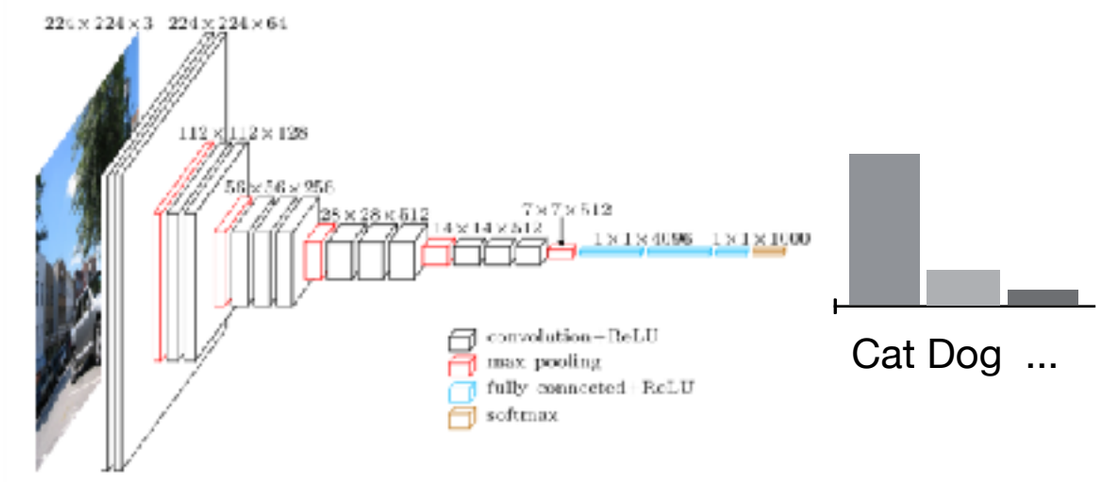
Unraveling the mysteries of stochastic gradient descent for deep neural networks

Pratik Chaudhari

The question



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

measures disagreement of predictions with ground truth

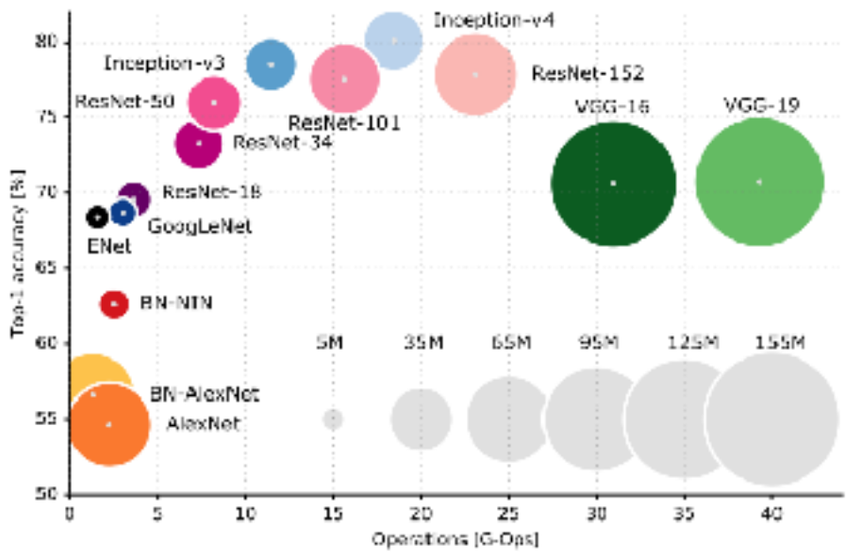
weights aka parameters

Stochastic gradient descent

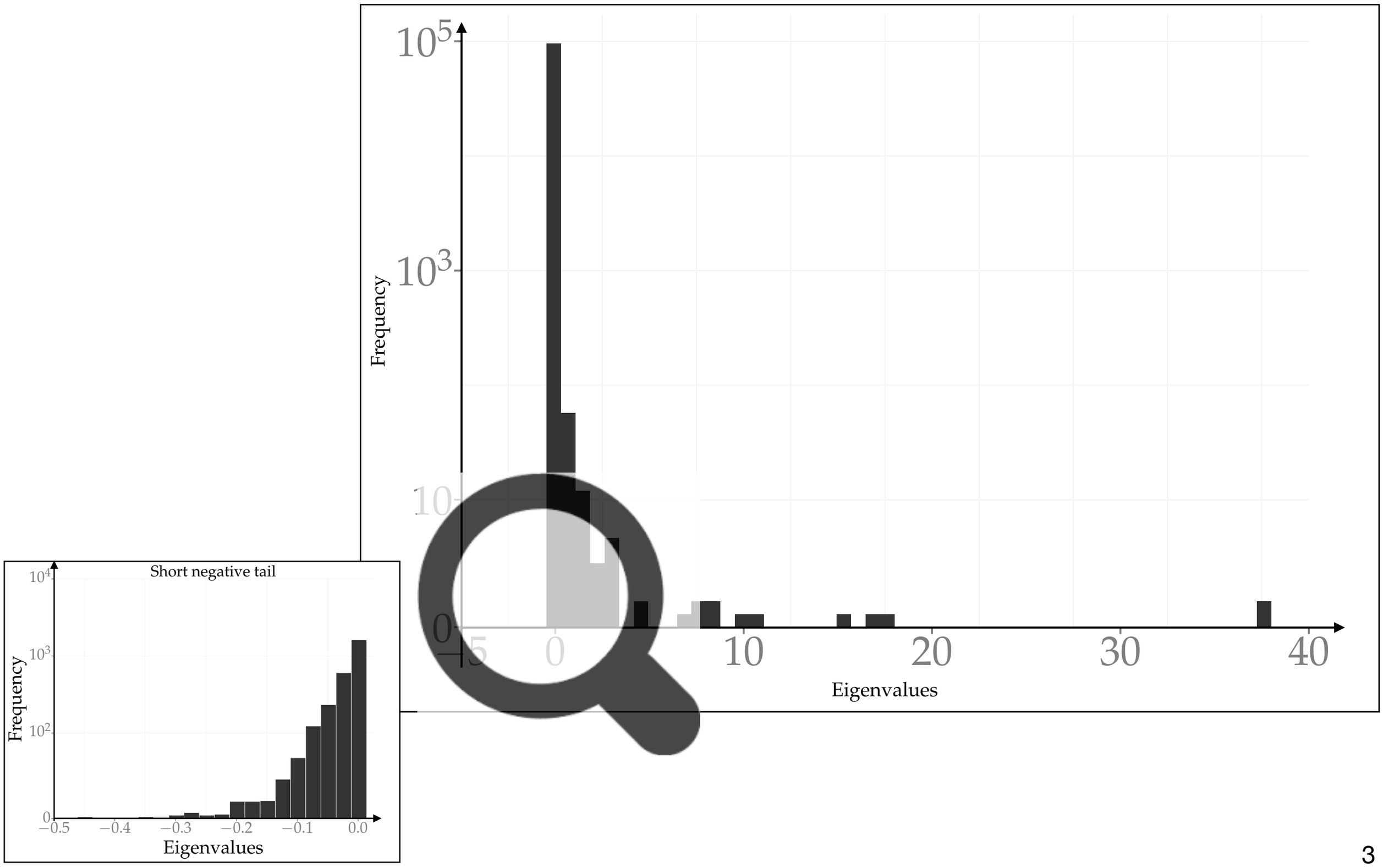
$$x_{k+1} = x_k - \eta \nabla f_{\theta}(x_k)$$

Many, many variants:
AdaGrad, rmsprop, Adam,
SAG, SVRG, Catalyst,
APPA, Natasha, Katyusha…

Why is SGD so special?

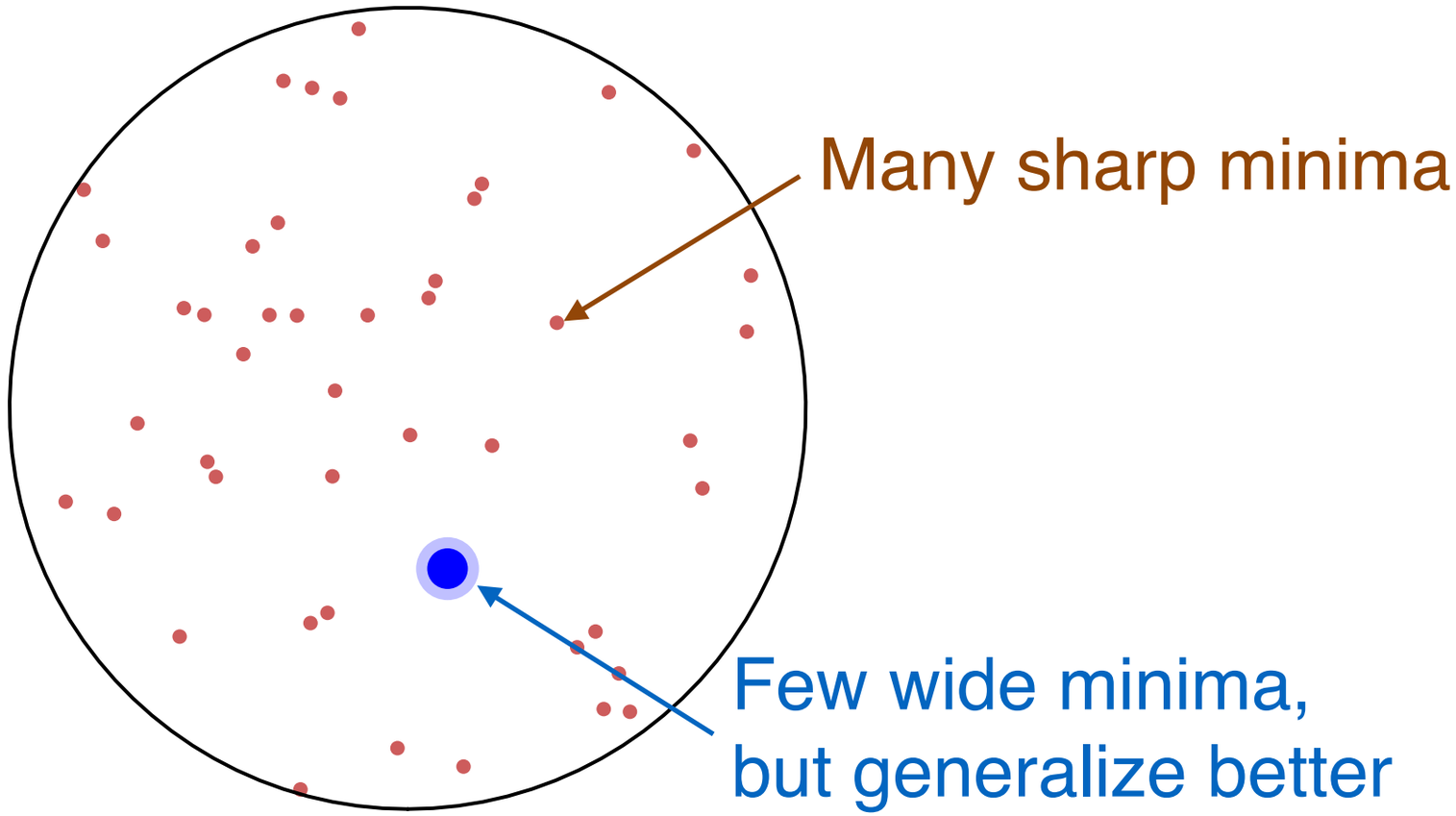


Empirical evidence: wide “minima”



A bit of statistical physics

- Energy landscape of a binary perceptron



- Wide “minima” are a large deviations phenomenon for this problem

Tilting the Gibbs measure

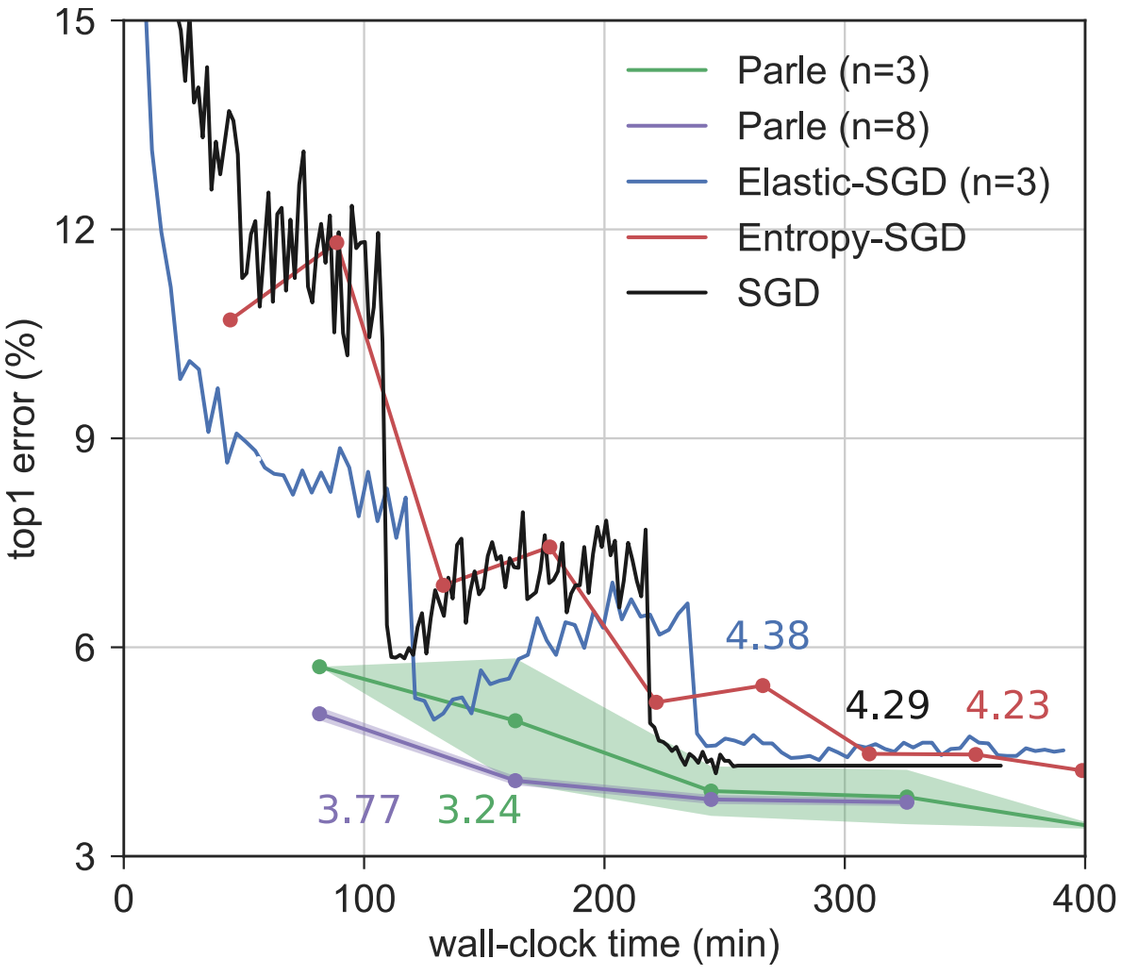
- Local Entropy [Chaudhari et al., ICLR '17]

$$
\begin{aligned}
x^* &= \operatorname*{argmin}_x \; f(x) \\
&= \operatorname*{argmax}_x \; e^{-f(x)} \\
&\approx \operatorname*{argmin}_x \; -\log\left(G_\gamma \; * \; e^{-f(x)}\right)
\end{aligned}
$$

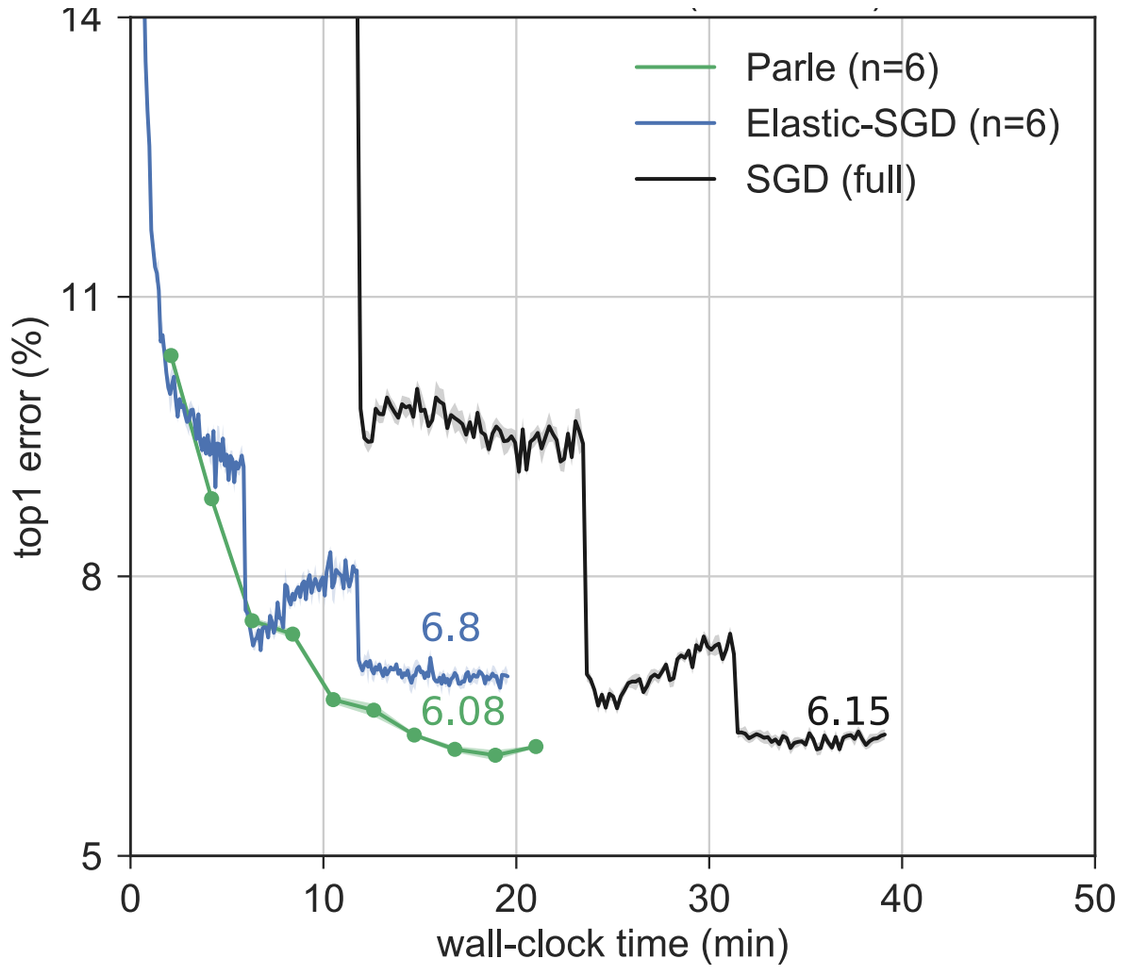
Gaussian kernel of variance γ

Parle: parallelization of SGD

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



Wide-ResNet: CIFAR-10



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

The question

Why is SGD so special?

A continuous-time view of SGD

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

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

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

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

A continuous-time view of SGD

- Continuous-time limit 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

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

Wasserstein gradient flow

- Heat equation $\rho_t = \operatorname{div}\left(I\ \nabla\rho\right)$ 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-Kinderleher-Otto (JKO) functional [Jordan et al., '97]

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

as the Lyapunov functional.

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

What happens for non-isotropic noise?

$$\rho_t = \mathrm{div}\Big(\underbrace{\nabla f \rho}_{\text{drift}} + \underbrace{\beta^{-1} \mathrm{div}\big(D\,\rho\big)}_{\text{diffusion}} \Big)$$

- FP monotonically minimizes the free energy

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

- Rewrite as

$$F(\rho) = \beta^{-1} \mathrm{KL}\left(\rho \,\|\, \rho^{\mathrm{ss}}\right)$$

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

SGD performs variational inference

Theorem [Chaudhari & Soatto, ICLR '18]

The functional

$$F(\rho) = \beta^{-1} \mathrm{KL}\left(\rho \,\|\, \rho^{\mathrm{ss}}\right)$$

is minimized monotonically by trajectories of the Fokker-Planck equation

$$\rho_t = \mathrm{div}\Big(\nabla f\, \rho + \beta^{-1} \mathrm{div}\left(D\ \rho\right)\Big)$$

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

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

up to a constant.

Some implications

- Learning rate should scale linearly with batch-size

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

- Sampling with replacement regularizes better than without

$$\beta^{-1}_{\text{w/o replacement}} = \frac{\eta}{2\mathscr{b}}\left(1 - \frac{\mathscr{b}}{N}\right)$$

also generalizes better.

Information Bottleneck Principle

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

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

- Minimizing these functionals is hard, SGD does it naturally

Potential Phi vs. original loss f

- The solution of the variational problem is

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

- Key point

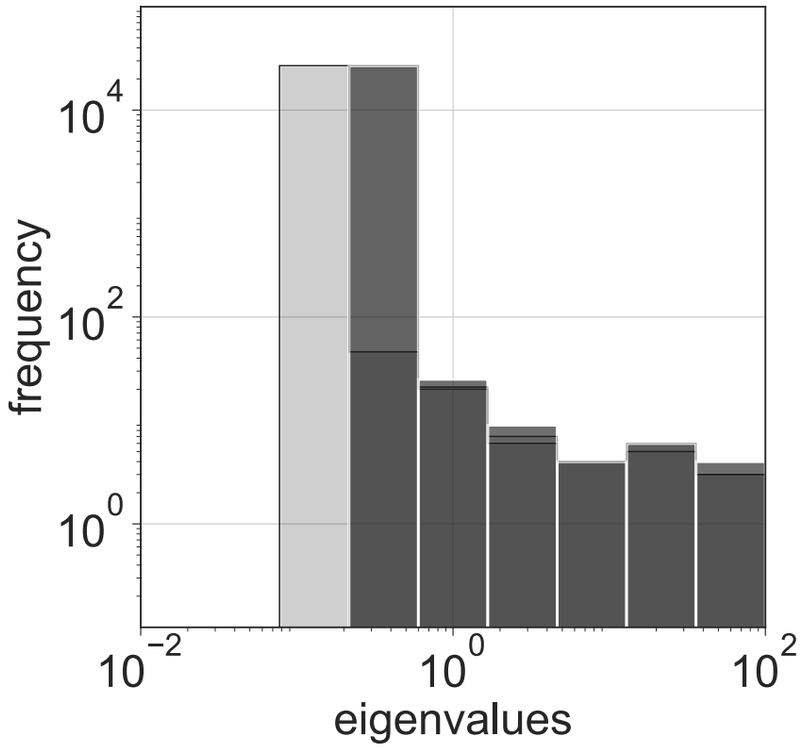
$$\rho^{ss}(x) \neq \frac{1}{Z'_\beta} e^{-\beta\, f(x)}$$

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

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

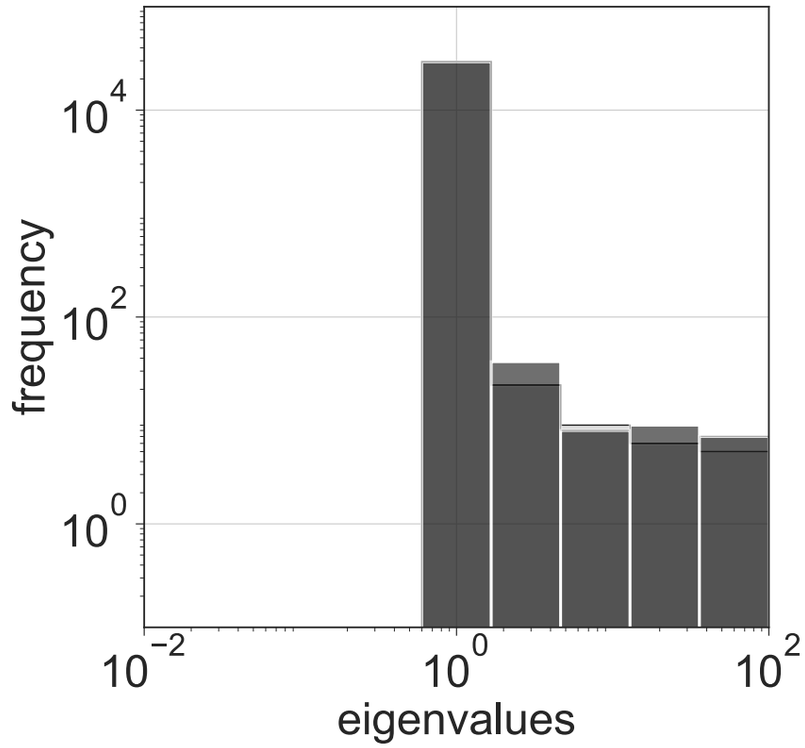
Deep networks have highly non-isotropic noise



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 I: Real data is not independent

Cause 2: Symmetries in deep architectures

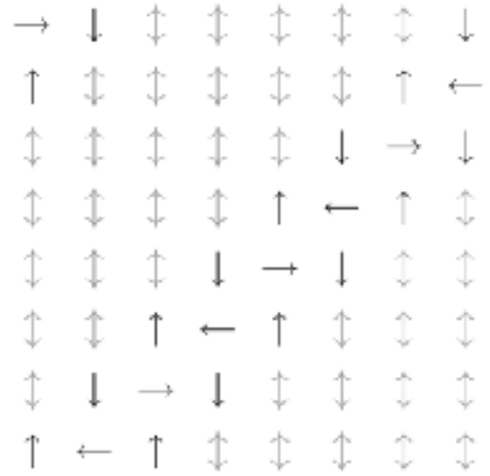
- Deep networks have permutation as well as continuous symmetries, e.g., from matrix factorization [Nouiehed & Razaviyayn '18],

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

- 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

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

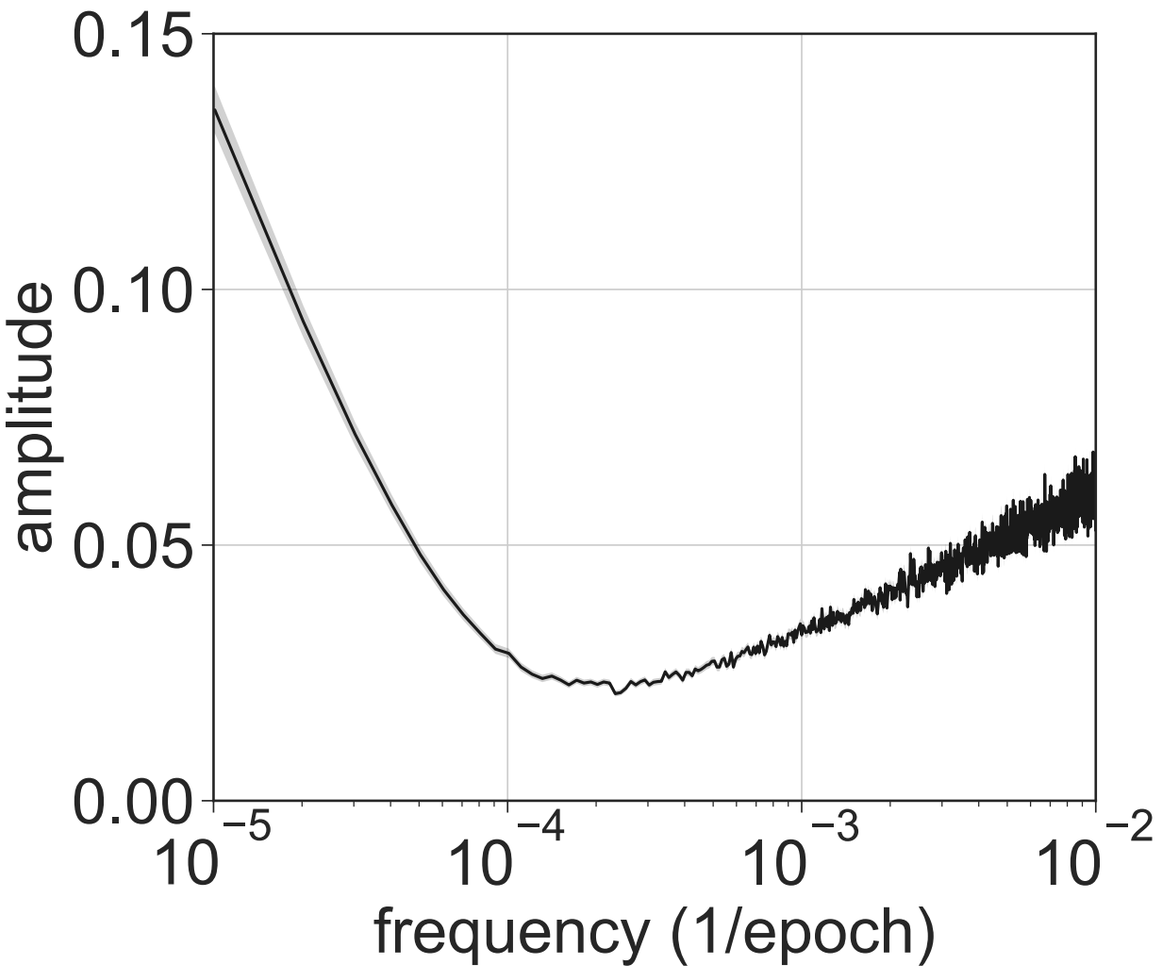
is such that

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

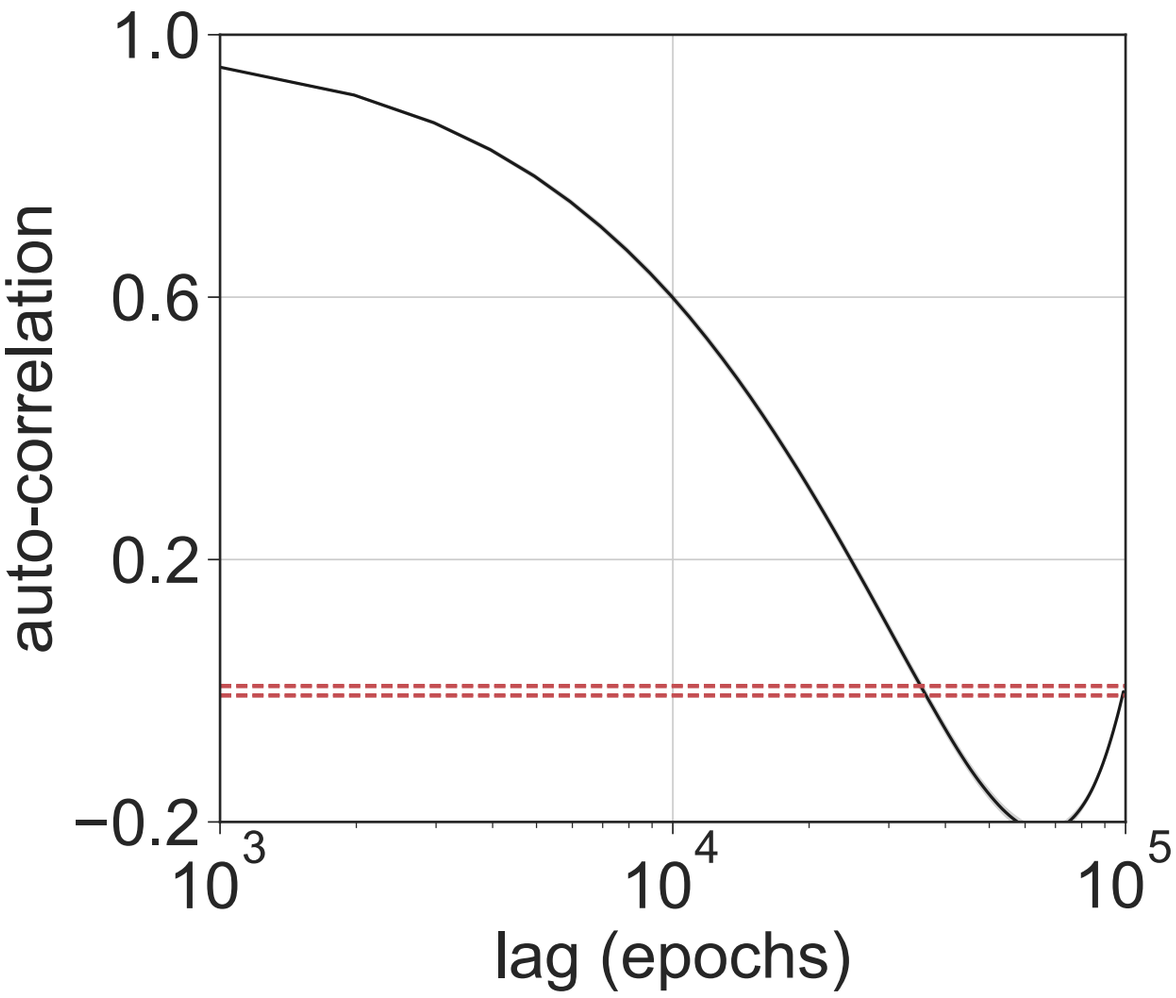
Most likely trajectories of SGD

- Run SGD for 10^5 epochs

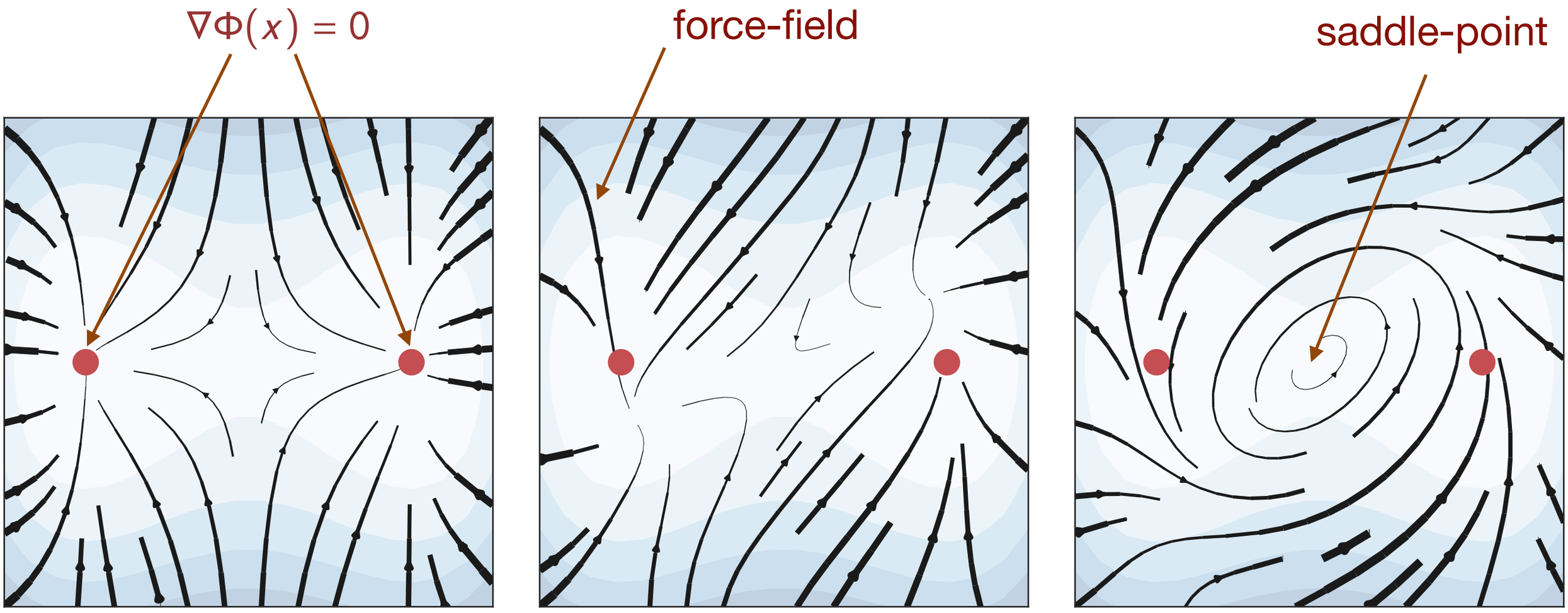
FFT of $x^i_{k+1} - x^i_k$



Auto-correlation of x^i_k



An example



$j(x) = 0$ $\quad$ $|j(x)|$ is small $\quad$ very large $|j(x)|$

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

Theorem [Chaudhari & Soatto, ICLR '18]

The Ito SDE

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

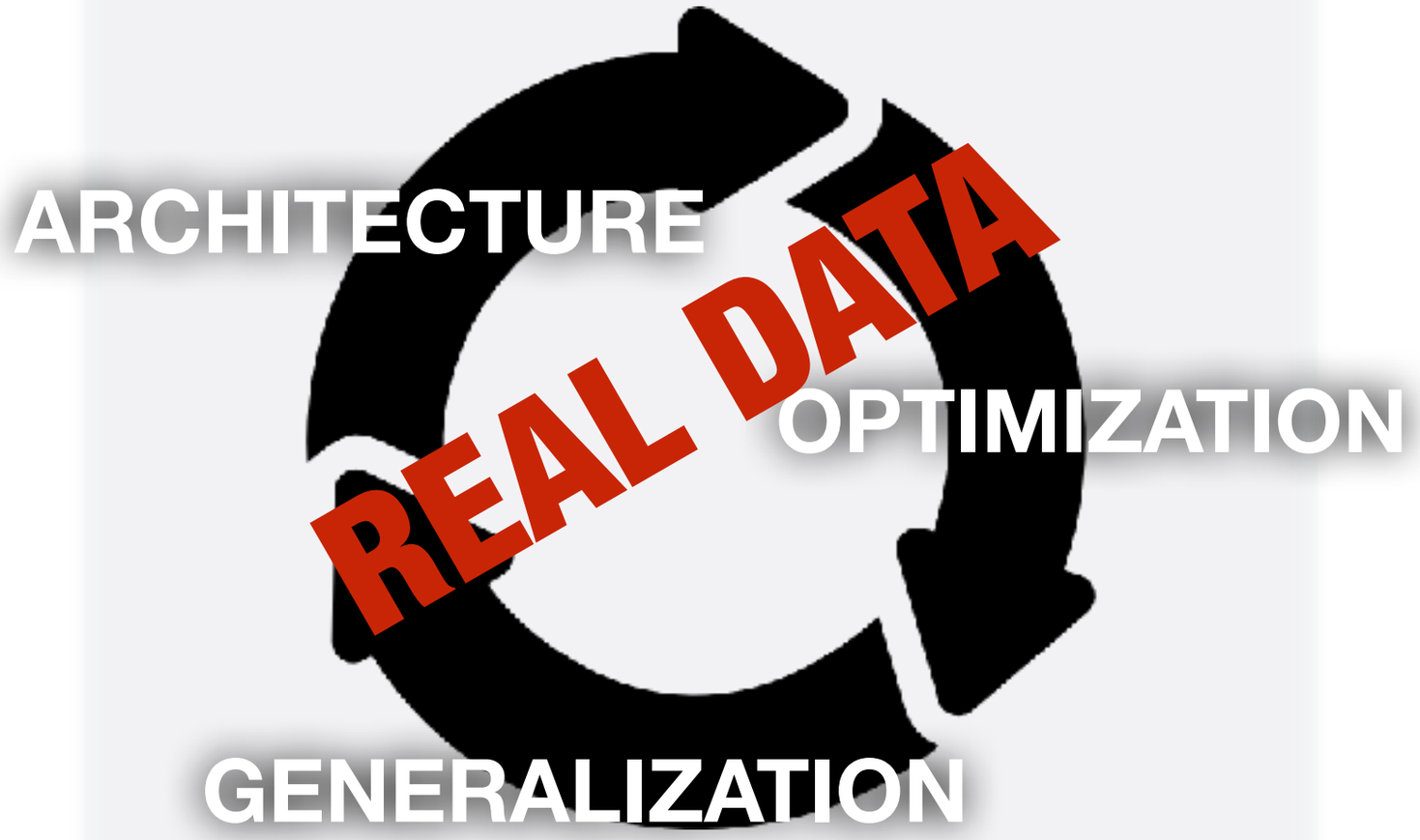
is equivalent to an A-type SDE

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

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

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

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 et al., [ICLR ’17]

3. Parle: parallel training of deep networks
   Pratik Chaudhari et al., [SysML ‘18]

www.pratikac.info

Thank you, questions?