Error dynamics of mini-batch gradient descent with random reshuffling for least squares

Jackie Lok[1] Rishi Sonthalia[2] Elizaveta Rebrova[1]

[1]Princeton University [2]Boston College

Introduction

Machine learning models are often trained using mini-batch gradient descent with random reshuffling: in each epoch, the dataset is randomly partitioned into mini-batches and iterated through.

Question: What are the *error dynamics* and *generalization capabilities* of the learned model?

Main difficulty: Introduction of dependencies complicates the analysis, compared to sampling with replacement.

Model. Observe n i.i.d. data samples $(\mathbf{x}_i, y_i)$, where $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_* + \eta_i$ for some $\boldsymbol{\beta}_* \in \mathbb{R}^p$ and noise η_i (i.e., $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_* + \boldsymbol{\eta}$).

Partition data $\mathbf{X} \in \mathbb{R}^{n \times p}$ into B mini-batches $\mathbf{X}_1, \dots, \mathbf{X}_B \in \mathbb{R}^{(n/B) \times p}$. In each epoch, mini-batches are ordered by a random permutation $\tau \in S_B$, and B iterations of GD with step size α for the least squares loss $L_b(\boldsymbol{\beta}) = \frac{B}{2n}\|\mathbf{X}_b \boldsymbol{\beta} - \mathbf{y}_b\|_2^2$ are performed:

$$\boldsymbol{\beta}_k^{(b)} = \boldsymbol{\beta}_k^{(b-1)} - \frac{B\alpha}{n}\mathbf{X}_{\tau(b)}^\top (\mathbf{X}_{\tau(b)} \boldsymbol{\beta}_k^{(b-1)} - \mathbf{y}_{\tau(b)}), \quad b = 1, 2, \dots, B.$$

Summary. By studying the discrete error dynamics of the *mean iterate* after k epochs, $\bar{\boldsymbol{\beta}}_k := \mathbb{E}_{\tau \sim \mathrm{Unif}(S_B)}\left[\boldsymbol{\beta}_k^{(B)}\right]$, we find that there are higher-order step size-dependent effects introduced by sampling without replacement, which result in subtly different trajectories under the linear scaling rule compared to full-batch gradient descent.

Definition of modified features

Let $\mathbf{W}_b = \frac{B}{n}\mathbf{X}_b^\top \mathbf{X}_b$ be the sample covariance of each mini-batch, and $\mathbf{W} = \frac{1}{n}\mathbf{X}^\top \mathbf{X} = \frac{1}{n}\sum_{b=1}^B \mathbf{X}_b^\top \mathbf{X}_b$. Define modified mini-batches $\widetilde{\mathbf{X}}_b := \mathbf{X}_b \Pi_b$, where $\Pi_b := \mathbb{E}_{\tau \sim \mathrm{Unif}(S_B)}\left[\prod_{j : j < \tau^{-1}(b)} (\mathbf{I} - \alpha \mathbf{W}_{\tau(j)})\right]$, and concatenate into $\widetilde{\mathbf{X}}$ (i.e., feature $\mathbf{x}_i \leftrightarrow \Pi_b \mathbf{x}_i$). Define the sample cross-covariance

$$\mathbf{Z} := \frac{1}{n}\widetilde{\mathbf{X}}^\top \mathbf{X} = \frac{1}{n}\sum_{b=1}^B \Pi_b \mathbf{X}_b^\top \mathbf{X}_b.$$

This can be shown to be symmetric, and $\mathbf{Z} \equiv \mathbf{Z}(\alpha) = \mathbf{W} + O(\alpha)$. *In general: $\mathbf{Z}$ is a non-commutative polynomial of $\mathbf{W}_1, \dots, \mathbf{W}_B$.*

Example (two-batch GD). For $B = 2$, $\widetilde{\mathbf{X}}_1 = \mathbf{X}_1\left(\mathbf{I} - \frac{1}{2}\alpha \mathbf{W}_2\right)$, and $\mathbf{Z} = \frac{1}{2}(\mathbf{W}_1 + \mathbf{W}_2) - \frac{1}{4}\alpha\left(\mathbf{W}_2\mathbf{W}_1 + \mathbf{W}_1\mathbf{W}_2\right)$.

Training error dynamics

Theorem 1.

$$\bar{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_* = (\mathbf{I} - B\alpha\mathbf{Z})^k(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_*) + \frac{1}{n}\left[\mathbf{I} - (\mathbf{I} - B\alpha\mathbf{Z})^k\right]\mathbf{Z}^\dagger \widetilde{\mathbf{X}}^\top \boldsymbol{\eta}.$$

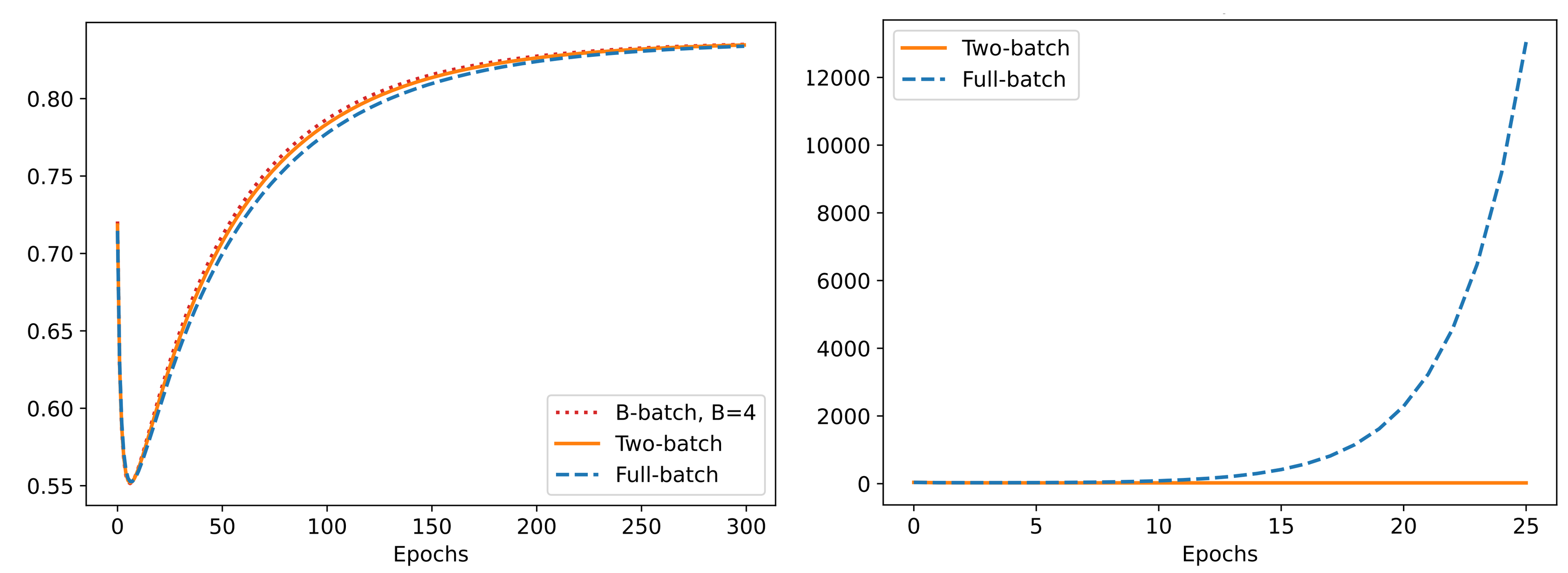
- Comparison with full-batch gradient descent:

$$\boldsymbol{\beta}_k^{\text{full}} - \boldsymbol{\beta}_* = (\mathbf{I} - \alpha\mathbf{W})^k(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_*) + \frac{1}{n}\left[\mathbf{I} - (\mathbf{I} - \alpha\mathbf{W})^k\right]\mathbf{W}^\dagger \mathbf{X}^\top \boldsymbol{\eta}.$$

Mini-batches sampled with replacement $\Rightarrow$ same expression for mean iterate as full-batch GD with $k \leftarrow Bk$ (time change).

- Linear scaling rule (i.e., using step size α/B with B batches) $\Rightarrow$ *mini-batch dynamics match full-batch GD up to first order in α* (left plot).

However, *full-batch can diverge while mini-batch converges* (right plot).



Generalization error dynamics

Assume that $\mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^\top\right] = \Sigma$, and η_i has mean zero and variance σ^2. We are interested in the dynamics of the risk

$$R_{\mathbf{X}}(\boldsymbol{\beta}) := \mathbb{E}\left[(\mathbf{x}^\top \boldsymbol{\beta} - \mathbf{x}^\top \boldsymbol{\beta}_*)^2 \mid \mathbf{X}\right] = \mathbb{E}\left[\|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_\Sigma^2 \mid \mathbf{X}\right].$$

Theorem 2. Let $\mathbf{P}_{\mathbf{Z},0}$, $\mathbf{P}_{\mathbf{Z}}$ be projectors onto $\mathrm{Null}(\mathbf{Z})$, $\mathrm{Range}(\mathbf{Z})$.

$$\begin{aligned} R_{\mathbf{X}}(\bar{\boldsymbol{\beta}}_k) &= (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_*)^\top \mathbf{P}_{\mathbf{Z},0}\Sigma\mathbf{P}_{\mathbf{Z},0}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_*) \\ &+ (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_*)^\top \mathbf{P}_{\mathbf{Z}}(\mathbf{I} - B\alpha\mathbf{Z})^k \Sigma (\mathbf{I} - B\alpha\mathbf{Z})^k \mathbf{P}_{\mathbf{Z}}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_*) \\ &+ \frac{\sigma^2}{n}\mathrm{Tr}\left(\left[\mathbf{I} - (\mathbf{I} - B\alpha\mathbf{Z})^k\right]\Sigma\left[\mathbf{I} - (\mathbf{I} - B\alpha\mathbf{Z})^k\right]\mathbf{Z}^\dagger\left(\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right)\mathbf{Z}^\dagger\right) \end{aligned}$$

Decomposition into fixed error in "frozen subspace" + bias component ($\to 0$) + variance component ($\to \frac{\sigma^2}{n}\mathrm{Tr}\left(\Sigma\mathbf{Z}^\dagger\left(\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right)\mathbf{Z}^\dagger\right)$).

Two-batch case: limiting variance $= (1 + O(\alpha))\frac{\sigma^2}{n}\mathrm{Tr}\left(\Sigma\mathbf{Z}^\dagger\right)$, which is highly reminiscent of $\frac{\sigma^2}{n}\mathrm{Tr}\left(\Sigma\mathbf{W}^\dagger\right)$ for full-batch GD.

Asymptotic analysis: p fixed, $n \to \infty$

Proposition 3. Suppose p is fixed. Then $\mathbf{Z}(\alpha/B) \to \Sigma(\mathbf{I} - p_{B,\alpha}(\Sigma))$ as $n \to \infty$, where $p_{B,\alpha}$ is a polynomial. (e.g., $p_{B,2}(\Sigma) = \frac{1}{4}\alpha\Sigma$, and $p_{B,3}(\Sigma) = \frac{1}{3}\alpha\Sigma - \frac{1}{27}\alpha^2\Sigma^2$).

Observation: if Σ has eigenvalues λ_i, then the limiting eigenvalues of $\mathbf{Z}$ are $\lambda_i(1 - p_{B,\alpha}(\lambda_i))$: shrinkage effect on spectrum.

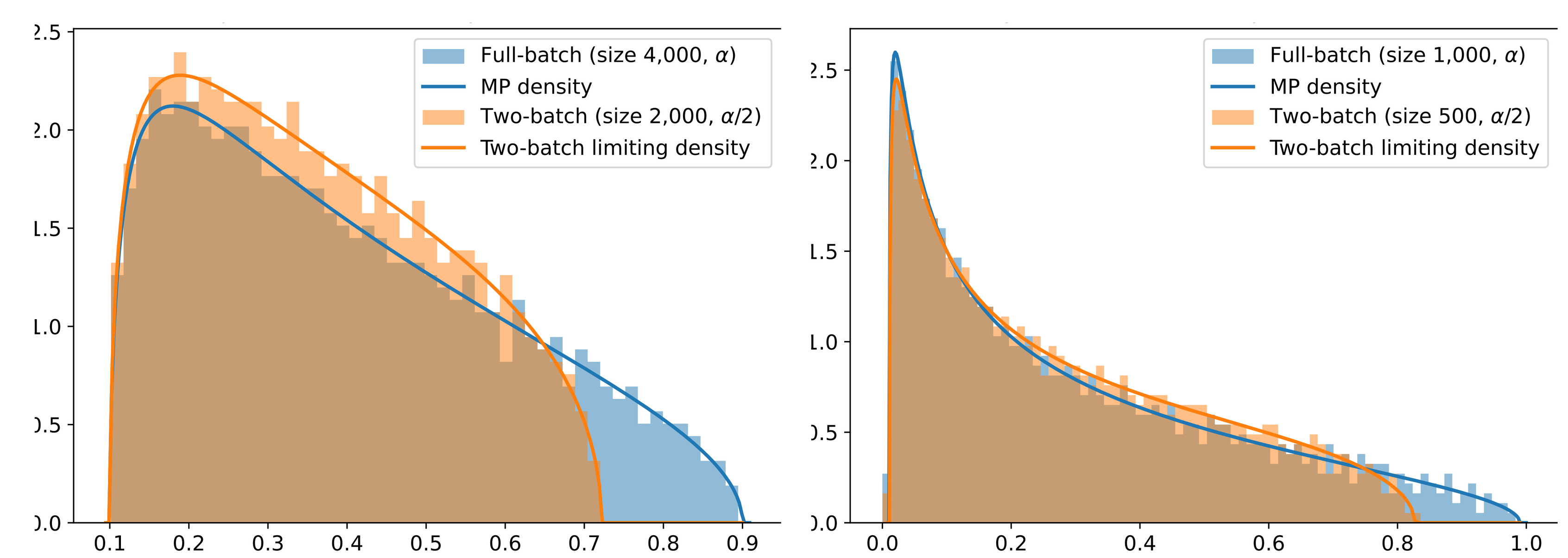
Idea: Features $\mathbf{x}_i$ i.i.d. with $\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\top\right] = \Sigma \Rightarrow \mathbf{W}_b \to \Sigma$ for all $b \in [B]$ a.s. by the law of large numbers. In particular, Π_b is independent of b asymptotically. If we explicitly assume this:

Proposition 4. Let $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ be a SVD of $\mathbf{X}$, and $\mathbf{W} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ have eigenvalues $\hat{\lambda}_i$. If $\Pi_b \equiv \mathbf{I} - p_{B,\alpha}(\mathbf{W})$ and Σ is isotropic, then

$$\begin{aligned} R_{\mathbf{X}}(\bar{\boldsymbol{\beta}}_k) &= \sum_{i=1}^p [1 - \alpha\hat{\lambda}_i(1 - p_{B,\alpha}(\hat{\lambda}_i))]^{2k}[\mathbf{V}^\top(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_*)]_i \\ &+ \frac{\sigma^2}{n}\sum_{i=1}^p \frac{1}{\hat{\lambda}_i}\left(1 - [1 - \alpha\hat{\lambda}_i(1 - p_{B,\alpha}(\hat{\lambda}_i))]^k\right)^2. \end{aligned}$$

Upshot: *explicit description* of how the *trajectory* differs from full-batch GD under linear scaling, based on the *spectrum of the features sample covariance* (same expression with $\hat{\lambda}_i \leftarrow \hat{\lambda}_i(1 - p_{B,\alpha}(\hat{\lambda}_i))$)!

Proportional regime: $p/n \to \gamma \in (0, \infty)$



Two-batch, Gaussian $\mathbf{X}$: $\alpha\mathbf{Z}(\alpha/2) = \frac{1}{2}\alpha(\mathbf{W}_1 + \mathbf{W}_2) - \frac{1}{8}\alpha^2(\mathbf{W}_2\mathbf{W}_1 + \mathbf{W}_1\mathbf{W}_2)$ *is a non-commutative polynomial of Wishart random matrices*: no (simple) analytical characterization of limiting spectral distribution.

Could *numerically compute* using an algorithm of Belinschi et al. (2017), based on operator-valued free probability. Figure shows results in (left) underparameterized $\gamma = 1/4$ and (right) overparameterized $\gamma = 3/2$ regimes. Shrinkage effect is again apparent.

arXiv preprint (2024): arXiv:2406.03696

https://jackielok.github.io/ jackie.lok@princeton.edu