Iteration	S-stat mean	Trace mean	Trace standard deviation
10.0	1.78×10^3	1.8×10^3	2.93×10^{-12}
100.0	1.69×10^3	1.71×10^3	2.29×10^{-12}
300.0	1.51×10^3	1.51×10^3	2.07×10^{-12}

Table 2 This table corresponds to Figure 6. For BayesCG with random search directions, it shows the S-statistic sample means, the trace means, and the trace standard deviations.

5.2 BayesCG with random search directions

By design, BayesCG with random search directions is a calibrated solver. The purpose is to establish a baseline for comparisons with BayesCG under the inverse and Krylov priors, and to demonstrate that the Z- and S-statistics perform as expected on a calibrated solver.

Summary of experiments below. Both, Z- and S-statistics strongly confirm that BayesCG with random search directions is indeed a calibrated solver, thereby corroborating the statements in Theorem 25 and Definition 33.

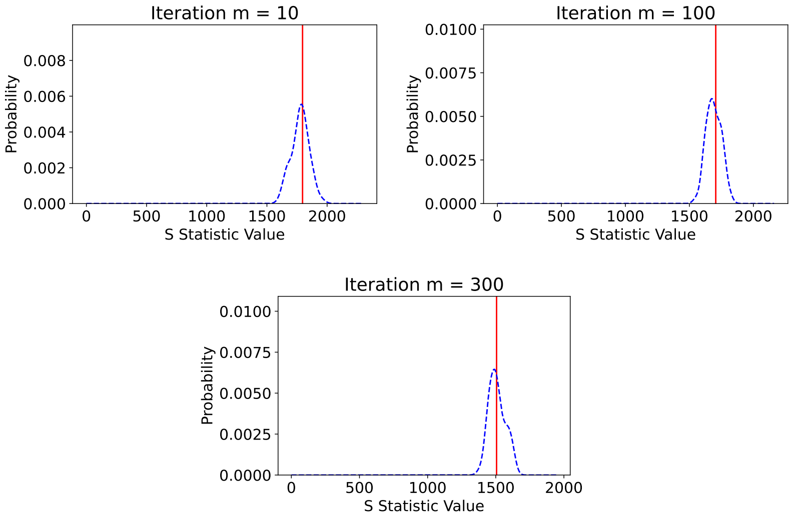

Figure 6 S-statistic samples and traces for BayesCG with random search directions after $m = 10, 100, 300$ iterations. The solid curve represents the traces and the dashed curve the S-statistic samples.

Figure 5 and Table 1. The Z-statistic samples in Figure 5 almost match the chi-squared distribution; and the Kolmogorov-Smirnov statistics in Table 1 are on the order of 10^{-1}, meaning close to zero. This confirms that BayesCG with random search directions is indeed calibrated.

Figure 6 and Table 2. The traces in Figure 6 are tightly concentrated around the empirical mean of the S-statistic samples. Table 2 confirms the strong clustering of the trace and S-statistic sample means around 10^{-3}, together with the very small deviation of the traces. Thus, the area in which the posteriors are concentrated is consistent with the error, confirming again that BayesCG with random search directions is calibrated.

5.3 BayesCG under the inverse prior $\mu_0 = \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$.

Summary of experiments below. Both, Z- and S-statics indicate that BayesCG under the inverse prior is pessimistic, and that the pessimism increases with the iteration count. This is consistent with the experiments in [8, Section 6.1].

Figure 7 and Table 3. The Z-statistic samples in Figure 7 are concentrated around smaller values than the predicted chi-squared distribution. The Kol--mogorov-Smirnov statistics in Table 3 are all equal to 1, indicating no overlap between Z-statistic samples and chi-squared distribution. The first two columns of Table 3 show that Z-statistic samples move further away from the chi-squared distribution as the iterations progress. Thus, BayesCG under

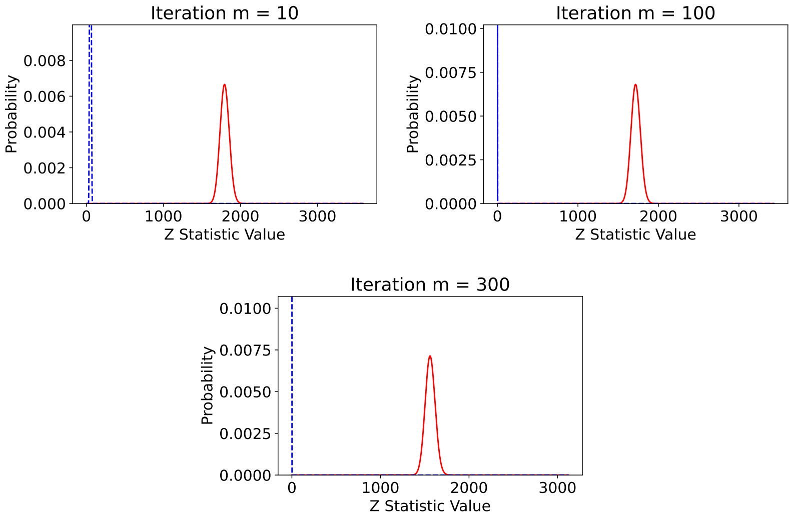

Figure 7 Z-statistic samples for BayesCG under the inverse prior after $m = 10, 100, 300$ iterations. The solid curve represents the chi-squared distribution and the dashed curve the Z-statistic samples.

Iteration	Z-stat mean	χ^2 mean	K-S statistic
10.0	51.9	1.8×10^3	1.0
100.0	0.545	1.72×10^3	1.0
300.0	1.33×10^{-5}	1.56×10^3	1.0

Table 3 This table corresponds to Figure 7. For BayesCG under the inverse prior, it shows the Z-statistic sample means; the chi-squared distribution means; and Kolmogorov-Smirnov statistic between the Z-statistic samples and the chi-squared and distribution.

Iteration	S-stat mean	Trace mean	Trace standard deviation
10.0	51.8	1.8×10^3	2.99×10^{-12}
100.0	0.574	1.71×10^3	0.164
300.0	3.57×10^{-6}	1.61×10^3	1.02

Table 4 This table corresponds to Figure 8. For BayesCG under the inverse prior, it shows the S-statistic sample means, the trace means, and the trace standard deviations.

the inverse prior is pessimistic, and the pessimism increases with the iteration count.

Figure 8 and Table 4. The S-statistic samples in Figure 8 are concentrated around smaller values than the traces. Table 4 indicates trace values at 10^3, while the S-statistic samples move towards zero as the iteration progress. Thus the errors are much smaller than the area in which the posteriors are concentrated, meaning the posteriors overestimate the error. This again confirms the pessimism of BayesCG under the inverse prior.

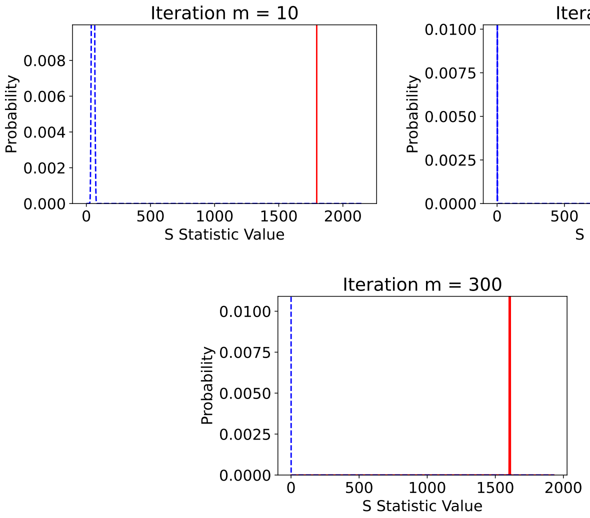

Figure 8 S-statistic samples and traces for BayesCG under the inverse prior after $m = 10, 100, 300$ iterations. The solid curve represents the traces and the dashed curve the S-statistic samples.

Iteration	Z-stat mean	χ^2 mean	K-S statistic
10.0	631.0	566.0	0.902
100.0	509.0	450.0	0.752
300.0	201.0	152.0	0.941

Table 5 This table corresponds to Figure 9. For BayesCG under the Krylov prior and full posteriors, it shows the Z-statistic sample means; the chi-squared distribution means; and the Kolmogorov-Smirnov statistic between the Z-statistic samples and the chi-squared distribution.

5.4 BayesCG under the Krylov prior

We consider full posteriors (Section 5.4.1), and then rank-50 approximate posteriors (Section 5.4.2).

5.4.1 Full Krylov posteriors

Summary of experiments below. The Z-statistic indicates that BayesCG under full Krylov posteriors is somewhat optimistic, while the S-statistic indicates resemblance to a calibrated solver.

Figure 9 and Table 5. The Z-statistic samples in Figure 9 are concentrated at somewhat larger values than the predicted chi-squared distribution. The Kolmogorov-Smirnov statistics in Table 5 are around .8 and .9, thus close to 1, and indicate very little overlap between Z-statistic samples and chi-squared distribution. Thus, BayesCG under full Krylov posteriors is somewhat optimistic.

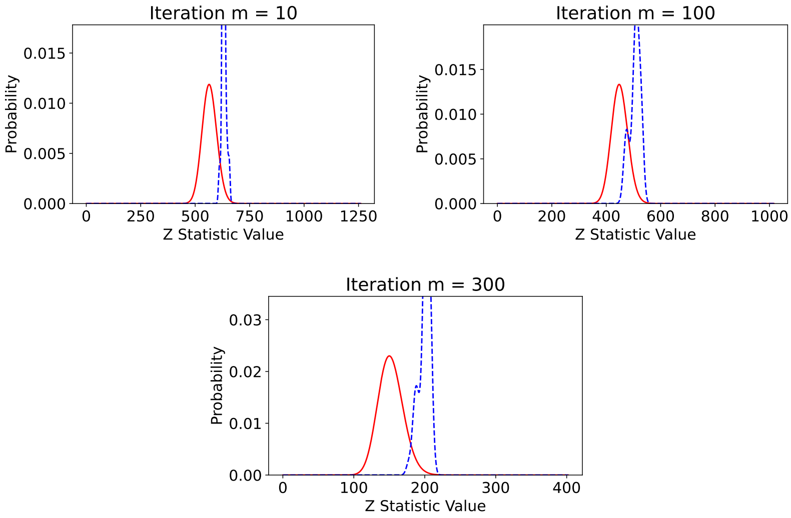

Figure 9 Z-statistic samples for BayesCG under the Krylov prior and full posteriors at $m = 10, 100, 300$ iterations. The solid curve represents the predicted chi-squared distribution and the dashed curve the Z-statistic samples.

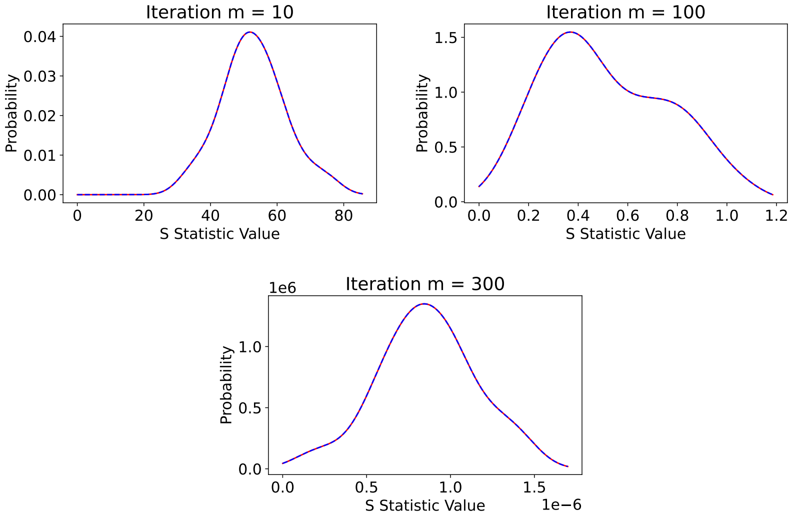

Figure 10 S-statistic samples and traces for BayesCG under the Krylov prior and full posteriors at $m = 10, 100, 300$ iterations. The solid curve represents the traces and the dashed curve the S-statistic samples.

Iteration	S-stat mean	Trace mean	Trace standard deviation
10.0	52.9	52.9	9.4
100.0	0.515	0.515	0.241
300.0	8.59×10^{-7}	8.59×10^{-7}	2.84×10^{-7}

Table 6 This table corresponds to Figure 10. For BayesCG under the Krylov prior and full posteriors, it shows the S-statistic sample means, the trace means, and the trace standard deviations.

Iteration	Z-stat mean	χ^2 mean	K-S statistic
10.0	319.0	50.0	1.0
100.0	375.0	50.0	1.0
300.0	194.0	50.0	1.0

Table 7 This table corresponds to Figure 11. For BayesCG under rank-50 approximate Krylov posteriors, it shows the Z-statistic sample means; chi-squared distribution means; and Kolmogorov-Smirnov statistic between the Z-statistic samples and the chi-squared distribution.

These numerical results differ from Theorem 29, which predicts Z-statistic samples equal to $g - m$. A possible reason might be that the rank of the Krylov prior computed by Algorithm 5 is smaller than the exact rank. In exact arithmetic, $\text{rank}(\mathbf{\Gamma}_0) = g = n = 1806$. However, in finite precision, $\text{rank}(\mathbf{\Gamma}_0)$ is determined by the convergence tolerance which is set to 10^{-12}, resulting in $\text{rank}(\mathbf{\Gamma}_0) < g$.

Figure 10 and Table 6. The S-statistic samples in Figure 10 match the traces extremely well, with Table 6 showing an agreement to 3 figures, as predicted in Section 4.3.4, Thus, the area in which the posteriors are concentrated is consistent with the error, as would be expected from a calibrated solver.

However, BayesCG under the Krylov prior does not behave exactly like a calibrated solver, such as BayesCG with random search directions in Section 5.2, where all traces are concentrated at the empirical mean of the S-statistic samples. Thus, BayesCG under the Krylov prior is not calibrated but has a performance similar to that of a calibrated solver.

5.4.2 Rank-50 approximate Krylov posteriors

Summary of the experiments below. Both, Z- and S-statistic indicate that BayesCG under rank-50 approximate Krylov posteriors is somewhat optimistic, and is not as close to a calibrated solver as BayesCG with full Krylov posteriors. In contrast to the Z-statistic, the respective S-statistic samples and traces for BayesCG under full and rank-50 posteriors are close.

Figure 11 and Table 7. The Z-statistic samples in Figure 11 are concentrated around larger values than the predicted chi-squared distribution, which is steady at 50. All Kolmogorov-Smirnov statistics in Table 7 are equal to 1, indicating no overlap between Z-statistic samples and chi-squared distribution. Thus, BayesCG under approximate Krylov posteriors is more optimistic than BayesCG under full posteriors.

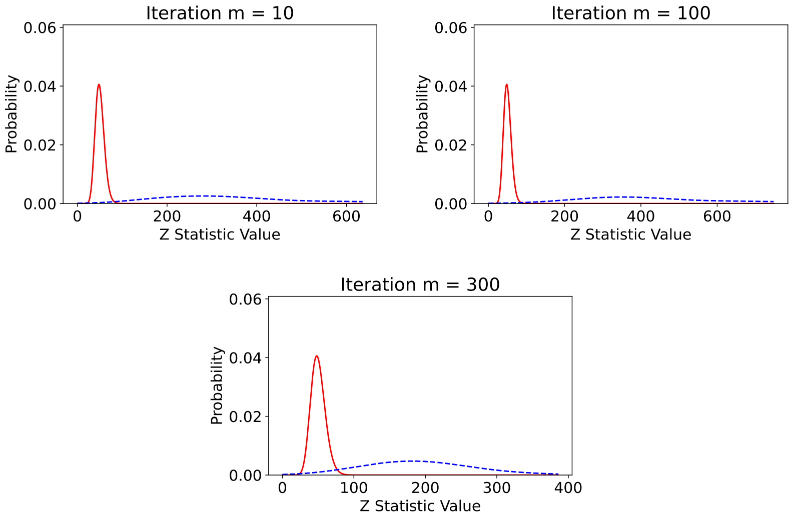

Figure 11 Z-statistic samples for BayesCG under rank-50 approximate Krylov posteriors at $m = 10, 100, 300$ iterations. The solid curve represents the predicted chi-squared distribution and the dashed curve the Z-statistic samples.

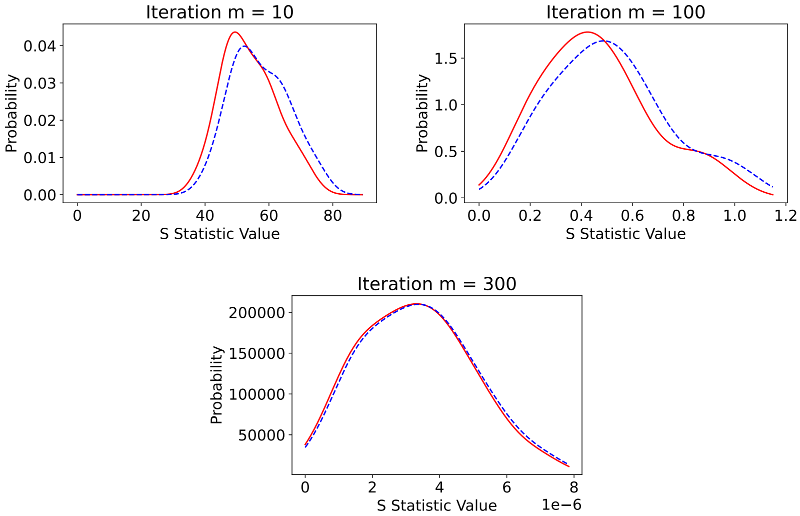

Figure 12 S-statistic samples and traces for BayesCG under rank-50 approximate Krylov posteriors at $m = 10, 100, 300$ iterations. The solid curve represents the traces and the dashed curve the S-statistic samples.

Iteration	S-stat mean	Trace mean	Trace standard deviation
10.0	57.2	53.9	8.32
100.0	0.517	0.467	0.214
300.0	3.37×10^{-6}	3.29×10^{-6}	1.6×10^{-6}

Table 8 This table corresponds to Figure 12. For BayesCG under rank-50 approximate Krylov posteriors, it shows the S-statistic sample means, trace means, and trace standard deviations.

Figure 12 and Table 8. The traces in Figure 12 are concentrated around slightly smaller values than the S-statistic samples, but they all have the same order of magnitude, as shown in Table 8. This means, the errors are slightly larger than the area in which the posteriors are concentrated; and the posteriors slightly underestimate the errors.

A comparison of Tables 6 and 8 shows that the S-statistic samples and traces, respectively, for full and rank-50 posteriors are close. From the point of view of the S-statistic, BayesCG under approximate Krylov posteriors is somewhat optimistic, and close to being a calibrated solver but not as close as BayesCG under full Krylov posteriors.

A Auxiliary Results

We present auxiliary results required for proofs in other sections.

The stability of Gaussian distributions implies that a linear transformation of a Gaussian random variable remains Gaussian.

Lemma 35 (Stability of Gaussian Distributions [27, Section 1.2]) *Let $X \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$ be a Gaussian random variable with mean $\mathbf{x} \in \mathbb{R}^n$ and covariance $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$. If $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{F} \in \mathbb{R}^{n\times n}$, then*

$$Z = \mathbf{y} + \mathbf{F}X \sim \mathcal{N}(\mathbf{y} + \mathbf{F}\mathbf{x}, \mathbf{F}\mathbf{\Sigma}\mathbf{F}^T).$$

The conjugacy of Gaussian distributions implies that the distribution of a Gaussian random variable conditioned on information that linearly depends on the random variable is a Gaussian distribution.

Lemma 36 (Conjugacy of Gaussian Distributions [30, Section 6.1], [37, Corollary 6.21]) *Let $X \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma}_x)$ and $Y \sim \mathcal{N}(\mathbf{y}, \mathbf{\Sigma}_y)$. The jointly Gaussian random variable $\begin{bmatrix} X^T & Y^T \end{bmatrix}^T$ has the distribution*

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_x & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{xy}^T & \mathbf{\Sigma}_y \end{bmatrix} \right),$$

where $\mathbf{\Sigma}_{xy} \equiv \mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbf{x})(Y - \mathbf{y})^T]$ and the conditional distribution of X given Y is

$$(X \mid Y) \sim \mathcal{N}(\overbrace{\mathbf{x} + \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_y^{\dagger}(Y - \mathbf{y})}^{mean},\ \overbrace{\mathbf{\Sigma}_x - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_y^{\dagger}\mathbf{\Sigma}_{xy}^T}^{covariance}).$$

We show how to transform a $\mathbf{B}$-orthogonal matrix into an orthogonal matrix.

Lemma 37 *Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ be symmetric positive definite, and let $\mathbf{H} \in \mathbb{R}^{n\times n}$ be a $\mathbf{B}$-orthogonal matrix with $\mathbf{H}^T\mathbf{B}\mathbf{H} = \mathbf{H}\mathbf{B}\mathbf{H}^T = \mathbf{I}$. Then*

$$\mathbf{U} \equiv \mathbf{B}^{1/2}\mathbf{H}$$

is an orthogonal matrix with $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$.

Proof The symmetry of $\mathbf{B}$ and the $\mathbf{B}$-orthogonality of $\mathbf{H}$ imply

$$\mathbf{U}^T\mathbf{U} = \mathbf{H}^T\mathbf{B}\mathbf{H} = \mathbf{I}.$$

From the orthonormality of the columns of $\mathbf{U}$, and the fact that $\mathbf{U}$ is square follows that $\mathbf{U}$ is an orthogonal matrix [21, Definition 2.1.3]. □

Definition 38 [21, Section 7.3] The *thin singular value decomposition* of the rank-p matrix $\mathbf{G} \in \mathbb{R}^{m\times n}$ is

$$\mathbf{G} = \mathbf{U}\mathbf{D}\mathbf{W}^T,$$

where $\mathbf{U} \in \mathbb{R}^{m\times p}$ and $\mathbf{W} \in \mathbb{R}^{n\times p}$ are matrices with orthonormal columns and $\mathbf{D} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with positive diagonal elements. The *Moore-Penrose inverse* of $\mathbf{G}$ is

$$\mathbf{G}^\dagger = \mathbf{W}\mathbf{D}^{-1}\mathbf{U}^T.$$

If a matrix has full column-rank or full row-rank, then its Moore-Penrose can be expressed in terms of the matrix itself. Furthermore, the Moore-Penrose inverse of a product is equal to the product of the Moore-Penrose inverses, provided the first matrix has full column-rank and the second matrix has full row-rank.

Lemma 39 ([5, Corollary 1.4.2]) *Let* $\mathbf{G} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{J} \in \mathbb{R}^{n\times p}$ *have full column and row rank respectively, so* $\operatorname{rank}(\mathbf{G}) = \operatorname{rank}(\mathbf{J}) = n$. *The Moore-Penrose inverses of* $\mathbf{G}$ *and* $\mathbf{J}$ *are*

$$\mathbf{G}^\dagger = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T \quad \textit{and} \quad \mathbf{J}^\dagger = \mathbf{J}^T(\mathbf{J}\mathbf{J}^T)^{-1}$$

respectively, and the Moore-Penrose inverse of the product equals

$$(\mathbf{G}\mathbf{J})^\dagger = \mathbf{J}^\dagger\mathbf{G}^\dagger.$$

Below is an explicit expression for the mean of a quadratic form of Gaussians.

Lemma 40 ([26, Sections 3.2b.1–3.2b.3]) *Let* $Z \sim \mathcal{N}(\mathbf{x}_z, \mathbf{\Sigma}_z)$ *be a Gaussian random variable in* $\mathbb{R}^n$, *and* $\mathbf{B} \in \mathbb{R}^{n\times n}$ *be symmetric positive definite. The mean of* $Z^T\mathbf{B}Z$ *is*

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

We show that the squared Euclidean norm of a Gaussian random variable with an orthogonal projector as its covariance matrix is distributed according to a chi-squared distribution.

Lemma 41 *Let* $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{P})$ *be a Gaussian random variable in* $\mathbb{R}^n$. *If the covariance matrix* $\mathbf{P}$ *is an orthogonal projector, that is, if* $\mathbf{P}^2 = \mathbf{P}$ *and* $\mathbf{P} = \mathbf{P}^T$, *then*

$$\|X\|_2^2 = (X^TX) \sim \chi_p^2,$$

where $p = \operatorname{rank}(\mathbf{P})$.

Proof We express the projector in terms of orthonormal matrices and then use the invariance of the 2-norm under orthogonal matrices and the stability of Gaussians.

Since $\mathbf{P}$ is an orthogonal projector, there exists $\mathbf{U}_1 \in \mathbb{R}^{n\times p}$ such that $\mathbf{U}_1\mathbf{U}_1^T = \mathbf{P}$ and $\mathbf{U}_1^T\mathbf{U} = \mathbf{I}_p$. Choose $\mathbf{U}_2 \in \mathbb{R}^{n\times(n-p)}$ so that $\mathbf{U} = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}$ is an orthogonal matrix. Thus,

$$X^TX = X^T\mathbf{U}\mathbf{U}^TX = X^T\mathbf{U}_1\mathbf{U}_1^TX + X^T\mathbf{U}_2\mathbf{U}_2^TX. \tag{31}$$

Lemma 35 implies that $Y = \mathbf{U}_1^TX$ is distributed according to a Gaussian distribution with mean $\mathbf{0}$ and covariance $\mathbf{U}_1^T\mathbf{U}_1\mathbf{U}_1^T\mathbf{U} = \mathbf{I}_p$. Similarly, $Z = \mathbf{U}_2^TX$ is distributed according to a Gaussian distribution with mean $\mathbf{0}$ and covariance $\mathbf{U}_2^T\mathbf{U}_1\mathbf{U}_1^T\mathbf{U}_2 = \mathbf{0}$, thus $Z = \mathbf{0}$.

Substituting Y and Z into (31) gives $X^TX = Y^TY + \mathbf{0}^T\mathbf{0}$. From $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ follows $(X^TX) \sim \chi_p^2$. □

Lemma 42 *If $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric positive definite, and $M \sim \mathcal{N}(\mathbf{x}_\mu \boldsymbol{\Sigma}_\mu)$ and $N \sim \mathcal{N}(\mathbf{x}_\nu, \boldsymbol{\Sigma}_\nu)$ are independent random variables in $\mathbb{R}^n$, then*

$$\mathbb{E}[\|M-N\|_{\mathbf{A}}^2] = \|\mathbf{x}_\mu - \mathbf{x}_\nu\|_{\mathbf{A}}^2 + \operatorname{trace}(\mathbf{A}\boldsymbol{\Sigma}_\mu) + \operatorname{trace}(\mathbf{A}\boldsymbol{\Sigma}_\nu).$$

Proof The random variable $M - N$ has mean $\mathbb{E}[M-N] = \mathbf{x}_\mu - \mathbf{x}_\nu$, and covariance

$$\begin{aligned}\boldsymbol{\Sigma}_{M-N} &\equiv \operatorname{Cov}(M-N, M-N)\\ &= \operatorname{Cov}(M,M) + \operatorname{Cov}(N,N) - \operatorname{Cov}(M,N) - \operatorname{Cov}(N,M)\\ &= \operatorname{Cov}(M,M) + \operatorname{Cov}(N,N) = \boldsymbol{\Sigma}_\mu + \boldsymbol{\Sigma}_\nu,\end{aligned}$$

where the covariances $\operatorname{Cov}(M,N) = \operatorname{Cov}(N,M) = 0$ because M and N are independent. Now apply Lemma 40 to $M - N$. □

B Algorithms

We present algorithms for the modified Lanczos method (Section B.1), BayesCG with random search directions (Section B.2), BayesCG with covariances in factored form (Section B.3), and BayesCG under the Krylov prior (Section B.4).

B.1 Modified Lanczos method

The Lanczos method [34, Algorithm 6.15] produces an orthonormal basis for the Krylov space $\mathcal{K}_g(\mathbf{A}, \mathbf{v}_1)$, while the modified version in Algorithm 5 produces an $\mathbf{A}$-orthonormal basis.

Algorithm 5 Modified Lanczos Method

```
 1: Input: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{v}_1 \in \mathbb{R}^n$, basis dimension $m$, convergence tolerance $\varepsilon$
 2: $\mathbf{v}_0 = \mathbf{0} \in \mathbb{R}^n$
 3: $i = 1$
 4: $\beta = (\mathbf{v}_i^T \mathbf{A} \mathbf{v}_i)^{1/2}$
 5: $\mathbf{v}_i = \mathbf{v}_i / \beta$
 6: while $i \leq m$ do
 7:     $\mathbf{w} = \mathbf{A}\mathbf{v}_i - \beta \mathbf{v}_{i-1}$
 8:     $\alpha = \mathbf{w}^T \mathbf{A} \mathbf{v}_i$
 9:     $\mathbf{w} = w - \alpha \mathbf{v}_i$
10:     $\mathbf{w} = \mathbf{w} - \sum_{j=1}^{i} \mathbf{v}_j \mathbf{v}_j^T \mathbf{A}\mathbf{w}$          ▷ Reorthogonalize $\mathbf{w}$
11:     $\mathbf{w} = \mathbf{w} - \sum_{j=1}^{i} \mathbf{v}_j \mathbf{v}_j^T \mathbf{A}\mathbf{w}$
12:     $\beta = (\mathbf{w}^T \mathbf{A} \mathbf{w})^{1/2}$
13:     if $\beta < \varepsilon$ then
14:         Exit while loop
15:     end if
16:     $i = i + 1$
17:     $\mathbf{v}_i = \mathbf{w} / \beta$
18: end while
19: $m = i - 1$                                   ▷ Number of basis vectors
20: Output: $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$          ▷ $\mathbf{A}$-orthonormal basis of $\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1)$
```

Algorithm 5 reorthogonalizes the basis vectors $\mathbf{v}_i$ with Classical Gram-Schmidt performed twice, see Lines 10 and 11. This reorthogonalization technique can be implemented efficiently and produces vectors that are orthogonal to machine precision [13, 14].

B.2 BayesCG with random search directions

The version of BayesCG in Algorithm 6 is designed to be calibrated because the search directions do not depend on $\mathbf{x}_*$, hence the posteriors do not depend on $\mathbf{x}_*$ either [7, Section 1.1].

After sampling an initial random search direction $\mathbf{s}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, Algorithm 6 computes an $\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}$-orthonormal basis for the Krylov space $\mathcal{K}_m(\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}, \mathbf{s}_1)$ with Algorithm 5. Then Algorithm 6 computes the BayesCG posteriors directly with (2) and (3) from Theorem 1. The numerical experiments in Section 5 run Algorithm 6 with the inverse prior $\mu_0 = \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$.

Algorithm 6 BayesCG with random search directions

```
1: Inputs: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, prior $\mu_0 = \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$, iteration count $m$
2: $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$                                  ▷ Initial residual
3: Sample $\mathbf{s}_1$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$                      ▷ Initial search direction
4: Compute columns of $\mathbf{S}$ with Algorithm 5
5: $\mathbf{\Lambda}_m = \mathbf{S}_m^T\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_m$                          ▷ $\mathbf{\Lambda}_m$ is diagonal
6: $\mathbf{x}_m = \mathbf{x}_0 + \mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_m\mathbf{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{r}_0$            ▷ Compute posterior mean with (2)
7: $\mathbf{\Sigma}_m = \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_m\mathbf{\Lambda}_m^{-1}\mathbf{S}_m^T\mathbf{A}\mathbf{\Sigma}_0$   ▷ Compute posterior covariance with (3)
8: Output: $\mu_m = \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$
```

B.3 BayesCG with covariances in factored form

Algorithm 7 takes as input a general prior covariance $\mathbf{\Sigma}_0$ in factored form, and subsequently maintains the posterior covariances $\mathbf{\Sigma}_m$ in factored form as well. Theorem 43 presents the correctness proof for Algorithm 7.

Theorem 43 *Under the conditions of Theorem 1, if* $\mathbf{\Sigma}_0 = \mathbf{F}_0\mathbf{F}_0^T$ *for* $\mathbf{F}_0 \in \mathbb{R}^{n\times\ell}$ *and some* $m \le \ell \le n$, *then* $\mathbf{\Sigma}_m = \mathbf{F}_m\mathbf{F}_m^T$ *with*

$$\mathbf{F}_m = \mathbf{F}_0\left(\mathbf{I} - \mathbf{F}_0^T\mathbf{A}\mathbf{S}_m(\mathbf{S}_m^T\mathbf{A}\mathbf{F}_0\mathbf{F}_0^T\mathbf{A}\mathbf{S}_m)^{-1}\mathbf{S}_m\mathbf{A}\mathbf{F}_0\right) \in \mathbb{R}^{n\times\ell}, \qquad 1 \le m \le n.$$

Proof Fix m. Substituting $\mathbf{\Sigma}_0 = \mathbf{F}_0\mathbf{F}_0^T$ into (3) and factoring out $\mathbf{F}_0$ on the left and $\mathbf{F}_0^T$ on the right gives $\mathbf{\Sigma}_m = \mathbf{F}_0\mathbf{P}\mathbf{F}_0^T$ where

$$\begin{aligned}\mathbf{P} &\equiv \mathbf{I} - \mathbf{F}_0^T\mathbf{A}\mathbf{S}_m(\mathbf{S}_m^T\mathbf{A}\mathbf{F}_0\mathbf{F}_0^T\mathbf{A}\mathbf{S}_m)^{-1}\mathbf{S}_m\mathbf{A}\mathbf{F}_0\\ &= (\mathbf{I} - \mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T) \qquad \text{where} \quad \mathbf{Q} \equiv \mathbf{F}_0^T\mathbf{A}\mathbf{S}_m.\end{aligned}$$

Show that $\mathbf{P}$ is a projector,

$$\begin{aligned}\mathbf{P}^2 &= \mathbf{I} - 2\mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T + \mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\\ &= \mathbf{I} - \mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T = \mathbf{P}.\end{aligned}$$

Hence $\mathbf{\Sigma}_m = \mathbf{F}_0\mathbf{P}\mathbf{F}_0^T = \mathbf{F}_0\mathbf{P}\mathbf{P}\mathbf{F}_0^T = \mathbf{F}_m\mathbf{F}_m^T$.

B.4 BayesCG under the Krylov prior

We present algorithms for BayesCG under full Krylov posteriors (Section B.4.1) and under approximate Krylov posteriors (Section B.4.2).

Algorithm 7 BayesCG with covariances in factored form

```
Input: spd $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}_0 \in \mathbb{R}^n$, $\mathbf{F}_0 \in \mathbb{R}^{n \times \ell}$    ▷ need $\mathbf{x}_* - \mathbf{x}_0 \in \mathrm{range}(\mathbf{\Sigma}_0)$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{s}_1 = \mathbf{r}_0$
$\mathbf{P} = \mathbf{0} \in \mathbb{R}^{n \times n}$
$m = 0$
while not converged do
    $m = m + 1$
    $\mathbf{P}(:, m) = \mathbf{F}_0^T \mathbf{A}\mathbf{s}_m$    ▷ Save column $m$ of $\mathbf{P}$
    $\mathbf{q} = \mathbf{F}_0 \mathbf{P}(:, m)$    ▷ Compute $\mathbf{q} = \mathbf{\Sigma}_0 \mathbf{A}\mathbf{s}_m$
    $\eta_m = \mathbf{s}_m^T \mathbf{A}\mathbf{q}$
    $\mathbf{P}(:, m) = \mathbf{P}(:, m)/\eta_m$    ▷ Normalize column $m$ of $\mathbf{P}$
    $\alpha_m = \left(\mathbf{r}_{m-1}^T \mathbf{r}_{m-1}\right)/\eta_m$
    $\mathbf{x}_m = \mathbf{x}_{m-1} + \alpha_m \mathbf{q}$
    $\mathbf{r}_m = \mathbf{r}_{m-1} - \alpha_m \mathbf{A}\mathbf{q}$
    $\beta_m = \left(\mathbf{r}_m^T \mathbf{r}_m\right) / \left(\mathbf{r}_{m-1}^T \mathbf{r}_{m-1}\right)$
    $\mathbf{s}_{m+1} = \mathbf{r}_m + \beta_m \mathbf{s}_m$
end while
$\mathbf{P} = \mathbf{P}(:, 1:m)$    ▷ Discard unused columns of $\mathbf{P}$
$\mathbf{F}_m = \mathbf{F}_0(\mathbf{I} - \mathbf{P}\mathbf{P}^T)$
Output: $\mathbf{x}_m$, $\mathbf{F}_m$    ▷ Final posterior
```

B.4.1 Full Krylov posteriors

Algorithm 8 computes the following: a matrix $\mathbf{V}$ whose columns are an $\mathbf{A}$-orthonormal basis for $\mathcal{K}_g(\mathbf{A}, \mathbf{r}_0)$; the diagonal matrix $\mathbf{\Phi}$ in (5); and the posterior mean $\mathbf{x}_m$ in (26). The output consists of the posterior mean $\mathbf{x}_m$, and the factors $\mathbf{V}_{m+1:g}$ and $\mathbf{\Phi}_{m+1:g}$ for the posterior covariance.

Algorithm 8 BayesCG under the Krylov prior with full posteriors

```
Inputs: spd $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}_0 \in \mathbb{R}^n$, iteration count $m$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$    ▷ Initial residual
$\mathbf{v}_1 = \mathbf{r}_0$    ▷ Initial search direction
Compute columns of $\mathbf{V}$ with Algorithm 5
$\mathbf{\Phi} = \mathrm{diag}((\mathbf{V}^T \mathbf{r}_0)^2)$    ▷ Compute $\mathbf{\Phi}$ with (5)
$\mathbf{x}_m = \mathbf{x}_0 + \mathbf{V}_{1:m} \mathbf{V}_{1:m}^T \mathbf{r}_0$    ▷ Compute posterior mean with (26)
Output: $\mathbf{x}_m$, $\mathbf{V}_{m+1:g}$, $\mathbf{\Phi}_{m+1:g}$
```

B.4.2 Approximate Krylov posteriors

Algorithm 9 computes rank-d approximate Krylov posteriors in two main steps: (i) posterior mean and iterates $\mathbf{x}_m$ in Lines 5-14; and (ii) factorization of the posterior covariance $\widehat{\mathbf{\Gamma}}_m$ in Lines 16-26.

References

1. BCSSTK14: BCS Structural Engineering Matrices (linear equations) Roof of the Omni Coliseum, Atlanta.

Algorithm 9 BayesCG under the Krylov prior [32, Algorithm 3.1]

```
Inputs: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}_0 \in \mathbb{R}^n$, iteration count $m$, posterior rank $d$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$                      ▷ Initial residual
$\mathbf{v}_1 = \mathbf{r}_0$                                  ▷ Initial search direction
$i = 0$                                        ▷ Initial iteration counter
while $i < m$ do                               ▷ CG recursions for posterior means
    $i = i + 1$                                ▷ Increment iteration count
    $\eta_i = \mathbf{v}_i^T \mathbf{A}\mathbf{v}_i$
    $\gamma_i = (\mathbf{r}_{i-1}^T \mathbf{r}_{i-1}) / \eta_i$                ▷ Next step size
    $\mathbf{x}_i = \mathbf{x}_{i-1} + \gamma_i \mathbf{v}_i$                   ▷ Next iterate
    $\mathbf{r}_i = \mathbf{r}_{i-1} - \gamma_i \mathbf{A}\mathbf{v}_i$                 ▷ Next residual
    $\delta_i = (\mathbf{r}_i^T \mathbf{r}_i) / (\mathbf{r}_{i-1}^T \mathbf{r}_{i-1})$
    $\mathbf{v}_{i+1} = \mathbf{r}_i + \delta_i \mathbf{v}_i$                   ▷ Next search direction
end while
$d = \min\{d, g - m\}$                         ▷ Compute full rank posterior if $d > g - m$
$\mathbf{V}_{m+1:m+d} = \mathbf{0}_{n\times d}$                 ▷ Initialize approximate posterior 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
    $\eta_j = \mathbf{v}_j^T \mathbf{A}\mathbf{v}_j$
    $\gamma_j = (\mathbf{r}_{j-1}^T \mathbf{r}_{j-1}) / \eta_j$
    $\mathbf{V}(:, j) = \mathbf{v}_j / \sqrt{\eta_j}$                 ▷ Next column of $\mathbf{V}_{m+1,m+d}$
    $\mathbf{\Phi}(j, j) = \gamma_j \|\mathbf{r}_{j-1}\|_2^2$              ▷ Next diagonal element of $\mathbf{\Phi}_{m+1,m+d}$
    $\mathbf{r}_j = \mathbf{r}_{j-1} - \gamma_j \mathbf{A}\mathbf{v}_j$
    $\delta_j = (\mathbf{r}_j^T \mathbf{r}_j) / (\mathbf{r}_{j-1}^T \mathbf{r}_{j-1})$
    $\mathbf{v}_{j+1} = \mathbf{r}_j + \delta_j \mathbf{v}_j$                   ▷ Next un-normalized column of $\mathbf{V}_{m+1,m+d}$
end for
Output: $\mathbf{x}_m$, $\mathbf{V}_{m+1:m+d}$, $\mathbf{\Phi}_{m+1:m+d}$
```

2. S. Bartels, J. Cockayne, I. C. F. Ipsen, and P. Hennig. Probabilistic linear solvers: a unifying view. *Stat. Comput.*, 29(6):1249–1263, 2019.
3. J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer New York, 1985.
4. M. Berljafa and S. Güttel. Generalized rational Krylov decompositions with an application to rational approximation. *SIAM J. Matrix Anal. Appl.*, 36(2):894–916, 2015.
5. S. L. Campbell and C. D. Meyer. *Generalized inverses of linear transformations*, volume 56 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2009.
6. J. Cockayne, M. M. Graham, C. J. Oates, and T. J. Sullivan. Testing whether a learning procedure is calibrated, 2021. 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):1–34, 2021.
8. J. Cockayne, C. J. Oates, I. C. F. Ipsen, and M. Girolami. A Bayesian conjugate gradient method (with discussion). *Bayesian Anal.*, 14(3):937–1012, 2019. Includes 6 discussions and a rejoinder from the authors.
9. J. Cockayne, C. J. Oates, I. C. F. Ipsen, and M. Girolami. Supplementary material for 'A Bayesian conjugate-gradient method'. *Bayesian Anal.*, 2019.
10. J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami. Bayesian probabilistic numerical methods. *SIAM Rev.*, 61(4):756–789, 2019.
11. Vladimir Fanaskov. Uncertainty calibration for probabilistic projection methods. *Stat. Comput.*, 31(5):Paper No. 56, 17, 2021.
12. M. Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Math. Nachr.*, 147:185–203, 1990.
13. L. Giraud, J. Langou, and M. Rozloznik. The loss of orthogonality in the Gram-Schmidt orthogonalization process. *Comput. Math. Appl.*, 50(7):1069–1075, 2005.

14. Luc Giraud, Julien Langou, Miroslav Rozložník, and Jasper van den Eshof. Rounding error analysis of the classical Gram-Schmidt orthogonalization process. *Numer. Math.*, 101(1):87–100, 2005.
15. Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, 4th edition, 2013.
16. 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(3):225–248, 2020.
17. P. Hennig. Probabilistic interpretation of linear solvers. *SIAM J. Optim.*, 25(1):234–260, 2015.
18. P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Proc. R. Soc. A.*, 471(2179):20150142, 17, 2015.
19. Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436, 1952.
20. N. J. Higham. *Functions of matrices. Theory and computation.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.
21. R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.
22. G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, second edition, 2021.
23. H.-M. Kaltenbach. *A concise guide to statistics.* Springer Briefs in Statistics. Springer, Heidelberg, 2012.
24. D. Kressner, J. Latz, S. Massei, and E. Ullmann. Certified and fast computations with shallow covariance kernels, 2020. arXiv:200109187.
25. J. Liesen and Z. Strakos. *Krylov Subspace Methods: Principles and Analysis.* Oxford University Press, 2013.
26. A. M. Mathai and S. B. Provost. *Quadratic forms in random variables: Theory and applications.* Dekker, 1992.
27. R. J. Muirhead. *Aspects of multivariate statistical theory.* John Wiley & Sons, Inc., New York, 1982. Wiley Series in Probability and Mathematical Statistics.
28. J. Nocedal and S. J. Wright. *Numerical optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.
29. C. J. Oates and T. J. Sullivan. A modern retrospective on probabilistic numerics. *Stat. Comput.*, 29(6):1335–1351, 2019.
30. D. V. Ouellette. Schur complements and statistics. *Linear Algebra Appl.*, 36:187–295, 1981.
31. N. Petra, H. Zhu, G. Stadler, T.J.R. 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(211):889–903, 2012.
32. T. W. Reid, I. C. F. Ipsen, J. Cockayne, and C. J. Oates. BayesCG as an uncertainty aware version of CG, 2022.
33. S. M. Ross. *Introduction to probability models.* Academic Press, Inc., Boston, MA, ninth edition, 2007.
34. Y. Saad. *Iterative Methods for Sparse Linear Systems.* SIAM, 2003.
35. 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.*, page e2364, 2021.
36. 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.*, 13:56–80, 2002.
37. A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010.
38. C. Villani. *Optimal Transport, Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].* Springer-Verlag, Berlin, 2009.
39. J. Wenger and P. Hennig. Probabilistic linear solvers for machine learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6731–6742. Curran Associates, Inc., 2020.