

Stochastic Gradient Descent in Correlated Settings: A Study on Gaussian Processes

Hao Chen¹, Lili Zheng¹, Raed AL Kontar², Garvesh Raskutti¹

¹ Department of Statistics, University of Wisconsin-Madison

² Department of Industrial & Operations Engineering, University of Michigan

Motivation

- SGD's success in independent sample settings.
- The lack of understanding of SGD's behavior in **correlated** sample settings.
- Naive exact GP inference is expensive, requiring $O(n^3)$ computation and $O(n^2)$ storage.
- Approximate GPs still suffer to scale beyond a couple million data points.
- Can we use SGD to accelerate GP inference and scale them far beyond what is currently possible?

Problem Setup

- GP model

$$f \sim \mathcal{GP}(0, \sigma_f^2 k(\cdot, \cdot)), \quad \mathbf{x}_1, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P},$$

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad 1 \leq i \leq n.$$

- Estimate $\theta^* = (\sigma_f^2, \sigma_\epsilon^2)$ by minimizing the negative marginal log-likelihood

$$\ell(\theta; \mathbf{X}_n, \mathbf{y}_n) = \frac{1}{2n} [\mathbf{y}_n^\top \mathbf{K}_n^{-1}(\theta) \mathbf{y}_n + \log |\mathbf{K}_n(\theta)| + n \log(2\pi)],$$

where

$$\mathbf{K}_n(\theta) = \theta_1 \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & & \vdots \\ \cdots & k(x_i, x_j) & \cdots \\ \vdots & & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix} + \theta_2 I_n$$

is the covariance matrix of \mathbf{y}_n .

Algorithm 1: Minibatch SGD

```

1 Input:  $\theta^{(0)} \in \mathbb{R}^2$ , initial step size  $\alpha_1 > 0$ .
2 for  $k = 1, 2, \dots, K$  do
3   Randomly sample a subset of indices  $\xi_k$  of size  $m \ll n$ ;
4   Compute the scaled stochastic gradient  $g(\theta^{(k)}; \mathbf{X}_{\xi_k}, \mathbf{y}_{\xi_k})$ ;
5    $\alpha_k \leftarrow \frac{\alpha_1}{k}$ ;
6    $\theta^{(k)} \leftarrow \theta^{(k-1)} - \alpha_k g(\theta^{(k-1)}; \mathbf{X}_{\xi_k}, \mathbf{y}_{\xi_k})$ ;
7 end for
```

- $O(m^3)$ computation and $O(m^2)$ storage per iteration.

Technical Challenges

- The inherent correlation among samples coded in $\mathbf{K}_n(\theta)$ leads to

- Loss function is highly **non-linear** w.r.t. data points

$$\ell(\theta; \mathbf{X}_n, \mathbf{y}_n) \neq \frac{1}{n} \sum_{i=1}^n \ell(\theta; \mathbf{x}_i, y_i).$$

- Stochastic gradient is a **biased** estimator of full gradient

$$\mathbb{E}(g(\theta^{(k-1)}; \mathbf{X}_{\xi_k}, \mathbf{y}_{\xi_k})) \neq \nabla \ell(\theta^{(k-1)}; \mathbf{X}_n, \mathbf{y}_n).$$

- Loss function is **non-convex** w.r.t hyper-parameter θ .

• Solution

Explore the connection between $g(\theta; \mathbf{X}_{\xi_k}, \mathbf{y}_{\xi_k})$ and its conditional expectation $\mathbb{E}(g(\theta; \mathbf{X}_{\xi_k}, \mathbf{y}_{\xi_k}) | \mathbf{X}_{\xi_k})$, and prove a relaxation of strong convexity for $\mathbb{E}(g(\theta; \mathbf{X}_{\xi_k}, \mathbf{y}_{\xi_k}) | \mathbf{X}_{\xi_k})$ by exploiting the eigendecay of $\mathbf{K}_n(\theta)$.

Theoretical Guarantee of Convergence

- Assumptions

- **Exponential decay**

The eigenvalues of kernel function $k(\cdot, \cdot)$ w.r.t. probability measure \mathbb{P} are $\{C e^{-bj}\}_{j=0}^\infty$.

- Bounded iterates

The true parameter θ^* and iterates $\theta^{(k)}$ lies in $[\theta_{\min}, \theta_{\max}]$.

- Bounded stochastic gradient

$$\|\nabla \ell(\theta^{(k)}; \mathbf{X}_{\xi_{k+1}}, \mathbf{y}_{\xi_{k+1}})\|_2 \leq G.$$

- **Convergence of parameter iterates**

$$(\theta_1^{(K)} - \theta_1^*)^2 \leq C \left[\frac{G^2}{(K+1)} + m^{-\frac{1}{2}+\epsilon} \right],$$

$$(\theta_2^{(K)} - \theta_2^*)^2 \leq C \left[\frac{G^2}{(K+1)} + (\log m)^{-\frac{1}{2}+\epsilon} \right].$$

- **Convergence of full gradient**

$$\|\nabla \ell(\theta^{(K)})\|_2^2 \leq C \left[\frac{G^2}{K+1} + m^{-\frac{1}{2}+\epsilon} \right].$$

Numerical Illustration of Convergence

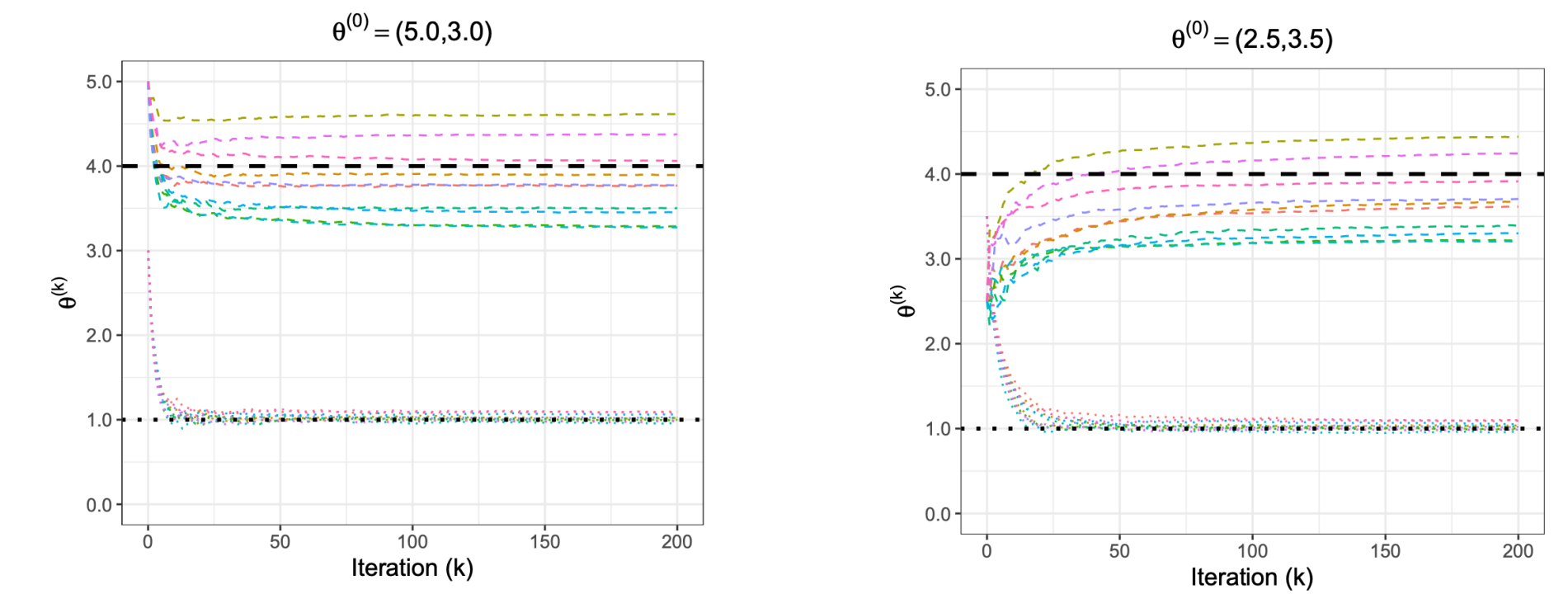


Figure 1: Convergence of parameters.

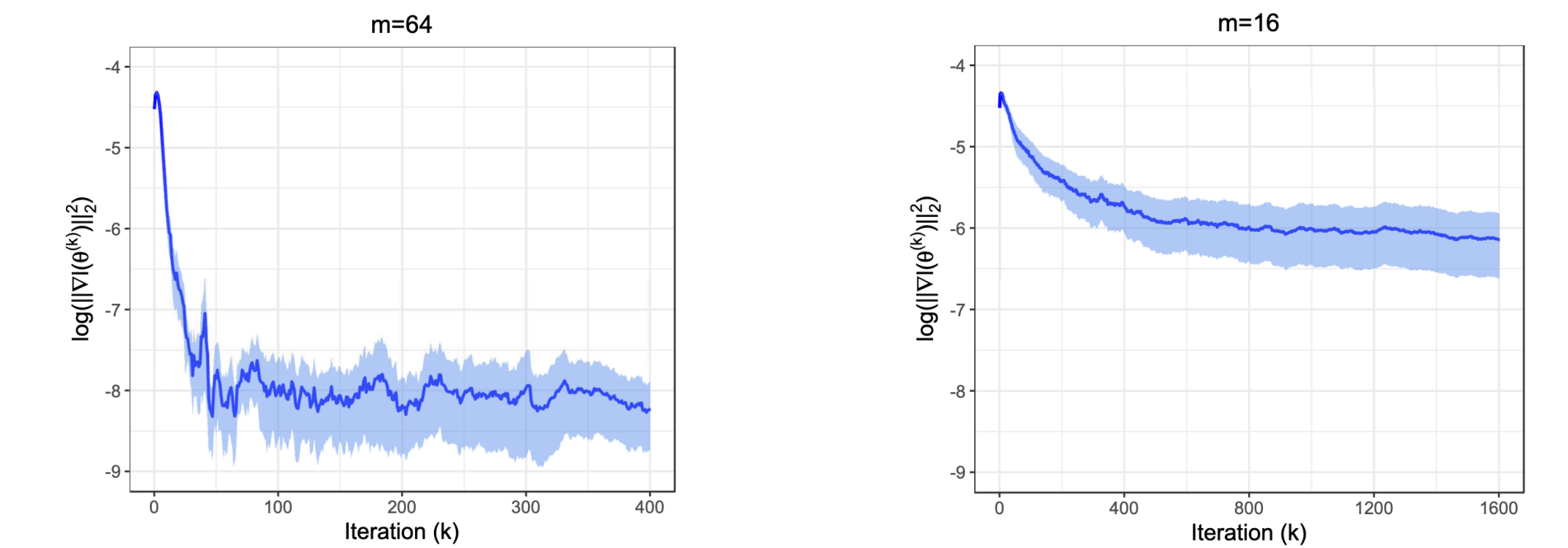


Figure 2: Convergence of full gradient.

Experiments

Table 1: Performance of different GPs on benchmark datasets. "—" indicates infeasible to train.

Dataset	Size	RMSE				Training Time (min)			
		sgGP	EGP	SGPR	SVGP	sgGP	EGP	SGPR	SVGP
Levy	10,000	0.27	0.31	0.56	0.58	0.51	11.48	4.04	14.58
Griewank	10,000	0.07	0.19	0.13	0.09	0.61	15.25	1.93	13.18
Bike	17,379	0.22	0.23	0.28	0.25	1.98	31.48	5.31	25.26
Energy	19,735	0.79	0.80	0.84	0.80	3.15	54.39	5.41	25.09
PM2.5	41,757	0.29	0.29	0.64	0.54	5.21	385.51	13.59	52.46
Protein	45,730	0.66	0.69	0.72	0.68	3.40	500.33	19.55	55.27
Query	100,000	0.05	—	0.06	0.06	6.40	—	20.73	124.73
Borehole	1,000,000	0.17	—	0.18	0.17	67.29	—	857.60	1380.86

Table 2: sgGP on toy datasets.

Dataset	Size	RMSE	Training Time (min)	Memory Usage (GB)
OTL Circuit	2,000,000	0.40	33.43	0.99
Wing Weight	2,000,000	0.07	78.78	1.22