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




Zhiyuan Li*
Princeton

Tianhao Wang*
Yale

Jason D. Lee
Princeton

Sanjeev Arora
Princeton

ICML, 2022

Background: Implicit Bias

- **<u>Implicit bias:</u>** special properties of the solution found by the optimization algorithm
  - *Not* implied by the value of the loss function
  - Arise from the trajectory taken in parameter space by the optimization
  - E.g., find sparse solutions without explicit ℓ_0 or ℓ_1 regularization

- Implicit bias is closely related to and can explain the generalization performance of algorithms
  - There are different sources of implicit bias: parametrization, step size, noise, etc.

- In this work, we study the following question:
  - *How do different parametrizations change the implicit bias of (continuous) gradient descent?*

Problem Setting: Reparametrized Gradient Flow

- Consider a model with loss $L : \mathbb{R}^d \to \mathbb{R}$ and parameter $w \in \mathbb{R}^d$

- $w = G(x)$ for a parametrization $G : \mathbb{R}^D \to \mathbb{R}^d$ with $x \in \mathbb{R}^D$ $(D \geq d)$
  - E.g., $w = G(x) = u^{\odot 2} - v^{\odot 2}$ where $x = \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^{2d}$

    Hadamard product

- $w(t) = G(x(t))$, where $x(t)$ is given by the gradient flow on $L \circ G$:

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

- Understand the implicit bias via the lens of (continuous) mirror descent

Understand Implicit Bias via Mirror Descent

- Gradient flow: $\mathrm{d}x(t) = -\nabla(L \circ G)(x(t))\mathrm{d}t = -\partial G(x(t))^\top \nabla L(G(x(t))\mathrm{d}t$

- $w(t) = G(x(t))$ admits the following dynamics:

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

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

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

- Then the dynamics of $w(t)$ satisfies

$$\mathrm{d}w(t) = -\nabla^2 R(w(t))^{-1} \nabla L(w(t))\mathrm{d}t \quad \text{(Riemannian gradient flow)}$$

$$\iff \mathrm{d}\nabla R(w(t)) = -\nabla L(w(t))\mathrm{d}t \quad \text{(Mirror flow)}$$

Understand Implicit Bias via Mirror Descent (cont.)

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

$$\mathrm{d}x(t) = -\nabla(L \circ G)(x(t))\mathrm{d}t \text{ (GF)} \quad \Longleftrightarrow \quad \mathrm{d}\nabla R(w(t)) = -\nabla L(w(t))\mathrm{d}t \text{ (MF)}$$

- Previous works presented several settings where the implicit bias of gradient flow can be described by the mirror flow

  Gunasekar et al. (2018); Vaskevicius et al. (2019); Woodworth et al. (2020); Amid & Warmuth (2020); Azulay et al. (2021); Yun et al. (2021) ……

  - Result (linear model): If as $t \to \infty$, $w(t)$ converges to some optimal solution w_∞, then w_∞ minimizes a convex regularizer among all optimal solutions:

$$w_\infty = \underset{w:\text{optimal}}{\arg\min} \; D_R(w, w(0))$$

- Question: When does $\nabla^2 R(w(t))^{-1} = \partial G(x(t)) \partial G(x(t))^\top$ hold?
- Our answer: *When G is a 'commuting parametrization'*

Notations

- Let $M \subseteq \mathbb{R}^D$ be a simply-connected open set (can be any smooth submanifold)
  - For $w = u^{\odot 2} - v^{\odot 2}$, can choose $M = \{(u, v) : u, v \in \mathbb{R}^d_+\}$

- For a parametrization $G : M \to \mathbb{R}^d$, $G(x) = [G_1(x), \ldots, G_d(x)]^\top$, Jacobian $\partial G(x) = [\nabla G_1(x), \ldots, \nabla G_d(x)]^\top$

- $\phi^t_{G_i}(x)$ denotes the solution at time t to $\mathrm{d}\phi^t_{G_i}(x) = -\nabla G_i(\phi^t_{G_i}(x))\mathrm{d}t$

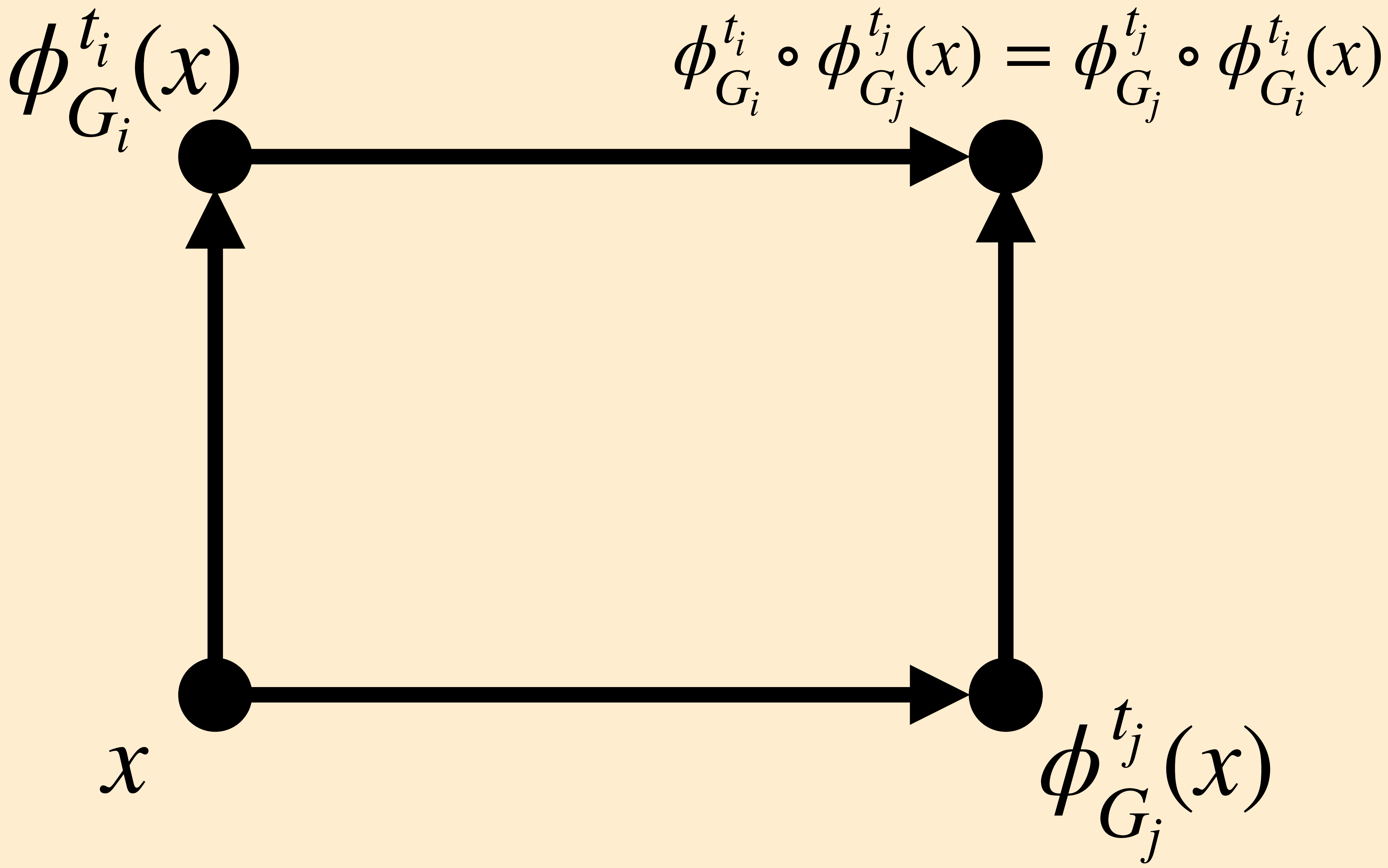
- Further define $\psi(x; \mu) = \phi^{\mu_1}_{G_1} \circ \phi^{\mu_2}_{G_2} \circ \cdots \circ \phi^{\mu_d}_{G_d}(x)$ for each $\mu \in \mathbb{R}^d$

Commuting Parametrization

Lie bracket $[\nabla G_i, \nabla G_j](x) = \nabla^2 G_j(x) \nabla G_i(x) - \nabla^2 G_i(x) \nabla G_j(x)$

<u>Def. (commuting parametrization):</u> Let $G : M \to \mathbb{R}^d$ be a parametrization. We say G is a *commuting parametrization* if $[\nabla G_i, \nabla G_j](x) = 0$ for all $x \in M$ and $i, j \in [d]$.

The commuting assumption implies:



Example: $w = G(x) = u^{\odot 2} - v^{\odot 2}$

- Each $G_i(x)$ only depends on (u_i, v_i)
- $\nabla G_i(x) = 2u_i \overrightarrow{e_i} - 2v_i \overrightarrow{e_{d+i}}$
- $\{\nabla G_i\}_{i=1}^d$ live in different subspaces
- $[\nabla G_i, \nabla G_j](x) \equiv 0, \ \forall i, j \in [d]$
- In this case, G is a commuting parametrization

Main Results: GF+Commuting $\Longrightarrow$ MF

Lemma 1 Let $G : M \to \mathbb{R}^d$ be a commuting parametrization. Let $x(t)$ follow the gradient flow on $L \circ G$ with $x(0) = x_{\text{init}}$, and define $\mu(t) = \int_0^t -\nabla L(G(x(s)))\mathrm{d}s$. Then $x(t) = \psi(x_{\text{init}}; \mu(t))$.

- *The gradient flow is determined by the integral of the negative gradient of the loss*

Lemma 2 Let $G : M \to \mathbb{R}^d$ be a commuting parametrization. Then for any $x_{\text{init}} \in M$, there exists a strictly convex 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 denoting $x = \psi(x_{\text{init}}; \mu)$, R satisfies

$$\nabla^2 R(w)^{-1} = \partial G(x) \partial G(x)^\top, \quad \text{where } w = G(x)$$

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

Theorem Every gradient flow with commuting parametrization is a mirror flow.

$$\mathrm{d}x(t) = -\nabla(L \circ G)(x(t))\mathrm{d}t \text{ (GF)} \quad \Longleftrightarrow \quad \mathrm{d}\nabla R(w(t)) = -\nabla L(w(t))\mathrm{d}t \text{ (MF)}$$

Commuting Param.

Main Results: MF$\Longrightarrow$GF+Commuting

Conversely, given any mirror flow, can it be reparametrized as a gradient flow?

- A similar question has been proposed by Amid & Warmuth (2020)

Our Answer: Yes! $\nabla^2 R(w(t))^{-1} \Longrightarrow \partial G(x(t)) \partial G(x(t))^\top$

Nash's embedding

<u>**Theorem**</u> For any smooth mirror map R, consider $w(t)$ admitting the mirror flow on loss L with respect to R. There exists a commuting parametrization $G : M \to \mathbb{R}^d$ such that $w(t) = G(x(t))$, where $x(t)$ admits the gradient flow on $L \circ G$.

- The embedding function does not admit analytic formula in general

Summary of Our Contributions

- We identify a notion of when a parametrization $w = G(x)$ is commuting, and use it to give a sufficient and (almost) necessary condition for when the gradient flow on x can be written as a mirror flow on w

- Using the above characterization, we recover and generalize existing implicit bias results for underdetermined linear regression

- Conversely, 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

Thank You!

arXiv: 2207.04036