Escaping Saddles with Stochastic Gradients

Hadi Daneshmand* Jonas Kohler* Aurélien Lucchi Thomas Hofmann

Institute for Machine Learning, ETH Zürich; Max Planck ETH Center for Learning Systems

Problem Setting

- ▷ Minimizing the training objective of a neural network, namely minimizing

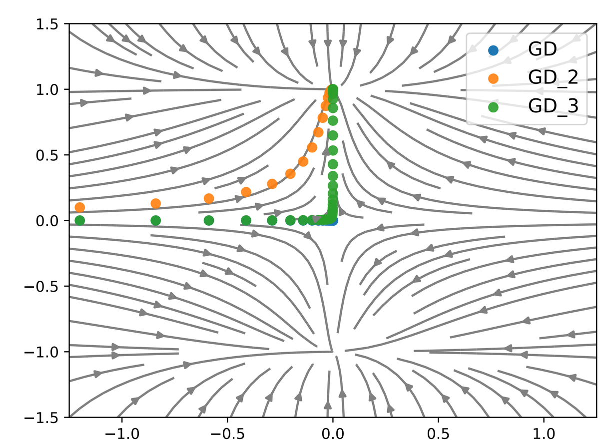
$$f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{z}_i}(\mathbf{w}), \quad \mathbf{w} \in \mathbb{R}^d,$$

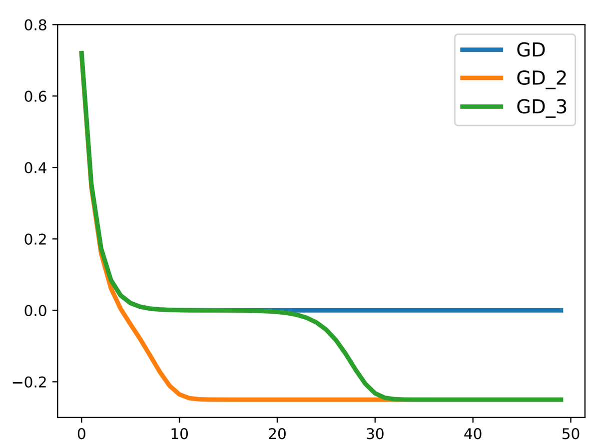
which potentially is a *non-convex* function.

- ▷ We assume that f is sufficiently smooth
  = (gradient is L-Lipschitz) + (Hessian is ρ-Lipschitz) + ($\|\nabla f_{\mathbf{z}_i}(\mathbf{w})\| \leq \ell$)
- ▷ Goal: convergence to a 2nd-order stationary points

$$\{\mathbf{w} \in \mathbb{R}^d | \|\nabla f(\mathbf{w})\| = 0, \nabla^2 f(\mathbf{w}) \succeq 0\}$$

Challenges of Strict Saddles





Gradient descent may converge to a strict saddle $\bar{\mathbf{w}}$, but

- ▷ GD is unstable around $\bar{\mathbf{w}}$
- ▷ And $P(\lim_t \mathbf{w}_t = \bar{\mathbf{w}}) = 0$ [LSJR16]
- ▷ Yet, it may take exponential time to escape [DJL+17]

Escaping Saddles with Isotropic Perturbations

Most of escape strategies rely on perturbation with an injective isotropic noise, for example

- ▷ PGD [JGN+17] $\mathbf{w}_+ = \mathbf{w} + \xi, \xi \sim B_r^d(0)$
- ▷ PSGD [GHJY15] $\mathbf{w}_+ = \mathbf{w} - \nabla f_{\mathbf{z}_i}(\mathbf{w}) + \xi, \xi \sim N(0, I)$

Escaping Saddles with Stochastic Gradients

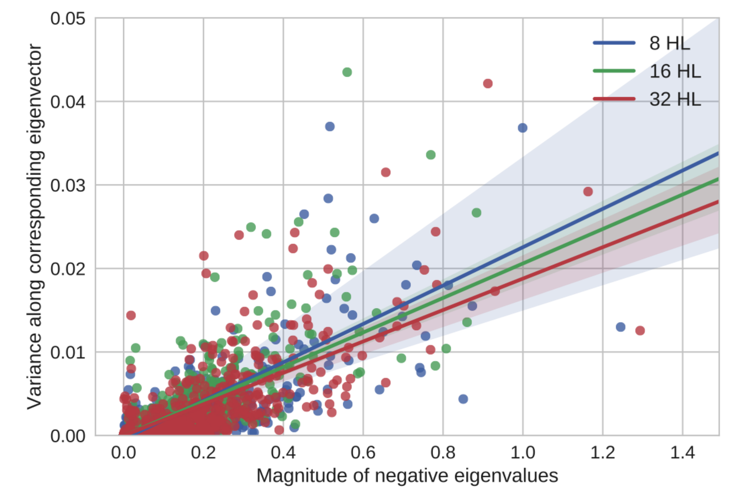
- ▷ Is noise isotropy necessary for escaping saddles?
- ▷ Is the inherent noise of stochastic gradients sufficient to escape from saddles of the training loss of neural networks?

$$\mathbf{w}_+ = \mathbf{w} - \nabla f_{\mathbf{z}_i}(\mathbf{w}) + \xi \quad \Rightarrow \quad \mathbf{w}_+ = \mathbf{w} - \nabla f_{\mathbf{z}_i}(\mathbf{w})$$

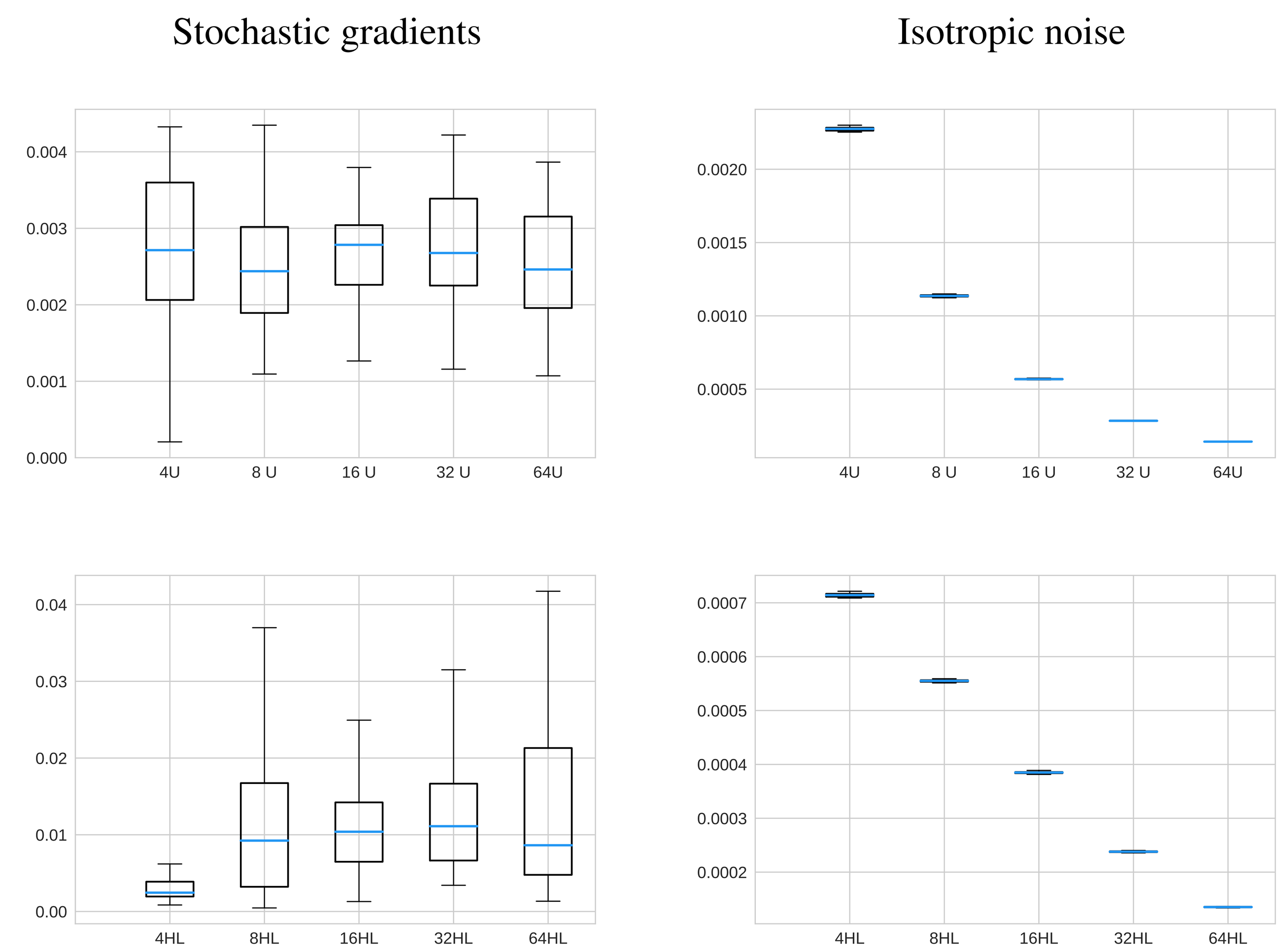
Non-isotropy of Stochastic Gradients

Let $\mathbf{v}_k$ be the eigenvector of $\nabla^2 f(\mathbf{w})$ associated with eigenvalue λ_k. The variance σ_k^2 is defined as

$$\sigma_k^2 \approx \mathbf{E_z}\left(\nabla f_{\mathbf{z}}(\mathbf{w})^\top \mathbf{v}_k(\mathbf{w})\right)^2$$



Variance Along the Extreme Curvature



A Theoretical Lower-bound: Learning Halfspaces

- ▷ Objective function: $f(\mathbf{w}) := \mathbf{E_z}\left[\varphi(\mathbf{w}^\top \mathbf{z})\right]$
- ▷ Loss assumption: $|\varphi''(t)| \leq c|\varphi'(t)|$ that holds for Sigmoid loss.
- ▷ Established lower-bound: We establish a lowerbound on the variance that depends on λ_k as $\mathbf{E_z}\left[(\nabla f_{\mathbf{z}}(\mathbf{w})^\top \mathbf{v}_k)^2\right] \geq (\lambda_k / c)^2$

Correlated Negative Curvature (CNC) Assumption

- ▷ We introduce Correlated Negative Curvature (CNC) as a relaxed noise condition for escaping saddles: There exists a constant $\gamma > 0$ such that $\mathbf{E_z}\langle \mathbf{v_w}, \nabla f_{\mathbf{z}}(\mathbf{w})\rangle^2 > \gamma$ holds for all $\mathbf{w}$ ($\mathbf{v_w}$ denotes the extreme negative curvature of $\nabla^2 f(\mathbf{w})$).
- ▷ The CNC condition is motivated by the variance of stochastic gradients on training objectives of neural networks.

Perturbed GD with Stochastic Gradients

Method	Identification	Perturbation	Termination
PGD [JGN+17]	$\|\nabla f(\mathbf{w}_t)\| < g_{\text{thres}}$	Isotropic perturbation $\mathbf{w}_t = \mathbf{w}_t + \xi$	Certified output
SGD-GD	$\|\nabla f(\mathbf{w}_t)\| < g_{\text{thres}}$	one SGD-step $\mathbf{w}_t = \mathbf{w}_t - r\nabla f_{\mathbf{z}_i}(\mathbf{w}_t)$	**Randomized output** [GL13]

Theorem 1. *There is a parameter choice for SGD-GD such that after $T = \mathcal{O}(\epsilon^{-2})$ steps, this method returns a $\mathbf{w}$ for which*

$$\|\nabla f(\mathbf{w})\| \leq \epsilon, \nabla^2 f(\mathbf{w}) \succeq -\sqrt{\rho}\epsilon^{2/5}\mathbf{I}$$

holds with high probability.

Vanilla SGD

$$\mathbf{w}_{t+1} = \begin{cases} \mathbf{w}_t - r\nabla_{\mathbf{z}} f(\mathbf{w}_t) & t \bmod t_{\text{thres}} = 0 \\ \mathbf{w}_t - \eta\nabla_{\mathbf{z}} f(\mathbf{w}_t) & \text{otherwise} \end{cases}, \text{return } \mathbf{w}_t, t \sim \text{Uniform}\{1, \ldots, T\}.$$

Theorem 2. *There is a parameter choice for SGD (i.e. T,t_{thres},r,η) such that $T = \mathcal{O}(\epsilon^{-4})$ steps of SGD return a $\mathbf{w}$ for which*

$$\|\nabla f(\mathbf{w})\| \leq \epsilon, \nabla^2 f(\mathbf{w}) \succeq -\sqrt{\rho}\epsilon^{2/5}\mathbf{I}$$

holds with high probability.

Comparison on Computational Complexity

Time complexity to obtain an (ϵ_g,ϵ_h)-approximate second-order stationary point, i.e. a parameter $\mathbf{w}$ for which the following holds:

$$\|\nabla f(\mathbf{w})\| \leq \epsilon_g, \nabla^2 f(\mathbf{w}) \succeq -\epsilon_h \mathbf{I}$$

Algorithm	1th-order Complexity	2nd-order Complexity	d-dependency
PSGD [GHJY15]	$\mathcal{O}(d^p\epsilon_g^{-4})$	$\mathcal{O}(d^p\epsilon_h^{-16})$	poly
PGD [JGN+17]	$\mathcal{O}(\log^4(d/\epsilon_g)\epsilon_g^{-2})$	$\mathcal{O}(\log^4(d/\epsilon_h)\epsilon_h^{-4})$	poly-log
SGD+NEON [XY17]	$\tilde{\mathcal{O}}(\epsilon_g^{-4})$	$\tilde{\mathcal{O}}(\epsilon_g^{-8})$	poly-log
CNC-GD	$\mathcal{O}(\epsilon_g^{-2}\log(1/\epsilon_g))$	$\mathcal{O}(\epsilon_h^{-5}\log(1/\epsilon_h))$	free
CNC-SGD	$\mathcal{O}(\epsilon_g^{-4}\log^2(1/\epsilon_g))$	$\mathcal{O}(\epsilon_h^{-10})\log^2(1/\epsilon_h))$	free

References

[DJL+17] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077, 2017.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[GL13] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[JGN+17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

[LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.

[XY17] Yi Xu and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. *arXiv preprint arXiv:1711.01944*, 2017.