Characterizing Implicit Bias in Terms of Optimization Geometry

Gunasekar et al., ICML 2018

Kunwoong Kim

May 13, 2021

Introduction

- This paper studies the implicit bias of generic optimization methods in linear model.
  e.g., Mirror descent, natural gradient descent and steepest descent
- We consider underdetermined (X is singular) linear regression or separable linear classification.
- How can **initial value, step size or momentum** implicitly bias the solutions to global minima?

Contents

Problem setting

- Suppose we observe a training dataset $\{(x_n, y_n) : n = 1, 2, \cdots, N\}$ with features $x_n \in \mathbb{R}^d$ and their corresponding labels $y_n \in \mathbb{R}$. We consider a linear model $f(x) = \langle w, x \rangle$ with parameters $w \in \mathbb{R}^d$.
- Here, our target objective to minimize is given by

$$\mathcal{L}(w) := \sum_{n=1}^{N} l(f(x_n), y_n) = \sum_{n=1}^{N} l(\langle w, x_n \rangle, y_n)$$

  where l is an appropriate loss function for target task.

We consider two cases: (1) loss with a unique finite root in regression problem and (2) strict monotone loss in classification problem.

Case 1

- For the first case, we consider the losses with a unique finite root. That is,

$$l(\hat{y}_t, y) \to \inf_{\hat{y}} l(\hat{y}, y) \iff \hat{y}_t \to y$$

  for any y and sequence $\hat{y}_t$.
- Assume $N < d$, then the observed feature matrix $X = [x_1, \cdots, x_n]$s do not span a full-rank subspace of $\mathbb{R}^d$ so that $\mathcal{L}(w)$ has multiple global minima denoted by

$$\mathcal{G} := \{w : \mathcal{L}(w) = 0\} = \{w : \forall n, \langle w, x_n \rangle = y_n\}.$$

Case 1

Here we ask a question: which specific global minima $w \in \mathcal{G}$ do different optimization algorithms reach when minimizing $\mathcal{L}(w)$? To figure out, we consider the following optimization methods.

- Gradient descent
- Mirror descent
- Natural gradient descent
- Steepest descent

Case 1: Gradient descent

With step size η_t at time step t,

$$w_{(t+1)} = w_{(t)} - \eta_t \bigtriangledown \mathcal{L}(w_{(t)})$$

- $w_{(t)} \rightarrow \arg\min_{w \in \mathcal{G}} ||w - w_{(0)}||_2$
- The iterated parameter converges to the unique global minimum that is closest to initialization $w_{(0)}$.
- We can verify the same consequence for the SGD (with momentum and acceleration).

Why? the gradients $\bigtriangledown \mathcal{L}(w) = \sum_n l'(\langle w, x_n \rangle, y_n) x_n$ are constrained to the fixed subspace spanned by $x_1, \cdots, x_N$. Thus $w_{(t)}$ are confined to low dimensional affine manifold $w_{(0)} + span(\{x_n\}_n)$.

Case 1: Mirror descent

Let ψ a strong convex and differentiable function, that we call it "potential".

$$\text{GD} : w_{(t+1)} = \arg\min_{w \in \mathcal{W}} \eta_t \langle w, \triangledown \mathcal{L}(w_{(t)}) \rangle + ||w - w_{(t)}||_2^2$$

$$\text{MD} : w_{(t+1)} = \arg\min_{w \in \mathcal{W}} \eta_t \langle w, \triangledown \mathcal{L}(w_{(t)}) \rangle + D_\psi(w, w_{(t)})$$

where $D_\psi(w, w') = \psi(w) - \psi(w') - \langle \triangledown \psi(w'), w - w' \rangle$ is the Bregman divergence and $\mathcal{W}$ be any constrained parameter set.
e.g.

- $\psi(w) = \frac{1}{2}||w||_2^2$: gradient descent
- $\psi(w) = \sum_i w[i] \log w[i] - w[i]$ under simplex constraint $\mathcal{W} = \{w : \sum_i w[i] = 1\}$: exponentiated gradient descent

Case 1: Mirror descent

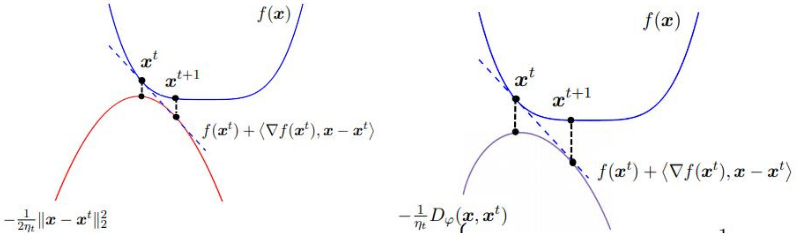


Figure: Gradient descent vs. Mirror descent, http://www.princeton.edu/~yc5/ele522_optimization/lectures/mirror_descent.pdf

Case 1: Mirror descent

(Theorem 1) Applying mirror descent algorithm with initial $w_{(0)}$ and step size η_t, assume the limit of iterated parameter $w_\infty = \lim_{t\to\infty} w_{(t)}$ satisfies $\mathcal{L}(w_\infty) = 0$. Then,

$$w_\infty = \underset{w\in\mathcal{G}}{\arg\min}\, D_\psi(w, w_{(0)})$$

- $w_{(0)} = \arg\min_w \psi(w) \to w_\infty = \arg\min_{w\in\mathcal{G}} \psi(w)$.

(Theorem 1a) Let a affine constraints $\mathcal{W} = \{w : Gw = h\}$ for some $G \in \mathbb{R}^{d'\times d}$ and $h \in \mathbb{R}^{d'}$ (in addition, we assume $\exists w \in \mathcal{W}$ such that $\mathcal{L}(w) = 0$) then,

$$w_\infty = \underset{w\in\mathcal{G}\cap\mathcal{W}}{\arg\min}\, D_\psi(w, w_{(0)})$$

- Let $\psi(w) = \sum_i w[i]\log w[i] - w[i]$ under simplex constraint $\mathcal{W} = \{w : \sum_i w[i] = 1\}$ and $w_{(0)} = \frac{1}{d}\mathbf{1} \to w_\infty = \arg\min_{w\in\mathcal{G}} \sum_i w[i]\log w[i]$

Case 1: Mirror descent

Here we consider with momentum.

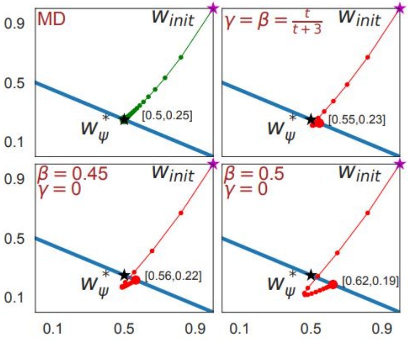
- Dual momentum:
  $\bigtriangledown\psi(w_{(t+1)}) = \bigtriangledown(w_{(t)}) + \beta_t \Delta z_{(t-1)} - \eta_t \bigtriangledown\mathcal{L}\left(w_{(t)} + \gamma_t \Delta w_{(t-1)}\right)$
- Primal momentum:
  $\bigtriangledown\psi(w_{(t+1)}) =$
  $\bigtriangledown(w_{(t)}) + \beta_t \Delta w_{(t-1)} - \eta_t \bigtriangledown \mathcal{L}\left(w_{(t)} + \gamma_t \Delta w_{(t-1)}\right)$

where $\Delta z_{(t-1)} = \bigtriangledown\psi(w_{(t)}) - \bigtriangledown\psi(w_{(t-1)})$ and $\Delta w_{(t-1)} = w_{(t)} - w_{(t-1)}$.

- With dual momentum, the same result holds (Theorem 2).
- However with primal momentum, $w_{(t)}$ strongly depends on the momentum parameters ((β_t, γ_t)) the step sizes $\{\eta_t\}$ (Example 2 and Proposition 2a : However with primal momentum only in the first step ($(\beta_t, \gamma_t) = (0, 0)$ for $t \geq 2$),).

Case 1: Mirror descent with primal momentum

This example shows the strong dependency of global minima to momentum parameters and step sizes. $l(u, y) = (u - y)^2$ and $x_1 = [1, 2], y_1 = 1$.



(a) Mirror descent primal momentum (Example 2)

Figure: Mirror descent with primal momentum

Case 1: Natural gradient descent

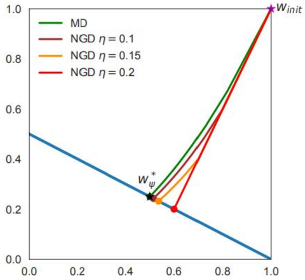
- Let a Riemannian metric tensor H that maps w to a positive definite local metric $H(w)$. In many instances, we consider $H = \bigtriangledown^2\psi$ for a strongly convex ψ.
- $$w_{(t+1)} = w_{(t)} - \eta_t H(w_{(t)})^{-1} \bigtriangledown \mathcal{L}(w_{(t)})$$
- For any positive definite D, if we consider a quadratic potential $\psi(w) = \frac{1}{2}||w||_D^2 = \frac{1}{2}w^\top Dw$,

$$\lim_{t\to\infty} w_{(t)} = \arg\min_{w\in\mathcal{G}} D_\psi(w, w_{(0)})$$

Case 1: Natural gradient descent

For non quadratic potential, it does not hold (Example 3 and Proposition 3a). $l(u, y) = (u - y)^2$ and $x_1 = [1, 2], y_1 = 1$. Let $\psi(w) = \sum_i w[i] \log w[i] - w[i]$. For $\eta_1 > 0$,

$$\lim_{t \to \infty} w_{(t)} = \arg\min_{w \in \mathcal{G}} D_\psi(w, w_{(1)}) \neq \arg\min_{w \in \mathcal{G}} D_\psi(w, w_{(0)})$$



(b) Natural gradient descent (Example 3)

Figure: Natural gradient descent with entropy potential

Case 1: Steepest descent

$$w_{(t+1)} = w_{(t)} - \eta_t \Delta w_{(t)}$$

where $\Delta w_{(t)} = \arg\min_v \langle \triangledown \mathcal{L}(w_{(t)}), v \rangle + \frac{1}{2}||v||^2$
e.g. w.r.t. l_2 norm: gradient descent, w.r.t. l_1 norm: coordinate descent.

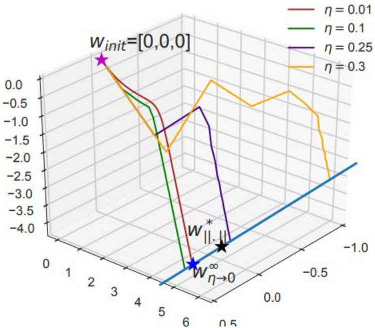
- For any positive definite D, when considering $||v||_D = \sqrt{v^\top Dv}$, then

$$\lim_{t\to\infty} w_{(t)} = \arg\min_{w\in\mathcal{G}} D_\psi(w, w_{(0)})$$

- However, for general norms, it does not. (e.g., $l_{4/3}$ norm, Example 4.) It strongly depends on the step size.

Case 1: Steepest descent

$l(u,y) = (u-y)^2$ and $x_1 = [1,1,1], x_2 = [1,2,0], y_1 = 1, y_2 = 10$



(c) Steepest descent w.r.t $||.||_{4/3}$ (Example 4)

Figure: Steepest descent with $l_{4/3}$ norm.

Case 2: classification

- Let consider the classification problem where $y \in \{-1, 1\}$ and $l(f(x), y)$ is a typically surrogate loss of the 0-1 loss. Here, we only consider the exponential loss $l(f(x), y) = \exp(-f(x) \cdot y)$ in this paper.
- That is, we consider a strict monotone loss as $l(\hat{y}, y)$ is strictly monotonically decreasing in $\hat{y}$. Let $\inf_y l(\hat{y}, y) = 0$ and $\lim_{\hat{y}y \to \infty} l(\hat{y}, y) = 0$.
  - Gradient descent
  - Steepest descent
  - Adaptive gradient descent (AdaGrad)

Case 2 : Gradient descent

- Let the dataset is linearly separable. That is, we assume $\exists : \forall n, y_n \langle w, x_n \rangle > 0$.
- Then, we cannot consider $\lim_{t\to\infty} w_{(t)}$ since $\mathcal{L}(w) = \sum_n \exp(-y_n \langle w, x_n \rangle) \to 0$ if $||w|| \to \infty$.
- Instead, we look at the direction

$$\bar{w}_\infty = \lim_{t\to\infty} \frac{w_{(t)}}{||w_{(t)}||}.$$

Case 2 : Gradient descent

Soudry et al. (2017) showed that

$$\bar{w}_\infty = \lim_{t\to\infty} \frac{w_{(t)}}{||w_{(t)}||_2} = w^*_{||\cdot||_2} = \underset{w:||w||_2\leq 1}{\arg\max}\, \min_n y_n\langle w, x_n\rangle$$

That is, gradient descent converges to maximum margin classifier with unit l_2 norm.

Case 2 : Steepest descent

For the steepest descent algorithm, we observe the similar consequence (Theorem 5).

$$\bar{w}_\infty = \underset{w:||w||\leq 1}{\arg\max}\, \min_n y_n \langle w, x_n \rangle$$

for any norm $||\cdot||$.

- This is independent to the initialization.
- The only requirement is the boundedness of step size: $\eta_t \leq C$ for some big C which only depends on $\max_n ||x_n||$ and $\mathcal{L}(w_{(t)})$.

Case 2 : AdaGrad

$$w_{(t+1)} = w_{(t)} - \eta \cdot \mathbf{G}_{(t)}^{-1/2} \bigtriangledown \mathcal{L}(w_{(t)})$$

where $\mathbf{G}_{(t)} \in \mathbb{R}^{d \times d}$ is a diagonal matrix with

$$\mathbf{G}_{(t)}[i,i] = \sum_{u=0}^{t} \left(\bigtriangledown \mathcal{L}(w_{(u)})[i] \right)^2 .$$

- Here, we need some requirements on the initialization of w and $\mathbf{G}$.
- (Theorem 6) If $\mathbf{G}$ converges, the limit direction depends on the initial conditions $w_{(0)}$ and $\mathbf{G}_{(0)}$.

$$\frac{\eta}{2}\mathcal{L}(w_{(0)}) < 1 \text{ and } ||\mathbf{G}_{(0)}^{-1/4} x_n||_2 \leq 1, \mathbf{G}_{(t)}[i,i] < \infty.$$

Gradient descent on the factorized parametrization

- For loss having finite unique root, Gunasekar et al (2017). have already done that the global minima depends on the initialization and step size.
- (Theorem 7) For monotone loss, this study shows the robustness of obtained global minima.

$$\bar{W}_\infty = \arg\max_{W \geq 0} \min_n y_n \langle W, X_n \rangle \text{ s.t. } ||W||_* \leq 1.$$

where $|| \cdot ||_*$ is unit nuclear norm.

Summary

Table: Implicit biases by various optimization algorithms on the linear model.

Unique root	Initial $w_{(0)}$	Step size η_t	Momentum (β_t, γ_t)
GD	O	X	X
MD	O	X (O if p.m.)	X (O if p.m.)
NGD	O	X (O if non-quad.)	-
SD	O	O (except for l_2 norm)	-

Monotone	Initial $w_{(0)}$	Step size η_t	Momentum (β_t, γ_t)
GD	X	X	-
SD	X	X	-
AdaGrad	$\triangle$	$\triangle$	-