

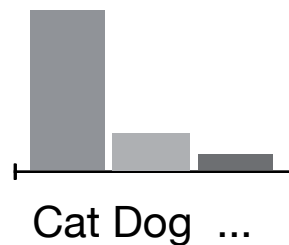
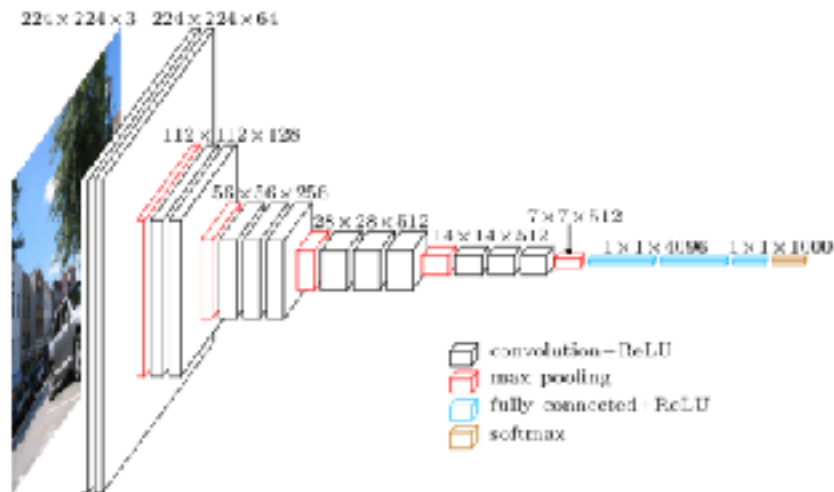
Unraveling the mysteries of stochastic gradient descent for deep neural networks

Pratik Chaudhari



University of Pennsylvania

The question



$$x^* = \underset{x}{\operatorname{argmin}} f(x)$$

measures disagreement of predictions with ground truth

weights aka parameters

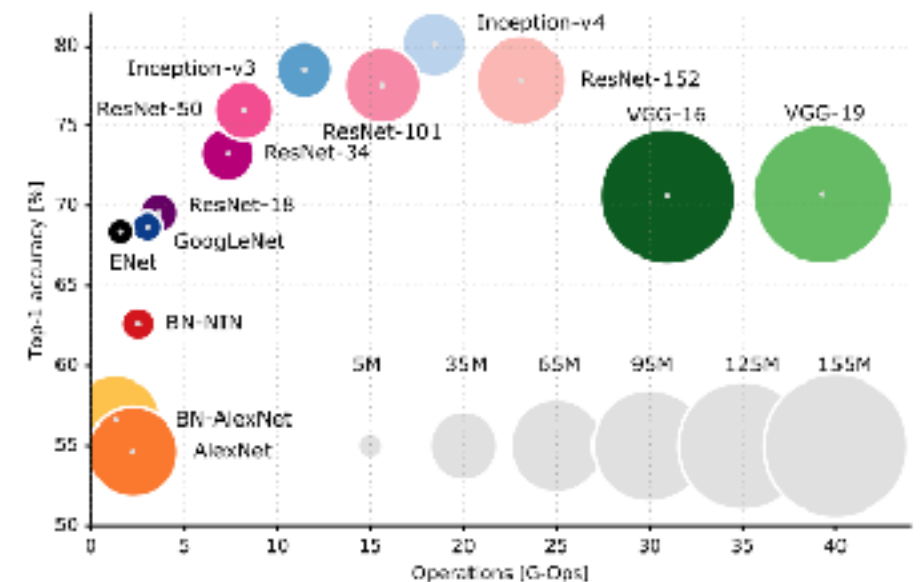
Stochastic gradient descent

$$x_{k+1} = x_k - \eta \nabla f_{\theta}(x_k)$$

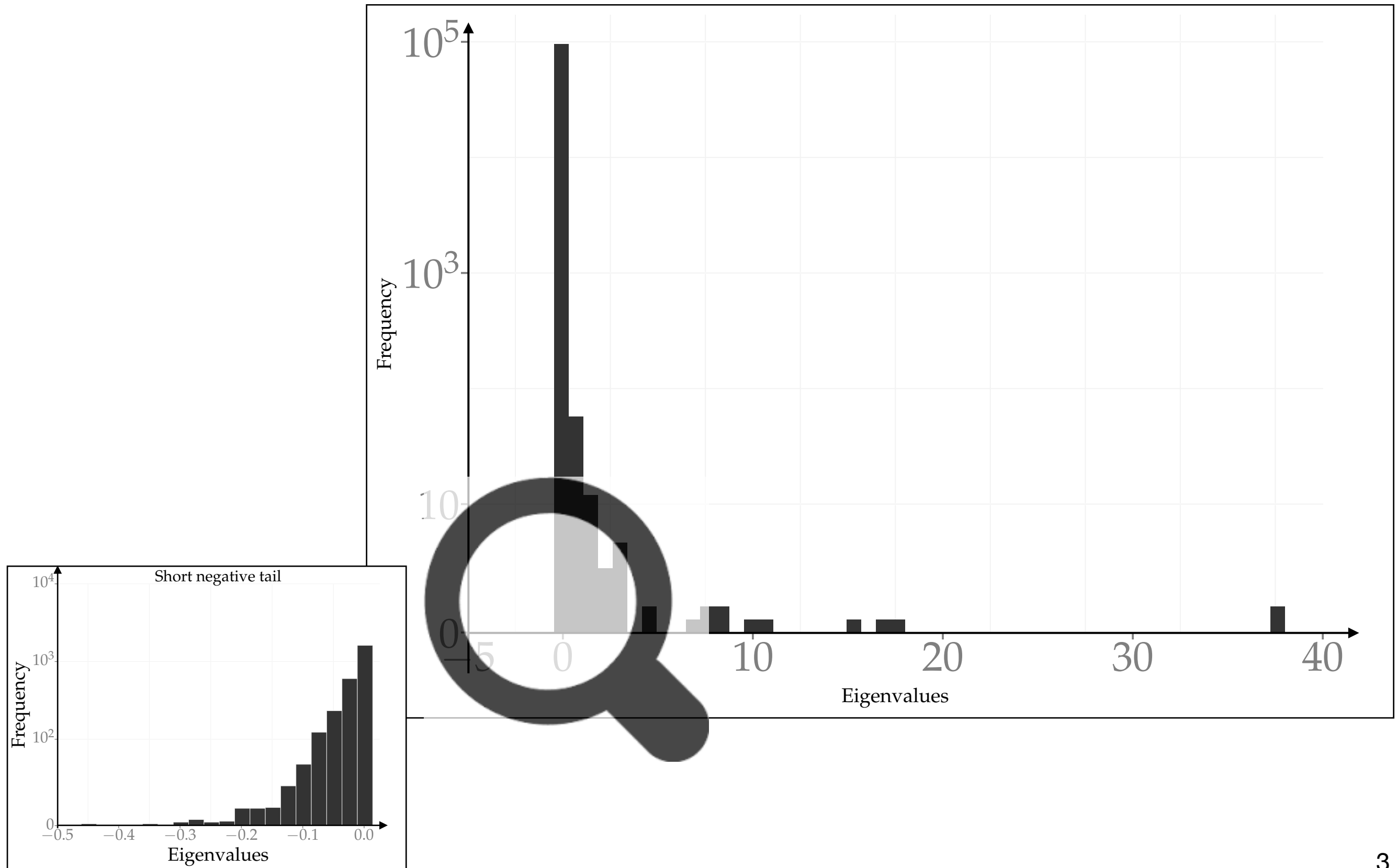
Many, many variants:

AdaGrad, rmsprop, Adam,
SAG, SVRG, Catalyst,
APPA, Natasha, Katyusha...

Why is SGD
so special?



Empirical evidence: wide “minima”



Seeking out wide regions

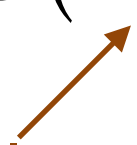
- ▶ Local Entropy [Chaudhari et al., ICLR '17]

$$x^* = \operatorname{argmin}_x f(x)$$

$$= \operatorname{argmax}_x e^{-f(x)}$$

$$\approx \operatorname{argmin}_x -\log \left(G_\gamma * e^{-f(x)} \right)$$

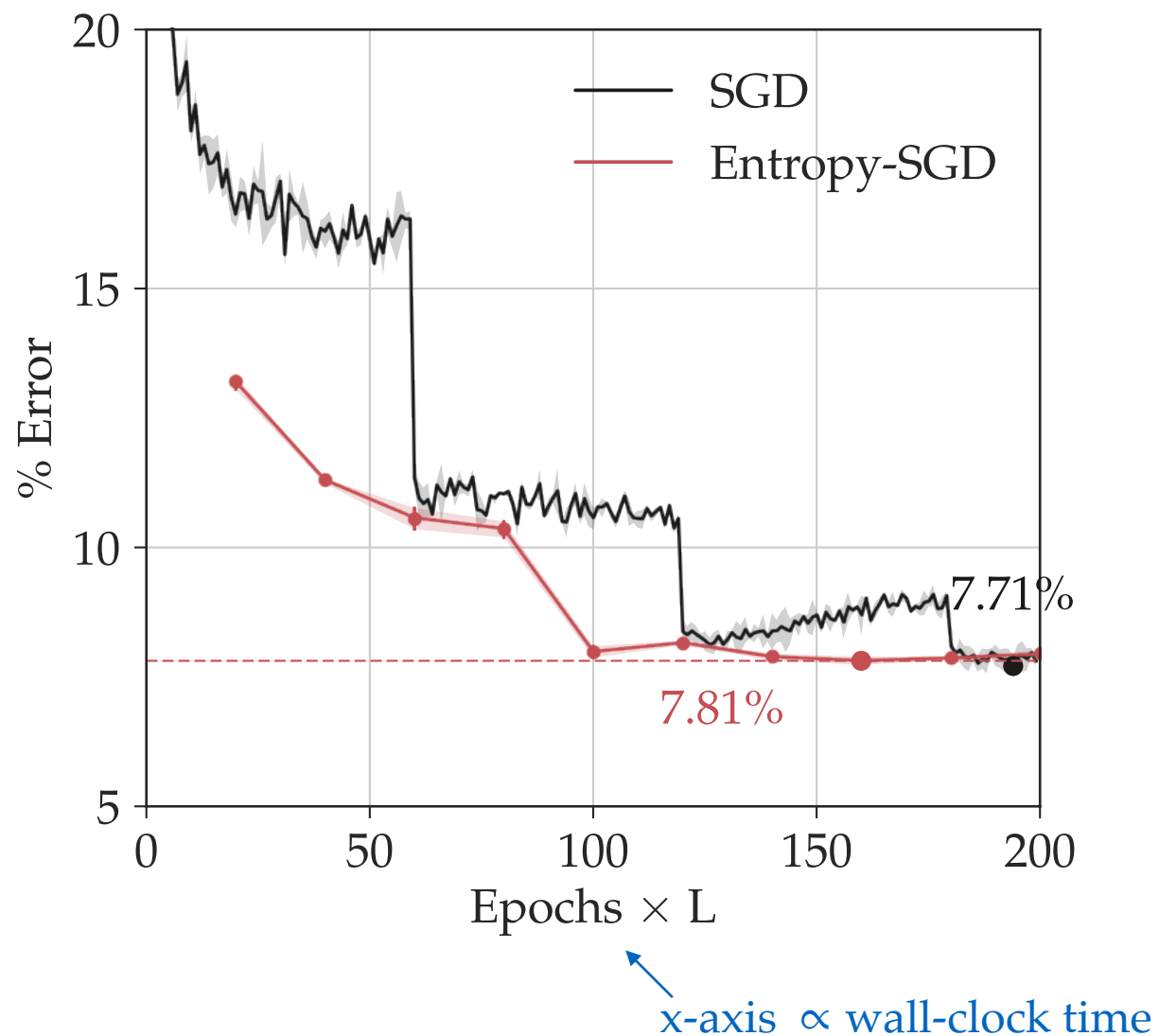
Gaussian kernel
of variance γ



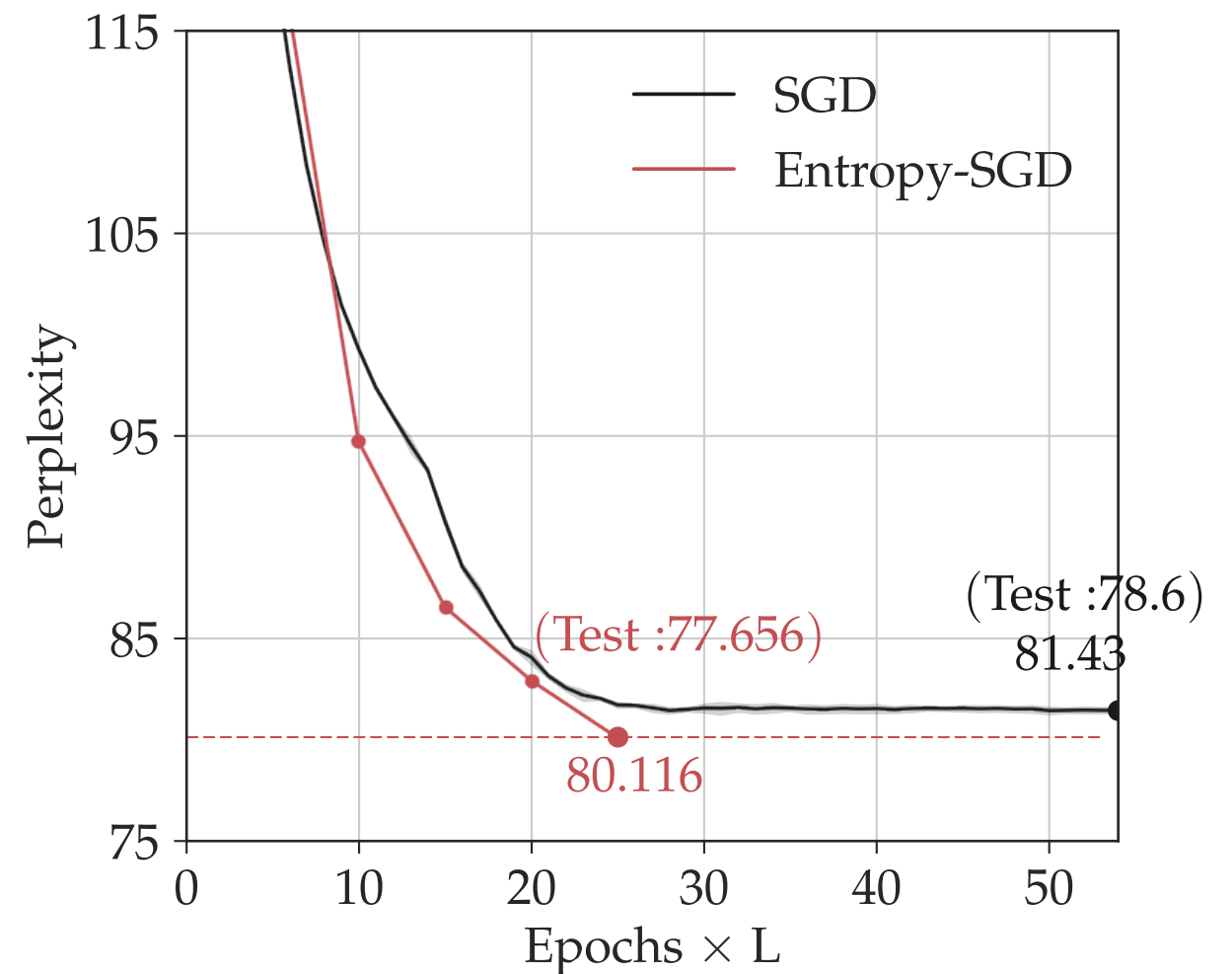
Entropy-SGD: biasing towards wide minima

- Minimize Local Entropy using Markov Chain Monte Carlo techniques [Chaudhari et al., ICLR '17]

CIFAR-10

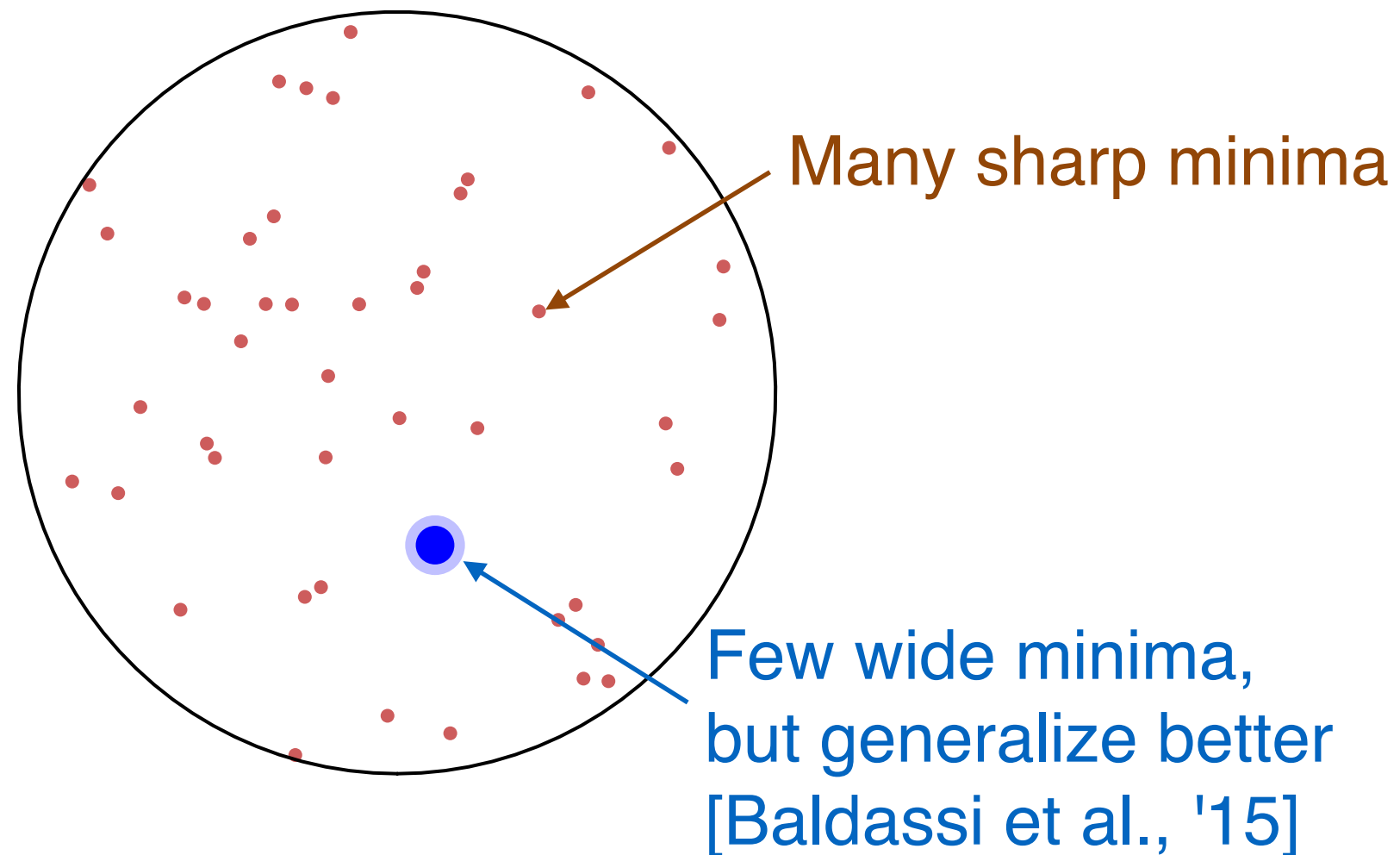


Penn Tree Bank



Diving deep I: statistical physics

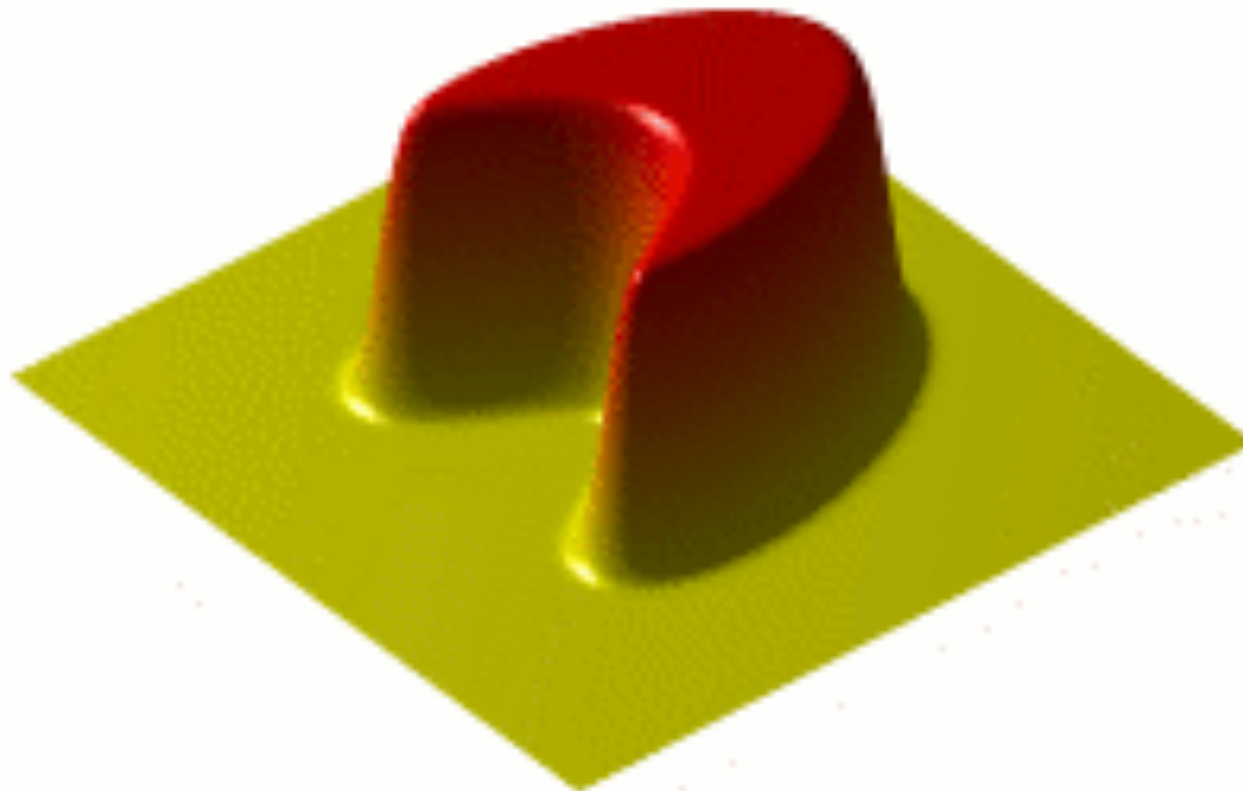
- ▶ Energy landscape of a binary perceptron



- ▶ Wide minima are a large deviations phenomenon

Diving deep II: partial differential equations

$$u_\gamma = \frac{1}{2} \Delta u$$
$$u(x, \gamma = 0) = f(x)$$



Diving deep II: partial differential equations

- ▶ Local Entropy is the solution of a Hamilton-Jacobi (HJ) equation [Chaudhari et al. '18, Research in Mathematical Sciences]

$$u_\gamma = -\frac{1}{2} |\nabla u|^2 + \frac{1}{2} \Delta u$$
$$u(x, \gamma = 0) = f(x)$$

- ▶ Stochastic optimal control interpretation

- ▶ Dynamics: $dx = -\nabla f(x) ds - \alpha(x, s) ds + dW(s)$

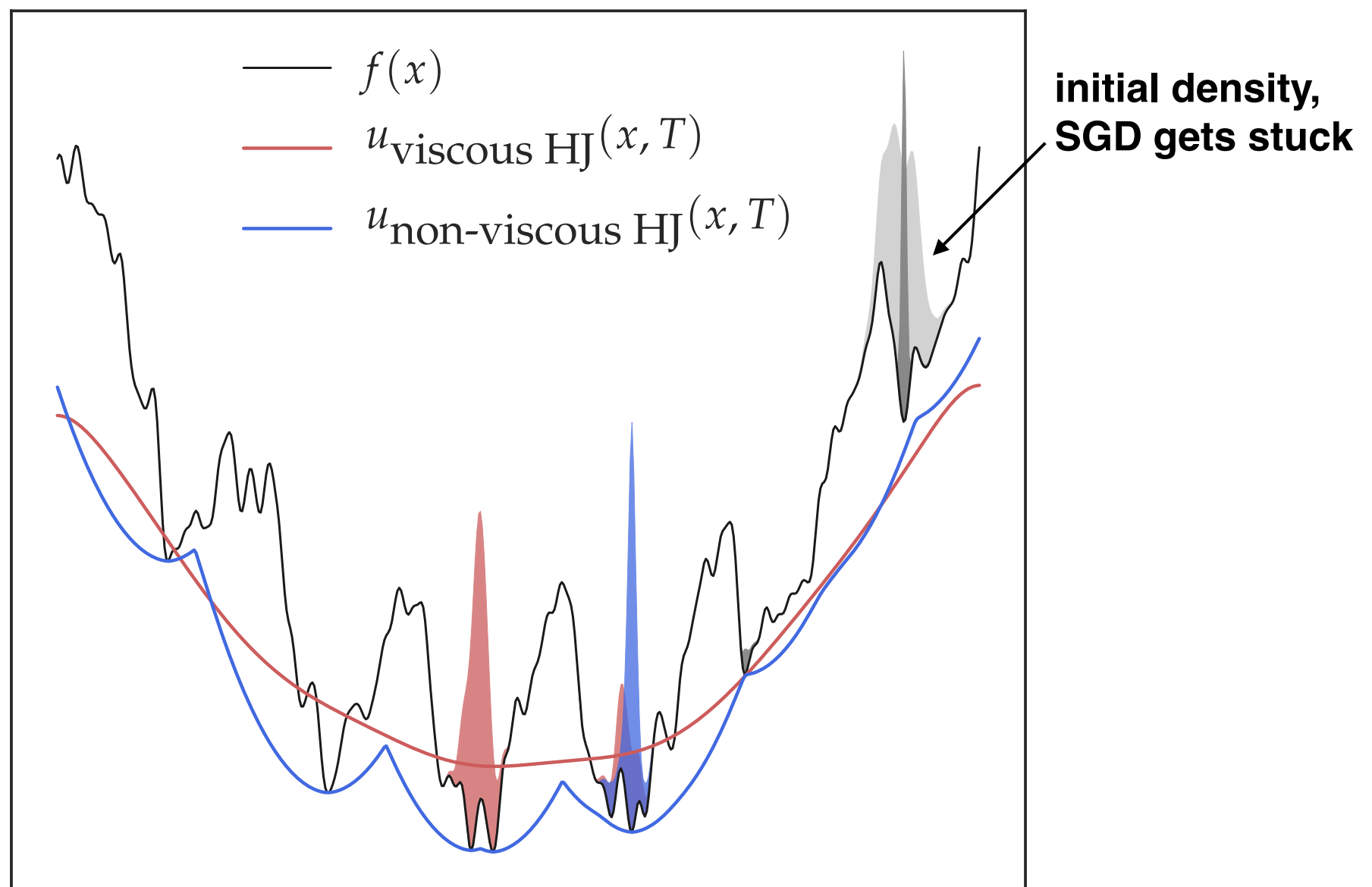
- ▶ Cost: $C = \mathbb{E} \left[f(x(T)) + \frac{1}{2} \int_t^T |\alpha(x(s), s)|^2 ds \right]$

- ▶ Control: $\alpha^*(x, s) = \nabla u(x, s)$

New PDEs for deep learning

- ▶ Non-viscous Hamilton-Jacobi equation

$$u_\gamma = -\frac{1}{2} |\nabla u|^2 + \cancel{\frac{1}{2} \Delta u}^0$$



New PDEs for deep learning

- ▶ Non-viscous Hamilton-Jacobi equation

$$u_\gamma = -\frac{1}{2} |\nabla u|^2 + \cancel{\frac{1}{2} \Delta u}^0$$

- ▶ Proximal Point Iteration [Moreau '65]

$$x_{k+1} = \operatorname{argmin}_y \left\{ f(y) + \frac{1}{2\gamma} \|y - x_k\|^2 \right\}$$

Distributed training of deep networks

- ▶ Difficult to exploit large servers
 - ▶ Communication: models are very large
 - ▶ Generalization: large batch sizes don't work well

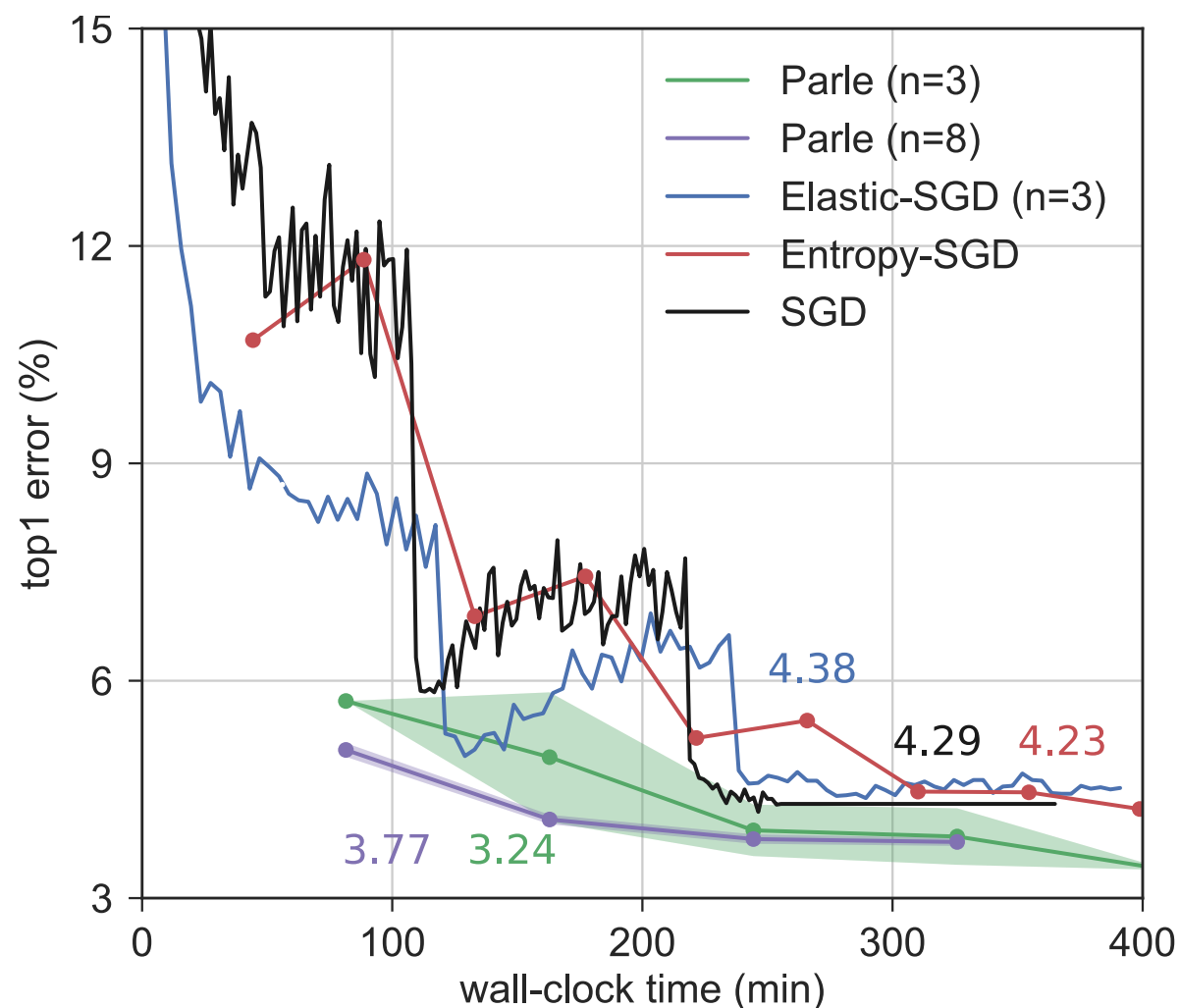
- ▶ Elastic-SGD [Zhang et al., '15]

$$\operatorname{argmin}_{\mathbf{x}, \{\mathbf{x}^k\}} \sum_{k=1}^n f(\mathbf{x}^k) + \frac{1}{2\gamma} \|\mathbf{x}^k - \mathbf{x}\|^2$$

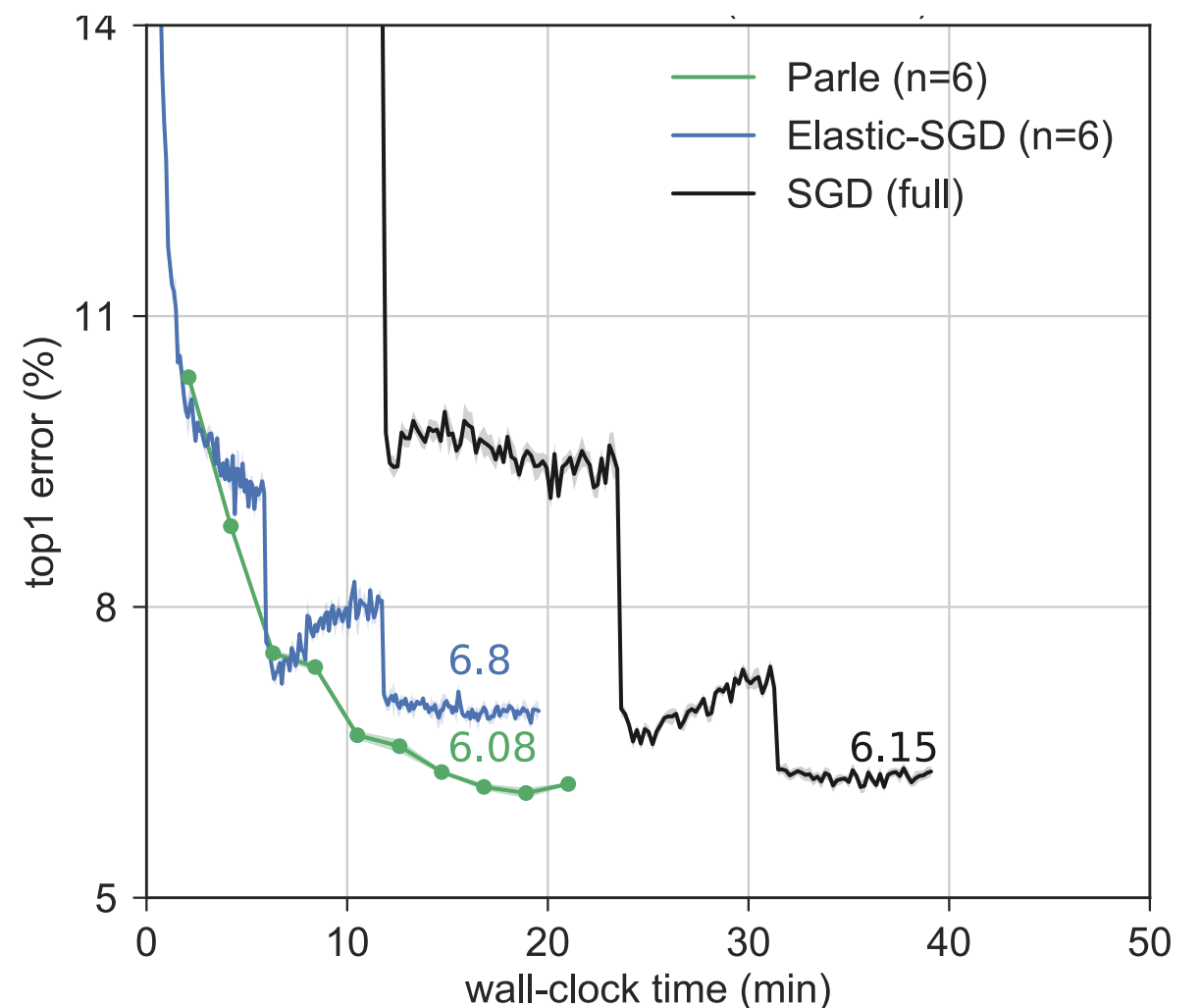
- ▶ Elastic-SGD minimizes local entropy
[Chaudhari et al. '18, Research in Mathematical Sciences]

Parle: parallelization of SGD

- State-of-the-art performance [Chaudhari et al., SysML '18]



Wide-ResNet: CIFAR-10



All-CNN: CIFAR-10 (25% data)

Homogenization of SDEs

Theorem [Chaudhari et al. '18, Research in Mathematical Sciences]

As $\epsilon \rightarrow 0$, the Entropy-SGD dynamics,

$$dx(s) = -\gamma^{-1}(x - y) ds$$

$$dy(s) = -\frac{1}{\epsilon} \left[\nabla f(y) + \gamma^{-1}(y - x) \right] ds + \frac{1}{\sqrt{\epsilon}} dW(s),$$

and the Elastic-SGD dynamics,

$$dx(s) = -\gamma^{-1}(x - y) ds$$

$$dy^k(s) = -\nabla f(y^k) ds - \gamma^{-1}(y^k - \bar{y}) ds + dW(s), \quad \forall k$$

converge to the same homogenized dynamics

$$dX(s) = \bar{h}(X) ds.$$

The question



**Why is SGD
so special?**

A continuous-time view of SGD

- ▶ **Diffusion matrix:** variance of mini-batch gradients

$$\begin{aligned}\text{var}\left(\nabla f_{\ell}(x)\right) &= \frac{D(x)}{\ell} \\ &= \frac{1}{\ell} \left(\frac{1}{N} \sum_{k=1}^N \nabla f_k(x) \nabla f_k(x)^{\top} - \nabla f(x) \nabla f(x)^{\top} \right)\end{aligned}$$

- ▶ **Temperature:** ratio of learning rate and step-size

$$\beta^{-1} = \frac{\eta}{2\ell}$$

A continuous-time view of SGD

- ▶ Continuous-time approximation of discrete-time updates

$$dx = -\nabla f(x) \underbrace{dt}_{\triangleq \eta} + \sqrt{2\beta^{-1} D(x)} dW(t)$$

will assume $x \in \Omega \subset \mathbb{R}^d$

- ▶ Fokker-Planck (FP) equation gives the distribution on the weight space induced by SGD

$$\rho_t = \operatorname{div} \left(\underbrace{\nabla f \rho}_{\text{drift}} + \underbrace{\beta^{-1} \operatorname{div}(D \rho)}_{\text{diffusion}} \right) \quad \text{where } x(t) \sim \rho(t)$$

Wasserstein gradient flow

- ▶ Heat equation $\rho_t = \operatorname{div}(\mathbf{I} \nabla \rho)$ performs steepest descent on the Dirichlet energy

$$\frac{1}{2} \int_{\Omega} |\nabla \rho(x)|^2 dx$$

- ▶ It is also the steepest descent in the Wasserstein metric for

$$-H(\rho) = \int_{\Omega} \log \rho d\rho$$

$$\rho_{k+1}^{\tau} \in \operatorname{argmin}_{\rho} \left\{ -H(\rho) + \frac{\mathbb{W}_2^2(\rho, \rho_k^{\tau})}{2\tau} \right\}$$

converges to trajectories
of the heat equation

- ▶ Negative entropy is a Lyapunov functional for Brownian motion

$$\rho_{\text{heat}}^{\text{ss}} = \operatorname{argmin}_{\rho} -H(\rho)$$

Wasserstein gradient flow: with drift

- ▶ If $D = I$, the Fokker-Planck equation

$$\rho_t = \operatorname{div} \left(\nabla f \rho + \beta^{-1} I \nabla \rho \right)$$

has the Jordan-Kinderlehrer-Otto (JKO) functional [Jordan et al., '97]

$$\rho^{\text{ss}}(x) = \operatorname{argmin}_{\rho} \underbrace{\mathbb{E}_{x \sim \rho} [f(x)]}_{\text{energetic term}} - \underbrace{\beta^{-1} H(\rho)}_{\text{entropic term}}$$

as the Lyapunov functional.

- ▶ FP is the steepest descent on JKO in the Wasserstein metric

What happens for non-isotropic noise?

$$\rho_t = \operatorname{div} \left(\underbrace{\nabla f \rho}_{\text{drift}} + \underbrace{\beta^{-1} \operatorname{div}(D \rho)}_{\text{diffusion}} \right)$$

- ▶ FP **monotonically** minimizes the free energy

$$\rho^{\text{ss}}(x) = \operatorname{argmin}_{\rho} \mathbb{E}_{x \sim \rho} [\Phi(x)] - \beta^{-1} H(\rho)$$

- ▶ Rewrite as

$$F(\rho) = \beta^{-1} \operatorname{KL}(\rho \parallel \rho^{\text{ss}})$$

compare with $|x - x^*|$ for deterministic optimization.

SGD performs variational inference

Theorem [Chaudhari & Soatto, ICLR '18]

The functional

$$F(\rho) = \beta^{-1} \text{KL}(\rho \parallel \rho^{\text{ss}})$$

is minimized monotonically by trajectories of the Fokker-Planck equation

$$\rho_t = \text{div}(\nabla f \rho + \beta^{-1} \text{div}(D \rho))$$

with ρ^{ss} as the steady-state distribution. Moreover,

$$\Phi = -\beta^{-1} \log \rho^{\text{ss}}$$

up to a constant.

Some straightforward implications

- ▶ Learning rate should scale linearly with batch-size

$$\beta^{-1} = \frac{\eta}{2\ell} \quad \text{should not be small}$$

- ▶ Sampling with replacement regularizes better than without

$$\beta_{\text{w/o replacement}}^{-1} = \frac{\eta}{2\ell} \left(1 - \frac{\ell}{N} \right)$$

also generalizes better.

Information Bottleneck Principle

- ▶ Minimize mutual information of the representation with the training data [Tishby et al. '99, Achille & Soatto '17]

$$\text{IB}_\beta(\theta) = \mathbb{E}_{x \sim \rho_\theta} [f(x)] - \beta^{-1} \text{KL}(\rho_\theta \parallel \text{prior})$$

- ▶ Minimizing these functionals is hard, SGD does it naturally

Potential Phi vs. original loss f

- ▶ The solution of the variational problem is

$$\rho^{\text{ss}}(x) = \frac{1}{Z_\beta} e^{-\beta \Phi(x)}$$

- ▶ Key point

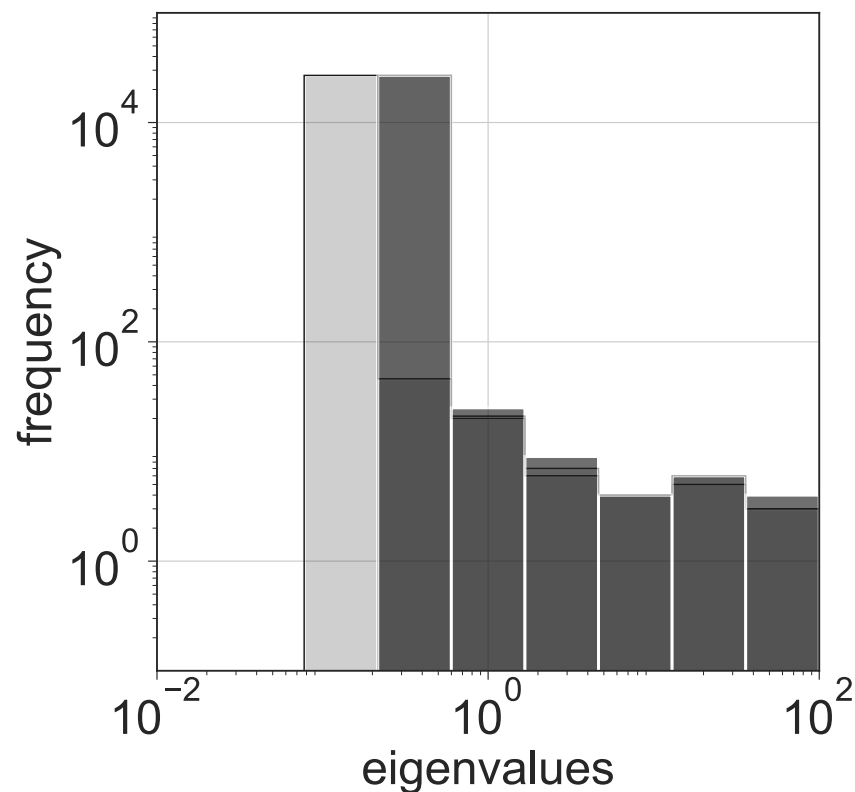
$$\rho^{\text{ss}}(x) \neq \frac{1}{Z'_\beta} e^{-\beta f(x)}$$

Most likely locations of SGD are not the critical points of the original loss

- ▶ The two losses are equal if and only if noise is isotropic

$$D(x) = I \quad \Leftrightarrow \quad \Phi(x) = f(x)$$

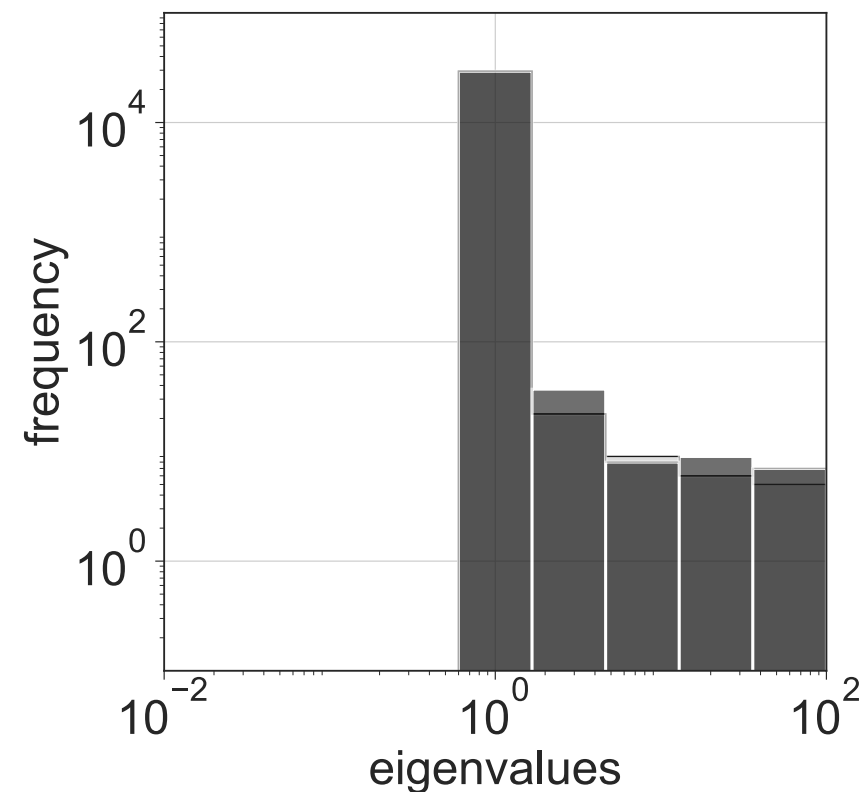
Deep networks have highly non-isotropic noise



CIFAR-10

$$\lambda(D) = 0.27 \pm 0.84$$

$$\text{rank}(D) = 0.34\%$$



CIFAR-100

$$\lambda(D) = 0.98 \pm 2.16$$

$$\text{rank}(D) = 0.47\%$$

- Evaluate neural architectures using the diffusion matrix

Cause 1: Distribution of real data



Cause 2: Symmetries in deep architectures

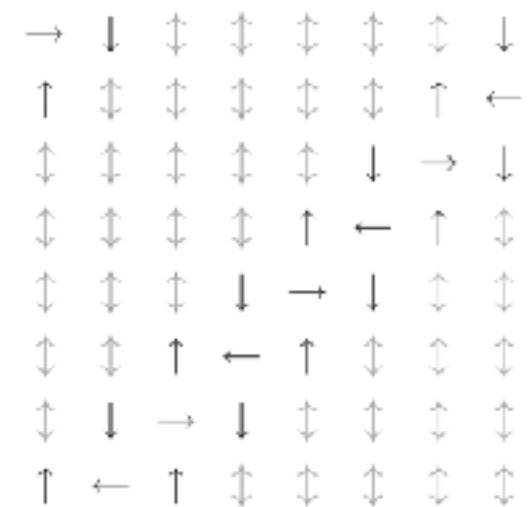
- ▶ Deep networks have permutation as well as continuous symmetries, e.g., from matrix factorization [Nouiehed & Razaviyayn '18],

$$X = A P P^{-1} B$$

- ▶ Over-parametrized two layer neural networks have non-isolated global minima, no local minima [Venturi et al., '18]

- ▶ XY-models

$$-H(\sigma) = \frac{1}{2} \sum_{(i,j) \in \text{neighbors}} \cos(\sigma_i - \sigma_j)$$



[Nerattini et al., '13]

Most likely trajectories of SGD

Theorem [Chaudhari & Soatto, ICLR '18]

The most likely trajectories of SGD are

$$\dot{x} = j(x),$$

where the "leftover" vector field

$$j(x) = -\nabla f(x) + D(x) \nabla \Phi(x) - \beta^{-1} \operatorname{div} D(x)$$

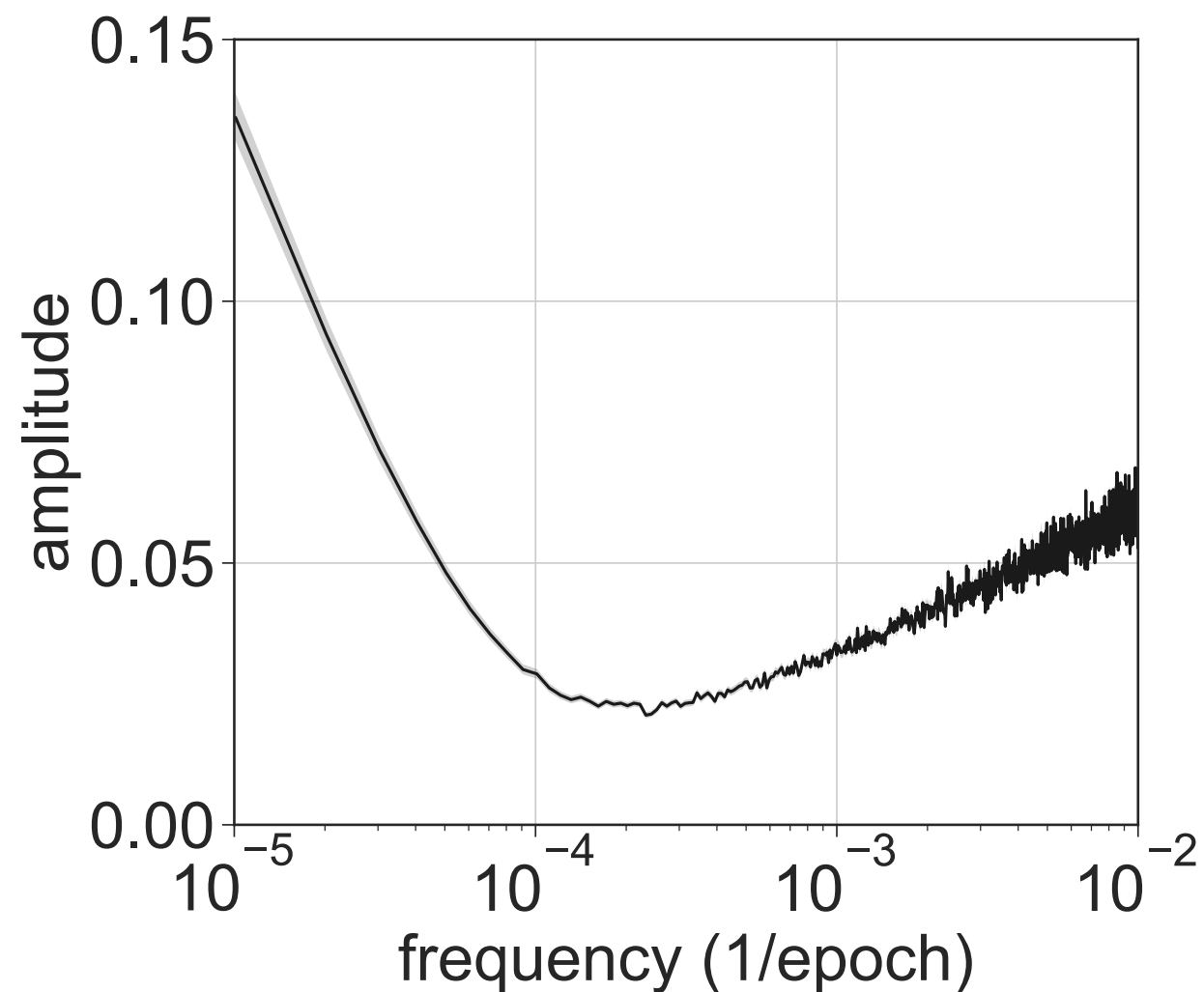
is such that

$$\operatorname{div} j(x) = 0.$$

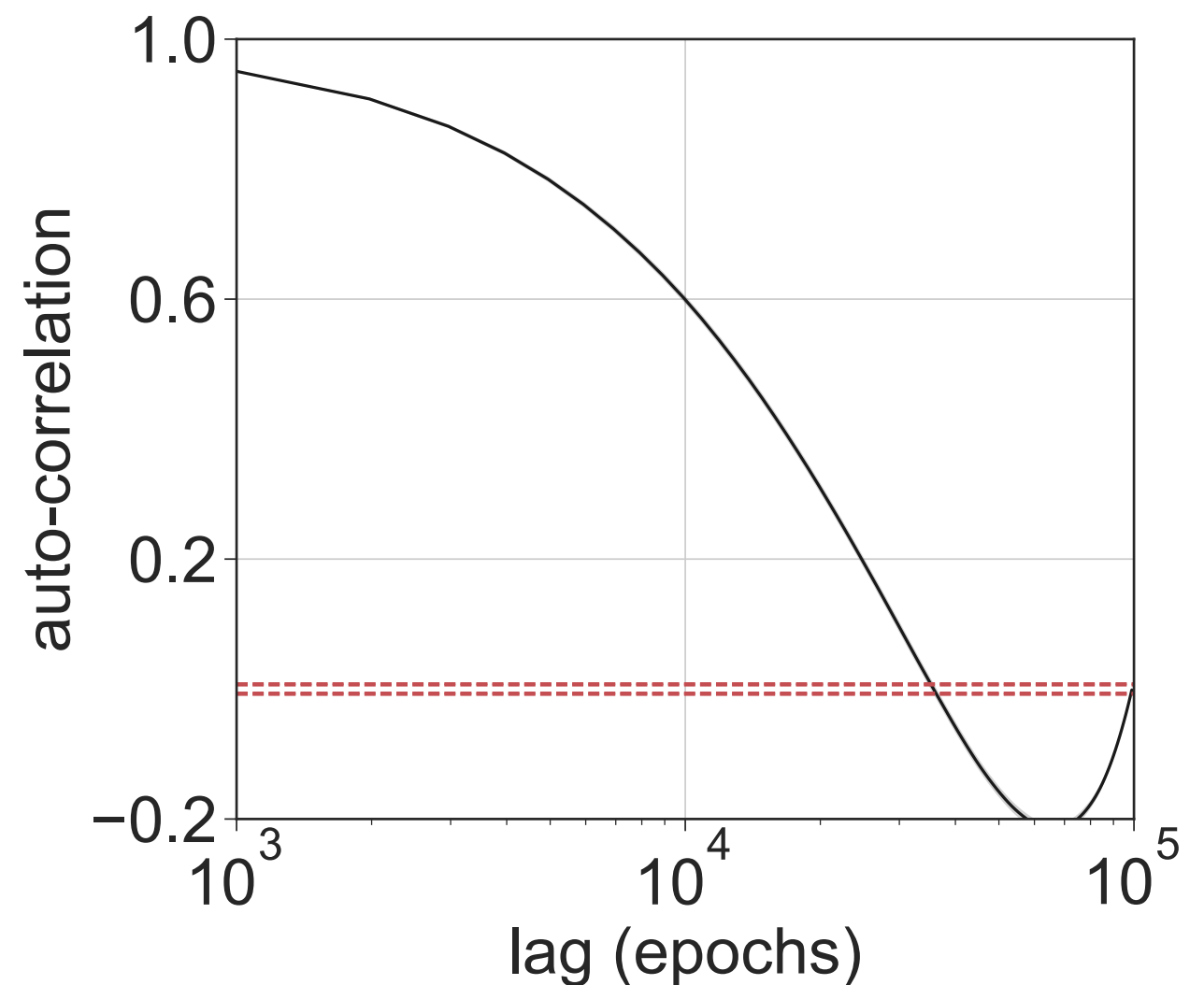
Most likely trajectories of SGD

- ▶ Run SGD for 10^5 epochs

FFT of $x_{k+1}^i - x_k^i$

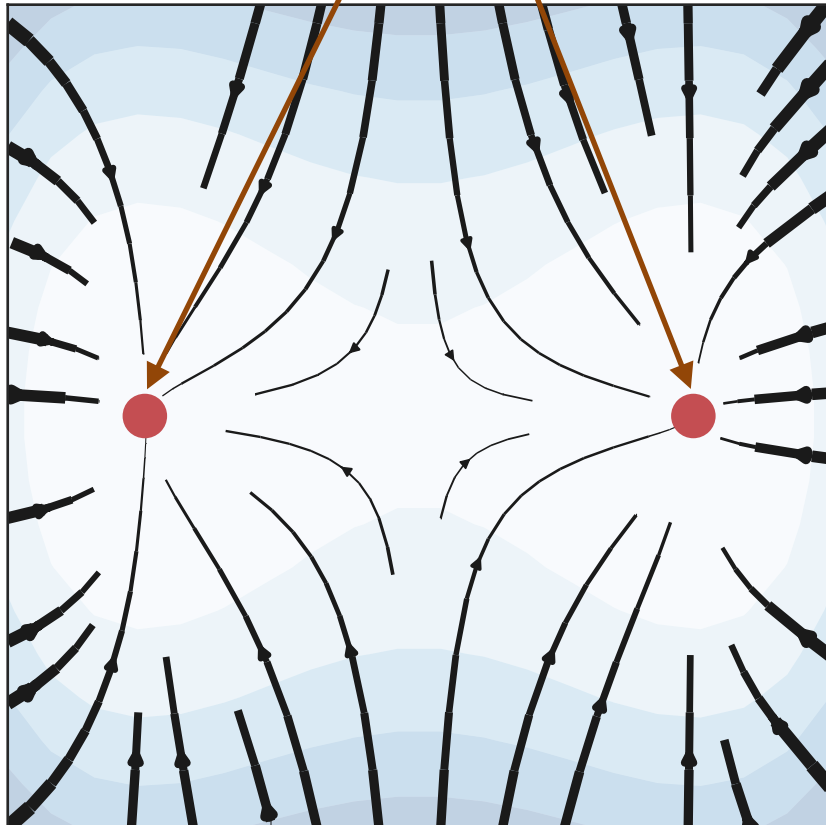


Auto-correlation of x_k^i



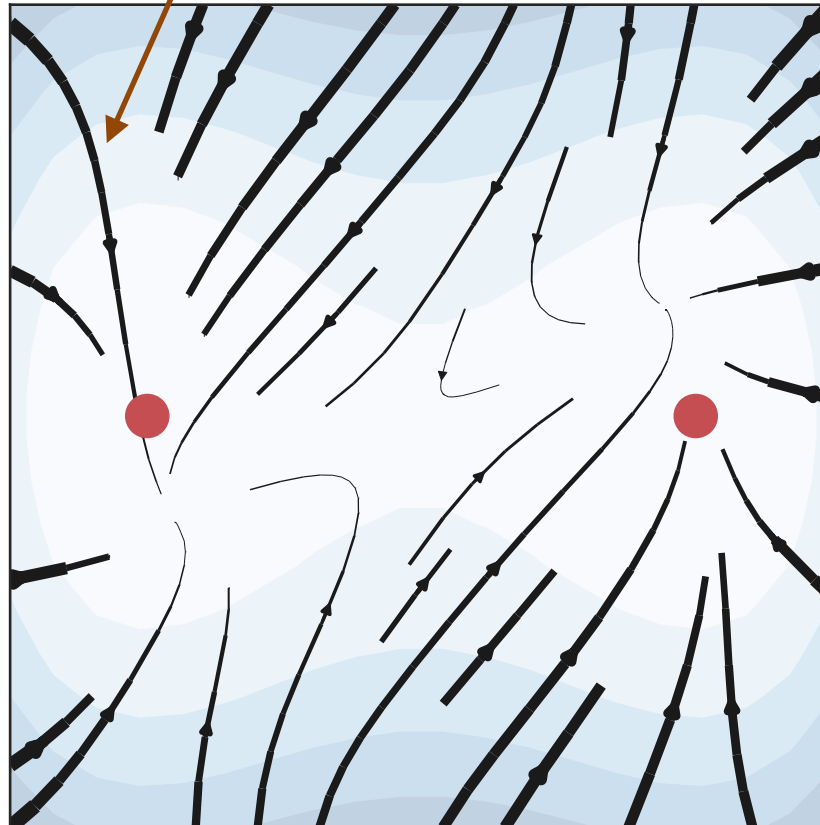
An example

$$\nabla\Phi(x) = 0$$



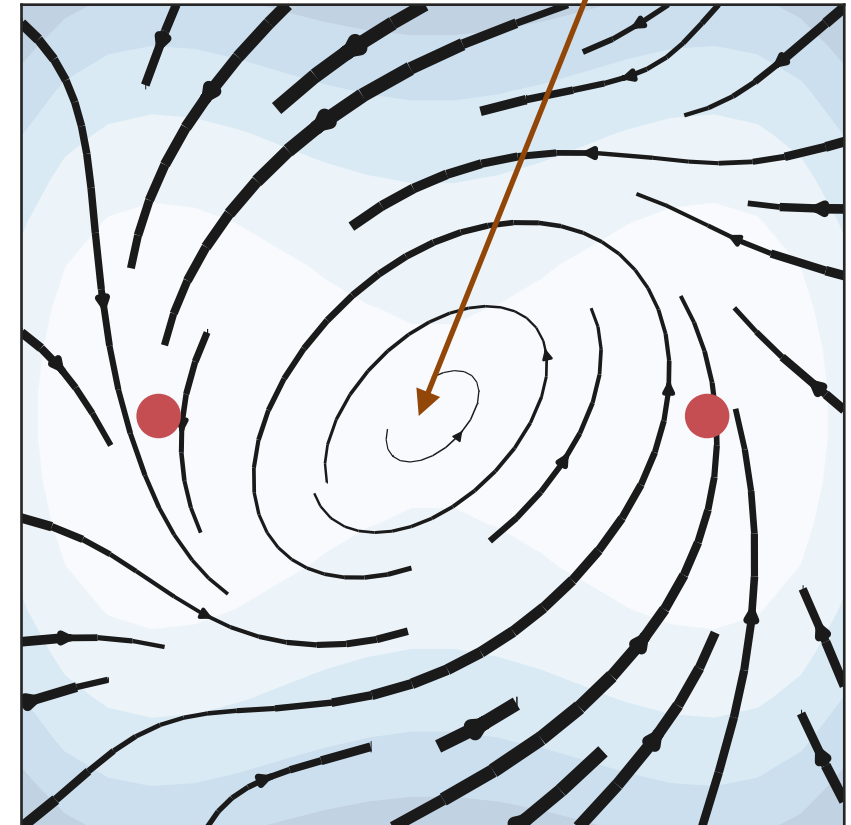
$$j(x) = 0$$

force-field



$|j(x)|$ is small

saddle-point



very large $|j(x)|$

Most likely locations are not the critical points of the original loss

Theorem [Chaudhari & Soatto, ICLR '18]

The Ito SDE

$$dx = -\nabla f \, dt + \sqrt{2\beta^{-1} D} \, dW(t)$$

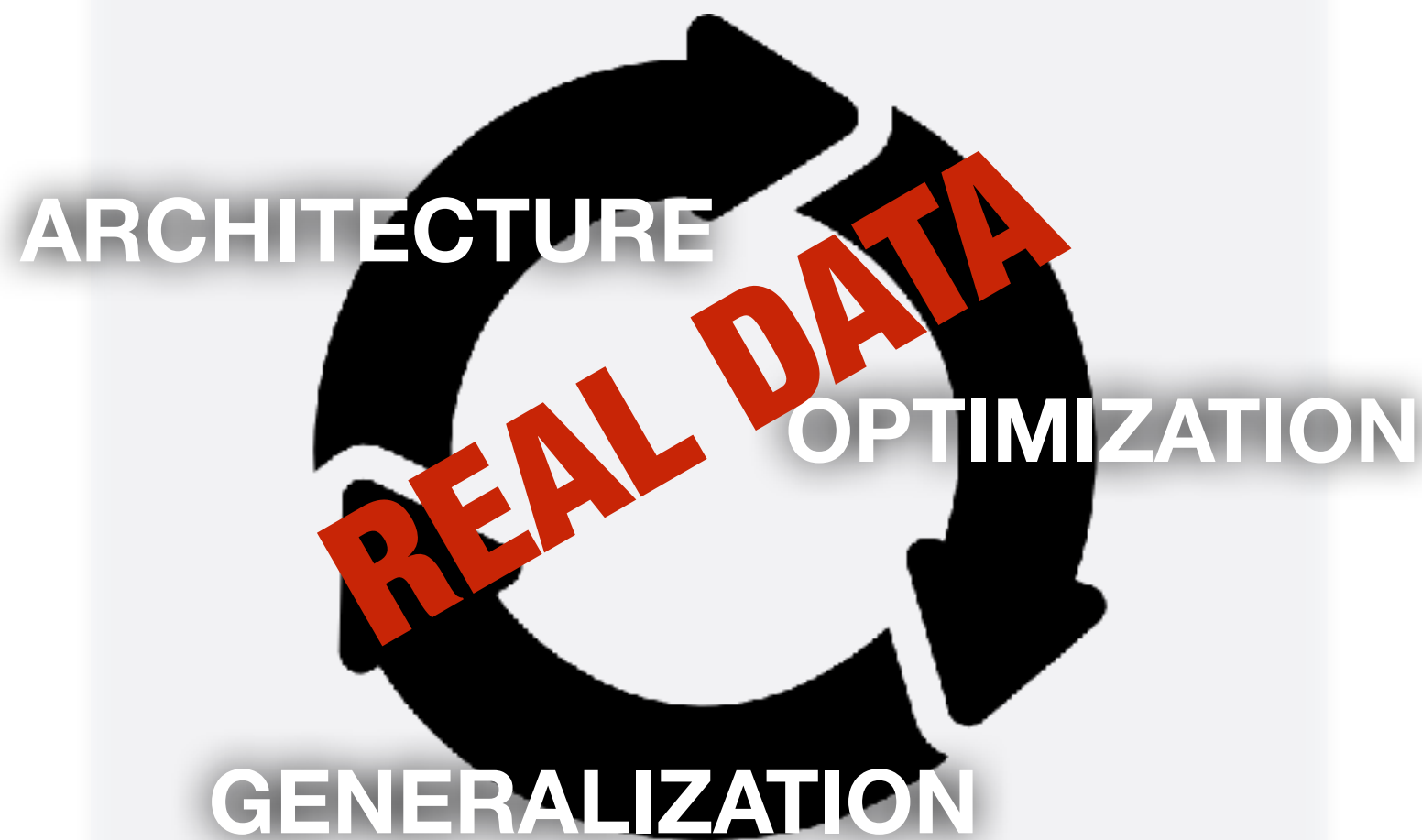
is equivalent to an A-type SDE

$$dx = -(D + Q) \nabla \Phi \, dt + \sqrt{2\beta^{-1} D} \, dW(t)$$

with the same steady-state $\rho^{ss} \propto e^{-\beta \Phi(x)}$ if

$$\nabla f = (D + Q) \nabla \Phi - \beta^{-1} \operatorname{div} (D + Q).$$

Knots in our understanding



1. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks.
Pratik Chaudhari and Stefano Soatto [ICLR '18]
2. Entropy-SGD: biasing SGD towards wide regions
Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun and Riccardo Zecchina [ICLR '17]
3. Parle: parallel training of deep networks
Pratik Chaudhari, Carlo Baldassi, Riccardo Zecchina, Stefano Soatto, Ameet Talwalkar and Adam Oberman [SysML '18]
4. Deep Relaxation: partial differential equations for optimizing deep neural networks
Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto and Guillaume Carlier
Research in Mathematical Sciences [RMS '19]



www.pratikac.info

Thank you, questions?