Numerische Mathematik (2023) 155:239–288
https://doi.org/10.1007/s00211-023-01375-7

Numerische Mathematik

Statistical properties of BayesCG under the Krylov prior

Tim W. Reid[1] · **Ilse C. F. Ipsen**[1] · **Jon Cockayne**[2] · **Chris J. Oates**[3]

Received: 23 July 2022 / Revised: 10 April 2023 / Accepted: 14 September 2023 /
Published online: 12 October 2023

Abstract
We analyse the calibration of BayesCG under the Krylov prior. BayesCG is a probabilistic numeric extension of the Conjugate Gradient (CG) method for solving systems of linear equations with real symmetric positive definite coefficient matrix. In addition to the CG solution, BayesCG also returns a posterior distribution over the solution. In this context, a posterior distribution is said to be 'calibrated' if the CG error is well-described, in a precise distributional sense, by the posterior spread. Since it is known that BayesCG is not calibrated, we introduce two related weaker notions of calibration, whose departures from exact calibration can be quantified. Numerical experiments confirm that, under low-rank approximate Krylov posteriors, BayesCG is only slightly optimistic and exhibits the characteristics of a calibrated solver, and is computationally competitive with CG.

Mathematics Subject Classification 65F10 · 62F15 · 65F50 · 15 A10

1 Introduction

We present a rigorous analysis of the probabilistic numeric solver BayesCG under the Krylov prior [1, 2] for solving systems of linear equations

✉ Ilse C. F. Ipsen
ipsen@ncsu.edu

Tim W. Reid
twreid@alumni.ncsu.edu

Jon Cockayne
jon.cockayne@soton.ac.uk

Chris J. Oates
chris.oates@ncl.ac.uk

[1] Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA

[2] Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK

[3] School of Mathematics and Statistics, Newcastle University, Newcastle-upon-Tyne NE1 7RU, UK

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

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

Probabilistic numerics.
This area [3–5] 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 [6–10], whose computational resources are limited by the impracticality of running the solver to completion. The solver's premature termination leaves the user with residual uncertainty about the *exact solution* x_* of the linear system (1).

Probabilistic numeric linear solvers.
Probabilistic numeric extensions of Krylov space and stationary iterative methods [1, 2, 6, 11–14] model the 'epistemic uncertainty' in a quantity of interest, which can be the matrix inverse $\mathbf{A}^{-1}$ [11, 13, 14] or the solution $\mathbf{x}_*$ [1, 6, 11, 12]. 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 a two–dimensional linear system.

Calibration.
An important criterion of probabilistic solvers is the statistical quality of their posterior distributions, in terms of accurately quantifying the error. A solver is considered 'calibrated' if its posterior distributions accurately model the users's uncertainty about $\mathbf{x}_*$ [1, Section 6.1]. Examples 4.4 and 4.5 in Sect. 4 provide verbal and visual intuition for the meaning of 'calibration'.

It turns out that 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 [11, Section 6.4], [1, Section 6.1]. Previous efforts for improving calibration have focused on scaling the posterior covariances [1, Section 4.2], [12, Section 7], [14, Section 3]. Calibration, that is proper quantification of errors, is a must for probabilistic solvers to become reliable components at the base of computational pipelines.

BayesCG.
We analyze the calibration of BayesCG under the Krylov prior [1, 2]. BayesCG was introduced in [1] as a probabilistic numeric extension of the Conjugate Gradient (CG) method [15] for solving the linear system (1). The Krylov prior proposed in [2] makes BayesCG competitive with CG. The numerical properties of BayesCG under the Krylov prior are analysed in [2], while here we analyse its statistical properties.

1.1 Contributions and overview

Our overall conclusion is that BayesCG under the Krylov prior, although not calibrated in the strict sense, has the desirable properties of a calibrated solver. Under the efficient

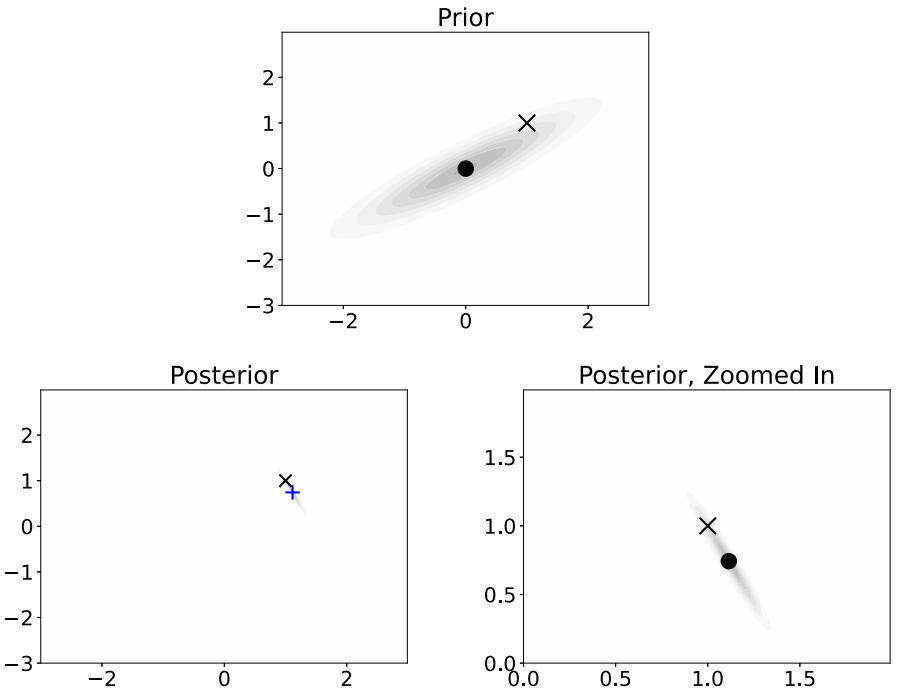

Fig. 1 Prior and posterior distributions for a linear system (1) with $n = 2$. Top plot: prior distribution. Bottom plots: posterior distributions, where the bottom right is a zoomed in version of the bottom left. The gray shaded contours represent the areas in which the distributions are concentrated,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

approximate Krylov posteriors, BayesCG is competitive with CG, and only slightly optimistic.

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

Approximate Krylov posteriors (Sect. 3).
We define the **A**-Wasserstein distance (Definition 3.3, Theorem 3.4); determine the error between Krylov posteriors and their low-rank approximations in the **A**-Wasserstein distance (Theorem 3.5); and present a statistical interpretation of a Krylov prior as an empirical Bayesian procedure (Theorem 3.7, Remark 3.8).

Calibration (Sect. 4).
We review the strict notion of calibration for probabilistic solvers (Definition 4.1, Lemma 4.2), and show that it does not apply to BayesCG under the Krylov prior (Remark 4.6).

We relax the strict notion and propose as an alternative assessment two test statistics that are necessary but not sufficient for calibration: the Z-statistic (Theorem 4.9) and the new S-statistic (Theorem 4.15, Definition 4.17). We present implementations for both statistics (Algorithms 4.1 and 4.2); and apply a Kolmogorov–Smirnov statistic (Definition 4.11) 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 4.13), while the S-statistic indicates that it is not calibrated (Sect. 4.3.4).

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

We assess calibration with the Z- and S-statistics for BayesCG with random search directions (Algorithms B.1 and B.2); BayesCG under the inverse prior (Algorithms 2.1 and B.3); and BayesCG under the Krylov prior with full posteriors (Algorithm B.4) and approximate posteriors (Algorithm B.5).

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, while BayesCG under rank-50 approximate posteriors does almost as well, being slightly optimistic.

Future research (Sect. 6).
We conclude with a few thoughts on possible future research directions.

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 identity matrix is $\mathbf{I}_n \in \mathbb{R}^{n\times n}$, or just $\mathbf{I}$ if the dimension is clear. The Moore–Penrose inverse of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is $\mathbf{A}^\dagger$, and the matrix square root is $\mathbf{A}^{1/2}$ [16, Chapter 6].

Probability distributions are represented in lowercase Greek letters, such as μ_m; and random variables in uppercase Roman, such as X. A random variable X with 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^2_f.

2 Review of existing work

We review BayesCG (Sect. 2.1), the ideal Krylov prior (Sect. 2.2), and practical approximations for Krylov posteriors (Sect. 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 2.1), and present a pseudo code for BayesCG (Algorithm 2.1).

Given an initial guess $\mathbf{x}_0$, BayesCG [1] solves symmetric positive definite linear systems (1) by computing iterates $\mathbf{x}_m$ that converge to the solution $\mathbf{x}_*$. In addi-

tion, 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 2.1 *(*[1, Proposition 1], [2, 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* $\operatorname{rank}(\mathbf{S}_m) = m$ *and* $\mathbf{\Lambda}_m \equiv \mathbf{S}_m^T\mathbf{A}\mathbf{\Sigma}_0\mathbf{A}\mathbf{S}_m$ *is non-singular. Then, the BayesCG posterior* $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$ *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 2.1 represents the iterative computation of the posteriors from [1, Propositions 6 and 7], [2, Theorem 2.7]. To illustrate the resemblance of BayesCG and the Conjugate Gradient method, we present the most common implementation of CG in Algorithm 2.2.

BayesCG (Algorithm 2.1) 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 [1, Section 2.3], thus easy to invert.
2. They form a basis for the Krylov space [17, Proposition S4]

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

Remark 2.2 The additional requirement $\mathbf{x}_* - \mathbf{x}_0 \in \operatorname{range}(\mathbf{\Sigma}_0)$ in Algorithm 2.1 ensures the nonsingularity of $\mathbf{\Lambda}_m$ as required by Theorem 2.1, even for singular prior covariance matrices $\mathbf{\Sigma}_0$ [2, Theorem 2.7].

2.2 The ideal Krylov Prior

After defining the Krylov space of maximal dimension (Definition 2.3), we review the ideal but impractical Krylov prior (Definition 2.4), and discuss its construction (Lemma 2.6) and properties (Theorem 2.7).

Definition 2.3 The Krylov space of *maximal dimension* for Algorithm 2.2 is

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

Algorithm 2.1 BayesCG [2, Algorithm 2.1]

Input: spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, prior $\mu_0 = \mathcal{N}(\mathbf{x}_0, \boldsymbol{\Sigma}_0)$ ▷ with $\mathbf{x}_* - \mathbf{x}_0 \in \text{range}(\boldsymbol{\Sigma}_0)$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ ▷ 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
  $\alpha_m = \left(\mathbf{r}_{m-1}^T \mathbf{r}_{m-1}\right) / \left(\mathbf{s}_m^T \mathbf{A}\boldsymbol{\Sigma}_0 \mathbf{A}\mathbf{s}_m\right)$
  $\mathbf{x}_m = \mathbf{x}_{m-1} + \alpha_m \boldsymbol{\Sigma}_0 \mathbf{A}\mathbf{s}_m$ ▷ Next posterior mean
  $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}_{m-1} - \boldsymbol{\Sigma}_0 \mathbf{A}\mathbf{s}_m \left(\boldsymbol{\Sigma}_0 \mathbf{A}\mathbf{s}_m\right)^T / (\mathbf{s}_m^T \mathbf{A}\boldsymbol{\Sigma}_0 \mathbf{A}\mathbf{s}_m)$ ▷ Next posterior covariance
  $\mathbf{r}_m = \mathbf{r}_{m-1} - \alpha_m \mathbf{A}\boldsymbol{\Sigma}_0 \mathbf{A}\mathbf{s}_m$ ▷ 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 $\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}$-orthogonal search direction
end while
Output: $\mu_m = \mathcal{N}(\mathbf{x}_m, \boldsymbol{\Sigma}_m)$ ▷ Final posterior

Algorithm 2.2 Conjugate Gradient Method (CG) [15, 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 $\mathbf{x}_*$

Here $g \leq n$ represents the *grade* of $\mathbf{r}_0$ with respect to $\mathbf{A} \in \mathbb{R}^{n\times n}$ [18, Definition 4.2.1], or the *invariance index* for $(\mathbf{A}, \mathbf{r}_0)$ [19, 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 2.4 [2, Definition 3.1] The *ideal Krylov prior* for $\mathbf{A}\mathbf{x}_* = \mathbf{b}$ is $\eta_0 \equiv \mathcal{N}(\mathbf{x}_0, \boldsymbol{\Gamma}_0)$ with symmetric positive semi-definite covariance

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

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 $\mathbf{\Phi} \equiv \operatorname{diag}\left(\phi_1 \cdots \phi_g\right) \in \mathbb{R}^{g\times g}$ has diagonal elements

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

Remark 2.5 The Krylov prior covariance satisfies the requirement of Algorithm 2.1 that $\mathbf{x}_* - \mathbf{x}_0 \in \operatorname{range}(\mathbf{\Gamma}_0)$. This follows from [18, 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 2.6 [2, Remark SM2.1] *The Krylov prior $\mathbf{\Gamma}_0$ can be constructed from quantities computed by CG (Algorithm 2.2),*

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

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

$$\begin{aligned} \mathbf{V}_{i:j} &\equiv \left[\mathbf{v}_i \cdots \mathbf{v}_j\right] \\ \mathbf{\Phi}_{i:j} &\equiv \operatorname{diag}\left(\phi_i \cdots \phi_j\right), \qquad 1 \le i \le j \le g, \end{aligned} \tag{6}$$

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

Under suitable assumptions, BayesCG (Algorithm 2.1) produces the same iterates as CG (Algorithm 2.2).

Theorem 2.7 [2, Theorem 3.3] *Let $\mathbf{x}_0$ be the starting vector for CG (Algorithm 2.2). Then BayesCG (Algorithm 2.1) 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*

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

are identical to the iterates in CG (Algorithm 2.2), and whose covariance matrices

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

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 2.1 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 2.8) by running d additional iterations. The corresponding CG-based implementation of BayesCG under approximate Krylov posteriors is relegated to Algorithm B.5 in "Appendix B".

Definition 2.8 [2, 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 [2, Section 3.4],

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

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 [20, 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 $\operatorname{trace}(\mathbf{A}\widehat{\mathbf{\Gamma}}_m) \approx \|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$ [20, Section 4].

3 Approximate Krylov posteriors

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

3.1 Error in approximate Krylov posteriors

We review the p-Wasserstein distance (Definition 3.1), extend the 2-Wasserstein distance to the $\mathbf{A}$-Wasserstein distance weighted by a symmetric positive definite matrix $\mathbf{A}$ (Theorem 3.4), and derive the $\mathbf{A}$-Wasserstein distance between approximate and full Krylov posteriors (Theorem 3.5).

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

Definition 3.1 [21, Definition 2.1], [22, Definition 6.1] The p-Wasserstein distance between probability distributions μ and ν on $\mathbb{R}^n$ is

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

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 3.2 [23, 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 3.3 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 3.2.

Theorem 3.4 *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 A.1 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

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

At last, apply Lemma 3.2 and the linearity of the trace. □

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

Theorem 3.5 *Let $\eta_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Gamma}_m)$ be a Krylov posterior from Theorem 2.7, 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 2.8. The* **A***-Wasserstein distance between η_m and $\widehat{\eta}_m$ is*

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

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

$$\widetilde{\mathbf{V}} \equiv \left[\mathbf{V}_{m+1:g} \;\; \mathbf{V}_m^{\perp} \right] \in \mathbb{R}^{n \times n}$$

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,

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

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

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

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 A.3 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

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

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

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

□

Theorem 3.5 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 [20, 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 3.6 The distance in Theorem 3.5 is a special case of the 2-Wasserstein distance between two distributions whose covariance matrices commute [21, Corollary 2.4].

To see this, consider the $\mathbf{A}$-Wasserstein distance between η_m and $\widehat{\eta}_m$ from Theorem 3.5, 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 3.7), and elucidate the connection between the random variables and the deterministic solution (Remark 3.8).

An *empirical Bayesian procedure* estimates the prior from data [24, 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 3.7 *Under the assumptions of Theorem 2.7, 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 2.1 in [17, Proof of Proposition 1], we exploit the stability and conjugacy of Gaussian distributions in Lemmas A.1 and A.2 in "Appendix A".

Prior.
Lemma A.1 implies that X in (18) is a Gaussian random variable with mean and covariance

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

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

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

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

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

Analogously,

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

From [25, Theorem 6.20] follows the expression for the posterior mean,

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

and for the posterior covariance

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

where

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

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 2.7, and the posterior covariance $\widehat{\mathbf{\Gamma}}_m$ is equal to the rank-d approximate Krylov posterior in Definition 2.8. □

Remark 3.8 The random variable X in Theorem 3.7 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, $1 \leq m \leq n$ 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 (Sect. 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 (Sect. 4.2) and the S-statistic (Sect. 4.3).

4.1 Calibration

We review the definition of calibration for probabilistic linear solvers (Definition 4.1, Lemma 4.2), discuss the difference between certain random variables (Remark 4.3), present two illustrations (Examples 4.4 and 4.5), and explain why this notion of calibration does not apply to BayesCG under the Krylov prior (Remark 4.6).

Informally, a probabilistic numerical solver is calibrated if its posterior distributions accurately model the uncertainty in the solution [6, 26].

Definition 4.1 [6, 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 $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$, $1 \leq m \leq n$. 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

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

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

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

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 4.2 *Under the conditions of Definition 4.1, a probabilistic linear solver is calibrated, if and only if*

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

where the notation $\mathbf{x}_m(X_*)$ highlights the dependence of $\mathbf{x}_m$ on X_* through its dependence on the projections $\mathbf{S}_m^T\mathbf{B} = \mathbf{S}_m^T\mathbf{A}X_*$, while $\boldsymbol{\Sigma}_m$ is independent of X_* as assumed in Definition 4.1.

Proof Let $\boldsymbol{\Sigma}_m = \mathbf{U}\mathbf{D}\mathbf{U}^T$ 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 4.1. Multiply the first equation of (21) on the left by $\mathbf{D}_m^{1/2} = (\mathbf{U}_m^T\boldsymbol{\Sigma}_m\mathbf{U}_m)^{1/2}$,

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

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

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

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

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

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

Since the covariance matrix $\boldsymbol{\Sigma}_m$ 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 $X_* = \mathbf{x}_m \in \ker(\boldsymbol{\Sigma}_m)$. Hence, this subspace must be $\ker(\boldsymbol{\Sigma}_m)$, and any remaining uncertainty about X_* lies in $\operatorname{range}(\boldsymbol{\Sigma}_m)$.

Remark 4.3 We discuss the difference between the random variable X_* in Definition 4.1 and the random variable X in Theorem 3.7.

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 4.2 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 4.1.

Example 4.4 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 4.5 This is the visual equivalent of Example 4.4, where Carol receives the images in Fig. 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.
- 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 4.6 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 4.1 and if the posterior means and covariances depend on the solution X_*, then the posterior means and covariances are also random variables.

Definition 4.1 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 [2], it does not apply to BayesCG under the Krylov prior, because Krylov posterior covariances depend nonlinearly on the right-hand side. In Sects. 4.2 and 4.3, we present a remedy for BayesCG in the form of test statistics that are motivated by Definition 4.1 and Lemma 4.2.

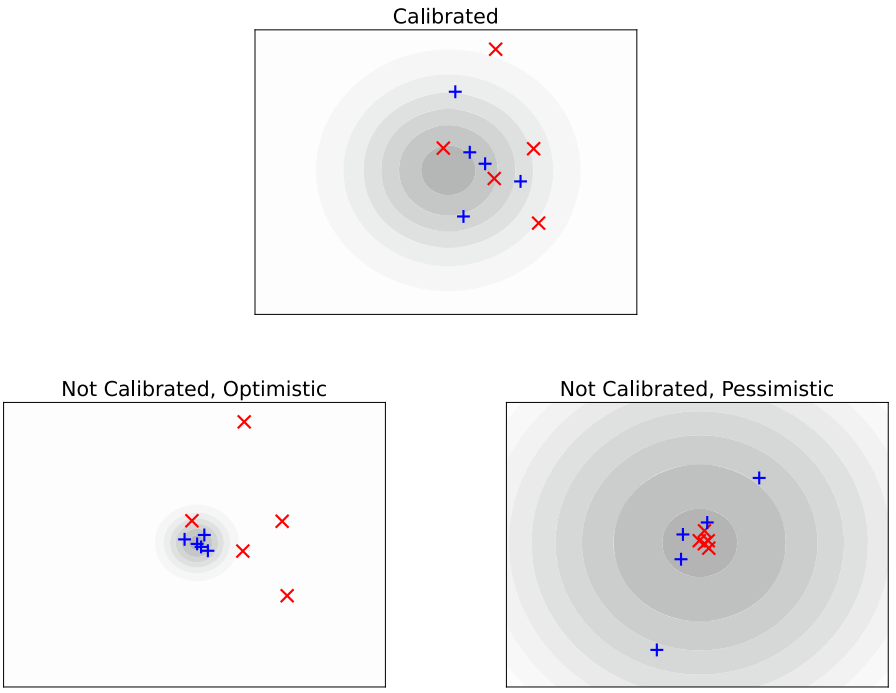

Fig. 2 Posterior distributions and solutions from three different probabilistic solvers: calibrated (top), optimistic (bottom left), and pessimistic (bottom right). The gray contours represent the posterior distributions, the red symbols "×" the solutions, and the blue symbols "+" samples from the posterior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

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 (Sect. 4.2.1), and apply it to BayesCG under the Krylov prior (Sect. 4.2.2).

4.2.1 Review of the *Z*-statistic

We review the Z-statistic (Definition 4.7), and the chi-square distribution (Definition 4.8), which links the Z-statistic to calibration (Theorem 4.9). Then we discuss how to generate samples from the Z-statistic (Algorithm 4.1), how to use the samples for the assessment of calibration (Remark 4.10), and then present the Kolmogorov–Smirnov statistic as a computationally inexpensive estimate for the difference between two general distributions (Definition 4.11).

The Z-statistic was introduced in [1, Section 6.1] as a means to assess the calibration of BayesCG, and has subsequently been applied to other probabilistic linear solvers [11, Section 6.4], [12, Section 9].

Definition 4.7 [1, 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 $X_* \sim \mu_0 \equiv \mathcal{N}(\mathbf{x}_0, \boldsymbol{\Sigma}_0)$. Let $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \boldsymbol{\Sigma}_m)$, $1 \leq m \leq n$, be the posterior distributions from a probabilistic solver under the prior μ_0 applied to $\mathbf{A}X_* = B$. The Z-statistic is

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

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

Definition 4.8 [27, 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 4.9 [17, 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 $X_* \sim \mu_0 \equiv \mathcal{N}(\mathbf{x}_0, \boldsymbol{\Sigma}_0)$. Assume that a probabilistic solver under the prior μ_0 applied to $\mathbf{A}X_* = B$ computes the posteriors $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \boldsymbol{\Sigma}_m)$ with $\operatorname{rank}(\boldsymbol{\Sigma}_m) = p_m$, $1 \leq m \leq n$.*

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 $M_m \equiv (\boldsymbol{\Sigma}_m^{\dagger})^{1/2}(X_* - \mathbf{x}_m)$. Lemma 4.2 implies that a calibrated solver produces posteriors with

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

With the eigenvector matrix $\mathbf{U}_m \in \mathbb{R}^{n\times p_m}$ as in Definition 4.1, Lemma A.1 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 A.7 implies $\mathrm{Z}_m(X_*) = (M_m^T M_m) \sim \chi_{p_m}^2$. □

Theorem 4.9 implies that BayesCG is calibrated if the Z-statistic is distributed according to a chi-squared distribution with $p_m = \operatorname{rank}(\boldsymbol{\Sigma}_0) - m$ 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 4.1 samples N_{test} solutions $\mathbf{x}_*$ from the prior distribution μ_0,

defines the linear systems $\mathbf{b} \equiv \mathbf{A}\mathbf{x}_*$, runs m iterations of the `solver` on $\mathbf{b} \equiv \mathbf{A}\mathbf{x}_*$, and computes $Z_m(\mathbf{x}_*)$ in (22).

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

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

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

Algorithm 4.1 Sampling from the Z-statistic

1: **Input:** spd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mu_0 = \mathcal{N}(\mathbf{x}_0, \boldsymbol{\Sigma}_0)$, `solver`, m, N_{test}
2: **for** $i = 1 : N_{\text{test}}$ **do**
3: Sample $\mathbf{x}_*$ from prior distribution μ_0 ▷ Sample solution vector
4: $\mathbf{b} = \mathbf{A}\mathbf{x}_*$ ▷ Define test problem
5: $[\mathbf{x}_m, \boldsymbol{\Sigma}_m] = \texttt{solver}(\mathbf{A}, \mathbf{b}, \mu_0, m)$ ▷ Compute posterior $\mu_m \equiv \mathcal{N}(\mathbf{x}_m, \boldsymbol{\Sigma}_m)$
6: $z_i = (\mathbf{x}_* - \mathbf{x}_m)^T \boldsymbol{\Sigma}_m^\dagger (\mathbf{x}_* - \mathbf{x}_m)$ ▷ Compute Z-statistic sample
7: **end for**
8: **Output:** Z-statistic samples z_i, $1 \le i \le N_{test}$.

Remark 4.10 We assess calibration of the `solver` by comparing the Z-statistic samples z_i from Algorithm 4.1 to the chi-squared distribution $\chi^2_{p_m}$ with $p_m \equiv \text{rank}(\boldsymbol{\Sigma}_0) - m$ degrees of freedom, based on the following criteria from [1, 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 [1, Section 6.1] and [11, Section 6.4], the Z-statistic samples and the predicted chi-squared distribution are compared visually. In Sect. 5, we make an additional quantitative comparison with the Kolmogorov–Smirnov test to estimate the difference between two probability distributions.

Definition 4.11 [28, 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 $0 \le KS(\mu, \nu) \le 1$.

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 3.1, the Kolmogorov–Smirnov statistic in Definition 4.11 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 4.12). Then we show that the Z-statistic for the full Krylov posteriors has the same *mean* as the corresponding chi-squared distribution (Theorem 4.13), but its *distribution* is different. Therefore the Z-statistic is inconclusive about the calibration of BayesCG under the Krylov prior (Remark 4.14).

Lemma 4.12 *In Definition 2.8, 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*

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

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 A.5 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 A.5 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 4.13 *Under the assumptions of Theorem 2.7, 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*

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

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

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

from Theorem 2.7 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 [18, Equation (5.6.5)], which relates the CG error to the search directions and step sizes of the remaining iterations.

With Lemma 4.12, this implies for the Z-statistic in Theorem 4.9

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

In other words,

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

where the last inequality follows from $\phi_j = (\mathbf{v}_j^T\mathbf{r}_0)^2$ in Definition 2.4. □

Remark 4.14 On the one hand, Theorem 4.13 shows that BayesCG is not calibrated under the Krylov prior, since it is distributed according to a Dirac distribution at $g - m$ rather than following a χ^2_{g-m} distribution. On the other hand, the assessment of calibration by means of the Z-statistics is not well-motivated for the Krylov prior.

In our empirical Bayesian construction, the prior depends upon the residual $\mathbf{r}_0$, so any linear system $\mathbf{A}\mathbf{x}_* = \mathbf{b}$ yields a different prior distribution. Arguments based on randomising the solution $\mathbf{x}_*$ according to the prior are therefore circular. This motivates the assessment of calibration by means of a weaker criterion that avoids this circular argument, which we will address in the next section.

Furthermore,Theorem 4.13 shows that the value of the Z-statistic is equal to the *mean* of the ideal χ^2_{g-m} distribution, suggesting that there is reason to believe that an appropriate notion of calibratedness might show that BayesCG is calibrated under the Krylov prior.

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 (Sect. 4.3.1), we define the S-statistic (Sect. 4.3.2), compare the S-statistic to the Z-statistic (Sect. 4.3.3), and then apply the S-statistic to BayesCG under the Krylov prior (Sect. 4.3.4).

4.3.1 Calibration and error estimation

We establish a relation between the error of the posterior means and the trace of posterior covariances (Theorem 4.15).

Theorem 4.15 *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

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

Proof For a calibrated solver Lemma 4.2 implies that $X_* \sim \mu_m$. Then apply Lemma A.8 from "Appendix A" to the error $\|X_* - \mathbf{x}_m\|_{\mathbf{A}}^2$. □

For a calibrated solver, Theorem 4.15 implies that the equality $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2 = \operatorname{trace}(\mathbf{A}\mathbf{\Sigma}_m)$ 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 4.15 relates the errors of a calibrated solver to the area in which its posteriors are concentrated.

Remark 4.16 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 [29, Section 12.2.1].

In analogy to viewing the $\mathbf{A}$-norm as the 2-norm weighted by $\mathbf{A}$, we can view $\operatorname{trace}(\mathbf{A}\mathbf{\Sigma}_m)$ as the trace of $\mathbf{\Sigma}_m$ weighted by $\mathbf{A}$. Theorem 4.15 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 4.17), present an algorithm for generating samples from the S-statistic (Algorithm 4.2), and discuss their use for assessing calibration of solvers (Remark 4.18).

The S-statistic represents a necessary condition for calibration, as established in Theorem 4.15.

Definition 4.17 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 4.15 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 4.2 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).

Algorithm 4.2 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_{test}
for $i = 1 : N_{\mathrm{test}}$ **do**
  Sample $\mathbf{x}_*$ from prior distribution μ_0 ▷ Sample solution vector
  $\mathbf{b} = \mathbf{A}\mathbf{x}_*$ ▷ Define test problem
  $[\mathbf{x}_m, \mathbf{\Sigma}_m] = \mathtt{solver}(\mathbf{A}, \mathbf{b}, \mu_0, m)$ ▷ 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
$h = \frac{1}{N_{test}} \sum_{i=1}^{N_{\mathrm{test}}} s_i$ ▷ Compute empirical mean of S-statistic samples
Output: S-statistic samples s_i and traces t_i; S-statistic mean h

Remark 4.18 We assess calibration of the solver by comparing the S-statistic samples s_i from Algorithm 4.2 to the traces t_i, $1 \leq i \leq N_{test}$. The following criteria are based on Theorem 4.15 and Remark 4.16.

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.

4.3.3 Comparison of the *Z*- and *S*-statistics

Both, Z- and S-statistic represent necessary conditions for calibration in (27) and (29); and both measure the norm of the error $X_* - \mathbf{x}_m$: The Z-statistic in the $\boldsymbol{\Sigma}_m^\dagger$-pseudo norm (Definition 4.7), and the S-statistic in the $\mathbf{A}$-norm (Definition 4.17). Deeper down, though, the Z-statistic projects errors onto a single dimension (Theorem 4.9), 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 Fig. 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.

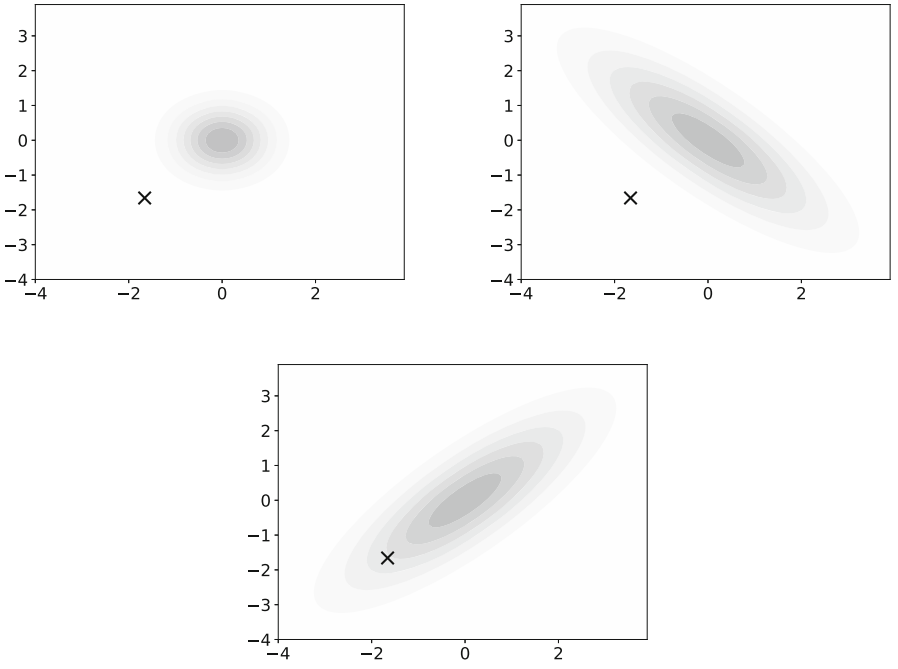

Fig. 3 Assessment of calibration from Z-statistic and S-statistic. The contour plots represent the posterior distributions, and the symbol '×' represents the solution. Top left: Both statistics decide that the solver is not calibrated. Top right: The S-statistic decides that the solver is calibrated,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

4.3.4 *S*-statistic for BayesCG under the Krylov prior

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

Calibration of BayesCG under full Krylov posteriors.
Theorem 2.7 implies that the *S*-statistic for any solution $\mathbf{x}_*$ is equal to

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

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 Krylov posteriors.
From (9) follows that $\mathrm{trace}(\mathbf{A}\widehat{\boldsymbol{\Gamma}}_m)$ 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 3.5. 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 (Sect. 5.1), we assess the calibration of three implementations of BayesCG: (i) BayesCG with random search directions (Sect. 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 (Sect. 5.3); and (iii) BayesCG under the Krylov prior (Sect. 5.4). Two additional experiments with BayesCG under approximate Krylov posteriors investigate: (i) the potential improvement in calibration when BayesCG returns $\mathbf{x}_{m+d}$ as the posterior mean instead of $\mathbf{x}_m$ (Sect. 5.5); and (ii) the effect of convergence on the calibration of BayesCG (Sect. 5.6). The experiments focus on the inverse and Krylov priors because their posterior means are identical to the CG iterates.

Conclusions from all 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 Krylov posteriors does almost as well, but 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 present the matrix $\mathbf{A}$ in the linear systems in Sects. 5.2–5.5 (Sect. 5.1.1); the setup of the Z- and S-statistic experiments (Sect. 5.1.2); and the three BayesCG implementations (Sect. 5.1.3).

5.1.1 The matrix A in the linear system

The symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ of dimension $n = 1806$ is a preconditioned version of the matrix BCSSTK14 from the Harwell–Boeing collection in [30]. Specifically, $\mathbf{B}$ is BCSSTK14, and

$$\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}, \qquad \text{where} \quad \mathbf{D} \equiv \operatorname{diag}\left(\mathbf{B}_{11} \cdots \mathbf{B}_{nn}\right).$$

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 in Algorithms 4.1 and 4.2, respectively. The calibration criteria for the Z-statistic are given in Remark 4.10, and for the S-statistic in Remark 4.18.

We sample from Gaussian distributions by exploiting their stability. According to Lemma A.1 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 in Definition 4.11 to the empirical cumulative distribution function of the Z-statistic samples and the analytical cumulative distribution function of the chi-squared distribution.

We choose the degree of freedom for the chi-squared distribution as the median numerical rank of the posterior covariances. To see why, note that the numerical rank of $\mathbf{\Sigma}_m$ can differ from

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

while the median rank represents an integer value equal to the rank of at least one of the posterior covariances.

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

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

where ε is machine epsilon and σ_i, $1 \le i \le n$, are the singular values of $\mathbf{\Sigma}_m$ [31, Section 5.4.1].

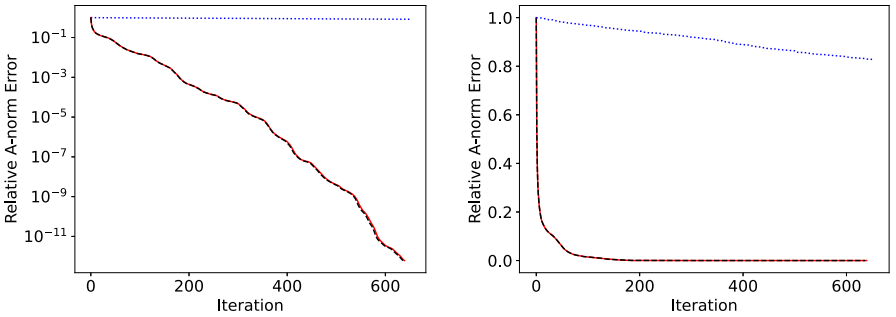

Fig. 4 Relative errors $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2 / \|\mathbf{x}_*\|_{\mathbf{A}}^2$ for BayesCG under the inverse prior (solid red curve), Krylov prior (dashed black curve), and with random search directions (dotted blue curve). Left panel: Vertical axis has a logarithmic scale. Right panel: Vertical axis has a linear scale

5.1.3 Three BayesCG implementations

We consider 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.
Algorithm B.2 in "Appendix B.2" computes posterior covariances that do not depend on the solution $\mathbf{x}_*$. To ensure that the search directions do not depend on $\mathbf{x}_*$ either [6, Section 1.1], we start with a random search direction $\mathbf{s}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ instead of the initial residual $\mathbf{r}_0 \equiv \mathbf{b}_0 - \mathbf{A}\mathbf{x}_0$. The prior is $\mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$.

This version of BayesCG is calibrated by design. However, it is also impractical due to its slow convergence, see Fig. 4, which takes n iterations. The random initial search direction $\mathbf{s}_1$ produces uninformative subspaces, so that BayesCG has to explore all of $\mathbb{R}^n$ before finding the solution.

BayesCG under the inverse prior $\mu_0 \equiv \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$.
Algorithm B.3 in "Appendix B.3" is a modified version of Algorithm 2.1 for general priors that maintains the posterior covariances in factored form.

BayesCG under the Krylov prior.
For full posteriors, the modified Lanczos solver Algorithm B.1 in "Appendix B.4" computes the full prior, followed by the direct computation of the posteriors in Algorithm B.2.

For approximate posteriors, Algorithm B.5 in "Appendix B.4" computes rank-d covariances at the same computational cost as $m + d$ iterations of CG.

For the Z- and S-statistic experiments, we cannot, as usual, sample the solutions $\mathbf{x}_*$ 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 are identical to the CG iterates [1, Section 3].

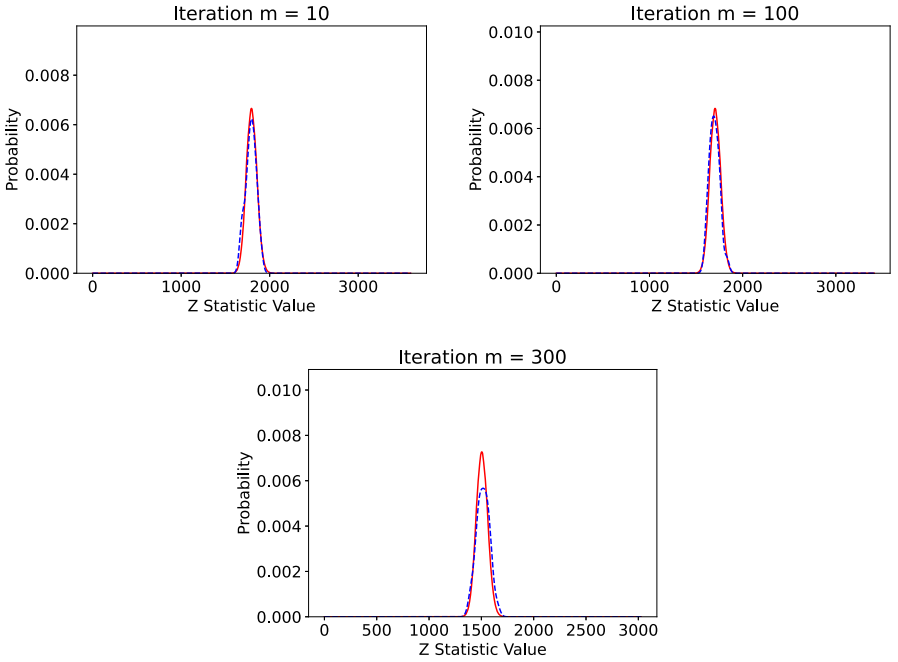

Fig. 5 Z-statistic samples for BayesCG with random search directions after $m = 10, 100, 300$ iterations. The solid red curve represents the chi-squared distribution and the dashed blue curve the Z-statistic samples (color figure online)

Table 1 This table corresponds to Fig. 5

Iteration	Z-stat mean	χ^2 mean	K–S statistic
10	1.79×10^3	1.8×10^3	7.91×10^{-2}
100	1.7×10^3	1.71×10^3	0.116
300	1.51×10^3	1.51×10^3	0.13

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

5.2 BayesCG with random search directions

By design, BayesCG with random search directions is a calibrated solver. Its 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 Theorem 4.9 and Definition 4.17.

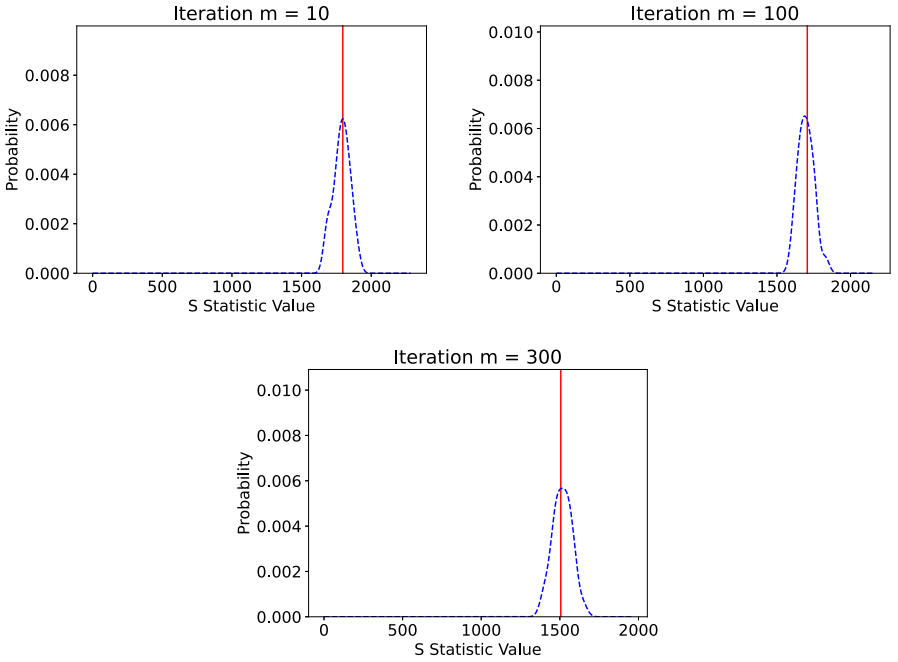

Fig. 6 S-statistic samples and traces for BayesCG with random search directions after $m = 10, 100, 300$ iterations. The solid red curve represents the traces and the dashed blue curve the S-statistic samples (color figure online)

Table 2 This table corresponds to Fig. 6

Iteration	S-stat mean	Trace mean	Trace standard deviation
10	1.79×10^3	1.8×10^3	4.83×10^{-12}
100	1.7×10^3	1.71×10^3	3.52×10^{-12}
300	1.52×10^3	1.51×10^3	2.43×10^{-12}

For BayesCG with random search directions, it shows the S-statistic sample means, the trace means, and the trace standard deviations

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

Figure 6 and Table 2.
The traces in Fig. 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.

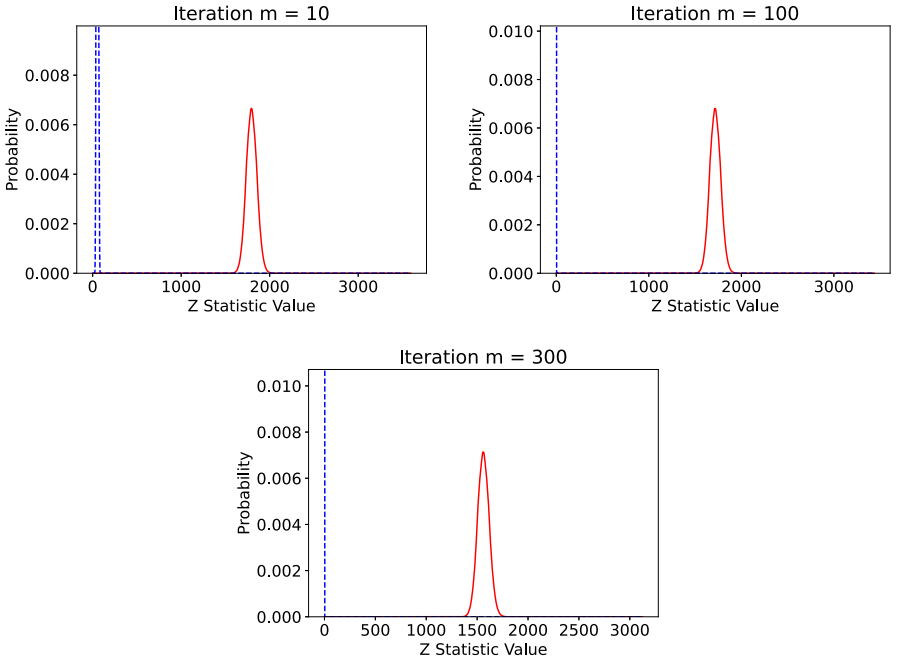

Fig. 7 Z-statistic samples for BayesCG under the inverse prior after $m = 10, 100, 300$ iterations. The solid red curve represents the chi-squared distribution and the dashed blue curve the Z-statistic samples (color figure online)

Table 3 This table corresponds to Fig. 7

Iteration	Z-stat mean	χ^2 mean	K–S statistic
10	52.5	1.8×10^3	1.0
100	0.509	1.72×10^3	1.0
300	7.61×10^{-6}	1.56×10^3	1.0

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

5.3 BayesCG under the inverse prior.

We assess calibration of BayesCG under the inverse prior $\mu_0 = \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$, by means of the Z- and S-statistics.

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 [1, Section 6.1].

Figure 7 and Table 3.
The Z-statistic samples in Fig. 7 are concentrated around smaller values than the predicted chi-squared distribution. The Kolmogorov–Smirnov statistics in Table 3

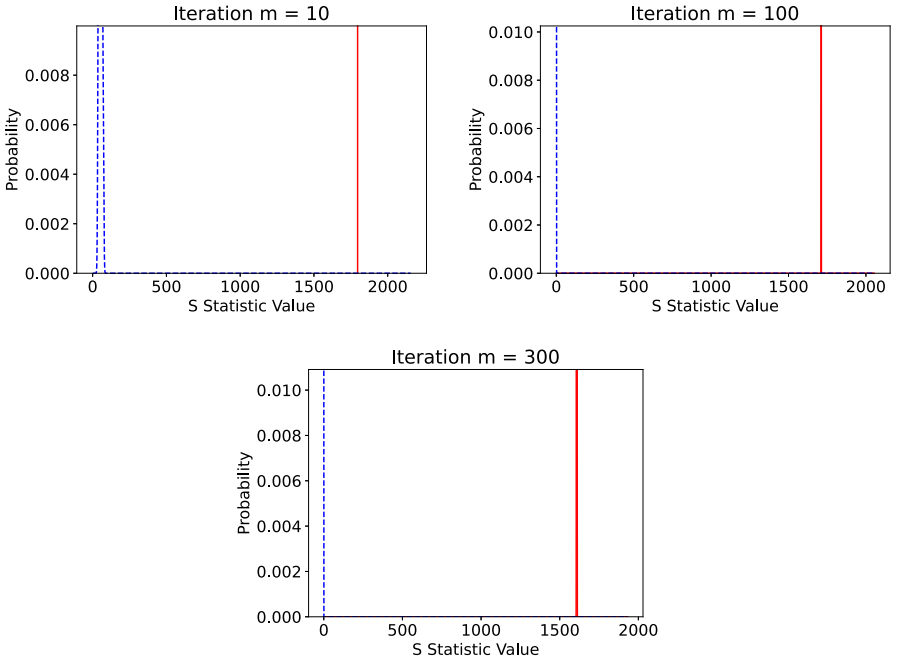

Fig. 8 S-statistic samples and traces for BayesCG under the inverse prior after $m = 10, 100, 300$ iterations. The solid red curve represents the traces and the dashed blue curve the S-statistic samples (color figure online)

Table 4 This table corresponds to Fig. 8

Iteration	S-stat mean	Trace mean	Trace standard deviation
10	53.0	1.8×10^3	5.22×10^{-12}
100	0.54	1.71×10^3	0.466
300	3.06×10^{-6}	1.61×10^3	1.19

For BayesCG under the inverse prior, it shows the S-statistic sample means, the trace means, and the trace standard deviations

are all equal to 1, indicating no overlap between Z-statistic samples and chi-squared distribution. The Z-statistic mean and χ^2 mean in Table 3 illustrate that Z-statistic samples move further away from the chi-squared distribution during the course of the iterations. Thus, BayesCG under the inverse prior is pessimistic, and the pessimism increases with the iteration count.

***Figure* 8 *and Table* 4.**
The S-statistic samples in Fig. 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 during the course of the iterations. 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.

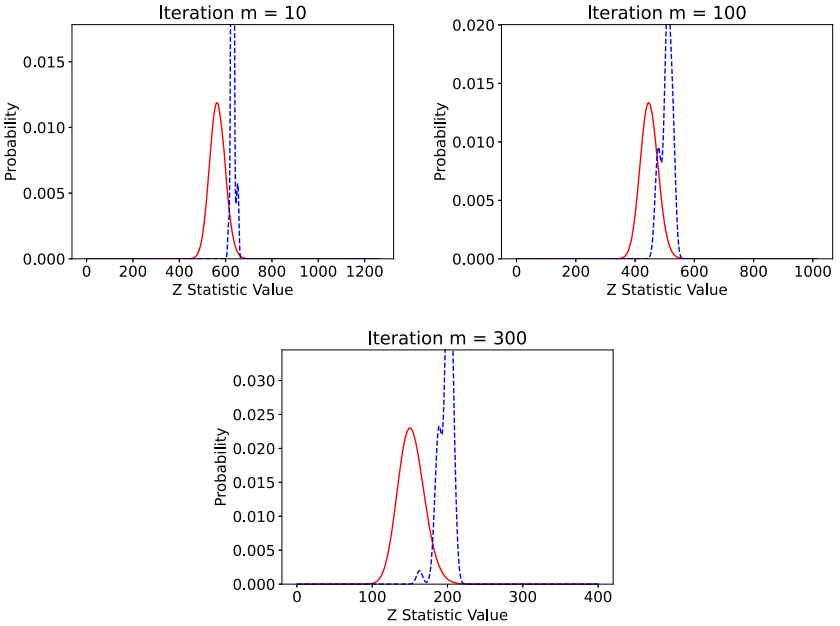

Fig. 9 Z-statistic samples for BayesCG under the Krylov prior and full posteriors at $m = 10, 100, 300$ iterations. The solid red curve represents the predicted chi-squared distribution and the dashed blue curve the Z-statistic samples (color figure online)

Table 5 This table corresponds to Fig. 9

Iteration	Z-stat mean	χ^2 mean	K–S statistic
10	631.0	565.0	0.925
100	509.0	448.0	0.774
300	201.0	152.0	0.93

For BayesCG under the Krylov prior and full Kryov 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 (Sect. 5.4.1), and rank-50 approximate posteriors (Sect. 5.4.2).

5.4.1 Full Krylov posteriors

We assess the calibration of BayesCG under full Krylov posteriors, with the help of the S- and Z-statisics.

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.

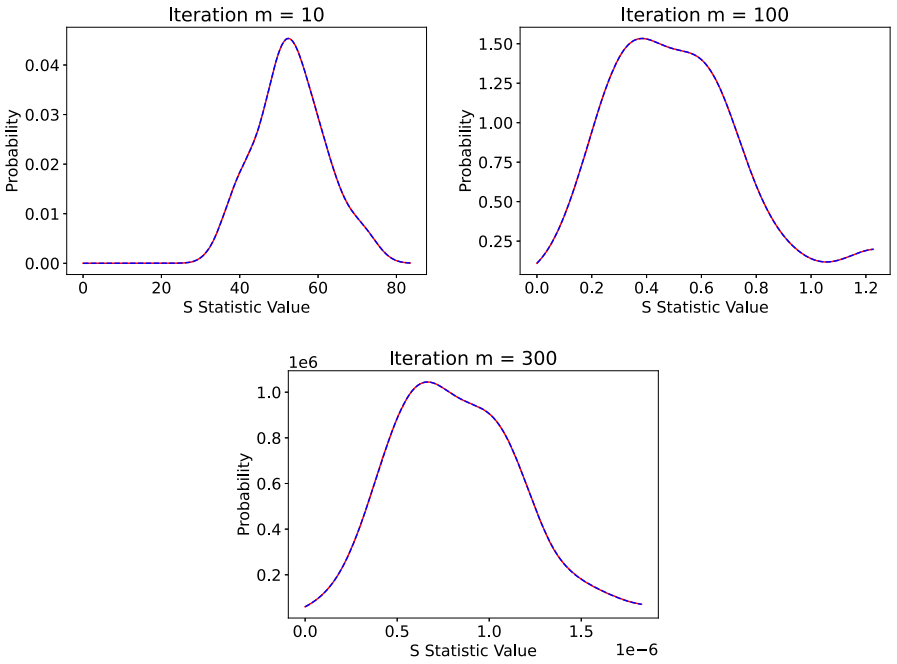

Fig. 10 S-statistic samples and traces for BayesCG under the Krylov prior and full Krylov posteriors at $m = 10, 100, 300$ iterations. The solid red curve represents the traces and the dashed blue curve the S-statistic samples (color figure online)

Table 6 This table corresponds to Fig. 10

Iteration	S-stat mean	Trace mean	Trace standard deviation
10	53.0	53.0	8.48
100	0.522	0.522	0.255
300	8.34×10^{-7}	8.34×10^{-7}	3.52×10^{-7}

For BayesCG under the Krylov prior and full Krylov posteriors, it shows the S-statistic sample means, the trace means, and the trace standard deviations

Figure 9 and Table 5.

The Z-statistic samples in Fig. 9 are concentrated at somewhat larger values than the predicted chi-squared distribution. The Kolmogorov–Smirnov statistics in Table 5 are around.75 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.

The numerical results in Table 5 differ from Theorem 4.13, which predicts Z-statistic samples equal to $g - m$. A possible reason might be that the computed rank of the Krylov prior from Algorithm B.1 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 of 10^{-12}, resulting in $\text{rank}(\mathbf{\Gamma}_0) < g$.

Table 7 This table corresponds to Fig. 11

Iteration	Z-stat mean	χ^2 mean	K–S statistic
10	314.0	50.0	1.0
100	340.0	50.0	1.0
300	164.0	50.0	1.0

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

***Figure* 10 *and Table* 6.**

The S-statistic samples in Fig. 10 match the traces extremely well, with Table 6 showing an agreement to 3 figures, as predicted in Sect. 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 Sect. 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 in the rigorous sense, but exhibits the performance of a calibrated solver.

5.4.2 Rank-50 approximate Krylov posteriors

We assess the calibration of BayesCG under rank-50 approximate Krylov posteriors, with the help of the S- and Z-statisics.

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 well calibrated as BayesCG with full Krylov posteriors. In contrast to the Z-statistic, the 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 Fig. 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* 12 *and Table* 8.**
The traces in Fig. 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 with the full Krylov posterior in Fig. 10 and Table 6 shows that the S-statistic samples and traces, respectively, for full and rank-50 posteriors are

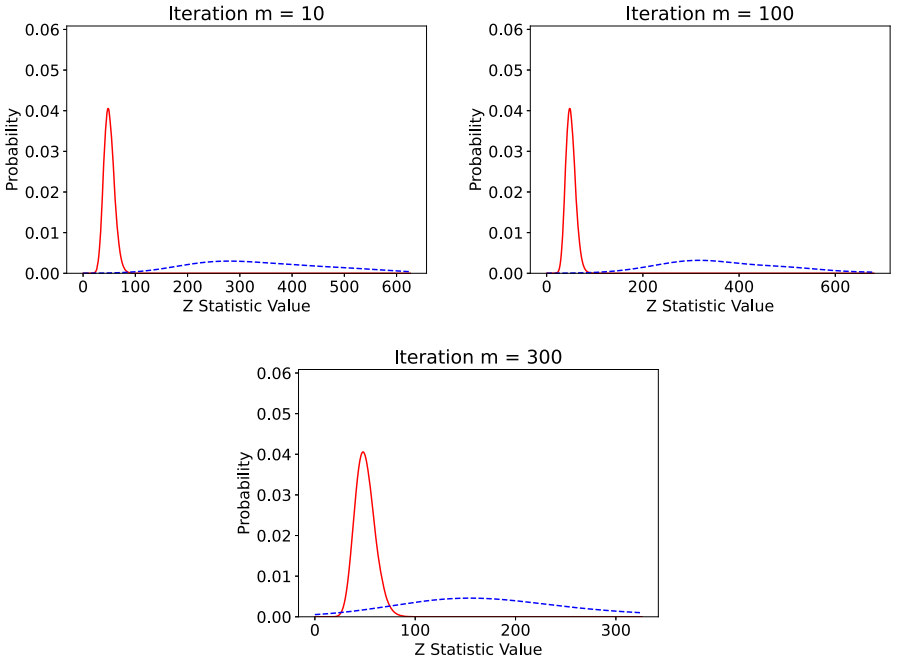

Fig. 11 Z-statistic samples for BayesCG under rank-50 approximate Krylov posteriors at $m = 10, 100, 300$ iterations. The solid red curve represents the predicted chi-squared distribution and the dashed blue curve the Z-statistic samples (color figure online)

Table 8 This table corresponds to Fig. 12

Iteration	S-stat mean	Trace mean	Trace standard deviation
10	53.0	49.8	7.96
100	0.538	0.489	0.251
300	2.97×10^{-6}	2.89×10^{-6}	1.51×10^{-6}

For BayesCG under rank-50 approximate Krylov posteriors, it shows the S-statistic sample means, trace means, and trace standard deviations

close. From the point of view of the S-statistic, BayesCG under approximate Krylov posteriors is somewhat optimistic, and close to a calibrated solver but not as close as BayesCG under full Krylov posteriors.

5.5 BayesCG under the Krylov prior with x_{m+d} as posterior mean

We investigate the effect of replacing the posterior mean $\mathbf{x}_m$ by $\mathbf{x}_{m+d}$ on the calibration of BayesCG under approximate priors, because Algorithm B.5 computes $\mathbf{x}_{m+d}$ at minimal additional cost. After comparing the choice of $\mathbf{x}_{m+d}$ as the posterior mean to a common practice in CG error estimation, we present the results of numerical experiments.

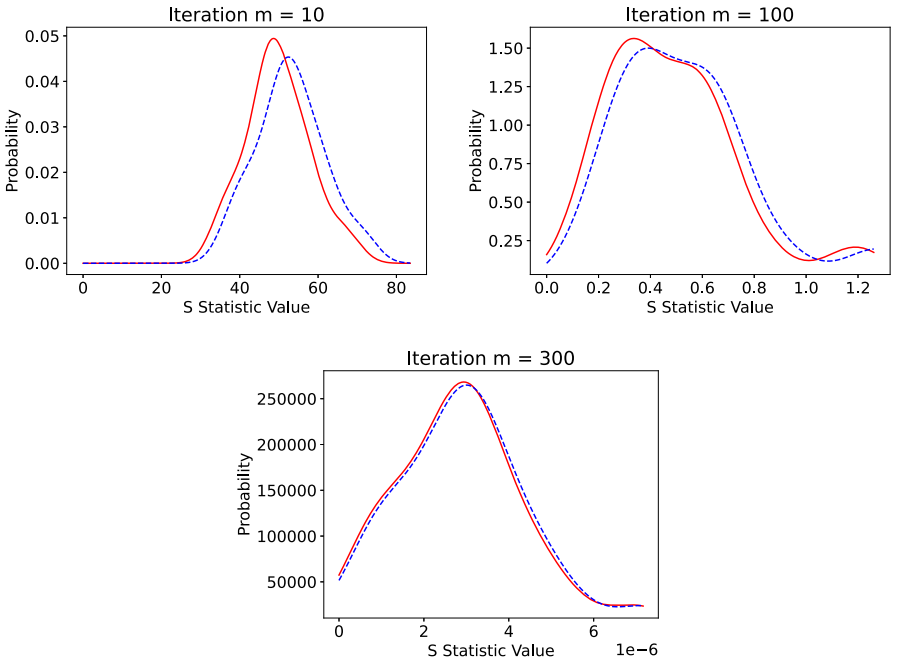

Fig. 12 S-statistic samples and traces for BayesCG under rank-50 approximate Krylov posteriors at $m = 10, 100, 300$ iterations. The solid red curve represents the traces and the dashed blue curve the S-statistic samples (color figure online)

Comparison with CG error estimation.
The strategy of 'bootstrapping' an error estimate with a less accurate method, so-called 'adaptive' step size control or local extrapolation, is popular in numerical analysis, i.e. the numerical solution of ODEs [32, Chapter II.4]. In the same vein, the output of $\mathbf{x}_{m+d}$ as the posterior mean is a common practice in CG error estimators that rely on information from $m + d$ iterations to estimate the error at iteration m [20, 33, 34]. There, the number of additional iterations d is usually referred to as the 'delay' [34, Section 1]. The corresponding algorithms return the error at iteration m, along with iterate x_{m+d} as an approximation to the solution.

Summary of the experiments below.
The Z- and S-statistic experiments suggest that BayesCG becomes more pessimistic when $\mathbf{x}_{m+d}$ replaces $\mathbf{x}_m$ as the posterior mean. Specifically, the Z-statistic indicates that BayesCG is less optimistic while the S-statistic indicates that it is more pessimistic.

***Figure* 13 *and Table* 9.**
The Z-statistic samples are concentrated around larger values than the predicted chi-square distribution, indicating that BayesCG is optimistic. At iteration $m = 300$, Table 9 lists Kolmogorov–Smirnov statistic values less than 1, implying some overlap between the Z-statistic and chi-square distributions. A comparison with Fig. 11 and Table 7 shows that the Z-statistic samples are closer to the chi-square distribution

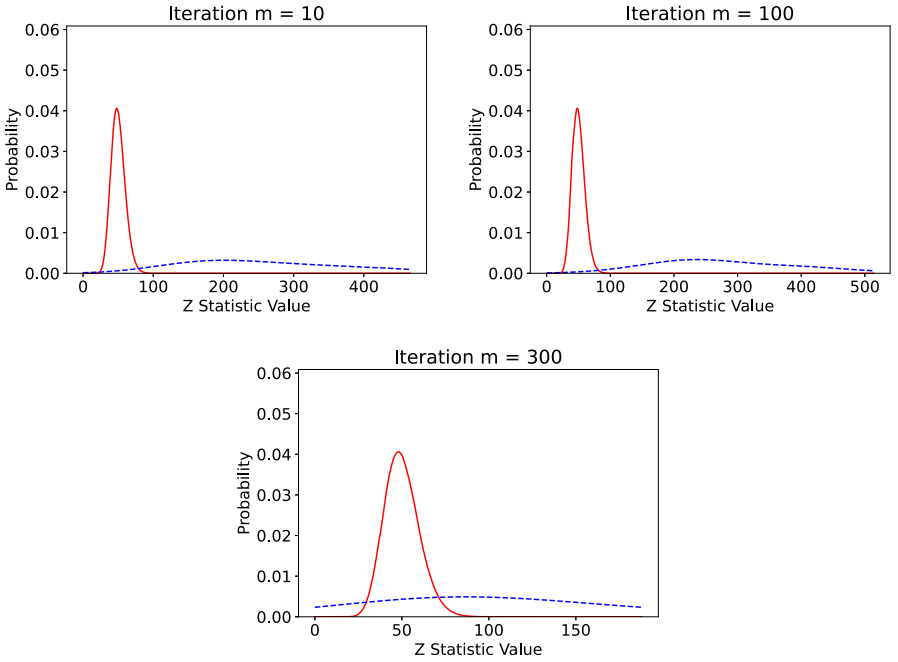

Fig. 13 Z-statistic samples for BayesCG under rank-50 approximate Krylov posteriors with $\mathbf{x}_{m+d}$ as the posterior mean at $m = 10, 100, 300$ iterations. The solid red curve represents the predicted chi-squared distribution and the dashed blue curve the Z-statistic samples (color figure online)

Table 9 This table corresponds to Fig. 13

Iteration	Z-stat mean	χ^2 mean	K–S statistic
10	233.0	50.0	1.0
100	258.0	50.0	0.999
300	94.4	50.0	0.692

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

when $\mathbf{x}_{m+d}$ replaces $\mathbf{x}_m$ as the posterior mean. Therefore, BayesCG with $\mathbf{x}_{m+d}$ as the posterior mean, looks less optimistic, thus better calibrated.

This is different than what we expected. In exact arithmetic, the Z-statistic samples are all zero, pointing to extreme pessimism of BayesCG. Intuitively, one might suspect that $\mathbf{x}_*$ is completely outside the support of $\mathcal{N}(\mathbf{x}_{m+d}, \widehat{\mathbf{\Gamma}}_m)$, which we justify below by showing that $\mathbf{x}_* - \mathbf{x}_{m+d} \in \ker(\widehat{\mathbf{\Gamma}}_m)$.

Lemma 5.1 *If BayesCG under the approximate Krylov posterior outputs* $\mathbf{x}_{m+d}$ *as the posterior mean, then* $\mathbf{x}_* - \mathbf{x}_{m+d} \in \ker(\widehat{\mathbf{\Gamma}}_m)$ *and* $\mathrm{Z}(\mathbf{x}_*) = 0$.

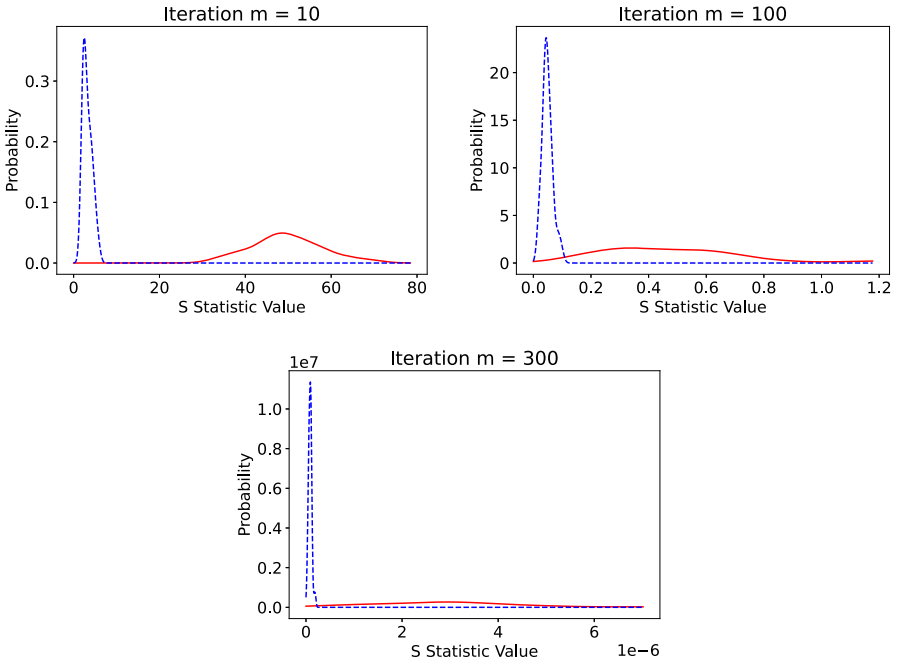

Fig. 14 S-statistic samples and traces for BayesCG under rank-50 approximate Krylov posteriors with $\mathbf{x}_{m+d}$ as the posterior mean at $m = 10, 100, 300$ iterations. The solid red curve represents the traces and the dashed blue curve the S-statistic samples (color figure online)

Proof From (26) follows

$$\mathbf{x}_* - \mathbf{x}_{m+d} = \mathbf{V}_{m+d+1:g}\mathbf{V}_{m+d+1:g}^T\mathbf{r}_0 \in \text{range}(\mathbf{V}_{m+d+1:g}).$$

The residuals $\mathbf{r}_i, 0 \leq i \leq g-1$, are an orthogonal basis for the Krylov space $\mathcal{K}_g(\mathbf{A}, \mathbf{r}_0)$ [2, Theorem 2.8], and in particular,

$$\text{range}(\mathbf{V}_{i:j}) = \text{span}\{\mathbf{r}_{i-1}, \ldots, \mathbf{r}_{j-1}\}, \qquad 1 \leq i < j \leq g.$$

Therefore $\text{range}(\mathbf{V}_{m+d+1:g}) \perp \text{range}(\mathbf{V}_{m+1:d})$, The symmetry of $\widehat{\mathbf{\Gamma}}_m$ implies

$$\text{range}(\mathbf{V}_{m+1:d}) = \text{range}(\widehat{\mathbf{\Gamma}}_m) \perp \ker(\widehat{\mathbf{\Gamma}}_m) = \ker(\widehat{\mathbf{\Gamma}}_m^\dagger).$$

Thus, $\text{range}(\mathbf{V}_{m+d+1:g}) \subseteq \ker(\widehat{\mathbf{\Gamma}}_m)$.

Since $\mathbf{x}_* - \mathbf{x}_{m+d} \in \text{range}(\mathbf{V}_{m+d+1:g}) \subseteq \ker(\widehat{\mathbf{\Gamma}}_m^\dagger) = \ker(\widehat{\mathbf{\Gamma}}_m^\dagger)$, we have that

$$\mathrm{Z}(\mathbf{x}_*) = (\mathbf{x}_* - \mathbf{x}_{m+d})^T\widehat{\mathbf{\Gamma}}_m^\dagger(\mathbf{x}_* - \mathbf{x}_{m+d}) = 0.$$

□

Table 10 This table corresponds to Fig. 14

Iteration	S-stat mean	Trace mean	Trace standard deviation
10	3.21	49.8	7.96
100	4.88×10^{-2}	0.489	0.251
300	8.66×10^{-8}	2.89×10^{-6}	1.51×10^{-6}

For BayesCG under rank-50 approximate Krylov posteriors with $\mathbf{x}_{m+d}$ as the posterior mean, it shows the S-statistic sample means, trace means, and trace standard deviations

***Figure* 14 *and Table* 10.**
With $\mathbf{x}_{m+d}$ as the posterior mean, the S-statistic samples are concentrated at much smaller values than the traces, suggesting that BayesCG is pessimistic with $\mathbf{x}_{m+d}$ as the posterior mean. This is confirmed by a comparison to Fig. 12 and Table 8 which suggest that BayesCG becomes more pessimistic when $\mathbf{x}_{m+d}$ replaces $\mathbf{x}_m$ as the posterior mean.

The S-statistic shows what we expected: $\mathrm{trace}(\mathbf{A}\widehat{\mathbf{\Gamma}}_m)$ in (9) underestimates the error $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2$. From $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2 \geq \|\mathbf{x}_* - \mathbf{x}_{m+d}\|_{\mathbf{A}}^2$ follows that $\mathrm{trace}(\mathbf{A}\widehat{\mathbf{\Gamma}}_m)$ either overestimates $\|\mathbf{x}_* - \mathbf{x}_{m+d}\|_{\mathbf{A}}^2$ or underestimates it by less. The S-statistic therefore suggests that BayesCG is either pessimistic, or else less optimistic.

5.6 Convergence rate and calibration

We investigate how the convergence rate affects the calibration of BayesCG under the approximate Krylov posterior. To this end, we perform the Z- and S-statistic experiments as in the previous section, but for a different linear system. After presenting the matrix $\mathbf{A}$ (Sect. 5.6.1), we discuss the results of the Z- and S- statistic experiments (Sect. 5.6.2).

5.6.1 The matrix A in the linear system

The symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ has dimension $n = 48$ and is defined as [20, Section 5]

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T, \tag{31}$$

where $\mathbf{Q}$ is a random orthogonal matrix and $\mathbf{D}$ is a diagonal matrix with diagonal elements

$$d_{ii} = 0.1 + \frac{i-1}{n-1}(1000 - 0.1)(0.9)^{n-i}, \qquad 1 \leq i \leq n.$$

This class of test matrices, first presented in [35], is popular for CG error estimation because it leads to significant accumulation of round off, and to distinct periods of slow and fast convergence, as illustrated in Fig. 15. This accumulation of roundoff slows the convergence CG, and BayesCG under the Krylov prior, requiring far more than $n = 48$ iterations before the error stops decreasing. This would not happen in exact arithmetic, where CG computes the solution in at most n iterations.

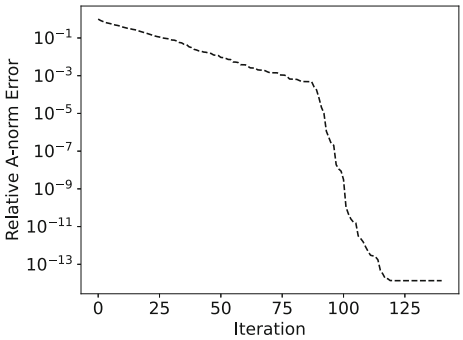

Fig. 15 Relative errors $\|\mathbf{x}_* - \mathbf{x}_m\|_{\mathbf{A}}^2 / \|\mathbf{x}_*\|_{\mathbf{A}}^2$ for BayesCG under the Krylov prior when applied to linear systems $\mathbf{A}\mathbf{x}_* = \mathbf{b}$ with $\mathbf{A}$ in (31) (color figure online)

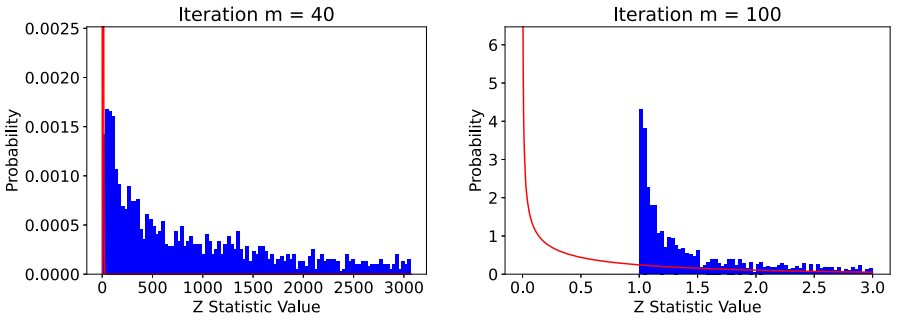

Fig. 16 Z-statistic samples for BayesCG under rank-50 approximate Krylov posteriors at $m = 40, 100$ iterations. The solid red curve represents the chi-squared distribution and the solid blue histogram the Z-statistic samples (color figure online)

We investigate the calibration of BayesCG under the Krylov prior at $m = 40$ iterations, which is in the period of slow convergence, and at $m = 100$ iterations, which is in the period of fast convergence. Although the iteration count of $m = 100$ exceeds the matrix dimension $n = 48$ by a factor of more than 2, we can still approximate Krylov posteriors as long as CG has not yet reached its maximum attainable accuracy.

5.6.2 Numerical experiments

We assess calibration of BayesCG under rank-4 approximate Krylov posteriors by means of the Z- and S-statistics.

Summary of the experiments below.
Both test statistics indicate that BayesCG under rank-4 approximate Krylov posteriors is better calibrated when convergence is fast. This behavior is expected: According to (9) and Theorem 3.5, the errors produced by the approximate and full-rank Krylov posteriors differ by the amount $\|\mathbf{x}_* - \mathbf{x}_{m+d}\|_{\mathbf{A}}^2$, which is small when convergence is fast.

***Figure* 16 *and Table* 11.**
To improve the visibility and interpretability of the figures for this particular linear system, we plot the Z-statistic as a histogram. Figure 16 shows that the Z-statistic sam-

Table 11 This table corresponds to Fig. 16

Iteration	Z-stat mean	χ^2 mean	K–S statistic
40 (slow convergence)	1.29×10^3	4.0	0.988
100 (fast convergence)	1.68	1.0	0.684

For BayesCG under rank-4 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

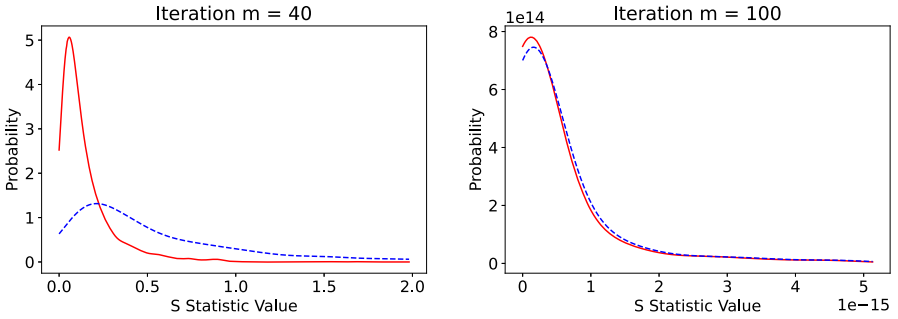

Fig. 17 S-statistic samples and traces for BayesCG under rank-50 approximate Krylov posteriors at $m = 40, 100$ iterations. The solid red curve represents the traces, and the dashed blue curve the S-statistic samples (color figure online)

Table 12 This table corresponds to Fig. 17

Iteration	S-stat mean	Trace mean	Trace std. dev
40.0 (slow convergence)	0.561	0.144	0.163
100.0 (fast convergence)	6.15×10^{-16}	5.47×10^{-16}	1.53×10^{-15}

For BayesCG under rank-4 approximate Krylov posteriors, it shows the S-statistic sample means, trace means, and trace standard deviations

ples are concentrated around larger values than the predicted chi-squared distribution, which suggests that the method is optimistic.

During the period of fast convergence at iteration $m = 100$, the Z-statistic samples are much closer to the predicted chi-squared distribution than they are during the period of slow convergence at iteration $m = 40$. The Kolmogorov–Smirnov statistic of 0.68 at iteration $m = 100$ indicates some overlap between the distributions, while the Kolmogorov–Smirnov statistic of 0.99 at iteration $m = 40$ indicates little overlap between the distributions. Thus, while BayesCG under approximate Krylov priors may not be calibrated in the strict sense, its calibration does improve with increasing convergence rate.

***Figure* 17 *and Table* 12.**
During slow convergence at iteration $m = 40$, S-statistic samples are concentrated around larger values than the traces, suggesting that the S-statistic views BayesCG under approximate Krylov posterior as optimistic. However, at iteration $m = 100$,

the traces and S-statistic samples nearly match each other, indicating that BayesCG is closer to being calibrated.

6 Future research

Although practical CG error estimates have been developed mostly for real linear systems, CG can be naturally extended to complex Hermitian positive definite matrices [18, Section 2.5], with an attendant increase in arithmetic cost.

Does this extension to complex Hermitian matrices also carry over to BayesCG? Preliminary numerical experiments suggest that the behaviour of BayesCG under the Krylov prior appears to be the same, regardless of whether the linear system is complex or real.

How about the Gaussian distributions? One option is to transform the complex linear system $\mathbf{A}\mathbf{x}_* = \mathbf{b}$ to a real system of twice the dimension. Specifically, if $\mathbf{A} = \mathbf{A}_1 + \imath\mathbf{A}_2 \in \mathbb{C}^{n\times n}$ with $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{n\times n}$, is Hermitian positive-definite, then the real matrix

$$\begin{bmatrix} \mathbf{A}_1 & -\mathbf{A}_2 \\ \mathbf{A}_2 & \mathbf{A}_1 \end{bmatrix} \in \mathbb{R}^{(2n)\times(2n)}$$

is symmetric positive-definite [36, P4.2.1]. This transformation would allow the continued use of Gaussians, albeit in the higher-dimensional space $\mathbb{R}^{2n}$. The posteriors then represent separate uncertainties about the real and complex parts of $\mathbf{x}_*$.

Another option is to design Gaussian distributions that can model uncertainty about complex-valued solutions. One could prescribe a matrix mean with separate columns for the real and complex parts, and a covariance tensor for the complex and real parts and their interactions. Extending all of the analysis in [1, 2] to such tensor-valued distributions looks non-trivial.

Acknowledgements We thank the reviewers for their helpful comments, and questions which lead to the addition of Sect. 6. We are also deeply indebted to Jonathan Wenger for his very careful reading of our paper and many perceptive suggestions that greatly improved the paper and widened its outlook through the addition of Sects. 5.5 and 5.6.

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

Appendix 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 A.1 (Stability of Gaussian Distributions [37, 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*

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

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 A.2 (Conjugacy of Gaussian Distributions [38, Section 6.1], [25, Corollary 6.21]) *Let* $X \sim \mathcal{N}(\mathbf{x}, \boldsymbol{\Sigma}_x)$ *and* $Y \sim \mathcal{N}(\mathbf{y}, \boldsymbol{\Sigma}_y)$. *The jointly Gaussian random variable* $\begin{bmatrix} X^T & Y^T \end{bmatrix}^T$ *has the distribution*

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

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

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

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

Lemma A.3 *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 [39, Definition 2.1.3]. □

Definition A.4 [39, 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 A.5 [40, 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*

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

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 A.6 [41, 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 A.7 *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,

$$X^TX = X^T\mathbf{U}\mathbf{U}^TX = X^T\mathbf{U}_1\mathbf{U}_1^TX + X^T\mathbf{U}_2\mathbf{U}_2^TX. \tag{A1}$$

Lemma A.1 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 (A1) 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 A.8 *If* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is symmetric positive definite, and* $M \sim \mathcal{N}(\mathbf{x}_\mu\mathbf{\Sigma}_\mu)$ *and* $N \sim \mathcal{N}(\mathbf{x}_\nu, \mathbf{\Sigma}_\nu)$ *are independent random variables in* $\mathbb{R}^n$, *then*

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

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

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

where the covariances $\mathrm{Cov}(M, N) = \mathrm{Cov}(N, M) = 0$ because M and N are independent. Now apply Lemma A.6 to $M - N$. □

Appendix B: Algorithms

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

B.1 Modified Lanczos method

The Lanczos method [42, 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 B.1 produces an $\mathbf{A}$-orthonormal basis.

Algorithm B.1 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 ε
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** $i \le m$ **do**
7: $\quad \mathbf{w} = \mathbf{A}\mathbf{v}_i - \beta\mathbf{v}_{i-1}$
8: $\quad \alpha = \mathbf{w}^T\mathbf{A}\mathbf{v}_i$
9: $\quad \mathbf{w} = w - \alpha\mathbf{v}_i$
10: $\quad \mathbf{w} = \mathbf{w} - \sum_{j=1}^{i} \mathbf{v}_j\mathbf{v}_j^T\mathbf{A}\mathbf{w}$ ▷ Reorthogonalize $\mathbf{w}$
11: $\quad \mathbf{w} = \mathbf{w} - \sum_{j=1}^{i} \mathbf{v}_j\mathbf{v}_j^T\mathbf{A}\mathbf{w}$
12: $\quad \beta = (\mathbf{w}^T\mathbf{A}\mathbf{w})^{1/2}$
13: $\quad$ **if** $\beta < \varepsilon$ **then**
14: $\quad\quad$ Exit while loop
15: $\quad$ **end if**
16: $\quad i = i + 1$
17: $\quad \mathbf{v}_i = \mathbf{w}/\beta$
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 B.1 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 [43, 44].

B.2 BayesCG with random search directions

The version of BayesCG in Algorithm B.2 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 [6, Section 1.1].

After sampling an initial random search direction $\mathbf{s}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, Algorithm B.2 computes an $\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}$-orthonormal basis for the Krylov space $\mathcal{K}_m(\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}, \mathbf{s}_1)$ with Algorithm B.1. Then Algorithm B.2 computes the BayesCG posteriors directly with (2) and (3) from Theorem 2.1. The numerical experiments in Sect. 5 run Algorithm B.2 with the inverse prior $\mu_0 = \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$.

Algorithm B.2 BayesCG with random search directions

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

B.3 BayesCG with covariances in factored form

Algorithm B.3 takes as input a general prior covariance $\boldsymbol{\Sigma}_0$ in factored form, and subsequently maintains the posterior covariances $\boldsymbol{\Sigma}_m$ in factored form as well. Theorem B.1 presents the correctness proof for Algorithm B.3.

Theorem B.1 *Under the conditions of Theorem 2.1, if* $\boldsymbol{\Sigma}_0 = \mathbf{F}_0\mathbf{F}_0^T$ *for* $\mathbf{F}_0 \in \mathbb{R}^{n\times\ell}$ *and some* $m \le \ell \le n$, *then* $\boldsymbol{\Sigma}_m = \mathbf{F}_m\mathbf{F}_m^T$ *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 $\boldsymbol{\Sigma}_0 = \mathbf{F}_0\mathbf{F}_0^T$ into (3) and factoring out $\mathbf{F}_0$ on the left and $\mathbf{F}_0^T$ on the right gives $\boldsymbol{\Sigma}_m = \mathbf{F}_0\mathbf{P}\mathbf{F}_0^T$ 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 $\boldsymbol{\Sigma}_m = \mathbf{F}_0\mathbf{P}\mathbf{F}_0^T = \mathbf{F}_0\mathbf{P}\mathbf{P}\mathbf{F}_0^T = \mathbf{F}_m\mathbf{F}_m^T$. □

Algorithm B.3 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 $\mathbf{x}_* - \mathbf{x}_0 \in \text{range}(\boldsymbol{\Sigma}_0)$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$
$\mathbf{s}_1 = \mathbf{r}_0$
$\mathbf{P} = \mathbf{0} \in \mathbb{R}^{n\times n}$
$m = 0$
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 $\mathbf{q} = \boldsymbol{\Sigma}_0\mathbf{A}\mathbf{s}_m$
  $\eta_m = \mathbf{s}_m^T\mathbf{A}\mathbf{q}$
  $\mathbf{P}(:, m) = \mathbf{P}(:, m)/\eta_m$ ▷ 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 BayesCG under the Krylov prior

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

B.4.1 Full Krylov posteriors

Algorithm B.4 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 $\boldsymbol{\Phi}$ 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 $\boldsymbol{\Phi}_{m+1:g}$ for the posterior covariance.

B.4.2 Approximate Krylov posteriors

Algorithm B.5 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 $\widehat{\boldsymbol{\Gamma}}_m$ in Lines 16–26.

Algorithm B.4 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 B.1
$\mathbf{\Phi} = \mathrm{diag}((\mathbf{V}^T\mathbf{r}_0)^2)$ ▷ 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: $\mathbf{x}_m$, $\mathbf{V}_{m+1:g}$, $\mathbf{\Phi}_{m+1:g}$

Algorithm B.5 BayesCG under the Krylov prior [2, 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: $\mathbf{x}_m$, $\mathbf{V}_{m+1:m+d}$, $\mathbf{\Phi}_{m+1:m+d}$

References

1. Cockayne, J., Oates, C.J., Ipsen, I.C.F., Girolami, M.: A Bayesian conjugate gradient method (with discussion). Bayesian Anal. **14**(3), 937–1012 (2019). https://doi.org/10.1214/19-BA1145. Includes 6 discussions and a rejoinder from the authors
2. Reid, T.W., Ipsen, I.C.F., Cockayne, J., Oates, C.J.: BayesCG as an uncertainty aware version of CG. arXiv:2008.03225 (2022)
3. Cockayne, J., Oates, C.J., Sullivan, T.J., Girolami, M.: Bayesian probabilistic numerical methods. SIAM Rev. **61**(4), 756–789 (2019). https://doi.org/10.1137/17M1139357
4. Hennig, P., Osborne, M.A., Girolami, M.: Probabilistic numerics and uncertainty in computations. Proc. R. Soc. A. **471**(2179), 20150142–17 (2015)
5. Oates, C.J., Sullivan, T.J.: A modern retrospective on probabilistic numerics. Stat. Comput. **29**(6), 1335–1351 (2019). https://doi.org/10.1007/s11222-019-09902-z
6. Cockayne, J., Ipsen, I.C.F., Oates, C.J., Reid, T.W.: Probabilistic iterative methods for linear systems. J. Mach. Learn. Res. **22**(232), 1–34 (2021)

7. Hart, J., van Bloemen Waanders, B., Herzog, R.: Hyperdifferential sensitivity analysis of uncertain parameters in PDE-constrained optimization. Int. J. Uncertain. Quantif. **10**(3), 225–248 (2020). https://doi.org/10.1615/Int.J.UncertaintyQuantification.2020032480
8. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer Series in Operations Research and Financial Engineering, p. 664. Springer, New York (2006)
9. Petra, N., Zhu, H., Stadler, G., Hughes, T.J.R., Ghattas, O.: An inexact Gauss-Newton method for inversion of basal sliding and rheology parameters in a nonlinear Stokes ice sheet model. J. Glaciol. **58**(211), 889–903 (2012). https://doi.org/10.3189/2012JoG11J182
10. Saibaba, A.K., Hart, J., van Bloemen Waanders, B.: Randomized algorithms for generalized singular value decomposition with application to sensitivity analysis. Numer. Linear Algebra Appl. **28**(4), 2364–27 (2021). https://doi.org/10.1002/nla.2364
11. Bartels, S., Cockayne, J., Ipsen, I.C.F., Hennig, P.: Probabilistic linear solvers: a unifying view. Stat. Comput. **29**(6), 1249–1263 (2019). https://doi.org/10.1007/s11222-019-09897-7
12. Fanaskov, V.: Uncertainty calibration for probabilistic projection methods. Stat. Comput. **31**(5), 56–17 (2021). https://doi.org/10.1007/s11222-021-10031-9
13. Hennig, P.: Probabilistic interpretation of linear solvers. SIAM J. Optim. **25**(1), 234–260 (2015). https://doi.org/10.1137/140955501
14. Wenger, J., Hennig, P.: Probabilistic linear solvers for machine learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 6731–6742. Curran Associates Inc, Red Hook (2020)
15. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand. **49**, 409–436 (1952). https://doi.org/10.6028/jres.049.044
16. Higham, N.J.: Functions of Matrices. Theory and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2008). https://doi.org/10.1137/1.9780898717778
17. Cockayne, J., Oates, C.J., Ipsen, I.C.F., Girolami, M.: Supplementary material for 'A Bayesian conjugate-gradient method'. Bayesian Anal. (2019). https://doi.org/10.1214/19-BA1145SUPP
18. Liesen, J., Strakos, Z.: Krylov Subspace Methods: Principles and Analysis. Oxford University Press, Oxford (2013)
19. Berljafa, M., Güttel, S.: Generalized rational Krylov decompositions with an application to rational approximation. SIAM J. Matrix Anal. Appl. **36**(2), 894–916 (2015). https://doi.org/10.1137/140998081
20. Strakoš, Z., Tichý, P.: On error estimation in the conjugate gradient method and why it works in finite precision computations. Electron. Trans. Numer. Anal. **13**, 56–80 (2002)
21. Kressner, D., Latz, J., Massei, S., Ullmann, E.: Certified and fast computations with shallow covariance kernels. arXiv:2001.09187 (2020)
22. Villani, C.: Optimal Transport, Old and New. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 338, p. 973. Springer, Berlin (2009). https://doi.org/10.1007/978-3-540-71050-9
23. Gelbrich, M.: On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. Math. Nachr. **147**, 185–203 (1990). https://doi.org/10.1002/mana.19901470121
24. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (1985). https://doi.org/10.1007/978-1-4757-4286-2
25. Stuart, A.M.: Inverse problems: a Bayesian perspective. Acta Numer **19**, 451–559 (2010). https://doi.org/10.1017/S0962492910000061
26. Cockayne, J., Graham, M.M., Oates, C.J., Sullivan, T.J.: Testing whether a Learning Procedure is Calibrated. arXiv:2012.12670 (2021)
27. Ross, S.M.: Introduction to Probability Models, 9th edn. Academic Press Inc, Boston (2007)
28. Kaltenbach, H.-M.: A Concise Guide to Statistics. Springer Briefs in Statistics, p. 111. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-23502-3
29. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, vol. 112, 2nd edn. Springer, New York (2021). https://doi.org/10.1007/978-1-0716-1418-1
30. BCSSTK14: BCS Structural Engineering Matrices (linear equations) Roof of the Omni Coliseum, Atlanta. https://math.nist.gov/MatrixMarket/data/Harwell-Boeing/bcsstruc2/bcsstk14.html
31. Golub, G.H., Van Loan, C.F.: Matrix Computations, 4th edn. The Johns Hopkins University Press, Baltimore (2013)
32. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations. I, 2nd edn. Springer Series in Computational Mathematics, vol. 8. Springer, Berlin (1993)

33. Golub, G.H., Meurant, G.: Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods. BIT **37**(3), 687–705 (1997). https://doi.org/10.1007/BF02510247
34. Meurant, G., Tichý, P.: Approximating the extreme Ritz values and upper bounds for the A-norm of the error in CG. Numer. Algorithms **82**(3), 937–968 (2019). https://doi.org/10.1007/s11075-018-0634-8
35. Strakoš, Z.: On the real convergence rate of the conjugate gradient method. Linear Algebra Appl. **154**(156), 535–549 (1991). https://doi.org/10.1016/0024-3795(91)90393-B
36. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. The Johns Hopkins University Press, Baltimore (1996)
37. Muirhead, R.J.: Aspects of Multivariate Statistical Theory. Wiley, New York (1982)
38. Ouellette, D.V.: Schur complements and statistics. Linear Algebra Appl. **36**, 187–295 (1981). https://doi.org/10.1016/0024-3795(81)90232-9
39. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
40. Campbell, S.L., Meyer, C.D.: Generalized Inverses of Linear Transformations. Classics in Applied Mathematics, vol. 56. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2009). https://doi.org/10.1137/1.9780898719048.ch0
41. Mathai, A.M., Provost, S.B.: Quadratic Forms in Random Variables: Theory and Applications. Marcel Dekker Inc, New York (1992)
42. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2003)
43. Giraud, L., Langou, J., Rozložník, M.: The loss of orthogonality in the Gram–Schmidt orthogonalization process. Comput. Math. Appl. **50**(7), 1069–1075 (2005). https://doi.org/10.1016/j.camwa.2005.08.009
44. Giraud, L., Langou, J., Rozložník, M., van den Eshof, J.: Rounding error analysis of the classical Gram–Schmidt orthogonalization process. Numer. Math. **101**(1), 87–100 (2005). https://doi.org/10.1007/s00211-005-0615-4