Learning the Optimal Step Size for Gradient Descent on Convex Quadratics

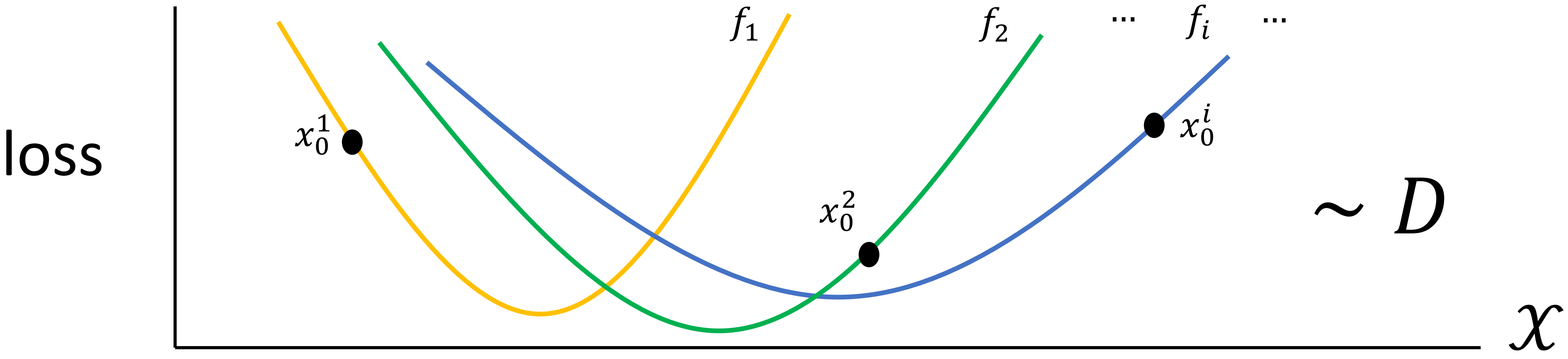
Alexandr Andoni, Daniel Hsu, Tim Roughgarden, **Kiran Vodrahalli**

Columbia University

NYAS ML Symposium, March 2020

A Distribution on Optimization Problems

Given: Distribution D over (f, x_0):



$f\colon \mathbb{R}^d \to \mathbb{R}$ is *convex quadratic*

Learning Gradient Descent Step Size

$$\boxed{x_{t+1} = x_t - \eta \nabla f(x_t)}$$

After L steps: $x_L(\eta, x_0)$

Goal: Learn optimal *single* step size η for distribution D.

$$\eta_L^* = \operatorname*{argmin}_{\eta} \mathbb{E}_{f, x_0 \sim D}[f(x_L(\eta, x_0))]$$

Motivation

- Gupta + Roughgarden ’17: Sample complexity of learning step size of GD
- How much does learning the step size help performance?
- Push the limits of performance of a single step size

Convex Quadratic Loss

$$f\left(x_{L}(\eta, x_{0})\right) = ||Ax_{L}(\eta, x_{0}) - b||_{2}^{2}$$

Convex Quadratic Loss

$$f\big(x_L(\eta, x_0)\big) = ||Ax_L(\eta, x_0) - b||_2^2$$

$$= \sum_{i=1}^{d} \sigma_i^2 z_i^2 \big(1 - \eta\sigma_i^2\big)^{2L}$$

Convex Quadratic Loss

$$f\big(x_L(\eta, x_0)\big) = ||Ax_L(\eta, x_0) - b||_2^2$$

$$= \sum_{i=1}^{d} \sigma_i^2 z_i^2 \left(1 - \eta\sigma_i^2\right)^{2L}$$

$$\boxed{\begin{array}{c} A = U\Sigma V^T \in \mathbb{R}^{n\times d};\ n \geq d;\ \ z = V^T(x_0 - x^*);\ Ax^* = b \\ \textbf{spectrum}(AA^T) = \sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_d^2 > 0 \end{array}}$$

spectral ratio: $c = \frac{\sigma_1^2}{\sigma_2^2} > 1$

Worst Case

As $c \to \infty$: No benefit over regular GD!

Worst Case

spectral ratio: $c = \frac{\sigma_1^2}{\sigma_2^2} > 1$

condition number: $\kappa = \frac{\sigma_1^2}{\sigma_d^2} > 1$

As $c \rightarrow \infty$: No benefit over regular GD!

But: If $c \rightarrow \infty$, with $L = \alpha \cdot c, \alpha \in \mathbb{R}$:

$$f\left(x_L(\eta_L^*, x_0)\right) - f(x^*) \leq \exp\left(-L\left(\frac{2}{c} + \frac{1}{\kappa}\right)\right)\left(f(x_0) - f(x^*)\right)$$

spectral ratio: $c = \frac{\sigma_1^2}{\sigma_2^2} > 1$

Worst Case

condition number: $\kappa = \frac{\sigma_1^2}{\sigma_d^2} > 1$

As $c \to \infty$: No benefit over regular GD!

But: If $c \to \infty$, with $L = \alpha \cdot c, \alpha \in \mathbb{R}$:

$$f\left(x_L(\eta_L^*, x_0)\right) - f(x^*) \leq \exp\left(-L\left(\frac{2}{c} + \frac{1}{\kappa}\right)\right)\left(f(x_0) - f(x^*)\right)$$

Improvement in the limit for large L! $(c < \kappa)$

Main Theorem

spectral ratio: $c = \frac{\sigma_1^2}{\sigma_2^2} > 1$

condition number: $\kappa = \frac{\sigma_1^2}{\sigma_d^2} > 1$

Theorem (informal): For L large enough, for fixed f, x_0 :

$$f\left(x_L(\eta_L^*, x_0)\right) - f(x^*) \leq \exp\left(-L\left(2\log(1 + 1/c) + \frac{1}{\kappa}\right)\right)\left(f(x_0) - f(x^*)\right)$$

spectral ratio: $c = \frac{\sigma_1^2}{\sigma_2^2} > 1$

Main Theorem

condition number: $\kappa = \frac{\sigma_1^2}{\sigma_d^2} > 1$

Theorem (informal): For L large enough, for fixed f, x_0 :

$$f\left(x_L(\eta_L^*, x_0)\right) - f(x^*) \leq \exp\left(-L\left(2\log(1 + 1/c) + \frac{1}{\kappa}\right)\right)\left(f(x_0) - f(x^*)\right)$$

Generalizes to **expectations** over $(f, x_0) \sim D$

Main Theorem

spectral ratio: $c = \frac{\sigma_1^2}{\sigma_2^2} > 1$

condition number: $\kappa = \frac{\sigma_1^2}{\sigma_d^2} > 1$

Theorem (informal): For L large enough, for fixed f, x_0 :

$$f\left(x_L(\eta_L^*, x_0)\right) - f(x^*) \leq \exp\left(-L\left(2\log(1 + 1/c) + \frac{1}{\kappa}\right)\right)\left(f(x_0) - f(x^*)\right)$$

Generalizes to **expectations** over $(f, x_0) \sim D$

Efficient to learn η^*:

constant pseudo-dim

binary search ERM to find η^*

Key Bound – Single Problem Instance

$$\left(\frac{Z_\alpha^{\frac{1}{2L-1}} c^{\frac{2L+1}{2L-1}}}{1 + Z_\alpha^{\frac{1}{2L-1}} c^{\frac{2L+1}{2L-1}}}\right)^{2L-1} \leq \frac{f\left(x_L(\eta^*, x_0)\right)}{f\left(x_L(1/\sigma_1^2, x_0)\right)} \leq \left(\frac{Z_\beta^{\frac{1}{2L-1}} c^{\frac{2L+1}{2L-1}}}{1 + Z_\beta^{\frac{1}{2L-1}} c^{\frac{2L+1}{2L-1}}}\right)^{2L-1}$$

$$Z_\alpha = \frac{z_1^2}{\sum_{i>1} z_i^2}; Z_\beta = \frac{z_1^2}{z_2^2}$$

Spectral ratio $c = \frac{\sigma_1^2}{\sigma_2^2} > 1$