A PRELIMINARIES ON STOCHASTIC PROCESSES

In this section we review a few basics of stochastic processes that will be useful for proving our results, so that our paper will be self-contained. We refer the reader to classics like Karatzas & Shreve (2014); Billingsley (2013); Pollard (2012) for more systematic derivations.

Throughout this section, let $\mathcal{E}$ be a Banach space equipped with norm $\|\cdot\|$, e.g., $(\mathbb{R}, |\cdot|)$ and $(\mathbb{R}^D, \|\cdot\|_2)$.

A.1 CADLAG FUNCTION AND METRIC

Definition A.1 (Cadlag function)**.** Let $T \in [0, \infty]$. A function $g : [0, T) \to E$ is *cadlag* if for all $t \in [0, T)$ it is right-continuous at t and its left limit $g(t-)$ exists. Let $D_{\mathcal{E}}[0, T)$ be the set of all cadlag function mapping $[0, T)$ into $\mathcal{E}$.

Definition A.2 (Continuity modulus)**.** For any function $f : [0, \infty) \to \mathcal{E}$ and any interval $I \subseteq [0, \infty)$, we define

$$\omega(f; I) = \sup_{s,t \in I} \|f(s) - f(t)\|.$$

For any $N \in \mathbb{N}$ and $\theta > 0$, we further define the continuity modulus of continuous f as

$$\omega_N(f, \theta) = \sup_{0 \le t \le t+\theta \le N} \{\omega(f; [t, t + \theta])\}.$$

Moreover, the continuity modulus of cadlag $f \in D_{\mathcal{E}}[0, \infty)$ is defined as

$$\omega'_N(f, \theta) = \inf \left\{ \max_{i \le r} \omega(f; [t_{i-1}, t_i) : 0 \le t_0 < \cdots < t_r = N, \inf_{i<r} (t_i - t_{i-1}) \ge \theta \right\}.$$

Definition A.3 (Jump)**.** For any $g \in D_{\mathcal{E}}[0, T)$, we define the jump of g at t to be

$$\Delta g(t) = g(t) - g(t-).$$

For any $\delta > 0$, we define $h_\delta : [0, \infty) \to [0, \infty)$ by

$$h_\delta(r) = \begin{cases} 0 & \text{if } r \le \delta \\ 1 - \delta/r & \text{if } r \ge \delta \end{cases}.$$

We then further define $J_\delta : D_{\mathbb{R}^D}[0, \infty) \to D_{\mathbb{R}^D}[0, \infty)$ (Katzenberger, 1991) as

$$J_\delta(g)(t) = \sum_{0 < s \le t} h_\delta(\|\Delta g(s)\|) \Delta g(s). \tag{19}$$

Definition A.4 (Skorohod metric on $D_{\mathcal{E}}[0, \infty)$)**.** For each finite $T > 0$ and each pair of functions $f, g \in D_{\mathcal{E}}[0, \infty)$, define $d_T(f, g)$ as the infimum of all those values of δ for which there exist grids $0 \le t_0 < t_1 < \cdots < t_m$ and $0 < s_0 < s_1 < \cdots < \cdots < s_m$, with $t_k, s_k \ge T$, such that $|t_i - s_i| \le \delta$ for $i = 0, \ldots, k$, and

$$\|f(t) - g(s)\| \le \delta \qquad \text{if } (t, s) \in [t_i, t_{i+1}) \times [s_i, s_{i+1})$$

for $i = 0, \ldots, k - 1$. The *Skorohod metric* on $D_{\mathcal{E}}[0, \infty)$ is defined to be

$$d(f, g) = \sum_{T=1}^{\infty} 2^{-T} \min\{1, d_T(f, g)\}.$$

A.2 STOCHASTIC PROCESSES AND STOCHASTIC INTEGRAL

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \ge 0}, \mathbb{P})$ be a filtered probability space.

Definition A.5 (Cross variation)**.** Let X and Y be two $\{\mathcal{F}_t\}_{t \ge 0}$-adapted stochastic processes such that X has sample paths in $D_{\mathbb{R}^{D \times e}}[0, \infty)$ and Y has samples paths in $D_{\mathbb{R}^e}[0, \infty)$, then the cross variation of X and Y on $(0, t]$, denoted by $[X, Y](t)$, is defined to be the limit of

$$\sum_{i=0}^{m-1} (X(t_{i+1}) - X(t_i))(Y(t_{i+1}) - Y(t_i))$$

in probability as the mesh size of $0 = t_0 < t_1 < \cdots < t_m = t$ goes to 0, if it exists. Moreover, for Y itself, we write

$$[Y] = \sum_{i=1}^{e} [Y^i, Y^i]$$

Definition A.6 (Martingale)**.** Let $\{X(t)\}_{t\geq 0}$ be a $\{\mathcal{F}_t\}_{t\geq 0}$-adapted stochastic process. If for all $0 \leq s \leq t$, it holds that

$$\mathbb{E}[X(t) \mid \mathcal{F}_s] = X(s),$$

then X is called a martingale.

Definition A.7 (Semimartingale)**.** Let $\{X(t)\}_{t\geq 0}$ be a $\{\mathcal{F}_t\}_{t\geq 0}$-adapted stochastic process. If there exists a sequence of $\{\mathcal{F}_t\}_{t\geq 0}$-stopping time, $\{\tau_k\}_{k\geq 0}$, such that

- $\mathbb{P}[\tau_k < \tau_{k+1}] = 1, \mathbb{P}[\lim_{k\to\infty} \tau_k = \infty] = 1,$
- and $\{X^{\tau_k}(t)\}_{t\geq 0}$ is a $\{\mathcal{F}_t\}_{t\geq 0}$-adapted martingale,

then X is called a semimartingale.

Lemma A.8 (Itô's Lemma)**.** *Let $\{X(t)\}_{t\geq 0}$ be defined through the following Itô drift-diffusion process:*

$$\mathrm{d}X(t) = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}W(t).$$

where $\{W(t)\}_{t\geq 0}$ is the standard Brownian motion. Then for any twice differentiable function f, it holds that

$$\mathrm{d}f(t, X(t)) = \left(\frac{\partial f}{\partial t} + (\nabla_x f)^\top \mu_t + \frac{1}{2}\operatorname{tr}[\sigma^\top \nabla_x^2 f \sigma]\right)\mathrm{d}t + (\nabla_x f)^\top \sigma(t)\mathrm{d}W(t).$$

A.3 Weak Convergence for Stochastic Processes

Let $(D_{\mathcal{E}}[0,\infty), \mathcal{A}, d)$ be a metric space equipped with a σ-algebra $\mathcal{A}$ and the Skorohod metric defined in the previous subsection.

Let $\{X_n\}_{n\geq 0}$ be a sequence of stochastic processes on a sequence of probability spaces $\{(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)\}_{n\geq 0}$ such that each X_n has sample paths in $D_{\mathcal{E}}[0,\infty)$. Also, let X be a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$ with sample paths on $D_{\mathcal{E}}[0,\infty)$.

Definition A.9 (Weak convergence)**.** A sequence of stochastic process$\{X_n\}_{n\geq 0}$ is said to *converge in distribution* or *weakly converge* to X (written as $X_n \Rightarrow X$) if and only if for all $\mathcal{A}$-measurable, bounded, and continuous function $f : D_{\mathcal{E}}[0,\infty) \to \mathbb{R}$, it holds that

$$\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[f(X)\right]. \tag{20}$$

Though we define weak convergence for a countable sequence of stochastic processes, but it is still valid if we index the stochastic processes by real numbers, e.g., $\{X_\eta\}_{\eta\geq 0}$, and consider the weak convergence of X_η as $\eta \to 0$. This is because the convergence in (20) is for a sequence of real numbers, which is also well-defined if we replace $\lim_{n\to\infty}$ by $\lim_{\eta\to 0}$.

Definition A.10 (δ-Prohorov distance)**.** Let $\delta > 0$. For any two probability measures P and Q on a metric space with metric d, let (X, Y) be a coupling such that P is the marginalized law of X and Q that of Y. We define

$$\rho^\delta(P, Q) = \inf\{\epsilon > 0 : \exists(X, Y), \mathbb{P}[d(X, Y) \geq \epsilon] \leq \delta\}.$$

Note this distance is not a metric because it does not satisfy triangle inequality.

Definition A.11 (Prohorov metric)**.** For any two probability measures P and Q on a metric space with metric d, let (X, Y) be a coupling such that P is the marginalized law of X and Q that of Y. Denote the marginal laws of X and Y by $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ respectively. We define the Prohorov metric as

$$\rho(P, Q) = \inf\{\epsilon > 0 : \exists(X, Y), \mathcal{L}(X) = P, \mathcal{L}(Y) = Q, \mathbb{P}[d(X, Y) \geq \epsilon] \leq \epsilon\}.$$

It can be shown that $X_n \Rightarrow X$ is equivalent to $\lim_{n\to\infty} \rho(X_n, X) = 0$.

Theorem A.12 (Skorohod Representation Theorem)**.** *Suppose $P_n, n = 1, 2, \ldots$ and P are probability measures on $\mathcal{E}$ such that $P_n \Rightarrow P$. Then there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which are defined $\mathcal{E}$-valued random variables $X_n, n = 1, 2, \ldots$ and X with distributions P_n and P respectively, such that $\lim_{n\to\infty} X_n = X$ a.s.*

B LIMITING DIFFUSION OF SGD

In this section, we give a complete derivation of the limiting diffusion of SGD. Here we use $\Rightarrow$ to denote the convergence in distribution. For any $U \subseteq \mathbb{R}^D$, we denote by $\mathring{U}$ its interior and $U^\perp$ its orthogonal complement.

First, as mentioned in Assumption 3.2, we verify that the mapping Φ is $\mathcal{C}^2$.

Lemma B.1 (Implication of Falconer (1983))**.** *Under Assumption 3.2, Φ is $\mathcal{C}^2$ on U.*

Proof of Lemma B.1. Applying Theorem 5.1 of Falconer (1983) with $f(\cdot) = \phi(\cdot, 1)$ suffices. □

Then we check the necessary conditions for applying the results in Katzenberger (1991) in Appendix B.1 and recap the corresponding theorem for the asymptotically continuous case in Appendix B.2. Finally, we provide a user-friendly interface for Katzenberger's theorem in Appendix B.3.

B.1 NECESSARY CONDITIONS

Below we collect the necessary conditions imposed on $\{Z_n\}_{n\geq 1}$ and $\{A_n\}_{n\geq 1}$ in Katzenberger (1991).

Condition B.2. The integrator sequence $\{A_n\}_{n\geq 1}$ is *asymptotically continuous*: $\sup_{t>0} |A_n(t) - A_n(t-)| \Rightarrow 0$ where $A_n(t-) = \lim_{s\to t-} A_n(s)$ is the left limit of A_n at t.

Condition B.3. The integrator sequence $\{A_n\}_{n\geq 1}$ *increases infinitely fast*: $\forall \epsilon > 0$, $\inf_{t\geq 0}(A_n(t+\epsilon) - A_n(t)) \Rightarrow \infty$.

Condition B.4 (Eq.(5.1), Katzenberger 1991)**.** For every $T > 0$, as $n \to \infty$, it holds that

$$\sup_{0<t\leq T\wedge\lambda_n(K)} \|\Delta Z_n(t)\|_2 \Rightarrow 0.$$

Condition B.5 (Condition 4.2, Katzenberger 1991)**.** For each $n \geq 1$, let Y_n be a $\{\mathcal{F}_t^n\}$-semimartingale with sample paths in $D_{\mathbb{R}^D}[0, \infty)$. Assume that for some $\delta > 0$ (allowing $\delta = \infty$) and every $n \geq 1$ there exist stopping times $\{\tau_n^m \mid m \geq 1\}$ and a decomposition of $Y_n - J_\delta(Y_n)$ into a local martingale M_n plus a finite variation process F_n such that $\mathbb{P}[\tau_n^m \leq m] \leq 1/m$, $\{[M_n](t \wedge \tau_n^m) + T_{t\wedge\tau_n^m}(F_n)\}_{n\geq 1}$ is uniformly integrable for every $t \geq 0$ and $m \geq 1$, and

$$\lim_{\gamma\to 0} \limsup_{n\to\infty} \mathbb{P}\left[\sup_{0\leq t\leq T} (T_{t+\gamma}(F_n) - T_t(F_n)) > \epsilon\right] = 0,$$

for every $\epsilon > 0$ and $T > 0$, where $T_t(\cdot)$ denotes total variation on the interval $[0, t]$.

Lemma B.6. *For SGD iterates defined using the notation in Lemma 4.2, the sequences $\{A_n\}_{n\geq 1}$ and $\{Z_n\}_{n\geq 1}$ satisfy Condition B.2, B.3, B.4 and B.5.*

Proof of Lemma B.6. Condition B.2 is obvious from the definition of $\{A_n\}_{n\geq 1}$.

Next, for any $\epsilon > 0$ and $t \in [0, T]$, we have

$$A_n(t+\epsilon) - A_n(t) = \eta_n \cdot \left\lfloor \frac{t+\epsilon}{\eta_n^2} \right\rfloor - \eta_n \cdot \left\lfloor \frac{t}{\eta_n^2} \right\rfloor \geq \frac{t+\epsilon-\eta_n^2}{\eta_n} - \frac{t}{\eta_n} = \frac{\epsilon-\eta_n^2}{\eta_n},$$

which implies that $\inf_{0\leq t\leq T}(A_n(t+\epsilon) - A_n(t)) > \epsilon/(2\eta_n)$ for small enough η_n. Then taking $n \to \infty$ yields the Condition B.3.

For Condition B.4, note that

$$\Delta Z_n(t) = \begin{cases} \eta_n(\mathbb{1}_{\xi_k} - \frac{1}{|\Xi|}\mathbb{1}) & \text{if } t = k \cdot \eta_n^2, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, we have $\|\Delta Z_n(t)\|_2 \le 2\eta_n$ for all $t > 0$. This implies that $\|\Delta Z_n(t)\|_2 \to 0$ uniformly over $t > 0$ as $n \to \infty$, which verifies Condition B.4.

We proceed to verify Condition B.5. By the definition of Z_n, we know that $\{Z_n(t)\}_{t\ge 0}$ is a jump process with independent increments and thus is a martingale. Therefore, by decomposing $Z_n = M_n + F_n$ with M_n being a local martingale and F_n a finite variation process, we must have $F_n = 0$ and M_n is Z_n itself. It then suffices to show that $[M_n](t \wedge \tau_n^m)$ is uniformly integrable for every $t \ge 0$ and $m \ge 1$. Since M_n is a pure jump process, we have

$$\begin{aligned}[M_n](t \wedge \tau_n^m) &= \sum_{0<s\le t\wedge\tau_n^m} \|\Delta M_n(s)\|_2^2 \le \sum_{0<s\le t} \|\Delta M_n(s)\|_2^2 \\ &= \sum_{k=1}^{\lfloor t/\eta_n^2 \rfloor} \left\| \eta_n \mathbb{1}_{\xi_k} - \frac{\eta_n}{|\Xi|}\mathbb{1} \right\|_2^2 \le 4 \sum_{k=1}^{\lfloor t/\eta_n^2 \rfloor} \eta_n^2 \le |\Xi| t.\end{aligned}$$

This implies that $[M_\eta](t \wedge \tau_\eta^m)$ is universally bounded by $4t$, and thus $[M_\eta](t \wedge \tau_\eta^m)$ is uniformly integrable. This completes the proof. □

Proof of Lemma 4.2. For any $n \ge 1$, it suffices to show that given $X_n(k\eta_n^2) = x_{\eta_n}(k)$, we further have $X_n((k+1)\eta_n^2) = x_{\eta_n}(k+1)$. By the definition of $X_n(t)$, we have

$$\begin{aligned}&X_n((k+1)\eta_n^2) - X_n(k\eta_n^2) \\ =& -\int_{k\eta_n^2}^{(k+1)\eta_n^2} \nabla L(X_n(t))\mathrm{d}A_n(t) + \int_{k\eta_n^2}^{(k+1)\eta_n^2} \sigma(X_n(t))\mathrm{d}Z_n(t) \\ =& -\nabla L(X_n(k\eta_n^2))(A_n((k+1)\eta_n^2) - A_n(k\eta_n^2)) + \sigma(X_n(k\eta_n^2))(Z_n((k+1)\eta_n^2) - Z_n(k\eta_n^2)) \\ =& -\eta_n \nabla L(X_n(k\eta_n^2)) + \eta_n \epsilon_{\xi_k}(X_n(k\eta_n^2)) \\ =& -\eta_n \nabla L(x_{\eta_n}(k)) + \eta_n \epsilon_{\xi_k}(x_{\eta_n}(k)) = x_{\eta_n}(k+1) - x_{\eta_n}(k)\end{aligned}$$

where the second equality is because $A_n(t)$ and $Z_n(t)$ are constant on interval $[k\eta_n^2, (k+1)\eta_n^2)$. This confirms the alignment between $\{X_n(k\eta_n^2)\}_{k\ge 1}$ and $\{x_{\eta_n}(k)\}_{k\ge 1}$.

For the second claim, note that $\sigma(x)\mathbb{E}Z_n(t) \equiv 0$ for all $x \in \mathbb{R}^D, t \ge 0$ (since the noise has zero-expectation) and that $\{Z_n(t) - \mathbb{E}Z_n(t)\}_{t\ge 0}$ will converge in distribution to a Brownian motion by the classic functional central limit theorem (see, for example, Theorem 4.3.5 in Whitt (2002)). Thus, the limiting diffusion of X_n as $n \to \infty$ can be obtained by substituting Z with the standard Brownian motion W in (22). This completes the proof. □

B.2 Katzenberger's Theorem for Asymptotically Continuous Case

The full Katzenberger's theorem deals with a more general case, which only requires the sequence of intergrators to be *asymptotically continuous*, thus including SDE (3) and SGD (1) with η goes to 0.

To describe the results in Katzenberger (1991), we first introduce some definitions. For each $n \ge 1$, let $(\Omega^n, \mathcal{F}^n, \{\mathcal{F}_t^n\}_{t\ge 0}, \mathbb{P})$ be a filtered probability space, Z_n an $\mathbb{R}^e$-valued cadlag $\{\mathcal{F}_t^n\}$-semimartingale with $Z_n(0) = 0$ and A_n a real-valued cadlag $\{\mathcal{F}_t^n\}$-adapted nondecreasing process with $A_n(0) = 0$. Let $\sigma_n : U \to \mathbb{M}(D, e)$ be continuous with $\sigma_n \to \sigma$ uniformly on compact subsets of U. Let X_n be an $\mathbb{R}^D$-valued cadlag $\{\mathcal{F}_t^n\}$-semimartingale satisfying, for all compact $K \subset U$,

$$X_n(t) = X(0) + \int_0^t \sigma(X_n)\mathrm{d}Z_n + \int_0^t -\nabla L(X_n)\mathrm{d}A_n \tag{21}$$

for all $t \le \lambda_n(K)$ where $\lambda_n(K) = \inf\{t \ge 0 \mid X_n(t-) \notin \mathring{K} \text{ or } X_n(t) \notin \mathring{K}\}$ is the stopping time of X_n leaving K.

Theorem B.7 (Theorem 6.3, Katzenberger 1991)**.** *Suppose $X(0) \in U$, Assumptions 3.1 and 3.2, Condition B.2, B.3, B.4 and B.5 hold. For any compact $K \subset U$, define $\mu_n(K) = \inf\{t \geq 0 \mid Y_n(t-) \notin \mathring{K} \text{ or } Y_n(t) \notin \mathring{K}\}$, then the sequence $\{(Y_n^{\mu_n(K)}, Z_n^{\mu_n(K)}, \mu_n(K)\}$ is relatively compact in $D_{\mathbb{R}^{D\times e}}[0,\infty) \times [0,\infty)$. If (Y, Z, μ) is a limit point of this sequence, then (Y, Z) is a continuous semimartingale, $Y(t) \in \Gamma$ for every $t \geq 0$ a.s., $\mu \geq \inf\{t \geq 0 \mid Y(t) \notin \mathring{K}\}$ a.s. and $Y(t)$ admits*

$$Y(t) = Y(0) + \int_0^{t\wedge\mu} \partial\Phi(Y(s))\sigma(Y(s))\mathrm{d}Z(s) + \frac{1}{2}\sum_{i,j=1}^{D}\sum_{k,l=1}^{e}\int_0^{t\wedge\mu} \partial_{i,j}\Phi(Y(s))\sigma(Y(s))^{ik}\sigma(Y(s))^{jl}\mathrm{d}[Z^k, Z^l](s). \tag{22}$$

B.3 A User-friendly Interface for Katzenberger's Theorem

Based on the Lemma B.6, we can immediately apply Theorem B.7 to obtain the following limiting diffusion of SGD.

Theorem B.8. *Let the manifold Γ and its open neighborhood U be as defined in (15). Let $K \subset U$ be any compact set and fix some $x_0 \in K$. Consider the SGD formulated in Lemma 4.2 where $X_n(0) \equiv x_0$. Define*

$$Y_n(t) = X_n(t) - \phi(X_n(0), A_n(t)) + \Phi(X_n(0))$$

and $\mu_n(K) = \min\{t \in \mathbb{N} \mid Y_n(t) \notin \mathring{K}\}$. Then the sequence $\{(Y_n^{\mu_n(K)}, Z_n, \mu_n(K))\}_{n\geq 1}$ is relatively compact in $D_{\mathbb{R}^{D}\times\mathbb{R}^n}[0,\infty) \times [0,\infty]$. Moreover, if (Y, Z, μ) is a limit point of this sequence, it holds that $Y(t) \in \Gamma$ a.s for all $t \geq 0$, $\mu \geq \inf\{t \geq 0 \mid Y(t) \notin \mathring{K}\}$ and $Y(t)$ admits

$$\mathrm{d}Y(t) = \partial\Phi(Y(t))\sigma(Y(t))\mathrm{d}W(t) + \frac{1}{2}\sum_{i,j}\partial_{ij}\Phi(Y(t))(\sigma(Y(t))\sigma(Y(t))^\top)^{ij}\mathrm{d}t \tag{23}$$

where $\{W(t)\}_{t\geq 0}$ is the standard Brownian motion and $\sigma(\cdot)$ is as defined in Lemma 4.2.

However, the above theorem is hard to parse and cannot be directly applied if we want to further study the implicit bias of SGD through this limiting diffusion. Therefore, we develop a user-friendly interface to it in below. In particular, Theorem 4.6 is the a special case of Theorem B.9. In Theorem 4.6, we replace $\partial\Phi(Y(t))\sigma(Y(t))$ to $\Sigma_{\|}^{\frac{1}{2}}(Y(t))$ to simplify the equation, since $\partial\Phi(Y(t))\sigma(Y(t))\left(\partial\Phi(Y(t))\sigma(Y(t))\right)^\top = \Sigma_{\|}(Y(t))$ and thus this change doesn't affect the distribution of the sample paths of the solution.

Since $\sigma(\cdot)$ is locally Lipschitz, when restricted on any compact $K \subset U$, the solution to (23) always exists and is unique. Let $\mu(K) = \inf\{t \geq 0 \mid Y(t) \notin \mathring{K}\}$ be the escaping time of the limiting diffusion Y.

Theorem B.9. *Under the same setting as B.8, we change the integer index back to $\eta > 0$ with a slight abuse of notation. For any compact set $K \subseteq U$ and $T > 0$, let $\delta = \mathbb{P}(\mu(K) \leq T)$. Then for any $\epsilon > 0$, it holds for all sufficiently small LR η that:*

$$\rho^{2\delta}(Y_\eta^{\mu_\eta(K)\wedge T}, Y^{\mu(K)\wedge T}) \leq \epsilon. \tag{24}$$

Moreover, when Y is a global solution of limiting diffusion Equation (23) and Y never leaves U, i.e. $\mathbb{P}[\forall t \geq 0, Y(t) \in U] = 1$, it holds that Y_η^T converges in distribution to Y^T as $\eta \to 0$ for any fixed $T > 0$.

Proof of the first claim of Theorem B.9. Let $\mathcal{E}_T$ be the event such that $\mu(K) > T$ on $\mathcal{E}_T$. Then restricted on $\mathcal{E}_T$, we have $Y(T \wedge \mu) = Y(T \wedge \mu(K))$ as $\mu \geq \mu(K)$ holds a.s. We first prove the claim for any convergent subsequence of η.

Now, let $\{\eta_m\}_{m\geq 1}$ be a sequence of LRs such that $\eta_m \to 0$ and $Y_{\eta_m}^{\mu_{\eta_m}(K)} \Rightarrow Y^\mu$ as $m \to \infty$. By applying the Skorohod representation theorem, we can put $\{Y_{\eta_m}\}_{m\geq 1}$ and Y under the same probability space and $Y_{\eta_m}^{\mu_{\eta_m}(K)} \to Y^\mu$ a.s. in the Skorohod metric, i.e.,

$$d(Y_{\eta_m}^{\mu_{\eta_m}(K)}, Y^\mu) \to 0, a.s.,$$

which further implies that for any $\epsilon > 0$, there exists some $N > 0$ such that for all $m > N$

$$\mathbb{P}\left[d(Y_{\eta_m}^{\mu_{\eta_m}(K)\wedge T}, Y^{\mu\wedge T}) \geq \epsilon\right] \leq \delta.$$

Restricted on $\mathcal{E}_T$, we have $d(Y_{\eta_m}^{\mu_{\eta_m}(K)\wedge T}, Y^{\mu\wedge T}) = d(Y_{\eta_m}^{\mu_{\eta_m}(K)\wedge T}, Y^{\mu(K)\wedge T})$, and it follows that for all $m > N$

$$\begin{aligned}
\mathbb{P}\left[d(Y_{\eta_m}^{\mu_{\eta_m}(K)\wedge T}, Y^{\mu(K)\wedge T}) \geq \epsilon\right] &\leq \mathbb{P}\left[\{d(Y_{\eta_m}^{\mu_{\eta_m}(K)\wedge T}, Y^{\mu(K)\wedge T}) \geq \epsilon\} \cap \mathcal{E}_T\right] + \mathbb{P}\left[\mathcal{E}_T^c\right] \\
&= \mathbb{P}\left[\{d(Y_{\eta_m}^{\mu_{\eta_m}(K)\wedge T}, Y^{\mu\wedge T}) \geq \epsilon\} \cap \mathcal{E}_T\right] + \mathbb{P}[\mathcal{E}_T^c] \\
&\leq \mathbb{P}\left[d(Y_{\eta_m}^{\mu_{\eta_m}(K)\wedge T}, Y^{\mu\wedge T}) \geq \epsilon\right] + \mathbb{P}[\mathcal{E}_T^c] \\
&\leq 2\delta
\end{aligned}$$

By the definition of the Prohorov metric in Definition A.11, we then get $\rho^{2\delta}(Y_{\eta_m}^{\mu_{\eta_m}(K)}, Y^{\mu(K)\wedge T}) \leq \epsilon$ for all $m > N$. Therefore, we have

$$\lim_{m\to\infty} \rho^{2\delta}(Y_{\eta_m}^{\mu_{\eta_m}(K)}, Y^{\mu(K)\wedge T}) = 0.$$

Now we claim that it indeed holds that $\lim_{\eta\to 0} \rho^{2\delta}(Y_{\eta}^{\mu_{\eta}(K)}, Y^{\mu(K)\wedge T}) = 0$. We prove this by contradiction. Suppose otherwise, then there exists some $\epsilon > 0$ such that for all $\eta_0 > 0$, there exists some $\eta < \eta_0$ with $\rho^{2\delta}(Y_{\eta}^{\mu_{\eta}(K)}, Y^{\mu(K)\wedge T}) > \epsilon$. Consequently, there is a sequence $\{\eta_m\}_{m\geq 1}$ satisfying $\lim_{m\to\infty} \eta_m = 0$ and $\rho^{2\delta}(Y_{\eta_m}^{\mu_{\eta_m}(K)}, Y^{\mu(K)\wedge T}) > \epsilon$ for all m. Since $\{(Y_{\eta_m}^{\mu_{\eta_m}(K)}, Z_{\eta_m}, \mu_{\eta_m}(K))\}_{m\geq 1}$ is relatively compact, there exists a subsequence (WLOG, assume it is the original sequence itself) converging to $(Y^{\mu\wedge T}, W, \mu)$ in distribution. However, repeat the exactly same argument as above, we would have $\rho^{2\delta}(Y_{\eta_m}^{\mu_{\eta_m}(K)}, Y^{\mu(K)\wedge T}) \leq \epsilon$ for all sufficiently large m, which is a contradiction. This completes the proof. □

Proof of the second claim of Theorem B.9. We will first show there exists a sequence of compact set $\{K_m\}_{m\geq 1}$ such that $\cup_{m=1}^{\infty} K_m = U$ and $K_m \subseteq K_{m+1}$. For $m \in \mathbb{N}^+$, we define $H_m = U \setminus (B_{1/m}(0) + \mathbb{R}^D \setminus U)$ and $K_m = \overline{H_m} \cap B_m(0)$. By definition it holds that $\forall m < m'$, $H_m \subseteq H_{m'}$ and $K_m \subseteq K_{m'}$. Moreover, since K_m is bounded and closed, K_m is compact for every m. Now we claim $\cup_{m=1}^{\infty} K_m = U$. Note that $\cup_{m=1}^{\infty} K_m = \cup_{m=1}^{\infty} \overline{H_m} \cap B_m(0) = \cup_{m=1}^{\infty} \overline{H_m}$. $\forall x \in U$, since U is open, we know $d(x, \mathbb{R}^D \setminus U) > 0$, thus there exists $m_0 \in \mathbb{N}^+$, such that $\forall m \geq m_0$, $x \notin (B_{1/m}(0) + \mathbb{R}^D \setminus U)$ and thus $x \in H_m$, which implies $x \in \cup_{m=1}^{\infty} \overline{H_m}$. On the other hand, $\forall x \in \mathbb{R}^D \setminus U$, it holds that $x \in (B_{1/m}(0) + \mathbb{R}^D \setminus U)$ for all $m \in \mathbb{N}^+$, thus $x \notin H_m \subset K_m$.

Therefore, since $Y \in U$ and is continuous almost surely, random variables $\lim_{m\to\infty} \mu(K_m) = \infty$ a.s., which implies $\mu(K_m)$ converges to ∞ in distribution, i,e,, $\forall \delta > 0, T > 0$, $\exists m \in \mathbb{N}^+$, such that $\forall K \supseteq K_m$, it holds $\mathbb{P}[\mu(K) \leq T] \leq \delta$.

Now we will show $\forall T > 0, \epsilon > 0$, there exists η_0 such that $\rho^{\epsilon}(Y^T, Y_{\eta}^T) \leq \epsilon$ for all $\eta \leq \eta_0$. For any fixed T, for all $\epsilon > 0$, let $\delta = \frac{\epsilon}{4}$, from above we know exists compact set K, such that $\mathbb{P}(\mu(K) \leq T) \leq \delta$. We further pick $K' = K + B_{2\epsilon'}(0)$, where ϵ' can be any real number satisfying $0 < \epsilon' < \epsilon$ and $K' \subseteq U$. Such ϵ' exists since U is open. Note $K \subseteq K'$, we have $\mathbb{P}(\mu(K') \leq T) \leq \mathbb{P}(\mu(K) \leq T) \leq \delta$. Thus by the first claim of Theorem B.9, there exists η_0, such that for all $\eta \leq \eta_0$, we have $\rho^{2\delta}(Y_{\eta}^{\mu_{\eta}(K')\wedge T}, Y^{\mu(K')\wedge T}) \leq \epsilon'$.

Note that $\rho^{\delta}(Y^{\mu(K)\wedge T}, Y^{\mu(K')\wedge T}) = 0$, we have for all $\eta \leq \eta_0$,

$$\rho^{3\delta}(Y^{\mu(K)\wedge T}, Y_{\eta}^{\mu_{\eta}(K')\wedge T}) \leq \epsilon'. \tag{25}$$

By the definition of δ-Prohorov distance, we can assume $Y^{\mu(K)\wedge T}, Y_{\eta}^{\mu_{\eta}(K')\wedge T}$ are already the coupling such that $\mathbb{P}[d(Y^{\mu(K)\wedge T}, Y_{\eta}^{\mu_{\eta}(K')\wedge T}) \geq \epsilon'] \leq 3\delta$. Note that for all $t \geq 0$, $Y^{\mu(K)\wedge T}(t) \in K$,

we know if $\mu_\eta(K') \le T$, then

$$\begin{aligned} d(Y^{\mu(K)\wedge T}, Y_\eta^{\mu_\eta(K')\wedge T}) &\ge \left\| Y^{\mu(K)\wedge T}(\mu_\eta(K')) - Y_\eta^{\mu_\eta(K')\wedge T}(\mu_\eta(K') \right\|_2 \\ &\ge d(K, \mathbb{R}^d/K') \\ &\ge \epsilon'. \end{aligned}$$

Thus we conclude $d(Y^{\mu(K)\wedge T}, Y_\eta^T) \ge 2\epsilon' \implies d(Y^{\mu(K)\wedge T}, Y_\eta^{\mu_\eta(K')\wedge T}) \ge 2\epsilon'$, which further implies that

$$\rho^{3\delta}(Y^{\mu(K)\wedge T}, Y_\eta^T) \le \epsilon'. \tag{26}$$

Note that $\rho^\delta(Y^T, Y^{\mu(K)\wedge T}) = 0$, we have for all $\eta \le \eta_0$,

$$\rho^\epsilon(Y^T, Y_\eta^T) = \rho^{4\delta}(Y^T, Y_\eta^T) \le \rho^{3\delta}(Y^{\mu(K)\wedge T}, Y_\eta^T) + \rho^\delta(Y^T, Y^{\mu(K)\wedge T}) \le \epsilon' \le \epsilon,$$

which completes the proof. □

C Explicit Formula of the Limiting Diffusion

In this section, we demonstrate how to compute the derivatives of Φ by relating to those of the loss function L, and then present the explicit formula of the limiting diffusion.

C.1 Explicit Expression of the Derivatives

For any $x \in \Gamma$, we choose an orthonormal basis of $T_x(\Gamma)$ as $\{v_1, \dots, v_{D-M}\}$. Let $\{v_{D-M+1}, \dots, v_D\}$ be an orthonormal basis of $T_x^\perp(\Gamma)$ so that $\{v_i\}_{i\in[D]}$ is an orthonormal basis of $\mathbb{R}^D$.

Lemma C.1. *For any $x \in \Gamma$ and any $v \in T_x(\Gamma)$, it holds that $\nabla^2 L(x)v = 0$.*

Proof. For any $x \in T_x(\Gamma)$, let $\{x(t)\}_{t\ge 0}$ be a parametrized smooth curve on Γ such that $x(0) = x$ and $\frac{\mathrm{d}x(t)}{\mathrm{d}t}\big|_{t=0} = v$. Then $\nabla L(x_t) = 0$ for all t. Thus $0 = \frac{\mathrm{d}\nabla L(x_t)}{\mathrm{d}t}\big|_{t=0} = \nabla^2 L(x)v$. □

Lemma C.2. *For any $x \in \mathbb{R}^D$, it holds that $\partial\Phi(x)\nabla L(x) = 0$ and*

$$\partial^2\Phi(x)[\nabla L(x), \nabla L(x)] = -\partial\Phi(x)\nabla^2 L(x)\nabla L(x).$$

Proof. Fixing any $x \in \mathbb{R}^D$, let $\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\nabla L(x(t))$ be initialized at $x(0) = x$. Since $\Phi(x(t)) = \Phi(x)$ for all $t \ge 0$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(x(t)) = -\partial\Phi(x(t))\nabla L(x(t)) = 0.$$

Evaluating the above equation at $t = 0$ yields $\partial\Phi(x)\nabla L(x) = 0$. Moreover, take the second order derivative and we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Phi(x_t) = -\partial^2\Phi(x(t))\left[\frac{\mathrm{d}x(t)}{\mathrm{d}t}, \nabla L(x(t))\right] + \partial\Phi(x(t))\nabla^2 L(x(t))\frac{\mathrm{d}x(t)}{\mathrm{d}t} = 0.$$

Evaluating at $t = 0$ completes the proof. □

Now we can prove Lemma 4.3, restated in below.

Lemma 4.3. *For any $x \in \Gamma$, $\partial\Phi(x) \in \mathbb{R}^{D\times D}$ is the projection matrix onto tangent space $T_x(\Gamma)$.*

Proof of Lemma 4.3. For any $v \in T_x(\Gamma)$, let $\{v(t), t \ge 0\}$ be a parametrized smooth curve on Γ such that $v(0) = x$ and $\frac{\mathrm{d}v(t)}{\mathrm{d}t}\big|_{t=0} = v$. Since $v(t) \in \Gamma$ for all $t \ge 0$, we have $\Phi(v(t)) = v(t)$, and thus

$$\left.\frac{\mathrm{d}v(t)}{\mathrm{d}t}\right|_{t=0} = \left.\frac{\mathrm{d}}{\mathrm{d}t}\Phi(v(t))\right|_{t=0} = \partial\Phi(x)\left.\frac{\mathrm{d}v(t)}{\mathrm{d}t}\right|_{t=0}.$$

This implies that $\partial\Phi(x)v = v$ for all $v \in T_x(\Gamma)$.

Next, for any $u \in T_x^\perp(\Gamma)$ and $t \geq 0$, consider expanding $\nabla L(x + t\nabla^2 L(x)^\dagger u)$ at $t = 0$:

$$\begin{aligned}\nabla L\left(x + t\nabla^2 L(x)^\dagger u\right) &= \nabla^2 L(x) \cdot t\nabla^2 L(x)^\dagger u + o(t)\\ &= tu + o(t)\end{aligned}$$

where the second equality follows from the assumption that $\nabla^2 L(x)$ is full-rank when restricted on $T_x^\perp(\Gamma)$. Then since $\partial\Phi$ is continuous, it follows that

$$\begin{aligned}\lim_{t\to 0} \frac{\partial\Phi(x + t\nabla^2 L(x)^\dagger u)\nabla L(x + t\nabla^2 L(x)^\dagger u)}{t} &= \lim_{t\to 0} \partial\Phi(x + t\nabla^2 L(x)^\dagger)(u + o(1))\\ &= \partial\Phi(x)u.\end{aligned}$$

By Lemma C.2, we have $\partial\Phi(x + t(\nabla^2 L(x))^\dagger u))\nabla L(x + t(\nabla^2 L(x))^\dagger u) = 0$ for all $t > 0$, which then implies that $\partial\Phi(x)u = 0$ for all $u \in T_x^\perp(\Gamma)$.

Therefore, under the basis $\{v_i, \ldots, v_N\}$, $\partial\Phi(x)$ is given by

$$\partial\Phi(x) = \begin{pmatrix} I_{D-M} & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{D\times D},$$

that is, the projection matrix onto $T_x(\Gamma)$. □

Lemma C.3. *For any $x \in \Gamma$, it holds that $\partial\Phi(x)\nabla^2 L(x) = 0$.*

Proof. It directly follows from Lemma C.1 and Lemma 4.3. □

Next, we proceed to compute the second-order derivatives.

Lemma C.4. *For any $x \in \Gamma$, $u \in \mathbb{R}^D$ and $v \in T_x(\Gamma)$, it holds that*

$$\partial^2\Phi(x)[v, u] = -\partial\Phi(x)\partial^2(\nabla L)(x)[v, \nabla^2 L(x)^\dagger u] - \nabla^2 L(x)^\dagger \partial^2(\nabla L)(x)[v, \partial\Phi(x)u].$$

Proof of Lemma C.4. Consider a parametrized smooth curve $\{v(t)\}_{t\geq 0}$ on Γ such that $v(0) = x$ and $\frac{\mathrm{d}v(t)}{\mathrm{d}t}\big|_{t=0} = v$. We define $P(t) = \partial\Phi(v(t))$, $P^\perp(t) = I_D - P(t)$ and $H(t) = \nabla^2 L(v(t))$ for all $t \geq 0$. By Lemma C.1 and 4.3, we have

$$P^\perp(t)H(t) = H(t)P^\perp(t) = H(t), \tag{27}$$

Denote the derivative of $P(t)$, $P^\perp(t)$ and $H(t)$ with respect to t as $P'(t)$, $(P^\perp)'(t)$ and $H'(t)$. Then differentiating with respect to t, we have

$$(P^\perp)'(t)H(t) = H'(t) - P^\perp(t)H'(t) = P(t)H'(t). \tag{28}$$

Then combining (27) and (28) and evaluating at $t = 0$, we have

$$P'(0)H(0) = -(P^\perp)'(0)H(0) = -P(0)H'(0) \tag{29}$$

We can decompose $P'(0)$ and $H(0)$ as follows

$$P'(0) = \begin{pmatrix} P'_{11}(0) & P'_{12}(0) \\ P'_{21}(0) & P'_{22}(0) \end{pmatrix}, \quad H(0) = \begin{pmatrix} 0 & 0 \\ 0 & H_{22}(0) \end{pmatrix}, \tag{30}$$

where $P'_{11}(0) \in \mathbb{R}^{(D-M)\times(D-M)}$ and H_{22} is the hessian of L restricted on $T_x^\perp(\Gamma)$. Also note that

$$\begin{aligned}P(0)H'(0)P^\perp(0) &= \begin{pmatrix} I_{D-M} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} H'_{11}(0) & H'_{12}(0) \\ H'_{21}(0) & H'_{22}(0) \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0 & I_M \end{pmatrix}\\ &= \begin{pmatrix} 0 & H'_{12}(0) \\ 0 & 0 \end{pmatrix},\end{aligned}$$

and thus by (30) we have

$$P'(0)H(0) = \begin{pmatrix} 0 & P'_{12}(0)H_{22}(0) \\ 0 & P'_{22}(0)H_{22}(0) \end{pmatrix} = \begin{pmatrix} 0 & -H'_{12}(0) \\ 0 & 0 \end{pmatrix}.$$

This implies that we must have $P'_{22}(0) = 0$ and $P'_{12}(0)H_{22}(0) = H'_{12}(0)$. Similarly, by taking transpose in (30), we also have $H_{22}(0)P'_{21}(0) = -H'_{21}(0)$.

It then remains to determine the value of $P'_{11}(0)$. Note that since $P(t)P(t) = P(t)$, we have $P'(t)P(t) + P(t)P'(t) = P'(t)$, evaluating at $t = 0$ yields

$$2P'_{11}(0) = P'_{11}(0).$$

Therefore, we must have $P'_{11}(0) = 0$. Combining the above results, we obtain

$$P'(0) = -P(0)H'(0)H(0)^\dagger - H(0)^\dagger H'(0)P(0).$$

Finally, recall that $P(t) = \partial\Phi(v(t))$, and thus

$$P'(0) = \frac{\mathrm{d}}{\mathrm{d}t}\partial\Phi(v(t))\bigg|_{t=0} = \partial^2\Phi(x)[v].$$

Similarly, we have $H'(0) = \partial^2(\nabla L)(x)[v]$, and it follows that

$$\partial^2\Phi(x)[v] = -\partial\Phi(x)\partial^2(\nabla L)(x)[v]\nabla^2 L(x)^\dagger - \nabla^2 L(x)^\dagger\partial^2(\nabla L)(x)[v]\partial\Phi(x).$$

□

Lemma C.5. *For any $x \in \Gamma$ and $u \in T_x^\perp(\Gamma)$, it holds that*

$$\partial^2\Phi(x)[uu^\top + \nabla^2 L(x)^\dagger uu^\top \nabla^2 L(x)] = -\partial\Phi(x)\partial^2(\nabla L)(x)[\nabla^2 L(x)^\dagger uu^\top].$$

Proof of Lemma C.5. For any $u \in T_x^\perp(\Gamma)$, we define $u(t) = x + t\nabla^2 L(x)^\dagger u$ for $t \geq 0$. By Taylor approximation, we have

$$\nabla L(u(t)) = t\nabla^2 L(x)\nabla^2 L(x)^\dagger u + o(t) = tu + o(t) \tag{31}$$

and

$$\nabla^2 L(u(t)) = \nabla^2 L(x) + t\partial^2(\nabla L)(x)[\nabla^2 L(x)^\dagger u] + o(t). \tag{32}$$

Combine (31) and (32) and apply Lemma C.2, and it follows that

$$\begin{aligned}
0 &= \partial^2\Phi(u(t))[\nabla L(u(t)), \nabla L(u(t))] + \partial\Phi(u(t))\nabla^2 L(u(t))\nabla L(u(t)) \\
&= t^2\partial^2\Phi(u(t))[u + o(1)](u + o(1)) + t^2\partial\Phi(u(t))\partial^2(\nabla L)(x)[\nabla^2 L(x)^\dagger u](u + o(1)) \\
&\quad + t^2\frac{\partial\Phi(u(t))}{t}\nabla^2 L(x)(u + o(1)) \\
&= t^2\partial^2\Phi(u(t))[u + o(1)](u + o(1)) + t^2\partial\Phi(u(t))\partial^2(\nabla L)(x)[\nabla^2 L(x)^\dagger u](u + o(1)) \\
&\quad + t^2\frac{\partial\Phi(u(t)) - \partial\Phi(x)}{t}\nabla^2 L(x)(u + o(1))
\end{aligned}$$

where the last equality follows from Lemma C.3. Dividing both sides by t^2 and letting $t \to 0$, we get

$$\partial^2\Phi(x)[u]u + \partial\Phi(x)\partial^2(\nabla L)(x)[\nabla^2 L(x)^\dagger u]u + \partial^2\Phi(x)[\nabla^2 L(x)^\dagger u]\nabla^2 L(x)u = 0.$$

Rearranging the above equation completes the proof. □

With the notion of Lyapunov Operator in Definition 4.4, Lemma C.5 can be further simplified into Lemma C.6.

Lemma C.6. *For any $x \in \Gamma$ and $\Sigma \in \mathrm{span}\{uu^\top \mid u \in T_x^\perp(\Gamma)\}$,*

$$\langle\partial^2\Phi(x), \Sigma\rangle = -\partial\Phi(x)\partial^2(\nabla L)(x)[\mathcal{L}^{-1}_{\nabla^2 L(x)}(\Sigma)]. \tag{33}$$

Proof of Lemma C.6. Let $A = uu^\top + \nabla^2 L(x)^\dagger uu^\top \nabla^2 L(x)$ and $B = \nabla^2 L(x)^\dagger uu^\top$. The key observation is that $A + A^\top = \mathcal{L}_{\nabla^2 L(x)}(B + B^\top)$. Therefore, by Lemma C.5, it holds that

$$\partial^2\Phi(x)[\mathcal{L}_{\nabla^2 L(x)}(B+B^\top)] = \partial^2\Phi(x)[A+A^\top] = 2\partial\Phi(x)\partial^2(\nabla L)(x)[B] = \partial\Phi(x)\partial^2(\nabla L)(x)[B+B^\top].$$

Since $\nabla^2 L(x)^\dagger$ is full-rank when restricted to $T_x^\perp(\Gamma)$, we have $\mathrm{span}\{\nabla^2 L(x)^\dagger uu^\top + uu^\top \nabla^2 L(x)^\dagger \mid u \in T_x^\perp(\Gamma)\} = \mathrm{span}\{uu^\top \mid u \in T_x^\perp(\Gamma)\}$. Thus by the linearity of above equation, we can replace $B + B^\top$ by any $\Sigma \in \mathrm{span}\{uu^\top \mid u \in T_x^\perp(\Gamma)\}$, resulting in the desired equation. □

Then Lemma 4.5 directly follows from Lemma C.4 and C.5.

Lemma C.7. *For any $x \in \Gamma$, suppose there exist a neighborhood U_x of x and two loss functions L and L' that define the same manifold Γ locally in U_x, i.e., $\Gamma \cap U_x = \{x \mid \nabla L(x) = 0\} = \{x \mid \nabla L'(x) = 0\}$. Then for any $v \in T_x(\Gamma)$, it holds that $(\nabla^2 L(x))^\dagger \partial^2(\nabla L)(x)[v,v] = (\nabla^2 L'(x))^\dagger \partial^2(\nabla L')(x)[v,v]$.*

Proof of Lemma C.7. Let $\{v(t)\}_{t\geq 0}$ be a smooth curve on Γ with $v(0) = x$ and $\frac{\mathrm{d}v(t)}{\mathrm{d}t}\big|_{t=0} = v$. Since $v(t)$ stays on Γ, we have $\nabla L(v(t)) = 0$ for all $t \geq 0$. Taking derivative for two times yields $\partial^2(\nabla L)(v(t))[\frac{\mathrm{d}v(t)}{\mathrm{d}t}, \frac{\mathrm{d}v(t)}{\mathrm{d}t}] + \nabla^2 L(v(t))\frac{\mathrm{d}^2 v(t)}{\mathrm{d}t^2} = 0$. Evaluating it at $t = 0$ and multiplying both sides by $\nabla^2 L(x)^\dagger$, we get

$$\nabla^2 L(x)^\dagger \partial^2(\nabla L)(x)[v,v] = -\nabla^2 L(x)^\dagger \nabla^2 L(x) \frac{\mathrm{d}^2 v(t)}{\mathrm{d}t^2}\bigg|_{t=0} = -\partial\Phi(x)\frac{\mathrm{d}^2 v(t)}{\mathrm{d}t^2}\bigg|_{t=0}.$$

Since $\partial\Phi(x)$ is the projection matrix onto $T_x(\Gamma)$ by Lemma 4.3, it does not depend on L, so analogously we also have $\nabla^2 L'(x)^\dagger \partial^2(\nabla L')(x)[v,v] = -\partial\Phi(x)\frac{\mathrm{d}^2 v(t)}{\mathrm{d}t^2}\big|_{t=0}$ as well. The proof is thus completed. Note that $\partial\Phi(x)\frac{\mathrm{d}^2 v(t)}{\mathrm{d}t^2}\big|_{t=0}$ is indeed the second fundamental form for v at x, and the value won't change if we choose another parametric smooth curve with a different second-order time derivative. (See Chapter 6 in Do Carmo (2013) for a reference.) □

C.2 Proof of results in Section 5

Now we are ready to give the missing proofs in Section 5 which yield explicit formula of the limiting diffusion for label noise and isotropic noise.

Corollary 5.1 (Limiting Diffusion for Isotropic Noise)**.** *If $\Sigma \equiv I_D$ on Γ, SDE (10) is then*

$$\mathrm{d}Y(t) = \underbrace{\partial\Phi(Y)\mathrm{d}W + \frac{1}{2}\nabla^2 L(Y)^\dagger \partial^2(\nabla L)(Y)[\partial\Phi(Y)]\,\mathrm{d}t}_{\text{Brownian Motion on Manifold}} + \underbrace{\frac{1}{2}\partial\Phi(Y)\nabla(\ln|\nabla^2 L(Y)|_+)\mathrm{d}t}_{\text{Normal Regularization}} \quad (11)$$

where $|\nabla^2 L(Y)|_+ = \lim_{\alpha\to 0}\frac{|\nabla^2 L(Y)+\alpha I_D|}{\alpha^{D-\mathrm{rank}(\nabla^2 L(Y))}}$ is the pseudo-determinant *of $\nabla^2 L(Y)$. $|\nabla^2 L(Y)|_+$ is also equal to the sum of log of non-zero eigenvalue values of $\nabla^2 L(Y)$.*

Proof of Corollary 5.1. Set $\Sigma_\parallel = \partial\Phi$, $\Sigma_\perp = I_D - \partial\Phi$ and $\Sigma_{\perp,\parallel} = \Sigma_{\parallel,\perp} = 0$ in the decomposition of Σ by Lemma 4.5, and we need to show $\partial\Phi\nabla(\ln|\Sigma|_+) = \partial^2(\nabla L)[(\nabla^2 L)^\dagger]$.

Holbrook (2018) shows that the gradient of pseudo-inverse determinant satisfies $\nabla|A|_+ = |A|_+ A^\dagger$. Thus we have for any vector $v \in \mathbb{R}^D$, $\langle v, \nabla\ln|\nabla^2 L|_+\rangle = \left\langle \frac{|\nabla^2 L|_+ \nabla^2 L}{|\nabla^2 L|_+}, \partial^2(\nabla L)[v]\right\rangle = \langle \nabla^2 L, \partial^2(\nabla L)[v]\rangle = \partial^2(\nabla L)[v, \nabla^2 L] = \langle v, \partial^2(\nabla L)[(\nabla^2 L)^\dagger]\rangle$, which completes the proof. □

Corollary 5.2 (Limiting Flow for Label Noise)**.** *If $\Sigma \equiv c\,\mathrm{tr}[\nabla^2 L]$ on Γ for some constant $c > 0$, SDE (10) can be simplified into (13) where the regularization is from the noise in normal space.*

$$\mathrm{d}Y(t) = -\frac{1}{4}\partial\Phi(Y(t))\nabla\,\mathrm{tr}[c\nabla^2 L(Y(t))]\mathrm{d}t. \quad (13)$$

Proof of Corollary 5.2. Since $\Sigma = c\nabla^2 L$, here we have $\Sigma_\perp = \Sigma$ and $\Sigma_\parallel, \Sigma_{\perp,\parallel}, \Sigma_{\parallel,\perp} = 0$. Thus it suffices to show that $2\partial^2(\nabla L)\left[\mathcal{L}^{-1}_{\nabla^2 L}(\Sigma_\perp)\right] = \nabla\,\mathrm{tr}[\nabla^2 L]$. Note that for any $v \in \mathbb{R}^D$,

$$v^\top \nabla\,\mathrm{tr}[\nabla^2 L] = \langle I_D, \partial^2(\nabla L)[v]\rangle = \langle I_D - \partial\Phi, \partial^2(\nabla L)[v]\rangle, \quad (34)$$

where the second equality is because the the tangent space of symmetric rank-n matrices at $\nabla^2 L$ is $\{A\nabla^2 L + \nabla^2 L A^\top \mid A \in \mathbb{R}^{D\times D}\}$, and every element in this tangent space has zero inner-product with $\partial\Phi$ by Lemma 4.3. Also note that $\mathcal{L}^{-1}_{\nabla^2 L}(\nabla^2 L) = \frac{1}{2}(I_D - \partial\Phi)$, thus $\langle I_D - \partial\Phi, \partial^2(\nabla L)[v]\rangle = 2\langle \mathcal{L}^{-1}_{\nabla^2 L}(\nabla^2 L), \partial^2(\nabla L)[v]\rangle = 2v^\top \partial^2(\nabla L)[\mathcal{L}^{-1}_{\nabla^2 L}(\nabla^2 L)]$. □

D PROOF OF RESULTS IN SECTION 6

In this section, we present the missing proofs in Section 6 regarding the overparametrized linear model.

For convenience, for any $p, r \geq 0$ and $u \in \mathbb{R}^D$, we denote by $B_r^p(u)$ the ℓ_p norm ball of radius r centered at u. We also denote $v^{i:j} = (v^i, v^{i+1}, \ldots, v^j)^\top$ for $i, j \in [D]$.

D.1 PROOF OF THEOREM 6.1

In this subsection, we provide the proof of Theorem 6.1.

Theorem 6.1. *In the setting of OLM, suppose the groundtruth is κ-sparse and $n \geq \Omega(\kappa \ln d)$ training data are sampled from either i.i.d. Gaussian or Boolean distribution. Then for any initialization x_{init} (except a zero-measure set) and any $\epsilon > 0$, there exist $\eta_0, T > 0$ such that for any $\eta < \eta_0$, OLM trained with label noise SGD (12) with LR equal to η for $\lfloor T/\eta^2 \rfloor$ steps returns an ϵ-optimal solution, with probability of $1 - e^{-\Omega(n)}$ over the randomness of the training dataset.*

Proof of Theorem 6.1. First, by Lemma 6.6, it holds with probability at least $1 - e^{-\Omega(n)}$ that the solution to (18), x_*, is unique up to and satisfies $|x_*| = \psi(w_*)$. Then on this event, for any $\epsilon > 0$, by Lemma 6.5, there exists some $T > 0$ such that x_T given by the Riemannian gradient flow (17) satisfies that x_T is an $\epsilon/2$-optimal solution of the OLM. For this T, by Theorem 4.6, we know that the $\lfloor T/\eta^2 \rfloor$-th SGD iterate, $x_\eta(\lfloor T/\eta^2 \rfloor)$, satisfies $\|x_\eta(\lfloor T/\eta^2 \rfloor) - x_T\|_2 \leq \epsilon/2$ with probability at least $1 - e^{-\Omega(n)}$ for all sufficiently small $\eta > 0$, and thus $x_\eta(\lfloor T/\eta^2 \rfloor)$ is an ϵ-optimal solution of the OLM. Finally, the validity of applying Theorem 4.6 is guaranteed by Lemma 6.2 and 6.3. This completes the proof. □

In the following subsections, we provide the proofs of all the components used in the above proof.

D.2 PROOF OF LEMMA 6.2

Recall that for each $i \in [n]$ $f_i(x) = f(u, v) = z_i^\top (u^{\odot 2} - v^{\odot 2})$, $\nabla f_i(x) = 2\binom{z_i \odot u}{z_i \odot v}$, and $K(x) = (K_{ij}(x))_{i,j \in [n]}$ where each $K_{ij}(x) = \langle \nabla f_i(x), \nabla f_j(x) \rangle$. Then

$$\nabla^2 \ell_i(x) = 2 \begin{pmatrix} z_i \odot u \\ -z_i \odot v \end{pmatrix} \begin{pmatrix} (z_i \odot u)^\top & -(z_i \odot v)^\top \end{pmatrix} + (f_i(u, v) - y_i) \cdot \mathrm{diag}(z_i, z_i).$$

So for any $x \in \Gamma$, it holds that

$$\nabla^2 L(x) = \frac{2}{n} \sum_{i=1}^n \begin{pmatrix} z_i \odot u \\ -z_i \odot v \end{pmatrix} \begin{pmatrix} (z_i \odot u)^\top & -(z_i \odot v)^\top \end{pmatrix}. \tag{35}$$

Lemma D.1. *For any fixed $x \in \mathbb{R}^D$, suppose $\{\nabla f_i(x)\}_{i \in [n]}$ is linearly independent, then $K(x)$ is full-rank.*

Proof of Lemma D.1. Suppose otherwise, then there exists some $\lambda \in \mathbb{R}^n$ such that $\lambda \neq 0$ and $\lambda^\top K(x) \lambda = 0$. However, note that

$$\begin{aligned} \lambda^\top K(x) \lambda &= \sum_{i,j=1}^n \lambda^i \lambda^j K_{ij}(x) \\ &= \sum_{i,j=1}^n \lambda^i \lambda^j \langle \nabla f_i(x), \nabla f_j(x) \rangle \\ &= \left\| \sum_{i=1}^n \lambda^i \nabla f_i(x) \right\|_2^2, \end{aligned}$$

which implies that $\sum_{i=1}^n \lambda^i \nabla f_i(x) = 0$. This is a contradiction since by assumption $\{\nabla f_i(x)\}_{i \in [n]}$ is linearly independent. □

Lemma 6.2. *Consider the loss function L defined in (14) and manifold Γ defined in (15). If data is full rank, i.e., $\operatorname{rank}(\{z_i\}_{i\in[n]}) = n$, then it holds that (a). Γ is a smooth manifold of dimension $D - n$; (b). $\operatorname{rank}(\nabla^2 L(x)) = n$ for all $x \in \Gamma$. In particular, it holds that $\operatorname{rank}(\{z_i\}_{i\in[n]}) = n$ holds with probability 1 for Gaussian distribution and with probability $1 - c^d$ for Boolean distribution for some constant $c \in (0, 1)$.*

Proof of Lemma 6.2. (1) By preimage theorem (Banyaga & Hurtubise, 2013), it suffices to check the jacobian $[\nabla f_1(x), \dots, \nabla f_n(x)] = 2[\binom{z_1 \odot u}{-z_1 \odot v}, \dots, \binom{z_n \odot u}{-z_n \odot v}]$ is full rank. Similarly, for the second claim, due to (35). it is also equivalent to show that $\{\binom{z_i \odot u}{-z_i \odot v}\}_{i\in[n]}$ is of rank n.

Since $\binom{u}{v} \in \Gamma \subset U$, each coordinate is non-zero, thus we only need to show that $\{z_i\}_{i\in[n]}$ is of rank n. This happens with probability 1 in the Gaussian case, and probability at least $1 - c^d$ for some constant $c \in (0, 1)$ by Kahn et al. (1995). This completes the proof. □

D.3 Proof of Lemma 6.3

We first establish some auxiliary results. The following lemma shows the PL condition along the trajectory of gradient flow.

Lemma D.2. $\|\nabla L(x_t)\|^2 \geq \lambda_{min}(ZZ^\top) \min_{i\in[d]} |u_0^i v_0^i| L(x_t)$.

To prove Lemma D.2, we need the following invariance along the gradient flow.

Lemma D.3. *Along the gradient flow generated by $\nabla L(x_t)$, $u_t^j v_t^j$ stays constant for all $j \in [d]$. Thus, $\operatorname{sign}(u_t^j) = \operatorname{sign}(u_0^j)$ and $\operatorname{sign}(v_t^j) = \operatorname{sign}(v_0^j)$ for any $j \in [d]$.*

Proof of Lemma D.3.

$$\begin{aligned}\frac{\partial}{\partial t}(u_t^j v_t^j) &= \frac{\partial u_t^j}{\partial t} \cdot v_t^j + u_t^j \cdot \frac{\partial v_t^j}{\partial t} \\ &= \nabla_u L(u_t, v_t)^j \cdot v_t^j + u_t^j \cdot \nabla_v L(u_t, v_t)^j \\ &= \frac{1}{4}\sum_{i=1}^n (f_i(u_t, v_t) - y_i) z_i^j u_t^j v_t^j - \frac{u_t^j}{4}\sum_{i=1}^n (f_i(u_t, v_t) - y_i) z_i^j v_t^j = 0.\end{aligned}$$

Therefore, any sign change of u^j, v^j would enforce $u_t^j = 0$ or $v_t^j = 0$ for some $t > 0$ since u_t^j, v_t^j are continuous in time t. This immediately leads to a contradiction to the invariance of $u_t^j v_t^j$. □

We then can prove Lemma D.2.

Proof of Lemma D.2. Note that

$$\begin{aligned}\|\nabla L(x)\|_2^2 =& \frac{1}{4}\sum_{i,j=1}^n (f_i(x) - y_i)(f_j(x) - y_j) \langle \nabla f_i(x), \nabla f_j(x)\rangle \\ \geq& \frac{1}{4}\sum_{i=1}^n (f_i(x) - y_i)^2 \lambda_{min}(K(x)) \\ =& L(x)\lambda_{min}(K(x)),\end{aligned}$$

where $K(x)$ is a $n \times n$ p.s.d. matrix with $K_{ij}(x) = \langle \nabla f_i(x), \nabla f_j(x)\rangle$. Below we lower bound $\lambda_{min}(K(x))$, the smallest eigenvalue of $K(x)$. Note that $K_{ij}(x) = \sum_{h=1}^n z_i^h z_j^h ((u_t^h)^2 + (v_t^h)^2)$, and we have

$$K(x) = Z\operatorname{diag}(u_t^{\odot 2} + v_t^{\odot 2})Z^\top \succeq Z\operatorname{diag}(|u_t v_t|)Z^\top \overset{(*)}{=} Z\operatorname{diag}(|u_0 v_0|)Z^\top \succeq \min_{i\in[d]} |u_0^i v_0^i| ZZ^T.$$

where $(*)$ is by Lemma D.3. Thus $\lambda_{\min}(K) \geq \min_{i\in[d]} |u_0^i v_0^i| \lambda_{min}(ZZ^T)$, which completes the proof. □

We also need the following characterization of the manifold Γ.

Lemma D.4. *All the stationary points in U are global minimizers, i.e., $\Gamma = \{x \in U \mid \nabla L(x) = 0\}$.*

Proof of Lemma D.4. Since Γ is the set of local minimizers, each x in Γ must satisfy $\nabla L(x) = 0$. The other direction is proved by noting that $\operatorname{rank}(\{z_i\}_{i\in[n]}) = n$, which implies $\operatorname{rank}(\{\nabla f_i(x)\}_{i\in[n]}) = n$. □

Now, we are ready to prove Lemma 6.3 which is restated below.

Lemma 6.3. *Consider the loss function L defined in (14), manifold Γ and its open neighborhood defined in (15). For gradient flow $\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla L(x_t)$ starting at any $x_0 \in U$, it holds that $\Phi(x_0) \in \Gamma$.*

Proof of Lemma 6.3. It suffices to prove gradient flow $\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla L(x_t)$ converges when $t \to \infty$, as long as $x_0 \in U$. Whenever it converges, it must converge to a stationary point in U. The proof will be completed by noting that all stationary point of L in U belongs to Γ (Lemma D.4).

Below we prove $\lim_{t\to\infty} x_t$ exists. By Lemma D.16, denote $C = \min_{i\in[d]} |u_0^i v_0^i| \lambda_{\min}(ZZ^\top)$, then $\lambda_{\min}(K(x_t)) \geq C$ for all $t \geq 0$. Thus,

$$\left\| \frac{\mathrm{d}x_t}{\mathrm{d}t} \right\| = \|\nabla L(x_t)\| \leq \frac{\|\nabla L(x_t)\|_2^2}{\sqrt{C L(x_t)}} = \frac{-\frac{\mathrm{d}L(x_t)}{\mathrm{d}t}}{\sqrt{L(x_t)}} = -\frac{1}{2\sqrt{C}} \frac{\mathrm{d}\sqrt{L(x_t)}}{\mathrm{d}t}.$$

Thus the total GF trajectory length is bounded by $\int_{t=0}^{\infty} \left\| \frac{\mathrm{d}x_t}{\mathrm{d}t} \right\| \mathrm{d}t \leq \int_{t=0}^{\infty} -\frac{1}{2\sqrt{C}} \frac{\mathrm{d}\sqrt{L}(x_t)}{\mathrm{d}t} \mathrm{d}t \leq \frac{L(x_0)}{2\sqrt{C}}$, where the last inequality uses that L is non-negative over $\mathbb{R}^D$. Therefore, the GF must converge. □

D.4 Proof of results in Section 6.2

To study the optimal solution to (18), we consider the corresponding d-dimensional convex program in terms of $w \in \mathbb{R}^d$, which has been studied in Tropp (2015):

$$\begin{aligned} \text{minimize} \quad & R(w) = \frac{1}{4n} \sum_{j=1}^{d} \left(\sum_{i=1}^{n} (z_i^j)^2 \right) |w^j|, \\ \text{subject to} \quad & Zw = Zw_*. \end{aligned} \tag{36}$$

Here we slightly abuse the notation of R and the parameter dimension will be clear from the context. We can relate the optimal solution to (18) to that of (36) via a canonical parametrization defined as follows.

Definition D.5 (Canonical Parametrization)**.** For any $w \in \mathbb{R}^d$, we define $\binom{u}{v} = \psi(w) = ([w^\top]_+^{\odot 1/2}, [-w^\top]_+^{\odot 1/2})^\top$ as the *canonical parametrization* of w. Clearly, it holds that $u^{\odot 2} - v^{\odot 2} = w$.

Indeed, we can show that if (36) has a unique optimal solution, it immediately follows that the optimal solution to (18) is also unique up to sign flips of each coordinate, as summarized in the lemma below.

Lemma D.6. *Suppose the optimal solution to (36) is unique and equal to w_*. Then the optimal solution to (18) is also unique up to sign flips of each coordinate. In particular, one of them is given by $(\tilde{u}_*, \tilde{v}_*) = \psi(w_*)$, that is, the canonical parametrization of w_*.*

Proof of Lemma D.6. Let $(\hat{u}, \hat{v})$ be any optimal solution of (18) and we define $\hat{w} = \hat{u}^{\odot 2} - \hat{v}^{\odot 2}$, which is also feasible to (36). By the optimality of w_*, we have

$$\sum_{j=1}^{d} \left(\sum_{i=1}^{n} (z_i^j)^2 \right) |w_*^j| \leq \sum_{j=1}^{d} \left(\sum_{i=1}^{n} (z_i^j)^2 \right) |\hat{w}^j| \leq \sum_{j=1}^{d} \left(\sum_{i=1}^{n} (z_i^j)^2 \right) [(\hat{u}^j)^2 + (\hat{v}^j)^2]. \tag{37}$$

On the other hand, $(\tilde{u}_*, \tilde{v}_*) = \psi(w_*)$ is feasible to (18). Thus, it follows from the optimality of $(\hat{u}, \hat{v})$ that

$$\sum_{j=1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)[(\hat{u}^j)^2+(\hat{v}^j)^2] \le \sum_{j=1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)[(\tilde{u}_*^j)^2+(\tilde{v}_*^j)^2] = \sum_{j=1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)|w_*^j|. \tag{38}$$

Combining (37) and (38) yields

$$\sum_{j=1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)[(\hat{u}^j)^2+(\hat{v}^j)^2] = \sum_{j=1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)|w_*^j| = \sum_{j=1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)|(\hat{u}^j)^2-(\hat{v}^j)^2| \tag{39}$$

which implies that $\hat{u}^{\odot 2} - \hat{v}^{\odot 2}$ is also an optimal solution of (36). Since w_* is the unique optimal solution to (36), we have $\hat{u}^{\odot 2} - \hat{v}^{\odot 2} = w_*$. Moreover, by (39), we must have $\hat{u}^{\odot 2} = [w_*]_+$ and $\hat{u}^{\odot 2} = [w_*]_+$, otherwise the equality would not hold. This completes the proof. □

Therefore, the unique optimality of (18) can be reduced to that of (36). In the sequel, we show that the latter holds for both Boolean and Gaussian random vectors. We divide Lemma 6.6 into to Lemma D.8 and D.7 for clarity.

Lemma D.7 (Boolean Case). *Let $z_1, \dots, z_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{\pm 1\}^d)$. There exist some constants $C, c > 0$ such that if the sample size n satisfies*

$$n \ge C[\kappa \ln(d/\kappa) + \kappa]$$

then with probability at least $1 - e^{-cn^2}$, the optimal solution of (18), $(\hat{u}, \hat{v})$, is unique up to sign flips of each coordinate and recovers the groundtruth, i.e., $\hat{u}^{\odot 2} - \hat{v}^{\odot 2} = w_*$.

Proof of Lemma D.7. By the assumption that $z_1, \dots, z_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{\pm 1\}^d)$, we have $\sum_{i=1}^n (z_i^j)^2 = n$ for all $j \in [d]$. Then (36) is equivalent to the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & g(w) = \|w\|_1, \\ \text{subject to} \quad & Zw = Z(u_*^{\odot 2} - v_*^{\odot 2}). \end{aligned} \tag{40}$$

This model exactly fits the Example 6.2 in Tropp (2015) with $\sigma = 1$ and $\alpha = 1/\sqrt{2}$. Then applying Equation (4.2) and Theorem 6.3 in Tropp (2015), (40) has a unique optimal solution equal to $u_*^{\odot 2} - v_*^{\odot 2}$ with probability at least $1 - e^{-ch^2}$ for some constant $c > 0$, given that the sample size satisfies

$$n \ge C(\kappa \ln(d/\kappa) + \kappa + h)$$

for some absolute constant $C > 0$. Choosing $h = \frac{n}{2C}$ and then adjusting the choices of C, c appropriately yield the desired result. Finally, applying Lemma D.6 finishes the proof. □

The Gaussian case requires more careful treatment.

Lemma D.8 (Gaussian Case). *Let $z_1, \dots, z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$. There exist some constants $C, c > 0$ such that if the sample size satisfies*

$$n \ge C\kappa \ln d,$$

then with probability at least $1 - (2d+1)e^{-cn}$, the optimal solution of (18), $(\hat{u}, \hat{v})$, is unique up to sign flips of each coordinate of $\hat{u}$ and $\hat{v}$ and recovers the groundtruth, i.e., $\hat{u}^{\odot 2} - \hat{v}^{\odot 2} = w_*$.

Proof of Lemma D.8. Since $z_1, \dots, z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, we have

$$\mathbb{P}\left[\sum_{i=1}^{n}(z_i^j)^2 \in [n/2, 3n/2], \forall j \in [d]\right] \ge 1 - 2de^{-cn}$$

for some constant $c > 0$, and we denote this event by $\mathcal{E}_n$. Therefore, on $\mathcal{E}_n$, we have

$$\frac{n}{2}\sum_{j=1}^{D}[(u^j)^2 + (v^j)^2] \leq R(x) \leq \frac{3n}{2}\sum_{j=1}^{D}[(u^j)^2 + (v^j)^2]$$

or equivalently,

$$\frac{n}{2}(\|u^{\odot 2}\|_1 + \|v^{\odot 2}\|_1) \leq R(x) \leq \frac{3n}{2}(\|u^{\odot 2}\|_1 + v^{\odot 2}\|_1).$$

Define $w_* = u_*^{\odot 2} - v_*^{\odot 2}$, and (36) is equivalent to the following convex optimization problem

$$\begin{aligned} \text{minimize} \quad & g(w) = \sum_{j=1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)\cdot|w^j + w_*^j|, \\ \text{subject to} \quad & Zw = 0. \end{aligned} \tag{41}$$

The point $w = 0$ is feasible for (41), and we claim that this is the unique optimal solution when n is large enough. In detail, assume that there exists a non-zero feasible point w for (41) in the descent cone (Tropp, 2015) $\mathcal{D}(g, w_*)$ of g, then

$$\lambda_{\min}(Z; \mathcal{D}(g, w_*)) \leq \frac{\|Zw\|_2}{\|w\|_2} = 0$$

where the equality follows from that w is feasible. Therefore, we only need to show that $\lambda_{\min}(Z; \mathcal{D}(g, x_*))$ is bounded from below for sufficiently large n.

On $\mathcal{E}_n$, it holds that g belongs to the following function class

$$\mathcal{G} = \left\{ h : \mathbb{R}^d \to \mathbb{R} \mid h(w) = \sum_{j=1}^{d}\xi^i|w^i|, \xi \in \Xi \right\} \text{ with } \Xi = \{\xi \in \mathbb{R}^d : \xi^j \in [0.5, 1.5], \forall j \in [d]\}.$$

We identify $g_\xi \in \mathcal{G}$ with $\xi \in \Xi$, then $\mathcal{D}(g, w_*) \subseteq \cup_{\xi\in\Xi}\mathcal{D}(g_\xi, w_*)) := \mathcal{D}_\Xi$, which further implies that

$$\lambda_{\min}(Z; \mathcal{D}(g, w_*)) \geq \lambda_{\min}(Z; \mathcal{D}_\Xi).$$

Recall the definition of minimum conic singular value (Tropp, 2015):

$$\lambda_{\min}(Z; \mathcal{D}_\Xi) = \inf_{p\in\mathcal{D}_\Xi\cap\mathcal{S}^{d-1}} \sup_{q\in\mathcal{S}^{n-1}} \langle q, Zp\rangle.$$

where $\mathcal{S}^{n-1}$ denotes the unit sphere in $\mathbb{R}^n$. Applying the same argument as in (Tropp, 2015) yields

$$\mathbb{P}\left[\lambda_{\min}(Z; \mathcal{D}_\Xi) \geq \sqrt{n-1} - w(\mathcal{D}_\Xi) - h\right] \geq 1 - e^{-h^2/2}.$$

Take the intersection of this event with $\mathcal{E}_n$, and we obtain from a union bound that

$$\lambda_{\min}(Z; \mathcal{D}(g, w_*)) \geq \sqrt{n-1} - w(\mathcal{D}_\Xi) - h \tag{42}$$

with probability at least $1 - e^{-h^2/2} - 2de^{-cn}$. It remains to determine $w(\mathcal{D}_\Xi)$, which is defined as

$$w(\mathcal{D}_\Xi) = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}\left[\sup_{p\in\mathcal{D}_\Xi\cap\mathcal{S}^{d-1}}\langle z, p\rangle\right] = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}\left[\sup_{\xi\in\Xi}\sup_{p\in\mathcal{D}(g_\xi,x_*)\cap\mathcal{S}^{d-1}}\langle z, p\rangle\right]. \tag{43}$$

Without loss of generality, we assume that $w_* = (w_*^1, \ldots, w_*^\kappa, 0, \ldots, 0)^\top$ with $w_*^1, \ldots, w_*^\kappa > 0$, otherwise one only needs to specify the signs and the nonzero set of w_* in the sequel. For any $\xi \in \Xi$ and any $p \in \mathcal{D}(g_\xi, w_*) \cap \mathcal{S}^{d-1}$, there exists some $\tau > 0$ such that $g_\xi(w_* + \tau \cdot p) \leq g_\xi(w_*)$, i.e.,

$$\sum_{j=1}^{d}\xi^j|w_*^j + \tau p^j| \leq \sum_{j=1}^{d}\xi^j|w_*^j|$$

which further implies that

$$\tau \sum_{j=\kappa+1}^{d} \xi^j |p^j| \le \sum_{j=1}^{\kappa} \xi^j (|w_*^j| - |w_*^j - \tau p^j|) \le \tau \cdot \sum_{j=1}^{\kappa} \xi^j |p^j|$$

where the second inequality follows from the triangle inequality. Then since each $\xi^j \in [1/2, 3/2]$, it follows that

$$\sum_{j=\kappa+1}^{d} |p^j| \le 3 \sum_{j=1}^{\kappa} |p^j|.$$

Note that this holds for all $\xi \in \Xi$ simultaneously. Now let us denote $p^{1:\kappa} = (p^1, \ldots, p^\kappa) \in \mathbb{R}^\kappa$ and $p^{(\kappa+1):d} = (p^{\kappa+1}, \ldots, p^d) \in \mathbb{R}^{d-\kappa}$, and similarly for other d-dimensional vectors. Then for all $p \in \mathcal{D}_\Xi \cap \mathcal{S}^{d-1}$, by Cauchy-Schwartz inequality, we have

$$\|p^{(\kappa+1):d}\|_1 \le 3 \cdot \|p^{1:\kappa}\|_1 \le 3\sqrt{\kappa} \cdot \|p^{1:\kappa}\|_2.$$

Thus, for any $z \in \mathbb{R}^d$ and any $p \in \mathcal{D}_\Xi \cap \mathcal{S}^{d-1}$, it follows that

$$\begin{aligned}
\langle z, p \rangle &= \langle z^{1:\kappa}, p^{1:\kappa} \rangle + \langle z^{(\kappa+1):d}, p^{(\kappa+1):d} \rangle \\
&\le \|z^{1:\kappa}\|_2 \|p^{1:\kappa}\|_2 + \|p^{(\kappa+1):d}\|_1 \cdot \max_{j \in \{\kappa+1,\ldots,d\}} |z_j| \\
&\le \|z^{1:\kappa}\|_2 \|p^{1:\kappa}\|_2 + 3\sqrt{\kappa} \|p^{1:\kappa}\|_2 \cdot \max_{j \in \{\kappa+1,\ldots,d\}} |z_j| \\
&\le \|z^{1:\kappa}\|_2 + 3\sqrt{\kappa} \cdot \max_{j \in \{\kappa+1,\ldots,d\}} |z_j|
\end{aligned}$$

where the last inequality follows from the fact that $p \in \mathcal{S}^{d-1}$. Therefore, combine the above inequality with (43), and we obtain that

$$\begin{aligned}
w(\mathcal{D}_\Xi) &\le \mathbb{E}\left[\|z_{1:\kappa}\|_2 + 3\sqrt{\kappa} \cdot \max_{j \in \{\kappa+1,\ldots,d\}} |z_j| \right] \\
&\le \sqrt{\kappa} + 3\sqrt{\kappa} \cdot \mathbb{E}\left[\max_{j \in \{\kappa+1,\ldots,d\}} |z_j| \right]. \qquad (44)
\end{aligned}$$

where the second inequality follows from the fact that $\mathbb{E}[\|z_{1:\kappa}\|_2] \le \sqrt{\mathbb{E}[\|z_{1:\kappa}\|_2^2]} = \sqrt{\kappa}$. To bound the second term in (44), applying Lemma D.9, it follows from (44) that

$$w(\mathcal{D}_\Xi) \le \sqrt{\kappa} + 3\sqrt{2\kappa \ln(2(d-\kappa))}. \qquad (45)$$

Therefore, combining (45) and (42), we obtain

$$\lambda_{\min}(Z; \mathcal{D}(g, w_*)) \ge \sqrt{n-1} - \sqrt{\kappa} - 3\sqrt{2\kappa \ln(2(d-\kappa))} - h.$$

Therefore, choosing $h = \sqrt{n-1}/2$, as long as n satisfies that $n \ge C(\kappa \ln d)$ for some constant $C > 0$, we have $\lambda_{\min}(Z; \mathcal{D}(g, w_*)) > 0$ with probability at least $1 - (2d+1)e^{-cn}$. Finally, the uniqueness of the optimal solution to (18) in this case follows from Lemma D.6. □

Lemma D.9. *Let $z \sim \mathcal{N}(0, I_d)$, then it holds that $\mathbb{E}\left[\max_{i \in [d]} |z^i|\right] \le \sqrt{2\ln(2d)}$.*

Proof of Lemma D.9. Denote $M = \max_{i \in [d]} |z^i|$. For any $\lambda > 0$, by Jensen's inequality, we have

$$e^{\lambda \cdot \mathbb{E}[M]} \le \mathbb{E}\left[e^{\lambda M}\right] = \mathbb{E}\left[\max_{i \in [d]} e^{\lambda |z^i|}\right] \le \sum_{i=1}^{d} \mathbb{E}\left[e^{\lambda |z^i|}\right].$$

Note that $\mathbb{E}[e^{\lambda |z^i|}] \le 2 \cdot \mathbb{E}[e^{\lambda z^i}]$. Thus, by the expression of the Gaussian moment generating function, we further have

$$e^{\lambda \cdot \mathbb{E}[M]} \le 2 \sum_{i=1}^{d} \mathbb{E}\left[e^{\lambda z^i}\right] = 2de^{\lambda^2/2},$$

from which it follows that

$$\mathbb{E}[M] \le \frac{\ln(2d)}{\lambda} + \frac{\lambda}{2}.$$

Choosing $\lambda = \sqrt{2\ln(2d)}$ yields the desired result. □

D.5 PROOF OF LEMMA 6.5

Instead of studying the convergence of the Riemannian gradient flow directly, it is more convenient to consider it in the ambient space $\mathbb{R}^D$. To do so, we define a Lagrange function $\mathcal{L}(x;\lambda) = R(x) + \sum_{i=1}^n \lambda^i (f_i(x) - y_i)$ for $\lambda \in \mathbb{R}^n$. Based on this Lagrangian, we can continuously extend $\partial\Phi(x)\nabla R(x)$ to the whole space $\mathbb{R}^D$. In specific, we can find a continuous function $F : \mathbb{R}^D \to \mathbb{R}^D$ such that $F(\cdot)|_\Gamma = \partial\Phi(\cdot)\nabla R(\cdot)$. Such an F can be implicitly constructed via the following lemma.

Lemma D.10. *The ℓ_2 norm has a unique minimizer among $\{\nabla_x \mathcal{L}(x;\lambda) \mid \lambda \in \mathbb{R}^n\}$ for any fixed $x \in \mathbb{R}^D$. Thus we can define $F : \mathbb{R}^D \to \mathbb{R}^D$ by $F(x) = \operatorname{argmin}_{g \in \{\nabla_x \mathcal{L}(x;\lambda) | \lambda \in \mathbb{R}^n\}} \|g\|_2$. Moreover, it holds that $\langle F(x), \nabla f_i(x)\rangle = 0$ for all $i \in [n]$.*

Proof of Lemma D.10. Fix any $x \in \mathbb{R}^D$. Note that $\{\nabla_x \mathcal{L}(x;\lambda) \mid \lambda \in \mathbb{R}^n\}$ is the subspace spanned by $\{\nabla f_i(x)\}_{i\in[n]}$ shifted by $\nabla R(x)$, thus there is unique minimizer of the ℓ_2 norm in this set. This implies that $F(x) = \operatorname{argmin}_{g \in \{\nabla_x \mathcal{L}(x;\lambda) | \lambda \in \mathbb{R}^n\}} \|g\|_2$ is well-defined.

To show the second claim, denote $h(\lambda) = \|\nabla_x \mathcal{L}(x;\lambda)\|_2^2/2$, which is a quadratic function of $\lambda \in \mathbb{R}^n$. Then we have

$$\nabla h(\lambda) = \begin{pmatrix} \langle \nabla R(x), \nabla f_1(x)\rangle \\ \vdots \\ \langle \nabla R(x), \nabla f_n(x)\rangle \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^n \lambda^i \langle \nabla f_1(x), \nabla f_i(x)\rangle \\ \vdots \\ \sum_{i=1}^n \lambda^i \langle \nabla f_n(x), \nabla f_i(x)\rangle \end{pmatrix} = \begin{pmatrix} \langle \nabla R(x), \nabla f_1(x)\rangle \\ \vdots \\ \langle \nabla R(x), \nabla f_n(x)\rangle \end{pmatrix} + K(x)\lambda.$$

For any λ such that $\nabla_x \mathcal{L}(x;\lambda) = F(x)$, we must have $\nabla h(\lambda) = 0$ by the definition of $F(x)$, which by the above implies

$$(K(x)\lambda)^i = -\langle \nabla R(x), \nabla f_i(x)\rangle \qquad \text{for all } i \in [n].$$

Therefore, we further have

$$\langle F(x), \nabla f_i(x)\rangle = \langle \nabla R(x), \nabla f_i(x)\rangle + \sum_{j=1}^n \lambda^j \langle \nabla f_i(x), \nabla f_j(x)\rangle = \langle \nabla R(x), \nabla f_i(x)\rangle + (K(x)\lambda)^i = 0$$

for all $i \in [n]$. This finishes the proof. □

Hence, with any initialization $x_0 \in \Gamma$, the limiting flow (17) is equivalent to the following dynamics

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -F(x_t). \tag{46}$$

Thus Lemma 6.5 can be proved by showing that the above x_t converges to x_* as $t \to \infty$. We first present a series of auxiliary results in below.

Lemma D.11 (Implications for $F(x) = 0$)**.** *Let $F : \mathbb{R}^D \to \mathbb{R}^D$ be as defined in Lemma D.10. For any $x = \binom{u}{v} \in \mathbb{R}^D$ such that $F(x) = 0$, it holds that for each $j \in [d]$, either $u^j = 0$ or $v^j = 0$.*

Proof. Since $F(x) = 0$, it holds for all $j \in [d]$ that,

$$0 = \frac{\partial R}{\partial u^j}(x) + \sum_{i=1}^n \lambda(x)^i \frac{\partial f_i}{\partial u^j}(x) = 2u^j \cdot \left[\sum_{i=1}^n (z_i^j)^2 + \sum_{i=1}^n \lambda(x)^i z_i^j\right],$$
$$0 = \frac{\partial R}{\partial v^j}(x) + \sum_{i=1}^n \lambda(x)^i \frac{\partial f_i}{\partial v^j}(x) = 2v^j \cdot \left[\sum_{i=1}^n (z_i^j)^2 - \sum_{i=1}^n \lambda(x)^i z_i^j\right].$$

If there exists some $j \in [d]$ such that $u^j \neq 0$ and $v^j \neq 0$, then it follows from the above two identities that

$$\sum_{i=1}^n (z_i^j)^2 = 0$$

which happens with probability 0 in both the Boolean and Gaussian case. Therefore, we must have $u^j = 0$ or $v^j = 0$ for all $j \in [d]$. □

Lemma D.12. *Let $F : \mathbb{R}^D \to \mathbb{R}^D$ be as defined in Lemma D.10. Then F is continuous on $\mathbb{R}^D$.*

Proof. **Case I.** We first consider the simpler case of any fixed $x_0 \in U = (\mathbb{R} \setminus \{0\})^D$, assuming that $K(x_0)$ is full-rank. Lemma D.10 implies that for any $\lambda \in \mathbb{R}^n$ such that $\nabla_x \mathcal{L}(x_0; \lambda) = F(x_0)$, we have

$$K(x_0)\lambda = (-\langle \nabla R(x_0), \nabla f_i(x_0) \rangle)_{i\in[n]}.$$

Thus such λ is unique and given by

$$\lambda(x_0) = K(x_0)^{-1}(-\langle \nabla R(x_0), \nabla f_i(x_0) \rangle)_{i\in[n]}.$$

Since $K(x)$ is continuous, there exists a sufficiently small $\delta > 0$ such that for any $x \in B_\delta(x_0)$, $K(x)$ is full-rank, which further implies that $K(x)^{-1}$ is also continuous in $B_\delta(x)$. Therefore, by the above characterization of λ, we see that $\lambda(x)$ is continuous for $x \in B_\delta(x_0)$, and so is $F(x) = \nabla R(x) + \sum_{i=1}^n \lambda(x)^i \nabla f_i(x)$.

Case II. Next, we consider all general $x \in \mathbb{R}^D$. Here for simplicity, we reorder the coordinates as $x = (u^1, v^1, u^2, v^2, \ldots, u^d, v^d)$ with a slight abuse of notation. Without loss of generality, fix any x_0 such that for some $q \in [d]$, $(u_0^i)^2 + (v_0^i)^2 > 0$ for all $i = 1, \ldots, q$ and $u_0^i = v_0^i = 0$ for all $i = q+1, \ldots, d$. Then $\nabla R(x_0)$ and $\{\nabla f_i(x_0)\}_{i\in[n]}$ only depend on $\{z_i^j\}_{i\in[n], j\in[q]}$, and for all $i \in [n]$, it holds that

$$\nabla R(x_0)^{(2q+1):D} = \nabla f_i(x_0)^{(2q+1):D} = 0.$$

Note that if we replace $\{\nabla f_i(x)\}_{i\in[n]}$ by any fixed and invertible linear transform of itself, it would not affect the definition of $F(x)$. In specific, we can choose an invertible matrix $Q \in \mathbb{R}^{n\times n}$ such that, for some $q' \in [q]$, $(\tilde{z}_1, \ldots, \tilde{z}_n) = (z_1, \ldots, z_n)Q$ satisfies that $\{\tilde{z}_i^{1:q}\}_{i\in[q']}$ is linearly independent and $\tilde{z}_i^{1:q} = 0$ for all $i = q'+1, \ldots, n$. We then consider $(\nabla \tilde{f}_1(x), \ldots, \nabla \tilde{f}_n(x)) = (\nabla f_1(x), \ldots, \nabla f_n(x))Q$ and the corresponding $\tilde{F}(x)$. For notational simplicity, we assume that Q can be chosen as the identity matrix, so that $(z_1, \ldots, z_n)$ itself satisfies the above property, and we repeat it here for clarity

$$\{z_i^{1:q}\}_{i\in[q']} \text{ is linearly independent and } \tilde{z}_i^{1:q} = 0 \text{ for all } i = q'+1, \ldots, n. \tag{47}$$

This further implies that

$$\nabla f_i(x)^{1:(2q)} = 0, \quad \text{for all } i \in \{q'+1, \ldots, n\} \text{ and } x \in \mathbb{R}^D. \tag{48}$$

In the sequel, we use λ for n-dimensional vectors and $\bar{\lambda}$ for q'-dimensional vectors. Denote[2]

$$\lambda(x) \in \operatorname*{argmin}_{\lambda\in\mathbb{R}^n} \left\| \nabla R(x) + \sum_{i=1}^n \lambda^i \nabla f_i(x) \right\|_2,$$
$$\bar{\lambda}(x) \in \operatorname*{argmin}_{\bar{\lambda}\in\mathbb{R}^{q'}} \left\| \nabla R(x)^{1:(2q)} + \sum_{i=1}^{q'} \bar{\lambda}^i \nabla f_i(x)^{1:(2q)} \right\|_2.$$

Then due to (47) and (48), we have

$$\left\| \nabla R(x_0)^{1:(2q)} + \sum_{i=1}^{q'} \bar{\lambda}(x_0)^i \nabla f_i(x_0)^{1:(2q)} \right\|_2 = \left\| \nabla R(x_0) + \sum_{i=1}^n \lambda(x)^i \nabla f_i(x_0) \right\|_2 = \|F(x_0)\|_2. \tag{49}$$

[2]We do not care about the specific choice of $\lambda(x)$ or $\bar{\lambda}(x)$ when there are multiple candidates, and we only need their properties according to Lemma D.10, so they can be arbitrary.

On the other hand, for any $x \in \mathbb{R}^D$, by (48), we have

$$\begin{aligned}\left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{q'} \bar{\lambda}(x)^i \nabla f_i(x)^{1:(2q)}\right\|_2 &= \min_{\lambda\in\mathbb{R}^n} \left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{n} \lambda^i \nabla f_i(x)^{1:(2q)}\right\|_2 \\ &\le \left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{n} \lambda(x)^i \nabla f_i(x)^{1:(2q)}\right\|_2 = \|F(x)^{1:(2q)}\|_2 \\ &\le \|F(x)\|_2 \le \left\|\nabla R(x) + \sum_{i=1}^{n} \lambda(x_0)^i \nabla f_i(x)\right\|_2 \end{aligned} \tag{50}$$

where the first and third inequalities follow from the definition of $F(x)$. Let $x \to x_0$, by the continuity of $\nabla R(x)$ and $\{\nabla f_i(x)\}_{i\in[n]}$, we have

$$\lim_{x\to x_0} \left\|\nabla R(x) + \sum_{i=1}^{n} \lambda(x_0)^i \nabla f_i(x)\right\|_2 = \left\|\nabla R(x_0) + \sum_{i=1}^{n} \lambda(x_0)^i \nabla f_i(x_0)\right\|_2 \tag{51}$$

Denote $\tilde{K}(x) = (\tilde{K}_{ij}(x))_{(i,j)\in[q']^2} = (\langle \nabla f_i(x)^{1:(2q)}, \nabla f_i(x)^{1:(2q)}\rangle)_{(i,j)\in[q']^2}$. By applying the same argument as in **Case I**, since $\tilde{K}(x_0)$ is full-rank, it also holds that $\lim_{x\to x_0} \bar{\lambda}(x) = \bar{\lambda}(x_0)$, and thus

$$\lim_{x\to x_0} \left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{q'} \bar{\lambda}(x)^i \nabla f_i(x)^{1:(2q)}\right\|_2 = \left\|\nabla R(x_0)^{1:(2q)} + \sum_{i=1}^{q'} \bar{\lambda}(x_0)^i \nabla f_i(x_0)^{1:(2q)}\right\|_2. \tag{52}$$

Combing (49), (50), (51) and (52) yields

$$\lim_{x\to x_0} \|F(x)^{1:(2q)}\|_2 = \lim_{x\to x_0} \min_{\lambda\in\mathbb{R}^n} \left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{n} \lambda^i \nabla f_i(x)^{1:(2q)}\right\|_2 = \|F(x_0)\|_2. \tag{53}$$

Moreover, since $\|F(x)^{(2q+1):D}\|_2 = \sqrt{\|F(x)\|_2^2 - \|F(x)^{1:(2q)}\|_2^2}$, we also have

$$\lim_{x\to x_0} \|F(x)^{(2q+1):D}\|_2 = 0. \tag{54}$$

It then remains to show that $\lim_{x\to x_0} F(x)^{1:(2q)} = F(x_0)^{1:(2q)}$, which directly follows from $\lim_{x\to x_0} \lambda(x)^{1:q'} = \lambda(x_0)^{1:q'} = \bar{\lambda}(x_0)$.

Now, for any $\epsilon > 0$, due to the convergence of $\bar{\lambda}(x)$ and that $\tilde{K}(x_0) \succ 0$, we can pick a sufficiently small δ_1 such that for some constant $\alpha > 0$ and all $x \in B_{\delta_1}(x_0)$, it holds that $\|\bar{\lambda}(x) - \bar{\lambda}(x_0)\|_2 \le \epsilon/2$ and

$$\left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{q'} \bar{\lambda}^i \nabla f_i(x)^{1:(2q)}\right\|_2^2 \ge \left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{q'} \bar{\lambda}(x)^i \nabla f_i(x)^{1:(2q)}\right\|_2^2 + \alpha\|\bar{\lambda} - \bar{\lambda}(x)\|_2^2. \tag{55}$$

for all $\bar{\lambda} \in \mathbb{R}^p$, where the inequality follows from the strong convexity. Meanwhile, due to (48), we have

$$\begin{aligned}\lim_{x\to x_0} \left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{q'} \lambda(x)^i \nabla f_i(x)^{1:(2q)}\right\|_2 &= \lim_{x\to x_0} \left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{n} \lambda(x)^i \nabla f_i(x)^{1:(2q)}\right\|_2 \\ &= \left\|\nabla R(x_0)^{1:(2q)} + \sum_{i=1}^{q'} \bar{\lambda}(x_0)^i \nabla f_i(x_0)^{1:(2q)}\right\|_2 \\ &= \lim_{x\to x_0} \left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{q'} \bar{\lambda}(x)^i \nabla f_i(x)^{1:(2q)}\right\|_2. \end{aligned}$$

where the second equality follows from (53) and the second equality is due to (52). Therefore, we can pick a sufficiently small δ_2 such that

$$\left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{q'} \lambda(x)^i \nabla f_i(x)^{1:(2q)}\right\|_2 \leq \left\|\nabla R(x)^{1:(2q)} + \sum_{i=1}^{q'} \bar{\lambda}(x)^i \nabla f_i(x)^{1:(2q)}\right\|_2 + \frac{\alpha\epsilon^2}{4} \tag{56}$$

for all $x \in B_{\delta_2}(x_0)$. Setting $\delta = \min(\delta_1, \delta_2)$, it follows from (55) and (56) that

$$\|\lambda(x)^{1:q'} - \bar{\lambda}(x)\|_2 \leq \frac{\epsilon}{2}, \quad \text{for all } x \in B_\delta(x_0).$$

Recall that we already have $\|\bar{\lambda}(x) - \bar{\lambda}(x_0)\| \leq \epsilon/2$, and thus

$$\|\lambda(x)^{1:q'} - \lambda(x_0)^{1:q'}\|_2 = \|\lambda(x)^{1:q'} - \bar{\lambda}(x_0)\|_2 \leq \|\lambda(x)^{1:q'} - \bar{\lambda}(x)\|_2 + \|\bar{\lambda}(x) - \bar{\lambda}(x_0)\|_2 \leq \epsilon$$

for all $x \in B_\delta(x_0)$. Therefore, we see that $\lim_{x\to x_0} \lambda(x)^{1:q'} = \lambda(x_0)^{1:q'}$.

Finally, it follows from the triangle inequality that

$$\begin{aligned}\|F(x) - F(x_0)\|_2 &\leq \|F(x)^{1:(2q)} - F(x_0)^{1:(2q)}\|_2 + \|F(x)^{(2q+1):D}\|_2 + \|F(x_0)^{(2q+1):D}\|_2 \\ &\leq \left\|\sum_{i=1}^{q'} \lambda(x)^i \nabla f_i(x) - \lambda(x_0)^i \nabla f_i(x_0)\right\|_2 + \|\nabla R(x) - \nabla R(x_0)\|_2 + \|F(x)^{(2q+1):D}\|_2\end{aligned}$$

where, as $x \to x_0$, the first term vanishes by the convergence of $\lambda(x)^{1:q'}$ and the continuity of each $\nabla f_i(x)$, the second term converges to 0 by the continuity of $\nabla R(x)$ and the third term vanishes by (54). Therefore, we conclude that

$$\lim_{x\to x_0} F(x) = F(x_0),$$

that is, F is continuous.

□

Lemma D.13. *For any initialization $x_0 \in \Gamma$, the Riemmanian Gradient Flow (17) (or equivalently, (46)) is defined on $[0, \infty)$.*

Proof of Lemma D.13. If the Riemannian gradient flow had stopped in finite time, we must have $u_t^j = 0$ or $v_t^j = 0$ for some $j \in [d]$ by Lemma D.11. Therefore, we only need to prove that all u_t^j's and v_t^j's are bounded away from 0 in finite time. Now, fix some $j \in [d]$, and we assume that $u_0^j v_0^j > 0$ without loss of generality. It then suffices to show that $u_t^j v_t^j > 0$ for all $t \in (0, \infty)$. By the definition of the projected gradient flow in (17), we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(u_t^j v_t^j) &= \begin{pmatrix} v_t^j e_j^\top & u_t^j e_j^\top \end{pmatrix} \frac{\partial x_t}{\partial t} \\ &= -\begin{pmatrix} v_t^j e_j^\top & u_t^j e_j^\top \end{pmatrix} F(x_t).\end{aligned}$$

By the expression of $F(x_t) = \nabla R(x_t) + \sum_{i=1}^n \lambda(x_t)^i \nabla f_i(x_t)$, we then have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(u_t^j v_t^j) &= -\left[\sum_{i=1}^n (z_i^j)^2 + \sum_{i=1}^n \lambda(x_t)^i z_i^t\right] u_t^j v_t^j - \left[\sum_{i=1}^n (z_i^j)^2 - \sum_{i=1}^n \lambda(x_t)^i z_i^t\right] u_t^j v_t^j \\ &= -\left(\sum_{i=1}^n (z_i^j)^2\right) u_t^j v_t^j.\end{aligned}$$

Denote $s_j = \sum_{i=1}^n (z_i^j)^2$. In either the Boolean or Gaussian case, we have $s_j \in (0, \infty)$ with probability 1. Therefore, it follows that $u_t^j v_t^j = u_0^j v_0^j e^{-s_j t} > 0$ for all $t \in (0, \infty)$. This finishes the proof. □

Before showing that F satisfies the PL condition, we need the following two intermediate results. Given two points u and v in $\mathbb{R}^d$, we say u weakly dominate v (written as $u \leq v$) if and only if $u^i \leq v^i$, for all $i \in [d]$. Given two subsets A and B of $\mathbb{R}^D$, we say A weakly dominates B if and only if for any point v in B, there exists a point $u \in A$ such that $u \leq v$.

Lemma D.14. *For some $q \in [D]$, let S be any q-dimensional subspace of $\mathbb{R}^D$ and $P = \{u \in \mathbb{R}^D \mid u^i \geq 0, \forall i \in [D]\}$. Let $u_\star$ be an arbitrary point in P and $Q = P \cap (u_\star + S)$. Then there exists a radius $r > 0$, such that $B_r^1(0) \cap Q$ weakly dominates Q, where $B_r^1(0)$ is the ℓ_1-norm ball of radius r centered at 0.*

As a direct implication, for any continuous function $f : P \to \mathbb{R}$, which is coordinate-wise non-decreasing, $\min_{x \in U} f(x)$ can always be achieved.

Proof of Lemma D.14. We will prove by induction on the environment dimension D. For the base case of $D = 1$, either $S = \{0\}$ or $S = \mathbb{R}$, and it is straight-forward to verify the desired for both scenarios.

Suppose the proposition holds for $D - 1$, below we show it holds for D. For each $i \in [D]$, we apply the proposition with $D - 1$ to $Q \cap \{u \in P \mid u^i = 0\}$ (which can be seen as a subset of $\mathbb{R}^{D-1}$), and let r_i be the corresponding ℓ_1 radius. Set $r = \max_{i \in [D]} r_i$, and we show that choosing the radius to be r suffices.

For any $v \in Q$, we take a random direction in S, denoted by ω. If $\omega \geq 0$ or $\omega \leq 0$, we denote by y the first intersection (i.e., choosing the smallest λ) between the line $\{v - \lambda|\omega|\}_{\lambda \geq 0}$ and the boundary of U, i.e., $\cup_{i=1}^D \{z \in \mathbb{R}^D \mid z^i = 0\}$. Clearly $y \leq v$. By the induction hypothesis, there exists a $u \in B_r^1(0) \cap Q$ such that $u \leq y$. Thus $u \leq v$ and meets our requirement.

If ω has different signs across its coordinates, we take y_1, y_2 to be the first intersections of the line $\{v - \lambda|\omega|\}_{\lambda \in \mathbb{R}}$ and the boundary of U in directions of $\lambda > 0$ and $\lambda < 0$, respectively. Again by the induction hypothesis, there exist $u_1, u_2 \in B_r^1(0) \cap Q$ such that $u_1 \leq y_1$ and $u_2 \leq y_2$. Since v lies in the line connecting u_1 and u_2, there exists some $h \in [0, 1]$ such that $v = (1 - h)u_1 + hu_2$. It then follows that $(1 - h)u_1 + hu_2 \leq (1 - h)y_1 + hy_2 = v$. Now since Q is convex, we have $(1 - h)u_1 + hu_2 \in Q$, and by the triangle inequality it also holds that $\|(1 - h)u_1 + hu_2\|_1 \leq r$, so $(1 - h)u_1 + hu_2 \in B_r^1(0) \cap Q$. Therefore, we conclude that $B_r^1(0) \cap Q$ weakly dominates Q, and thus the proposition holds for D. This completes the proof by induction. □

Lemma D.15. *For some $q \in [D]$, let S be any q-dimensional subspace of $\mathbb{R}^D$ and $P = \{u \in \mathbb{R}^D \mid u^i \geq 0, \forall i \in [D]\}$. Let $u_\star$ be an arbitrary point in P and $Q = P \cap (u_\star + S)$. Then there exists a constant $c \in (0, 1]$ such that for any sufficiently small radius $r > 0$, $c \cdot Q$ weakly dominates $P \cap (u_\star + S + B_r^2(0))$, where $B_r^2(0)$ is the ℓ_2-norm ball of radius r centered at 0.*

Proof of Lemma D.15. We will prove by induction on the environment dimension D. For the base case of $D = 1$, either $S = \{0\}$ or $S = \mathbb{R}$. $S = \mathbb{R}$ is straight-forward; for the case $S = \{0\}$, we just need to ensure $c|u_\star| \leq |u_\star| - r$, and it suffices to pick $r = |u_\star|$ and $c = 0.5$.

Suppose the proposition holds for $D - 1$, below we show it holds for D. For each $i \in [D]$, we first consider the intersection between $P \cap (u_\star + S + B_r^2(0))$ and $H_i := \{u \in \mathbb{R}^D \mid u_i = 0\}$. Let u_i be an arbitrary point in $P \cap (u_\star + S) \cap H_i$, then $P \cap (u_\star + S) \cap H_i = P \cap (u_i + S) \cap H_i = P \cap (u_i + S \cap H_i)$. Furthermore, there exists $\{\alpha_i\}_{i \in [D]}$ which only depends on S and satisfies $P \cap (u_* + S + B_r^2(0)) \cap H_i \subset P \cap (u_i + S \cap H_i + B_{\alpha_i r}^2(0) \cap H_i)$. Applying the induction hypothesis to $P \cap (u_i + S \cap H_i + B_{\alpha_i r}^2(0) \cap H_i)$, we know there exists a $c > 0$ such that for sufficiently small r, $c(P \cap (u_\star + S) \cap H_i) = c(P \cap (u_i + S \cap H_i))$ weakly dominates $P \cap (u_i + S \cap H_i + B_{\alpha_i r}^2(0) \cap H_i)$.

For any point v in Q and any $z \in B_r^2(0)$, we take a random direction in S, denoted by ω. If $\omega \geq 0$ or $\omega \leq 0$, we denote by y the first intersection between $\{v + z - \lambda|\omega|\}_{\lambda \geq 0}$ and the boundary of U. Clearly $y \leq v$. Since $y \in P \cap (u_\star + S + B_r^2(0)) \cap H_i \subset P \cap (u_i + S \cap H_i + B_{\alpha_i r}^2(0) \cap H_i)$, by the induction hypothesis, there exists a $u \in c(P \cap (u_\star + S) \cap H_i)$ such that $u \leq y$. Thus $z \leq v + z$ and $z \in c(P \cap (u_\star + S)) = cQ$.

If ω has different signs across its coordinates, we take y_1, y_2 to be the first intersections of the line $\{v + z - \lambda|\omega|\}_{\lambda \in \mathbb{R}}$ and the boundary of U in directions of $\lambda > 0$ and $\lambda < 0$, respectively. By the induction hypothesis, there exist $u_1, u_2 \in c \cdot Q$ such that $u_1 \leq y_1$ and $u_2 \leq y_2$. Since $v + z$ lies

in the line connecting u_1 and u_2, there exists some $h \in [0,1]$ such that $v+z = (1-h)y_1 + hy_2$. It then follows that $(1-h)u_1 + hu_2 \le (1-h)y_1 + hy_2 = v + z$. Since Q is convex, we have $(1-h)u_1 + hu_2 \in cQ$. Therefore, we conclude that $cQ \cap Q$ weakly dominates $P \cap (u_\star + S + B_r^2(0))$ for all sufficiently small r, and thus the proposition holds for D. This completes the proof by induction.

□

Lemma D.16. *(Polyak-Łojasiewicz condition for F.) For any x_* such that $L(x_*) = 0$, i.e., $x_* \in \overline{\Gamma}$, there exist a neighbourhood U' of x_* and a constant $c > 0$, such that $\|F(x)\|_2^2 \ge c \cdot \max(R(x) - R(x_*), 0)$ for all $x \in U' \cap \overline{\Gamma}$. Note this requirement is only non-trivial when $\|F(x_*)\|_2 = 0$ since F is continuous.*

Proof of Lemma D.16. It suffices to show the PL condition for $\{x \mid F(x) = 0\}$. We need to show for any x_* satisfying $F(x_*) = 0$, there exist some $\epsilon > 0$ and $C > 0$, such that for all $x \in \overline{\Gamma} \cap B_\epsilon^2(x_*)$ with $R(x) > R(x_*)$, it holds that $\|F(x)\|_2^2 \ge C(R(x) - R(x_*))$.

Case I. We first prove the case where $x = \binom{u}{v}$ itself is a canonical parametrization of $w = u^{\odot 2} - v^{\odot 2}$, i.e., $u^j v^j = 0$ for all $j \in [d]$. Since x_* satisfies $\nabla F(x_*) = 0$, by Lemma D.11, we have $x_* = \psi(w_*)$ where $w_* = u_*^{\odot 2} - v_*^{\odot 2}$. In this case, we can rewrite both R and F as functions of $w \in \mathbb{R}^d$. In detail, we define $R'(w) = R(\psi(w))$ and $F'(w) = F(\psi(w))$ for all $w \in \mathbb{R}^d$. For any w in a sufficiently small neighbourhood of w_*, it holds that $\operatorname{sign}(w^j) = \operatorname{sign}(w_*^j)$ for all $j \in [q]$. Below we show that for each possible sign pattern of $w^{(q+1):d}$, there exists some constant C which admits the PL condition in the corresponding orthant. Then we take the minimum of all C from different orthant and the proof is completed. W.L.O.G., we assume that $w^j \ge 0$, for all $j = q+1 \dots, d$.

We temporarily reorder the coordinates as $x = (u^1, v^1, u^2, v^2, \dots, u^d, v^d)^\top$. Recall that $Z = [z_1, \dots, z_n]^\top$ is a n-by-d matrix, and we have

$$\|F'(w)\|_2^2 = \min_{\lambda \in \mathbb{R}^n} \left\langle (a - \operatorname{sign}(w) \odot Z^\top \lambda)^{\odot 2}, |w| \right\rangle,$$

where $a = \frac{1}{n}\sum_{i=1}^n z_i^{\odot 2} \in \mathbb{R}^d$. Since $F(x_*) = 0$, there must exist $\lambda_* \in \mathbb{R}^n$, such that the first $2q$ coordinates of $\nabla R(x_*) + \sum_{i=1}^n \lambda_*^i \nabla f_i(x_*)$ are equal to 0. As argued in the proof of Lemma D.12, we can assume the first q' rows of Z are linear independent on the first q coordinates for some $q' \in [q]$. In other words, Z can be written as $\begin{bmatrix} Z_A & Z_B \\ 0 & Z_D \end{bmatrix}$ where $Z_A \in \mathbb{R}^{q' \times q}$. We further denote $\lambda_1 := \lambda^{1:q'}$, $\lambda_2 := \lambda^{(q'+1):n}$, $w_1 := w^{1:q}$ and $w_2 := w^{(q+1):d}$ for convenience, then we have

$$\|F'(w)\|_2^2 = \min_{\lambda \in \mathbb{R}^n} \left\langle (a_1 + \operatorname{sign}(w_1) \odot Z_A^\top \lambda_1)^{\odot 2}, |w_1| \right\rangle + \left\langle (a_2 + Z_B^\top \lambda_1 + Z_D^\top \lambda_2)^{\odot 2}, w_2 \right\rangle. \quad (57)$$

Since every w in $\overline{\Gamma}$ is a global minimizer, $R'(w) = R'(w) + \sum_{i=1}^n \lambda_*^i (z_i^\top w - y_i) := g^\top w$, where $g = \operatorname{sign}(w) \odot a + Z^\top \lambda_*$. Similarly we define $g_1 := g^{1:q}$ and $g_2 := g^{(q+1):d}$. It holds that $g_1 = 0$ and we assume $Z_D g_2 = 0$ without loss of generality, because this can always be done by picking suitable λ_*^i for $i = q'+1, \dots, n$.

We denote $\lambda_1 - \lambda_{*,1}$ by $\Delta\lambda_1$, then since $0 = g_1 = \operatorname{sign}(w_1) \odot a_1 + Z_A^\top \lambda_{*,1}$, we further have

$$\begin{aligned} \left\langle (a_1 + \operatorname{sign}(w_1) \odot Z_A^\top \lambda_1)^{\odot 2}, |w_1| \right\rangle &= \left\langle (a_1 + \operatorname{sign}(w_1) \odot Z_A^\top \lambda_{*,1} + \operatorname{sign}(w_1) \odot Z_A^\top \Delta\lambda_1)^{\odot 2}, |w_1| \right\rangle \\ &= \left\langle (\operatorname{sign}(w_1) \odot Z_A^\top \Delta\lambda_1)^{\odot 2}, |w_1| \right\rangle. \end{aligned}$$

On the other hand, we have $g_2 = \operatorname{sign}(w_2) \odot a_2 + Z_B^\top \lambda_{*,1} + Z_D^\top \lambda_{*,2} = a_2 + Z_B^\top \lambda_{*,1} + Z_D^\top \lambda_{*,2}$ by the assumption that each coordinate of w_2 is non-negative. Combining this with the above identity, we can rewrite Equation (57) as:

$$\|F'(w)\|_2^2 = \min_{\lambda \in \mathbb{R}^D} \left\langle (Z_A^\top \Delta\lambda_1)^{\odot 2}, |w_1| \right\rangle + \left\langle (g_2 + Z_B^\top \Delta\lambda_1 + Z_D^\top \lambda_2)^{\odot 2}, w_2 \right\rangle. \quad (58)$$

Now suppose $R'(w) - R'(w_*) = g_2^\top w_2 = \delta$ for some sufficiently small δ (which can be controlled by ϵ). We will proceed in the following two cases separately.

- **Case I.1**: $\|\Delta\lambda_1\|_2 = \Omega(\sqrt{\delta})$. Since Z_A has full row rank and every coordinate of w_1 is non-zero, the first term of Equation (58) is $\Omega(\delta) = \Omega(R'(w) - R'(w_*))$.
- **Case I.2**: $\|\Delta\lambda_1\|_2 = O(\sqrt{\delta})$. Let $u = g_2 + Z_B^\top \Delta\lambda_1 + Z_D^\top \lambda_2$, then we have $u \in S + B^2_{c\sqrt{\delta}}(0)$ for some constant $c > 0$, where $S = \{g_2 + Z_D^\top \lambda_2 \mid \lambda_2 \in \mathbb{R}^{n-q'}\}$. By Lemma D.14, there exists some constant $c_0 \geq 1$, such that $\frac{1}{c_0} \cdot S$ weakly dominates $S + B^2_{c\sqrt{\delta}}(0)$. Thus we have $\|F'(w)\|_2^2 \geq \inf_{u \in S + B^2_{c\sqrt{\delta}}(0)} \langle u^{\odot 2}, w_2 \rangle \geq \inf_{u \in \frac{1}{c_0} \cdot S} \langle s^{\odot 2}, w_2 \rangle$, where the last step is because each coordinate of w_2 is non-negative.

  Let A be the orthogonal complement of $\mathrm{span}(Z_D, g_2)$, we know $w_2 \in \frac{\delta}{\|g_2\|_2^2} g_2 + A$, since $Z_D w_2 = Z_D w_{*,2} = 0$ and $g_2^\top w_2 = \delta$. Therefore,

$$\begin{aligned}\inf_{w: R'(w) - R'(w_*) = \delta > 0} \frac{\|F'(w)\|_2^2}{R'(w) - R'(w_*)} \geq & \inf_{w_2: R'(w) - R'(w_*) = \delta > 0} \inf_{u \in \frac{1}{c_0} \cdot S} \left\langle u^{\odot 2}, \frac{w_2}{\delta} \right\rangle \\ \geq & \frac{1}{c_0^2} \inf_{w_2 \in \frac{\delta}{\|g_2\|_2^2} g_2 + A, w_2 \geq 0, u \in S} \langle u^{\odot 2}, w_2 \rangle . \end{aligned} \tag{59}$$

  Note $\langle u^{\odot 2}, w_2 \rangle$ is a monotone non-decreasing function in the first joint orthant, i.e., $\{(u, w_2) \in \mathbb{R}^d \times \mathbb{R}^{d-q'} \mid u \geq 0, w_2 \geq 0\}$, thus by Lemma D.15 the infinimum can be achieved by some finite (u, w_2) in the joint first orthant. Applying the same argument to each other orthant of $u \in \mathbb{R}^d$, we conclude that the right-hand-side of (59) can be achieved.

  On the other hand, we have $u^\top w_2 = \delta > 0$ for all $w_2 \in \frac{\delta}{\|g_2\|_2^2} g_2 + A$ and $u \in S$, by $Z_D g_2 = 0$ and the definition of A. This implies there exists at least one $i \in [d - q']$ such that $w_2^i u^i > 0$, which further implies $\langle u^{\odot 2}, w_2 \rangle > 0$. Therefore, we conclude that $\|F'(w)\|_2^2 = \Omega(R'(w) - R'(w_0))$.

Case II. Next, for any general $x = \binom{u}{v}$, we define $w = u^{\odot 2} - v^{\odot 2}$ and $m = \min\{u^{\odot 2}, v^{\odot 2}\}$, where $\min$ is taken coordinate-wise. Then we can rewrite $\|F(x)\|_2^2$ as

$$\begin{aligned}\|F(x)\|_2^2 &= \min_{\lambda \in \mathbb{R}^n} \left\| \left(\begin{bmatrix} a \\ a \end{bmatrix} + \begin{bmatrix} Z \\ -Z \end{bmatrix} \lambda \right) \odot \begin{bmatrix} u \\ v \end{bmatrix} \right\|_2^2 \\ &= \min_{\lambda \in \mathbb{R}^n} \left\| \left(\begin{bmatrix} a \\ a \end{bmatrix} + \begin{bmatrix} Z \\ -Z \end{bmatrix} \lambda \right)^{\odot 2} \odot \begin{bmatrix} u^{\odot 2} \\ v^{\odot 2} \end{bmatrix} \right\|_1 \\ &= \min_{\lambda \in \mathbb{R}^n} \left\| \left(\begin{bmatrix} a \\ a \end{bmatrix} + \begin{bmatrix} Z \\ -Z \end{bmatrix} \lambda \right)^{\odot 2} \odot \left(\psi(w)^{\odot 2} + \begin{bmatrix} m \\ m \end{bmatrix} \right) \right\|_1 \\ &\geq \min_{\lambda \in \mathbb{R}^n} \left\| \left(\begin{bmatrix} a \\ a \end{bmatrix} + \begin{bmatrix} Z \\ -Z \end{bmatrix} \lambda \right)^{\odot 2} \odot \psi(w)^{\odot 2} \right\|_1 + \min_{\lambda \in \mathbb{R}^n} \left\| \left(\begin{bmatrix} a \\ a \end{bmatrix} + \begin{bmatrix} Z \\ -Z \end{bmatrix} \lambda \right)^{\odot 2} \odot \begin{bmatrix} m \\ m \end{bmatrix} \right\|_1 \\ &= \min_{\lambda \in \mathbb{R}^n} \left\| \left(\begin{bmatrix} a \\ a \end{bmatrix} + \begin{bmatrix} Z \\ -Z \end{bmatrix} \lambda \right) \odot \psi(w) \right\|_2^2 + \min_{\lambda \in \mathbb{R}^n} \left\| \left(\begin{bmatrix} a \\ a \end{bmatrix} + \begin{bmatrix} Z \\ -Z \end{bmatrix} \lambda \right) \odot \begin{bmatrix} \sqrt{m} \\ \sqrt{m} \end{bmatrix} \right\|_2^2 . \end{aligned}$$

Then applying the result for the previous case yields the following for some constant $C \in (0, 1)$:

$$\begin{aligned}\|F(x)\|_2^2 &\geq C(R(\psi(w)) - R(\psi(w_*)) + \min_{\lambda \in \mathbb{R}^n} \left\| \left(\begin{bmatrix} a \\ a \end{bmatrix} + \begin{bmatrix} Z \\ -Z \end{bmatrix} \lambda \right) \odot \begin{bmatrix} \sqrt{m} \\ \sqrt{m} \end{bmatrix} \right\|_2^2 \\ &= C(R(\psi(w)) - R(x_*) + 2 \langle a^{\odot 2}, m \rangle \\ &\geq C(R(\psi(w)) - R(x_*) + 2 \min_{i \in [d]} a_i \langle a, m \rangle \\ &= C(R(\psi(w)) - R(x_*) + \min_{i \in [d]} a_i (R(x) - R(\psi(w))) \\ &\geq \min \left\{ C, \min_{i \in [d]} a_i \right\} (R(x) - R(x_*)), \end{aligned}$$

where the first equality follows from the fact that $x_* = \psi(w_*)$ and the last inequality is due to the fact that both $R(\psi(w)) - R(\psi(w_*))$ and $R(x) - R(\psi(w))$ are non-negative. This completes the proof. □

Now, based on the PL condition, we can show that (17) indeed converges.

Lemma D.17. *The trajectory of the flow defined in (17) has finite length, i.e., $\int_{t=0}^{\infty} \|\frac{\mathrm{d}x}{\mathrm{d}t}\|_2 \mathrm{d}t < \infty$ for any $x_0 \in \Gamma$. Moreover, x_t converges to some x_∞ when $t \to \infty$ with $F(x_\infty) = 0$.*

Proof of Lemma D.17. Note that along the Riemannian gradient flow, $R(x_t)$ is non-increasing, thus $\|x_t\|_2$ is bounded over time and $\{x_t\}_{t\geq 0}$ has at least one limit point, which we will call x_*. Therefore, $R(x_*)$ is a limit point of $R(x_t)$, and again since $R(x_t)$ is non-increasing, it follows that $R(x_t) \geq R(x_*)$ and $\lim_{t\to\infty} R(x_t) = R(x_*)$. Below we will show $\lim_{t\to\infty} x_t = x_*$.

Note that $\frac{\mathrm{d}R(x_t)}{\mathrm{d}t} = \langle \nabla R(x_t), \frac{\mathrm{d}x_t}{\mathrm{d}t} \rangle = -\langle \nabla R(x_t), F(x_t) \rangle = -\|F(x_t)\|_2^2$ where the last inequality applies Lemma D.10. By Lemma D.16, there exists a neighbourhood of x_*, U', in which PL condition holds of F. Since x_* is a limit point, there exists a time T_0, such that $x_{T_0} \in U$. Let $T_1 = \inf_{t\geq T_0}\{x(t) \notin U'\}$ (which is equal to ∞ if $x(t) \in U'$ for all $t \geq T_0$). Since x_t is continuous in t and U is open, we know $T_1 > T_0$ and for all $t \in [T_0, T_1)$, we have $\|F(x_t)\|_2 \geq \sqrt{c}(R(x_t) - R(x_*))^{1/2}$.

Thus it holds that for $t \in [T_0, T_1)$,

$$\frac{\mathrm{d}(R(x_t) - R(x_*))}{\mathrm{d}t} \leq -\sqrt{c}(R(x_t) - R(x_*))^{1/2} \|F(x_t)\|_2 ,$$

that is,

$$\frac{\mathrm{d}(R(x_t) - R(x_*))^{1/2}}{\mathrm{d}t} \leq -\frac{\sqrt{c}}{2} \|F(x_t)\|_2 .$$

Therefore, we have

$$\int_{t=T_0}^{T_1} \|F(x_t)\|_2 \,\mathrm{d}t \leq \frac{1}{2\sqrt{c}}(R(x_{T_0}) - R(x_*)). \tag{60}$$

Thus if we pick T_0 such that $R(x_{T_0}) - R(x_*)$ is sufficiently small, $R(T_1)$ will remain in U, which implies that T_1 cannot be finite and has to be ∞. Therefore, Equation (60) shows that the trajectory of x_t is of finite length, so $x_\infty = \lim_{t\to\infty} x_t$ exists and is equal to x_*. As a by-product, $F(x_*)$ must be 0. □

Finally, collecting all the above lemmas, we are able to prove Lemma 6.5.

Lemma 6.5. *Let $\{x_t\}_{t\geq 0} \subseteq \mathbb{R}^D$ be generated by the flow defined in (17) with any initialization $x_0 \in \Gamma$. Then $x_\infty = \lim_{t\to\infty} x_t$ exists. Moreover, $x_\infty = x_*$ is the optimal solution of (18).*

Proof of Lemma 6.5. We will prove by contradiction. Suppose $x_\infty = \binom{u_\infty}{v_\infty} = \lim_{t\to\infty} x_t$ is not the optimal solution to (18). Denote $w_t = u_t^{\odot 2} - v_t^{\odot 2}$, then $w_\infty = \lim_{t\to\infty} w_t$ is not the optimal solution to (36). Thus we have $R(w_t) > R(w_*)$. Without loss of generality, suppose there is some $q \in [d]$ such that $(u_\infty^i)^2 + (v_\infty^i)^2 > 0$ for all $i = 1, \ldots, q$ and $u_\infty^i = v_\infty^i = 0$ for all $i = q+1, \ldots, d$. Again, as argued in the proof of Lemma D.12, we can assume that, for some $q' \in [q]$,

$$\{z_i^{1:q}\}_{i\in[q']} \text{ is linearly independent and } z_i^{1:q} = 0 \text{ for all } i = q'+1, \ldots, n. \tag{61}$$

Since both w_∞ and w_* satisfy the constraint that $Zw_\infty = Zw_* = Y$, we further have

$$0 = \langle z_i, w_\infty \rangle = \langle z_i, w_* \rangle = \langle z_i^{(q+1):d}, w_*^{(q+1):d} \rangle, \quad \text{for all } i = q'+1, \ldots, n. \tag{62}$$

Consider a potential function $\varphi : U \to \mathbb{R}$ defined as

$$\varphi(x) = \varphi(u, v) = \sum_{j=q+1}^{d} w_*^j \left[\ln(u^j)^2 \mathbb{1}\{w_*^j > 0\} - \ln(v^j)^2 \mathbb{1}\{w_*^j < 0\}\right].$$

Clearly $\lim_{t\to\infty}\varphi(x_t) = -\infty$ if $\lim_{t\to\infty} x_t = x_\infty$. Below we will show contradiction if x_∞ is suboptimal. Consider the dynamics of $\varphi(x)$ along the Riemannian gradient flow:

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t}(x_t) = \left\langle \nabla\varphi(x_t), \frac{\mathrm{d}x_t}{\mathrm{d}t} \right\rangle = -\left\langle \nabla\varphi(x_t), F(x_t)\right\rangle \tag{63}$$

where F is defined previously in Lemma D.10. Recall the definition of F, and we have

$$\left\langle \nabla\varphi(x_t), F(x_t)\right\rangle = \underbrace{\left\langle \nabla\varphi(x_t), \nabla R(x_t) + \sum_{i=1}^{q'} \lambda(x_t)^i \nabla f_i(x_t)\right\rangle}_{\mathcal{I}_1} + \underbrace{\left\langle \nabla\varphi(x_t), \sum_{i=q'+1}^{n} \lambda(x_t)^i \nabla f_i(x_t)\right\rangle}_{\mathcal{I}_2}. \tag{64}$$

To show $\langle \nabla\varphi(x_t), F(x_t)\rangle < 0$, we analyze $\mathcal{I}_1$ and $\mathcal{I}_2$ separately. By the definition of $\varphi(x)$, we have

$$\nabla\varphi(x) = \sum_{j=q+1}^{d} 2w_*^j \left[\frac{\mathbb{1}\{w_*^j > 0\}}{u^j}\cdot e_j - \frac{\mathbb{1}\{w_*^j < 0\}}{v^j}\cdot e_{D+j}\right]$$

where e_j is the j-th canonical base of $\mathbb{R}^d$. Recall that $\nabla f_i(x) = \binom{z_i\odot u}{-z_i\odot v}$, and we further have

$$\begin{aligned}
\mathcal{I}_2 &= \sum_{i=q'+1}^{n} \lambda(x_t)^i \sum_{j=q+1}^{d} 2w_*^j \left[\frac{\mathbb{1}\{w_*^j > 0\}}{u^j}\langle e_j, z_i\odot u\rangle + \frac{\mathbb{1}\{w_*^j < 0\}}{v^j}\langle e_j, z_i\odot v\rangle\right] \\
&= \sum_{i=q'+1}^{n} \lambda(x_t)^i \sum_{j=q+1}^{d} 2w_*^j \left[\frac{\mathbb{1}\{w_*^j > 0\}}{u^j} z_i^j u^j + \frac{\mathbb{1}\{w_*^j < 0\}}{v^j} z_i^j v^j\right] \\
&= \sum_{i=q'+1}^{n} \lambda(x_t)^i \sum_{j=q+1}^{d} w_*^j z_i^j = \sum_{i=q'+1}^{n} \lambda(x_t)^i \langle z_i^{(q+1):d}, w_*^{(q+1):d}\rangle = 0
\end{aligned} \tag{65}$$

where the last equality follows from (62).

Next, we show that $\mathcal{I}_1 < 0$ by utilizing the fact that $w_* - w_\infty$ is a descent direction of $R'(w)$. For $w\in\mathbb{R}^d$, define $\tilde{f}_i(w) = z_i^\top w$ and

$$\tilde{R}(w) = R(w) + \sum_{i=1}^{q'} \lambda(x_\infty)^i (\tilde{f}_i(w) - y_i).$$

Clearly, for any $w\in\mathbb{R}^D$ satisfying $Zw = Y$, it holds that $\tilde{f}_i(w) - y_i = 0$ for each $i\in[n]$, and thus $R(w) = \tilde{R}(w)$. In particular, we have $\tilde{R}(w_\infty) = R(w_\infty) > R(w_*) = \tilde{R}(w_*)$. Since $\tilde{R}(w)$ is a convex function, it follows that $\tilde{R}(w_\infty + s(w_* - w_\infty)) < \tilde{R}(w_\infty)$ for all sufficiently small $s > 0$, which implies $\frac{\mathrm{d}\tilde{R}}{\mathrm{d}t}(w_\infty + s(w_* - w_\infty))|_{s=0} < -2c < 0$ for some constant $c > 0$. Note that, for small enough $s > 0$, we have

$$\begin{aligned}
R(w_\infty + s(w_* - w_\infty)) &= \sum_{j=1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)|w_\infty^j + s(w_*^j - w_\infty^j)| \\
&= \sum_{j=1}^{q}\left(\sum_{i=1}^{n}(z_i^j)^2\right)\mathrm{sign}(w_\infty^j)(w_\infty^j + s(w_*^j - w_\infty^j)) \\
&\quad + \sum_{j=q+1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right) s|w_*^j|.
\end{aligned}$$

Therefore, we can compute the derivative with respect to s at $s = 0$ as

$$
\begin{aligned}
-2c > \frac{\mathrm{d}\tilde{R}}{\mathrm{d}t}(w_\infty + s(w_* - w_\infty))\Big|_{s=0} &= \sum_{j=1}^{q}\left(\sum_{i=1}^{n}(z_i^j)^2\right)\operatorname{sign}(w_\infty^j)(w_*^j - w_\infty^j) + \sum_{j=q+1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)|w_*^j| \\
&\quad + \sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^\top (w_* - w_\infty^j) \\
&= \sum_{j=1}^{q}\left(\sum_{i=1}^{n}(z_i^j)^2\right)\operatorname{sign}(w_\infty^j)(w_*^j - w_\infty) + \sum_{j=q+1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)|w_*^j| \\
&\quad + \sum_{j=1}^{q}(w_*^j - w_\infty^j)\sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^j + \sum_{j=q+1}^{d} w_*^j \sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^j
\end{aligned} \tag{66}
$$

where the second equality follows from the fact that $w_\infty^{(q+1):d} = 0$. Since x_t converges to x_∞, we must have $F(x_\infty) = 0$, which implies that for each $j \in \{1, \ldots, q\}$,

$$
\begin{aligned}
0 = \frac{\partial R}{\partial u^j}(x_\infty) + \sum_{i=1}^{q'}\lambda(x_\infty)^i \frac{\partial f_i}{\partial u^j}(x_\infty) &= 2u_\infty^j \cdot \left[\sum_{i=1}^{n}(z_i^j)^2 + \sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^j\right], \\
0 = \frac{\partial R}{\partial v^j}(x_\infty) + \sum_{i=1}^{q'}\lambda(x_\infty)^i \frac{\partial f_i}{\partial v^j}(x_\infty) &= 2v_\infty^j \cdot \left[\sum_{i=1}^{n}(z_i^j)^2 - \sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^j\right].
\end{aligned}
$$

Combining the above two equalities yields

$$
\sum_{i=1}^{n}(z_i^j)^2 = -\operatorname{sign}(w_\infty^j)\sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^j, \quad \text{for all } j \in [q].
$$

Apply the above identity together with (66), and we obtain

$$
\begin{aligned}
-2c > \sum_{j=1}^{q} -\operatorname{sign}(w_\infty^j)^2(w_*^j - w_\infty)\sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^j + \sum_{j=q+1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)|w_*^j| \\
+ \sum_{j=1}^{q}(w_*^j - w_\infty^j)\sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^j + \sum_{j=q+1}^{d} w_*^j \sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^j \\
= \sum_{j=q+1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)|w_*^j| + \sum_{j=q+1}^{d} w_*^j \sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^j
\end{aligned} \tag{67}
$$

On the other hand, by directly evaluating $\nabla R(x_t)$ and each $\nabla f_i(x_t)$, we can compute $\mathcal{I}_1$ as

$$\begin{aligned}
\mathcal{I}_1 &= 2\sum_{j=q+1}^{d}\frac{w_*^j\mathbb{1}\{w_*^j>0\}}{u_t^j}\left[\sum_{i=1}^{n}(z_i^j)^2u_t^j+\sum_{i=1}^{q'}\lambda(x_t)^i z_i^j u_t^j\right] \\
&\quad -2\sum_{j=q+1}^{d}\frac{w_*^j\mathbb{1}\{w_*^j<0\}}{v_t^j}\left[\sum_{i=1}^{n}(z_i^j)^2v_t^j-\sum_{i=1}^{q'}\lambda(x_t)^i z_i^j v_t^j\right] \\
&= 2\sum_{j=q+1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)|w_*^j|+2\sum_{j=q+1}^{d}w_*^j\sum_{i=1}^{q'}\lambda(x_t)^i z_i^j \\
&= 2\sum_{j=q+1}^{d}\left(\sum_{i=1}^{n}(z_i^j)^2\right)|w_*^j|+2\sum_{j=q+1}^{d}w_*^j\sum_{i=1}^{q'}\lambda(x_\infty)^i z_i^j \\
&\quad +2\sum_{j=q+1}^{d}w_*^j\sum_{i=1}^{q'}\left(\lambda(x_t)^i-\lambda(x_\infty)^i\right)z_i^j.
\end{aligned}$$

We already know that $\lambda(x)^{1:q'}$ is continuous at x_∞ by the proof of Lemma D.12, so the third term converges to 0 as x_t tends to x_∞. Now, applying (67), we immediately see that there exists some $\delta>0$ such that $\mathcal{I}_1<-c$ for $x_t\in B_\delta(x_\infty)$. As we have shown in the above that $\mathcal{I}_2=0$, it then follows from (63) and (64) that

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t}(x_t)>c,\quad \text{for all } x_t\in B_\delta(x_\infty). \tag{68}$$

Since $\lim_{t\to\infty}x_t=x_\infty$, there exists some $T>0$ such that $x_t\in B_\delta(x_\infty)$ for all $t>T$. By the proof ofLemma D.13, we know that $\varphi(x_T)>-\infty$, then it follows from (68) that

$$\lim_{t\to\infty}\varphi(x_t)=\varphi(x_T)+\int_T^\infty\frac{\mathrm{d}\varphi(x_t)}{\mathrm{d}t}\mathrm{d}t>\varphi(x_T)+\int_T^\infty c\mathrm{d}t=\infty$$

which is a contradiction. This finishes the proof. □

D.6 Proof of Theorem 6.7

Here we present the lower bound on the sample complexity of GD in the kernel regime.

Theorem 6.7. *Assume $z_1,\dots,z_n\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,I_d)$ and $y_i=w_*^\top z_i$, for all $i\in[n]$. Define the loss with linearized model as $L(x)=\sum_{i=1}^n(f_i(x_0)+\langle\nabla f_i(x_0),x-x_0\rangle-y_i)^2$, where $x=\binom{u}{v}$ and $x_0=\binom{u_0}{v_0}=\alpha\binom{\mathbb{1}}{\mathbb{1}}$. Then for any ground truth w_*, any learning rate schedule $\{\eta_t\}_{t\geq1}$, and any fixed number of steps T, the expected ℓ_2 loss of x_T is at least $(1-\frac{n}{d})\|w_*\|_2^2$, where x_T is the T-th iterate of GD on L, i.e., $x_{t+1}=x_t-\eta_t\nabla L(x_t)$, for all $t\geq0$.*

Proof of Theorem 6.7. We first simplify the loss function by substituting $x'=x-x_0$, so correspondingly $x_0'=0$ and we consider $L'(x'):=L(x)=(\langle\nabla f_i(x_0),x'\rangle-y_i)^2$. We can think as if GD is performed on $L'(x')$. For simplicity, we still use the x and $L(x)$ notation in below.

In order to show test loss lower bound against a single fixed target function, we must take the properties of the algorithm into account. The proof is based on the observation that GD is rotationally equivariant (Ng, 2004; Li et al., 2020c) as an iterative algorithm, i.e., if one rotates the entire data distribution (including both the training and test data), the expected loss of the learned function remains the same. Since the data distribution and initialization are invariant under any rotation, it means the expected loss of x_T with ground truth being w_* is the same as the case where the ground truth is uniformly randomly sampled from all vectors of ℓ_2-norm $\|w_*\|_2$.

Thus the test loss of x_T is

$$\mathbb{E}_z\left[(\langle\nabla f_z(x_0),x_T\rangle-\langle z,w_*\rangle)^2\right]=\mathbb{E}_z\left[(\langle z,w_*-(u_T-v_T)\rangle)^2\right]=\|w_*-(u_T-v_T)\|_2^2. \tag{69}$$

Note $x_T \in \text{span}\{\nabla f_x(x_0)\}$, which is at most an n-dimensional space spanned by the gradients of model output at x_0, so is $u_T - v_T$. We denote the corresponding space for $u_T - v_T$ by S, so $\dim(S) \leq n$ and it holds that $\|w_* - (u_T - v_T)\|_2^2 \geq \|(I_D - P_S)w^*\|_2^2$, where P_S is projection matrix onto space S.

The expected test loss is lower bounded by

$$\begin{aligned}
\mathbb{E}_{w_*}\left[\mathbb{E}_{z_i}\left[\|w_* - (u_T - v_T)\|_2^2\right]\right] &= \mathbb{E}_{z_i}\left[\mathbb{E}_{w_*}\left[\|w_* - (u_T - v_T)\|_2^2\right]\right] \\
&\geq \min_{z_i} \mathbb{E}_{w_*}\left[\|(I_D - P_S)w^*\|_2^2\right] \\
&\geq \left(1 - \frac{n}{d}\right)\|w_*\|_2^2 .
\end{aligned}$$

□