Continuous vs. Discrete Optimization of Deep Neural Networks

Nadav Cohen

Joint with Omer Elkabetz, NeurIPS 2021 Spotlight

AI Week

7 February 2022

Motivation

Success of deep neural networks (DNNs) is driven by Gradient Descent (**GD**)

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla f(\boldsymbol{\theta}_k)$$

Approaches for theoretical analysis:

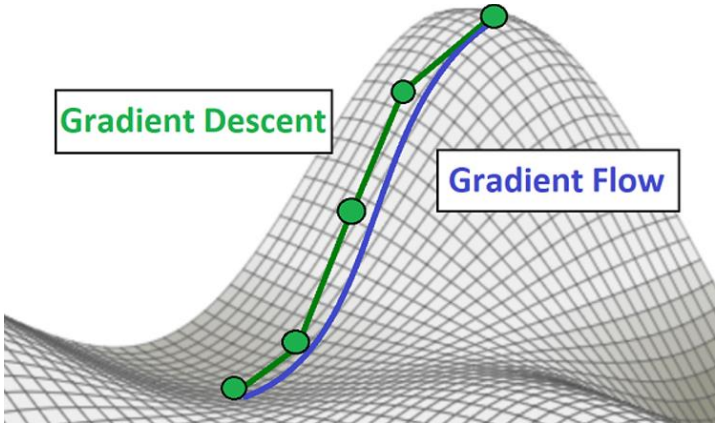
- Analyze continuous version of GD – Gradient Flow (**GF**)
- Directly analyze GD

$$\frac{d}{dt}\boldsymbol{\theta}(t) = -\nabla f(\boldsymbol{\theta}(t))$$

GF easier to analyze than GD, but unrealistic

Open Question

Does GF over DNNs represent GD?



Background: Numerical Integration

Differential Equation $\frac{d}{dt}\boldsymbol{\theta}(t) = \boldsymbol{g}\big(\boldsymbol{\theta}(t)\big)$ $\quad \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$

step size (predetermined)

Numerical approximation:

Euler's method

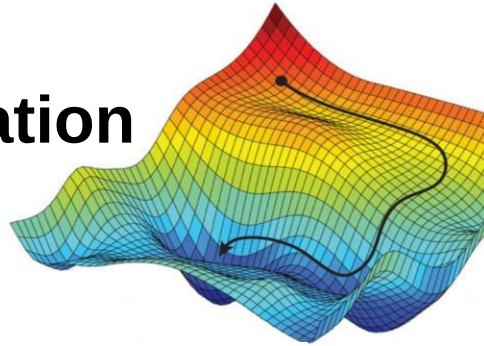
$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta \boldsymbol{g}(\boldsymbol{\theta}_k) \qquad \boldsymbol{\theta}_k \approx \boldsymbol{\theta}(k\eta)$$

Fundamental Theorem [Hairer et al. 1993]

Numerical error

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_{k=t/\eta}\| \lesssim \eta t \exp\Big(\int_0^t \lambda_{\max}\big(J_{\boldsymbol{g}(\boldsymbol{\theta}(t'))}\big)\, dt' \Big)$$

Numerical Integration $g = -\nabla f$ Optimization



Differential Equation

$$\frac{d}{dt}\boldsymbol{\theta}(t) = \boldsymbol{g}\big(\boldsymbol{\theta}(t)\big)$$

Euler's method

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta \boldsymbol{g}\big(\boldsymbol{\theta}_k\big)$$

Numerical error

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_{k=t/\eta}\| \lesssim \eta t \exp\Big(\int_0^t \lambda_{\max}\big(J_{\boldsymbol{g}(\boldsymbol{\theta}(t'))}\big)\, dt'\Big)$$

GF

$$\frac{d}{dt}\boldsymbol{\theta}(t) = -\nabla f\big(\boldsymbol{\theta}(t)\big)$$

GD

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla f\big(\boldsymbol{\theta}_k\big)$$

GF-GD distance

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_{k=t/\eta}\| \lesssim \eta t \exp\Big(-\int_0^t \lambda_{\min}\big(\nabla^2 f(\boldsymbol{\theta}(t'))\big)\, dt'\Big)$$

GF-GD Distance Depends on Convexity

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_{k=t/\eta}\| \lesssim \eta t \exp\left(-\int_0^t \lambda_{\min}\left(\nabla^2 f(\boldsymbol{\theta}(t'))\right) dt'\right)$$

Small enough step size η guarantees ϵ-distance. How small?

Coarsely taking $\lambda_{\min} := \inf_q \lambda_{\min}\left(\nabla^2 f(\boldsymbol{q})\right)$ instead of $\lambda_{\min}\left(\nabla^2 f\left(\boldsymbol{\theta}(t')\right)\right)$ yields:

Strongly convex	*Non-strongly convex*	*Non-convex*
$(\lambda_{\min} > 0)$	$(\lambda_{\min} = 0)$	$(\lambda_{\min} < 0)$
$\eta \lesssim \epsilon$	$\eta = \epsilon / t$	$\eta \lesssim \epsilon / t e^{\lvert\lambda_{\min}\rvert t}$

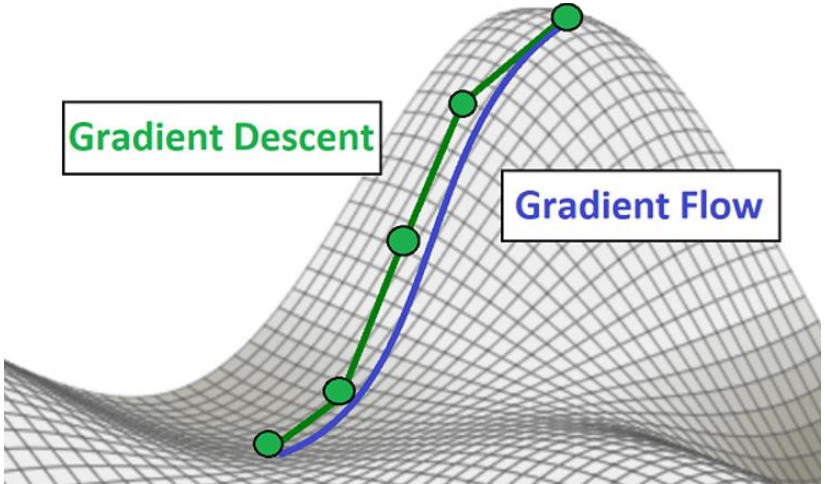
Claim: exist settings where non-convex bound on η is tight

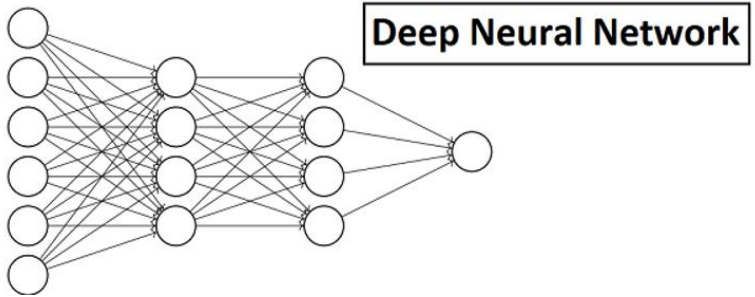
Problem: in worst case, exp small η needed for non-convex objective

DNN Optimization is Roughly Convex

Theorem: min eigenvalue of Hessian along GF trajectory over (homogeneous) DNN init near zero is only slightly negative

Suggests GF ≈ GD





Translating Continuous Analysis to Discrete Result

Setup: linear DNN (arbitrarily deep), scalar output

Proposition: GF → global min almost surely under random near zero init

⇓

GF-GD Translation Machinery

Theorem: min eigenvalue of Hessian along GF trajectory over (homogeneous) DNN init near zero is only slightly negative

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_{k=t/\eta}\| \lesssim \eta t \exp\left(-\int_0^t \lambda_{\min}\left(\nabla^2 f(\boldsymbol{\theta}(t'))\right) dt'\right)$$

⇓

Theorem: GD efficiently → global min almost surely under random near zero init

Translating Continuous Analysis to Discrete Result

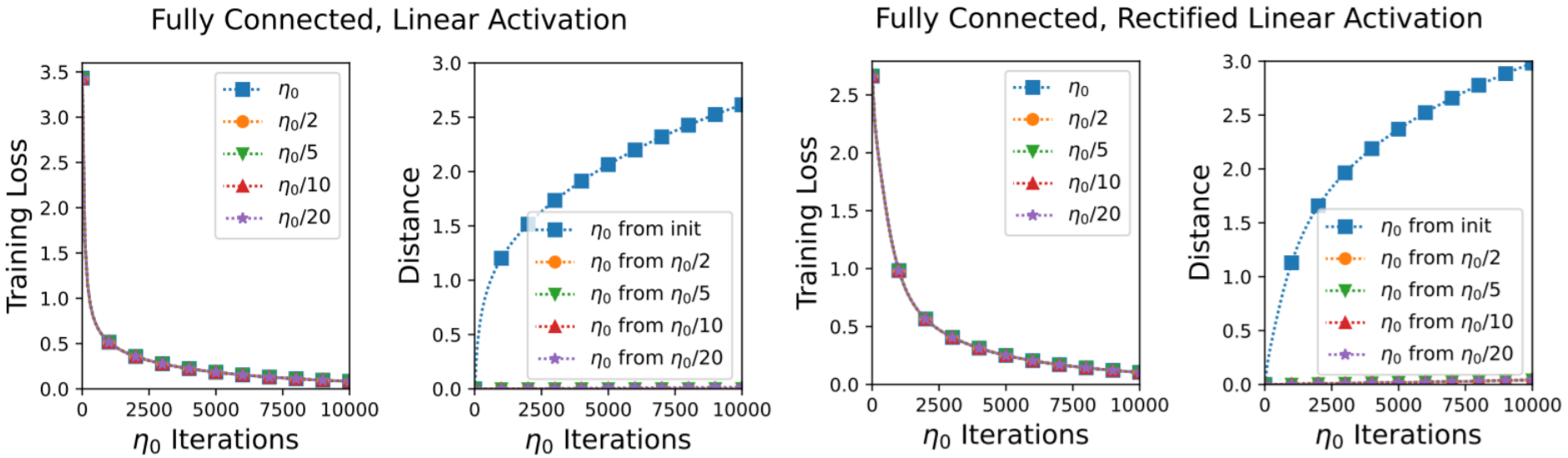
Theorem: GD efficiently → global min almost surely under random near zero init

First guarantee of GD over fixed size DNN (depth ≥ 3) efficiently converging to global min *almost surely* under random init!

We not only know GD reaches global min, but also its path (sheds light on implicit regularization)

Experiments

Over simple DNNs, indeed GF ≈ GD



Similar results for convolutional networks

$(\text{MNIST}, \eta_0 = 0.001)$

Ongoing Work: Large Step Size Regime

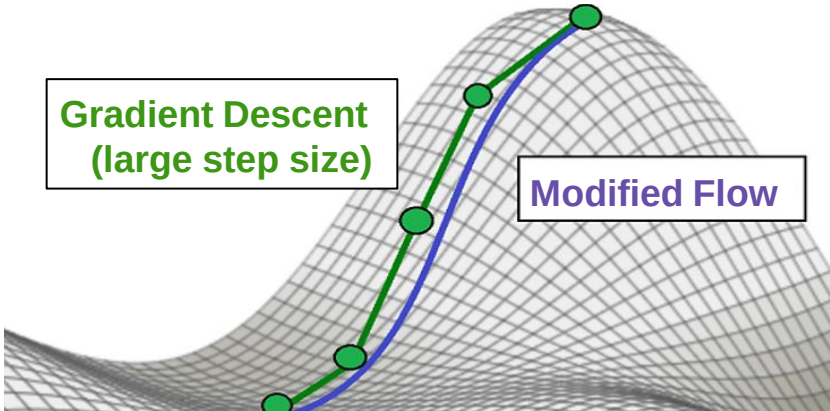
Recent evidence: large step size for GD can improve generalization

New variants of GF aim to capture GD with large step size

$$\frac{d}{dt}\boldsymbol{\theta}(t) = -\nabla f\big(\boldsymbol{\theta}(t)\big) - \frac{\eta}{2}\frac{d^2}{dt^2}\boldsymbol{\theta}(t)$$

[Smith et al. 2021, Barrett & Dherin 2020, Kunin et al. 2020]

Ongoing work: adapt our analysis to account for such variants



Conclusion

- GF-GD distance is small if landscape along GF trajectory is “roughly convex”
- “Rough convexity” holds along GF trajectories over (homogeneous) DNNs
- Translation of GF analysis to GD ⇒ first convergence guarantee of its kind!
- Experiments with simple DNNs verify GF ≈ GD

Hypothesis: GF will unravel mysteries behind deep learning

Thank you!

Work supported by a Google Research Scholar Award, a Google Research Gift, the Yandex Initiative in Machine Learning, the Israel Science Foundation (grant 1780/21), Len Blavatnik and the Blavatnik Family Foundation, and Amnon and Anat Shashua.