GD and Large Linear Regression

Concentration and Asymptotics for a Spiked Model

Hugo Latourelle-Vigeant, Jiajun Yu, Nicolas Fertout

McGill Unversity, Summer 2021

Motivation: Curse of Dimensionality

Why are high dimensions so important?

Motivation: Curse of Dimensionality

Why are high dimensions so important?

What is unique about high-dimensional problems?

Motivation: Curse of Dimensionality

Why are high dimensions so important?

What is unique about high-dimensional problems?

$$
\begin{aligned}
\text{High-dimensional data} &\Leftrightarrow \text{large number of features and samples} \\
&\Leftrightarrow \text{sparseness in the data} \\
&\Leftrightarrow \text{fall in accuracy} \\
&\Leftrightarrow \text{deterioration in performance}
\end{aligned}
$$

Motivation: Curse of Dimensionality

Why are high dimensions so important?

What is unique about high-dimensional problems?

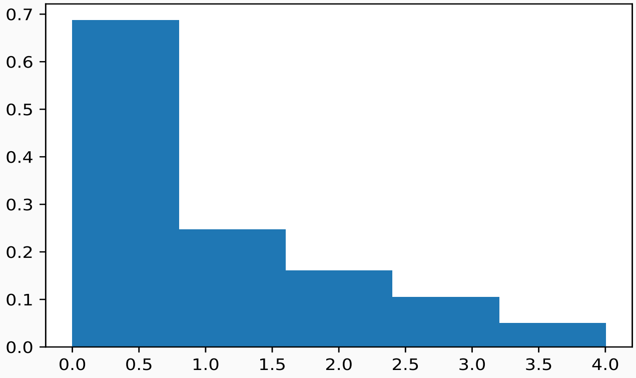
$$\begin{aligned} \text{High-dimensional data} &\Leftrightarrow \text{large number of features and samples} \\ &\Leftrightarrow \text{sparseness in the data} \\ &\Leftrightarrow \text{fall in accuracy} \\ &\Leftrightarrow \text{deterioration in performance} \end{aligned}$$

Solution: Random Matrix Theory!

Spectrum of Random Matrix

Let $\boldsymbol{M}$ be a 4000×4000 random matrix.

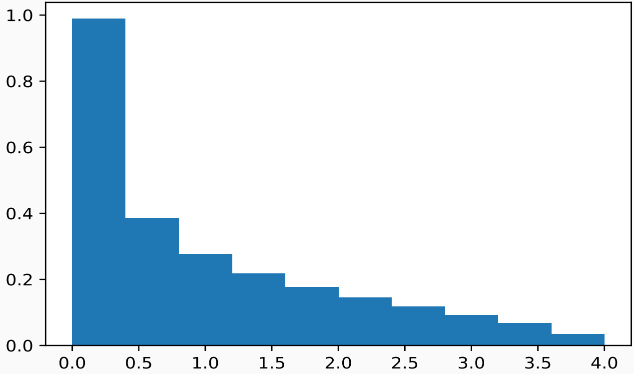
Spectra of $\frac{\boldsymbol{M}\boldsymbol{M}^T}{4000}$ with 5 bins



Spectrum of Random Matrix

Let $\boldsymbol{M}$ be a 4000×4000 random matrix.

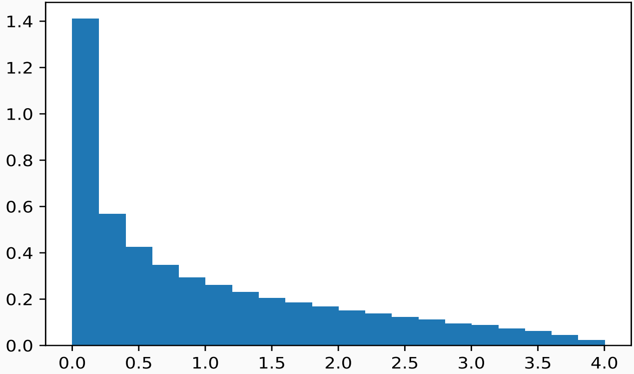
Spectra of $\frac{\boldsymbol{M}\boldsymbol{M}^T}{4000}$ with 10 bins



Spectrum of Random Matrix

Let $\boldsymbol{M}$ be a 4000×4000 random matrix.

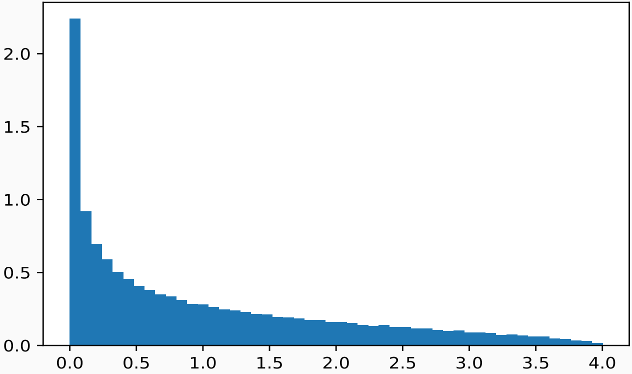
Spectra of $\frac{\boldsymbol{M}\boldsymbol{M}^T}{4000}$ with 20 bins



Spectrum of Random Matrix

Let $\boldsymbol{M}$ be a 4000×4000 random matrix.

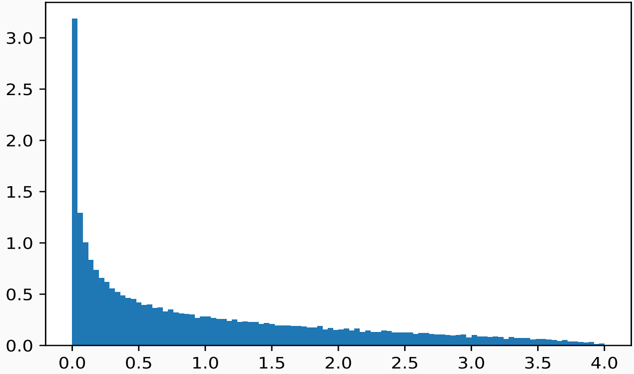
Spectra of $\frac{\boldsymbol{M}\boldsymbol{M}^T}{4000}$ with 50 bins



Spectrum of Random Matrix

Let $\boldsymbol{M}$ be a 4000×4000 random matrix.

Spectra of $\frac{\boldsymbol{M}\boldsymbol{M}^T}{4000}$ with 100 bins



Marchenko-Pastur

Marchenko-Pastur Law: $\mu_{MP}(\lambda) \stackrel{\text{def}}{=} \underbrace{\nu_{MP}(\lambda)}_{\textit{density}} + \underbrace{\omega_0 \delta_0(\lambda)}_{\textit{pointmass}}$

Marchenko-Pastur

Marchenko-Pastur Law: $\mu_{MP}(\lambda) \stackrel{\text{def}}{=} \underbrace{\nu_{MP}(\lambda)}_{\textit{density}} + \underbrace{\omega_0 \delta_0(\lambda)}_{\textit{pointmass}}$

- $\nu_{MP}(\lambda) \stackrel{\text{def}}{=} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi r \lambda}, \quad \lambda_\pm = \left(1 \pm \sqrt{r}\right)^2$

Marchenko-Pastur

Marchenko-Pastur Law: $\mu_{MP}(\lambda) \stackrel{\text{def}}{=} \underbrace{\nu_{MP}(\lambda)}_{\textit{density}} + \underbrace{\omega_0 \delta_0(\lambda)}_{\textit{pointmass}}$

- $\nu_{MP}(\lambda) \stackrel{\text{def}}{=} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi r \lambda}, \qquad \lambda_\pm = \left(1 \pm \sqrt{r}\right)^2$
- $\omega_0 \stackrel{\text{def}}{=} \max\left(0, 1 - \frac{1}{r}\right)$

Marchenko-Pastur

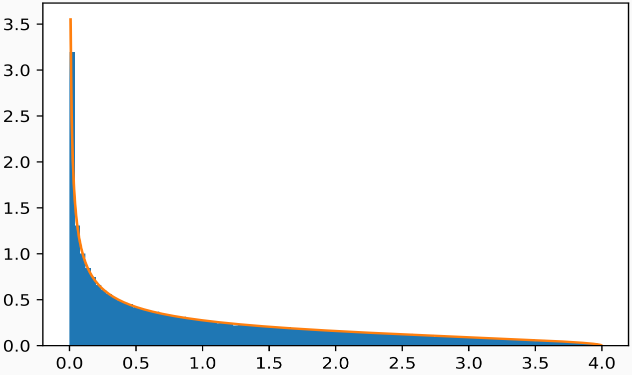
Marchenko-Pastur Law: $\mu_{MP}(\lambda) \stackrel{\text{def}}{=} \underbrace{\nu_{MP}(\lambda)}_{\textit{density}} + \underbrace{\omega_0 \delta_0(\lambda)}_{\textit{pointmass}}$

- $\nu_{MP}(\lambda) \stackrel{\text{def}}{=} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi r \lambda}, \quad \lambda_\pm = \left(1 \pm \sqrt{r}\right)^2$
- $\omega_0 \stackrel{\text{def}}{=} \max\left(0, 1 - \frac{1}{r}\right)$

Spectra of large random matrix $\Longrightarrow$ Marchenko-Pastur Law

Convergence to MP

Spectra of $\frac{\boldsymbol{MM}^T}{4000}$ vs Marchenko-Pastur Law (r=1)



MNIST

MNIST Databse

- Database of handwritten digits
- 60,000 training samples
- Each sample is 28 by 28

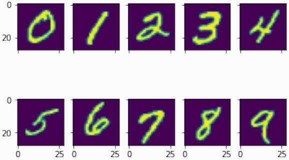
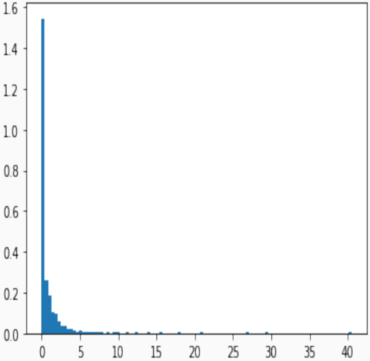


Figure 1: Example of MNIST samples

The database is (relatively) huge! What part of the data is interesting?

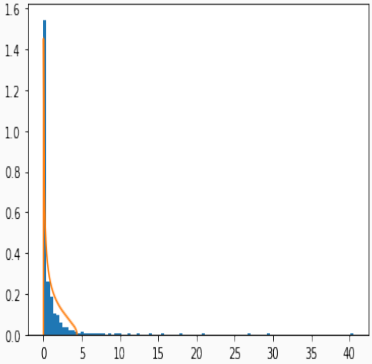
Spectral Distribution of MNIST



Let $\mathcal{M} \in \mathbb{R}^{60000 \times 784}$ represents the MNIST database.

- Plot the spectrum of $\frac{\mathcal{M}^T \mathcal{M}}{60000}$

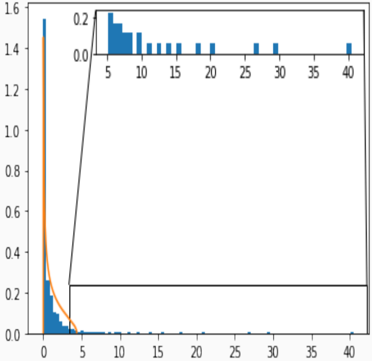
Spectral Distribution of MNIST



Let $\mathcal{M} \in \mathbb{R}^{60000 \times 784}$ represents the MNIST database.

- Plot the spectrum of $\frac{\mathcal{M}^T \mathcal{M}}{60000}$
- Overlap the Marchenko-Pastur Law

Spectral Distribution of MNIST



Let $\mathcal{M} \in \mathbb{R}^{60000\times 784}$ represents the MNIST database.

- Plot the spectrum of $\frac{\mathcal{M}^T\mathcal{M}}{60000}$
- Overlap the Marchenko-Pastur Law
- Large sets of data are interesting in the way they **do not** look random

Random least squares

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2}||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||_2^2, \quad \text{with data matrix } \boldsymbol{A} \in \mathbb{R}^{n\times d}, \quad \text{target vector } \boldsymbol{b} \in \mathbb{R}^n$$

Random least squares

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2}||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||_2^2, \quad \text{with data matrix } \boldsymbol{A} \in \mathbb{R}^{n\times d}, \quad \text{target vector } \boldsymbol{b} \in \mathbb{R}^n$$

What structure should we assume for $\boldsymbol{A}$ and $\boldsymbol{b}$ to accurately model real world large instances?

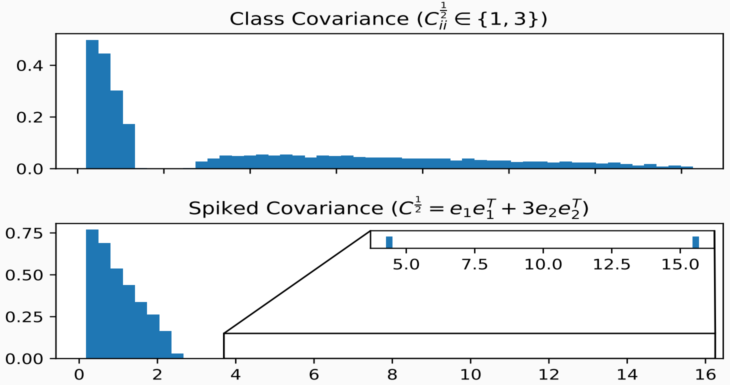
Data matrix

$\boldsymbol{A} \stackrel{\text{def}}{=} \boldsymbol{C}^{\frac{1}{2}} \boldsymbol{Z}$ is a sample covariance matrix

- $\boldsymbol{C} \succeq 0$ is $n \times n$ covariance matrix
- $\boldsymbol{Z}$ is $n \times d$ standardized random matrix
- independent features with covariance between samples

Covariance matrices

Many types of covariance matrices.



We will assume that $\boldsymbol{C} \in \mathbb{R}^{n \times n} = \boldsymbol{I}_n + \sum_{i=1}^{\xi} s_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ is a low rank perturbation of the identity ($\xi << n$).

Target vector

$$\boldsymbol{b} \stackrel{\text{def}}{=} \left(\boldsymbol{C}^{\frac{1}{2}} - \boldsymbol{I}_n\right) \boldsymbol{y} + \boldsymbol{\eta}$$

- $\boldsymbol{y}$ is a signal vector in the direction of the low rank perturbation
- $\boldsymbol{\eta}$ is a noise vector

Note: Since $\mathbb{E}[||\boldsymbol{\eta}||_2^2] = n$, we need $\mathbb{E}[||\boldsymbol{b}||_2^2] \approx n$ to be competitive!

Gradient descent

GD $\quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma \nabla f(\boldsymbol{x}_k)$

- Fixed step size γ
- Fixed initialization $\boldsymbol{x}_0 = 0$
- Guarantees convergence to local minimum

Set-up

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{2n}||\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}||_2^2, \quad \boldsymbol{A} \stackrel{\text{def}}{=} \boldsymbol{C}^{\frac{1}{2}}\boldsymbol{Z}, \quad \boldsymbol{b} \stackrel{\text{def}}{=} \left(\boldsymbol{C}^{\frac{1}{2}} - \boldsymbol{I}_n\right)\boldsymbol{y} + \boldsymbol{\eta}$$

Teacher-Student Model

- $\boldsymbol{A}$ is $n \times d$ data matrix
  - $\boldsymbol{Z}$ is $n \times d$ random matrix, standardized
  - $\boldsymbol{C} \stackrel{\text{def}}{=} \boldsymbol{I}_n + \sum_{i=1}^{\xi} s_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ is $n \times n$ a covariance matrix, $\boldsymbol{C} \succeq 0$
- $\boldsymbol{b}$ is a target vector
  - $\boldsymbol{y} \stackrel{\text{def}}{=} \sqrt{\frac{nR}{\xi}} \sum_{i=1}^{\xi} \boldsymbol{u}_i$ is a signal vector in the direction of the spikes, competitive in norm with the noise
  - $\boldsymbol{\eta}$ is noise vector
- $n =$ samples, $d =$ model size or features, $\frac{n}{d} \to r \in (0, \infty)$

Limiting empirical distribution

Spikeless model:

- Marchenko-Pastur law
- no stray eigenvalues

Spiked model:

- Limiting empirical distribution is still Marchenko-Pastur
- But there may be some stray eigenvalues

Spiked models

- Assume the 4th moment is finite
- Population eigenvalues within $[(1-\sqrt{r})^2, (1+\sqrt{r})^2]$ have no effect on the sample eigenvalues
- # stray sample eigenvalue = # population eigenvalues outside $[(1-\sqrt{r})^2, (1+\sqrt{r})^2]$

Theorem (Baik-Silverstein '06)

If $s_1 \geq s_2 \geq \cdots \geq s_\xi$ *and* $\lambda_1(\frac{1}{d}AA^T) \geq \lambda_2(\frac{1}{d}AA^T) \geq \cdots \geq \lambda_n(\frac{1}{d}AA^T)$,

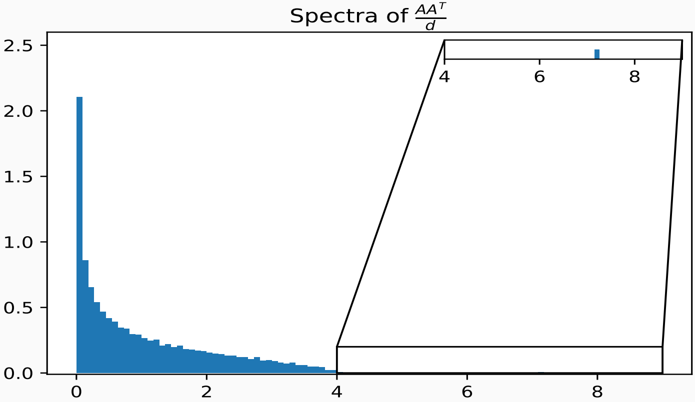
$$\lambda_i(\frac{1}{d}AA^T) \xrightarrow{a.s.} \begin{cases} \lambda_{loc}^{(i)} > \lambda_+, & s_i > \sqrt{r} \\ \lambda_+, & s_i \leq \sqrt{r} \end{cases}$$

for all $1 \leq i \leq \xi$.

Eigenvalues isolate from support of MP

Let $n = d = 2000$ $(r = 1)$, $s_1 = 0.5$, $s_2 = 1.5$. Then,

- $s_1 \leq r \implies$ **does not** isolate
- $s_2 > r \implies$ **does** isolates



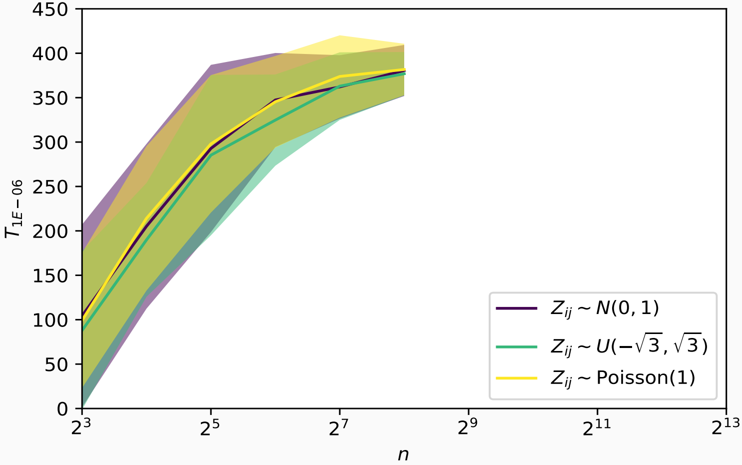
Main results: Concentration

Theorem (Fertout-Latourelle-Yu '21)

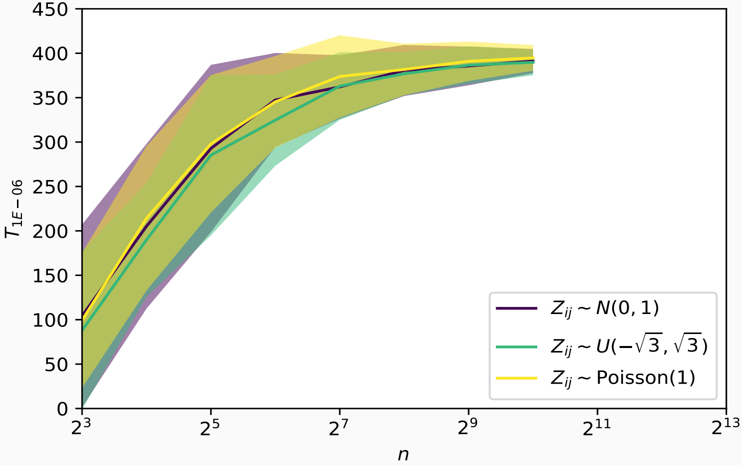
$$f(\boldsymbol{x}_k) \xrightarrow[n,d\to\infty]{\frac{n}{d}\to r\in\mathbb{R}_{>0}} \frac{1}{2}\int \left(1-\frac{\gamma}{r}\lambda\right)^{2k} \mu(d\lambda) + \int_{\mathbb{R}} \left(1-\frac{\gamma}{r}\lambda\right)^{2k} \mu_{MP}(d\lambda)$$

- $\mu(d\lambda) \overset{\text{def}}{=} \underbrace{\nu(\lambda)}_{\text{density}} d\lambda + \underbrace{\sum_{i=1}^{\xi} \omega_i \delta_{\lambda_{loc}^{(i)}}(\lambda)}_{\text{spikes}} + \underbrace{\omega_0\delta_0(\lambda)}_{\text{rank deficiency}}$
  - $\operatorname{supp}\nu = \operatorname{supp}\nu_{MP} = [\lambda_-, \lambda_+]$
  - $\omega_i \neq 0 \iff$ ith dominant eigenvalue separates from the support of *MP*
  - $\omega_0 \neq 0 \iff$ $\boldsymbol{A}$ is skinny ($\boldsymbol{AA}^T$ is rank deficient)
- μ_{MP} is the Marchenko-Pastur law

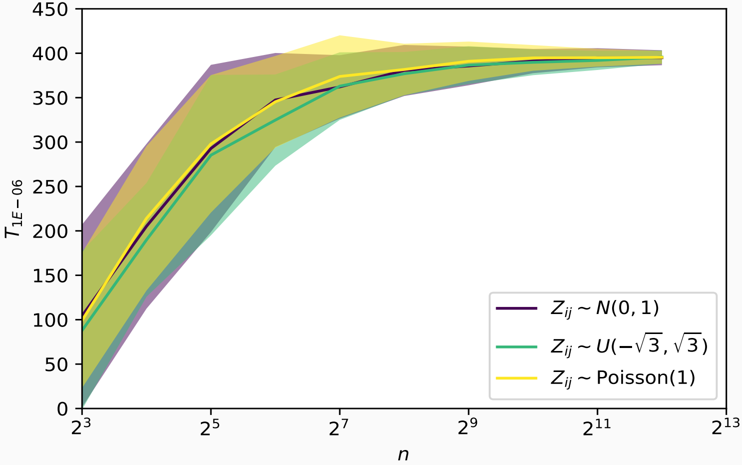
Halting time is predictable



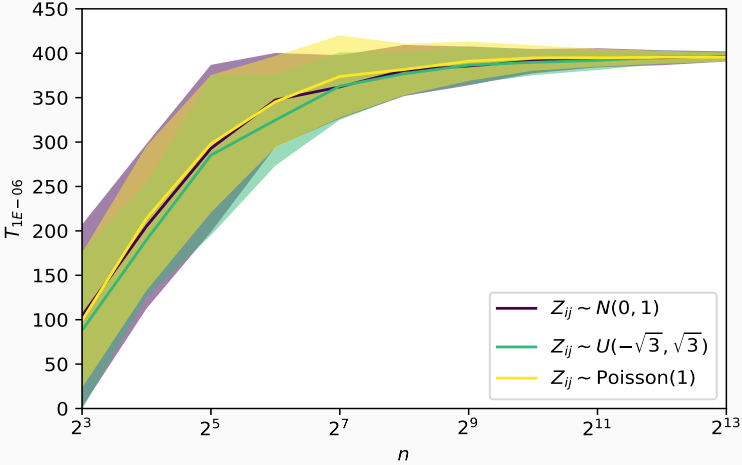
Halting time is predictable



Halting time is predictable



Halting time is predictable



We observe **universality**: halting time independent of distributions!

Main results: Asymptotics

Let $\Psi(k;\gamma) := \frac{1}{2}\int \left(1 - \frac{\gamma}{r}\lambda\right)^{2k} \mu(d\lambda) + \int_{\mathbb{R}} \left(1 - \frac{\gamma}{r}\lambda\right)^{2k} \mu_{MP}(d\lambda)$.

Theorem (Fertout-Latourelle-Yu '21)

If $r \neq 1$ (strongly convex) and $0 < \gamma \leq \frac{r}{\lambda_+ + \frac{1}{2}\max\{0, s_{max} - (3+2\sqrt{2})\sqrt{r}\}}$

$$\Psi(k;\gamma) - \Psi^\star \sim \rho \frac{\sqrt{r(\lambda_+ - \lambda_-)}}{16\sqrt{2\pi}\lambda_-\gamma^{\frac{3}{2}} k^{\frac{3}{2}}} \left(1 - \frac{\gamma}{r}\lambda_-\right)^{2k+\frac{3}{2}}$$

for some $\rho \in \mathbb{R}_{>0}$.

Theorem (Fertout-Latourelle-Yu '21)

If $r = 1$ (non strongly convex) and $0 < \gamma \leq \frac{2}{4+\max\{0, s_{\max} - 1\}}$,

$$\Psi(k;\gamma) - \Psi^\star \sim \rho \frac{1}{2\sqrt{2\pi\gamma k}}$$

for some $\rho \in \mathbb{R}_{>0}$.

Effects of r and spike strengths on halting time

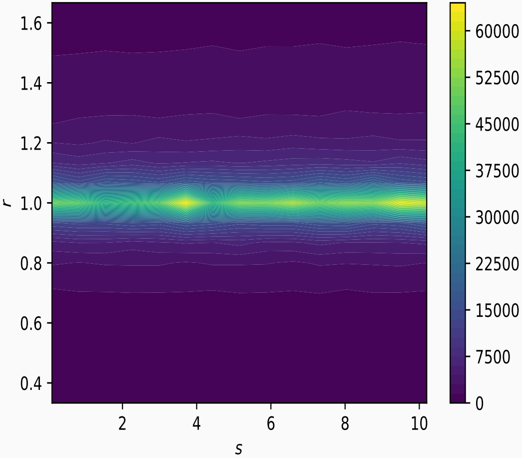


Figure 2: Halting time ($T_{0.001}$) w.r.t. s and r

Effects of r and spike strengths on halting time

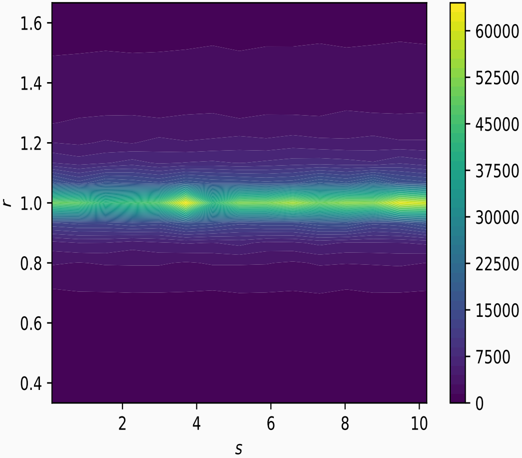


Figure 2: Halting time ($T_{0.001}$) w.r.t. s and r

- $r \approx 1 \implies$ slower

Effects of r and spike strengths on halting time

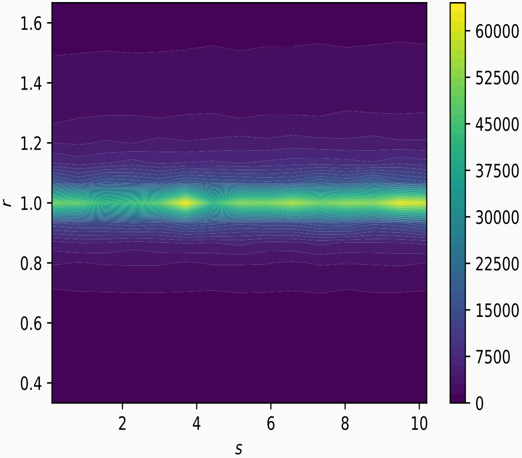


Figure 2: Halting time ($T_{0.001}$) w.r.t. s and r

- $r \approx 1 \implies$ slower
- Strong spikes $\implies$ slower

Effects of r and spike strengths on halting time

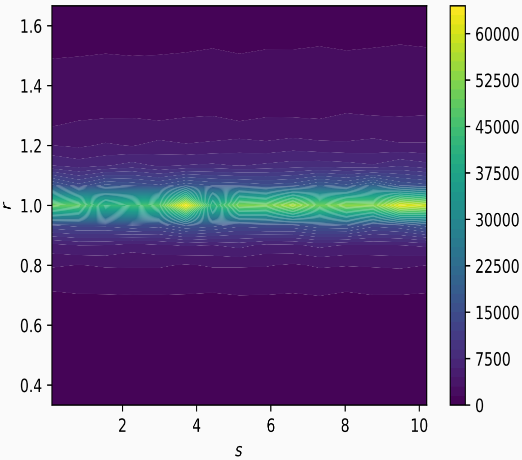


Figure 2: Halting time ($T_{0.001}$) w.r.t. s and r

- $r \approx 1 \implies$ slower
- Strong spikes $\implies$ slower
- **Caveat**: Step size depend on r AND spikes
  - $\uparrow$spike $\implies \downarrow \gamma \implies \downarrow$speed
  - $\downarrow r \implies \downarrow \gamma \implies \downarrow$speed

Possible extensions

Open Problems

- Analyzing feature covariance models that may more accurately represent some real world instances
- Extensions beyond gradient descent to other algorithm

The end!

C. Paquette, B. van Merriënboer, E. Paquette, F. Pedregosa. *Halting Time is Predictable for Large Models: Universality Property and Average-case Analysis*, arxiv.org/pdf/2006.04299.pdf

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 (accepted at COLT 2021)

Z. Liao. *A random matrix framework for large dimensional machine learning and neural networks* , tel.archives-ouvertes.fr/tel-02397287/document