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

Anonymous Author(s)
Affiliation
Address
email

Abstract

As part of the effort to understand implicit bias of gradient descent in overparametrized models, several results have shown how the training trajectory on the overparametrized model can be understood as mirror descent on a different objective. The main result here is a characterization of this phenomenon under a notion termed *commuting parametrization*, which encompasses all the previous results in this setting. It is shown that gradient flow with any commuting parametrization is equivalent to continuous mirror descent with a related mirror map. Conversely, continuous mirror descent with any mirror map can be viewed as gradient flow with a related commuting parametrization. The latter result relies upon Nash's embedding theorem.

1 Introduction

Implicit bias refers to the phenomenon in machine learning that the solution obtained from loss minimization has special properties that were not implied by value of the loss function and instead arise from the trajectory taken in parameter space by the optimization. Quantifying implicit bias necessarily has to go beyond the traditional black-box convergence analyses of optimization algorithms. Implicit bias can explain how choice of optimization algorithm can affect generalization [57, 39, 38].

Many existing results about implicit bias treat training (in the limit of infinitesimal step size) as a differential equation or process $\{x(t)\}_{t\geq 0} \subset \mathbb{R}^D$. To show the implicit bias of $x(t)$, the idea is to show for another (more intuitive or better understood) process $\{w(t)\}_{t\geq 0} \subset \mathbb{R}^d$ that $x(t)$ is *simulating* $w(t)$, in the sense that there exists a mapping $G : \mathbb{R}^D \to \mathbb{R}^d$ such that $w(t) = G(x(t))$. Then the implicit bias of $x(t)$ can be characterized by translating the special properties of $w(t)$ back to $x(t)$ through G. A related term *implicit regularization* refers to a handful of such results where particular update rules are shown to lead to regularized solutions; specifically, $x(t)$ is simulating $w(t)$ where $w(t)$ is solution to a regularized version of the original loss.

The current paper develops a general framework involving optimization in the continuous-time regime of a loss $L : \mathbb{R}^d \to \mathbb{R}$ that has been re-parametrized before optimization[1] as $w = G(x)$ for some $G : \mathbb{R}^D \to \mathbb{R}^d$. Then the original loss $L(w)$ in the w-space induces the implied loss $L(G(x))$ in the x-space, and the gradient flow in the x-space is given by

$$\mathrm{d}x(t) = -\nabla_x L(G(x(t)))\mathrm{d}t, \tag{1}$$

[1]Two examples from recent years, where G does not change expressiveness of the model, involve (a) overparametrized linear regression where the parameter vector w is reparametrized (for example as $w = u^{\odot 2} - v^{\odot 2}$ [57]) and (b) deep linear nets [28] where a matrix W is factorized as $W = W_1 W_2 \cdots W_L$ where each W_ℓ is the weight matrix for the ℓ-th layer.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

where ∇_x denotes the gradient with respect to x. Using $w(t) = G(x(t))$ and the fact that $\nabla_x L(G(x)) = \partial G(x)^\top \nabla L(G(x))$ where $\partial G(x) \in \mathbb{R}^{d\times D}$ denotes the Jacobian of G at x, the corresponding dynamics of (1) in the w-space is

$$\mathrm{d}w(t) = \partial G(x(t))\mathrm{d}x(t) = -\partial G(x(t))\partial G(x(t))^\top \nabla L(w(t))\mathrm{d}t. \tag{2}$$

Our framework is developed to fully understand phenomena in recent papers [22, 54, 60, 4, 57, 5, 8], which give examples suggesting that gradient flow in the x-space could end up simulating a more classical algorithm, mirror descent (specifically, the continuous analog, mirror flow) in the w-space. Recall that mirror flow is continuous-time limit of the classical mirror descent, written as $\mathrm{d}\nabla R(w(t)) = -\nabla L(w(t))\mathrm{d}t$ where $R : \mathbb{R}^d \to \mathbb{R}$ is a strictly convex function called *mirror map* [46, 10]. Equivalently it is *Riemannian gradient flow* with metric tensor $\nabla^2 R$, an old notion in geometry:

$$\mathrm{d}w(t) = -\nabla^2 R(w(t))^{-1}\nabla L(w(t))\mathrm{d}t. \tag{3}$$

If $\partial G(x(t))\partial G(x(t))^\top = \nabla^2 R(w(t))^{-1}$ for all t, then (2) becomes a simple mirror flow in the w-space as noted in [23]. Many existing results about implicit bias indeed concern reparametrizations G that satisfy $\partial G(x)\partial G(x)^\top = \nabla^2 R(w)^{-1}$ for a stricly convex function R, and the implicit bias/regularization is demonstrated by showing that the convergence point satisfies the KKT conditions needed for minimizing R among all minimizers of L. A concrete example is $w_i(t) = G_i(x(t)) = (x_i(t))^2$ for all $i \in [d]$, so here $D = d$. In this case, the mirror map R must satisfy $(\nabla^2 R(w(t)))^{-1} = \partial G(x(t))\partial G(x(t))^\top = 4\mathrm{diag}((x_1(t))^2, \dots, (x_d(t))^2) = 4\mathrm{diag}(w_1(t), \dots, w_d(t))$ which suggests R is the classical negative entropy function, *i.e.*, $R(w) = \sum_{i=1}^d w_i(\ln w_i - 1)$.

But in general, it is hard to decide *whether gradient flow for a given parametrization G can be written as mirror flow for some mirror map R*, especially when $D > d$ and G is not an injective map. In such cases, there could be multiple x's mapping to the same $G(x)$ yet having different $\partial G(x)\partial G(x)^\top$. If more than one of such x can be reached by gradient flow, then the desired mirror map cannot exist. However, if only one of such x can be reached by gradient flow, we must decide which x it is in order to decide the value of $\nabla^2 R$ using $\partial G \partial G^\top$. Conversely, [5] raises the following question: *for what mirror map R can the corresponding mirror flow be the result of gradient flow after some reparametrization G?* Answering the questions in both directions requires a deeper understanding of the impact of parametrizations.

The following are the main contributions of the current paper:

- In Section 4, building on classic study of commuting vector fields we identify a notion of when a parametrization $w = G(x)$ is *commuting* (Definition 4.1) and use it to give a sufficient condition (Theorem 4.4) of when the gradient flow in the x-space governed by $-\nabla_x L(G(x))$ is simulating a mirror flow in the w-space with respect to some mirror map $R : \mathbb{R}^d \to \mathbb{R}$, which encompasses all the previous results [22, 54, 60, 4, 57, 5, 8]. Moreover, the mirror map is independent of the loss L and depends only on the initialization x_{init} and the parametrization G.
- We recover and generalize existing implicit regularization results for underdetermined linear regression as implications of the above characterization (Corollary 4.12). We also give new convergence analysis in such settings (Corollary 4.10), filling the gap in previous works [22, 57, 8] where parameter convergence is only assumed but not proved.
- In the reverse direction, we use the famous Nash embedding theorem to show that every mirror flow in the w-space with respect to some mirror map R simulates a commutative gradient flow under some embedding $x = F(w)$ where $F : \mathbb{R}^d \to \mathbb{R}^D$ and the parametrization G is the inverse of F. (Theorem 5.3) This provides an affirmative and fully general answer to the question of when such reparametrization functions exist, giving a full answer to questions raised in a more restricted setting in [5].

2 Related work

Implicit bias. With high overparametrization as used in modern machine learning, there usually exist multiple optima, and it is crucial to understand which particular solutions are found by the

optimization algorithm. Implicit bias of gradient descent for classification tasks with separable data was studied in [52, 20, 43, 32, 42, 31] and for non-separable data in [29, 30], where the implicit bias appears in the form of margin maximization. The implicit bias for regression problems has also been analyzed by leveraging tools like mirror descent [57, 20, 60, 54, 4, 5], later generalized in [8].

The sharp contrast between the so-called *kernel* and *rich* regimes [57] reflects the importance of the initialization scale, where a large initialization often leads to the kernel regime with features barely changing during training [26, 14, 18, 17, 2, 1, 61, 7, 58, 27], while with a small initialization, the solution exhibits richer behavior with the resulting model having lower complexity [21, 22, 36, 49, 6, 13, 38, 40, 41, 50, 53, 19]. Recently [59] give a complete characterization on the relationship between initialization scale, parametrization and learning rate in order to avoid kernel regime.

There are also papers on the implicit bias of other types of optimization algorithms, e.g., stochastic gradient descent [37, 11, 25, 39, 16, 62] and adaptive and momentum-based methods [48, 56, 55, 33], to name a few.

Understanding mirror descent. In the continuous-time limit as step size goes to 0, the mirror flow is equivalent to the Riemannian gradient flow. [23] showed that a partial discretization of the latter gives rise to the classical mirror descent. Assuming the existence of some reparametrization function, [5] showed that a particular mirror flow can be reparametrized as a gradient flow. Our paper shows that such reparametrization always exists by using Nash's embedding theorem.

3 Preliminaries and notations

For any positive integer n, we denote $\{1, 2, \ldots, n\}$ by $[n]$. For any vector $u \in \mathbb{R}^D$, we denote its i-th coordinate by u_i. For any vector $u, v \in \mathbb{R}^D$ and $\alpha \in \mathbb{R}$, we define $u \odot v = (u_1 v_1, \ldots, u_D v_D)^\top$ and $u^{\odot\alpha} = ((u_1)^\alpha, \ldots, (u_D)^\alpha)^\top$. For any integer k, we say a function f is $\mathcal{C}^k$ if it is k times continuously differentiable. For any convex function $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, we denote its domain by $\operatorname{dom} R = \{w \in \mathbb{R}^d \mid R(w) < \infty\}$. For any set S, we use $\operatorname{int}(S)$ to denote its interior. For any submanifold $M \subset \mathbb{R}^D$ and $x \in M$, we denote by $T_x(M)$ the tangent space of M at x.

We assume the model has parameter vector $w \in \mathbb{R}^d$ and loss function $L : \mathbb{R}^d \to \mathbb{R}$ which is a $\mathcal{C}^1$ function. Training involves a reparametrized vector $x \in \mathbb{R}^D$, which is a reparametrization of w such that $w = G(x)$ for some differentiable parametrization function G. More precisely, we assume that the reparametrized vector x takes value in a smooth submanifold of $\mathbb{R}^D$ denoted by M, so whenever we say $G : M \to \mathbb{R}^d$ is a parametrization function, by default M is a smooth submanifold of $\mathbb{R}^D$. Given an initialization x_{init}, we consider running gradient flow on x:

$$\mathrm{d}x(t) = -\nabla_x L(x(t))\mathrm{d}t.$$

From now on, we follow the convention that d is the dimension of the original parameter w and D is the dimension of the reparametrized x. We also refer to $\mathbb{R}^d$ as the w-space and $\mathbb{R}^D$ as the x-space.

Some of the proofs and counterexamples also use the following notion of time-dependent loss.

Definition 3.1 (Time-dependent loss)**.** A time-dependent loss is a function $L_t(w)$ that maps $(w, t) \in \mathbb{R}^d \times \mathbb{R}$ to a real number such that for any $t \in \mathbb{R}$, $L_t(w)$ is continuously differentiable in w, and $\nabla L_t(w)$ is piece-wise continuous in (w, t). The set of all time-dependent losses is denoted by $\mathcal{L}$.

3.1 Vector fields

Vector fields are a natural way to formalize continuous-time gradient descent (a good reference is [34]). Given any smooth submanifold of $\mathbb{R}^D$, denoted by M, a *vector field* X on M is a continuous map from M to $\mathbb{R}^D$ such that for any $x \in M$, $X(x)$ is in the tangent space of M at x, $T_x(M)$.

Definition 3.2 (Complete vector fields, p.215, [34])**.** We say X is a *complete vector field* on M if and only if for any initialization $x_{\text{init}} \in M$, the differential equation $\mathrm{d}x(t) = X(x(t))\mathrm{d}t$ has a solution on $(-\infty, \infty)$ with $x(0) = x_{\text{init}}$.

For any differentiable function $f : M \to \mathbb{R}$, we denote by ∇f its gradient vector field. For example, in the special case when $M = \mathbb{R}^D$, ∇f is simply the ordinary gradient of f. For any $x \in M$, we denote by $\phi_f^t(x)$ the point on M reached after time t by following the vector field $-\nabla f$ starting at x:

(here we allow t to be negative and use the convention $\int_a^b = -\int_b^a$)

$$\phi_f^t(x) = x - \int_0^t \nabla f(\phi_f^s(x))\mathrm{d}s.$$

Now, let $G : M \to \mathbb{R}^d$ be a continuously differentiable parametrization function. We denote its Jacobian by $\partial G(x) = (\nabla G_1(x), \nabla G_2(x), \ldots, \nabla G_d(x))^\top$. For any $x \in M$ and $\mu \in \mathbb{R}^d$, we define

$$\psi(x; \mu) = \phi_{G_1}^{\mu_1} \circ \phi_{G_2}^{\mu_2} \circ \cdots \circ \phi_{G_d}^{\mu_d}(x), \tag{4}$$

where $\circ$ denotes the composition of functions.

3.2 Parametrizations

We focus on a class of parametrization functions that do not change the expressiveness of the original model. So when it comes to implicit bias, we require by default the original model is expressive enough to have multiple optimal solutions.

Definition 3.3 (Canonical parametrizations)**.** A *canonical parametrization* $G : M \to \mathbb{R}^d$ is a continuously differentiable function that satisfies: (1) the domain M is a smooth submanifold of $\mathbb{R}^D$, (2) the gradient vector fields $\nabla G_1, \nabla G_2, \ldots, \nabla G_d$ are all complete vector fields on M and (3) $\partial G(x)$ is of rank d for all $x \in M$.

Given a canonical parametrization $G : M \to \mathbb{R}^d$ and any initialization $x_{\text{init}} \in M$, the set of points that are reachable via gradient flow with respect to some time-dependent loss is a subset of M that depends on G and x_{init}. Since "reachable" is a transitive relationship, the notion of reachable sets divide the domain M into equivalent classes.

Definition 3.4 (Reachable sets)**.** For a canonical parametrization $G : M \to \mathbb{R}^d$ and initialization $x_{\text{init}} \in M$, we define the reachable set $\Omega_x(x_{\text{init}}; G)$ as

$$\Omega_x(x_{\text{init}}; G) = \Big\{ x(t) \;\Big|\; \forall t \in \mathbb{R}, L_\cdot \in \mathcal{L} \text{ and continuous } x(\cdot) : \mathbb{R} \to \mathbb{R}^d \\ \text{s.t. } \mathrm{d}x(s)/\mathrm{d}t = -\nabla_x L_s(G(x(s))),\ \forall s \in \mathbb{R}, x(0) = x_{\text{init}} \Big\}.$$

The above reachable set in the x-space further induces the corresponding reachable set in the w-space given by $\Omega_w(x_{\text{init}}; G) = G(\Omega_x(x_{\text{init}}; G))$.

3.3 Mirror descent and mirror flow

Below we introduce some basics notions for mirror descent [46, 10].

Definition 3.5 (Mirror maps, p.298, [12])**.** A function $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is called a *mirror map* if and only if (1) R is strictly convex and differentiable on $\mathrm{int}(\mathrm{dom}\, R)$; (2) the gradient map $\nabla R : \mathrm{int}(\mathrm{dom}\, R) \to \mathbb{R}^d$ is surjective; and (3) for any sequence $\{w_n\}_{n\geq 1}$ going to the boundary of $\mathrm{dom}\, R$, $\lim_{n\to\infty} \|\nabla R(w_n)\|_2 = \infty$.

Then for initialization $w_0 \in \mathrm{int}(\mathrm{dom}\, R)$, mirror descent with step size η updates as follows:

$$\nabla R(w_{k+1}) = \nabla R(w_k) - \eta \nabla L(w_k). \tag{5}$$

In the limit of $\eta \to 0$, the above discrete update becomes the continuous mirror flow:

$$\mathrm{d}\nabla R(w(t)) = -\nabla L(w(t))\mathrm{d}t. \tag{6}$$

Recall that given a differentiable function R, the corresponding Bregman divergence D_R is defined as

$$D_R(w, w') = R(w) - R(w') - \langle \nabla R(w'), w - w' \rangle.$$

We recall a well-known implicit bias result for mirror flow (which holds for mirror descent as well), which shows that for a specific type of loss, if mirror flow converges to some optimal solution, then the convergence point achieves some implicit regularization among all optimal solutions.

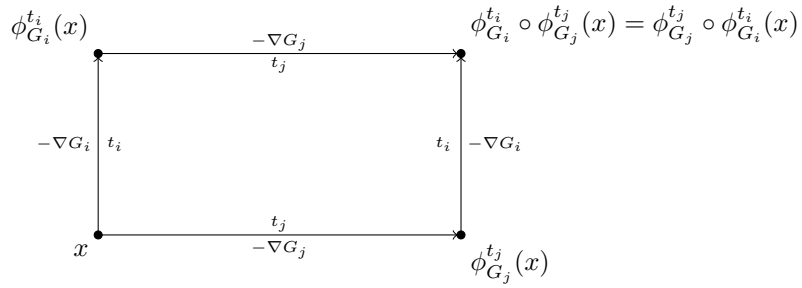

Figure 1: Illustration of commuting parametrizations. Suppose $G : M \to \mathbb{R}^d$ is a commuting parametrization, then starting from any $x \in M$, first moving along $-\nabla G_i$ for time t_i then moving along $-\nabla G_j$ for time t_j yields the same result as first moving along $-\nabla G_j$ for time t_j then moving along $-\nabla G_i$ for time t_i does, i.e., $\phi_{G_i}^{t_i} \circ \phi_{G_j}^{t_j}(x) = \phi_{G_j}^{t_j} \circ \phi_{G_i}^{t_i}(x)$.

Theorem 3.6. *Given any data $Z \in \mathbb{R}^{n \times d}$ and the corresponding label $Y \in \mathbb{R}^n$, suppose the loss $L(w)$ is in the form of $L(w) = \widetilde{L}(Zw)$ for some differentiable $\widetilde{L} : \mathbb{R}^n \to \mathbb{R}$. Assume that initialized at w_0, the mirror flow (6) converges and the convergence point $w_\infty = \lim_{t\to\infty} w(t)$ satisfies $Zw_\infty = Y$, then $D_R(w_\infty, w_0) = \min_{w:Zw=Y} D_R(w, w_0)$.*

See Appendix B for a proof. The above theorem is the building block for proving the implicit bias induced by any commuting parametrization in overparametrized linear models (see Theorem 4.11).

4 Every commutative gradient flow is a mirror flow

We now formalize the notion of commuting parametrization. We remark that M is a general smooth submanifold of $\mathbb{R}^D$ that the reparametrized vector x takes value in.

Definition 4.1 (Commuting parametrizations)**.** A *commuting parametrization* $G : M \to \mathbb{R}^d$ is a canonical parametrization such that the gradient vector fields $\nabla G_1, \nabla G_2, \ldots, \nabla G_d$ on M satisfy that for any $i, j \in [d]$ and any $t_i, t_j \in \mathbb{R}$, $\phi_{G_i}^{t_i} \circ \phi_{G_j}^{t_j}(x) = \phi_{G_j}^{t_j} \circ \phi_{G_i}^{t_i}(x)$ for all $x \in M$.

When G is a commuting parametrization, we call the corresponding gradient flow (1) a *commutative gradient flow*. We illustrate the concept of commuting parametrization in Figure 1.

The meaning of commuting in the above definition is built upon the flows generated by the vector fields, which is further equivalent to that the Lie bracket of any pair of the vector fields satisfies $[\nabla G_i, \nabla G_j] \equiv 0$. In the latter sense, we say ∇G_i and ∇G_j are commuting vector fields (see p.231, [34]). Intuitively, the Lie bracket $[\nabla G_i, \nabla G_j]$ characterizes the change of ∇G_j along the flow generated by ∇G_i and $[\nabla G_i, \nabla G_j] \equiv 0$ means ∇G_j is invariant along $\phi_{G_i}^{t}(x)$, thus justifying the commutability between $\phi_{G_i}^{t_i}$ and $\phi_{G_j}^{t_j}$. The equivalence is summarized in the following theorem, which can be applied to check whether a parametrization is a commuting parametrization or not.

Theorem 4.2 (Adapted from Theorem 9.44, [34])**.** *Given a differentiable map $G : M \to \mathbb{R}^d$ where M is a smooth submanifold of $\mathbb{R}^D$, for any $i, j \in [d]$, ∇G_i and ∇G_j are commuting vector fields, i.e., $[\nabla G_i, \nabla G_j] \equiv 0$, if and only if any $t_i, t_j \in \mathbb{R}$, $\phi_{G_i}^{t_i} \circ \phi_{G_j}^{t_j}(x) = \phi_{G_j}^{t_j} \circ \phi_{G_i}^{t_i}(x)$ for all $x \in M$.*

In particular, when $\nabla G_1, \nabla G_2, \ldots, \nabla G_d$ are ordinary gradient vectors in $\mathbb{R}^D$, the Lie bracket between any ∇G_i and ∇G_j is given by $[\nabla G_i, \nabla G_j](x) = \nabla^2 G_i(x) \nabla G_j(x) - \nabla^2 G_j(x) \nabla G_i(x)$. The above condition clearly holds when each G_i only depends on a different subset of coordinates of x, because we have $\nabla^2 G_i(x) \nabla G_j(x) \equiv 0$, for all $i \neq j \in [d]$ and $x \in M$ as $\nabla^2 G_i$ and ∇G_j live in different spaces. We call this setting *separable case* and it covers all the previous examples [22, 54, 4, 57, 5]. For concreteness, we give the example below.

Example 4.3 (Quadratically overparametrized model)**.** Consider the case of the overparametrized linear model (also known as diagonal linear network) [57] where $w \in \mathbb{R}^d$ is parametrized by $w = u^{\odot 2} - v^{\odot 2}$. In this case, we set $D = 2d$ and $M = \{x = \binom{u}{v} \in \mathbb{R}^D \mid u_i, v_i > 0, \forall i \in [d]\}$, and $G : M \to \mathbb{R}^d$ is given by $G(x) = u^{\odot 2} - v^{\odot 2}$. Since each G_i involves u_i, v_i, this is a separable case and thus $\nabla G_1, \nabla G_2, \ldots, \nabla G_d$ are commuting vector fields.

Next, we present our result on characterizing any commutative gradient flow by a mirror flow.

Theorem 4.4. *Let $G : M \to \mathbb{R}^d$ be a commuting parametrization on some smooth submanifold $M \subseteq \mathbb{R}^D$ as defined in Definition 4.1. For any $\mathcal{C}^1$ loss function $L : \mathbb{R}^d \to \mathbb{R}$ and any initialization $x_{init} \in M$, let $x(t)$ be given by the commutative gradient flow governed by $-\nabla_x L(G(x))$:*

$$\mathrm{d}x(t) = -\nabla_x L(G(x(t)))\mathrm{d}t, \qquad x(0) = x_{init}.$$

Denote $w(t) = G(x(t))$, then there exists some mirror map R such that the dynamics of $w(t)$ is a mirror flow with respect to R, i.e.,

$$\mathrm{d}\nabla R(w(t)) = -\nabla L(w(t)), \qquad w(0) = G(x_{init}).$$

Moreover, this mirror map R only depends on the initialization x_{init} and the commuting parametrization G, and is independent of the loss function L.

The proof of the above theorem relies on two lemmas that highlight the special properties of commuting parametrizations. Lemma 4.5 tells us that the point reached by a commutative gradient flow is determined by the integral of the negative gradient of the loss along the trajectory.

Lemma 4.5. *Let $G : M \to \mathbb{R}^d$ be a commuting parametrization and consider any time-dependent loss L_t. For any initialization $x_{init} \in M$, let the solution to the gradient flow governed by $-\nabla_x L_t(G(x))$ at time t be $x(t)$, i.e., $x(t) = x_{init} + \int_0^t -\nabla_x L_t(G(x(s)))\mathrm{d}s$. Define $\mu(t) = \int_0^t -\nabla L_t(G(x(s)))\mathrm{d}s$, then it holds that $x(t) = \psi(x_{init}; \mu(t))$.*

Based on Lemma 4.5, the following key lemma reveals the essential approach to find the mirror map.

Lemma 4.6. *Let $G : M \to \mathbb{R}^d$ be a commuting parametrization. For any initialization $x_{init} \in M$, there exists a twice differentiable strictly convex function $Q : \mathbb{R}^d \to \mathbb{R}$ such that $\nabla Q(\mu) = G(\psi(x_{init}; \mu))$. Moreover, the convex conjugate of Q, denoted by R, is a mirror map such that $\mathrm{int}(\mathrm{dom}\, R) = \Omega_w(x_{init}; G)$ and $\nabla^2 R(G(\psi(x_{init}; \mu))) = (\partial G(\psi(x_{init}; \mu))\partial G(\psi(x_{init}; \mu))^\top)^{-1}$ for all $\mu \in \mathbb{R}^d$.*

Now, we are ready to prove Theorem 4.4.

Proof of Theorem 4.4. Recall that the gradient flow in the x-space governed by $-\nabla_x L(G(x))$ is

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

Using $w(t) = G(x(t))$, the corresponing dynamics in the w-space is

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

Next, by Lemma 4.5, we know that the solution to the gradient flow satisfies $x(t) = \psi(x_{\text{init}}; \mu(t))$ where $\mu(t) = \int_0^t -\nabla L_t(G(x(s)))\mathrm{d}s$. Therefore, applying Lemma 4.6, we get a mirror map R with domain $\Omega_w(x_{\text{init}}; G)$ such that

$$\nabla^2 R(w(t)) = \nabla^2 R(G(\psi(x_{\text{init}}; \mu(t)))) = (\partial G(\psi(x_{\text{init}}; \mu(t)))\partial G(\psi(x_{\text{init}}; \mu(t))))^{-1}$$

for all $t \geq 0$. Then the dynamics of $w(t)$ satisfies $\mathrm{d}w(t) = -\nabla^2 R(w(t))^{-1}\nabla L(w(t))\mathrm{d}t$ which is exactly the mirror flow with respect to R initialized at $w(0) = G(x_{\text{init}})$. Note that the result of Lemma 4.6 is completely independent of the loss function L, thus the mirror map R only depends on the initialization x_{init} and the commuting parametrization G. This finishes the proof. □

Remark 4.7 (Non-commuting flow cannot be mirror flow, when dimension matches.)**.** *Suppose $G : M \to \mathbb{R}^d$ is a canonical parametrization with $\dim(\Omega_x(x_0; G)) = d$ and for any time-dependent loss $L_\cdot \in \mathcal{L}$, gradient flow governed by $-\nabla_x L_t(G(x))$ can be written as a mirror flow, then G must be commuting parametrization. To see why, let $\delta > 0$ and consider a time-dependent loss given by*

$$L_t(w) = \begin{cases} \langle e_i, w\rangle & 0 \leq t < \delta, \\ \langle e_j, w\rangle & t \leq \delta < t + 2\delta, \\ -\langle e_i, w\rangle & 2\delta \leq t < 3\delta, \\ -\langle e_j, w\rangle & 3\delta \leq t < 4\delta, \\ 0 & t \geq 4\delta. \end{cases}$$

Suppose for any initialization $x_{init} \in M$, $\mathrm{d}x(t) = -\nabla_x L_t(G(x(t))\mathrm{d}t$ with $x(0) = x_{init}$ can be written as $\mathrm{d}\nabla R(w(t)) = -\nabla L_t(w(t))\mathrm{d}t$ for some mirror map R where $w(t) = G(x(t))$. Then $\nabla R(4\delta)) = \nabla R(w(0)) + \int_0^t -\nabla L_s(w(s))\mathrm{d}s = \nabla R(w(0))$, which further implies that $w(4\delta) = w(0)$, and thus $G(x(4\delta)) = G(x_{init})$. One can show that $x(4\delta) = x_{init} + \delta^2[-\nabla G_j, -\nabla G_i](x_{init}) + \mathcal{O}(\delta^3)$, which implies $[-\nabla G_j, -\nabla G_i](x_{init}) \in T_{x_{init}}(\Omega_x(x_{init}; G))$. Since $\dim(\Omega_x(x_{init}; G)) = d$ and G is canonical, $\{\nabla G_k(x_{init})\}_{k=1}^d$ form a basis of $T_{x_{init}}(\Omega_x(x_{init}; G))$. Note that $0 = \lim_{t\to 0} \frac{G_k(x(4t)) - G_k(x_{init})}{t^2} = \langle \nabla G_i(x_{init}), [-\nabla G_j, -\nabla G_i](x_{init})\rangle$ for all $k \in [d]$, we conclude $[-\nabla G_j, -\nabla G_i](x_{init})$ must be 0. Thus by Theorem 4.2, we conclude G is a commuting parametrization since x_{init} is arbitrary.

Next, we establish the convergence of $w(t) = G(x(t))$ when $x(t)$ is given by some commutative gradient flow. Here we require that the mirror map R given by Theorem 4.4 is a Bregman function (see definition in Appendix A). The proofs of Theorem 4.8 and Corollary 4.9 are in Appendix C.

Theorem 4.8. *Under the setting of Theorem 4.4, further suppose that the loss L is quasi-convex and ∇L is locally Lipschitz. If the strictly convex function R given by Theorem 4.4 is a Bregman function satisfying $\operatorname{argmin}\{L(w) \mid w \in \operatorname{dom} R\}$ is non-empty, then as $t \to \infty$, $w(t)$ converges to some w^* such that $\nabla L(w^*)^\top(w - w^*) \geq 0$ for all $w \in \operatorname{dom} R$. Moreover, if the loss function L is convex, then $w(t)$ converges to a global minimizer.*

Corollary 4.9. *Under the setting of Theorem 4.8, if the reachable set in the w-space satisfies $\Omega_w(x_{init}; G) = \mathbb{R}^d$, then R is a Bregman function and all the statements in Theorem 4.8 hold.*

Corollary 4.10. *Under the setting of Theorem 4.8, if $M = \mathbb{R}^d_+$ and $G_i(x) = x^\top A_i x$ for all $i \in [d]$ where $A_i \in \mathbb{R}^{D\times D}$ are symmetric matrices and commute with each other,* i.e., *$A_iA_j - A_jA_i = 0$, $\forall i, j \in [d]$, then R is a Bregman function and all the statements in Theorem 4.8 hold.*

4.1 Case study: underdetermined linear regression

Next, we specialize to underdetermined linear regression problems to showcase our framework.

Setting: underdetermined linear regression. Let $\{(z_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ be a dataset of size n. Given any parametrization $w = G(x)$, the output of the linear model on the i-th data is $z_i^\top G(x)$. The goal is to solve the regression for the label vector $Y = (y_1, y_2, \ldots, y_n)^\top$. For notational convenience, we define $Z = (z_1, z_2, \ldots, z_n) \in \mathbb{R}^{d\times n}$.

We can apply Theorem 3.6 to obtain the implicit bias of a commutative gradient flow.

Theorem 4.11. *Suppose $G : M \to \mathbb{R}^d$ is a commuting parametrization satisfying Definition 4.1 and the loss function L satisfies $L(w) = \widetilde{L}(Zw)$ for some differentiable $\widetilde{L} : \mathbb{R}^n \to \mathbb{R}$. Given any initialization $x_{init} \in M$, consider the gradient flow $\mathrm{d}x(t) = -\nabla_x L(G(x(t))\mathrm{d}t$ initialized at $x(0) = x_{init}$. Then there exists a convex function R (depending on x_{init} and the parametrization G), such that for any dataset $\{(z_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, if $w(t) = G(x(t))$ converges and the convergence point $w_\infty = \lim_{t\to\infty} w(t)$ satisfies $Zw_\infty = Y$, then $R(w_\infty) = \min_{w:Zw=Y} R(w)$, that is, gradient flow implicitly minimizes the convex regularizer R among all interpolating solutions.*

Proof of Theorem 4.11. By Theorem 4.4, $w(t)$ obeys the following mirror flow:

$$\mathrm{d}\nabla R(w(t)) = -\nabla L(w(t))\mathrm{d}t, \qquad w(0) = G(x_{\text{init}}).$$

Applying Theorem 3.6 yields $D_R(w_\infty, G(x_{\text{init}})) = \min_{w:Zw=Y} D_R(w, G(x_{\text{init}}))$. Therefore, for any $w \in \operatorname{int}(\operatorname{dom} R)$ such that $Zw = Y$, we have

$$\begin{aligned} &R(w_\infty) - R(G(x_{\text{init}})) - \langle \nabla R(G(x_{\text{init}}), w_\infty - G(x_{\text{init}})\rangle \\ &\quad \leq R(w) - R(G(x_{\text{init}})) - \langle \nabla R(G(x_{\text{init}}), w - G(x_{\text{init}})\rangle \end{aligned}$$

which can be reorganized as

$$R(w_\infty) \leq R(w) - \langle \nabla R(G(x_{\text{init}})), w - w_\infty\rangle.$$

Note that by Lemma 4.6, we have $\nabla R(G(x_{\text{init}})) = \nabla R(G(\psi(x_{\text{init}}; 0))) = \nabla R(\nabla Q(0)) = 0$ where the last equality follows from the property of convex conjugate. Combining this with the above inequality, we get $R(w_\infty) \leq R(w)$ for all $w \in \operatorname{int}(\operatorname{dom} R)$ such that $Zw = Y$. By the continuity of R, this property can be further extended to the whole $\operatorname{dom} R$, and for any $w \notin \operatorname{dom} R$, we have $R(w) = \infty$ by definition, so $R(w_\infty) \leq R(w)$ holds trivially. This finishes the proof. $\square$

Note that the identity parametrization $w = G(x) = x$ is a commuting parametrization. Therefore, if we run the ordinary gradient flow on w itself and it converges to some interpolating solution, then the convergence point is closest to the initialization in Euclidean distance among all interpolating solutions. This recovers the well-known implicit bias of gradient flow for underdetermined regression.

Furthermore, we can recover the results on the quadratically overparametrized linear model in a series of paper [22, 57, 8] and strenghthen their results by removing the convergence assumption in their theorems by Corollary 4.10. Corollary 4.10 is also more general than convergence results on Example 4.3 in [47, 39].

Corollary 4.12 (Quadratically overparametrized linear model)**.** *Consider the quadratically overparametrized linear model where the parametrization is given by $w = G(x) = u^{\odot 2} - v^{\odot 2}$ with $x = \binom{u}{v} \in \mathbb{R}^{2d}$. Suppose the loss function L is quasi-convex, ∇L is locally Lipschitz and $L(w) = \widetilde{L}(Zw)$ for some differentiable $\widetilde{L} : \mathbb{R}^n \to \mathbb{R}$. For any initialization $x_{init} = \binom{u_0}{v_0}$ such that all coordinates of x_{init} are non-zero, let $x(t)$ be the solution to the gradient flow (1) at time t. Then as $t \to \infty$, $w(t) = G(x(t))$ converges to some w_∞ with $Zw_\infty = Y$, and $R(w_\infty) = \min_{w:Zw=Y} R(w)$, where the mirror map R is given by*

$$R(w) = \frac{1}{4}\sum_{i=1}^{d}\Big(w_i \operatorname{arcsinh}\Big(\frac{w_i}{2u_{0,i}v_{0,i}}\Big) - \sqrt{w_i^2 + 4u_{0,i}^2 v_{0,i}^2} - w_i \ln \frac{u_{0,i}}{v_{0,i}}\Big).$$

5 Every mirror flow is a commutative gradient flow

Now we consider any $\mathcal{C}^2$ mirror map $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Recall the corresponding mirror flow:

$$\mathrm{d}\nabla R(w(t)) = -\nabla L(w(t))\mathrm{d}t.$$

We denote the interior of the domain of R by C, then C is a convex open set of $\mathbb{R}^d$, hence a smooth manifold (see Example 1.26 in [34]). Then $\nabla^2 R$ is a continuous positive-definite metric on C. As discussed previously, the above mirror flow can be further rewritten as the Riemannian gradient flow on the Riemannian manifold $(C, \nabla^2 R)$, i.e.,

$$\mathrm{d}w(t) = -\nabla^2 R(w(t))^{-1}\nabla L(w(t))\mathrm{d}t.$$

The goal is to find a commuting parametrization $G : M \to \mathbb{R}^d$ where M is smooth submanifold of $\mathbb{R}^D$ such that the dynamics of $w(t) = G(x(t))$ is given by the commutative gradient flow on $x(t)$ governed by $-\nabla_x L(G(x))$.

To illustrate the idea, let us first suppose such a G exists and is a bijection, whose inverse is denoted by F. Then it satisfies that $\nabla^2 R(w) = (\partial G(F(w))\partial G(F(w))^\top)^{-1}$, which further suggests that $\nabla^2 R(w) = \partial F(w)^\top \partial F(w)$. Note that this corresponds to expressing the metric tensor using an explicit map F, which is further equivalent to embed the Riemannian manifold $(\mathbb{R}^d, \nabla^2 R)$ into a Euclidean space $\mathbb{R}^D$ in a way that preserves its geometry. This refers to a notion called isometric embedding in differential geometry.

Definition 5.1 (Isometric embedding, p.15, [35])**.** Let (M, g) and $(\widetilde{M}, \tilde{g})$ be Riemannian manifolds. An *isometric embedding* is an embedding $F : M \to \widetilde{M}$ which preserves the metric in the sense that g is equal to the pullback of $\tilde{g}$ by F. More specifically, for any two tangent vectors $v, w \in T_x(M)$ we have $g(v, w) = \tilde{g}(\mathrm{d}F(v), \mathrm{d}F(w))$.

Nash's embedding theorem is a classic result in differential geometry that guarantees the existence of isometric embedding of any Riemannian manifold into a Euclidean space with a plain geometry.

Theorem 5.2 (Nash's embedding theorem, [44, 45, 24])**.** *Any d-dimensional Riemannian manifold has an isometric embedding in $\mathbb{R}^D$ for some $D \geq d$.*

By Nash's embedding theorem, we have a map $F : C \to \mathbb{R}^D$ that isometrically embeds $(C, \nabla^2 R)$ into $(\mathbb{R}^D, I_D)$. Denote $M = F(C)$, i.e., the embedding of C in $\mathbb{R}^D$. We denote the inverse of F by $G : M \to \mathbb{R}^d$, which satisfies the following (see Lemma D.1 and its proof):

$$\partial F(w)(\partial F(w)^\top \partial F(w))^{-1} = \partial G(F(w))^\top, \qquad \partial G(F(w))\partial G(F(w))^\top = \nabla^2 R(w)^{-1} \tag{7}$$

for all $w \in C$. This G is exactly the commuting parametrization that we seek for.

Theorem 5.3. *For any $\mathcal{C}^2$ mirror map $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, consider the corresponding mirror flow:*

$$\mathrm{d}\nabla R(w(t)) = -\nabla L(w(t))\mathrm{d}t.$$

There exists a commuting parametrization $G : M \to \mathbb{R}^d$ such that for any initialization of the mirror flow $w(0) = w_{init}$, it holds that $w(t) = G(x(t))$ for all t where $x(t)$ is given by the gradient flow governed by $-\nabla_x L(G(x))$ initialized at some x_{init}, i.e.,

$$\mathrm{d}x(t) = -\nabla_x L(G(x(t)))\mathrm{d}t, \quad x(0) = x_{init}.$$

Proof of Theorem 5.3. We follow the notation used in the previous discussion in this section and consider a more general case where we have a time-dependent loss $L_\cdot \in \mathcal{L}$. Given initialization w_{init}, let $w(t)$ be the solution to

$$\mathrm{d}w(t) = -\nabla^2 R(w(t))^{-1}\nabla L_t(w(t))\mathrm{d}t, \qquad w(0) = w_0.$$

Since F is an isometric embedding, we have $\nabla^2 R(w)_{ij} = \sum_{k=1}^{D} \frac{\partial F_k(w)}{\partial w_i}\frac{\partial F_k(w)}{\partial w_j}$ for all $i, j \in [d]$, which can be rewritten in the matrix form as $\nabla^2 R(w) = \partial F(w)^\top \partial F(w)$. Then we have

$$\mathrm{d}w(t) = -(\partial F(w(t))^\top \partial F(w(t)))^{-1}\nabla L_t(w(t))\mathrm{d}t.$$

Define $x(t) = F(w(t))$, and it follows that

$$\begin{aligned}\mathrm{d}x(t) &= \partial F(w(t))\mathrm{d}w(t) = -\partial F(w(t))(\partial F(w(t))^\top \partial F(w(t)))^{-1}\nabla L_t(w(t))\mathrm{d}t \\ &= -\partial G(F(w(t)))^\top \nabla L_t(w(t))\mathrm{d}t = -\nabla_x L_t(G(x(t)))\mathrm{d}t\end{aligned}$$

where the third equality follows from (7). Specializing $L_t \equiv L$ yields the first claim.

Next, we verify that $G : M \to \mathbb{R}^d$ is a commuting parametrization. First, for any $x \in M$, we have $\partial G(x)^\top = \partial F(G(x))(\partial F(G(x))^\top \partial F(G(x)))^{-1}$ by (7). Since $\nabla^2 R(w) = \partial F(w)^\top \partial F(w)$ is of rank d for all $w \in C$, it follows that $\partial F(w)$ is also of rank d for all $w \in C$, thus $\partial G(x)$ is of rank d for all $x \in M$. Second, we show that each ∇G_j is a complete vector field on M. For any $x_{\text{init}} \in M$, consider a loss $L(w) = \langle e_j, w\rangle$ and the corresponding gradient flow is

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

so $x(t) = \phi^t_{G_j}(x_{\text{init}})$ for all $t \geq 0$. On the other hand, $w(t) = G(x(t))$ satisfies that

$$\begin{aligned}\mathrm{d}w(t) &= \partial G(x(t))\mathrm{d}x(t) = -\partial G(x(t))\partial G(x(t))^\top \nabla L_t(w(t))\mathrm{d}t \\ &= -\nabla^2 R(w(t))^{-1}\nabla L_t(w(t))\mathrm{d}t = -\nabla^2 R(w(t))^{-1} e_j \mathrm{d}t\end{aligned}$$

where the last equality follows from (7). Therefore, rewriting the above as a Mirror Flow yields

$$\mathrm{d}\nabla R(w(t)) = -e_j \mathrm{d}t,$$

to which the solution exists for all $t \in \mathbb{R}$ and is given by $\nabla R(w(t)) = e_j t$, so $w(t) = (\nabla R)^{-1}(e_j t)$. This further implies that $x(t) = F(w(t))$ is well-defined for all $t \in \mathbb{R}$, hence ∇G_j is a complete vector field. Finally, to show the commutability of G, it suffices to use the same time-dependent loss as in Remark 4.7. This finishes the proof. □

6 Conclusion and future work

We presented a framework that characterizes when gradient descent with proper paramterization becomes equivalent to mirror descent. In the limit of infinitesimal step size, we identify a notion named commuting parametrization such that any gradient flow (i.e., the continuous analog of gradient descent) with a commuting parametrization is equivalent to a mirror flow (i.e., the continuous analog of mirror descent) in the original parameter space with respect to a mirror map that depends only on the initialization and the parametrization. Conversely, we use Nash's embedding theorem to show that any mirror flow can be characterized by a gradient flow in the reparametrized space with a commuting parametrization. Using our framework, we recover and generalize results on the implicit bias of gradient descent in a series of existing works, including a rigorous and general proof of convergence. However, it remains open whether gradient flow with a non-commuting parametrization can also be written as a mirror flow, which, in other words, requires a precise necessary condition.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 2019.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[3] Felipe Alvarez, Jérôme Bolte, and Olivier Brahic. Hessian riemannian gradient flows in convex programming. *SIAM journal on control and optimization*, 43(2):477–501, 2004.

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

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

[6] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[8] 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.

[9] Heinz H Bauschke, Jonathan M Borwein, et al. Legendre functions and the method of random bregman projections. *Journal of convex analysis*, 4(1):27–67, 1997.

[10] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[11] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

[12] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[13] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

[14] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

[15] Jean-Pierre Crouzeix. A relationship between the second derivatives of a convex function and of its conjugate. *Mathematical Programming*, 13(1):364–365, 1977.

[16] Alex Damian, Tengyu Ma, and Jason Lee. Label noise sgd provably prefers flat global minimizers. *arXiv preprint arXiv:2106.06530*, 2021.

[17] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[19] Rong Ge, Yunwei Ren, Xiang Wang, and Mo Zhou. Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[21] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[22] 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.

[23] Suriya Gunasekar, Blake Woodworth, and Nathan Srebro. Mirrorless mirror descent: A natural derivation of mirror descent. In *International Conference on Artificial Intelligence and Statistics*, pages 2305–2313. PMLR, 2021.

[24] Matthias Gunther. Isometric embeddings of riemannian manifolds, kyoto, 1990. In *Proc. Intern. Congr. Math.*, pages 1137–1143. Math. Soc. Japan, 1991.

[25] Jeff Z HaoChen, Colin Wei, Jason D Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. *arXiv preprint arXiv:2006.08680*, 2020.

[26] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[27] Arthur Jacot, François Ged, Franck Gabriel, Berfin Şimşek, and Clément Hongler. Deep linear networks dynamics: Low-rank biases induced by initialization scale and l2 regularization. *arXiv preprint arXiv:2106.15933*, 2021.

[28] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

[29] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[30] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

[31] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17176–17186. Curran Associates, Inc., 2020.

[32] Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

[33] Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.

[34] John M Lee. *Introduction to Smooth Manifolds*. Springer, 2013.

[35] John M Lee. *Introduction to Riemannian manifolds*. Springer, 2018.

[36] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

[37] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *arXiv preprint arXiv:1907.04595*, 2019.

[38] Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2020.

[39] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=siCt4xZn5Ve.

[40] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[41] Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34, 2021.

[42] Edward Moroshko, Blake E Woodworth, Suriya Gunasekar, Jason D Lee, Nati Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *Advances in neural information processing systems*, 33:22182–22193, 2020.

[43] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[44] John Nash. C1 isometric imbeddings. *Annals of mathematics*, pages 383–396, 1954.

[45] John Nash. The imbedding problem for riemannian manifolds. *Annals of mathematics*, pages 20–63, 1956.

[46] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[47] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34, 2021.

[48] Qian Qian and Xiaoyuan Qian. The implicit bias of adagrad on separable data. *Advances in Neural Information Processing Systems*, 32, 2019.

[49] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

[50] Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. *arXiv preprint arXiv:2201.11729*, 2022.

[51] Ralph Tyrell Rockafellar. Convex analysis. In *Convex analysis*. Princeton university press, 2015.

[52] 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(1):2822–2878, 2018.

[53] Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[55] Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pages 10849–10858. PMLR, 2021.

[56] Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, and Zhi-Ming Ma. Momentum doesn't change the implicit bias. *arXiv preprint arXiv:2110.03891*, 2021.

[57] 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.

[58] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[59] Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

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

[61] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

[62] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P Foster, and Sham Kakade. The benefits of implicit regularization from sgd in least squares problems. *Advances in Neural Information Processing Systems*, 34:5456–5468, 2021.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section X.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This paper only studies the theoretical properties of optimization algorithms.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We explicitly clarify the assumptions for each result.
   (b) Did you include complete proofs of all theoretical results? [Yes] Part of the proofs appears in the main context and others are deferred to appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] We do not have any experiments.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A] We do not use any existing assets.

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We do not use crowdsourcing nor conduct research with human subjects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]