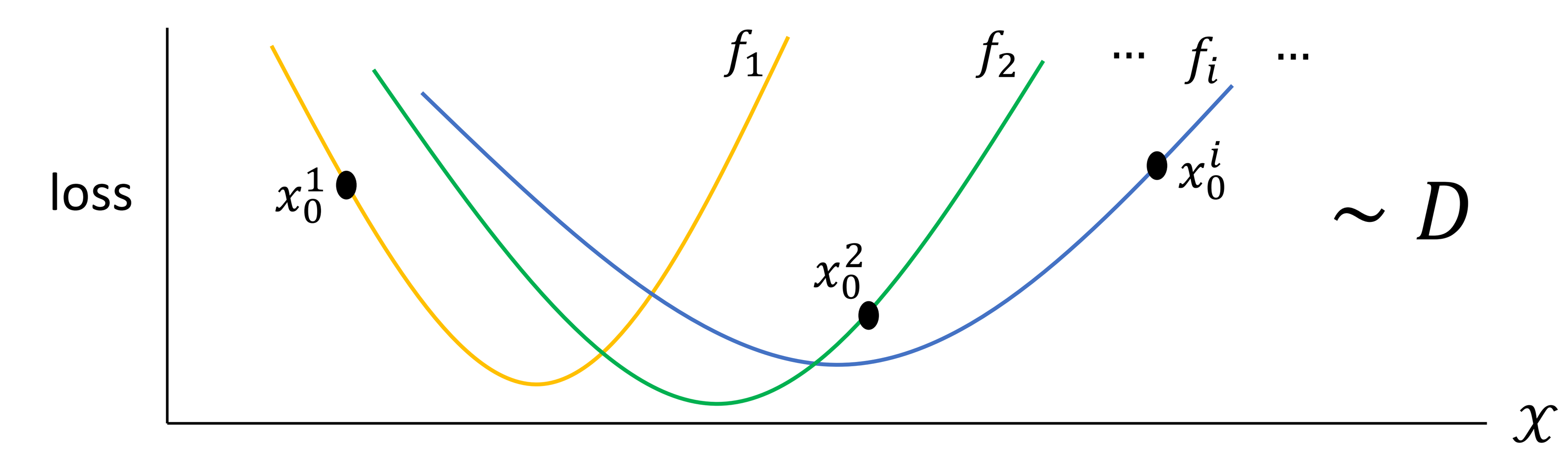
# Learning the Optimal Step Size for Gradient Descent on Convex Quadratics

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

**Given:** Distribution $D$ over $(f, x_0)$:




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

$$f(x_L(\eta, x_0)) = \|Ax_L(\eta, x_0) - b\|_2^2 = \boxed{\sum_{i=1}^d \sigma_i^2 z_i^2 (1 - \eta\sigma_i^2)^{2L}}$$

$$A = U\Sigma V^T \in \mathbb{R}^{n\times d};\ n \geq d;\ z = V^T(x_0 - x^*);\ Ax^* = b$$

$$\mathbf{spectrum}(AA^T) = \sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_d^2 > 0$$

Optimization Algorithm: Gradient Descent

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

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

**Goal:** Learn optimal *single* step size $\eta$ for distribution $D$:

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

Solve with ERM: **constant pseudo-dim binary search ERM** to find $\eta^*$

## Motivation

Gupta + Roughgarden '17[1]: 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

1. Rishi Gupta and Tim Roughgarden. A pac approach to application-specific algorithm selection. *Siam Journal on Computing (SIJCOMP)*, 46(3), 2017.

$$\underbrace{\kappa = \frac{\sigma_1^2}{\sigma_d^2}}_{\textbf{condition number}} > \underbrace{c = \frac{\sigma_1^2}{\sigma_2^2}}_{\textbf{spectral ratio}} > 1$$

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

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

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

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

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

**Proof sketch:**

1. Bound $\left(\frac{Z_\alpha^{\frac{1}{2L-1}} c^{\frac{2L+1}{2L-1}}}{1 + Z_\alpha^{\frac{1}{2L-1}} c^{\frac{2L+1}{2L-1}}}\right)^{2L-1} \leq \frac{f(x_L(\eta^*, x_0))}{f(x_L(1/\sigma_1^2, x_0))} \leq \left(\frac{Z_\beta^{\frac{1}{2L-1}} c^{\frac{2L+1}{2L-1}}}{1 + Z_\beta^{\frac{1}{2L-1}} c^{\frac{2L+1}{2L-1}}}\right)^{2L-1}$ ; $Z_\alpha = \frac{z_1^2}{\sum_{i>1} z_i^2}$ ; $Z_\beta = \frac{z_1^2}{z_2^2}$
   a) By analyzing ratio of spectral decompositions of $f$
2. Same proof generalizes to expectations