We thank the reviewers for their helpful feedback, and we hope to address your remarks here and in the revision. We note that R4 questioned the correctness of Thm. 1 (specifically, the claim made in Lemma 2). While we thank R4 for a careful review of Appx. G, we are *strongly* convinced that Thm. 1 is correct and hope to clarify the misunderstanding.

Correctness of Thm. 1 (R4). Your question revolves around a specific inequality in the proof of Lemma 2:

$$\left(\frac{\sqrt{\kappa(\mathbf{K}+t_q\mathbf{I})}-1}{\sqrt{\kappa(\mathbf{K}+t_q\mathbf{I})}+1}\right)^J \|\mathbf{b}\|_2 \le \left(\frac{\sqrt{\kappa(\mathbf{K})}-1}{\sqrt{\kappa(\mathbf{K})}+1}\right)^J \|\mathbf{b}\|_2. \tag{1}$$

Your counter example. "Set $\mathbf{K} = 4\mathbf{I}$ and $t = 5$—then on the left we have $(1/2)^J$ and on the right $(1/3)^J$—the former is clearly larger." We believe you may accidentally used the max eigenvalue alone in these calculations, rather than $\kappa(\mathbf{K})$.

For symmetric positive definite $\mathbf{K}$, the condition number $\kappa(\mathbf{K}) \triangleq \lambda_{\max}/\lambda_{\min}$ (max and min eigenvalues). Adding $t\mathbf{I}$ increases both the max and min eigenvalues; thus $\kappa(\mathbf{K} + t\mathbf{I}) = (\lambda_{\max} + t)/(\lambda_{\min} + t)$. Using your specific numbers, we have $\kappa(\mathbf{K}) = 4/4 = 1$ and $\kappa(\mathbf{K} + t_q\mathbf{I}) = 9/9 = 1$. Plugging these into Eq. (1) we have that both sides equal 0.

Proof. Here we carefully show that Eq. (1) holds for any symmetric positive definite matrix $\mathbf{K}$ and $t \geq 0$. Note that $\kappa(\mathbf{K} + t\mathbf{I}) = \frac{\lambda_{\max}+t}{\lambda_{\min}+t} \leq \frac{\lambda_{\max}}{\lambda_{\min}} = \kappa(\mathbf{K})$, which holds as long as $\lambda_{\max} \geq \lambda_{\min} > 0$ (true here as $\mathbf{K} \succ 0$). From here, note that $\frac{a-1}{a+1} \leq \frac{b-1}{b+1} < 1$ whenever $0 \leq a \leq b$. Setting $a = \sqrt{\kappa(\mathbf{K} + t_q\mathbf{I})}$ and $b = \sqrt{\kappa(\mathbf{K})}$ and noting that condition numbers are always at least 1 (implying the square root preserves the ordering), we have $a \leq b$, and thus Eq. (1) holds.

"Krylov methods often suffer from a high degree of numerical instability" (R4). Our method has two key advantages that improve stability. First, we only use Krylov methods to solve linear systems rather than eigenvalue problems. Common numerical pitfalls that hinder Krylov eigen-solvers (e.g. loss of orthogonality between Lanczos vectors) have been shown to have little empirical effect on linear system solvers like MINRES and CG [e.g. Trefethen and Bau, 1997; Fong and Saunders, SQUJS 2012]. Second, each solve is inherently a shifted system $\mathbf{K} + t_q\mathbf{I}$. While Thm. 1 is in terms of the conditioning of $\mathbf{K}$ (because $\min_q t_q \to 0$ as $Q \to \infty$), in practice these shifts *dramatically* improve the conditioning of $\mathbf{K}$, with $\min_q t_q \geq 1\text{e}{-3}$ when $Q < 20$. This allows us to work directly with the matrix $\mathbf{K}$ rather than having to first add diagonal jitter. We will discuss this more in the revision.

Accuracy of square roots (R3). (*"In Fig. 1 the relative error appears to level off as Q increases."*) For these experiments, we stopped msMINRES at a tolerance of $1\text{e}{-4}$, as we viewed 0.01% error as sufficient for most tasks. Consequentially, as Q increases the error converges to the solver tolerance. (*Is 4 or 5 decimal places enough for predictive means/variances?*) We believe this is often sufficient; we tried tighter tolerances and found no difference.

Running time and storage (R3, R4). **R3:** (*"I didn't understand… storage being reduced from $\mathcal{O}(N^2)$."*) Using our method, we can avoid storing the $\mathcal{O}(N^2)$ kernel matrix if the MVMs are computed in a map-reduce fashion. While the Cholesky factorization can be performed in-place, the artifact it produces still requires $\mathcal{O}(N^2)$ storage. We will clarify this in the revision. **R4:** (*"The cost from CIQ must be linear in Q."*) We agree that the quadrature running time depends on Q, but view this as negligible since it crucially does not impact the number of MVMs performed with $\mathbf{K}$. Thank you for pointing this out; we will clarify this.

Missing citations (R1). Thank you, we will add these citations. While many communities have extensively studied Krylov methods, as you note—the $\mathbf{K}^{-1/2}\mathbf{b}$ problem we are interested in has received far less attention than Krylov-based matrix solves. We would again highlight our novel contributions in this space: a simple vector recurrence for $\mathbf{K}^{-1/2}\mathbf{b}$, a detailed error analysis, efficient NGD updates, a simple backward pass, and a mechanism for preconditioning multiple shifted solves up to an orthogonal transformation.

Comparisons (R2). (*"It will be necessary to compare to [Wilson et. al, ICML 2020]."*) This was not published as of the NeurIPS submission, as you point out. Wilson et al. use RFFs to sample from the prior and an inducing point approximation of the conditional to convert prior samples into posterior samples. Our method could augment their approach, allowing for more inducing points and/or replacing RFFs for prior sampling. (*"Missing a comparison to the related work that used CG."*) We would argue that the pros/cons of CG versus MINRES has been studied exhaustively (e.g. [Fong and Saunders, SQUJS 2012]). Our use of MINRES enables the proof of our main convergence result. CG's error bound uses a different norm and cannot be easily combined with the "quadrature error" term in Thm. 1.

Empirical evaluation (R2). ("*The number of MINRES iterations J has been fixed to 200.*") The stopping criteria is a specified residual tolerance (10^{-3} or 10^{-4}) or 200 iterations, whichever comes first. In practice the tolerance is almost always met before $J = 200$; we will note this in the revision. ("*How many MINRES iterations… before the approach becomes slower than Cholesky?*") This depends on matrix size. For $N = 5{,}000$ it would take approximately $J = 1{,}000$ iterations, after which CIQ has more than converged.