High-Dimensional Dynamics of SGD on Structured Data

Inbar Seroussi

Joint work with Elizabeth Colins-Woodfin (University of Oregon),

Courtney Paquette (McGill University) and Elliot Paquette (McGill University)

Statistical Physics and Machine Learning: Moving forward, Cargèse 2025, 6.8.2025

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Goal and motivation for today

- Understand the **high-dimensional** dynamics of online **SGD** on **real** data

- What is the role of structure/high dimensionality in the dynamics?

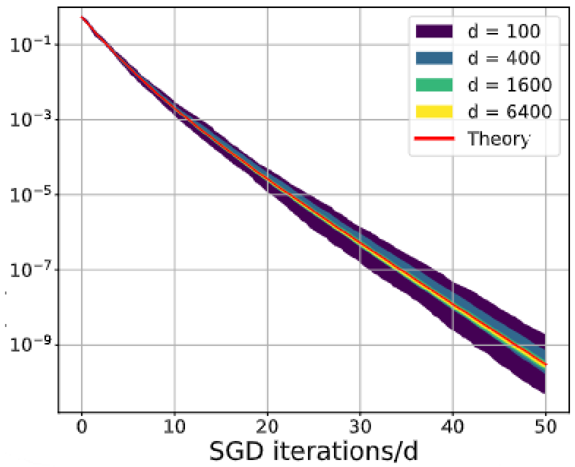
Given data - $\{(\boldsymbol{a}_i; \boldsymbol{y}_i)\}_{i=1}^{n}$, $X \in \mathbb{R}^p$ is a set of learnable parameters with SGD

$$\min_{X\in\mathbb{R}^p} \{L(X) = \mathbb{E}_{(\boldsymbol{a},\boldsymbol{y})}[f(X; \boldsymbol{a}, \boldsymbol{y})]\} \quad \overset{?}{\Longleftrightarrow} \quad \min_{X\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} f(X; \boldsymbol{a}_i; \boldsymbol{y}_i)$$

- Exact prediction of the dynamics

- Condition on feature learning, scaling, stability, classification capabilities

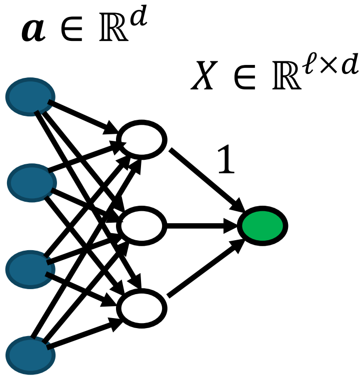
$L(X)$



SGD dynamics - One class

Deterministic equivalence and more

Data and estimator setup



"Neural Network"

$$g(X\boldsymbol{a}) = \sum_{i=1}^{\ell} \sigma\,(X_i\boldsymbol{a})$$

<u>Target (Teacher) model</u>:

$$\boldsymbol{y}_i = \phi(X^\star \boldsymbol{a}_i; \boldsymbol{\varepsilon}_i), \qquad \boldsymbol{\varepsilon}_i \text{ - i.i.d. noise with bounded variance}$$

with a true matrix $X^\star \in \mathbb{R}^{\ell^\star \times d}$ and $\phi: \mathbb{R}^{\ell^\star} \to \mathbb{R}^m$, with $\ell^\star = O_d(1)$

$\boldsymbol{a}_i \sim \mathcal{N}(0, K)$, with $K \in \mathbb{R}^{d \times d}$, $\|K\|_{\text{op}}$ bounded

$\phi(X^\star \boldsymbol{a}_i; \boldsymbol{\varepsilon}_i)$ → $\{(\boldsymbol{a}_i, \boldsymbol{y}_i)\}_{i=1}^n$

<u>Estimator (Student) model</u> Given $\{(\boldsymbol{a}_i, \boldsymbol{y}_i)\}_{i=1}^n$, choose a function $g: \mathbb{R}^\ell \to \mathbb{R}^m$, **estimate** the matrix $X \in \mathbb{R}^{\ell \times d}$ with $\ell = O_d(1)$

$\{(\boldsymbol{a}_i, \boldsymbol{y}_i)\}_{i=1}^n$ →  → $g(X\boldsymbol{a}_i)$

<u>Optimization problem</u>: $L(X) = \mathbb{E}[\text{dist}(g(X\boldsymbol{a}); \boldsymbol{y}] = \mathbb{E}_{(\boldsymbol{a},\boldsymbol{\varepsilon})}[f(X\boldsymbol{a}; X^\star \boldsymbol{a}, \boldsymbol{\varepsilon})]$

Stochastic Gradient Descent (SGD)

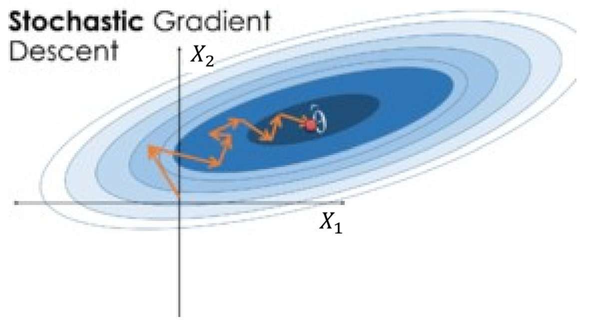
- One pass single batch (online learning) with fixed step size

$$X_{k+1} = X_k - \frac{\gamma_k}{d} \nabla_{X_k} f(X_k; \boldsymbol{a}_{k+1}, \boldsymbol{y}_{k+1})$$

Learning rate (step size)

- Initialization $X_0 \in \mathbb{R}^{\ell \times d}$ with a bounded norm
- High-dimensional limit, n, d large, $\frac{n}{d} \to T \in (0, \infty)$

Samples (iteration) parameters



One class - No structure ($K = I_d$)

- Recall the features $\boldsymbol{a}_i \sim \mathcal{N}(0, K)$, $K = I_d$ - corresponds to isotropic data – **no structure**
- Iterates of SGD:

$$X_{k+1} = X_k - \frac{\gamma}{d} \nabla_r f \otimes \boldsymbol{a}_{k+1}, \text{ where } \nabla_{\mathrm{r}} f = \nabla_{\boldsymbol{r}_k} f(\boldsymbol{r}_k; \boldsymbol{r}^\star, \boldsymbol{\varepsilon}_{k+1})$$

with $\boldsymbol{r}_k = X_k \boldsymbol{a}_{k+1}$, $\boldsymbol{r}^\star = X^\star \boldsymbol{a}_{k+1}$

- Order Parameters: "**Norm**" $\langle X_k, X_k \rangle = X_k^\top X_k$, and "**Overlap**" $\langle X^\star, X_k \rangle = (X^\star)^\top X_k$

<u>Theorem:</u> (E. Collins-Woodfin, C&E. Paquette, **SI**): Fix $T = \frac{n}{d} \in [0, \infty)$ and for some $\varepsilon > 0$ with overwhelming probability,

$$\sup_{0 \le t \le T} \left\| \mathcal{B}(t) - B_{\lfloor td \rfloor} \right\| \le d^{-\varepsilon}$$

where $B_k = \begin{bmatrix} \langle X_k, X_k \rangle & \langle X_k, X^\star \rangle \\ \langle X^\star, X_k \rangle & \langle X^\star, X^\star \rangle \end{bmatrix}$

↑

Time scale: k iterates of SGD $= td$

continues time: $d \to \infty$ instead of $\gamma \to 0$

Limiting ODEs - No structure ($K = I_d$)

Given the deterministic B-matrix:

$$\mathcal{B}(t) = \begin{bmatrix} \mathcal{B}_{xx}(t) & \mathcal{B}_{x\star}(t) \\ \mathcal{B}_{x\star}(t) & \langle X^\star, X^\star \rangle \end{bmatrix}$$

Fisher matrix:

$$I(\mathcal{B}(t)) = \mathbb{E}[\nabla_{\mathrm{r}} f^{\otimes 2}]$$

Gradient of the loss:

$$H = \begin{bmatrix} \nabla_{\mathcal{B}_{xx}} \mathcal{L} & 0 \\ \nabla_{\mathcal{B}_{x\star}} \mathcal{L} & 0 \end{bmatrix}$$

The limiting ODEs:

"Gradient" term "Noise" term

$$\frac{\mathrm{d}\mathcal{B}(t)}{\mathrm{d}t} = -\gamma\left(\mathcal{B}(t) H(\mathcal{B}(t)) + H(\mathcal{B}(t))^\top \mathcal{B}(t)\right) + \gamma^2 \begin{bmatrix} I(\mathcal{B}(t)) & 0 \\ 0 & 0 \end{bmatrix}$$

Order Parameters: "**Norm**" $B_{xx,k} = \langle X_k, X_k \rangle$, and "**Overlap**" $B_{x\star,k} = \langle X_k, X^\star \rangle$

Related literature - SGD in high dimension

<u>Isotropic data ($K = I_d$):</u>

- Two-layer neural net (Saad & Solla *Phys. Rev. E* '95, Riegler & Biehl *Physica A* '95…)
- Phase retrieval (Tan & Veryshynin *JMLR* '23, Mignacco et al. *NeurIPS* '20)
- Tensor PCA (Ben Arous et al. *NeurIPS* '22, Liang et al. *Inf. Inference* '23)
- Gaussian mixture models (Ben Arous et al. *ICLR* '24, 25')
- Generalized linear model (Gerbelot et al '22, Celentano et al. '21)
- Two-layer neural net (Goldt et al. *NeurIPS* '19)
- ….

<u>Structured data (general K):</u>

- Linear regression – Balasubramanian et al. '23, Wang et al. *J. Stat. Mech.* '19, Paquette et al. '22-25'
- Two-layer neural net – Yoshida et al. *NeuriPS* '19, Goldt et. al. *PRX* '20

Structural data (general K)– Resolvent trick

Issue: One **cannot** write an autonomous set of equations,

$$\frac{\mathrm{d}\mathcal{B}(t)}{\mathrm{d}t} = \mathcal{F}(\mathcal{B}(t))$$

Higher powers of K appears!

Terms of the form
$X^T K X, X^T K^2 X, \ldots$

Solution: Random matrix theory trick!

Resolvent: $R(z;K) = (K - zI_d)^{-1}$ for $z \in \mathbb{C}$

Some nice resolvent identities:

- $KR(z;K) = I_d + zR(z;K)$
- $R(z;K) = -\frac{1}{z}\left(I_d - \frac{K}{z}\right)^{-1} = \sum_{j=1}^{\infty}(K/z)^j$

This allows us to represent **any** polynomial of K!

$$p(K) = -\frac{1}{2\pi i}\oint_{\Gamma} p(z)R(z;K)\mathrm{d}z$$

For any contour $\Gamma \subset \mathbb{C}$ enclosing the eigenvalues of K.

Structural data (general K)

Order Parameters:

"Resolvent Norm" $\langle X_k, X_k \rangle_R = X_k R(z; K) X_k^\top$

"Resolvent Overlap" $\langle X_k, X^\star \rangle_R = X_k R(z; K) \, (X^\star)^\top$

- Define the S matrix of "Order Parameters" : $S_k(z) = \begin{bmatrix} \langle X_k, X_k \rangle_R & \langle X_k, X^\star \rangle_R \\ \langle X^\star, X_k \rangle_R & \langle X^\star, X^\star \rangle_R \end{bmatrix} = \begin{bmatrix} X_k \\ X^\star \end{bmatrix} R(z; K) [X_k^\top \quad (X^\star)^\top]$

Theorem (E. Collins-Woodfin, C&E. Paquette, **SI**): For any $T = \frac{n}{d} \in [0, \infty)$ and for some $\varepsilon > 0$ with

overwhelming probability

$$\sup_{0 \le t \le T} \left\| \mathcal{S}(t, z) - S_{\lfloor td \rfloor}(z) \right\| \le d^{-\varepsilon}.$$

This then allows us to derive a limiting ODEs:

$$\frac{\mathrm{d}\mathcal{S}(t, z)}{\mathrm{d}t} = \mathcal{F}(z, \mathcal{S}(t, z))$$

Explicit risk curves

A large class of functions (statistics):

$$\varphi(X) = h\big(\langle [X, X^\star]^{\otimes 2}, p(K)\rangle\big) \to h\Big(-\frac{1}{2\pi i}\oint_\Gamma p(z)\mathcal{S}(t,z)\mathrm{d}z\Big)$$

h is α pseudo-Lipchitz function, and p is a polynomial

$$\varphi(X_k) = \mathcal{L}(X_k) = \mathbb{E}_{(\boldsymbol{a},\boldsymbol{\varepsilon})} f(X\boldsymbol{a}, X^\star\boldsymbol{a}; \boldsymbol{\varepsilon})$$

By our theorem

$$\mathcal{L}(X_k) = h(\langle [X, X^\star]^{\otimes 2}, K\rangle) \to \mathcal{L}(t)$$

- Other functions: $\|X\|^2, \|X - X^\star\|^2$ ...

$$X_k\boldsymbol{a} \sim \mathcal{N}(0, \langle X_k^{\otimes 2}, K\rangle),$$
$$X^\star\boldsymbol{a} \sim \mathcal{N}\big(0, \langle (X^\star)^{\otimes 2}, K\rangle\big),$$
$$\mathbb{E}[\langle X^\star\boldsymbol{a}, X_k\boldsymbol{a}\rangle] = \langle X^\star \otimes X_k, K\rangle$$

Main result: Limiting process - Homogenized SGD

Homogenized SGD: The process $\mathcal{X}_t$ satisfies the following SDE: "Noise" term

$$\mathrm{d}\mathcal{X}_t = -\gamma \nabla \mathcal{L}(\mathcal{X}_t)\mathrm{d}t + \boxed{\frac{\gamma}{\sqrt{d}}\sqrt{\mathbb{E}[(\nabla_{\mathrm{r}} f)^{\otimes 2}] \otimes K}\,\mathrm{d}\mathcal{W}_t}$$

where $\mathcal{W}_t$ denotes d-dimensional Brownian motion

"Fisher matrix"

Time scale:
k iterates of SGD = td,

<u>Theorem</u> (E. Collins-Woodfin, C&E. Paquette, **SI**):

Fix $T = \frac{n}{d} \in [0, \infty)$, the process $\mathcal{X}_t$ for $t \in [0, T]$ and some $\varepsilon > 0$ with overwhelming probability

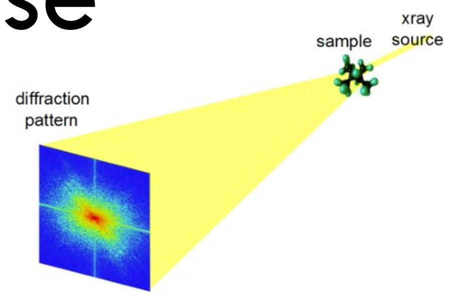
$$\sup_{0 \le t \le T} \left| \varphi\left(X_{\lfloor td \rfloor}\right) - \varphi(\mathcal{X}_t) \right| \le d^{-\varepsilon}$$

Note $X_{\lfloor td \rfloor} \nRightarrow \mathcal{X}_t$

Recall SGD iterates: $X_{k+1} = X_k - \frac{\gamma}{d} \nabla_r f \otimes \boldsymbol{a}_{k+1}^{\top}$, with the population loss $\mathcal{L}(\mathcal{X}) = \mathbb{E}[f]$

Example 1: Phase retrieval – Hard phase

Candes et al., '11



- **Task**: Recover $X^\star \in \mathbb{R}^{1\times d}$, from modulo of projections on the vectors $\boldsymbol{a}$:

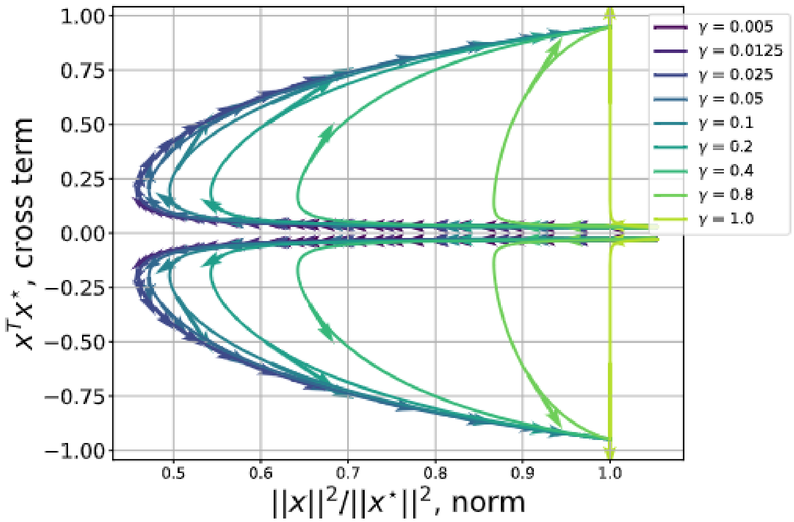
$$\mathcal{L}(X) = \mathbb{E}_{\boldsymbol{a}}[(|X\boldsymbol{a}| - |X^\star \boldsymbol{a}|)^2]$$

Student: $g(X\boldsymbol{a}) = |X\boldsymbol{a}|$, and teacher: $\phi(X^\star\boldsymbol{a}) = |X^\star\boldsymbol{a}|$

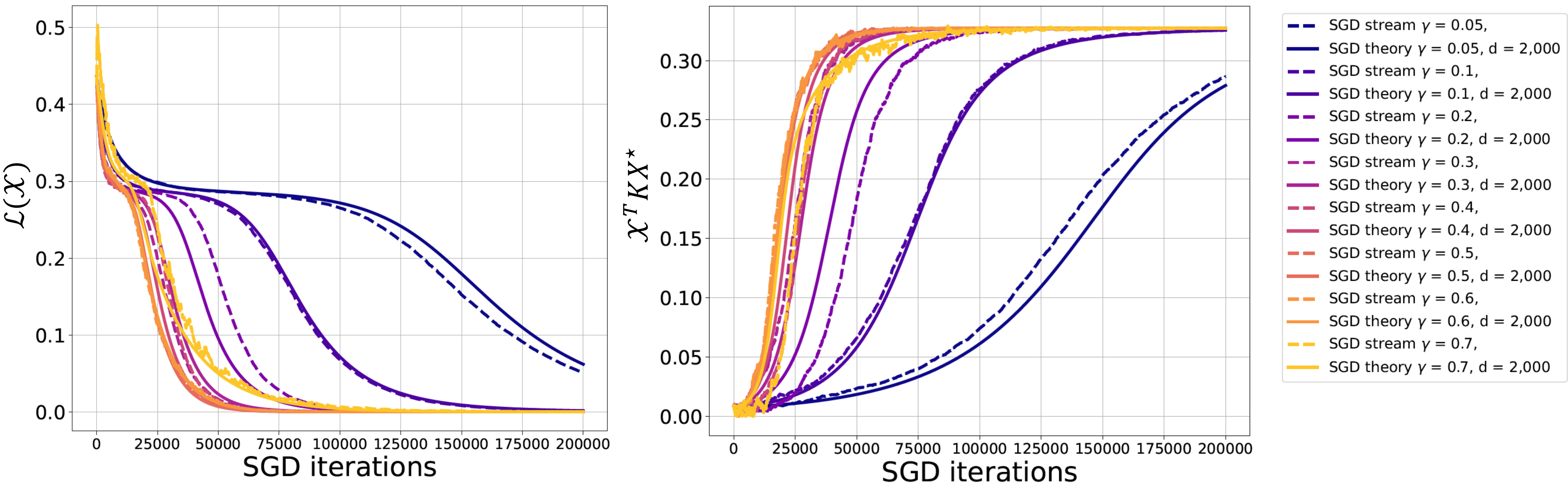
- Random initialization is problematic, suppose $X_0 \sim \mathcal{N}(0, I_d)$

$$\text{Initial overlap} = \langle X_0, X^\star\rangle \sim \frac{1}{\sqrt{d}}$$

- If $\text{Initial overlap} \sim \frac{1}{\sqrt{d}}$ SGD converges in $n = O(d\,\log d)$
- If $\text{Initial overlap} \sim O(1)$ SGD converges in $n = O(d)$
- This can be seen directly from our equation of the norm and overlap



Example: Phase retrieval – risk and alignment



What learning rate ensures descent?

- **Distance to optimality,** by our theorem $\left\|X_{\lfloor td \rfloor} - X^\star\right\|^2 \to \mathcal{D}(t)^2$:

$$\frac{\mathrm{d}\mathcal{D}(t)^2}{\mathrm{dt}} = -2\gamma A(t) + \frac{\gamma^2}{d}\mathrm{Tr}(K) I(t)$$

where $A(t), I(t)$ are functions of the limiting norm and overlap

Thus, $\mathcal{D}(t)^2$ is decreasing when: $\gamma \leq \gamma_t^{\text{stable}} = \frac{2}{\frac{1}{d}\mathrm{Tr}(K)} \frac{A(t)}{I(t)}$

- If for some $m > 0$, $mI(t) \leq A(t)$ (convexity and smoothness assumption):

$$\gamma \leq \frac{2m}{\frac{1}{d}\mathrm{Tr}(K)}$$

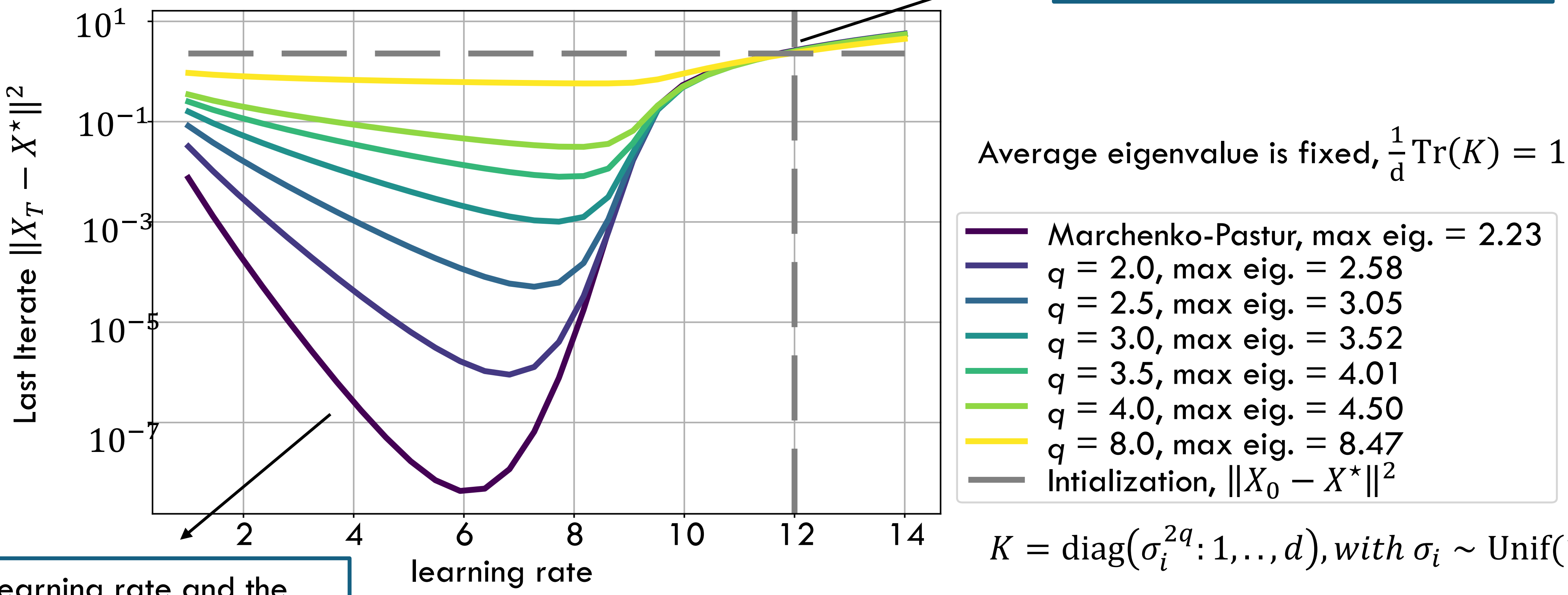
Average eigenvalue rather than largest!

- E..g. if ∇f is Lipschitz with constant L then $m = 1/2L$
- Convergence rate will now depend on $\frac{\lambda_{\min}(K)}{\frac{1}{d}\mathrm{Tr}(K)}$

Dynamical threshold
Motivate ideas such as line search, and Polyak step size

Descent and critical learning rate

$\mathcal{D}(t)^2$ in binary noiseless logistic regression for various γ and K



Threshold learning rate depends only on $\frac{1}{\mathrm{d}}\mathrm{Tr}(K)$

Average eigenvalue is fixed, $\frac{1}{\mathrm{d}}\mathrm{Tr}(K) = 1$

$K = \mathrm{diag}\left(\sigma_i^{2q}: 1,..,d\right), with\ \sigma_i \sim \mathrm{Unif}(1,2)$

Optimal learning rate and the rate of convergence do vary as the max/min eigenvalue changes

Example 2: Stochastic adaptive methods - AdaGrad Norm

- Algorithm setup $X, X^\star \in \mathbb{R}^d$, with $\gamma_0 = \frac{\gamma}{b_0} > 0$:

$$X_{k+1} = X_k - \frac{\gamma_k}{d} \nabla_{X_k} f(X_k; \boldsymbol{a_{k+1}}, \boldsymbol{y_{k+1}})$$

$$\gamma_k = \frac{\gamma}{\sqrt{b_0^2 + \sum_{j=1}^{k} \left\| \nabla_{X_k} f \right\|^2}}$$

- Deterministic equivalence $\gamma_{\lfloor td \rfloor} \to \gamma_t$:

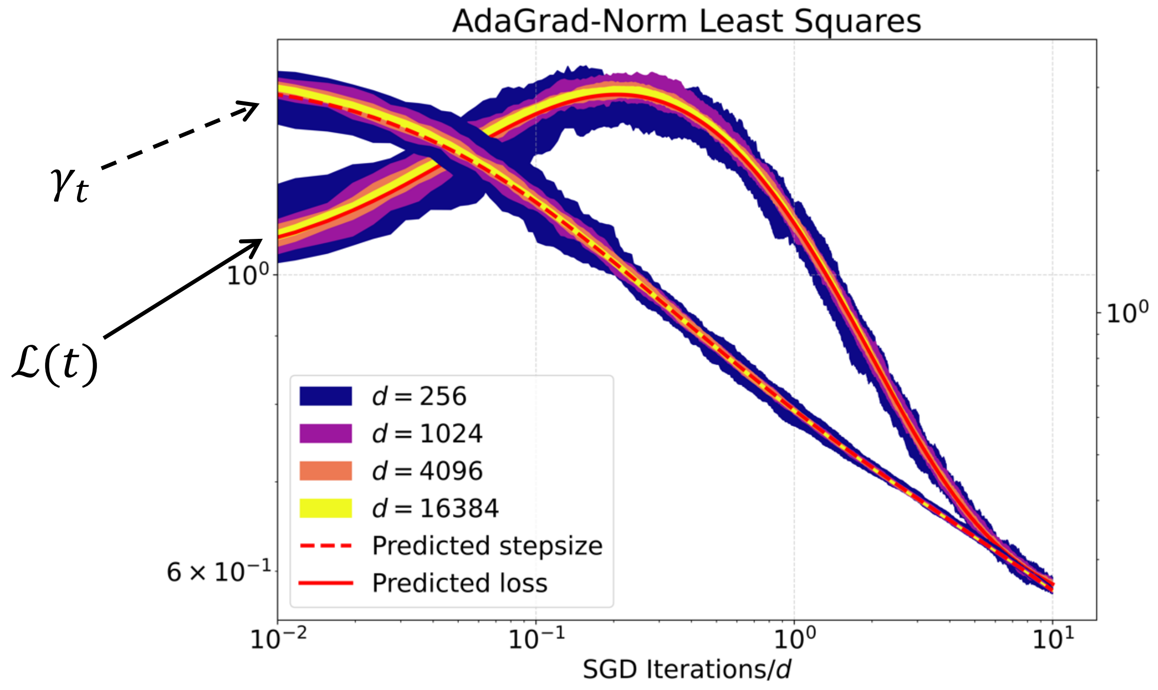
$$\gamma_t = \frac{\gamma}{\sqrt{b_0^2 + \frac{\mathrm{Tr}(K)}{d} \int_0^t I(s) \mathrm{d}s}}$$

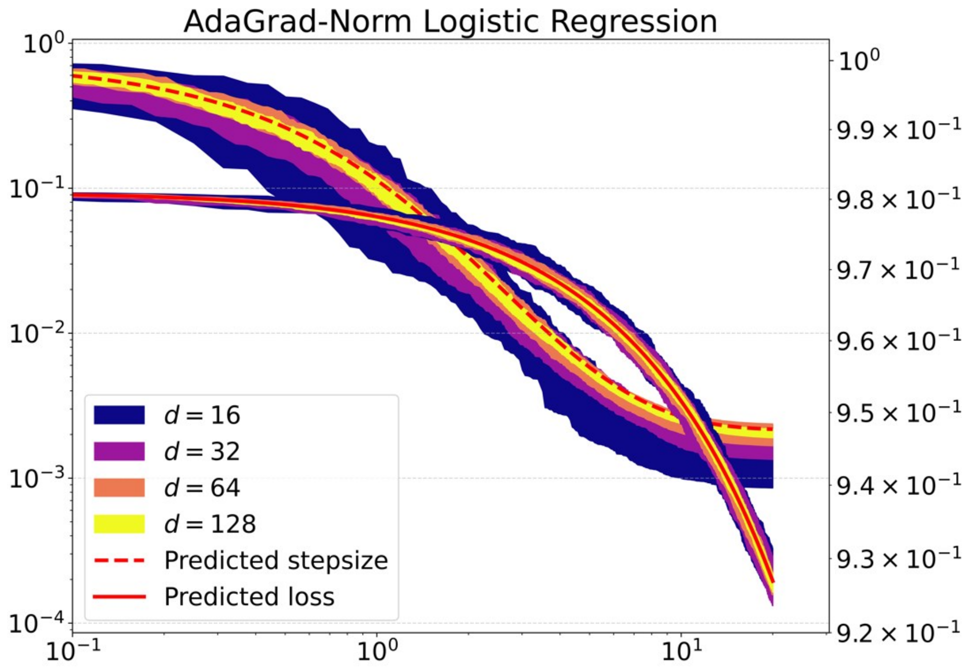
with $I(s) = \mathbb{E}[f'(X^T a)^2]$

- For Least square: $I(s) = 2\mathcal{L}(s)$.

R. Ward, X.Wu, and L.Bottou, JMLR 21(1):9047–9076, 2020

Stochastic adaptive methods – Exact Adagrad dynamics



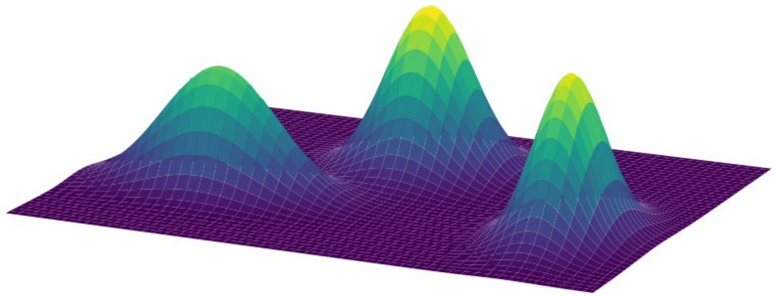


Collins-Woodfin, E., **SI.**, Malaxechebarría, B.G., Mackenzie, A.W., Paquette, E. & Paquette, C., arXiv preprint arXiv:2405.19585. *NeurIPS 2024*

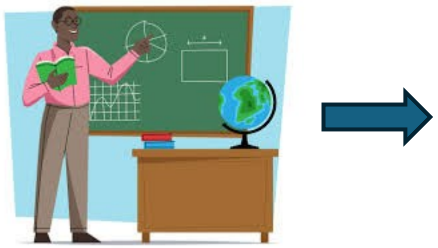
How can we extend this to the multiclass setting?

Gaussian Mixture Model (GMM)

Joint work (in progress) with Elizabeth Colins-Woodfin



Model-setup - Data distribution



$\{(\boldsymbol{a}_{i,I}, \boldsymbol{y}_{i,I})\}_{i=1}^{n}$

Target (teacher) model: Data from $\ell^\star$ classes:

$$\boldsymbol{a}_{i,I} \sim \mathcal{N}(\mu_I, K_I), \text{ with } K_I \in \mathbb{R}^{d\times d}, \|K_I\|_{\text{op}} \text{ bounded, all } \{K_I\} \text{ commute}$$

$c \in [\ell^\star] = O\big(\log(d)\big)$, and $p_c = \mathbb{P}(c = I)$

Our setup allow for the following two settings:

"Hard label" - $\boldsymbol{y}_{i,I} = \boldsymbol{I}$ or one - hot encoding of $\boldsymbol{I}$

"Soft label" : $y_{i,I} = \phi_I\big(X^\star \boldsymbol{a}_{i,I}; \boldsymbol{\varepsilon}_i\big)$ e.g. $y_i = \phi_I\big(X^\star \boldsymbol{a}_{i,I}, \boldsymbol{\varepsilon}\big) = \text{softmax}(X^\star \boldsymbol{a}_i)$ $\quad \text{softmax}(\mathrm{r})_{\mathrm{i}} = \frac{\mathrm{e}^{r_i}}{\sum_j e^{r_j}}$

Classifier and optimization problem GMM

$\{(\boldsymbol{a}_{i,I}, \boldsymbol{y}_{i,I})\}_{i=1}^{n}$ →  → $g_I(X\boldsymbol{a}_{i,I})$

Classifier (student) model:

Choose a function $g\colon \mathbb{R}^{\ell} \to \mathbb{R}^{m}$, **estimate using online SGD** the matrix $X \in \mathbb{R}^{\ell\times d}$

$\ell = O(\log(d))$

Optimization problem:

$$\min_{X\in\mathbb{R}^{d\times\ell}} L(X) = \mathbb{E}_{(\boldsymbol{a},I)}[f_I(X\boldsymbol{a}_I; \boldsymbol{y}_I)]$$

Related work: Seddik et al ICLR 2020, Loureiro et al NIPS 2021, Mai &Liao 2019, Ben-Arous et al 2025

Main result - deterministic equivalence

<u>Theorem</u> (Collins-Woodfin, **SI** '25) Fix $T = \frac{n}{d} > 0$. For any $\epsilon \in (0, \frac{1}{2})$, with overwhelming probability,

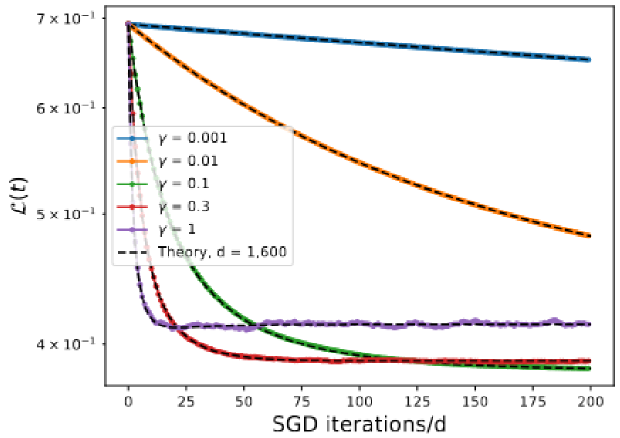
$$\sup_{0 \leq t \leq T} |L(X_{\lfloor td \rfloor}) - \mathcal{L}(t)| < Cd^{-\epsilon}$$

where $\mathcal{L}(t)$ is the "deterministic equivalent" of the risk, expressible in terms of a system of autonomous ODEs.

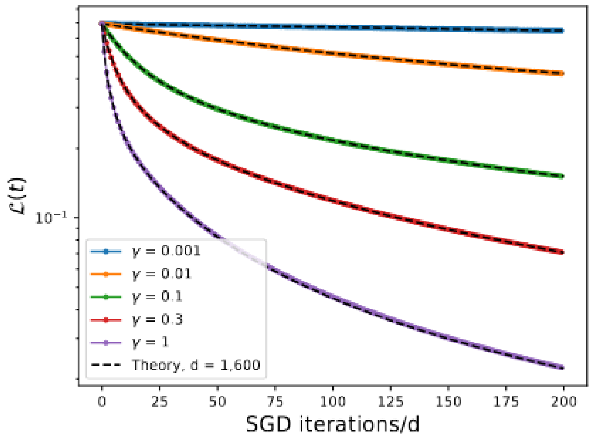
- This holds for other statistics of X, not just risk.

How does the structure of the classes affect
the SGD dynamics?

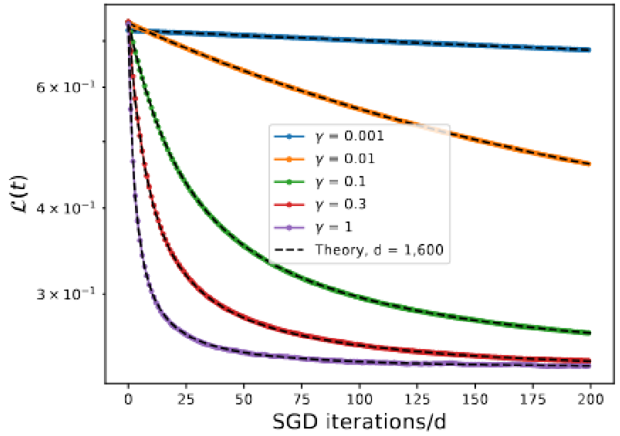
SGD vs theory for different data models



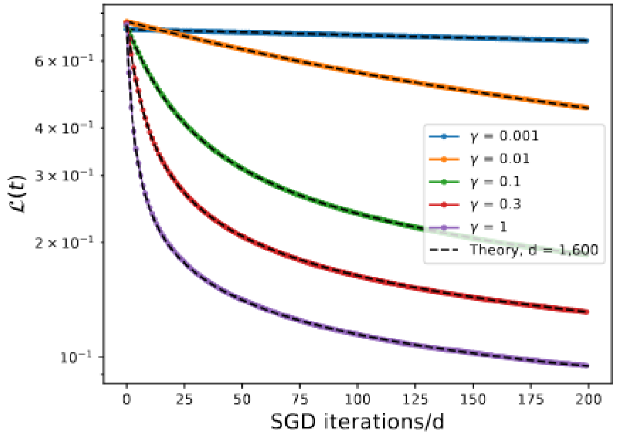
(a) Identity covariance



(b) Zero-one model

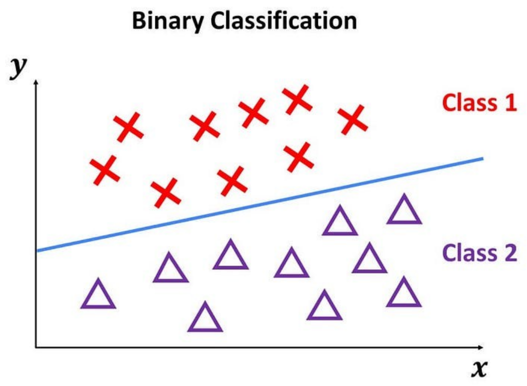


(c) Mild Power-law $\beta = 1$



(d) Extreme power-law $\beta = 0.2$

Binary logistic regression



Two classes: $y_i = 1_{i=1\_}$ and $a \mid i = 1 \sim N(\mu_1, K_1)$, and $a \mid i = 2 \sim N(\mu_2, K_2)$, wit

$$L(X) = \mathbb{E}_{(a,y)}\left[-a^\top X y + \log\left(e^{a^\top X} + 1\right)\right]$$

For simplicity: $\mu_1 = -\mu_2 = \mu \qquad K_1 = \mathrm{diag}\left(\lambda_1^{(1)}, \ldots, \lambda_d^{(1)}\right) \qquad K_2 = \mathrm{diag}(\lambda_1^{(2)}, \ldots, \lambda_d^{(2)})$

Identity model: $K_1 = K_2 = I_d$ $(|I_{11}| = d)$

Zero-One model - All eigenvalues in {0, 1}

- Partition indices $\{1, \cdot \ \cdot \ \cdot, d\} = I_{00} \sqcup I_{01} \sqcup I_{10} \sqcup I_{11}$
- where $I_{jk} := \{ i \leq d \ \lambda_i^{(1)} = j, \lambda_i^{(2)} = k\}$

Example $d = 4$:

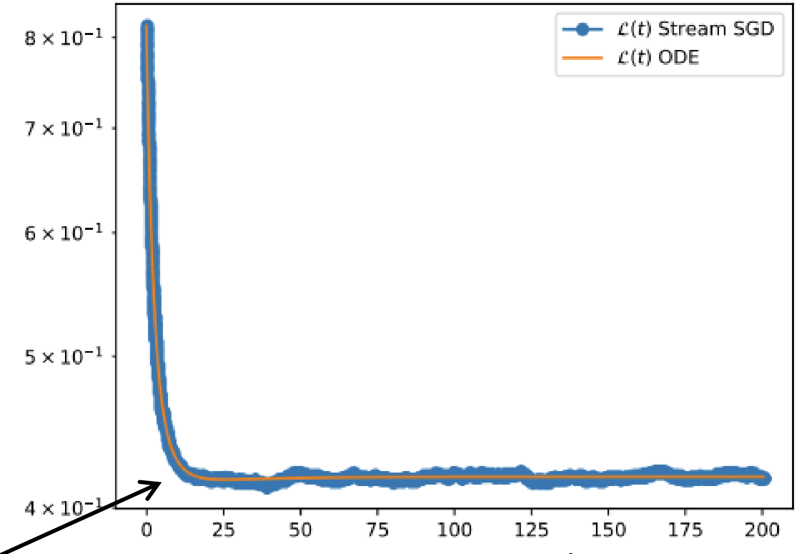
$$K_1 = \begin{bmatrix} 0 & & & \\ & 0 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \qquad K_2 = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & 0 & \\ & & & 1 \end{bmatrix}$$

Comparing Identity and Zero-One models - Risk

Does SGD find the "perfect" subspace?

How "clean" directions (I_{00}) affect the classification?

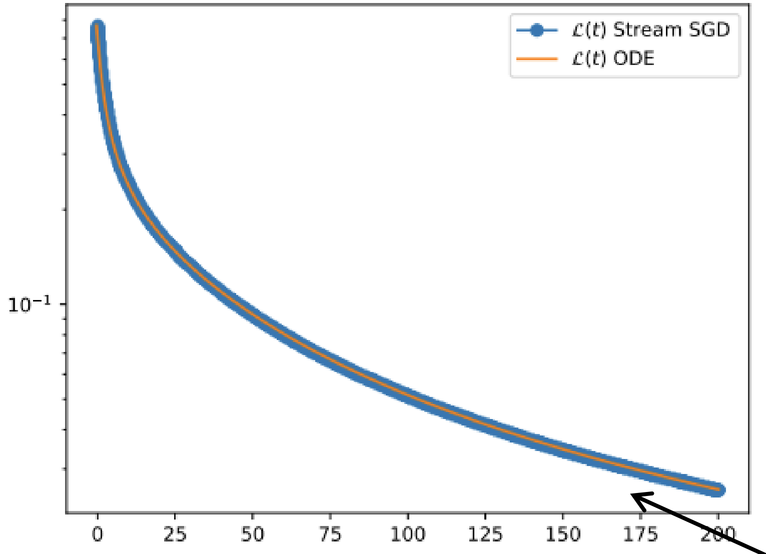
Learning curve - identity



SGD iterations/d

flattens out at around 0.4

Learning curve - Zero-one



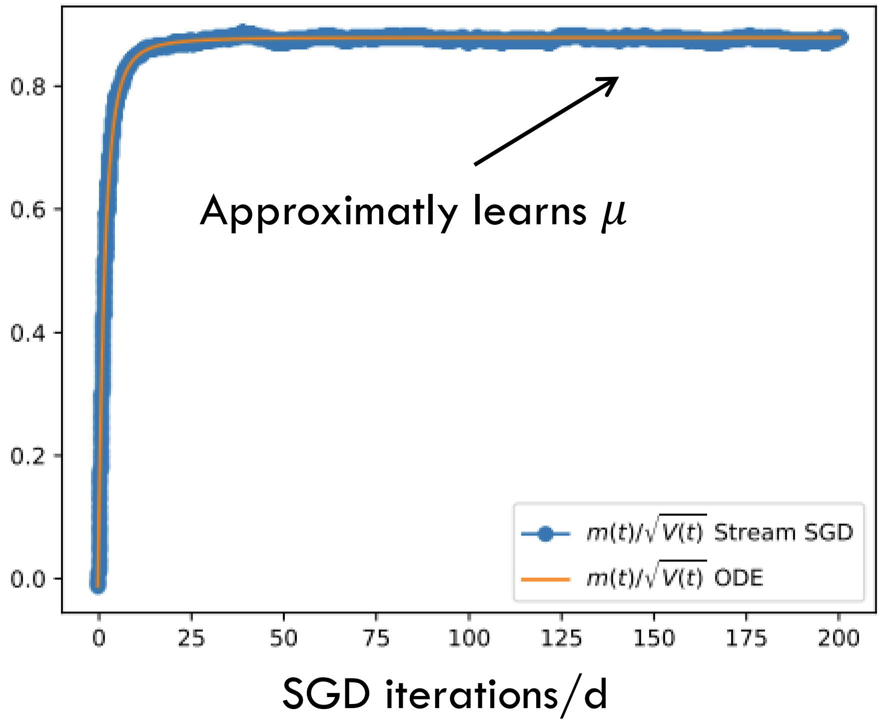
SGD iterations/d

Continue decreasing

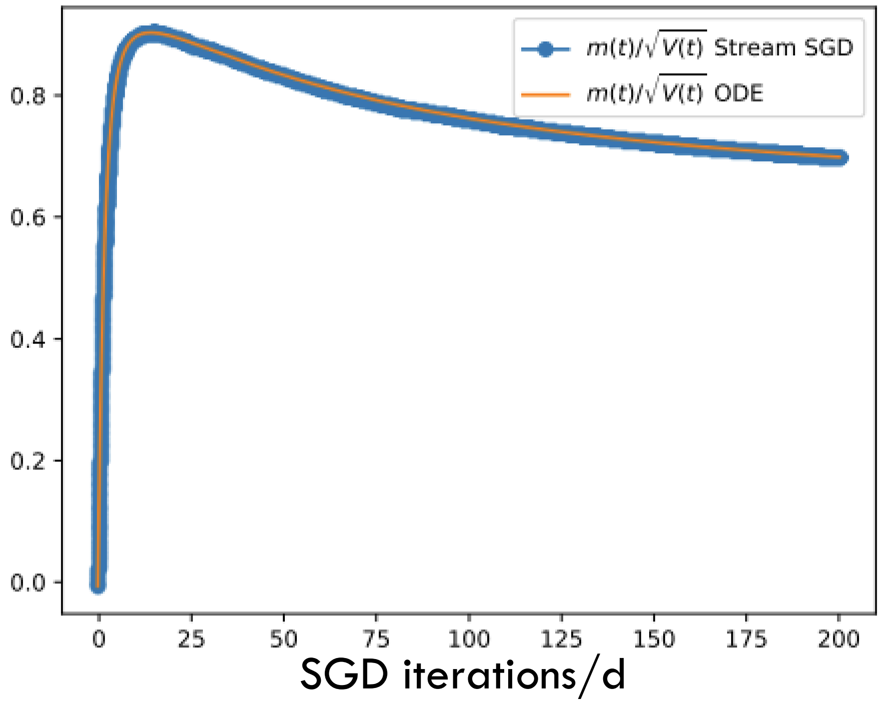
Comparing Identity and Zero-One models – Alignment

$$\text{Alignment} := \frac{\mu^\top X_{\lfloor td \rfloor}}{\|X_{\lfloor td \rfloor}\|} \to \frac{m(t)}{\sqrt{\mathcal{V}(t)}}$$

Alignment with μ - Identity



Alignment with μ – Zero-One



Zero-One asymptotic

Proposition (Collins-Woodfin, **SI** '25) : For $\gamma < 1, p_1 = \frac{1}{2}, |I_{00}| = |I_{01}| = |I_{10}| = |I_{11}| = d/4$. There exist $C_1(\gamma), C_2(\gamma)$ such that

$$t^{-C_1(\gamma)} \leq \mathcal{L}(t) \leq t^{-C_2(\gamma)}$$

where the alignment with μ:

$$\mathcal{m}(t) \asymp \log t\,, \qquad \frac{\mathcal{m}(t)}{\sqrt{\mathcal{V}(t)}} = \frac{1}{2}\left(1 + O((\log t)^{-1/2})\right)$$

Remarks:

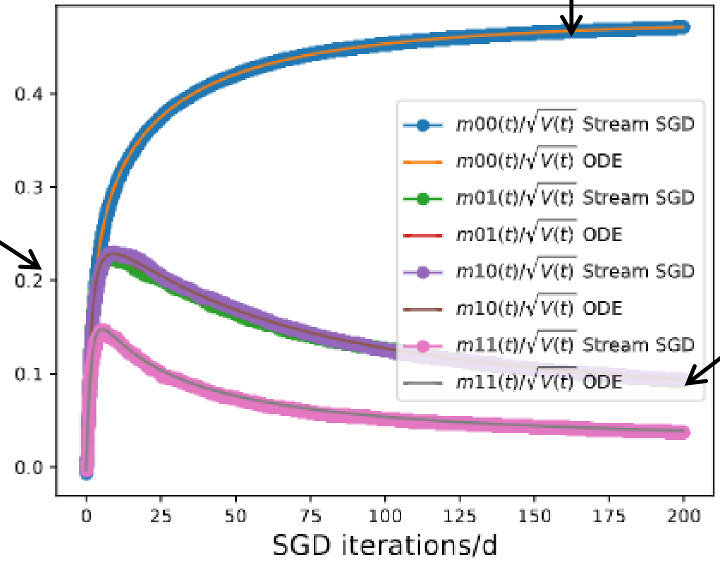
- This implies analogous bounds on the **original loss** in high dimension by our Theorem
- $\mathcal{V}(t) \approx \left\|X_{\lfloor kd \rfloor}\right\|^2$ and the $\mathcal{m}(t) \approx \mu^\top X_{\lfloor kd \rfloor}$ grows logarithmically with n! (very different than the identity setting!)
- The covariance matrices has no power law structure.

Perfect classification vs clean directions

- What direction do we learn?
  - **Largest distance** between class means (μ)
  - **Smallest variance** in data (eigenspace of I_{00})
- Project into each eigenspace I_{jk}, denote by $m_{jk}(t), \mathcal{V}_{ik}(t)$

X continues growing in the I_{00} of μ direction

Initially X aligns in the μ direction



X remain bounded in other directions

$d = 1000$ and $\gamma = 0.9$

In particular, we can prove that:

$$m_{00}(t) \asymp \log t\,,$$

$$m_{01}(t), m_{10}, (t), m_{11}(t) = O(\sqrt{\log t}),$$

All in one - Power law covariance and mean

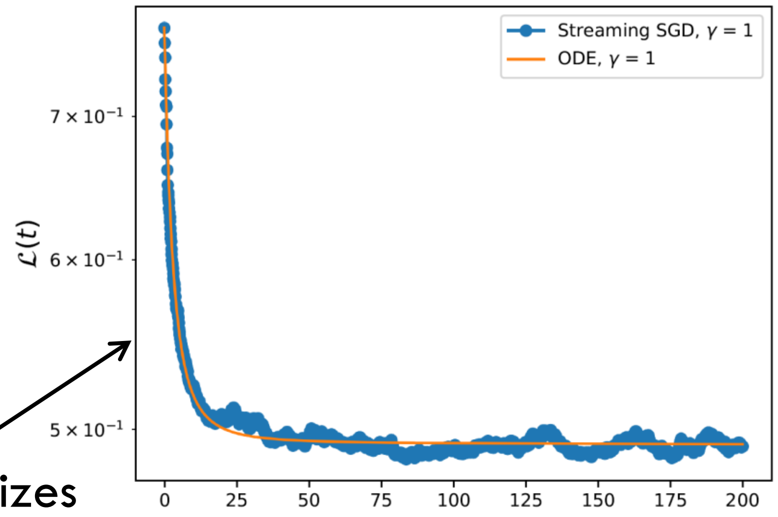
Power law model: $K_1 = K_2 = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \mu_1 = -\mu_2 = \mu$

$$\lambda_i = \left(\frac{i}{d}\right)^{\alpha} \quad \text{and } \mu_i^2 = \frac{1}{d}\left(\frac{i}{d}\right)^{\beta}, \text{ for } \beta \geq 0, \alpha > 1$$

Spectrum with eigenvalues accumulating near zero

- There is a **phase transition** at $\alpha = 1 + \beta$

Mild power law

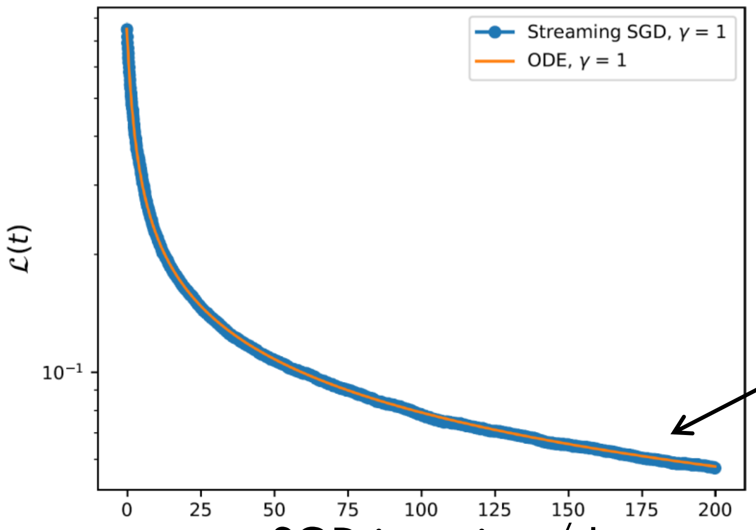


Risk stabilizes as in the identity case

SGD iterations/d

$1 + \beta = 2.2 > \alpha = 1.2$

Extreme power law



Risk continues to go down

SGD iterations/d

$1 + \beta = 1 < \alpha = 1.2$

Mild power law and identity regime

<u>Proposition</u>: Suppose $X_0 = 0, \gamma < 1, p_1 = \frac{1}{2}$. Then for $t \geq 1$,

Mild power law $(\alpha < \beta + 1)$

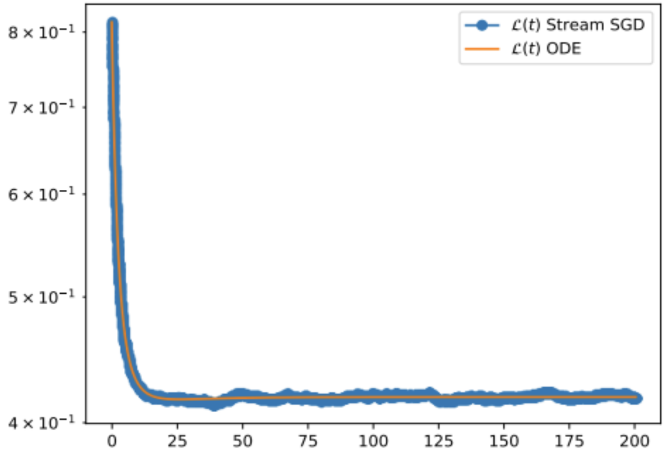
- $m(t) \asymp \mu^\top [K]^{-1}\mu$
- $\mathcal{L}(t) \asymp \mathcal{L}_{\min}$>0

Identity covariance: $K = I_d$ with $\|\mu\| = O(1)$

- $m(t) \asymp \mu^\top [K]^{-1}\mu$
- $\mathcal{L}(t) \asymp \mathcal{L}_{\min}$>0

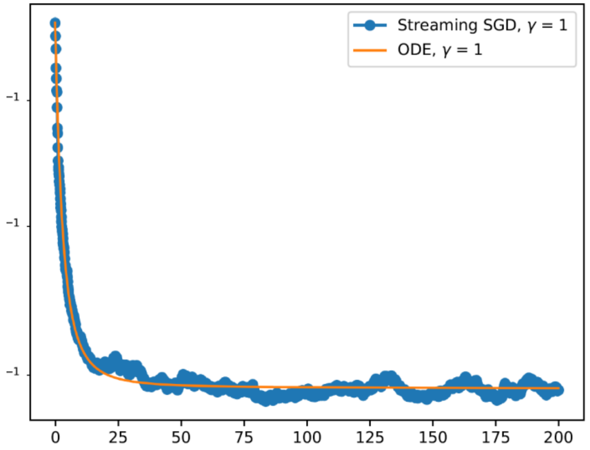
Rate of convergence are different!

identity



SGD iterations/d

Mild power law, $\alpha < \beta + 1$



SGD iterations/d

Extreme power law ($\alpha \geq \beta + 1$)

Proposition: Suppose $X_0 = 0, \gamma < 1, p_1 = \frac{1}{2}$. Then for $t \geq 1$,

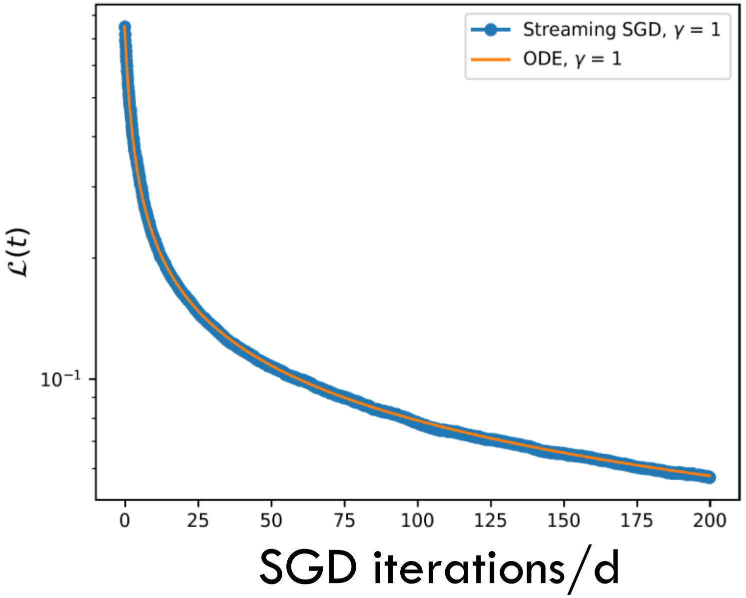
- $m(t)$ **grows** with t at a polylog rate
- $\mathcal{L}(t) \to 0$ faster than polynomial decay,

  but still slower than exponential decay.

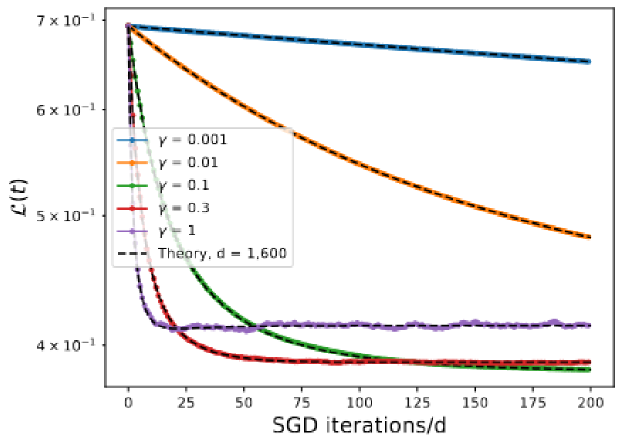
Remarks:

- Closely related to the Zero-One thought rates are different!
- **Small variance** directions contribute the **most to the learning.**

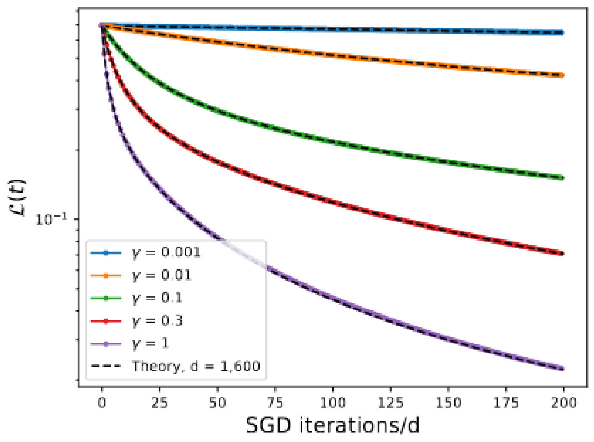
Extreme power law, $\alpha > \beta + 1$



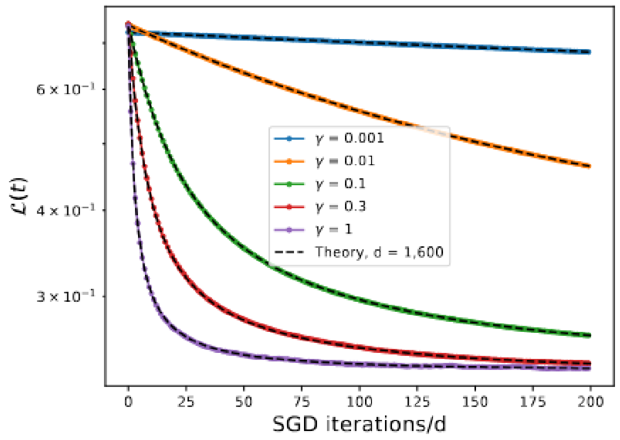
SGD vs theory for different data models



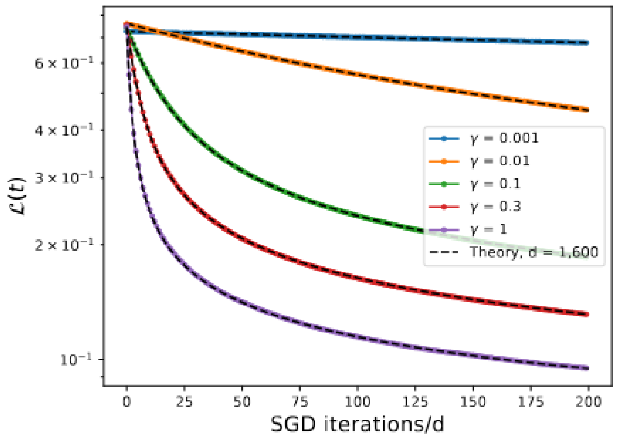
(a) Identity covariance



(b) Zero-one model



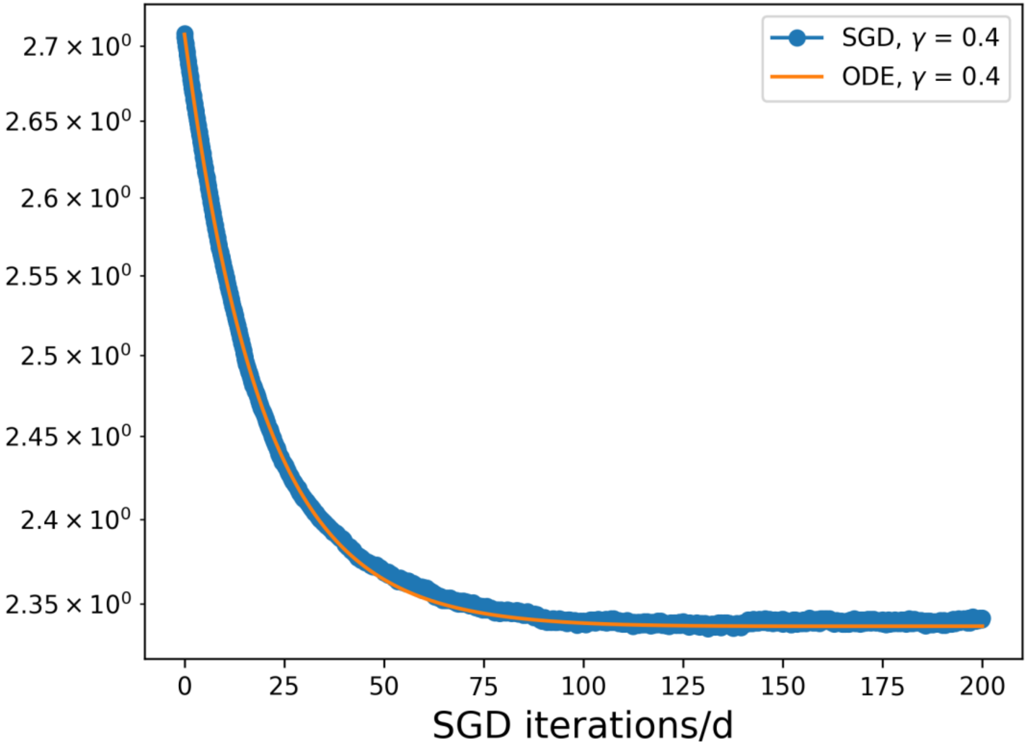
(c) Mild Power-law $\beta = 1$



(d) Extreme power-law $\beta = 0.2$

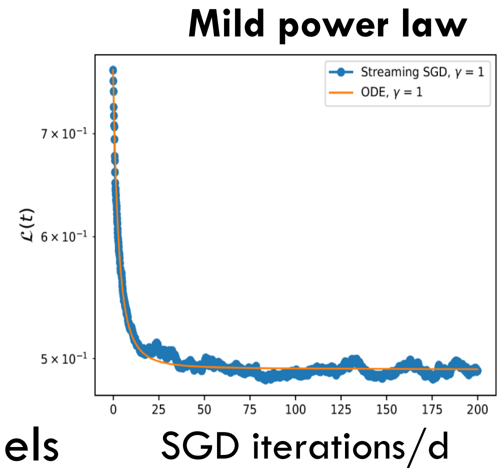
Large number of classes

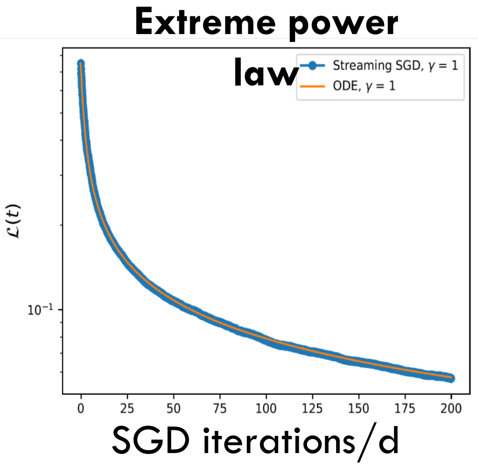
- We allow the number of classes to grow as $\ell = \ell^\star = O(\log(d))$



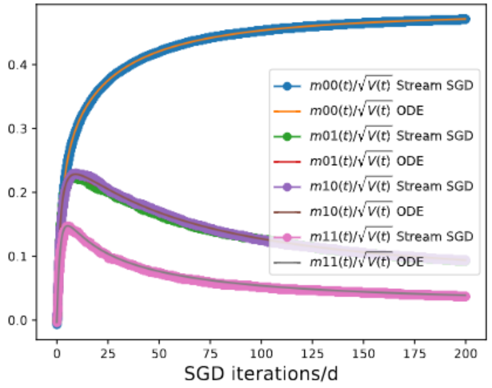
$\ell = \ell^\star = 15, d = 500$

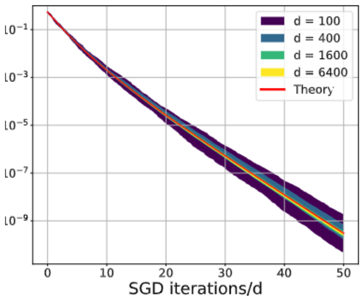
Conclusions

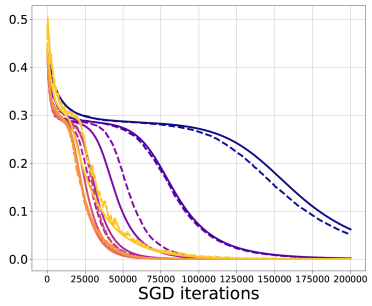


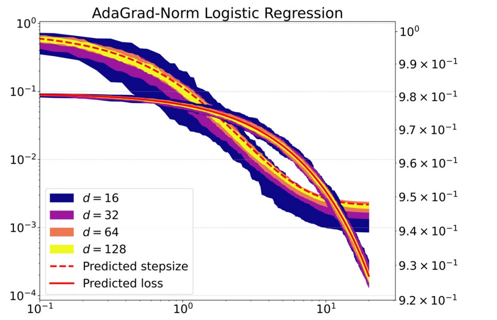


- An **exact** asymptotic theory of online SGD for multi-index models
- Applies to **stochastic adaptive methods,** such as Adagrad
- Extension of the theory for **nonisotropic** Gaussian mixture models
- Algorithm-dependent scaling laws and **phase transition** as a function of the structure
- Asymptotic analysis show the exact scaling behavior of the loss and other statistics
- Allow for growing number of classes $\ell^\star = O(\log d)$









Thank You!

Questions?

- Collins-Woodfin, E & **Seroussi, I** ”SGD dynamics for Gaussian Mixtures models with non-isotropic Covariance and mean” (To appear soon!)
- Collins-Woodfin, E., Paquette, C., Paquette, E., & **Seroussi, I**. (2024). Hitting the high-dimensional notes: An ode for SGD learning dynamics on GLMs and multi-index models. *Information and Inference: A Journal of the IMA*, *13*(4), iaae028.
- Collins-Woodfin, E., **Seroussi, I.**, Malaxechebarría, B.G., Mackenzie, A.W., Paquette, E. and Paquette, C., 2024. The High Line: Exact Risk and Learning Rate Curves of Stochastic Adaptive Learning Rate Algorithms. arXiv preprint arXiv:2405.19585. *NeurIPS 2024*