# Escaping Saddles with Stochastic Gradients

Hadi Daneshmand*, Jonas Kohler*
Aurelien Lucchi, Thomas Hofmann

March 20th, 2018

data analytics lab

ETH zürich

# Setting

We consider minimizing empirical risk functions

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

where $f$ is sufficiently smooth (but possibly *non-convex*), namely

- gradient map and Hessian are Lipschitz continuous
- stochastic gradients $\nabla f_{\mathbf{z}_i}$ have bounded norms

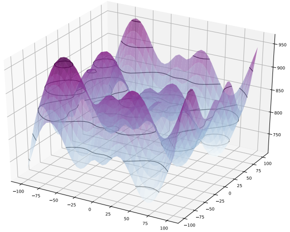
# 2nd-order stationary points

Goal:

convergence to 2nd-order stationary points

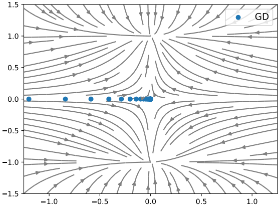
$$\{\mathbf{w} \in \mathbb{R}^d | \nabla f(\mathbf{w}) = 0,\ \nabla^2 f(\mathbf{w}) \succeq 0\}$$

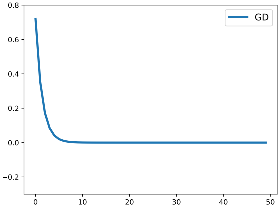
- local minima ✓
- avoid strict saddles $(\nabla^2 f(\mathbf{w}) \not\succeq 0)$



# First order methods and saddle points

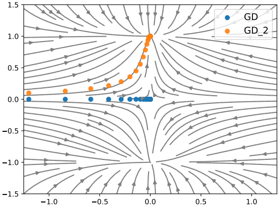


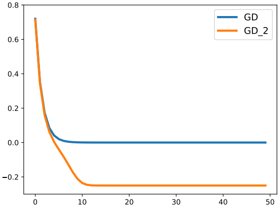



Gradient descent may converge

to strict saddle $\bar{\mathbf{w}}$

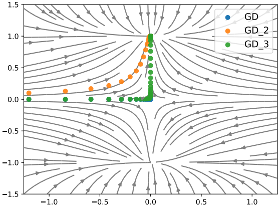


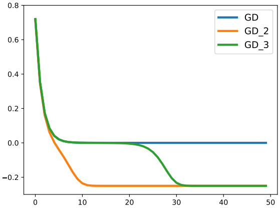




Gradient descent may converge to strict saddle $\bar{\mathbf{w}}$

but

GD is unstable around $\bar{\mathbf{w}}$




Gradient descent may converge

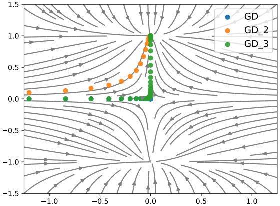
to strict saddle $\bar{\mathbf{w}}$

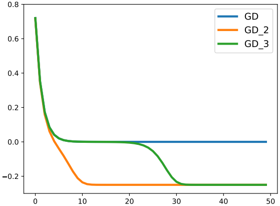
but

attractor set of $\bar{\mathbf{w}}$ is an unstable manifold

$$P(\lim_t \mathbf{w}_t = \bar{\mathbf{w}}) = 0$$

[Lee et al., 2016]




Gradient descent may not converge

to strict saddle $\bar{\mathbf{w}}$

but

it may take exponential time to escape

[Du et al., 2017]

# Escaping saddles with isotropic perturbation



| Method | Perturbation | Noise | Opt. strategy |
|---|---|---|---|
| (1) Cubic Reg. | - | - | 2nd-order |
| (2) PGD | $\triangle \mathbf{w} = \xi$ | $\xi \sim B_r^d(0)$ | 1st-order |
| (3) NGD | $\triangle \mathbf{w} = -\eta_{\mathbf{w}} \nabla f(\mathbf{w}) + \xi$ | $\xi \sim N(0, I)$ | 1st-order |
| (4) PSGD | $\triangle \mathbf{w} = -\nabla f_{\mathbf{z}_i}(\mathbf{w}) + \xi$ | $\xi \sim N(0, I)$ | stochastic 1st |
| (5) SGLD | $\triangle \mathbf{w} = -\eta_{\mathbf{w}} \nabla f_{\mathbf{z}_i}(\mathbf{w}) + \xi$ | $\xi \sim N(0, I)$ | stochastic 1st |

(1) [Nesterov and Polyak, 2006]
(2) [Jin et al., 2017]
(3) [Levy, 2016]
(4) [Ge et al., 2015]
(5) [Zhang et al., 2017]

# The question that we consider here



Is the inherent noise of SGD sufficient to escape from saddles?

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

vs.

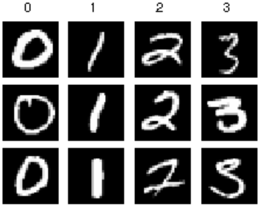
$$\triangle \mathbf{w} = -\nabla f_{\mathbf{z}_i}(\mathbf{w})$$

# Training objective of neural networks

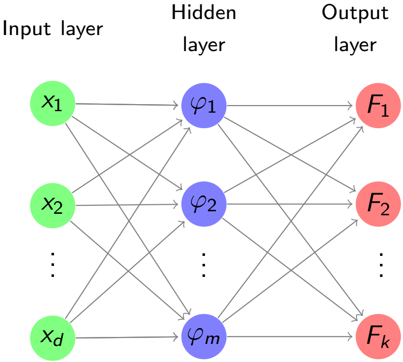


$f($  $)$

MLP with cross-entropy loss




MNIST dataset
(downsized to $10 \times 10$ images)



# Spectral decompostion of the Hessian



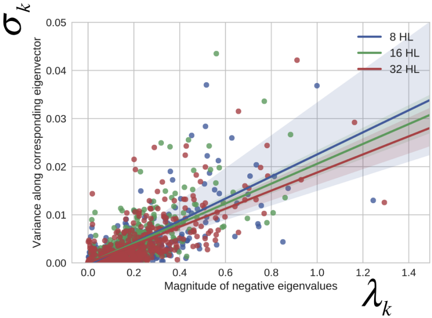
**A**  **with random weights** $w$

$$\nabla^2 f(w) = \begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{n1} & \cdots & v_{nn} \end{pmatrix} \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} v_{11} & \cdots & v_{n1} \\ \vdots & \ddots & \vdots \\ v_{1n} & \cdots & v_{nn} \end{pmatrix}$$

$$v_1(w)$$

# Stochastic gradients are **not** spectrally isotropic

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

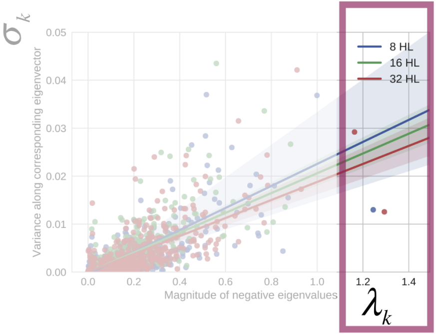



Also observed in [Chaudhari and Soatto, 2017]

# Stochastic gradients are **not** spectrally isotropic

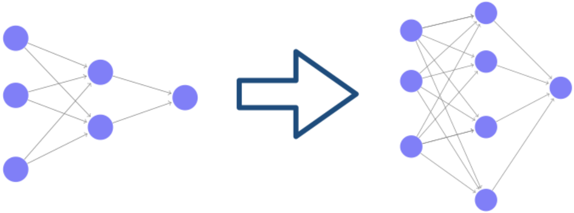


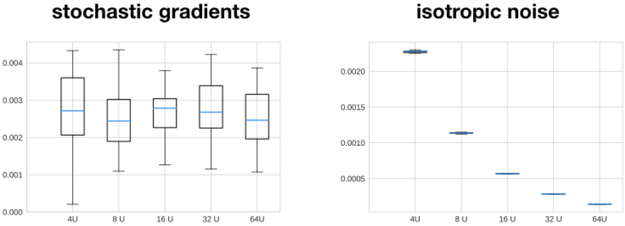
$$\sigma_k \approx \mathbf{E}_{\mathbf{z}} \left( \nabla f_{\mathbf{z}}(\mathbf{w})^\top \mathbf{v}_k(\mathbf{w}) \right)^2$$



# Variance as a function of network width

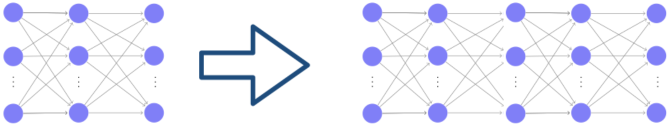


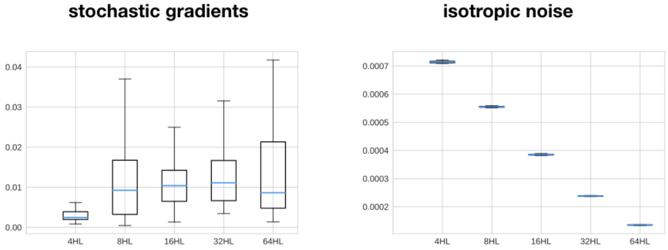





Dependency of the variance along the extreme curvature on the width of NNs

# Variance as a function of network depth






Dependency of the variance along the extreme curvature on the depth of NNs

# A theoretical lower-bound on the variance



Special case: learning halfspaces $f(\mathbf{w}) := \mathbf{E}_{\mathbf{z}}[f_{\mathbf{z}}(\mathbf{w})],\ f_z(\mathbf{w}) := \varphi(\langle \mathbf{w}, \mathbf{z} \rangle)$.

Key assumption on loss: $|\varphi''| \leq c|\varphi'|$ for some $c > 0$

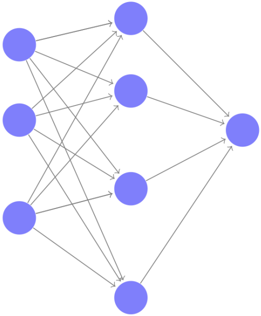
- holds true for Sigmoid loss

## Lemma

*Suppose that $|\varphi''| \leq c|\varphi'|$ and $\|\mathbf{z}\| \leq 1$. Let $\mathbf{v}$ be a unit eigenvector of $\nabla^2 f(\mathbf{w})$ associated with eigenvalue $\lambda < 0$. The second-moment of stochastic gradients along direction $\mathbf{v}$ is lower-bounded as*

$$\mathbf{E}\left[\langle \nabla f_{\mathbf{z}}(\mathbf{w}), \mathbf{v} \rangle^2\right] \geq \left(\frac{\lambda}{c}\right)^2, \quad \forall \mathbf{w}\,.$$

Lower-bound for simple feed-forward network: to be published soon.



$$\mathbf{E}\left[\langle \nabla f_{\mathbf{z}}(\mathbf{w}), \mathbf{v}\rangle^2\right] \geq C\lambda^2$$

# A (relaxed) sufficient escape condition



Isotropy is a too strong requirement, only specific direction(s) of negative curvature matter!

Relaxed sufficient condition for escaping saddles (CNC-condition):

$$\boxed{\exists \gamma > 0 \text{ s.t. } \forall \mathbf{w} : \mathbf{E}_z \left[ \langle \nabla f_z(\mathbf{w}), \mathbf{v}(\mathbf{w}) \rangle^2 \right] > \gamma \,,}$$

where $\mathbf{v}(\mathbf{w})$ is a unit eigenvector corresponding to $\lambda_{min}(\nabla^2 f(\mathbf{w}))$.

**Identification**

$\|\nabla f(\mathbf{w}_t)\| < g_{\text{thres}}$

**Perturbation**

Perturb parameter vector as $\mathbf{w}_t = \mathbf{w}_t + \xi$. Then run $t_{\text{thres}}$ GD steps.

**Certification**

Certifying second-order stationarity as $f(\mathbf{w}_{t+t_{\text{thres}}}) - f(\mathbf{w}_t) > -f_{\text{thres}}$

# Perturbation with stochastic gradients(SGD-GD)



**Identification**

$\|\nabla f(\mathbf{w}_t)\| < g_{\text{thres}}$

**Perturbation**

Take one SGD-step $\mathbf{w}_t = \mathbf{w}_t - r\nabla f_{\mathbf{z}_i}(\mathbf{w}_t)$. Then run $t_{\text{thres}}$ GD steps.

**Randomized output** [Ghadimi and Lan, 2013]

Return one of visited parameters uniformly at random.

# Convergence guarantee



**Theorem**

*There is a choice for parameters of SGD-GD such that after $T = \mathcal{O}((\gamma\epsilon)^{-2}\log(1/(\gamma\epsilon))$ 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.*

# Comparison with the isotropic perturbation



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

| Algorithm | First-order Complexity | Second order Complexity | $d$ Dependency |
|---|---|---|---|
| (1) PGD | $\mathcal{O}(\log^4(d/\epsilon_g)\epsilon_g^{-2})$ | $\mathcal{O}(\log^4(d/\epsilon_h)\epsilon_h^{-4})$ | poly-log |
| SGD-GD | $\mathcal{O}(\log(1/\epsilon_g)\epsilon_g^{-2})$ | $\mathcal{O}(\log(1/\epsilon_h)\epsilon_h^{-5})$ | free |

(1) [Jin et al., 2017]

# Vanilla SGD



SGD($T$,$t_{\text{thres}}$,$r$,$\eta$):

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

return $\mathbf{w}_t,\, t \sim \text{Uniform}\{1, \ldots, T\}$

# Convergence guarantee



**Theorem**

*There is a choice for parameters of SGD($T$,$t_{thres}$,$r$,$\eta$) such that $T = \mathcal{O}((\gamma\epsilon)^{-4}\log^2(1/(\gamma\epsilon)))$ steps of SGD 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.*

# Comparison with isotropic noise perturbation



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

| Algorithm | First-order Comp. | Second order Comp. | $d$ Depend. |
|---|---|---|---|
| (1) PSGD | $\mathcal{O}(d^p \epsilon_g^{-4})$ | $\mathcal{O}(d^p \epsilon_h^{-16})$ | poly |
| (2) SGD+NEON | $\mathcal{O}(\log(d)^p \epsilon_g^{-4})$ | $\mathcal{O}(\log(d)^p \epsilon_g^{-8})$ | poly-log |
| SGD | $\mathcal{O}(\log^2(1/\epsilon_g)\epsilon_g^{-4})$ | $\mathcal{O}(\log^2(1/\epsilon_h)\epsilon_h^{-10})$ | free |

(1) [Ge et al., 2015]
(2) [Xu and Yang, 2017]

# Take-home-message



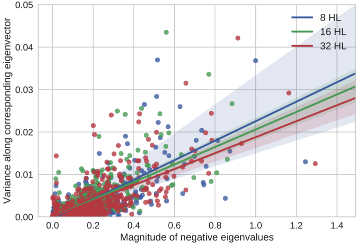
*Training objectives of neural networks (NNs) are special*

*Stochastic gradients signal the negative curvature on NNs*

*Additional perturbation of SGD is not necessary on NNs*

# Future work






Does such a property relate to generalization?

# Thank you!



Poster #206.

I am looking for a post-doc position (Email:hadi.daneshmand@inf.ethz.ch).

Latex template credit: Lilyana Vankova.
Acknowledgements: We would like to thank Antonio Orvieto, Kfir Levy, Gary Becigneul, Yannic Kilcher and Kevin Roth for their helpful discussions.

# References I



[Chaudhari and Soatto, 2017] Chaudhari, P. and Soatto, S. (2017).
Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks.
*arXiv preprint arXiv:1710.11029*.

[Du et al., 2017] Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. (2017).
Gradient descent can take exponential time to escape saddle points.
In *NIPS*, pages 1067–1077.

[Ge et al., 2015] Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015).
Escaping from saddle points—online stochastic gradient for tensor decomposition.
In *COLT*.

# References II



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

[Jin et al., 2017] Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. (2017).
How to escape saddle points efficiently.
*NIPS*.

[Lee et al., 2016] Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016).
Gradient descent converges to minimizers.
*COLT*.

[Levy, 2016] Levy, K. Y. (2016).
The power of normalization: Faster evasion of saddle points.
*arXiv preprint arXiv:1611.04831*.

# References III



[Nesterov and Polyak, 2006] Nesterov, Y. and Polyak, B. T. (2006).
Cubic regularization of newton method and its global performance.
*Mathematical Programming*, 108(1).

[Xu and Yang, 2017] Xu, Y. and Yang, T. (2017).
First-order stochastic algorithms for escaping from saddle points in almost linear time.
*arXiv preprint arXiv:1711.01944*.

[Zhang et al., 2017] Zhang, Y., Liang, P., and Charikar, M. (2017).
A hitting time analysis of stochastic gradient langevin dynamics.
*COLT*.