Probabilistic linear solvers: a unifying view

Simon Bartels[1,2] · Jon Cockayne[3] · Ilse C. F. Ipsen[4] · Philipp Hennig[1,2]

Abstract

Several recent works have developed a new, probabilistic interpretation for numerical algorithms solving linear systems in which the solution is inferred in a Bayesian framework, either directly or by inferring the unknown action of the matrix inverse. These approaches have typically focused on replicating the behaviour of the conjugate gradient method as a prototypical iterative method. In this work, surprisingly general conditions for equivalence of these disparate methods are presented. We also describe connections between probabilistic linear solvers and projection methods for linear systems, providing a probabilistic interpretation of a far more general class of iterative methods. In particular, this provides such an interpretation of the generalised minimum residual method. A probabilistic view of preconditioning is also introduced. These developments unify the literature on probabilistic linear solvers and provide foundational connections to the literature on iterative solvers for linear systems.

Keywords Probabilistic linear solvers · Projection methods · Iterative methods · Preconditioning

1 Introduction

Consider the linear system

$$Ax^* = b \tag{1}$$

where $A \in \mathbb{R}^{d \times d}$ is an invertible matrix, $b \in \mathbb{R}^d$ is a given vector, and $x^* \in \mathbb{R}^d$ is an unknown to be determined. Recent work (Hennig 2015; Cockayne et al. 2018) has constructed iterative solvers for this problem which output *probability*

✉ Simon Bartels
  sbartels@tue.mpg.de

  Jon Cockayne
  j.cockayne@warwick.ac.uk

  Ilse C. F. Ipsen
  ipsen@ncsu.edu

  Philipp Hennig
  ph@tue.mpg.de

[1] University of Tübingen, 72076 Tübingen, Germany

[2] Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany

[3] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

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

measures, constructed to quantify uncertainty due to terminating the algorithm before the solution has been identified completely. On the surface the approaches in these two works appear different: in the matrix-based inference (MBI) approach of Hennig (2015), a posterior is constructed on the matrix A^{-1}, while in the solution-based inference (SBI) method of Cockayne et al. (2018) a posterior is constructed on the solution vector x^*.

These algorithms are instances of *probabilistic numerical methods* (PNM) in the sense of Hennig et al. (2015) and Cockayne et al. (2017). PNM are numerical methods which output posterior distributions that quantify uncertainty due to discretisation error. An interesting property of PNM is that they often result in a posterior distributions whose mean element coincides with the solution given by a classical numerical method for the problem at hand. The relationship between PNM and classical solvers has been explored for integration (e.g. Karvonen and Sarkka 2017), ODE solvers (Schober et al. 2014, 2019; Kersting et al. 2018) and PDE solvers (Cockayne et al. 2016) in some generality. For linear solvers, attention has thus far been restricted to the conjugate gradient (CG) method. Since CG is but a single member of a larger class of iterative solvers, and applicable only if the matrix A is symmetric and positive definite, extending the probabilistic interpretation is an interesting endeavour. Probabilistic interpretations provide an alternative perspec-

tive on numerical algorithms and can also provide extensions such as the ability to exploit noisy or corrupted observations. The probabilistic view has also been used to the develop new numerical methods (Xi et al. 2018), and *Bayesian* PNM can be incorporated rigorously into pipelines of computation (Cockayne et al. 2017).

Preconditioning—mapping Eq. (1) to a better conditioned system with the same solution—is key to the fast convergence of iterative linear solvers, particularly those based upon Krylov methods (Liesen and Strakos 2012). The design of preconditioners has been referred to as "a combination of art and science" (Saad 2003, p. 283). In this work, we also provide a new, probabilistic interpretation of preconditioning as a form of prior information.

1.1 Contribution

This text contributes three primary insights:

1. It is shown that, for particular choices of the generative model, matrix-based inference (MBI) and solution-based inference (SBI) can be equivalent (Sect. 2).
2. A general probabilistic interpretation of projection methods (Saad 2003) is described (Sect. 3.1), leading to a probabilistic interpretation of the generalised minimum residual method (GMRES; Saad and Schultz (1986), Sect. 6). The connection to CG is expanded and made more concise in Sect. 5.
3. A probabilistic interpretation of preconditioning is presented in Sect. 4.

Most of the proofs are presented inline; lengthier proofs are deferred to "Appendix B". While an important consideration, the predominantly theoretical contributions of this paper will not consider the impact of finite numerical precision.

1.2 Notation

For a symmetric positive-definite matrix $M \in \mathbb{R}^{d \times d}$ and two vectors $v, w \in \mathbb{R}^d$, we write $\langle v, w \rangle_M = v^\top M w$ for the inner product induced by M, and $\|v\|_M^2 = \langle v, v \rangle_M$ for the corresponding norm.

A set of vectors $s_1, \ldots, s_m$ is called *M-orthogonal* or *M-conjugate* if $\langle s_i, s_j \rangle_M = 0$ for $i \neq j$, and *M-orthonormal* if, in addition, $\|s_i\|_M = 1$ for $1 \leq i \leq m$.

For a square matrix $A = \begin{bmatrix} a_1 \ldots a_d \end{bmatrix}^\top \in \mathbb{R}^{d \times d}$, the *vectorisation operator* vec : $\mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d^2}$ stacks the rows[1] of A into one long vector:

$$\vec{A} \equiv \text{vec}(A) = \begin{bmatrix} a_1 \\ \vdots \\ a_d \end{bmatrix}, \quad \text{with} \quad \left[\vec{A}\right]_{(ij)} = [A]_{ij}.$$

The *Kronecker product* of two matrices $A, B \in \mathbb{R}^{d \times d}$ is $A \otimes B$ with $[A \otimes B]_{(ij),(k\ell)} = [A]_{ik}[B]_{j\ell}$. A list of its properties is provided in "Appendix A".

The Krylov space of order m generated by the matrix $A \in \mathbb{R}^{d \times d}$ and the vector $b \in \mathbb{R}^d$ is

$$K_m(A, b) = \text{span}\left(b, Ab, A^2 b, \ldots, A^{m-1} b\right).$$

We will slightly abuse notation to describe shifted and scaled subspaces of $\mathbb{R}^d$: let $\mathbb{S}$ be an m-dimensional linear subspace of $\mathbb{R}^d$ with basis $\{s_1, \ldots, s_m\}$. Then, for a vector $v \in \mathbb{R}^d$ and a matrix $M \in \mathbb{R}^{d \times d}$, let

$$v + M\mathbb{S} = \text{span}(v + M s_1, \ldots, v + M s_m).$$

2 Probabilistic linear solvers

Several probabilistic frameworks describing the solution of Eq. (1) have been constructed in recent years. They primarily differ in the subject of inference: SBI approaches such as Cockayne et al. (2018), of which *BayesCG* is an example, place a prior distribution on the solution x^* of Eq. (1). Conversely, the MBI approach of Hennig (2015) and Bartels and Hennig (2016) places a prior on A^{-1}, treating the action of the inverse operator as an unknown to be inferred.[2] This section reviews each approach and adds some new insights. In particular, SBI can be viewed as strict special case of MBI (Sect. 2.4).

Throughout this section, we will assume that the search directions S_m in $S_m^\top A x^* = S_m^\top b$ are independent of x^*. Generally speaking, this is not the case for projection methods, in which the solution space often depends strongly on b, as described in Sects. 5 and 6. This disconnect is the source of the poor uncertainty quantification reported in Cockayne et al. (2018) and shown also to hold for the methods in this work in Sect. 6.4. This will not be examined in further detail in this work, though it remains an important area of development for probabilistic linear solvers.

---

[1] Stacking the columns is equivalently possible and common. It is associated with a permutation in the definition of the Kronecker product, but the resulting inferences are equivalent.

[2] Hennig (2015) also discusses inference over A. This model class will not be discussed further in the present work. It has the disadvantage that the associated marginal on x^* is nonanalytic, but more easily lends itself to situations with noisy or otherwise perturbed matrix-vector products as observations.

2.1 Background on Gaussian conditioning

The propositions in this section follow from the following two classic properties of Gaussian distributions.

Lemma 1 *Let $x \in \mathbb{R}^d$ be Gaussian distributed with density $p(x) = \mathcal{N}(x; x_0, \Sigma)$ for $x_0 \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ a positive semi-definite matrix. Let $M \in \mathbb{R}^{n \times d}$ and $z \in \mathbb{R}^n$. Then, $v = Mx + z$ is also Gaussian, with*

$$p(v) = \mathcal{N}(v; Mx_0 + z, M\Sigma M^\top).$$

Lemma 2 *Let $x \in \mathbb{R}^d$ be distributed as in Lemma 1, and let observations $y \in \mathbb{R}^n$ be generated from the conditional density*

$$p(y \mid x) = \mathcal{N}(y; Mx + z, \Lambda)$$

with $M \in \mathbb{R}^{n \times d}$, $z \in \mathbb{R}^n$, and $\Lambda \in \mathbb{R}^{n \times n}$ again positive semi-definite. Then, the associated conditional distribution on x after observing y is again Gaussian, with

$$p(x \mid y) = \mathcal{N}(x; \bar{x}, \bar{\Sigma}) \quad where$$
$$\bar{x} = x_0 + \Sigma M^\top (M\Sigma M^\top + \Lambda)^{-1}(y - Mx_0 - z)$$
$$\bar{\Sigma} = \Sigma - \Sigma M^\top (M\Sigma M^\top + \Lambda)^{-1} M\Sigma).$$

This formula also applies if $\Lambda = 0$, i.e. observations are made without noise, with the caveat that if $M\Sigma M^\top$ is singular, the inverse should be interpreted as a pseudo-inverse.

2.2 Solution-based inference

To phrase the solution of Eq. (1) as a form of probabilistic inference, Cockayne et al. (2018) consider a Gaussian prior over the solution x^*, and condition on observations provided by a set of *search directions* $s_1, \ldots, s_m$, $m < d$. Let $S_m \in \mathbb{R}^{d \times m}$ be given by $S_m = [s_1, \ldots, s_m]$, and let information be given by $y_m := S_m^\top A x^* = S_m^\top b$. Since the information is a linear projection of x^*, the posterior distribution is a Gaussian distribution on x^*:

Lemma 3 (Cockayne et al. 2018) *Assume that the columns of S_m are linearly independent. Consider the prior*

$$p(x) = \mathcal{N}(x; x_0, \Sigma_0).$$

The posterior from SBI is then given by

$$p(x \mid y_m) = \mathcal{N}(x; x_m, \Sigma_m)$$

where

$$x_m = x_0 + \Sigma_0 A^\top S_m (S_m^\top A \Sigma_0 A^\top S_m)^{-1} S_m^\top r_0$$

$$\Sigma_m = \Sigma_0 - \Sigma_0 A^\top S_m (S_m^\top A \Sigma_0 A^\top S_m)^{-1} S_m^\top \Sigma_0, \qquad (2)$$

and $r_0 = b - Ax_0$.

The following proposition establishes an optimality property of the posterior mean x_m. This is a relatively well-known property of Gaussian inference, which will prove useful in subsequent sections.

Proposition 4 *If $\mathbb{S}_m = range(S_m)$, then the posterior mean in Lemma 3 satisfies the optimality property*

$$x_m = \arg \min_{x \in x_0 + \Sigma_0 A^\top \mathbb{S}_m} \|x - x^*\|_{\Sigma_0^{-1}}.$$

Proof With the abbreviations $X = \Sigma_0 A^\top S_m$ and $y = x^* - x_0$ the mean in Lemma 3 can be written as

$$x_m = x_0 + X c_m,$$

where

$$c_m = (X^\top \Sigma_0^{-1} X)^{-1} X^\top \Sigma_0^{-1} y$$

is the solution of the weighted least squares problem (Golub and Van Loan 2013, Section 6.1)

$$c_m = \arg \min_{c \in \mathbb{R}^m} \|Xc - y\|_{\Sigma_0^{-1}}$$
$$= \arg \min_{c \in \mathbb{R}^m} \|x_0 + \Sigma_0 A^\top S_m c - x^*\|_{\Sigma_0^{-1}}.$$

This is equivalent to the desired statement. □

2.3 Matrix-based inference

In contrast to SBI, the MBI approach of Hennig (2015) treats the matrix inverse A^{-1} as the unknown in the inference procedure. As in the previous section, search directions S_m yield matrix-vector products $Y_m \in \mathbb{R}^{d \times m}$. In Hennig (2015), these arise from *right*-multiplying[3] A with S_m, i.e. $Y_m = AS_m$. Note that

$$S_m = A^{-1} Y_m, \text{ or, equivalently } \overrightarrow{S_m} = (I \otimes Y_m^\top) \overrightarrow{A^{-1}}. \qquad (3)$$

Thus, S_m is a linear transformation of A^{-1} and Lemma 2 can again be applied:

Lemma 5 (Lemma 2.1 in Hennig (2015))[4] *Consider the prior*

$$p\left(\overrightarrow{A^{-1}}\right) = \mathcal{N}\left(\overrightarrow{A^{-1}}; \overrightarrow{A_0^{-1}}, \Sigma_0 \otimes W_0\right). \qquad (4)$$

---

[3] This work also considers a model class that explicitly encodes *symmetry* of A, such that the distinction between left multiplication and right multiplication vanishes. See Sect. 5.2 and Proposition 16 for more.

[4] This corrects a printing error in Hennig (2015). The notation has been adapted to fit the context.

Then, the posterior given the observations $\overrightarrow{S_m} = A^{-1}Y_m$ *is given by*

$$p\left(\overrightarrow{A^{-1}} \middle| \overrightarrow{S_m}\right) = \mathcal{N}\left(\overrightarrow{A^{-1}}; \overrightarrow{A_m^{-1}}, \Sigma_0 \otimes W_m\right)$$

with

$$A_m^{-1} = A_0^{-1} + (S_m - A_0^{-1}Y_m)(Y_m^\top W_0 Y_m)^{-1} Y_m^\top W_0$$
$$W_m = W_0 - W_0 Y_m (Y_m^\top W_0 Y_m)^{-1} Y_m^\top W_0.$$

For linear solvers, the object of interest is $x^* = A^{-1}b$. Writing $A^{-1}b = (I \otimes b^\top)\overrightarrow{A^{-1}}$, and again using Lemma 1, we see that the associated marginal is also Gaussian and given by

$$p(x \mid S, Y) = \mathcal{N}(x; A_m^{-1}b, b^\top W_m b \cdot \Sigma_0). \tag{5}$$

In the Kronecker product specification for the prior covariance on A^{-1} from Eq. (4), the matrix Σ_0, describes the dependence between the columns of A^{-1}. The matrix W_0 captures the dependency between the rows of A^{-1}. Note that in Lemma 5, the posterior covariance has the form $\Sigma_0 \otimes W_m$. When compared to the prior covariance, $\Sigma_0 \otimes W_0$, it is clear that the observations have conveyed no new information to the first term of the Kronecker product covariance.

The Kronecker structure of the prior covariance matrix in Eq. (4) is by no means the only option that facilitates tractable inference.[5] However, in the absence of the literature exploring other approaches within MBI, we will assume throughout that MBI refers to the use of the Kronecker produce prior covariance.

2.4 Equivalence of MBI and SBI

In practice, Hennig (2015) notes that inference on A^{-1} should be performed only implicitly, avoiding the d^2 storage cost and the mathematical complexity of the operations involved in Lemma 5. This raises the question of when MBI is equivalent to SBI. Although, based on Lemma 1, one might suspect SBI and MBI to be equivalent, in fact the posterior from Lemma 5 is structurally different to the posterior in Lemma 3: after projecting into solution space, the posterior covariance in Lemma 5 is a scalar multiple of the matrix Σ_0, which is not the case in general in Lemma 3.

However, the implied posterior over the solution vector can be made to coincide with the posterior from SBI if one considers observations in MBI as

$$S_m^\top = Y_m^\top A^{-1}. \tag{6}$$

That is, as *left*-multiplications of A. We will refer to the observation model of Eq. (3) as *right-multiplied information*, and to Eq. (6) as *left-multiplied information*.

Proposition 6 *Consider a Gaussian MBI prior*

$$p(A^{-1}) = \mathcal{N}\left(A^{-1}; \overrightarrow{A_0^{-1}}, \Sigma_0 \otimes W_0\right),$$

conditioned on the left-multiplied information of Eq. (6). The associated marginal on x is identical to the posterior on x arising in Lemma 3 from $p(x) = \mathcal{N}(x; x_0, \Sigma_0)$ under the conditions

$$A_0^{-1}b = x_0 \quad and \quad b^\top W_0 b = 1.$$

Proof See "Appendix B". □

The first of the two conditions requires that the prior mean on the matrix inverse be consistent with the prior mean on the solution, which is natural. The second condition demands that, after projection into solution space, the relationship between the rows of A^{-1} modelled by W_0 does not inflate the covariance Σ_0. Note that this condition is trivial to enforce for an arbitrary covariance $\bar{W}_0$ by setting $W_0 = (b^\top \bar{W}_0 b)^{-1}\bar{W}_0$.

2.5 Remarks

The result in Proposition 6 shows that any result proven for SBI applies immediately to MBI with left-multiplied observations. Though MBI has more model parameters than SBI, there are situations in which this point of view is more appropriate. Unlike in SBI, the information obtained in MBI need not be specific to a particular solution vector x^* and thus can be propagated and recycled over several linear problems, similar to the notion of subspace recycling (Soodhalter et al. 2014). Secondly, MBI is able to utilise both left- and right-multiplied information, while SBI is restricted to left-multiplied information. This additional generality may prove useful in some applications.

3 Projection methods as inference

This section discusses a connection between probabilistic numerical methods for linear systems and the classic framework of projection methods for the iterative solution of linear problems. Section 3.1 reviews this established class of solvers, while Sect. 3.2 presents the novel results.

3.1 Background

Many iterative methods for linear systems, including CG and GMRES, belong to the class of projection methods

---

[5] A diagonal covariance matrix would allow efficient inference, as well.

(Saad [2003], p. 130f.). Saad describes a projection method as an iterative scheme in which, at each iteration, a solution vector $\boldsymbol{x}_m$ is constructed by projecting $\boldsymbol{x}^*$ into a *solution space* $\mathbb{X}_m \subset \mathbb{R}^d$, subject to the restriction that the residual $\boldsymbol{r}_m = \boldsymbol{b} - A\boldsymbol{x}_m$ is orthogonal to a *constraint space* $\mathbb{U}_m \subset \mathbb{R}^d$.

More formally, each iteration of a projection method is defined by two matrices $X_m, U_m \in \mathbb{R}^{d \times m}$, and by a starting point $\boldsymbol{x}_0$. The matrices X_m and U_m each encode the solution and constraint spaces as $\mathbb{X}_m = \text{range}(X_m)$ and $\mathbb{U}_m = \text{range}(U_m)$. The projection method then constructs $\boldsymbol{x}_m$ as $\boldsymbol{x}_m = \boldsymbol{x}_0 + X_m \boldsymbol{\alpha}_m$ with $\boldsymbol{\alpha}_m \in \mathbb{R}^m$ determined by the constraint $U_m^\top \boldsymbol{r}_m = \boldsymbol{0}$. This is possible only if $U_m^\top A X_m$ is nonsingular, in which case one obtains

$$\boldsymbol{\alpha}_m = (U_m^\top A X_m)^{-1} U_m^\top \boldsymbol{r}_0, \text{ and thus} \tag{7}$$

$$\boldsymbol{x}_m = \boldsymbol{x}_0 + X_m (U_m^\top A X_m)^{-1} U_m^\top \boldsymbol{r}_0. \tag{8}$$

From this perspective, CG and GMRES perform only a single step with the number of iterations m fixed and determined in advance. For CG, the spaces are $\mathbb{U}_m = \mathbb{X}_m = K_m(A, \boldsymbol{b})$, while for GMRES they are $\mathbb{X}_m = K_m(A, \boldsymbol{b})$ and $\mathbb{U}_m = A K_m(A, \boldsymbol{b})$ (Saad [2003], Proposition 5.1).

3.2 Probabilistic perspectives

In this section, we first show, in Proposition [7], that the conditional mean from SBI after m steps corresponds to some projection method. Then, in Proposition [8] we prove the converse: that each projection method is also the posterior mean of a probabilistic method, for some prior covariance and choice of information.

Proposition 7 *Let the columns of S_m be linearly independent. Consider SBI under the prior*

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{x}_0, \Sigma_0),$$

and with observations $\boldsymbol{y}_m = S_m^\top \boldsymbol{b}$. Then, the posterior mean $\boldsymbol{x}_m$ in Lemma [3] is identical to the iterate from a projection method defined by the matrices $U_m = S_m$ and $X_m = \Sigma_0 A^\top S_m$, and the starting vector $\boldsymbol{x}_0$.

Proof Substituting $U_m = S_m$ and $X_m = \Sigma_0 A^\top S_m$ into Lemma [3] gives Eq. [8], as required. □

The converse to this also holds:

Proposition 8 *Consider a projection method defined by the matrices $X_m, U_m \in \mathbb{R}^{d \times m}$, each with linearly independent columns, and the starting vector $\boldsymbol{x}_0 \in \mathbb{R}^d$. Then, the iterate $\boldsymbol{x}_m$ in Eq. [8] is identical to the SBI posterior mean in Lemma [3] under the prior*

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{x}_0, X_m X_m^\top) \tag{9}$$

when search directions $S_m = U_m$ are used.

Proof Abbreviate $Z = X_m^\top A^\top U_m$ and write the projection method iterate from Eq. [8] as

$$\boldsymbol{x}_m = \boldsymbol{x}_0 + X_m Z^{-T} U_m^\top \boldsymbol{r}_0.$$

Multiply the middle matrix by the identity,

$$\begin{aligned} Z^{-T} &= Z Z^{-1} Z^{-T} = Z(Z^\top Z)^{-1} \\ &= X_m^\top A^\top U_m (U_m^\top A \Sigma_0 A^\top U_m)^{-1}, \end{aligned}$$

and insert this into the expression for $\boldsymbol{x}_0$,

$$\boldsymbol{x}_m = \boldsymbol{x}_0 + \Sigma_0 A^\top U_m (U_m^\top A \Sigma_0 A^\top U_m)^{-1} U_m^\top \boldsymbol{r}_0.$$

Setting $U_m = S_m$ gives the mean in Lemma [3]. □

A direct way to enforce the posterior occupying the solution space is by placing a prior on the coefficients $\boldsymbol{\alpha}$ in $\boldsymbol{x} = \boldsymbol{x}_0 + X_m \boldsymbol{\alpha}$. Under a unit Gaussian prior $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{0}, I)$, the implied prior on $\boldsymbol{x}$ naturally has the form of Eq. [9].

However, this prior is unsatisfying since it requires the solution space to be specified a priori, precluding adaptivity in the algorithm and perhaps more worryingly, the posterior uncertainty over the solution is a matrix of zeros even though the solution is not fully identified. Again taking $Z = X_m^\top A^\top U_m$:

$$\begin{aligned} \Sigma_m &= \Sigma_0 - \Sigma_0 A^\top U_m (U_m^\top A \Sigma_0 A^\top U_m)^{-1} U_m^\top A \Sigma_0 \\ &= X_m X_m^\top - X_m Z(Z^\top Z)^{-1} Z^\top X_m^\top \\ &= X_m X_m^\top - X_m X_m^\top \\ &= 0. \end{aligned}$$

Concerning this issue, Hennig ([2015]) and Bartels and Hennig ([2016]) each proposed to adding additional uncertainty in the null space of X_m. This empirical uncertainty calibration step has not yet been analysed in detail. Such analysis is left for future work.

Including the solution space X_m in the prior covariance matrix requires it to be specified a priori. For solvers like CG and GMRES which construct X_m adaptively, this assumption may appear problematic—a probabilistic interpretation should use for inference only quantities that have already been computed. The computation of X_m could be seen as part of the initialisation, but this requires that the number of iterations m to be fixed a priori, whereas typically such methods choose m adaptively by examining the norm of the residual.[6] Nevertheless, the proposition provides a probabilistic view for *arbitrary* projection methods and does not

---

[6] Sometimes m is fixed a priori, due to memory or computation time limits.

involve A^{-1}, unlike the results presented in Hennig (2015), Cockayne et al. (2017).

The above prior is not unique. The next proposition establishes probabilistic interpretations of projection methods under priors that are independent of solution- and constraint space, albeit under more restrictive conditions. The benefit of this is that m need not be fixed a priori.

Proposition 9 *Consider a projection method defined by* $X_m, U_m \in \mathbb{R}^{d \times m}$ *and the starting vector* $\boldsymbol{x}_0$. *Further suppose that* $U_m = RX_m$ *for some invertible* $R \in \mathbb{R}^{d \times d}$, *and that* $A^\top R$ *is symmetric positive definite. Then, under the prior*

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{x}_0, (A^\top R)^{-1}\right)$$

and the search directions $S_m = U_m = RX_m$, *the iterate in the projection method is identical to the posterior mean in Lemma* 3.

Proof First substitute $X_m = R^{-1}U_m$ into Eq. (8) to obtain

$$\begin{aligned} \boldsymbol{x}_m &= \boldsymbol{x}_0 + R^{-1}U_m(U_m^\top AR^{-1}U_m)^{-1}U_m^\top \boldsymbol{r}_0 \\ &= \boldsymbol{x}_0 + R^{-1}A^{-\top}A^\top U_m(U_m^\top AR^{-1}A^{-\top}A^\top U_m)^{-1}U_m^\top \boldsymbol{r}_0 \\ &= \boldsymbol{x}_0 + \Sigma_0 A^\top U_m(U_m^\top A\Sigma_0 A^\top U_m)^{-1}U_m^\top \boldsymbol{r}_0. \end{aligned}$$

The third line uses $\Sigma_0 = (A^\top R)^{-1} = R^{-1}A^{-T}$. This is equivalent to the posterior mean in Eq. (2) with $S_m = U_m$. $\square$

A corollary which provides further insight arises when one considers the *polar decomposition* of A. Recall that an invertible matrix A has a unique polar decomposition $A = PH$, where $P \in \mathbb{R}^{d \times d}$ is orthogonal and $H \in \mathbb{R}^{d \times d}$ is symmetric positive definite.

Corollary 10 *Consider a projection method defined by* $X_m, U_m \in \mathbb{R}^{d \times m}$ *and the starting vector* $\boldsymbol{x}_0$, *and suppose that* $U_m = PX_m$, *where* P *arises from the polar decomposition* $A = PH$. *Then, under the prior*

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{x}_0, H^{-1}\right)$$

and the search directions $S_m = U_m = PX_m$, *the iterate in the projection method is identical to the posterior mean in Lemma* 3.

Proof This follows from Proposition 9. Setting $R = P$ aligns the search directions in Corollary 10 with those in Proposition 9. Since P is orthogonal, $P^{-1} = P^\top$, and since H is symmetric positive definite, $A^\top P = P^\top A = H$ by definition of the polar decomposition, which gives the prior covariance required for Proposition 9. $\square$

This is an intuitive analogue of similar results in Hennig (2015) and Cockayne et al. (2017) which show that

CG is recovered under certain conditions involving a prior $\Sigma_0 = A^{-1}$. When A is not symmetric and positive definite, it cannot be used as a prior covariance. This corollary suggests a natural way to select a prior covariance still linked to the linear system, though this choice is still not computationally convenient. Furthermore, in the case that A is symmetric positive definite, this recovers the prior which replicates CG described in Cockayne et al. (2018). Note that each of H and P can be stated explicitly as $H = (A^\top A)^{\frac{1}{2}}$ and $P = A(A^\top A)^{-\frac{1}{2}}$. Thus, in the case of symmetric positive-definite A we have that $H = A$ and $P = I$, so that the prior covariance $\Sigma_0 = A^{-1}$ arises naturally from this interpretation.

4 Preconditioning

This section discusses probabilistic views on preconditioning. Preconditioning is a widely used technique accelerating the convergence of iterative methods (Saad 2003, Sections 9 and 10). A preconditioner P is a nonsingular matrix satisfying two requirements:

1. Linear systems $Pz = c$ can be solved at low computational cost
2. P is "close" to A in some sense.

In this sense, solving systems based upon a preconditioner can be viewed as approximately inverting A, and indeed, many preconditioners are constructed based upon this intuition. One distinguishes between *right preconditioners* P_r and *left preconditioners* P_l, depending on whether they act on A from the left or the right. Two-sided preconditioning with nonsingular matrices P_l and P_r transforms implicitly Eq. (1) into a new linear problem

$$P_l A P_r \, z^* = P_l \boldsymbol{b}, \qquad \text{with} \quad \boldsymbol{x}^* = P_r z^*. \tag{10}$$

The preconditioned system can then be solved using arbitrary projection methods as described in Sect. 3.1, from the starting point z_0 defined by $\boldsymbol{x}_0 = P_r z_0$. The probabilistic view can be used to create a nuanced description of preconditioning as a form of prior information. In the SBI framework, Proposition 11 below shows that solving a right-preconditioned system is equivalent to modifying the prior, while Proposition 12 shows that left preconditioning is equivalent to making a different choice of observations.

Proposition 11 (Right preconditioning) *Consider the right-preconditioned system*

$$AP_r z^* = \boldsymbol{b} \quad \text{where} \quad \boldsymbol{x}^* = P_r z^*. \tag{11}$$

SBI on Eq. (11) *under the prior*

$$z \sim \mathcal{N}(z; z_0, \Sigma_0) \tag{12}$$

is equivalent to solving Eq. (1) *under the prior*

$$x \sim \mathcal{N}(x; P_r z_0, P_r \Sigma_0 P_r^\top).$$

Proof Let $p(x) = \mathcal{N}(x; x_0, \Sigma_r)$. Lemma 3 implies that after observing information from search directions S_m, the posterior mean equals

$$x_m = x_0 + \Sigma_r A^\top S_m (S_m^\top A \Sigma_r A^\top S_m)^{-1} S_m^\top r_0$$

where $r_0 = b - A x_0$. Setting $x_0 = P_r z_0$ and letting $\Sigma_r = P_r \Sigma_0 P_r^\top$ gives

$$x_m = P_r z_0 + P_r \Sigma_0 B^\top S_m (S_m^\top B \Sigma_0 B^\top S_m)^{-1} S_m^\top \hat{r}_0$$

where $B := A P_r$ and $\hat{r}_0 = b - B z_0$. Left multiplying by P_r^{-1} shows that this is equivalent to

$$\begin{aligned} z_m &:= P_r^{-1} x_m \\ &= z_0 + \Sigma_0 B^\top S_m (S_m^\top B \Sigma_0 B^\top S_m)^{-1} S_m^\top \hat{r}_0. \end{aligned}$$

Thus, z_m is the posterior mean of the system $B z^* = b$ with prior Eq. (12) after observing search directions S_m. $\square$

Proposition 12 (Left preconditioning) *Consider the left-preconditioned system*

$$P_l A x^* = P_l b \tag{13}$$

And the SBI prior

$$p(x) = \mathcal{N}(x; x_0, \Sigma_0).$$

Then, the posterior from SBI on Eq. (13) *under search directions S_m is equivalent to the posterior from SBI applied to the system Eq.* (1) *under search directions $P_l^\top S_m$.*

Proof Lemma 3 implies that after observing search directions T_m, the posterior mean over the solution of Eq. (1) equals

$$x_m = x_0 + \Sigma_0 A^\top T_m (T_m^\top A \Sigma_0 A^\top T_m)^{-1} T_m^\top r_0$$

where $r_0 = b - A x_0$. Setting $T_m = P_l^\top S_m$ gives

$$x_m = x_0 + \Sigma_0 B^\top S_m (S_m^\top B \Sigma_0 B^\top S_m)^{-1} S_m^\top P_l \hat{r}_0$$

where $B := P_l A$ and $\hat{r}_0 = P_l b - P_l A x_0$. Thus, x_m is the posterior mean of the system $B x^* = P_l b$ after observing search directions S_m. $\square$

If a probabilistic linear solver has a posterior mean which coincides with a projection method (as discussed in Sect. 3.1), the Propositions 11 and 12 show how to obtain a probabilistic interpretation of the *preconditioned* version of that algorithm. Furthermore, the equivalence demonstrated in Sect. 2.4 shows that the reasoning from Propositions 11 and 12 carries over to MBI based on left-multiplied observations: right preconditioning corresponds to a change in prior belief, while left-preconditioning corresponds to a change in observations.

We do not claim that this probabilistic interpretation of preconditioning is unique. For example, when using MBI with right-multiplied observations, the same line of reasoning can be used to show the converse: right preconditioning corresponds to a change in the observations and left preconditioning to a change in the prior.

5 Conjugate gradients

Conjugate gradients have been studied from a probabilistic point of view before by Hennig (2015) and Cockayne et al. (2018). This section generalises the results of Hennig (2015) and leverages Proposition 6 for new insights into BayesCG. For this section (but not thereafter), assume that A is a symmetric and positive definite matrix.

5.1 Left-multiplied view

The BayesCG algorithm proposed by Cockayne et al. (2018) encompasses conjugate gradients as a special case. BayesCG uses left-multiplied observations and was derived in the solution-based perspective.

The posterior in Lemma 3 does not immediately result in a practical algorithm as it involves the solution of a linear system based on the matrix $S_m^\top A \Sigma_0 A^\top S_m \in \mathbb{R}^{m \times m}$, which requires $\mathcal{O}(m^3)$ arithmetic operations. BayesCG avoids this cost by constructing search directions that are $A \Sigma_0 A^\top$-orthonormal, as shown below, see (Cockayne et al. 2018, Proposition 7).

Proposition 13 [Proposition 7 of Cockayne et al. 2018 (BayesCG)] *Let $\tilde{s}_1 = b - A x_0$, and let $s_1 = \tilde{s}_1 / \|\tilde{s}_1\|$. For $j = 2, \ldots, m$ let*

$$\tilde{s}_j = b - A x_{j-1} - \langle b - A x_{j-1}, s_{j-1} \rangle_{A \Sigma_0 A^\top} s_{j-1}$$
$$s_j = \tilde{s}_j / \|\tilde{s}_j\|_{A \Sigma_0 A^\top}.$$

Then, the set $\{s_1, \ldots, s_m\}$ is $A \Sigma_0 A^\top$-orthonormal, and consequently $S_m^\top A \Sigma_0 A^\top S_m = I$.

With these search directions constructed, BayesCG becomes an iterative method:

Proposition 14 (Proposition 6 of Cockayne et al. 2018) *Using the search directions from Proposition 13, the posterior from Lemma 3 reduces to:*

$$\boldsymbol{x}_m = \boldsymbol{x}_{m-1} + \Sigma_0 A^\top \boldsymbol{s}_m (\boldsymbol{s}_m^\top (\boldsymbol{b} - A\boldsymbol{x}_{m-1}))$$
$$\Sigma_m = \Sigma_{m-1} - \Sigma_0 A^\top \boldsymbol{s}_m \boldsymbol{s}_m^\top A \Sigma_0$$

In Proposition 4 of Cockayne et al. (2018), it was shown that the BayesCG posterior mean corresponds to the CG solution estimate when the prior covariance is taken to be $\Sigma_0 = A^{-1}$, though this is not a practical choice of prior covariance as it requires access to the unavailable A^{-1}. Furthermore, in Proposition 9 it was shown that when using the search directions from Proposition 13, the posterior mean from BCG has the following optimality property:

$$x_m = \arg\min_{\boldsymbol{x} \in K_m(\Sigma_0 A^\top A, \, \Sigma_0 A^\top \boldsymbol{b})} \|\boldsymbol{x} - \boldsymbol{x}^*\|_{\Sigma_0^{-1}}$$

Note that this is now a trivial special case of Proposition 4.

The following proposition leverages these results along with Proposition 6 to show that there exists an MBI method which, under a particular choice of prior and with a particular methodology for the generation of search directions, is consistent with CG.

Proposition 15 *Consider the MBI prior*

$$p(\overrightarrow{A^{-1}}) = \mathcal{N}(\overrightarrow{A^{-1}}; \overrightarrow{A_0^{-1}}, A^{-1} \otimes W_0)$$

where $W_0 \in \mathbb{R}^d$ is symmetric positive definite and so that $\boldsymbol{b}^\top W_0 \boldsymbol{b} = 1$. Suppose left-multiplied information is used, and that the search directions are generated sequentially according to:

$$\tilde{\boldsymbol{s}}_1 = (I - AA_0^{-1})\boldsymbol{b}$$
$$\boldsymbol{s}_1 = \frac{\tilde{\boldsymbol{s}}_1}{\|\tilde{\boldsymbol{s}}_1\|_A}$$

and for $j = 2, \ldots, m$

$$\tilde{\boldsymbol{s}}_j = (I - AA_{j-1}^{-1})\boldsymbol{b} - \boldsymbol{b}^\top (I - AA_{j-1}^{-1})^\top A\boldsymbol{s}_{j-1} \cdot \boldsymbol{s}_{j-1}$$
$$\boldsymbol{s}_j = \frac{\tilde{\boldsymbol{s}}_j}{\|\tilde{\boldsymbol{s}}_j\|_A}.$$

Then, it holds that the implied posterior mean on solution space, given by $A_m^{-1}\boldsymbol{b}$, corresponds to the CG solution estimate after m iterations, with starting point $\boldsymbol{x}_0 = A_0^{-1}\boldsymbol{b}$.

Proof First note that, by Proposition 6, since left-multiplied observations are used and since $\boldsymbol{b}^\top W_0 \boldsymbol{b} = 1$, the implied posterior distribution on solution space from MBI is identical to the posterior distribution from SBI under the prior

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; A_0^{-1}\boldsymbol{b}, A^{-1}).$$

It thus remains to show that the sequence of search directions generated is identical to those in Proposition 13 for this prior. For $\tilde{\boldsymbol{s}}_1$:

$$\tilde{\boldsymbol{s}}_1 = (I - AA_0^{-1})\boldsymbol{b} = \boldsymbol{b} - A\boldsymbol{x}_0$$

as required. For $\tilde{\boldsymbol{s}}_j$:

$$\begin{aligned}
\tilde{\boldsymbol{s}}_j &= (I - AA_{j-1}^{-1})\boldsymbol{b} - \boldsymbol{b}^\top (I - AA_{j-1}^{-1})^\top A\boldsymbol{s}_{j-1} \cdot \boldsymbol{s}_{j-1} \\
&= \boldsymbol{b} - A\boldsymbol{x}_{m-1} - (\boldsymbol{b} - A\boldsymbol{x}_{j-1})^\top A\boldsymbol{s}_{j-1} \cdot \boldsymbol{s}_{j-1} \\
&= \boldsymbol{b} - A\boldsymbol{x}_{m-1} - \langle \boldsymbol{b} - A\boldsymbol{x}_{j-1}, \boldsymbol{s}_{j-1} \rangle_A \cdot \boldsymbol{s}_{j-1}
\end{aligned}$$

where the second line uses that $A_{j-1}^{-1}\boldsymbol{b} = \boldsymbol{x}_{j-1}$. Thus, the search directions coincide with those in Proposition 13. It therefore holds that the implied posterior mean on solution space, $A_m^{-1}\boldsymbol{b}$, coincides with the solution estimate produced by CG.   □

5.2 Right-multiplied view

Interpretations of CG (and general projection methods) that use right-multiplied observations seem to require more care than those based on left-multiplied observations. Nevertheless, Hennig (2015) provided an interpretation for CG in this framework, essentially showing[7] that Algorithm 1 reproduces both the search directions and solution estimates from CG under the prior

$$p(A^{-1}) = \mathcal{N}(\overrightarrow{A^{-1}}; \overrightarrow{\alpha I}, \beta A^{-1} \otimes A^{-1}).$$

where $\alpha \in \mathbb{R} \setminus \{0\}$, $\beta \in \mathbb{R}^+$ and $\otimes$ denotes the symmetric Kronecker product (see Section A.1). The posterior under such a prior is described in Lemma 2.2 of Hennig (2015) (see Lemma 21), though we note that the sense in which the solution estimate $\boldsymbol{x}_m$ output by this algorithm is related to the posterior over A^{-1} differs from that in the previous section, in the sense that $A_m^{-1}\boldsymbol{b} \neq \boldsymbol{x}_m$. (More precisely, $\boldsymbol{x}_m = A_m^{-1}(\boldsymbol{b} - A\boldsymbol{x}_0) - \boldsymbol{x}_0 - (1 - \alpha_m)\boldsymbol{d}_m$, as the CG estimate is corrected by the step size computed in line 6. Fixing this rank-1 discrepancy would complicate the exposition of Algorithm 1 and yield a more cumbersome algorithm.) The following proposition generalises this result.

Proposition 16 *Consider the prior*

$$p(A^{-1}) = \mathcal{N}(\overrightarrow{A^{-1}}; \overrightarrow{\alpha I}, (\beta I + \gamma W) \otimes (\beta I + \gamma W))$$

where $W := A^{-1}$. For all choices $\alpha \in \mathbb{R} \setminus \{0\}$ and $\beta, \gamma \in \mathbb{R}_{+,0}$ with $\beta + \gamma > 0$, Algorithm 1 is equivalent to CG, in the sense that it produces the exact same sequence of estimates $\boldsymbol{x}_i$ and scaled search directions $\boldsymbol{s}_i$.

---

[7] Algorithm 1 is not included in this form in the op.cit.

Proof The proof is extensive and has been moved to "Appendix B". □

Algorithm 1 The algorithm referred to by Proposition 16, which reproduces the search directions and solution estimates from CG.

1 $x_0 \leftarrow A_0^{-1} b$	// initial guess
2 $r_0 \leftarrow A x_0 - b$	
3 **for** $i = 1, \ldots, m$ **do**	
4 $\quad d_i \leftarrow -A_{i-1}^{-1} r_{i-1}$	// compute optimisation direction
5 $\quad z_i \leftarrow A d_i$	// **observe**
6 $\quad \alpha_i \leftarrow -\dfrac{d_i^\top r_{i-1}}{d_i^\top z_i}$	// optimal step size
7 $\quad s_i \leftarrow \alpha_i d_i$	// re-scale step
8 $\quad y_i \leftarrow \alpha_i z_i$	// re-scale observation
9 $\quad x_i \leftarrow x_{i-1} + s_i$	// update estimate for x
10 $\quad r_i \leftarrow r_{i-1} + y_i$	// new gradient at x_i
11 $\quad A_i^{-1} \leftarrow \mathbb{E}_{p(A^{-1}|S,Y)} A^{-1}$	// estimate A^{-1}
12 **end for**	
13 **return** x_m	

Note that, unlike previous propositions, Proposition 16 proposes a prior that does not involve A^{-1} for the case when $\gamma = 0$.

6 GMRES

The *generalised minimal residual method* (Saad 2003, Section 6.5) applies to general nonsingular matrices A. At iteration m, GMRES minimises the residual over the affine space $x_0 + K_m(A, r_0)$. That is, $r_m = r_0 - A x_m$ satisfies

$$\|r_m\|_2 = \min_{x \in K_m(A,r_0)} \|Ax - r_0\|_2$$
$$= \min_{x \in x_0 + K_m(A,r_0)} \|Ax - b\|_2. \qquad (14)$$

Since $Ax - b = A(x - x^*)$, this corresponds to minimising the error in the $A^\top A$ norm.

We present a brief development of GMRES, starting with Arnoldi's method (Sect. 6.1) and the GMRES algorithm (Sect. 6.2), before presenting our Bayesian interpretation (Sect. 6.3).

6.1 Arnoldi's method

GMRES uses Arnoldi's method (Saad 2003, Section 6.3) to construct orthonormal bases for Krylov spaces of general, nonsingular matrices A. Starting with $q_1 = r_0/\|r_0\|_2$, Arnoldi's method recursively computes the orthonormal basis

$$Q_m = [q_1 \ldots q_m] \in \mathbb{R}^{d \times m}$$

for $K_m(A, r_0)$. The basis vectors satisfy the relations

$$A Q_m = Q_{m+1} \tilde{H}_m = Q_m H_m + h_{m+1,m} q_{m+1} e_m^\top \qquad (15)$$

and $Q_m^\top A Q_m = H_m$, where the *upper Hessenberg* matrix H_m is defined as

$$H_m = \begin{bmatrix} h_{11} & h_{12} & h_{13} & \ldots & h_{1,m-1} & h_{1m} \\ h_{21} & h_{22} & h_{23} & \ldots & h_{2,m-1} & h_{2m} \\ 0 & h_{32} & h_{33} & \ldots & h_{3,m-1} & h_{3m} \\ \vdots & 0 & h_{43} & \ldots & h_{4,m-1} & h_{3m} \\ \vdots & & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & \ldots & 0 & h_{m,m-1} & h_{mm} \end{bmatrix} \in \mathbb{R}^{m \times m}$$

and

$$\tilde{H}_m = \begin{bmatrix} H_m \\ h_{m+1,m} e_m^\top \end{bmatrix} \in \mathbb{R}^{(m+1) \times m}.$$

6.2 GMRES

GMRES computes the iterate

$$x_m = x_0 + Q_m c_m$$

based on the optimality condition in Eq. (14), which can equivalently be expressed as

$$c_m = \arg\min_{c \in \mathbb{R}^m} \|A Q_m c - r_0\|_2$$
$$= \left((A Q_m)^\top (A Q_m) \right)^{-1} (A Q_m)^\top r_0. \qquad (16)$$

Thus,

$$x_m = x_0 + Q_m \left(Q_m^\top A^\top A Q_m \right)^{-1} Q_m^\top A^\top r_0, \qquad (17)$$

confirming that GMRES is a projection method with $X_m = Q_m$ and $U_m = A Q_m$.

GMRES solves the least squares problem in Eq. (16) efficiently by projecting it to a lower-dimensional space via Arnoldi's method. To this end, express the starting vector in the Krylov basis,

$$r_0 = \|r_0\|_2 q_1 = \|r_0\|_2 Q_{m+1} e_1,$$

and exploit the Arnoldi recursion from Eq. (15),

$$A Q_m c - r_0 = Q_{m+1} \left(\tilde{H}_{m+1} c - \|r_0\|_2 e_1 \right),$$

followed by the unitary invariance of the two-norm,

$$\|A Q_m c - r_0\|_2 = \|\tilde{H}_m c - \|r_0\|_2 e_1\|_2.$$

Thus, instead of solving the least squares problem Equation (16) with d rows, GMRES solves instead a problem with only $m + 1$ rows,

$$c_m = \arg\min_{c \in \mathbb{R}^m} \| \tilde{H}_m c - \| r_0 \|_2 e_1 \|_2. \qquad (18)$$

The computations are summarised in Algorithm 2.

Algorithm 2 GMRES (Saad 2003, Algorithm 6.9)

---
1 $r_0 \leftarrow b - A x_0$, $\beta \leftarrow \| r_0 \|_2$, $q_1 \leftarrow r_0 / \beta$
2 **for** $j = 1, \dots, m$ **do**
3 | $w_j \leftarrow A q_j$
4 | **for** $i = 1, \dots, j$ **do**
5 | | $h_{ij} \leftarrow \langle w_j, A q_i \rangle$
6 | | $w_j \leftarrow w_j - h_{ij} q_i$
7 | **end for**
8 | $h_{j+1,j} \leftarrow \| w_j \|_2$
9 | **if** $h_{j+1,j} = 0$ **then**
10 | | $m \leftarrow j$, go to 14
11 | **end if**
12 | $q_{j+1} \leftarrow w_j / h_{j+1,j}$
13 **end for**
14 Define $\tilde{H}_m \in \mathbb{R}^{(m+1) \times m}$ with elements h_{ij}
15 $c_m \leftarrow \arg\min_c \| \tilde{H}_m c - \beta e_1 \|_2$
16 $x_m \leftarrow x_0 + Q_m c_m$

6.3 Bayesian interpretation of GMRES

We now present probabilistic linear solvers with posterior means that coincide with the solution estimate from GMRES.

6.3.1 Left-multiplied view

Proposition 17 *Under the SBI prior*

$$p(x) = \mathcal{N}(x; x_0, \Sigma_0) \quad where \quad \Sigma_0 = (A^\top A)^{-1}$$

and the search directions $U_m = A Q_m$, the posterior mean is identical to the GMRES iterate x_m in Eq. (17).

Proof Substitute $R = A$ and $U_m = A Q_m$ into Proposition 9. □

Proposition 17 is intuitive in the context of Proposition 4: setting $\Sigma_0 = (A^\top A)^{-1}$ ensures that the norm being minimised coincides with that of GMRES, as does the solution space $X_m = A Q_m$. This interpretation exhibits an interesting duality with CG for which $\Sigma_0 = A^{-1}$.

Another probabilistic interpretation follows from Proposition 8.

Corollary 18 *Under the prior*

$$p(x) = \mathcal{N}(x; x_0, \Sigma_0) \quad where \quad \Sigma_0 = Q_m Q_m^\top, \qquad (19)$$

and with observations $y_m = Q_m^\top b$, the posterior mean from SBI is identical to the GMRES iterate x_m in Eq. (17).

Note that Proposition 17 has a posterior covariance which is not practical, as it involves A^{-1}. (Cockayne et al. 2017) proposed replacing A^{-1} in the prior covariance with a preconditioner to address this, which does yield a practically computable posterior, but this extension was not explored here. Furthermore, that approach yields unsatisfactorily calibrated posterior uncertainty, as described in that work. Corollary 18 does not have this drawback, but the posterior covariance is a matrix of zeroes.

6.3.2 Right-multiplied view

As for CG in Sect. 5.2, finding interpretations of GMRES that use right-multiplied observations appears to be more difficult.

Proposition 19 *Under the prior*

$$p(A^{-1}) = \mathcal{N}(0, \Sigma \otimes I) \qquad (20)$$

and given $Y_m = A Q_m$, the implied posterior mean on the solution space given by $A_m^{-1} b$ is equivalent to the GMRES solution. This correspondence breaks when $x_0 \neq 0$.

Proof Under this prior, b applied to the posterior mean is

$$\begin{aligned}
A_m^{-1} b &= A_0^{-1} b + (Q_m - A_m^{-1} Y_m)(Y_m^\top Y_m)^{-1} Y_m^\top b \\
&= Q_m (Y_m^\top Y_m)^{-1} Y_m^\top b \\
&= Q_m (Q_m^\top A^\top A Q_m)^{-1} Q_m^\top A^\top b
\end{aligned}$$

which is the GMRES projection step if $x_0 = 0$. □

6.4 Simulation study

In this section, the simulation study of Cockayne et al. (2018) will be replicated to demonstrate that the uncertainty produced from GMRES in Proposition 17 is similarly poorly calibrated to CG, owing to the dependence of Q_m on x^* by way of its dependence on b. Throughout the size of the test problems is set to $d = 100$. The eigenvalues of A were drawn from an exponential distribution with parameter $\gamma = 10$ and eigenvectors uniformly from the Haar-measure over rotation matrices (see Diaconis and Shahshahani 1987). In contrast to Cockayne et al. (2018), the entries of b are drawn from a standard Gaussian distribution, rather than x_*. By Lemma 1, the prior is then perfectly calibrated for this scenario, providing justification for the expectation that the posterior should be equally well calibrated for $m \geq 1$.

Cockayne et al. (2018) argue that if the uncertainty is well calibrated, then x^* can be considered as a draw from the

posterior. Under this assumption, i.e. $\Sigma_m^{-1/2}(x^* - x_m) \sim \mathcal{N}(0, I)$ they derive the test statistic:

$$Z(x^*) := \|\Sigma_m^{-1/2}(x^* - x_m)\| \sim \chi_{d-m}^2.$$

Figure 1 shows on the left the convergence of GMRES and on the right the convergence rate of the trace of the posterior covariance. Figure 2 displays the test statistic. It can be

seen that the same poor uncertainty quantification occurs in BayesGMRES; even after just 10 iterations, the empirical distribution of the test statistic exhibits a profound left shift, indicating an overly conservative posterior distribution. Producing well-calibrated posteriors remains an open issue in the field of probabilistic linear solvers.
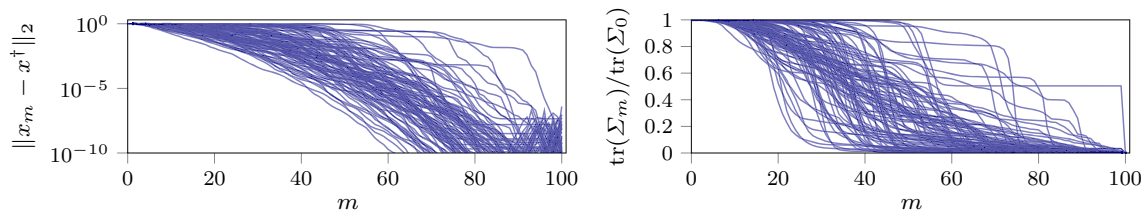
Fig. 1 Convergence of posterior mean and variance of the probabilistic interpretation of GMRES from Proposition 17
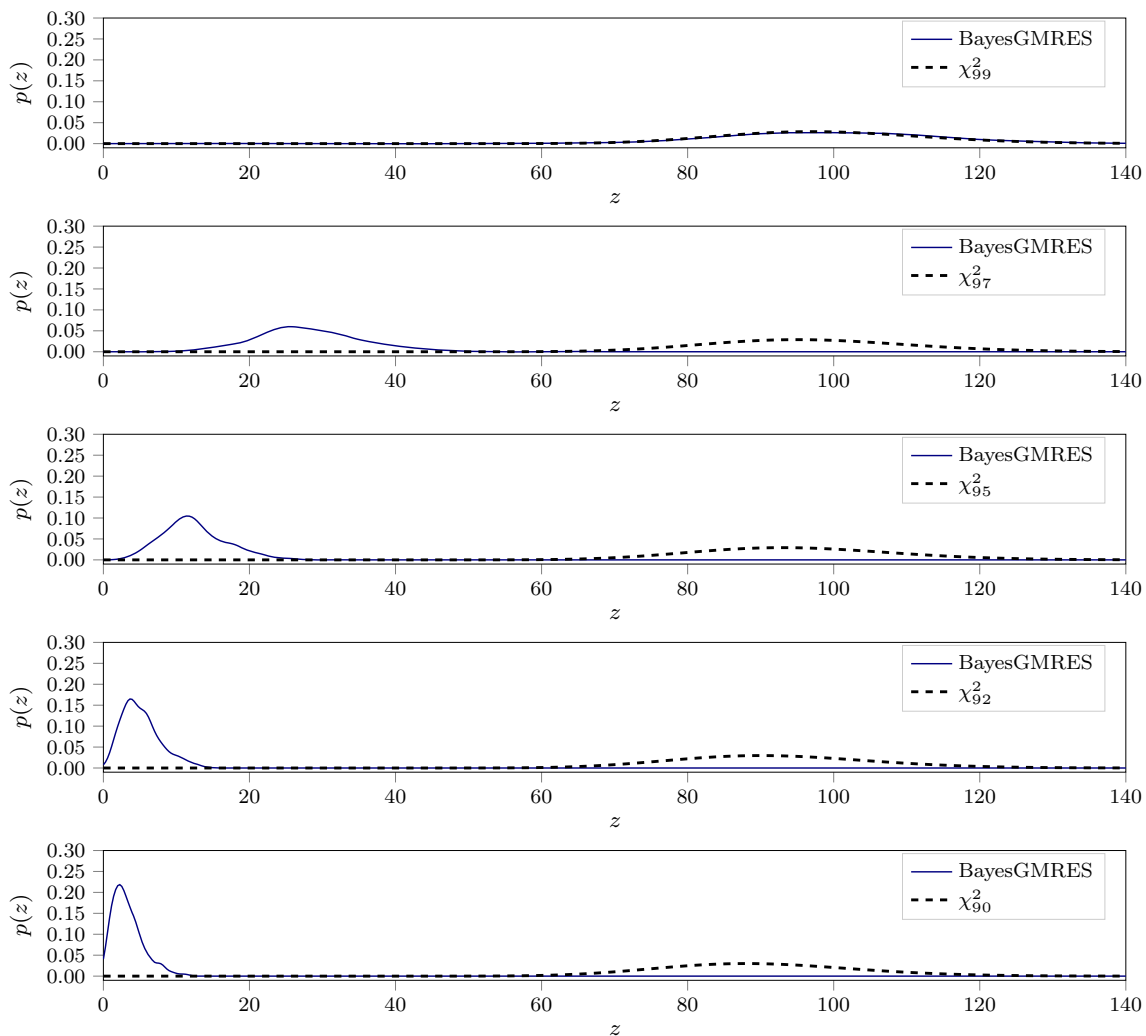
Fig. 2 Assessment of the uncertainty quantification. Plotted are kernel density estimates for the statistic Z based on 500 randomly sampled test problems for steps $m = \{1, 3, 5, 8, 10\}$. These are compared with the theoretical distribution of Z when the posterior distribution is well calibrated

7 Discussion

We have established many new connections between probabilistic linear solvers and a broad class of iterative methods. Matrix-based and solution-based inference were shown to be equivalent in a particular regime, showing that results from SBI transfer to MBI with left-multiplied observations. Since SBI is a special case of MBI, future research will establish what additional benefits the increased generality of MBI can provide.

We also established a connection between the wide class of projection methods and probabilistic linear solvers. The common practice of preconditioning has an intuitive probabilistic interpretation, and all probabilistic linear solvers can be interpreted as projection methods. While the converse was shown to hold, the conditions under which generic projection methods can be reproduced are somewhat restrictive; however, GMRES and CG, which are among the most commonly used projection methods, have a well-defined probabilistic interpretation. Probabilistic interpretations of other widely used iterative methods can, we anticipate, be established from the results presented in this work.

Posterior uncertainty remains a challenge for probabilistic linear solvers. Direct probabilistic interpretations of CG and GMRES yield posterior covariance matrices which are not always computable, and even when the posterior can be computed, the uncertainty remains poorly calibrated. This is owed to the dependence of the search directions in Krylov methods on $Ax^* = b$, resulting in an algorithm which is not strictly Bayesian. Mitigating this issue without sacrificing the fast rate of convergence provided by Krylov methods remains an important focus for future work.

Appendix A: Properties of Kronecker products

The following identities about Kronecker products and the vectorisation operator are easily derived, but recalled here for the convenience of the reader:

$$(A \otimes B) \overrightarrow{C} = \overrightarrow{ACB^\top} \tag{K1}$$

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \tag{K2}$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \tag{K3}$$

$$(A \otimes B)^\top = A^\top \otimes B^\top \tag{K4}$$

$$(A + B) \otimes C = A \otimes C + B \otimes C \tag{K5}$$

A.1 The symmetric Kronecker product

Definition 20 (*symmetric Kronecker product*) The *symmetric* Kronecker product for two square matrices $A, B \in \mathbb{R}^{N \times N}$ of equal size is defined as

$$A \underline{\otimes} B := \Gamma (A \otimes B) \Gamma$$

where $[\Gamma]_{ij,kl} := \tfrac{1}{2}\delta_{ik}\delta_{jl} + \tfrac{1}{2}\delta_{il}\delta_{jk}$ satisfies

$$\Gamma \overrightarrow{C} = \tfrac{1}{2}\overrightarrow{C} + \tfrac{1}{2}\overrightarrow{C^\top}$$

for all square matrices $C \in \mathbb{R}^{N \times N}$.

Proposition 21 (Theorem 2.3 in Hennig 2015) *Let* $W \in \mathbb{R}^{d \times d}$ *be symmetric and positive definite. Assume a Gaussian prior of symmetric mean* A_0^{-1} *and covariance* $W \underline{\otimes} W$ *on the elements of a symmetric matrix* A^{-1}. *After* m *linearly independent noise-free observations of the form* $S = A^{-1}Y$, $Y \in \mathbb{R}^{d \times m}$, $\mathrm{rk}(Y) = m$, *the posterior belief over* A^{-1} *is a Gaussian with mean*

$$\begin{aligned}
A_m^{-1} = {} & A_0^{-1} + (S - A_0^{-1}Y)GY^\top W \\
& + WYG(S - A_0^{-1}Y)^\top \\
& + WYGY^\top(S - A_0^{-1}Y)GY^\top W
\end{aligned} \tag{21}$$

and posterior covariance

$$V_m = (W - WYGY^\top W) \underline{\otimes} (W - WYGY^\top W) \tag{22}$$

where $G := (Y^\top WY)^{-1}$.

Remark 22 Since A_0^{-1} is symmetric and the symmetric prior places mass only on symmetric matrices, the posterior mean A_m^{-1} is also symmetric.

Appendix B: Proofs

B.1 Proposition 6

Proof of Proposition 6 Let $H = A^{-1}$ and let $A_0^{-1} = H_0$. First note that by right-multiplying the information in Eq. (6) by H:

$$Y_m^\top H = S_m^\top$$

$$\implies \overrightarrow{Y_m^\top H} = \overrightarrow{S_m^\top}$$
$$\implies (Y_m \otimes I)\overrightarrow{H} = \overrightarrow{S_m^\top} \quad \text{(from K1)}$$

Now the implied posterior on $\overrightarrow{H}$ can be computed using the standard laws of Gaussian conditioning:

$$\overrightarrow{H} \sim \mathcal{N}(\overrightarrow{H_0}, \Sigma \otimes W)$$
$$\implies \overrightarrow{H}|\mathcal{D} \sim \mathcal{N}(\overrightarrow{H_m}, \Omega_m).$$

Let $\Omega_0 = \Sigma_0 \otimes W_0$ and let $P = Y_m^\top \otimes I$. Then,

$$\overrightarrow{H_m} = \overrightarrow{H_0} + [P\Omega_0]^\top [P\Omega_0 P^\top]^{-1}(\overrightarrow{S_m^\top} - \overrightarrow{Y_m^\top H_0})$$
$$\Omega_m = \Omega_0 - [P\Omega_0]^\top [P\Omega_0 P^\top]^{-1}(P\Omega_0)$$

Now note that

$$P\Omega_0 = (Y_m^\top \otimes I)(\Sigma_0 \otimes W)$$
$$= (Y_m^\top \Sigma_0) \otimes W$$
$$\implies (P\Omega_0)^\top = (\Sigma_0 Y_m) \otimes W$$

where the second line uses Eq. (K2) and the third uses Eq. (K4). Thus,

$$P\Omega_0 P^\top = (Y_m^\top \otimes I)(\Sigma_0 \otimes W_0)(Y_m^\top \otimes I)^\top$$
$$= (Y_m^\top \Sigma_0 Y_m) \otimes W_0$$
$$\implies (P\Omega_0 P^\top)^{-1} = (Y_m^\top \Sigma_0 Y_m)^{-1} \otimes W_0^{-1}$$

where the second line is again using Eq .(K2) and Eq. (K4), while the third line uses Eq. (K3). We conclude that

$$(P\Omega_0)^\top (P\Omega_0 P^\top)^{-1}$$
$$= [(\Sigma_0 Y_m) \otimes W][(Y_m^\top \Sigma_0 Y_m)^{-1} \otimes W^{-1}]$$
$$= (\Sigma_0 Y_m (Y_m^\top \Sigma_0 Y_m)^{-1}) \otimes I$$
$$\implies (P\Omega_0)^\top (P\Omega_0 P^\top)^{-1}(P\Omega_0)$$
$$= (\Sigma_0 Y_m (Y_m^\top \Sigma_0 Y_m)^{-1} Y_m^\top \Sigma_0) \otimes W_0.$$

From these expressions, it is straightforward to simplify the expressions for $\overrightarrow{H_m}$:

$$\overrightarrow{H_m} = \overrightarrow{H_0} + (\Sigma_0 Y_m (Y_m^\top \Sigma_0 Y_m)^{-1} \otimes I)(\overrightarrow{S_m^\top} - \overrightarrow{Y_m^\top H_0})$$
$$= \text{vec}\left(H_0 + \Sigma_0 Y^m (Y_m^\top \Sigma_0 Y_m)^{-1}(S_m^\top - Y_m^\top H_0) \right)$$

where the last line follows from K1. For Ω_m:

$$\Omega_m = \Sigma_0 \otimes W - (\Sigma_0 Y_m (Y_m^\top \Sigma_0 Y_m)^{-1} Y_m^\top \Sigma_0) \otimes W_0$$

$$= (\Sigma_0 - \Sigma_0 Y_m (Y_m^\top \Sigma_0 Y_m)^{-1} Y_m^\top \Sigma_0) \otimes W_0$$

where the last line is from application of K5.

It remains to project the posterior into $\mathbb{R}^d$ by performing the matrix-vector product $H\boldsymbol{b}$.

$$\boldsymbol{x} = \overrightarrow{H\boldsymbol{b}} = (I \otimes \boldsymbol{b}^\top)H. \quad \text{(from K1)}$$

Thus, the implied posterior is $\boldsymbol{x} \sim \mathcal{N}(\bar{\boldsymbol{x}}_m, \bar{\Sigma}_m)$, with

$$\bar{\boldsymbol{x}}_m = (I \otimes \boldsymbol{b}^\top)\text{vec}$$
$$\left(H_0 + \Sigma_0 Y_m (Y_m^\top \Sigma_0 Y_m)^{-1}(S_m^\top - Y_m^\top H_0) \right)$$
$$= \text{vec}\left(H_0 \boldsymbol{b} + \Sigma_0 Y_m (Y_m^\top \Sigma_0 Y_m)^{-1}(S_m^\top \boldsymbol{b} - Y_m^\top H_0 \boldsymbol{b}) \right)$$
$$= \boldsymbol{x}_0 + \Sigma_0 A^\top S_m (S_m^\top A \Sigma_0 A^\top S_m)^{-1} S_m^\top (\boldsymbol{b} - A\boldsymbol{x}_0)$$

where in the last line we have used that $H_0 \boldsymbol{b} = \boldsymbol{x}_0$ and that $Y_m = A^\top S_m$. Furthermore,

$$\bar{\Sigma}_m = (I \otimes \boldsymbol{b}^\top)$$
$$\cdot \left[(\Sigma_0 - \Sigma_0 Y_m (Y_m^\top \Sigma_0 Y_m)^{-1} Y_m^\top \Sigma_0) \otimes W_0 \right]$$
$$\cdot (I \otimes \boldsymbol{b}^\top)^\top$$
$$= (\Sigma_0 - \Sigma_0 Y_m (Y_m^\top \Sigma_0 Y_m)^{-1} Y_m^\top \Sigma_0) \times \boldsymbol{b}^\top W_0 \boldsymbol{b}$$
$$= \Sigma_0 - \Sigma_0 A^\top S_m (S_m^\top A \Sigma_0 A^\top S_m)^{-1} S_m^\top A \Sigma_0$$

where in the second line we have used K2 and the fact that $\boldsymbol{b}^\top W_0 \boldsymbol{b}$ is a scalar, while in the third line we have used that $\boldsymbol{b}^\top W_0 \boldsymbol{b} = 1$ and that $Y_m = A^\top S_m$.

Note that $\boldsymbol{x}_m = \bar{\boldsymbol{x}}_m$ and $\Sigma_m = \bar{\Sigma}_m$, as defined in Cockayne et al. (2018). Thus, the proof is complete. $\square$

B.2 Theorem 16

Proof of Theorem 16. Denote by $\boldsymbol{x}_i^{CG}$ the conjugate gradient estimate in iteration i and with $\boldsymbol{p}_i$ the search direction in that iteration. From one iteration to the next, the update to the solution can be written as (Nocedal and Wright 1999, p. 108)

$$\boldsymbol{x}_{i+1}^{CG} = \boldsymbol{x}_i^{CG} + \frac{\boldsymbol{r}_i^\top \boldsymbol{p}_i}{\boldsymbol{p}_i^\top A \boldsymbol{p}_i} \boldsymbol{p}_i. \tag{K23}$$

Comparing this update to lines 7 to 10 in Algorithm 1, it is sufficient to show that $\boldsymbol{d}_i \propto \boldsymbol{p}_i$ which follows from Lemma 23. $\square$

Lemma 23 *Assume that CG does not terminate before d iterations. Using the prior of Theorem 16 in Algorithm 1, the directions $\boldsymbol{d}_i$ are scaled conjugate gradients search directions, i.e.*

$$\boldsymbol{d}_i = \gamma_i \boldsymbol{p}_i^{CG}$$

where p_i^{CG} is the CG search direction in iteration i and $\gamma_i \in \mathbb{R} \setminus \{0\}$.

Proof The proof proceeds by induction. Throughout we will suppress the superscript CG on the CG search directions, i.e. $p_i^{CG} = p_i$. For $i = 1$, $A_{i-1}^{-1} = \alpha I$ by assumption and therefore $d_i = \alpha r_0$ which is the first CG search direction scaled by $\gamma_1 = \alpha \neq 0$.

For the inductive step, suppose that the search directions $s_1, \ldots, s_{i-1}$ are scaled CG directions and that the vectors $x_1, \ldots, x_{i-1}$ are the same as the first $i - 1$ solution estimates produced by CG. We will prove that s_i is the ith CG search direction, and that x_i is the ith solution estimate from CG. Lemma 25 states that d_i can be written as

$$d_i = A_{i-1}^{-1} r_{i-1} = \sum_{j<i} v_j s_j + v_i r_{i-1}. \tag{K24}$$

where $v_j \in \mathbb{R}$, $j = 1, \ldots, i$. Under the prior, the posterior mean A_i^{-1} is always symmetric as stated in Remark 22. This allows application of Lemma 24, so that $\{s_1, \ldots, s_{i-1}, d_i\}$ is an A-conjugate set. Thus, we have, for $\ell < i$:

$$0 = s_\ell^\top A d_i = v_\ell s_\ell^\top A s_\ell + v_i s_\ell^\top A r_{i-1}$$
$$= v_\ell s_\ell^\top A s_\ell + v_i y_\ell^\top r_{i-1}. \tag{K25}$$

Now note that

$$y_\ell^\top r_{i-1} = (r_\ell - r_{\ell-1})^\top r_{i-1}.$$

This follows from line 10 of Algorithm 1, from which it is clear that $y_\ell = r_\ell - r_{\ell-1}$. Recall that the CG residuals r_j are orthogonal (Nocedal and Wright 1999, p. 109) and that from the inductive assumption, Algorithm 1 is equivalent to CG up to iteration $i - 1$). Thus, for $\ell < i - 1$ we have that

$$y_\ell^\top r_{i-1} = 0$$
$$\implies s_\ell A d_i = v_\ell s_\ell^\top A s_\ell = 0 \quad \forall \ell < i - 1$$

where the second line is from application of the first line in Eq. (K25). However, A is positive definite and by assumption the algorithm has not converged, so $d_\ell \neq 0$. Furthermore, clearly $s_\ell^\top A s_\ell \neq 0$. Hence, we must have that

$$v_\ell = 0 \quad \forall j < i - 1.$$

Equation (K24) thus simplifies to

$$d_i = v_{i-1} s_{i-1} + v_i r_{i-1} = v_{i-1} \alpha_{i-1} d_{i-1} + v_i r_{i-1}. \tag{K26}$$

Now, again by Lemma 24, d_i must be conjugate to s_{i-1} which implies $v_i \neq 0$. Pre-multiplying Eq. (K26) by $s_{i-1}^\top A$ gives

$$0 = v_{i-1} \alpha_{i-1} s_{i-1}^\top A d_{i-1} + v_i s_{i-1}^\top A r_{i-1}$$
$$\implies v_{i-1} \alpha_{i-1} = -v_i \frac{s_{i-1}^\top A r_{i-1}}{s_{i-1}^\top A d_{i-1}}.$$

Thus, d_i can be written as

$$d_i = v_i \left(r_{i-1} - \frac{s_{i-1}^\top A r_{i-1}}{s_{i-1}^\top A d_{i-1}} d_{i-1} \right)$$
$$= v_i \left(r_{i-1} - \frac{p_{i-1}^\top A r_{i-1}}{p_{i-1}^\top A p_{i-1}} p_{i-1} \right) \tag{K27}$$

where the second line again applies the inductive assumption that d_{i-1} and s_{i-1} are proportional to the CG search direction p_{i-1}, noting that the proportionality constants on numerator and denominator cancel. The term inside the brackets is precisely the ith CG search direction. This completes the result. $\square$

Lemma 24 *If the belief over A_m^{-1} is symmetric for all $m = 0, \ldots, d$ and A is symmetric and positive definite, then Algorithm 1 produces A-conjugate directions.*

Proof The proof is by induction. Note that the case $i = 1$ is irrelevant since a set consisting of one element is trivially A-conjugate. On many occasions, the proof relies on the consistency of the MBI belief, i.e. $A_i^{-1} z_k = d_k$ for $k \leq i$ and by symmetry $z_k^\top A_i^{-1} = d_k^\top$. Thus, for the base case $i = 2$ we have:

$$d_1^\top A d_2 = -d_1^\top A(A_1^{-1} r_1)$$
$$= -d_1^\top A(A_1^{-1}(y_1 + r_0))$$
$$= -d_1^\top A(s_1 + A_1^{-1} r_0)$$

where the second line is by line 10 of Algorithm 1. Now recall that $\alpha_1 = -d_1^\top r_0 / d_1^\top A d_1$ to give:

$$d_1^\top A d_2 = -\alpha_1 d_1^\top A d_1 - d_1^\top A A_1^{-1} r_0$$
$$= d_1^\top r_0 - d_1^\top A A_1^{-1} r_0$$
$$= d_1^\top r_0 - z_1^\top A_1^{-1} r_0$$
$$= d_1^\top r_0 - d_1^\top r_0$$
$$= 0. \tag{K28}$$

Here, the symmetry of the estimator A_i^{-1} is used in Eq. (K28). For the inductive step, assume $\{d_0, \ldots, d_{i-1}\}$ are pairwise A-conjugate. For any $k < i$, we have:

$$d_k^\top A d_i = -d_k^\top A(A_i^{-1} r_i)$$
$$= -d_k^\top A A_i^{-1} \left(\sum_{j \leq i} y_j + r_0 \right)$$

where the second line follows from the fact that $r_i = r_{i-1} + y_i$. Thus, we have:

$$
\begin{aligned}
d_k^\top A d_i &= -d_k^\top A \left(\sum_{j \le i} s_j + A_i^{-1} r_0 \right) \\
&= -d_k^\top A \left(\sum_{j \le i} \alpha_j d_j + A_i^{-1} r_0 \right).
\end{aligned}
$$

Now, applying the conjugacy from the inductive assumption:

$$
\begin{aligned}
d_k^\top A d_i &= -\alpha_k d_k^\top A d_k - d_k^\top A (A_i^{-1} r_0) \\
&= d_k^\top r_{k-1} - d_k^\top r_0 \\
&= d_k^\top \left(\sum_{j < k} y_j + r_0 \right) - d_k^\top r_0 = 0 \\
&= \sum_{j < k} \alpha_j d_k^\top A d_j = 0.
\end{aligned}
$$

where the second line rearranges line 6 of the algorithm to obtain $\alpha_i d_i^\top z_i = -d_i^\top r_{i-1}$. The third line again uses that $r_i = r_{i-1} + y_i$, while the fourth line is from the assumed conjugacy. $\square$

Lemma 25 *Under the prior in Theorem* 16 *and given scaled CG search directions* $p_1, ..., p_i$, *it holds that* $A_i^{-1} r_i \in$ span$\{p_1, ..., p_i, r_i\}$.

Proof Recall first that under the prior in Theorem 16, $A_0^{-1} = \alpha I$. Then, by inspection of Eq. (21) we have $A_i^{-1} r_i \in \mathcal{S}$ where

$$
\mathcal{S} = \text{span}\{r_i, p_1, ..., p_i, y_1, ..., y_i, W y_1, ..., W y_i\}
$$

By choice of $W = \beta I + \gamma A^{-1}$, $\mathcal{S} = \text{span}\{r_i, p_1, ..., p_i, y_1, ..., y_i\}$. From line 10 of Algorithm 1, $y_i = r_i - r_{i-1}$ and therefore $\mathcal{S} = \text{span}\{r_1, ..., r_i, p_1, ..., p_i\}$. By Theorem 5.3 in (Nocedal and Wright 1999, p. 109), the span of the conjugate gradients residuals and search directions are equivalent. Therefore, $\mathcal{S} \subseteq \{r_i, p_1, ..., p_i\}$. $\square$

References

Bartels, S., Hennig, P.: Probabilistic approximate least-squares. In: Proceedings of Artificial Intelligence and Statistics (AISTATS) (2016)

Cockayne, J., Oates, C., Sullivan, T.J., Girolami, M.: Probabilistic numerical methods for partial differential equations and Bayesian inverse problems. arXiv:1605.07811 (2016)

Cockayne, J., Oates, C., Sullivan, T.J., Girolami, M.: Bayesian probabilistic numerical methods. 1702.03673 (2017)

Cockayne, J., Oates, C., Ipsen, I.C.F., Girolami, M.: A Bayesian conjugate gradient method. arXiv:1801.05242 (2018)

Diaconis, P., Shahshahani, M.: The subgroup algorithm for generating uniform random variables. Probab. Eng. Inf. Sci. **1**(01), 15 (1987). https://doi.org/10.1017/s0269964800000255

Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences, 4th edn. Johns Hopkins University Press, Baltimore (2013)

Hennig, P.: Probabilistic interpretation of linear solvers. SIAM J. Optim. **25**(1), 234–260 (2015). https://doi.org/10.1137/140955501

Hennig, P., Osborne, M.A., Girolami, M.: Probabilistic numerics and uncertainty in computations. Proc. R. Soc. Lond. A Math. Phys. Eng. Sci. **471**, 20150142 (2015)

Karvonen, T., Sarkka, S.: Classical quadrature rules via gaussian processes. In: IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE (2017). https://doi.org/10.1109/mlsp.2017.8168195

Kersting, H., Sullivan, T.J., Hennig, P.: Convergence rates of Gaussian ODE filters. arXiv:1807.09737, 7 (2018)

Liesen, J., Strakos, Z.: Krylov Subspace Methods. Principles and Analysis. Oxford University Press, Oxford (2012). https://doi.org/10.1093/acprof:oso/9780199655410.001.0001

Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, Berlin (1999)

Saad, Y.: Iterative Methods for Sparse Linear Systems. Society for Industrial and Applied Mathematics, 2nd edn. SIAM, Philadelphia (2003)

Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**(3), 856–869 (1986). https://doi.org/10.1137/0907058

Schober, M., Duvenaud, D., Hennig, P.: Probabilistic ODE solvers with Runge–Kutta means. In: Advances in Neural Information Processing Systems, vol. 27, pp. 739–747. Curran Associates, Inc. (2014). http://papers.nips.cc/paper/5451-probabilistic-ode-solvers-with-runge-kutta-means.pdf

Schober, M., Särkkä, S., Hennig, P.: A probabilistic model for the numerical solution of initial value problems. Stat. Comput. **29**(1), 99–122 (2019)

Soodhalter, K.M., Szyld, D.B., Xue, F.: Krylov subspace recycling for sequences of shifted linear systems. Appl. Numer. Math. **81**, 105–118 (2014). https://doi.org/10.1016/j.apnum.2014.02.006

Xi, X., Briol, F.-X., Girolami, M.: Bayesian quadrature for multiple related integrals. In: Proceedings of the 35th International Conference on Machine Learning (ICML) (2018). arXiv:8010.4153