https://arxiv.org/abs/1710.11029

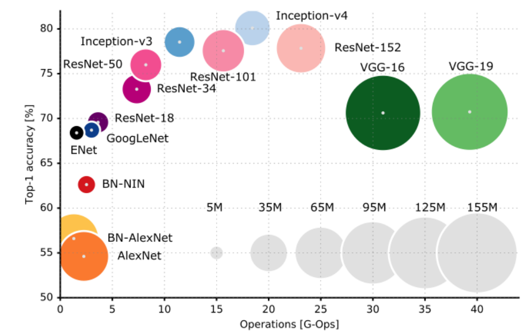
Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks

Pratik Chaudhari and Stefano Soatto

- **SGD performs variational inference**
  - **minimizes an average potential**
  - **with entropic regularization**
- **Deep networks introduce highly non-isotropic mini-batch gradient noise**
- **Potential is different from the original loss**
- **Most likely trajectories of SGD are limit cycles in the weight space**

Despite numerous variants, state of the art networks are trained with SGD

Why is SGD so special?



Continuous-time SGD

- Stochastic differential equation

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

- Statistics and magnitude of noise

$$\mathrm{var}\left(\nabla f_{b}(x)\right) = \frac{D(x)}{b} \qquad \beta^{-1} = \frac{\eta}{2b}\left(1-\frac{b}{N}\right)$$

SGD performs variational inference

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

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

Explains —

- Learning rate **scales linearly** with batch-size
- Sampling **with replacement** is better than without
- SGD has an **information bottleneck**
- Generalization of **Wasserstein gradient flow**

Non-equilibrium thermodynamics

$$j(x) = -\nabla f(x) + D(x)\,\nabla\Phi(x)$$

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

sufficient to satisfy the 2nd law of thermodynamics

Most likely trajectories of SGD are **limit cycles** in the weight space

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

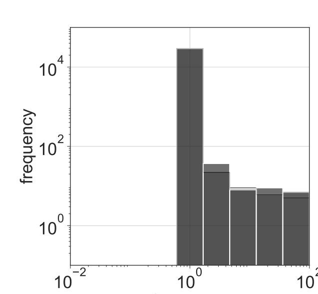
with deviation

$$\nabla f - (D+Q)\,\nabla\Phi = -\beta^{-1}\left(\operatorname{div}\cdot Q\right)^{T}$$

Generalization performance and local entropy

$$f_{\gamma}(x) = -\log\left(G_{\gamma} * e^{-f(x)}\right)$$

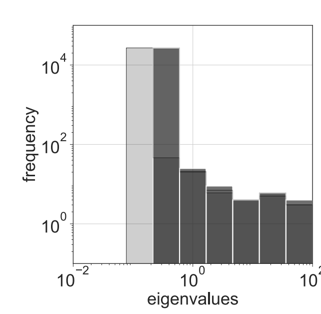
Noise covariance



CIFAR-10

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

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

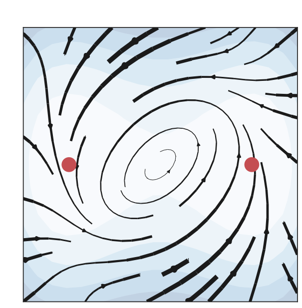


CIFAR-100

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

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

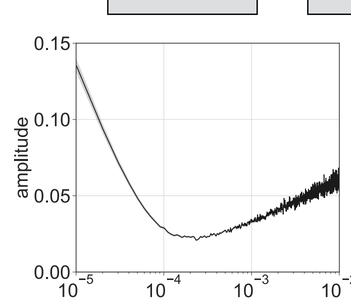
An example



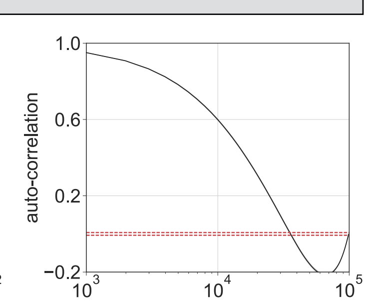
very large $|j(x)|$

SGD may converge around saddle points

FFT



Auto-correlation



Full gradient