

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

Filip Kovačević¹, Hong Chang Ji², Denny Wu^{3,*}, Mahdi Soltanolkotabi^{4,*}, Marco Mondelli^{1,*}

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

*Equal contribution

Problem

When does reusing data help **optimization**?

Full batch GD

VS

One pass SGD

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

$$\mathcal{D} \rightarrow \mathcal{D}_{\mathcal{I}_1} \rightarrow \mathcal{D}_{\mathcal{I}_2} \rightarrow \dots$$

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} l(\theta_t, (x_t, y_t))$$

$$\mathcal{D}^1 \rightarrow \mathcal{D}^2 \rightarrow \mathcal{D}^3 \rightarrow \dots$$

Setting

Single index model

$$y_i = \sigma(\langle x_i, \theta^* \rangle) \quad \|\theta^*\|_2 = 1$$

$$x_i \sim \mathcal{N}(0, I_d)$$

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

Goal

Weak recovery

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

Strong recovery

$$\|\theta(\mathcal{D}) - \theta^*\| \rightarrow 0$$

SGD and spectral method performance

Theorem (Ben Arous et. al., 2021)

For $\sigma(z) = z^2$ **one pass SGD** on $\mathbb{S}^{d-1}$ achieves

with $n \ll d \log d$, $|\langle \theta^*, \theta_n \rangle| \rightarrow 0$,

with $n \gg d \log^2 d$, $\|\theta^* - \theta_n\| \rightarrow 0$.

This also applies to truncated quadratic activations.

Corollary of (Mondelli & Montanari, 2018; Lu & Li, 2020)

For $\sigma(z) = z^2$, let $A^* = \sum_{i=1}^n \tau(y_i) x_i x_i^\top$, then **spectral method**

with $n \sim d$, for $\tau = id$, $|\langle \theta^*, v_1(A^*) \rangle| \rightarrow 0$,

with $n \sim d$, for $\tau = \min\{id, c\}$, $|\langle \theta^*, v_1(A^*) \rangle| \rightarrow c > 0$.

Full-batch does not improve

Full batch Gradient Flow with **correlation loss**

$$\mathcal{L}(\theta, \mathcal{D}) = -\frac{1}{n} \sum_{i=1}^n y_i \sigma(\langle x, \theta \rangle), \quad \theta \in \mathbb{S}^{d-1}.$$

The flow is a power method on A^* .

$$\frac{d\theta_t}{dt} = -\nabla_{\mathbb{S}^{d-1}} \mathcal{L}(\theta_t, \mathcal{D}) \rightarrow \frac{d\theta_t}{dt} = -(I - \theta_t^\top \theta_t) A^* \theta_t$$

Theorem (ours)

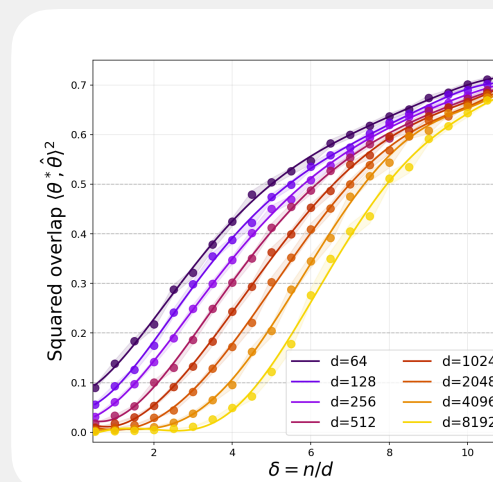
For $\sigma(z) = z^2$ **full-batch GF**

$n \ll d \log d$, $|\langle \theta^*, \theta \rangle| \rightarrow 0$.

Unbounded quadratic link

A^* dominated by rare large samples

Top eigenvector uninformative



Truncation helps weak recovery

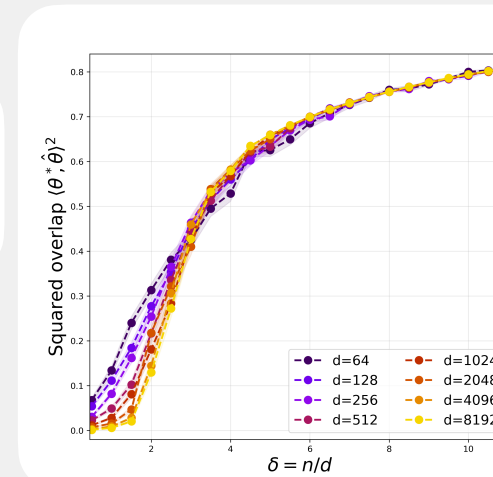
Theorem (ours)

For $\sigma(z) = \min\{z^2, C\}$ **full-batch GF**

$n \sim d$, $|\langle \theta^*, \theta \rangle| \rightarrow c > 0$.

Bounded quadratic link

Uniform BBP spike A_t^*



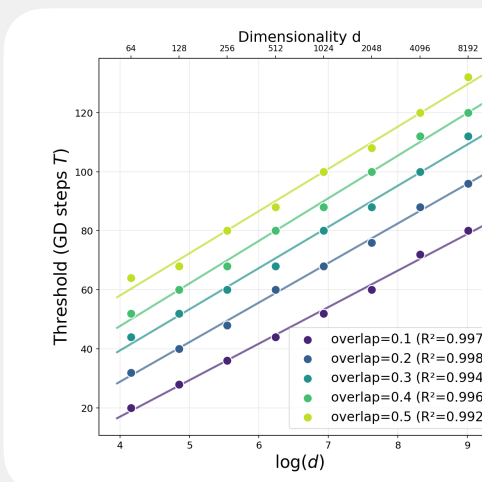
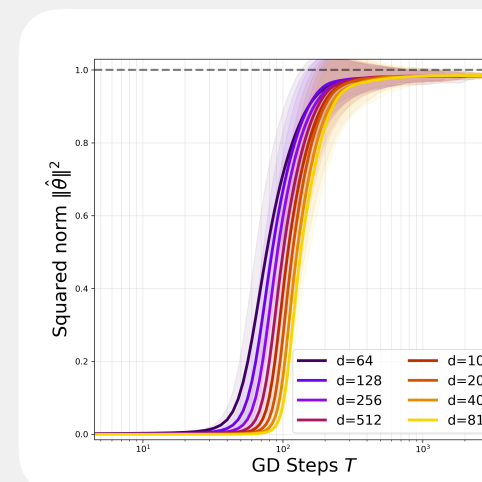
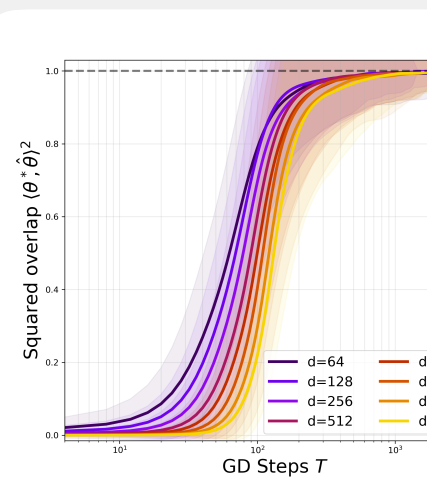
Strong recovery and training dynamics

Full batch Gradient Descent with **square loss**

$$\theta_{t+1} = \theta_t - \eta \nabla_{\mathbb{R}^d} \mathcal{L}(\theta_t, \mathcal{D}), \quad \theta \in \mathbb{R}^d, \quad \mathcal{L}(\theta_t, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n (y_i - \sigma(\langle x, \theta_t \rangle))^2$$

Theorem (ours)

For $\sigma(z) = \min\{z^2, C\}$ **full-batch GD** with $n \sim d$, $\|\hat{\theta} - \theta^*\| \rightarrow 0$, and it takes $T \sim \log d$ steps.



Two phase dynamics

Small random init

$$\theta_0 \sim r_0 \mathbb{S}^{d-1}, \quad r_0 \ll 1$$

Phase I: search

Early **GD** $\approx$ power method on $A^* \rightarrow |\langle \theta^*, \theta_t \rangle| \geq c$

Phase II: refinement

$$\langle \nabla \mathcal{L}(\theta_t, \mathcal{D}), \theta^* - \theta_t \rangle \geq \alpha \|\theta^* - \theta_t\|^2$$

Strong recovery

$\|\theta^* - \theta_t\|^2$ decays geometrically

Method	Activation	Samples	Recovery
One pass SGD	(truncated) quadratic	$d \log d$	(strong) weak
Full batch GF	quadratic	$d \log d$	weak
Full batch GF	truncated quadratic	d	weak
Full batch GD	truncated quadratic	$d, T \sim \log d$	strong