Optimization Algorithms in the Large

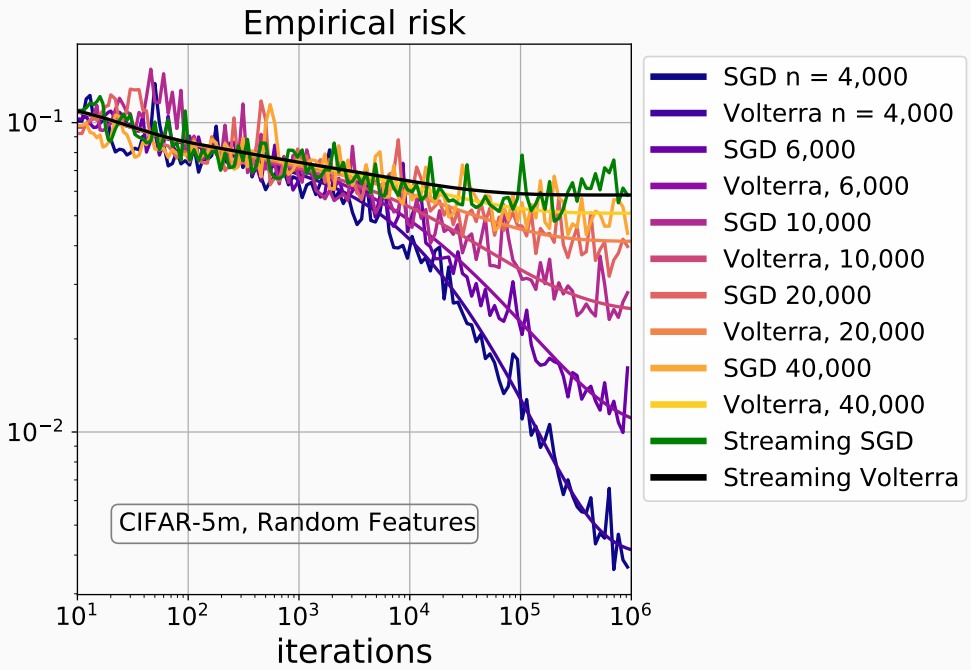
Exact Dynamics, Average-case Analysis, and Stepsize Criticality

Courtney Paquette

Joint work: Fabian Pedregosa, Elliot Paquette, Bart van Merriënboer, Kiwon Lee, Jeffrey Pennington, Ben Adlam, and Andrew Cheng

Erice, May 2022

Theory meets practice: CIFAR-5m



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 Problem

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

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 Problem

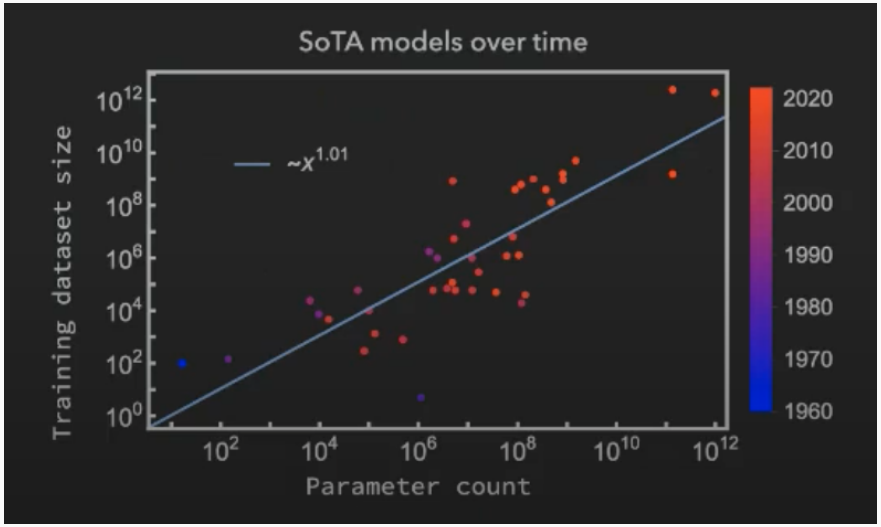
$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

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) are proportional



Jeffrey Pennington, data from "Parameter Counts in Machine Learning", lesswrong.com, June 2021

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: 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)

Stochastic algorithms and random least squares

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2 + \frac{\delta}{2}\|\boldsymbol{x}\|^2 = \min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}) \overset{\text{def}}{=} \sum_{i=1}^{n} \underbrace{\frac{1}{2}(\boldsymbol{a}_i\boldsymbol{x}-b_i)^2 + \frac{\delta}{2n}\|\boldsymbol{x}\|^2}_{f_i(\boldsymbol{x})} \right\},$$

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

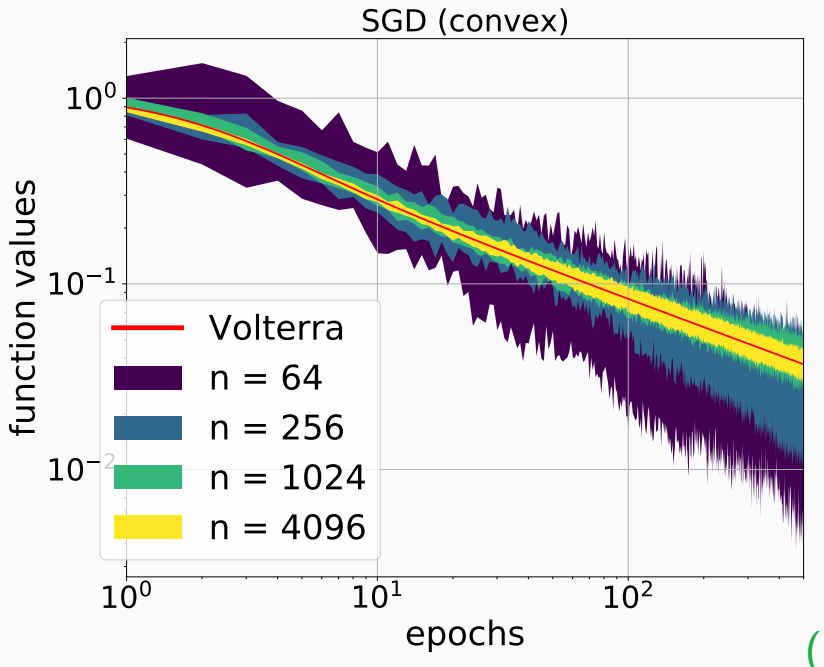
Stochastic Gradient Descent (SGD) $\quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma(t)\nabla f_{i_k}(\boldsymbol{x}_k)$

Stochastic algorithms and random least squares

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax-b\|^2+\frac{\delta}{2}\|x\|^2 = \min_{x\in\mathbb{R}^d}\left\{ f(x) \overset{\text{def}}{=} \sum_{i=1}^{n} \underbrace{\frac{1}{2}(a_i x - b_i)^2 + \frac{\delta}{2n}\|x\|^2}_{f_i(x)} \right\},$$

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

Stochastic Gradient Descent (SGD) $\quad x_{k+1} = x_k - \gamma(t)\nabla f_{i_k}(x_k)$



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

- $f(x_k) \xrightarrow{\Pr}$(smooth function)
- Analyze this smooth function
- Determined by the spectrum of the Hessian

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

Data matrices A

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2 \qquad \text{Hessian } \boldsymbol{H}=\boldsymbol{A}^T\boldsymbol{A}$$

Distributions on data matrices A and Random matrices

Many classes of random matrices share characteristics which depend on the low moments of their entries, called **Universality**

Example

- **De-localization of eigenvectors of $\boldsymbol{H}$**: eigenvectors are not aligned with the unit vectors

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

Detour into random matrix theory...

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 λ_{max} and λ_{min} of $\boldsymbol{H}$, $\|\boldsymbol{H}\|_2 < C$

Detour into random matrix theory...

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 $\boldsymbol{H}$, $\|\boldsymbol{H}\|_2 < C$

3. De-localization of eigenvectors of $\boldsymbol{H}$:
   Ω contour enclosing eigenvalues of $\boldsymbol{H}$, $\theta < 1/2$
   (i) $\max_{z\in\Omega} \max_{1\leq i\neq j\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_j^T| \leq n^{\theta-1/2}$.
   (ii) $\max_{z\in\Omega} \max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i - \frac{1}{n}\operatorname{tr} R(z; \boldsymbol{A}\boldsymbol{A}^T)| \leq n^{\theta-1/2}$.
   (iii) Examples
   - Isotropic features. Entries of $\boldsymbol{A} \sim N(0,1)$
   - Sample covariance matrices. independent samples w/ covariance between features
   - Random features. $\boldsymbol{A} = \sigma(\boldsymbol{X}\boldsymbol{W})$ where σ is an activation function

Our framework

$$\min_{x \in \mathbb{R}^d} f(x) \stackrel{\text{def}}{=} \underbrace{\frac{1}{2}\|Ax - b\|^2}_{\mathcal{L}(x)} + \frac{\delta}{2}\|x\|^2, \quad \mathcal{L}(x) \stackrel{\text{def}}{=} \frac{1}{2}\|Ax - b\|^2$$

Model

- (random) initialization $x_0 \in \mathbb{R}^d$ is $\|x_0\|^2 \leq R$ (across dimensions)
- b is target vector, $\|b\|^2 \leq \widetilde{R}$ (across dimensions)
- b and x_0 are "independent" of the eigenvectors of A

Our framework

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \stackrel{\text{def}}{=} \underbrace{\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2}_{\mathcal{L}(\boldsymbol{x})} + \frac{\delta}{2}\|\boldsymbol{x}\|^2, \quad \mathcal{L}(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2$$

Model

- (random) initialization $\boldsymbol{x}_0 \in \mathbb{R}^d$ is $\|\boldsymbol{x}_0\|^2 \leq R$ (across dimensions)
- $\boldsymbol{b}$ is target vector, $\|\boldsymbol{b}\|^2 \leq \widetilde{R}$ (across dimensions)
- $\boldsymbol{b}$ and $\boldsymbol{x}_0$ are "independent" of the eigenvectors of $\boldsymbol{A}$

Algorithmic considerations

- Small batch: batch size $\to 0$ as $n \to \infty$
- Multi-pass SGD
- Step size: bounded, continuous function s.t. $\gamma(t) \to \widetilde{\gamma}$

  (if $\tilde{\gamma} = 0$, then $\int_0^\infty \gamma(s)\,\mathrm{d}s = \infty$, Robbins-Monro conditions)

Exact Dynamics Idea: Diffusion Approximation

Homogenized SGD (**C.P.**-E. Paquette, NeurIPS '21 & Mori '21)

$$\mathrm{d}\boldsymbol{X}_t = -\gamma(t)\nabla f(\boldsymbol{X}_t)\,\mathrm{d}t + \sqrt{\tfrac{2}{n}\mathcal{L}(\boldsymbol{X}_t)\nabla^2\mathcal{L}(\boldsymbol{X}_t)}\,\mathrm{d}\boldsymbol{B}_t$$

$\boldsymbol{X}_0 = \boldsymbol{x}_0$ and $(\boldsymbol{B}_t : t \geq 0)$ is a d-dimen. standard Brownian motion

- New diffusion process (c.f. Li et al., Mandt et al.):
  - Dimension $n \to \infty$ instead of stepsize $\gamma \to 0$ to create it
  - Explicitly solvable and don't need to send stepsize to 0

Exact Dynamics Idea: Diffusion Approximation

Homogenized SGD (**C.P.**-E. Paquette, NeurIPS '21 & Mori '21)

$$\mathrm{d}\boldsymbol{X}_t = -\gamma(t)\nabla f(\boldsymbol{X}_t)\,\mathrm{d}t + \sqrt{\tfrac{2}{n}\mathcal{L}(\boldsymbol{X}_t)\nabla^2\mathcal{L}(\boldsymbol{X}_t)}\,\mathrm{d}\boldsymbol{B}_t$$

$\boldsymbol{X}_0 = \boldsymbol{x}_0$ and $(\boldsymbol{B}_t : t \geq 0)$ is a d-dimen. standard Brownian motion

- New diffusion process (c.f. Li et al., Mandt et al.):
  - Dimension $n \to \infty$ instead of stepsize $\gamma \to 0$ to create it
  - Explicitly solvable and don't need to send stepsize to 0

Theorem: High dimensional equivalence of SGD

(**C.P.**-E. Paquette-B. Adlam-J.Pennington)

For any quadratic statistic q,

$$\Pr\Big(\sup_{0\leq t\leq T} |q(\boldsymbol{x}_t) - q(\underbrace{\boldsymbol{X}_t}_{\text{diffusion}})| > d^{-C}\Big) \leq d^{-D}$$

concentration around mean: $q(\underbrace{\boldsymbol{X}_t}_{\text{diffusion}}) \to \mathbb{E}\left[q(\underbrace{\boldsymbol{X}_t}_{\text{diffusion}}) \mid \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{x}_0\right]$

Dynamics of SGD: Loss function

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

Let t = # of passes through data. If $\gamma(s) \to \widetilde{\gamma} < 2\left(\frac{1}{n}\mathrm{tr}((\boldsymbol{H})^2(\boldsymbol{H}+\delta \boldsymbol{I})^{-1})\right)^{-1}$,

$$\Pr\Big(\sup_{0\le t\le T}\left|\mathcal{L}\big(\boldsymbol{x}_{\lfloor nt\rfloor}\big)-\psi_t\right| > d^{-C}\Big) \le d^{-D}$$

where ψ_t is solution to a **Volterra equation**

$$\psi_t = \underbrace{\mathcal{L}(\boldsymbol{X}^{\mathrm{gf}}_{\Gamma(t)})}_{\text{gradient flow}} + \int_0^t \underbrace{\gamma^2(s) r h_2(t,s)\psi_s}_{\text{noise term}}\,\mathrm{d}s$$

- $h_2(t,s) = \frac{1}{d}\sum_{i=1}^d \lambda_i^2 e^{-2(\Gamma(t)-\Gamma(s))(\lambda_i+\delta)}$, $\quad \lambda_i$ eigenvalues of $\boldsymbol{H}$
- Integrated learning $\Gamma(t) = \int_0^t \gamma(s)\,\mathrm{d}s$

Dynamics of SGD: Loss function

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

Let t = # of passes through data. If $\gamma(s) \to \widetilde{\gamma} < 2\left(\frac{1}{n}\mathrm{tr}((\boldsymbol{H})^2(\boldsymbol{H}+\delta \boldsymbol{I})^{-1})\right)^{-1}$,

$$\Pr\Big(\sup_{0\le t\le T}\left|\mathcal{L}\big(\boldsymbol{x}_{\lfloor nt\rfloor}\big)-\psi_t\right| > d^{-C}\Big) \le d^{-D}$$

where ψ_t is solution to a **Volterra equation**

$$\psi_t = \underbrace{\mathcal{L}(\boldsymbol{X}^{\mathrm{gf}}_{\Gamma(t)})}_{\text{gradient flow}} + \int_0^t \underbrace{\gamma^2(s) r h_2(t,s)\psi_s}_{\text{noise term}}\,\mathrm{d}s$$

- $h_2(t,s) = \frac{1}{d}\sum_{i=1}^d \lambda_i^2 e^{-2(\Gamma(t)-\Gamma(s))(\lambda_i+\delta)}$, $\quad \lambda_i$ eigenvalues of $\boldsymbol{H}$
- Integrated learning $\Gamma(t) = \int_0^t \gamma(s)\,\mathrm{d}s$

Rate of Convergence: Competition

- For small γ, same rate as gradient descent
- For large γ, noise term dominates rate

Phase transition & Asymptotics: Fixed Stepsize

Critical stepsize

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

Theorem

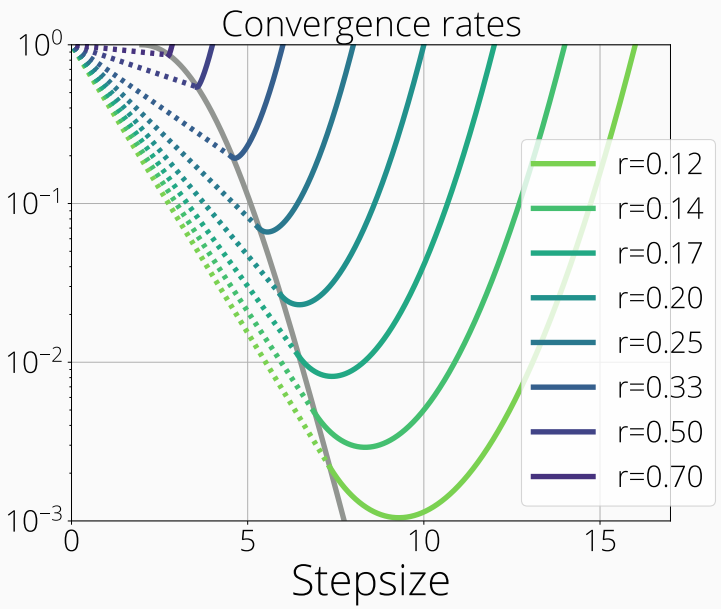
(**C.P.**-Lee-E. Paquette-Pedregosa, COLT '21)

For small $\gamma < \gamma_*$,

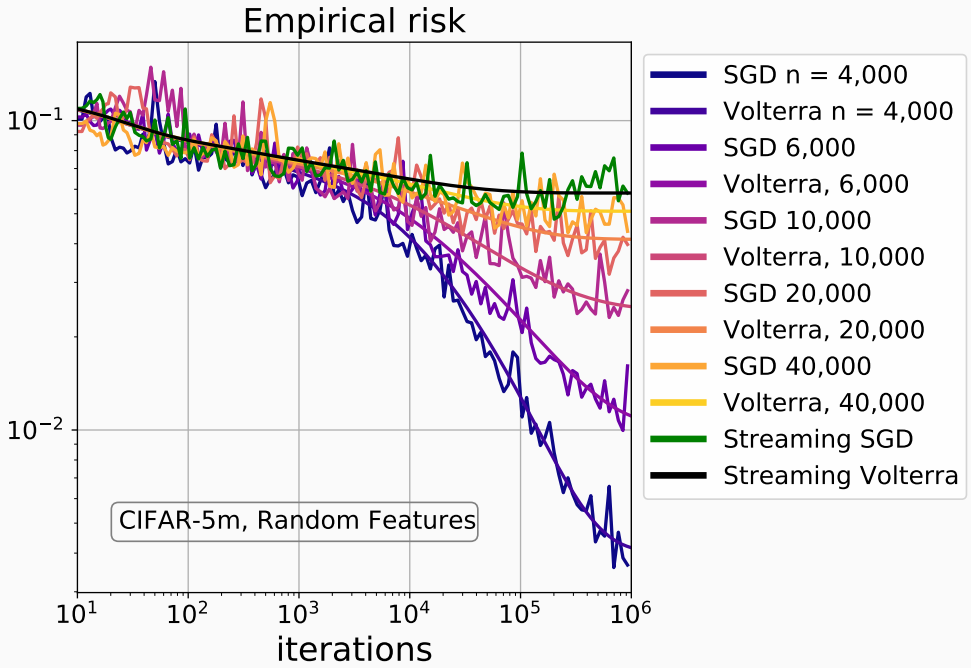
$$\psi_{\boldsymbol{H}}(t) - \psi_{\boldsymbol{H}}(\infty) \sim \frac{1}{t^\alpha} e^{-2\gamma t \lambda_{\min}}.$$

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

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



Real world predictions: CIFAR-5m



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.

Statistic: Expected risk

$$\mathcal{R}(\boldsymbol{x}_t) = \frac{1}{2}\mathbb{E}[(b - \boldsymbol{x}_t^T \boldsymbol{a})^2 | \boldsymbol{x}_t] \quad \text{where} \quad (\boldsymbol{a}, b) \sim \mathcal{D}$$

Theorem (**C.P.**-Adlam-E. Paquette-Pennington)

Let t = # of passes through data. If $\gamma(t) \to \widetilde{\gamma} < 2\left(\frac{1}{n}\text{tr}((\boldsymbol{H})^2(\boldsymbol{H} + \delta \boldsymbol{I})^{-1})\right)^{-1}$,

$$\Pr\left(\sup_{0 \le t \le T} |\mathcal{R}(\boldsymbol{x}_{\lfloor tn \rfloor}) - \Omega_t| > d^{-C}\right) \le d^{-D}$$

where Ω_t equals

$$\Omega_t = \underbrace{\mathcal{R}(\boldsymbol{X}^{\text{gf}}_{\Gamma(t)})}_{\text{gradient flow}} + \int_0^t \underbrace{\gamma^2(s) \cdot K(t, s, \nabla^2 \mathcal{R}) \cdot \psi_s}_{\text{noise term}} \, \mathrm{d}s,$$

and ψ_s limiting loss function $\mathcal{L}$, $K(t, s, \nabla^2 \mathcal{R})$ explicit kernel

Statistic: Expected risk

$$\mathcal{R}(\boldsymbol{x}_t) = \frac{1}{2}\mathbb{E}[(b - \boldsymbol{x}_t^T \boldsymbol{a})^2 | \boldsymbol{x}_t] \quad \text{where} \quad (\boldsymbol{a}, b) \sim \mathcal{D}$$

Theorem (**C.P.**-Adlam-E. Paquette-Pennington)

Let t = # of passes through data. If $\gamma(t) \to \widetilde{\gamma} < 2\left(\frac{1}{n}\mathrm{tr}((\boldsymbol{H})^2(\boldsymbol{H} + \delta \boldsymbol{I})^{-1})\right)^{-1}$,

$$\Pr\Big(\sup_{0 \le t \le T} |\mathcal{R}(\boldsymbol{x}_{\lfloor tn \rfloor}) - \Omega_t| > d^{-C}\Big) \le d^{-D}$$

where Ω_t equals

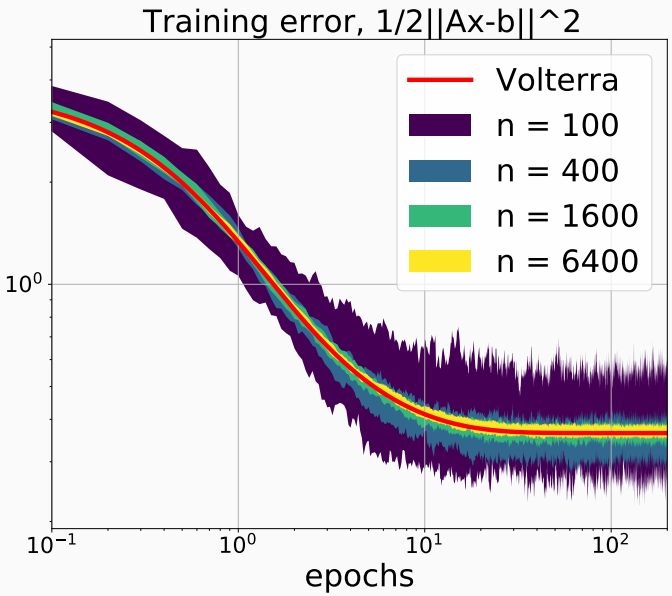
$$\Omega_t = \underbrace{\mathcal{R}(\boldsymbol{X}^{\mathrm{gf}}_{\Gamma(t)})}_{\text{gradient flow}} + \int_0^t \underbrace{\gamma^2(s) \cdot K(t, s, \nabla^2 \mathcal{R}) \cdot \psi_s}_{\text{noise term}} \, \mathrm{d}s,$$

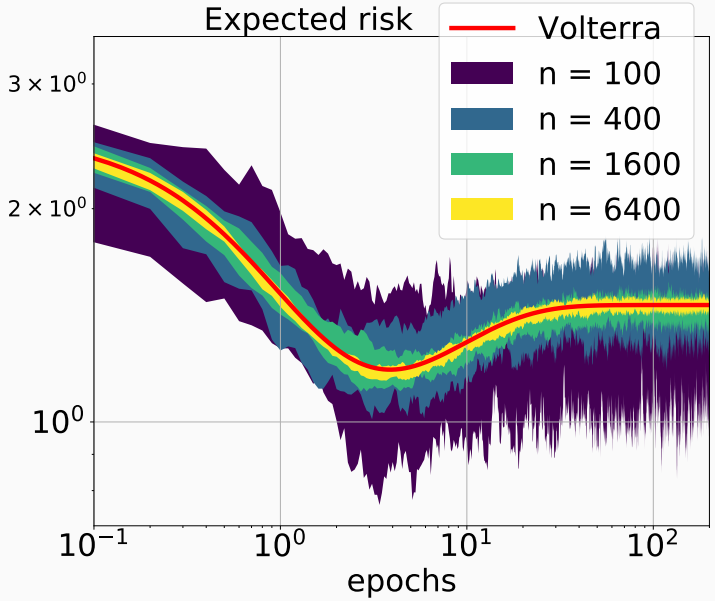
and ψ_s limiting loss function $\mathcal{L}$, $K(t, s, \nabla^2 \mathcal{R})$ explicit kernel

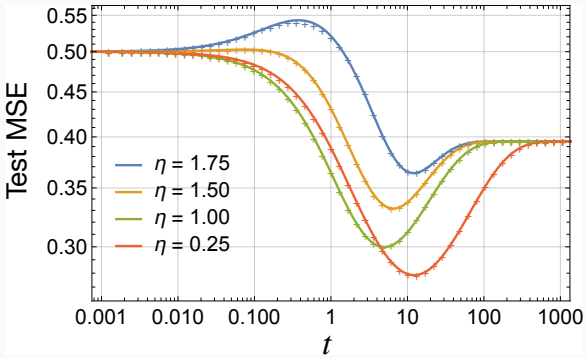
Consequences

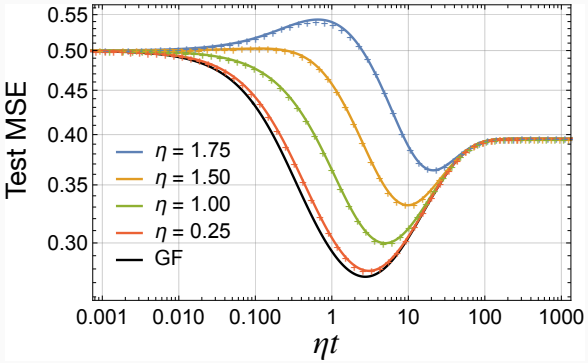
- Expected risk of SGD $\geq$ Expected risk of gradient flow
  $\Rightarrow$ no implicit regularization
- If $\mathcal{L}(\boldsymbol{x}_t) \to 0$ or $\gamma(t) \to 0$, recover gradient flow at $t \to \infty$

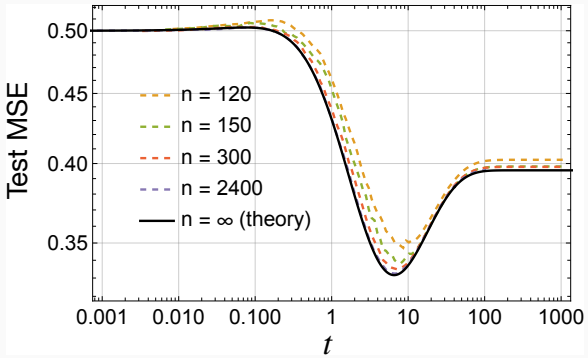
Concentration effect



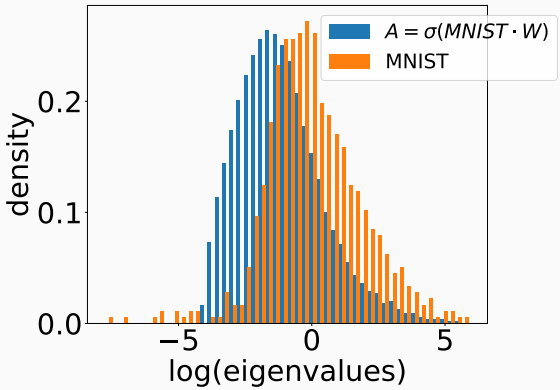


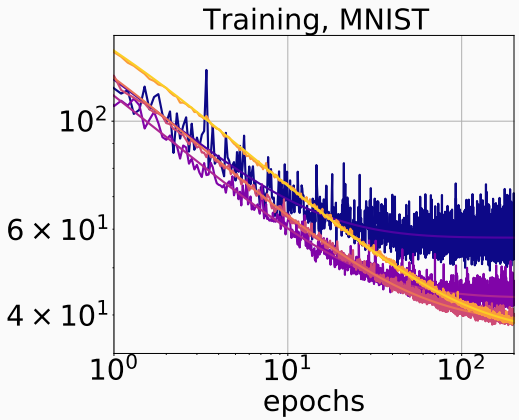


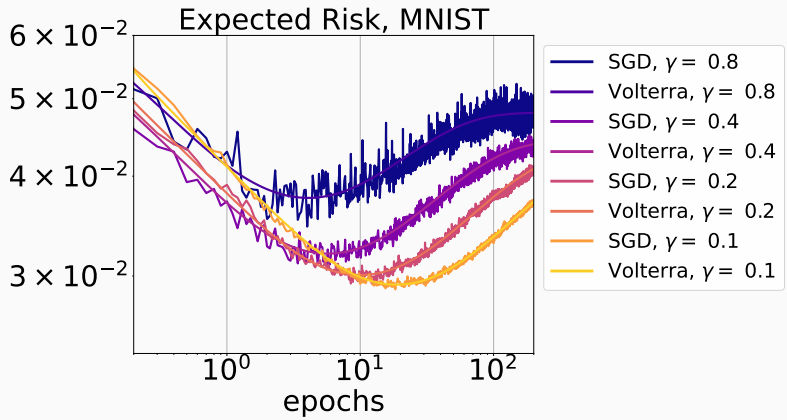




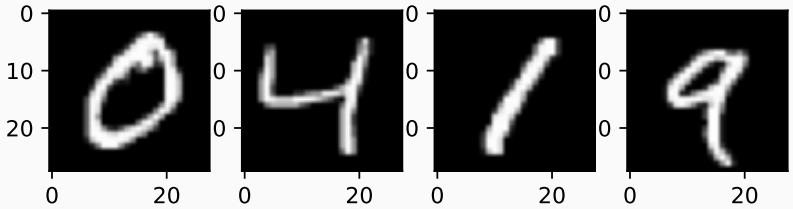
Real world predictions: MNIST







Using a random features model to predict MNIST digits, the **Volterra equation** (using the Hessian spectra an input) gives **good predictions** for behavior of SGD.



Question: Can you go faster?

SGD+M vanishing batch size

Stochastic gradient descent with momentum (SGD+M)

$$\boldsymbol{y}_k = (1-\theta)\boldsymbol{y}_{k-1} + \boldsymbol{\Gamma}_1 \nabla f_{i_k}(\boldsymbol{x}_k)$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{y}_k$$

✓ $\boldsymbol{\Gamma}_1$ step size ✓ $1-\theta$ momentum parameter

Homogenized SGD (Diffusion Process) (**C.P.**-E. Paquette, NeurIPS '21)

$$\mathrm{d}\boldsymbol{X}_t = -\boldsymbol{\Gamma}_1 \int_0^t e^{-n\theta(t-s)}\, \mathrm{d}\boldsymbol{Z}_t$$

$$\text{where } \mathrm{d}\boldsymbol{Z}_t = \nabla f(\boldsymbol{X}_t)\,\mathrm{d}t + \sqrt{\tfrac{2}{n}\mathcal{L}(\boldsymbol{X}_t)\nabla^2\mathcal{L}(\boldsymbol{X}_t)}\,\mathrm{d}\boldsymbol{B}_t$$

SGD+M vanishing batch size

Stochastic gradient descent with momentum (SGD+M)

$$\boldsymbol{y}_k = (1-\theta)\boldsymbol{y}_{k-1} + \boldsymbol{\Gamma}_1 \nabla f_{i_k}(\boldsymbol{x}_k)$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{y}_k$$

✓ $\boldsymbol{\Gamma}_1$ step size ✓ $1-\theta$ momentum parameter

Homogenized SGD (Diffusion Process) (**C.P.**-E. Paquette, NeurIPS '21)

$$\mathrm{d}\boldsymbol{X}_t = -\boldsymbol{\Gamma}_1 \int_0^t e^{-n\theta(t-s)}\, \mathrm{d}\boldsymbol{Z}_t$$

$$\text{where } \mathrm{d}\boldsymbol{Z}_t = \nabla f(\boldsymbol{X}_t)\,\mathrm{d}t + \sqrt{\tfrac{2}{n}\mathcal{L}(\boldsymbol{X}_t)\nabla^2\mathcal{L}(\boldsymbol{X}_t)}\,\mathrm{d}\boldsymbol{B}_t$$

Under **homogenized SGD**, convolution-type Volterra equation $\psi_{\boldsymbol{H}}$ (**C.P.**-E. Paquette, NeurIPS '21):

$$\psi_t = F(t) + \int_0^t \mathcal{I}(t-s)\psi_s\, \mathrm{d}s.$$

Practical considerations: SGD+M = SGD and SDAHB $\neq$ SGD

Experiment

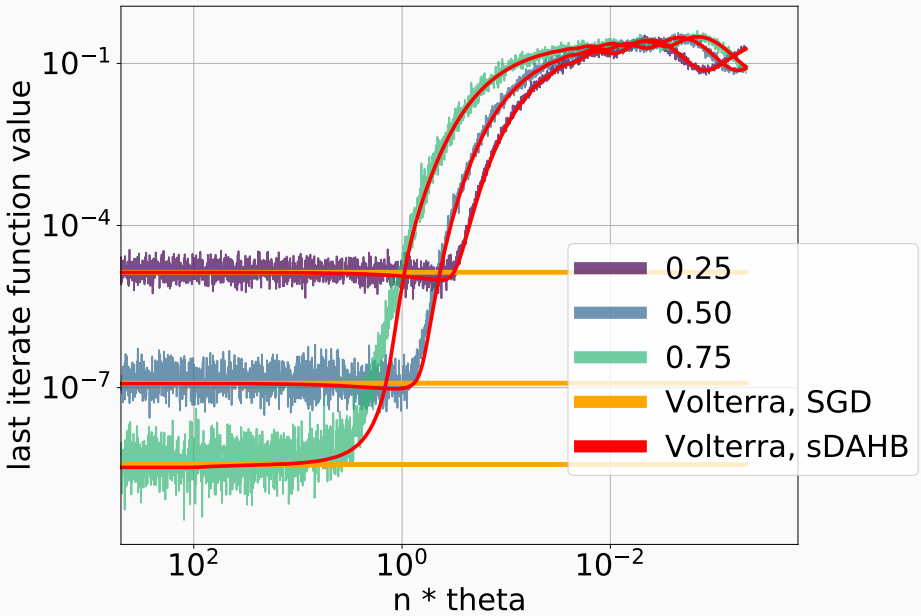
- Fix matrix $\boldsymbol{A}$; Run SGD+M for a fixed time and output function value
- Vary the θ and γ holding $\frac{\gamma}{\theta} = c$ fixed; $c \in \{0.25, 0.5, 0.75\}$

Conclusions

- For fixed θ, n large,

$$\text{SGD+M}(\gamma, \theta) = \text{SGD}(\gamma/\theta)$$

- $\theta \sim 1/n$ barely faster than SGD, but not equivalent



Batching + SGD+M

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

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

SGD+M with batches

$$\boldsymbol{y}_k = \Delta \cdot \boldsymbol{y}_{k-1} + \gamma \cdot \zeta \sum_{i_k \in B} \nabla f_{i_k}(\boldsymbol{x}_k)$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{y}_k$$

Batching + SGD+M

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

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

SGD+M with batches

$$\boldsymbol{y}_k = \Delta \cdot \boldsymbol{y}_{k-1} + \gamma \cdot \zeta \sum_{i_k \in B} \nabla f_{i_k}(\boldsymbol{x}_k)$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{y}_k$$

Results

- When $\zeta \to 0$ as $n \to \infty$, SGD + M $\equiv$ SGD
- When $\zeta = 1$, GD + M is faster than GD
  - $\mathcal{O}(\sqrt{\kappa})$ versus $\mathcal{O}(\kappa)$, $\kappa = \frac{\lambda_{\max}(\boldsymbol{H})}{\lambda_{\min}(\boldsymbol{H})}$

What happens in between?

Concentration of SGD+M with batches

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

If $\gamma < \min\left\{ \frac{1+\Delta}{\zeta \lambda_{\max}(\boldsymbol{H})}, \frac{n(1-\Delta)}{(1-\zeta)} \mathrm{tr}(\boldsymbol{H})^{-1} \right\}$,

$$\Pr\Big(\sup_{0 \le t \le T} \big| \mathcal{L}(\boldsymbol{x}_t) - \psi(t) \big| > d^{-D} \Big) \le d^{-C}$$

where $\psi(t)$ is solution to **(discrete) convolution-type Volterra**

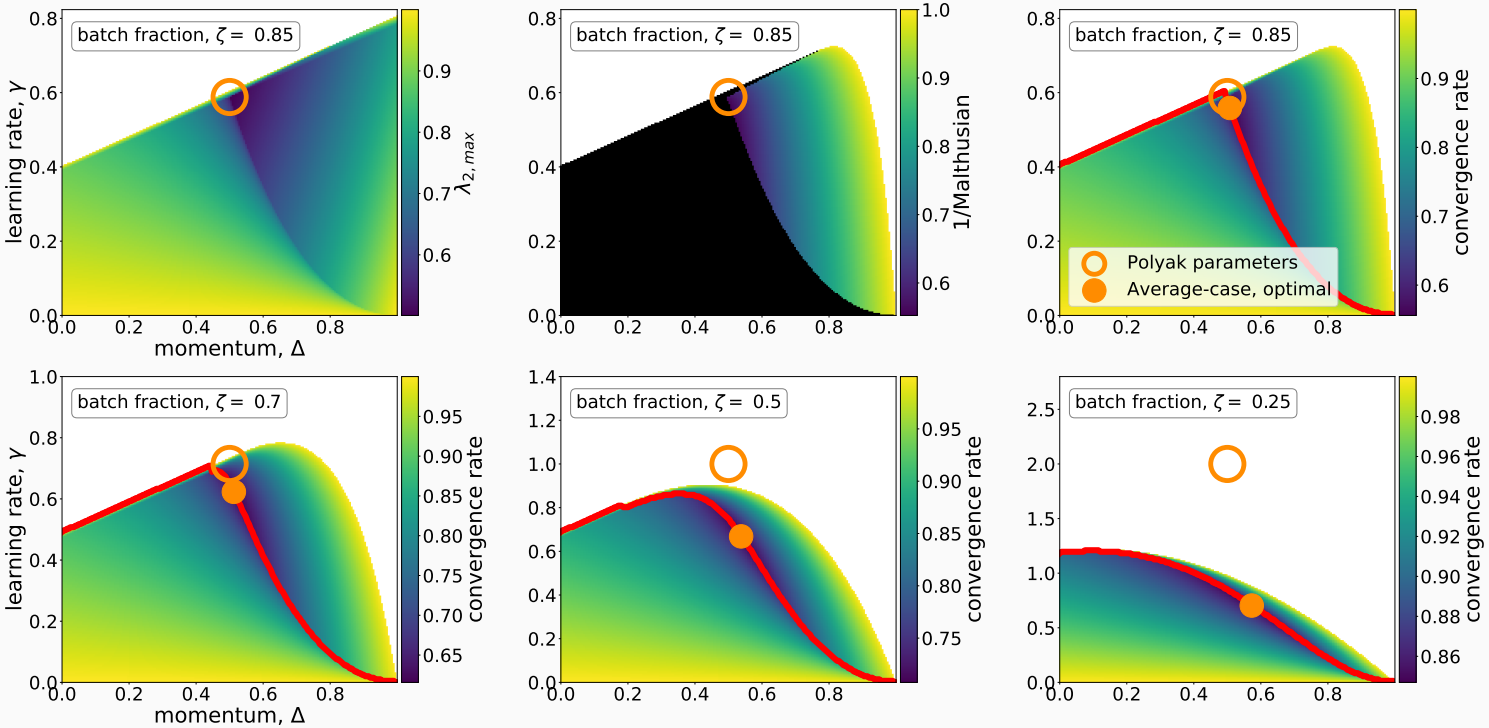
$$\psi(t+1) = \mathcal{L}(\boldsymbol{X}^{\mathrm{gd+M}}) + \sum_{k=0}^{t} \underbrace{\gamma^2 \zeta (1-\zeta) H(t-k) \psi(k)}_{\text{noise}}$$

- H(t-k) explicit function of eigenvalues of $\boldsymbol{H}$
- Λ, rate at which $\mathcal{L}(\boldsymbol{X}^{\mathrm{gd+M}})$ decrease
- For some γ and Δ, noise term slows you down

Convergence of SGD+M with batches

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

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



Large vs Small batch: Convergence

$$\gamma = \frac{(1-\sqrt{\Delta})^2}{\zeta \lambda_{\min}}, \qquad \Delta = \max\left\{ \left(1 - \frac{\zeta}{(1-\zeta)\bar{\kappa}}\right)^2, \left(1 - \frac{1}{\sqrt{\kappa}}\right)^2 \right\}$$

(average) $\bar{\kappa} \overset{\text{def}}{=} \frac{\frac{1}{n}\operatorname{tr}(\boldsymbol{H})}{\lambda_{\min}(\boldsymbol{H})}$, *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

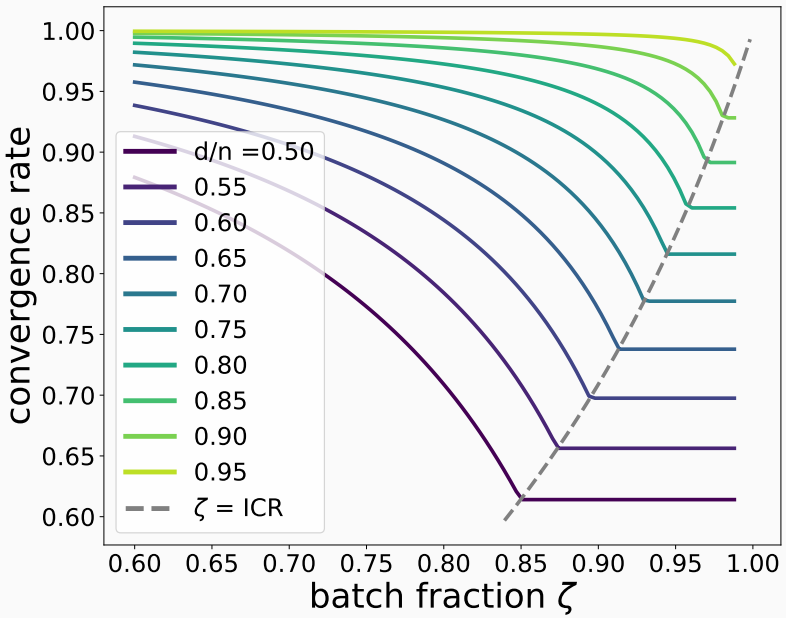
  SGD+M linearly at rate $\mathcal{O}(1/\sqrt{\kappa})$

  and SGD+M accelerates

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

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

  SGD+M $\Leftrightarrow$ SGD



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*, (submitted to NeurIPS '22)

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)

C. Paquette, E. Paquette. *Dynamics of Stochastic Momentum Methods on Large-scale, Quadratic Models*, arxiv.org/pdf/2106.03696.pdf (accepted at NeurIPS 2021)