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

Hao Chen[1], Lili Zheng[1], Raed AL Kontar[2], Garvesh Raskutti[1]

[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

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

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

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

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

where

$$\mathbf{K}_n(\boldsymbol{\theta}) = \theta_1 \begin{pmatrix} k(\mathbf{x}_1,\mathbf{x}_1) & \cdots & k(\mathbf{x}_1,\mathbf{x}_n) \\ \vdots & & \\ k(\mathbf{x}_i,\mathbf{x}_1) & \cdots & k(\mathbf{x}_i,\mathbf{x}_j) \\ \vdots & & \\ k(\mathbf{x}_n,\mathbf{x}_1) & \cdots & k(\mathbf{x}_n,\mathbf{x}_n) \end{pmatrix} + \theta_2 \mathbf{I}_n$$

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

**Algorithm 1:** Minibatch SGD

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

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

## Technical Challenges

- The inherent correlation among samples coded in $\mathbf{K}_n(\boldsymbol{\theta})$ leads to
  - Loss function is highly non-linear w.r.t. data points

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

  - Stochastic gradient is a biased estimator of full gradient

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

  - Loss function is non-convex w.r.t hyper-parameter $\boldsymbol{\theta}$.

- **Solution**

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

## Theoretical Guarantee of Convergence

- Assumptions
  - Exponential decay

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

    The true parameter $\boldsymbol{\theta}^*$ and iterates $\boldsymbol{\theta}^{(k)}$ lie in $[\boldsymbol{\theta}_{\min}, \boldsymbol{\theta}_{\max}]$.
  - Bounded stochastic gradient

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

- **Convergence of parameter iterates**

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

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

- **Convergence of full gradient**

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

## Numerical Illustration of Convergence

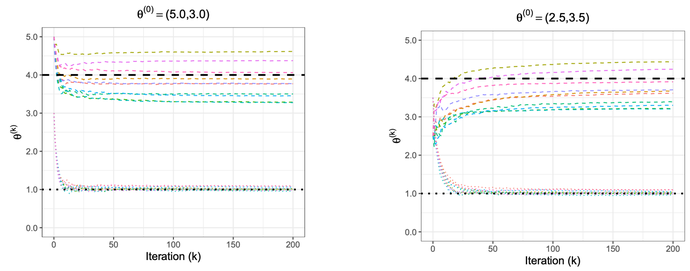


Figure 1: Convergence of parameters.

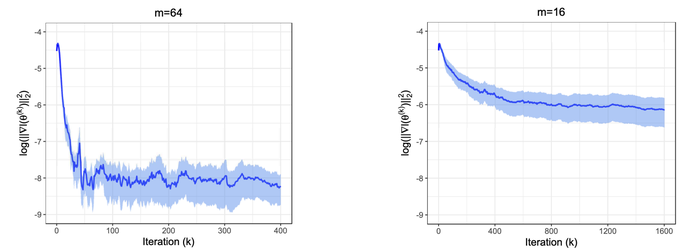


Figure 2: Convergence of full gradient.

## Experiments

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

| | | RMSE | | | | Training Time (min) | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | Size | 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 |