Exact Gaussian Processes on a Million Data Points

Ke Alexander Wang*[1], Geoff Pleiss*[1], Jacob R. Gardner[2], Stephen Tyree[3], Kilian Q. Weinberger[1], Andrew Gordon Wilson[4]

[1]Cornell University, [2]Uber AI Labs, [3]NVIDIA, [4]New York University

Summary

1. We **scale exact GPs to over 10^6 training points** using multi-GPU parallelism and conjugate gradients-based inference. On 10^6 data points, **training takes less than 2 hours**.
2. We demonstrate that **computing predictive distributions with exact GPs is fast and practical** on consumer-grade GPUs.
3. We perform the **first-ever comparison of exact GPs against GP approximations** on datasets with $10^4 - 10^6$ data points.

Background: matrix-multiplication inference

Given n training data points $(X, \mathbf{y})$, GP training requires optimizing model hyperparameters θ (e.g. kernel lengthscale, observed noise):

$$\min_{\theta} \; y^\top \widehat{K}(\theta)_{XX}^{-1} y + \log \left| \widehat{K}(\theta)_{XX} \right| \tag{1}$$

($\widehat{K}(\theta)_{XX}$ is the $n \times n$ covariance matrix with observational noise.)
We use Black-Box Matrix-Matrix (BBMM) inference [1] to reduce the time per optimization iteration from $O(n^3)$ to $O(n^2)$ by relying only on matrix multiplications and conjugate gradients (CG).

Scaling GPs from $n = 10^4$ to $n = 10^6$

Reducing BBMM's memory from $O(n^2)$ to $O(n)$:

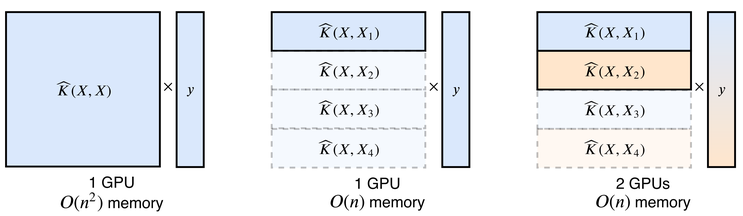

Figure 1:Partitioned kernel matrix-vector multiplication.

Additional techniques to speed up BBMM training:

1. Distribute the multiplication partitions across multiple GPUs.
2. Initializing hyperparameters θ from a GP trained on $<10\%$ of data.
3. Increasing the rank of the pivoted Cholesky preconditioner (from rank-5 to rank-100) to compute $\widehat{K}(\theta)_{XX}^{-1}\mathbf{y}$
4. Using a looser CG convergence criterion at train time.

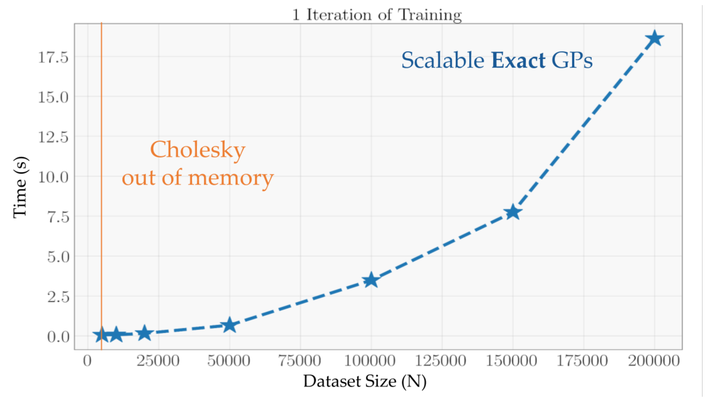

Figure 2:Conventional Cholesky-based inference versus multi-GPU CG-based inference.

Comparisons against approximate methods

We compare against approximate GP methods SGPR [2] and SVGP [3], two popular inducing point methods for large scale GP regression.

			RMSE			NLL		
Dataset	n	d	**Exact GP** (BBMM)	**SGPR** (m=512)	**SVGP** (m=1,024)	**Exact GP** (BBMM)	**SGPR** (m=512)	**SVGP** (m=1,024)
PoleTele	9,600	26	**0.088** ± 0.003	0.113 ± 0.005	0.109 ± 0.002	−0.660 ± 0.081	−0.817 ± 0.005	**−0.644** ± 0.008
Elevators	10,623	18	0.399 ± 0.011	0.426 ± 0.007	**0.388** ± 0.010	0.626 ± 0.043	0.528 ± 0.015	**0.486** ± 0.019
Bike	11,122	17	**0.043** ± 0.012	0.094 ± 0.010	0.077 ± 0.005	**−1.323** ± 0.170	−0.805 ± 0.005	−0.984 ± 0.021
Kin40K	25,600	8	**0.080** ± 0.001	0.225 ± 0.026	0.240 ± 0.007	**−0.755** ± 0.009	−0.073 ± 0.055	0.091 ± 0.033
Protein	29,267	9	**0.511** ± 0.009	0.619 ± 0.003	0.613 ± 0.011	0.960 ± 0.033	**0.915** ± 0.004	0.952 ± 0.018
KeggDirected	31,248	20	**0.083** ± 0.001	0.104 ± 0.002	0.105 ± 0.003	−0.838 ± 0.031	**−1.163** ± 0.005	−0.853 ± 0.033
CTslice	34,240	385	0.497 ± 0.029	**0.217** ± 0.009	1.004 ± 0.005	0.939 ± 0.004	**−0.037** ± 0.060	1.423 ± 0.005
KEGGU	40,708	27	**0.120** ± 0.001	0.130 ± 0.001	0.126 ± 0.002	−0.540 ± 0.035	**−1.049** ± 0.010	−0.653 ± 0.013
3DRoad	278,319	3	**0.110** ± 0.017	0.578 ± 0.001	0.390 ± 0.005	1.239 ± 0.025	0.791 ± 0.033	**0.486** ± 0.010
Song	329,820	90	**0.774** ± 0.001	0.816 ± 0.038	0.998 ± 0.000	**1.162** ± 0.002	1.243 ± 0.083	1.417 ± 0.000
Buzz	373,280	77	0.279 ± 0.002	0.289 ± 0.001	**0.270** ± 0.012	0.161 ± 0.026	**0.092** ± 0.017	0.119 ± 0.042
HouseElectric	1,311,539	9	**0.054** ± 0.000	—	0.127 ± 0.046	**−0.207** ± 0.001	—	0.024 ± 0.984

Table 1:Exact GP vs SGPR vs SVGP using a Matern 3/2 kernel with independent lengthscales.

Training times and prediction times

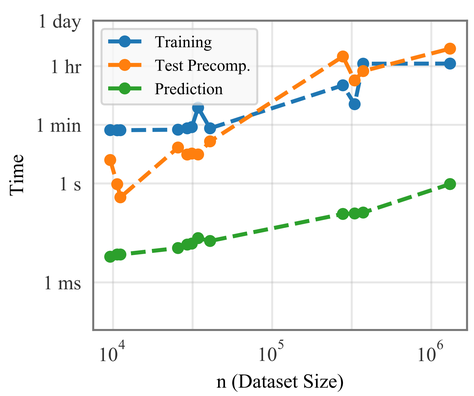

Figure 3:Training, test-time precomputation, and prediction times as a function of dataset size. All predictions for exact GPs can be done in less than a second.

How much does pretraining help?

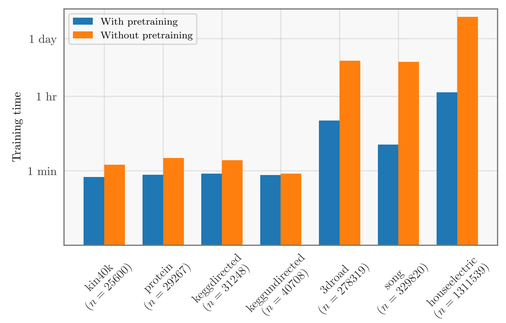

Figure 4:Fine-tuning on the full training set after pretraining on a smaller subset significantly reduces training time. Models with and without pretraining achieve similar errors with less than 5% difference (not shown here).

Do GPs need the entire dataset?

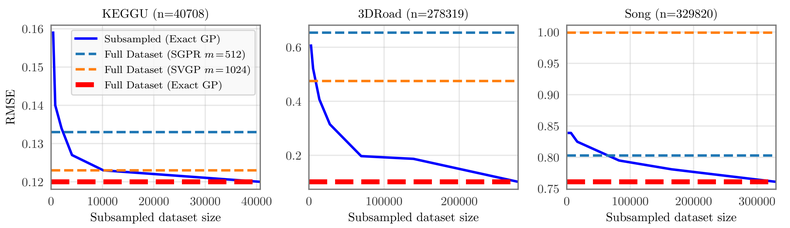

Figure 5:Test root-mean-square error (RMSE) vs. subsampled dataset size.

Would more inducing points help?

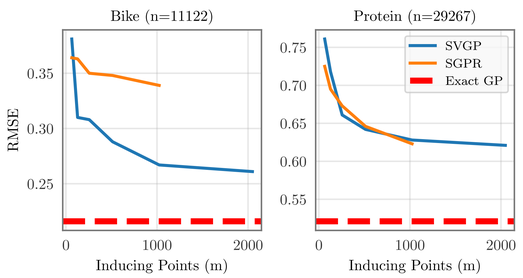

Figure 6:Test root-mean-square error (RMSE) of SVGP and SGPR methods as a function of the number of inducing points.

References

[1] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *NeurIPS*, pages 7587–7597, 2018.

[2] Michalis K Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *AISTATS*, pages 567–574, 2009.

[3] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. In *UAI*, 2013.