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

Zhiyuan Li[*]
Princeton University
Princeton, NJ 08540
zhiyuanli@cs.princeton.edu

Tianhao Wang[*]
Yale University
New Haven, CT 06511
tianhao.wang@yale.edu

Jason D. Lee
Princeton University
Princeton, NJ 08540
jasonlee@princeton.edu

Sanjeev Arora
Princeton University
Princeton, NJ 08540
arora@cs.princeton.edu

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 Legendre function. Conversely, continuous mirror descent with any Legendre function 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 whereby the solution obtained from loss minimization has special properties that were not implied by value of the loss function and instead arose from the optimization's trajectory through the parameter space. 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 [61, 42, 41].

Many existing results about implicit bias view 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 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

[*]Equal contribution

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

$(L \circ G)(x) \equiv L(G(x))$ in the x-space, and the gradient flow in the x-space is given by[2]

$$\mathrm{d}x(t) = -\nabla(L \circ G)(x(t))\mathrm{d}t. \tag{1}$$

Using $w(t) = G(x(t))$ and the fact that $\nabla(L \circ G)(x) = \partial G(x)^\top \nabla L(G(x))$ 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 [26, 58, 64, 4, 61, 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 $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a strictly convex function [49, 10], which is called *mirror map* or *Legendre function* in literature. 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 there exists a Legendre function R such that $\partial G(x(t))\partial G(x(t))^\top = \nabla^2 R(w(t))^{-1}$ for all t, then (2) becomes a simple mirror flow in the w-space. 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 strictly 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 the loss L. A concrete example is that $w_i(t) = G_i(x(t)) = (x_i(t))^2$ for all $i \in [d]$, so here $D = d$. In this case, the Legendre function 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.*, $R(w) = \sum_{i=1}^d w_i(\ln w_i - 1)$.

However, in general, it is hard to decide *whether gradient flow for a given parametrization G can be written as mirror flow for some Legendre function 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 Legendre function cannot exist.[3] 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 Legendre function 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 4.8) and a slightly weaker necessary condition (Theorem 4.9) of when the gradient flow in the x-space governed by $-\nabla(L \circ G)$ is simulating a mirror flow in the w-space with respect to some Legendre function $R : \mathbb{R}^d \to \mathbb{R}$. This condition encompasses all the previous results [26, 58, 64, 4, 61, 5, 8]. Moreover, the Legendre function is independent of the loss L and depends only on the initialization x_{init} and the parametrization G.
- We recover and generalize existing implicit bias results for underdetermined linear regression as implications of the above characterization (Corollary 4.17). We also give new convergence analysis in such settings (Theorem 4.15), filling the gap in previous works [26, 61, 8] where parameter convergence is only assumed but not proved.
- In the reverse direction, we use the famous Nash's embedding theorem to show that every mirror flow in the w-space with respect to some Legendre function R simulates a gradient flow with commuting parametrization under some embedding $x = F(w)$ where $F : \mathbb{R}^d \to \mathbb{R}^D$ and the

[2] 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 $w = u^{\odot 2} - v^{\odot 2}$ [61]) and (b) deep linear nets [6] 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.

[3] To avoid such an issue, [5] has to assume all the preimages of G at w have the same $\partial G(\partial G)^\top$ and a recent paper [23] assumes that G is injective.

parametrization G is the inverse of F (Theorem 5.1). 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 [55, 24, 46, 35, 45, 34] and for non-separable data in [32, 33], 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 [61, 24, 64, 58, 4, 5], later generalized in [8].

The sharp contrast between the so-called *kernel* and *rich* regimes [61] reflects the importance of the initialization scale, where a large initialization often leads to the kernel regime with features barely changing during training [30, 16, 20, 19, 2, 1, 65, 7, 62, 31], while with a small initialization, the solution exhibits richer behavior with the resulting model having lower complexity [25, 26, 39, 52, 6, 15, 41, 43, 44, 53, 56, 22]. Recently [63] gave 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 [40, 11, 29, 42, 18, 66] and adaptive and momentum methods [51, 60, 59, 36], to name a few.

Understanding mirror descent. In the continuous-time regime, the mirror flow is equivalent to a Riemannian gradient flow with the metric tensor induced by the Legendre function. [27] 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. [23] generalized the equivalence result of [5] to discrete updates.

3 Preliminaries and notations

Notations. We denote $\mathbb{N}$ as the set of natural numbers. For any $n \in \mathbb{N}$, we denote $\{1, 2, \dots, n\}$ 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 $u \odot v = (u_1 v_1, \dots, u_D v_D)^\top$ and $u^{\odot \alpha} = ((u_1)^\alpha, \dots, (u_D)^\alpha)^\top$. For any $k \in \mathbb{N} \cup \{\infty\}$, we say a function f is $\mathcal{C}^k$ if it is k times continuously differentiable, and use $\mathcal{C}^k(M)$ to denote the set of all C^k functions from M to $\mathbb{R}$. We use $\circ$ to denote the composition of functions, *e.g.*, $f \circ g(x) = f(g(x))$. For any convex function $R : \mathbb{R}^D \to \mathbb{R} \cup \{\infty\}$, we denote its domain by $\operatorname{dom} R = \{w \in \mathbb{R}^D | R(w) < \infty\}$. For any set S, we denote its interior by $\operatorname{int}(S)$ and its closure by $\overline{S}$.

We assume that the model has parameter vector $w \in \mathbb{R}^d$ and $\mathcal{C}^1$ loss function $L : \mathbb{R}^d \to \mathbb{R}$. 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, and the objective is $L(G(x))$. 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.

In particular, we are interested in understanding the dynamics of gradient flow under the objective $L \circ G$ on some submanifold $M \subseteq \mathbb{R}^D$. Most of our results also generalize to the following notion of *time-dependent* loss.

Definition 3.1 (Time-dependent loss)**.** A time-dependent loss $L_t(w)$ is a function piecewise constant in time t and continuously differentiable in $w \in \mathbb{R}^d$, that is, there exist $k \in \mathbb{N}$, $0 = t_1 < t_2 < \cdots < t_{k+1} = \infty$ and $\mathcal{C}^1$ loss functions $L^{(1)}, L^{(2)}, \dots, L^{(k)}$ such that for each $i \in [k]$ and all $t \in [t_i, t_{i+1})$,

$$L_t(w) = L^{(i)}(w), \qquad \forall w \in \mathbb{R}^d.$$

We denote the set of such time-dependent loss functions by $\mathcal{L}$.

3.1 Manifold and vector field

Vector fields are a natural way to formalize the continuous-time gradient descent (a good reference is [38]). Let M be any smooth submanifold of $\mathbb{R}^D$. 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, which is denoted by T_xM. Formally, $T_xM := \{\frac{\mathrm{d}\gamma}{\mathrm{d}t}\big|_{t=0} \mid \forall \text{ smooth curves } \gamma : \mathbb{R} \to M, \gamma(0) = x\}$.

Definition 3.2 (Complete vector field; p.215, [38])**.** Let M be a smooth submanifold of $\mathbb{R}^D$ and X be a vector field on M. We say X is a *complete vector field* on M 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 $x(0) = x_{\text{init}}$.

Equipping the smooth submanifold $M \subseteq \mathbb{R}^D$ with a metric tensor g, we then have a Riemannian manifold (M, g), where for each $x \in M$, $g_x : T_xM \times T_xM \to \mathbb{R}$ is a positive definite bilinear form. In particular, the standard Euclidean metric $\overline{g}$ corresponds to $\overline{g}_x(u, v) = u^\top v$ for each $x \in M$ and $u, v \in T_xM$, under which the length of any arc on M is given by its length as a curve in $\mathbb{R}^D$.

For any differentiable function $f : M \to \mathbb{R}$, we denote by $\nabla_g f$ its gradient vector field with respect to metric tensor g. More specifically, $\nabla_g f(x)$ is defined as the unique vector in $\mathbb{R}^D$ such that $\nabla_g f(x) \in T_xM$ and $\frac{\mathrm{d}f(\gamma(t))}{\mathrm{d}t}\big|_{t=0} = g_x\big(\nabla f(x), \frac{\mathrm{d}\gamma(t)}{\mathrm{d}t}\big|_{t=0}\big)$ for any smooth curve $\gamma : \mathbb{R} \to M$ with $\gamma(0) = x$. Throughout the paper, we assume by default that the metric on the submanifold $M \subseteq \mathbb{R}^D$ is inherited from $(\mathbb{R}^D, \overline{g})$, and we will use ∇f as a shorthand for $\nabla_{\overline{g}} f$. If M is an open set of $\mathbb{R}^D$, ∇f is then simply the ordinary gradient of f.

For any $x \in M$ and $\mathcal{C}^1$ function $f : M \to \mathbb{R}$, 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, *i.e.*, the solution at time t (when it exists) of

$$\mathrm{d}\phi_f^t = -\nabla f(\phi_f^t)\mathrm{d}t, \qquad \phi_f^0(x) = x.$$

We say $\phi_f^t(x)$ is *well-defined* at time t when the above differential equation has a solution at time t. Moreover, for any differentiable function $X : M \to \mathbb{R}^d$, we define its Jacobian by

$$\partial X(x) = (\nabla X_1(x), \nabla X_2(x), \ldots, \nabla X_d(x))^\top.$$

Definition 3.3 (Lie bracket)**.** Let M be a smooth submanifold of $\mathbb{R}^D$. Given two C^1 vector fields X, Y on M, we define the *Lie bracket* of X and Y as $[X, Y](x) := \partial Y(x)X(x) - \partial X(x)Y(x)$.

3.2 Parametrizations

We use the term *parametrization* to refer to differentiable maps from a smooth submanifold of $\mathbb{R}^D$ (x-space) to $\mathbb{R}^d$ (w-space). We reserve G to denote parametrizations, and omit the dependence on G for notations of objects related to G when it is clear from the context.

The following notion of regular parametrization plays an important role in our analysis, and it is necessary for our main equivalence result between mirror flow and gradient flow with commuting parametrization. This is because if the null space of $\partial G(x)$ is non-trivial, *i.e.,* it contains some vector $u \neq 0$, then the gradient flow with parametrization G obviously cannot simulate any mirror flow with nonzero velocity in the direction of u.

Definition 3.4 (Regular parametrization)**.** Let M be a smooth submanifold of $\mathbb{R}^D$. A *regular parametrization* $G : M \to \mathbb{R}^d$ is a $\mathcal{C}^1$ parametrization such that $\partial G(x)$ is of rank d for all $x \in M$.

Note that a regular parametrization G can become irregular when its domain is changed. For example, $G(x) = x^2$ is regular on $\mathbb{R}_+$, but it is not regular on $\mathbb{R}$ as $\partial G(0) = 0$.

Given a $\mathcal{C}^2$ parametrization $G : M \to \mathbb{R}^d$, for any $x \in M$ and $\mu \in \mathbb{R}^d$, we define

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

when it is well-defined, *i.e.*, the corresponding integral equation has a solution. For any $x \in M$, we define the domain of $\psi(x; \cdot)$ as

$$\mathcal{U}(x) = \big\{\mu \in \mathbb{R}^d \mid \psi(x; \mu) \text{ is well-defined}\big\}. \tag{5}$$

When every ∇G_i is a complete vector field on M as in Definition 3.2, we have $\mathcal{U}(x) = \mathbb{R}^d$. However, such completeness assumption is relatively strong, and most polynomials would violate it. For

example, consider $G(x) = x^{\odot 3}$ for $x \in \mathbb{R}^d$, then the solution to $\mathrm{d}x_i(t) = 3x_i(t)^2\mathrm{d}t$ explodes in finite time for each $i \in [d]$. To relax this, we consider parametrizations such that the domain of the flows induced by its gradient vector fields is pairwise symmetric. More specifically, we define

$$\mathcal{U}_{ij}(x) = \left\{(s,t) \in \mathbb{R}^2 \mid \phi^s_{G_i} \circ \phi^t_{G_j}(x) \text{ is well-defined}\right\}$$

for any $x \in M$ and $i, j \in [d]$, and we make the following assumption.

Assumption 3.5. Let M be a smooth submanifold of $\mathbb{R}^D$ and $G : M \to \mathbb{R}^d$ be a parametrization. We assume that for any $x \in M$ and $i \in [d]$, $\phi^t_x(x)$ is well-defined for $t \in (T_-, T_+)$ such that either $\lim_{t \to T_+} \|\phi^t_x(x)\|_2 = \infty$ or $T_+ = \infty$ and similarly for T_-. Also, we assume that for any $x \in M$ and $i, j \in [d]$, it holds that $\mathcal{U}_{ji}(x) = \{(t,s) \in \mathbb{R}^2 \mid (s,t) \in \mathcal{U}_{ij}(x)\}$, i.e., $\phi^s_{G_i} \circ \phi^t_{G_j}(x)$ is well-defined if and only if $\phi^t_{G_j} \circ \phi^s_{G_i}(x)$ is.

Indeed, under Assumption 3.5, we can show that for any $x \in M, \mathcal{U}(x)$ is a hyperrectangle in $\mathbb{R}^d$, i.e,

$$\mathcal{U}(x) = \mathcal{I}_1(x) \times \mathcal{I}_2(x) \times \cdots \times \mathcal{I}_d(x) \quad \text{where each } \mathcal{I}_j(x) \subset \mathbb{R} \text{ is an open interval.} \tag{6}$$

See Lemma C.1 and its proof in Appendix C. Next, for any initialization $x_{\text{init}} \in M$, the set of points that are reachable via gradient flow under some time-dependent loss (see Definition 3.1) with parametrization G is a subset of M that depends on G and x_{init}.

Definition 3.6 (Reachable set)**.** Let M be a smooth submanifold of $\mathbb{R}^D$. For any $\mathcal{C}^2$ parametrization $G : M \to \mathbb{R}^d$ and any initialization $x_{\text{init}} \in M$, the reachable set $\Omega_x(x_{\text{init}}; G)$ is defined as

$$\Omega_x(x_{\text{init}}; G) = \left\{ \phi^{\mu_1}_{L_1 \circ G} \circ \phi^{\mu_2}_{L_2 \circ G} \circ \cdots \circ \phi^{\mu_k}_{L_k \circ G}(x_{\text{init}}) \;\middle|\; \forall k \in \mathbb{N}, \forall i \in [k], L_i \in \mathcal{C}^1(\mathbb{R}^d), \mu_i \geq 0 \right\}.$$

It is clear that the above definition induces a transitive "reachable" relationship between points on M, and it is also reflexive since for all $L \in \mathcal{C}^1(\mathbb{R}^d)$ and $t > 0$, $\phi^t_{L \circ G} \circ \phi^t_{(-L) \circ G}$ is the identity map on the domain of $\phi^t_{-L \circ G}$. In this sense, the reachable sets are orbits of the family of gradient vector fields $\{\nabla(L \circ G) \mid L \in \mathcal{C}^1(\mathbb{R}^d)\}$, *i.e.*, the reachable sets divide the domain M into equivalent classes. The above reachable set in the x-space further induces the corresponding reachable set in the w-space given by $\Omega_w(x_{\text{init}}; G) = G(\Omega_x(x_{\text{init}}; G))$.

In most natural examples, the parametrization G is smooth (though this is not necessary for our results), and by Sussman's Orbit Theorem [57], each reachable set $\Omega_x(x_{\text{init}}; G)$ is an immersed submanifold of M. Moreover, it follows that $\Omega_x(x_{\text{init}}; G)$ can be generated by $\{\nabla G_i\}_{i=1}^d$, *i.e.*, $\Omega_x(x_{\text{init}}; G) = \{\phi^{\mu_1}_{G_{j_1}} \circ \phi^{\mu_2}_{G_{j_2}} \circ \cdots \circ \phi^{\mu_k}_{G_{j_k}}(x_{\text{init}}) \mid \forall k \in \mathbb{N}, \forall i \in [k], j_i \in [d], \mu_i \geq 0\}$.

3.3 Mirror descent and mirror flow

Next, we introduce some basic notions for mirror descent [49, 10]. We refer the readers to Appendix B for more preliminaries on convex analysis.

Definition 3.7 (Legendre function and mirror map)**.** Let $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a differentiable convex function. We say R is a *Legendre function* when it satisfies that (1) R is strictly convex on $\operatorname{int}(\operatorname{dom} R)$, and (2) for any sequence $\{w_i\}_{i=1}^\infty$ going to the boundary of $\operatorname{dom} R$, $\lim_{i \to \infty} \|\nabla R(w_i)\|_2 = \infty$. In particular, we call R a *mirror map* if R further satisfies that the gradient map $\nabla R : \operatorname{int}(\operatorname{dom} R) \to \mathbb{R}^d$ is surjective (see p.298 in [13]).

Given a Legendre function $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, for any initialization $w_0 = w_{\text{init}} \in \operatorname{int}(\operatorname{dom} R)$, mirror descent with step size η updates as follows:

$$\nabla R(w_{k+1}) = \nabla R(w_k) - \eta \nabla L(w_k). \tag{7}$$

Usually ∇R is required to be surjective so that after a discrete descent step in the dual space, it can be projected back to the primal space via $(\nabla R)^{-1}$. Nonetheless, as long as $\nabla R(w_k) - \eta \nabla L(w_k)$ is in the range of ∇R, the above discrete update is well-defined. In the limit of $\eta \to 0$, (7) becomes the continuous mirror flow:

$$\mathrm{d}\nabla R(w(t)) = -\nabla L(w(t))\mathrm{d}t. \tag{8}$$

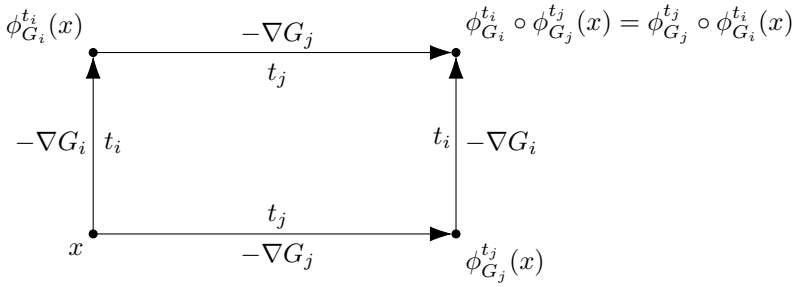

Figure 1: Illustration of commuting parametrizations. Suppose $G : M \to \mathbb{R}^d$ is a commuting parametrization satisfying Assumption 3.5, then starting from any $x \in M$, first moving along $-\nabla G_i$ for time t_i then moving along $-\nabla G_j$ for time t_j yields the same result as first moving along $-\nabla G_j$ for time t_j then moving along $-\nabla G_i$ for time t_i does, i.e., $\phi^{t_i}_{G_i} \circ \phi^{t_j}_{G_j}(x) = \phi^{t_j}_{G_j} \circ \phi^{t_i}_{G_i}(x)$.

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 [24] (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 minimizes some convex regularizer among all optimal solutions.

Theorem 3.8. *Given any data $Z \in \mathbb{R}^{n \times d}$ and 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 $w(0) = w_{\text{init}}$, the mirror flow (8) 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 C 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 4.16).

4 Every gradient flow with commuting parametrization is a mirror flow

4.1 Commuting parametrization

We now formalize the notion of commuting parametrization. We remark that M is a smooth submanifold of $\mathbb{R}^D$, and it is the domain of the parametrization G.

Definition 4.1 (Commuting parametrization)**.** Let M be a smooth submanifold of $\mathbb{R}^D$. A $\mathcal{C}^2$ parametrization $G : M \to \mathbb{R}^d$ is *commuting* in a subset $S \subseteq M$ if and only if for any $i, j \in [d]$, the Lie bracket $[\nabla G_i, \nabla G_j](x) = 0$ for all $x \in S$. Moreover, we say G is a *commuting parametrization* if it is commuting in the entire M.

In particular, when M is an open subset of $\mathbb{R}^d$, $\{\nabla G_i\}_{i=1}^d$ are ordinary gradients in $\mathbb{R}^D$, and the Lie bracket between any pair of ∇G_i and ∇G_j is given by $[\nabla G_i, \nabla G_j](x) = \nabla^2 G_j(x) \nabla G_i(x) - \nabla^2 G_i(x) \nabla G_j(x)$. This provides an easy way to check whether G is commuting or not.

The above definition of commuting parametrizations builds upon the differential properties of the gradient vector fields $\{\nabla G_i\}_{i=1}^d$, where each Lie bracket $[\nabla G_i, \nabla G_j]$ quantifies the change of ∇G_j along the flow generated by ∇G_i. Indeed, the above characterization of 'commuting' is further equivalent to another characterization in the integral form (Theorem 4.2), as illustrated in Figure 1.

Theorem 4.2. *Let M be a smooth submanifold of $\mathbb{R}^D$ and $G : M \to \mathbb{R}^d$ be a $\mathcal{C}^2$ parametrization satisfying Assumption 3.5. For any $i, j \in [d]$, $[\nabla G_i, \nabla G_j](x) = 0$ for all $x \in M$ if and only if for any $x \in M$, it holds that $\phi^s_{G_i} \circ \phi^t_{G_j}(x) = \phi^t_{G_j} \circ \phi^s_{G_i}(x)$ for all $(s, t) \in \mathcal{I}_i(x) \times \mathcal{I}_j(x)$, where $\mathcal{I}_i(x)$ and $\mathcal{I}_j(x)$ are the time domains of $\phi^s_{G_i}(x)$ and $\phi^t_{G_j}(x)$ as defined in (6).*

The commuting condition clearly holds when each G_i only depends on a different subset of coordinates of x, because we then have $\nabla^2 G_i(\cdot) \nabla G_j(\cdot) \equiv 0$ for any distinct $i, j \in [d]$ as $\nabla^2 G_i$ and ∇G_j live in different subspaces of $\mathbb{R}^D$. We call such G *separable parametrizations*[4], and this case covers all the previous examples [26, 58, 4, 61, 5]. Another interesting example is the *quadratic parametrization*: We parametrize $w \in \mathbb{R}^d$ by $G : \mathbb{R}^D \to \mathbb{R}^d$ where for each $i \in [d]$, there is a symmetric matrix

[4]We further discuss the existence of non-separable commuting parametrizations in Appendix A.2.

$A_i \in \mathbb{R}^{D\times D}$ such that $G_i(x) = \frac{1}{2}x^\top A_i x$. Then each $[\nabla G_i, \nabla G_j](x) = (A_jA_i - A_iA_j)x$, and thus G is a commuting parametrization if and only if matrices $\{A_i\}_{i=1}^d$ commute.

For concreteness, we analyze two examples below. The first one is both a separable parametrization and a commuting quadratic parametrization, while the second one is quadratic but non-commuting.

Example 4.3 ($u^{\odot 2} - v^{\odot 2}$ parametrization, [61])**.** Parametrize $w \in \mathbb{R}^d$ by $w = u^{\odot 2} - v^{\odot 2}$. Here $D = 2d$, and the parametrization G is given by $G(x) = u^{\odot 2} - v^{\odot 2}$ for $x = \binom{u}{v} \in \mathbb{R}^D$. Since each $G_i(x)$ involves only u_i and v_i, G is a separable parametrization and hence a commuting parametrization. Meanwhile, each $G_i(x)$ is a quadratic form in x, and it can be directly verified that the matrices underlying these quadratic forms commute with each other.

Example 4.4 (Matrix factorization)**.** As a counter-example, consider two parametrizations for matrix factorization: $G(U) = UU^\top$ and $G(U, V) = UV^\top$, where $U, V \in \mathbb{R}^{d\times r}$ and $d \geq 2, r \geq 1$. These are both *non-commuting* quadratic parametrizations. Here we only demonstrate for the parametrization $G(U) = UU^\top$, and $G(U, V) = UV^\top$ follows a similar argument. For each $i, j \in [d]$, we define $E_{ij} \in \mathbb{R}^d$ as the one-hot matrix with the (i, j)-th entry being 1 and the rest being 0, and denote $\overline{E}_{ij} = \frac{1}{2}(E_{ij} + E_{ji})$. For $r = 1$, we have $G_{ij}(U) = U_iU_j = U^\top \overline{E}_{ij}U$ for any $i, j \in [d]$, so G is a quadratic parametrization. Note that $\overline{E}_{ii}\overline{E}_{ij} = \frac{1}{2}E_{ij} \neq \frac{1}{2}E_{ji} = \overline{E}_{ij}\overline{E}_{ii}$ for all distinct $i, j \in [d]$, which implies that $[\nabla G_{ij}, \nabla G_{ii}] \neq 0$, so G is non-commuting. More generally, we can reshape U as a vector $\overrightarrow{U} := [U_{:1}^\top, \ldots, U_{:r}^\top]^\top \in \mathbb{R}^{rd}$ where each $U_{:j}$ is the j-th column of U, and the resulting quadratic form for the (i, j)-entry of $G(U)$ corresponds to a block-diagonal matrix:

$$G_{ij}(U) = (\overrightarrow{U})^\top \mathrm{diag}(\overline{E}_{ij}, \ldots, \overline{E}_{ij})\overrightarrow{U}.$$

Therefore, $\nabla^2 G_{ij}$ does not commute with $\nabla^2 G_{ii}$ due to the same reason as in the rank-1 case.

Remark 4.5. *This non-commuting issue for general matrix factorization does not conflict with the theoretical analysis in [26] where the measurements are commuting, or equivalently, only involve diagonal elements, as $\{G_{ii}\}_{i=1}^d$ are indeed commuting parametrizations. [26] is the first to identify the above non-commuting issue and conjectured that the implicit bias result for diagonal measurements can be extended to the general case.*

4.2 Main equivalence result

Next, we proceed to present our analysis for gradient flow with commuting parametrization. The following two lemmas highlight the special properties of commuting parametrizations. Lemma 4.6 shows that the point reached by gradient flow with any commuting parametrization is determined by the integral of the negative gradient of the loss along the trajectory.

Lemma 4.6. *Let M be a smooth submanifold of $\mathbb{R}^D$ and $G : M \to \mathbb{R}^d$ be a commuting parametrization. For any initialization $x_{\text{init}} \in M$, consider the gradient flow for any time-dependent loss $L_t \in \mathcal{L}$ as in Definition 3.1: $\mathrm{d}x(t) = -\nabla(L_t \circ G)(x(t))\mathrm{d}t$, $x(0) = x_{\text{init}}$. Further define $\mu(t) = \int_0^t -\nabla L_t(G(x(s)))\mathrm{d}s$. Suppose $\mu(t) \in \mathcal{U}(x_{\text{init}})$ for all $t \in [0, T)$ where $T \in \mathbb{R} \cup \{\infty\}$, then it holds that $x(t) = \psi(x_{\text{init}}; \mu(t))$ for all $t \in [0, T)$.*

Based on Lemma 4.6, the next key lemma reveals the essential approach to find the Legendre function.

Lemma 4.7. *Let M be a smooth submanifold of $\mathbb{R}^D$ and $G : M \to \mathbb{R}^d$ be a commuting and regular parametrization satisfying Assumption 3.5. Then for any $x_{\text{init}} \in M$, there exists a Legendre function $Q : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ such that $\nabla Q(\mu) = G(\psi(x_{\text{init}}; \mu))$ for all $\mu \in \mathcal{U}(x_{\text{init}})$. Moreover, let R be the convex conjugate of Q, then R is also a Legendre function and $\mathrm{int}(\mathrm{dom}\, R) = \Omega_w(x_{\text{init}}; G)$ and $\nabla^2 R(G(\psi(x_{\text{init}}; \mu))) = \big(\partial G(\psi(x_{\text{init}}; \mu))\partial G(\psi(x_{\text{init}}; \mu))^\top\big)^{-1}$ for all $\mu \in \mathcal{U}(x_{\text{init}})$.*

Next, we present our main result on characterization of gradient flow with commuting parametrization.

Theorem 4.8. *Let M be a smooth submanifold of $\mathbb{R}^D$ and $G : M \to \mathbb{R}^d$ be a commuting and regular parametrization satisfying Assumption 3.5. For any initialization $x_{\text{init}} \in M$, consider the gradient flow for any time-dependent loss function $L_t : \mathbb{R}^d \to \mathbb{R}$:*

$$\mathrm{d}x(t) = -\nabla(L_t \circ G)(x(t))\mathrm{d}t, \qquad x(0) = x_{\text{init}}.$$

Define $w(t) = G(x(t))$ for all $t \geq 0$, then the dynamics of $w(t)$ is a mirror flow with respect to the Legendre function R given by Lemma 4.7, i.e.,

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

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

Theorem 4.8 provides a sufficient condition for when a gradient flow with certain parametrization G is simulating a mirror flow. The next question is then: What are the necessary conditions on the parametrization G so that it enables the gradient flow to simulate a mirror flow? We provide a (partial) characterization of such G in the following theorem.

Theorem 4.9 (Necessary condition on smooth parametrization to be commuting)**.** *Let M be a smooth submanifold of $\mathbb{R}^D$ and $G : M \to \mathbb{R}^d$ be a smooth parametrization. If for any $x_{\text{init}} \in M$, there is a Legendre function R such that for all time-dependent loss $L_t \in \mathcal{L}$, the gradient flow under $L_t \circ G$ initialized at x_{init} can be written as the mirror flow under L_t with respect to R, then G must be a regular parametrization, and it also holds that for each $x \in M$,*

$$\text{Lie}^{\geq 2}(\partial G)\big|_x \subseteq \ker(\partial G(x)), \tag{9}$$

where $\text{Lie}^{\geq K}(\partial G) := \text{span}\big\{[[[[\nabla G_{j_1}, \nabla G_{j_2}], \ldots], \nabla G_{j_{k-1}}], \nabla G_{j_k}] \mid k \geq K, \forall i \in [k], j_i \in [d]\big\}$ is the subset of the Lie algebra generated by $\{\nabla G_i\}_{i=1}^d$ only containing elements of order higher than K, and $\ker(\partial G(x))$ is the orthogonal complement of $\text{span}(\{\nabla G_i(x)\}_{i=1}^d)$ in $\mathbb{R}^D$.

Note the necessary condition in (9) is weaker than assuming that G is a commuting parametrization, and we conjecture that it is indeed sufficient.

Conjecture 4.10. The claim in Theorem 4.8 still holds, if we relax the commuting assumption to that $\text{Lie}^{\geq 2}(\partial G)\big|_x \subseteq \ker(\partial G(x))$ for all $x \in M$.

With the above necessary condition (9), we can formally refute the possibility that one can use mirror flow to characterize the implicit bias of gradient flow for matrix factorization in general settings, as summarized in Corollary 4.11. It is also worth mentioning that [40] constructed a concrete counter example showing that the implicit bias for commuting measurements, that gradient flow finds the solution with minimal nuclear norm, does not hold for the general case, where gradient flow could prefer the solution with minimal rank instead.

Corollary 4.11 (Gradient flow for matrix factorization cannot be written as mirror flow)**.** *For any $d, r \in \mathbb{N}$, let M be an open set in $\mathbb{R}^{d\times r}$ and $G : M \to \mathbb{R}^{d\times d}$ be a smooth parametrization given by $G(U) = UU^\top$. Then there exists a initial point $U_{init} \in M$ and a time-dependent loss L_t such that the gradient flow under $L_t \circ G$ starting from U_{init} cannot be written as a mirror flow with respect to any Legendre function R under the loss L_t.*

The following corollary shows that gradient flow with non-commuting parametrization cannot be mirror flow, when the dimension of the reachable set matches that of the w-space.

Corollary 4.12. *Let M be a smooth submanifold of $\mathbb{R}^D$ whose dimension is at least d. Let $G : M \to \mathbb{R}^d$ be a regular parametrization such that for any $x_{\text{init}} \in M$, (1) $\Omega_x(x_{\text{init}}; G)$ is a submanifold of dimension d, and (2) there is a Legendre function R such that for any time-dependent loss $L_t \in \mathcal{L}$, the gradient flow governed by $-\nabla(L_t \circ G)$ with initialization x_{init} can be written as a mirror flow with respect to R. Then G must be a commuting parametrization.*

Next, we establish the convergence of $w(t) = G(x(t))$ when $x(t)$ is given by some gradient flow with the commuting parametrization G. Here we require that the convex function R given by Lemma 4.7 is a Bregman function (see definition in Appendix B). The proofs of Theorem 4.13, Corollary 4.14 and Theorem 4.15 are in Appendix D.

Theorem 4.13. *Under the setting of Theorem 4.8, further assume that the loss L is quasi-convex, ∇L is locally Lipschitz and $\text{argmin}\{L(w) \mid w \in \text{dom } R\}$ is non-empty where $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is the convex function given by Lemma 4.7. Suppose R is a Bregman function, then as $t \to \infty$, $w(t)$ converges to some w^* such that $\nabla L(w^*)^\top (w - w^*) \geq 0$ for all $w \in \text{dom } R$. Moreover, if the loss function L is convex, then $w(t)$ converges to a minimizer in $\overline{\text{dom } R}$.*

Corollary 4.14. *Under the setting of Theorem 4.13, if the reachable set in the w-space satisfies $\Omega_w(x_{\text{init}}; G) = \mathbb{R}^d$, then R is a Bregman function and all the statements in Theorem 4.13 hold.*

Theorem 4.15. *Under the setting of Theorem 4.13, consider the commuting quadratic parametrization $G : \mathbb{R}^D \to \mathbb{R}^d$ where each $G_i(x) = \frac{1}{2} x^\top A_i x$, for symmetric matrices $A_1, A_2, \ldots, A_d \in \mathbb{R}^{D\times D}$ that commute with each other,* i.e., *$A_i A_j - A_j A_i = 0$ for all $i, j \in [d]$. For any $x_{\text{init}} \in \mathbb{R}^D$, if $\{\nabla G_i(x_{\text{init}})\}_{i=1}^d = \{A_i x_{\text{init}}\}_{i=1}^d$ are linearly independent, then the following holds:*

(a) *For all* $\mu \in \mathbb{R}^d$, $\psi(x_{\text{init}}; \mu) = \exp(\sum_{i=1}^d \mu_i A_i) x_{\text{init}}$ *where* $\exp(\cdot)$ *is the matrix exponential defined as* $\exp(A) := \sum_{k=0}^{\infty} \frac{A^k}{k!}$.

(b) *For each* $j \in [d]$ *and all* $\mu \in \mathbb{R}^d$, $G_j(\psi(x_{\text{init}}; \mu)) = \frac{1}{2} x_{\text{init}}{}^\top \exp(\sum_{i=1}^d 2\mu_i A_i) A_j x_{\text{init}}$.

(c) $Q(\mu) = \frac{1}{4} \|\psi(x_{\text{init}}; \mu)\|_2^2$ *is a Legendre function with domain* $\mathbb{R}^d$.

(d) R *is a Bregman function with* $\operatorname{dom} R = \overline{\operatorname{range} \nabla Q}$ *where* $\operatorname{range} \nabla Q$ *is the range of* ∇Q, *and thus all the statements in Theorem 4.13 hold.*

4.3 Solving underdetermined linear regression with commuting parametrization

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 G, 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 3.8 to show the implicit bias of gradient flow with commuting parametrization.

Theorem 4.16. *Let* M *be a smooth submanifold of* $\mathbb{R}^d$ *and* $G : M \to \mathbb{R}^d$ *be a commuting and regular parametrization satisfying Assumption 3.5. Suppose the loss function* L *satisfies* $L(w) = \widetilde{L}(Zw)$ *for some differentiable* $\widetilde{L} : \mathbb{R}^n \to \mathbb{R}$. *For any initialization* $x_{\text{init}} \in M$, *consider the gradient flow*

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

There exists a convex function R *(given by Lemma 4.7, depending only on* x_{init} *and* 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 as* $t \to \infty$ *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.*

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 studied in a series of papers [26, 61, 8], as summarized in the following Corollary 4.17. Note that their results assumed convergence in order to characterize the implicit bias, whereas our framework enables us to directly prove the convergence as in Theorem 4.15. The convergence guarantee here is also more general than existing convergence results for Example 4.3 in [50, 42].

Corollary 4.17. *Consider the underdetermined linear regression problem with data* $Z \in \mathbb{R}^{d \times n}$ *and* $Y \in \mathbb{R}^n$. *Let* $\widetilde{L} : \mathbb{R}^n \to \mathbb{R}$ *be a differentiable loss function such that* $\widetilde{L}$ *is quasi-convex,* $\nabla \widetilde{L}$ *is locally Lipschitz, and* $Y \in \mathbb{R}^n$ *is its unique global minimizer. Consider solving* $\min_w \widetilde{L}(Zw)$ *by running gradient flow on* $L(w) = \widetilde{L}(Zw)$ *with the quadratic parametrization* $w = G(x) = u^{\odot 2} - v^{\odot 2}$ *where* $x = \binom{u}{v} \in \mathbb{R}_+^{2d}$, *for any initialization* $x_{\text{init}} \in \mathbb{R}_+^{2d}$: $\mathrm{d}x(t) = -\nabla (L \circ G)(x(t)) \mathrm{d}t$, $x(0) = x_{\text{init}}$. *Then as* $t \to \infty$, $w(t) = G(x(t))$ *converges to some* w_∞ *such that* $Zw_\infty = Y$ *and* $R(w_\infty) = \min_{w: Zw = Y} R(w)$ *where* R *is given by*

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

5 Every mirror flow is a gradient flow with commuting parametrization

For any smooth Legendre function $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, recall the corresponding mirror flow:

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

Note that $\operatorname{int}(\operatorname{dom} R)$ is a convex open set of $\mathbb{R}^d$, hence a smooth manifold (see Example 1.26 in [38]), and $\nabla^2 R$ is a continuous positive-definite metric on $\operatorname{int}(\operatorname{dom} R)$. As discussed previously in (3), the above mirror flow is the Riemannian gradient flow on the Riemannian manifold $(\operatorname{int}(\operatorname{dom} R), \nabla^2 R)$. The goal is to find a parametrization $G : U \to \mathbb{R}^d$, where U is an open set of $\mathbb{R}^D$, such that the dynamics of $w(t) = G(x(t))$ can be induced by the gradient flow on $x(t)$ governed by $-\nabla (L \circ G)(x)$. Formally, we have the following result:

Theorem 5.1. *Let $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a smooth Legendre function. There exist a smooth submanifold of $\mathbb{R}^D$ denoted by M, an open neighborhood U of M and a smooth and regular parametrization $G : U \to \mathbb{R}^d$ such that for the mirror flow under any time-dependent loss L_t with any initialization $w_{\rm init} \in \mathrm{int}(\mathrm{dom}\, R)$*

$$\mathrm{d}\nabla R(w(t)) = -\nabla L_t(w(t))\mathrm{d}t, \quad w(0) = w_{\rm init}, \tag{10}$$

it holds that $w(t) = G(x(t))$ for all $t \geq 0$ where $x(t)$ is given by the gradient flow under $L_t \circ G$:

$$\mathrm{d}x(t) = -\nabla(L_t \circ G)(x(t))\mathrm{d}t, \quad x(0) = x_{\rm init} \tag{11}$$

where $x_{\rm init}$ satisfies $G(x_{\rm init}) = w_{\rm init}$. Moreover, let $G|_M$ be the restriction of G on M, then $G|_M$ is a commuting and regular parametrization and $\partial G = \partial G|_M$ on M, which implies $x(t) \in M$ for all $t \geq 0$. If R is further a mirror map, then $\{\nabla G_i|_M\}_{i=1}^d$ are complete vector fields on M.

The proof of Theorem 5.1 can be found in Appendix E. To illustrate the idea, let us first suppose such a smooth and regular parametrization G exists and is a bijection between the reachable set $\Omega_x(x_{\rm init}; G) \subset \mathbb{R}^D$ and $\mathrm{int}(\mathrm{dom}\, R)$, and denote its inverse by F. It turns out that we can show

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

where the second equality follows from the relationship between R and G as discussed in the introduction on (2). Note that this corresponds to expressing the metric tensor $\nabla^2 R$ using an explicit map F, which is further equivalent to embedding the Riemannian manifold $(\mathrm{int}(\mathrm{dom}\, R), \nabla^2 R)$ into a Euclidean space $(\mathbb{R}^D, \overline{g})$ in a way that preserves its metric. This refers to a notion called isometric embedding in differential geometry.

Definition 5.2 (Isometric embedding)**.** Let (M, g) be a Riemannian submanifold of $\mathbb{R}^d$. An *isometric embedding* from (M, g) to $(\mathbb{R}^D, \overline{g})$ is a differentiable injective map $F : M \to \mathbb{R}^D$ that preserves the metric, i.e., for any two tangent vectors $v, w \in T_x M$ it holds that $g_x(v, w) = \overline{g}_x(\partial F(x)v, \partial F(x)w)$.

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. See Appendix A.1 for additional discussion on construction of G given a Legendre function R.

Theorem 5.3 (Nash's embedding theorem, [47, 48, 28])**.** *Any d-dimensional Riemannian manifold has an isometric embedding to $(\mathbb{R}^D, \overline{g})$ for $D = \max\{d(d+5)/2, d(d+3)/2 + 5\}$.*

As another way to understand Theorem 4.8, note that $\nabla^2 R(w)^{-1}\nabla L(w)$ is the Riemannian gradient of L on the Riemannian manifold $(\mathrm{int}(\mathrm{dom}\, R), \nabla^2 R)$. It is well-known that gradient flow is invariant under isometric embedding, and thus we can use Nash's embedding theorem to rewrite the Riemannian gradient flow on $(\mathrm{int}(\mathrm{dom}\, R), g^R)$ as that on $(\mathbb{R}^D, \overline{g})$.

6 Conclusion

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 Legendre function 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. We also provide a necessary condition for the parametrization such that gradient flow in the reparametrized space is equivalent to a mirror flow in the original space. However, the necessary condition is slightly weaker than the commuting condition and it is left for future work to close the gap.

Acknowledgement

This work was supported by NSF, DARPA/SRC, Simons Foundation, and ONR. ZL acknowledges support of Microsoft Research PhD Fellowship and JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, ONR Young Investigator Award, and NSF CAREER Award 2144994.

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] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

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

[14] Yair Censor and Arnold Lent. An iterative row-action method for interval convex programming. *Journal of Optimization theory and Applications*, 34(3):321–353, 1981.

[15] 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.

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

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

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

[19] 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.

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

[21] Robert L Foote. Regularity of the distance function. *Proceedings of the American Mathematical Society*, 92(1):153–155, 1984.

[22] 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.

[23] Udaya Ghai, Zhou Lu, and Elad Hazan. Non-convex online learning via algorithmic equivalence. *arXiv preprint arXiv:2205.15235*, 2022.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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.

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

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

[34] 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.

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

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

[37] Serge Lang. *Introduction to differentiable manifolds*. Springer Science & Business Media, 2006.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

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

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

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

[50] 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.

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

[52] 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.

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

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

[55] 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.

[56] 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.

[57] Héctor J Sussmann. Orbits of families of vector fields and integrability of distributions. *Transactions of the American Mathematical Society*, 180:171–188, 1973.

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

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

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

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

[66] 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]

A Additional Results

We provide additional results summarized as follows. We discuss how to construct the parametrization G from a given Legendre function R in Appendix A.1. We discuss the existence of non-separable commuting parametrization in Appendix A.2.

A.1 Examples of constructing G from a given Legendre function R

Since the construction of the isometric embedding in Nash's embedding theorem is not explicit and infeasible to compute in general, the corresponding parametrization G given by Theorem 5.1 does not admit an analytic formula for general mirror map R. However, in many cases involving standard convex functions R which are separable, it is indeed tractable to explicitly compute the corresponding parametrization G.

For example, for Burg entropy ($R(w) = -\sum_{i=1}^d \log w_i$) and negative entropy ($R(w) = \sum_{i=1}^d w_i \log w_i$), it is easy to verify that we can choose $w = G(x) = (e^{x_1}, \dots, e^{x_d})$ and $w = G(x) = (x_1^2/4, \dots, x_d^2/4)$ respectively, where in both cases x has the same dimension as w does. (See also Example 2 and 4 in [5].) More generally, suppose R satisfies that $\nabla^2 R(w)$ is always diagonal, it suffices to find a $F_i : \mathbb{R} \to \mathbb{R}$ such that $F_i'(w_i) = \sqrt{\partial_{ii} R(w_i)}, \forall w_i \in \mathbb{R}$, which can be solved easily by integral. Once we have F_i, the parametrization $G : \mathbb{R}^d \to \mathbb{R}^d$ defined by $G_i = F_i^{-1}$ is the desired commuting parametrization with respect to which gradient flow can be written as mirror flow with respect to R. (Note G_i is well-defined because F_i is monotone increasing) This is because $\partial G(G^{-1}(w)) \partial G(G^{-1}(w))^\top = (\partial G^{-1}(w) \partial G^{-1}(w)^\top)^{-1} = (\nabla^2 R(w))^{-1}$.

Finally, we also want to remark that in our application of Nash's embedding theorem, the Riemannian metric is given by the Hessian of a mirror map, and it is not clear if this would endow a more explicit and tractable construction of the isometric embedding. We are not aware of such results to the best of our knowledge.

A.2 Existence of non-separable commuting parametrization

Despite the recent line of works on the connection between mirror descent and gradient descent [24, 4, 5, 8, 23], so far we have not seen any concrete example of non-separable parametrization (in the sense of Definition A.1) such that the reparametrized gradient flow can be written as a mirror flow. In this subsection, we discuss how we can use Theorem 5.1 to construct non-separable, yet commuting parametrizations.

Definition A.1 (Generalized separable parametrization)**.** Let M be an open subset of $\mathbb{R}^D$. We say a function $G : M \to \mathbb{R}^d$ is a *generalized separable parametrization* if and only if there exist d projection matrices $\{P_i\}_{i=1}^d$ satisfying $\sum_{i=1}^d P_i = I_d$, $P_i P_j = \mathbb{1}\{i = j\} \cdot P_i$, a function $\widehat{G} : M \to \mathbb{R}^d$ satisfying $\widehat{G}_i(x) = \widehat{G}_i(P_i x)$, a matrix $A \in \mathbb{R}^{d \times d}$ and a vector $b \in \mathbb{R}^d$, such that

$$G(x) = A\widehat{G}(x) + b, \qquad \forall x \in M.$$

Given the above definition, it is easy to check that $\widehat{G}$ is a commuting parametrization as $\nabla^2 \widehat{G}_i \nabla \widehat{G}_j = P_i \nabla^2 \widehat{G}_i P_i \cdot P_j \nabla \widehat{G}_j \equiv 0$ for all $i \neq j$, so each Lie bracket $[\nabla G_i, \nabla G_j]$ is also 0 by the linearity.

As a concrete example, for matrix sensing with commutable measurement $A_1, \dots, A_m \in \mathbb{R}^{d \times d}$ (see Example 4.4 and Remark 4.5), let $V = (v_1, \dots, v_d) \in \mathbb{R}^{d \times d}$ be a common eigenvector matrix for $\{A_i\}_{i=1}^m$ such that we can write $A_i = V \Sigma_i V^\top = \sum_{j=1}^d \sigma_{i,j} v_j v_j^\top$ for each $i \in [m]$. For parametrization $G : \mathbb{R}^{d \times r} \to d$ where each $G_i(U) = v_i^\top U U^\top v_i$, we can write $\langle A_i, UU^\top \rangle = \sum_{j=1}^d \sigma_{i,j} G_j(U)$.

However, the bad news is that separable commuting parametrizations can express only a restricted class of Legendre functions. It is easy to see $\partial \hat{G}(x) \partial \hat{G}(x)^\top$ must be diagonal for every x. Thus $\partial G(x) \partial G(x)^\top$ is simultaneously diagonalizable for all x, and so is the Hessian of the corresponding Legendre function (given by Lemma 4.7). Yet there are interesting Legendre functions whose

Hessians are not simultaneously diagonalizable, such as

$$R(w) = \sum_{i=1}^{d} w_i(\ln w_i - 1) + \left(1 - \sum_{i=1}^{d} w_i\right)\left(\ln\left(1 - \sum_{i=1}^{d} w_i\right) - 1\right),$$

where each $w_i > 0$ and $\sum_{i=1}^{d} w_i < 1$. We can check that $\nabla R(w) = \sum_{i=1}^{d} \ln \frac{w_i}{1-\sum_{i=1}^{d} w_i}$ and $\nabla^2 R(w) = \operatorname{diag}(w^{\odot(-1)}) + \mathbb{1}_d \mathbb{1}_d^\top$. Indeed, it is proposed as an open problem by [5] whether we can find a parametrization G such that the reparametrized gradient flow in the x-space simulates the mirror flow in the w-space with respect to the aforementioned Legendre function R.

Our Theorem 5.1 answers the open problem by [5] affirmatively since it shows every mirror flow can be written as some reparametrized gradient flow. According to the previous discussion, every mirror flow for Legendre function whose Hessian cannot be simultaneously diagonalized always induces a non-separable commuting parametrization. But this type of construction has two caveats: First, the construction of the Legendre function uses Nash's Embedding theorem, which is implicit and hard to implement; second, the parametrization given by Theorem 5.1, though defined on an open set in $\mathbb{R}^D$, is only commuting on the reachable set, which is a d-dimensional submanifold of $\mathbb{R}^D$. This is different from all the natural examples of commuting parametrizations that are commuting on an open set, leading to the following open question.

Open Question: Is there any smooth, regular, commuting, yet non-separable (in the sense of Definition A.1) parametrization from an open subset of $\mathbb{R}^D$ to $\mathbb{R}^d$, for some integers D and d?

Theorem A.2. *All smooth, regular and commuting parametrizations are non-separable when $D = 1$.*

Proof of Theorem A.2. Note that $[\nabla G_i, \nabla G_j] \equiv 0$ implies that all G_i share the same set of stationary points, *i.e.*, $\{x \in \mathbb{R} \mid \nabla G_i(x) = 0\}$ is the same for all $i \in [d]$. Since $D = 1$, without loss of generality, we can assume $G_i'(x) = \nabla G_i(x) > 0$ for all $x \in M$ and $i \in [d]$ since G is regular. Then it holds that $\operatorname{sign}(G_i')(\ln |G_i'|)' = \operatorname{sign}(G_j')(\ln |G_j'|)'$, which implies that $|G_i'|/|G_j'|$ is equal to some constant independent of x. This completes the proof. □

Remark A.3. *We note that the assumption that the parametrization is regular is necessary for the open question to be non-trivial. Otherwise, consider the following example with $D = 1$ and $d = 2$: Let $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ be any smooth function supported on $(0, 1)$ and $(1, 2)$ respectively. Define $G_i(x) = \int_0^x f_i(t)\mathrm{d}t$ for all $x \in \mathbb{R}$. Then parametrization G is non-separable.*

B Related basics for convex analysis

We first introduce some additional notations. For any function f, we denote its range (or image) by $\operatorname{range} f$. For any set S, we use $\overline{S}$ to denote its closure. For any matrix $\Lambda \in \mathbb{R}^{d\times D}$ and set $S \subseteq \mathbb{R}^D$, we define $\Lambda S = \{\Lambda x \mid x \in S\} \subseteq \mathbb{R}^d$.

Below we collect some related basic definitions and results in convex analysis. We refer the reader to [54] and [9] as main reference sources. In particular, Sections 2, 3 and 4 in [9] provide a clear summary of the related concepts.

Here we consider a convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ whose domain is $\operatorname{dom} f = \{w \in \mathbb{R}^d \mid f(w) < \infty\}$. **From now on, we assume by default** that f is continuous on $\operatorname{dom} f$, the interior of its domain $\operatorname{int}(\operatorname{dom} f)$ is non-empty, and f is differentiable on $\operatorname{int}(\operatorname{dom} f)$.

The notions of essential smoothness and essential strict convexity defined below describe certain nice properties of a convex function (see Section 26 in [54]).

Definition B.1 (Essential smoothness and essential strict convexity)**.** If for any sequence $\{w_n\}_{n=1}^{\infty} \subset \operatorname{int}(\operatorname{dom} f)$ going to the boundary of $\operatorname{dom} f$ as $n \to \infty$, it holds that $\|\nabla f(w_n)\| \to \infty$, then we say f is *essentially smooth*. If f is strictly convex on every convex subset of $\operatorname{int}(\operatorname{dom} f)$, then we say f is *essentially strictly convex*.

The concept of *convex conjugate* is critical in our derivation. Specifically, given a convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, its convex conjugate f^* is defined as

$$f^*(w) = \sup_{y\in\mathbb{R}^d} \langle w, y\rangle - f(y).$$

The following results characterize the relationship between a convex function and its conjugate.

Theorem B.2 (Theorem 26.3, [54]). *A convex function f is essentially strictly convex if and only if its convex conjugate f^* is essentially smooth.*

Proposition B.3 (Proposition 2.5, [9]). If f is essentially strictly convex, then $\operatorname{range} \partial f = \operatorname{int}(\operatorname{dom} f^*) = \operatorname{dom} \nabla f^*$, where ∂f is the subgradient of f.

Lemma B.4 (Corollary 2.6, [9]). *If f is essentially strictly convex, then it holds for all $w \in \operatorname{int}(\operatorname{dom} f)$ that $\nabla f(w) \in \operatorname{int}(\operatorname{dom} f^*)$ and $\nabla f^*(\nabla f(w)) = w$.*

The class of Legendre functions defined in Definition 3.7 contains convex functions that are both essentially smooth and essentially strictly convex.

Theorem B.5 (Theorem 26.5, [54]). *A convex function f is a Legendre function if and only if its conjugate f^* is. In this case, the gradient mapping $\nabla f : \operatorname{int}(\operatorname{dom} f) \to \operatorname{int}(\operatorname{dom} f^*)$ satisfies $(\nabla f)^{-1} = \nabla f^*$.*

Next, we introduce the notion of Bregman function [12, 14]. It has been shown in [9] that the properties of Bregman functions are crucial to prove the trajectory convergence of Riemannian gradient flow where the metric tensor is given by the Hessian of some Bregman function f.

Definition B.6 (Bregman functions; Definition 4.1, [3]). A function f is called a *Bregman function* if it satisfies the following properties:

(a) $\operatorname{dom} f$ is closed. f is strictly convex and continuous on $\operatorname{dom} f$. f is $\mathcal{C}^1$ on $\operatorname{int}(\operatorname{dom} f)$.

(b) For any $w \in \operatorname{dom} f$ and $\alpha \in \mathbb{R}$, $\{y \in \operatorname{dom} f \mid D_R(w, y) \le \alpha\}$ is bounded.

(c) For any $w \in \operatorname{dom} f$ and sequence $\{w_i\}_{i=1}^{\infty} \subset \operatorname{int}(\operatorname{dom} f)$ such that $\lim_{i\to\infty} w_i = w$, it holds that $\lim_{i\to\infty} D_R(w, w_i) \to 0$.

The following theorem provides a special sufficient condition for f to be a Bregman function.

Theorem B.7 (Theorem 4.7, [3]). *If f is a Legendre function with $\operatorname{dom} f = \mathbb{R}^d$, then $\operatorname{dom} f^* = \mathbb{R}^d$ implies that f is a Bregman function.*

The following theorem from [3] provides a convenient tool for proving the convergence of a Riemannian gradient flow.

Theorem B.8 (Theorem 4.2, [3]). *Suppose $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a Bregman function and also a Legendre function, and satisfies that f is twice continuously differentiable on $\operatorname{int}(\operatorname{dom} f)$ and $\nabla^2 f$ is locally Lipschitz. Consider the following Riemannian gradient flow:*

$$\mathrm{d}w(t) = -\nabla^2 f(w(t))^{-1} \nabla L(w(t)) \mathrm{d}t, \qquad w(0) = w_{\text{init}} \in \operatorname{int}(\operatorname{dom} f)$$

where the loss $L : \mathbb{R}^d \to \mathbb{R}$ satisfies that L is quasi-convex, ∇L is locally Lipschitz, and $\operatorname{argmin}\{L(w) \mid w \in \operatorname{dom} f\}$ is non-empty. Then as $t \to \infty$, $w(t)$ converges to some $w^* \in \operatorname{dom} f$ such that $\langle \nabla L(w^*), w - w^* \rangle \ge 0$ for all $w \in \operatorname{dom} f$. If the loss L is further convex, then w^* is a minimizer of L on $\operatorname{dom} f$.

C Omitted proofs in Section 3

Here we first present the result and its proof for the domain of the flow induced by G.

Lemma C.1. *Let M be a smooth submanifold of $\mathbb{R}^D$ and $G : M \to \mathbb{R}^d$ be a $\mathcal{C}^2$ parametrization satisfying Assumption 3.5. Then for any $x \in M$, $\mathcal{U}(x)$ is a hyperrectangle, i.e., $\mathcal{U}(x)$ can be decomposed as*

$$\mathcal{U}(x) = \mathcal{I}_1(x) \times \mathcal{I}_2(x) \times \cdots \times \mathcal{I}_d(x)$$

where $\mathcal{I}_j(x) := \{x'_j \mid x' \in \mathcal{U}(x)\}$ is an open interval.

Proof of Lemma C.1. Fix any $x \in M$. For each $i \in [d]$, let $\mathcal{I}_i(x)$ be the domain of $\phi^t_{G_j}(x)$ in terms of t. If ∇G_i is a complete vector field on M as in Definition 3.2, then $\mathcal{I}_i(x) = \mathbb{R}^d$, otherwise $\phi^t_{G_j}(x)$ is defined for t in an open interval containing 0 (see, e.g., Theorem 2.1 in [37]). Then we claim

that for any distinct $j_1, j_2, \ldots, j_k \in [d]$ where $k \in [d]$, the set of all $(\mu_{j_1}, \ldots, \mu_{j_k}) \in \mathbb{R}^k$ such that $\phi^{\mu_{j_1}}_{G_{j_1}} \circ \cdots \circ \phi^{\mu_{j_k}}_{G_{j_k}}(x)$ is well-defined is a hyperrectangle given by $\mathcal{I}_{j_1}(x) \times \mathcal{I}_{j_2}(x) \times \cdots \times \mathcal{I}_{j_k}(x)$. Then the desired result can be obtained by letting $(j_1, j_2, \ldots, j_d) = (1, 2, \ldots, d)$. We prove the claim by induction over $k \in [d]$.

The base case for $k = 1$ has already been established above. Next, assume the claim holds for $1, 2, \ldots, k-1$ where $k \geq 3$, and we proceed to show it for k. By the claim for $k-2$, $\phi^{\mu_{j_3}}_{G_{j_3}} \circ \cdots \circ \phi^{\mu_{j_k}}_{G_{j_k}}(x)$ is well-defined for $(\mu_{j_3}, \ldots, \mu_{j_k}) \in \mathcal{I}_{j_3}(x) \times \cdots \times \mathcal{I}_{j_k}(x)$. For any such $(\mu_{j_3}, \ldots, \mu_{j_k})$, $\phi^{t}_{G_{j_1}} \circ \phi^{\mu_3}_{G_{j_3}} \circ \cdots \circ \phi^{\mu_{j_k}}_{G_{j_k}}(x)$ is well-defined for t in and only in the open interval $\mathcal{I}_{j_1}(x)$ by applying the claim for $k-1$, and similarly $\phi^{t}_{G_{j_2}} \circ \phi^{\mu_3}_{G_{j_3}} \circ \cdots \circ \phi^{\mu_{j_k}}_{G_{j_k}}(x)$ is also well-defined for t in and only in the open interval $\mathcal{I}_{j_2}(x)$. Note that for any $(s, t) \in \mathcal{I}_{j_1}(x) \times \mathcal{I}_{j_2}(x)$,

$$\phi^{s}_{G_{j_1}} \circ \phi^{-t}_{G_{j_2}} \circ \phi^{t}_{G_{j_2}} \circ \phi^{\mu_{j_3}}_{G_{j_3}} \circ \cdots \circ \phi^{\mu_{j_k}}_{G_{j_k}}(x)$$

is well-defined, so by Assumption 3.5, we see that

$$\phi^{-t}_{G_{j_2}} \circ \phi^{s}_{G_{j_1}} \circ \phi^{t}_{G_{j_2}} \circ \phi^{\mu_{j_3}}_{G_{j_3}} \circ \cdots \circ \phi^{\mu_{j_k}}_{G_{j_k}}(x)$$

is also well-defined, which further implies that $\phi^{s}_{G_{j_1}} \circ \phi^{t}_{G_{j_2}} \circ \phi^{\mu_{j_3}}_{G_{j_3}} \circ \cdots \circ \phi^{\mu_{j_k}}_{G_{j_k}}(x)$ is well-defined. Therefore, we conclude that $\phi^{\mu_{j_1}}_{G_{j_1}} \circ \cdots \circ \phi^{\mu_{j_k}}_{G_{j_k}}(x)$ is well-defined for and only for $(\mu_{j_1}, \ldots, \mu_{j_k}) \in \mathcal{I}_{j_1}(x) \times \cdots \times \mathcal{I}_{j_k}(x)$. This completes the induction and hence finishes the proof. □

Next, we provide the proof for the implicit bias of mirror flow summarized in Theorem 3.8. We need the following lemma that characterizes the KKT conditions for minimizing a convex function R in a linear subspace.

Lemma C.2. *For any convex function $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and $Z \in \mathbb{R}^{n \times d}$, suppose $\nabla R(w^*) = Z^\top \lambda$ for some $\lambda \in \mathbb{R}^n$, then*

$$R(w^*) = \min_{w : Z(w - w^*) = 0} R(w).$$

Proof of Lemma C.2. Consider another convex function defined as $\widetilde{R}(w) = R(w) - w^\top Z^\top \lambda$, then $\nabla \widetilde{R}(w^*) = \nabla R(w^*) - Z^\top \lambda = 0$, which implies that

$$\begin{aligned}
\tilde{R}(w^*) &= \min_{w \in \mathbb{R}^d} R(w) - w^\top Z^\top \lambda \\
&\leq \min_{w : Z(w - w^*) = 0} R(w) - w^\top Z^\top \lambda \\
&= \min_{w : Z(w - w^*) = 0} R(w) - w^{*\top} Z^\top \lambda.
\end{aligned}$$

Since $\widetilde{R}(w^*) = R(w^*) - w^{*\top} Z^\top \lambda$, it follows that

$$R(w^*) \leq \min_{w : Z(w - w^*) = 0} R(w),$$

and the equality is achieved at $w = w^*$. This finishes the proof. □

We can then prove Theorem 3.8 by using Lemma C.2.

Proof of Theorem 3.8. Since $L(w) = \widetilde{L}(Zw - Y)$, the mirror flow (8) can be further written as

$$\mathrm{d} \nabla R(w(t)) = -Z^\top \nabla \widetilde{L}(Zw(t) - Y) \mathrm{d}t.$$

Integrating the above yields that for any $t \geq 0$,

$$\nabla R(w(t)) - \nabla R(w_0) = -Z^\top \int_0^t \nabla \widetilde{L}(Zw(s) - Y) \mathrm{d}s \in \mathrm{span}(X^\top),$$

which further implies that $\nabla R(w_\infty) - \nabla R(w_0) \in \mathrm{span}(Z^\top)$. Therefore,

$$\nabla D_R(w, w_0)|_{w = w_\infty} = \nabla R(w_\infty) - \nabla R(w_0) \in \mathrm{span}(Z^\top).$$

Then applying Lemma C.2 yields

$$D_R(w_\infty, w_0) = \min_{w : Z(w - w_\infty) = 0} D_R(w, w_0).$$

This finishes the proof. □

D Omitted proofs in Section 4

Here we provide the omitted proofs in Section 4, including four main parts:

(1) Properties of commuting parametrizations (Appendix D.1);
(2) Necessary condition for a smooth parametrization to be commuting (Appendix D.2);
(3) Convergence for gradient flow with commuting parametrization (Appendix D.3);
(4) Results for the underdetermined linear regression (Appendix D.4).

D.1 Properties of commuting parametrizations

We first recall the following result on the characterization of commuting vector fields from [38].

Theorem D.1 (Adapted from Theorem 9.44 in [38]). *Let M be a smooth submanifold of $\mathbb{R}^D$ and $G : M \to \mathbb{R}^d$ be a $\mathcal{C}^2$ parametrization. For any $i, j \in [d]$, $[\nabla G_i, \nabla G_j](x) = 0$ for all $x \in M$ if and only if for any $x \in M$, whenever both $\phi^s_{G_i} \circ \phi^t_{G_j}(x)$ and $\phi^t_{G_j} \circ \phi^s_{G_i}(x)$ are well-defined for all (s,t) in some rectangle $\mathcal{I}_1 \times \mathcal{I}_2$ where $\mathcal{I}_1, \mathcal{I}_2 \subseteq \mathbb{R}$ are open intervals, it holds that $\phi^s_{G_i} \circ \phi^t_{G_j}(x) = \phi^t_{G_j} \circ \phi^s_{G_i}(x)$ for all $(s,t) \in \mathcal{I}_1 \times \mathcal{I}_2$.*

Proof of Theorem 4.2. Note that under Assumption 3.5, Lemma C.1 implies that the domain of $\phi^s_{G_i} \circ \phi^t_{G_j}(x)$ is exactly $\mathcal{I}_i(x) \times \mathcal{I}_j(x)$, and the statement of Theorem 4.2 immediately follows. □

Next, we prove the representation formula for gradient flow with commuting parametrization given in Lemma 4.6.

Proof of Lemma 4.6. Let $\mu(t)$ be given by the following differential equation:

$$\mathrm{d}\mu(t) = -\nabla L_t(G(\psi(x_{\text{init}}; \mu(t))))\mathrm{d}t, \qquad \mu(0) = 0.$$

For any $\mu \in \mathcal{U}(x)$ and $j \in [d]$, $\mu + \delta e_j \in \mathcal{U}(x)$ for all sufficiently small δ, thus

$$\begin{aligned}\frac{\partial}{\partial \mu_j}\psi(x_{\text{init}}; \mu) &= \lim_{\delta \to 0} \frac{\psi(x_{\text{init}}; \mu + \delta e_j) - \psi(x_{\text{init}}; \mu)}{\delta} \\ &= \lim_{\delta \to 0} \frac{\phi^\delta_{G_j}(\psi(x_{\text{init}}; \mu)) - \psi(x_{\text{init}}; \mu)}{\delta} \\ &= \nabla G_j(\psi(x_{\text{init}}; \mu))\end{aligned}$$

where the second equality follows from the assumption that G is a commuting parametrization and Theorem 4.2. Then we have $\frac{\partial \psi(x_{\text{init}}; \mu)}{\partial \mu} = \partial G(\psi(x_{\text{init}}; \mu))^\top$ for all $\mu \in \mathcal{U}(x_{\text{init}})$, and thus when $\mu(t) \in \mathcal{U}(x_{\text{init}})$,

$$\begin{aligned}\mathrm{d}\psi(x_{\text{init}}; \mu(t)) &= \frac{\partial \psi(x_{\text{init}}; \mu(t))}{\partial \mu(t)}\mathrm{d}\mu(t) \\ &= -\partial G(x_{\text{init}}; \mu(t))\nabla L_t(G(\psi(x_{\text{init}}; \mu(t))))\mathrm{d}t \\ &= -\nabla(L_t \circ G)(\psi(x_{\text{init}}; \mu(t)))\mathrm{d}t.\end{aligned}$$

Then since $\psi(x_{\text{init}}; \mu(0)) = x_{\text{init}}$ and $\psi(x_{\text{init}}; \mu(t))$ follows the same differential equation and has the same initialization as $x(t)$, we have $x(t) \equiv \psi(x_{\text{init}}; \mu(t))$ for all $t \in [0, T)$. Therefore,

$$\mu(t) = \mu(0) + \int_0^t -\nabla L_t(G(\psi(x_{\text{init}}; \mu(s))))\mathrm{d}s = \int_0^t -\nabla L_t(G(x(s)))\mathrm{d}s$$

for all $t \in [0, T)$, which completes the proof. □

Next, to prove Lemma 4.7, we need the following lemma which provides a sufficient condition for a vector function to be gradient of some other function.

Lemma D.2. *Let $\Psi : C \to \mathbb{R}^d$ be a differentiable function where C is a simply connected open subset of $\mathbb{R}^d$. If for all $w \in C$ and any $i, j \in [d]$, $\frac{\partial}{\partial w_j}\Psi_i(w) = \frac{\partial}{\partial w_i}\Psi_j(w)$, then there exists some function $Q : C \to \mathbb{R}$ such that $\Psi = \nabla Q$.*

Proof of Lemma D.2. This follows from a direct application of Corollary 16.27 in [38]. □

Based on the above results, we proceed to prove Lemma 4.7.

Proof of Lemma 4.7. By Lemma C.1, $\mathcal{U}(x_{\text{init}})$ is hyperrectangle, and hence is convex. Next, recall that by the proof of Lemma 4.6, we have $\frac{\partial \psi(x_{\text{init}};\mu)}{\partial \mu} = \partial G(\psi(x_{\text{init}};\mu))^\top$ for all $\mu \in \mathcal{U}(x_{\text{init}})$. Denoting $\Psi(\mu) = G(\psi(x_{\text{init}};\mu))$, we further have

$$\partial \Psi(\mu) = \frac{\partial G(\psi(x_{\text{init}};\mu))}{\partial \psi(x_{\text{init}};\mu)} \frac{\partial \psi(x_{\text{init}};\mu)}{\partial \mu} = \partial G(\psi(x_{\text{init}};\mu)) \partial G(\psi(x_{\text{init}};\mu))^\top, \quad \forall \mu \in \mathcal{U}(x).$$

Since G is regular, $\partial G(\psi(x_{\text{init}};\mu))$ is of full-rank for all $\mu \in \mathcal{U}(x_{\text{init}})$, so $\partial \Psi$ is symmetric and positive definite for all $\mu \in \mathcal{U}(x_{\text{init}})$, which implies that Ψ is the gradient of some strictly convex function $Q : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ by Lemma D.2. This Q satisfies that $\nabla Q(\mu) = \Psi(\mu) = G(\psi(x_{\text{init}};\mu))$ for all $\mu \in \mathcal{U}(x_{\text{init}})$. Therefore, Q is a strictly convex function with $\operatorname{dom} \nabla Q = \mathcal{U}(x_{\text{init}})$ and $\operatorname{range} \nabla Q = \Omega_w(x_{\text{init}}; G)$.

Next, we show that Q is essentially smooth. If $\mathcal{U}(x_{\text{init}}) = \mathbb{R}^d$, then $\operatorname{dom} Q = \mathbb{R}^d$ and the boundary of $\operatorname{dom} Q$ is empty, so it is trivial that Q is essentially smooth. Otherwise, it suffices to show that for any μ on the boundary of $\operatorname{dom} Q$ and any sequence $\{\mu_k\}_{k=1}^\infty \subset \mathcal{U}(x_{\text{init}})$ such that $\lim_{k\to\infty} \mu_k = \mu_\infty$, we have $\lim_{k\to\infty} \|\nabla Q(\mu_k)\|_2 = \infty$. Since each $\nabla Q(\mu_k) = G(\psi(x_{\text{init}};\mu_k))$, we only need to show that $\lim_{k\to\infty} \|G(\psi(x_{\text{init}};\mu_k))\|_2 = \infty$. Suppose otherwise, then $\{G(\psi(x_{\text{init}};\mu_k)\}_{k=1}^\infty$ is bounded. Note that by Lemma 4.6, let $H_k(x) = \langle \mu_k, G(x)\rangle$, and we have

$$\psi(x_{\text{init}};\mu_k) = \phi^1_{-H_k}(x_{\text{init}}) = x_{\text{init}} + \int_0^1 \nabla H_k(\phi^s_{-H_k}(x_{\text{init}}))\mathrm{d}s.$$

Therefore,

$$\|\psi(x_{\text{init}};\mu_k) - x_{\text{init}}\|_2 \le \int_0^1 \left\|\nabla H_k(\phi^s_{-H_k}(x_{\text{init}}))\right\|_2 \mathrm{d}s \le \sqrt{\int_0^1 \left\|\nabla H_k(\phi^s_{-H_k}(x_{\text{init}}))\right\|_2^2 \mathrm{d}s}. \tag{12}$$

where the second inequality follows from Cauchy-Schwarz inequality. Further note that

$$\begin{aligned} H_k(\psi(x_{\text{init}};\mu_k)) - H_k(x_{\text{init}}) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s} H_k(\phi^s_{-H_k}(x_{\text{init}}))\mathrm{d}s \\ &= \int_0^1 \left\langle \nabla H_k(\phi^s_{-H_k}(x_{\text{init}})), \frac{\mathrm{d}\phi^s_{-H_k}(x_{\text{init}})}{\mathrm{d}s} \right\rangle \mathrm{d}s \\ &= \int_0^1 \|\nabla H_k(\phi^s_{-H_k}(x_{\text{init}}))\|_2^2 \mathrm{d}s. \end{aligned} \tag{13}$$

Then combining (12) and (13), we get

$$\begin{aligned} \|\psi(x_{\text{init}};\mu_k) - x_{\text{init}}\|_2 &\le \sqrt{\langle \mu_k, G(\psi(x_{\text{init}};\mu_k)) - G(x_{\text{init}})\rangle} \\ &\le \sqrt{\|\mu_k\|_2 \cdot \|G(\psi(x_{\text{init}};\mu_k)) - G(x_{\text{init}})\|_2}, \end{aligned}$$

which implies that $\{\psi(x_{\text{init}};\mu_k)\}_{k=1}^\infty$ is bounded. Then there exists a convergent subsequence of $\{\psi(x_{\text{init}};\mu_k)\}_{k=1}^\infty$, and without loss of generality we assume that $\psi(x_{\text{init}};\mu_k)$ itself converges to some $x_\infty \in M$ as $k \to \infty$. Note that $\psi(x_\infty;\mu)$ is well-defined for μ in a small open neighborhood of 0, and since $\lim_{k\to\infty} \psi(x_{\text{init}};\mu_k) = x_\infty$, for sufficiently large k, $\psi(\psi(x_{\text{init}};\mu_k);\mu)$ is well-defined for μ in a small neighborhood of 0 that does not depend on k. Thus there exists some $\mu \in \mathbb{R}^d$ such that $\mu_k + \mu \notin \mathcal{U}(x_{\text{init}})$ but $\psi(\psi(x_{\text{init}};\mu_k);\mu)$ is well-defined for sufficiently large k. But by Lemma C.1 and Theorem D.1, $\psi(\psi(x_{\text{init}};\mu_k);\mu) = \psi(x_{\text{init}};\mu_k + \mu)$ and thus $\mu_k + \mu \in \mathcal{U}(x_{\text{init}})$, which leads to a contradiction. Hence, we conclude that Q is essentially smooth.

Combining the above, it follows that Q is a Legendre function. Let $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be the convex conjugate of Q. Then by Theorem B.5, R is also a Legendre function. Note that for any $\mu \in \mathcal{U}(x_{\text{init}})$, by the result in [17], we have

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

Therefore, R and Q are both Legendre functions, and by Proposition B.3, we further have $\operatorname{range}\nabla R = \operatorname{int}(\operatorname{dom} Q) = \operatorname{dom}\nabla Q = \mathcal{U}(x)$ and conversely $\operatorname{dom}\nabla R = \operatorname{range}\nabla Q = \Omega_w(x_{\text{init}}; G)$. This finishes the proof. □

Then using Lemma 4.6 and Lemma 4.7, we can prove Theorem 4.8.

Proof of Theorem 4.8. Recall that the gradient flow in the x-space governed by $-\nabla(L_t \circ G)(x)$ is

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

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

$$\mathrm{d}w(t) = \partial G(x(t))\mathrm{d}x(t) = -\partial G(x(t))\partial G(x(t))^\top \nabla L_t(w(t))\mathrm{d}t. \tag{14}$$

By Lemma 4.6, 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 4.7, we get a Legendre function $R : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ with domain $\Omega_w(x_{\text{init}}; G)$ such that

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

for all $t \geq 0$. Then the dynamics of $w(t)$ in (14) can be rewritten as

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

or equivalently,

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

which is exactly the mirror flow with respect to R initialized at $w(0) = G(x_{\text{init}})$. Further note that the result of Lemma 4.7 is completely independent of the loss function L_t, and thus R only depends on the initialization x_{init} and the parametrization G. This finishes the proof. □

D.2 Necessary condition for a smooth parametrization to be commuting

Proof of Theorem 4.9. Fix any initialization $x_{\text{init}} \in M$, and let the Legendre function R be given such that for all time-dependent loss L_t, the gradient flow under $L_t \circ G$ initialized at x can be written as the mirror flow under L_t with respect to the Legendre function R. We first introduce a few notations that will be useful for the proof. For any $s \in \mathbb{R}$, we define a time-shifting operator $\mathcal{T}_s$ such that for any time-dependent loss $L_t(\cdot)$, $(\mathcal{T}_s L)_t(\cdot) = L_{t-s}(\cdot)$. We say a time-dependent loss L_t is supported on finite time if $L_t = \sum_{i=1}^k \mathbb{1}_{t\in[t_i,t_{i+1})} L^{(i)}$ for some $k \geq 1$ where $t_1 = 0$, $t_{k+1} = \infty$ and $L^{(k)} \equiv 0$, and we denote $\operatorname{len}(L) = t_k$. We further define the concatenation of two time-dependent loss L_t, L'_t supported on finite time as $L \parallel L' = L + \mathcal{T}_{\operatorname{len}(L)} L'$. We also use $\overline{L}$ to denote the time-reverse of the time-dependent loss L which is supported on finite time, that is, $\overline{L}_t = L_{\operatorname{len}(L)-t}$ for all $t \geq 0$. For any $j \in [d]$ and $\delta > 0$, we define the following loss function

$$\ell_t^{j,\delta}(w) = \mathbb{1}_{0\leq t\leq\delta} \cdot \langle e_j, w\rangle \tag{15}$$

where e_j is the j-th canonical base of $\mathbb{R}^d$.

Now for any $k \geq 2$, let $\{j_i\}_{i=1}^k$ be any sequence where each $j_i \in [d]$. Then we recursively define a sequence of time-dependent losses as follows: First define $L^{1,\delta} = -\ell^{j_1,\delta}$, then sequentially for each $i = 2, 3, \ldots, k$, we define

$$L^{i,\delta} = L^{i-1,\sqrt{\delta}} \parallel \left(-\ell^{j_i,\sqrt{\delta}}\right) \parallel \left(-\overline{L}^{i-1,\sqrt{\delta}}\right) \parallel \ell^{j_i,\sqrt{\delta}} \tag{16}$$

where we write $\overline{L}^{i-1,\sqrt{\delta}} = \overline{L^{i-1,\sqrt{\delta}}}$ for convenience. Denote $\iota_i(\delta) = \operatorname{len}(L^{i,\delta})$ for each $i \in [k]$. Then $\iota_1(\delta) = \delta$ and $\iota_i(\delta) = 2\sqrt{\delta} + 2\iota_{i-1}(\sqrt{\delta})$ for $i = 2, 3, \ldots, k$, which further implies

$$\iota_i(\delta) = \sum_{m=1}^{i-1} 2^m \delta^{1/2^m} + 2^{i-1}\delta^{1/2^{i-1}} \text{ for all } i \in [k].$$

Moreover, for each $i = 2, 3, \ldots, k$, the gradient of $L^{i,\delta}$ with respect to w is given by

$$\nabla L_t^{i,\delta}(w) = \begin{cases} \nabla L_t^{i-1,\sqrt{\delta}}(w) & 0 \le t \le \iota_{i-1}(\sqrt{\delta}), \\ -e_{j_i} & \iota_{i-1}(\sqrt{\delta}) < t \le \iota_{i-1}(\sqrt{\delta}) + \sqrt{\delta}, \\ -\nabla \overline{L}_t^{i-1,\sqrt{\delta}}(w) & \iota_{i-1}(\sqrt{\delta}) + \sqrt{\delta} < t \le 2\iota_{i-1}(\sqrt{\delta}) + \sqrt{\delta}, \\ e_{j_i} & 2\iota_{i-1}(\sqrt{\delta}) + \sqrt{\delta} < t \le 2\iota_{i-1}(\sqrt{\delta}) + 2\sqrt{\delta}, \\ 0 & t > 2\iota_{i-1}(\sqrt{\delta}) + 2\sqrt{\delta}. \end{cases} \tag{17}$$

This inductively implies that for any $t \in [0, \iota_k(\delta)]$, $\nabla L_t^{k,\delta}(w) \in \{e_j\}_{j=1}^d$ does not depend on w and is only determined by t. Therefore, for any initialization $x \in M$, for all sufficiently small $\delta > 0$, the gradient flow under $L^{k,\delta}$ for $\iota_k(\delta)$ time, i.e., $\phi_{L^{k,\delta}}^{\iota_k(\delta)}(x)$, is well-defined. Moreover, it follows from (17) that

$$\begin{aligned} \int_0^{\iota_{k-1}(\delta)} \nabla L_t^{k,\delta}(w(t))\mathrm{d}t &= \int_0^{\iota_{k-1}(\sqrt{\delta})} \nabla L^{k-1,\sqrt{\delta}}(w(t))\mathrm{d}t + \int_{\iota_{k-1}(\sqrt{\delta})}^{\iota_{k-1}(\sqrt{\delta})+\sqrt{\delta}} -e_{j_k}\mathrm{d}t \\ &\quad + \int_{\iota_{k-1}(\sqrt{\delta})+\sqrt{\delta}}^{2\iota_{k-1}(\sqrt{\delta})+\sqrt{\delta}} -\nabla \overline{L}^{k-1,\sqrt{\delta}}(w(t))\mathrm{d}t + \int_{2\iota_{k-1}(\sqrt{\delta})+\sqrt{\delta}}^{2\iota_{k-1}(\sqrt{\delta})2\sqrt{\delta}} e_{j_k}\mathrm{d}t \\ &= \int_0^{\iota_{k-1}(\sqrt{\delta})} \left(\nabla L_t^{k-1,\sqrt{\delta}}(w(t)) - \nabla \overline{L}_t^{k-1,\sqrt{\delta}}(w(t)) \right) \mathrm{d}t = 0 \end{aligned}$$

where the last two equalities follow from the fact that $\nabla L_t^{k-1,\sqrt{\delta}}(w)$ does not depend on w and is only determined by t by our construction.

Hence, the mirror flow with respect to the Legendre function R for the time-dependent loss $L^{k,\delta}$ will return to the initialization after $\iota_k(\delta)$ time since

$$\nabla R(w(\iota_k(\delta))) - \nabla R(w(0)) = \int_0^{\iota_k(\delta)} -\nabla L^{k,\delta}(w(t))\mathrm{d}t = 0.$$

This further implies that

$$G(x_{\text{init}}) = G\big(\phi_{L^{k,\delta} \circ G}^{\iota_k(\delta)}(x_{\text{init}})\big)$$

for all sufficiently small δ. Then differentiating with δ on both sides yields

$$\partial G(x) \cdot \frac{\mathrm{d}\phi_{L^{k,\delta} \circ G}^{\iota_k(\delta)}(x_{\text{init}})}{\mathrm{d}\delta}\bigg|_{\delta=0} = 0. \tag{18}$$

Note that if the following holds:

$$\frac{\mathrm{d}\phi_{L^{k,\delta} \circ G}^{\iota_k(\delta)}(x_{\text{init}})}{\mathrm{d}\delta}\bigg|_{\delta=0} = [[[[\nabla G_{j_1}, \nabla G_{j_2}], \ldots], \nabla G_{j_{k-1}}], \nabla G_{j_k}](x_{\text{init}}), \tag{19}$$

then combining (18) and (19) completes the proof, so it remains to verify (19).

We will prove by induction over k, and now let $\{j_i\}_{i=1}^\infty$ be an arbitrary sequence where each $j_i \in [d]$. For notational convenience, we denote for each $k \ge 1$,

$$\pi_{k,\delta}(\cdot) := \phi_{-\ell^{j_k},\delta}^{\delta}(\cdot) \quad \text{and} \quad \Pi_{k,\delta}(\cdot) := \phi_{L^{k,\delta}}^{\iota_k(\delta)}(\cdot).$$

Then their inverse maps are given by $\pi_{k,\delta}^{-1}(\cdot) = \phi_{\ell^{j_k},\delta}^{\delta}(\cdot)$ and $\Pi_{k,\delta}^{-1}(\cdot) = \phi_{-\overline{L}^{k,\delta}}^{\iota_k(\delta)}(\cdot)$ respectively. Since G is smooth, each $\Pi_{k,\sqrt{\delta}}$ is a $\mathcal{C}^\infty$ function of $\delta^{1/2^k}$, and we can expand it in $\delta^{1/2^k}$ as

$$\Pi_{k,\sqrt{\delta}}(x) = x + \sum_{i=1}^{2^k} \frac{\delta^{i/2^k}}{i!} \Delta_{k,i}(x) + r_{k,\delta}(x) \tag{20}$$

where the remainder term $r_{k,\delta}(x)$ is continuous in x and for each $x \in M$, $r_{k,\delta}(x) = o(\delta)$ (i.e., $\lim_{\delta\to 0} \frac{r_{k,\delta}(x)}{\delta} = 0$), and each $\Delta_{k,i}$ is defined as

$$\Delta_{k,i}(x) = \frac{\mathrm{d}^i \Pi_{k,\sqrt{\delta}}(x)}{\mathrm{d}(\delta^{1/2^k})^i}\bigg|_{\delta=0}.$$

In particular, for $k = 1$, we have

$$\Pi_{1,\sqrt{\delta}}(x) = \pi_{1,\sqrt{\delta}}(x) = x + \sqrt{\delta}\nabla G_{j_1}(x) + \frac{\delta}{2}\partial(\nabla G_{j_1})(x)\nabla G_{j_1}(x) + r_{1,\delta}(x) \tag{21}$$

where the second equality holds as well for any other G_j in place of G_{j_1}, with a different but similar remainder term. For any fixed $K \geq 2$, there is a small open neighborhood of x_{init} on M, denoted by $\mathcal{N}_{x_{\text{init}}} \subseteq M$, such that for all $k \in [K]$, we have $r_{k,\delta}(x) = o(\delta)$ uniformly over all $x \in \mathcal{N}_{x_{\text{init}}}$, so we can replace all $r_{k,\delta}(x)$ by $o(\delta)$ when $x \in \mathcal{N}_{x_{\text{init}}}$. Then we claim that for each $k = 2, 3, \ldots, K$,

$$\lim_{\delta\to\infty} \frac{1}{\sqrt{\delta}} \sum_{i=1}^{2^{k-1}} \frac{\delta^{i/2^k}}{i!}\Delta_{k,i}(x) = [[[\nabla G_{j_1}, \nabla G_{j_2}], \ldots], \nabla G_{j_k}](x), \quad \forall x \in \mathcal{N}_{x_{\text{init}}}, \tag{22}$$

which directly implies (19). With a slight abuse of notation, the claim is also true for $k = 1$ since $\Delta_{1,1}(x) = \nabla G_{j_1}(x)$ by (21), so we use this as the base case of the induction. Then, assuming (22) holds for $k - 1 < K$, we proceed to prove it for k. For convenience, further define $\mathrm{Lie}_G(j_{1:k}) = [[[\nabla G_{j_1}, \nabla G_{j_2}], \ldots], \nabla G_{j_k}]$.

Combining the Taylor expansion in (20) and (22) for $k - 1$, we obtain for all $x \in \mathcal{N}_{x_{\text{init}}}$ that

$$\Pi_{k-1,\sqrt{\delta}}(x) = x + \sqrt{\delta} \cdot \mathrm{Lie}_G(j_{1:(k-1)})(x) + \sum_{i=2^{k-2}+1}^{2^{k-1}} \frac{\delta^{i/2^{k-1}}}{i!}\Delta_{k-1,i}(x) + o(\delta)$$

for sufficiently small δ. Further apply (21) with G_{j_k} in place of G_{j_1} for sufficiently small δ, and then

$$\begin{aligned}
&\Pi_{k-1,\sqrt{\delta}}\big(\pi_{k,\sqrt{\delta}}(x)\big) \\
&\quad= \Pi_{k-1,\sqrt{\delta}}\bigg(x + \sqrt{\delta}\nabla G_{j_k}(x) + \frac{\delta}{2}\partial(\nabla G_{j_k})(x)\nabla G_{j_k}(x) + o(\delta)\bigg) \\
&\quad= x + \sqrt{\delta}\nabla G_{j_k}(x) + \frac{\delta}{2}\partial(\nabla G_{j_k})(x)\nabla G_{j_k}(x) + o(\delta) \\
&\qquad+ \sqrt{\delta} \cdot \mathrm{Lie}_G(j_{1:(k-1)})\bigg(x + \sqrt{\delta}\nabla G_{j_k}(x) + \frac{\delta}{2}\partial(\nabla G_{j_k})(x)\nabla G_{j_k}(x) + o(\delta)\bigg) \\
&\qquad+ \sum_{i=2^{k-2}+1}^{2^{k-1}} \frac{\delta^{i/2^{k-1}}}{i!}\Delta_{k-1,i}\bigg(x + \sqrt{\delta}\nabla G_{j_k}(x) + \frac{\delta}{2}\partial(\nabla G_{j_k})(x)\nabla G_{j_k}(x) + o(\delta)\bigg) \\
&\qquad+ r_{k-1,\delta}\bigg(x + \sqrt{\delta}\nabla G_{j_k}(x) + \frac{\delta}{2}\partial(\nabla G_{j_k})(x)\nabla G_{j_k}(x) + o(\delta)\bigg)
\end{aligned}$$

where the second equality follows from the Taylor expansion of $\Pi_{k-1,\sqrt{\delta}}$ and that $\pi_{k,\sqrt{\delta}}(x) \in \mathcal{N}_{x_{\text{init}}}$ for sufficiently small δ. Then by the Taylor expansion of $\mathrm{Lie}_G(j_{1:(k-1)})$ and each $\Delta_{k-1,i}$, we have for all $x \in \mathcal{N}_{x_{\text{init}}}$,

$$\begin{aligned}
\Pi_{k-1,\sqrt{\delta}}\big(\pi_{k,\sqrt{\delta}}(x)\big) = x + \sqrt{\delta}\nabla G_{j_k}(x) + \sqrt{\delta} \cdot \mathrm{Lie}_G(j_{1:(k-1)})(x) + \frac{\delta}{2}\partial(\nabla G_{j_k})(x)\nabla G_{j_k}(x) \\
+ \delta \cdot \partial\mathrm{Lie}_G(j_{1:(k-1)})(x)\nabla G_{j_k}(x) + \sum_{i=2^{k-2}+1}^{2^{k-1}} \frac{\delta^{i/2^{k-1}}}{i!}\Delta_{k-1,i}(x) + o(\delta)
\end{aligned} \tag{23}$$

for sufficiently small δ. For the other way around, we similarly have

$$\begin{aligned}\pi_{k,\sqrt{\delta}}\big(\Pi_{k-1,\sqrt{\delta}}(x)\big) &= \pi_{k,\sqrt{\delta}}\bigg(x + \sqrt{\delta}\cdot \mathrm{Lie}_G(j_{1:(k-1)})(x) + \sum_{i=2^{k-2}+1}^{2^{k-1}} \frac{\delta^{i/2^{k-1}}}{i!}\Delta_{k-1,i}(x) + o(\delta)\bigg)\\ &= x + \sqrt{\delta}\nabla G_{j_k}(x) + \sqrt{\delta}\cdot \mathrm{Lie}_G(j_{1:(k-1)}) + \frac{\delta}{2}\partial(\nabla G_{j_k})(x)\nabla G_{j_k}(x)\\ &\quad + \delta\partial(\nabla G_{j_k})(x)\mathrm{Lie}_G(j_{1:(k-1)})(x) + \sum_{i=2^{k-2}+1}^{2^{k-1}} \frac{\delta^{i/2^{k}}}{i!}\Delta_{k-1,i}(x) + o(\delta)\end{aligned} \tag{24}$$

for all $x \in \mathcal{N}_{x_{\text{init}}}$, when δ is sufficiently small. Note that $x = \pi^{-1}_{k,\sqrt{\delta}} \circ \Pi^{-1}_{k-1,\sqrt{\delta}} \circ \Pi_{k-1,\sqrt{\delta}} \circ \pi_{k,\sqrt{\delta}}(x)$, thus

$$\begin{aligned}\Pi_{k,\delta}(x) - x &= \pi^{-1}_{k,\sqrt{\delta}} \circ \Pi^{-1}_{k-1,\sqrt{\delta}} \circ \pi_{k,\sqrt{\delta}} \circ \Pi_{k-1,\sqrt{\delta}}(x) - x\\ &= \pi^{-1}_{k,\sqrt{\delta}} \circ \Pi^{-1}_{k-1,\sqrt{\delta}} \circ \pi_{k,\sqrt{\delta}} \circ \Pi_{k-1,\sqrt{\delta}}(x) - \pi^{-1}_{k,\sqrt{\delta}} \circ \Pi^{-1}_{k,\sqrt{\delta}} \circ \Pi_{k,\sqrt{\delta}} \circ \pi_{k,\sqrt{\delta}}(x)\\ &= \pi^{-1}_{k,\sqrt{\delta}} \circ \Pi^{-1}_{k-1,\sqrt{\delta}} \circ \pi_{k,\sqrt{\delta}} \circ \Pi_{k-1,\sqrt{\delta}}(x) - \pi_{k,\sqrt{\delta}} \circ \Pi_{k-1,\sqrt{\delta}}(x)\\ &\quad + \pi_{k,\sqrt{\delta}} \circ \Pi_{k-1,\sqrt{\delta}}(x) - \Pi_{k,\sqrt{\delta}} \circ \pi_{k,\sqrt{\delta}}(x)\\ &\quad + \Pi_{k-1,\sqrt{\delta}}(x) \circ \pi_{k,\sqrt{\delta}} - \pi^{-1}_{k,\sqrt{\delta}} \circ \Pi^{-1}_{k,\sqrt{\delta}} \circ \Pi_{k,\sqrt{\delta}} \circ \pi_{k,\sqrt{\delta}}(x)\\ &= \Pi_{k-1,\sqrt{\delta}} \circ \pi_{k,\sqrt{\delta}}(x) - \pi_{k,\sqrt{\delta}} \circ \Pi_{k-1,\sqrt{\delta}}(x) + o(\delta)\end{aligned} \tag{25}$$

where the last equality follows from the Taylor expansion of $\pi^{-1}_{k,\sqrt{\delta}} \circ \Pi^{-1}_{k-1,\sqrt{\delta}}(\cdot)$ in terms of $\sqrt{\delta}$. Now, combining (23), (24) and (25), we obtain

$$\begin{aligned}\Pi_{k,\delta}(x) - x &= \delta\left(\partial(\nabla G_{j_k})(x)\mathrm{Lie}_G(j_{1:(k-1)})(x) - \partial\mathrm{Lie}_G(j_{1:(k-1)})(x)\nabla G_{j_k}(x)\right) + o(\delta)\\ &= \delta\cdot[\mathrm{Lie}_G(j_{1:(k-1)}), \nabla G_{j_k}](x) + o(\delta)\end{aligned} \tag{26}$$

where the second equality follows from the definition of Lie bracket. Comparing (26) with (20) yields (22). This completes the induction for $k \in [K]$ and hence finishes the proof as K is arbitrary. □

Proof of Corollary 4.11. It turns out that the necessary condition in Theorem 4.9 is already violated by only considering the Lie algebra spanned by $\{\nabla G_{11}, \nabla G_{12}\}$. We follow the notation in Example 4.4 to define each $E_{ij} \in \mathbb{R}^d$ as the one-hot matrix with the (i,j)-th entry being 1, and denote $\overline{E}_{ij} = \frac{1}{2}(E_{ij} + E_{ji})$ and $\Delta_{ij} = E_{ij} - E_{ji}$. Then $[\nabla G_{11}, \nabla G_{12}](U) = 4(\overline{E}_{11}\overline{E}_{12} - \overline{E}_{12}\overline{E}_{11})U = \Delta_{12}U$ and $[\nabla G_{11}, [\nabla G_{11}, \nabla G_{12}]](U) = (\overline{E}_{11}\Delta_{12} - \Delta_{12}\overline{E}_{11})U = \overline{E}_{12}U$. Further noting that $\langle[\nabla G_{11}, [\nabla G_{11}, \nabla G_{12}]], \nabla G_{12}\rangle = 2\left\|\overline{E}_{12}U\right\|_F^2 = \frac{1}{2}\sum_{i=1}^r (U_{1i}^2 + U_{2i}^2)$ must be positive at some U in every open set M, by Theorem 4.9, we know such U_{init} and L_t exist. Moreover, L_t will only depend on $G_{11}(U)$ and $G_{12}(U)$. □

Proof of Corollary 4.12. By the condition (b) and Theorem 4.9, we know that each Lie bracket $[\nabla G_i, \nabla G_j] \in \ker(\partial G)$. By the condition (a), we know that each Lie bracket $[\nabla G_i, \nabla G_j] \in \mathrm{span}\{\nabla G_i\}_{i=1}^d$. Combining these two facts, we conclude that each $[\nabla G_i, \nabla G_j] \equiv 0$, so G is a commuting parametrization. □

D.3 Convergence for gradient flow with commuting parametrization

Proof of Theorem 4.13. Recall that the dynamics of $w(t)$ is given by

$$\mathrm{d}w(t) = -\nabla^2 R(w(t))^{-1}\nabla L(w(t))\mathrm{d}t, \qquad w(0) = G(x_{\text{init}}).$$

By Lemma 4.7, we know that R is a Legendre function. Therefore, when R is further a Bregman function, we can apply Theorem B.8 to obtain the convergence of $w(t)$. This finishes the proof. □

Based on Theorem B.7, we can prove the trajectory convergence of $w(t)$ for the special case where $\Omega_w(x_{\text{init}}; G) = \mathbb{R}^d$ as summarized in Corollary 4.14.

Proof of Corollary 4.14. It suffices to verify that R is a Bregman function in this case. By Lemma 4.7, we know that R is a Legendre function and satisfies that $\mathbb{R}^d = \Omega_w(x_{\text{init}}; G) = \operatorname{dom} \nabla R \subseteq \operatorname{dom} R \subseteq \mathbb{R}^d$, which implies $\operatorname{dom} R = \mathbb{R}^d$. Moreover, the domain of its convex conjugate Q is also $\mathbb{R}^d$. Then by Theorem B.7, we see that R is a Bregman function. This finishes the proof. □

Next, we prove that for a class of commuting quadratic parametrizations, the corresponding Legendre function is also a Bregman function, thus guaranteeing the trajectory convergence.

Proof of Theorem 4.15. Since $A_1, A_2, \ldots, A_d$ commute with each other, these matrices can be simultaneously diagonalized. Thus we can assume without loss of generality that each $A_i = \operatorname{diag}(\lambda_i)$ where $\lambda_i \in \mathbb{R}^D$, then $G_i(x) = \lambda_i^\top x^{\odot 2}$. For convenience, we denote $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_d)^\top \in \mathbb{R}^{d\times D}$, so the parametrization is given by $G(x) = \Lambda x^{\odot 2}$. Note that for each $i \in [d]$, $\nabla G_i(x) = 2\lambda_i \odot x$ and $\nabla^2 G_i(x) = 2\operatorname{diag}(\lambda_i)$, so for any $i, j \in [d]$, we have

$$[\nabla G_i, \nabla G_j](x) = 4\operatorname{diag}(\lambda_i)\lambda_j \odot x - 4\operatorname{diag}(\lambda_j)\lambda_i \odot x = 0.$$

Therefore, we see that $G : \mathbb{R}_+^D \to \mathbb{R}^d$ is a commuting parametrization. Also, for any $t \in \mathbb{R}$, $x(t) = x_{\text{init}} - \int_0^t \nabla G_i(x(s))\mathrm{d}s = x_{\text{init}} \odot e^{-2\lambda_i t}$, which proves the first and the second claims. Moreover, if the sign of each coordinate of x will not change from that of initialization, (sign means $+$,$-$ or 0). Without loss of generality, below we will assume every coordinate is non-zero at initialization (otherwise we just ignore it). We can also assume the coordinates at initialization are all positive, as the negatives will induce the same trajectory in terms of $G(x)$. By Theorem 4.8, the dynamics of $w(t) = G(x(t))$ is given by

$$\mathrm{d}w(t) = -\nabla^2 R(w(t))^{-1}\nabla L(w(t))\mathrm{d}t, \qquad w(0) = G(x_{\text{init}})$$

for some Legendre function R whose conjugate is denoted by Q. To apply the results in Theorem 4.13, it suffices to show that this R is a Bregman function.

To do so, we further denote $\widetilde{w} = x^{\odot 2}$ and $\widetilde{G}(x) = x^{\odot 2}$, then $w = \Lambda\widetilde{w}$ and in this case $\widetilde{G}$ is a commuting parametrization for $\widetilde{w}$ defined on $M = \mathbb{R}_+^D$. Also, we have $\partial G(x) = \Lambda\partial\widetilde{G}(x)$. Let $\widetilde{L} : \mathbb{R}^d \to \mathbb{R}$ be defined by $\widetilde{L}(\widetilde{w}) = L(\Lambda\widetilde{w})$, which satisfies that $\nabla\widetilde{L}(\widetilde{w}) = \Lambda^\top\nabla L(\Lambda\widetilde{w})$. Then the gradient flow with parametrization $\widetilde{G}$ governed by $-\nabla(\widetilde{L}\circ\widetilde{G})(x)$ is given by

$$\begin{aligned}\mathrm{d}x(t) &= -\nabla(\widetilde{L}\circ\widetilde{G})(x)\mathrm{d}t = -\partial\widetilde{G}(x(t))^\top\nabla\widetilde{L}(\widetilde{G}(x(t))\mathrm{d}t\\ &= -\partial\widetilde{G}(x(t))^\top\Lambda^\top\nabla L(\Lambda\widetilde{G}(x(t))\mathrm{d}t\\ &= -\partial G(x(t))^\top\nabla L(G(x(t))\mathrm{d}t,\end{aligned}$$

which yields the same dynamics of the gradient flow with parametrization G governed by $-\nabla(L\circ G)(x)$. Therefore, we have $w(t) = G(x(t)) = \Lambda\widetilde{G}(x(t)) = \Lambda\widetilde{w}(t)$, where again by Theorem 4.8, the dynamics of $\widetilde{w}(t)$ is

$$\mathrm{d}\widetilde{w}(t) = -\nabla^2\widetilde{R}(\widetilde{w}(t))^{-1}\nabla\widetilde{L}(\widetilde{w}(t))\mathrm{d}t, \qquad \widetilde{w}(0) = \widetilde{G}(x_{\text{init}})$$

for some Legendre function $\widetilde{R}$ whose conjugate is denoted by $\widetilde{Q}$. For any $x \in M$ and $\widetilde{\mu} \in \mathbb{R}^D$, we define $\widetilde{\psi}(x;\widetilde{\mu}) = \phi^{\widetilde{\mu}_1}_{\widetilde{G}_1} \circ \phi^{\widetilde{\mu}_2}_{\widetilde{G}_2} \circ \cdots \circ \phi^{\widetilde{\mu}_D}_{\widetilde{G}_D}(x)$. We need the following lemma.

Lemma D.3. *In the setting of the proof of Theorem 4.15, for any $\mu \in \mathbb{R}^d$ and $x \in M$, we have $\psi(x;\mu) = \widetilde{\psi}(x;\Lambda^\top\mu)$.*

Recall from Lemma 4.7 that $\nabla Q(\mu) = G(\psi(x_{\text{init}};\mu))$ for any $\mu \in \mathbb{R}^d$ and $\nabla\widetilde{Q}(\widetilde{\mu}) = \widetilde{G}(\widetilde{\psi}(x_{\text{init}};\widetilde{\mu}))$ for any $\widetilde{\mu} \in \mathbb{R}^D$. Note that

$$\nabla Q(\mu) = \Lambda\psi(x_{\text{init}};\mu)^{\odot 2} = \Lambda\widetilde{\psi}(x_{\text{init}};\Lambda^\top\mu)^{\odot 2} = \Lambda\widetilde{G}(\widetilde{\psi}(x_{\text{init}};\Lambda^\top\mu)) = \Lambda\nabla\widetilde{Q}(\Lambda^\top\mu) \tag{27}$$

where the second equality follows from Lemma D.3. This implies that $Q(\mu) = \widetilde{Q}(\Lambda^\top\mu) + C$ for some constant C. Recall the definition of convex conjugate, and we have

$$\widetilde{R}(\widetilde{w}) = \sup_{\widetilde{\mu}\in\mathbb{R}^D} \langle\widetilde{\mu},\widetilde{w}\rangle - \widetilde{Q}(\widetilde{\mu}), \qquad R(w) = \sup_{\mu\in\mathbb{R}^d} \langle\mu, w\rangle - Q(\mu).$$

Then for any $\widetilde{w} \in \mathbb{R}^D$, we have

$$\begin{aligned} R(\Lambda\widetilde{w}) &= \sup_{\mu\in\mathbb{R}^d} \langle \mu, \Lambda\widetilde{w}\rangle - Q(\mu) = \sup_{\mu\in\mathbb{R}^d} \langle \Lambda^\top\mu, \widetilde{w}\rangle - \widetilde{Q}(\Lambda^\top\mu) - C \\ &= \sup_{\widetilde{\mu}\in\Lambda^\top\mathbb{R}^d} \langle \widetilde{\mu}, \widetilde{w}\rangle - \widetilde{Q}(\widetilde{\mu}) - C \le \sup_{\widetilde{\mu}\in\mathbb{R}^D} \langle \widetilde{\mu}, \widetilde{w}\rangle - \widetilde{Q}(\widetilde{\mu}) - C = \widetilde{R}(\widetilde{w}) - C \end{aligned} \tag{28}$$

Therefore, for any $\widetilde{w} \in \operatorname{dom}\widetilde{R}$, it holds that $R(\Lambda\widetilde{w}) \le \widetilde{R}(\widetilde{w}) - C < \infty$, so $\Lambda \operatorname{dom}\widetilde{R} \subseteq \operatorname{dom} R$, where $\Lambda \operatorname{dom}\widetilde{R}$ On the other hand, by (27) and Proposition B.3, we have

$$\operatorname{dom}\nabla R = \operatorname{range}\nabla Q \subseteq \Lambda \operatorname{range}\nabla\widetilde{Q} = \Lambda \operatorname{dom}\nabla\widetilde{R}$$

and it follows that

$$\operatorname{int}(\operatorname{dom} R) = \operatorname{dom}\nabla R \subseteq \Lambda \operatorname{dom}\nabla\widetilde{R} = \Lambda \operatorname{int}(\operatorname{dom}\widetilde{R}).$$

Combining the above, we see that $\operatorname{dom} R = \Lambda \operatorname{dom}\widetilde{R}$. As discussed in Section 1, here it is straightforward to verify that $\widetilde{R}(\widetilde{w}) = \sum_{i=1}^D \widetilde{w}_i(\ln\frac{\widetilde{w}_i}{x_{\text{init},i}^2} - 1)$, which is indeed a Bregman function with domain $\operatorname{dom}\widetilde{R} = \overline{\mathbb{R}^D_+}$. Thus $\operatorname{dom} R = \Lambda\overline{\mathbb{R}^D_+}$ is also a closed set. This yields the first condition in Definition B.6.

Next, we verify the second condition in Definition B.6. For any $\mu \in \mathbb{R}^d$, we have

$$\nabla R(G(\psi(x_{\text{init}};\mu))) = \nabla R(\nabla Q(\mu)) = \mu$$

and

$$\nabla\widetilde{R}(\widetilde{G}(\psi(x_{\text{init}};\mu))) = \nabla\widetilde{R}(\widetilde{G}(\widetilde{\psi}(x_{\text{init}};\Lambda^\top\mu))) = \nabla\widetilde{R}(\nabla\widetilde{Q}(\Lambda^\top\mu)) = \Lambda^\top\mu.$$

Comparing the above two equalities, we get

$$\nabla\widetilde{R}(\widetilde{w}) = \Lambda^\top\nabla R(\Lambda\widetilde{w}) \tag{29}$$

for all $\widetilde{w} \in \mathbb{R}^D_+$. Then for any $\widetilde{w} \in \overline{\mathbb{R}^D_+}$ and $y = \Lambda\widetilde{y} \in \operatorname{int}(\operatorname{dom} R)$, we have

$$\begin{aligned} D_R(\Lambda\widetilde{w}, y) &= R(\Lambda\widetilde{w}) - R(y) - \langle \nabla R(y), \Lambda\widetilde{w} - y\rangle \\ &= R(\Lambda\widetilde{w}) - R(\Lambda\widetilde{y}) - \langle \Lambda^\top\nabla R(\Lambda\widetilde{y}), \widetilde{w} - \widetilde{y}\rangle \\ &= R(\Lambda\widetilde{w}) - R(\Lambda\widetilde{y}) - \langle \nabla\widetilde{R}(\widetilde{y}), \widetilde{w} - \widetilde{y}\rangle \\ &= R(\Lambda\widetilde{w}) - R(\Lambda\widetilde{y}) - \widetilde{R}(\widetilde{w}) + \widetilde{R}(\widetilde{y}) + D_{\widetilde{R}}(\widetilde{w}, \widetilde{y}) \\ &\ge R(\Lambda\widetilde{w}) - \widetilde{R}(\widetilde{w}) + C + D_{\widetilde{R}}(\widetilde{w}, \widetilde{y}) \end{aligned} \tag{30}$$

where the inequality follows from (28). Therefore, we further have for any $\alpha \in \mathbb{R}$

$$\{y \in \operatorname{int}(\operatorname{dom} R) \mid D_R(\Lambda\widetilde{w}, y) \le \alpha\} \subseteq \Lambda\{\widetilde{y} \in \mathbb{R}^D_+ \mid D_{\widetilde{R}}(\widetilde{w}, \widetilde{y}) \le \alpha - R(\Lambda\widetilde{w}) + \widetilde{R}(\widetilde{w}) - C\}$$

where the right-hand side is bounded since $\widetilde{R}$ is a Bregman function, and so is the left-hand side.

Finally, we verify the third condition in Definition B.6. Consider any $w \in \operatorname{dom} R$ and sequence $\{w_i\}_{i=1}^\infty \subset \operatorname{int}(\operatorname{dom} R)$ such that $\lim_{i\to\infty} w_i = w$. Since $\operatorname{dom} R = \Lambda \operatorname{dom}\widetilde{R}$, there is some $\widetilde{w} \in \overline{\mathbb{R}^D_+}$ such that $w = \Lambda\widetilde{w}$ and some $\widetilde{w}_i \in \mathbb{R}^D_+$ for each $i \in \mathbb{N}^+$ such that $w_i = \Lambda\widetilde{w}_i$. We have that

$$\begin{aligned} R(w) - R(w_i) &= \int_0^1 \langle \nabla R((1-t)w_i + tw), w - w_i\rangle \mathrm{d}t \\ &= \int_0^1 \langle \Lambda^\top\nabla R(\Lambda((1-t)\widetilde{w}_i + t\widetilde{w})), \widetilde{w} - \widetilde{w}_i\rangle \mathrm{d}t \\ &= \int_0^1 \langle \nabla\widetilde{R}((1-t)\widetilde{w}_i + t\widetilde{w}), \widetilde{w} - \widetilde{w}_i\rangle \mathrm{d}t \\ &= \widetilde{R}(\widetilde{w}) - \widetilde{R}(\widetilde{w}_i). \end{aligned}$$

Combining this with (30), we get $D_R(w, w_i) = D_{\widetilde{R}}(\widetilde{w}, \widetilde{w}_i)$. Note that we can always choose each $\widetilde{w}_i$ properly such that $\lim_{i\to\infty}\widetilde{w}_i = \widetilde{w}$. Then since $\widetilde{R}$ is a Bregman function, we have

$$\lim_{i\to\infty} D_R(w, w_i) = \lim_{i\to\infty} D_{\widetilde{R}}(\widetilde{w}, \widetilde{w}_i) = 0.$$

Therefore, we conclude that R is also a Bregman function. This finishes the proof. □

Proof of Lemma D.3. For each $i \in [D]$ and any $t > 0$, we have

$$\phi^t_{G_i}(x) = x + \int_{s=0}^{t} -\nabla G_i(\phi^s_{f_i}(x))\mathrm{d}s = x + \int_{s=0}^{t} -\sum_{j=1}^{D} \lambda_{i,j} \nabla \widetilde{G}_j(\phi^s_{f_i}(x))\mathrm{d}s = \widetilde{\psi}(x; t\lambda_i)$$

where the last equality follows from Lemma 4.6. Therefore, for any $\mu \in \mathbb{R}^d$, we further have

$$\begin{aligned}
\psi(x;\mu) &= \phi^{\mu_1}_{G_1} \circ \phi^{\mu_2}_{G_2} \circ \cdots \circ \phi^{\mu_d}_{G_d}(x) \\
&= \phi^{\mu_1\lambda_{1,1}}_{\widetilde{G}_1} \circ \cdots \circ \phi^{\mu_1\lambda_{1,D}}_{\widetilde{G}_D} \circ \cdots \circ \phi^{\mu_d\lambda_{d,1}}_{\widetilde{G}_1} \circ \cdots \circ \phi^{\mu_d\lambda_{d,D}}_{\widetilde{G}_D}(x) \\
&= \phi^{\sum_{i=1}^d \mu_i\lambda_{i,1}}_{\widetilde{G}_1} \circ \cdots \circ \phi^{\sum_{i=1}^d \mu_i\lambda_{i,D}}_{\widetilde{G}_D}(x) \\
&= \phi^{(\Lambda^\top \mu)_1}_{\widetilde{G}_1} \circ \cdots \phi^{(\Lambda^\top \mu)_D}_{\widetilde{G}_D}(x) = \widetilde{\psi}(x; \Lambda^\top \mu).
\end{aligned}$$

where the third equality follows from the assumption that $\widetilde{G}$ is a commuting parametrization. This finishes the proof. □

D.4 Results for underdetermined linear regression

Here we provide the proof for the implicit bias result for the quadratically overparametrized linear model.

Proof of Theorem 4.16. By Theorem 4.8, $w(t)$ 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 3.8 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

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

which can be reorganized as

$$R(w_\infty) \le R(w) - \langle \nabla R(G(x_{\text{init}})), w - w_\infty \rangle. \tag{31}$$

Note that by Lemma 4.7, we also have

$$\nabla R(G(x_{\text{init}})) = \nabla R(G(\psi(x_{\text{init}}; 0))) = \nabla R(\nabla Q(0)) = 0 \tag{32}$$

where the last equality follows from the property of convex conjugate. Combining (31) and (32), we get $R(w_\infty) \le R(w)$ 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 entire $\operatorname{dom} R$, and for any $w \notin \operatorname{dom} R$, we have $R(w) = \infty$ by definition, so $R(w_\infty) \le R(w)$ holds trivially. This finishes the proof. □

Proof of Corollary 4.17. By symmetry, we assume without loss of generality that all coordinates of x_{init} are positive. Note that for $M = \mathbb{R}^D_+$ with $D = 2d$, $G : M \to \mathbb{R}^d$ can be written as $G_i(x) = x^\top A_i x$ where each $A_i = e_i e_i^\top - e_{d+i} e_{d+i}^\top$. Therefore, this parametrization G satisfies the conditions in Theorem 4.15, which then implies the convergence of $w(t)$.

Next, we identify the function R given by Theorem 4.8. we have $\psi(x_{\text{init}}; \mu) = \binom{u_0 \odot e^{-2\mu}}{v_0 \odot e^{2\mu}}$ and thus

$$\begin{aligned}
G(\psi(x_{\text{init}}; \mu)) &= u_0^{\odot 2} \odot e^{-4\mu} - v_0^{\odot 2} \odot e^{4\mu} \\
&= (u_0^{\odot 2} + v_0^{\odot 2}) \odot \sinh(4\mu) + (u_0^{\odot 2} - v_0^{\odot 2}) \odot \cosh(4\mu).
\end{aligned}$$

So $G(\psi(x_{\text{init}}; \mu))$ is the gradient of $Q(\mu) = \frac{1}{4}(u_0^{\odot 2} + v_0^{\odot 2}) \odot \cosh(4\mu) + \frac{1}{4}(u_0^{\odot 2} - v_0^{\odot 2}) \odot \sinh(4\mu) + C$ where C is an arbitrary constant. Also note that $(\nabla Q(\mu))_i$ only depends on μ_i, then we have

$$\begin{aligned}
(\nabla R(w))_i = (\nabla Q(\mu))_i^{-1}(w) &= \frac{1}{4} \ln \left(\sqrt{1 + \left(\frac{w_i}{2u_{0,i}v_{0,i}} \right)^2} + \frac{w_i}{2u_{0,i}v_{0,i}} \right) + \frac{1}{4} \ln \frac{v_{0,i}}{u_{0,i}} \\
&= \frac{1}{4} \operatorname{arcsinh} \left(\frac{w_i}{2u_{0,i}v_{0,i}} \right) + \frac{1}{4} \ln \frac{v_{0,i}}{u_{0,i}}
\end{aligned}$$

which further implies that

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

This finishes the proof. □

E Omitted proofs in Section 5

We first prove the following intermediate result that will be useful in the proof of Theorem 5.1.

Lemma E.1. *Under the setting of Theorem 5.1, let F be the smooth map that isometrically embeds $(\operatorname{int}(\operatorname{dom} R), g^R)$ into $(\mathbb{R}^D, \overline{g})$. Let $M = \operatorname{range}(F)$, and denote the inverse of F by $\widetilde{G} : M \to \mathbb{R}^d$. Then for any $w \in \operatorname{int}(\operatorname{dom} R)$, it holds that*

$$\partial F(w)(\partial F(w)^\top \partial F(w))^{-1} = \partial\widetilde{G}(F(w))^\top \quad \textit{and} \quad \partial\widetilde{G}(F(w))\partial\widetilde{G}(F(w))^\top = \nabla^2 R(w)^{-1}.$$

Proof of Lemma E.1. For any $x \in M$ and $v \in T_x(M)$, consider a parametrized curve $\{x(t)\}_{t\geq 0} \subset M$ such that $x(0) = x$ and $\frac{\mathrm{d}x(t)}{\mathrm{d}t}\big|_{t=0} = v$. Since $x(t) = F(\widetilde{G}(x(t)))$ for any $t \geq 0$, differentiating with respect to t on both sides and evaluating at $t = 0$ yield

$$v = \partial F(\widetilde{G}(x))\partial\widetilde{G}(x)v. \tag{33}$$

Now, for any $w \in \operatorname{int}(\operatorname{dom} R)$, let $x = F(w)$, then for any $v \in T_x(M)$, it follows from (33) that

$$v^\top \partial F(w) = v^\top(\partial F(w)\partial\widetilde{G}(F(w)))^\top \partial F(w) = v^\top \partial\widetilde{G}(F(w))^\top \partial F(w)^\top \partial F(w).$$

Note that the span of the column space of $\partial F(w)$ is exactly $T_x(M)$, so for any v in the orthogonal complement of $T_x(M)$, it holds that

$$v^\top \partial F(w) = 0 = v^\top \partial\widetilde{G}(F(w))^\top \partial F(w)^\top \partial F(w)$$

where the second equality follows from the fact that for any $i \in [d]$, $\nabla\widetilde{G}_i(x) \in T_x(M)$. Therefore, combining the above two cases, we conclude that

$$\partial F(w) = \partial\widetilde{G}(F(w))^\top \partial F(w)^\top \partial F(w).$$

Since $\partial F(w)^\top \partial F(w) = \nabla^2 R(w)$ is invertible, we then get

$$\partial\widetilde{G}(F(w))^\top = \partial F(w)(\partial F(w)^\top \partial F(w))^{-1}.$$

Next, for any $w \in \operatorname{int}(\operatorname{dom} R)$, since $\widetilde{G}(F(w)) = w$, differentiating on both sides yields

$$\partial\widetilde{G}(F(w))\partial F(w) = I_d.$$

Therefore, using the identity proved above, we have

$$\begin{aligned}\partial\widetilde{G}(F(w))\partial\widetilde{G}(F(w))^\top &= \partial\widetilde{G}(F(w))\partial F(w)(\partial F(w)^\top \partial F(w))^{-1} \\ &= (\partial F(w)^\top \partial F(w))^{-1} = \nabla^2 R(w)^{-1}.\end{aligned}$$

This finishes the proof. □

Proof of Theorem 5.1. By Nash's embedding theorem, there is a smooth map $F : \operatorname{int}(\operatorname{dom} R) \to \mathbb{R}^D$ that isometrically embeds $(\operatorname{int}(\operatorname{dom} R), g^R)$ into $(\mathbb{R}^D, \overline{g})$. Denote $M = \operatorname{range}(F)$, *i.e.*, the embedding of $\operatorname{int}(\operatorname{dom} R)$ in $\mathbb{R}^D$. We further denote the inverse of F on M by $\widetilde{G} : M \to \mathbb{R}^d$. Note $(M, \widetilde{G})$ is a global atlas for M, we have that $T_x(M) = \operatorname{span}(\{\nabla\widetilde{G}_i(x)\}_{i=1}^d)$ for all $x \in M$. This $\widetilde{G}$ is almost the commuting parametrization that we seek for, except now it is only defined on M but not on an open neighborhood of M. Yet we can extend $\widetilde{G}$ to an open neighbourhood of M in the following way: First by [21], for each $x \in M$, there is an open neighbourhood U_x of x such that projection function P defined by

$$P(y) = \operatorname*{argmin}_{y' \in M} \|y - y'\|_2$$

is smooth in U_x. Then we define $U = \cup_{x\in M} U_x$, and extend $\widetilde{G}$ to U by defining $G(x) := \widetilde{G}(P(x))$ for all $x \in U$. We have $G(x) = \widetilde{G}(x)$ for all $x \in M$, and we can verify that $\partial G \equiv \partial \widetilde{G}$ on M as well. For any $v \in T_x(M)$, let $\{\gamma(t)\}_{t\geq 0}$ be a parametrized curve on M such that $\gamma(0) = x$ and $\frac{\mathrm{d}\gamma(t)}{\mathrm{d}t}\big|_{t=0} = v$, then for sufficiently small t, by Taylor expansion we have

$$\begin{aligned}\gamma(t) = P(\gamma(t)) &= P(x) + \partial P(x)(\gamma(t) - x) + o(\|\gamma(t) - x\|_2)\\ &= x + \partial P(x)(\gamma(t) - x) + o(\|\gamma(t) - x\|_2)\end{aligned}$$

which implies that $v = \partial P(x)v$ by letting $t \to 0$. While for any v in the orthogonal complement of $T_x(M)$, for sufficiently small $\delta > 0$, we have $P(x + \delta v)$ is smooth in δ. Then since $P(x + \delta v) \in M$ for all sufficiently small δ by its definition, we have

$$\partial P(x)v = \frac{\mathrm{d}P(x+\delta v)}{\mathrm{d}\delta}\bigg|_{\delta=0} = \lim_{\delta\to 0}\frac{P(x+\delta v) - P(x)}{\delta} =: u \in T_x(M). \tag{34}$$

Note that $\|x + \delta v - P(x + \delta v)\|_2 \leq \|x + \delta v - P(x)\|_2 = \delta\|v\|_2$, and by Taylor expansion, we have

$$\|x + \delta v - P(x + \delta v)\|_2 = \|x + \delta v - \delta\partial P(x)v + O(\delta^2)\|_2 = \|x + \delta v - \delta u + O(\delta^2)\|_2$$

where $O(\delta^2)$ denotes a term whose norm is bounded by $C\delta^2$ for a constant $C > 0$ for all sufficiently small δ, and the second equality follows from (34). Then dividing both sides by δ and letting $\delta \to 0$, we have $\|v\|_2 \geq \|v - u\|_2$. Since u is orthogonal to v, we must have $u = 0$. As v is arbitrary, we conclude that $\partial P(x)$ is the orthogonal projection matrix onto $T_x(M)$. Then differentiating both sides of $G(x) = \widetilde{G}(P(x))$ with x yields

$$\partial G(x) = \partial\widetilde{G}(P(x))\partial P(x) = \partial\widetilde{G}(x) \tag{35}$$

where the second equality follows from the fact that $T_x(M) = \mathrm{span}(\{\nabla\widetilde{G}_i(x)\}_{i=1}^d)$. This further implies that the solution of Equation (11) satisfies $\mathrm{d}x/\mathrm{d}t = -\nabla(L \circ \widetilde{G})(x) \in T_x(M)$, and thus $x(t) \in M$ for all $t \geq 0$.

Now we consider the mirror flow

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

Since $\nabla^2 R(w) = \partial F(w)^\top \partial F(w)$ by the fact that F is an isometric embedding, we further have

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

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

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

where the third equality follows from Lemma E.1 and (35).

Next, we verify that G restricted on M, $\widetilde{G}$, is a commuting and regular parametrization. First, for any $x \in M$, we have $\partial\widetilde{G}(x)^\top = \partial F(\widetilde{G}(x))(\partial F(\widetilde{G}(x))^\top \partial F(\widetilde{G}(x)))^{-1}$ by Lemma E.1 and (35). Since $\nabla^2 R(w) = \partial F(w)^\top \partial F(w)$ is of rank d for all $w \in \mathrm{int}(\mathrm{dom}\, R)$, it follows that $\partial F(w)$ is also of rank d for all $w \in \mathrm{int}(\mathrm{dom}\, R)$, thus $\partial\widetilde{G}(x)$ is of rank d for all $x \in M$. The commutability of $\{\nabla\widetilde{G}_i\}_{i=1}^d$ follows directly from Corollary 4.12. Here we just need to show $\mathrm{rank}(\Omega_x(x;\widetilde{G})) = \mathrm{rank}(M)$. This is because on one hand $\mathrm{rank}(\Omega_x(x;\widetilde{G})) \geq \mathrm{rank}(\mathrm{span}(\{\nabla\widetilde{G}_i(x)\}_{i=1}^d)) = \mathrm{rank}(M)$, and on the other hand, $\mathrm{rank}(\Omega_x(x;\widetilde{G})) \leq \mathrm{rank}(M)$ since $\Omega_x(x;\widetilde{G}) \subset M$, for any $x \in M$.

Finally, we show that when R is a mirror map, each $\nabla\widetilde{G}_j$ is a complete vector field on M. For any $x_{\mathrm{init}} \in M$, consider loss $L_t(w) = \langle e_j, w\rangle$, and the corresponding gradient flow is

$$\mathrm{d}x(t) = -\nabla(L_t \circ \widetilde{G})(x(t))\mathrm{d}t = -\partial\widetilde{G}(x(t))^\top \nabla L_t(\widetilde{G}(x(t)))\mathrm{d}t = -\nabla\widetilde{G}_j(x(t)),$$

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

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

where the third equality follows from Lemma E.1 and Equation (35). Therefore, rewriting the above as a mirror Flow yields

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

the solution to which exists for all $t \in \mathbb{R}$ and is given by $\nabla R(w(t)) = e_j t$, so $w(t) = (\nabla R)^{-1}(e_j t)$ is defined for all $t \in \mathbb{R}$ as ∇R is surjective. This further implies that $x(t) = F(w(t))$ is well-defined for all $t \in \mathbb{R}$, hence $\nabla \widetilde{G}_j$ is a complete vector field. □