Supplementary Material for "A Bayesian Conjugate-Gradient Method"

Jon Cockayne[*], Chris J. Oates[†], Ilse C.F. Ipsen[‡] and Mark Girolami[§]

October 11, 2018

S1 Proof of Theoretical Results

Proof of Proposition 1. Note that the joint distribution of $\boldsymbol{x}$ and $\boldsymbol{y}_m$ is given by

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y}_m \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_0 \\ S_m^\top A \boldsymbol{x}_0 \end{bmatrix}, \begin{bmatrix} \Sigma_0 & \Sigma_0 A^\top S_m \\ S_m^\top A \Sigma_0 & S_m^\top A \Sigma_0 A^\top S_m \end{bmatrix} \right)$$

from which the stated conditional distribution is deduced. □

Proof of Proposition 2. Consider an arbitrary vector $\boldsymbol{\ell} \in \mathbb{R}^d$. Now

$$\begin{aligned} \boldsymbol{\ell}^\top \boldsymbol{x}_m - \boldsymbol{\ell}^\top \boldsymbol{x}^* &= \boldsymbol{\ell}^\top (\boldsymbol{x}_0 - \boldsymbol{x}^*) + \boldsymbol{\ell}^\top \Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top A (\boldsymbol{x}^* - \boldsymbol{x}_0) && \text{(from Eq. 5)} \\ &= \boldsymbol{\ell}^\top (\Sigma_0 - \Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top A \Sigma_0) \Sigma_0^{-1} (\boldsymbol{x}_0 - \boldsymbol{x}^*) \\ &= \langle \Sigma_m \boldsymbol{\ell}, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle_{\Sigma_0^{-1}} && \text{(from Eq. 6)} \end{aligned}$$

and so:

$$\begin{aligned} |\boldsymbol{\ell}^\top \boldsymbol{x}_m - \boldsymbol{\ell}^\top \boldsymbol{x}^*| &= \left| \langle \Sigma_m \boldsymbol{\ell}, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle_{\Sigma_0^{-1}} \right| \\ &\leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\Sigma_0^{-1}} \underbrace{\|\Sigma_m \boldsymbol{\ell}\|_{\Sigma_0^{-1}}}_{(*)} . \end{aligned} \tag{S1}$$

[*]University of Warwick (j.cockayne@warwick.ac.uk).
[†]Newcastle University and the Alan Turing Institute (chris.oates@ncl.ac.uk).
[‡]North Carolina State University (ipsen@ncsu.edu)
[§]Imperial College London and the Alan Turing Institute (m.girolami@imperial.ac.uk).

where the last line follow from Cauchy–Schwarz. Now, by expanding the term $(*)$ and simplifying, we see that

$$\begin{aligned}
\|\Sigma_m \boldsymbol{\ell}\|_{\Sigma_0^{-1}}^2 &= \boldsymbol{\ell}^\top (\Sigma_0 - \Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top A \Sigma_0)^\top \Sigma_0^{-1} (\Sigma_0 - \Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top A \Sigma_0) \boldsymbol{\ell} \\
&= \boldsymbol{\ell}^\top \big(\Sigma_0 - 2\Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top A \Sigma_0 \\
&\qquad + \Sigma_0 A^\top S_m \Lambda_m^{-1} \underbrace{S_m^\top A \Sigma_0 A^\top S_m}_{=\Lambda_m} \Lambda_m^{-1} S_m^\top A \Sigma_0 \big) \boldsymbol{\ell} \\
&= \boldsymbol{\ell}^\top (\Sigma_0 - \Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top A \Sigma_0) \boldsymbol{\ell} \qquad \text{(S2)} \\
&= \boldsymbol{\ell}^\top \Sigma_m \boldsymbol{\ell}
\end{aligned}$$

which follows from Eq. 6

Finally let $\boldsymbol{e}_i$ denote the vector whose j^{th} entry is δ_{ij} and note that

$$\begin{aligned}
\|\boldsymbol{x}_m - \boldsymbol{x}^*\|_{\Sigma_0^{-1}} &= \|\Sigma_0^{-\frac{1}{2}} (\boldsymbol{x}_m - \boldsymbol{x}^*)\|_2 \\
&= \left(\sum_{i=1}^d \left| \boldsymbol{e}_i^\top \Sigma_0^{-\frac{1}{2}} \boldsymbol{x}_m - \boldsymbol{e}_i^\top \Sigma_0^{-\frac{1}{2}} \boldsymbol{x}^* \right|^2 \right)^{\frac{1}{2}} \\
&\leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\Sigma_0^{-1}} \left(\sum_{i=1}^d \boldsymbol{e}_i^\top \Sigma_0^{-\frac{1}{2}} \Sigma_m \Sigma_0^{-\frac{1}{2}} \boldsymbol{e}_i \right)^{\frac{1}{2}} \qquad \text{(from Eq. S1, S2)} \\
&= \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\Sigma_0^{-1}} \sqrt{\operatorname{tr}\left(\Sigma_0^{-\frac{1}{2}} \Sigma_m \Sigma_0^{-\frac{1}{2}} \right)} \\
&= \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\Sigma_0^{-1}} \sqrt{\operatorname{tr}(\Sigma_m \Sigma_0^{-1})}
\end{aligned}$$

where the last line uses the fact that the trace is invariant under cyclic permutation of the argument. □

Proof of Proposition 3. Note that

$$\begin{aligned}
\operatorname{tr}(\Sigma_m \Sigma_0^{-1}) &= \operatorname{tr}(I - \Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top A) \\
&= \operatorname{tr}(I) - \operatorname{tr}(\Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top A) \\
&= \operatorname{tr}(I) - \operatorname{tr}(\underbrace{S_m^\top A \Sigma_0 A^\top S_m}_{=\Lambda_m} \Lambda_m^{-1}) \\
&= d - m
\end{aligned}$$

where the third line uses the fact that the trace is invariant under cyclic permutation of the argument. □

Proof of Proposition 4. First, note that

$$\Lambda_m = (S_m^{\text{CG}})^\top A \Sigma_0 A^\top S_m^{\text{CG}} = (S_m^{\text{CG}})^\top A S_m^{\text{CG}} = I$$

since the columns of S_m^{CG} are A-orthonormal. Then, from Proposition 1 we have

$$\begin{aligned}\boldsymbol{x}_m &= \boldsymbol{x}_0 + \Sigma_0 A^\top S_m^{\mathrm{CG}} \Lambda_m^{-1} (S_m^{\mathrm{CG}})^\top \boldsymbol{r}_0 \\ &= S_m^{\mathrm{CG}} (S_m^{\mathrm{CG}})^\top \boldsymbol{r}_0 \\ &\equiv \boldsymbol{x}_m^{\mathrm{CG}}\end{aligned}$$

as required.

□

Proof of Proposition 5. We first introduce the concept of an average-case optimal algorithm and average-case optimal information. The *information space* B and the *solution space* X are, informally, the spaces in which the right-hand-side and the solution of the system live, respectively. We wish to computationally approximate an intractable *solution operator* $\mathcal{A}(b)$, based upon a finite amount of information provided by the *information operator* $S_m : B \to \mathbb{R}^m$. This is accomplished by an *algorithm* $\psi(S_m(b))$, which we hope approximates $\mathcal{A}(b)$ well in a way which will now be made formal.

For a reference measure ν on B, denote the *average-case error* of an algorithm ψ with information S_m as

$$e_M^{\mathrm{avg}}(S_m, \psi) := \left[\int_{\mathbb{R}^d} \|\mathcal{A}(b) - \psi(S_m(b))\|_M^2 \, \mu(\mathrm{d}\boldsymbol{x}) \right]^{\frac{1}{2}}.$$

An algorithm ψ^* which minimises $e^{\mathrm{avg}}(\cdot, \psi)$ for arbitrary S_m is said to be *average-case-optimal*. An S_m^* which minimises $e^{\mathrm{avg}}(S_m, \psi^*)$ is said to be *average-case optimal information*.

By Theorem 3.3 of Cockayne et al. [2017], in the present setting optimal information for the average risk in Eq. (10) is identical to average-case optimal information. This is by virtue of the fact that, for any symmetric positive-definite M, $(\mathbb{R}^d, \langle \cdot, \cdot \rangle_M)$ forms an inner-product space.

Now recall two relevant theorems from Novak and Woźniakowski [2008]. For measurable spaces $(B, \mathcal{F}_B)$ and $(X, \mathcal{F}_B)$, an operator $\mathcal{A} : B \to X$ and a measure μ on B, let $\mathcal{A}_{\#}\mu$ denote the *pushforward* of μ through $\mathcal{A}$, a measure on X defined as

$$[\mathcal{A}_{\#}\mu](C) = \mu(\mathcal{A}^{-1}(C))$$

for each $C \in \mathcal{F}_X$.

Theorem S1 (Theorem 4.28 of Novak and Woźniakowski [2008])**.** *Let B be a separable real Banach space equipped with a zero-mean Gaussian measure ν with covariance operator C_ν. Let the solution operator $\mathcal{A} : B \to X$ be a bounded linear operator into a separable real Hilbert space X with inner product $\langle \cdot, \cdot \rangle_X$. Let $\eta = \mathcal{A}_{\#}\nu$ be a Gaussian measure on solution elements. Consider linear information $S_m = [s_1, \ldots, s_m]$ where $s_i : B \to \mathbb{R}$ and $s_i(C_\nu s_j) = \delta_{ij}$, and consider information $y_i = s_i(b)$. Then the algorithm*

$$\psi(b) = \sum_{i=1}^m y_i \mathcal{A}(C_\nu s_i)$$

is average-case optimal.

Denote by C_η the covariance operator of η, and let $\{(\gamma_i^*, \phi_i^*) : i \in I\}$ for $I \subseteq \mathbb{N}$ denote its eigensystem, ordered so that $\gamma_1^* \geq \gamma_2^* \geq \ldots$. Note that if X is finite-dimensional with dimension d then $I = \{1, \ldots, d\}$, while otherwise $I = \mathbb{N}$.

Theorem S2 (Theorem 4.30 of Novak and Woźniakowski [2008])**.** *Under the assumptions of Theorem S1, for $b \in B$ the optimal information S_m^* is given by*

$$S_m^*(b) = [L_1^*(b), \ldots, L_m^*(b)]$$

where

$$L_i^*(b) := \frac{\langle \mathcal{A}(b), \phi_i^* \rangle_X}{(\gamma_i^*)^{\frac{1}{2}}}.$$

We will first establish that the posterior mean from Proposition 1 represents an average-case optimal algorithm, by applying Theorem S1. In the notation of that theorem, $B = X = \mathbb{R}^d$, which satisfies the required assumptions as $\mathbb{R}^d$ is separable. The measure ν is given by $\nu = A_\# \mu \sim \mathcal{N}(\mathbf{0}, A\Sigma_0 A^\top)$, so that $C_\nu = A\Sigma_0 A^\top$. Furthermore the information operator S_m is simply a matrix in $\mathbb{R}^{d \times m}$, which is subject to the restriction from Theorem S1 that $\Lambda_m = S_m^\top A \Sigma_0 A^\top S_m = I$. Note that this is markedly similar to the conjugacy requirement in Section 2.2.

Now we seek the optimal algorithm $\psi(\boldsymbol{b})$ which minimises

$$\begin{aligned}\int_{\mathbb{R}^d} \|A^{-1}\boldsymbol{b} - \psi(S_m^\top \boldsymbol{b})\|_M^2 \, \nu(\mathrm{d}\boldsymbol{b}) &= \int_{\mathbb{R}^d} \|M^{\frac{1}{2}} A^{-1}\boldsymbol{b} - M^{\frac{1}{2}} \psi(S_m^\top \boldsymbol{b})\|_2^2 \, \nu(\mathrm{d}\boldsymbol{b}) \\ &= \int_{\mathbb{R}^d} \|M^{\frac{1}{2}} A^{-1}\boldsymbol{b} - \bar{\psi}(S_m^\top \boldsymbol{b})\|_2^2 \, \nu(\mathrm{d}\boldsymbol{b}) \qquad \text{(S3)}\end{aligned}$$

where $\bar{\psi} = M^{\frac{1}{2}} \psi$. Eq. (S3) is of the form required by Theorem S1, with the solution operator $\mathcal{A} = M^{\frac{1}{2}} A^{-1}$, which is a bounded linear operator as required. For any S_m conjugate to $A\Sigma_0 A^\top$, the optimal algorithm is therefore given by

$$\begin{aligned}\bar{\psi}(\boldsymbol{b}) &= \sum_{i=1}^m (\boldsymbol{s}_i^\top \boldsymbol{b}) M^{\frac{1}{2}} A^{-1} A \Sigma_0 A^\top \boldsymbol{s}_i \\ &= M^{\frac{1}{2}} \Sigma_0 A^\top S_m S_m^\top \boldsymbol{b} \\ \implies \psi(\boldsymbol{b}) &= \Sigma_0 A^\top S_m S_m^\top \boldsymbol{b}.\end{aligned}$$

In this conjugate setting with $\boldsymbol{x}_0 = \mathbf{0}$, this is identical to the expression for $\boldsymbol{x}_m$ in Proposition 1.

Theorem S2 can now be applied to determine the optimal information S_m^*. Note that since A is a bijection, $\eta = [M^{\frac{1}{2}} A^{-1}]_\# [A_\# \mu] = M^{\frac{1}{2}}_\# \mu$, so the required eigensystem is that of $M^{\frac{1}{2}} \Sigma_0 M^{\frac{\top}{2}}$. As before, denote this (ordered) eigensystem by $\{(\gamma_i^*, \boldsymbol{\phi}_i^*)\}_{i=1}^d$, with the eigenvectors normalised so that $(\boldsymbol{\phi}_i^*)^\top \boldsymbol{\phi}_i^* = 1$. It then holds from Theorem S2 that the optimal search directions are given by

$$\boldsymbol{s}_i^* = (\gamma_i^*)^{-\frac{1}{2}} A^{-\top} M^{\frac{\top}{2}} \boldsymbol{\phi}_i^*.$$

Lastly, noting that the scaling by $(\gamma_i^*)^{-\frac{1}{2}}$ does not affect the output yields the result that the optimal information is given by

$$S_m = A^{-\top} M^{\frac{\top}{2}} \Phi_m.$$

□

Proof of Proposition 6. First, note that $\Lambda_m = I$ as the search directions $\{\boldsymbol{s}_i\}$, $i = 1, \ldots, m$ are Q-orthonormal, where $Q = A\Sigma_0 A^\top$. Then, from Eq. 5:

$$\begin{aligned}
\boldsymbol{x}_m &= \boldsymbol{x}_0 + \Sigma_0 A^\top S_m S_m^\top \boldsymbol{r}_0 \\
&= \boldsymbol{x}_0 + \Sigma_0 A^\top \begin{bmatrix} S_{m-1} & \boldsymbol{s}_m \end{bmatrix} \begin{bmatrix} S_{m-1}^\top \\ \boldsymbol{s}_m^\top \end{bmatrix} \boldsymbol{r}_0 \\
&= \boldsymbol{x}_0 + \underbrace{\Sigma_0 A^\top S_{m-1} S_{m-1}^\top \boldsymbol{r}_0}_{=\boldsymbol{x}_{m-1}} + \Sigma_0 A^\top \boldsymbol{s}_m \boldsymbol{s}_m^\top \boldsymbol{r}_0.
\end{aligned}$$

It therefore remains to show that $\boldsymbol{s}_m^\top \boldsymbol{r}_0 = \boldsymbol{s}_m^\top \boldsymbol{r}_{m-1}$. To this end, from Eq. 5 we have

$$\begin{aligned}
\boldsymbol{s}_m^\top \boldsymbol{r}_{m-1} &= \boldsymbol{s}_m^\top \boldsymbol{b} - \boldsymbol{s}_m^\top A \boldsymbol{x}_{m-1} \\
&= \boldsymbol{s}_m^\top \boldsymbol{b} - \boldsymbol{s}_m^\top \boldsymbol{x}_0 - \underbrace{\boldsymbol{s}_m^\top A \Sigma_0 A^\top S_{m-1}^\top}_{=0} \boldsymbol{r}_0 \\
&= \boldsymbol{s}_m^\top \boldsymbol{r}_0
\end{aligned}$$

which completes the proof. □

Lemma S3. *Assume that the search directions $\{\boldsymbol{s}_i\}$ are $A\Sigma_0 A^\top$-orthogonal. At iteration m, the residual $\boldsymbol{r}_m = \boldsymbol{b} - A\boldsymbol{x}_m$ satisfies $\boldsymbol{r}_m^\top \boldsymbol{s}_i = 0$ for $i = 1, \ldots, m$.*

Proof of Lemma S3. By definition of $\boldsymbol{r}_m$ and $\boldsymbol{x}_m$

$$\begin{aligned}
\boldsymbol{s}_i^\top \boldsymbol{r}_m &= \boldsymbol{s}_i^\top \boldsymbol{b} - \boldsymbol{s}_i^\top A \boldsymbol{x}_m \\
&= \boldsymbol{s}_i^\top \boldsymbol{b} - \boldsymbol{s}_i^\top A \boldsymbol{x}_0 - \boldsymbol{s}_i^\top A \Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top \boldsymbol{r}_0
\end{aligned}$$

Note that $\boldsymbol{s}_i^\top A \Sigma_0 A^\top S_m \Lambda_m^{-1} = \boldsymbol{e}_i^\top$, the vector with $[\boldsymbol{e}_i]_j = \delta_{ij}$, since $\boldsymbol{s}_i^\top A \Sigma_0 A^\top S_m$ is the i^{th} row of Λ_m, whenever $i \leq m$. Thus, $\boldsymbol{s}_i^\top \boldsymbol{r}_m = \boldsymbol{s}_i^\top \boldsymbol{r}_0 - \boldsymbol{e}_i^\top S_m^\top \boldsymbol{r}_0 = 0$, as required. □

Proof of Proposition 7. Let $\tilde{\boldsymbol{t}}_1 := \boldsymbol{r}_0$, and for each $m > 1$, define $\tilde{\boldsymbol{t}}_m$ as

$$\tilde{\boldsymbol{t}}_m := \boldsymbol{r}_{m-1} - \sum_{i=1}^{m-1} \left(\boldsymbol{r}_{m-1}^\top Q \boldsymbol{t}_i \right) \boldsymbol{t}_i. \tag{S4}$$

where $Q = A\Sigma_0 A^\top$. Let $\boldsymbol{t}_m = \tilde{\boldsymbol{t}}_m / \|\tilde{\boldsymbol{t}}_m\|_Q$. We will show, inductively, that for each m the set of search directions $\{\boldsymbol{t}_i\}_{i=1}^m$ is Q-orthonormal, and further that each $\boldsymbol{t}_i = \boldsymbol{s}_i$, as defined in the proposition statement.

For $m = 1$ the set $\{\boldsymbol{t}_1\}$ is trivially Q-orthonormal and $\boldsymbol{t}_1 = \boldsymbol{s}_1$. For $m > 1$ suppose $\{\boldsymbol{t}_i\}_{i=1}^{m-1}$ is Q-orthonormal and such that $\boldsymbol{t}_i = \boldsymbol{s}_i$, for $i = 1, \ldots, m-1$. Then, for $j < m$

$$\begin{aligned}\boldsymbol{t}_j^\top Q\tilde{\boldsymbol{t}}_m &= \boldsymbol{t}_j^\top Q\boldsymbol{r}_{m-1} - \sum_{i=1}^{m-1} \boldsymbol{r}_{m-1}^\top Q\boldsymbol{t}_i \cdot \underbrace{\boldsymbol{t}_j^\top Q\boldsymbol{t}_i}_{=\delta_{ij}} \qquad \text{(by the inductive assumption)} \\ &= \boldsymbol{t}_j^\top Q\boldsymbol{r}_{m-1} - \boldsymbol{t}_j^\top Q\boldsymbol{r}_{m-1} = 0 \end{aligned} \tag{S5}$$

which shows that the set $\{\boldsymbol{t}_i\}_{i=1}^{m}$ is Q-orthonormal.

As a result we can apply Proposition 6 to show that

$$\begin{aligned}\boldsymbol{r}_j &= \boldsymbol{b} - A\boldsymbol{x}_j \\ &= \boldsymbol{b} - A\boldsymbol{x}_{j-1} - Q\boldsymbol{t}_j(\boldsymbol{t}_j^\top \boldsymbol{r}_{j-1}) \\ \implies Q\boldsymbol{t}_j &= \frac{\boldsymbol{r}_{j-1} - \boldsymbol{r}_j}{\boldsymbol{t}_j^\top \boldsymbol{r}_{j-1}} \\ \implies \boldsymbol{r}_{m-1}^\top Q\boldsymbol{t}_j &= \frac{\boldsymbol{r}_{m-1}^\top \boldsymbol{r}_{j-1} - \boldsymbol{r}_{m-1}^\top \boldsymbol{r}_j}{\boldsymbol{t}_j^\top \boldsymbol{r}_{j-1}}. \end{aligned} \tag{S6}$$

Since the set $\{\boldsymbol{t}_i\}_{i=1}^{m}$ is Q-orthonormal, we have from Lemma S3 that for each $j \leq m$, $\boldsymbol{r}_m^\top \boldsymbol{t}_j = 0$. Thus, from Eq. (S5) for each $j \leq m$:

$$0 = \boldsymbol{r}_m^\top \tilde{\boldsymbol{t}}_j := \boldsymbol{r}_m^\top \boldsymbol{r}_{j-1} - \sum_{i=1}^{m-1} \boldsymbol{r}_{m-1}^\top Q\boldsymbol{t}_i \cdot \underbrace{\boldsymbol{r}_m^\top \boldsymbol{t}_i}_{=0}. \tag{S7}$$

from which we conclude that $\boldsymbol{r}_m^\top \boldsymbol{r}_j = 0$ whenever $j < m$. It follows that Eq. (S6) is zero for all $j < m - 1$. Thus, all terms in the summation in Eq. (S4) vanish apart from the last, and we are left with

$$\tilde{\boldsymbol{t}}_m = \boldsymbol{r}_{m-1} - (\boldsymbol{r}_{m-1}^\top Q\boldsymbol{t}_{m-1})\boldsymbol{t}_{m-1}$$

which is equal to $\tilde{\boldsymbol{s}}_m$ for each $m > 1$, completing the proof. □

Proof of Proposition 11. First the posterior marginal for ν is computed. Note that

$$p(\nu|\boldsymbol{y}) \propto p(\boldsymbol{y}|\nu)p(\nu)$$

where

$$\begin{aligned}\boldsymbol{y}|\nu &\sim \mathcal{N}(S_m^\top A\boldsymbol{x}_0, \nu\Lambda_m) \\ \implies p(\nu|\boldsymbol{y}) &\propto \nu^{-\frac{m}{2}-1} \exp\left(-\frac{1}{2\nu}\boldsymbol{r}_0^\top S_m \Lambda_m^{-1} S_m^\top \boldsymbol{r}_0\right)\end{aligned}$$

which is $\mathsf{IG}\left(\frac{m}{2}, \frac{1}{2}\boldsymbol{r}_0^\top S_m \Lambda_m^{-1} S_m^\top \boldsymbol{r}_0\right)$. Now to determine the posterior marginal for $\boldsymbol{x}$

$$\begin{aligned}p(\boldsymbol{x}|\boldsymbol{y}) &= \int_0^\infty p(\boldsymbol{x}|\nu, \boldsymbol{y})p(\nu|\boldsymbol{y})\, \mathrm{d}\nu \\ &\propto \int_0^\infty \nu^{-1-(m+d)/2} \exp\left(-\nu^{-1}K(\boldsymbol{x})\right)\, \mathrm{d}\nu \end{aligned} \tag{S8}$$

where

$$K(\boldsymbol{x}) := \frac{1}{2}\left[\boldsymbol{r}_0^\top S_m \Lambda_m^{-1} S_m^\top \boldsymbol{r}_0 + (\boldsymbol{x}-\boldsymbol{x}_m)^\top \Sigma_m^{-1}(\boldsymbol{x}-\boldsymbol{x}_m)\right]$$

Eq. S8 is recognised as the integral of an unnormalised inverse-Gamma density, so that

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) &\propto \Gamma(m+d)K(\boldsymbol{x})^{-\frac{1}{2}(m+d)} \\ &\propto \left[1 + \frac{1}{m}(\boldsymbol{x}-\boldsymbol{x}_m)^\top \left\{\frac{\boldsymbol{r}_0^\top S_m \Lambda_m^{-1} S_m^\top \boldsymbol{r}_0}{m}\Sigma_m\right\}^{-1}(\boldsymbol{x}-\boldsymbol{x}_m)\right]^{-\frac{1}{2}(m+d)} \end{aligned}$$

and therefore

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathsf{MVT}_m\left(\boldsymbol{x}_m, \frac{\boldsymbol{r}_0^\top S_m \Lambda_m^{-1} S_m^\top \boldsymbol{r}_0}{m}\Sigma_m\right)$$

□

Proposition S4. *It holds that* $\boldsymbol{x}_m \in \boldsymbol{x}_0 + K_{m-1}(\Sigma_0 A^\top A, \Sigma_0 A^\top \boldsymbol{r}_0)$.

Proof of Proposition S4. Let $\bar{K}_m = K_m(\Sigma_0 A^\top A, \Sigma_0 A^\top \boldsymbol{r}_0)$. Proof is by induction, with the additional inductive claims that

$$\Sigma_0 A^\top \boldsymbol{s}_m \in \bar{K}_{m-1} \tag{S9}$$

$$\Sigma_0 A^\top \boldsymbol{r}_m \in \bar{K}_m. \tag{S10}$$

Note that Eq. (S9) implies the required result by Proposition 1. Let $Q = A\Sigma_0 A^\top$. For $m = 1$, the first search direction is given by

$$\boldsymbol{s}_1 = \frac{\boldsymbol{r}_0}{\|\boldsymbol{r}_0\|_Q}$$

from which Eq. (S9) is clear. Further,

$$\begin{aligned} \boldsymbol{r}_1 &= \boldsymbol{b} - A\boldsymbol{x}_1 \\ &= \boldsymbol{b} - A\boldsymbol{x}_0 - \frac{A\Sigma_0 A^\top \boldsymbol{r}_0(\boldsymbol{r}_0^\top \boldsymbol{r}_0)}{\|\boldsymbol{r}_0\|_Q^2} \\ &= \boldsymbol{r}_0 - \frac{A\Sigma_0 A^\top \boldsymbol{r}_0(\boldsymbol{r}_0^\top \boldsymbol{r}_0)}{\|\boldsymbol{r}_0\|_Q^2} \\ \implies \Sigma_0 A^\top \boldsymbol{r}_1 &= \Sigma_0 A^\top \boldsymbol{r}_0 - \frac{(\Sigma_0 A^\top A)\Sigma_0 A^\top \boldsymbol{r}_0(\boldsymbol{r}_0^\top \boldsymbol{r}_0)}{\|\boldsymbol{r}_0\|_Q^2} \end{aligned}$$

from which it is clear that $\Sigma_0 A^\top \boldsymbol{r}_1 \in \bar{K}_1$.

Now for the inductive step. Assume that Equations (S9) and (S10) hold true up to $m-1$. From Proposition 7 we have that

$$\begin{aligned} \tilde{\boldsymbol{s}}_m &= \boldsymbol{r}_{m-1} - (\boldsymbol{r}_{m-1}^\top Q \boldsymbol{s}_{m-1})\boldsymbol{s}_{m-1} \\ \implies \Sigma_0 A^\top \tilde{\boldsymbol{s}}_m &= \underbrace{\Sigma_0 A^\top \boldsymbol{r}_{m-1}}_{\in \bar{K}_{m-1}} - (\boldsymbol{r}_{m-1}^\top Q \boldsymbol{s}_{m-1}) \underbrace{\Sigma_0 A^\top \boldsymbol{s}_{m-1}}_{\in \bar{K}_{m-2}} \end{aligned}$$

where inclusion in the Krylov subspaces is by the inductive assumption. It follows that $\Sigma_0 A^\top \boldsymbol{s}_m \in \bar{K}_{m-1}$. Lastly, observe that

$$\begin{aligned}
\boldsymbol{r}_m &= \boldsymbol{b} - A\boldsymbol{x}_m \\
&= \boldsymbol{r}_{m-1} - A\Sigma_0 A^\top \boldsymbol{s}_m (\boldsymbol{s}_m^\top \boldsymbol{r}_m) \\
\implies \Sigma_0 A^\top \boldsymbol{r}_m &= \underbrace{\Sigma_0 A^\top \boldsymbol{r}_{m-1}}_{\in \bar{K}_{m-1}} - \underbrace{(\Sigma_0 A^\top A)\Sigma_0 A^\top \boldsymbol{s}_m}_{\in \bar{K}_m} (\boldsymbol{s}_m^\top \boldsymbol{r}_m)
\end{aligned}$$

which by the inductive assumption is in $\bar{K}_m$, as required. □

Proof of Proposition 9. Let $Q = A\Sigma_0 A^\top$. Begin with $m = 1$. Any $\boldsymbol{x} \in K_0^*$ can be represented as $\boldsymbol{x}_0 + \alpha_1 \Sigma_0 A^\top \boldsymbol{r}_0$ for some α_1. Thus, when $\boldsymbol{x} \in K_1^*$:

$$\begin{aligned}
\|\boldsymbol{x} - \boldsymbol{x}^*\|_{\Sigma_0^{-1}}^2 &= \|\boldsymbol{x}_0 + \alpha_1 \Sigma_0 A^\top \boldsymbol{r}_0 - \boldsymbol{x}^*\|_{\Sigma_0^{-1}}^2 \\
&= \boldsymbol{x}_0^\top \Sigma_0^{-1} \boldsymbol{x}_0 + 2\alpha_1 \boldsymbol{x}_0^\top A^\top \boldsymbol{r}_0 - 2\boldsymbol{x}_0^\top \Sigma_0^{-1} \boldsymbol{x}^* \\
&\quad + \alpha_1^2 \boldsymbol{r}_0^\top A\Sigma_0 A^\top \boldsymbol{r}_0 - 2\alpha_1 \boldsymbol{r}_0^\top A\boldsymbol{x}^* \\
&\quad + (\boldsymbol{x}^*)^\top \Sigma_0^{-1} \boldsymbol{x}^* \\
\implies \frac{\mathrm{d}}{\mathrm{d}\alpha_1} \|\boldsymbol{x} - \boldsymbol{x}^*\|_{\Sigma_0^{-1}}^2 &= 2\boldsymbol{x}_0^\top A^\top \boldsymbol{r}_0 + 2\alpha_1 \boldsymbol{r}_0^\top A\Sigma_0 A^\top \boldsymbol{r}_0 - 2\boldsymbol{r}_0^\top A\boldsymbol{x}^*.
\end{aligned}$$

Setting this to zero, we obtain:

$$\alpha_1 = \frac{\boldsymbol{r}_0^\top (\boldsymbol{b} - A\boldsymbol{x}_0)}{\|\boldsymbol{r}_0\|_Q^2} = \frac{\boldsymbol{r}_0^\top \boldsymbol{r}_0}{\|\boldsymbol{r}_0\|_Q^2}.$$

From Proposition 6, this corresponds to $\boldsymbol{x} = \boldsymbol{x}_1$. It is further clear that

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha_1^2} \|\boldsymbol{x} - \boldsymbol{x}^*\|_{\Sigma_0^{-1}}^2 = 2\|\boldsymbol{r}_0\|_Q^2 > 0$$

so that $\boldsymbol{x}_1$ is optimal in K_0^*.

Now observe that $\Sigma_0 A^\top \boldsymbol{s}_m$ is orthogonal to $\boldsymbol{x}_{m-1} - \boldsymbol{x}_0$ in the Σ_0^{-1}-inner-product:

$$\begin{aligned}
\left\langle \Sigma_0 A^\top \boldsymbol{s}_m, \boldsymbol{x}_{m-1} - \boldsymbol{x}_0 \right\rangle_{\Sigma_0^{-1}} &= \boldsymbol{s}_m^\top A\boldsymbol{x}_0 + \underbrace{\boldsymbol{s}_m^\top A\Sigma_0 A^\top S_{m-1}}_{=0} (S_{m-1}^\top \boldsymbol{r}_0) - \boldsymbol{s}_m^\top A\boldsymbol{x}_0 \\
&= 0
\end{aligned}$$

As a result, for $m > 1$ it suffices to determine α_m in

$$\begin{aligned}
\boldsymbol{x} &= \boldsymbol{x}_0 + (\boldsymbol{x}_{m-1} - \boldsymbol{x}_0) + \alpha_m \Sigma_0 A^\top \boldsymbol{s}_m \\
&= \boldsymbol{x}_{m-1} + \alpha_m \Sigma_0 A^\top \boldsymbol{s}_m
\end{aligned}$$

where $\boldsymbol{x} \in K^*_{m-1}$. Again, α_m is determined directly, much as above:

$$\begin{aligned}
&\|\boldsymbol{x} - \boldsymbol{x}^*\|^2_{\Sigma_0^{-1}} = \|\boldsymbol{x}_{m-1} + \alpha_m \Sigma_0 A^\top \boldsymbol{s}_m - \boldsymbol{x}^*\|^2_{\Sigma_0^{-1}} \\
\implies &\frac{\mathrm{d}}{\mathrm{d}\alpha_m}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2_{\Sigma_0^{-1}} = 2\boldsymbol{x}_{m-1}^\top A^\top \boldsymbol{s}_m + 2\alpha_m \boldsymbol{s}_m^\top A \Sigma_0 A^\top \boldsymbol{s}_m - 2\boldsymbol{s}_m^\top A \boldsymbol{x}^*. \\
\implies &\alpha_m = \frac{\boldsymbol{s}_m^\top(\boldsymbol{b} - A\boldsymbol{x}_{m-1})}{\|\boldsymbol{s}_m\|^2_Q} = \boldsymbol{s}_m^\top \boldsymbol{r}_{m-1}
\end{aligned}$$

which is also a minimum. Thus, we have that

$$\operatorname*{arg\,min}_{\boldsymbol{x} \in K^*_{m-1}} \|\boldsymbol{x} - \boldsymbol{x}^*\|^2_{\Sigma_0^{-1}} = \boldsymbol{x}_{m-1} + \Sigma_0 A^\top \boldsymbol{s}_m (\boldsymbol{s}_m^\top \boldsymbol{r}_m) \equiv \boldsymbol{x}_m$$

from Proposition 6, which completes the proof. □

Proof of Proposition 10. We begin by introducing the *operator norm* induced by the energy norm $\|\cdot\|_A$, which is a norm on matrices $M \in \mathbb{R}^{d\times d}$

$$\|M\|^{\mathrm{op}}_A = \sup\left\{\|M\boldsymbol{v}\|_A : \|\boldsymbol{v}\|_A = 1\right\}.$$

From Proposition S4 it holds that there exists a polynomial $\tilde{P}_{m-1}$ of degree $m-1$ such that

$$\begin{aligned}
\boldsymbol{e}_m := \boldsymbol{x}_m - \boldsymbol{x}^* &= \boldsymbol{x}_0 - \boldsymbol{x}^* + \tilde{P}_{m-1}(\Sigma_0 A^\top A)\Sigma_0 A^\top \boldsymbol{r}_0 \\
&= \boldsymbol{e}_0 + \tilde{P}_{m-1}(\Sigma_0 A^\top A)\Sigma_0 A^\top A \boldsymbol{e}_0 \\
&= P_m(\Sigma_0 A^\top A)\boldsymbol{e}_0
\end{aligned}$$

where P_m is some polynomial of degree m. Thus

$$\begin{aligned}
\|\boldsymbol{e}_m\|_{\Sigma_0^{-1}} &\leq \|P_m(\Sigma_0 A^\top A)\|^{\mathrm{op}}_{\Sigma_0^{-1}} \cdot \|\boldsymbol{e}_0\|_{\Sigma_0^{-1}} \\
&= \|\Sigma_0^{-\frac{1}{2}} P_m(\Sigma_0 A^\top A)\Sigma_0^{\frac{1}{2}}\|^{\mathrm{op}}_I \cdot \|\boldsymbol{e}_0\|_{\Sigma_0^{-1}} \\
&= \|P_m(\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}})\|^{\mathrm{op}}_I \cdot \|\boldsymbol{e}_0\|_{\Sigma_0^{-1}}
\end{aligned}$$

Now, note that $\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}}$ is symmetric, and can thus be represented as $\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}} = V\Gamma V^\top$, where Γ is the matrix with the eigenvalues of $\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}}$ on its diagonal, and V is the orthonormal matrix of its eigenvectors. Furthermore note that $\Sigma_0 A^\top A = \Sigma_0^{\frac{1}{2}}[\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}}]\Sigma_0^{-\frac{1}{2}}$. Hence, $\Sigma_0 A^\top A$ is similar to $\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}}$, and so the matrices share the same eigenvalues.

Now, clearly $P_m(V\Gamma V^\top) = V P_m(\Gamma) V^\top$ since V is orthonormal. Thus

$$\begin{aligned}
\|\boldsymbol{e}_m\|_I &\leq \underbrace{\|V\|^{\mathrm{op}}_I \|V^\top\|^{\mathrm{op}}_I}_{=1} \|P_m(\Gamma)\|^{\mathrm{op}}_I \cdot \|\boldsymbol{e}_0\|_{\Sigma_0^{-1}} \\
&= \|P_m(\Gamma)\|^{\mathrm{op}}_I \cdot \|\boldsymbol{e}_0\|_{\Sigma_0^{-1}} \qquad \text{(S11)}
\end{aligned}$$

where $\|V\|_I^{\text{op}}\|V^\top\|_I^{\text{op}} = 1$ follows since V is unitary. Let $\mathbb{P}_m$ denote the set of all polynomials of order m with the property that $P(0) = 1$ for each $P \in \mathbb{P}_m$. This requirement ensures that if A is singular, $\|\boldsymbol{e}_m\|_{\Sigma_0^{-1}} = \|\boldsymbol{e}_0\|_{\Sigma_0^{-1}}$ for all m. Now, from Proposition 9 we have that $P_m \in \mathbb{P}_m$ is constructed to minimise the error $\boldsymbol{e}_m$. Let $\bar{\Gamma}$ denote the set of eigenvalues of $\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}}$. Then

$$\begin{aligned}\|P_m(\Gamma)\|_I^{\text{op}} &= \min_{P\in\mathbb{P}_m} \max_{\gamma\in\bar{\Gamma}} \sup_{\|\boldsymbol{v}\|_2=1} \|P(\gamma)\boldsymbol{v}\|_2 \\ &= \min_{P\in\mathbb{P}_m} \max_{\gamma\in\bar{\Gamma}} |P(\gamma)| \\ &\leq \min_{P\in\mathbb{P}_m} \max_{\gamma\in[\gamma_{\min},\gamma_{\max}]} |P(\gamma)| \end{aligned} \tag{S12}$$

Lemma S5, proven below, establishes that the polynomial minimising this expression is

$$P(\gamma) = \frac{T_m\left(\frac{\gamma_{\max}+\gamma_{\min}-2\gamma}{\gamma_{\max}-\gamma_{\min}}\right)}{T_m\left(\frac{\gamma_{\max}+\gamma_{\min}}{\gamma_{\max}-\gamma_{\min}}\right)}$$

where $T_m(\cdot)$ is the m^{th} Chebyshev polynomial of the first kind.

Let $\kappa = \gamma_{\max}/\gamma_{\min}$. Now, $T_m(z) \in [-1, 1]$ for all m and all $z \in [-1, 1]$; thus the numerator takes maximum value 1. Therefore

$$\|P_m(\Gamma)\|_{\Sigma_0^{-1}}^{\text{op}} \leq \left|T_m\left(\frac{\kappa+1}{\kappa-1}\right)\right|^{-1}.$$

Lastly, note that by definition

$$T_m(z) = \frac{1}{2}\left[\left(z+\sqrt{z^2-1}\right)^m + \left(z-\sqrt{z^2-1}\right)^m\right]$$

so that

$$\begin{aligned}\|P_m(\Gamma)\|_2^{\text{op}} &\leq 2\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^m + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m\right]^{-1} \\ &\leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m.\end{aligned}$$

Inserting this into Eq. (S11) and recalling that since $\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}}$ has the same eigenvalues as $\Sigma_0 A^\top A$, it also has the same condition number, completes the proof. □

Lemma S5 (Appendix S3 of Shewchuk [1994])**.** *Eq.* (S12) *is minimised by*

$$P(\gamma) = \frac{T_m\left(\frac{\gamma_{\max}+\gamma_{\min}-2\gamma}{\gamma_{\max}-\gamma_{\min}}\right)}{T_m\left(\frac{\gamma_{\max}+\gamma_{\min}}{\gamma_{\max}-\gamma_{\min}}\right)}$$

where T_m is the m^{th} Chebyshev polynomial of the first kind.

Proof. For convenience let

$$\gamma_0 := \frac{\gamma_{\max} + \gamma_{\min}}{\gamma_{\max} - \gamma_{\min}}$$

and note that $\gamma_0 > 1$. Further, observe that

$$\gamma \in [\gamma_{\min}, \gamma_{\max}] \implies \frac{\gamma_{\max} + \gamma_{\min} - 2\gamma}{\gamma_{\max} - \gamma_{\min}} \in [-1, 1].$$

Now recall the following properties of Chebyshev polynomials:

C1 $T_m(z) \in [-1, 1]$ for all $z \in [-1, 1]$.

C2 $T_m(1) = 1$, and $T_m(-1) = (-1)^m$.

C3 Let $Z = \{z_i\}, i = 1, \ldots, m$ denote the ordered zeros of $T_m(z)$. Then, $Z \subset [-1, 1]$.

C4 $T_m(z)$ attains the value $(-1)^{m+i}$ in the range $[z_i, z_{i+1}]$ for $i = 1, \ldots, m-1$.

First, note that clearly $P(0) = 1$ as $T_m(\gamma_0) \neq 0$. This is because $\gamma_0 > 1$ and so $T_m(\gamma_0) > 1$ by **C2** and **C3**. Thus, $P(\gamma) \in \mathbb{P}_m$ as required. Further, note that

$$\max_{\gamma \in [\gamma_{\min}, \gamma_{\max}]} |P(\gamma)| = T_m(\gamma_0)^{-1}$$

by **C1**. Proof that $P(\gamma)$ minimizes Eq. (S12) is by contradiction. Suppose there is a $Q(\gamma) \in \mathbb{P}_m$ with

$$\max_{\gamma \in [\gamma_{\min}, \gamma_{\max}]} |Q(\gamma)| < T_m(\gamma_0)^{-1} \tag{S13}$$

Consider the polynomial $P(\gamma) - Q(\gamma)$. From **C1**, $P(\gamma) \in [-T_m(\gamma_0)^{-1}, T_m(\gamma_0)^{-1}]$, and $P(\gamma)$ has m zeros in $[\gamma_{\min}, \gamma_{\max}]$. From Eq. S13 it is clear that $P(\gamma) - Q(\gamma)$ also has m zeros in $[\gamma_{\min}, \gamma_{\max}]$, as to prevent $P(\gamma)$ from crossing zero between its extrema in this range would require $|Q(\gamma)| > T_m(\gamma_0)^{-1}$ (by **C4**).

However, since $P(0) = Q(0) = 1$, $P - Q$ has an additional zero outside $[\gamma_{\min}, \gamma_{\max}]$. Therefore, $P - Q$ is a polynomial of degree m with at least $m + 1$ zeros, which is a contradiction. Thus $P(\gamma)$ minimises Eq. S12.

□

S2 Further Simplication of BayesCG

In this section, the simplifications mentioned in Section 5 and exploited in Algorithm 1 are described in detail. First, two coefficients must be calculated, one to update $\boldsymbol{x}_m$ and one to update $\tilde{\boldsymbol{s}}_m$. Note that for stability reasons we work with un-normalized rather than normalized search directions where possible. As usual let $Q = A\Sigma_0 A^\top$, and express these quantities as

$$\begin{aligned}
\boldsymbol{x}_m &= \boldsymbol{x}_{m-1} + \alpha_m \Sigma_0 A^\top \tilde{\boldsymbol{s}}_m \\
\tilde{\boldsymbol{s}}_m &= \boldsymbol{r}_{m-1} + \beta_{m-1} \tilde{\boldsymbol{s}}_{m-1}
\end{aligned}$$

where

$$\alpha_m = \frac{\tilde{\boldsymbol{s}}_m^\top \boldsymbol{r}_{m-1}}{\|\tilde{\boldsymbol{s}}_m\|_Q^2}$$
$$\beta_m = -\frac{\boldsymbol{r}_m Q \tilde{\boldsymbol{s}}_m}{\|\tilde{\boldsymbol{s}}_m\|_Q^2}$$

Now, using the expression for $\tilde{\boldsymbol{s}}_m$, note that

$$\begin{aligned}\alpha_m &= \frac{\boldsymbol{r}_{m-1}^\top (\boldsymbol{r}_{m-1} - \beta_m \tilde{\boldsymbol{s}}_{m-1})}{\|\tilde{\boldsymbol{s}}_m\|_Q^2} \\ &= \frac{\boldsymbol{r}_{m-1}^\top \boldsymbol{r}_{m-1}}{\|\tilde{\boldsymbol{s}}_m\|_Q^2}\end{aligned}$$

since, from Lemma S3, $\tilde{\boldsymbol{s}}_m^\top \boldsymbol{r}_m = 0$. Furthermore, from the proof of Proposition 7, we have

$$\begin{aligned}\boldsymbol{r}_m^\top Q \boldsymbol{s}_m &= \frac{\boldsymbol{r}_m^\top \boldsymbol{r}_{m-1} - \boldsymbol{r}_m^\top \boldsymbol{r}_m}{\boldsymbol{s}_m^\top \boldsymbol{r}_{m-1}} \\ &= -\frac{\boldsymbol{r}_m^\top \boldsymbol{r}_m}{\boldsymbol{s}_m^\top \boldsymbol{r}_{m-1}} \\ &= -\frac{\boldsymbol{r}_m^\top \boldsymbol{r}_m}{\boldsymbol{r}_{m-1}^\top \boldsymbol{r}_{m-1}} \|\tilde{\boldsymbol{s}}_m\|_Q^2\end{aligned}$$

so that

$$\beta_m = \frac{\boldsymbol{r}_m^\top \boldsymbol{r}_m}{\boldsymbol{r}_{m-1}^\top \boldsymbol{r}_{m-1}}$$

These two simplifications allow rearranging the expressions in Proposition 6 into Algorithm 1.

S3 Additional Numerical Results for Simulation Study

In this section we discuss an empirical procedure for calibrating the scale of the posterior covariance, in an attempt to compensate for the fact that search directions depend upon $\boldsymbol{x}^*$, and show that this results in better calibrated UQ. The proposed approach is to construct an error indicator over the course of the algorithm, and then use this to adjust an appropriate measure of spread of the posterior to match that error prediction.

1. Constructing the Error Indicator The aim here is to construct a proxy for the true error by measuring the convergence of the BayesCG mean. Let

$$z_i := \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\|_2 \,.$$

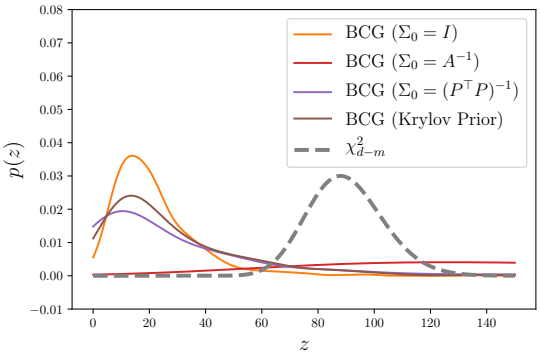

Figure S1: Uncertainty quantification provided in the new proposal. This should be compared against Fig. 3 in the main text.

The idea is to perform a simple regression on the values $\{z_i\}_{i=1}^m$ and use the fitted model $\nu(i)$ to extrapolate the error forward. Justified by the exponential convergence rate of BayesCG, as well as its simplicity, a log-linear function $\nu(i) = \exp(a+bi)$ has been used.

To derive our error indicator we use the following triangle inequality bound:

$$\begin{aligned}\|\boldsymbol{x}_m - \boldsymbol{x}^*\|_2 &\leq \sum_{i=m+1}^{d} \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\|_2 \\ &\approx \sum_{i=m+1}^{d} \nu(i) =: \alpha_m.\end{aligned}$$

Thus α_m provides an approximate upper-bound for $\|\boldsymbol{x}_m - \boldsymbol{x}^*\|_2$.

2. Fitting the Posterior Next we adjust the spread of the posterior, in a somewhat *ad-hoc* manner, based on the approximate upper-bound α_m on the true error. This requires the posterior spread to be quantified, and for the ease of computability we used $\text{trace}(\nu_m \Sigma_m)$. Thus, to be concrete, we would like to select ν_m so that the spread

$$\begin{aligned}\text{trace}(\nu_m \Sigma_m) &= \alpha_m \\ \implies \nu_m &= \frac{\alpha_m}{\text{trace}(\Sigma_m)}.\end{aligned}$$

Note that, since α_m appears in the numerator and provides an approximate upper bound for the true error, the UQ provided will still be conservative in general.

3. Results The UQ computed in the main text, Fig. 3, can be compared with the UQ under this proposal shown in Fig. S1. The UQ under this proposal is substantially better calibrated than when the Jeffrey's prior is used in all cases apart from the case when $\Sigma_0 = A^{-1}$. However since this performs well for all the practical choices of prior covariance suggested, the results indicate that approaches based on heuristic calibration

of posterior spread could be used to compensate for the fact that the search directions have been constructed in a data-driven manner.

S4 Experimental Results for Higher-Dimensional Systems

In this section additional experimental results are reported for higher-dimensional systems than those considered in Section 6.1. The experimental protocol adopted in that section is challenging to adapt to higher-dimensional systems, as the method for generating sparse positive-definite matrices has empirically been found to produce numerically singular matrices when d is increased. As a result, in this section we will adopt a more structured approach to generating random systems, based on discretisation of a simple elliptic PDE.

Specifically, the PDE considered is the following PDE with random boundary conditions:

$$\begin{aligned} -\nabla u(\boldsymbol{z}) &= 0 && \boldsymbol{z} \in (0,1)^2 \\ \frac{\partial u}{\partial z_2}(\boldsymbol{z}) &= 0 && z_1 \in (0,1),\ z_2 \in \{0,1\} \\ u(\boldsymbol{z}) &= f(\boldsymbol{z}) && z_1 \in \{0,1\},\ z_2 \in [0,1] \end{aligned} \tag{S14}$$

where $\log f(\boldsymbol{z}) \sim \mathcal{GP}(0, k(\boldsymbol{z}, \boldsymbol{z}'; \rho))$ and $k(\boldsymbol{z}, \boldsymbol{z}'; \rho)$ is a Matérn covariance function:

$$k(\boldsymbol{z}, \boldsymbol{z}'; \rho) = \left(1 + \frac{\sqrt{3}}{\rho}\right) \exp\left(-\frac{\sqrt{3}\|\boldsymbol{z} - \boldsymbol{z}'\|_2}{\rho}\right).$$

For the purposes of this experiment the length-scale was fixed to $\rho = 0.1$.

Eq. (S14) was discretised with the finite-element method using standard piecewise linear basis functions as implemented in `FEniCS`, as in Section 6.2. The domain was discretised using a simple regular triangular mesh over the domain resulting in d elements. To investigate the performance of BayesCG as a function of the dimension of the system, three different discretisation levels were used: $d = 121$, $d = 1089$ and $d = 10201$.

A subset of the priors from Section 6.1 were considered. The prior $\Sigma_0 = A^{-1}$ was not used, as for $d = 10201$ computing this is impractical. Similarly, the procedure we have used for calculating the Krylov subspace prior in that section requires knowledge of $\kappa(A)$, which is also impractical. As a result we have focussed on the prior $\Sigma_0 = (P^\top P)^{-1}$ where P is, as in Section 6.1, a preconditioner based on an incomplete Cholesky factorisation of A. Results for $\Sigma_0 = I$ are also included.

Results for convergence of the posterior mean are reported in Fig. S2. Note that the number of iterations required for both CG and BCG with the preconditioner prior seems to increase sub-linearly with dimension, while with $\Sigma_0 = I$ there is qualitatively little difference as a function of dimension. Furthermore note that the rate of convergence of CG appears to overtake that of the preconditioner prior as the dimension increases, suggesting that the quality of the incomplete Cholesky preconditioner decays with dimension. However, we note that this is only one choice of preconditioner, and others preconditioners may behave better.

The quality of the UQ as a function of dimension is displayed in Fig. S3, with the statistic Z as described in Section 6.1. The UQ after $\lfloor \frac{d}{10} \rfloor$ iterations was considered. The quality of UQ provided is seen to decrease as a function of dimension d.

S5 Experimental Set-Up for EIT

The results presented in this paper used experimental data provided by EIDORS[1] and due to Isaacson et al. [2004]. In the experiment, depicted in Figure 8a, three targets were placed into a tank filled with saline, two of which are lung-shaped and one of which is heart-shaped. The lung-shaped targets have lower conductivity than the surrounding saline, while the heart-shaped target has higher conductivity. A total of 32 electrodes were placed around the boundary of the domain, and stimulated with 31 distinct stimulation patterns as described in Isaacson et al. [2004]. For each stimulation, the voltage induced at every electrode was recorded, and there are thus 32×31 distinct measurements on which the prior must be conditioned. The inducing currents and measured voltages were each supplied in the referenced dataset.

In the simulations the circular tank was modelled as a unit circular domain, and the electrodes were assumed to occupy precisely $1/64^{\text{th}}$ of the boundary. Thus, each electrode had length $\pi/32$ and there was a distance of $\pi/32$ between each neighbouring pair of electrodes on the boundary. Since no information is known on the quality of the electrode contact, we set the contact impedances to an arbitrary value, $\zeta_l = 1$ for each l.

The trangulations required to discretise the PDE were generated using the Python package `meshpy`, configured to ensure that there were N_d equally sized elements on the boundary. N_d was chosen to be a multiple of the number of boundary electrodes, so that each electrode corresponds to the same number of boundary elements, and other boundary elements are disjoint from all electrodes. Figure S4 shows an illustration of a triangulation of the domain used to discretise the PDE, with $N_d = 64$.

S6 Additional Numerical Results for EIT

Figure S5 shows the posterior distribution obtained from BayesCG for different values of ϵ. The linear system solved was generated for $N_d = 128$ and with the conductivity field $\hat{\sigma}(\boldsymbol{z})$. Plotted is the posterior mean from BayesCG, along with samples from the posterior distribution, over the spatial domain of the PDE. That is, the voltage field $v(\boldsymbol{z})$ has been plotted rather than the conductivity field from the inverse problem. The top row has the largest value of ϵ, and here clearly the posterior mean deviates far from the true solution, depicted in the bottom row. However by $\epsilon = 5$ the mean from BayesCG appears close to the truth. The second, third and fourth column show samples from the posterior distribution, and while there is significantly more noise in these columns the main characteristics of the true solution are visible even at $\epsilon = 20$, suggesting that the use of BayesCG within a Bayesian approach to EIT can be qualitatively justified.

[1] At time of writing this data can be found at the EIDORS website.

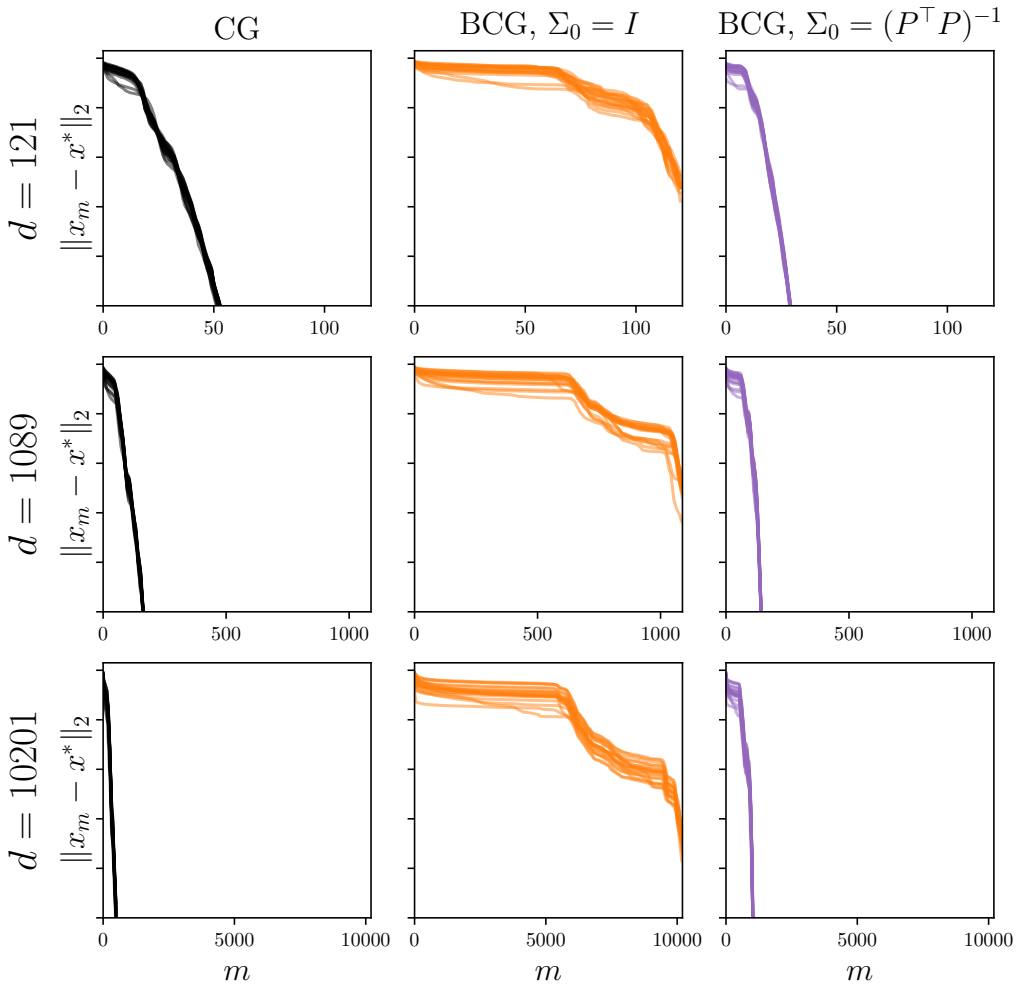

Figure S2: Convergence of the posterior mean from BayesCG, for the PDE example from Section S4, as dimension is varied. The error $\|\boldsymbol{x}_m = \boldsymbol{x}^*\|_2$ is reported for CG (left) as well as BayesCG with different prior covariances Σ_0.

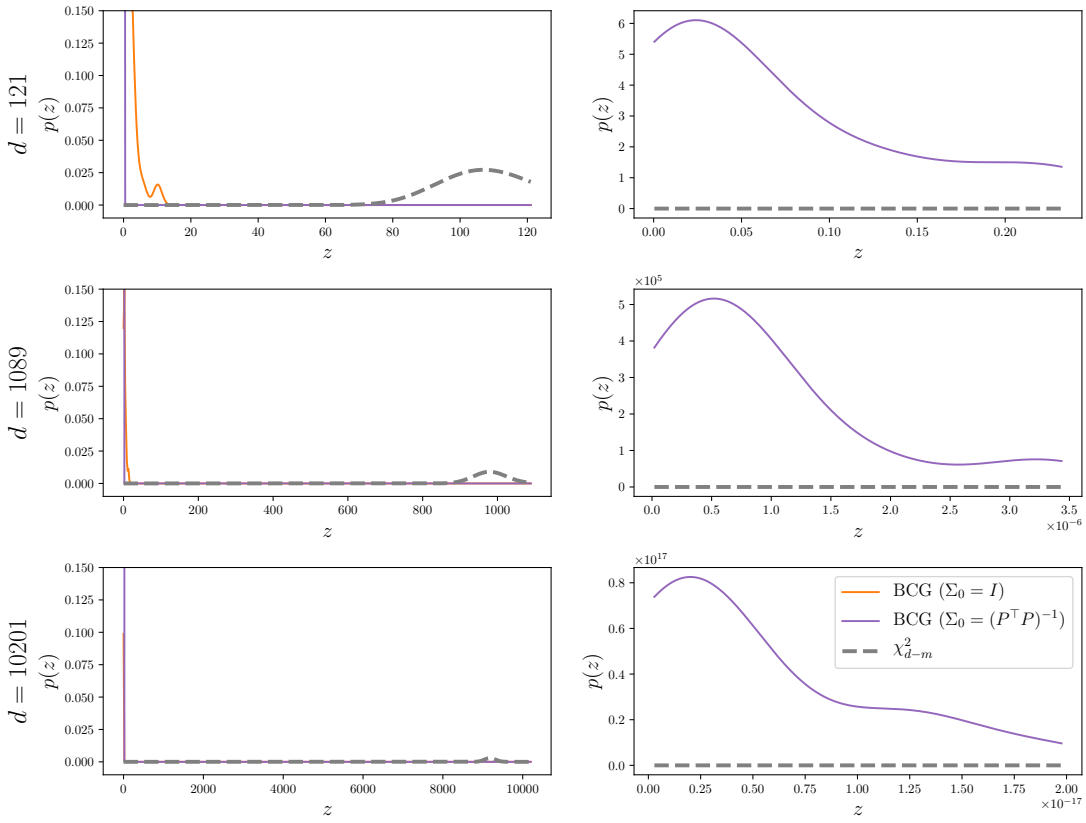

Figure S3: Assessment of the uncertainty quantification provided by the Gaussian version of BayesCG, for the PDE example from Section S4, as dimension is varied. The statistic Z is reported based on 100 independently sampled test problems, for different prior covariances Σ_0. These are compared to the theoretical distribution of Z when the posterior distribution is well-calibrated. The right panel zooms in on the portion of the x-axis occupied by the statistic for $\Sigma_0 = (P^\top P)^{-1}$.

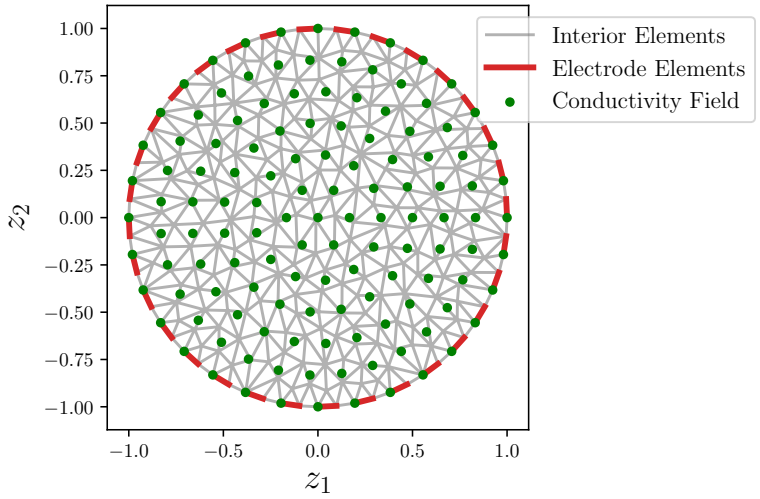

Figure S4: Finite-element discretisation used for the EIT experiment described in Section 6.2. The mesh was generated using the Python package `meshpy`, configured to ensure that there were 64 equally spaced boundary elements. Red lines indicate the elements which correspond to electrodes. Green dots show the locations at which the posterior conductivity field was sampled.

Fig. S6 shows the behavior of the posterior distribution when a standard CG forward solver is used but with m held fixed. At $m = 40$ the computed mean bears no resemblance to the actual posterior mean. Moreover, the computed distribution is over-confident, as reflected by the uniformly *lower* computed standard deviation, compared to the actual posterior. At $m = 60$ the qualitative features of the posterior mean have been recovered, though the recovery still differs noticeably from that of the actual posterior, and the computed standard deviation remains lower. This provides further motivation for the use of BayesCG, as a means to constrain the solver to fewer iterations while still obtaining estimates that are statistically meaningful.

Fig. S7 repeats this experiment for the BayesCG forward solver when the preconditioner prior is used. Again, as m is increased the posterior mean exhibits clear structure in the conductivity field. While the posterior variance does not visibly appear to decrease in the bottom row, the integrated standard deviation is nevertheless decreasing, starting at 0.0586 at $m = 40$, decreasing to 0.0459 at $m = 60$ and 0.0365 at $m = 80$.

References

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

D. Isaacson, J. L. Mueller, J. C. Newell, and S. Siltanen. Reconstructions of chest phantoms by the d-bar method for electrical impedance tomography. *IEEE*

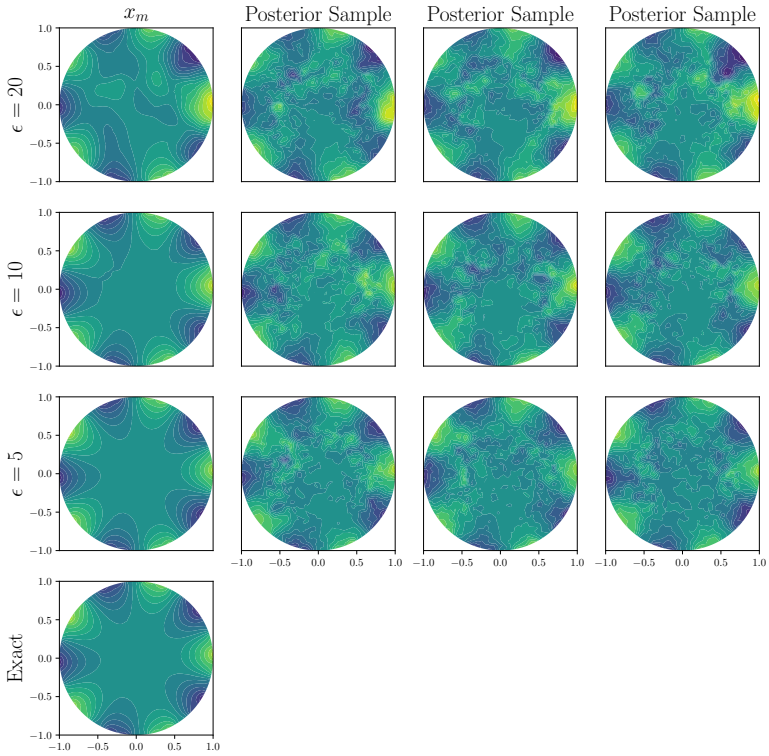

Figure S5: The posterior mean (left column), and samples from the posterior distribution (other columns) produced by BayesCG. The bottom panel represents the actual posterior mean, obtained when the defining linear systems are exactly solved.

Transactions on Medical Imaging, 23(7):821–828, July 2004. ISSN 0278-0062. doi:10.1109/TMI.2004.827482.

E. Novak and H. Woźniakowski. *Tractability of Multivariate Problems Volume I: Linear Information*. European Mathematical Society, 2008. doi:10.4171/026.

J. R. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. Technical report, 1994.

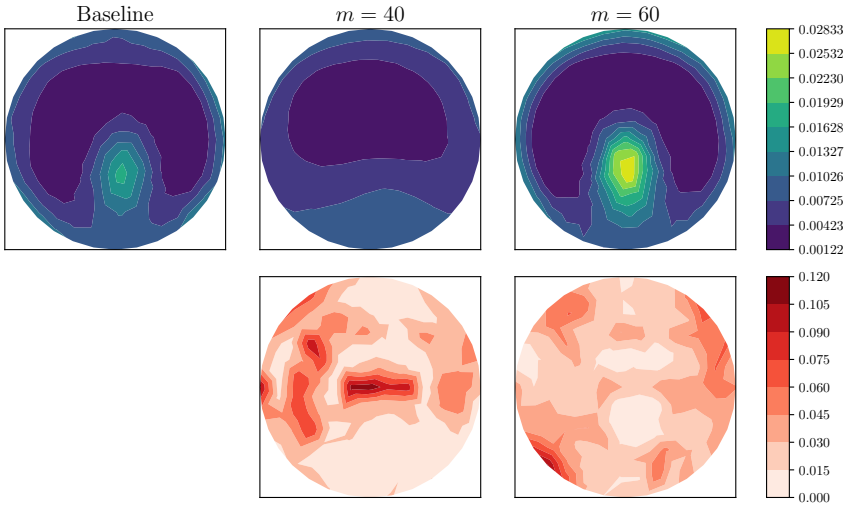

Figure S6: Posterior means (top row) and standard deviations relative to the actual ("baseline") posterior (bottom row) for the EIT problem with a standard CG forward solver, as m is varied.

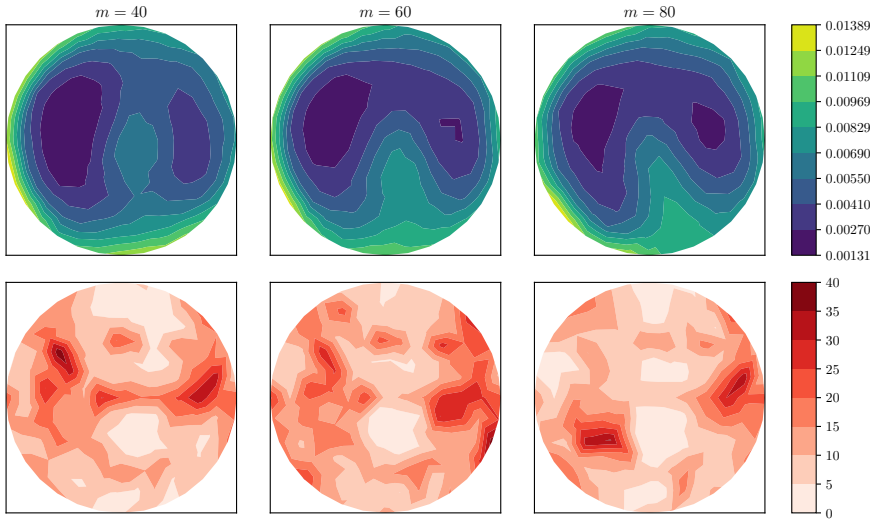

Figure S7: Posterior means (top row) and standard deviations relative to the actual ("baseline") posterior (bottom row) for the EIT problem using a BayesCG forward solver, as m is varied.