See PDF

Implicit Bias of Gradient Descent on Reparametrized Models: On Equivalence to Mirror Descent

Zhiyuan Li[†,*] Tianhao Wang[‡,*] Jason D. Lee[†] Sanjeev Arora[†]
[†]Princeton University [‡]Yale University

Introduction

Given a loss L with multiple minimizers, the generalization error depends not only on the capacity of the function class, but also on the special property of the solution found by the training algorithm, which is also called *implicit bias* of the algorithm.

In this work, we study the following question: ***How do different parametrizations change the implicit bias of GD?***

Continuous gradient descent on reparametrized models:

- The model consists of a loss $L : \mathbb{R}^d \to \mathbb{R}$ with parameter $w \in \mathbb{R}^d$.
- Let $w = G(x)$ for a parametrization $G : \mathbb{R}^D \to \mathbb{R}^d$ with $x \in \mathbb{R}^D$.
  - E.g., $w = G(x) = u^{\odot 2} - v^{\odot 2}$ where $x = \binom{u}{v} \in \mathbb{R}^{2d}$.
- Consider the gradient flow

$$\mathrm{d}x(t) = -\nabla(L \circ G)(x(t))\mathrm{d}t. \qquad \text{(GF)}$$

Previous works [GWB+18, VKR19, YKM20, AW20a, WGL+20, AW20b, AMN+21] presents several settings where the implicit bias of GF can be described by mirror flow (MF, a.k.a. continuous mirror descent) with different convex functions.

Understanding the implicit bias via mirror descent:

- Let $w(t) = G(x(t))$, then $w(t)$ admits the following dynamics

$$\mathrm{d}w(t) = \partial G(x(t))\mathrm{d}x(t) = -\partial G(x(t))\partial G(x(t))^\top \nabla L(w(t))\mathrm{d}t. \qquad (1)$$

- If there is some strictly convex function $R : \mathbb{R}^d \to \mathbb{R}$ such that

$$\nabla^2 R(w(t))^{-1} = \partial G(x(t))\partial G(x(t))^\top, \qquad (2)$$

  then the dynamics of $w(t)$ satisfies

$$\mathrm{d}\nabla R(w(t)) = -\nabla L(w(t))\mathrm{d}t \qquad \text{(MF)}$$

  which is the mirror flow.
- Then the implicit bias of $w(t) = G(x(t))$ can be characterized via properties of mirror flow, in the sense that if as $t \to \infty$ $w(t)$ converges to some optimal solution w_∞, then w_∞ minimizes a convex regularizer given by the Bregman divergence of R among all optimal solutions.
- But when does the identity in (2) hold? Or,

When can a gradient flow with parametrization G be written as a mirror flow? (Q)

Our Contributions

The main contributions are summarized as follows:

1. We identify a notion of when a parametrization $w = G(x)$ is *commuting*, and use it to give a sufficient and (almost) necessary condition when (Q) has an affirmative answer.
2. Using the above characterization, we recover and generalize existing implicit bias results for underdetermined linear regression.
3. For the reverse direction of (Q), we use Nash's embedding theorem to show that every mirror flow can be written as a gradient flow with some reparametrization in a possibly higher-dimensional space.

GF w/ commuting parametrization is a MF

Notation:

- M is a simply-connected open subset of $\mathbb{R}^D$ (can be generalized to any smooth submanifold of $\mathbb{R}^D$).
- For parametrization $G : M \to \mathbb{R}^d$, $\{\nabla G_i\}_{i=1}^d$ are the gradients of the coordinate functions, and $\partial G(x) = (\nabla G_1(x), \ldots, \nabla G_d(x))$.
- Lie bracket $[\nabla G_i, \nabla G_j](x) = \nabla^2 G_j(x)\nabla G_i(x) - \nabla^2 G_i(x)\nabla G_j(x)$
- For any $x \in M$ and $i \in [d]$, $\phi^t_{G_i}(x)$ denotes the solution at time t to $\mathrm{d}\phi^t_{G_i}(x) = -\nabla G_i(\phi^t_{G_i}(x))\mathrm{d}t$. For any $\mu \in \mathbb{R}^d$, denote $\psi(x;\mu) = \phi^{\mu_1}_{G_1} \circ \phi^{\mu_2}_{G_2} \circ \cdots \circ \phi^{\mu_d}_{G_d}(x)$.

Definition (Commuting parametrization): Let $G : M \to \mathbb{R}^d$ be a parametrization satisfying $\mathrm{rank}(\partial G(x)) = d$ for all $x \in M$. We say G is a *commuting parametrization* if $[\nabla G_i, \nabla G_j](x) = 0$ for all $x \in M$ and any $i, j \in [d]$. (This implies that $\phi^{\mu_i}_{G_i} \circ \phi^{\mu_j}_{G_j}(x) = \phi^{\mu_j}_{G_j} \circ \phi^{\mu_i}_{G_i}(x)$.)

E.g., quadratic parametrization: $G_i(x) = \frac{1}{2}x^\top A_i x$ for all $i \in [d]$, where the matrices $\{A_i\}_{i=1}^d$ commute with each other ($A_iA_j = A_jA_i$). Note that the $w = u^{\odot 2} - v^{\odot 2}$ parametrization is a special case of quadratic parametrization.

Lemma 1. Let $G : M \to \mathbb{R}^d$ be a commuting parametrization. Then for any $x_{\text{init}} \in M$, there exists a strictly function Q such that $\nabla Q(\mu) = G(\psi(x_{\text{init}};\mu))$ for all μ. Moreover, let R be the convex conjugate of Q, then R satisfies $\nabla^2 R(G(\psi(x_{\text{init}};\mu))) = \left(\partial G(\psi(x_{\text{init}};\mu))\partial G(\psi(x_{\text{init}};\mu))^\top\right)^{-1}$.

Theorem 2. Let $G : M \to \mathbb{R}^d$ be a commuting parametrization, and for any $x_{\text{init}} \in M$, let R be the strictly convex function given by Lemma 1. Let $x(t)$ admit the (GF) with $x(0) = x_{\text{init}}$, then $w(t) = G(x(t))$ satisfies the (MF) with $w(0) = G(x_{\text{init}})$.

Remark: This R depends only on the initialization x_{init} and the parametrization G, and is independent of the loss L.

MF is a GF w/ commuting parametrization

Conversely, given any (MF), we can use Nash's embedding theorem to show that it can be given by a gradient flow with some commuting parametrization G.

Theorem 3. For any smooth R, consider $w(t)$ admitting the (MF). There exists a commuting parametrization $G : M \to \mathbb{R}^d$ such that $w(t) = G(x(t))$, where $x(t)$ follows the (GF) with parametrization G.

References

[AMN+21] Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477. PMLR, 2021.

[AW20a] Ehsan Amid and Manfred K Warmuth. Winnowing with gradient descent. In *Conference on Learning Theory*, pages 163–182. PMLR, 2020.

[AW20b] Ehsan Amid and Manfred KK Warmuth. Reparameterizing mirror descent as gradient descent. *Advances in Neural Information Processing Systems*, 33:8430–8439, 2020.

[GWB+18] Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

[VKR19] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32:2972–2983, 2019.

[WGL+20] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

[YKM20] Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. *arXiv preprint arXiv:2010.02501*, 2020.