Preconditioning for Scalable Gaussian Process Hyperparameter Optimization

Jonathan Wenger [1 2 3] **Geoff Pleiss** [3] **Philipp Hennig** [1 2] **John P. Cunningham** [3] **Jacob R. Gardner** [4]

Abstract

Gaussian process hyperparameter optimization requires linear solves with, and log-determinants of, large kernel matrices. Iterative numerical techniques are becoming popular to scale to larger datasets, relying on the conjugate gradient method (CG) for the linear solves and stochastic trace estimation for the log-determinant. This work introduces new algorithmic and theoretical insights for preconditioning these computations. While preconditioning is well understood in the context of CG, we demonstrate that it can also accelerate convergence and reduce variance of the estimates for the log-determinant and its derivative. We prove general probabilistic error bounds for the preconditioned computation of the log-determinant, log-marginal likelihood and its derivatives. Additionally, we derive specific rates for a range of kernel-preconditioner combinations, showing that up to exponential convergence can be achieved. Our theoretical results enable provably efficient optimization of kernel hyperparameters, which we validate empirically on large-scale benchmark problems. There our approach accelerates training by up to an order of magnitude.

1. Introduction

Gaussian processes (GPs) are a theoretically well-founded and powerful probabilistic model (Rasmussen & Williams, 2006). However, conditioning a GP on data is often computationally prohibitive for large datasets. This problem is amplified when optimizing kernel hyperparameters. Gradient-based optimization requires repeated evaluation of the log-marginal likelihood $\mathcal{L}$ and its derivatives. These computations both have cubic complexity in the size n of the data.

[1]University of Tübingen [2]Max Planck Institute for Intelligent Systems, Tübingen [3]Columbia University [4]University of Pennsylvania. Correspondence to: Jonathan Wenger <jonathan.wenger@uni-tuebingen.de>.

Preprint.

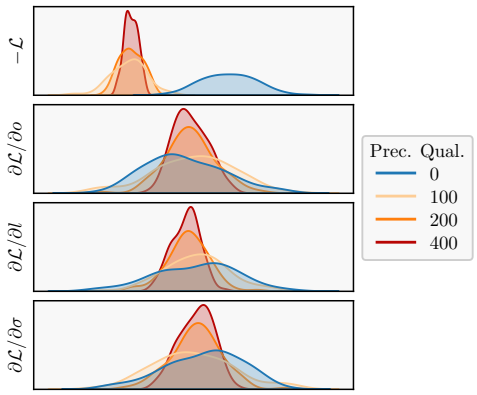


Figure 1: *Preconditioning reduces not only bias but also variance in stochastic approximations to the* log*-marginal likelihood $\mathcal{L}$ and its derivatives.* GP hyperparameter optimization for large datasets requires cheap estimates of $\mathcal{L}$ and its gradient. Preconditioning makes these more precise and less noisy as is shown here for increasing preconditioner quality on the "Elevators" dataset using a Matérn$(\frac{3}{2})$ kernel.

GP Inference via Matrix-Vector Multiplication Recently, Krylov methods (Golub & van Loan, 2013), based on iterative matrix-vector multiplication with the kernel matrix, have become popular for GP inference (Murray, 2009; Cutajar et al., 2016; Ubaru et al., 2017; Dong et al., 2017; Gardner et al., 2018; Wang et al., 2019). Methods primarily relying on matrix-vector products are advantageous. They can leverage structure in the kernel matrix (Nocedal & Wright, 2006), and importantly, they make effective use of modern hardware and parallelization (Gardner et al., 2018; Wang et al., 2019; Charlier et al., 2021). When optimizing GP hyperparameters one needs to repeatedly solve linear systems with, and compute log-determinants of, the kernel matrix. Both can be done using Krylov methods. The linear systems are solved via the conjugate gradient method (CG) (Hestenes & Stiefel, 1952), which reduces the cost of kernel matrix solves from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 m)$ for m iterations. The log-determinant can be approximated via stochastic trace estimation (STE) (Hutchinson, 1989) combined with another Krylov method, the Lanczos algorithm (Lanczos, 1950) (see (Ubaru et al., 2017)). Its derivative may also be estimated via STE combined with CG (Gardner et al., 2018).

Challenges with this Approach Despite the advantages of combining Krylov methods with stochastic trace estimation, there are considerable challenges in practice. These essentially reduce to bias and variance of the numerical approximations. First, the convergence of CG depends on the conditioning of the kernel matrix, which can grow rapidly with n (e.g. for the RBF kernel). Many iterations may be needed to achieve a desired error, and stopping the solver early can result in biased solutions (Potapczynski et al., 2021). However, if a preconditioner – i.e. an approximation of the kernel matrix – is available, convergence can be accelerated substantially (Golub & van Loan, 2013). Second, stochastic approximations of the log-determinant and its derivative introduce variance into hyperparameter optimization. While the estimates are unbiased (assuming sufficient Krylov iterations), variance can significantly slow down optimization. Reducing variance either requires further approximation at the cost of more bias (Artemev et al., 2021), or a larger number of samples ℓ which only reduces error at a rate of $\mathcal{O}(\ell^{-\frac{1}{2}})$ (Avron & Toledo, 2011). Now, while preconditioning is known to accelerate CG, it has not yet been explored for stochastic trace estimation in this context.

Contributions We demonstrate that, with only a small algorithmic modification, preconditioning can be exploited for highly efficient log-determinant estimation, and in turn GP hyperparameter optimization. We show that

(a) *preconditioning reduces variance* – or equivalently accelerates convergence – of the stochastic estimate of the log-determinant and its derivative (Theorem 1).

We leverage this result, illustrated in Figure 1, to prove

(b) *stronger theoretical guarantees* for the computation of the log-determinant (Theorems 2 and 3) and log-marginal likelihood (Theorem 4) than previously known (Ubaru et al., 2017; Gardner et al., 2018) and a *novel error bound for the derivative* (Theorem 5).

To make these general results concrete, we derive

(c) *specific rates for important combinations of kernels and preconditioners* (Table 1), making preconditioner choice for GP inference rigorous rather than heuristic.

Finally, using our approach, we empirically observe

(d) *up to twelvefold speedup in training* of GP regression models applied to large-scale benchmark problems with up to $n \approx 325{,}000$ datapoints.

2. Background

We want to infer a map $h : \mathcal{X} \to \mathcal{Y}$ from an input space $\mathcal{X} \subset \mathbb{R}^d$ to an output space $\mathcal{Y} \subset \mathbb{R}$, given a dataset $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ of n training inputs $\boldsymbol{x}_i \in \mathbb{R}^d$ and outputs $\boldsymbol{y} \in \mathbb{R}^n$.

2.1. Gaussian Processes

A stochastic process $f \sim \mathcal{GP}(\mu, k)$ with mean function μ and kernel k is called a *Gaussian process* if $\mathbf{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))^\intercal \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$ is jointly Gaussian with mean $\boldsymbol{\mu}_i = \mu(\boldsymbol{x}_i)$ and covariance $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Assuming $\boldsymbol{y} \mid \mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2 \boldsymbol{I})$, the posterior distribution for test inputs $\boldsymbol{x}_\star$ is also Gaussian with

$$\mathbb{E}[\mathbf{f}_\star] = \mu(\boldsymbol{x}_\star) + k(\boldsymbol{x}_\star, \boldsymbol{X})\hat{\boldsymbol{K}}^{-1}\boldsymbol{y},$$
$$\mathrm{Cov}(\mathbf{f}_\star) = k(\boldsymbol{x}_\star, \boldsymbol{x}_\star) - k(\boldsymbol{x}_\star, \boldsymbol{X})\hat{\boldsymbol{K}}^{-1}k(\boldsymbol{X}, \boldsymbol{x}_\star),$$

where $\hat{\boldsymbol{K}} = \boldsymbol{K} + \sigma^2 \boldsymbol{I}$.

Hyperparameter Optimization The computational bottleneck when optimizing kernel hyperparameters $\boldsymbol{\theta}$ is the repeated evaluation of the log-*marginal likelihood*

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}) \\ &= -\frac{1}{2}\big(\boldsymbol{y}^\intercal \hat{\boldsymbol{K}}^{-1}\boldsymbol{y} + \underbrace{\log\det(\hat{\boldsymbol{K}})}_{=\mathrm{tr}(\log(\hat{\boldsymbol{K}}))} + n\log(2\pi)\big) \quad (1)\end{aligned}$$

and its *derivative* with respect to the hyperparameters

$$\frac{\partial}{\partial\theta}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{y}^\intercal \hat{\boldsymbol{K}}^{-1}\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\hat{\boldsymbol{K}}^{-1}\boldsymbol{y} - \frac{1}{2}\,\mathrm{tr}\big(\hat{\boldsymbol{K}}^{-1}\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\big). \quad (2)$$

Computing (1) and (2) via a Cholesky decomposition has complexity $\mathcal{O}(n^3)$ which is prohibitive for large n. In response, many methods for approximate GP inference were developed (Titsias, 2009; Hensman et al., 2013; Wilson et al., 2015; Wilson & Nickisch, 2015; Pleiss et al., 2018).

In contrast, we aim for tractable *numerically exact* GP inference in the large-scale setting (Gardner et al., 2018; Wang et al., 2019). To achieve this, we focus on efficient computation of the log-determinant in (1) and its derivative in (2) (see Section 3). This allows us to theoretically (Section 4) and empirically (Section 5) accelerate GP hyperparameter optimization.

2.2. Numerical Toolbox for Inference

We will use the following established numerical techniques.

Stochastic Trace Estimation (STE) The trace $\mathrm{tr}(\boldsymbol{A})$ of a matrix can be approximated by drawing ℓ independent random vectors $\boldsymbol{z}_i$ with $\mathbb{E}[\boldsymbol{z}_i] = \boldsymbol{0}$ and $\mathrm{Cov}(\sqrt{n}\boldsymbol{z}_i) = \boldsymbol{I}$ and computing *Hutchinson's estimator* (Hutchinson, 1989)

$$\tau_\ell^{\mathrm{STE}}(\boldsymbol{A}) = \tfrac{n}{\ell}\textstyle\sum_{i=1}^{\ell} \boldsymbol{z}_i^\intercal \boldsymbol{A}\boldsymbol{z}_i \approx \mathrm{tr}(\boldsymbol{A}). \quad (3)$$

Here, we additionally assume the random vectors are normalized[1] $\boldsymbol{z}_i = \tilde{\boldsymbol{z}}_i / \|\tilde{\boldsymbol{z}}_i\|_2$ and that $\sqrt{n}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\ell)^\intercal \in \mathbb{R}^{\ell n}$

[1] Normalization is necessary for Lanczos quadrature (see Golub & Meurant, 2009, Chap. 7.2).

satisfies the convex concentration property[2] (see Definition S1). These assumptions are fulfilled by Rademacher-distributed random vectors with entries $\{+1, -1\}$.[3] Evaluating the log-marginal likelihood (1) requires computing $\mathrm{tr}(\log(\hat{\boldsymbol{K}}))$. To use Hutchinson's estimator, we need to efficiently compute ℓ quadratic terms $\{\boldsymbol{z}_i^\intercal \log(\hat{\boldsymbol{K}})\boldsymbol{z}_i\}_{i=1}^{\ell}$.

Stochastic Lanczos Quadrature (SLQ) Given a matrix function f, one can approximate bilinear forms $\boldsymbol{z}^\intercal f(\hat{\boldsymbol{K}})\boldsymbol{z}$ using quadrature (Golub & Meurant, 2009, Chap. 7). The nodes and weights of the quadrature rule can be computed efficiently via m iterations of the *Lanczos algorithm* (Lanczos, 1950) (or equivalently via CG (Gardner et al., 2018)). The combination with Hutchinson's estimator

$$\tau_{\ell,m}^{\mathrm{SLQ}}(f(\hat{\boldsymbol{K}})) \approx \tau_{\ell}^{\mathrm{STE}}(f(\hat{\boldsymbol{K}})) \approx \mathrm{tr}(f(\hat{\boldsymbol{K}})), \tag{4}$$

is called *stochastic Lanczos quadrature* (Ubaru et al., 2017).

To compute the linear solves $\boldsymbol{v} \mapsto \hat{\boldsymbol{K}}^{-1}\boldsymbol{v}$ with the kernel matrix in (1) and (2), we use the *conjugate gradient method*.

Conjugate Gradient Method (CG) CG (Hestenes & Stiefel, 1952) is an iterative method for solving linear systems with symmetric positive definite matrix. It is particularly suited for large systems since it is matrix-free, and relies primarily on matrix-vector multiplication with $\hat{\boldsymbol{K}}$.

Preconditioning It is well-known that CG can be accelerated via a symmetric positive definite preconditioner $\hat{\boldsymbol{P}} \approx \hat{\boldsymbol{K}}$, by solving an equivalent linear system with matrix $\hat{\boldsymbol{P}}^{-\frac{1}{2}}\hat{\boldsymbol{K}}\hat{\boldsymbol{P}}^{-\frac{\intercal}{2}} \approx \boldsymbol{I}$ (Trefethen & Bau, 1997). CG's convergence is then determined by the condition number

$$\kappa := \kappa(\hat{\boldsymbol{P}}^{-\frac{1}{2}}\hat{\boldsymbol{K}}\hat{\boldsymbol{P}}^{-\frac{\intercal}{2}}) \ll \kappa(\hat{\boldsymbol{K}}) = \frac{|\lambda_{\max}(\hat{\boldsymbol{K}})|}{|\lambda_{\min}(\hat{\boldsymbol{K}})|}. \tag{5}$$

Suppose the approximation quality of a sequence of preconditioners $\{\hat{\boldsymbol{P}}_\ell\}_\ell$ indexed by ℓ is given by[4]

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq \mathcal{O}(g(\ell))\|\hat{\boldsymbol{K}}\|_F. \tag{6}$$

If $g(\ell) \to 0$ quickly, a small amount of precomputation can significantly accelerate CG, since by Lemma S4

$$\kappa \leq (1 + \mathcal{O}(g(\ell))\|\hat{\boldsymbol{K}}\|_F)^2. \tag{7}$$

Preconditioners must be cheap to obtain and allow efficient linear solves $\boldsymbol{v} \mapsto \hat{\boldsymbol{P}}^{-1}\boldsymbol{v}$.[5] As an example, *diagonal-plus-low-rank* preconditioners $\hat{\boldsymbol{P}}_\ell = \sigma^2\boldsymbol{I} + \boldsymbol{L}_\ell\boldsymbol{L}_\ell^\intercal$, with $\boldsymbol{L}_\ell \in \mathbb{R}^{n\times\ell}$, admit linear solves in $\mathcal{O}(n\ell^2)$ via the matrix inversion lemma.

[2]Concentration enables us to prove probabilistic error bounds.

[3]We conjecture they also hold for vectors $\tilde{\boldsymbol{z}}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

[4]The use of ℓ for the number of random vectors and the preconditioner sequence is deliberate. Setting them to the same value enables variance reduction as we prove in Theorem 1.

[5]While CG (and Lanczos) assume a symmetric pos. definite matrix, both can be implemented using only $\hat{\boldsymbol{P}}^{-1}$, not $\hat{\boldsymbol{P}}^{-\frac{1}{2}}\hat{\boldsymbol{K}}\hat{\boldsymbol{P}}^{-\frac{\intercal}{2}}$.

3. Log-Determinant Estimation

Our goal is to compute $\log\det(\hat{\boldsymbol{K}}) = \mathrm{tr}(\log(\hat{\boldsymbol{K}}))$ and its derivative via matrix-vector multiplication. As described, we can use stochastic trace estimation to do so. Now assume we additionally have access to a preconditioner $\hat{\boldsymbol{P}} \approx \hat{\boldsymbol{K}}$. As we will show, we can then not just accelerate the convergence of CG, but *also* more efficiently compute the forward and backward pass for the log-determinant.

By the properties of the matrix logarithm we can decompose the log-determinant into a *deterministic approximation* based on the preconditioner and a *residual trace* computed via stochastic trace estimation.[6] It holds by Lemma S3, that

$$\log\det(\hat{\boldsymbol{K}}) = \log\det(\hat{\boldsymbol{P}}_\ell) + \mathrm{tr}(\underbrace{\log(\hat{\boldsymbol{K}}) - \log(\hat{\boldsymbol{P}}_\ell)}_{=\boldsymbol{\Delta}_{\log}}) \tag{8}$$

where $\mathrm{tr}(\boldsymbol{\Delta}_{\log}) = \mathrm{tr}\big(\log\big(\hat{\boldsymbol{P}}_\ell^{-\frac{1}{2}}\hat{\boldsymbol{K}}\hat{\boldsymbol{P}}_\ell^{-\frac{\intercal}{2}}\big)\big)$ and we assume $\log\det(\hat{\boldsymbol{P}}_\ell)$ is efficient to compute. Equation (8) has two crucial benefits we can exploit. First and foremost, the faster $\hat{\boldsymbol{P}}_\ell \to \hat{\boldsymbol{K}}$, i.e. $g(\ell) \to 0$, the less the stochastic approximation of $\mathrm{tr}(\boldsymbol{\Delta}_{\log})$ affects the estimate. Since its contribution to the overall error decreases the better $\log\det(\hat{\boldsymbol{P}}_\ell)$ approximates $\log\det(\hat{\boldsymbol{K}}_\ell)$, significantly fewer random vectors are needed to achieve a desired error with high probability. Second, we can now run Lanczos on the preconditioned matrix accelerating its convergence. As we will show later, one can also exploit (8) for the backward pass.

3.1. Variance-reduced Stochastic Trace Estimation

This intuitive argument for the log-determinant also holds generally, assuming a similar decomposition exists.

Theorem 1 (Variance-reduced Stochastic Trace Estimation) *Let $\hat{\boldsymbol{K}}, \hat{\boldsymbol{P}}_\ell \in \mathbb{R}^{n\times n}_{\mathrm{spd}}, \boldsymbol{\Delta}_f \in \mathbb{R}^{n\times n}$ and $f : \mathbb{R}^{n\times n}_{\mathrm{spd}} \to \mathbb{R}$ such that $\mathrm{tr}(f(\hat{\boldsymbol{K}})) = \mathrm{tr}(f(\hat{\boldsymbol{P}}_\ell)) + \mathrm{tr}(\boldsymbol{\Delta}_f)$, and define the estimator $\tau_* = \mathrm{tr}(f(\hat{\boldsymbol{P}}_\ell)) + \tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_f)$. Now, assume there exist $c_{\boldsymbol{\Delta}} > 0$ and $g : \mathbb{N} \to (0, \infty)$ such that*

$$\|\boldsymbol{\Delta}_f\|_F \leq c_{\boldsymbol{\Delta}} g(\ell)\|f(\hat{\boldsymbol{K}})\|_F. \tag{9}$$

Then there exists $c_{\boldsymbol{z}} > 0$ dependent on the choice of random vectors, such that, if $\ell \geq c_{\boldsymbol{z}}\log(\delta^{-1})$, it holds with probability $1 - \delta \in [\frac{1}{2}, 1)$ that

$$|\tau_* - \mathrm{tr}(f(\hat{\boldsymbol{K}}))| \leq \varepsilon_{\mathrm{STE}}\|f(\hat{\boldsymbol{K}})\|_F. \tag{10}$$

where for $C_1 = c_{\boldsymbol{\Delta}}\sqrt{c_{\boldsymbol{z}}}$ the relative error is given by

$$\varepsilon_{\mathrm{STE}}(\delta, \ell) = C_1\sqrt{\log(\delta^{-1})}\ell^{-\frac{1}{2}}g(\ell). \tag{11}$$

Proof. See Section S2. □

[6]Similar approaches have been suggested by Adams et al. (2018); Meyer et al. (2021). Our work is notably different in that it a) uses preconditioning, b) also considers the backward pass and c) gives stronger theoretical guarantees.

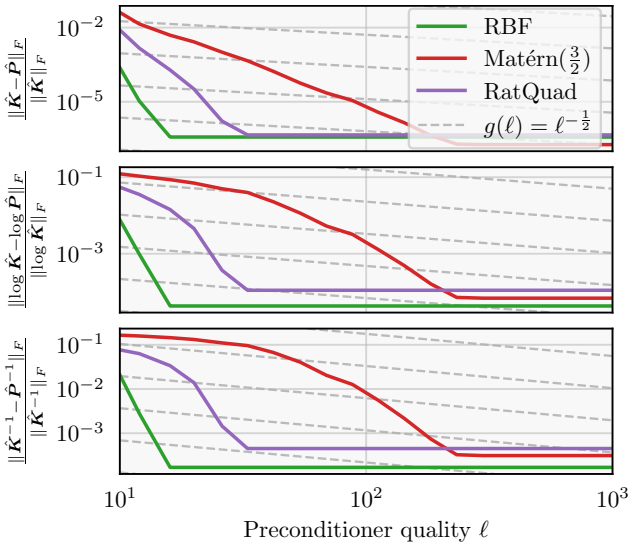


Figure 2: *Relative error of matrix functions.* For analytic functions $f(\hat{\boldsymbol{P}}_\ell) \to f(\hat{\boldsymbol{K}})$ at the asymptotic rate of the preconditioner $\hat{\boldsymbol{P}}_\ell \to \hat{\boldsymbol{K}}$. Here, we use an incomplete Cholesky preconditioner on a synthetic dataset ($n = 1{,}000$).

Notice that Theorem 1 assumes that the sequence $\{f(\hat{\boldsymbol{P}}_\ell)\}_\ell$ approximates $f(\hat{\boldsymbol{K}})$ sufficiently fast with ℓ in (9). Intuitively, if $\hat{\boldsymbol{P}}_\ell \to \hat{\boldsymbol{K}}$ quickly, one might expect the same for $f(\hat{\boldsymbol{P}}_\ell) \to f(\hat{\boldsymbol{K}})$ under certain conditions on f. Indeed, one obtains the same asymptotic rate $g(\ell)$ of the preconditioner $\hat{\boldsymbol{P}}_\ell$ for the approximation of $f(\hat{\boldsymbol{K}})$ by $f(\hat{\boldsymbol{P}})$ (see Proposition S1). This is illustrated in Figure 2. Therefore, the error of the variance-reduced stochastic trace estimate is *determined by the quality $g(\ell)$ of the preconditioner*.

Comparison of Theorem 1 and Existing Results Consider the case where $f = \mathrm{id}$. If we are not using a preconditioner, i.e. $\hat{\boldsymbol{P}} = \boldsymbol{0}$ and thus $c_{\boldsymbol{\Delta}} = g(\ell) = 1$, we recover the well-known convergence rate $\mathcal{O}(\ell^{-\frac{1}{2}})$ of Hutchinson's estimator (Avron & Toledo, 2011; Roosta-Khorasani & Ascher, 2015). If instead we choose a randomized low rank approximation as a preconditioner with $g(\ell) = \ell^{-\frac{1}{2}}$, then Theorem 1 recovers the convergence rate $\varepsilon_{\mathrm{STE}} \in \mathcal{O}(\ell^{-1})$ of Hutch++ (Meyer et al., 2021; Persson et al., 2021; Jiang et al., 2021) as a special case. However, as we will show, using preconditioning one can achieve polynomial – even exponential – convergence rates for common kernels. Such a drastic improvement is possible since neither variants of Hutchinson's make any assumptions about the kernel matrix, whereas preconditioners are designed to leverage structure.

3.2. Forward Pass

We can now analyze the error of the preconditioned stochastic log-determinant estimate. Combining Theorem 1 with Lanczos quadrature error analysis, the following holds.

Theorem 2 (Error Bound for $\log\det(\hat{\boldsymbol{K}})$)
Let $f = \log$, $\boldsymbol{\Delta}_{\log} = \log\left(\hat{\boldsymbol{P}}^{-\frac{1}{2}}\hat{\boldsymbol{K}}\hat{\boldsymbol{P}}^{-\intercal\frac{1}{2}}\right)$ and assume the conditions of Theorem 1 hold. Then, with probability $1-\delta$, it holds for $\tau_*^{\log} = \log(\det(\hat{\boldsymbol{P}})) + \tau_{\ell,m}^{\mathrm{SLQ}}(\boldsymbol{\Delta}_{\log})$, that

$$\left|\tau_*^{\log} - \log\det(\hat{\boldsymbol{K}})\right| \leq (\varepsilon_{\mathrm{Lanczos}} + \varepsilon_{\mathrm{STE}})\|\log(\hat{\boldsymbol{K}})\|_F,$$

where the individual errors are bounded by

$$\varepsilon_{\mathrm{Lanczos}}(\kappa, m) \leq K_1 \left(\tfrac{\sqrt{2\kappa+1}-1}{\sqrt{2\kappa+1}+1}\right)^{2m} \tag{12}$$

$$\varepsilon_{\mathrm{STE}}(\delta, \ell) \leq C_1\sqrt{\log(\delta^{-1})}\ell^{-\frac{1}{2}}g(\ell) \tag{13}$$

and $K_1 = \frac{5\kappa\log(2(\kappa+1))}{2\|\log(\hat{\boldsymbol{K}})\|_F\sqrt{2\kappa+1}}$.

Proof. See Section S3.2. □

Corollary 1
Assume the conditions of Theorem 2 hold. If the number of random vectors ℓ satisfies (11) *with $\varepsilon_{\mathrm{STE}} = \frac{\varepsilon}{2}$ and we run*

$$m \geq \tfrac{\sqrt{3}}{4}\sqrt{\kappa}\log\left(2K_1\varepsilon^{-1}\right) \tag{14}$$

iterations of Lanczos, then it holds that

$$\boxed{\mathbb{P}\left(\left|\tau_*^{\log} - \log\det(\hat{\boldsymbol{K}})\right| \leq \varepsilon\|\log(\hat{\boldsymbol{K}})\|_F\right) \geq 1-\delta.}$$

Proof. See Section S3.2. □

We note two major improvements over the bound by Ubaru et al. (2017, Corollary 4.5). First, the number of Lanczos steps now depends on the condition number κ of the *preconditioned* matrix, implying faster convergence. Second, depending on the preconditioner quality $g(\ell)$, we need significantly fewer random vectors by Theorem 1.

3.3. Backward Pass

By differentiating through (8), we obtain a decomposition into a *deterministic approximation* based on the preconditioner and a *residual trace* for the backward pass. For $\boldsymbol{\Delta}_{\mathrm{inv}\partial} = \hat{\boldsymbol{K}}^{-1}\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta} - \hat{\boldsymbol{P}}^{-1}\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}$, we have

$$\tfrac{\partial}{\partial\theta}\log\det(\hat{\boldsymbol{K}}) = \mathrm{tr}\left(\hat{\boldsymbol{P}}^{-1}\tfrac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\right) + \mathrm{tr}(\boldsymbol{\Delta}_{\mathrm{inv}\partial}), \tag{15}$$

Therefore the stochastic trace estimator

$$\tau_{\ell,m}^{\mathrm{SCG}}(\boldsymbol{\Delta}_{\mathrm{inv}\partial}) \approx \mathrm{tr}(\boldsymbol{\Delta}_{\mathrm{inv}\partial}) \tag{16}$$

requires solves $\boldsymbol{z}_i^\intercal\hat{\boldsymbol{K}}^{-1}\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\boldsymbol{z}_i$ and $\boldsymbol{z}_i^\intercal\hat{\boldsymbol{P}}^{-1}\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\boldsymbol{z}_i$. The former can be computed with m iterations of preconditioned CG, while the latter is simply a solve with the preconditioner. Note that the deterministic term $\mathrm{tr}\left(\hat{\boldsymbol{P}}^{-1}\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\right)$ is efficient to calculate for many types of preconditioners. For example, if $\hat{\boldsymbol{P}}$ is a diagonal-plus-low-rank preconditioner it can be computed in $\mathcal{O}(n\ell^2)$ (see Section S3.3). Using Theorem 1, we obtain a probabilistic error bound for the derivative estimate.

Theorem 3 (Error Bound for $\operatorname{tr}\big(\hat{\boldsymbol{K}}^{-1}\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\big)$)
Let $f(\hat{\boldsymbol{K}}) = \hat{\boldsymbol{K}}^{-1}\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}$, $\boldsymbol{\Delta}_{\mathrm{inv}\partial} = \hat{\boldsymbol{K}}^{-1}\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta} - \hat{\boldsymbol{P}}^{-1}\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}$ and assume the conditions of Theorem 1 hold. If we solve $\hat{\boldsymbol{K}}^{-1}\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\boldsymbol{z}_i$ with m iterations of preconditioned CG, initialized at $\mathbf{0}$ or better, then it holds with probability $1-\delta$ for $\tau_*^{\mathrm{inv}\partial} = \operatorname{tr}\big(\hat{\boldsymbol{P}}^{-1}\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\big) + \tau_{\ell,m}^{\mathrm{SCG}}(\boldsymbol{\Delta}_{\mathrm{inv}\partial})$, that

$$\Big|\tau_*^{\mathrm{inv}\partial} - \operatorname{tr}\big(\hat{\boldsymbol{K}}^{-1}\tfrac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\big)\Big| \leq (\varepsilon_{\mathrm{CG}'} + \varepsilon_{\mathrm{STE}})\Big\|\boldsymbol{K}^{-1}\tfrac{\partial\boldsymbol{K}}{\partial\theta}\Big\|_F,$$

where the individual errors are bounded by

$$\varepsilon_{\mathrm{CG}'}(\kappa, m) \leq K_2\left(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \tag{17}$$

$$\varepsilon_{\mathrm{STE}}(\delta, \ell) \leq C_1\sqrt{\log(\delta^{-1})}\ell^{-\frac{1}{2}}g(\ell) \tag{18}$$

and $K_2 = 2\sqrt{\kappa(\hat{\boldsymbol{K}})}n$.

Proof. See Section S3.3. □

Corollary 2
Assume the conditions of Theorem 3 hold. If the number of random vectors ℓ satisfies (11) *with $\varepsilon_{\mathrm{STE}} = \frac{\varepsilon}{2}$, and we run*

$$m \geq \tfrac{1}{2}\sqrt{\kappa}\log\big(2K_2\varepsilon^{-1}\big) \tag{19}$$

iterations of CG, then

$$\boxed{\mathbb{P}\left(\Big|\tau_*^{\mathrm{inv}\partial} - \operatorname{tr}\big(\hat{\boldsymbol{K}}^{-1}\tfrac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\big)\Big| \leq \varepsilon\Big\|\hat{\boldsymbol{K}}^{-1}\tfrac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\Big\|_F\right) \geq 1-\delta.}$$

Proof. See Section S3.3. □

4. Efficient GP Hyperparameter Optimization

Having established an efficient way to compute the forward and backward pass for the log-determinant, we can use these results to accelerate GP hyperparameter optimization by fully exploiting preconditioning not just for the linear solves, but *also* for the log-determinant and its derivative.

4.1. Log-Marginal Likelihood

We obtain a bound on the log-marginal likelihood by combining Theorem 2 with standard CG convergence analysis.

Theorem 4 (Error Bound for the log-Marginal Likelihood)
Assume the conditions of Theorem 2 hold and we solve $\hat{\boldsymbol{K}}\boldsymbol{u} = \boldsymbol{y}$ via preconditioned CG initialized at $\boldsymbol{u}_0$ and terminated after m iterations. Then with probability $1-\delta$, the error in the estimate $\eta = -\frac{1}{2}(\boldsymbol{y}^\intercal\boldsymbol{u}_m + \tau_*^{\log} + n\log(2\pi))$ of the log-marginal likelihood $\mathcal{L}$ satisfies

$$|\eta - \mathcal{L}| \leq \varepsilon_{\mathrm{CG}} + \tfrac{1}{2}(\varepsilon_{\mathrm{Lanczos}} + \varepsilon_{\mathrm{STE}})\|\log(\hat{\boldsymbol{K}})\|_F,$$

where the individual errors are bounded by

$$\varepsilon_{\mathrm{CG}}(\kappa, m) \leq K_3\left(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \tag{20}$$

$$\varepsilon_{\mathrm{Lanczos}}(\kappa, m) \leq K_1\left(\tfrac{\sqrt{2\kappa+1}-1}{\sqrt{2\kappa+1}+1}\right)^{2m} \tag{21}$$

$$\varepsilon_{\mathrm{STE}}(\delta, \ell) \leq C_1\sqrt{\log(\delta^{-1})}\ell^{-\frac{1}{2}}g(\ell) \tag{22}$$

for $K_3 = \sqrt{\kappa(\hat{\boldsymbol{K}})}\|\boldsymbol{y}\|_2\|\boldsymbol{u}_0 - \boldsymbol{u}\|_2$.

Proof. See Section S4.1. □

4.2. Derivative of the Log-Marginal Likelihood

Similarly, we can leverage Theorem 3 for the derivative.

Theorem 5 (Error Bound for the Derivative)
Assume the conditions of Theorem 3 hold and we solve $\hat{\boldsymbol{K}}\boldsymbol{u} = \boldsymbol{y}$ via preconditioned CG initialized at $\mathbf{0}$ or better and terminated after m iterations. Then with probability $1-\delta$, the error in the estimate $\phi = \frac{1}{2}(\boldsymbol{u}_m^\intercal\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\boldsymbol{u}_m - \tau_*^{\mathrm{inv}\partial})$ of the derivative of the log-marginal likelihood $\frac{\partial}{\partial\theta}\mathcal{L}$ satisfies

$$\Big|\phi - \tfrac{\partial}{\partial\theta}\mathcal{L}\Big| \leq \varepsilon_{\mathrm{CG}}(\kappa, m) + \tfrac{1}{2}(\varepsilon_{\mathrm{CG}'} + \varepsilon_{\mathrm{STE}})\Big\|\hat{\boldsymbol{K}}^{-1}\tfrac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\Big\|_F,$$

where the individual errors are bounded by

$$\varepsilon_{\mathrm{CG}}(\kappa, m) \leq K_4\left(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \tag{23}$$

$$\varepsilon_{\mathrm{CG}'}(\kappa, m) \leq K_2\left(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \tag{24}$$

$$\varepsilon_{\mathrm{STE}}(\delta, \ell) \leq C_1\sqrt{\log(\delta^{-1})}\ell^{-\frac{1}{2}}g(\ell) \tag{25}$$

for $K_4 = 6\kappa(\hat{\boldsymbol{K}})\max(\|\boldsymbol{u}\|_2, \|\boldsymbol{u}\|_2^3)\big\|\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\big\|_2$.

Proof. See Section S4.2. □

4.3. Preconditioner Choice

Our theoretical convergence results fundamentally depend on how quickly the preconditioner approximates the kernel matrix, either directly via $g(\ell)$, or indirectly via the condition number improvement (7). This leaves the question which preconditioners should be chosen in practice and what rates $g(\ell)$ they attain. In Table 1, we give an extensive list of kernel-preconditioner combinations with associated rates (see Section S5 for proofs). This includes the commonly used RBF and Matérn(ν) kernels for which the incomplete Cholesky (Kershaw, 1978) and QFF (Mutnỳ & Krause, 2018) preconditioners result in exponential and polynomial convergence rates, respectively. For STE in this context this is a substantial improvement over the rate of Hutchinson's estimator $\mathcal{O}(\ell^{-\frac{1}{2}})$ (Avron & Toledo, 2011; Roosta-Khorasani & Ascher, 2015; Skorski, 2021) and Hutch++ with $\mathcal{O}(\ell^{-1})$ (Meyer et al., 2021; Persson et al., 2021; Jiang et al., 2021). Depending on the problem this can mean a difference of tens vs. thousands of random vectors. To the best of our knowledge, for the use of CG in GP inference, only the one-dimensional RBF kernel and incomplete Cholesky preconditioner have been previously analyzed theoretically (Gardner et al., 2018). In contrast, Table 1 gives convergence rates for arbitrary d-dimensional kernels and multiple preconditioners. In fact, our results also apply to any kernel

Table 1: *Error rates for combinations of kernels and preconditioners.* The rate $g(\ell)$ measures how fast a sequence of preconditioners $\{\hat{\boldsymbol{P}}_\ell\}_\ell$ approaches the kernel matrix $\hat{\boldsymbol{K}}$ constructed from data $\boldsymbol{X} \in \mathbb{R}^{n\times d}$. Thus it determines both the convergence speed of Krylov methods and the preconditioned stochastic trace estimator. Or, equivalently, the faster $g(\ell) \to 0$ the fewer CG iterations m and random vectors ℓ are needed to approximate the log-marginal likelihood and its gradient.

Kernel	d	Preconditioner	$g(\ell)$	Condition	Proof
RBF	1	incomplete Cholesky	$\exp(-c\ell)$	for some $c > 0$	Proposition S2
RBF	$\mathbb{N}$	QFF	$\exp(-b\ell^{\frac{1}{d}})$	for some $b > 0$ if $\ell^{\frac{1}{d}} > 2\gamma^{-2}$	Proposition S3
Matérn(ν)	$\mathbb{N}$	incomplete Cholesky	$\ell^{-(\frac{2\nu}{d}+1)}$	$2\nu \in \mathbb{N}$ and maximin ordering	(Schäfer et al., 2021)
Matérn(ν)	1	QFF	$\ell^{-(s(\nu)+1)}$	where $s(\nu) \in \mathbb{N}$	Proposition S4
mod. Matérn(ν)	$\mathbb{N}$	QFF	$\ell^{-\frac{s(\nu)+1}{d}}$	where $s(\nu) \in \mathbb{N}$	Proposition S4
any	$\mathbb{N}$	truncated SVD	$\ell^{-\frac{1}{2}}$		Proposition S5
any	$\mathbb{N}$	randomized SVD	$\ell^{-\frac{1}{2}} + \mathcal{O}(\ell^{\frac{1}{4}} s^{-\frac{1}{4}})$	w/ high probability for s samples	Proposition S6
any	$\mathbb{N}$	randomized Nyström	$\ell^{-\frac{1}{2}} + \mathcal{O}(\ell^{\frac{1}{4}} s^{-\frac{1}{4}})$	w/ high probability for s samples	Proposition S7
any	$\mathbb{N}$	RFF	$\ell^{-\frac{1}{2}}$	w/ high probability	Proposition S8
additive	$\mathbb{N}$	any	$dg(\ell)$	all summands have rate $g(\ell)$	Lemma S5
any	$\mathbb{N}$	any approximation	$g(\ell)$	$\exists$ uniform convergence bound	Lemma S6

approximation with a uniform convergence bound (such as RFF (Rahimi et al., 2007)). All the while for many, e.g. diagonal-plus-low-rank preconditioners, the amount of precomputation needed *amortizes with more data*, i.e. the cost of preconditioning becomes negligible the larger the dataset.

4.4. Algorithms

The above leads to Algorithms 1 and 2 computing $\mathcal{L}$ and $\frac{\partial}{\partial\theta}\mathcal{L}$ for GP hyperparameter optimization.[7] Our algorithms are similar to those presented in prior work by Cutajar et al. (2016); Ubaru et al. (2017); Gardner et al. (2018), yet crucially they leverage preconditioning for faster CG convergence *and* variance reduction of the log-determinant estimate and its derivative. In the following, $\mathrm{CG}(\hat{\boldsymbol{K}}, \boldsymbol{y}, \hat{\boldsymbol{P}}, m)$ denotes a CG solve of $\hat{\boldsymbol{K}}\boldsymbol{u} = \boldsymbol{y}$ with preconditioner $\hat{\boldsymbol{P}}$ run for m iterations. Here, we equivalently use CG instead of Lanczos, as suggested by Gardner et al. (2018).

Computational Complexity Algorithm 1 has complexity $\mathcal{O}(n^2 m\ell + P_{\log\det})$ and Algorithm 2 has complexity $\mathcal{O}((n^2 m + P_{\mathrm{solve}})\ell + P_{\mathrm{tr\,inv}\partial})$, where $P_{(\cdot)}$ denotes the cost of an operation with the preconditioner.[8] Assuming $m, \ell \ll n$, this is asymptotically faster than Cholesky-based inference with complexity $\mathcal{O}(n^3)$. Due to the reduction to matrix-vector multiplication, if $\boldsymbol{v} \mapsto \hat{\boldsymbol{K}}\boldsymbol{v}$ is more efficient than $\mathcal{O}(n^2)$ (e.g. for structured or sparse matrices) the complexity reduces further. Finally, the **for**-loops are embarrassingly parallel, giving additional speedup in practice.

[7] While presented sequentially for clarity, in practice one would presample all random vectors and run a single call of (parallelized) CG with multiple right hand sides, as in (Gardner et al., 2018).

[8] For diagonal-plus-low-rank preconditioners, such as the incomplete Cholesky, P_{solve}, $P_{\log\det}$, and $P_{\mathrm{tr\,inv}\partial}$ are in $\mathcal{O}(n\ell^2)$ by the matrix inversion and determinant lemmas.

Algorithm 1: log-Marginal Likelihood

Input: $\boldsymbol{y}$ (labels), $\hat{\boldsymbol{K}}$ (kernel matrix), $\hat{\boldsymbol{P}}$ (preconditioner), ℓ (# of random STE vectors), m (# of CG iterations)

procedure LogMargLikelihood($\boldsymbol{y}, \hat{\boldsymbol{K}}, \hat{\boldsymbol{P}}, \ell, m$)
  $\boldsymbol{u} \leftarrow \mathrm{CG}(\hat{\boldsymbol{K}}, \boldsymbol{y}, \hat{\boldsymbol{P}}, m)$ ▷ $\approx \hat{\boldsymbol{K}}^{-1}\boldsymbol{y}$
  $\tau^{\log}_{\hat{\boldsymbol{P}}} \leftarrow \log\det(\hat{\boldsymbol{P}})$
  for $i = 1, \ldots, \ell$ **do**
    $\boldsymbol{z}_i \leftarrow \tilde{\boldsymbol{z}}_i / \|\tilde{\boldsymbol{z}}_i\|_2$ for rand. vector $\tilde{\boldsymbol{z}}_i$
    $\boldsymbol{T} \leftarrow \mathrm{CG}(\hat{\boldsymbol{K}}, \boldsymbol{z}_i, \hat{\boldsymbol{P}}, m)$ ▷ equiv. to Lanczos
    $[\boldsymbol{W}, \boldsymbol{\lambda}] \leftarrow \mathrm{EigenDecomp}(\boldsymbol{T})$ ▷ $\boldsymbol{T}$ tridiagonal
    $\omega_j \leftarrow (\boldsymbol{e}_1^\intercal \boldsymbol{w}_j)^2$ for $j = 0, \ldots, m$ ▷ quad. weights
    $\gamma_i \leftarrow \sum_{j=0}^{m} \omega_j \log(\lambda_j)$ ▷ $\approx \boldsymbol{z}_i^\intercal \boldsymbol{\Delta}_{\log} \boldsymbol{z}_i$
  $\tau^{\log}_* \leftarrow \tau^{\log}_{\hat{\boldsymbol{P}}} + \frac{n}{\ell}\sum_{i=1}^{\ell}\gamma_i$ ▷ $\approx \log\det(\hat{\boldsymbol{K}})$
  return $-\frac{1}{2}(\boldsymbol{y}^\intercal\boldsymbol{u} + \tau^{\log}_* + n\log(2\pi))$ ▷ $\approx \mathcal{L}(\boldsymbol{\theta})$

Algorithm 2: Derivative of the log-Marginal Likelihood

Input: $\boldsymbol{y}$ (labels), $\hat{\boldsymbol{K}}$ (kernel matrix), $\hat{\boldsymbol{P}}$ (preconditioner), ℓ (# of random STE vectors), m (# of CG iterations), $\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}$ / $\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}$ (functions for computing kernel / preconditioner derivatives)

procedure Derivative($\boldsymbol{y}, \hat{\boldsymbol{K}}, \frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}, \hat{\boldsymbol{P}}, \frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}, \ell, m$)
  $\boldsymbol{u} \leftarrow \mathrm{CG}(\hat{\boldsymbol{K}}, \boldsymbol{y}, \hat{\boldsymbol{P}}, m)$ ▷ $\approx \hat{\boldsymbol{K}}^{-1}\boldsymbol{y}$
  $\tau^{\mathrm{inv}\partial}_{\hat{\boldsymbol{P}}} \leftarrow \mathrm{tr}\big(\hat{\boldsymbol{P}}^{-1}\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\big)$
  for $i = 1, \ldots, \ell$ **do**
    $\boldsymbol{z}_i \leftarrow \tilde{\boldsymbol{z}}_i / \|\tilde{\boldsymbol{z}}_i\|_2$ for rand. vector $\tilde{\boldsymbol{z}}_i$
    $\boldsymbol{w}_i \leftarrow \mathrm{CG}(\hat{\boldsymbol{K}}, \frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\boldsymbol{z}_i, \hat{\boldsymbol{P}}, m)$ ▷ $\approx \hat{\boldsymbol{K}}^{-1}\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\boldsymbol{z}_i$
    $\tilde{\boldsymbol{w}}_i \leftarrow \hat{\boldsymbol{P}}^{-1}\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\boldsymbol{z}_i$
    $\gamma_i \leftarrow \boldsymbol{z}_i^\intercal(\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i)$ ▷ $\approx \boldsymbol{z}_i^\intercal\boldsymbol{\Delta}_{\mathrm{inv}\partial}\boldsymbol{z}_i$
  $\tau^{\mathrm{inv}\partial}_* \leftarrow \tau^{\mathrm{inv}\partial}_{\hat{\boldsymbol{P}}} + \frac{n}{\ell}\sum_{i=1}^{\ell}\gamma_i$ ▷ $\approx \mathrm{tr}\big(\boldsymbol{K}^{-1}\frac{\partial\boldsymbol{K}}{\partial\theta}\big)$.
  return $\frac{1}{2}(\boldsymbol{u}^\intercal\frac{\partial\boldsymbol{K}}{\partial\theta}\boldsymbol{u} - \tau^{\mathrm{inv}\partial}_*)$ ▷ $\approx \frac{\partial}{\partial\theta}\mathcal{L}(\boldsymbol{\theta})$

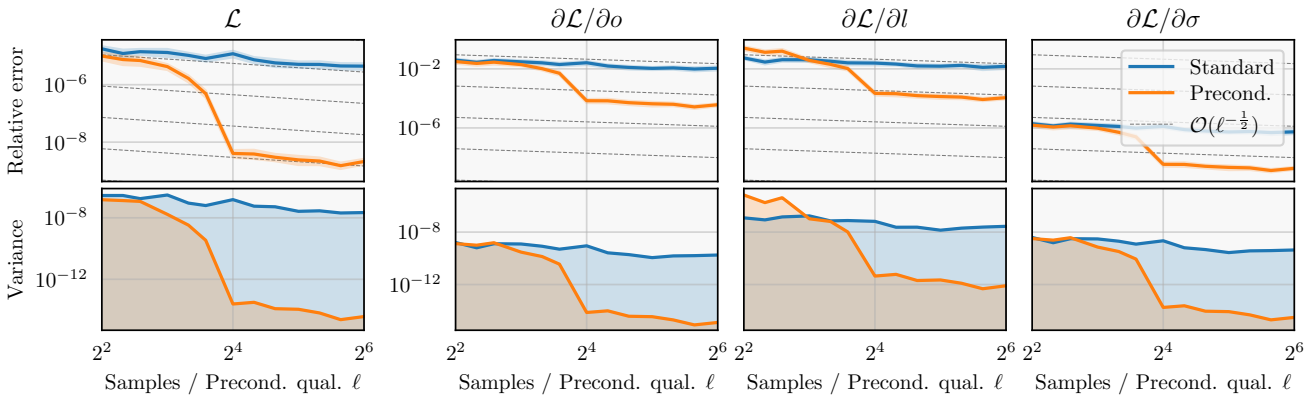


Figure 3: *Bias and variance of the estimators for the* log*-marginal likelihood* $\mathcal{L}$ *and its derivatives.* The relative error and variance decrease faster with the number of random vectors ℓ when using a preconditioner $\hat{\boldsymbol{P}}_\ell$. The decrease rate $\mathcal{O}(\ell^{-\frac{1}{2}}g(\ell))$, determined by the preconditioner, significantly improves upon the standard Hutchinson's rate $\mathcal{O}(\ell^{-\frac{1}{2}})$.

4.5. Related Work

Krylov methods have been used for GP inference since the work of Gibbs (1997). While these methods were primarily relegated to structured GPs that afford fast matrix-vector products (Cunningham et al., 2008; Saatçi, 2012; Wilson & Nickisch, 2015), they have seen growing use as a general purpose method, especially when coupled with parallel hardware (Murray, 2009; Cutajar et al., 2016; Gardner et al., 2018; Wang et al., 2019; Artemev et al., 2021). Some recent works propose variance-free (but biased) estimates of the GP log-determinant (e.g. Artemev et al., 2021), though many works compute this term by combining STE (Hutchinson, 1989; Bekas et al., 2007; Avron & Toledo, 2011; Roosta-Khorasani & Ascher, 2015) with SLQ (Golub & Meurant, 2009; Ubaru et al., 2017; Dong et al., 2017; Gardner et al., 2018; Cortinovis & Kressner, 2020). Our work builds on ideas for variance-reduced STE (Adams et al., 2018; Meyer et al., 2021; Persson et al., 2021; Jiang et al., 2021), but, by leveraging preconditioning, requires significantly fewer random vectors than existing approaches. When applied to GP hyperparameter optimization, we obtain stronger theoretical guarantees for the forward pass than previously known (Ubaru et al., 2017) and novel guarantees for the backward pass. Finally, our results on preconditioners for kernel matrices (Table 1) give a rigorous foundation to their proposed use by Cutajar et al. (2016).

5. Experiments

We validate our theoretical findings empirically via GP hyperparameter optimization on synthetic and benchmark datasets with and without preconditioning. We find that

(a) *preconditioning reduces bias and variance* in the forward and backward pass, which results in
(b) *less noisy search directions* and *fewer* log*-likelihood and gradient evaluations for the line search.*

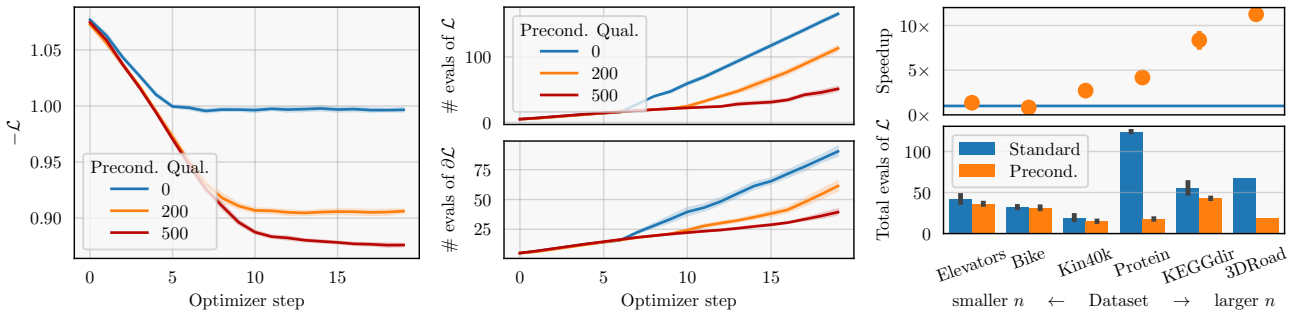
This allows the use of rapidly converging optimizers, and

(c) *accelerates training significantly.*.

Experimental Setup We consider a one-dimensional synthetic dataset of $n = 10{,}000$ iid standard normal samples, as well as a range of UCI datasets (Dua & Graff, 2017) with training set sizes ranging from $n = 12{,}449$ to $326{,}155$ (see Table 2). All experiments were performed on single NVIDIA GPUs, a GeForce RTX 2080 and Titan RTX, respectively. We perform GP regression using an RBF and Matérn$(\frac{3}{2})$ kernel with output scale o, lengthscales l_j – one per input dimension – and noise σ^2. Hyperparameters were optimized with L-BFGS using an Armijo-Wolfe line search and early stopping via a validation set. We use an incomplete Cholesky preconditioner throughout. An implementation of our method is available as part of GPyTorch (Gardner et al., 2018).[9]

Preconditioning reduces bias & variance in $\mathcal{L}$ and $\frac{\partial}{\partial\theta}\mathcal{L}$ Figure 3 shows the relative error of the marginal log-likelihood and its derivatives on synthetic data. Already for $\ell \geq 16$ random samples *bias and variance are reduced by several orders of magnitude*. We observe exponential decrease and then a return to the standard Hutchinson's rate of $\mathcal{O}(\ell^{-\frac{1}{2}})$. After $\ell = 16$ iterations the algorithm computing the preconditioner has reached a specified tolerance and terminates, invalidating the approximation quality assumption (6) for $\ell \geq 16$. Similar observations hold for the Matérn and RatQuad kernel (see Table S1 and Figure S1). As predicted by Theorem 1 and illustrated by Figure 3, the variance reduction is determined by the preconditioner. For higher dimensions, the rate $g(\ell)$ generally slows (see Table 1), which in turn reduces the bias and variance reduction achieved by

[9] github.com/cornellius-gp/gpytorch



(a) Training loss for "Protein". (b) Line search computations for "Protein". (c) Speedup on UCI datasets.

Figure 4: *Preconditioning reduces noise and in turn accelerates hyperparameter optimization.* Variance reduction improves optimization via better search directions and fewer evaluations of $\mathcal{L}$ and $\frac{\partial}{\partial\theta_i}\mathcal{L}$ for the line search. (a) Training loss and (b) model evaluations for line search decrease with preconditioner size, as shown for the "Protein" dataset. (c) The reduction in loss (and gradient) evaluations and of noise in the gradients results in an order of magnitude speedup on UCI datasets.

Table 2: *Hyperparameter optimization on UCI datasets.* GP regression using a Matérn$(\frac{3}{2})$ kernel and incomplete Cholesky preconditioner of size 500 with $\ell = 50$ random samples. Hyperparameters were optimized with L-BFGS for at most 20 steps using early stopping. All results, but "3DRoad", are averaged over 10 runs. Differences ≥ 1 standard deviation in bold.

Dataset	n	d	$-\mathcal{L}_{\text{train}} \downarrow$		$-\mathcal{L}_{\text{test}} \downarrow$		RMSE $\downarrow$		Runtime (s)	
			Standard	Precond.	Standard	Precond.	Standard	Precond.	Standard	Precond.
Elevators	12,449	18	0.4647	**0.4377**	0.4021	0.4022	0.3484	0.3482	53	**39**
Bike	13,034	17	–0.9976	–0.9985	–0.9934	–0.9877	0.0446	0.0454	**31**	37
Kin40k	30,000	8	–0.3339	**–0.4332**	–0.3141	–0.3135	**0.0929**	0.0949	187	**45**
Protein	34,297	9	0.9963	**0.9273**	0.8869	0.8835	0.5722	**0.5577**	893	**43**
KEGGdir	36,620	20	–0.9501	**–1.0043**	–0.9459	–0.9490	0.0861	0.0864	1450	**174**
3DRoad	326,155	3	0.7733	**0.1284**	1.4360	**1.1690**	0.2982	**0.1265**	82,200	**7306**

our method (see Table S1). However, on real datasets we still see strong variance reduction via our method, possibly since real data often lies on a low-dimensional manifold.

Preconditioning accelerates hyperparameter optimization On datasets from the UCI repository, we find that preconditioning results in lower training loss $-\mathcal{L}(\boldsymbol{\theta})$ (illustrated in Figure 4(a)) on almost all datasets and essentially identical generalization error (see Table 2). Reducing stochasticity via preconditioning significantly lowers the number of $\mathcal{L}$ and $\frac{\partial}{\partial\theta}\mathcal{L}$ evaluations for the line search during optimization (see Figure 4(b)) and results in less noisy search directions. In fact, the noise in the loss and gradients caused by stochastic trace estimation previously necessitated the use of slower converging, but more noise-robust optimizers (Wang et al., 2019), such as Adam (Kingma & Ba, 2015). As these experiments show, our variance-reduced estimators make the use of L-BFGS possible, which significantly outperforms Adam (c.f. Table 2 and Table S2). These combined effects due to preconditioning *accelerate training up to twelvefold*, as Figure 4(c) shows. We observe that the speedup increases with the size of the dataset. This is partly explained by the amortization of the cost of computing and applying the preconditioner with increasing n.

6. Conclusion

One might reasonably hope that structural knowledge about the kernel matrix can accelerate GP hyperparameter optimization. Preconditioning is a way to encode and exploit such structure. As we showed, it can be used to great effect – not only for the solution of linear systems – but importantly *also* for stochastic approximation of the log-determinant and its derivative. Our convergence results combined with the rates for kernel-preconditioner pairs in Table 1 rigorously explain why preconditioning has been observed empirically to be so effective for large-scale GP inference (Cutajar et al., 2016; Gardner et al., 2018; Wang et al., 2019).

In fact, our work implies that software packages for GPs, which make use of Krylov methods for inference, should not use a fixed preconditioner. Instead, the preconditioner should be automatically chosen based on the specified model. While we derive a range of such kernel and preconditioner combinations, it is likely that better preconditioners exist for certain kernels or types of data. Other scientific fields invest substantial research effort into the design of preconditioners, e.g. for PDEs (Saad, 2003). Our work strongly suggests that, similarly, developing specialized preconditioners is a promising approach to scale Gaussian processes.

Acknowledgements

JW and PH gratefully acknowledge financial support by the European Research Council through ERC StG Action 757275 / PANAMA; the DFG Cluster of Excellence "Machine Learning - New Perspectives for Science", EXC 2064/1, project number 390727645; the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A); and funds from the Ministry of Science, Research and Arts of the State of Baden-Württemberg. JW is grateful to the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support. GP and JPC are supported by the Simons Foundation, McKnight Foundation, the Grossman Center, and the Gatsby Charitable Trust.

The authors would like to thank Marius Hobbhahn, Lukas Tatzel and Felix Dangel for helpful feedback on an earlier version of this manuscript.

References

Adamczak, R. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:1 – 13, 2015.

Adams, R. P., Pennington, J., Johnson, M. J., Smith, J., Ovadia, Y., Patton, B., and Saunderson, J. Estimating the spectral density of large implicit matrices. *arXiv preprint*, 2018. URL http://arxiv.org/abs/1802.0345.

Artemev, A., Burt, D. R., and van der Wilk, M. Tighter bounds on the log marginal likelihood of Gaussian process regression using conjugate gradients. In *International Conference on Machine Learning (ICML)*, 2021.

Avron, H. and Toledo, S. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):1–34, 2011.

Bekas, C., Kokiopoulou, E., and Saad, Y. An estimator for the diagonal of a matrix. *Applied numerical mathematics*, 57(11-12):1214–1229, 2007.

Charlier, B., Feydy, J., Glaunès, J. A., Collin, F.-D., and Durif, G. Kernel operations on the GPU, with autodiff, without memory overflows. *Journal of Machine Learning Research (JMLR)*, 22 (74):1–6, 2021. URL http://jmlr.org/papers/v22/20-275.html.

Cortinovis, A. and Kressner, D. On randomized trace estimates for indefinite matrices with an application to determinants. *arXiv preprint*, 2020. URL http://arxiv.org/abs/2005.10009.

Cunningham, J. P., Shenoy, K. V., and Sahani, M. Fast Gaussian process methods for point process intensity estimation. In *International Conference on Machine Learning (ICML)*, 2008.

Cutajar, K., Osborne, M., Cunningham, J., and Filippone, M. Preconditioning kernel matrices. In *International Conference on Machine Learning (ICML)*, 2016.

Dong, K., Eriksson, D., Nickisch, H., Bindel, D., and Wilson, A. G. Scalable log determinants for Gaussian process kernel learning. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6330–6340, 2017.

Drineas, P., Mahoney, M. W., and Cristianini, N. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research (JMLR)*, 6(12), 2005.

Drineas, P., Kannan, R., and Mahoney, M. W. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing (SICOMP)*, 36:158–183, 2006.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., and Wilson, A. G. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018:7576–7586, 2018.

Gibbs, M. Bayesian Gaussian processes for classification and regression. *University of Cambridge, Cambridge*, 1997.

Golub, G. H. and Meurant, G. *Matrices, moments and quadrature with applications*, volume 30. Princeton University Press, 2009.

Golub, G. H. and van Loan, C. F. *Matrix Computations*. JHU Press, fourth edition, 2013.

Harbrecht, H., Peters, M., and Schneider, R. On the low-rank approximation by the pivoted Cholesky decomposition. *Applied numerical mathematics*, 62(4):428–440, 2012.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 282–290, 2013.

Hestenes, M. R. and Stiefel, E. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49, 1952.

Hutchinson, M. F. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Jiang, S., Pham, H., Woodruff, D. P., Qiuyi, and Zhang. Optimal sketching for trace estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kasiviswanathan, S. P. and Rudelson, M. Restricted isometry property under high correlations. *arXiv preprint*, 2019. URL http://arxiv.org/abs/1904.05510.

Kershaw, D. S. The incomplete Cholesky—conjugate gradient method for the iterative solution of systems of linear equations. *Journal of Computational Physics*, 26(1):43–65, 1978.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Kittaneh, F. On Lipschitz functions of normal operators. *Proceedings of the American Mathematical Society*, 94(3):416–418, 1985.

Lanczos, C. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Government Press Office Los Angeles, CA, 1950.

Ledoux, M. *The concentration of measure phenomenon*. Number 89. American Mathematical Society, 2001.

Meyer, R. A., Musco, C., Musco, C., and Woodruff, D. P. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, pp. 142–155. SIAM, 2021.

Mitrinovic, D. S. and Vasic, P. M. *Analytic inequalities*, volume 1. Springer, 1970.

Murray, I. Gaussian processes and fast matrix-vector multiplies. In *Numerical Mathematics in Machine Learning Workshop (ICML)*, 2009.

Mutnỳ, M. and Krause, A. Efficient high dimensional Bayesian optimization with additivity and quadrature Fourier features. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9005–9016, 2018.

Nocedal, J. and Wright, S. *Numerical optimization*. Springer Science & Business Media, 2006.

Persson, D., Cortinovis, A., and Kressner, D. Improved variants of the Hutch++ algorithm for trace estimation. *arXiv preprint*, 2021. URL http://arxiv.org/abs/2109.10659.

Pleiss, G., Gardner, J., Weinberger, K., and Wilson, A. G. Constant-time predictive distributions for Gaussian processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 4114–4123, 2018.

Potapczynski, A., Wu, L., Biderman, D., Pleiss, G., and Cunningham, J. P. Bias-free scalable Gaussian processes via randomized truncations. In *International Conference on Machine Learning (ICML)*, 2021.

Rahimi, A., Recht, B., et al. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Roosta-Khorasani, F. and Ascher, U. Improved bounds on sample size for implicit matrix trace estimators. *Foundations of Computational Mathematics*, 15(5):1187–1212, 2015.

Saad, Y. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics (SIAM), 2003.

Saatçi, Y. *Scalable inference for structured Gaussian process models*. PhD thesis, University of Cambridge, 2012.

Samson, P.-M. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *The Annals of Probability*, 28 (1):416 – 461, 2000.

Schäfer, F., Sullivan, T., and Owhadi, H. Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. *Multiscale Modeling and Simulation*, 19(2):688–730, 2021.

Skorski, M. Modern analysis of Hutchinson's trace estimator. In *55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–5. IEEE, 2021.

Sriperumbudur, B. K. and Szabó, Z. Optimal rates for random Fourier features. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 567–574. PMLR, 2009.

Trefethen, L. N. and Bau, D. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics (SIAM), 1997.

Ubaru, S., Chen, J., and Saad, Y. Fast estimation of $\mathrm{tr}(f(A))$ via stochastic Lanczos quadrature. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1075–1099, 2017.

Wang, K. A., Pleiss, G., Gardner, J. R., Tyree, S., Weinberger, K. Q., and Wilson, A. G. Exact Gaussian processes on a million data points. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Wilson, A. and Nickisch, H. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Wilson, A. G., Dann, C., and Nickisch, H. Thoughts on massively scalable Gaussian processes. *arXiv preprint*, 2015. URL http://arxiv.org/abs/1511.01870.

This supplementary material is structured as follows. Section S1 contains background on Krylov methods, such as known convergence results. Section S2 contains the main result and proof for variance-reduced stochastic trace estimation. Section S3 gives proofs for the forward and backward pass of the approximation to the log-determinant. In turn, Section S4 contains the error bounds for the log-marginal likelihood and its derivative. Error rates for specific preconditioners are given in Section S5 and finally, additional experimental results can be found in Section S7.

References referring to sections, equations or theorem-type environments within the supplement are prefixed with 'S', while references to, or results from, the main paper are stated as is.

S1. Background on Krylov Methods

S1.1. Conjugate Gradient Method

Theorem S6 (Convergence Rate of Preconditioned CG (Trefethen & Bau, 1997))
Let $\boldsymbol{A}, \boldsymbol{P} \in \mathbb{R}^{n \times n}$ be symmetric positive definite. The error of the conjugate gradient method with preconditioner $\boldsymbol{P}$ after $m \in \mathbb{N}$ steps is given by

$$\|\boldsymbol{x}_k - \boldsymbol{x}\|_{\boldsymbol{A}} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^m \|\boldsymbol{x}_0 - \boldsymbol{x}\|_{\boldsymbol{A}} \tag{S26}$$

and in euclidean norm by

$$\|\boldsymbol{x}_k - \boldsymbol{x}\|_2 \leq 2\sqrt{\kappa(\boldsymbol{A})} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^m \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 \tag{S27}$$

where $\kappa = \kappa(\boldsymbol{P}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{P}^{-\frac{\intercal}{2}})$ is the condition number of the preconditioned system matrix.

Proof. Preconditioned CG is equivalent to running CG on the transformed problem

$$\tilde{\boldsymbol{A}} \tilde{\boldsymbol{x}} = \boldsymbol{P}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{P}^{-\frac{\intercal}{2}} \tilde{\boldsymbol{x}} = \boldsymbol{P}^{-\frac{1}{2}} \boldsymbol{b}$$

with the substitution $\tilde{\boldsymbol{x}} = \boldsymbol{P}^{\intercal\frac{1}{2}}\boldsymbol{x}$. By Trefethen & Bau (1997), the convergence rate of CG on the problem is given by

$$\|\tilde{\boldsymbol{x}}_m - \tilde{\boldsymbol{x}}\|_{\tilde{\boldsymbol{A}}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \|\tilde{\boldsymbol{x}}_0 - \tilde{\boldsymbol{x}}\|_{\tilde{\boldsymbol{A}}}$$

The first equation follows by recognizing that

$$\|\tilde{\boldsymbol{x}}_m - \tilde{\boldsymbol{x}}\|^2_{\tilde{\boldsymbol{A}}} = (\tilde{\boldsymbol{x}}_m - \tilde{\boldsymbol{x}})^\intercal \boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{P}^{-\frac{\intercal}{2}}(\tilde{\boldsymbol{x}}_m - \tilde{\boldsymbol{x}}) = (\boldsymbol{x}_m - \boldsymbol{x})\boldsymbol{A}(\boldsymbol{x}_m - \boldsymbol{x}) = \|\boldsymbol{x}_m - \boldsymbol{x}\|^2_{\boldsymbol{A}}.$$

Now it holds by the min-max principle, that

$$\sqrt{\lambda_{\min}(\tilde{\boldsymbol{A}})}\|\boldsymbol{x}_k - \boldsymbol{x}\|_2 \leq \|\boldsymbol{x}_m - \boldsymbol{x}\|_{\boldsymbol{A}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \|\boldsymbol{x}_0 - \boldsymbol{x}\|_{\boldsymbol{A}} \leq 2\sqrt{\lambda_{\max}(\boldsymbol{A})}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2.$$

□

Corollary S3
Let $\varepsilon \in (0, 1]$, then preconditioned CG has relative error $\|\boldsymbol{x}_k - \boldsymbol{x}\|_{\boldsymbol{A}} \leq \varepsilon\|\boldsymbol{x}_0 - \boldsymbol{x}\|_{\boldsymbol{A}}$ after

$$m \geq \frac{\sqrt{\kappa}}{2}\log\left(2\varepsilon^{-1}\right) \tag{S28}$$

iterations, where κ is the condition number of the preconditioned system matrix. In euclidean norm $\|\cdot\|_2$ relative error ε is achieved after

$$m \geq \frac{\sqrt{\kappa}}{2}\log\left(2\sqrt{\kappa(\boldsymbol{A})}\varepsilon^{-1}\right) \tag{S29}$$

iterations.

Proof. It holds by Lemma S7 and the assumption on the number of iterations m, that

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \leq 2\exp\left(-\frac{2}{\sqrt{\kappa}}m\right) \leq 2\exp\left(-\log\left(\frac{2}{\varepsilon}\right)\right) = \varepsilon$$

Using Theorem S6 proves the statement. The proof for the euclidean norm is analogous. □

S1.2. Lanczos Algorithm

The *Lanczos algorithm* (Lanczos, 1950) is a Krylov method, which for a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ iteratively builds an approximate tridiagonalization

$$\boldsymbol{A} \approx \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{T}}\tilde{\boldsymbol{Q}}$$

where $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{n\times m}$ orthonormal and $\tilde{\boldsymbol{T}} \in \mathbb{R}^{m\times m}$ tridiagonal. For an initial probe vector $\boldsymbol{b} \in \mathbb{R}^n$, Gram-Schmidt orthogonalization is applied to the Krylov subspace basis. The orthogonalized vectors form $\tilde{\boldsymbol{Q}}$, while the Gram-Schmidt coefficients form $\tilde{\boldsymbol{T}}$. This low-rank approximation becomes an exact tridiagonalization $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^\intercal$ for $m = n$. The Lanczos process is often used to compute (approximate) eigenvalues and eigenvectors, which is done by computing an eigendecomposition of the tridiagonal matrix $\tilde{\boldsymbol{T}}$ at cost $\mathcal{O}(m^2)$. The tridiagonal matrix $\tilde{\boldsymbol{T}}$ can also be formed by running CG on the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and by collecting the step lengths α_i and conjugacy corrections β_i used in the solution and search direction updates (Saad, 2003, Section 6.7.3).

S1.3. Stochastic Lanczos Quadrature

One can approximate $\operatorname{tr}(f(\boldsymbol{A}))$ for symmetric positive definite $\boldsymbol{A}$ via *stochastic Lanczos quadrature* (SLQ) (Golub & Meurant, 2009; Ubaru et al., 2017) by combining Hutchinson's estimator with quadrature and the Lanczos algorithm. It holds that

$$\operatorname{tr}(f(\boldsymbol{A})) \approx \tau^{\mathrm{STE}}_\ell(f(\boldsymbol{A})) = \frac{n}{\ell}\sum_{i=1}^{\ell} \boldsymbol{z}_i^\intercal f(\boldsymbol{A})\boldsymbol{z}_i \approx \frac{n}{\ell}\sum_{i=1}^{\ell} I^{(i)}_m = \tau^{\mathrm{SLQ}}_{\ell,m}(f(\boldsymbol{A}))$$

The quadratic terms $\boldsymbol{z}_i^\intercal f(\boldsymbol{A})\boldsymbol{z}_i$ are approximated by quadrature $I^{(i)}_m$ where the weights and nodes of the quadrature rule are computed via m iterations of the Lanczos algorithm. For the log-determinant the following bound for the error incurred by Lanczos quadrature holds.

Corollary S4 (Section 4.3 of Ubaru et al. (2017))
Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be symmetric positive definite with condition number $\kappa = \kappa(\boldsymbol{A})$. Then it holds that

$$\left|\tau_\ell^{\mathrm{STE}}(\log(\boldsymbol{A})) - \tau_{\ell,m}^{\mathrm{SLQ}}(\log(\boldsymbol{A}))\right| \leq K\left(\frac{\sqrt{2\kappa+1}-1}{\sqrt{2\kappa+1}+1}\right)^{2m} \tag{S30}$$

where $K = \frac{5\kappa\log(2(\kappa+1))}{2\sqrt{2\kappa+1}}$.

S2. Stochastic Trace Estimation

Definition S1 (Convex Concentration Property (Ledoux, 2001))
Let $\boldsymbol{x} \in \mathbb{R}^n$ be a random vector. We say $\boldsymbol{x}$ has the *convex concentration property* (c.c.p.) with constant $K \in \mathbb{R}$ if for every 1-Lipschitz convex function $\phi : \mathbb{R}^n \to \mathbb{R}$, we have $\mathbb{E}[|\phi(\boldsymbol{x})|] < \infty$ and for every $t > 0$,

$$\mathbb{P}[|\phi(\boldsymbol{x}) - \mathbb{E}[\phi(\boldsymbol{x})]| \geq t] \leq 2\exp\left(-\frac{t^2}{K^2}\right).$$

Some common examples of random vectors having the c.c.p. are

- random vectors with independent and almost surely bounded entries $|\boldsymbol{x}_i| \leq 1$, where $K = 2\sqrt{2}$ (Samson, 2000);
- Gaussian random vectors $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $K^2 = 2\|\boldsymbol{\Sigma}\|_2$ (Kasiviswanathan & Rudelson, 2019); and
- random vectors which are uniformly distributed on the sphere $\sqrt{n}\mathbb{S}^{n-1}$, where $K = 2$ (Kasiviswanathan & Rudelson, 2019).

Remark S1
Note, that since Rademacher random vectors have iid entries $\{+1, -1\}$, they satisfy $\|\tilde{\boldsymbol{z}}_i\|_2 = \sqrt{n}$. In particular, it holds that $\sqrt{n}\boldsymbol{z}_i = \tilde{\boldsymbol{z}}_i$. Therefore the random vectors $\sqrt{n}\boldsymbol{z}_i$ and $\boldsymbol{z}' = \sqrt{n}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\ell)^\intercal \in \mathbb{R}^{\ell n}$ all have independent entries bounded by 1 and thus satisfy the convex concentration property with $K = 2\sqrt{2}$.

Theorem S7 (Hanson-Wright Inequality for Random Vectors with the Convex Concentration Property (Adamczak, 2015))
Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$ a zero-mean random vector with the convex concentration property with constant K. Then for all $t > 0$, it holds that

$$\mathbb{P}(|\boldsymbol{x}^\intercal\boldsymbol{A}\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}^\intercal\boldsymbol{A}\boldsymbol{x}]| \geq t) \leq 2\exp\left(-c\min\left(\frac{t^2}{\|\boldsymbol{A}\|_F^2}, \frac{t}{\|\boldsymbol{A}\|_2}\right)\right) \tag{S31}$$

where $c = c(K) > 0$ is a constant only dependent on the distribution of the random vectors.

Proof. By Theorem 2.5 of Adamczak (2015) we have

$$\mathbb{P}(|\boldsymbol{x}^\intercal\boldsymbol{A}\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}^\intercal\boldsymbol{A}\boldsymbol{x}]| > t) \leq 2\exp\left(-\frac{1}{c'}\min\left(\frac{t^2}{2K^4\|\boldsymbol{A}\|_F^2}, \frac{t}{K^2\|\boldsymbol{A}\|_2}\right)\right)$$

where $c' > 0$ is a universal constant. Now it holds that

$$\min\left(\frac{t^2}{2K^4\|\boldsymbol{A}\|_F^2}, \frac{t}{K^2\|\boldsymbol{A}\|_2}\right) \geq \frac{1}{K^2\max(2K^2, 1)}\min\left(\frac{t^2}{\|\boldsymbol{A}\|_F^2}, \frac{t}{\|\boldsymbol{A}\|_2}\right)$$

Choosing $c = \frac{1}{c'K^2\max(2K^2,1)}$ concludes the proof. □

Lemma S1
Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\ell \in \mathbb{N}$. Consider ℓ random vectors $\tilde{\boldsymbol{z}}_i \in \mathbb{R}^n$ with zero mean and unit covariance, such that for $\boldsymbol{z}_i = \tilde{\boldsymbol{z}}_i/\|\tilde{\boldsymbol{z}}_i\|_2$ the stacked random vector $\sqrt{n}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\ell)^\intercal \in \mathbb{R}^{\ell n}$ has the convex concentration property.[10] *Then there exists $c_{\boldsymbol{z}} > 0$ such that if $\ell \geq c_{\boldsymbol{z}}\log(\delta^{-1})$, then Hutchinson's trace estimator τ_ℓ^{STE} satisfies*

$$\mathbb{P}(\left|\tau_\ell^{\mathrm{STE}}(\boldsymbol{A}) - \mathrm{tr}(\boldsymbol{A})\right| \leq \sqrt{c_{\boldsymbol{z}}\log(\delta^{-1})\ell^{-1}}\|\boldsymbol{A}\|_F) \geq 1 - \delta.$$

[10] See Remark S1 for an explanation why this is satisfied for Rademacher random vectors.

Proof. Note that the proof strategy used here is the same as in Meyer et al. (2021, Lemma 2) with a different assumption on the distribution of the random vectors. To begin, define

$$\boldsymbol{A}' = \begin{pmatrix} \boldsymbol{A} & 0 & \dots & 0 \\ 0 & \boldsymbol{A} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \boldsymbol{A} \end{pmatrix} \in \mathbb{R}^{\ell n \times \ell n} \qquad \text{and} \qquad \boldsymbol{z}' = \sqrt{n} \begin{pmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \\ \boldsymbol{z}_\ell \end{pmatrix} \in \mathbb{R}^{\ell n}.$$

By assumption the random vector $\boldsymbol{z}'$ has the convex concentration property and therefore Theorem S7 holds. We obtain

$$\mathbb{P}(|(\boldsymbol{z}')^\intercal \boldsymbol{A}' \boldsymbol{z}' - \mathbb{E}[(\boldsymbol{z}')^\intercal \boldsymbol{A}' \boldsymbol{z}']| \geq t) \leq 2 \exp\left(-c \cdot \min\left(\frac{t^2}{\|\boldsymbol{A}'\|_F^2}, \frac{t}{\|\boldsymbol{A}'\|_2} \right) \right). \tag{S32}$$

Now, we have $(\boldsymbol{z}')^\intercal \boldsymbol{A}' \boldsymbol{z}' = n \sum_{i=1}^{\ell} \boldsymbol{z}_i^\intercal \boldsymbol{A} \boldsymbol{z}_i = \ell \tau_\ell^{\mathrm{STE}}(\boldsymbol{A})$ and

$$\begin{aligned} \mathbb{E}[(\boldsymbol{z}')^\intercal \boldsymbol{A}' \boldsymbol{z}'] &= n \sum_{i=1}^{\ell} \mathbb{E}[\boldsymbol{z}_i^\intercal \boldsymbol{A} \boldsymbol{z}_i] = n\ell \, \mathbb{E}[\operatorname{tr}(\boldsymbol{z}_i^\intercal \boldsymbol{A} \boldsymbol{z}_i)] = n\ell \, \mathbb{E}[\operatorname{tr}(\boldsymbol{A} \boldsymbol{z}_i \boldsymbol{z}_i^\intercal)] \\ &= n\ell \operatorname{tr}(\boldsymbol{A} \, \mathbb{E}[\boldsymbol{z}_i \boldsymbol{z}_i^\intercal]) = n\ell \operatorname{tr}(\boldsymbol{A} \operatorname{Cov}(\boldsymbol{z}_i)) = \ell \operatorname{tr}\big(\boldsymbol{A} \operatorname{Cov}(\sqrt{n} \boldsymbol{z}_i)\big) = \ell \operatorname{tr}(\boldsymbol{A}) \end{aligned}$$

Therefore by setting $t = \sqrt{\frac{\log(2\delta^{-1})}{c}} \ell \|\boldsymbol{A}\|_F$, we obtain

$$\begin{aligned} &\mathbb{P}\left(\ell \big| \tau_\ell^{\mathrm{STE}}(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A}) \big| \geq \sqrt{\frac{\log(2\delta^{-1})}{c}} \ell \|\boldsymbol{A}\|_F \right) \\ &\qquad \leq 2 \exp\left(-c \cdot \min\left(\frac{\log(2\delta^{-1})}{c} \frac{\ell \|\boldsymbol{A}\|_F^2}{\|\boldsymbol{A}'\|_F^2}, \sqrt{\frac{\log(2\delta^{-1})}{c}} \ell \frac{\|\boldsymbol{A}\|_F}{\|\boldsymbol{A}'\|_2} \right) \right) \end{aligned}$$

Further, it holds that $\|\boldsymbol{A}'\|_F^2 = \ell \|\boldsymbol{A}\|_F^2$ and $\|\boldsymbol{A}'\|_2 = \|\boldsymbol{A}\|_2$, thus we have

$$= 2 \exp\left(-\min\left(\log(2\delta^{-1}), \sqrt{c \log(2\delta^{-1}) \ell} \frac{\|\boldsymbol{A}\|_F}{\|\boldsymbol{A}\|_2} \right) \right).$$

Now assume $\ell \geq \frac{1}{c} \log(2\delta^{-1})$. Then since $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F$, the minimum is given by

$$\min\left(\log(2\delta^{-1}), \sqrt{c \log(2\delta^{-1}) \ell} \frac{\|\boldsymbol{A}\|_F}{\|\boldsymbol{A}\|_2} \right) = \log(2\delta^{-1}).$$

Further setting $c_{\boldsymbol{z}} = 2c^{-1}$, it holds that $\ell \geq c_{\boldsymbol{z}} \log(\delta^{-1}) = 2 \log(\delta^{-1}) c^{-1} \geq \log(2\delta^{-1}) c^{-1}$ since $0 < \delta \leq \frac{1}{2}$. Combining the above we obtain

$$\mathbb{P}\left(\ell \big| \tau_\ell^{\mathrm{STE}}(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A}) \big| \geq \sqrt{c_{\boldsymbol{z}} \log(\delta^{-1}) \ell} \|\boldsymbol{A}\|_F \right) \leq 2 \exp\big(-\log(2\delta^{-1}) \big) = \delta,$$

which is equivalent to

$$\mathbb{P}\left(\big| \tau_\ell^{\mathrm{STE}}(\boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A}) \big| \leq \sqrt{c_{\boldsymbol{z}} \log(\delta^{-1}) \ell^{-1}} \|\boldsymbol{A}\|_F \right) \geq 1 - \delta.$$

This proves the statement. □

Theorem 1 (Variance-reduced Stochastic Trace Estimation)
Let $\hat{\boldsymbol{K}}, \hat{\boldsymbol{P}}_\ell \in \mathbb{R}^{n \times n}_{\mathrm{spd}}, \boldsymbol{\Delta}_f \in \mathbb{R}^{n \times n}$ and $f : \mathbb{R}^{n \times n}_{\mathrm{spd}} \to \mathbb{R}$ such that $\operatorname{tr}(f(\hat{\boldsymbol{K}})) = \operatorname{tr}(f(\hat{\boldsymbol{P}}_\ell)) + \operatorname{tr}(\boldsymbol{\Delta}_f)$, and define the estimator $\tau_* = \operatorname{tr}(f(\hat{\boldsymbol{P}}_\ell)) + \tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_f)$. Now, assume there exist $c_{\boldsymbol{\Delta}} > 0$ and $g : \mathbb{N} \to (0, \infty)$ such that

$$\|\boldsymbol{\Delta}_f\|_F \leq c_{\boldsymbol{\Delta}} g(\ell) \|f(\hat{\boldsymbol{K}})\|_F. \tag{9}$$

Then there exists $c_{\boldsymbol{z}} > 0$ dependent on the choice of random vectors, such that, if $\ell \geq c_{\boldsymbol{z}} \log(\delta^{-1})$, it holds with probability $1 - \delta \in [\frac{1}{2}, 1)$ that

$$|\tau_* - \operatorname{tr}(f(\hat{\boldsymbol{K}}))| \leq \varepsilon_{\mathrm{STE}} \|f(\hat{\boldsymbol{K}})\|_F. \tag{10}$$

where for $C_1 = c_{\boldsymbol{\Delta}} \sqrt{c_{\boldsymbol{z}}}$ the relative error is given by

$$\varepsilon_{\mathrm{STE}}(\delta, \ell) = C_1 \sqrt{\log(\delta^{-1})} \ell^{-\frac{1}{2}} g(\ell). \tag{11}$$

Proof. By assumption $|\tau_* - \operatorname{tr}(f(\hat{\boldsymbol{K}}))| = \left|\tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_f) - \operatorname{tr}(\boldsymbol{\Delta}_f)\right|$. By Lemma S1 it holds with probability $\geq 1 - \delta$, that

$$\begin{aligned}
\left|\tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_f) - \operatorname{tr}(\boldsymbol{\Delta}_f)\right| &\leq \sqrt{c_{\boldsymbol{z}} \log(\delta^{-1}) \ell^{-1}} \|\boldsymbol{\Delta}_f\|_F \\
&\leq \sqrt{c_{\boldsymbol{z}} \log(\delta^{-1})} c_{\boldsymbol{\Delta}_f} \ell^{-\frac{1}{2}} g(\ell) \|f(\hat{\boldsymbol{K}})\|_F \qquad \text{Assumption (9).} \\
&= \varepsilon_{\mathrm{STE}}(\delta, \ell) \|f(\hat{\boldsymbol{K}})\|_F
\end{aligned}$$

This concludes the proof. □

Corollary S5
Let $\varepsilon \in (0, 1]$ be a desired error. If the conditions of Theorem 1 hold and the number of random vectors ℓ satisfies

$$\ell^{\frac{1}{2}} g(\ell)^{-1} \geq C_1 \varepsilon^{-1} \sqrt{\log(\delta^{-1})}, \tag{S33}$$

then it holds that

$$\boxed{\mathbb{P}(|\tau_* - \operatorname{tr}(\boldsymbol{A})| \leq \varepsilon \|\boldsymbol{A}\|_F) \geq 1 - \delta.}$$

Proof. Follows from Theorem 1 given (S33). □

S3. Log-Determinant Estimation

S3.1. Approximation of a Matrix Function

Lemma S2 (Lipschitz Continuity)
Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$ be symmetric. Assume $f : \Omega \to \mathbb{R}$ is globally Lipschitz continuous with Lipschitz constant $L > 0$ on the combined spectrum $\Omega = \lambda(\boldsymbol{A}) \cup \lambda(\boldsymbol{B}) \subset \mathbb{R}$, then there exists $c_p > 0$ such that

$$\|f(\boldsymbol{A}) - f(\boldsymbol{B})\|_p \leq c_p L \|\boldsymbol{A} - \boldsymbol{B}\|_p, \tag{S34}$$

where $\|\cdot\|_p$ denotes any matrix norm. In particular $c_2 = 1$ and $c_F = \sqrt{n}$.

Proof. Since $\boldsymbol{A}, \boldsymbol{B}$ are symmetric, they are normal. By Kittaneh (1985), it holds that

$$\|f(\boldsymbol{A}) - f(\boldsymbol{B})\|_2 \leq L \|\boldsymbol{A} - \boldsymbol{B}\|_2.$$

The result now follows by equivalence of norms on finite dimensional spaces. For the Frobenius norm we have $\frac{1}{\sqrt{n}} \|\boldsymbol{M}\|_F \leq \|\boldsymbol{M}\|_2 \leq \|\boldsymbol{M}\|_F$, and therefore $c_F = \sqrt{n}$.

□

Proposition S1
Let $\hat{\boldsymbol{K}} \in \mathbb{R}^{n \times n}$ be symmetric positive definite and assume f is analytic in a domain containing the spectrum $\lambda(\hat{\boldsymbol{K}})$. Let $\{\hat{\boldsymbol{P}}_\ell\}_\ell$ be a sequence of preconditioners with approximation quality (6)*. Then it holds that*

$$\|f(\hat{\boldsymbol{K}}) - f(\hat{\boldsymbol{P}}_\ell)\|_F \leq c(n, \hat{\boldsymbol{K}}, f) g(\ell) \|f(\hat{\boldsymbol{K}})\|_F \tag{S35}$$

where $c(n, \hat{\boldsymbol{K}}, f) = \frac{L \|\hat{\boldsymbol{K}}\|_F}{c_{f(\lambda)}}$, $L > 0$ is the Lipschitz constant of f and $c_{f(\lambda)} = \max\{\min_i |f(\lambda_i(\hat{\boldsymbol{K}}))|, \frac{\max_i |f(\lambda_i(\hat{\boldsymbol{K}}))|}{\sqrt{n}}\}$.

Proof. It holds that

$$\|f(\hat{\boldsymbol{K}})\|_F = \sqrt{\sum_{i=1}^{n} f(\lambda_i)^2} \geq \begin{cases} \sqrt{n \min_i f(\lambda_i)^2} = \sqrt{n} \min_i |f(\lambda_i)| \\ \|f(\hat{\boldsymbol{K}})\|_2 = \sigma_{\max}(f(\hat{\boldsymbol{K}})) = \sqrt{\lambda_{\max}(f(\hat{\boldsymbol{K}})^2)} = \max_i |f(\lambda_i)| \end{cases}$$

and therefore $\|f(\hat{\boldsymbol{K}})\|_F \geq \sqrt{n} c_{f(\lambda)}$. Since f is analytic and therefore Lipschitz, it holds that

$$\begin{aligned} \|f(\hat{\boldsymbol{K}}) - f(\hat{\boldsymbol{P}}_\ell)\|_F &\leq L\sqrt{n}\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F && \text{Lemma S2} \\ &\leq L\sqrt{n} g(\ell) \|\hat{\boldsymbol{K}}\|_F && \text{Preconditioner quality (6)} \\ &= L\sqrt{n} g(\ell) \frac{\|\hat{\boldsymbol{K}}\|_F}{\|f(\hat{\boldsymbol{K}})\|_F} \|f(\hat{\boldsymbol{K}})\|_F \\ &\leq L\sqrt{n} g(\ell) \frac{\|\hat{\boldsymbol{K}}\|_F}{\sqrt{n} c_{f(\lambda)}} \|f(\hat{\boldsymbol{K}})\|_F \\ &\leq \frac{L\|\hat{\boldsymbol{K}}\|_F}{c_{f(\lambda)}} g(\ell) \|f(\hat{\boldsymbol{K}})\|_F. \end{aligned}$$

This proves the claim. □

S3.2. Approximation of the Log-Determinant

Lemma S3 (Decomposition of the log-determinant)
For $\hat{\boldsymbol{K}}, \hat{\boldsymbol{P}} \in \mathbb{R}^{n \times n}$ symmetric positive definite, it holds that

$$\log\det(\hat{\boldsymbol{K}}) = \log\det(\hat{\boldsymbol{P}}) + \operatorname{tr}(\log(\hat{\boldsymbol{K}}) - \log(\hat{\boldsymbol{P}})) \tag{S36}$$

$$= \log\det(\hat{\boldsymbol{P}}) + \operatorname{tr}\Big(\log\Big(\hat{\boldsymbol{P}}^{-\frac{1}{2}} \hat{\boldsymbol{K}} \hat{\boldsymbol{P}}^{-\frac{\intercal}{2}}\Big)\Big). \tag{S37}$$

Proof. Note that for symmetric positive definite matrices $\boldsymbol{A}, \boldsymbol{B}$, the matrix logarithm satisfies the following

$$\begin{aligned} \log\det(\boldsymbol{A}) &= \operatorname{tr}(\log(\boldsymbol{A})), \\ \operatorname{tr}(\log(\boldsymbol{A}\boldsymbol{B})) &= \operatorname{tr}(\log(\boldsymbol{A})) + \operatorname{tr}(\log(\boldsymbol{B})), \\ \log\big(\boldsymbol{A}^{-1}\big) &= -\log(\boldsymbol{A}). \end{aligned}$$

Using the above properties, we obtain

$$\begin{aligned} \log\det(\hat{\boldsymbol{K}}) &= \operatorname{tr}\big(\log\big(\hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{K}}\big)\big) \\ &= \operatorname{tr}(\log(\hat{\boldsymbol{P}})) - \operatorname{tr}(\log(\hat{\boldsymbol{P}})) + \operatorname{tr}(\log(\hat{\boldsymbol{K}})) \\ &= \log\det(\hat{\boldsymbol{P}}) + \operatorname{tr}(\log(\hat{\boldsymbol{K}}) - \log(\hat{\boldsymbol{P}})) \end{aligned}$$

Now since $\hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{K}}$ and $\hat{\boldsymbol{P}}^{-\frac{1}{2}}\hat{\boldsymbol{K}}\hat{\boldsymbol{P}}^{-\frac{\intercal}{2}}$ are similar, they have the same determinant. Therefore we have

$$\operatorname{tr}(\log(\hat{\boldsymbol{K}}) - \log(\hat{\boldsymbol{P}})) = \operatorname{tr}\big(\log\big(\hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{K}}\big)\big) = \log\det\Big(\hat{\boldsymbol{P}}^{-\frac{1}{2}}\hat{\boldsymbol{K}}\hat{\boldsymbol{P}}^{-\frac{\intercal}{2}}\Big).$$

This completes the proof. □

Theorem 2 (Error Bound for $\log\det(\hat{\boldsymbol{K}})$)
Let $f = \log$, $\boldsymbol{\Delta}_{\log} = \log\big(\hat{\boldsymbol{P}}^{-\frac{1}{2}}\hat{\boldsymbol{K}}\hat{\boldsymbol{P}}^{-\frac{\intercal}{2}}\big)$ and assume the conditions of Theorem 1 hold. Then, with probability $1-\delta$, it holds for $\tau_*^{\log} = \log(\det(\hat{\boldsymbol{P}})) + \tau_{\ell,m}^{\mathrm{SLQ}}(\boldsymbol{\Delta}_{\log})$, that

$$\Big|\tau_*^{\log} - \log\det(\hat{\boldsymbol{K}})\Big| \leq (\varepsilon_{\mathrm{Lanczos}} + \varepsilon_{\mathrm{STE}})\|\log(\hat{\boldsymbol{K}})\|_F,$$

where the individual errors are bounded by

$$\varepsilon_{\mathrm{Lanczos}}(\kappa, m) \leq K_1 \left(\frac{\sqrt{2\kappa+1}-1}{\sqrt{2\kappa+1}+1}\right)^{2m} \tag{12}$$

$$\varepsilon_{\mathrm{STE}}(\delta, \ell) \leq C_1 \sqrt{\log(\delta^{-1})} \ell^{-\frac{1}{2}} g(\ell) \tag{13}$$

and $K_1 = \frac{5\kappa\log(2(\kappa+1))}{2\|\log(\hat{\boldsymbol{K}})\|_F\sqrt{2\kappa+1}}$.

Proof. Using the decomposition (8), we have

$$\begin{aligned}\left|\tau_*^{\log} - \log\det(\hat{\boldsymbol{K}})\right| &= \left|\tau_{\ell,m}^{\mathrm{SLQ}}(\boldsymbol{\Delta}_{\log}) - \mathrm{tr}(\boldsymbol{\Delta}_{\log})\right| \\ &\leq \left|\mathrm{tr}(\boldsymbol{\Delta}_{\log}) - \tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_{\log})\right| + \left|\tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_{\log}) - \tau_{\ell,m}^{\mathrm{SLQ}}(\boldsymbol{\Delta}_{\log})\right| \\ &= \underbrace{\left|\mathrm{tr}(\log(\hat{\boldsymbol{K}})) - (\mathrm{tr}(\log(\hat{\boldsymbol{P}})) + \tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_{\log}))\right|}_{e_{\mathrm{STE}}} + \underbrace{\left|\tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_{\log}) - \tau_{\ell,m}^{\mathrm{SLQ}}(\boldsymbol{\Delta}_{\log})\right|}_{e_{\mathrm{Lanczos}}}.\end{aligned}$$

Now the individual absolute errors are bounded as follows. By the error bound for stochastic trace estimation in Theorem 1, we have

$$e_{\mathrm{STE}} \leq \varepsilon_{\mathrm{STE}}(\delta,\ell)\|\log(\hat{\boldsymbol{K}})\|_F = C_1\sqrt{\log(\delta^{-1})}\ell^{-\frac{1}{2}}g(\ell)\|\log(\hat{\boldsymbol{K}})\|_F$$

and by Corollary S4, it follows that

$$e_{\mathrm{Lanczos}} \leq K\left(\frac{\sqrt{2\kappa+1}-1}{\sqrt{2\kappa+1}+1}\right)^{2m} = K_1\left(\frac{\sqrt{2\kappa+1}-1}{\sqrt{2\kappa+1}+1}\right)^{2m}\|\log(\hat{\boldsymbol{K}})\|_F.$$

□

Corollary 1
Assume the conditions of Theorem 2 hold. If the number of random vectors ℓ satisfies (11) *with $\varepsilon_{\mathrm{STE}} = \frac{\varepsilon}{2}$ and we run*

$$m \geq \tfrac{\sqrt{3}}{4}\sqrt{\kappa}\log\left(2K_1\varepsilon^{-1}\right) \tag{14}$$

iterations of Lanczos, then it holds that

$$\boxed{\mathbb{P}\left(\left|\tau_*^{\log} - \log\det(\hat{\boldsymbol{K}})\right| \leq \varepsilon\|\log(\hat{\boldsymbol{K}})\|_F\right) \geq 1-\delta.}$$

Proof. By assumption Theorem 1 is satisfied and therefore $\varepsilon_{\mathrm{STE}} = \frac{\varepsilon}{2}$ with probability $1-\delta$. Now for the error of Lanczos it holds by Theorem 2 in combination with Lemma S7, that

$$\begin{aligned}\varepsilon_{\mathrm{Lanczos}} &\leq K_1\left(\frac{\sqrt{2\kappa+1}-1}{\sqrt{2\kappa+1}+1}\right)^{2m} \\ &\leq K_1\exp\left(-\frac{4}{\sqrt{2\kappa+1}}m\right) && \text{Lemma S7} \\ &\leq K_1\exp\left(-\frac{\sqrt{3\kappa}}{\sqrt{2\kappa+1}}\log\left(2K_1\varepsilon^{-1}\right)\right) && \text{By assumption (14).} \\ &\leq K_1\exp\left(-\log\left(2K_1\varepsilon^{-1}\right)\right) \\ &= \frac{\varepsilon}{2}\end{aligned}$$

The result now follows by Theorem 2. □

S3.3. Approximation of the Derivative of the Log-Determinant

Computation of $\mathrm{tr}\left(\hat{\boldsymbol{P}}^{-1}\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\right)$ Algorithm 1 and Algorithm 2 primarily rely on matrix-vector multiplication, except for computation of $\tau_{\boldsymbol{P}}^{\mathrm{inv}\partial} = \mathrm{tr}\left(\hat{\boldsymbol{P}}^{-1}\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\right)$. Efficient computation of this term depends on the structure of $\hat{\boldsymbol{P}}^{-1}$. If $\hat{\boldsymbol{P}}$ is the pivoted-Cholesky preconditioner, or any other diagonal-plus-low-rank preconditioner $\sigma^2\boldsymbol{I} + \boldsymbol{L}_\ell\boldsymbol{L}_\ell^\intercal$, we can rewrite this term using the matrix inversion lemma

$$\begin{aligned}\mathrm{tr}\left(\hat{\boldsymbol{P}}^{-1}\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\right) &= \sigma^{-2}\,\mathrm{tr}\left(\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\right) - \sigma^{-2}\,\mathrm{tr}\left(\boldsymbol{L}_\ell\left(\sigma^2\boldsymbol{I} + \boldsymbol{L}_\ell^\intercal\boldsymbol{L}_\ell\right)^{-1}\boldsymbol{L}_\ell^\intercal\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\right) \\ &= \sigma^{-2}\sum_{i=1}^{n}\frac{\partial\hat{\boldsymbol{P}}_{ii}}{\partial\theta} - \sigma^{-2}\left(\left(\boldsymbol{L}_\ell\left(\sigma^2\boldsymbol{I} + \boldsymbol{L}_\ell^\intercal\boldsymbol{L}_\ell\right)^{-1}\right)\circ\left(\frac{\partial\hat{\boldsymbol{P}}}{\partial\theta}\boldsymbol{L}_\ell\right)\right)\mathbf{1},\end{aligned} \tag{S38}$$

where $\circ$ denotes elementwise multiplication. The second term requires ℓ matrix-vector multiplies with $\frac{\partial \hat{\boldsymbol{P}}}{\partial \theta}$ and $\mathcal{O}(n\ell^2)$ additional work. The first term is simply the derivative of the kernel diagonal which will take $\mathcal{O}(n)$ time. We note that similar efficient procedures exists for other types of preconditioners, such as when $\hat{\boldsymbol{P}}^{-1}$ has banded structure.

Theorem 3 (Error Bound for $\operatorname{tr}\big(\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\big)$)
Let $f(\hat{\boldsymbol{K}}) = \hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}$, $\boldsymbol{\Delta}_{\mathrm{inv}\partial} = \hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta} - \hat{\boldsymbol{P}}^{-1}\frac{\partial \hat{\boldsymbol{P}}}{\partial \theta}$ and assume the conditions of Theorem 1 hold. If we solve $\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\boldsymbol{z}_i$ with m iterations of preconditioned CG, initialized at $\mathbf{0}$ or better, then it holds with probability $1-\delta$ for $\tau_*^{\mathrm{inv}\partial} = \operatorname{tr}\big(\hat{\boldsymbol{P}}^{-1}\frac{\partial \hat{\boldsymbol{P}}}{\partial \theta}\big) + \tau_{\ell,m}^{\mathrm{SCG}}(\boldsymbol{\Delta}_{\mathrm{inv}\partial})$, that

$$\big|\tau_*^{\mathrm{inv}\partial} - \operatorname{tr}\big(\hat{\boldsymbol{K}}^{-1}\tfrac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\big)\big| \leq (\varepsilon_{\mathrm{CG}'} + \varepsilon_{\mathrm{STE}})\big\|\boldsymbol{K}^{-1}\tfrac{\partial \boldsymbol{K}}{\partial \theta}\big\|_F,$$

where the individual errors are bounded by

$$\varepsilon_{\mathrm{CG}'}(\kappa, m) \leq K_2\left(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \tag{17}$$

$$\varepsilon_{\mathrm{STE}}(\delta, \ell) \leq C_1\sqrt{\log(\delta^{-1})}\ell^{-\frac{1}{2}}g(\ell) \tag{18}$$

and $K_2 = 2\sqrt{\kappa(\hat{\boldsymbol{K}})}n$.

Proof. Using the decomposition (15), we have

$$\begin{aligned}
\left|\tau_*^{\mathrm{inv}\partial} - \operatorname{tr}\left(\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\right)\right| &= \left|\tau_{\ell,m}^{\mathrm{SCG}}(\boldsymbol{\Delta}_{\mathrm{inv}\partial}) - \operatorname{tr}(\boldsymbol{\Delta}_{\mathrm{inv}\partial})\right| \\
&\leq \left|\operatorname{tr}(\boldsymbol{\Delta}_{\mathrm{inv}\partial}) - \tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_{\mathrm{inv}\partial})\right| + \left|\tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_{\mathrm{inv}\partial}) - \tau_{\ell,m}^{\mathrm{SCG}}(\boldsymbol{\Delta}_{\mathrm{inv}\partial})\right| \\
&= \underbrace{\left|\operatorname{tr}\left(\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\right) - \left(\operatorname{tr}\left(\hat{\boldsymbol{P}}^{-1}\frac{\partial \hat{\boldsymbol{P}}}{\partial \theta}\right) + \tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_{\mathrm{inv}\partial})\right)\right|}_{e_{\mathrm{STE}}} + \underbrace{\left|\tau_\ell^{\mathrm{STE}}(\boldsymbol{\Delta}_{\mathrm{inv}\partial}) - \tau_{\ell,m}^{\mathrm{SCG}}(\boldsymbol{\Delta}_{\mathrm{inv}\partial})\right|}_{e_{\mathrm{CG}}}.
\end{aligned}$$

Now the individual absolute errors are bounded as follows. By the error bound for stochastic trace estimation in Theorem 1, we have

$$e_{\mathrm{STE}} \leq \varepsilon_{\mathrm{STE}}(\delta, \ell)\left\|\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\right\|_F = C_1\sqrt{\log(\delta^{-1})}\ell^{-\frac{1}{2}}g(\ell)\left\|\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\right\|_F.$$

Now, let $\boldsymbol{w}_i = \hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\boldsymbol{z}_i$, $\tilde{\boldsymbol{w}}_i = \hat{\boldsymbol{P}}^{-1}\frac{\partial \hat{\boldsymbol{P}}}{\partial \theta}\boldsymbol{z}_i$ and $\boldsymbol{w}_{m,i} \approx \boldsymbol{w}_i$ be the solution computed via preconditioned CG with m iterations. Then we have by Theorem S6, that

$$\begin{aligned}
e_{\mathrm{CG}} &= \left|\frac{n}{\ell}\sum_{i=1}^{\ell}\boldsymbol{z}_i^\intercal(\boldsymbol{w}_{m,i} - \tilde{\boldsymbol{w}}_i - (\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i))\right| \\
&\leq \frac{n}{\ell}\sum_{i=1}^{\ell}\underbrace{\|\boldsymbol{z}_i\|_2}_{=1}\|\boldsymbol{w}_{m,i} - \boldsymbol{w}_i\|_2 \\
&\leq \frac{n}{\ell}\sum_{i=1}^{\ell}2\sqrt{\kappa(\hat{\boldsymbol{K}})}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m\|\boldsymbol{w}_{0,i} - \boldsymbol{w}_i\|_2 && \text{CG convergence by Theorem S6.} \\
&\leq 2n\sqrt{\kappa(\hat{\boldsymbol{K}})}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m\|\boldsymbol{w}_i\|_2 && \text{CG initialized at } \boldsymbol{w}_{0,i} = \mathbf{0} \text{ or better.} \\
&\leq K_2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m\left\|\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\boldsymbol{z}_i\right\|_F \\
&\leq K_2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m\left\|\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\right\|_F
\end{aligned}$$

This completes the argument. □

Corollary 2
Assume the conditions of Theorem 3 hold. If the number of random vectors ℓ satisfies (11) *with $\varepsilon_{\mathrm{STE}} = \frac{\varepsilon}{2}$, and we run*

$$m \geq \tfrac{1}{2}\sqrt{\kappa}\log\big(2K_2\varepsilon^{-1}\big) \tag{19}$$

iterations of CG, then

$$\boxed{\mathbb{P}\left(\left|\tau_*^{\mathrm{inv}\partial} - \mathrm{tr}\big(\hat{\boldsymbol{K}}^{-1}\tfrac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\big)\right| \leq \varepsilon\left\|\hat{\boldsymbol{K}}^{-1}\tfrac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\right\|_F\right) \geq 1-\delta.}$$

Proof. By assumption Theorem 1 is satisfied and therefore $\varepsilon_{\mathrm{STE}} = \frac{\varepsilon}{2}$ with probability $1-\delta$. Now for the error of CG, it holds by Theorem 3 in combination with Lemma S7, that

$$\begin{aligned}
\varepsilon_{\mathrm{CG}} &\leq K_2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \\
&\leq K_2\exp\left(-\frac{2m}{\sqrt{\kappa}}\right) && \text{Lemma S7} \\
&\leq K_2\exp\big(-\log\big(2K_2\varepsilon^{-1}\big)\big) && \text{Assumption (19)} \\
&\leq \frac{\varepsilon}{2}.
\end{aligned}$$

The result now follows by Theorem 3. □

S4. GP Hyperparameter Optimization

S4.1. Approximation of the Log-Marginal Likelihood

Theorem 4 (Error Bound for the log-Marginal Likelihood)
Assume the conditions of Theorem 2 hold and we solve $\hat{\boldsymbol{K}}\boldsymbol{u} = \boldsymbol{y}$ via preconditioned CG initialized at $\boldsymbol{u}_0$ and terminated after m iterations. Then with probability $1-\delta$, the error in the estimate $\eta = -\frac{1}{2}(\boldsymbol{y}^\intercal\boldsymbol{u}_m + \tau_*^{\log} + n\log(2\pi))$ of the log*-marginal likelihood $\mathcal{L}$ satisfies*

$$|\eta - \mathcal{L}| \leq \varepsilon_{\mathrm{CG}} + \tfrac{1}{2}(\varepsilon_{\mathrm{Lanczos}} + \varepsilon_{\mathrm{STE}})\|\log(\hat{\boldsymbol{K}})\|_F,$$

where the individual errors are bounded by

$$\varepsilon_{\mathrm{CG}}(\kappa, m) \leq K_3\left(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \tag{20}$$

$$\varepsilon_{\mathrm{Lanczos}}(\kappa, m) \leq K_1\left(\tfrac{\sqrt{2\kappa+1}-1}{\sqrt{2\kappa+1}+1}\right)^{2m} \tag{21}$$

$$\varepsilon_{\mathrm{STE}}(\delta, \ell) \leq C_1\sqrt{\log(\delta^{-1})}\ell^{-\frac{1}{2}}g(\ell) \tag{22}$$

for $K_3 = \sqrt{\kappa(\hat{\boldsymbol{K}})}\|\boldsymbol{y}\|_2\|\boldsymbol{u}_0 - \boldsymbol{u}\|_2$.

Proof. It holds by assumption that

$$\begin{aligned}
|\eta - \mathcal{L}| &= \frac{1}{2}\big|\boldsymbol{y}^\intercal\boldsymbol{u}_m + \tau_*^{\log} - (\boldsymbol{y}^\intercal\hat{\boldsymbol{K}}^{-1}\boldsymbol{y} + \log\det(\hat{\boldsymbol{K}}))\big| \\
&\leq \frac{1}{2}\big(\underbrace{|\boldsymbol{y}^\intercal\boldsymbol{u}_m - \boldsymbol{y}^\intercal\boldsymbol{u}|}_{e_{\mathrm{CG}}} + \underbrace{\big|\tau_*^{\log} - \log\det(\hat{\boldsymbol{K}})\big|}_{e_{\mathrm{SLQ}}}\big).
\end{aligned}$$

For the error of CG when solving $\hat{\boldsymbol{K}}\boldsymbol{u} = \boldsymbol{y}$, we have by Theorem S6

$$e_{\mathrm{CG}} = |\boldsymbol{y}^\intercal\boldsymbol{u}_m - \boldsymbol{y}^\intercal\boldsymbol{u}| \leq \|\boldsymbol{y}\|_2\|\boldsymbol{u}_m - \boldsymbol{u}\|_2 \leq \|\boldsymbol{y}\|_2 2\sqrt{\kappa(\hat{\boldsymbol{K}})}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m\|\boldsymbol{u}_0 - \boldsymbol{u}\|_2 = 2\varepsilon_{\mathrm{CG}},$$

and for the absolute error in the log-determinant estimate via preconditioned stochastic Lanczos quadrature, we obtain by Theorem 2, that

$$e_{\mathrm{SLQ}} \leq (\varepsilon_{\mathrm{Lanczos}} + \varepsilon_{\mathrm{STE}})\|\log(\hat{\boldsymbol{K}})\|_F.$$

This proves the statement. □

Corollary S6
Assume the conditions of Theorem 4 hold. If the number of random vectors ℓ satisfies (11) *with $\varepsilon_{\mathrm{STE}} = \varepsilon$, and we run $m \geq \max(m_{\mathrm{CG}}, m_{\mathrm{Lanczos}})$ iterations of CG and Lanczos, where*

$$m_{\mathrm{CG}} \geq \frac{1}{2}\sqrt{\kappa}\log\big(2K_3\varepsilon^{-1}\big), \tag{S39}$$

$$m_{\mathrm{Lanczos}} \geq \frac{\sqrt{3}}{4}\sqrt{\kappa}\log\big(K_1\varepsilon^{-1}\big), \tag{S40}$$

then it holds that

$$\boxed{\mathbb{P}\left(|\eta - \mathcal{L}| \leq \varepsilon(1 + \|\log(\hat{\boldsymbol{K}})\|_F)\right) \geq 1 - \delta.}$$

Proof. We begin with the error of CG, it holds by Theorem 4 in combination with Lemma S7, that

$$\begin{aligned}
\varepsilon_{\mathrm{CG}} &\leq K_3 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \\
&\leq K_3 \exp\left(-\frac{2m}{\sqrt{\kappa}}\right) && \text{Lemma S7} \\
&\leq K_3 \exp\big(-\log\big(2K_3\varepsilon^{-1}\big)\big) && \text{Assumption (S39)} \\
&= \frac{\varepsilon}{2}.
\end{aligned}$$

Now for the error of the estimate of the log-determinant. By assumption Theorem 1 is satisfied and therefore $\varepsilon_{\mathrm{STE}} = \varepsilon$ with probability $1 - \delta$. For the error of Lanczos, it holds by Theorem 4, that

$$\begin{aligned}
\varepsilon_{\mathrm{Lanczos}} &\leq K_1 \left(\frac{\sqrt{2\kappa+1}-1}{\sqrt{2\kappa+1}+1}\right)^{2m} \\
&\leq K_1 \exp\left(-\frac{4}{\sqrt{2\kappa+1}}m\right) && \text{Lemma S7} \\
&\leq K_1 \exp\left(-\frac{\sqrt{3\kappa}}{\sqrt{2\kappa+1}}\log\big(K_1\varepsilon^{-1}\big)\right) && \text{Assumption (S40).} \\
&\leq K_1 \exp\big(-\log\big(K_1\varepsilon^{-1}\big)\big) \\
&= \varepsilon
\end{aligned}$$

The result now follows by Theorem 4. □

S4.2. Approximation of the Derivative of the Log-Marginal Likelihood

Theorem 5 (Error Bound for the Derivative)
Assume the conditions of Theorem 3 hold and we solve $\hat{\boldsymbol{K}}\boldsymbol{u} = \boldsymbol{y}$ via preconditioned CG initialized at $\mathbf{0}$ or better and terminated after m iterations. Then with probability $1 - \delta$, the error in the estimate $\phi = \frac{1}{2}(\boldsymbol{u}_m^\intercal \frac{\partial \hat{\boldsymbol{K}}}{\partial \theta}\boldsymbol{u}_m - \tau_*^{\mathrm{inv}\partial})$ of the derivative of the log*-marginal likelihood $\frac{\partial}{\partial\theta}\mathcal{L}$ satisfies*

$$\left|\phi - \tfrac{\partial}{\partial\theta}\mathcal{L}\right| \leq \varepsilon_{\mathrm{CG}}(\kappa, m) + \tfrac{1}{2}(\varepsilon_{\mathrm{CG}'} + \varepsilon_{\mathrm{STE}})\left\|\hat{\boldsymbol{K}}^{-1}\tfrac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\right\|_F,$$

where the individual errors are bounded by

$$\varepsilon_{\mathrm{CG}}(\kappa, m) \leq K_4 \left(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \tag{23}$$

$$\varepsilon_{\mathrm{CG}'}(\kappa, m) \leq K_2 \left(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \tag{24}$$

$$\varepsilon_{\mathrm{STE}}(\delta, \ell) \leq C_1\sqrt{\log(\delta^{-1})}\ell^{-\frac{1}{2}}g(\ell) \tag{25}$$

for $K_4 = 6\kappa(\hat{\boldsymbol{K}})\max(\|\boldsymbol{u}\|_2, \|\boldsymbol{u}\|_2^3)\left\|\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\right\|_2$.

Proof. It holds that

$$\begin{aligned}\left|\phi - \frac{\partial}{\partial\theta}\mathcal{L}\right| &= \frac{1}{2}\left|\boldsymbol{u}_m^\intercal \frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\boldsymbol{u}_m - \tau_*^{\mathrm{inv}\partial} - \left(\boldsymbol{y}^\intercal \hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\hat{\boldsymbol{K}}^{-1}\boldsymbol{y} - \mathrm{tr}\left(\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\right)\right)\right| \\ &\leq \Bigg(\underbrace{\left|\boldsymbol{u}_m^\intercal \frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\boldsymbol{u}_m - \boldsymbol{u}^\intercal \frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\boldsymbol{u}\right|}_{e_{\mathrm{CG}}} + \underbrace{\left|\tau_*^{\mathrm{inv}\partial} - \mathrm{tr}\left(\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\right)\right|}_{e_{\mathrm{CGSTE}}}\Bigg)\end{aligned}$$

Now by Theorem 3, we have

$$e_{\mathrm{CGSTE}} \leq (\varepsilon_{\mathrm{CG}'} + \varepsilon_{\mathrm{STE}})\left\|\hat{\boldsymbol{K}}^{-1}\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\right\|_F.$$

For the absolute error of the quadratic term, it holds that

$$\begin{aligned}e_{\mathrm{CG}} &= \left|\|\boldsymbol{u}\|^2_{\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}} - \|\boldsymbol{u}_m - \boldsymbol{u} + \boldsymbol{u}\|^2_{\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}}\right| \\ &\leq \left|\|\boldsymbol{u}\|^2_{\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}} - (\|\boldsymbol{u}_m - \boldsymbol{u}\|_{\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}} + \|\boldsymbol{u}\|_{\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}})^2\right| \\ &= \|\boldsymbol{u}_m - \boldsymbol{u}\|_{\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}} + 2\|\boldsymbol{u}_m - \boldsymbol{u}\|_{\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}}\|\boldsymbol{u}\|_{\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}} \\ &\leq \left\|\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\right\|_2 (\|\boldsymbol{u}_m - \boldsymbol{u}\|_2^2 + 2\|\boldsymbol{u}_m - \boldsymbol{u}\|_2\|\boldsymbol{u}\|_2) \\ &\leq \left\|\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\right\|_2 g(\|\boldsymbol{u}_m - \boldsymbol{u}\|_2)(1 + 2\|\boldsymbol{u}\|_2)\end{aligned}$$

for $g(t) = \max(t, t^2)$. Now it holds by Theorem S6 and monotoncity of g, that

$$\leq \left\|\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\right\|_2 g\left(2\sqrt{\kappa(\hat{\boldsymbol{K}})}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \|\boldsymbol{u}_0 - \boldsymbol{u}\|_2\right)(1 + 2\|\boldsymbol{u}\|_2)$$

Since for $a \leq 1$, it holds that $g(at) \leq ag(t)$ and for $a > 1 : g(at) \leq a^2 g(t)$, we have

$$\begin{aligned}&\leq \left\|\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\right\|_2 4\kappa(\hat{\boldsymbol{K}})\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m g(\|\boldsymbol{u}_0 - \boldsymbol{u}\|_2)(1 + 2\|\boldsymbol{u}\|_2) \\ &\leq \left\|\frac{\partial \hat{\boldsymbol{K}}}{\partial\theta}\right\|_2 4\kappa(\hat{\boldsymbol{K}})\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m 3\max(\|\boldsymbol{u}\|_2, \|\boldsymbol{u}\|_2^3) \\ &= 2K_4\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m\end{aligned}$$

where we used that CG was initialized at $\boldsymbol{u}_0 = \boldsymbol{0}$ or better. □

Corollary S7

Assume the conditions of Theorem 5 hold. If the number of random vectors ℓ satisfies (11) *with $\varepsilon_{\mathrm{STE}} = \varepsilon$, and we run $m \geq \max(m_{\mathrm{CG}}, m_{\mathrm{CG}'})$ iterations of CG and Lanczos, where*

$$m_{\mathrm{CG}} \geq \frac{1}{2}\sqrt{\kappa}\log\big(2K_4\varepsilon^{-1}\big), \tag{S41}$$

$$m_{\mathrm{CG}'} \geq \frac{1}{2}\sqrt{\kappa}\log\big(K_2\varepsilon^{-1}\big), \tag{S42}$$

then it holds that

$$\boxed{\mathbb{P}\left(\left|\phi - \tfrac{\partial}{\partial\theta}\mathcal{L}(\boldsymbol{\theta})\right| \leq \varepsilon(1 + \left\|\boldsymbol{K}^{-1}\tfrac{\partial \boldsymbol{K}}{\partial\theta}\right\|_F)\right) \geq 1 - \delta.}$$

Proof. We begin with the error of CG's estimate of $\boldsymbol{y}^\intercal \hat{\boldsymbol{K}}^{-1} \frac{\partial \hat{\boldsymbol{K}}}{\partial \theta} \hat{\boldsymbol{K}}^{-1} \boldsymbol{y}$, it holds by Theorem 5 in combination with Lemma S7, that

$$\begin{aligned}
\varepsilon_{\mathrm{CG}} &\leq K_4 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^m \\
&\leq K_4 \exp\left(-\frac{2m}{\sqrt{\kappa}} \right) && \text{Lemma S7} \\
&\leq K_4 \exp\big(-\log\big(2 K_4 \varepsilon^{-1} \big) \big) && \text{Assumption (S41)} \\
&= \frac{\varepsilon}{2}.
\end{aligned}$$

Now for the error of the stochastic trace estimator. By assumption Theorem 1 is satisfied and therefore $\varepsilon_{\mathrm{STE}} = \varepsilon$ with probability $1 - \delta$. For the error of CG used in the stochastic trace estimate, we obtain by Theorem 5

$$\begin{aligned}
\varepsilon_{\mathrm{CG}'} &\leq K_2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^m \\
&\leq K_2 \exp\left(-\frac{2m}{\sqrt{\kappa}} \right) && \text{Lemma S7} \\
&\leq K_2 \exp\big(-\log\big(K_2 \varepsilon^{-1} \big) \big) && \text{Assumption (S42)} \\
&= \varepsilon.
\end{aligned}$$

The result now follows by Theorem 5. □

S5. Preconditioning

Lemma S4 (Condition Number and Preconditioner Quality)
Let $\hat{\boldsymbol{K}}, \hat{\boldsymbol{P}}_\ell \in \mathbb{R}^{n \times n}$ symmetric positive-definite such that (6) *holds and assume that there exists $c > 0$ such that $c \geq \max(\lambda_{\min}(\hat{\boldsymbol{P}}), \lambda_{\min}(\hat{\boldsymbol{K}}))$ for all ℓ. Then it holds that*

$$\kappa = \kappa(\hat{\boldsymbol{P}}_\ell^{-\frac{1}{2}} \hat{\boldsymbol{K}} \hat{\boldsymbol{P}}_\ell^{-\frac{1}{2}}) \leq (1 + \mathcal{O}(g(\ell)) \|\hat{\boldsymbol{K}}\|_F)^2 \tag{S43}$$

Proof. Part of the strategy for this proof is adapted from Gardner et al. (2018). First note, that the matrices $\hat{\boldsymbol{P}}^{-1} \hat{\boldsymbol{K}}$, $\hat{\boldsymbol{K}} \hat{\boldsymbol{P}}^{-1}$ and $\hat{\boldsymbol{P}}_\ell^{-\frac{1}{2}} \hat{\boldsymbol{K}} \hat{\boldsymbol{P}}_\ell^{-\frac{1}{2}}$ are similar, and thus have the same eigenvalues. Now, we have:

$$\begin{aligned}
\kappa(\hat{\boldsymbol{P}}_\ell^{-\frac{1}{2}} \hat{\boldsymbol{K}} \hat{\boldsymbol{P}}_\ell^{-\frac{1}{2}}) &= \frac{\lambda_{\max}(\hat{\boldsymbol{P}}^{-1} \hat{\boldsymbol{K}})}{\lambda_{\min}(\hat{\boldsymbol{K}} \hat{\boldsymbol{P}}^{-1})} = \left\| \hat{\boldsymbol{P}}^{-1} \hat{\boldsymbol{K}} \right\|_2 \left\| \hat{\boldsymbol{P}} \hat{\boldsymbol{K}}^{-1} \right\|_2 \\
&= \left\| \hat{\boldsymbol{P}}^{-1} (\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}} + \hat{\boldsymbol{P}}) \right\|_2 \left\| (\hat{\boldsymbol{P}} - \hat{\boldsymbol{K}} + \hat{\boldsymbol{K}}) \hat{\boldsymbol{K}}^{-1} \right\|_2 \\
&= \left\| 1 + \hat{\boldsymbol{P}}^{-1} (\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}) \right\|_2 \left\| 1 - \hat{\boldsymbol{K}}^{-1} (\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}) \right\|_2
\end{aligned}$$

Applying Cauchy-Schwarz and the triangle inequality, we obtain:

$$\begin{aligned}
&\leq (1 + \left\| \hat{\boldsymbol{P}}^{-1} \right\|_2 \|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}\|_2)(1 + \left\| \hat{\boldsymbol{K}}^{-1} \right\|_2 \|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}\|_2) \\
&\leq (1 + c \|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}\|_F)^2 \\
&\leq (1 + \mathcal{O}(g(\ell)) \|\hat{\boldsymbol{K}}\|_F)^2
\end{aligned}$$

□

Note since typically $\hat{\boldsymbol{P}} = \sigma^2 \boldsymbol{I} + \boldsymbol{P}$ with $\lambda_{\min}(\boldsymbol{P}) \approx 0$ and $\lambda_{\min}(\hat{\boldsymbol{K}}) \leq \sigma^2 + \lambda_{\min}(\boldsymbol{K})$, where $\lambda_{\min}(\boldsymbol{K})$ small for most kernels, usually $c \approx \sigma^2$.

S5.1. Additive Kernels

Lemma S5 (Additive Kernels)
Let $k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{d} k_j(\boldsymbol{x}_j, \boldsymbol{y}_j)$ be an additive kernel and $\{(\hat{\boldsymbol{P}}_\ell)_j\}_{j=1}^{d}$ a set of preconditioners indexed by ℓ, such that for all $j = 1, \dots, d$, we have

$$\|\hat{\boldsymbol{K}}_j - (\hat{\boldsymbol{P}}_\ell)_j\|_F \leq c_j g(\ell) \|\hat{\boldsymbol{K}}_j\|_F. \tag{S44}$$

Then it holds for $\hat{\boldsymbol{P}}_\ell = \sum_{j=1}^{d} (\hat{\boldsymbol{P}}_\ell)_j$ and $c = \max_j c_j$ that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq c d g(\ell) \|\hat{\boldsymbol{K}}\|_F \tag{S45}$$

Proof. It holds by assumption, that

$$\begin{aligned}
\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F &\leq \sum_{j=1}^{d} \|\hat{\boldsymbol{K}}_j - (\hat{\boldsymbol{P}}_\ell)_j\|_F && \text{Cauchy-Schwarz} \\
&\leq \sum_{j=1}^{d} c_j g(\ell) \|\hat{\boldsymbol{K}}_j\|_F \\
&\leq \max_j c_j g(\ell) \sum_{j=1}^{d} \sqrt{\sum_{i=1}^{n} \lambda_i(\hat{\boldsymbol{K}}_j)^2} \\
&\leq c g(\ell) \sum_{j=1}^{d} \sqrt{\sum_{i=1}^{n} \lambda_i(\hat{\boldsymbol{K}})^2} && \boldsymbol{A}, \boldsymbol{B} \text{ spd} \implies \lambda_i(\boldsymbol{A}) \leq \lambda_i(\boldsymbol{A} + \boldsymbol{B}) \\
&\leq c g(\ell) d \|\hat{\boldsymbol{K}}\|_F
\end{aligned}$$

□

S5.2. Kernels with a Uniformly Converging Approximation

Lemma S6 (Preconditioner Quality from Uniform Convergence)
Let $\Omega \subset \mathbb{R}^d$ be the data domain and $k(\cdot, \cdot)$ a positive-definite kernel such that for all $\boldsymbol{x} \in \Omega$ it holds that $k(\boldsymbol{x}, \boldsymbol{x}) \leq o^2$. Let $P_\ell(\cdot, \cdot)$ be a kernel approximation, such that a uniform convergence bound of the form

$$\sup_{\boldsymbol{x}, \boldsymbol{y} \in \Omega} |k(\boldsymbol{x}, \boldsymbol{y}) - P_\ell(\boldsymbol{x}, \boldsymbol{y})| \leq g(\ell) c_{unif}(d, \Omega, k) \tag{S46}$$

holds. Then it holds for the preconditioner $\hat{\boldsymbol{P}}_\ell = \sigma^2 \boldsymbol{I} + P_\ell(\boldsymbol{X}, \boldsymbol{X})$, that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq g(\ell) c(d, \Omega, k, n) \|\hat{\boldsymbol{K}}\|_F, \tag{S47}$$

where $c(d, \Omega, k, n) = \frac{\sqrt{n}}{o^2} c_{unif}(d, \Omega, k)$.

Proof. It holds that

$$\begin{aligned}
\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F = \|\boldsymbol{K} - \boldsymbol{P}_\ell\|_F &= \sqrt{\sum_{i,j=1}^{n} (k(\boldsymbol{x}_i, \boldsymbol{x}_j) - P_\ell(\boldsymbol{x}_i, \boldsymbol{x}_j))^2} \\
&\leq \sqrt{\sum_{i,j=1}^{n} (g(\ell) c_{\text{unif}}(d, \Omega, k))^2} && \text{uniform convergence bound (S46)} \\
&= n g(\ell) c_{\text{unif}}(d, \Omega, k) \\
&= g(\ell)\sqrt{n}\left(\sum_{i=1}^{n} \frac{k(\boldsymbol{x}_i, \boldsymbol{x}_i)^2}{k(\boldsymbol{x}_i, \boldsymbol{x}_i)^2}\right)^{\frac{1}{2}} && k \text{ bounded} \\
&\leq g(\ell)\frac{\sqrt{n}}{o^2}\left(\sum_{i=1}^{n} \boldsymbol{K}_{ij}^2\right)^{\frac{1}{2}} \\
&= g(\ell) c(d, \Omega, k, n) \|\boldsymbol{K}\|_F \\
&\leq g(\ell) c(d, \Omega, k, n) \|\hat{\boldsymbol{K}}\|_F
\end{aligned}$$

□

S5.3. Incomplete Cholesky Decomposition

Proposition S2 (Incomplete Cholesky Approximation Quality)
Let $k(\boldsymbol{x}, \boldsymbol{y}) = k(\|\boldsymbol{x} - \boldsymbol{y}\|)$ be a stationary kernel with output scale $o^2 = k(0) > 0$. Assume the kernel matrix spectrum decays at least exponentially, i.e. $\lambda_i(\boldsymbol{K}) \leq c\exp(-bi)$ for $c > 0$ and $b > \log(4)$. Then the incomplete Cholesky preconditioner $\hat{\boldsymbol{P}}_\ell = \sigma^2 \boldsymbol{I} + \boldsymbol{L}_\ell \boldsymbol{L}_\ell^\intercal$ satisfies

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq \frac{c\sqrt{n}}{o^2} \exp(-b'\ell) \|\boldsymbol{K}\|_F \tag{S48}$$

where $b' = b - \log(4) > 0$.

Proof. First note that since $k(\cdot, \cdot)$ is a positive definite kernel, the choice $o^2 = k(0)$ is no restriction. Now, it holds that

$$\begin{aligned}
&\lambda_i \leq c\exp(-bi) = c\exp(-(b' + \log(4))i) = c\exp(-b'i)4^{-i} \\
\iff\ &4^i \lambda_i \leq c\exp(-b'i).
\end{aligned}$$

Therefore by Theorem 3.2 of Harbrecht et al. (2012), we have $\operatorname{tr}(\boldsymbol{K} - \boldsymbol{L}_\ell \boldsymbol{L}_\ell^\intercal) \leq cn\exp(-b'\ell)$. Now it holds since $\boldsymbol{K} - \boldsymbol{L}_\ell \boldsymbol{L}_\ell^\intercal$ positive definite, that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F = \|\boldsymbol{K} - \boldsymbol{L}_\ell \boldsymbol{L}_\ell^\intercal\|_F \leq \operatorname{tr}(\boldsymbol{K} - \boldsymbol{L}_\ell \boldsymbol{L}_\ell^\intercal) \leq cn\exp(-b'\ell),$$

and with $\boldsymbol{K}_{ii} = o^2$ that

$$n = \sqrt{n}\left(\sum_{i=1}^{n} \frac{\boldsymbol{K}_{ii}^2}{o^4}\right)^{\frac{1}{2}} \leq \frac{\sqrt{n}}{o^2}\left(\sum_{i,j=1}^{n} \boldsymbol{K}_{ij}^2\right)^{\frac{1}{2}} = \frac{\sqrt{n}}{o^2}\|\boldsymbol{K}\|_F.$$

This concludes the argument. □

S5.4. Quadrature Fourier Features (QFF)

Proposition S3 (QFF Approximation Quality)
Assume $\Omega = [0, 1]^d$, k a kernel with Fourier transform $p(\omega) = \exp\left(-\frac{1}{2}\sum_{j=1}^{d} \omega_j^2 \gamma_j^2\right)$ such that Assumption 1 of Mutný & Krause (2018) is satisfied and let $\hat{\boldsymbol{P}}_\ell = \sigma^2 \boldsymbol{I} + \boldsymbol{P}_\ell$, where $\boldsymbol{P}_\ell$ is the QFF approximated kernel matrix. Let $\ell^{\frac{1}{d}} > \frac{2}{\gamma^2}$, then for $b = \frac{1}{2}(\log(4) - 1)$, it holds that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq c(d, n, k) \exp\left(-b\ell^{\frac{1}{d}}\right) \|\hat{\boldsymbol{K}}\|_F. \tag{S49}$$

Proof. By Theorem 1 of Mutnỳ & Krause (2018), replacing $\ell = (2\bar{m})^d$ it holds that

$$\begin{aligned}
\sup_{\boldsymbol{x},\boldsymbol{y}\in\Omega} |k(\boldsymbol{x},\boldsymbol{y}) - P_\ell(\boldsymbol{x},\boldsymbol{y})| &\leq d2^{d-1}\sqrt{\frac{\pi}{2}}\left(\frac{e}{2\gamma^2\ell^{\frac{1}{d}}}\right)^{\frac{1}{2}\ell^{\frac{1}{d}}} \\
&\leq c(d)\left(\frac{e}{4}\right)^{\frac{1}{2}\ell^{\frac{1}{d}}} && \ell^{\frac{1}{d}} > 2\gamma^{-2} \\
&= c(d)\exp\left(\frac{1}{2}\ell^{\frac{1}{d}} - \log(4)\frac{1}{2}\ell^{\frac{1}{d}}\right) \\
&= c(d)\exp\left(-\frac{1}{2}(\log(4)-1)\ell^{\frac{1}{d}}\right) \\
&= c(d)\exp\left(-b\ell^{\frac{1}{d}}\right)
\end{aligned}$$

Now by Lemma S6, we have

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq c(d,n)\exp\left(-b\ell^{\frac{1}{d}}\right)\|\hat{\boldsymbol{K}}\|_F, \tag{S50}$$

where $c(d,n) = \sqrt{n}c(d)$ by Assumption 1 of Mutnỳ & Krause (2018), which assumes $k(\boldsymbol{x},\boldsymbol{y}) \leq 1$. □

Proposition S4 (General QFF Approximation Quality)
Assume $\Omega = [0,1]^d$, k a kernel with Fourier transform $p(\omega)$ such that Assumption 1 of Mutnỳ & Krause (2018) is satisfied and $f_\delta(\phi) = p(\cot(\phi))\frac{\cos(\delta\cot(\phi))}{\sin(\phi)^2}$ is $(s-1)$-times absolutely continuous. Let $\hat{\boldsymbol{P}}_\ell = \sigma^2\boldsymbol{I} + \boldsymbol{P}_\ell$, where $\boldsymbol{P}_\ell$ is the QFF approximated kernel matrix. Then it holds that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq c(d,\Omega,k,n)\ell^{-\frac{s+1}{d}}\|\hat{\boldsymbol{K}}\|_F. \tag{S51}$$

Proof. By Theorem 4 of (Mutnỳ & Krause, 2018), replacing $\ell = (2\bar{m})^d$, it holds that

$$\begin{aligned}
\sup_{\boldsymbol{x},\boldsymbol{y}\in\Omega} |k(\boldsymbol{x},\boldsymbol{y}) - P_\ell(\boldsymbol{x},\boldsymbol{y})| &\leq d2^{d-1}\frac{(s+2)^{s+1}}{s!}\max_{\delta\in\Omega}\mathrm{TV}(f_\delta^{(s)})2^{s+1}\ell^{-\frac{s+1}{d}} \\
&= c(d,\Omega,k)\ell^{-\frac{s+1}{d}}
\end{aligned}$$

Now by Lemma S6, we have

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq c(d,\Omega,k,n)\ell^{-\frac{s+1}{d}}\|\hat{\boldsymbol{K}}\|_F.$$

□

Remark S2 (Modified Matérn Kernel)
Proposition S4 is satisfied for example for the modified Matérn(ν) kernel, defined via its spectral density

$$p(\omega) = \prod_{j=1}^{d}\frac{1}{(1+\gamma_j^2\omega_j^2)^{\nu+\frac{1}{2}}}. \tag{S52}$$

See the appendix of Mutnỳ & Krause (2018) for a detailed definition and motivation.

S5.5. Truncated Singular Value Decomposition

Proposition S5 (Truncated SVD Approximation Quality)
Let $\boldsymbol{K}$ be a kernel matrix and $\boldsymbol{P}_\ell = \boldsymbol{V}_\ell\boldsymbol{\Lambda}_\ell\boldsymbol{V}_\ell$ its truncated singular value decomposition consisting of the eigenvectors of the largest ℓ eigenvalues. Then it holds for $\hat{\boldsymbol{P}} = \sigma^2\boldsymbol{I} + \boldsymbol{P}_\ell$, that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq c(n)\ell^{-\frac{1}{2}}\|\hat{\boldsymbol{K}}\|_F. \tag{S53}$$

where $c(n) = \sqrt{n}$.

Proof. Since the optimal rank-ℓ approximation in Frobenius norm is given by the truncated SVD (Eckart & Young, 1936), we have for $\lambda_1(\boldsymbol{K}) \geq \cdots \geq \lambda_n(\boldsymbol{K})$, that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F^2 = \|\boldsymbol{K} - \boldsymbol{P}_\ell\|_F^2 = \sum_{i=\ell+1}^{n} \lambda_i(\boldsymbol{K})^2$$

Now since $\lambda_{\ell+1}(\boldsymbol{K}) \leq \frac{1}{\ell}\sum_{i=1}^{\ell} \lambda_i(\boldsymbol{K}) \leq \frac{1}{\ell}\operatorname{tr}(\boldsymbol{K})$, we have

$$\leq \lambda_{\ell+1}(\boldsymbol{K}) \sum_{i=\ell+1}^{n} \lambda_i(\boldsymbol{K}) \leq \frac{1}{\ell}\operatorname{tr}(\boldsymbol{K})^2 = \frac{1}{\ell}\|\boldsymbol{\lambda}\|_1^2 \leq \frac{n}{\ell}\|\boldsymbol{\lambda}\|_2^2 = \frac{n}{\ell}\|\boldsymbol{K}\|_F^2 \leq \frac{n}{\ell}\|\hat{\boldsymbol{K}}\|_F^2$$

where $\boldsymbol{\lambda} = (\lambda_1(\boldsymbol{K}), \ldots, \lambda_n(\boldsymbol{K}))^\intercal \in \mathbb{R}^n$. □

S5.6. Randomized Singular Value Decomposition

Proposition S6 (Randomized SVD Approximation Quality)
Let $\boldsymbol{K}$ be a kernel matrix and $\boldsymbol{P}_\ell = \boldsymbol{H}_\ell \boldsymbol{H}_\ell^\intercal \boldsymbol{K}$ its randomized singular value decomposition constructed via the LINEARTIMESVD *algorithm (Drineas et al., 2006) with $s \in \mathbb{N}$ samples drawn according to probabilities $\{p_i\}_{i=1}^n$. Then for $\hat{\boldsymbol{P}} = \sigma^2\boldsymbol{I} + \boldsymbol{P}_\ell$, it holds with probability $1-\delta$, that*

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq c(n)(\ell^{-\frac{1}{2}} + \mathcal{O}(\ell^{\frac{1}{4}} s^{-\frac{1}{4}}))\|\hat{\boldsymbol{K}}\|_F, \tag{S54}$$

where $c(n) = \sqrt{n}$.

Proof. By Drineas et al. (2006, Theorem 4) it holds with probability $1-\delta$, that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F^2 = \|\boldsymbol{K} - \boldsymbol{P}_\ell\|_F^2 \leq \|\boldsymbol{K} - \boldsymbol{K}_\ell\|_F^2 + c(p_i, \delta)\ell^{\frac{1}{2}} s^{-\frac{1}{2}}\|\boldsymbol{K}\|_F^2$$

By the same argument as in the proof of Proposition S5 for the error of the optimal rank-ℓ approximation $\|\boldsymbol{K} - \boldsymbol{K}_\ell\|_F^2$, we obtain

$$\begin{aligned}
&\leq \frac{n}{\ell}\|\boldsymbol{K}\|_F^2 + c(p_i, \delta)\ell^{\frac{1}{2}} s^{-\frac{1}{2}}\|\boldsymbol{K}\|_F^2 \\
&\leq n(\ell^{-1} + \mathcal{O}(\ell^{\frac{1}{2}} s^{-\frac{1}{2}}))\|\hat{\boldsymbol{K}}\|_F^2
\end{aligned}$$

This completes the proof. □

S5.7. Randomized Nyström Method

Proposition S7 (Randomized Nyström Approximation Quality)
Let $\boldsymbol{K}$ be a kernel matrix and $\boldsymbol{P}_\ell = \boldsymbol{C}\boldsymbol{W}_\ell^+\boldsymbol{C}^\intercal$ its randomized Nyström approximation constructed via Algorithm 3 of Drineas et al. (2005) with $s \in \mathbb{N}$ columns drawn according to probabilities $p_i = \frac{\boldsymbol{K}_{ii}^2}{\sum_{i=1}^n \boldsymbol{K}_{ii}^2}$. Then for $\hat{\boldsymbol{P}} = \sigma^2\boldsymbol{I} + \boldsymbol{P}_\ell$, it holds with probability $1-\delta$, that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq c(n)(\ell^{-\frac{1}{2}} + \mathcal{O}(\ell^{\frac{1}{4}} s^{-\frac{1}{4}}))\|\hat{\boldsymbol{K}}\|_F, \tag{S55}$$

where $c(n) = \sqrt{n}$.

Proof. By Drineas et al. (2005, Theorem 3) it holds with probability $1-\delta$, that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F = \|\boldsymbol{K} - \boldsymbol{P}_\ell\|_F \leq \|\boldsymbol{K} - \boldsymbol{K}_\ell\|_F + c(\delta)\ell^{\frac{1}{4}} s^{-\frac{1}{4}} \sum_{i=1}^{n} \boldsymbol{K}_{ii}^2$$

By the same argument as in the proof of Proposition S5 for the error of the optimal rank-ℓ approximation $\|\boldsymbol{K} - \boldsymbol{K}_\ell\|_F$, we obtain

$$\begin{aligned}
&\leq n^{\frac{1}{2}}\ell^{-\frac{1}{2}}\|\boldsymbol{K}\|_F^2 + c(\delta)\ell^{\frac{1}{4}}s^{-\frac{1}{4}}\sum_{i=1}^{n}\boldsymbol{K}_{ii}^2 \\
&\leq n^{\frac{1}{2}}\ell^{-\frac{1}{2}}\|\boldsymbol{K}\|_F^2 + c(\delta)\ell^{\frac{1}{4}}s^{-\frac{1}{4}}\sqrt{\sum_{i=1}^{n}\boldsymbol{K}_{ii}^2}\|\boldsymbol{K}\|_F \\
&\leq n^{\frac{1}{2}}(\ell^{-\frac{1}{2}} + \mathcal{O}(\ell^{\frac{1}{4}}s^{-\frac{1}{4}}))\|\hat{\boldsymbol{K}}\|_F
\end{aligned}$$

This completes the proof. □

S5.8. Random Fourier Features (RFF)

Proposition S8 (RFF Approximation Quality)
Let $k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{x} - \boldsymbol{y})$ be a positive-definite kernel with compact data domain $\Omega \subset \mathbb{R}^d$. Let $\boldsymbol{P}_\ell = \boldsymbol{Z}_\ell\boldsymbol{Z}_\ell^\intercal$ be the random Fourier feature approximation (Rahimi et al., 2007), where $\boldsymbol{Z}_\ell \in \mathbb{R}^{n\times\ell}$. Then for $\hat{\boldsymbol{P}}_\ell = \sigma^2\boldsymbol{I} + \boldsymbol{P}_\ell$ it holds with probability $1-\delta$, that

$$\|\hat{\boldsymbol{K}} - \hat{\boldsymbol{P}}_\ell\|_F \leq c(d, \Omega, k, \delta, n)\ell^{-\frac{1}{2}}\|\hat{\boldsymbol{K}}\|_F. \tag{S56}$$

Proof. By Theorem 1 of Sriperumbudur & Szabó (2015) with probability $1-\delta$, we obtain the uniform convergence bound

$$\sup_{\boldsymbol{x},\boldsymbol{y}\in\Omega}|k(\boldsymbol{x}, \boldsymbol{y}) - P_\ell(\boldsymbol{x}, \boldsymbol{y})| \leq c_{\text{unif}}(d, \Omega, k, \delta)\ell^{-\frac{1}{2}}. \tag{S57}$$

Now, applying Lemma S6 completes the proof. □

S6. Technical Results

Lemma S7
Let $x \in \mathbb{R}$ such that $x > 1$, then it holds that

$$\frac{x-1}{x+1} \leq \exp\left(-\frac{2}{x}\right). \tag{S58}$$

Proof. By Mitrinovic & Vasic (1970, Section 3.6.18) it holds for $y > 0$, that

$$\frac{\log(y+1)}{y} \geq \frac{2}{2+y}$$

Substituting $y = \frac{x+1}{x-1} - 1$ we obtain

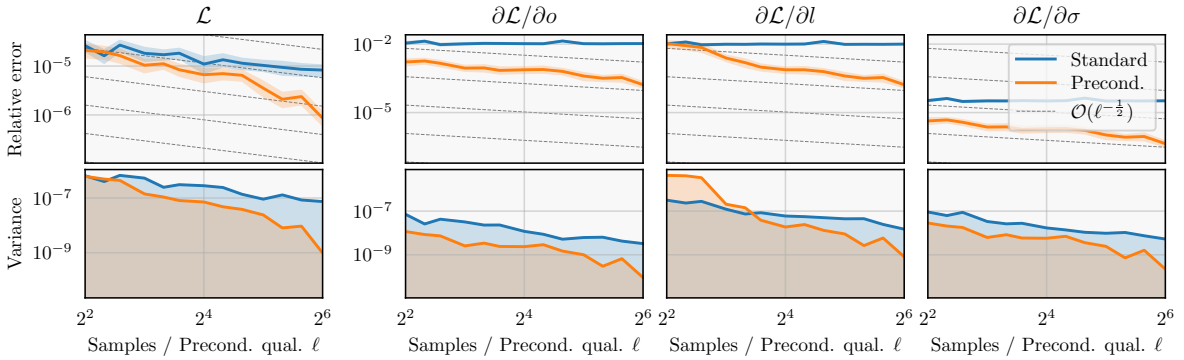
$$\begin{aligned}
\log\left(\frac{x+1}{x-1}\right) \geq \frac{2(\frac{x+1}{x-1} - 1)}{1 + \frac{x+1}{x-1}} = \frac{2(x+1-x+1)}{2x} = \frac{2}{x} \\
\iff \frac{x+1}{x-1} \geq \exp\left(\frac{2}{x}\right)
\end{aligned}$$

This proves the claim. □

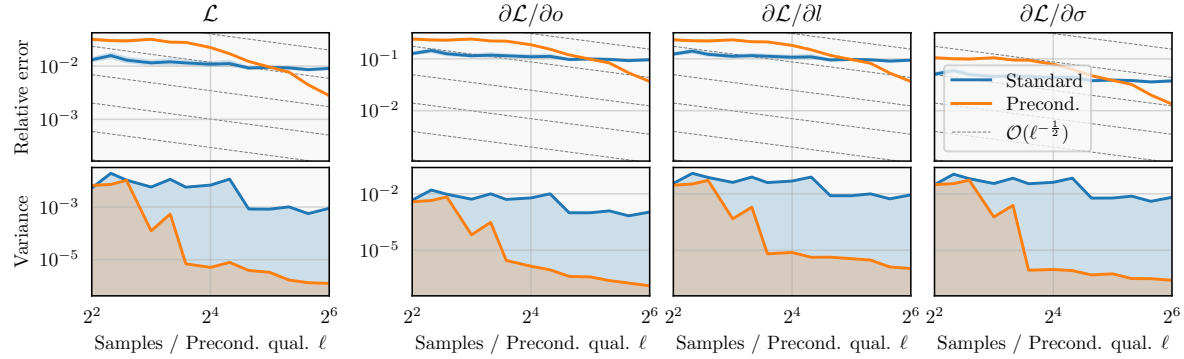
S7. Additional Experimental Results

S7.1. Synthetic Data

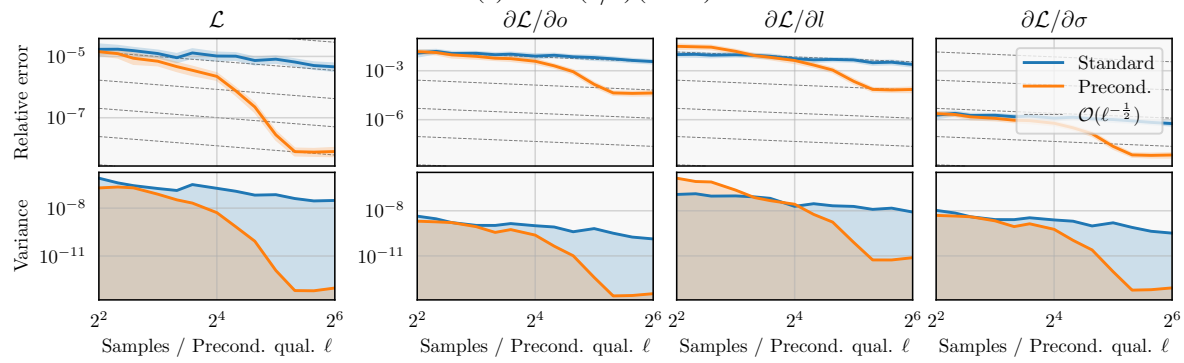
We report bias and variance of the stochastic estimators for the log-maginal likelihood and its derivatives for the exponentiated quadratic, Matérn($\frac{3}{2}$) and rational quadratic kernel on a synthetic dataset of size $n = 10{,}000$ with varying dimensionality $d \in \{1, 2, 3\}$ in Table S1. Using $\ell = 128$ random samples with a preconditioner of the same size bias and variance are reduced by several orders of magnitude across different kernels. Note, that the variance reduction tends to decline with dimensionality, even though this is not necessarily universal across kernels. We show the bias and variance reduction for increasing number of random samples, respectively preconditioner size in Figure S1.



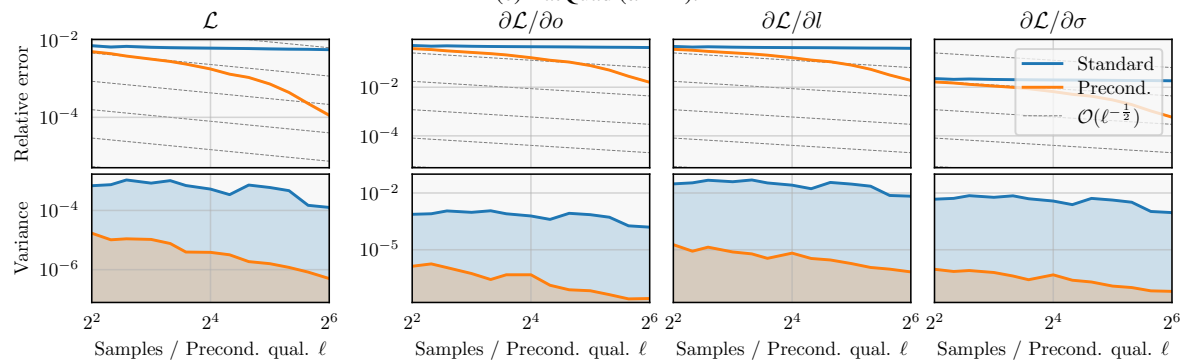
(a) Matérn(3/2) ($d = 1$).



(b) Matérn(3/2) ($d = 3$).



(c) RatQuad ($d = 1$).



(d) RatQuad ($d = 3$).

Figure S1: *Bias and variance decrease on synthetic datasets for different kernels.* Relative error and variance of the stochastic estimators of the log-marginal likelihood and its derivative for increasing number of random vectors, equivalently preconditioner size. Experiments were performed for different kernels on a synthetic dataset of size $n = 10{,}000$ with dimension $d \in \{1, 2, 3\}$. Plots show mean and 95% confidence intervals for the relative error computed over 25 repetitions.

Table S1: *Bias and variance reduction for different kernels.* Bias and variance of the stochastic estimators for the log-marginal likelihood and its derivative(s) computed for synthetic data ($n = 10{,}000$, $\sigma^2 = 10^{-2}$) with 25 repetitions.

			$\mathcal{L}$		$\partial\mathcal{L}/\partial o$		$\partial\mathcal{L}/\partial l$		$\partial\mathcal{L}/\partial\sigma$	
			Bias	Var.	Bias	Var.	Bias	Var.	Bias	Var.
Kernel	d	Prec. Qual.								
Matérn$(3/2)$	1	0	3e−4	3e−8	9e−4	1e−9	2e−3	8e−9	2e−3	2e−9
		128	9e−6	4e−11	4e−6	8e−12	1e−5	7e−11	7e−6	2e−11
	2	0	3e−1	3e−3	4e−1	4e−3	1	3e−2	9e−1	3e−2
		128	3e−4	5e−8	7e−5	3e−9	2e−4	3e−8	1e−4	7e−9
	3	0	3e−1	6e−4	4e−1	7e−4	1	6e−3	8e−1	4e−3
		128	7e−3	7e−7	2e−2	3e−8	5e−2	4e−7	3e−2	8e−8
RBF	1	0	1e−4	1e−8	1e−5	4e−11	1e−4	3e−9	2e−5	1e−10
		128	5e−8	1e−15	3e−8	4e−16	7e−7	2e−13	4e−8	1e−15
	2	0	3e−3	7e−8	5e−3	2e−9	8e−2	1e−7	1e−2	2e−9
		128	1e−6	1e−12	5e−7	2e−13	1e−5	1e−10	8e−7	5e−13
	3	0	1e−1	2e−7	2e−1	4e−8	2	9e−7	4e−1	8e−9
		128	3e−4	5e−8	5e−5	2e−9	7e−4	3e−7	8e−5	4e−9
RatQuad	1	0	1e−4	1e−8	1e−5	1e−10	1e−4	5e−9	2e−5	3e−10
		128	3e−7	6e−14	2e−7	2e−14	2e−6	4e−12	2e−7	4e−14
	2	0	3e−2	2e−7	4e−2	3e−8	3e−1	2e−7	1e−1	7e−9
		128	8e−5	3e−9	2e−5	2e−10	2e−4	1e−8	3e−5	5e−10
	3	0	2e−1	2e−4	3e−1	3e−4	2	1e−2	8e−1	2e−3
		128	4e−4	1e−7	2e−4	1e−8	4e−3	4e−7	4e−4	4e−8

Table S2: *Hyperparameter optimization using Adam.* GP regression using a Matérn$(\frac{3}{2})$ kernel and incomplete Cholesky preconditioner of size 500 with $\ell = 50$ random samples. Hyperparameters were optimized with Adam for at most 20 steps using early stopping via a validation set.

Dataset	n	d	Prec. Size	$-\mathcal{L}_{\text{train}} \downarrow$	$-\mathcal{L}_{\text{test}} \downarrow$	RMSE $\downarrow$	Runtime (s)
Elevators	12,449	18	500	0.4803	0.4593	0.3684	109
Bike	13,034	17	500	0.2265	0.3473	0.2300	64
Kin40k	30,000	8	500	0.4392	−0.1200	0.0982	159
Protein	34,297	9	500	0.9438	0.9319	0.5681	92
KEGGdir	36,620	20	500	-1.0070	−1.0390	0.0810	239

S7.2. UCI Datasets

For the experiments we conducted on UCI datasets, we report the full experimental results with their deviation across 10 runs in Table S3. Test errors with and without preconditioning did not differ by more than two standard deviations. However, model evaluations of the optimizer were significantly reduced when using a preconditioner of size 500, leading to substantial speedup. Note, that the experiment on the "3DRoad" dataset was only carried out once due to the prohibitive runtime without preconditioning.

We used the L-BFGS optimizer in our experiments due to its favorable convergence properties. As an ablation experiment we compared to the Adam optimizer as sometimes used for its robustness to noise, when using stochastic approximations of the log-marginal likelihood (Gardner et al., 2018; Wang et al., 2019). We find that with preconditioning optimization with L-BFGS significantly outperformed optimization with Adam, both in terms of training and test error, except for the "KEGGdir" dataset (cf. Table 2 and Table S2). Additionally, L-BFGS converged faster across all experiments. This shows that variance reduction via preconditioning makes the use of second-order optimizers not only possible, but preferred for GP hyperparameter optimization when using stochastic approximations.

Table S3: *Hyperparameter optimization on UCI datasets.* GP regression using a Matérn($\frac{3}{2}$) kernel and incomplete Cholesky preconditioner of size 500 with $\ell = 50$ random samples. Hyperparameters were optimized with L-BFGS for at most 20 steps using early stopping via a validation set. All results, but "3DRoad", are averaged over 10 runs.

					Model evals.		Runtime (s)		Speedup		$-\mathcal{L}_{\text{train}} \downarrow$		$-\mathcal{L}_{\text{test}} \downarrow$		RMSE $\downarrow$	
					mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Dataset	n	d	Prec.	Qual. Opt. Steps												
Elevators	12,449	18	0	19	42.2	7.5	53.0	7.7	1.0	0.0	0.4647	0.0035	0.4021	0.0102	0.3484	0.0052
			500	19	36.4	1.2	39.2	0.8	1.4	0.2	0.4377	0.0030	0.4022	0.0097	0.3482	0.0053
Bike	13,034	17	0	19	32.3	1.3	30.6	1.1	1.0	0.0	–0.9976	0.0114	–0.9934	0.0137	0.0446	0.0040
			500	19	31.4	2.1	37.1	1.5	0.8	0.0	–0.9985	0.0186	–0.9877	0.0180	0.0454	0.0032
Kin40k	30,000	8	0	8	19.0	4.4	186.5	68.0	1.0	0.0	–0.3339	0.0018	–0.3141	0.0016	0.0929	0.0014
			500	6	15.0	0.4	44.6	1.3	2.7	0.1	–0.4332	0.0055	–0.3135	0.0038	0.0949	0.0015
Protein	34,297	9	0	15	124.2	0.4	892.6	19.1	1.0	0.0	0.9963	0.0031	0.8869	0.0081	0.5722	0.0065
			500	7	17.8	0.6	42.5	5.0	4.2	0.4	0.9273	0.0044	0.8835	0.0053	0.5577	0.0079
KEGGdir	36,620	20	0	19	55.8	11.7	1450.3	253.4	1.0	0.0	–0.9501	0.0087	–0.9459	0.0343	0.0861	0.0043
			500	19	42.9	0.7	173.7	2.9	8.4	1.5	–1.0043	0.0093	–0.9490	0.0312	0.0864	0.0042
3DRoad	326,155	3	0	9	68.0		82,200.0		1.0		0.7733		1.4360		0.2982	
			500	9	19.0		7306.0		11.3		0.1284		1.1690		0.1265	