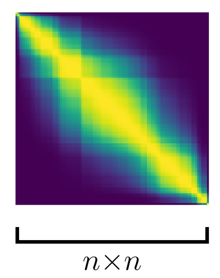
# Preconditioning for Scalable Gaussian Process Hyperparameter Optimization

Jonathan Wenger[1,2,3] Geoff Pleiss[3] Philipp Hennig[1,2] John Cunningham[3] Jacob Gardner[4]

[1] University of Tübingen
[2] Max Planck Institute for Intelligent Systems, Tübingen
[3] Columbia University
[4] University of Pennsylvania

## Goal: Scalable GP Hyperparam. Optim.

Need to: Evaluate log-marginal likelihood $\mathcal{L}$ and its derivative $\frac{\partial}{\partial\theta}\mathcal{L}$ repeatedly.

$\hat{\boldsymbol{K}} =$ 


Challenge: Costly $\mathcal{O}(n^3)$ operations with the kernel matrix.

- ▷ linear solves $\hat{\boldsymbol{K}}^{-1}(\cdot)$
- ▷ matrix traces $\log\det(\hat{\boldsymbol{K}}) = \operatorname{tr}(\log(\hat{\boldsymbol{K}}))$ and $\operatorname{tr}\left(\hat{\boldsymbol{K}}^{-1}\frac{\partial\hat{\boldsymbol{K}}}{\partial\theta}\right)$

## Known: Reducable to Matrix-Vector Mult.

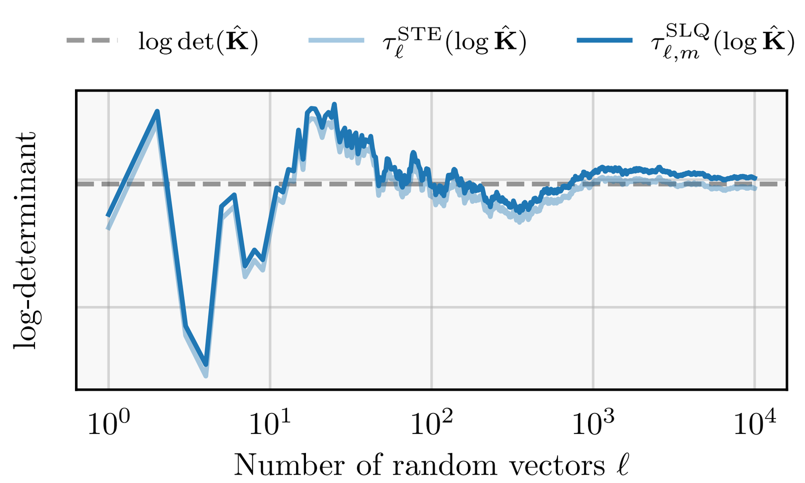
**Linear Solves**: Preconditioned CG **Matrix Traces**: Stochastic trace estimator

Great because . . .

- ▷ matrix-vector multiplies cost at most $\mathcal{O}(n^2)$
- ▷ no need to store kernel matrix in memory
- ▷ can exploit parallelization and modern hardware (GPUs)

lower time and space complexity

## Problem: *Stochastic* Trace Estimators




$$\operatorname{tr}(f(\hat{\boldsymbol{K}})) = n\mathbb{E}[\boldsymbol{z}_i^\intercal f(\hat{\boldsymbol{K}})\boldsymbol{z}_i]$$
$$\approx \tau_\ell^{\mathrm{STE}}(f(\hat{\boldsymbol{K}})) = \frac{n}{\ell}\sum_{i=1}^{\ell}\boldsymbol{z}_i^\intercal f(\hat{\boldsymbol{K}})\boldsymbol{z}_i$$
$$\approx \tau_{\ell,m}^{\mathrm{SLQ}}(f(\hat{\boldsymbol{K}}))$$

Bad because. . .

- ▷ slow $\mathcal{O}(\ell^{-\frac{1}{2}})$ convergence in number of random vectors
- ▷ adds noise into hyperparameter optimization

slows down training

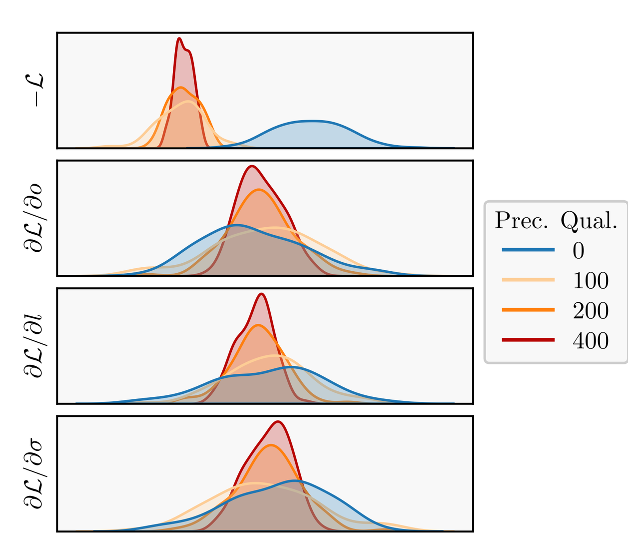
## Our work: Precondition Trace Estimators

### Insight

Can precondition not only linear solves but also stochastic trace estimators!

### Contributions

- ▷ Preconditioning reduces variance of the STE, i.e. accelerates convergence.
- ▷ Theoretical guarantees.
- ▷ Preconditioner choices for given kernels.
- ▷ Up to twelvefold training speedup.

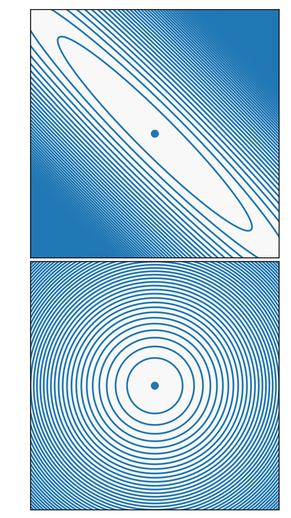



## Background: Preconditioning

$$\hat{\boldsymbol{P}} \approx \hat{\boldsymbol{K}}$$

such that $\kappa(\hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{K}}) \ll \kappa(\hat{\boldsymbol{K}})$ and $\hat{\boldsymbol{P}}$ is tractable.

- ▷ Computing and storing $\hat{\boldsymbol{P}}$ is cheap.
- ▷ Linear solves $\boldsymbol{v} \mapsto \hat{\boldsymbol{P}}^{-1}\boldsymbol{v}$ are efficient.
- ▷ Derived properties (determinant, spectrum, . . . ) known.



Asymptotic approx. error $g(\ell) \to 0$ of precond. seq. $\hat{\boldsymbol{P}}_\ell \to \hat{\boldsymbol{K}}$:

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

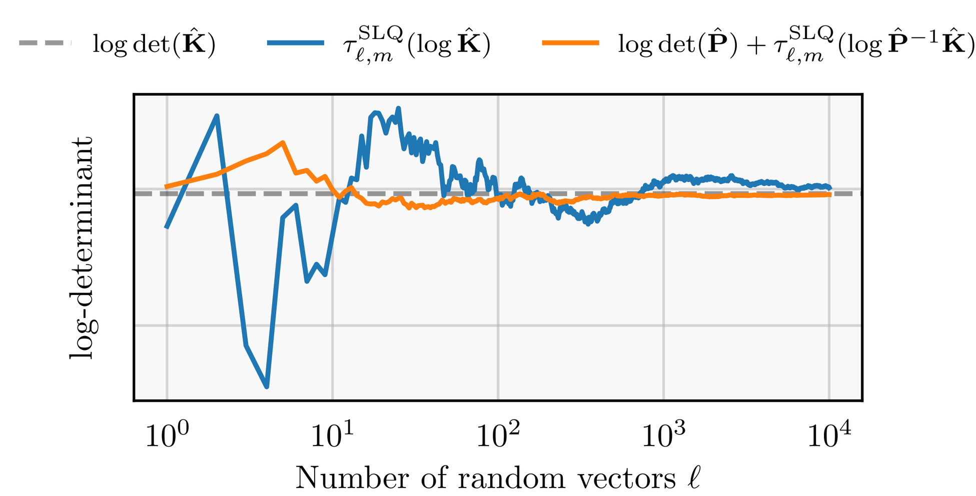
Known Use: Accelerate and stabilize linear solves via CG ⇒ bias reduction

## Precond. Log-Determinant Estimation

**Idea:** Decompose $\log$-determinant into deterministic and stochastic approximation.

$$\log\det(\hat{\boldsymbol{K}}) = \log\det\left(\hat{\boldsymbol{P}}_\ell\hat{\boldsymbol{P}}_\ell^{-1}\hat{\boldsymbol{K}}\right) = \underbrace{\log\det(\hat{\boldsymbol{P}}_\ell)}_{\text{known}} + \underbrace{\operatorname{tr}(\log(\hat{\boldsymbol{K}}) - \log(\hat{\boldsymbol{P}}_\ell))}_{\approx \text{ stochastic trace estimate (STE)}}$$

Better preconditioner ⇒ smaller stochastic approximation ⇒ variance reduction




- ▷ Backward pass analogously via automatic differentiation.
- ▷ If we compute a preconditioner for CG, we can simply reuse it at negligible overhead.

If $\hat{\boldsymbol{P}}_\ell \to \hat{\boldsymbol{K}}$ at rate $g(\ell)$, then the STE only requires $\mathcal{O}(\ell^{-\frac{1}{2}}g(\ell))$ random vectors.

## Theoretical Results

### Probabilistic Error Bounds

Preconditioning not only reduces bias, but crucially also reduces variance.

***Theorem (Log-marg. likelihood)***
*[. . . ] Then with probability $1-\delta$, the error in the estimate $\eta$ 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 errors are bounded by*

$$\varepsilon_{\mathrm{CG}}(\kappa, m) \leq K_3\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \quad (1)$$
$$\varepsilon_{\mathrm{Lanczos}}(\kappa, m) \leq K_1\left(\frac{\sqrt{2\kappa+1}-1}{\sqrt{2\kappa+1}+1}\right)^{2m} \quad (2)$$
$$\boxed{\varepsilon_{\mathrm{STE}}(\delta, \ell) \leq C_1\sqrt{\log\left(\delta^{-1}\right)}\ell^{-\frac{1}{2}}g(\ell)} \quad (3)$$

***Theorem (Derivative)***
*[. . . ] Then with probability $1-\delta$, the error in the estimate $\phi$ of the derivative of the $\log$-marginal likelihood $\frac{\partial}{\partial\theta}\mathcal{L}$ satisfies*

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

*where the errors are bounded by*

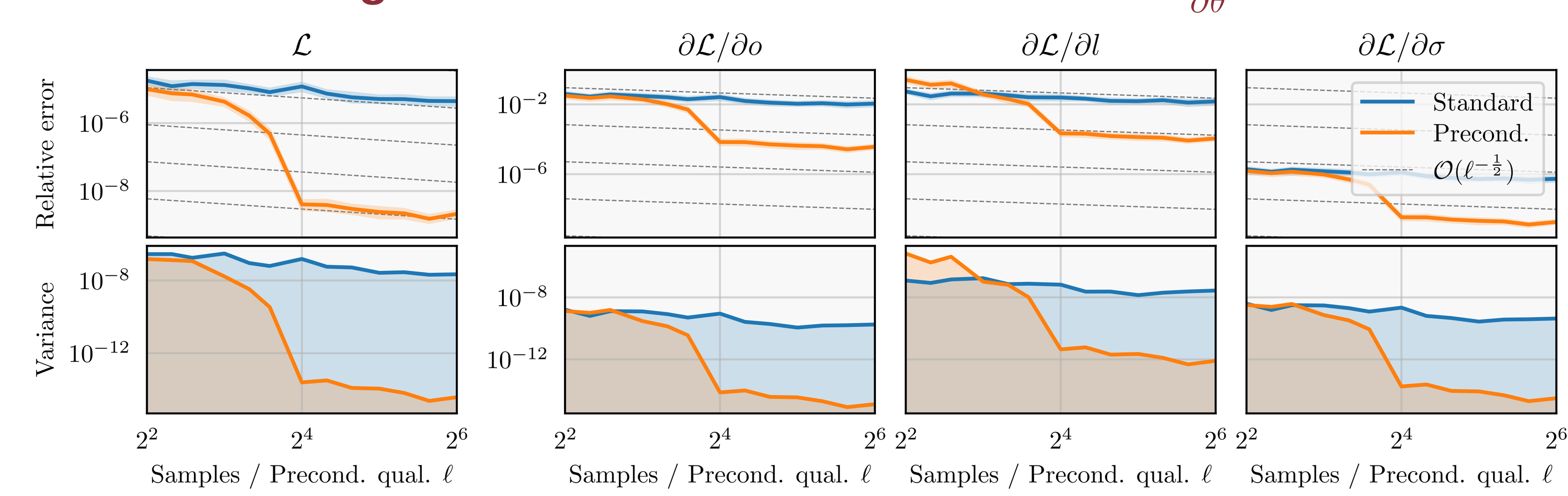
$$\varepsilon_{\mathrm{CG}}(\kappa, m) \leq K_4\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \quad (4)$$
$$\varepsilon_{\mathrm{CG}'}(\kappa, m) \leq K_2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m \quad (5)$$
$$\boxed{\varepsilon_{\mathrm{STE}}(\delta, \ell) \leq C_1\sqrt{\log\left(\delta^{-1}\right)}\ell^{-\frac{1}{2}}g(\ell)} \quad (6)$$

### Convergence rates for combinations of kernels and preconditioners

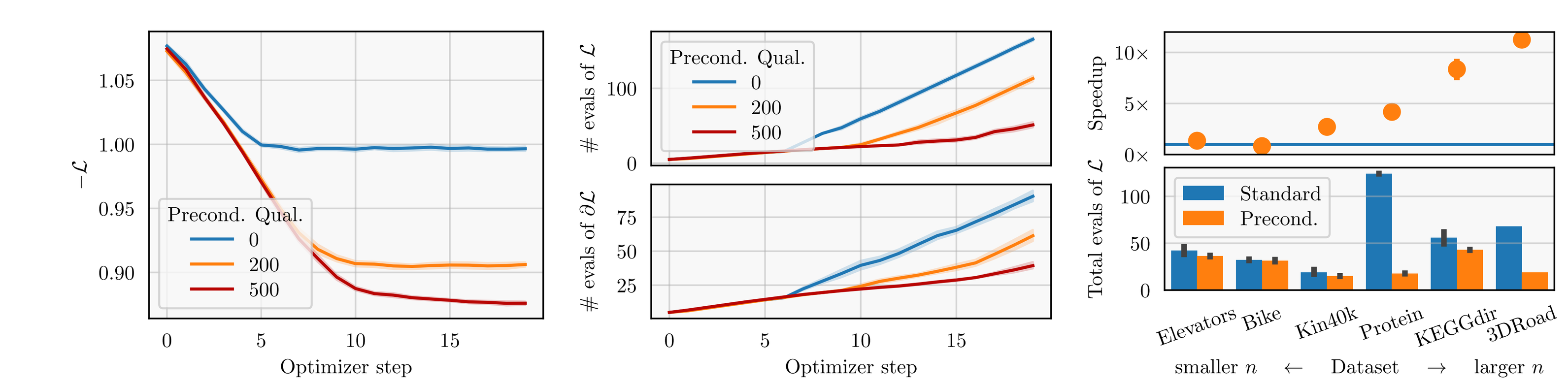
| Kernel | $d$ | Preconditioner | $g(\ell)$ | Condition |
|---|---|---|---|---|
| any | $\mathbb{N}$ | none | $1$ | |
| . . . | . . . | . . . | . . . | . . . |
| any | $\mathbb{N}$ | RFF | $\ell^{-\frac{1}{2}}$ | w/ high probability |
| RBF | $1$ | partial Cholesky | $\exp(-c\ell)$ | for some $c > 0$ |
| RBF | $\mathbb{N}$ | QFF | $\exp(-b\ell^{\frac{1}{d}})$ | for some $b > 0$ if $\ell^{\frac{1}{d}} > 2\gamma^{-2}$ |
| Matérn($\nu$) | $\mathbb{N}$ | partial Cholesky | $\ell^{-(\frac{2\nu}{d}+1)}$ | $2\nu \in \mathbb{N}$ and maximin ordering |
| Matérn($\nu$) | $1$ | QFF | $\ell^{-(s(\nu)+1)}$ | where $s(\nu) \in \mathbb{N}$ |
| mod. Matérn($\nu$) | $\mathbb{N}$ | QFF | $\ell^{-\frac{s(\nu)+1}{d}}$ | where $s(\nu) \in \mathbb{N}$ |
| additive | $\mathbb{N}$ | any | $dg(\ell)$ | all summands have rate $g(\ell)$ |
| any | $\mathbb{N}$ | any kernel approx. | $g(\ell)$ | $\exists$ uniform convergence bound |

## Experiments

### Preconditioning reduces bias and variance in $\mathcal{L}$ and $\frac{\partial}{\partial\theta}\mathcal{L}$




### Preconditioning reduces noise ⇒ accelerates hyperparam. optim.




(a) Training loss (“Protein”). (b) Line search comput. (“Protein”). (c) Speedup on UCI datasets.

| Dataset | $n$ | $d$ | $-\mathcal{L}_{\mathrm{train}}\downarrow$ Standard | $-\mathcal{L}_{\mathrm{train}}\downarrow$ Precond. | $-\mathcal{L}_{\mathrm{test}}\downarrow$ Standard | $-\mathcal{L}_{\mathrm{test}}\downarrow$ Precond. | RMSE ↓ Standard | RMSE ↓ Precond. | Runtime (s) Standard | Runtime (s) 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** |

**Paper** on ARXIV

**Implementation** as part of GPYTORCH