Algorithmic Regularization: Bias Us Toward “Simple” Models

Shihua Zhang

October 23, 2024

Table of Contents

Deep learning achieves big successes

- The rise of deep learning in various applications
  - Image classification, semantic segmentation
  - Natural language processing
  - AlphaGo
  - AlphaFold
  - ...

Deep learning achieves big successes

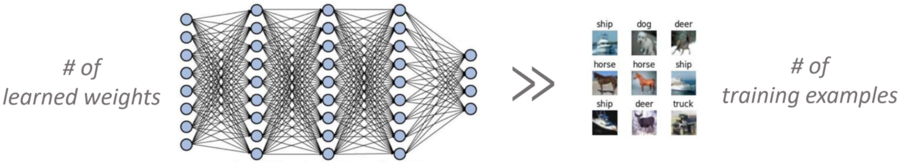
- The rise of deep learning in various applications
  - Image classification, semantic segmentation
  - Natural language processing
  - AlphaGo
  - AlphaFold
  - ...

- Deep models often generalize well even without explicit regularization

- Algorithmic regularization: the optimization algorithm biases us toward a "simple" model that generalize well

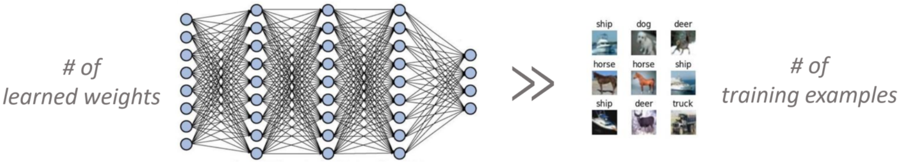
Generalaization in deep learning

Deep neural networks are typically over-parameterized

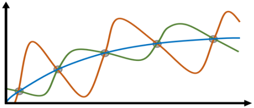


Generalaization in deep learning

Deep neural networks are typically over-parameterized

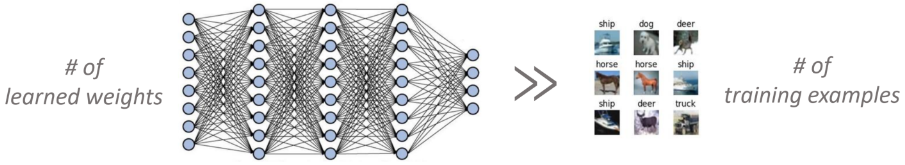


Many possible solutions fit training data

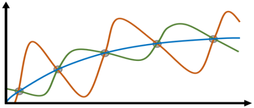


Generalaization in deep learning

Deep neural networks are typically over-parameterized



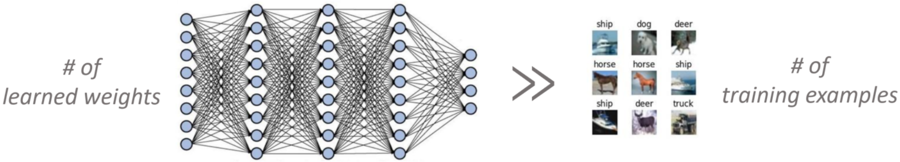
Many possible solutions fit training data



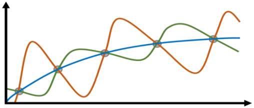
Variants of gradient descent (GD) usually find solutions that generalize well

Generalaization in deep learning

Deep neural networks are typically over-parameterized



Many possible solutions fit training data



Variants of gradient descent (GD) usually find solutions that generalize well

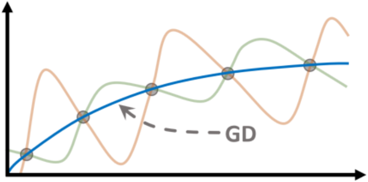
↑

Even without explicit regularization!

Implicit regularization

Imiplict regularization prefers “simpler” models

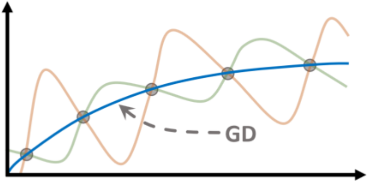
- GD fits trainning data with predictors of lowest possible complexity



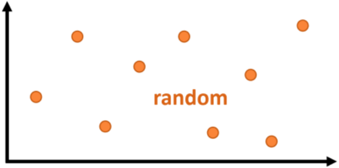
Implicit regularization

Imiplict regularization prefers “simpler” models

- GD fits trainning data with predictors of lowest possible complexity



- Natural data can be fit with low complexity, other data cannot



Challenge: how to formalize the implicit regularization?

Goal

Mathematically formalize implicit regularization in deep learning

Challenge: how to formalize the implicit regularization?

Goal

Mathematically formalize implicit regularization in deep learning

Approach

- Start with simple models and standard GD algorithms
- Investigate the implicit bias for variants of GD on general models

Table of Contents

Let's start with a simple model

Consider linear regression with the squared loss function

Empirical risk minimization

$$L(w) = \sum_{i=1}^{n} (w^T x^{(i)} - y^{(i)})^2$$

- $n < d$ and the objective function is realizable, i.e., $\min_w L(w) = 0$

Let's start with a simple model

Consider linear regression with the squared loss function

Empirical risk minimization

$$L(w) = \sum_{i=1}^{n} (w^T x^{(i)} - y^{(i)})^2$$

- $n < d$ and the objective function is realizable, i.e., $\min_w L(w) = 0$
- The objective function has multiple global minima

$$\mathcal{G} = \{w : \forall i, w^T x^{(i)} = y^{(i)}\}$$

GD induces a unique minimum

Proposition 1 ([GLSS18])

Consider GD updates w_t starting with w_0. For any step-size schedule that minimizes $L(w)$, the algorithm returns a special global minimizer that implicitly also minimizes the Euclidean distance to w_0:

$$w_t \to \arg\min_{w \in \mathcal{G}} \|w - w_0\|_2^2 \tag{1}$$

GD induces a unique minimum

Proposition 1 ([GLSS18])

Consider GD updates w_t starting with w_0. For any step-size schedule that minimizes $L(w)$, the algorithm returns a special global minimizer that implicitly also minimizes the Euclidean distance to w_0:

$$w_t \to \arg\min_{w \in \mathcal{G}} \|w - w_0\|_2^2 \tag{1}$$

- GD implicitly induces a unique minimum that also minimizes the Euclidean distance to w_0

Proof sketch

Proof.

Note that $\forall w$, $\nabla L(w) = \sum_i (w^T x^{(i)} - y^{(i)}) x^{(i)} \in \text{span}(x^{(i)})$.
The gradients are restricted to a n dimensional subspace that is independent of w. The GD updates from initialization w_0, thus $w_t - w_0 = \sum_{t'<t} \eta \nabla L(w_{t'})$ are also constrained to the n dimensional subspace.

Proof sketch

Proof.

Note that $\forall w$, $\nabla L(w) = \sum_i (w^T x^{(i)} - y^{(i)}) x^{(i)} \in \text{span}(x^{(i)})$.
The gradients are restricted to a *n* dimensional subspace that is independent of w. The GD updates from initialization w_0, thus $w_t - w_0 = \sum_{t'<t} \eta \nabla L(w_{t'})$ are also constrained to the n dimensional subspace.
There exists a unique global minimizer that both fits the data ($w \in \mathcal{G}$) and is reachable by GD $w \in w_0 + \text{span}(x^{(i)})$. It is exactly the KKT condition of

$$\min_{w \in \mathcal{G}} ||w - w_0||_2^2 \tag{2}$$

which completes the proof $\square$

Table of Contents

Geometry induced by updates of local search algorithm

GD iterations can be alternatively specified as a local approximation while constraining the step length

$$w_{t+1} = \arg\min_{w} \langle w, \nabla L(w_t) \rangle + \frac{1}{2\eta} \|w - w_t\|_2^2 \tag{3}$$

Motivated by this connection, we can study other families of algorithms that work under different geometries

Geometry induced by updates of local search algorithm

GD iterations can be alternatively specified as a local approximation while constraining the step length

$$w_{t+1} = \arg\min_{w} \langle w, \nabla L(w_t) \rangle + \frac{1}{2\eta} \|w - w_t\|_2^2 \tag{3}$$

Motivated by this connection, we can study other families of algorithms that work under different geometries

- Mirror descent w.r.t. Bregman divergence with potential ψ
- Steepest descent w.r.t. general norms

Mirror descent

Mirror descent w.r.t. Bregman divergence with potential ψ

Mirror descent updates are defined for any strongly convex and differentiable potential ψ as

$$\begin{aligned} w_{t+1} &= \arg\min_{w} \eta\langle w, \nabla L(w_t)\rangle + D_{\psi}(w, w_t) \\ &\Rightarrow \nabla\psi(w_{t+1}) = \nabla\psi(w_t) - \eta\nabla L(w_t) \end{aligned} \tag{4}$$

where $D_{\psi}(w, w') = \psi(w) - \psi(w') - \langle\nabla\psi(w'), w - w'\rangle$ is the Bregman divergence.

- $\psi(w) = \frac{1}{2}||w||_2^2$ leads to gradient descent
- Entropy potential $\psi(w) = \sum_i w[i]\log w[i] - w[i]$

Mirror update induced minima

Theorem 1 ([BT03])

For any realizable dataset $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$, and any strongly convex potential ψ, consider the mirror descent iterates w_t that minimizes $L(w)$. For w_0, if the step-size schedule minimizes $L(w)$, then then the asymptotic solution of the algorithm is given by

$$w_t \to \arg\min_{w \in \mathcal{G}} D_{\psi}(w, w_0) \tag{5}$$

Steepest descent

GD is also a special case of steepest descent (SD) w.r.t. a generic norm $||\cdot||$

Steepest descent w.r.t. general norms

$$w_{t+1} = w_t + \eta_t \Delta w_t, \text{ where } \Delta w_t = \arg\min_v \langle \nabla L(w_t), v \rangle + \frac{1}{2}||v||^2 \quad (6)$$

- ℓ_2 norm leads to gradient descent
- ℓ_1 norm leads to coordinate descent

Steepest descent

GD is also a special case of steepest descent (SD) w.r.t. a generic norm $||\cdot||$

Steepest descent w.r.t. general norms

$$w_{t+1} = w_t + \eta_t \Delta w_t, \text{ where } \Delta w_t = \arg\min_v \langle \nabla L(w_t), v \rangle + \frac{1}{2}||v||^2 \quad (6)$$

- ℓ_2 norm leads to gradient descent
- ℓ_1 norm leads to coordinate descent

- We may expect the steepest descent iterates to converge to the solution closest to w_0 in the corresponding norm
- It is only true for quadratic norms $||v||_D = \sqrt{v^T D v}$
- Unfortunately, it does not hold for general norms

Example: the global minimum depends on the step size

Consider the dataset
$\{(x^{(1)} = [1, 1, 1], y^{(1)} = 1), (x^{(2)} = [1, 2, 0], y^{(2)} = 10)\}$ using steepest descent updates w.r.t. $\ell_{4/3}$ norm

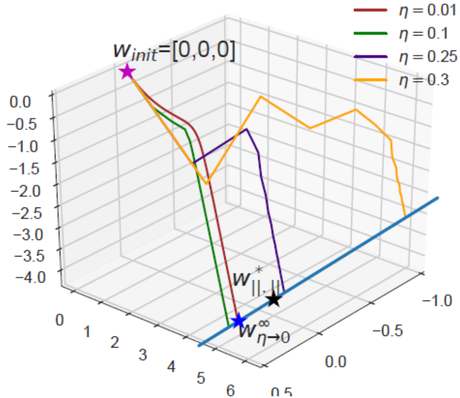


Table of Contents

Matrix factorization as a prediction problem

Matrix completion: recover an unknown matrix given its subset of entries

Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?

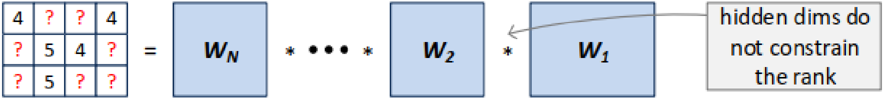
observations $\{y_{ij}\}_{(i,j)\in\Omega}$

$n \times p$ matrix completion $\Longleftrightarrow$ prediction from $\{1, \cdots, n\} \times \{1, \cdots, p\}$ to $\mathbb{R}$

Matrix Factorization $\longleftrightarrow$ Linear Neural Network

Matrix Factorization (MF)

Parameterize solution as product of matrices and fit observations via GD



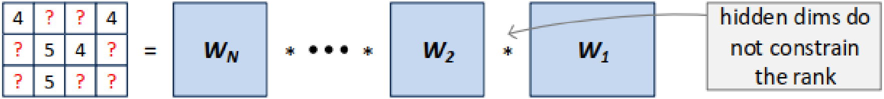
$$\min_{W_1,\ldots,W_N} \sum_{(i,j)\in\Omega} \left([W_N W_{N-1} \cdots W_1]_{ij} - y_{ij} \right)^2$$

MF $\longleftrightarrow$ matrix completion via linear NN (with no explicit regularization)

Matrix Factorization $\longleftrightarrow$ Linear Neural Network

Matrix Factorization (MF)

Parameterize solution as product of matrices and fit observations via GD



$$\min_{W_1,\dots,W_N} \sum_{(i,j)\in\Omega} \left([W_N W_{N-1} \cdots W_1]_{ij} - y_{ij}\right)^2$$

MF $\longleftrightarrow$ matrix completion via linear NN (with no explicit regularization)

Empirical phenomenon [GWB$^+$18]

MF (with small init and step size) accurately recovers low rank matrices

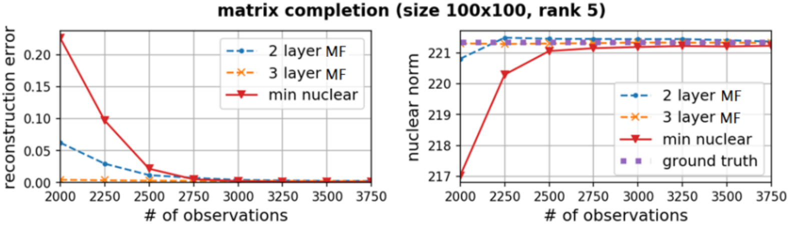
Implicit regularization of GD for MF

Classic results [CR09]

If (i) unknown matrix has low rank; (ii) observations are sufficiently many, then minimizing nuclear norm yields accurate recovery

Conjecture [GWB$^+$18]

MF of depth 2 (with small init and step size) fits observations while minimizing nuclear norm



Dynamical analysis of implicit regularization

Denote: $W_e := W_d \cdots W_1$ – end matrix of MF, $\{\sigma_r\}_r$ – singular vals of W_e

Theorem 2 ([ACHL19])

In training MF of depth d (with small init and step size): $\frac{d}{dt}\sigma_r \propto \sigma_r^{2-2/d}$

Depth speeds up (slows down) large (small) singular vals!

Dynamical analysis of implicit regularization

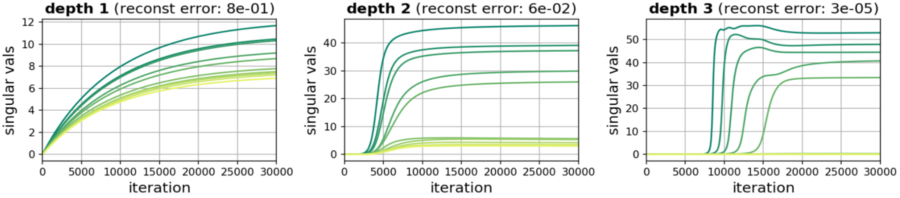
Denote: $W_e := W_d \cdots W_1$ – end matrix of MF, $\{\sigma_r\}_r$ – singular vals of W_e

Theorem 2 ([ACHL19])

In training MF of depth d (with small init and step size): $\frac{d}{dt}\sigma_r \propto \sigma_r^{2-2/d}$

Depth speeds up (slows down) large (small) singular vals!

Completion of low rank matrix via MF



MF depth leads to larger gaps between singular vals (lower rank)!

Table of Contents

Linear models in classification

- Consider linear classification with exponential loss
  $\ell(u, y) = \exp(-uy)$

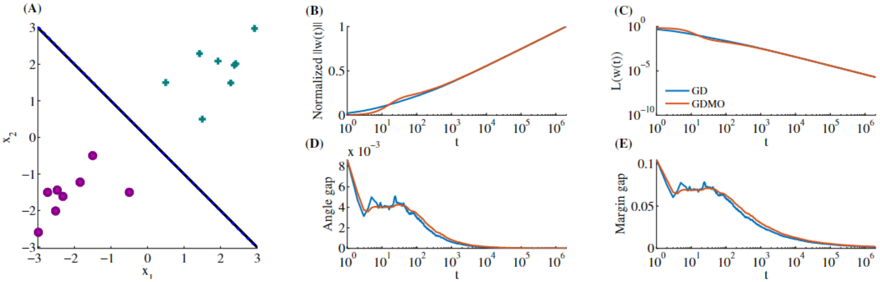
Empirical risk minimization

$$L(w) = \sum_{i=1}^{n} \exp\left(-y^{(i)} w^T x^{(i)}\right)$$

where $y^{(i)} \in \{-1, 1\}$

- Similarly, we consider the gradient descent and steepest descent

Empirical phenomena of GD



- (A) The asymptotic solution of GD coincides with the Max-Margin separator
- (B) $\|w(t)\|$ increases logarithmically
- (C) The loss decrease as t^{-1}
- GD with momentum (GDMO) behaviors similarly

Gradient descent induces ℓ_2 max-margin vector

Theorem 3 ([SHN$^+$18])

For any dataset which is linearly separable, any β-smooth decreasing loss function with an exponential tail, any stepsize $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(X)$, where X is the data matrix and any starting point $w(0)$, the GD iterates will behave as:

$$w(t) = \hat{w}\log t + \rho(t),$$

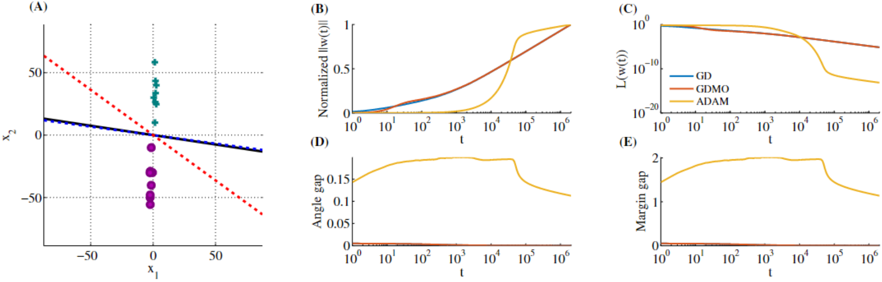
where $\hat{w}$ is the L_2 max margin vector (the solution to hard margin SVM):

$$\hat{w} = \arg\max_{w\in\mathbb{R}^d} \|w\|^2 \ s.t.\ w^T x_n \geqslant 1$$

and the residual grows at most as $\|\rho\| = O(\log\log(t))$, and so

$$\lim_{t\to\infty} \frac{w(t)}{\|w(t)\|} = \frac{\hat{w}}{\|\hat{w}\|}$$

Different algorithms behaves differently



- (A) ADAM [KB15] does not converges to the Max-Margin solution
- GD and GDMO converges to the Max-Margin solution

Implicit bias of steepest descent

Theorem 4 ([GLSS18])

For any separable dataset and any norm $||\cdot||$, consider the steepest descent updates for minimizing $L(w)$ with the exponential loss $\ell(u,y) = exp(-uy)$. For all initialization w_0, and all bounded step-sizes satisfying $\eta_t \leqslant \min\{\eta_+, \frac{1}{B^2 L(w_t)}\}$ where $B := \max_n ||x_n||_*$, $||x||_* := \sup_{||y||\leqslant 1} ||x^T y||$ and $\eta_+ < \infty$. The iterates w_t satisfy

$$\lim_{t\to\infty} \min_n \frac{y_i\langle w_t, y_i\rangle}{||w_t||} = \max_{w:||w||\leqslant 1} \min_n y_i\langle w, x_i\rangle =: \gamma.$$

If the maximum-$||\cdot||$ margin solution $w^* = \arg\max_{||w||\leqslant 1} \min_i y_i\langle w_t, y_i\rangle$ exist, then the direction satisfy $\lim_{t\to\infty} \frac{w_t}{||w_t||_2} = w^*$

- It is a generalization of Theorem 3

The implicit bias of GD for importance weighting

- Assigning importance weights to instances is common practice

$$L(\theta; w) = \frac{1}{N}\sum_{i=1}^{N} w_i \ell(y_i f(\theta, x_i)),$$

  where θ is the parameter of the network and $w_i \in [1/M, M]$ is the bounded importance weight

- [BL19] observes that the effect of importance weights diminishes as the training proceeds

- Question: What is the implicit bias of GD in the presence of importance weights?

The effect of importance weights diminishes

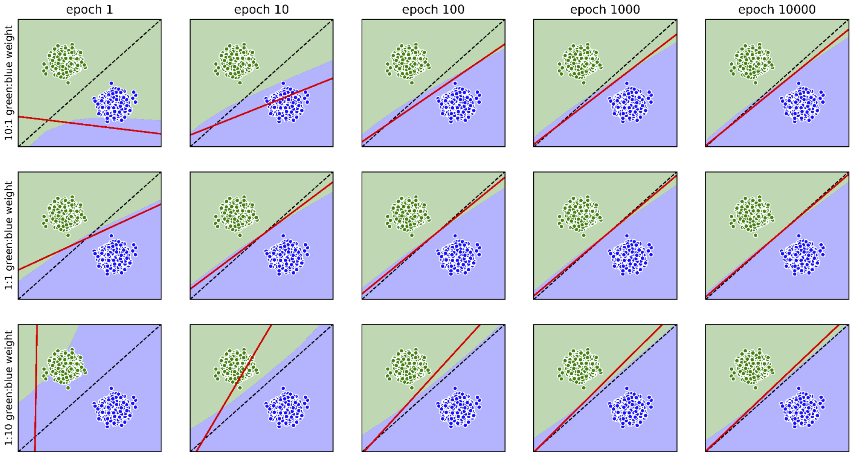


Figure: The decision boundaries are single-layered MLP with 64 hidden units [BL19]. **Black dashed line** shows the max-margin separator and the red dashed line shows the boundary of MLP

Implicit bias of GD for importance weighting

Theorem 5 (informal [XYR21])

For a separable data, with a sufficiently small constant rate η_t, for any $w \in [1/M, M]^n$, we have

$$\left| \frac{\theta^{(t)}}{\|\theta^{(t)}\|} - \theta^* \right| \precsim \frac{\log N + D_{KL}(p^* \| w) + M}{\gamma^* \log t},$$

where $p^* = [p_1^*, \ldots, p_N^*] \geqslant 0$ and $\sum_{i=1}^N p_i^*$ is the dual optimal for the hard margin SVM where $\theta^* = \sum_{i=1}^N y_i x_i p_i^*$, and D_{KL} is the Kullback-Leibler divergence.

- Importance weights does not change the convergence result as well as the convergence rate
- GD still induces the Max-Margin separator

Table of Contents

Homogeneous models with exponential tailed loss

Consider the asymptotic behavior of GD when the prediction is a homogeneous function

Definition 6 (α-homogeous)

$$L(w) = \sum_{i=1}^{n} \exp(-y_i f_i(w)),$$

where $f_i(cw) = c^\alpha f_i(w)$ is α-homogeous. $f_i(w)$ is the output of the prediction.

The associated non-linear margin maximization

$$\min \|w\|^2 \text{ s.t. } y_i f_i(w) \geqslant \gamma$$

First-order stationary point

- The max-margin problem itself is a constrained non-convex problem
- Instead, we show that GD iterates converge to the first-order stationary points of the max-margin problem

Definition 7 (First-order stationary point)

The first-order optimality conditions of Max-Margin are:

- $\forall i, y_i f_i(w) \geqslant \gamma$
- There exists Lagrange multipliers $\lambda \in R_+^N$ such that $w = \sum_n \lambda_n \nabla f_n(w)$ and $\lambda_n = 0$ for $n \notin S_m(w) := \{i : y_i f_i(w) = \gamma\}$, where $S_m(w)$ is the set of support vectors.

$\mathcal{W}^*$ indicates the set of firs-order stationary points

Implicit bias of GD for α-homogeneous function

Theorem 8

Define $\bar{w} = \lim_{t\to\infty} \frac{w_t}{||w_t||}$. *Suppose that* $f_i(w)$ *is a* C^2, $L(w_t) \to 0$, $\lim_{t\to\infty} \frac{w_t}{||w_t||}$ *and* $\lim_{t\to\infty} \frac{\ell_t}{||\ell_t||_1}$ *exist where* ℓ_t *is a vector whose* i*-th entry is* $\exp(-f_i(w_t))$, *and the linear independence constraint qualification (LICQ) holds, i.e.,* $\nabla\{f_i(w)\}_{i\in S_m(w)}$ *are linearly independent.* $\hat{w} \in \mathcal{W}$ *is a first-order stationary point of Max-Margin*

- Theorem 8 extends the result of linear models to α-homogeneous functions
- GD also converges to the Max-Margin solution in a sense

Table of Contents

Stability of GD on quadratic function

- Consider a convex quadratic function, gradient descent with step size η

$$f(x) = \frac{1}{2}x^T A x + b^T x + c$$

- **Sharpness**: the $\lambda_1 := A_{\max}$ is the largest eigenvalue of the Hessian of the objective function
  - If $\eta < 2/\lambda_1$, GD converges
  - If $\eta > 2/\lambda_1$, GD diverges
- In deep learning where the objective is nonconvex, similar analyses can show convergence towards stationary points and local minima
- However, recent empirical studies [CKL+21] showed compelling evidence to the contrary: Edge of Stability across various datasets and net structures

Dynamics of GD on neural networks training

Edge of Stability (EoS)

- Sharpness rises beyond $2/\eta$
- Sharpness stops rising but hovers noticeably above $2/\eta$ and even decreases a little
- Training loss behaves non-monotonically over individual iterations, yet consistently decreases over long timescales.

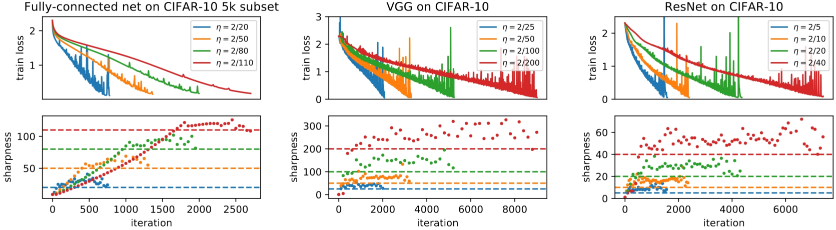


Figure: GD typically occurs at the Edge of Stability. [CKL+21] empirically observed this phenomenon.

GD enters EoS after sharpness reaches $2/\eta$

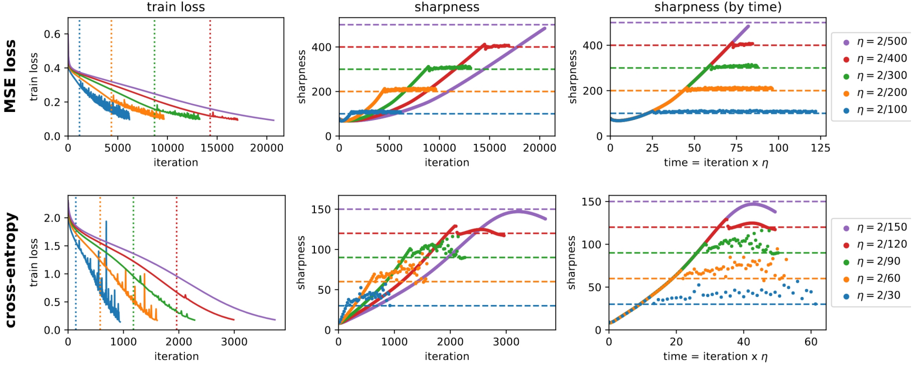
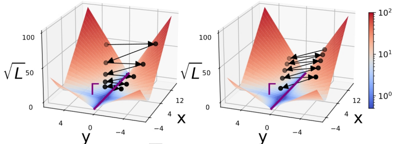


Figure: After progressive sharpening, GD enters the edge of stability

What happens after EoS?

Bias Towards Flattened Solutions

[ALP22] shows that normalized GD provably enters EoS, with associated flow on the manifold minimizing $\lambda_1(\nabla^2 L)$



(a) GD on $\sqrt{L}$ (b) Normalized GD on L

Explains SGD outperforms GD

[WS23] shows in a 2-layered neural network.

- SGD solution is linearly stable $\to Tr(\nabla^2 L) < 2/\eta$
- GD solution is linearly stable $\to \lambda_1(\nabla^2 L) < 2/\eta$

SGD often generalizes better than GD.

Summary

- Survey the recent advance on the implicit bias of gradient descent and other optimization algorithms
  - Linear regression model with squared loss
  - Matrix factorization
  - Linear classification model with exponential-tailed loss
  - Edge of stability
  - ...
- The implicit bias implies that those gradient descent prefers a "simpler" model
  - Bias towards max-margin solution
  - Bias towards "flatten" solution
- The implicit bias may partially explain why deep learning models trained with gradient descent generalize well

References I

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo, *Implicit regularization in deep matrix factorization*, Advances in Neural Information Processing Systems **32** (2019), 7413–7424.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi, *Understanding gradient descent on the edge of stability in deep learning*, International Conference on Machine Learning, PMLR, 2022, pp. 948–1024.

Jonathon Byrd and Zachary Lipton, *What is the effect of importance weighting in deep learning?*, International Conference on Machine Learning, PMLR, 2019, pp. 872–881.

Amir Beck and Marc Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters **31** (2003), no. 3, 167–175.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar, *Gradient descent on neural networks typically occurs at the edge of stability*, International Conference on Learning Representations, 2021.

Emmanuel J Candès and Benjamin Recht, *Exact matrix completion via convex optimization*, Foundations of Computational mathematics **9** (2009), no. 6, 717–772.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro, *Characterizing implicit bias in terms of optimization geometry*, International Conference on Machine Learning, PMLR, 2018, pp. 1832–1841.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro, *Implicit regularization in matrix factorization*, 2018 Information Theory and Applications Workshop (ITA), IEEE, 2018, pp. 1–10.

Diederik P Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, ICLR (Poster), 2015.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro, *The implicit bias of gradient descent on separable data*, The Journal of Machine Learning Research **19** (2018), no. 1, 2822–2878.

References II

Lei Wu and Weijie J Su, *The implicit regularization of dynamical stability in stochastic gradient descent*, International Conference on Machine Learning, PMLR, 2023, pp. 37656–37684.

Da Xu, Yuting Ye, and Chuanwei Ruan, *Understanding the role of importance weighting for deep learning*, International Conference on Learning Representations, 2021.