Training Dynamics of Softmax Self-Attention: Fast Global Convergence via Preconditioning

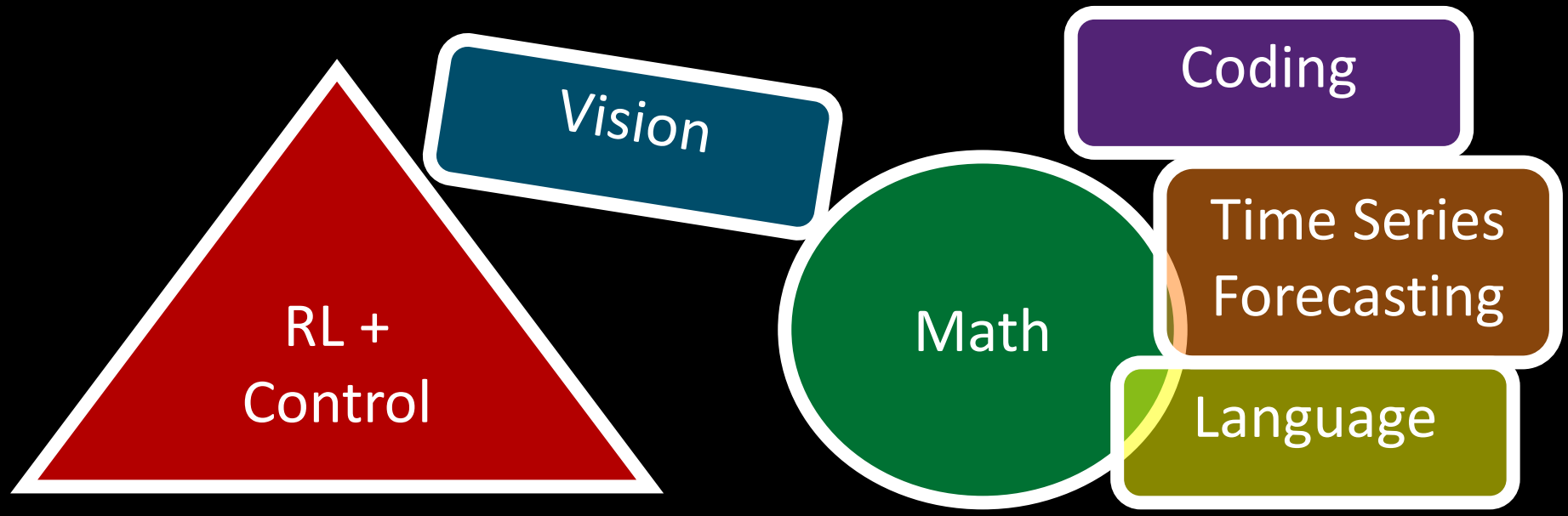
Gautam Goel

Simons Institute, UC Berkeley

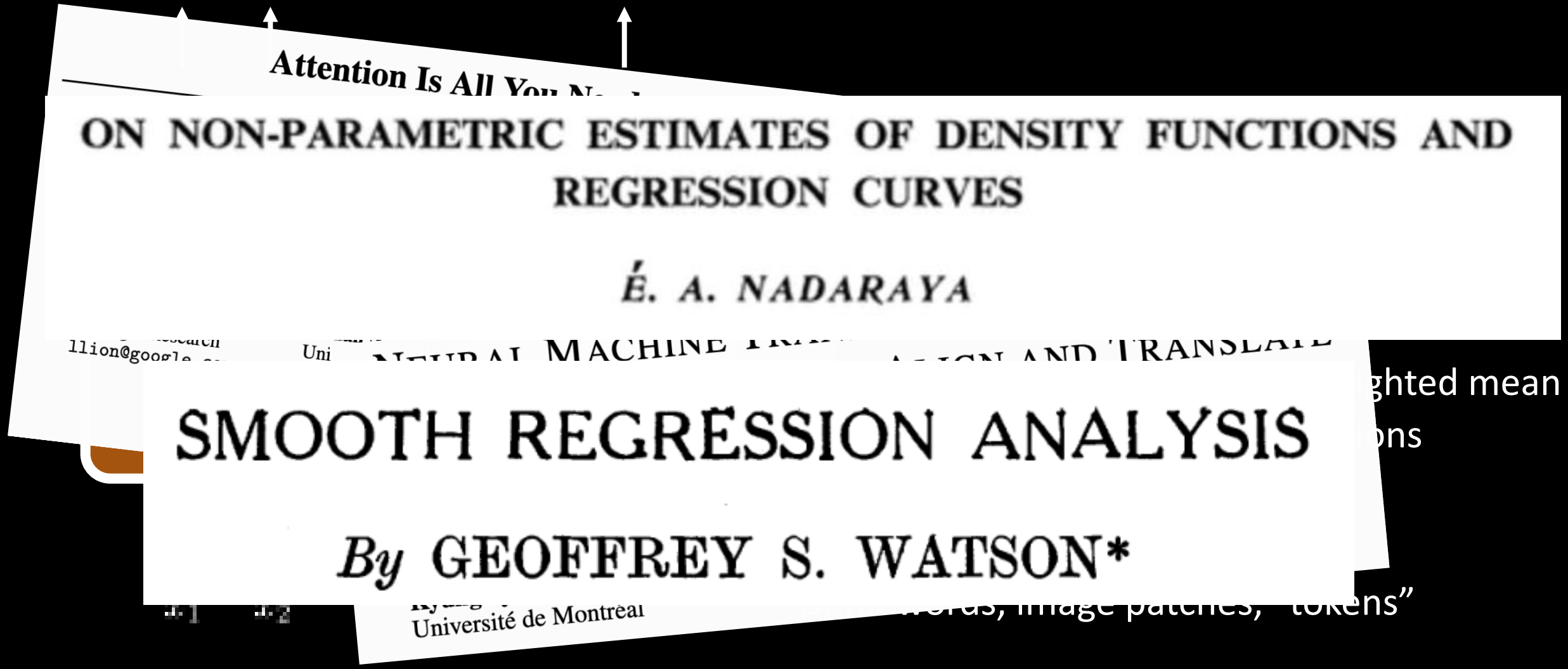
Based on joint work with Mahdi Soltanolkotabi (USC) and Peter Bartlett (Simons Institute)

One architecture to rule them all…

~~RNNs~~, ~~CNNs~~, ~~LSTMs~~, ~~VAEs~~, ~~GANs~~



Transformers + Self-Attention



ghted mean
ons

words, image patches, "tokens"

What can we learn from linear regression?

Optimization: how to solve find global optima despite nonconvexity?

Generalization: does test loss decrease with more samples and GD steps?

This talk:

GD algorithm with linear global convergence (1000x speedup vs SGD)

polylog(T) regret bound for online stochastic regression

Linear Regression using Self-Attention

Data Model: Given data $\{x_i, y_i\}_{i=1}^n$, where

- (covariates $\in \mathbb{R}^d$) $\quad x_i \sim \mathcal{N}(0, \Sigma)$ i.i.d.
- (targets $\in \mathbb{R}^p$) $\quad y_i = Mx + z_i$
- (noise $\in \mathbb{R}^p$) $\quad z_i \sim \mathcal{N}(0, \Omega)$ i.i.d.

$$\hat{y}_i = A \sum_{j=1}^{n} \frac{\exp(x_i^\top B x_j) x_j}{\sum_{j=1}^{n} \exp(x_i^\top B x_j)}$$

Training objective: choose $A \in \mathbb{R}^{p \times d}$, $B \in \mathbb{R}^{d \times d}$ to minimize the empirical loss

$$\hat{L}(A, B) = \frac{1}{2n} \sum_{i=1}^{n} \|\hat{y}_i - y_i\|_2^2$$

Vanilla GD:

Initialize:

$$A_0^{(i)} \sim \mathcal{N}(0, \sigma_A), \qquad B_0^{(i)} \sim \mathcal{N}(0, \sigma_B)$$

Update:

$$A_{t+1} = A_t - \eta \nabla_A \hat{L}(A, B), \quad B_{t+1} = B_t - \eta \nabla_B \hat{L}(A, B)$$

Population Loss as mean-field limit of Empirical Loss

Classical ERM/OLS setting:

$$\hat{L}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \|\hat{y}_i - y_i\|_2^2 = \frac{1}{2n} \sum_{i=1}^{n} \|\theta x_i - y_i\|_2^2$$

$$L(\theta) = \lim_{n \to +\infty} \frac{1}{2n} \sum_{i=1}^{n} \|\theta x_i - y_i\|_2^2 = \frac{1}{2} \mathop{\mathbb{E}}_{x,\varepsilon} \|\theta x - (Mx + \varepsilon)\|_2^2$$

Our setting:

$$\hat{L}(A, B) = \frac{1}{2n} \sum_{i=1}^{n} \|\hat{y}_i - y_i\|_2^2 = \frac{1}{2n} \sum_{i=1}^{n} \left\| A \sum_{j=1}^{n} \frac{\exp(x_i^\top B x_j) x_j}{\sum_{j=k}^{n} \exp(x_i^\top B x_k)} - y_i \right\|_2^2$$

Population Loss as mean-field limit of Empirical Loss

$$\lim_{n\to+\infty}\sum_{j=1}^{n}\frac{\exp(x_i^\top Bx_j)x_j}{\sum_{j=k}^{n}\exp(x_i^\top Bx_k)}=\lim_{n\to+\infty}\frac{\sum_{j=1}^{n}\exp(x_i^\top Bx_j)x_j}{\sum_{j=k}^{n}\exp(x_i^\top Bx_k)}=\lim_{n\to+\infty}\frac{\frac{1}{n}\sum_{j=1}^{n}\exp(x_i^\top Bx_j)x_j}{\frac{1}{n}\sum_{j=k}^{n}\exp(x_i^\top Bx_k)}$$

$$=\frac{\lim_{n\to+\infty}\left[\frac{1}{n}\sum_{j=1}^{n}\exp(x_i^\top Bx_j)x_j\right]}{\lim_{n\to+\infty}\left[\frac{1}{n}\sum_{j=k}^{n}\exp(x_i^\top Bx_k)\right]}=\frac{\mathbb{E}_{x_j}\left[\exp(x_i^\top Bx_j)x_j\right]}{\mathbb{E}_{x_k}\left[\exp(x_i^\top Bx_k)\right]}=\Sigma B^\top x_i$$

Population Loss as mean-field limit of Empirical Loss

$$L(A,B) = \lim_{n\to\infty} \frac{1}{2n} \sum_{i=1}^{n} \left\| A \lim_{n\to\infty} \left[\sum_{j=1}^{n} \frac{\exp(x_i^\top B x_j) x_j}{\sum_{j=k}^{n} \exp(x_i^\top B x_k)} \right] - y_i \right\|_2^2$$

$$= \lim_{n\to\infty} \frac{1}{2n} \sum_{i=1}^{n} \|A\Sigma B^\top x_i - y_i\|_2^2$$

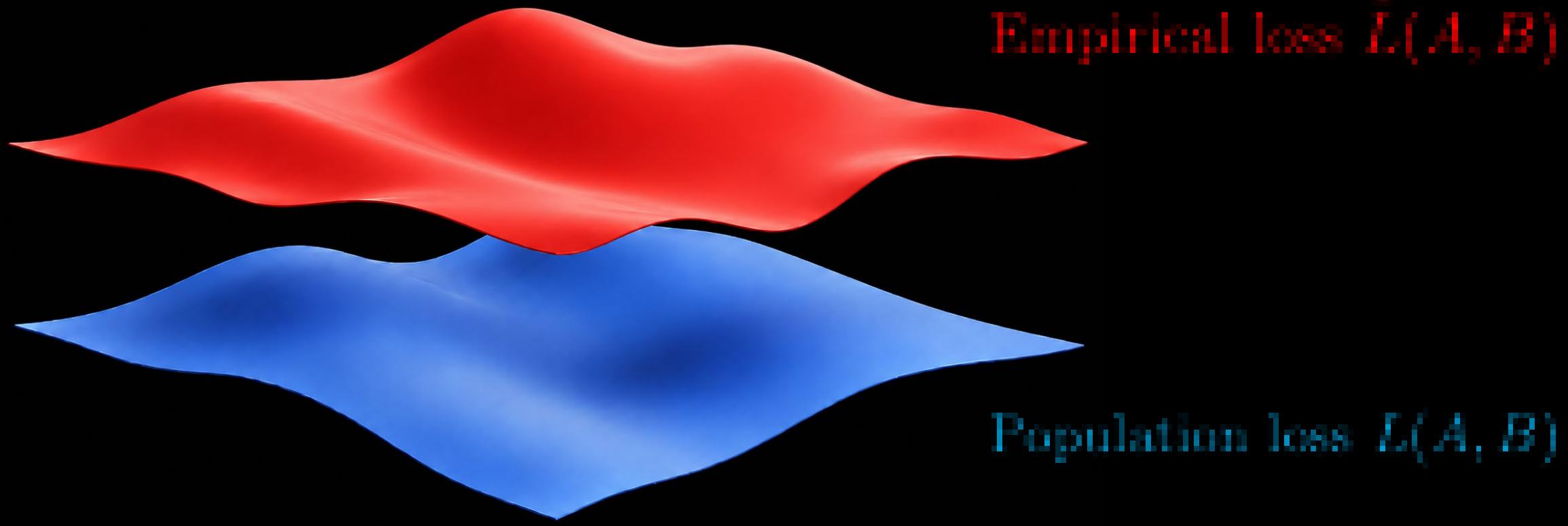
$$= \frac{1}{2} \mathop{\mathbb{E}}_{x,z} \left[\|A\Sigma B^\top x - (Mx + z)\|_2^2 \right] = \boxed{\frac{1}{2}\operatorname{Tr}(\Omega) + \frac{1}{2}\|A\Sigma B^\top \Sigma^{1/2} - M\Sigma^{1/2}\|_F^2}$$

Noise floor L^*

How well do params fit the signal?

Remark: $\mathbb{E}[\hat{L}(A,B)] \neq L(A,B)$

Idea: GD on Empirical Loss behaves like GD on Population Loss



Population Loss

$$L(A,B) = L^* + \frac{1}{2}\|A\Sigma B^\top \Sigma^{1/2} - M\Sigma^{1/2}\|_F^2$$

Observation: Looks like a "weighted" matrix factorization loss!

$$F(U,V) = \frac{1}{2}\|UV^\top - M\Sigma^{1/2}\|_F^2$$

Idea: add regularizer to make population loss strongly convex!

Low-rank Solutions of Linear Matrix Equations via Procrustes Flow [Tu et al '16]

Global minima of regularized loss

$$L(A,B) = L^* + \frac{1}{2}\left\|A\Sigma B^\top \Sigma^{1/2} - M\Sigma^{1/2}\right\|_F^2$$

$$R(A,B) = \frac{1}{8}\left\|\Sigma^{1/2}\left(A^\top A - B^\top \Sigma B\right)\Sigma^{1/2}\right\|_F^2$$

$$Q(A,B) = L(A,B) + R(A,B)$$

Theorem

The set of global minima of the regularized loss $Q(A,B) = L(A,B) + R(A,B)$ form a smooth connected manifold

$$S = \left\{\left(U\Gamma^{1/2}J^\top\Sigma^{-1/2}, \Sigma^{-1/2}V\Gamma^{1/2}J^\top\Sigma^{-1/2}\right)\right\},$$

where $U\Gamma V^\top$ is a singular value decomposition of $M\Sigma^{1/2}$ and $J \in \mathbb{R}^{d\times d}$ is orthogonal.

Regularized loss is one-point strongly convex

Theorem

Define the inner product

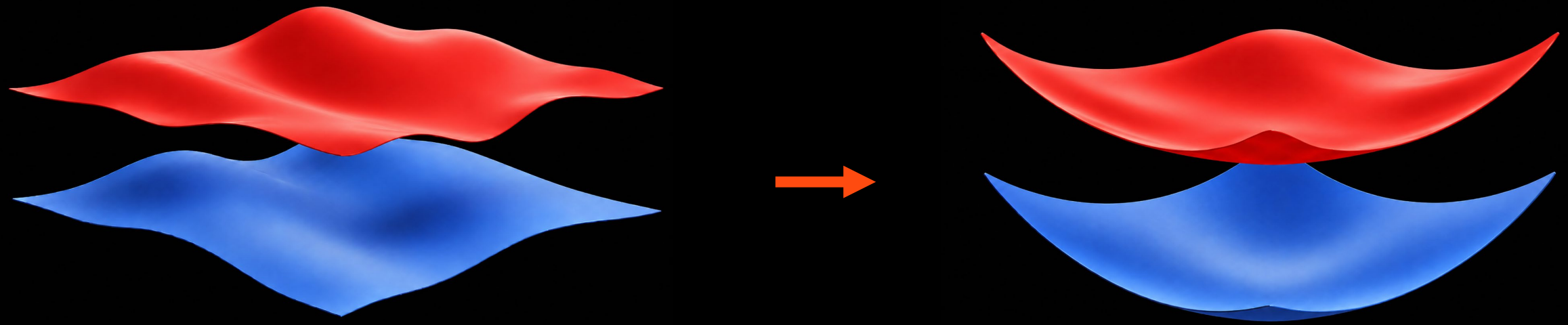
$$((A_1, B_1)(A_2, B_2))_\Sigma = \mathrm{Tr}(A_1^\top A_2) + \mathrm{Tr}(B_1^\top \Sigma B_2).$$

Let $\|\cdot\|_\Sigma$ be the norm generated by this inner product. Given any $\theta = (A, B)$ which is ν-close to $\mathcal{S}$, let $\theta^* = (A^*, B^*)$ be the projection in the Σ-norm of θ onto $\mathcal{S}$. The following inequality holds:

$$\langle P^{-1}\nabla_\theta Q(\theta), \theta - \theta^* \rangle_\Sigma \geq \alpha \|\theta - \theta^*\|_\Sigma^2,$$

where we define

$$P = \begin{bmatrix} I & 0 \\ 0 & \Sigma \end{bmatrix}.$$



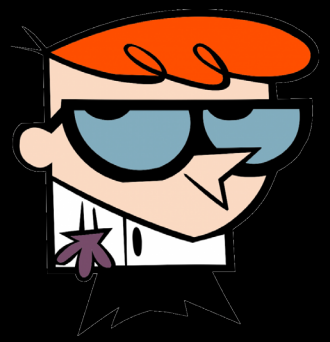
Preconditioned + Regularized GD:

Initialize:

$$(A_0, B_0) \text{ near } S: \quad (A_0, B_0) = (\hat{U}\hat{\Gamma}^{1/2}J^\top \hat{\Sigma}^{-1/2}, \hat{\Sigma}^{-1/2}\hat{V}\hat{\Gamma}^{1/2}J^\top \hat{\Sigma}^{-1/2})$$

Update:

$$A_{t+1} = A_t - \eta \nabla_A \hat{Q}(A, B), \quad B_{t+1} = B_t - \eta \hat{\Sigma}^{-1}\nabla_B \hat{Q}(A, B)$$



$$\hat{U}\hat{\Gamma}\hat{V}^\top = \text{SVD}(\hat{M}\hat{\Sigma}^{1/2})$$

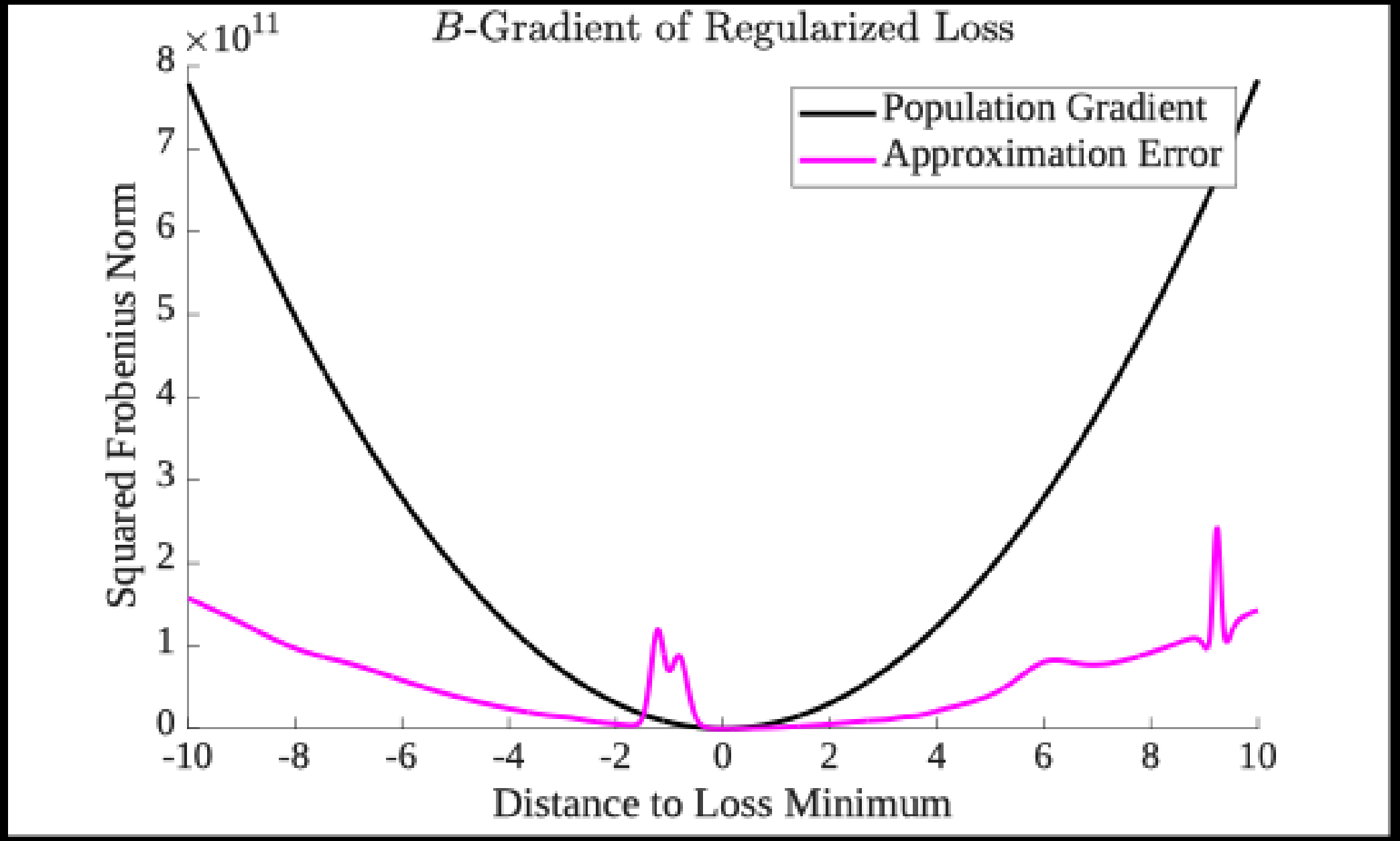
$$\hat{Q}(A, B) = \hat{L}(A, B) + \hat{R}(A, B)$$

$$\hat{R}(A, B) = \frac{1}{8}\left\|\hat{\Sigma}^{1/2}(A^\top A - B^\top \hat{\Sigma} B)\hat{\Sigma}^{1/2}\right\|_F^2$$

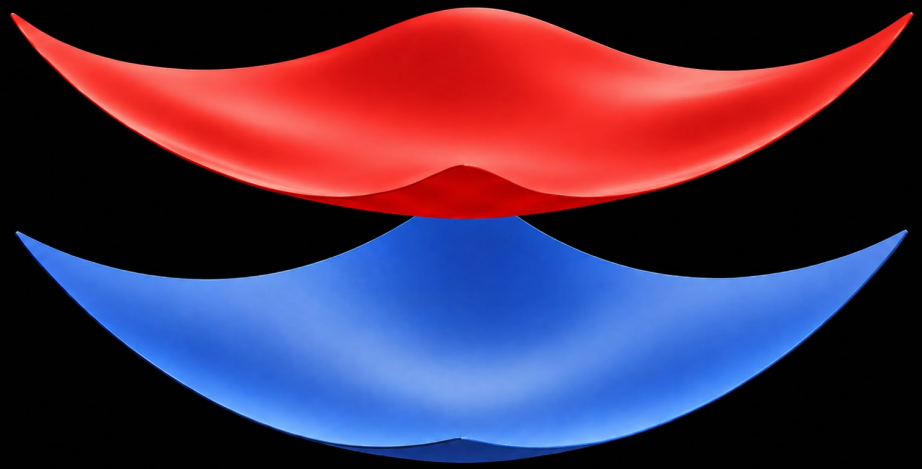
Does this actually work?

Spectral initialization: yes

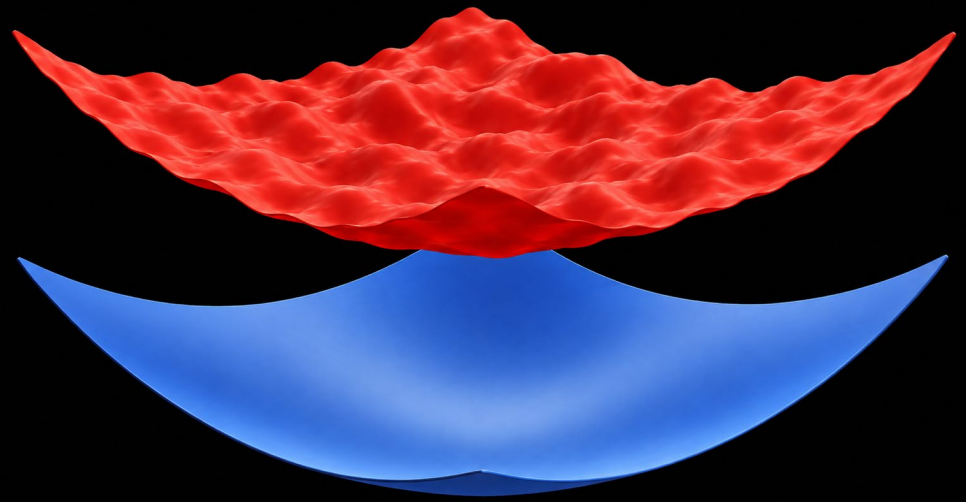
Random initialization: no


B-Gradient of Regularized Loss
$\times 10^{11}$
Squared Frobenius Norm
0
1
2
3
4
5
6
7
8
-10
-8
-6
-4
-2
0
2
4
6
8
10
Distance to Loss Minimum
Population Gradient
Approximation Error

want



have



Empirical Gradient

$$\nabla_A \hat{L}(A,B) = \frac{1}{n}\sum_{i=1}^{n}(A\mu_i - Mx_i)\,\mu_i^\top, \qquad \nabla_B \hat{L}(A,B) = \frac{1}{n}\sum_{i=1}^{n} x_i(A\mu_i - Mx_i)^\top A\Sigma_i,$$

$$\mu_i = \sum_{j=1}^{n} p_{ij}x_j, \qquad \Sigma_i = \sum_{j=1}^{n} p_{ij}x_j x_j^\top - \mu_i\mu_i^\top, \qquad p_{ij} = \frac{\exp(x_j^\top Bx_i)}{\sum_{k=1}^{n}\exp(x_k^\top Bx_i)}$$

Population Gradient

$$\nabla_A L(A,B) = (A\Sigma B^\top - M)\Sigma B\Sigma, \qquad \nabla_B L(A,B) = \Sigma(A\Sigma B - M)^\top A\Sigma,$$

Concentration of Softmax mean

$$\mu_i - \Sigma B^\top x_i = \underbrace{\mu_i - \mathbb{E}_{-i}[\mu_i]}_{\leq \frac{K}{n} \exp\left(\frac{1}{2} x_i^\top B \Sigma B^\top x_i\right)} + \underbrace{\mathbb{E}_{-i}[\mu_i] - \Sigma B^\top x_i}_{\leq \frac{K}{n} \exp\left(\frac{1}{2} x_i^\top B \Sigma B^\top x_i\right)}$$

(concentration inequalities) **(Gaussian IBP)**

$$\frac{1}{2} x_i^\top B \Sigma B^\top x_i \leq \frac{1}{2} \|\Sigma\|_{\mathrm{op}} \|B\|_{\mathrm{op}}^2 \cdot \sup_{i \in [n]} \|x_i\|_2^2 \leq 1 \quad \text{if} \quad \|B\|_{\mathrm{op}} \leq \frac{K}{\|\Sigma\|_{\infty} \sqrt{\log(n)}}$$

Set of θ's where GD on $L(\theta)$ finds L^*
(θ's which are ν-close to S)

Set of θ's where $\nabla_\theta \hat{L}(\theta) \approx \nabla_\theta L(\theta)$

$$\left(\|B\|_{\text{op}} \leq \frac{K}{\|\Sigma\|_{\text{op}} \sqrt{\log(n)}} \right)$$

Reparameterization to the rescue!

Fix $\gamma > 0$. Let $(A, B) = \phi(\tilde{A}, \tilde{B}) = \left(\gamma^{-1}\|\tilde{B}\|_F \tilde{A}, \gamma\|\tilde{B}\|_F^{-1}\tilde{B}\right)$.

$$L^\phi(\tilde{A}, \tilde{B}) = L(\phi(\tilde{A}, \tilde{B}))$$
$$\hat{L}^\phi(\tilde{A}, \tilde{B}) = \hat{L}(\phi(\tilde{A}, \tilde{B}))$$

This changes the gradients!
New gradients from Chain Rule

$$\begin{aligned} L^\phi(\tilde{A}, \tilde{B}) &= L(A, B) \\ &= \frac{1}{2}\|A\Sigma B^\top \Sigma^{1/2} M \Sigma^{1/2}\|_F^2 \\ &= \frac{1}{2}\left\|\left(\gamma^{-1}\|\tilde{B}\|_F \tilde{A}\right)\Sigma\left(\gamma\|\tilde{B}\|_F^{-1}\tilde{B}\right)^\top \Sigma^{1/2} M \Sigma^{1/2}\right\|_F^2 \\ &= \frac{1}{2}\|\tilde{A}\Sigma\tilde{B}^\top \Sigma^{1/2} M \Sigma^{1/2}\|_F^2 \end{aligned}$$

Same functional form as old L!

Preconditioned + Regularized GD (Take Two):

Initialize:

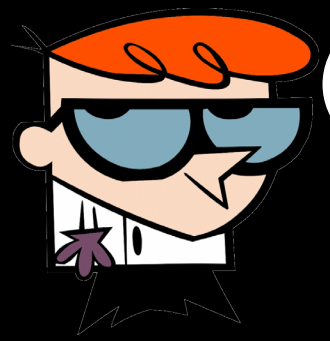
$$(\tilde{A}_0, \tilde{B}_0) \text{ near } S: \quad (\tilde{A}_0, \tilde{B}_0) = (\hat{U}\hat{\Gamma}^{1/2} J^\top \hat{\Sigma}^{-1/2},\ \hat{\Sigma}^{-1/2}\hat{V}\hat{\Gamma}^{1/2} J^\top \hat{\Sigma}^{-1/2})$$

Update:

$$\tilde{A}_{t+1} = \tilde{A}_t - \eta \nabla_A Q(\tilde{A}, \tilde{B}), \quad \tilde{B}_{t+1} = \tilde{B}_t - \eta \Sigma^{-1} \nabla_B Q(\tilde{A}, \tilde{B})$$

Output:

$$(A_m, B_m) = \phi(\tilde{A}_m, \tilde{B}_m)$$



$$\hat{Q}^\rho(\tilde{A}, \tilde{B}) = \hat{L}^\rho(\tilde{A}, \tilde{B}) + \hat{R}(\tilde{A}, \tilde{B})$$

Theorem

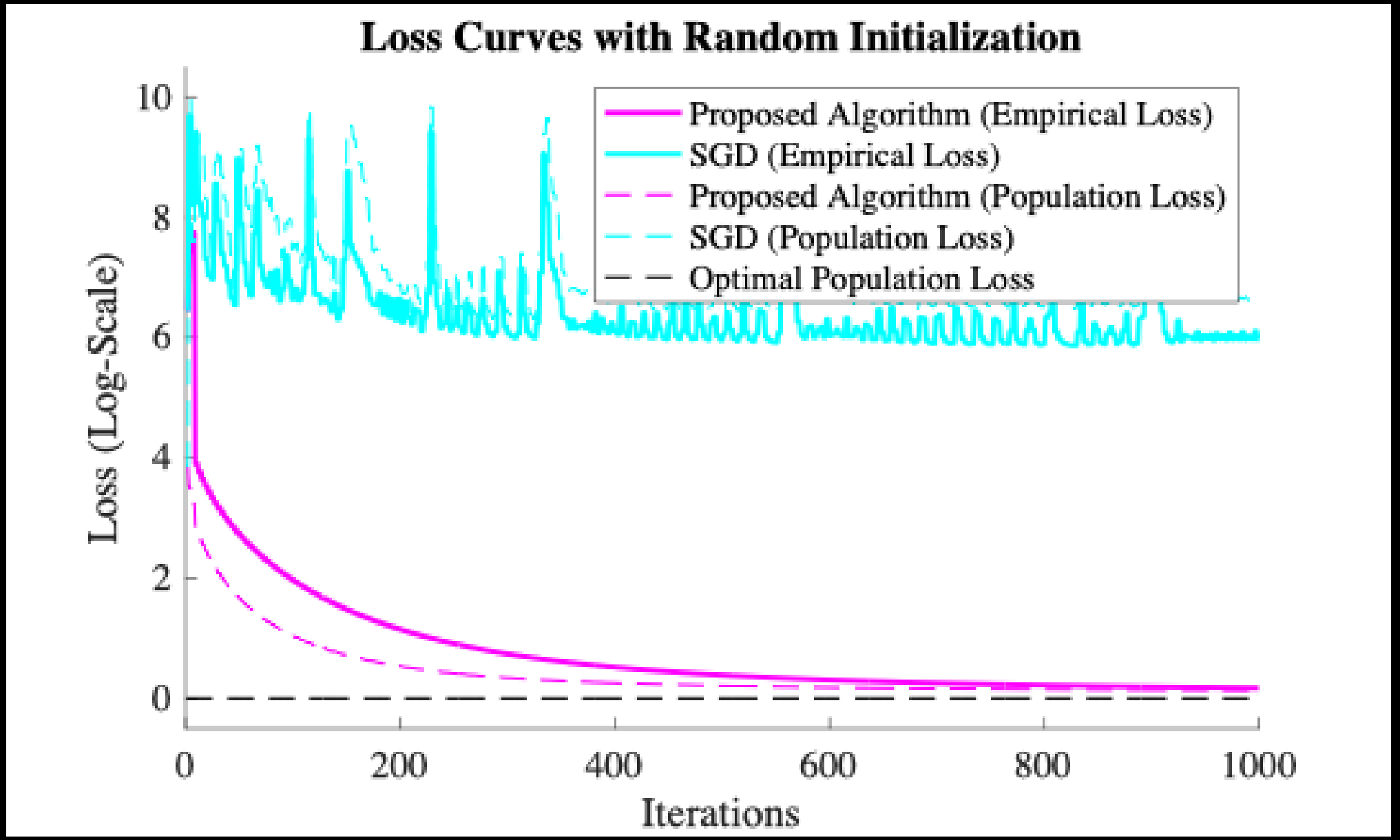
Run the algorithm for m steps with $\gamma = K\|\Sigma\|_{\mathrm{op}}^{-1}\log^{-1/2}(n)$. Fix any $\delta \in (0,1)$. If n is sufficiently large and η is sufficiently small, then with probability $1-\delta$,

$$L(\theta_m) - L^* \leq \frac{K_1 \log^4(n)\log(mn/\delta)}{n} + K_2\exp(-K_3 m).$$

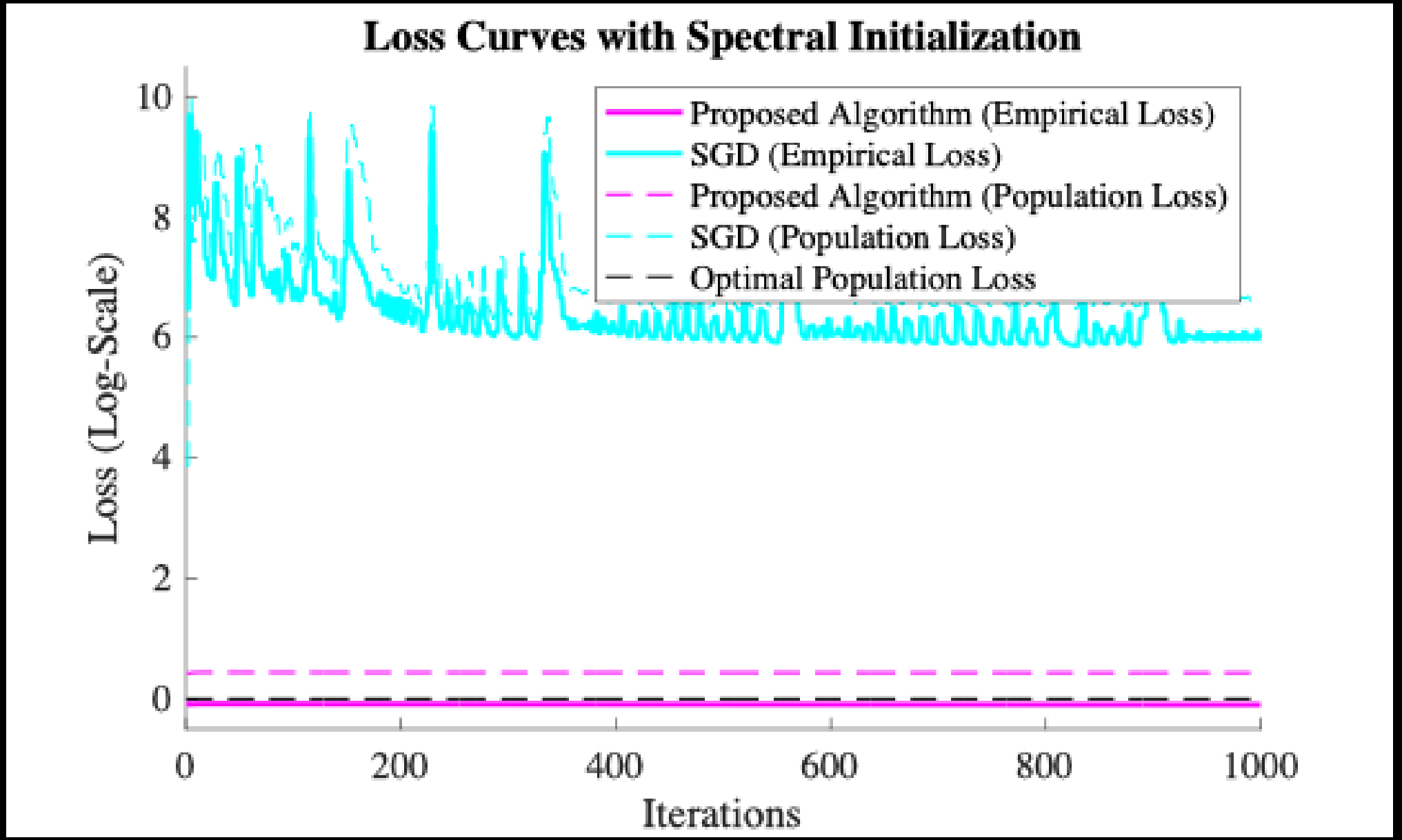
excess risk | statistical error | optimization error

This matches the excess risk of the OLS estimator!

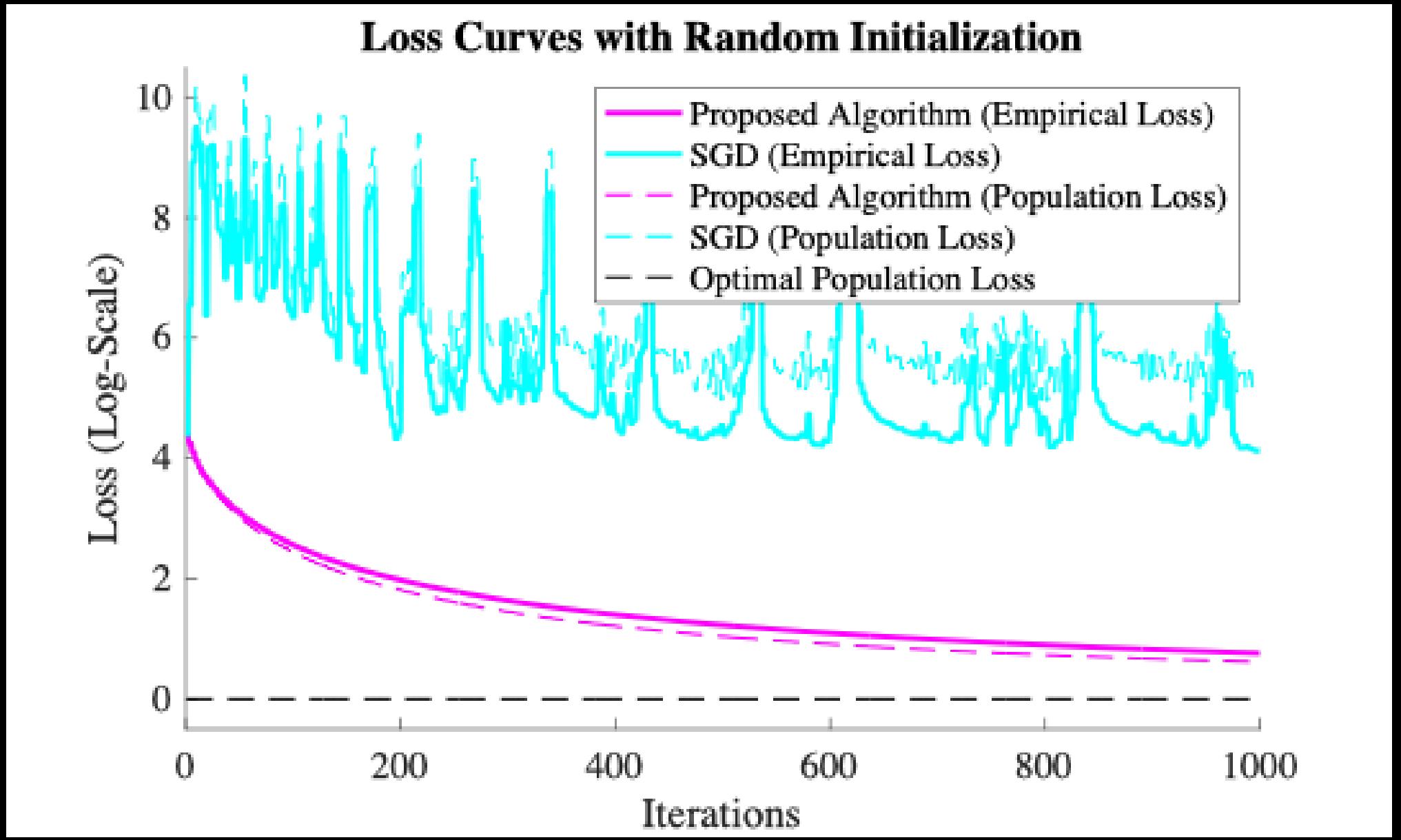
Experiments



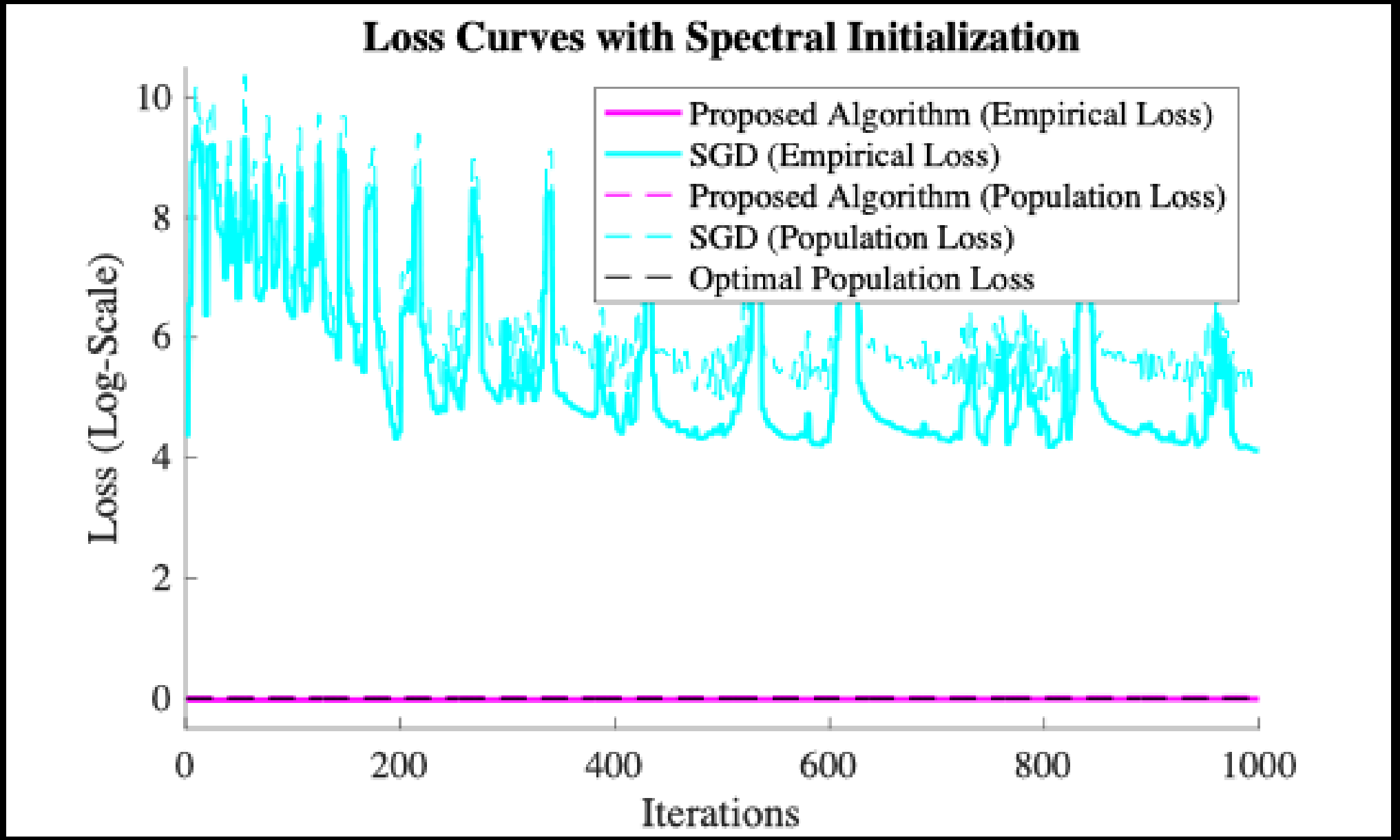
$n = 50,\ k = 50,\ p = 100,\ d = 10,\ \eta = 0.001,\ \gamma = 1.0$



$n = 50,\ k = 50,\ p = 100,\ d = 10,\ \eta = 0.001,\ \gamma = 1.0$



$n = 500,\ k = 50,\ p = 100,\ d = 10,\ \eta = 0.001,\ \gamma = 1.0$



$n = 500,\ k = 50,\ p = 100,\ d = 10,\ \eta = 0.001,\ \gamma = 1.0$

Online Learning

Online Stochastic Linear Regression

1. Learner recieves $x_t \sim \mathcal{N}(0, \Sigma)$
2. Learner predicts $\hat{y}_t$
3. Learner pays cost $\|\hat{y}_t - y_t\|_2^2$, where $y_t = Mx_t + z_t$

Follow-the-Leader/Online Least Squares achieves log(T) regret against M

$$\hat{y}_t = A_t \sum_{j=1}^{t} \frac{\exp(x_t^\top B_t x_j)\, x_j}{\sum_{j=1}^{t} \exp(x_t^\top B_t x_j)}$$

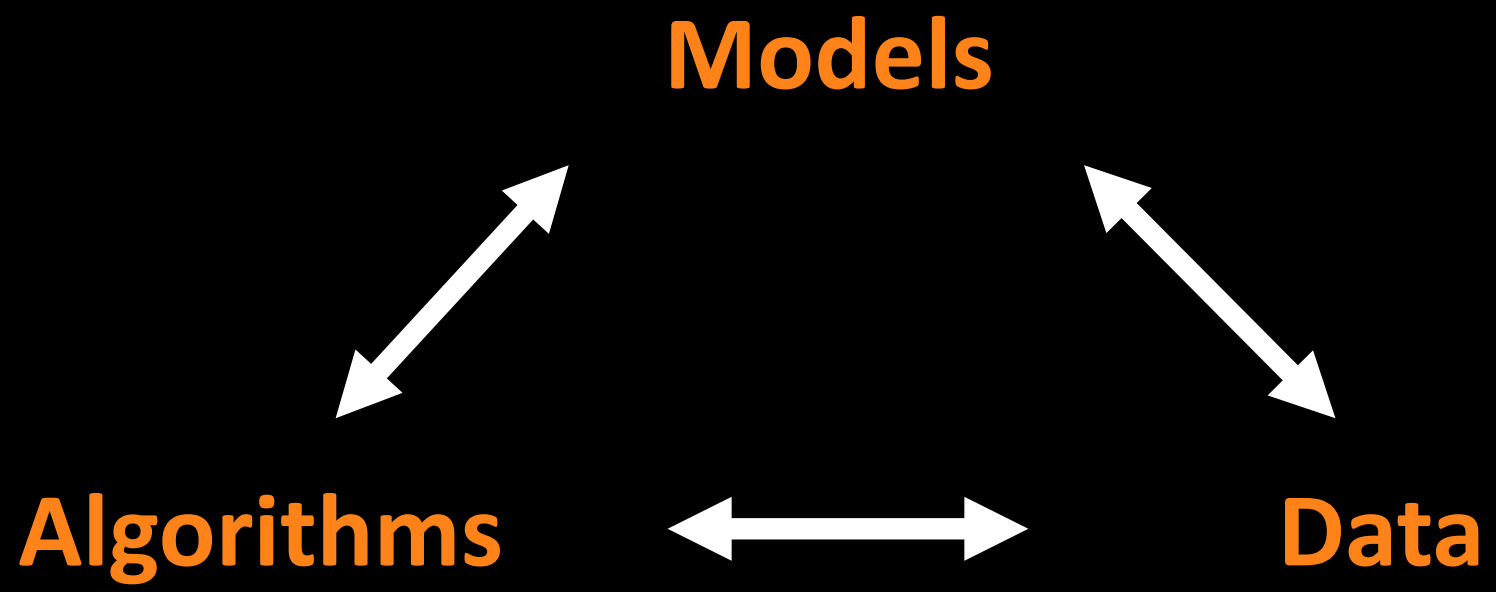
Theorem

Fix any $\delta_0 \in (0,1)$. Set $\gamma = K\|\Sigma\|_{\mathrm{op}}^{-1}\log^{-1/2}(T)$. Run FTL with $m = K\log(T)$ steps in each round, where the failure probability is set to $\delta = \delta_0/T$. If T is sufficiently large and η is sufficiently small, then with probability $1-\delta_0$,

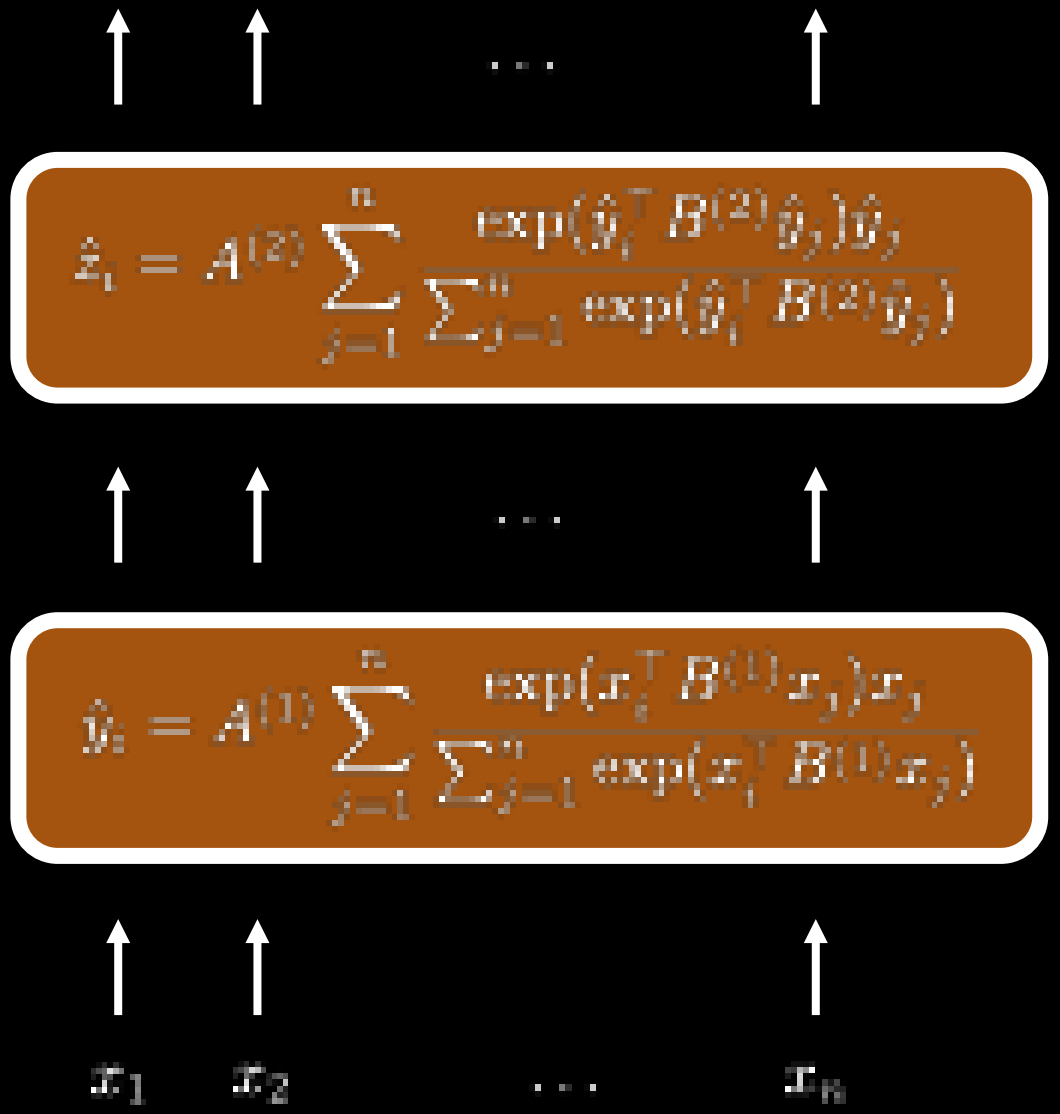
$$\mathrm{Regret}(M) \le \log^5(T)\log(T^2/\delta_0).$$

Intuition: pay 1/t cost in round t, sum harmonic series

Future Directions



Models



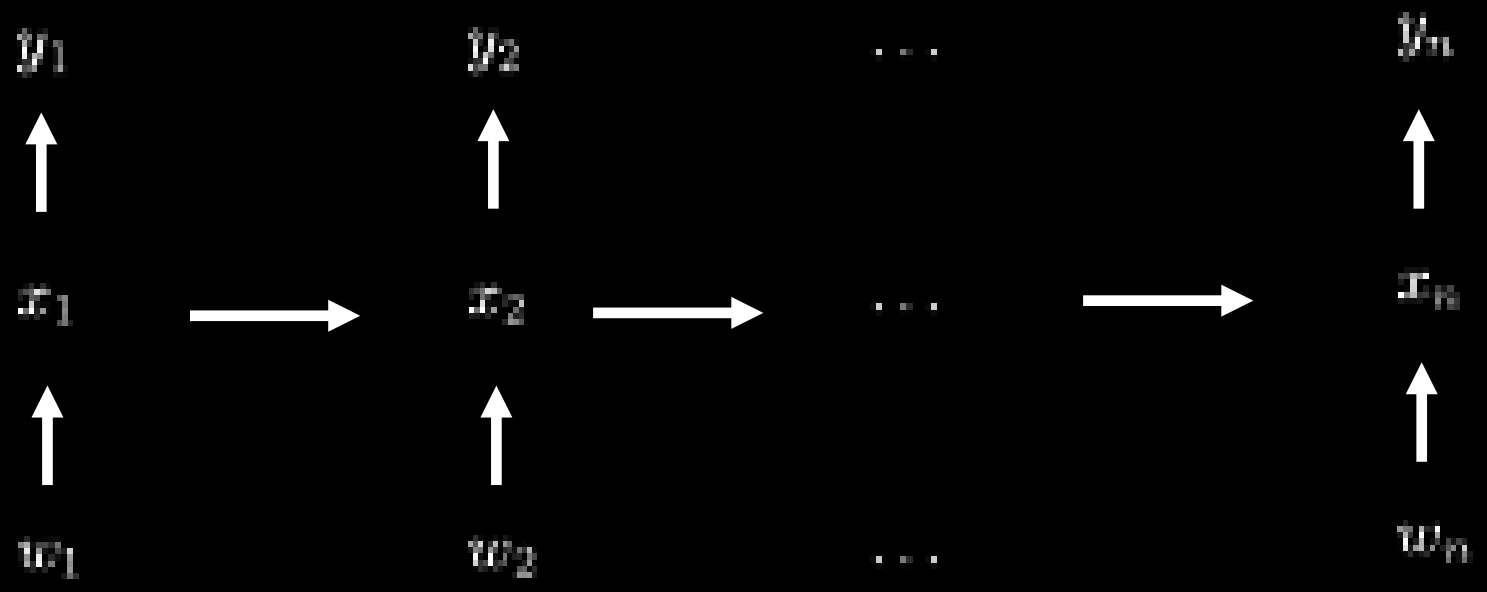
$$L(A^{(1)}, B^{(1)}, A^{(2)}, B^{(2)})$$

$$= L^* + \tfrac{1}{2}\|A^{(2)}\tilde{\Sigma}B^{(2)T}\tilde{\Sigma}^{1/2} - M\Sigma^{1/2}\|_F^2$$

where

$$\tilde{\Sigma} = A^{(1)}\Sigma B^{(1)T}\Sigma B^{(1)}\Sigma A^{(1)T}$$

Data



$$x_{t+1} = f(x_t, w_t) \qquad y_t = g(x_t)$$

Algorithms

Usual questions

- Adaptive stepsize
- Implicit Bias / Overparameterization

Replace FTL with OGD

- Less memory, fast updates

Learning in nonstationary environments

- Discounting, Changepoint detection
- Adaptive regret

Bandits, RL/Control

- What if the learner only sees loss?
- What if learner's actions affect future loss? (nonlinear effects compound)

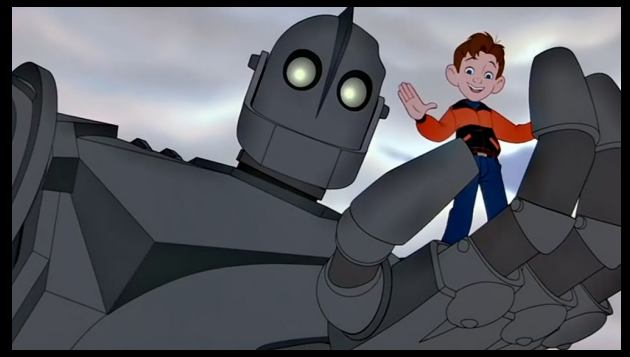
Speed up SGD

- Classical SGD: O(n) to O(k)
- Softmax SGD: O(n^2) to O(kn)
- Space-partitioning algorithms

Linear Regression
(iid data)
(one layer)

Nonlinear Regression
(dependent data)
(many layers)

Online Control
Kalman Filtering
Bandits
RL
…



???