The High Line: Exact Risk and Learning Rate Curves of Stochastic Adaptive Learning Rate Algorithms

Elizabeth Collins-Woodfin, McGill University

Inbar Seroussi, McGill University, Tel-Aviv University

Begoña García Malaxechebarría, University of Washington

Andrew W. Mackenzie, McGill University

Elliot Paquette, McGill University

Courtney Paquette, McGill University, Google DeepMind

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

Setting

- **Stochastic Gradient Descent (SGD)**: We study streaming SGD with batch size 1. At each iteration, the algorithm computes a stochastic gradient based on a single data point and moves one step in the decreasing direction
- **High-Dimensional Linear (High Line) Composite Models:** Our theorem applies to various models including linear regression, logistic regression, and simple neural nets

Goal

Analyze the dynamics of SGD with adaptive learning rates (SGD+AL) in high dimensions

Main Contributions

- Training dynamics of **SGD+AL converge** to the solution of a deterministic system of ODEs
- Greed can be **arbitrarily bad** in the presence of strong anisotropy
- AdaGrad-Norm selects the **optimal learning rate**, provided it has a warm start
- AdaGrad-Norm can use **overly pessimistic** decaying schedules on hard problems

Model Setup

Optimization Problem

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left\{\mathcal{R}(\mathbf{x})\stackrel{\text{def}}{=}\mathbb{E}_{\mathbf{a},\epsilon}\left[f(\mathbf{a}^\top\mathbf{x};\mathbf{a}^\top\mathbf{x}^\star,\epsilon)\right]\right\}$$

- $\mathbf{a}\in\mathbb{R}^d$, $\mathbf{a}\sim\mathcal{N}(0,\mathbf{K})$
- $\epsilon\in\mathbb{R}$, $\epsilon\sim\mathcal{N}(0,\omega^2)$
- $\|\mathbf{K}\|_{\text{op}}, \|\mathbf{x}^\star\|_2$ bounded independent of d
- Includes problems like: least squares, logistic regression, one-neuron networks
- Our goal is to classify limiting behavior as $d\to\infty$

SGD+AL Algorithm

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\frac{\gamma_k}{d}\nabla f(\mathbf{a}_k^\top\mathbf{x}_k;\mathbf{a}_k^\top\mathbf{x}^\star,\epsilon_k)$$

- $\|\mathbf{x}_0\|_2$ is bounded independent of d
- γ_k can depend on historical norms of gradients $\|\nabla f\|_2$, losses $\mathcal{R}(\mathbf{x})$, and iterate norms $\|\mathbf{x}_k\|_2$
- γ_k is bounded in its arguments
- Includes algorithms like: AdaGrad-Norm, RMSProp, DoG, D-Adaptation

Specific Algorithms

Example: AdaGrad-Norm

$$\gamma_k=\frac{\eta}{\sqrt{b^2+\frac{1}{d^2}\sum_{j=0}^k\left\|\nabla f(\mathbf{a}_k^\top\mathbf{x};\mathbf{a}_k^\top\mathbf{x}^\star,\epsilon_k)\right\|^2}}$$

- AdaGrad, but with a global learning rate rather than adjusted on a per-weight basis.
- Stepsize is automatically bounded by $\frac{\eta}{b}$.
- Depends only on the norms $\|\nabla f\|_2$ of past gradients.

Require: $\eta>0$, $\mathbf{x}_0\in\mathbb{R}^d$, $b\in\mathbb{R}$, $b_0=bd$
for $k=1,2,\ldots,$ **do**
  Take $\mathbf{a}_k\sim\mathcal{N}(0,\mathbf{K})$, $\epsilon_k\sim\mathcal{N}(0,\omega^2)$;
  $\nabla_k\leftarrow\nabla f(\mathbf{a}_k^\top\mathbf{x};\mathbf{a}_k^\top\mathbf{x}^\star,\epsilon_k)$
  $b_k^2\leftarrow b_{k-1}^2+\|\nabla_k\|^2$;
  $\gamma_{k-1}\leftarrow d\times\frac{\eta}{|b_k|}$; ▷ update stepsize
  $\mathbf{x}_k\leftarrow\mathbf{x}_{k-1}-\frac{\gamma_{k-1}}{d}\nabla_k$; ▷ weights
end for

Main Concentration Result

We define $S:\mathbb{R}^d\to\mathbb{R}^{2\times2}$ given by

$$S(\mathbf{x};z)=\begin{bmatrix}\mathbf{x}^\top R(z,\mathbf{K})\,\mathbf{x} & \mathbf{x}^\top R(z,\mathbf{K})\,\mathbf{x}^\star\\ \mathbf{x}^\top R(z,\mathbf{K})\,\mathbf{x}^\star & (\mathbf{x}^\star)^\top R(z,\mathbf{K})\,\mathbf{x}^\star\end{bmatrix}$$

where $R(z,\mathbf{K})=(\mathbf{K}-z\cdot I_d)^{-1}$ for $z\in\mathbb{C}\setminus\sigma(\mathbf{K})$ is the **resolvent of K**

Theorem [Informal]

S along SGD concentrates around the deterministic solution to the system of ODEs

$$\mathrm{d}\mathscr{S}(t;z)=\mathscr{F}(\mathscr{S}(t;z))$$

We consider $\varphi:\mathbb{R}^d\to\mathbb{R}$ given by

$$\varphi(\mathbf{x})=g\left(\begin{bmatrix}\mathbf{x}^\top q(\mathbf{K})\,\mathbf{x} & \mathbf{x}^\top q(\mathbf{K})\,\mathbf{x}^\star\\ \mathbf{x}^\top q(\mathbf{K})\,\mathbf{x}^\star & (\mathbf{x}^\star)^\top q(\mathbf{K})\,\mathbf{x}^\star\end{bmatrix}\right)$$

where g is **α-pseudo-Lipschitz** with $\alpha\le1$ and q is a **polynomial**

- Can recover $\mathcal{R}(\mathbf{x})$ and $D(\mathbf{x})\stackrel{\text{def}}{=}\|\mathbf{x}-\mathbf{x}^\star\|$ from φ
- Using **Cauchy's integral formula**,

$$\varphi(\mathbf{x})=g\left(\frac{1}{2\pi i}\oint_\Gamma q(z)S(\mathbf{x};z)\,\mathrm{d}z\right)$$

where Γ is a fixed contour around spectrum of $\mathbf{K}$

Corollary [Informal]

φ along SGD concentrates around the deterministic function

$$\phi(t)=g\left(\frac{1}{2\pi i}\oint_\Gamma q(z)\mathscr{S}(t,z)\,\mathrm{d}z\right)$$

- We refer to ϕ as **deterministic equivalent** of φ
- In particular, we define $\mathscr{R}$ and $\mathscr{D}$ as deterministic equivalents of $\mathcal{R}$ and D, respectively
- We can derive an ODE $\mathrm{d}\phi(t)=\mathscr{G}(\mathscr{S}(t;z))$

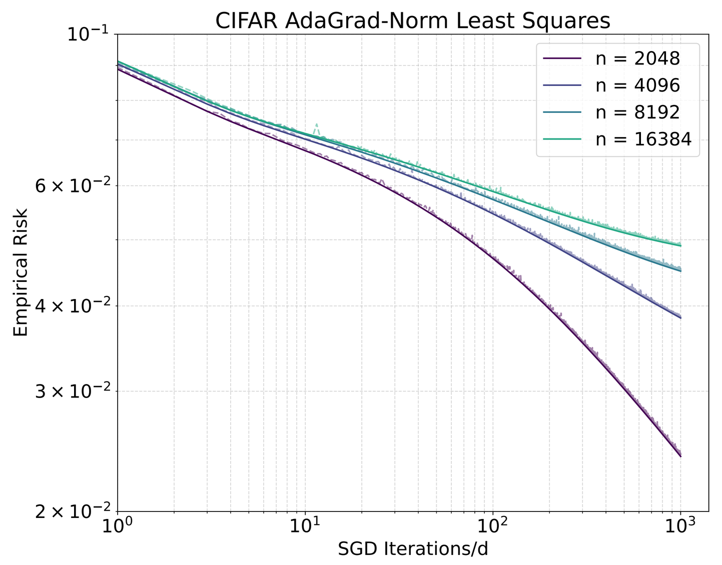
Beyond Gaussian Data: CIFAR-5m

Discrete Problem

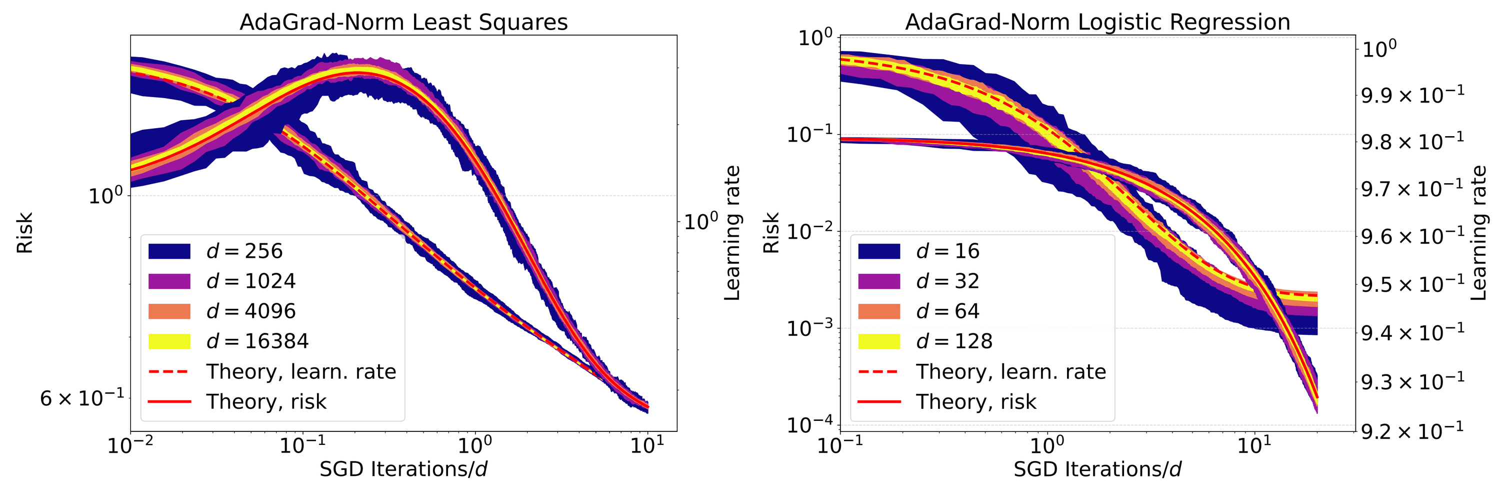
$$\min_{\mathbf{x}\in\mathbb{R}^d}\left\{\mathcal{R}(\mathbf{x})\stackrel{\text{def}}{=}\frac{1}{2n}\|\mathbf{F}\mathbf{x}-\mathbf{b}\|^2=\frac{1}{2n}\sum_{i=1}^n(\mathbf{f}_i\cdot\mathbf{x}-b_i)^2\right\}$$

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\gamma_k(\mathbf{f}_{i_{k+1}}\cdot\mathbf{x}_k-b_{i_{k+1}})\mathbf{f}_{i_{k+1}},\qquad \{i_k\}\text{ iid Unif}(\{1,2,\cdots,n\})$$

- Binary classification with least squares; γ_k is AdaGrad-Norm learning rate.
- Take n images from two classes of CIFAR-5m, reshape into a matrix $\mathbf{A}\in\mathbb{R}^{n\times1024}$ (preconditioned to have centered rows with norm 1.) $\mathbf{b}\in\mathbb{R}^n$ has $b_i=1$ if the corresponding image is an airplane and $b_i=0$ otherwise.
- Generate matrix $\mathbf{W}\in R^{1024\times d}$ with iid Gaussian entries, set features $\mathbf{F}=\text{relu}(AW)$.
- Shown: AdaGrad-Norm true vs predicted loss for $d=2000$. Concentration is nearly perfect.
- For small n, SGD can overfit and learn quickly; for larger n, a general mapping must be learned, so loss decreases more slowly.



Main Result Examples



Concentration of learning rate and risk for AdaGrad-Norm on least squares with label noise $\omega=1$ *(left)* and logistic regression with no noise *(right)*. As dimension increases, both risk and learning rate concentrate around a deterministic limit *(red)* described by our ODE. The initial risk increase *(left)* suggests the learning rate started too high, but AdaGrad-Norm adapts. Our ODEs predict this behavior.

Classical Idealized Algorithms Analysis

Two main interests for choosing the learning rate at each iteration: **maximize the decrease** in risk or in distance to optimality

- For stochastic algorithms, this is not feasible

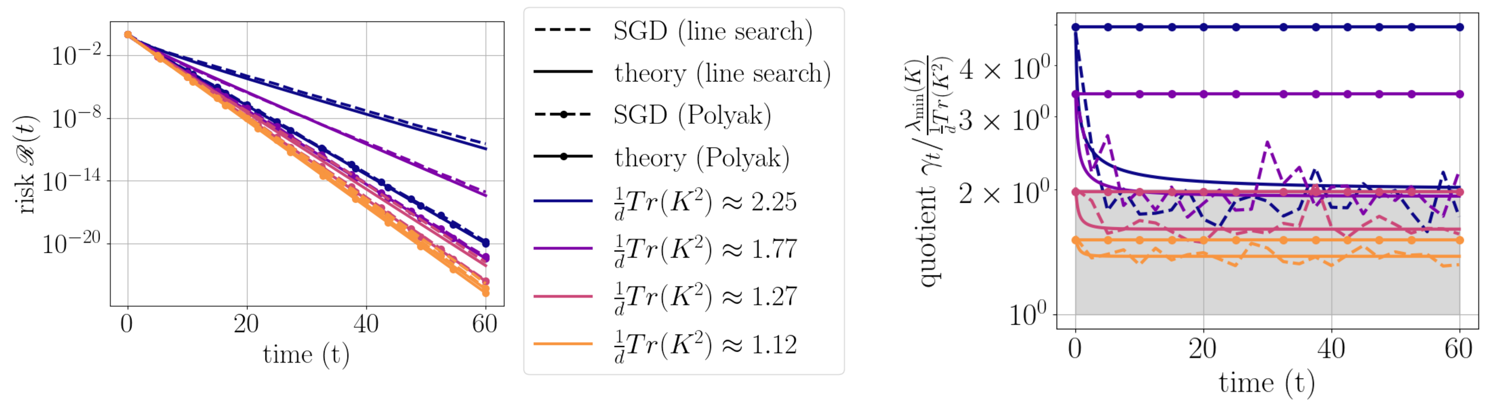
Consider the stochastic **idealized** algorithms whose deterministic equivalents satisfy

$$\gamma_t^{\text{Line Search}}\in\arg\min_\gamma \mathrm{d}\mathscr{R}(t)\qquad\textsc{Exact Line Search}$$

$$\gamma_t^{\text{Polyak}}\in\arg\min_\gamma \mathrm{d}\mathscr{D}(t)\qquad\textsc{Polyak}$$

In the **noiseless least squares** problem with $\lambda_{\min}(\mathbf{K})>C>0$,

$$\gamma_t^{\text{Polyak}}=\frac{1}{\frac{1}{d}\operatorname{tr}(\mathbf{K})}\quad\text{and}\quad\gamma_t^{\text{Line Search}}\asymp\frac{\lambda_{\min}(\mathbf{K})}{\frac{1}{d}\operatorname{tr}(\mathbf{K}^2)}$$



Comparison for Exact Line Search and Polyak on a noiseless least squares problem. The left plot illustrates the convergence of the risk function, while the right plot depicts the convergence of the quotient $\gamma_t/\frac{\lambda_{\min}(\mathbf{K})}{\frac{1}{d}\operatorname{Tr}(\mathbf{K}^2)}$ for Polyak and Exact Line Search. Both plots highlight that, in high-dimensional settings, a broader spectrum of $\mathbf{K}$ results in $\frac{\lambda_{\min}(\mathbf{K})}{\frac{1}{d}\operatorname{Tr}(\mathbf{K}^2)}\ll\frac{1}{\frac{1}{d}\operatorname{Tr}(\mathbf{K})}$, indicating slower risk convergence and poorer performance of Exact Line Search (unmarked) as it deviates from Polyak (circle markers).

AdaGrad-Norm Analysis

We analyze the behavior of AdaGrad-Norm in the **least squares** setting. In the presence of additive **noise**, the learning rate decays like $t^{-1/2}$, regardless of the data covariance $\mathbf{K}$. In contrast, the model with **no noise** exhibits a learning rate that depends on the spectrum of $\mathbf{K}$. We consider **three cases**:

Spectrum of K bounded below

In the noiseless least squares problem with $\lambda_{\min}(\mathbf{K})>C>0$, integrable risk, $\frac{1}{d}\operatorname{tr}(\mathbf{K})\le\frac{b}{\eta}$

$$\gamma_t^{\text{AdaGrad-Norm}}\asymp\frac{\eta^2}{\frac{b}{\eta}+\frac{1}{4d}\operatorname{tr}(\mathbf{K})\|\mathbf{x}_0-\mathbf{x}^\star\|^2}.$$

$o(d)$ eigenvalues below fixed threshold

With $o(d)$ eigenvalues below some fixed threshold, $\mathbf{x}^\star$ not aligned with eigenvectors, $\mathbf{x}_0=0$, there exists $\tilde\gamma\ge0$ such that

$$\gamma_t^{\text{AdaGrad-Norm}}\ge\tilde\gamma\qquad\text{for all }t\ge0.$$

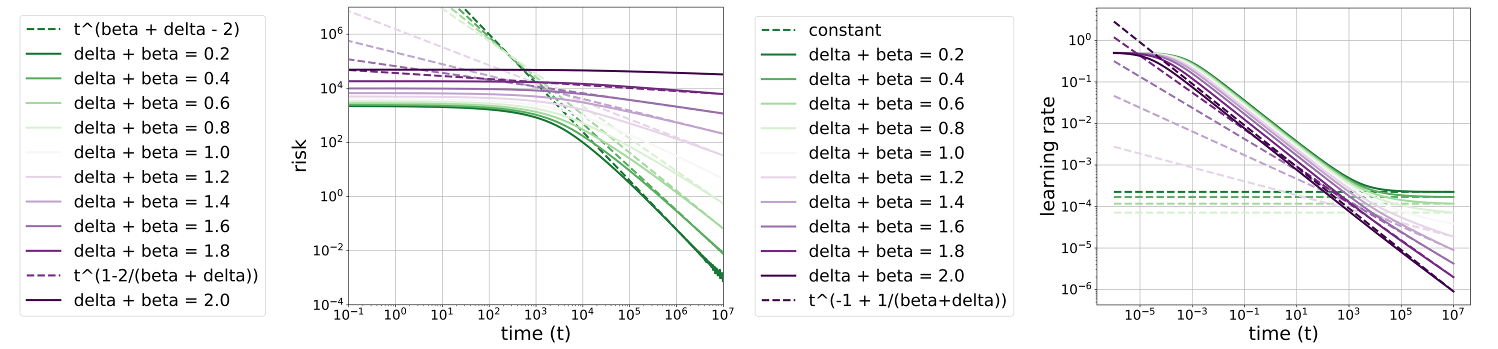
Power law covariance supported on $(0,1)$ at $d\to\infty$

When the spectrum of $\mathbf{K}$ and $\mathbf{x}^\star$ converge to the power law measures $\rho(\lambda)=(1-\beta)\lambda^{-\beta}\mathbb{1}_{(0,1)}$ and $\left((\mathbf{x}_0-\mathbf{x}^\star)^\top\omega_i\right)^2\sim\lambda_i^{-\delta}$, then, for all $t\ge1$,

$$\text{if }0<\beta+\delta<1,\quad\text{there exists }\tilde\gamma>0\text{ such that }\gamma_t^{\text{AdaGrad-Norm}}\ge\tilde\gamma$$

$$\text{if }\beta+\delta=1,\quad\gamma_t^{\text{AdaGrad-Norm}}\asymp_{\alpha,\beta}\frac{1}{\log(t+1)}$$

$$\text{if }1<\beta+\delta<2,\quad\gamma_t^{\text{AdaGrad-Norm}}\asymp_{\alpha,\beta}t^{-1+\frac{1}{\beta+\delta}}$$



Phase transition as $\delta+\beta$ varies. When $\delta+\beta<1$ *(green)*, the learning rate *(right)* is constant as $t\to\infty$. In contrast, when $2>\delta+\beta>1$ *(purple)*, the learning rate decreases at a rate $t^{-1+1/(\beta+\delta)}$ with $\delta+\beta=1$ (white) where the change occurs. Same phase transition occurs in the sublinear rate of the risk decay *(left)*.

Future Questions

- Can we extend our analysis to ...
  - **D-adaptation?**
  - **DoG?**
  - **RMSProp?**
- Conclusions about **catapult mechanism?**
- Can we generalize our theorem to ...
  - **non-Gaussian data?**
  - **non-convex problems?**
  - **different risk structures?**
- Analogous result for **multi-pass SGD?**

References

1 E. Collins-Woodfin, I. Seroussi, B. García Malaxechebarría, A. W. Mackenzie, E. Paquette, C. Paquette. *The High Line: Exact Risk and Learning Rate Curves of Stochastic Adaptive Learning Rate Algorithms*. Proceedings of NeurIPS, 2024.

2 E. Collins-Woodfin, C. Paquette, E. Paquette, I. Seroussi. *Hitting the High-dimensional notes: an ODE for SGD learning dynamics on GLMs and multi-index models*. IMA Inf. Inference, 13(4), 2024.

Arxiv: 2405.19585