Supplementary Material for "A Bayesian Conjugate-Gradient Method"

Jon Cockayne, Chris J. Oates and Mark Girolami

January 16, 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

$$\begin{aligned} \boldsymbol{\ell}^\top \boldsymbol{x}_m - \boldsymbol{\ell}^\top \boldsymbol{x}^\dagger &= \boldsymbol{\ell}^\top (\boldsymbol{x}_0 - \boldsymbol{x}^\dagger) + \boldsymbol{\ell}^\top \Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top A (\boldsymbol{x}^\dagger - \boldsymbol{x}_0) && \text{(from Eq. 4)} \\ &= \boldsymbol{\ell}^\top (\Sigma_0 - \Sigma_0 A^\top S_m \Lambda_m^{-1} S_m^\top A \Sigma_0) \Sigma_0^{-1} (\boldsymbol{x}_0 - \boldsymbol{x}^\dagger) \\ &= \left\langle \Sigma_m \boldsymbol{\ell}, \boldsymbol{x}_0 - \boldsymbol{x}^\dagger \right\rangle_{\Sigma_0^{-1}} && \text{(from Eq. 5)} \end{aligned}$$

and so:

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

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. 5

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

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

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

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

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

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

An algorithm ψ^* which minimises $e^{\text{avg}}(\cdot, \psi)$ for arbitrary S_m is said to be *average-case-optimal.* An S_m^* which minimises $e^{\text{avg}}(S_m, \psi^*)$ 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. (8) 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. 4:

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

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. 4 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 $j < m$

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

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 $j \le m$, $\boldsymbol{r}_m^\top \boldsymbol{t}_j = 0$. Thus, from Eq. (S5) for each $j \le m$:

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

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 12. 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 6. 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

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

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 $\boldsymbol{x} \in K_1^*$:

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

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

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

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:

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

which is also a minimum. Thus, we have that

$$\arg\min_{\boldsymbol{x} \in K_{m-1}^*} \|\boldsymbol{x} - \boldsymbol{x}^\dagger\|_{\Sigma_0^{-1}}^2 = \boldsymbol{x}_{m-1} + \Sigma_0 A^\top \boldsymbol{s}_m (\boldsymbol{s}_m^\top \boldsymbol{r}_m) \equiv \boldsymbol{x}_m$$

from Proposition 6, which completes the proof. □

Proof of Lemma 10. 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 $\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}}$. As similar matrices share the same eigenvalues and since $\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}}$ is positive definite, the eigenvalues of $\Sigma_0 A^\top A$ must also be strictly positive. □

Proof of Proposition 11. We begin by introducing the *operator norm* induced by the energy norm $\|\cdot\|_A$, which is a norm on matrices $M \in \mathbb{R}^{d \times d}$

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

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

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

where P_m is some polynomial of degree m. Thus

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

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 recall from the proof of Lemma 10 that $\Sigma_0^{\frac{1}{2}} A^\top A \Sigma_0^{\frac{1}{2}}$ is similar to $\Sigma_0 A^\top A$ and hence the two matrices have the same eigenvalues.

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

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

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

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

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.

□

Algorithm 1 Computation of the posterior distribution described in Proposition 6. The implementation is optimised compared to that given in Proposition 6; see Supplement S2 for detail. Further note that, for clarity, all required matrix-vector multiplications have been left explicit, but for efficiency these should be calculated once-per-loop and stored. Σ_m can be computed from this output as $\Sigma_m = \Sigma_0 - \Sigma_F \Sigma_F^\top$.

procedure $\mathrm{BCG}(A, \boldsymbol{b}, \boldsymbol{x}_0, \Sigma_0, \epsilon, m_{\min}, m_{\max})$ ▷ (ϵ the tolerance)
    Σ_F initialised to a matrix of size $(d \times 0)$ ▷ ($m_{\min}$ the minimum # iterations)
    $\boldsymbol{r}_0 \leftarrow \boldsymbol{b} - A\boldsymbol{x}_0$ ▷ ($m_{\max}$ the maximum # iterations)
    $\tilde{\boldsymbol{s}}_1 \leftarrow \boldsymbol{r}_0$
    $\tilde{\nu}_0 \leftarrow 0$
    for $m = 1, \ldots, m_{\max}$ **do**
        $E^2 \leftarrow \tilde{\boldsymbol{s}}_m^\top A \Sigma_0 A^\top \tilde{\boldsymbol{s}}_m$
        $\alpha_m \leftarrow \frac{\boldsymbol{r}_{m-1}^\top \boldsymbol{r}_{m-1}}{E^2}$
        $\boldsymbol{x}_m \leftarrow \boldsymbol{x}_{m-1} + \alpha_m \Sigma_0 A^\top \tilde{\boldsymbol{s}}_m$
        $\boldsymbol{r}_m \leftarrow \boldsymbol{r}_{m-1} - A\boldsymbol{x}_m$
        $\Sigma_F \leftarrow [\Sigma_F, \Sigma_0 A^\top \tilde{\boldsymbol{s}}_m / E]$
        $\tilde{\nu}_m \leftarrow \tilde{\nu}_{m-1} + \frac{(\boldsymbol{r}_{m-1}^\top \boldsymbol{r}_{m-1})^2}{E^2}$
        $\sigma_m \leftarrow \frac{\sqrt{(d-m)\tilde{\nu}_m}}{\sqrt{m}}$
        if $\sigma_m < \epsilon$ & $m \geq m_{\min}$ **then**
            break
        end if
        $\beta_m \leftarrow \frac{\boldsymbol{r}_m^\top \boldsymbol{r}_m}{\boldsymbol{r}_{m-1}^\top \boldsymbol{r}_{m-1}}$
        $\tilde{\boldsymbol{s}}_{m+1} \leftarrow \boldsymbol{r}_m + \beta_m \tilde{\boldsymbol{s}}_m$
    end for
    return $\boldsymbol{x}_m, \Sigma_{F,m}, \nu_m := \tilde{\nu}_m / m$
end procedure

S2 Pseudocode for the Bayesian Conjugate Gradient Method

Full pseudocode for the BCG method is provided in Algorithm 1. Two algebraic simplifications have been exploited here relative to the presentation in the main text. 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

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

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 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

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

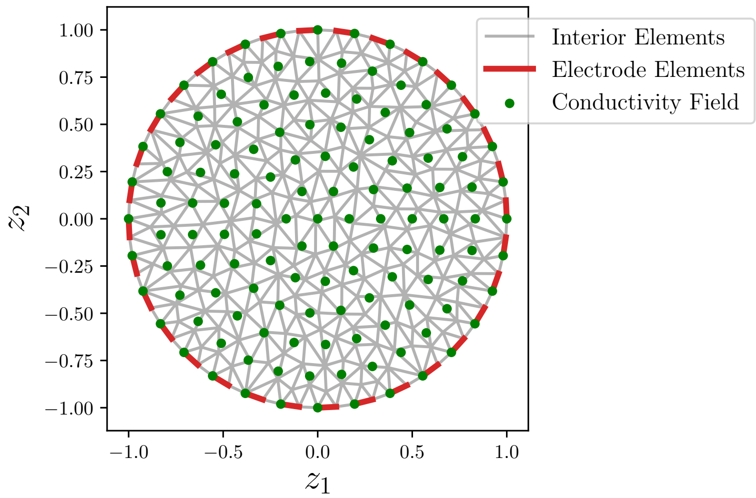

Figure 1: Finite-element discretisation used for the EIT experiment described in Section 6.2. The mesh was generated using the Python package `meshpy`, configured to ensure that there were 64 equally spaced boundary elements. Red lines indicate the elements which correspond to electrodes. Green dots show the locations at which the posterior conductivity field was sampled.

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 1 shows an illustration of a triangulation of the domain used to discretise the PDE, with $N_d = 64$.

S4 Additional Numerical Results for EIT

Figure 2 shows the posterior distribution obtained from BCG 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 BCG, along with samples from the posterior distribution, over the spatial domain of the PDE. That is, the voltage field $v(\boldsymbol{z})$ has

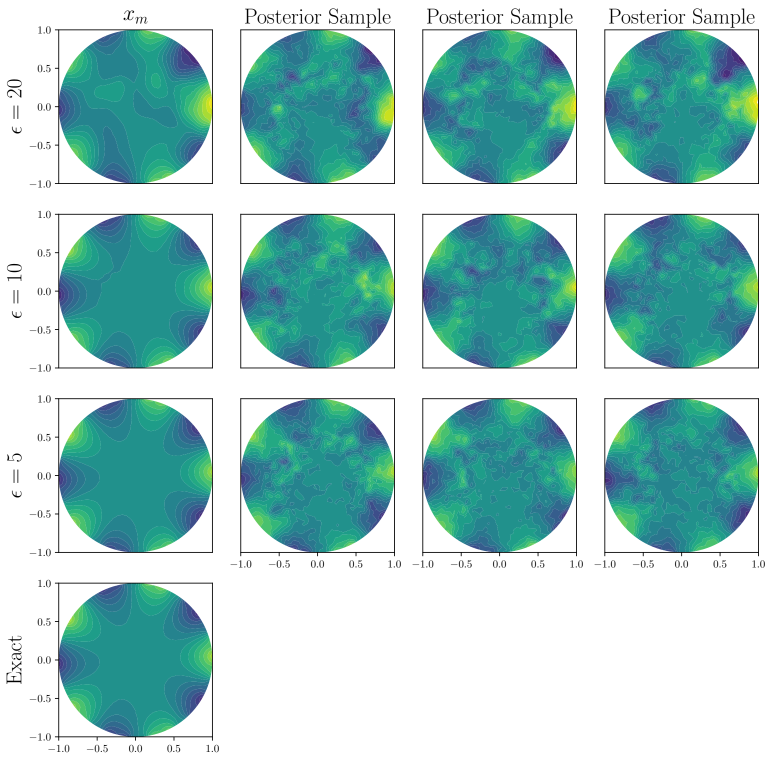

Figure 2: The posterior mean (left column), and samples from the posterior distribution (other columns) produced by BCG. The bottom panel represents the true posterior mean obtained when the defining linear systems are exactly solved.

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 BCG 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 BCG within a Bayesian approach to EIT can be qualitatively justified.

References

Jon Cockayne, Chris Oates, Tim Sullivan, and Mark 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 Transactions on Medical Imaging*, 23(7):821–828, July 2004. ISSN 0278-0062. doi:10.1109/TMI.2004.827482.

Erich Novak and Henryk Woźniakowski. *Tractability of Multivariate Problems Volume I: Linear Information.* European Mathematical Society, 2008.

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