Hitting the High-D(imensional) Notes:

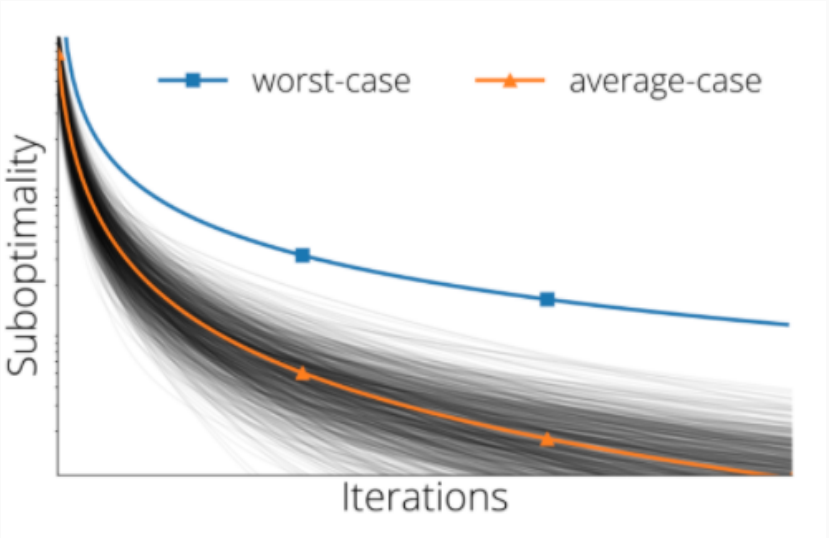
An ODE for SGD learning dynamics

Courtney Paquette (McGill & Google DeepMind)

Joint work: Elliot Paquette (McGill), Kiwon Lee (McGill), Elizabeth Collins-Woodfin (McGill), Inbar Seroussi (Tel-Aviv), Jeffrey Pennington (Google DeepMind), Ben Adlam (Google DeepMind), Andrew Cheng (PhD, Harvard)

Challenges/Problems for Optimization in Machine Learning

- ✓ Theory vs. Practice
- ✓ Mismatch with assumptions
  $\rightarrow$ too general in optimization theory
- ✓ Less convergence of algorithm

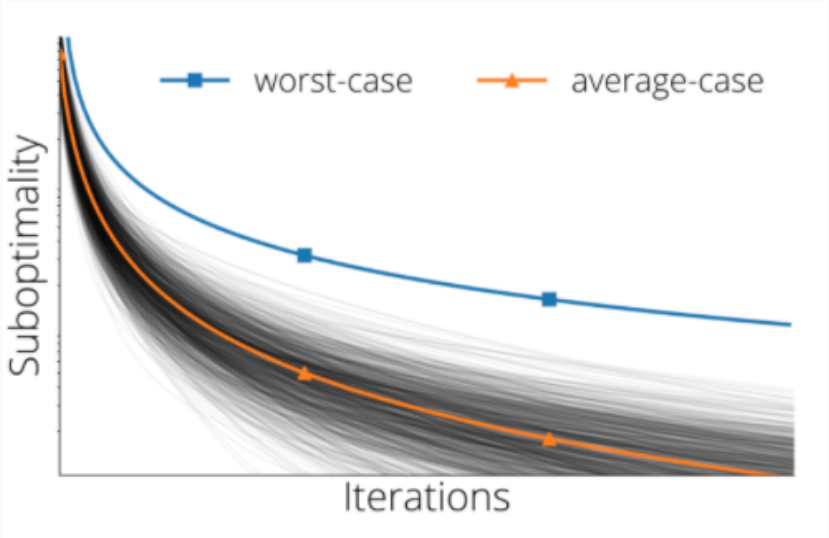


Challenges/Problems for Optimization in Machine Learning

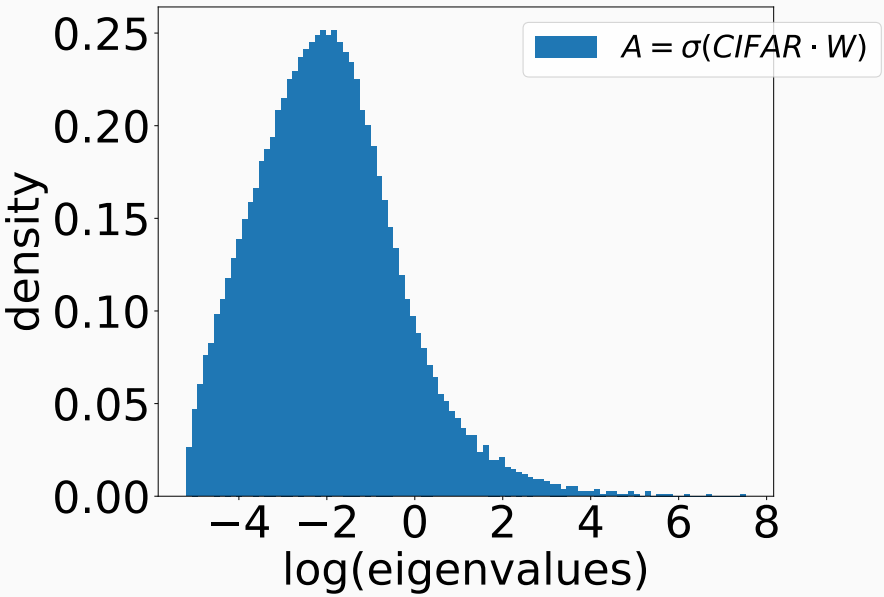
✓ Theory vs. Practice

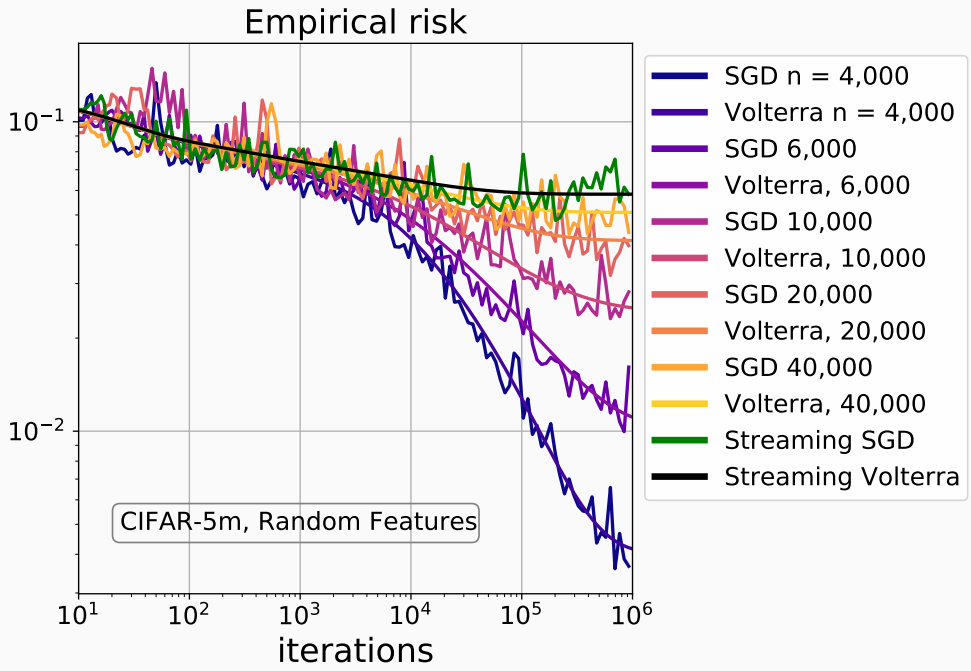
✓ Mismatch with assumptions
→ too general in optimization theory

✓ Less convergence of algorithm



Theory meets practice: CIFAR-5m





$$\min_{\boldsymbol{x}} \ \|\sigma(\text{CIFAR} \cdot \boldsymbol{W})\boldsymbol{x} - \boldsymbol{b}\|^2$$

Using a random features model to predict CIFAR-5m (Nakkiran et al., '21) car/plane, the **Volterra equation** (using the Hessian spectra an input) gives **good predictions** for behavior of SGD.

Typical Machine Learning Problems

$$\min_{X \in \mathbb{R}^d} \mathcal{L}(X) := \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{x})$$

High dimensional $\Leftrightarrow$ large number of features (d) and samples (n)

✓ State-of-the-art models have millions/billions parameters

- Meena: 2.6 billion, Turing NLG: 17 billion, GPT-3: 175 billion

Typical Machine Learning Problems

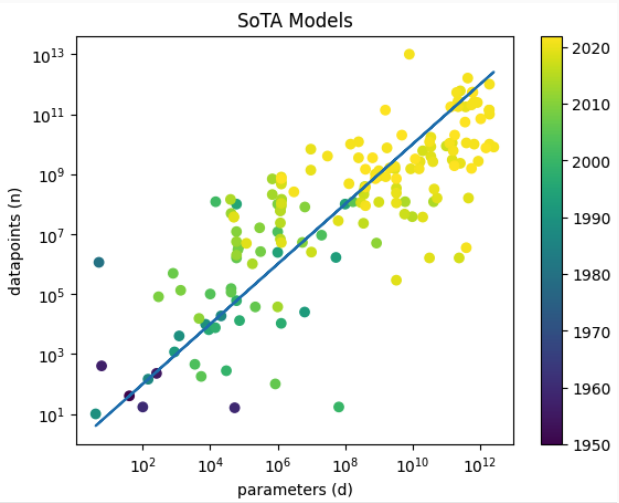
$$\min_{X \in \mathbb{R}^d} \mathcal{L}(X) := \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{x})$$

High dimensional $\Leftrightarrow$ large number of features (d) and samples (n)

✓ State-of-the-art models have millions/billions parameters

- Meena: 2.6 billion, Turing NLG: 17 billion, GPT-3: 175 billion

✓ Ratio of features (d) to samples (n) is constant, $d^{\alpha} \leq n \leq d^{1/\alpha}$



Andrew Cheng, data from "Parameter Counts in Machine Learning", lesswrong.com, 2023

High-dimensions

What's different about **high-dimensions**?

Input which generates worst complexity can be far from typical

"more room = more possibilities"

High-dimensions

What's different about **high-dimensions**?

Input which generates worst complexity can be far from typical

"more room = more possibilities"

How do we capture high-dimensional structure?

Probability distribution on the inputs

Remark: Some results will hold for deterministic designs

Statistical learning (Mei & Montanari '19, Adlam & Pennington '21, Louart & Liao & Couillet '18)

Numerical Methods (Trogdon & Deift '19, Chandrasekher '21)

Set-up

High-dimensional linear composites

$$\min_{X \in \mathbb{R}^{d \times m}} \left\{ \mathcal{R}(X) = \mathbb{E}_{a,\epsilon}[f(a^T X; a^T X^\star, \epsilon)] \right\}$$

- $a \sim N(0, K)$ data
- Covariance $K = \mathbb{E}[aa^T]$, $\|K\|_{\text{op}}$ bounded, independent of d
- ϵ is label noise, $X^\star \in \mathbb{R}^{d \times m^\star}$
- **Idea:** Think of $f : \mathbb{R}^m \to \mathbb{R}$ as pseudo-Lipschitz and low dimensional, i.e. $m, m^\star \ll d$ large

Set-up

High-dimensional linear composites

$$\min_{X \in \mathbb{R}^{d \times m}} \left\{ \mathcal{R}(X) = \mathbb{E}_{a,\epsilon}[f(a^T X; a^T X^\star, \epsilon)] \right\}$$

- $a \sim N(0, K)$ data
- Covariance $K = \mathbb{E}[aa^T]$, $\|K\|_{\text{op}}$ bounded, independent of d
- ϵ is label noise, $X^\star \in \mathbb{R}^{d \times m^\star}$
- **Idea:** Think of $f : \mathbb{R}^m \to \mathbb{R}$ as pseudo-Lipschitz and low dimensional, i.e. $m, m^\star \ll d$ large

What does this allow?

- GLMs, multi-index models
- e.g., multi-class logistic regression $m =$ number of classes
- $X^\star$ – ground truth signal, mean of data, fixed vector

One-pass SGD

$$\min_{X \in \mathbb{R}^{d \times m}} \left\{ \mathcal{R}(X) = \mathbb{E}_{a,\epsilon}[f(a^T X; a^T X^\star, \epsilon)] \right\}$$

One-pass SGD: Generate new $(a_{k+1}, \epsilon_{k+1})$

$$X_{k+1} = X_k - \frac{\gamma_k}{d} a_{k+1} \otimes \nabla f(a_{k+1}^T X_k),$$

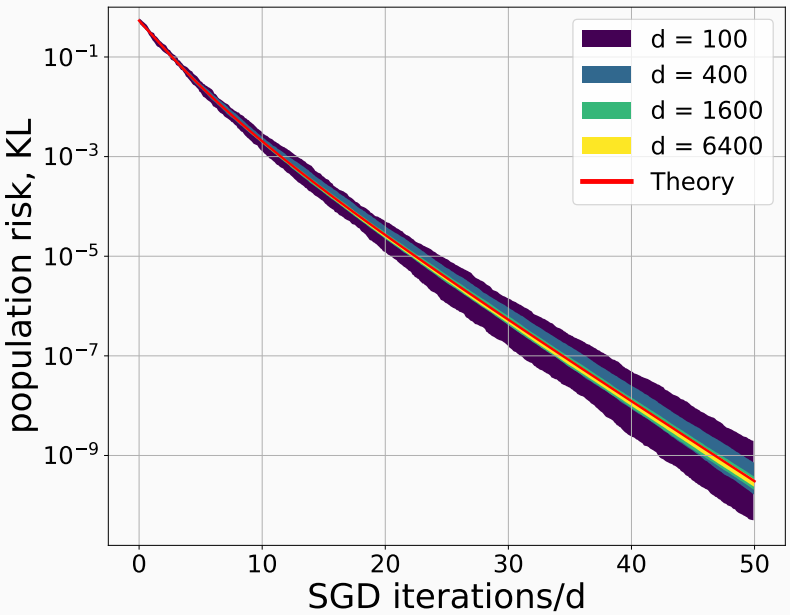
One-pass SGD

$$\min_{X \in \mathbb{R}^{d \times m}} \left\{ \mathcal{R}(X) = \mathbb{E}_{a,\epsilon}[f(a^T X; a^T X^\star, \epsilon)] \right\}$$

One-pass SGD: Generate new $(a_{k+1}, \epsilon_{k+1})$

$$X_{k+1} = X_k - \frac{\gamma_k}{d} a_{k+1} \otimes \nabla f(a_{k+1}^T X_k),$$

For **large models**, as



$$\frac{\text{parameters}}{\text{samples}} = \frac{d}{n} \to r,$$

- $\mathcal{R}(X_k) \xrightarrow{\Pr}$(smooth function)
- Analyze this smooth function
- Determined by the spectrum of the covariance matrix $K = \mathbb{E}[aa^T]$

Deterministic Statistics:

Why?

- High-dimensional compositional structure, $a^T X$ averages out d

Deterministic Statistics:

Why?

- High-dimensional compositional structure, $a^T X$ averages out d
- Let $W \stackrel{\text{def}}{=} [X|X^\star]$. Lots of statistics can be represented by

$$\varphi(X) = g(W^T q(K) W) = g\left(\begin{bmatrix} X^T q(K) X & X^T q(K) X^\star \\ (X^\star)^T q(K) X & (X^\star)^T q(K) X^\star \end{bmatrix}\right),$$

where q is a polynomial, $K = \mathbb{E}[aa^T]$.

Examples: Risk $\mathcal{R}(X)$, $\|\nabla \mathcal{R}\|^2$, distance to optimality $\|X - X^\star\|^2$

Examples

Least squares:

$$\mathcal{R}(X) = \mathbb{E}_{a,\epsilon}[\mathrm{tr}((a^T X - (a^T X^\star + \epsilon))^2)] = \mathrm{tr}((X - X^\star)^T K (X - X^\star)) + \mathbb{E}[\mathrm{tr}(\epsilon \epsilon^T)]$$

Examples

Least squares:

$$\mathcal{R}(X) = \mathbb{E}_{a,\epsilon}[\mathrm{tr}((a^T X - (a^T X^\star + \epsilon))^2)] = \mathrm{tr}((X - X^\star)^T K (X - X^\star)) + \mathbb{E}[\mathrm{tr}(\epsilon \epsilon^T)]$$

Logistic Loss (binary, noiseless): Student-teacher generate targets $y = \frac{\exp(a^T X^\star)}{\exp(a^T X^\star)+1}$. Then the risk becomes

$$\mathcal{R}(X) = \mathbb{E}_a \left[-a^T X \cdot \frac{\exp(a^T X^\star)}{\exp(a^T X^\star) + 1} + \log\left(\exp(a^T X) + 1\right) \right].$$

Examples

Least squares:

$$\mathcal{R}(X) = \mathbb{E}_{a,\epsilon}[\mathrm{tr}((a^T X - (a^T X^\star + \epsilon))^2)] = \mathrm{tr}((X - X^\star)^T K (X - X^\star)) + \mathbb{E}[\mathrm{tr}(\epsilon\epsilon^T)]$$

Logistic Loss (binary, noiseless): Student-teacher generate targets $y = \frac{\exp(a^T X^\star)}{\exp(a^T X^\star)+1}$. Then the risk becomes

$$\mathcal{R}(X) = \mathbb{E}_a\left[- a^T X \cdot \frac{\exp(a^T X^\star)}{\exp(a^T X^\star) + 1} + \log\left(\exp(a^T X) + 1\right)\right].$$

Why? Let $a^T X \sim N(0, X^T K X)$

$$\mathbb{E}_a[\log\left(\exp(a^T X) + 1\right)] = \mathbb{E}_w\left[\log(\exp(\sqrt{X^T K X}\, w) + 1)\right], \quad w \sim N(0,1)$$

Dynamics of Covariance Matrix

Goal: Understand the impact of SGD noise

Lots of statistics can be represented by

$$\varphi(X) = g(W^T q(K) W) = g\left(\begin{bmatrix} X^T q(K) X & X^T q(K) X^\star \\ (X^\star)^T q(K) X & (X^\star)^T q(K) X^\star \end{bmatrix}\right),$$

where q is a polynomial, $K = \mathbb{E}[aa^T]$.

Dynamics of Covariance Matrix

Goal: Understand the impact of SGD noise

Lots of statistics can be represented by

$$\varphi(X) = g(W^T q(K) W) = g\left(\begin{bmatrix} X^T q(K) X & X^T q(K) X^\star \\ (X^\star)^T q(K) X & (X^\star)^T q(K) X^\star \end{bmatrix}\right),$$

where q is a polynomial, $K = \mathbb{E}\left[aa^T\right]$.

Take away: To understand dynamics of SGD amounts to understanding **low-dimensional covariance matrix**

for any polynomial q, $\quad W^T q(K) W \quad \Rightarrow \quad S(W; z) \stackrel{\text{def}}{=} W^T R(z; K) W$

where $R(z; K) = (z - K)^{-1}$ resolvent of K, $z \in \mathbb{C} \setminus (\text{spectrum } K)$.

Dynamics of Covariance Matrix

Goal: Understand the impact of SGD noise

Lots of statistics can be represented by

$$\varphi(X) = g(W^T q(K) W) = g\left(\begin{bmatrix} X^T q(K) X & X^T q(K) X^\star \\ (X^\star)^T q(K) X & (X^\star)^T q(K) X^\star \end{bmatrix}\right),$$

where q is a polynomial, $K = \mathbb{E}[aa^T]$.

Take away: To understand dynamics of SGD amounts to understanding **low-dimensional covariance matrix**

for any polynomial q, $\quad W^T q(K) W \quad \Rightarrow \quad S(W; z) \overset{\text{def}}{=} W^T R(z; K) W$

where $R(z; K) = (z - K)^{-1}$ resolvent of K, $z \in \mathbb{C} \setminus (\text{spectrum } K)$.

Remark:

$$W^T q(K) W = \frac{1}{2\pi i} \oint_\Gamma q(z) W^T R(z; K) W \, \mathrm{d}z$$

Main Result

Continuous time scale: iterates of SGD $k = td$, where $t \in \mathbb{R}$

$S(W; z) \stackrel{\text{def}}{=} W^T R(z; K) W$, where $R(z; K)$ is resolvent of K, $W = [X|X^*]$

Main Result

Continuous time scale: iterates of SGD $k = td$, where $t \in \mathbb{R}$

$S(W; z) \stackrel{\text{def}}{=} W^T R(z; K) W$, where $R(z; K)$ is resolvent of K, $W = [X|X^*]$

Theorem: High dimensional concentration of SGD

(**C.P.**-E.Collins-Woodfin-I. Seroussi-E. Paquette)

For a fixed $T > 0$,

$$\Pr\Big(\sup_{0 \le t \le T} \| S(\underbrace{W_{\lfloor td \rfloor}}_{\text{SGD}}; z) - \underbrace{\mathscr{S}(t; z)}_{\substack{\text{deterministic} \\ \text{function}}} \| > d^{-C} \Big) \le d^{-D}$$

and, for any $\varphi(X) = g(W^T q(K) W)$,

$$\Pr\Big(\sup_{0 \le t \le T} |\varphi(\underbrace{X_{\lfloor td \rfloor}}_{\text{SGD}}) - \underbrace{\phi(t)}_{\substack{\text{deterministic} \\ \text{function}}} | > d^{-C} \Big) \le d^{-D}$$

Main Result

Continuous time scale: iterates of SGD $k = td$, where $t \in \mathbb{R}$

$S(W; z) \stackrel{\text{def}}{=} W^T R(z; K) W$, where $R(z; K)$ is resolvent of K, $W = [X|X^*]$

Theorem: High dimensional concentration of SGD

(**C.P.**-E.Collins-Woodfin-I. Seroussi-E. Paquette)

For a fixed $T > 0$,

$$\Pr\Big(\sup_{0 \le t \le T} \| S(\underbrace{W_{\lfloor td \rfloor}}_{\text{SGD}}; z) - \underbrace{\mathscr{S}(t; z)}_{\substack{\text{deterministic} \\ \text{function}}} \| > d^{-C} \Big) \le d^{-D}$$

and, for any $\varphi(X) = g(W^T q(K) W)$,

$$\Pr\Big(\sup_{0 \le t \le T} |\varphi(\underbrace{X_{\lfloor td \rfloor}}_{\text{SGD}}) - \underbrace{\phi(t)}_{\substack{\text{deterministic} \\ \text{function}}} | > d^{-C} \Big) \le d^{-D}$$

✓ **Deterministic** function $\mathscr{S}$ defined by ODE **(see later)**

✓ Statistic

$$\phi(t) = g\left(\oint_\Gamma q(z) \mathscr{S}(t; z) \, \mathrm{d}z \right)^*$$

* ϕ will also satisfy an ODE.

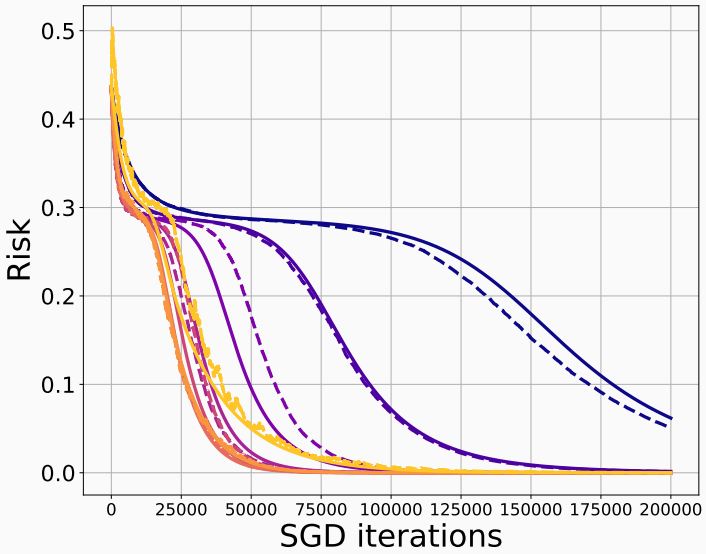
Example: Phase Retrieval

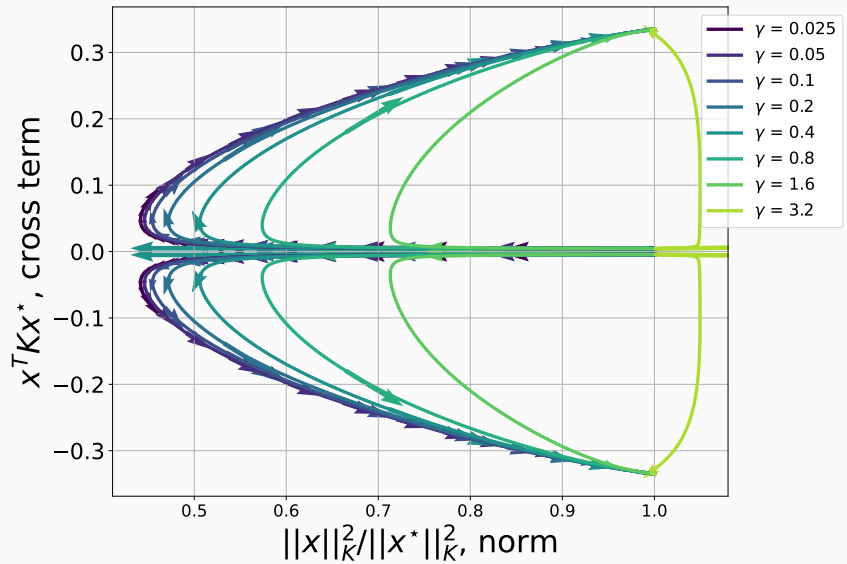
Problem: Find $\boldsymbol{x} \in \mathbb{R}^d$ such that

$$(a^T X)^2 \approx (a^T X^\star)^2, \quad a \in \mathbb{R}^d \quad \text{and} \quad X^\star \in \mathbb{R}^d \text{ true signal}$$

Optimization formulation:

$$\min_{X \in \mathbb{R}^d} \left\{ \mathcal{R}(X) \overset{\text{def}}{=} \mathbb{E}_a[((a^T X)^2 - (a^T X^\star)^2)^2] \right\}$$





Exact Dynamics Idea: Diffusion Approximation

Time scale: k iterates of SGD $= td$, where $t \in \mathbb{R}$

Homogenized SGD (**C.P.**-E.Collins-Woodfin-I. Seroussi-E. Paquette)

$$\mathrm{d}\mathscr{X}_t = -\gamma_t \nabla_X \mathcal{R}(\mathscr{X}_t)\,\mathrm{d}t + \gamma_t \langle \sqrt{K/d} \otimes \sqrt{\mathbb{E}_{a,\epsilon}[\nabla_x f(a^T X; a^T X^\star, \epsilon)^{\otimes 2}]}, \mathrm{d}B_t \rangle_{\mathcal{A}\otimes\mathcal{O}}$$

$\mathscr{X}_0 = X_0$ and $(B_t : t \geq 0)$ is a d-dimen. standard Brownian motion

- New diffusion process (Li et al., Mandt et al.)
- Continuous time is made by $d \to \infty$ instead of stepsize $\gamma \to 0$, $t = 1$ means n SGD updates
- ~~$\mathscr{X}_t$ mean/covariance same as $X^{\mathrm{SGD}}_{\lfloor td \rfloor}$~~, Goal: $\varphi(\mathscr{X}_t) \approx \varphi(X^{\mathrm{SGD}}_{\lfloor td \rfloor})$

Exact Dynamics Idea: Diffusion Approximation

Time scale: k iterates of SGD $= td$, where $t \in \mathbb{R}$

Homogenized SGD (**C.P.**-E.Collins-Woodfin-I. Seroussi-E. Paquette)

$$\mathrm{d}\mathscr{X}_t = -\gamma_t \nabla_X \mathcal{R}(\mathscr{X}_t)\,\mathrm{d}t + \gamma_t \langle \sqrt{K/d} \otimes \sqrt{\mathbb{E}_{a,\epsilon}[\nabla_x f(a^T X; a^T X^\star, \epsilon)^{\otimes 2}]}, \mathrm{d}B_t \rangle_{\mathcal{A}\otimes\mathcal{O}}$$

$\mathscr{X}_0 = X_0$ and $(B_t : t \geq 0)$ is a d-dimen. standard Brownian motion

- New diffusion process (Li et al., Mandt et al.)
- Continuous time is made by $d \to \infty$ instead of stepsize $\gamma \to 0$, $t = 1$ means n SGD updates
- ~~$\mathscr{X}_t$ mean/covariance same as $X^{\mathrm{SGD}}_{\lfloor td \rfloor}$~~, Goal: $\varphi(\mathscr{X}_t) \approx \varphi(X^{\mathrm{SGD}}_{\lfloor td \rfloor})$

Theorem: High dimensional equivalence of SGD

(**C.P.**-E.Collins-Woodfin-I. Seroussi-E. Paquette)

For any $\varphi(X) = g(W^T q(K) W)$,

$$\Pr\Big(\sup_{0 \leq t \leq T} |\varphi(X_{\lfloor td \rfloor}) - \varphi(\underbrace{\mathscr{X}_t}_{\text{diffusion}})| > d^{-C} \Big) \leq d^{-D}$$

Deterministic equivalent of φ

Assumption: Statistic $\varphi(X) = g(W^T q(K) W)$

Homogenized SGD

$$\mathrm{d}\mathscr{X}_t = -\gamma_t \nabla \mathcal{R}(\mathscr{X}_t)\,\mathrm{d}t + \gamma_t \langle \sqrt{K/d} \otimes \sqrt{\mathbb{E}_{a,\epsilon}[\nabla f(a^T \mathscr{X}_t)^{\otimes 2}]}, \mathrm{d}B_t \rangle_{\mathcal{A}\otimes\mathcal{O}}$$

Deterministic equivalent of φ

Assumption: Statistic $\varphi(X) = g(W^T q(K) W)$

Homogenized SGD

$$\mathrm{d}\mathscr{X}_t = -\gamma_t \nabla \mathcal{R}(\mathscr{X}_t)\,\mathrm{d}t + \gamma_t \langle \sqrt{K/d} \otimes \sqrt{\mathbb{E}_{a,\epsilon}[\nabla f(a^T \mathscr{X}_t)^{\otimes 2}]}, \mathrm{d}B_t \rangle_{\mathcal{A}\otimes\mathcal{O}}$$

Intuition: Apply Itô Calculus

$$\mathrm{d}\varphi(\mathscr{X}_t) = -\gamma_t \langle \nabla\varphi(\mathscr{X}_t), \nabla \mathcal{R}(\mathscr{X}_t)\rangle\,\mathrm{d}t \quad \textbf{(grad. flow)}$$

$$+ \frac{\gamma_t^2}{2d} \langle \nabla^2 \varphi(\mathscr{X}_t), K \otimes \mathbb{E}_{a,\epsilon}[\nabla f(a^T \mathscr{X}_t)^{\otimes 2}]\rangle\,\mathrm{d}t \quad \textbf{(SGD noise)}$$

$$+ \text{incremental martingale}$$

$$\stackrel{\text{def}}{=} \mathcal{F}(S(\mathscr{W}_t; z)) + \text{incremental martingale}$$

All the quantities are functions

$$S(\mathscr{W}_t; z) \stackrel{\text{def}}{=} \mathscr{W}_t^T R(z; K) \mathscr{W}_t, \quad \mathscr{W}_t = [\mathscr{X}_t | X^\star].$$

Deterministic equivalent

Idea: Set $\varphi(\mathscr{X}_t) = S(\mathscr{W}_t; z)$ in Ito (and drop martingale)

$$\mathrm{d}S(\mathscr{W}_t; z) = \mathcal{G}(S(\mathscr{W}_t; z)) + \cancel{\text{incremental martingale}}$$

$\Rightarrow$ ODE for deterministic equivalent $\mathscr{S}(t, z)$ for $S(\mathscr{W}_t, z)$

$$\mathrm{d}\mathscr{S}(t, z) = \mathcal{G}(\mathscr{S}(t, z))$$

Deterministic equivalent

Idea: Set $\varphi(\mathscr{X}_t) = S(\mathscr{W}_t; z)$ in Ito (and drop martingale)

$$\mathrm{d}S(\mathscr{W}_t; z) = \mathcal{G}(S(\mathscr{W}_t; z)) + \cancel{\text{incremental martingale}}$$

$\Rightarrow$ ODE for deterministic equivalent $\mathscr{S}(t, z)$ for $S(\mathscr{W}_t, z)$

$$\mathrm{d}\mathscr{S}(t, z) = \mathcal{G}(\mathscr{S}(t, z))$$

ODE for Deterministic Equivalent:

$$\begin{aligned} \mathrm{d}S(\mathscr{W}_t; z) = \mathcal{G}(S(\mathscr{W}_t; z)) \quad &\Rightarrow \quad \mathrm{d}\mathscr{S}(t; z) = \mathcal{G}(\mathscr{S}(t; z)) \quad \Leftarrow \text{ solve numerically} \\ \mathrm{d}\varphi(\mathscr{X}_t) = \mathcal{F}(S(\mathscr{W}_t; z)) \quad &\Rightarrow \quad \mathrm{d}\phi(t) = \mathcal{F}(\mathscr{S}(t; z)) \end{aligned}$$

where ϕ is the deterministic equivalent of $\varphi(X)$

Main Result

Theorem: High dimensional equivalence of SGD

(**C.P.**-E.Collins-Woodfin-I. Seroussi-E. Paquette)

For any $\varphi(X) = g(W^T q(K)W)$,

$$\Pr\Big(\sup_{0\le t\le T} |\varphi(X_{\lfloor td \rfloor}) - \varphi(\underbrace{\mathscr{X}_t}_{\text{diffusion}})| > d^{-C}\Big) \le d^{-D}$$

and

$$\Pr\Big(\sup_{0\le t\le T} |\varphi(\underbrace{\mathscr{X}_t}_{\text{diffusion}}) - \phi(t)| > d^{-C}\Big) \le d^{-D}$$

Why Interesting?

Optimization Question

What choice of learning rate ensures distance to optimality decreases at each iteration of SGD?

$\rightarrow$ Can't do this with SGD because of the stochasity in the gradients
$\rightarrow$ Can ask on the **deterministic equivalent** of the distance to optimality, $\|X - X^\star\|^2$,

What stepsize is needed for $\|X - X^\star\|^2$ to be a decreasing function?

Intuition-Critical Threshold

Deterministic equivalent of $\|X - X^\star\|^2$:

$$\sup_{0\le t\le T} \left| \|X_{\lfloor td \rfloor} - X^\star\|^2 - \mathscr{D}^2(t) \right| \le d^{-\varepsilon}$$

where

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^2 = -2\gamma_t A(\mathscr{S}) + \frac{\gamma_t^2}{d}\,\mathrm{tr}(K) I(\mathscr{S}), \quad \begin{cases} A(\mathscr{S}) = \mathbb{E}_{a,\epsilon}[\langle x - x^\star, \nabla f(x; x^\star)\rangle], \\ I(\mathscr{S}) = \mathbb{E}_{a,\epsilon}[\|\nabla f(x; x^\star)\|^2], \\ \text{where } (x \oplus x^\star) \sim N(0, W^T K W). \end{cases}$$

Intuition-Critical Threshold

Deterministic equivalent of $\|X - X^\star\|^2$:

$$\sup_{0 \le t \le T} \left| \|X_{\lfloor td \rfloor} - X^\star\|^2 - \mathscr{D}^2(t) \right| \le d^{-\varepsilon}$$

where

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^2 = -2\gamma_t A(\mathscr{S}) + \frac{\gamma_t^2}{d}\operatorname{tr}(K) I(\mathscr{S}), \quad \begin{cases} A(\mathscr{S}) = \mathbb{E}_{a,\epsilon}[\langle x - x^\star, \nabla f(x; x^\star)\rangle], \\ I(\mathscr{S}) = \mathbb{E}_{a,\epsilon}[\|\nabla f(x; x^\star)\|^2], \\ \text{where } (x \oplus x^\star) \sim N(0, W^T K W). \end{cases}$$

Critical learning rate $(\mathrm{d}\mathscr{D}^2 < 0)$

$$\gamma_t^{\text{critical}} = \frac{2A(\mathscr{S}(t;z))}{\frac{\operatorname{tr}(K)}{d} I(\mathscr{S}(t;z))}$$

Intuition-Critical Threshold

Deterministic equivalent of $\|X - X^\star\|^2$:

$$\sup_{0 \le t \le T} \left| \|X_{\lfloor td \rfloor} - X^\star\|^2 - \mathscr{D}^2(t) \right| \le d^{-\varepsilon}$$

where

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^2 = -2\gamma_t A(\mathscr{S}) + \frac{\gamma_t^2}{d}\operatorname{tr}(K) I(\mathscr{S}), \quad \begin{cases} A(\mathscr{S}) = \mathbb{E}_{a,\epsilon}[\langle x - x^\star, \nabla f(x; x^\star)\rangle], \\ I(\mathscr{S}) = \mathbb{E}_{a,\epsilon}[\|\nabla f(x; x^\star)\|^2], \\ \text{where } (x \oplus x^\star) \sim N(0, W^T K W). \end{cases}$$

Critical learning rate $(\mathrm{d}\mathscr{D}^2 < 0)$

$$\gamma_t^{\text{critical}} = \frac{2A(\mathscr{S}(t;z))}{\frac{\operatorname{tr}(K)}{d} I(\mathscr{S}(t;z))} \ge \frac{2q}{\frac{\operatorname{tr}(K)}{d}}, \quad \text{where } \frac{A(\mathscr{S}(t;z))}{I(\mathscr{S}(t;z))} \ge q$$

- Functions $A(\mathscr{S})$ and $I(\mathscr{S})$ – don't carry K or d
- Lower bound A and I based on convexity/smoothness assumptions of f
- Critical stepsize depends on average eigenvalue of K

Convergence rate

Theorem: Convergence of strongly convex
(**C.P.**-E.Collins-Woodfin-I. Seroussi-E. Paquette)

Suppose f is $\hat{\mu}$-strongly convex and $\hat{L}$-Lipschitz gradients. for some $0 < \zeta < 1$, then for all $t \geq 0$

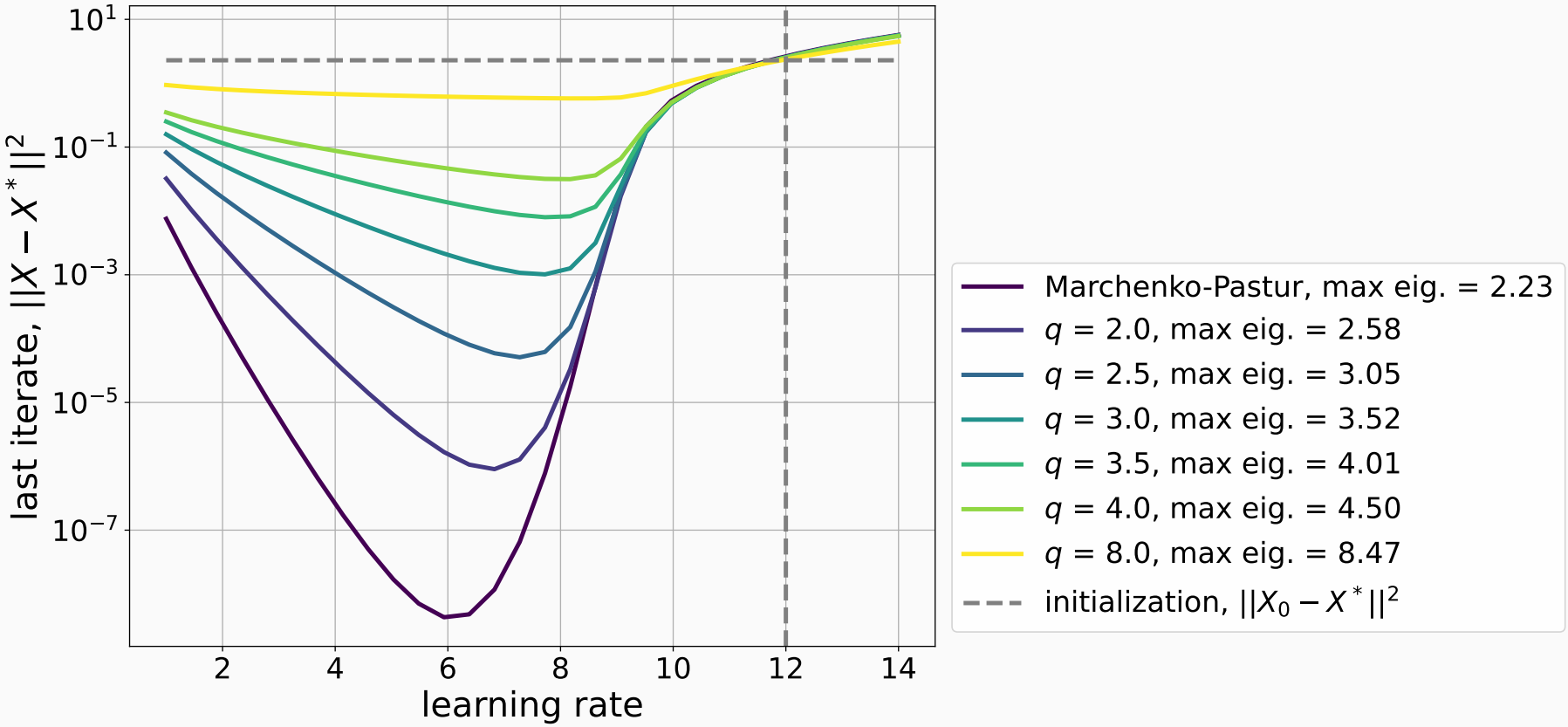
$$\mathscr{D}^2(t) \leq e^{-at}\mathscr{D}^2(0),$$

where $\gamma < \gamma^{\text{critical}}$

$$\textbf{(convergence rate)} \quad a = \frac{\hat{\mu}^2}{\hat{L}^2} \cdot \frac{\lambda_{\min}(K)}{\frac{1}{d}\operatorname{tr}(K)}$$

Logistic regression

$$\gamma_t^{\text{critical}} = \frac{2A(\mathscr{S}(t;z))}{\frac{\text{tr}(K)}{d} I(\mathscr{S}(t;z))}$$



Caption: Covariance matrix $K = \text{diag}(\sigma_i^{2q} : i = 1, \ldots, 1000)$, $\text{tr}(K)/d = 1$.

Least Squares

Focus on the **Least Squares** Problem with extensions
(e.g., multi-pass)

Our framework

$$\min_{X\in\mathbb{R}^d} \frac{1}{2}\|AX-b\|^2 = \min_{X\in\mathbb{R}^d} \left\{ \mathcal{L}(X) \overset{\text{def}}{=} \sum_{i=1}^{n} \underbrace{\frac{1}{2}(a_i^T X - b_i)^2}_{f_i(x)} \right\},$$

with *random* $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$ *random* vector

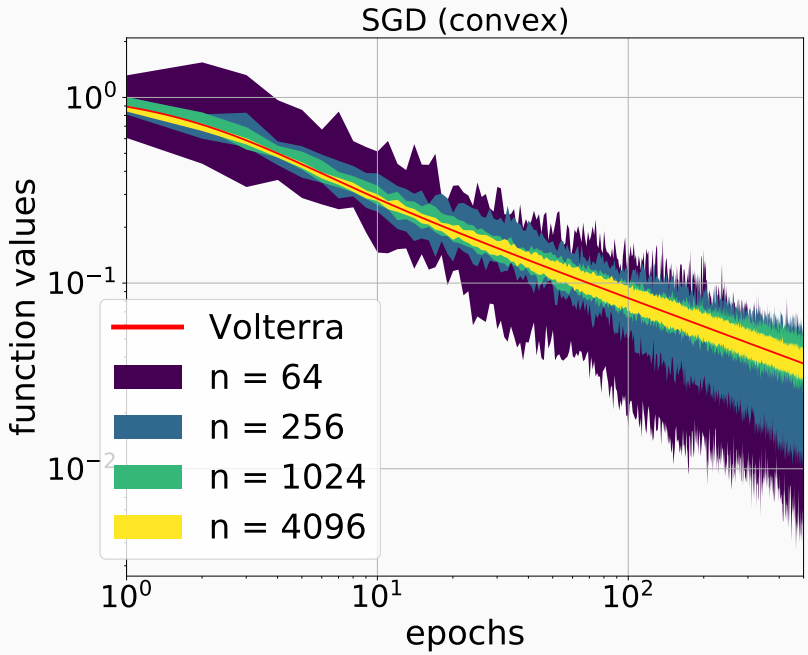
Multi-pass SGD Select index i_k $X_{k+1} = X_k - \gamma_k \nabla f_{i_k}(X_k)$

Our framework

$$\min_{X\in\mathbb{R}^d} \frac{1}{2}\|AX-b\|^2 = \min_{X\in\mathbb{R}^d} \left\{ \mathcal{L}(X) \overset{\text{def}}{=} \sum_{i=1}^{n} \underbrace{\frac{1}{2}(a_i^T X - b_i)^2}_{f_i(x)} \right\},$$

with *random* $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$ *random* vector

Multi-pass SGD Select index i_k $\quad X_{k+1} = X_k - \gamma_k \nabla f_{i_k}(X_k)$



For **large models**, as $\frac{d}{n} \to r$,

- $\mathcal{L}(X_k) \xrightarrow{\Pr} \mathscr{L}(t)$ (smooth function)
- Determined by the spectrum of the Hessian
- **Homogenized SGD** (**C.P.**-E. Paquette, NeurIPS '21 & Mori '21)

$$\mathrm{d}\mathscr{X}_t = -\gamma(t)\nabla\mathcal{L}(\mathscr{X}_t)\,\mathrm{d}t + \gamma_t\sqrt{\tfrac{2}{n}\mathcal{L}(\mathscr{X}_t)A^TA}\,\mathrm{d}B_t$$

Remove Gaussian Assumption...

Hessian of least squares: $\boldsymbol{H} = \boldsymbol{A}^T \boldsymbol{A}$

Assumptions on data matrix (Bai & Silverstein '10, Benigni & Peche '19)

1. model size (d) and # of samples (n) polynomially related

$$d^\alpha \leq n \leq d^{1/\alpha} \quad \text{for some } \alpha \in (0,1)$$

2. Mild assumptions on eigenvalues $\lambda_{\max}$ and $\lambda_{\min}$ of $\boldsymbol{H}$

3. De-localization of eigenvectors of AA^T: eigenvectors are not aligned with the unit vectors

   e.g., if $A_{i,j} \sim N(0,1)$, then eigenvectors of $H \sim \text{Unif}(\mathbb{S}^{d-1})$

   - Isotropic features. Entries of $A \sim N(0,1)$
   - Sample covariance matrices. independent samples w/ covariance between features
   - Random features. $A = \sigma(ZV)$ where σ is an activation function

Real world predictions: MNIST

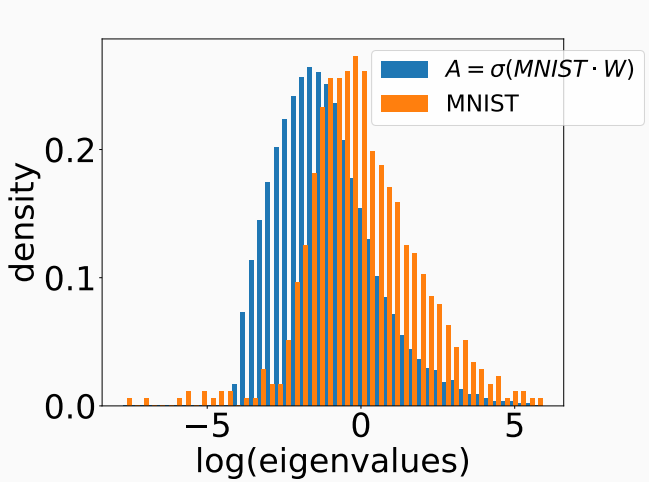
Expected Risk:

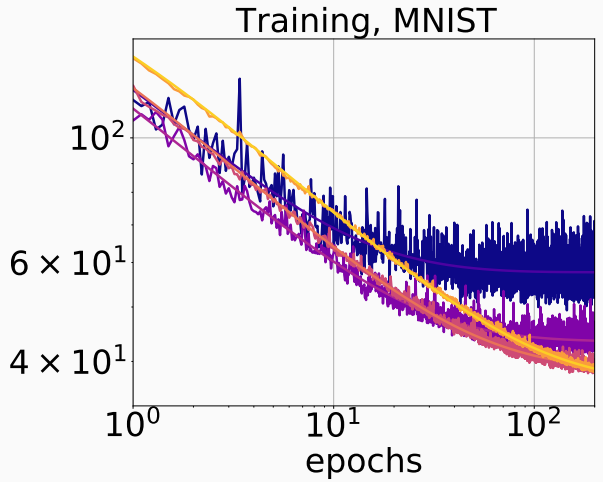
$$\mathcal{R}(X_k) = \frac{1}{2}\mathbb{E}[(b - X_k^T a)^2 | X_k] \quad \text{where} \quad (a, b) \sim \mathcal{D}, \quad X_k = \text{ SGD iterate on } \mathcal{L}$$

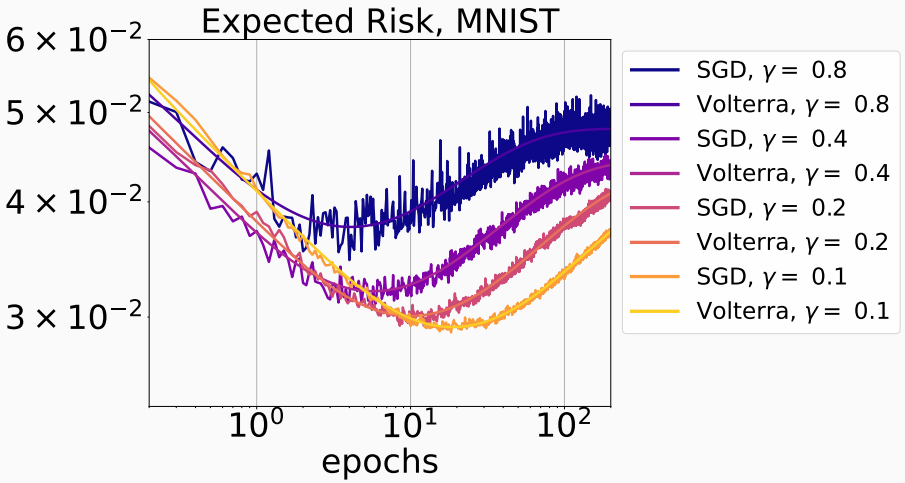
Real world predictions: MNIST

Expected Risk:

$$\mathcal{R}(X_k) = \frac{1}{2}\mathbb{E}[(b - X_k^T a)^2 | X_k] \quad \text{where} \quad (a, b) \sim \mathcal{D}, \quad X_k = \text{ SGD iterate on } \mathcal{L}$$







Phase transition & Asymptotics

Phase transition stepsize

$$\gamma_* = \frac{1}{\frac{d}{2n}\int_0^\infty \frac{x^2}{x-\lambda_{\min}(A^TA)}\,\mathrm{d}\mu(x)}$$

Theorem

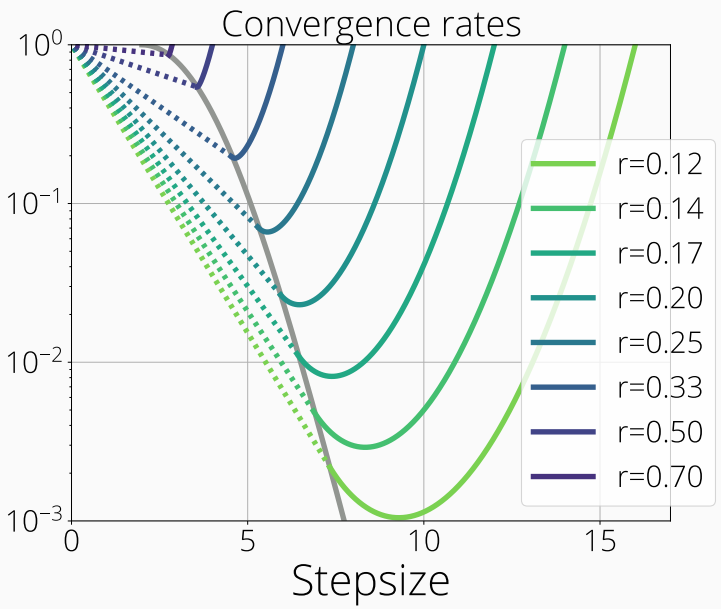
(**C.P.**-Lee-E. Paquette-Pedregosa, COLT '21)

For small $\gamma < \gamma_*$,

$$\mathscr{L}(t) - \mathscr{L}(\infty) \sim \frac{1}{t^\alpha} e^{-2\gamma t \lambda_{\min}}.$$

For large $\gamma > \gamma_*$, $\exists$ non-linear $\lambda^*(\gamma)$

and $$\mathscr{L}(t) - \mathscr{L}(\infty) \sim \frac{1}{\gamma} e^{-2\gamma t \lambda^*(\gamma)}.$$



Large batch: SGD+M

$$\min_{X\in\mathbb{R}^d} \frac{1}{2}\|AX-b\|^2 = \min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ \mathcal{L}(X) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \underbrace{\frac{1}{2}(a_i^T X - b_i)^2}_{f_i(\boldsymbol{x})} \right\},$$

with *batch* $B \subset [n]$, *batch fraction* $\zeta \stackrel{\text{def}}{=} \dfrac{|B|}{n}$

Multi-pass SGD
+ momentum (SGD+M)

$$Y_k = \Delta \cdot Y_{k-1} + \gamma \cdot \zeta \sum_{i_k \in B} \nabla f_{i_k}(X_k)$$
$$X_{k+1} = X_k - Y_k$$

Large batch: SGD+M

$$\min_{X\in\mathbb{R}^d} \frac{1}{2}\|AX-b\|^2 = \min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ \mathcal{L}(X) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \underbrace{\frac{1}{2}(a_i^T X - b_i)^2}_{f_i(\boldsymbol{x})} \right\},$$

with *batch* $B \subset [n]$, *batch fraction* $\zeta \stackrel{\text{def}}{=} \dfrac{|B|}{n}$

Multi-pass SGD
+ momentum (SGD+M)

$$Y_k = \Delta \cdot Y_{k-1} + \gamma \cdot \zeta \sum_{i_k \in B} \nabla f_{i_k}(X_k)$$
$$X_{k+1} = X_k - Y_k$$

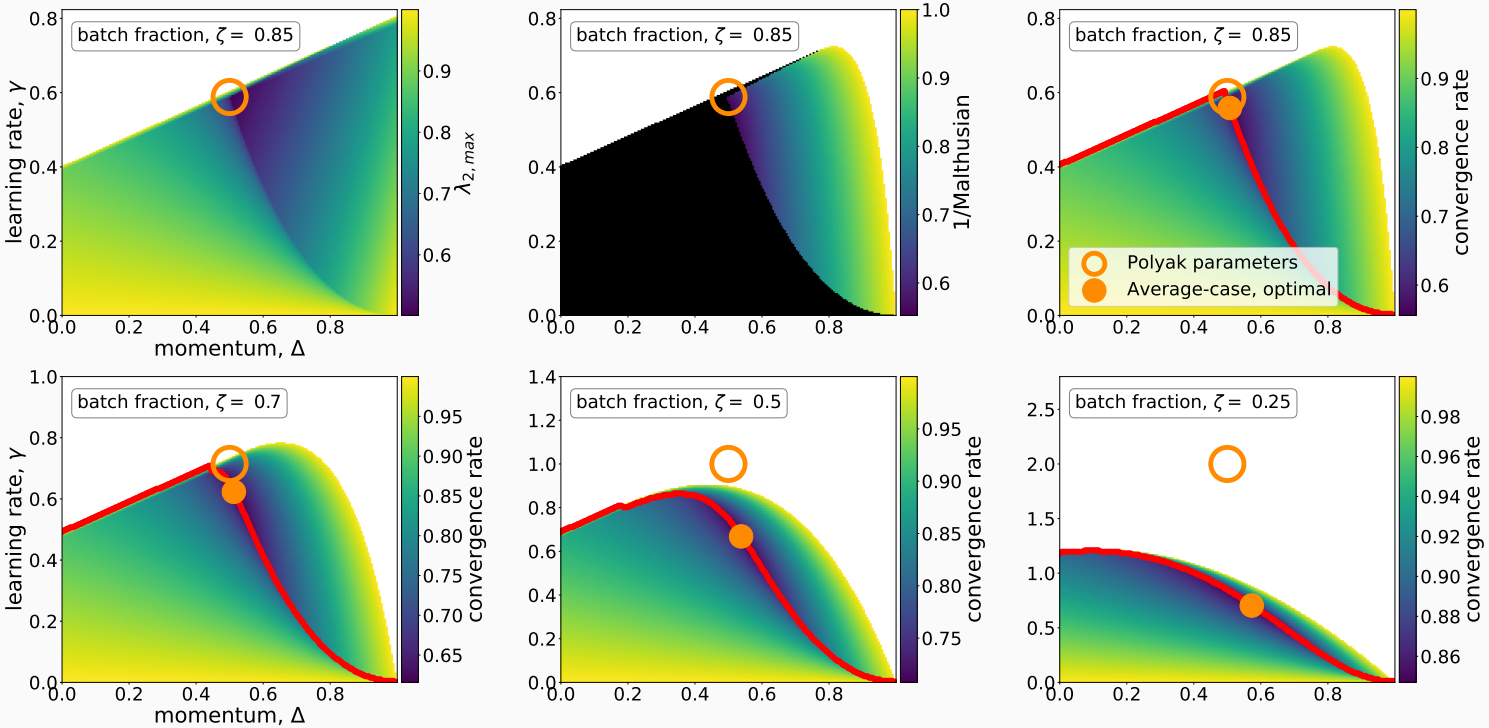
- Concentration of $\mathcal{L}(X_k) \to \mathscr{L}(k)$ deterministic, discrete **not** continuous

(**C.P.**-Lee-Cheng-E. Paquette)

Convergence of SGD+M with batches

Theorem (**C.P.**-Lee-Cheng-E. Paquette)

$$\lim_{k\to\infty} (\mathscr{L}(k) - \mathscr{L}(\infty))^{1/k} = \max\left\{ \underbrace{\Lambda}_{\text{GD+M}}, \underbrace{\Xi^{-1}}_{\text{noise}} \right\}$$



Nearly-optimal parameters

Condition numbers

$$\text{(average)} \quad \bar{\kappa} \stackrel{\text{def}}{=} \frac{\frac{1}{n}\operatorname{tr}(A^T A)}{\lambda_{\min}(A^T A)} \quad < \quad \frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)} \stackrel{\text{def}}{=} \kappa \quad \text{(classic)}$$

Nearly-optimal parameters

Condition numbers

$$\text{(average)} \quad \bar{\kappa} \stackrel{\text{def}}{=} \frac{\frac{1}{n}\operatorname{tr}(A^T A)}{\lambda_{\min}(A^T A)} \quad < \quad \frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)} \stackrel{\text{def}}{=} \kappa \quad \text{(classic)}$$

Theorem (**C.P.**-Lee-Cheng-E. Paquette)
Suppose (near optimal parameters)

$$\gamma = \frac{(1-\sqrt{\Delta})^2}{\zeta \lambda_{\min}(A^T A)}, \qquad \Delta = \max\left\{ \underbrace{\left(1 - \frac{\zeta}{(1-\zeta)\bar{\kappa}}\right)^2}_{\text{rate of SGD}}, \underbrace{\left(1 - \frac{1}{\sqrt{\kappa}}\right)^2}_{\text{rate of GD+M}} \right\}$$

Then

$$\lim_{k\to\infty} (\mathscr{L}(k) - \mathscr{L}(\infty))^{1/k} = \Delta$$

Large vs Small batch: Convergence

$$\lim_{k\to\infty} (\mathscr{L}(k) - \mathscr{L}(\infty))^{1/k} = \max\left\{ \left(1 - \frac{\zeta}{(1-\zeta)\bar{\kappa}}\right)^2, \left(1 - \frac{1}{\sqrt{\kappa}}\right)^2 \right\}$$

(average) $\bar{\kappa} \overset{\text{def}}{=} \frac{\frac{1}{n}\operatorname{tr}(A^T A)}{\lambda_{\min}(A^T A)}$, *implicit conditioning ratio, ICR* $\overset{\text{def}}{=} \frac{\bar{\kappa}}{\sqrt{\kappa}} = \frac{\text{average}}{\sqrt{\text{classic}}}$.

Large vs Small batch: Convergence

$$\lim_{k\to\infty} (\mathscr{L}(k) - \mathscr{L}(\infty))^{1/k} = \max\left\{ \left(1 - \frac{\zeta}{(1-\zeta)\bar{\kappa}}\right)^2, \left(1 - \frac{1}{\sqrt{\kappa}}\right)^2 \right\}$$

(average) $\bar{\kappa} \overset{\text{def}}{=} \frac{\frac{1}{n}\operatorname{tr}(A^T A)}{\lambda_{\min}(A^T A)}$, *implicit conditioning ratio, ICR* $\overset{\text{def}}{=} \frac{\bar{\kappa}}{\sqrt{\kappa}} = \frac{\text{average}}{\sqrt{\text{classic}}}$.

Phase transition

(C.P.-Lee-Cheng-E. Paquette)

- **Large batch**: $\zeta \geq$ ICR

  SGD+M linearly at rate $\mathcal{O}(1/\sqrt{\kappa})$

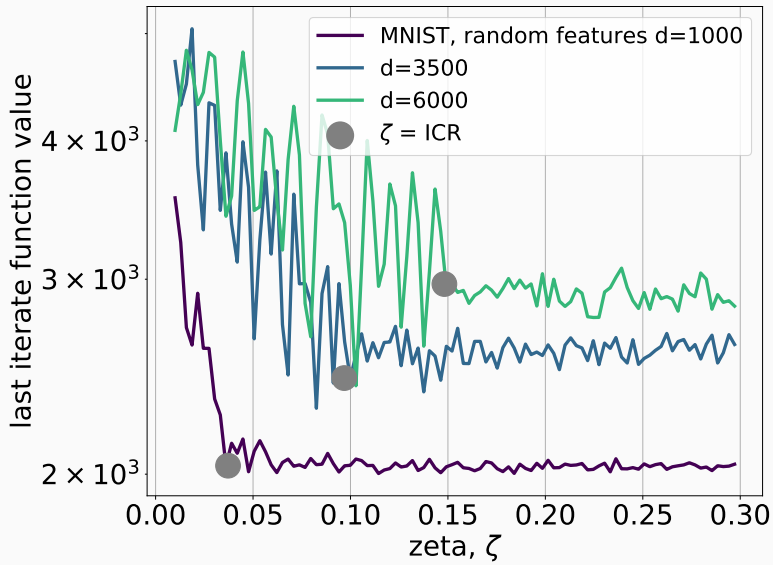
  and SGD+M accelerates

- **Small batch**: $\zeta \leq$ ICR

  SGD+M linearly at rate $\mathcal{O}(\zeta/\bar{\kappa})$

  SGD+M $\Leftrightarrow$ SGD

Related work: Bollapragada-Chen-Ward, '23



Saturating batch fraction – after which increasing the batch fraction does not improve convergence.

Thank you!

C. Paquette, E. Paquette, B. Adlam, J. Pennington. *Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties*, arxiv.org/pdf/2205.07069.pdf

K. Lee, A. Cheng, E. Paquette, C. Paquette. *Trajectory of Mini-Batch Momentum: Batch Size Saturation and Convergence in High Dimensions*, arxiv.org/pdf/2206.01029.pdf (NeurIPS 2022)

C. Paquette, K. Lee, F. Pedregosa, E. Paquette. *SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality*, arxiv.org/pdf/2102.04396.pdf (COLT 2021)

C. Paquette, E. Paquette. *Hitting the High-Dimensional Notes: An ODE for SGD Learning dynamics on GLMs and multi-index models*, arxiv.org/pdf/2308.08977.pdf