Full-Batch Gradient Descent Outperforms One-Pass SGD: Sample Complexity Separation in Single-Index Learning

Filip Kovačević[1], Hong Chang Ji[2,*], Denny Wu[3,*], Mahdi Soltanolkotabi[4,*], Marco Mondelli[1,*]

[1]Institute of Science and technology Austria
[2]Sung Kyun Kwan University
[3]New York University and Flatiron Institute
[4]University of Southern California

*Equal contribution

Problem setup

Dataset

$$\mathscr{D} = \{x_i, y_i\}_{i=1}^n$$

True model

$$y = f_{\theta^*}(x)$$

Hypothesis class

$$\{f_\theta \,|\, \theta \in \mathbb{R}^d\}$$

Empirical loss

$$\mathscr{L}(\theta, \mathscr{D}) = \frac{1}{n} \sum_{i=1}^{n} l(\theta, (x_i, y_i))$$

Main question

When does reusing data help optimization?

Optimization

Gradient descent

Full batch GD

$$\theta_t = \theta_{t-1} - \eta \nabla_\theta \mathscr{L}(\theta_{t-1}, \mathscr{D})$$

VS

One pass SGD

$$\theta_t = \theta_{t-1} - \eta \nabla_\theta l(\theta_{t-1}, (x_{t-1}, y_{t-1}))$$

Data model

Dataset

$$\mathscr{D} = \{x_i, y_i\}_{i=1}^n$$

True model

$$y = \sigma(\langle x, \theta^* \rangle)$$

$$\theta^* \in \mathbb{S}^{d-1}$$

Hypothesis class

$$\{\sigma(\langle x, \theta \rangle) \,|\, \theta \in \mathbb{S}^{d-1}\}$$

Empirical loss

$$\mathscr{L}(\theta, \mathscr{D}) = -\frac{1}{n} \sum_{i=1}^{n} y_i \, \sigma(\langle x, \theta \rangle)$$

Goal

weak/strong recovery

$$|\langle \theta^*, \theta(\mathscr{D}) \rangle| \to c > 0$$

$$\|\theta(\mathscr{D}) - \theta^*\| \to 0$$

Previous results

Theorem (informal)[Ben Arous et. al., 2021]

For $\sigma(z) = z^2$

SGD on $\mathbb{S}^{d-1}$ with $n \ll d \log d, \quad |\langle \theta^*, \theta_n \rangle| \to 0.$

SGD on $\mathbb{S}^{d-1}$ with $n \gg d \log^2 d, \quad |\langle \theta^*, \theta_n \rangle| \to 1.$

"GAP"

Theorem (informal)[Mondelli & Montanari, 2018, Barbier et. al., 2019]

For $\sigma(z) = z^2$

$\forall$ algorithm with $n \ll d, \quad |\langle \theta^*, \theta(\mathscr{D}) \rangle| \to 0.$

$\exists$ algorithm with $n \sim d, \quad |\langle \theta^*, \theta(\mathscr{D}) \rangle| \to c > 0.$

Ben Arous, G., Gheissari, R., and Jagannath, A. Online stochastic gradient descent on non-convex losses from high-dimensional inference. Journal of Machine Learning Research, 22(106):1–51, 2021.

Mondelli, M. and Montanari, A. Fundamental limits of weak recovery with applications to phase retrieval. In Conference On Learning Theory, pp. 1445–1450. PMLR, 2018.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborova, L. Optimal errors and phase transitions in high- ´ dimensional generalized linear models. Proceedings of the National Academy of Sciences, 116(12):5451–5460, 2019.

Full batch GF - phase retrieval

Full batch Gradient flow

$$\frac{d\theta_t}{dt} = -\nabla_{\mathbb{S}^{d-1}} \mathscr{L}(\theta_{t-1}, \mathscr{D})$$

Theorem (informal)[ours]

For $\sigma(z) = z^2$ full-batch GF with $n \ll d \log d, \quad |\langle \theta^*, \theta \rangle| \to 0.$

Full batch GF - truncated phase retrieval

Full batch Gradient flow

$$\frac{d\theta_t}{dt} = -\nabla_{\mathbb{S}^{d-1}} \mathscr{L}(\theta_{t-1}, \mathscr{D})$$

Theorem (informal)[ours]

For $\sigma(z) = \min\{z^2, C\}$ full-batch GF with $n \sim d, \quad |\langle \theta^*, \theta \rangle| \to c > 0.$

Theorem (informal)[Ben Arous et. al., 2021]

For $\sigma(z) = \min\{z^2, C\}$

SGD with $n \ll d \log d, \quad |\langle \theta^*, \theta_n \rangle| \to 0.$

SGD with $n \gg d \log^2 d, \quad |\langle \theta^*, \theta_n \rangle| \to 1.$

Ben Arous, G., Gheissari, R., and Jagannath, A. Online stochastic gradient descent on non-convex losses from high-dimensional inference. Journal of Machine Learning Research, 22(106):1–51, 2021.

Full batch GD - truncated phase retrieval

Hypothesis class

$$\{\sigma(\langle x, \theta\rangle) \,|\, \theta \in \mathbb{R}^d\}$$

Empirical loss

$$\mathscr{L}(\theta, \mathscr{D}) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \sigma(\langle x, \theta\rangle))^2$$

Full batch Gradient Descent

$$\theta_t = \theta_{t-1} - \eta \nabla_{\mathbb{R}^d} \mathscr{L}(\theta_{t-1}, \mathscr{D})$$

Theorem (informal)[ours]

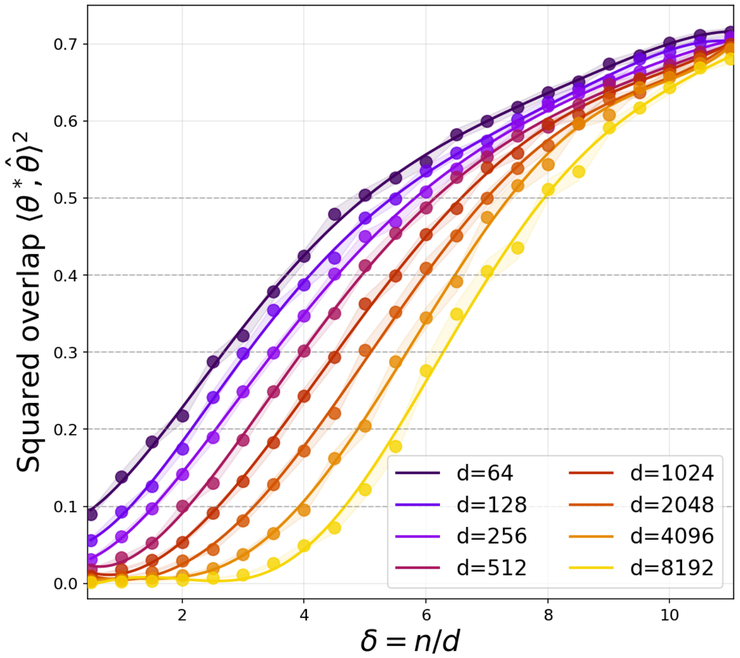
For $\sigma(z) = \min\{z^2, C\}$ full-batch GD with $n \sim d, \quad \|\hat{\theta} - \theta^*\| \to 0.$

Moreover, strong recovery is achieved after $\sim \log d$ steps.

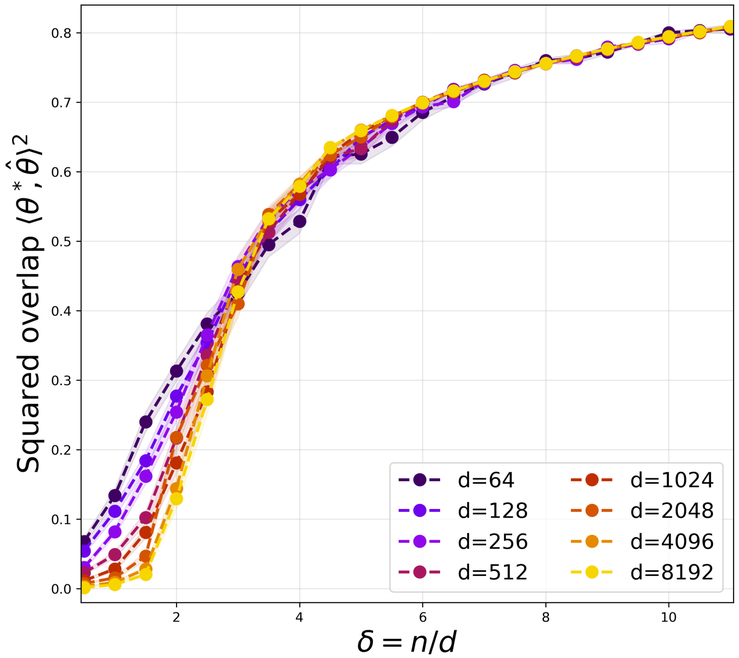
Thank you for your time!

Experiments - Full batch GD

$\sigma(z) = z^2$



$\sigma(z) = \min\{z^2, C\}$



Spherical GD with learning rate $\eta = 0.1$ for $1000 \log^2 d$ steps.