Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Probabilistic Iterative Methods for Linear Systems

Jon Cockayne JCOCKAYNE@TURING.AC.UK
Alan Turing Institute
96 Euston Road
London, NW1 2DB, UK

Ilse C.F. Ipsen IPSEN@NSCU.EDU
Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8205, USA

Chris. J. Oates CHRIS.OATES@NCL.AC.UK
School of Mathematics and Statistics
Newcastle University
Newcastle-upon-Tyne, NE1 7RU, UK

Tim W. Reid TWREID@NCSU.EDU
Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8205, USA

Editor: TBD

Abstract

This paper presents a probabilistic perspective on iterative methods for approximating the solution $\mathbf{x} \in \mathbb{R}^d$ of a nonsingular linear system $\mathbf{Ax} = \mathbf{b}$. Classically, an iterative method produces a sequence $\mathbf{x}_m$ of approximations that converge to $\mathbf{x}$ in $\mathbb{R}^d$. Our approach, instead, lifts a standard iterative method to act on the set of probability distributions, $\mathcal{P}(\mathbb{R}^d)$, outputting a sequence of probability distributions $\mu_m \in \mathcal{P}(\mathbb{R}^d)$. The output of a probabilistic iterative method can provide both a "best guess" for $\mathbf{x}$, for example by taking the mean of μ_m, and also probabilistic uncertainty quantification for the value of $\mathbf{x}$ when it has not been exactly determined. A comprehensive theoretical treatment is presented in the case of a stationary linear iterative method, where we characterise both the rate of contraction of μ_m to an atomic measure on $\mathbf{x}$ and the nature of the uncertainty quantification being provided. We conclude with an empirical illustration that highlights the potential for probabilistic iterative methods to provide insight into solution uncertainty.

Keywords: linear algebra, probabilistic numerical methods, uncertainty quantification

1. Introduction

The focus of this paper is on the numerical solution of a linear systems of equations

$$\mathbf{Ax} = \mathbf{b}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a given non-singular matrix, $\mathbf{b} \in \mathbb{R}^d$ is a non-zero vector and $\mathbf{x} \in \mathbb{R}^d$ is an unknown vector to be computed. The problem of solving linear systems is central to

scientific computation (Golub and Van Loan, 2013, p103). Solvers can broadly be categorized as either *direct*, meaning they compute $\mathbf{x}$ by factorizing the matrix $\mathbf{A}$, or as *iterative*, meaning they output a sequence of approximations to $\mathbf{x}$. The focus of the present paper is on a probabilistic version of iterative methods.

There exist a wide variety of iterative methods, with the two main classes being the stationary iterative methods (Young, 1971), such as Richardson's method and Jacobi's method, and Krylov subspace methods (Liesen and Strakos, 2012) such as the conjugate gradient method (CG; Hestenes and Stiefel, 1952). In each case, the output of an iterative method is a sequence $\mathbf{x}_m$ of approximations to $\mathbf{x}$, that one hopes will converge to $\mathbf{x}$ as m is increased. In practice the error $\mathbf{e}_m = \mathbf{x} - \mathbf{x}_m$ is unknown but can be estimated. Error estimation for linear systems has a long history, with von Neumann and Goldstine (1947) among the earliest works in this now vast literature. For CG applied to a symmetric positive definite matrix $\mathbf{A}$, one typically estimates a bound for the $\mathbf{A}$-norm of the error $\|\mathbf{e}_m\|_{\mathbf{A}} = \sqrt{\mathbf{e}_m^\top \mathbf{A} \mathbf{e}_m}$ (e.g. Strakoš and Tichý, 2002, 2005; Meurant and Tichý, 2013, 2019; Meurant, 1997; Golub and Meurant, 1997, 1994). Norm-wise estimates such as this may be of limited utility for three reasons: they are often conservative, they may be complicated to compute and, being a scalar-valued summary, they cannot capture all of the structure that may be present in the error $\mathbf{e}_m$.

The purpose of this paper is to lift standard iterative methods into probability space, replacing iterates $\mathbf{x}_m \in \mathbb{R}^d$ with iterates $\mu_m \in \mathcal{P}(\mathbb{R}^d)$, where $\mathcal{P}(\mathbb{R}^d)$ denotes the set of probability measures on $\mathbb{R}^d$. The output of such a method then simultaneously provides an approximation to $\mathbf{x}$, for example by taking the mean of μ_m, and probabilistic error assessment. To motivate why such a method may be useful, suppose that the value of $\mathbf{x}$ is the input to some further computation, denoted abstractly as $F(\mathbf{x})$ for $F : \mathbb{R}^d \to \mathbb{R}$, and suppose that one wishes to characterise the error $F(\mathbf{x}) - F(\mathbf{x}_m)$ in replacing the unknown $\mathbf{x}$ with the numerical approximation $\mathbf{x}_m$. It is not trivial to transfer a bound on a derived quantity such as $\|\mathbf{e}_m\|_{\mathbf{A}}$ into a practically useful estimate of this error, particularly when F is not analytically tractable. For example, if F depends only on a subset of the entries of $\mathbf{x}$ for which the iterative method converges rapidly, while the other entries converge slowly, a bound on $F(\mathbf{x}) - F(\mathbf{x}_m)$ that is a function only of $\|\mathbf{e}_m\|_{\mathbf{A}}$ can be too conservative to be useful. In contrast, a probabilistic representation μ_m of uncertainty regarding $\mathbf{x}$ can be directly propagated through F by repeatedly sampling $\boldsymbol{X} \sim \mu_m$ and computing $F(\boldsymbol{X})$. The resulting probability distribution provides probabilistic *uncertainty quantification* (UQ) for the unknown quantity of interest $F(\mathbf{x})$, and may not suffer the same degree of conservatism of the norm-based estimators that we briefly described.

The methods described herein can be viewed as *probabilistic numerical methods* (PNM; Larkin, 1972; Diaconis, 1988; Hennig et al., 2015; Cockayne et al., 2019b; Oates and Sullivan, 2019). PNMs for linear systems are numerical methods that take as input the quantities $\mathbf{A}$ and $\mathbf{b}$, together with an *initial distribution* $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$, and return a probability distribution $\mu_m \in \mathcal{P}(\mathbb{R}^d)$ as their output. The role of μ_0 is to encode any *a priori* information that can be provided to the PNM. This is achieved by assigning probability mass to subsets of $\mathbb{R}^d$ in which $\mathbf{x}$ is believed to be located, prior to any computations being performed. This information may be elicited from a domain expert or obtained in an objective manner, for instance by performing additional computations related to the numerical task. While such applications to numerics have a different flavour to traditional applications of UQ (e.g.

Smith, 2014), the use of probabilities to describe uncertainty is philosophically similar; see further discussion in Hennig et al. (2015) and Cockayne et al. (2019b).

1.1 Related Work

Probabilistic Linear Solvers There has been recent interest in the construction of PNM for the solution of Eq. (1), with contributions in Hennig et al. (2015); Bartels and Hennig (2016); Bartels et al. (2019); Cockayne et al. (2019a); Reid et al. (2020); Wenger and Hennig (2020). With the exception of Bartels et al. (2019), these works have predominantly focused on replicating CG, and so a positive-definite $\mathbf{A}$ is assumed. Each of these works constructed a PNM in the Bayesian statistical framework, where the distribution μ_0 has the interpretation of a *prior* posited over some quantity related to Eq. (1) at the outset, and this distribution is updated based on the limited computations that are performed. The updating is achieved using Bayes' theorem and the result is a *posterior* or *conditional* distribution μ_m that forms the output of the method; it is a distribution over the unknown $\mathbf{x}$ that quantifies uncertainty given the limited computation performed. In Hennig et al. (2015); Bartels and Hennig (2016); Wenger and Hennig (2020) the prior was placed on the entries of $\mathbf{A}^{-1}$ (or jointly on $\mathbf{A}$ and $\mathbf{A}^{-1}$), while in Bartels et al. (2019); Cockayne et al. (2019a) the prior was placed directly on the unknown solution of Eq. (1). In each case, computation consisted of projecting Eq. (1) against a set of search directions $\mathbf{s}_i$, $i = 1, \ldots, m$ (i.e. by computing $\mathbf{s}_i^\top \mathbf{A}\mathbf{x} = \mathbf{s}_i^\top \mathbf{b}$) and the output of the PNM was a distribution that contracts to a point mass at $\mathbf{x}$ in an appropriate computational limit.

Each of these methods exploited conjugacy of Gaussian distributions under linear transformations to condition on the linear information provided by the pairs $(\mathbf{s}_i, \mathbf{s}_i^\top \mathbf{b})$, $i = 1, \ldots, m$. This conditioning is justified *only* when the search directions $\mathbf{s}_i$ are not themselves dependent on $\mathbf{x}$, the solution of Eq. (1). However, in practice these authors advocated the use of search directions generated using a Lanczos-style recursion (Liesen and Strakos, 2012, Section 2.4), meaning that the $\mathbf{s}_i$ depend on $\mathbf{x}$ via $\mathbf{b}$ and the required assumption is violated. As remarked in Bartels et al. (2019); Cockayne et al. (2019a), this violation leads to PNM that are neither Bayesian nor *calibrated*, with the latter understood to mean that the "width" of the probability distribution μ_m produced by the PNM can be a gross over-estimate of the actual error, as quantified by the difference between the mean of μ_m and $\mathbf{x}$. Reid et al. (2020) addressed this deficiency by constructing a prior which corrects for the over-confidence in an empirical Bayesian fashion, though with such a prescribed prior it is difficult for other problem-specific information to be incorporated. It therefore remains an open problem to develop a PNM for the solution of Eq. (1) that allows a generic initial distribution $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ to be used and ensures the distributional output $\mu_m \in \mathcal{P}(\mathbb{R}^d)$ of the PNM is calibrated.

Calibration A central concern of this paper is the idea of *calibration* of a PNM. As discussed above, informally this is the idea that $\mathbf{x}$ should look like a "typical sample" from the output distribution μ_m. This is of critical importance for PNM, since their primary aim is to provide meaningful UQ for the unknown solution $\mathbf{x}$. Cockayne et al. (2020) recently introduced a criterion for calibration, building on earlier work such as Dawid (1982); Monahan and Boos (1992), which is appropriate for PNM. Two definitions of calibration were introduced: in *strong calibration*, calibration is assessed by drawing multiple values of $\mathbf{x}$

from μ_0, computing the corresponding μ_m and testing whether, on average, $\mathbf{x}$ is distributed according to μ_m. Conversely in *weak calibration* only a marginal comparison is performed, that is, we check whether the average of μ_m over different values of $\mathbf{x}$ drawn from μ_0 is equal to μ_0. Naturally strong calibration is stronger, but as noted in Cockayne et al. (2020) it is significantly more difficult to verify than weak calibration.

1.2 Contributions

This paper adopts a profoundly different approach to the existing literature on PNM for linear systems. To our knowledge all existing methods seek to apply Bayes' theorem as described above. In this work we instead posit an initial distribution μ_0 and iteratively update this distribution by transforming it according to a standard iterative method for solving Eq. (1). Thus, rather than μ_m being the *conditional distribution* of μ_0 on some prescribed data as in all existing PNM for linear systems, in a probabilistic iterative method it is the *pushforward distribution* of μ_0 through the sequence of maps that define a standard iterative method.

The initial distribution μ_0 is loosely analogous to the prior in a Bayesian approach, but since no analogue of the Bayesian update occurs these methods are not Bayesian and we refrain from using the terms *prior* and *posterior* in this work. We thus refer to μ_m as a *belief distribution*, following the contemporary literature on generalised Bayesian inference (Bissiri et al., 2016). In departing from an established statistical paradigm one is required to justify, mathematically, the sense in which the uncertainty quantification provided by μ_m is meaningful. For this purpose we leverage the recent work of Cockayne et al. (2020), who argued that non-Bayesian procedures can be justified if they are *calibrated*, meaning that the unknown true solution $\mathbf{x}$ is indistinguishable in a certain, statistical sense, from any other sample drawn independently from μ_m. The contributions of this paper are therefore as follows:

- We introduce *probabilistic iterative methods*, a class of PNM derived from iterative methods for solving linear systems such as Eq. (1). These methods can be interpreted as a lifting of standard iterative methods into probability space, and are equivalent to randomising the initial iterate in a standard iterative method.

- A detailed theoretical analysis of the convergence properties of these new PNM is conducted for the class of linear stationary iterative methods, in which the next iterate is obtained by an affine transformation of the previous iterate. We prove that in this case the iterates produced are *strongly calibrated* in the sense of Cockayne et al. (2020) and hence provide meaningful uncertainty quantification despite not existing in the Bayesian paradigm.

- We test the bounds of our theory by describing and implementing an empirical test for calibration of more complex and general iterative methods, such as Krylov methods.

- We study application of probabilistic iterative methods to a toy regression problem. Here we examine the convergence and calibration of both linear and nonlinear probabilistic iterative methods, and highlight how their output may be used to gain insight into the impact of numerical uncertainty in the context of the regression task.

1.3 Structure of the Paper

In Section 2 we introduce iterative methods for linear systems and describe how these may be lifted into algorithms that operate on probability space. Theoretical results concerning the convergence and calibration of a class of analytically tractable probabilistic iterative methods are presented in Section 3, and in Section 4 we consider the general case, presenting a statistical test that can be used to assess whether the output from a probabilistic iterative method is calibrated. In Section 5 we apply probabilistic iterative methods to solve a linear system arising in a regression problem. Lastly, in Section 6 we discuss the results presented and the outlook for this new class of methods.

1.4 Notation

Here the notation for the paper is established. We will work in the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ where $\mathcal{B}(\mathbb{R}^d)$ is the standard Borel sigma-algebra for $\mathbb{R}^d$. Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of all probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Bold lower-case roman letters (e.g. $\mathbf{v}$) will be used to denote vectors in $\mathbb{R}^d$ and bold capital roman letters to denote matrices in $\mathbb{R}^{d\times d}$ (e.g. $\mathbf{M}$). Bold capital italic letters will denote random variables on $\mathbb{R}^d$ (e.g. $\boldsymbol{X}$) and lower-case Greek letters (e.g. μ) will be used to denote elements of $\mathcal{P}(\mathbb{R}^d)$.

Throughout it will be assumed that $\|\cdot\|$ is a fixed but arbitrary norm on $\mathbb{R}^d$. One important example is the vector p-norm, given by

$$\|\mathbf{v}\|_p = \left(\sum_{i=1}^{d} |v_i|^p\right)^{\frac{1}{p}},$$

though we note that many of the results presented herein do not assume any particular norm, and where a specific norm *is* required this will be emphasised. This notation will also be used for the induced norm on $\mathbb{R}^{d\times d}$, given by

$$\|\mathbf{M}\| = \sup_{\|\mathbf{x}\|=1} \|\mathbf{M}\mathbf{x}\|.$$

Recall that all induced norms are sub-multiplicative, meaning that $\|\mathbf{M}\mathbf{v}\| \leq \|\mathbf{M}\|\|\mathbf{v}\|$. Let $\rho(\mathbf{M})$ denote the spectral radius of $\mathbf{M}$, let $\mathbf{M}^\dagger$ denote the Moore-Penrose pseudo-inverse of $\mathbf{M}$, let range($\mathbf{M}$) denote its range and ker($\mathbf{M}$) its kernel or null space. For a symmetric matrix $\mathbf{M}$, let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ denote the smallest and largest eigenvalue of $\mathbf{M}$. For a positive-definite matrix $\mathbf{M}$ we define the weighted norm $\|\mathbf{v}\|_{\mathbf{M}} = (\mathbf{v}^\top \mathbf{M}\mathbf{v})^{1/2}$. Let $\mathbf{M}^{1/2}$ denote a matrix for which $\mathbf{M} = (\mathbf{M}^{1/2})^\top \mathbf{M}^{1/2}$. Note that this is not the typical notion of a square root, in that it will *not* be required that $(\mathbf{M}^{1/2})^\top = \mathbf{M}^{1/2}$.

For a measurable map $S : \mathbb{R}^d \to \mathbb{R}^d$ and a set $B \subset \mathbb{R}^d$, $S^{-1}[B]$ will be used to denote the preimage of B under S, i.e.

$$S^{-1}[B] = \{\mathbf{v} \in \mathbb{R}^d \text{ s.t. } S(\mathbf{v}) \in B\}.$$

The notation $\mathcal{N}(\mathbf{v}, \boldsymbol{\Sigma})$ will be used to denote the multivariate Gaussian distribution with mean $\mathbf{v}$ and positive semi-definite covariance $\boldsymbol{\Sigma}$.

2. Probabilistic Iterative Methods

In this section we start by recalling standard iterative methods, using the taxonomy of Young (1971), before then presenting our new concept of a probabilistic iterative method.

2.1 Iterative Methods

A general iterative method $\mathcal{I}$ is defined (Young, 1971, Section 3.1) as a sequence of maps $\mathcal{I} = (P_m)_{m\geq 1}$, for which $\mathbf{x}_m = P_m(\mathbf{x}_0, \ldots, \mathbf{x}_{m-1}; \mathbf{A}, \mathbf{b})$. The notation $\mathcal{I}(\mathbf{A}, \mathbf{b})$ will occasionally be used to make the dependence of the iterative method on $\mathbf{A}$ and $\mathbf{b}$ explicit. The iterative method $\mathcal{I}$ is said to be *linear* if each P_m is linear in $\mathbf{x}_0, \ldots, \mathbf{x}_{m-1}$. It is said to be *of degree* s if for all $m \geq s$ we have that P_m depends only on the s previous iterates, i.e. $P_m(\mathbf{x}_0, \ldots, \mathbf{x}_{m-1}; \mathbf{A}, \mathbf{b}) = P_m(\mathbf{x}_{m-s}, \ldots, \mathbf{x}_{m-1}; \mathbf{A}, \mathbf{b})$. Lastly, a degree s method is said to be *stationary* if the maps $(\mathbf{x}_{m-s}, \ldots, \mathbf{x}_{m-1}) \mapsto P_m(\mathbf{x}_{m-s}, \ldots, \mathbf{x}_{m-1})$ are independent of m for all $m \geq s$. In what follows we tend to suppress dependence of $\mathcal{I}$ and the P_m on $\mathbf{A}$ and $\mathbf{b}$ to reduce notational overhead.

Many of the most widely used iterative methods can be expressed as methods of degree $s = 1$. For simplicity, we present the majority of the material in this paper in these terms, though the core ideas readily generalise to higher degree methods as will be explored in Sections 4 and 5. Any iterative method $\mathcal{I}$ of degree $s = 1$ implies a map P^m that acts only on the first iterate $\mathbf{x}_0$ to produce iterate $\mathbf{x}_m$, as follows:

$$P^m(\mathbf{x}_0) = (P_m \circ \cdots \circ P_1)(\mathbf{x}_0).$$

In such cases each P_m is generally a *contraction map* with fixed point $\mathbf{x}$, i.e. $P_m(\mathbf{x}) = \mathbf{x}$. Thus, when the iterative method is stationary it amounts to applying a single fixed contraction map to an initial iterate until convergence.

We now present several examples of first degree iterative methods; for each see Young (1971, Section 3.3). These methods are seldom used as linear solvers in contemporary applications, but are still sometimes used in conjunction with other methods (Saad, 2003, p103).

Example 1 (Stationary Richardson method)**.** *This method adopts the following iteration*

$$\mathbf{x}_m = \mathbf{x}_{m-1} + \omega(\mathbf{b} - \mathbf{A}\mathbf{x}_{m-1}), \qquad m \geq 1$$

where $\omega > 0$ is a parameter of the method. The method is stationary and linear, with each map P_m of the form

$$P_m(\mathbf{v}) = P(\mathbf{v}) = \mathbf{G}\mathbf{v} + \mathbf{f} \tag{2}$$

where $\mathbf{G} = \mathbf{I}_d - \omega\mathbf{A}$ and $\mathbf{f} = \omega\mathbf{b}$.

Example 2 (Jacobi's method)**.** *In* Jacobi's method *it is assumed that the diagonal elements of $\mathbf{A}$ are nonzero. The iteration takes the form*

$$\mathbf{x}_m = \mathbf{D}^{-1}(\mathbf{b} - (\mathbf{A} - \mathbf{D})\mathbf{x}_{m-1}) + \mathbf{x}_{m-1}, \qquad m \geq 1$$

where $\mathbf{D} = \mathrm{diag}(\mathbf{A})$. The method is again stationary and linear. In the notation of Eq. (2). we have that $\mathbf{G} = \mathbf{I}_d - \mathbf{D}^{-1}\mathbf{A}$ and $\mathbf{f} = \mathbf{D}^{-1}\mathbf{b}$.

The next method, CG, sees significantly more use, particularly in the solution of large sparse linear systems. Whereas the above two methods are based on matrix splittings, in CG the solution $\mathbf{x}$ is instead projected into a sequence of *Krylov subspaces* (Liesen and Strakos, 2012, Section 2.2) of increasing dimension. As a result it is not traditionally viewed within the classification of Young (1971)[1]. Nevertheless CG is currently seen as an iterative method and may be categorised within the taxonomy presented above, albeit rather degenerately since CG converges (in exact arithmetic) in a finite number $m' \leq d$ of iterations and so P_m is undefined for $m > m'$.

Example 3 (Conjugate gradient method)**.** *In CG, for a symmetric positive-definite matrix* $\mathbf{A}$ *the iteration is of the form*

$$\mathbf{x}_m = \mathbf{x}_{m-1} + \alpha_m \mathbf{s}_m \qquad \alpha_m = \frac{\mathbf{s}_m^\top \mathbf{r}_m}{\mathbf{s}_m^\top \mathbf{A}\mathbf{s}_m}$$

$$\mathbf{s}_{m+1} = \mathbf{r}_m + \beta_m \mathbf{s}_m \qquad \beta_m = \frac{\mathbf{r}_m^\top \mathbf{r}_m}{\mathbf{r}_{m-1}^\top \mathbf{r}_{m-1}}$$

where the initial direction $\mathbf{s}_0$ *is taken to be the initial residual* $\mathbf{r}_0$*, and we recall that recall that* $\mathbf{r}_m = \mathbf{b} - \mathbf{A}\mathbf{x}_m$*. From Saad (2003, Algorithm 6.19), CG may be expressed as a three-term recurrence. Examining this, we see that CG is neither stationary nor linear, and is of second degree. Nevertheless in terms of its* implementation*, the algorithm requires only the storage of* $\mathbf{x}_m$*,* $\mathbf{r}_m$ *and* $\mathbf{s}_m$ *to compute* $\mathbf{x}_{m+1}$*.*

2.2 Lifting to Probability Space

Now we introduce the central definition of this paper, that of a probabilistic iterative method. We define probabilistic iterative methods in terms of degree $s = 1$ methods since, as noted above, many degree $s > 1$ methods can be expressed as degree $s = 1$ methods. We extend our definition to higher degree methods in Section 3.3 . First, recall that for a distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$, the *pushforward distribution* $S_\# \mu$ is the element of $\mathcal{P}(\mathbb{R}^d)$ defined as $(S_\# \mu)(B) = \mu(S^{-1}[B])$ for each $B \in \mathcal{B}(\mathbb{R}^d)$. Effectively $S_\# \mu$ is the image of μ under the map S.

Definition 1. *Let* $\mathcal{I} = (P_m)_{m \geq 1}$ *be an iterative method of first degree. Then the maps* $P_m : \mathbb{R}^d \to \mathbb{R}^d$ *can be lifted to maps* $(P_m)_\# : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$ *operating on elements* $\mu \in \mathcal{P}(\mathbb{R}^d)$ *by computing the pushforward distribution* $(P_m)_\# \mu$*. We say that* $\mathcal{I}_\# = ((P_m)_\#)_{m \geq 1}$ *is a* probabilistic iterative method*.*

Thus probabilistic iterative methods are a class of PNMs that take as input an initial distribution $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ and return a sequence of iterates $\mu_m = (P_m)_\# \mu_{m-1}$. Again we note that $\mathcal{I}_\#$, and therefore μ_m, each formally depend on $\mathbf{A}$ and $\mathbf{b}$, but this dependence is notationally suppressed. The distribution μ_0 should be thought of as an initial belief

1. The discussion in Liesen and Strakos (2012, Section 2.5.7) highlights that, when Young (1971) was written, CG was still often considered a direct method owing to its convergence in $m' \leq d$ iterations; its attractive properties as an iterative method were not understood by the community until Reid (1971), who studied its use as an iterative method for large sparse linear systems. This likely explains why Young (1971) does not attempt to categorise it within his taxonomy.

about where the solution $\mathbf{x}$ to the linear system might lie in $\mathbb{R}^d$. Thus μ_0 has a similar role to the prior distribution in the Bayesian setting. However, unlike existing PNMs for linear systems the iterates μ_m do not arise as a conditional distribution, and so the output from probabilistic iterative methods does not have the same Bayesian interpretation as existing methods. It is therefore crucial to ensure that the UQ provided by the method is meaningful. Indeed, in contrast to a Bayesian approach, it is straightforward to construct an example showing that the support of μ_m need not be contained in the support of μ_0. Thus, even if μ_0 encodes properties of the solution that are expected to hold with probability one (for example, positivity of the elements) μ_m is not guaranteed to inherit those properties. This emphasises the need for careful analysis of probabilistic iterative methods, which we present in detail in Section 3.2 (for stationary linear methods) and Section 4.2 (for general methods).

Compared to earlier attempts to construct PNM for solution of Eq. (1), probabilistic iterative methods are significantly easier to implement. For example, an algorithm for producing a sample from μ_m is to sample $\boldsymbol{X} \sim \mu_0$ and compute $P^m(\boldsymbol{X})$. Thus sampling from the output of a probabilistic iterative method inherits the computational efficiency and stability of the underlying iterative method, only multiplying the cost by the number of samples required. Conversely, earlier approaches to PNM (which had a Bayesian flavour) generally required new algorithms and corresponding code to be developed, whose numerical stability must then be independently tested and verified[2].

Our first theoretical result shows that if the classical iterates $\mathbf{x}_m$ converge to the true solution $\mathbf{x}$, then the distributions μ_m contract to an atomic mass centred on $\mathbf{x}$ under weak regularity conditions on μ_0, namely that the integral $\int \|\mathbf{x} - \mathbf{v}\|^k \mathrm{d}\mu_0(\mathbf{v})$ is finite for some $k > 0$.

Proposition 2. *Let $\mathcal{I}$ be an iterative method of first degree for solution of Eq.* (1)*. Suppose that each P_m has error controlled by the bound*

$$\|\mathbf{x} - P_m(\mathbf{x}_0)\| \leq \varphi(m)\|\mathbf{x} - \mathbf{x}_0\|, \qquad m \geq 1 \tag{3}$$

where $\varphi : \mathbb{N} \to \mathbb{R}$ is some function independent of $\mathbf{x}_0$, such that $\varphi(m) \to 0$ as $m \to \infty$. Then for any $k > 0$ and $\delta > 0$,

$$\mu_m(B_\delta^{\mathsf{c}}(\mathbf{x})) \leq \left(\frac{\varphi(m)}{\delta}\right)^k \int_{\mathbb{R}^d} \|\mathbf{x} - \mathbf{v}\|^k \, \mathrm{d}\mu_0(\mathbf{v}),$$

where $B_\delta(\mathbf{x})$ represents a $\|\cdot\|$-ball of radius δ about $\mathbf{x}$, i.e. $B_\delta(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{v}\| < \delta\}$, and $B_\delta^{\mathsf{c}}(\mathbf{x})$ its complement in $\mathbb{R}^d$.

2. This is particularly true of existing PNM for solving linear systems such as those methods discussed in Section 1.1. The Lanczos-style recursions exploited to construct the search directions of those methods are well known to lead to accumulation of round off error in algorithms such as CG, and the impact of this on the posterior covariance matrices computed in those methods has, to our knowledge, not yet been analysed.

Proof For any $k > 0$,

$$\begin{aligned}\int_{\mathbb{R}^d} \|\mathbf{x}-\mathbf{v}\|^k \,\mathrm{d}\mu_m(\mathbf{v}) &= \int_{\mathbb{R}^d} \|\mathbf{x}-\mathbf{v}\|^k \,\mathrm{d}P^m_{\#}\mu_0(\mathbf{v}) \\ &= \int_{\mathbb{R}^d} \|\mathbf{x}-P^m(\mathbf{v})\|^k \,\mathrm{d}\mu_0(\mathrm{d}\mathbf{v}) \\ &\leq \varphi(m)^k \int_{\mathbb{R}^d} \|\mathbf{x}-\mathbf{v}\|^k \,\mathrm{d}\mu_0(\mathrm{d}\mathbf{v})\end{aligned}$$

where the second line follows from the change of variables $v \mapsto P^m(v)$, and the third line follows from Eq. (3) and extracting terms independent of $\mathbf{v}$ from the integral. Now, recall from Chebyshev's inequality (Kallenberg, 2002, Lemma 3.1) we have that for a measure μ on $\mathbb{R}^d$, a μ-measurable function $f : \mathbb{R}^d \to [0,\infty)$ and scalars $\delta \in [0,\infty)$, $k \in (0,\infty)$ it holds that

$$\mu(\{\mathbf{v} \in \mathbb{R}^d : f(\mathbf{v}) \geq \delta\}) = \mu(\{\mathbf{v} \in \mathbb{R}^d : f(\mathbf{v})^k \geq \delta^k\}) \leq \frac{1}{\delta^k}\int_{\mathbb{R}^d} f(\mathbf{v})^k \,\mathrm{d}\mu(\mathbf{v}).$$

Applying this in the present setting with $f(\mathbf{v}) = \|\mathbf{x}-\mathbf{v}\|$ we therefore have

$$\mu_m(B^{\mathsf{c}}_\delta(\mathbf{x})) \leq \left(\frac{\varphi(m)}{\delta}\right)^k \int_{\mathbb{R}^d} \|\mathbf{x}-\mathbf{v}\|^k \,\mathrm{d}\mu_0(\mathbf{v})$$

as required. ■

Thus the probability mass assigned by μ_m to the region outside of a ball $B_\delta(\mathbf{x})$ centred on the true solution $\mathbf{x}$ vanishes as $m \to \infty$. Moreover, and again asymptotically as $m \to \infty$, the probability mass outside $B_\delta(\mathbf{x})$ vanishes more rapidly when high-order moments of μ_0 exist (i.e. for large k). However, Proposition 2 does not imply that the UQ provided by μ_m is meaningful, or even that $\mathbf{x}$ is in the support of μ_m. For the UQ to be meaningful further assumptions are required on $\mathcal{I}$, such as those made in Section 3.2.

3. Linear Probabilistic Iterative Methods

In this section we restrict attention to *linear, stationary* iterative methods of first degree, as a richer set of theoretical results can be developed for this restricted set of methods. In Section 3.1 we recall some classical results and describe how the probabilistic iterates μ_m can be exactly computed when μ_0 is Gaussian. In Section 3.2 we prove that these methods are strongly calibrated in the sense of Cockayne et al. (2020), and in Section 3.3 we discuss relaxing the stationarity and first degree assumptions.

3.1 Linear and Stationary Probabilistic Iterative Methods

For a linear stationary iterative methods of first degree, $P_m(\mathbf{x}_0,\ldots,\mathbf{x}_{m-1}) = P(\mathbf{x}_{m-1})$, as described in Example 1, where

$$P(\mathbf{v}) = \mathbf{G}\mathbf{v} + \mathbf{f} \tag{4}$$

for some $\mathbf{G} \neq \mathbf{0} \in \mathbb{R}^{d\times d}$ and $\mathbf{f} \in \mathbb{R}^d$. It follows that $P^m(\mathbf{x}_0) = \mathbf{G}^m\mathbf{x}_0 + \sum_{i=0}^{m-1}\mathbf{G}^m\mathbf{f}$, where $\mathbf{G}^0 = \mathbf{I}$. We now recall a classical result for linear stationary iterative methods of first

degree which will later be useful. The following is based on Young (1971, Section 3.3.2 and 3.3.5) and Saad (2003, Section 4.2).

Proposition 3. *Let* $\mathbf{A}$ *be nonsingular, suppose that* $\mathbf{G} \in \mathbb{R}^{d\times d}$ *is such that* $\|\mathbf{G}\| < 1$ *and*

$$\mathbf{f} = (\mathbf{I}_d - \mathbf{G})\mathbf{A}^{-1}\mathbf{b} = (\mathbf{I}_d - \mathbf{G})\mathbf{x}. \tag{5}$$

Then the iterative method

$$\mathbf{x}_{m+1} = \mathbf{G}\mathbf{x}_m + \mathbf{f} \qquad m \geq 1$$

converges to $\mathbf{x}$ *for all* $\mathbf{x} \in \mathbb{R}^d$*. Furthermore the error in* $\mathbf{x}_m$ *is controlled by the bound*

$$\|\mathbf{x} - \mathbf{x}_m\| \leq \|\mathbf{G}\|^m \|\mathbf{x} - \mathbf{x}_0\|.$$

Now we consider lifting linear stationary iterative methods of first degree into $\mathcal{P}(\mathbb{R}^d)$. Our main observation is that if μ_0 is Gaussian, then the distribution μ_m can be computed in closed form using standard formulae for linear transforms of Gaussian distributions.

Proposition 4. *Let* $\mathcal{I}$ *be a linear, stationary, first degree iterative method. Let* $\mu_0 = \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$*. Then* $\mu_m = \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$*, where* $\mathbf{x}_m = P^m(\mathbf{x}_0)$ *coincides with the iterate from the underlying method, and*

$$\mathbf{\Sigma}_m = \mathbf{G}^m \mathbf{\Sigma}_0 (\mathbf{G}^m)^\top.$$

Furthermore we have the following bounds:

$$\|\mathbf{x} - \mathbf{x}_m\| \leq \|\mathbf{G}\|^m \|\mathbf{x} - \mathbf{x}_0\| \qquad\qquad \|\mathbf{\Sigma}_m\| \leq \|\mathbf{G}\|^m \|\mathbf{G}^\top\|^m \|\mathbf{\Sigma}_0\|.$$

Proof From elementary properties of Gaussian distributions (Tong, 1990, Theorem 3.3.3) we have that $\mu_1 = \mathcal{N}(\mathbf{x}_1, \mathbf{\Sigma}_1)$ where $\mathbf{x}_1 = \mathbf{G}\mathbf{x}_0 + \mathbf{f}$ and $\mathbf{\Sigma}_1 = \mathbf{G}\mathbf{\Sigma}_0\mathbf{G}^\top$. This can be continued inductively to achieve the form stated in the proposition for all $m \geq 1$. The bound on $\|\mathbf{x} - \mathbf{x}_m\|$ is a consequence of $\mathbf{x}_m$ coinciding with the classical iterate and Proposition 3. The bound on $\mathbf{\Sigma}_m$ is direct by applying submultiplicativity of the norm $\|\cdot\|$ to $\|\mathbf{G}^m \mathbf{\Sigma}_0 (\mathbf{G}^m)^\top\|$. ■

Remark 5. *The bound on* $\mathbf{\Sigma}_m$ *in Proposition 4 does not require that* $\|\cdot\|$ *be the induced norm, only that it is submultiplicative. As a result, this applies to other matrix norms such as the Frobenius norm, which is submultiplicative but not induced.*

3.2 Evaluation of Uncertainty Quantification

The crucial point that must be addressed in order for probabilistic iterative methods to be useful is whether the covariance matrix $\mathbf{\Sigma}_m$ relates meaningfully to the error $\mathbf{e}_m = \mathbf{x} - \mathbf{x}_m$. It is not possible to provide a satisfactory answer to this question by considering just one linear system; this would be akin to asking whether the number 3 is meaningfully related to the distribution $\mathcal{N}(0, 1)$. Therefore a collection of linear systems is required so that average-case properties can be discussed.

The criteria for meaningful UQ introduced in Cockayne et al. (2020) imply the calibration of the PNM can be assessed using an ensemble of linear systems obtained by replacing

the right hand side $\mathbf{b}$ with realisations of a random vector $\boldsymbol{B} = \mathbf{A}\boldsymbol{X}$, $\boldsymbol{X} \sim \mu_0$. The PNM is then said to be *strongly calibrated* if the true solution $\boldsymbol{X}$ is statistically "plausible" as a sample from $\mathcal{N}(\mathbf{x}_m, \boldsymbol{\Sigma}_m)$, on average with respect to $\boldsymbol{X}$, a notion that will be formalised in Definition 6. Note that when $\boldsymbol{X}$ is randomised in this way both the mean $\mathbf{x}_m$ and covariance $\boldsymbol{\Sigma}_m$ of μ_m will themselves be random in general[3], as a consequence of the fact that $\mathcal{I}_\# = \mathcal{I}_\#(\mathbf{A}, \boldsymbol{B})$. A strongly calibrated PNM provides meaningful UQ, since its output provides a probabilistic representation of uncertainty whose credible sets have correct coverage with respect to realisations of $\boldsymbol{X}$.

In this section we will show that linear, stationary, first-degree probabilistic iterative methods are strongly calibrated when a Gaussian μ_0 is used. This is in contrast to earlier work, where empirical studies in Cockayne et al. (2019a) found that the PNM proposed in that work (called *BayesCG*) failed to be calibrated, though we note that Reid et al. (2020) proposed a particular prior under which BayesCG is calibrated. Initially we assume that $\boldsymbol{\Sigma}_m$ is nonsingular, which implies that $\mathbf{G}$ must also be nonsingular.

Definition 6 (Strong calibration, nonsingular case)**.** *Fix $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$. Suppose that a PNM for the solution of Eq.* (1) *produces output of the form $\mu_m = \mathcal{N}(\mathbf{x}_m, \boldsymbol{\Sigma}_m)$ where $\boldsymbol{\Sigma}_m$ is a symmetric positive-definite matrix. Then the PNM is said to be* strongly calibrated *for $(\mu_0, \mathbf{A})$ if, when applied to solve a random linear system defined by $\mathbf{A}$ and $\boldsymbol{B} = \mathbf{A}\boldsymbol{X}$, $\boldsymbol{X} \sim \mu_0$, it holds for all $m > 0$ that*

$$\boldsymbol{\Sigma}_m^{-\frac{1}{2}}(\boldsymbol{X} - \mathbf{x}_m) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d). \tag{6}$$

Similar notions of calibration have recently been exploited for verifying the correctness of algorithms for Bayesian computation in Cook et al. (2006); Talts et al. (2018); see Cockayne et al. (2020) for detail. Similar ideas have also been explored in the literature on PNM, such as in Cockayne et al. (2019a); Bartels et al. (2019); Reid et al. (2020). Those works explored calibration through a statistic referred to as the *Z-statistic*. Definition 6 is strictly more general than the Z-statistic, which is obtained by simply taking the squared 2-norm of Eq. (6).

The next proposition proves that when $\mathbf{G}$ is nonsingular, probabilistic iterative methods are strongly calibrated.

Proposition 7. *Let the assumptions of Definition 6 hold, with $\mu_0 = \mathcal{N}(\mathbf{x}_0, \boldsymbol{\Sigma}_0)$ and $\boldsymbol{\Sigma}_0$ a positive definite matrix. Additionally assume that $\mathcal{I}$ is a linear first degree stationary iterative method with nonsingular $\mathbf{G}$, and that Eq.* (5) *holds with probability one when $\mathcal{I}$ is applied to solve a system defined by the right hand side $\boldsymbol{B} = \mathbf{A}\boldsymbol{X}$, $\boldsymbol{X} \sim \mu_0$. Then $\mathcal{I}_\#$ is strongly calibrated for $(\mu_0, \mathbf{A})$.*

Proof First we consider $\boldsymbol{\Sigma}_m^{-1/2}(\mathbf{x} - \mathbf{x}_m)$ for a *fixed* true solution $\mathbf{x}$; we will complete the proof by randomising $\mathbf{x}$ to obtain the result. Note that $\boldsymbol{\Sigma}_m$ is nonsingular since $\mathbf{G}$ and $\boldsymbol{\Sigma}_0$ are nonsingular, and recall that since square-roots are not required to be *symmetric* in this

3. A possible exception occurs if $\mathcal{I}$ is a linear, stationary, first-degree iterative method, when $\boldsymbol{\Sigma}_m$ depends only on $\mathbf{G}$, and for such methods $\mathbf{G}$ is often independent of $\mathbf{b}$. In this case $\boldsymbol{\Sigma}_m$ is not random when $\boldsymbol{X}$ is randomised.

work we have that $\boldsymbol{\Sigma}_m^{\frac{1}{2}} = \boldsymbol{\Sigma}_{m-1}^{\frac{1}{2}}\mathbf{G}$. Now, for each fixed $\mathbf{x}$ and all $m > 0$ we have:

$$\begin{aligned}\boldsymbol{\Sigma}_m^{-\frac{1}{2}}(\mathbf{x} - \mathbf{x}_m) &= \boldsymbol{\Sigma}_{m-1}^{-\frac{1}{2}}(\mathbf{G}^{-1}(\mathbf{x} - \mathbf{G}\mathbf{x}_{m-1} - \mathbf{f})) \\ &= \boldsymbol{\Sigma}_{m-1}^{-\frac{1}{2}}(\mathbf{G}^{-1}(\mathbf{x} - \mathbf{f}) - \mathbf{x}_{m-1}).\end{aligned}$$

Now we have $\mathbf{G}^{-1}(\mathbf{x} - \mathbf{f}) = \mathbf{x}$, from nonsingularity of $\mathbf{G}$ and Eq. (5). It follows inductively over m that

$$\begin{aligned}\boldsymbol{\Sigma}_m^{-\frac{1}{2}}(\mathbf{x} - \mathbf{x}_m) &= \boldsymbol{\Sigma}_{m-1}^{-\frac{1}{2}}(\mathbf{x} - \mathbf{x}_{m-1}) \\ &= \boldsymbol{\Sigma}_0^{-\frac{1}{2}}(\mathbf{x} - \mathbf{x}_0).\end{aligned}$$

Thus, if we now randomise $\mathbf{x}$ according to $\boldsymbol{X} \sim \mu_0 = \mathcal{N}(\mathbf{x}_0, \boldsymbol{\Sigma}_0)$, we obtain $\boldsymbol{\Sigma}_m^{-1/2}(\boldsymbol{X} - \mathbf{x}_m) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, completing the proof. ■

Remark 8. *The only demand Proposition 7 makes of $\mathcal{I}$ is that Eq. (5) is almost surely satisfied; it does not require that $\|\mathbf{G}\| < 1$. Thus strong calibration of a PNM does not imply that μ_m contracts to the truth, only that μ_m should be a fair reflection of the size of the error. For example, if $\mathcal{I}$ diverges for some $\mathbf{x}_0$ it is natural that μ_m should tend to a distribution with infinite variance as $m \to \infty$.*

The assumption of nonsingular $\mathbf{G}$ permits a straightforward proof for Proposition 7, but unfortunately $\mathbf{G}$ may be singular even for such elementary methods as the Jacobi iterations. The next definition adapts Definition 6 to the case where $\mathbf{G}$, and therefore also $\boldsymbol{\Sigma}_m$, are singular. It simplifies the subsequent presentation to focus on the case where $\boldsymbol{\Sigma}_m$ does not depend on $\mathbf{b}$. To the best of our knowledge this is the case for the majority of stationary iterative methods.

Definition 9 (Strongly calibrated, singular case)**.** *Fix $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$. Suppose that a PNM for the solution of Eq. (1) produces output of the form $\mu_m = \mathcal{N}(\mathbf{x}_m, \boldsymbol{\Sigma}_m)$ where $\boldsymbol{\Sigma}_m$ is a positive semidefinite matrix with rank $0 < r < d$, with $\boldsymbol{\Sigma}_m$ not depending on the right hand side $\mathbf{b}$. Then the PNM is said to be* strongly calibrated *for $(\mu_0, \mathbf{A})$ if, when applied to solve a random linear system defined by $\mathbf{A}$ and $\boldsymbol{B} = \mathbf{A}\boldsymbol{X}$, $\boldsymbol{X} \sim \mu_0$, for each $m > 0$ there exist $\mathbf{R} \in \mathbb{R}^{d\times r}$ and $\mathbf{N} \in \mathbb{R}^{d\times(d-r)}$, with $\mathrm{range}(\mathbf{R}) = \mathrm{range}(\boldsymbol{\Sigma}_m)$ and $\mathrm{range}(\mathbf{N}) = \ker(\boldsymbol{\Sigma}_m)$, such that the following two conditions are satisfied:*

1. $(\mathbf{R}^\top \boldsymbol{\Sigma}_m \mathbf{R})^{-\frac{1}{2}} \mathbf{R}^\top (\boldsymbol{X} - \mathbf{x}_m) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$.
2. $\mathbf{N}^\top (\boldsymbol{X} - \mathbf{x}_m) = \mathbf{0}$.

This definition is an intuitive extension of Definition 6 to the case of singular $\boldsymbol{\Sigma}_m$; it demands that in any subspace of $\mathbb{R}^d$ in which $\boldsymbol{\Sigma}_m$ is nonzero, the PNM is strongly calibrated as in Definition 6, and in any subspace in which it is zero and thus no uncertainty remains, $\mathbf{x}_m$ is identically equal to the true solution $\boldsymbol{X}$. Note that in the special case $r = d$, Definition 9 reduces to Definition 6 since $\mathrm{range}(\boldsymbol{\Sigma}_m) = \mathrm{range}(\mathbf{I}_d)$.

We then have the following result, the proof of which is provided in Appendix A. The intuition behind the proof in the singular case is the same as in the nonsingular case, but additional technical effort is required to project into the null space of $\boldsymbol{\Sigma}_m$.

Proposition 10. *Let $\mu_0 = \mathcal{N}(\mathbf{x}_0, \mathbf{\Sigma}_0)$ where $\mathbf{\Sigma}_0$ is a positive definite matrix. Let $\mathcal{I}$ be a linear first degree stationary iterative method such that Eq.* (5) *holds with probability one when $\mathcal{I}$ is applied to solve a system defined by the right hand side $\boldsymbol{B} = \mathbf{A}\boldsymbol{X}$, $\boldsymbol{X} \sim \mu_0$. Suppose that $\mathbf{G}$ is independent of $\boldsymbol{B}$, and that $\mathbf{G}$ is diagonalisable with rank $0 < r \leq d$. Then the probabilistic iterative method $\mathcal{I}_\#$ is strongly calibrated for $(\mu_0, \mathbf{A})$.*

Remark 11. *Since in Definition 9 the matrix $\mathbf{\Sigma}_m$ does not depend on $\boldsymbol{B}$, both $\mathbf{R}$ and $\mathbf{N}$ can be fixed matrices independent of $\boldsymbol{X}$. Furthermore while the columns of $\mathbf{R}$ and $\mathbf{G}$ must be bases of the range and kernel of $\mathbf{\Sigma}_m$ respectively, Definition 9 and Proposition 10 are basis-independent.*

Propositions 7 and 10 provide a clear and defensible sense in which the output μ_m from a probabilistic iterative method $\mathcal{I}_\#$, arising from a linear first degree stationary iterative method $\mathcal{I}$, can be considered to be meaningful. Specifically, one has a guarantee that the unknown solution is indistinguishable, in a statistical sense, from samples drawn from μ_m. Thus one may interpret μ_m as quantifying uncertainty with respect to the unknown true value of $\mathbf{x}$ in Eq. (1).

3.3 Generalisations

Here we discuss generalisations to both non-stationary and higher degree iterative methods, while remaining in the linear framework.

3.3.1 Non-Stationary Methods

In a non-stationary linear iterative method of first degree (Young, 1971, Chapter 9), the iteration is of the form:

$$\mathbf{x}_m = \mathbf{G}_m \mathbf{x}_{m-1} + \mathbf{f}_m \tag{7}$$

where $\mathbf{f}_m \in \mathbb{R}^d$ and $\mathbf{G}_m \in \mathbb{R}^{d \times d}$ for all $m \geq 0$. The map P^m is then of the form:

$$P^m(\mathbf{x}_0) = \hat{\mathbf{G}}_m \mathbf{x}_0 + \hat{\mathbf{f}}_m$$

$$\hat{\mathbf{G}}_m = \prod_{i=1}^{m} \mathbf{G}_m \qquad \hat{\mathbf{f}}_m = \mathbf{f}_m + \sum_{i=1}^{m} \left(\prod_{j=i+1}^{m} \mathbf{G}_j \right) \mathbf{f}_i.$$

From this it follows by an identical argument to Proposition 4 that $\mu_m = \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$, with $\mathbf{x}_m = \hat{\mathbf{G}}_m \mathbf{x}_0 + \hat{\mathbf{f}}_m$ and $\mathbf{\Sigma}_m = \hat{\mathbf{G}}_m \mathbf{\Sigma}_0 \hat{\mathbf{G}}_m$.

Considering the consistency of the implied probabilistic iterative method, as in the stationary setting, $\mathbf{x}_m$ coincides with the classical iterate. Furthermore, Young (1971) notes that the iteration from Eq. (7) converges to $\mathbf{x}$ only if $\hat{\mathbf{G}}_m \to \mathbf{0}$. In this event clearly $\mathbf{\Sigma}_m \to \mathbf{0}$, and so provided the underlying iterative method converges, μ_m converges to an atomic mass on $\mathbf{x}$ as $m \to \infty$.

From the perspective of calibration of UQ, the proofs in Section 3.2 do not apply to non-stationary iterative methods $\mathcal{I}$ since those proofs exploit that $\mathbf{\Sigma}_m = \mathbf{G}^m \mathbf{\Sigma}_0 (\mathbf{G}^m)^\top$, which no longer holds in the non-stationary setting. However if one instead directly assumes $\hat{\mathbf{G}}_m$ to be diagonalisable for each m, the proof of Proposition 7 would need only minor modifications to establish that the associated probabilistic iterative method $\mathcal{I}_\#$ is strongly calibrated in the non-stationary setting.

3.3.2 Higher Degree Methods

Modifying Definition 1 to allow methods of higher degree requires changing the space on which μ is defined, and the domain of P_m (and by extension $(P_m)_\#$), to a Cartesian product of s instances of $\mathbb{R}^d$.

In terms of such methods, when $s = 2$ (Young, 1971, Chapter 16) the iteration takes the form

$$\mathbf{x}_m = \mathbf{G}\mathbf{x}_{m-1} + \mathbf{H}\mathbf{x}_{m-2} + \mathbf{k} \tag{8}$$

where $\mathbf{G}, \mathbf{H} \in \mathbb{R}^{d\times d}$ and $\mathbf{k} \in \mathbb{R}^d$. While second degree methods are seldom used in practice, higher order methods can accelerate convergence and raise some interesting statistical questions. These methods are analysed by augmenting the space as follows, to obtain a first degree linear stationary iterative method on $\mathbb{R}^{2d}$:

$$\begin{pmatrix} \mathbf{x}_{m-1} \\ \mathbf{x}_m \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_d \\ \mathbf{H} & \mathbf{G} \end{pmatrix} \begin{pmatrix} \mathbf{x}_{m-2} \\ \mathbf{x}_{m-1} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{k} \end{pmatrix} = \tilde{\mathbf{G}} \begin{pmatrix} \mathbf{x}_{m-2} \\ \mathbf{x}_{m-1} \end{pmatrix} + \tilde{\mathbf{k}}.$$

Convergence of the iterate, and hence the covariance in Proposition 4, then requires $\rho(\tilde{\mathbf{G}}) < 1$. Similarly, provided $\tilde{\mathbf{G}}$ satisfies the assumptions in Propositions 7 and 10, μ_m will provide meaningful UQ according to Definitions 6 and 9.

An interesting technicality for higher degree methods is that, whereas in first degree methods only an initial iterate $\mathbf{x}_0$ must be supplied, in second degree methods both the iterates $\mathbf{x}_0$ and $\mathbf{x}_1$ are required. This raises a challenge in the probabilistic framework because it is not clear how one should specify an initial distribution jointly over $\mathbf{x}_0$ and $\mathbf{x}_1$. While expert knowledge may be exploited to build a distribution over $\mathbf{x}_0$, the same is not true of $\mathbf{x}_1$. Several possible approaches are considered experimentally in Section 5.

4. Beyond Linearity

In the non-Gaussian and non-linear setting it is significantly more difficult to formulate an appropriate sense in which a PNM can be considered to be strongly calibrated. Instead, in this section we adopt a strictly weaker notion called *weak calibration*, which is simply defined and can be empirically tested. In Section 4.1 we present that definition and in Section 4.2 discuss statistical tests for weak calibration which will be applied in Section 5 when nonlinear iterative methods are assessed.

4.1 Weakly Calibrated Probabilistic Iterative Methods

The chief issue with Definitions 6 and 9 is that in order to define strong calibration we require that μ_m is Gaussian. This is problematic because Gaussian distributions are unable to express all initial beliefs about components of $\mathbf{x}$, and because the linear iterative methods which result in a Gaussian μ_m are less widely-used compared to nonlinear iterative methods, such as CG. Therefore we turn to an alternative, weaker sense in which the output μ_m from a (possibly nonlinear) probabilistic iterative method can be considered to be meaningful.

Our notion of weak calibration is also due to Cockayne et al. (2020), and will now be defined. In the same setting as Section 3.2, we fix $\mathbf{A}$ and randomly generate a right hand side $\boldsymbol{B} = \mathbf{A}\boldsymbol{X}$, $\boldsymbol{X} \sim \mu_0$. Then, conditional on $\boldsymbol{X}$ and for each $m > 0$, we introduce a second

random variable $\boldsymbol{Y}^{(m)}|\boldsymbol{X} \sim \mu_m$ that is sampled from the output μ_m of the PNM applied to solve the linear system defined by $\mathbf{A}$ and $\boldsymbol{B}$. Let $\boldsymbol{Y}^{(m)}$ denote the random variable obtained by marginalising $\boldsymbol{Y}^{(m)}|\boldsymbol{X}$ over realisations of $\boldsymbol{X}$.

Definition 12 (Weakly calibrated)**.** *Fix $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$. A PNM for the solution of Eq.* (1) *is said to be* weakly calibrated *to $(\mu_0, \mathbf{A})$ if, when applied to solve a random linear system defined by $\mathbf{A}$ and $\boldsymbol{B} = \mathbf{A}\boldsymbol{X}$, $\boldsymbol{X} \sim \mu_0$, and when $\boldsymbol{Y}^{(m)}|\boldsymbol{X} \sim \mu_m$, it holds for all $m > 0$ that $\boldsymbol{Y}^{(m)}$ has marginal distribution*

$$\boldsymbol{Y}^{(m)} \sim \mu_0. \tag{9}$$

Eq. (9) is sometimes called the *self-consistency property* and, as with strong calibration, the notion of weak calibration has previously been exploited to verify the correctness of algorithms for Bayesian computation (Geweke, 2004). Cockayne et al. (2020, Lemma 2.19) establishes that strong calibration implies weak calibration. Although weaker than strong calibration, Definition 12 allows for statistical tests of distributional equality to be used to assess the quality of the uncertainty quantification provided by a PNM whose output is non-Gaussian.

Remark 13 (Strong versus weak calibration)**.** *From a simulation perspective, we can intuitively think about strong and weak calibration in the following terms:*

1. *draw $\boldsymbol{X} \sim \mu_0$,*
2. *compute output μ_m from the probabilistic iterative method $\mathcal{I}_{\#}(\mathbf{A}, \mathbf{A}\boldsymbol{X})$,*
3. *draw $\boldsymbol{X}' \sim \mu_m$,*

then, in strong calibration we

4. *compare $\boldsymbol{X}$ to $\boldsymbol{X}'$.*

while in weak calibration we

4. *independently draw $\boldsymbol{X}'' \sim \mu_0$ and compare $\boldsymbol{X}''$ to $\boldsymbol{X}'$.*

Thus in strong calibration a conditional comparison is performed, while in weak calibration only a marginal comparison is performed.

4.2 Testing for Weak Calibration

We now present a statistical test to determine whether a PNM is weakly calibrated. For convenience we let ν_m denote the distribution of $\boldsymbol{Y}^{(m)}$, so that we aim to test whether $\nu_m = \mu_0$. Since ν_m does not necessarily have a closed form but it is possible to access samples from ν_m, we aim to perform a goodness-of-fit test to determine whether such samples are consistent with being drawn from μ_0. In this work we adopt a general purpose goodness-of-fit test based on *maximum mean discrepancy* (MMD), due to Gretton et al. (2012), which we briefly describe next.

Definition 14 (Maximum mean discrepancy). *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and let $\mathcal{F}$ be a set of real-valued, μ and ν-integrable functions on $\mathbb{R}^d$. Then the MMD between μ and ν, based on $\mathcal{F}$, is given by*

$$\textit{MMD}_{\mathcal{F}}(\mu, \nu) := \sup_{f \in \mathcal{F}} \left(\int f(\mathbf{v}) \, \mu(\mathrm{d}\mathbf{v}) - \int f(\mathbf{v}) \, \nu(\mathrm{d}\mathbf{v}) \right).$$

Gretton et al. (2012) considered taking $\mathcal{F}$ to be a unit ball in a reproducing kernel Hilbert space (RKHS), showing that when the RKHS is chosen judiciously then MMD is a metric on $\mathcal{P}(\mathbb{R}^d)$. Moreover, this choice ensures that an unbiased estimator for MMD can be constructed, as will now be explained. Recall that an RKHS is associated with a symmetric positive definite *kernel* $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$; we emphasise this using the notation $\mathcal{F} \equiv \mathcal{F}_k = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\}$ where $\mathcal{H}_k$ is the unique RKHS with kernel k and $\|\cdot\|_{\mathcal{H}_k}$ is the norm in $\mathcal{H}_k$. Define the *kernel mean embedding* of μ in $\mathcal{H}_k$ as $\mu[k]$ where $\mu[k](\mathbf{v}) := \int k(\mathbf{v}, \mathbf{v}')\mu(\mathrm{d}\mathbf{v}')$. Then Gretton et al. (2012, Lemma 4) asserts that $\mathrm{MMD}_{\mathcal{F}_k}(\mu, \nu)$ can be expressed as a difference between the kernel mean embeddings of μ and ν:

$$\mathrm{MMD}_{\mathcal{F}_k}(\mu, \nu) := \|\mu[k] - \nu[k]\|_{\mathcal{H}_k}. \tag{10}$$

For convenient choices of k and μ it may be possible to compute $\mu[k]$ in closed-form, but in general one must resort to approximating Eq. (10) based on samples from one or both of μ and ν. Given independent samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N \sim \mu_0$ and $\boldsymbol{Y}_1^{(m)}, \ldots, \boldsymbol{Y}_N^{(m)} \sim \nu_m$, we define an estimator

$$\widehat{\mathrm{MMD}}^2_{\mathcal{F}_k} := \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} k(\boldsymbol{X}_i, \boldsymbol{X}_j) + k(\boldsymbol{Y}_i^{(m)}, \boldsymbol{Y}_j^{(m)}) - k(\boldsymbol{X}_i, \boldsymbol{Y}_i^{(m)}) - k(\boldsymbol{Y}_i^{(m)}, \boldsymbol{X}_j), \tag{11}$$

which can be verified to be an unbiased estimator of $\mathrm{MMD}_{\mathcal{F}_k}(\mu, \nu_m)^2$ provided that, in addition to having the stated distribution, the samples $\boldsymbol{Y}_1^{(m)}, \ldots, \boldsymbol{Y}_N^{(m)}$ are generated independently from the samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$.

The statistic in Eq. (11) enables a goodness-of-fit test to be performed, and the distribution of this test statistic under the null hypothesis $\nu_m = \mu_0$ may be estimated using a standard bootstrap procedure as described in Gretton et al. (2012, Section 5). Having obtained M approximate samples from the distribution of Eq. (11) using the bootstrap, we determine a threshold for a prescribed power level $\alpha \in (0, 1)$ by computing a $(1-\alpha)$-quantile of this empirical distribution. This procedure will be used in Section 5, next, to empirically test whether PNM are weakly calibrated.

5. Empirical Assessment

The aim of this section is to empirically assess our proposed probabilistic iterative methods. For this purpose we consider the problem of inverting a linear system that arises when building a kernel interpolant. Our aim is not to address the problem of computing kernel interpolants *per se*, as many powerful methods exist for this task, but this problem serves as a convenient test-bed in which probabilistic iterative methods can be examined. The code to reproduce these results is available on GitHub[4].

4. https://github.com/jcockayne/probabilistic_iterative_methods_code

5.1 Problem Definition

Consider a dataset consisting of pairs (z_i, y_i), $i = 1, \ldots, d$, $d \in \mathbb{N}$, where the $z_i \in [0, 1]$ are distinct locations at which observations $y_i \in \mathbb{R}$ of some physical phenomenon were obtained. The aim is to compute a interpolant of this dataset, that is, a function $g : [0, 1] \to \mathbb{R}$ which is such that $g(z_i) = y_i$ for all $i = 1, \ldots, d$. For a given symmetric positive definite kernel $c : [0, 1] \times [0, 1] \to \mathbb{R}$, we consider an interpolant of the form

$$g(z) := \sum_{i=1}^{d} x_i c(z, z_i) \tag{12}$$

and note that there is a unique set of weights $x_i \in \mathbb{R}$ such that the interpolation equations

$$g(z_i) = y_i, \qquad i = 1, \ldots, d$$

are satisfied. The vector $\mathbf{x} = (x_1, \ldots, x_d)^\top$ of such weights satisfies the d-dimensional linear system in Eq. (1) with $A_{i,j} = c(z_i, z_j)$ and $\mathbf{b} = (y_1, \ldots, y_d)^\top$.

This linear system is representative of linear systems that are widely encountered in statistics and machine learning, and naturally a variety of methods have been proposed to circumvent the need to solve them; for example, based on reducing the degrees of freedom of the parametric function g so that the dataset is only approximately interpolated. Our aim is to use a finite number of iterations, m, of a probabilistic iterative method on the full problem in Eq. (1) and to lift the distribution μ_m over the unknown solution vector $\mathbf{x}$ into the function space spanned by functions of the form in Eq. (12). This enables uncertainty due to limited computation to be interpreted in the domain on which the interpolation problem was defined.

The condition number of $\mathbf{A}$ depends on the spectrum of the kernel c and the closeness of the elements in $\{z_1, \ldots, z_d\}$. For kernels with rapidly decaying spectrum, such as the squared exponential kernel

$$c(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\ell^2}\right) \tag{13}$$

with length-scale parameter $\ell > 0$, it is common for $\mathbf{A}$ to be badly conditioned. Thus even when d is small, solution of Eq. (1) can be difficult, which motivates the use of probabilistic methods that return a measure of error.

A dataset of size $d = 520$ was generated, with $(z_i)_{i=1,\ldots,d}$ consisting of 60 evenly spaced points in $[0, 0.1]$, 400 evenly spaced points in $[0.2, 0.8]$ and 60 evenly spaced points in $[0.9, 1]$, and $y_i = f(z_i)$ where $f(z) = 1_{z<0.5} \sin(2\pi z) + 1_{z \geq 0.5} \sin(4\pi z)$. The parameter $\ell = 0.0012$ was used, which produces a system for which a direct solver can be used, so that a ground-truth is accessible, but which is not entirely trivial.

5.2 Choice of μ_0

For the initial distribution μ_0 several candidates were considered. Firstly a DEFAULT choice given by $\mu_0 = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. This can be interpreted as a lack of *a priori* insight, since under this prior the components of $\mathbf{x}$ are independent and identically distributed. Secondly the

NATURAL choice $\mu_0 = \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$ which incorporates the structure of $\mathbf{A}$ into the initial distribution. This choice of prior covariance and has been noted to have desirable theoretical properties in the related work of Cockayne et al. (2019a); Hennig (2015); when used as a prior in those works it was shown to yield a posterior mean that coincides with the iterates from CG. We note that the NATURAL initial distribution is not a practical choice in general as it requires computation of $\mathbf{A}^{-1}$.

The third initial distribution we consider is applicable only in settings where a small number of *ansatz* solutions (i.e. guesses) are provided, perhaps obtained by expert knowledge of the system at hand. Let $\mathbf{x}_i$, $i = 1, \ldots, N$, be these *ansatz* solutions; we use these to estimate the scaling parameter ν^2 for an initial distribution $\mu_0 = \mathcal{N}(\mathbf{0}, \nu^2 \boldsymbol{\Sigma}_0)$ where $\boldsymbol{\Sigma}_0$ is fixed. Maximum likelihood estimation yields the estimator

$$\nu^2_{\text{opt}} := \frac{1}{Nd} \sum_{i=1}^{N} \|\mathbf{x}_i\|^2_{\boldsymbol{\Sigma}_0^{-1}},$$

which can be seen to adapt to the scale of the problem at hand; we call this approach OPT. In the experiments below where this approach is used we assume that $\boldsymbol{\Sigma}_0 = \mathbf{I}_d$, so that this choice is effectively DEFAULT with a scaling parameter that allows the prior to adapt to the scale of the problem. We used $N = 5$ *ansatz* solutions, obtained by sampling 5 right-hand-sides $\mathbf{B}_1, \ldots, \mathbf{B}_5 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and computing $\mathbf{x}_i = \mathbf{A}^{-1}\mathbf{B}_i$.

We note that this does not result in an entirely fair comparison since 5 exact solutions to the linear system are used to construct the initial distribution. One could consider instead using only approximate solutions, but this introduces additional degrees of freedom into the assessment. Since the focus of this paper goes beyond selecting μ_0, we simply use exact solutions within OPT for the assessment. Furthermore note that this choice of μ_0 violates the assumption from Proposition 10 that $\boldsymbol{\Sigma}_m$ is independent of $\mathbf{b}$, since $\boldsymbol{\Sigma}_0$ implicitly depends on $\mathbf{b}$ through the $\mathbf{x}_i$. Thus, this choice of prior tests the bounds of our theoretical results.

5.3 Results in Function Space

In this section we examine the resulting distributions μ_m from application of a number of probabilistic iterative methods to the problem above, for each choice of initial distribution from Section 5.2.

Stationary Iterative Methods We first consider Richardson's iteration with a constant step size. Since this method is stationary and linear, the theoretical results obtained in Section 3 apply. Fig. 1 displays samples (grey curves) from each of the probabilistic iterative methods that we considered and the blue curve represents the exact kernel interpolant. The step size ω was set to either the optimal value in Fig. 1b, $\omega = 2/(\lambda_{\min}(\mathbf{A}) + \lambda_{\max}(\mathbf{A}))$, that minimises the spectral radius of $\mathbf{G}$, or a default value $\omega = 2/3$ in Fig. 1a. Note that the optimal ω is not practical, since computing $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ is at least as expensive as solving the linear system. It is included to provide a point of comparison. Jacobi's method was also considered, but in our simulations the results were virtually identical for this problem, so they are not presented.

Examining first Fig. 1a, the output is seen to contract around the exact solution as the number m of iterations is increased, which is to be expected as the underlying iterative

methods are known to converge for this problem for any initial $\mathbf{x}_0$. Interestingly, very little variation is observed in the intervals $[0.1, 0.2]$ and $[0.8, 0.9]$ where no data points are located. In this region the posterior mean reverts to the prior mean, and so it is somewhat natural to expect fast convergence of the solver in this portion of the domain and slower convergence where more data points are concentrated. The low variation seen in the output of the probabilistic iterative method in these regions suggests that this is indeed the case, and therefore that the distributional output can act as a local error indicator.

The results for each of the three priors chosen appear to be very similar apart from changes in the posterior width, with DEFAULT the narrowest and OPT the widest. The increased width of OPT is to be expected since this prior is simply DEFAULT inflated by the parameter ν_{OPT}, which for this problem was computed to be 5.48.

Turning now to Fig. 1b, we note that both DEFAULT and NATURAL appear to exhibit some bias away from the true solution at $m = 3, 5$ and 10. We believe this is due to a violation of one of the central assumptions of Propositions 7 and 10, namedly that $\mathbf{x} \sim \mu_0$. Since in this case $\mathbf{x}$ depends upon pointwise values of the interpolant, it inherits smoothness properties that none of the priors considered encodes. This view is supported by results in Section 5.4, wherein Richardson iteration with optimal ω is shown to be weakly calibrated when $\mathbf{x}_0 \sim \mu_0$. This highlights the importance of prior selection. Note that this bias does not visually appear to be present for OPT, but we suggest that this is due to the increased width of the posterior.

Non-Stationary and Higher-Order Methods We now consider non-stationary and higher-order iterative methods. As discussed in Section 3.3, these methods are expected to be strongly calibrated as they are still linear, though calibration has not been rigorously established. For the non-stationary scheme we considered Richardson iteration again but with the step-size chosen adaptively, with $\omega_m = \mathbf{r}_m^\top \mathbf{A}\mathbf{r}_m / \|\mathbf{A}\mathbf{r}_m\|_2^2$ minimising the Euclidean norm of the residual $\mathbf{r}_{m+1} = \mathbf{b} - \mathbf{A}\mathbf{x}_{m+1}$. Results for the non-stationary scheme are presented in Fig. 2, with qualitative behaviour appearing to be similar to that with the default ω from Fig. 1a. Since the non-stationary scheme is better able to adapt to the problem at hand, this seems a more prudent choice than an arbitrary $\omega = 2/3$, though we note that the calibration of this method remains to be assessed empirically; this will be considered in Section 5.4.

As an example of a higher-order iterative method, we consider a second-degree version of Richardson iteration presented in Young (1972). In this method the iteration is of the form

$$\mathbf{x}_m = \gamma\sigma\left(\frac{2}{\beta-\alpha}\mathbf{G} - \frac{\beta+\alpha}{\beta-\alpha}\right)\mathbf{x}_{m-1} + (1-\gamma)\mathbf{x}_{m-2} + \frac{2\gamma\sigma}{\beta-\alpha}\mathbf{f}$$

where $\mathbf{G}$ and $\mathbf{f}$ are as given in the classical first-order Richardson iteration from Example 1, with optimal $\omega = 2/(\lambda_{\min}(\mathbf{A}) + \lambda_{\max}(\mathbf{A}))$, while $\alpha = \lambda_{\min}(\mathbf{G}), \beta = \lambda_{\max}(\mathbf{G})$ and

$$\sigma = \frac{\beta-\alpha}{2-(\beta-\alpha)} \qquad \gamma = \frac{2}{1+\sqrt{1-\sigma^2}}.$$

Recall that for a second degree probabilistic iterative method, a joint initial distribution must be specified for $\mathbf{x}_0$ and $\mathbf{x}_1$. The distribution assigned to $\mathbf{x}_0$ was fixed to OPT, since, in the results for $s = 1$ (to follow), this appeared to provide better UQ across different

choices of ω. Three choices were considered for initial distributions for $\mathbf{x}_1$: IID, in which $\mathbf{x}_1$ is an independent copy of $\mathbf{x}_0$, CORR, in which $\mathbf{x}_1$ is identical to $\mathbf{x}_0$ and RICH, in which $\mathbf{x}_1$ is obtained from $\mathbf{x}_0$ by performing one iteration of Richardson iteration with optimal ω. Note that both IID and CORR yield the same marginal distribution for $\mathbf{x}_1$, but the joint distributions differ.

Fig. 3 displays samples from the output of the probabilistic iterative methods just described. Qualitatively, the results appear to be similar to those from Fig. 1b with initial distribution OPT, as one would expect given that μ_0 in all three rows is that same distribution. Of the three choices for μ_1, RICH appears to contract marginally faster, though in all three methods the improvement over the first order method from Fig. 1b appears to be negligible.

Nonlinear Methods Here we consider a probabilistic iterative method based on CG, which is one of the most widely used iterative methods, but for which our theoretical results on strong calibration do not hold. Results are displayed in Fig. 4. Convergence is clearly seen to be faster than in the other methods considered, though qualitatively the samples obtained otherwise seem to be similar. This hints at the results from the next section, in which we will see that CG is weakly calibrated for this problem and for the initial distributions that we considered.

5.4 Testing Calibration

We now test for evidence against weak calibration for all of the probabilistic iterative methods and initial distributions considered. Recall that, according to the results in Section 3.2, stationary Richardson iterations give rise to probabilistic iterative methods that are strongly calibrated when ω is fixed (irrespective of whether the optimal ω or a fixed ω is used). Non-stationary Richardson iteration with adaptive parameter ω_m is conjectured to also give rise to a probabilistic iterative method that is strongly calibrated, as is the higher order method described above, but these strong calibration results have not been established. It is unknown whether probabilistic iterative methods based on CG are strongly or weakly calibrated. In addition to probabilistic iterative methods, we also include BayesCG from Cockayne et al. (2019a), which is not a probabilistic iterative method in the sense of this paper and is not expected to be strongly calibrated owing to the negative results presented in Cockayne et al. (2019a) and in Reid et al. (2020). It was hitherto unknown whether BayesCG is weakly calibrated.

To test the hypothesis that probabilistic iterative methods are weakly calibrated, we apply the MMD-based test described in Section 4.2. For each initial distribution and each iterative method we generated $N = 100$ independent samples from μ_0 and ν_m from which the test statistic Eq. (11) was computed. Significance was assessed using the bootstrap method with $M = 1000$. The kernel k used was the squared exponential kernel from Eq. (13), with the length-scale set using the *median heuristic* as recommended in Gretton et al. (2012). For each method, $m = 10$ iterations were performed. For the second order method, we opted to use the RICH initial distribution for $\mathbf{x}_1$.

Table 1 shows test statistics obtained for each of these methods arising in the test for weak calibration described in Section 4.2, for each choice of initial distribution from Section 5.2. Reported are the value of Eq. (11) (as MMD in Table 1). Note that while

		Rich. (default)	Rich. (optimal)	Rich. (adaptive)	Rich. (2o)	CG	BayesCG
DEFAULT	$\text{MMD}^2_{\mathcal{F}_k}$	1.90e-04	-3.11e-05	9.76e-06	5.36e-05	-2.80e-05	**1.14e-03**
	q	0.34	0.52	0.45	0.43	0.49	**0.03**
NATURAL	$\text{MMD}^2_{\mathcal{F}_k}$	-1.72e-04	-2.71e-04	-2.44e-04	-3.20e-04	-2.98e-04	**4.18e-03**
	q	0.60	0.64	0.64	0.68	0.68	**0.00**
OPT	$\text{MMD}^2_{\mathcal{F}_k}$	3.59e-05	1.00e-05	4.30e-06	-6.62e-06	3.57e-05	**6.57e-03**
	q	0.48	0.47	0.48	0.49	0.47	**0.00**

Table 1: Results from applying the maximum mean discrepancy (MMD)-based test from Section 4.2 to the methods described in Section 5. The abbreviation "Rich." refers to Richardson iteration. "2o" refers to the second order method. The test does not reject the null that each of the methods assessed is weakly calibrated, with the exception of BayesCG where the null is rejected. Results that are statistically significant at the 5% level, indicating that the method is not weakly calibrated, are highlighted in bold.

strictly speaking MMD ought to be positive, due to sampling error it may be negative; this was also observed in Gretton et al. (2012). Also reported is the statistic q, which is analogous to a p-value in a classical statistical test. To compute this we again used a bootstrap-based method. In detail, we bootstrapped a sample of size M of Eq. (11) by pooling the samples from μ_0 and ν_m, sampling with replacement two samples of size N from the pooled samples and computing Eq. (11). We then computed the empirical quantile q' of the obtained value of MMD within this sample, by placing the sample in ascending order, computing the rank r of the value immediately below MMD within the bootstrapped sample and letting $q' = r/M$. We then took $q = 1 - q'$. We used the value $\alpha = 0.05$, representing a 5% significance level, as a threshold in Table 1; thus, if a value of q below 0.05 was obtained this constitutes evidence that the method is *not* weakly calibrated. Note that owing to the fact that q is based on a sample from the bootstrapped distribution, it is possible to obtain $q = 0$; we would expect the true p-value to be small but positive.

Examining the results, Richardson iteration with both default and optimal ω is seen to be weakly calibrated. This provides support for our testing methodology, since from Cockayne et al. (2020, Lemma 2.19) any strongly calibrated PNM must be weakly calibrated. Similarly the second order method is weakly calibrated, which is to be expected since the proof of strong calibration for this method would require only a small extension relative to the case of a first order method. Richardson iteration with the adaptive ω appears to be weakly calibrated for all initial distributions considered, suggesting that the non-stationarity implied by the adaptive parameter does not affect the weak calibration of the method. Also note that the fact that for OPT, $\boldsymbol{\Sigma}_m$ implicitly depends upon $\mathbf{b}$ through the parameter ν_{opt}, does not appear to affect the calibration of any of the methods considered, suggesting that this theorem may be generalisable.

Perhaps more surprisingly, owing to its high degree of nonlinearity, CG also appears to be weakly calibrated. This hints at the possibility of a more fundamental result regarding the calibration of probabilistic iterative methods in the general setting, though we leave study of this conjecture to future work.

Concerning BayesCG (which we emphasise again is *not* a probabilistic iterative method in the same sense as the other methods considered), the results show that BayesCG is *not* weakly calibrated for either the NATURAL or OPT initial distributions μ_0 even when the prior distribution, required in BayesCG, is set equal to μ_0 itself. This is to be expected, considering that this method is known *not* to produce meaningful posteriors apart from in special cases (e.g. Reid et al., 2020). One other noteworthy point is that for the DEFAULT initial distribution the MMD obtained for BayesCG has a slightly higher value of $q = 0.03$. This is perhaps due to the fact that, with such an uninformative prior, BayesCG is known to converge quite slowly. Thus the posterior after 10 iterations may not have deviated far from the prior.

5.5 Spectral Behaviour

Lastly we examine the spectral behaviour of one of the methods above by performing a principal component analysis, to illustrate how the output of a probabilistic iterative method can provide a richer description of error compared to a classical error bound. In this section we fixed the distribution μ_0 to NATURAL.

Here we consider principal components (leading eigenvectors) of the covariance matrix $\mathbf{A}\boldsymbol{\Sigma}_m\mathbf{A}^\top$, which describes covariance in the domain of the function Eq. (12). The six leading principal components for the probabilistic iterative method based on Richardson iteration with default parameter $\omega = 2/3$ are displayed in Figure 5. At each of the values of m considered, the low frequency variation over the interval $[0.2, 0.8]$ is seen to be the dominant principal component (more so as m is increased), which accords with the result of Figure 1a in that the error of NATURAL is mainly manifest in a low-frequency vertical shift between the exact interpolant and the sampled output. At $m = 100$ the first six components account for over 50% of the variability in the distributional output, with the remaining variability dedicated to higher-frequency aspects of the solution.

The detailed nature of these error indicators may be useful to shed light on the aspects of the exact solution $\mathbf{x}$ that we are most uncertain about, having run a finite number of iterations of a probabilistic iterative method. This rich description of numerical uncertainty can trivially be propagated through subsequent computation $F(\mathbf{x})$, e.g. by sampling from μ_m and then applying F, in order to probabilistically assess the impact of numerical uncertainty on any subsequent computational output.

6. Conclusion

In this paper we have introduced probabilistic iterative methods, a new class of probabilistic numerical methods for solving linear systems. We have provided theoretical results concerning the convergence and calibration of these methods in the stationary and linear setting, and examined their empirical performance using a synthetic test-bed. Finally, we alluded to how the output of a probabilistic iterative method could be used represent *numerical uncertainty* and how such a representation could be propagated through subsequent computational output.

Several interesting avenues for future related work are now highlighted:

6.1 Generalisation to Nonlinear Methods

The generalisation of this work to nonlinear iterative methods, such as CG (Hestenes and Stiefel, 1952) and other Krylov methods is of interest. These methods are more widely used than stationary iterative methods in modern applications, owing both to their faster convergence and that they only require access to the action of $\mathbf{A}$, rather than needing to interrogate and modify the elements of $\mathbf{A}$.

The definition that we proposed for probabilistic iterative methods in Definition 1, and the sampling algorithm for accessing the output of a probabilistic iterative method described in Section 3, do not require the generating iterative method to be linear. However, with the exception of Proposition 2, the theoretical results presented in this paper depend strongly on linearity. The experimental results in Section 5.4 indicate that CG, a prototypical nonlinear iterative method, may be weakly calibrated. The goal of theoretically establishing the calibration properties of nonlinear probabilistic iterative methods represents interesting future work.

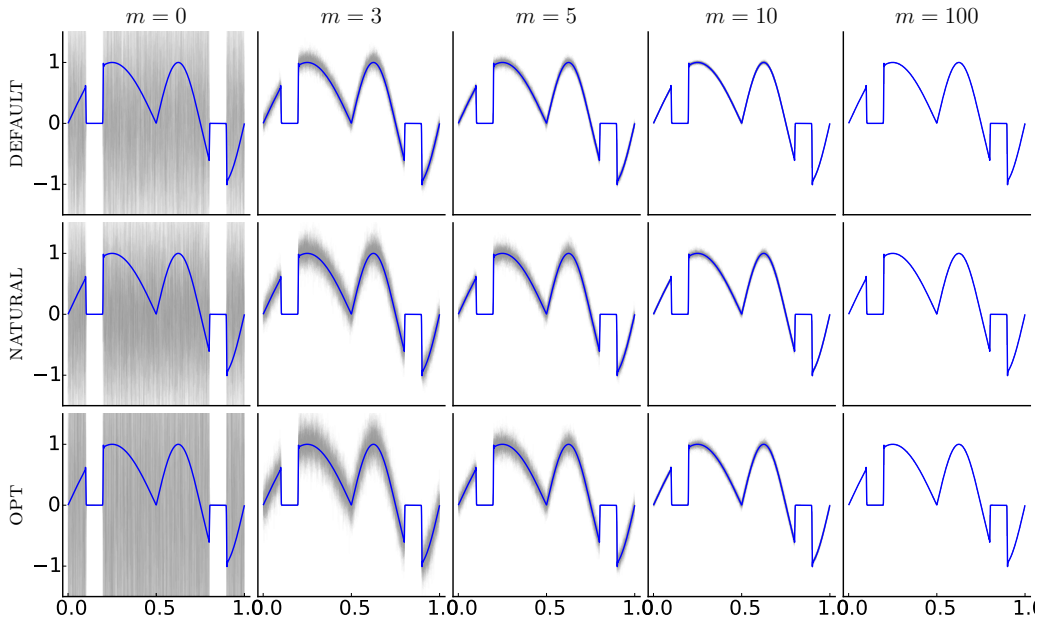
6.2 Gradient Flow Interpretation

Recent work in the numerical analysis community highlights that iterative methods for linear systems may be interpreted as the discrete-time solution of an underlying dynamical system on $\mathbb{R}^d$ (Chu, 2008). Insight may then be gained by studying the original dynamical system. In parallel, recent work in the statistics and machine learning communities has provided gradient flow interpretations of various sampling and variational inference algorithms on $\mathcal{P}(\mathbb{R}^d)$ (e.g. Arbel et al., 2019; Liu et al., 2019) An interesting avenue for future work would be to consider whether the methods presented in this paper may be interpreted as a discretisation of a gradient flow on $\mathcal{P}(\mathbb{R}^d)$, and whether insight can be gained by performing analysis of the continuous flow.

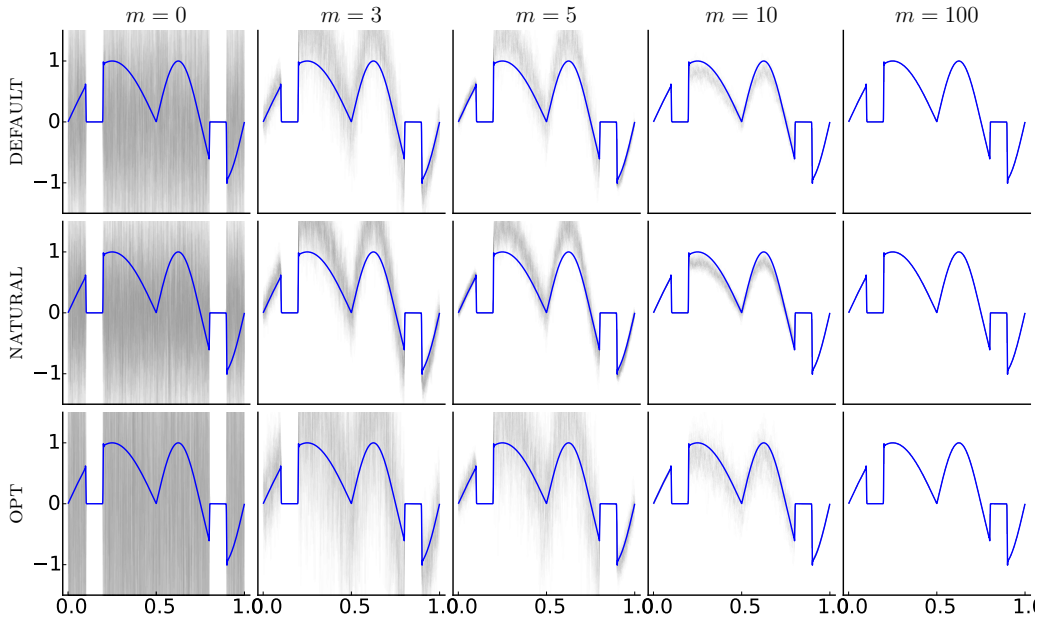
6.3 Wider Applications

In this paper we have focussed on iterative methods for solving linear systems. However, the assumption that $\mathcal{I}$ was an iterative method for solving such systems was not essential to Definition 1. Provided an initial distribution μ_0 can be constructed in the domain of $\mathcal{I}_{\#}$, probabilistic iterative methods could be applied to *any* classical problem for which iterative methods are used, such as solvers for eigenproblems, numerical optimisation problems or even solvers for nonlinear differential equations. Proposition 2 also applies to this general case, provided a suitable bound of the form in Eq. (3) can be derived in a norm adapted to the problem and, when the iteration is an affine map, we expect that the proof techniques from Section 3.2 could be applied.

Acknowledgements JC was supported by Wave 1 of the UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Digital Twins for Complex Engineering Systems" theme within that grant, and the Alan Turing Institute. The work of ICFI was supported in part by National Science Foundation grants DMS-1760374 and DMS-1745654. CJO was supported by the Lloyd's Register Foundation programme on data-centric engineering at the Alan Turing Institute, UK. The work of TWR was supported in part by National Science Foundation grant DMS-1745654.



(a) Default ω



(b) Optimal ω

Figure 1: Samples from the distributional output of a probabilistic iterative method based on Richardson iteration, used to solve an interpolation problem and visualised in the physical domain in which the interpolant is defined. The rows in each figure represent the three choices of initial distribution described in Section 5.2. In each panel we present 50 samples (grey curves) from the output of the probabilistic iterative method after m iterations have been performed. The interpolant, corresponding to the exact solution of the linear system, is also shown in blue.

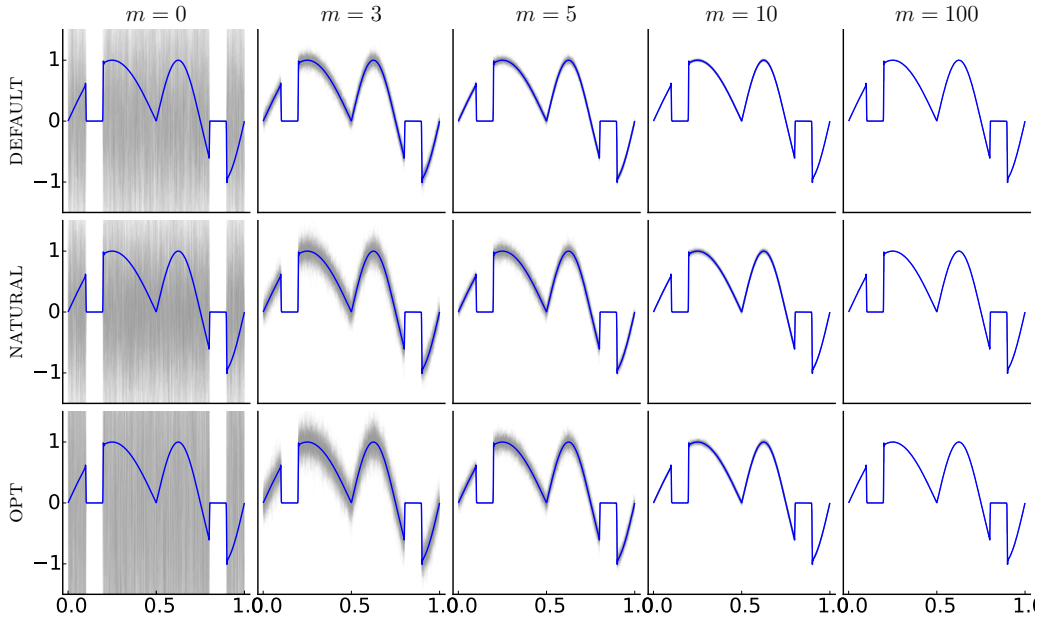


Figure 2: As in Fig. 1, but with ω chosen adaptively.

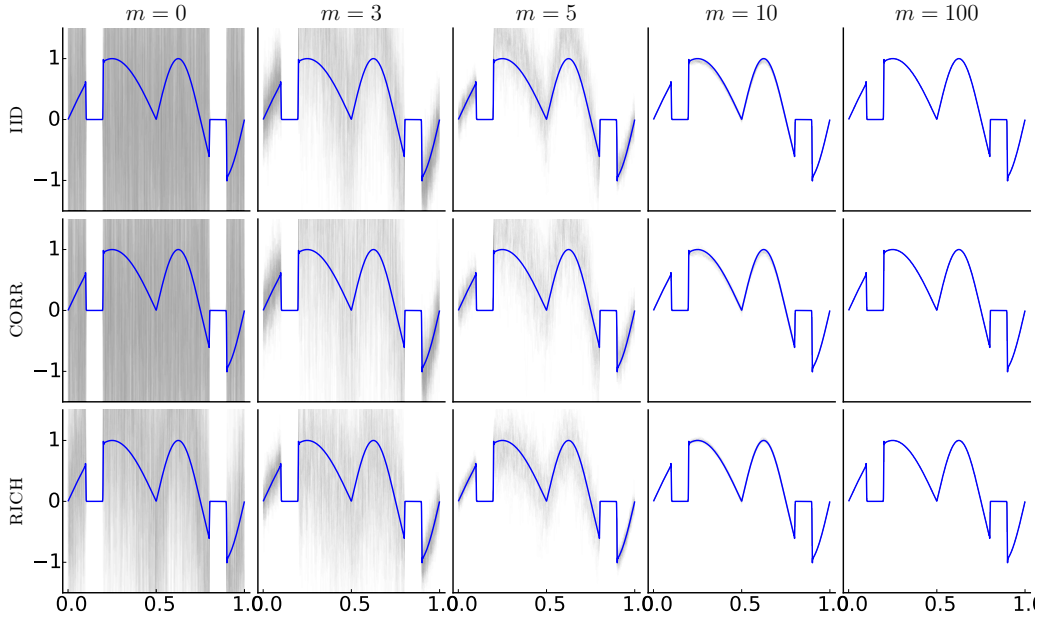


Figure 3: A probabilistic iterative method based on a second degree version of Richardson iteration, as described in Section 5. Each row uses OPT as the initial distribution for $\mathbf{x}_0$ and a different initial distribution for $\mathbf{x}_1$, as described in the main text.

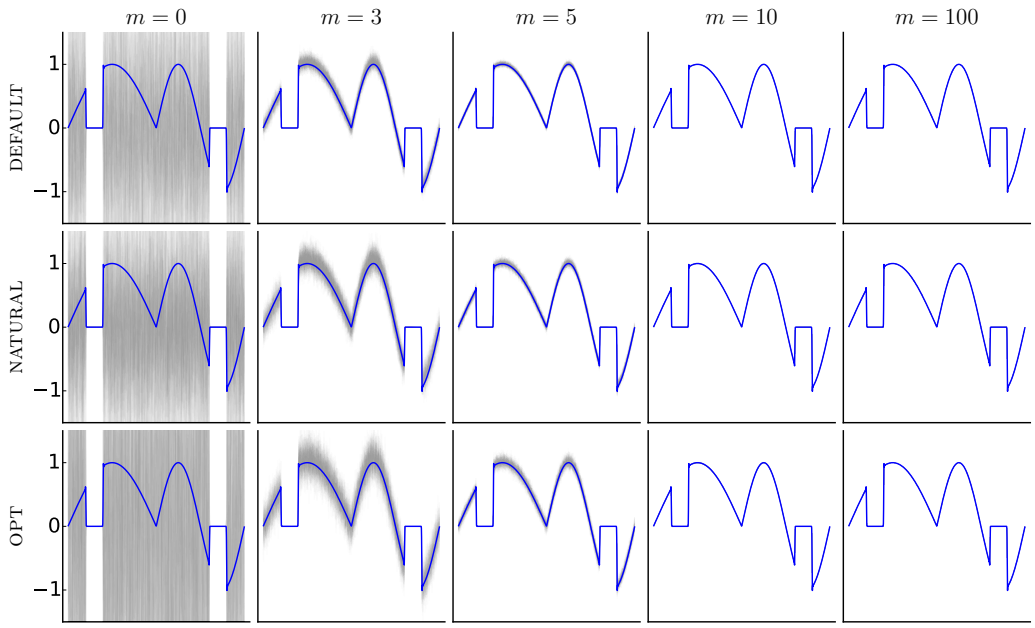


Figure 4: Samples from the distributional output of the probabilistic iterative method implied by using the conjugate gradient method as the underlying iterative method.

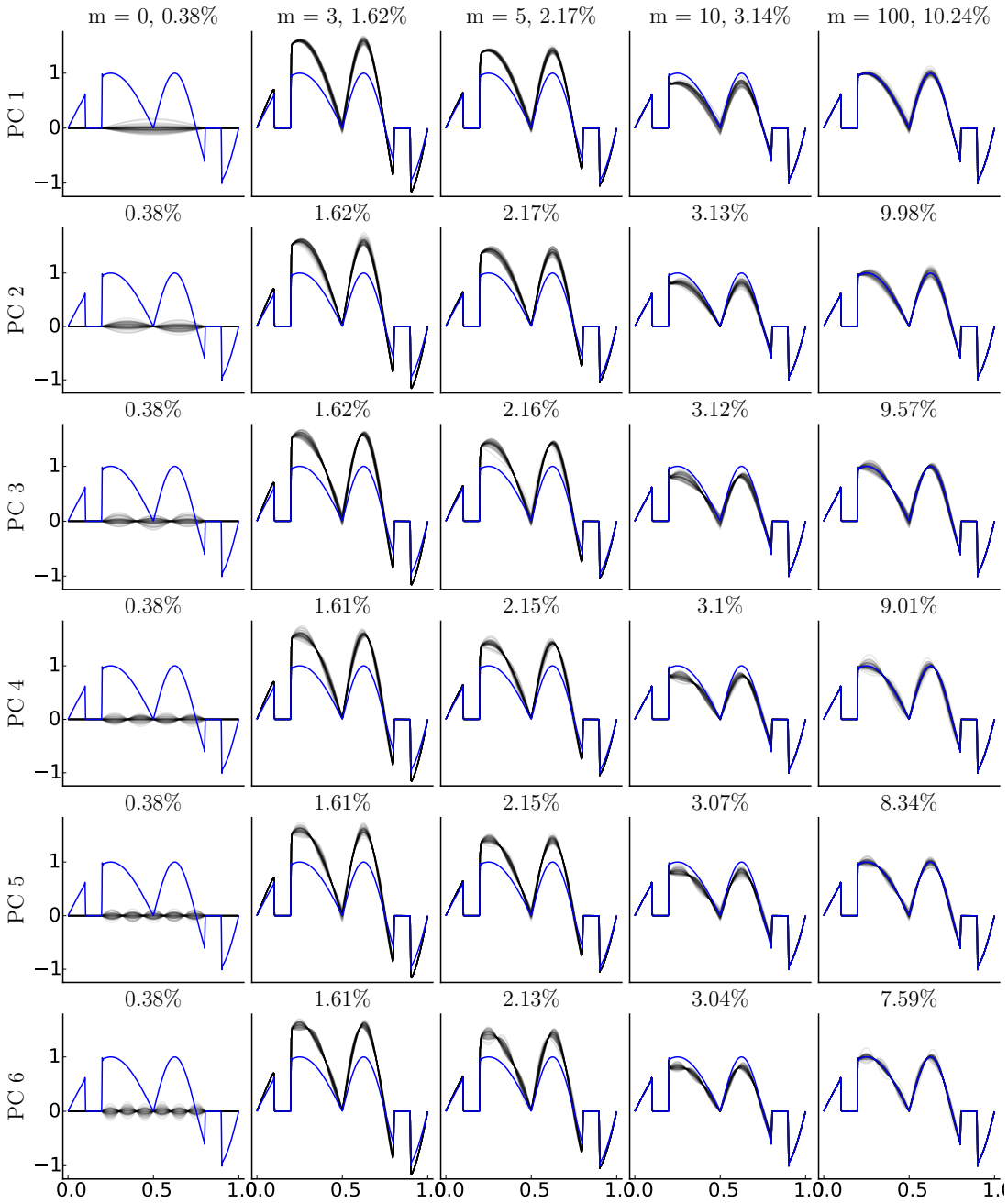


Figure 5: A closer look at the distributional output: principal components from a probabilistic iterative method based on Richardson iteration with the default parameter and initial distribution NATURAL. Here the first 6 principal components (PC) are displayed for the same values of m used in Figure 1. The percentages indicate the percentage of the total variation that is explained by that component. Each grey line is constructed as the mean of μ_m, plus a sample in the direction of the relevant principal component, re-scaled to improve visualisation, with 50 samples shown in total.

Appendix A. Proof of Proposition 10

In order to prove Proposition 10, we need several results from linear algebra about the range and kernel of products of matrices, as well as decomposition of a diagonalizable matrix.

Lemma 15 (Ipsen (2009, Fact 6.3)). *Let $\mathbf{Y}, \mathbf{W} \in \mathbb{R}^{d\times d}$. If $\mathbf{Y}$ is non-singular, then* $\ker(\mathbf{YW}) = \ker(\mathbf{W})$.

Lemma 16 (Ipsen (2009, Facts 6.3 and 6.4)). *Let $\mathbf{Y}, \mathbf{\Omega}, \mathbf{W} \in \mathbb{R}^{d\times d}$ where $\mathbf{Y}$ and $\mathbf{W}$ are non-singular. If $\mathbf{Y}$, $\mathbf{\Omega}$, and $\mathbf{W}$ have the partitions*

$$\mathbf{Y} = \begin{bmatrix}\mathbf{Y}_1 & \mathbf{Y}_2\end{bmatrix}, \quad \mathbf{\Omega} = \begin{bmatrix}\mathbf{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \quad \text{and} \quad \mathbf{W} = \begin{bmatrix}\mathbf{W}_1^\top \\ \mathbf{W}_2^\top\end{bmatrix},$$

with $\mathbf{Y}_1, \mathbf{W}_1 \in \mathbb{R}^{d\times r}$, $\mathbf{Y}_2, \mathbf{W}_2 \in \mathbb{R}^{d\times(d-r)}$, and $\mathbf{\Omega}_{11} \in \mathbb{R}^{r\times r}$ nonsingular, then

$$\text{range}(\mathbf{Y\Omega W}) = \text{range}(\mathbf{Y}_1) \quad \text{and} \quad \ker(\mathbf{Y\Omega W}) = \text{range}(\mathbf{W}_2).$$

Lemma 17. *Horn and Johnson (2009, Lemma 3.4.1.10) Let $\mathbf{G} \in \mathbb{R}^{d\times d}$ be diagonalisable and of rank $r < d$. Then $\mathbf{G}$ may be represented in its real Jordan canonical form as*

$$\mathbf{G} = \mathbf{Y\Omega Y}^{-1}$$

where $\mathbf{Y} \in \mathbb{R}^{d\times d}$ is invertible, while $\mathbf{\Omega} \in \mathbb{R}^{d\times d}$ is of the form

$$\mathbf{\Omega} = \begin{pmatrix}\mathbf{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}.$$

Here $\mathbf{\Omega}_{11} \in \mathbb{R}^{r\times r}$ is nonsingular and block-diagonal, with ℓ 2×2 blocks and s 1×1 blocks, where ℓ is the number of nonzero conjugate pairs of complex eigenvalues of $\mathbf{G}$ and s is the number of nonzero real eigenvalues of $\mathbf{G}$, so that $r = 2\ell + s$.

With these results stated we proceed to the main proof:

Proof [Proof of Proposition 10] First note that if $\text{rank}(\mathbf{G}) = d$ then $\mathbf{G}$ is invertible, so the probabilistic iterative method is strongly calibrated as a result of Proposition 7. Thus we focus on the case that $\text{rank}(\mathbf{G}) < d$.

We complete this proof in multiple steps:

Step 1 We express the range and kernel of $\mathbf{\Sigma}_m$ in terms of the matrices forming the real Jordan canonical form of $\mathbf{G}$, thus identifying the matrices $\mathbf{R}$ and $\mathbf{N}$ from Proposition 10.

Step 2 We compute $(\mathbf{R}^\top\mathbf{\Sigma}_m\mathbf{R})^{1/2}$, $(\mathbf{R}^\top\mathbf{\Sigma}_m\mathbf{R})^{1/2}\mathbf{R}^\top(\mathbf{x}_m - \mathbf{x})$ and $\mathbf{N}^\top(\mathbf{x}_m - \mathbf{x})$.

Step 3 We combine these results to show that stationary iterative methods are strongly calibrated when $\mathbf{G}$ is diagonalisable.

Step 1 We first compute the range and kernel of $\mathbf{\Sigma}_m$. This covariance matrix is defined as

$$\mathbf{\Sigma}_m = \mathbf{G}^m \mathbf{\Sigma}_0 (\mathbf{G}^m)^\top.$$

From Lemma 17 we have that

$$\mathbf{G}^i = \mathbf{Y} \mathbf{\Omega}^i \mathbf{Y}^{-1}, \qquad 0 \leq i \leq m.$$

We partition the factors above as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 \end{bmatrix}, \quad \mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \text{and} \quad \mathbf{Y}^{-1} = \begin{bmatrix} \mathbf{W}_1^\top \\ \mathbf{W}_2^\top \end{bmatrix},$$

where $\mathbf{Y}_1, \mathbf{W}_1 \in \mathbb{R}^{d \times r}$, $\mathbf{Y}_2, \mathbf{W}_2 \in \mathbb{R}^{d \times (d-r)}$, and $\mathbf{\Omega}_{11} \in \mathbb{R}^{r \times r}$ is nonsingular. With this partitioning and Lemma 16 we have

$$\operatorname{range}(\mathbf{G}^i) = \operatorname{range}(\mathbf{Y}_1) \quad \text{and} \quad \ker((\mathbf{G}^i)^\top) = \operatorname{range}(\mathbf{W}_2), \qquad 0 \leq i \leq m. \tag{14}$$

We now express the range and kernel of $\mathbf{\Sigma}_m$ in terms of $\mathbf{Y}_1$ and $\mathbf{W}_2$. Express $\mathbf{\Sigma}_m$ as the product $\mathbf{\Sigma}_m = \mathbf{Q}\mathbf{Q}^\top$, where $\mathbf{Q} = \mathbf{G}^m \mathbf{\Sigma}_0^{1/2}$. For any $\mathbf{v} \in \ker(\mathbf{\Sigma}_m)$ we have

$$\mathbf{\Sigma}_m \mathbf{v} = \mathbf{0} \iff \mathbf{v}^\top \mathbf{\Sigma}_m \mathbf{v} = \mathbf{0} \iff (\mathbf{Q}^\top \mathbf{v})^\top \mathbf{Q}^\top \mathbf{v} = \mathbf{0} \iff \mathbf{Q}^\top \mathbf{v} = \mathbf{0}$$

where the first equivalence above holds because Σ_m is symmetric positive-semi-definite. Thus $\ker(\mathbf{\Sigma}_m) = \ker(\mathbf{Q}^\top)$. Because $\mathbf{\Sigma}_0^{1/2}$ is the non-singular square root of the non-singular matrix $\mathbf{\Sigma}_0$, we can apply Lemma 15 to $\mathbf{Q}^\top = \mathbf{\Sigma}_0^{1/2} (\mathbf{G}^m)^\top$ to obtain

$$\ker(\mathbf{\Sigma}_m) = \ker(\underbrace{\mathbf{\Sigma}_0^{1/2} (\mathbf{G}^m)^\top}_{\mathbf{Q}^\top}) = \ker((\mathbf{G}^m)^\top). \tag{15}$$

By the fundamental theorem of linear algebra, $\ker((\mathbf{G}^m)^\top)$ is the orthogonal complement of $\operatorname{range}(\mathbf{G}^m)$ and $\ker(\mathbf{\Sigma}_m)$ is the orthogonal complement of $\operatorname{range}(\mathbf{\Sigma}_m^\top) = \operatorname{range}(\mathbf{\Sigma}_m)$. This combined with Eq. (15) implies

$$\operatorname{range}(\mathbf{\Sigma}_m) = \operatorname{range}(\mathbf{G}^m). \tag{16}$$

Applying Lemma 16 with $\mathbf{W} = \mathbf{Y}^{-1}$ gives

$$\operatorname{range}(\mathbf{\Sigma}_m) = \operatorname{range}(\mathbf{Y}_1) \quad \text{and} \quad \ker(\mathbf{\Sigma}_m) = \operatorname{range}(\mathbf{W}_2). \tag{17}$$

Therefore, referring to Proposition 10, we have that $\mathbf{R} = \mathbf{Y}_1$ and $\mathbf{N} = \mathbf{W}_2$.

Step 2 We begin by computing $(\mathbf{Y}_1^\top \mathbf{\Sigma}_m \mathbf{Y}_1)^{1/2}$. We have that

$$\begin{aligned} \mathbf{Y}_1^\top \mathbf{\Sigma}_m \mathbf{Y}_1 &= \mathbf{Y}_1^\top \mathbf{G}^m \mathbf{\Sigma}_0 (\mathbf{G}^m)^\top \mathbf{Y}_1 \\ &= \mathbf{Y}_1^\top \mathbf{Y} \mathbf{\Omega}^m \mathbf{Y}^{-1} \mathbf{\Sigma}_0 \mathbf{Y}^{-\top} (\mathbf{\Omega}^m)^\top \mathbf{Y}^\top \mathbf{Y}_1 \\ &= \mathbf{Y}_1^\top \mathbf{Y}_1 \mathbf{\Omega}_{11}^m \mathbf{W}_1^\top \mathbf{\Sigma}_0 \mathbf{W}_1 (\mathbf{\Omega}_{11}^m)^\top \mathbf{Y}_1^\top \mathbf{Y}_1. \end{aligned}$$

The product $\mathbf{Y}_1^\top \mathbf{Y}_1$ is Hermitian positive definite because $\mathbf{Y}_1$ is full rank. Additionally, $\mathbf{Y}_1 \mathbf{W}_1^\top = \mathbf{I}_r$ because $\mathbf{Y}\mathbf{Y}^{-1} = \mathbf{I}_d$. Therefore the inverse square root[5] is,

$$(\mathbf{Y}_1^\top \boldsymbol{\Sigma}_m \mathbf{Y}_1)^{-1/2} = \mathbf{B}\boldsymbol{\Omega}_{11}^{-m}(\mathbf{Y}_1^\top \mathbf{Y}_1)^{-1}, \tag{18}$$

where $\mathbf{B} = (\mathbf{W}_1^\top \boldsymbol{\Sigma}_0 \mathbf{W}_1)^{-1/2} \in \mathbb{R}^{r \times r}$.

Next, we compute $(\mathbf{Y}_1^\top \boldsymbol{\Sigma}_m \mathbf{Y}_1)^{1/2} \mathbf{Y}_1^\top (\mathbf{x} - \mathbf{x}_m)$. Left-multiplying $\mathbf{x} - \mathbf{x}_m$ by $\mathbf{Y}_1^\top$ yields

$$\begin{aligned}
\mathbf{Y}_1^\top (\mathbf{x} - \mathbf{x}_m) &= \mathbf{Y}_1^\top \left(\mathbf{x} - \mathbf{G}^m \mathbf{x}_0 - \sum_{i=0}^{m-1} \mathbf{G}^i \mathbf{f} \right) && (19) \\
&= \mathbf{Y}_1^\top \left(\mathbf{x} - \mathbf{Y}\boldsymbol{\Omega}^m \mathbf{Y}^{-1} \mathbf{x}_0 - \mathbf{f} - \sum_{i=1}^{m-1} \mathbf{Y}\boldsymbol{\Omega}^i \mathbf{Y}^{-1} \mathbf{f} \right) \\
&= \mathbf{Y}_1^\top \mathbf{x} - \mathbf{Y}_1^\top \mathbf{Y}_1 \boldsymbol{\Omega}_{11}^m \mathbf{W}_1^\top \mathbf{x}_0 - \mathbf{Y}_1^\top \mathbf{f} - \sum_{i=1}^{m-1} \mathbf{Y}_1^\top \mathbf{Y}_1 \boldsymbol{\Omega}_{11}^i \mathbf{W}_1^\top \mathbf{f}. && (20)
\end{aligned}$$

Now left-multiplying by Eq. (18) gives

$$\begin{aligned}
&(\mathbf{Y}_1^\top \boldsymbol{\Sigma}_m \mathbf{Y}_1)^{-1/2} \mathbf{Y}_1^\top (\mathbf{x} - \mathbf{x}_m) \\
&\qquad = \underbrace{\mathbf{B}\boldsymbol{\Omega}_{11}^{-m}(\mathbf{Y}_1^\top \mathbf{Y}_1)^{-1} \left(\mathbf{Y}_1^\top (\mathbf{x} - \mathbf{f}) - \sum_{i=1}^{m-1} \mathbf{Y}_1^\top \mathbf{Y}_1 \boldsymbol{\Omega}_{11}^i \mathbf{W}_1^\top \mathbf{f} \right)}_{(\star)} - \mathbf{B}\mathbf{W}_1^\top \mathbf{x}_0. \qquad (21)
\end{aligned}$$

We now focus on simplifying $(\star)$. Left-multiplying Eq. (5) by $\mathbf{Y}_1^\top$ gives

$$\begin{aligned}
\mathbf{Y}_1^\top \mathbf{x} &= \mathbf{Y}_1^\top \mathbf{Y}_1 \boldsymbol{\Omega}_{11} \mathbf{W}_1^\top \mathbf{x} + \mathbf{Y}_1^\top \mathbf{f}. \\
\implies \mathbf{W}_1^\top \mathbf{x} &= \boldsymbol{\Omega}_{11}^{-1} (\mathbf{Y}_1^\top \mathbf{Y}_1)^{-1} \mathbf{Y}_1^\top (\mathbf{x} - \mathbf{f}) && (22)
\end{aligned}$$

while left-multiplying by $\mathbf{W}_1^\top$ gives

$$\begin{aligned}
\mathbf{W}_1^\top \mathbf{x} &= \boldsymbol{\Omega}_{11} \mathbf{W}_1^\top \mathbf{x} + \mathbf{W}_1^\top \mathbf{f} \\
\implies \mathbf{W}_1^\top \mathbf{x} &= \boldsymbol{\Omega}_{11}^{-1} \mathbf{W}_1^\top (\mathbf{x} - \mathbf{f}). && (23)
\end{aligned}$$

Substituting Eq. (22) into $(\star)$ results in

$$\begin{aligned}
(\star) &= \mathbf{B}\boldsymbol{\Omega}_{11}^{-m}(\mathbf{Y}_1^\top \mathbf{Y}_1)^{-1} \left(\mathbf{Y}_1^\top (\mathbf{x} - \mathbf{f}) - \sum_{i=1}^{m-1} \mathbf{Y}_1^\top \mathbf{Y}_1 \boldsymbol{\Omega}_{11}^i \mathbf{W}_1^\top \mathbf{f} \right) \\
&= \mathbf{B}\boldsymbol{\Omega}_{11}^{-(m-1)} \left(\boldsymbol{\Omega}_{11}^{-1}(\mathbf{Y}_1^\top \mathbf{Y}_1)^{-1} \mathbf{Y}_1^\top (\mathbf{x} - \mathbf{f}) - \boldsymbol{\Omega}_{11}^{-1}(\mathbf{Y}_1^\top \mathbf{Y}_1)^{-1} \sum_{i=1}^{m-1} \mathbf{Y}_1^\top \mathbf{Y}_1 \boldsymbol{\Omega}_{11}^i \mathbf{W}_1^\top \mathbf{f} \right) \\
&= \mathbf{B}\boldsymbol{\Omega}_{11}^{-(m-1)} \left(\mathbf{W}_1^\top \mathbf{x} - \mathbf{W}_1^\top \mathbf{f} - \sum_{i=1}^{m-2} \boldsymbol{\Omega}_{11}^i \mathbf{W}_1^\top \mathbf{f} \right).
\end{aligned}$$

5. This is a square root in the sense of Section 1.4, a matrix $\mathbf{T}^{1/2}$ such that $\mathbf{T}^{1/2}(\mathbf{T}^{1/2})^\top = \mathbf{T}$.

Repeatedly substituting Eq. (23) into the previous equation gives

$$\begin{aligned}
(\star) &= \mathbf{B}\boldsymbol{\Omega}_{11}^{-(m-1)}\left(\mathbf{W}_1^\top(\mathbf{x}-\mathbf{f}) - \sum_{i=1}^{m-2}\boldsymbol{\Omega}_{11}^{i}\mathbf{W}_1^\top\mathbf{f}\right)\\
&= \mathbf{B}\boldsymbol{\Omega}_{11}^{-(m-2)}\left(\boldsymbol{\Omega}_{11}^{-1}\mathbf{W}_1^\top(\mathbf{x}-\mathbf{f}) - \boldsymbol{\Omega}_{11}^{-1}\sum_{i=1}^{m-2}\boldsymbol{\Omega}_{11}^{i}\mathbf{W}_1^\top\mathbf{f}\right)\\
&= \mathbf{B}\boldsymbol{\Omega}_{11}^{-(m-2)}\left(\mathbf{W}_1^\top(\mathbf{x}-\mathbf{f}) - \sum_{i=1}^{m-3}\mathbf{W}_1^\top\mathbf{f}\right)\\
&\;\;\vdots\\
&= \mathbf{B}\left(\boldsymbol{\Omega}_{11}^{-1}\mathbf{W}_1^\top(\mathbf{x}-\mathbf{f})\right)\\
&= \mathbf{B}\mathbf{W}_1^\top\mathbf{x}.
\end{aligned}$$

Finally substituting this back into Eq. (21) shows

$$(\mathbf{Y}_1^\top\boldsymbol{\Sigma}_m\mathbf{Y}_1)^{-1/2}\mathbf{Y}_1^\top(\mathbf{x}-\mathbf{x}_m) = \mathbf{B}\mathbf{W}_1^\top(\mathbf{x}-\mathbf{x}_0). \tag{24}$$

Lastly we compute $\mathbf{W}_2^\top(\mathbf{x}-\mathbf{x}_m)$. This follows a similar argument to the above. We have

$$\mathbf{W}_2^\top(\mathbf{x}-\mathbf{x}_m) = \mathbf{W}_2^\top(\mathbf{x}-\mathbf{f}) \tag{25}$$

since $\mathbf{W}_2^\top\mathbf{G} = \mathbf{0}$. Similarly, left-multiplying the fixed-point equation Eq. (5) by $\mathbf{W}_2^\top$ gives

$$\mathbf{W}_2^\top\mathbf{x} = \mathbf{W}_2^\top\mathbf{f}$$

Substituting this into Eq. (25) gives

$$\mathbf{W}_2^\top(\mathbf{x}-\mathbf{x}_m) = \mathbf{0}. \tag{26}$$

Step 3 Eq. (26) validates the second requirement of Definition 9, since $\mathbf{N} = \mathbf{W}_2$. It remains to establish the first requirement. To accomplish this replace $\mathbf{x}$ with $\boldsymbol{X} \sim \mathcal{N}(\mathbf{x}_0, \boldsymbol{\Sigma}_0)$ in Eq. (24). Since $\mathbf{W}_1^\top\boldsymbol{X} \sim \mathcal{N}(\mathbf{W}_1^\top\mathbf{x}_0, \mathbf{W}_1^\top\boldsymbol{\Sigma}_0\mathbf{W}_1^\top)$, it follows that

$$\mathbf{B}\mathbf{W}_1^\top(\boldsymbol{X}-\mathbf{x}_0) = (\mathbf{W}_1^\top\boldsymbol{\Sigma}_0\mathbf{W}_1^\top)^{-\frac{1}{2}}\mathbf{W}_1^\top(\boldsymbol{X}-\mathbf{x}_0) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r).$$

which verifies the first requirement and completes the proof. ■

References

M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum mean discrepancy gradient flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6484–6494. Curran Associates, Inc., 2019.

S. Bartels and P. Hennig. Probabilistic approximate least-squares. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 676–684, May 2016.

S. Bartels, J. Cockayne, I. C. F. Ipsen, and P. Hennig. Probabilistic linear solvers: a unifying view. *Stat. Comput.*, 29(6):1249–1263, 2019.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *J. R. Stat. Soc. B*, 78(5):1103–1130, 2016.

M. T. Chu. Linear algebra algorithms as dynamical systems. *Acta Numer.*, 17:1–86, 2008. ISSN 0962-4929. doi: 10.1017/S0962492906340019.

J. Cockayne, C. J. Oates, I. C. F. Ipsen, and M. Girolami. A Bayesian conjugate gradient method (with discussion). *Bayesian Anal.*, 14(3):937–1012, 2019a.

J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami. Bayesian probabilistic numerical methods. *SIAM Rev.*, 61(4):756–789, 2019b.

J. Cockayne, M. Graham, C. Oates, and T. Sullivan. Testing whether a learning procedure is calibrated. *arXiv preprint arXiv:2012.12670*, 2020.

S. R. Cook, A. Gelman, and D. B. Rubin. Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Stat.*, 15(3):675–692, 2006.

A. P. Dawid. The well-calibrated Bayesian. *J. Amer. Statist. Assoc.*, 77(379):605–610, 1982. doi: 10.1080/01621459.1982.10477856.

P. Diaconis. Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, 1:163–175, 1988.

J. Geweke. Getting it right: Joint distribution tests of posterior simulators. *J. Am. Stat. Assoc.*, 99(467):799–804, 2004.

G. H. Golub and G. Meurant. Matrices, moments and quadrature. In *Numerical analysis 1993 (Dundee, 1993)*, volume 303 of *Pitman Res. Notes Math. Ser.*, pages 105–156. Longman Sci. Tech., Harlow, 1994.

G. H. Golub and G. Meurant. Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods. *BIT*, 37(3):687–705, 1997. Direct methods, linear algebra in optimization, iterative methods (Toulouse, 1995/1996).

G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, fourth edition, 2013.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(25):723–773, 2012.

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

P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *J. R. Stat. Soc. A Stat.*, 471(2179):20150142, 17, 2015. ISSN 1364-5021. doi: 10.1098/rspa.2015.0142.

M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.*, 49(6), December 1952.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2009. doi: 10.1017/cbo9781139020411.

I. C. F. Ipsen. *Numerical Matrix Analysis*. Society for Industrial and Applied Mathematics, Jan. 2009. doi: 10.1137/1.9780898717686.

O. Kallenberg. *Foundations of Modern Probability*. Springer New York, 2002. doi: 10.1007/978-1-4757-4015-8.

F. Larkin. Gaussian measure in Hilbert space and applications in numerical analysis. *Rocky Mt. J. Math.*, pages 379–421, 1972.

J. Liesen and Z. Strakos. *Krylov Subspace Methods*. Oxford University Press, Oct. 2012. doi: 10.1093/acprof:oso/9780199655410.001.0001.

C. Liu, J. Zhuo, P. Cheng, R. Zhang, and J. Zhu. Understanding and accelerating particle-based variational inference. volume 97 of *Proceedings of Machine Learning Research*, pages 4082–4092. PMLR, 2019.

G. Meurant. The computation of bounds for the norm of the error in the conjugate gradient algorithm. *Numer. Algorithms*, 16(1):77–87 (1998), 1997. Sparse matrices in industry (Lille, 1997).

G. Meurant and P. Tichý. On computing quadrature-based bounds for the A-norm of the error in conjugate gradients. *Numer. Algorithms*, 62(2):163–191, 2013.

G. Meurant and P. Tichý. Approximating the extreme Ritz values and upper bounds for the A-norm of the error in CG. *Numer. Algorithms*, 82(3):937–968, 2019.

J. F. Monahan and D. D. Boos. Proper likelihoods for Bayesian analysis. *Biometrika*, 79 (2):271–278, 1992.

C. J. Oates and T. J. Sullivan. A modern retrospective on probabilistic numerics. *Stat. Comput.*, 29(6):1335–1351, 2019.

J. K. Reid. On the method of conjugate gradients for the solution of large sparse systems of linear equations. *Large Sparse Sets of Linear Equations (Proc. Conf. St. Catherine's Coll., Oxford, 1970)*, pages 231–254, 1971.

T. W. Reid, I. C. F. Ipsen, J. Cockayne, and C. J. Oates. BayesCG as an uncertainty aware version of CG. *arXiv preprint arXiv:2008.03225*, 2020.

Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, jan 2003. doi: 10.1137/1.9780898718003.

R. C. Smith. *Uncertainty Quantification*, volume 12 of *Computational Science & Engineering*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2014.

Z. Strakoš and P. Tichý. On error estimation in the conjugate gradient method and why it works in finite precision computations. *Electron. Trans. Numer. Anal.*, 13:56–80, 2002.

Z. Strakoš and P. Tichý. Error estimation in preconditioned conjugate gradients. *BIT*, 45 (4):789–817, 2005.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Y. L. Tong. *The Multivariate Normal Distribution*. Springer New York, 1990. doi: 10.1007/978-1-4613-9655-0.

J. von Neumann and H. H. Goldstine. Numerical inverting of matrices of high order. *Bulletin of the American Mathematical Society*, 53(11):1021–1100, Nov. 1947. doi: 10.1090/s0002-9904-1947-08909-6. URL `https://doi.org/10.1090/s0002-9904-1947-08909-6`.

J. Wenger and P. Hennig. Probabilistic linear solvers for machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

D. M. Young. *Iterative Solution of Large Linear Systems*. Elsevier, 1971. doi: 10.1016/c2013-0-11733-3.

D. M. Young. Second-degree iterative methods for the solution of large linear systems. *J. Approx. Theory*, 5:137–148, 1972.