Continuous vs. Discrete Optimization of Deep Neural Networks

Nadav Cohen

Tel Aviv University

International Conference on Machine Learning (ICML)

Workshop on Continuous Time Perspectives in Machine Learning

23 July 2022

Source

Continuous vs. Discrete Optimization of Deep Neural Networks

Omer Elkabetz
Tel Aviv University
omer.elkabetz@cs.tau.ac.il

Nadav Cohen
Tel Aviv University
cohennadav@cs.tau.ac.il

Abstract

Existing analyses of optimization in deep learning are either continuous, focusing on (variants of) gradient flow, or discrete, directly treating (variants of) gradient descent. Gradient flow is amenable to theoretical analysis, but is stylized and disregards computational efficiency. The extent to which it represents gradient descent is an open question in the theory of deep learning. The current paper studies this question. Viewing gradient descent as an approximate numerical solution to the initial value problem of gradient flow, we find that the degree of approximation depends on the curvature around the gradient flow trajectory. We then show that over deep neural networks with homogeneous activations, gradient flow trajectories enjoy favorable curvature, suggesting they are well approximated by gradient descent. This finding allows us to translate an analysis of gradient flow over deep linear neural networks into a guarantee that gradient descent efficiently converges to global minimum *almost surely* under random initialization. Experiments suggest that over simple deep neural networks, gradient descent with conventional step size is indeed close to gradient flow. We hypothesize that the theory of gradient flows will unravel mysteries behind deep learning.[1]

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2021

SPOTLIGHT

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

Motivation

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

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

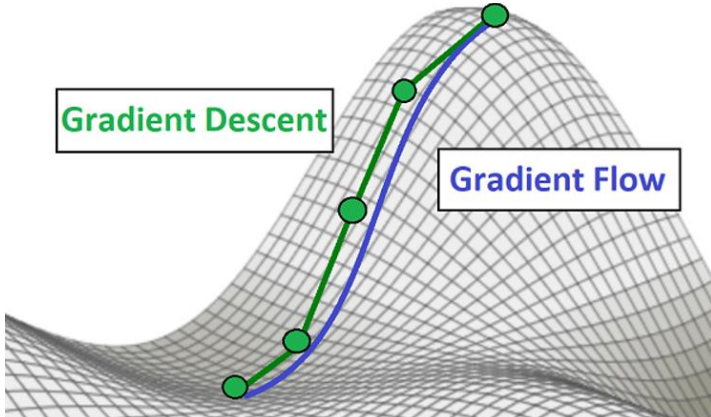
Approaches for theoretical analysis:

- **Continuous:** analyze Gradient Flow (**GF**), i.e. take $\eta \to 0$

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

- **Discrete:** directly analyze GD, i.e. treat $\eta > 0$

Often more tractable, but unrealistic!



Open Question

Does GF over DNNs represent GD?

Background: Numerical Integration

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

Numerical approximation:

Euler's method

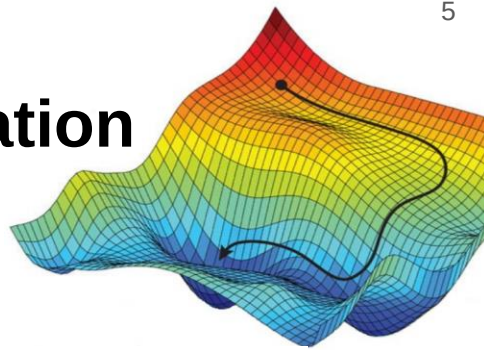
step size (predetermined)

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

Fundamental Theorem [Hairer et al. 1993]

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

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, exponentially small η needed for non-convex objective

Worst Case Scenario: Proof Sketch

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

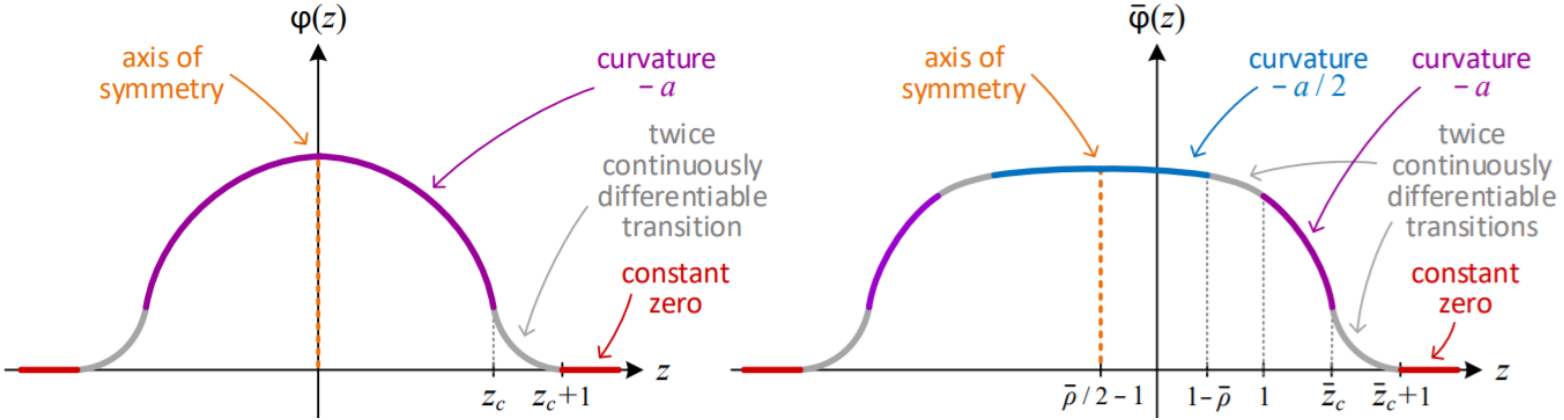
$$\eta \lesssim \epsilon / t e^{|\lambda_{\min}| t}$$

$$(\lambda_{\min} < 0)$$

Consider:

$$f(\theta_1, \theta_2) = \varphi(\theta_1) + \bar{\varphi}(\theta_2)$$

where:



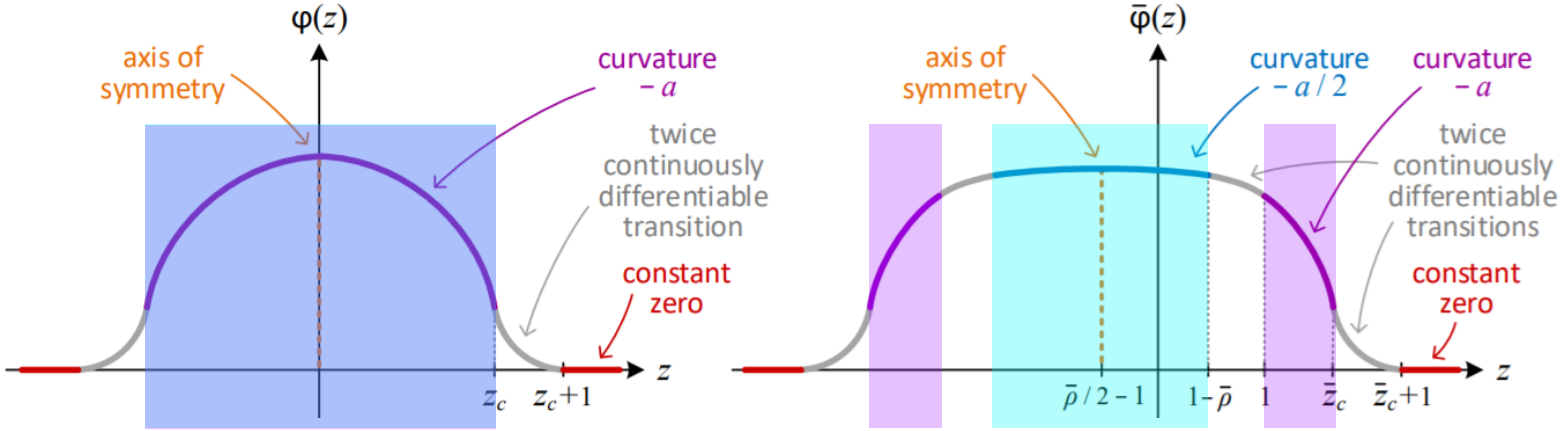
Worst Case Scenario: Proof Sketch

$$f(\theta_1, \theta_2) = \varphi(\theta_1) + \bar{\varphi}(\theta_2)$$

Regions in weight space:

anisotropic

isotropic



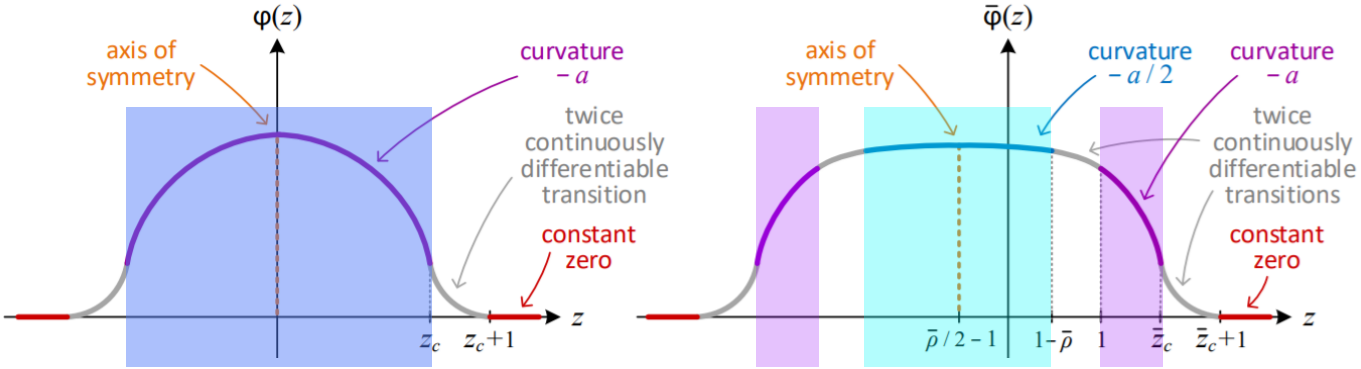
Worst Case Scenario: Proof Sketch

When init in anisotropic region, both GF and GD continue to isotropic one

At entrance to isotropic region, GF-GD discrepancy is proportional to step size η

Throughout isotropic region, discrepancy grows exponentially with time, i.e. as e^{at}

For discrepancy at time t less than ϵ, must have $\eta \in \mathcal{O}(e^{-at}\epsilon)$



anisotropic region

isotropic region

DNN Optimization is Roughly Convex

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

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

What about DNNs?

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

$\Longrightarrow$ for (homogeneous) DNN init near zero:

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

GF ≈ GD

DNN Optimization is Roughly Convex: Proof Sketch*

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

** For linear DNNs; non-linear DNNs treated similarly (where differentiable)*

Linear DNN:

$$\boldsymbol{\theta} = (W_1, W_2, \ldots, W_n) \qquad h_{\boldsymbol{\theta}}(x) = W_{n:1}x$$

end-to-end matrix
$W_{n:1} := W_n W_{n-1} \cdots W_1$

Training objective:

$$f(\boldsymbol{\theta}) = \phi(W_{n:1})$$

end-to-end objective

DNN Optimization is Roughly Convex: Proof Sketch*

Hessian:

$$W_{j':j} := W_{j'} W_{j'-1} \cdots W_1$$

$$\nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \Delta W_2, \ldots, \Delta W_n] = \nabla^2 \phi(W_{n:1}) \left[\sum_{j=1}^{n} W_{n:j+1} (\Delta W_j) W_{j-1:1} \right]$$
$$+ 2\mathrm{Tr}\left(\nabla \phi(W_{n:1})^\top \sum_{1 \le j < j' \le n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \right)$$

Implies:

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \gtrsim -\|\nabla \phi(W_{n:1})\|_{\text{Frobenius}} \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}| = n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_{\text{spectral}}$$

GF init near zero maintains balancedness: [Du et al. 2018]

$$\forall j : W_j W_j^\top \approx W_{j+1}^\top W_{j+1}$$

Under balancedness:

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \gtrsim -\underbrace{\|\nabla \phi(W_{n:1})\|_{\text{Frobenius}}}_{\text{small at convergence}} \underbrace{\|W_{n:1}\|_{\text{spectral}}^{1-2/n}}_{\text{small at init}}$$

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

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

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

⇓

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)

GF Convergence: Proof Sketch

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

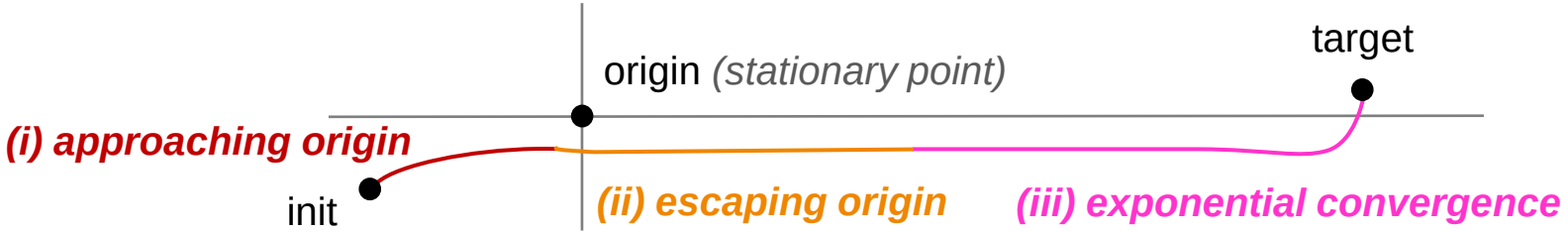
End-to-end matrix:

$$W_{n:1} := W_n W_{n-1} \cdots W_1$$

Dynamics induced by GF init near zero: [Arora et al. 2018]

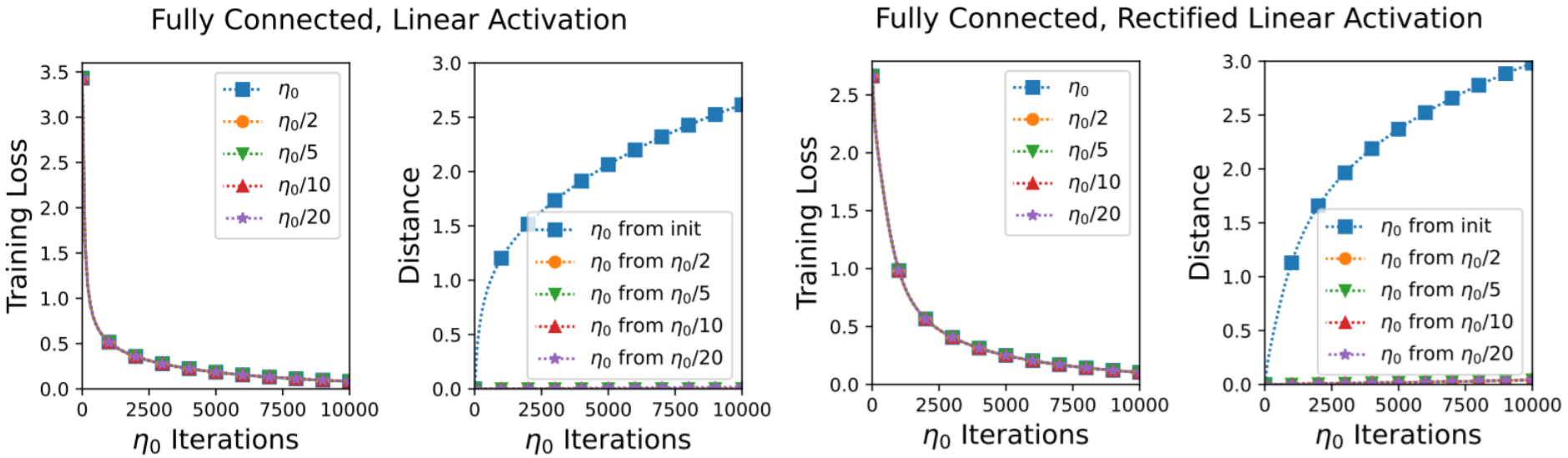
$$\frac{d}{dt} W_{n:1}(t) = -\nabla \phi\big(W_{n:1}(t)\big) \Big(\|W_{n:1}(t)\|_{Frobenius}^{2-2/n} I_{d_0} + (n-1) \big[W_{n:1}^{\top}(t) W_{n:1}(t)\big]^{1-1/n} \Big)$$

Using dynamics, we establish three phases of optimization:



Experiments

Over simple DNNs, indeed GF ≈ GD



Similar results for convolutional networks

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

Future Work: Large Step Size Regime

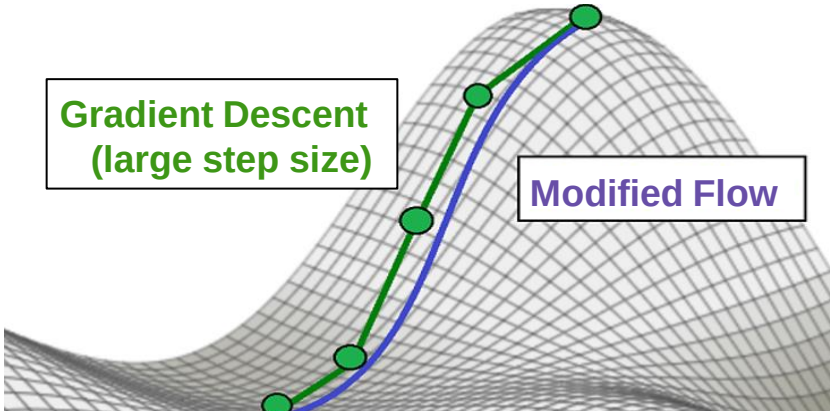
Recent evidence: large step size for GD can improve generalization

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]

Future 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!