Turbocharging Gaussian Process Inference with Approximate Sketch-and-Project

Pratik Rathore[1], Zachary Frangella[1], Sachin Garg[2], Shaghayegh Fazliani[1], Michał Dereziński[2], Madeleine Udell[1]

November 5, 2025

[1]Stanford University [2]University of Michigan

Motivation

- Gaussian processes (GPs) are a mainstay of modern ML and scientific computing
- GP inference requires solving the linear system $(K + \lambda I)w = y$
- Algorithms for large-scale GP inference require trade-offs between quality of inference, convergence speed, ease of setting hyperparameters, and scalability
- Our solution: **a**pproximate, **d**istributed, **a**ccelerated **s**ketch-**a**nd-**p**roject (ADASAP)

Algorithm 1 ADASAP

Require: distribution of row indices $\mathcal{D}$ over $[n]^b$, distribution over Nyström sketching matrices $\mathcal{D}_{\mathrm{Nyström}}$ over $\mathbb{R}^{b\times r}$, number of iterations T_{max}, initialization $W_0 \in \mathbb{R}^{n\times n_{\mathrm{rhs}}}$, workers $\{\mathcal{W}_1, \ldots, \mathcal{W}_{\mathrm{N_{workers}}}\}$, acceleration parameters μ, ν

Initialize acceleration parameters
$\beta \leftarrow 1 - \sqrt{\mu/\nu}, \quad \gamma \leftarrow 1/\sqrt{\mu\nu}, \quad \alpha \leftarrow 1/(1+\gamma\nu)$
$V_0 \leftarrow W_0, Z_0 \leftarrow W_0$

for $t = 0, 1, \ldots, T_{\mathrm{max}} - 1$ **do**

Step I: Compute matrix-matrix product $K_{\mathcal{B}n}W_t$ ▷ Costs $\mathcal{O}(bnn_{\mathrm{rhs}}/\mathrm{N_{workers}})$
Sample row indices $\mathcal{B}$ of size b from $[n]$ according to $\mathcal{D}$
$K_{\mathcal{B}n}W_t \leftarrow$ `ColDistMatMat`$(W_t, \mathcal{B}, \{\mathcal{W}_1, \ldots, \mathcal{W}_{\mathrm{N_{workers}}}\})$

Step II: Compute Nyström sketch $K_{\mathcal{B}\mathcal{B}}\Omega$ ▷ Costs $\mathcal{O}(rb^2/\mathrm{N_{workers}})$
Sample $\Omega \in \mathbb{R}^{b\times n}$ from $\mathcal{D}_{\mathrm{Nyström}}$
$K_{\mathcal{B}\mathcal{B}}\Omega \leftarrow$ `RowDistMatMat`$(\Omega, \mathcal{B}, \{\mathcal{W}_1, \ldots, \mathcal{W}_{\mathrm{N_{workers}}}\})$

Step III: Compute Nyström approximation and get stepsize ▷ Costs $\mathcal{O}(b^2 + br^2)$
$U, S \leftarrow$ `RandNysAppx`$(K_{\mathcal{B}\mathcal{B}}\Omega, \Omega, r)$
$P_{\mathcal{B}} \leftarrow USU^T + (S_{rr} + \lambda)I$ ▷ Get preconditioner. Never explicitly formed!
$\eta_{\mathcal{B}} \leftarrow$ `GetStepsize`$(P_{\mathcal{B}}, K_{\mathcal{B}\mathcal{B}} + \lambda I)$

Step IV: Compute updates using acceleration ▷ Costs $\mathcal{O}(brn_{\mathrm{rhs}} + nn_{\mathrm{rhs}})$
$D_t \leftarrow I_{\mathcal{B}}^T P_{\mathcal{B}}^{-1}(K_{\mathcal{B}n}W_t + \lambda I_{\mathcal{B}}W_t - Y_{\mathcal{B}})$
$W_{t+1}, V_{t+1}, Z_{t+1} \leftarrow$ `NestAcc`$(W_t, V_t, Z_t, D_t, \eta_{\mathcal{B}}, \beta, \gamma, \alpha)$

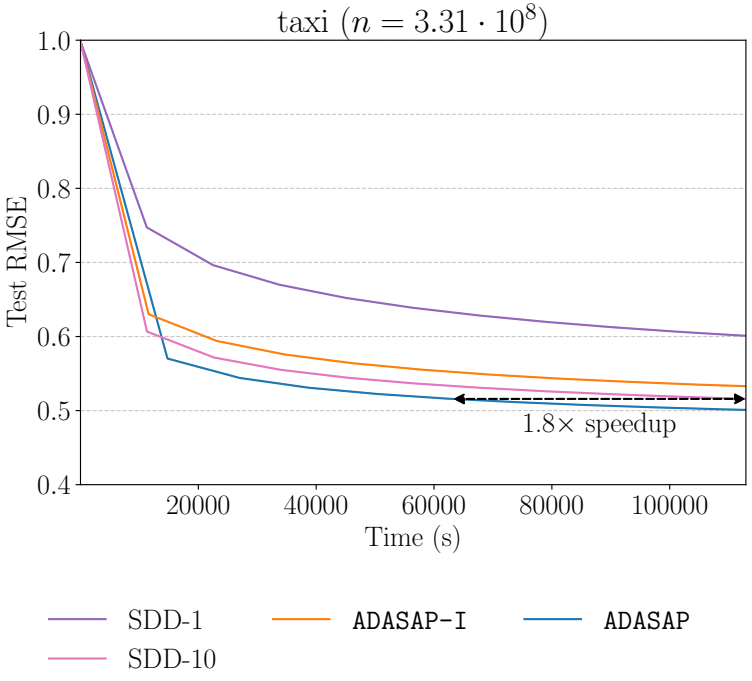
end for
return W_T ▷ Approximate solution to $K_\lambda W = Y$

ADASAP outperforms state-of-the-art on benchmarks

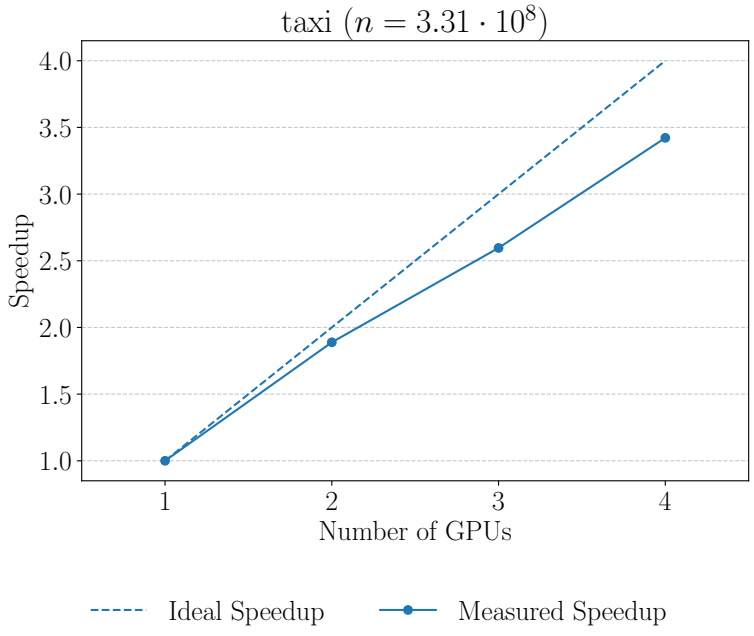
Table 1: RMSE and mean negative log-likelihood (NLL) obtained by **ADASAP** and competitors on the test set.

	Dataset n	yolanda $3.60 \cdot 10^5$	song $4.64 \cdot 10^5$	benzene $5.65 \cdot 10^5$	malonaldehyde $8.94 \cdot 10^5$	acsincome $1.50 \cdot 10^6$	houseelec $1.84 \cdot 10^6$
Test RMSE	**ADASAP**	**0.795**	**0.752**	**0.012**	**0.015**	**0.789**	**0.027**
	ADASAP-I	0.808	0.782	0.168	0.231	**0.795**	0.066
	SDD-1	0.833	0.808	0.265	0.270	0.801	0.268
	SDD-10	**0.801**	0.767	0.112	Diverged	**0.792**	0.119
	SDD-100	Diverged	Diverged	Diverged	Diverged	Diverged	Diverged
	PCG	**0.795**	**0.752**	0.141	0.273	0.875	1.278
Test Mean NLL	**ADASAP**	**1.179**	**1.121**	**-2.673**	**-2.259**	**1.229**	**-2.346**
	ADASAP-I	1.196	1.170	-0.217	0.466	**1.235**	-2.185
	SDD-1	1.225	1.203	0.531	0.903	1.242	-0.281
	SDD-10	**1.187**	1.149	-0.762	Diverged	**1.232**	-1.804
	SDD-100	Diverged	Diverged	Diverged	Diverged	Diverged	Diverged
	PCG	**1.179**	**1.121**	-0.124	0.925	1.316	2.674

ADASAP scales to $3 \cdot 10^8$ samples



(a) Comparison between **ADASAP** and competitors.



(b) Multi-GPU scaling of **ADASAP**.

Figure 1: Results on huge-scale transportation data analysis (NYC taxi dataset).

Theoretical result

Corollary (Corollary 3.3 in paper)
Suppose that the matrix K exhibits polynomial spectral decay, i.e., $\lambda_i(K) = \Theta(i^{-\beta})$ for some $\beta > 1$. Then for any $\ell \in \{1, \ldots, n\}$ and $\epsilon \in (0, 1)$, choosing block size $b = 4\ell$ we can find $\hat{m}$ that with probability at least 0.99 satisfies $\|\mathrm{proj}_\ell(\hat{m}) - \mathrm{proj}_\ell(m_n)\|_{\mathcal{H}}^2 \leq \epsilon \cdot \|\mathrm{proj}_\ell(m_n)\|_{\mathcal{H}}^2$ in

$$\tilde{\mathcal{O}}\left((n^2 + n\ell^2)\min\left\{\frac{1}{\epsilon}, \left(1 + \frac{\ell(\lambda_\ell(K) + \lambda)}{n(\lambda_n(K) + \lambda)}\right)\log\left(\frac{1}{\epsilon}\right)\right\}\right) \quad \textit{time.}$$

- We analyze a simpler version of ADASAP to obtain a theoretical result
- Takeaway: sketch-and-project can attain moderately accurate estimates of the posterior mean with a rate that is independent of the conditioning of K

Conclusion

- ADASAP = Approximation + distribution + acceleration + sketch-and-project
- ADASAP outperforms state-of-the-art methods for GP inference
- Empirical success of ADASAP is supported by theory: we show that sketch-and-project rapidly converges to a moderately accurate estimate of the posterior mean

Thanks for listening!

Email: pratikr@stanford.edu

Paper: https://arxiv.org/abs/2505.13723