Understanding Gradient Descent on the Edge of Stability in Deep Learning

Sanjeev Arora [*1] **Zhiyuan Li** [*1] **Abhishek Panigrahi** [*1]

Abstract

Deep learning experiments by Cohen et al. (2021) using deterministic Gradient Descent (GD) revealed an *Edge of Stability (EoS)* phase when learning rate (LR) and sharpness (*i.e.*, the largest eigenvalue of Hessian) no longer behave as in traditional optimization. Sharpness stabilizes around $2/\text{LR}$ and loss goes up and down across iterations, yet still with an overall downward trend. The current paper mathematically analyzes a new mechanism of implicit regularization in the EoS phase, whereby GD updates due to non-smooth loss landscape turn out to evolve along some deterministic flow on the manifold of minimum loss. This is in contrast to many previous results about implicit bias either relying on infinitesimal updates or noise in gradient. Formally, for any smooth function L with certain regularity condition, this effect is demonstrated for (1) *Normalized GD*, i.e., GD with a varying LR $\eta_t = \frac{\eta}{\|\nabla L(x(t))\|}$ and loss L; (2) GD with constant LR and loss $\sqrt{L - \min_x L(x)}$. Both provably enter the Edge of Stability, with the associated flow on the manifold minimizing $\lambda_1(\nabla^2 L)$. The above theoretical results have been corroborated by an experimental study.

1. Introduction

Traditional convergence analyses of gradient-based algorithms assume learning rate η is set according to the basic relationship $\eta < 2/\lambda$ where λ is the largest eigenvalue of the Hessian of the objective, called *sharpness*. *Descent Lemma* says that if this relationship holds along the trajectory of Gradient Descent, loss drops during each iteration. In deep learning where objectives are nonconvex and have multiple optima, similar analyses can show convergence towards stationary points and local minima. In practice, sharpness is unknown and η is set by trial and error. Since deep learning works, it has been generally assumed that this trial and error allows η to adjust to sharpness so that the theory applies. But recent empirical studies (Cohen et al., 2021; Ahn et al., 2022) showed compelling evidence to the contrary. On a variety of popular architectures and training datasets, GD with fairly small values of η displays following phenomena that they termed *Edge of Stability (EoS)*: (a) Sharpness rises beyond $2/\eta$, thus violating the above-mentioned relationship. (b) Thereafter sharpness stops rising but hovers noticeably above $2/\eta$ and even decreases a little. (c) Training loss behaves non-monotonically over individual iterations, yet consistently decreases over long timescales.

Note that (a) was already pointed out by Li et al. (2020b). Specifically, in modern deep nets, which use some form of normalization combined with weight decay, training to near-zero loss must lead to arbitrarily high sharpness. (However, Cohen et al. (2021) show that the EoS phenomenon appears even without normalization.) Phenomena (b), (c) are more mysterious, suggesting that GD with finite η is able to continue decreasing loss despite violating $\eta < 2/\lambda$, while at the same time regulating further increase in value of sharpness and even causing a decrease. These striking inter-related phenomena suggest a radical overhaul of our thinking about optimization in deep learning. At the same time, it appears mathematically challenging to analyze such phenomena, at least for realistic settings and losses (as opposed to toy examples with 2 or 3 layers). The current paper introduces frameworks for doing such analyses.

We start by formal definition of stableness, ensuring that if a point + LR combination is stable then a gradient step is guaranteed to decrease the loss by the local version of Descent Lemma.

Definition 1.1 (Stableness)**.** Given a loss function L, a parameter $x \in \mathbb{R}^d$ and LR $\eta > 0$ we define the *stableness* of L at (x, η) be $S_L(x, \eta) := \eta \cdot \sup_{0 \le s \le \eta} \lambda_1(\nabla^2 L(x - s\nabla L(x)))$. We say L is *stable* at (x, η) iff the stableness of L at (x, η) is smaller than or equal to 2; otherwise we say L is *unstable* at (x, η).

The above defined stableness is a better indicator for EoS than only using the sharpness at a specific point x, *i.e.* $\eta\lambda_1(\nabla^2 L(x)) < 2$, because the loss can still oscillate in the

[*]Alphabetical order [1] Princeton University
. Correspondence to: Sanjeev Arora <arora@cs.princeton.edu>, Zhiyuan Li <zhiyuanli@cs.princeton.edu>, Abhishek Panigrahi <ap34@cs.princeton.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

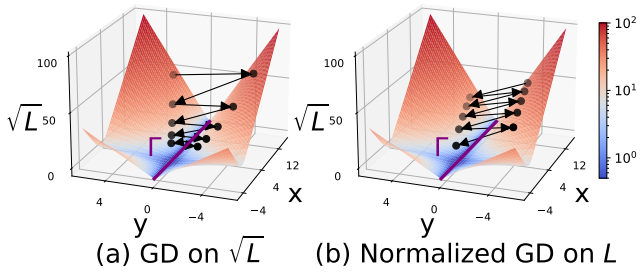

Figure 1: GD operating on EoS oscillates around the zero loss manifold $\Gamma = \{(x, y) \mid y = 0\}$ while slowly moving towards flatter local minima. Here $L(x, y) = (1 + x^2)y^2$ and the sharpness of L decreases as $|x|$ decreases.

latter case. [1] A concrete example is $L(x) = |x|, x \in \mathbb{R}$. For any $c \in (0, 1)$ and LR $\eta > 0$, the GD iterates $x(2k) = c\eta$ and $x(2k+1) = -(1-c)\eta$, always have zero sharpness for all $k]in\mathbb{N}$, but Descent Lemma doesn't apply because the gradient is not continuous around $x = 0$ (*i.e.* the sharpness is infinity when $x = 0$). As a result, the loss is not stable and oscillates between c and $1 - c$.

1.1. Two Provable Mechanisms for Edge of Stability: Non-smoothness and Adaptivity

In this paper we identify two settings where GD provably operates on Edge of Stability. The intuition is from Definition 1.1, which suggests that either sharpness or learning rate has to increase to avoid convergence and to ensure that GD stays on Edge of Stability.

The first setting, which is simple yet quite general, is to consider a modified training loss $f(L)$ where $f : \mathbb{R} \to \mathbb{R}$ is a monotone increasing but non-smooth function. For concreteness, assume GD is performed on $\widetilde{L} := \sqrt{L}$ where L is a smooth loss function with $\min_x L(x) = 0$ and $\nabla^2 L \neq 0$ at its minimizers. Note that $\nabla\widetilde{L} = \frac{\nabla L}{2\sqrt{L}}$ and $\nabla^2\widetilde{L} = \frac{2L\nabla^2 L - \nabla L \nabla L^\top}{4\sqrt{L^3}}$, which implies $\nabla^2\widetilde{L}$ must diverge whenever x converges to any minimizer where $\nabla^2 L$ has rank at least 2, since $\nabla L \nabla L^\top$ is rank-1. (An analysis is also possible when $\nabla^2 L$ is rank-1, which is the reason for Definition 1.1.)

The second setting assumes that the loss is smooth but learning rate is effectively adaptive. We focus a concrete example, *Normalized Gradient Descent*, $x \leftarrow x - \eta\nabla L/\|\nabla L\|$, which exhibits EoS behavior as $\nabla L \to 0$. We can view Normalized GD as GD with a varying LR $\eta_t = \frac{\eta}{\|\nabla L(x(t))\|}$, which goes to infinity when $\nabla L \to 0$.

These analyses will require (1) $\Gamma = \{x \mid L(x) = 0\}$ [2] is a $(D - M)$ dimensional submanifold of $\mathbb{R}^D$ for some $1 \leq M \leq D$ and (2) $\nabla^2 L(x)$ is rank-M for any $x \in \Gamma$. Note that while modern deep learning evolved using non-differentiable losses, the recent use of activations such as Swish (Ramachandran et al., 2017) instead of ReLU has allowed differentiable losses without harming performance.

Our Contribution: We show that Normalized GD on L (Section 4.3) and GD on $\sqrt{L}$ (Section 4.4) exhibit similar two-phase dynamics with sufficiently small LR η. In the first phase, GD tracks gradient flow (GF), with a monotonic decrease in loss until getting $\mathcal{O}(\eta)$-close to the manifold (Theorems 4.3 and 4.5) and the stableness becomes larger than 2. In the second phase, GD no longer tracks GF and loss is not monotone decreasing due to the high stableness. Repeatedly overshooting, GD iterate jumps back and forth across the manifold while moving slowly along the direction in the tangent space of the manifold which decreases the sharpness. (See Figure 1 for a graphical illustration) Formally, we prove when $\eta \to 0$, the trajectory of GD converges to some limiting flow on the manifold (Theorems 4.4 and 4.6). We further prove that in both settings GD in the second phase operates on EOS, and loss decreases in a non-monotone manner. Formally, we show that the average stableness over any two consecutive steps is at least 2 and that the average of $\sqrt{L}/\eta$ over two consecutive is proportional to sharpness or square root of sharpness (Theorems 4.7 and 4.8).

Though many works have suggested (primarily via experiments and some intuition) that the training algorithm in deep learning implicitly selects out solutions of low sharpness in some way, we are not aware of a formal setting where this had ever been made precise. Note that our result requires no stochasticity as in SGD (Li et al., 2022), though we need to inject tiny noise (*e.g.,* of magnitude $O(\eta^{100})$) to GD iterates occasionally (Algorithms 1 and 3). We believe that this is due the technical limitation of our current analysis and can be relaxed with a more advanced analysis. Indeed, in experiments, our theoretical predictions hold for the deterministic GD directly without any perturbation.

Novelty of Our Analysis: Our analysis is inspired by the mathematical framework of studying limiting dynamics of SGD around manifold of minimizers by Li et al. (2022), where the high-level idea is to introduce a projection function Φ mapping the current iterate x_t to the manifold and it suffices to understand the dynamics of $\Phi(x_t)$. It turns out that the one-step update of $\Phi(x_t)$ depends on the second moment of (stochastic) gradient at x_t, $\mathbb{E}[\nabla L(x_t)(\nabla L(x_t))^\top]$. While for SGD the second moment converges to the covariance matrix of stochastic gradient (Li et al., 2022) as

[1] See such experiments (e.g., ReLU CNN (+BN), Figure 75) in Appendix of in Cohen et al. (2021).

[2] Without loss of generality, we assume $\min_{x'} L(x') = 0$ throughout the paper. The main results for Normalized GD still hold if we relax the assumption and only assume Γ to be a manifold of local minimizers. For GD on $\sqrt{L}$, we need to replace $\sqrt{L}$ by $\sqrt{L - L_{\min}}$ where $L_{\min}$ is the local minimum.

x_t gets close to the manifold when $\eta \to 0$, for GD operating on EOS, $\nabla\sqrt{L}(x_t)$ or $\frac{\nabla L(x_t)}{\|\nabla L(x_t)\|}$ is non-smooth and not even defined at the manifold of the minimizers! To show $\Phi(x_t)$ moves in the direction which decreases the sharpness, the main technical difficulty is to show that $\nabla\sqrt{L}(x_t)$ or $\frac{\nabla L(x_t)}{\|\nabla L(x_t)\|}$ aligns to the top eigenvector of the Hessian $\nabla^2 L(x_t)$ and then the analysis follows from the framework by Li et al. (2022).

To prove the alignment between the gradient and the top eigenvector of Hessian, it boils down to analyzing Normalized GD on quadratic functions (2), which to the best of our knowledge has not been studied before. The dynamics is like chaotic version of power iteration, and we manage to show that the iterate will always align to the top eigenvector of Hessian of the quadratic loss. The proof is based on identifying a novel potential (Section 3) and might be of independent interest.

2. Related Works

Sharpness: Low sharpness has long been related to flat minima and thus to good generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2016). Recent study on predictors of generalization (Jiang* et al., 2020) does show sharpness-related measures as being good predictors, leading to SAM algorithm that improves generalization by explicitly controlling a parameter related to sharpness (Foret et al., 2021). However, Dinh et al. (2017) show that due to the positive homogeneity in the network architecture, networks with rescaled parameters can have very different sharpness yet be the same to the original one in function space. This observation weakens correlation between sharpness and and generalization gap and makes the definition of sharpness ambiguous. In face of this challenge, multiple notions of scale-invariant sharpness have been proposed (Yi et al., 2019a;b; Tsuzuku et al., 2020; Rangamani et al., 2021). Especially, Yi et al. (2021); Kwon et al. (2021) derived new algorithms with better generalization by explicitly regularizing new sharpness notions aware of the symmetry and invariance in the network. He et al. (2019) goes beyond the notion of sharpness/flatness and argues that the local minima of modern deep networks can be asymmetric, that is, sharp on one side, but flat on the other side.

Limiting Diffusion/Flow around Manifold of Minimizers: The idea of analyzing the behavior of SGD with small LR along the the manifold originates from (Blanc et al., 2020), which gives a local analysis on a special noise type named label noise, *i.e.* noise covariance is equal to Hessian at minimizers. Damian et al. (2021) extends this analysis and show SGD with label noise finds approximate stationary point for original loss plus some Hessian-related regularizer. The formal mathematical framework of approximating the limiting dynamics of SGD with arbitrary noise by Stochastic Differential Equations is later established by Li et al. (2022), which is built on the convergence result for solutions of SDE with large-drift (Katzenberger, 1991).

Implicit Bias: The notion that training algorithm plays an active role in selecting the solution (when multiple optima exist) has been termed the *implicit bias* of the algorithm (Gunasekar et al., 2018c) and studied in a large number of papers (Soudry et al., 2018; Li et al., 2018; Arora et al., 2018; 2019a; Gunasekar et al., 2018b;a; Lyu & Li, 2020; Li et al., 2020a; Woodworth et al., 2020; Razin & Cohen, 2020; Lyu et al., 2021; Azulay et al., 2021; Gunasekar et al., 2021). In the infinite width limit, the implicit bias of Gradient Descent is shown to be the solution with the minimal RKHS norm with respect to the Neural Tangent Kernel (NTK) (Jacot et al., 2018; Li & Liang, 2018; Du et al., 2019; Arora et al., 2019b;c; Allen-Zhu et al., 2019b;a; Zou et al., 2020; Chizat et al., 2019; Yang, 2019). The implicit bias results from these papers are typically proved by performing a trajectory analysis for (Stochastic) Gradient Descent. Most of the results can be directly extended to the continuous limit (i.e., GD infinitesimal LR) and even some heavily relies on the conservation property which only holds for the continuous limit. In sharp contrast, the implicit bias shown in this paper – reducing the sharpness along the minimizer manifold – requires finite LR and doesn't exist for the corresponding continuous limit. Other implicit bias results that fundamentally relies on the finiteness of LR includes stability analysis (Wu et al., 2017; Ma & Ying, 2021) and implicit gradient regularization (Barrett & Dherin, 2021), which is a special case of approximation results for stochastic modified equation by Li et al. (2017; 2019).

3. Warm-up: Quadratic Loss Functions

To introduce ideas that will be used in the main results, we sketch analysis of Normalized GD (1) on quadratic loss function $L(x) = \frac{1}{2}x^\top Ax$ where $A \in \mathbb{R}^{D\times D}$ is positive definite with eigenvalues $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_D$ and $v_1, \ldots, v_D$ are the corresponding eigenvectors.

$$x(t+1) = x(t) - \eta \frac{\nabla L(x(t))}{\|\nabla L(x(t))\|} = x(t) - \eta \frac{Ax(t)}{\|Ax(t)\|}. \quad (1)$$

Our main result (3.1) is that the iterates of Normalized GD $x(t)$ will converge to v_1 in direction, from which the loss oscillation (3.2) follows. Define $\widetilde{x}(t) = \frac{Ax(t)}{\eta}$, and the following update rule (2) holds. The convergence of $\widetilde{x}_t$ to v_1 in direction implies the convergence of x_t as well.

$$\widetilde{x}(t+1) = \widetilde{x}(t) - A\frac{\widetilde{x}(t)}{\|\widetilde{x}(t)\|}. \quad (2)$$

Theorem 3.1. *If* $|\langle v_1, \widetilde{x}(t)\rangle| \neq 0$, $\forall t \geq 0$, *then there exists* $0 < C < 1$ *and* $s \in \{\pm 1\}$ *such that* $\lim_{t\to\infty} \widetilde{x}(2t) = Cs\lambda_1 v_1$ *and* $\lim_{t\to\infty} \widetilde{x}(2t+1) = (C-1)s\lambda_1 v_1$.

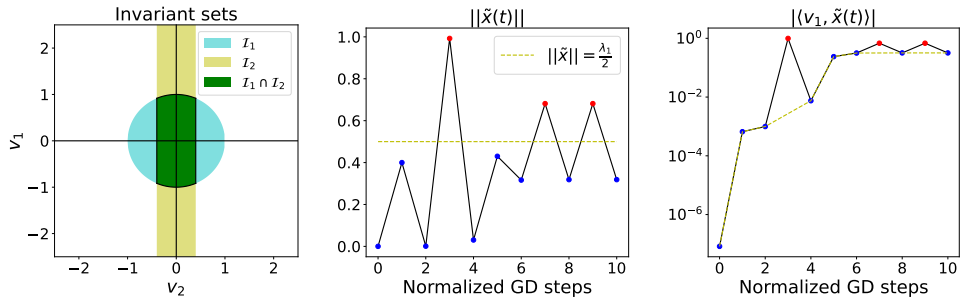

Figure 2: Visualization of key concepts and lemmas in the analysis for Normalized GD on a 2D quadratic loss with $\lambda_1 = 1, \lambda_2 = 0.4$. **Left**: invariant sets (defined in Lemma 3.3). **Middle**: $\|\widetilde{x}(t)\|$ drops below $\frac{\lambda_1}{2}$ in the next step whenever it is above $\frac{\lambda_1}{2}$ (Lemma 3.5). **Right**: $|\langle v_1, \widetilde{x}(t)\rangle|$ monotone increases among all the steps with norm below $\frac{\lambda_1}{2}$. (Lemma 3.6)

As a direct corollary, the loss oscillates as between time step $2t$ and time step $2t+1$ as $t \to \infty$. This shows that the behavior of loss is not monotonic and hence indicates the edge of stability phenomena for the quadratic loss.

Corollary 3.2. *If* $|\langle v_1, \widetilde{x}(t)\rangle| \neq 0, \forall t \geq 0$, *then there exists* $0 < C < 1$ *such that* $\lim_{t\to\infty} L(x(2t)) = \frac{1}{2}C^2\lambda_1\eta^2$ *and* $\lim_{t\to\infty} L(x(2t+1)) = \frac{1}{2}(C-1)^2\lambda_1\eta^2$.

We analyse the trajectory of the iterate $\widetilde{x}(t)$ in two phases. For convenience, we define $P^{(j:D)}$ as the projection matrix into the space spanned by $\{v_i\}_{i=j}^D$, *i.e.*, $P^{(j:D)} := \sum_{i=j}^D v_i v_i^\top$. In the first *preparation phase*, $\widetilde{x}(t)$ enters the intersection of D invariant sets $\{\mathcal{I}_j\}_{j=1}^D$ around the origin. (Lemma 3.3) In the second *alignment phase*, the projection of $\widetilde{x}(t)$ on the top eigenvector, $|\langle \widetilde{x}(t), v_1\rangle|$, is shown to increase monotonically among the steps among the steps $\{t \in \mathbb{N} \mid \|\widetilde{x}(t)\| \leq 0.5\lambda_1\}$. Since it is bounded, it must converge. The vanishing increment over steps turns out to suggest the $\widetilde{x}(t)$ must converge to v_1 in direction.

Lemma 3.3 (Preparation Phase)**.** *For any* $j \in [D]$ *and* $t \geq \frac{\lambda_1}{\lambda_j}\ln\frac{\lambda_1}{\lambda_j} + \max\{\frac{\|\widetilde{x}(0)\|-\lambda_1}{\lambda_D}, 0\}$, *it holds that* $\widetilde{x}(t) \in \mathcal{I}_j$, *where* $\mathcal{I}_j := \{\widetilde{x} \mid \left\|P^{(j:D)}\widetilde{x}\right\| \leq \lambda_j\}$.

Proof of Lemma 3.3. First, we show for any $j \in [D]$, $\mathcal{I}_j$ is indeed an invariant set for update rule (2) via Lemma A.1. With straightforward calculation, one can show that for any $j \in [D]$, $\left\|P^{(j:D)}\widetilde{x}(t)\right\|$ decreases by $\frac{\lambda_D\left\|P^{(j:D)}\widetilde{x}(t)\right\|}{\|\widetilde{x}(t)\|}$ if $\left\|P^{(j:D)}\widetilde{x}(t)\right\| \geq \lambda_j$ (Lemma A.2). Setting $j = 1$, we have $\|\widetilde{x}(t)\|$ decreases by λ_D if $\|\widetilde{x}(t)\| \geq \lambda_1$ (Corollary A.3). Thus for all $t \geq \max\{\frac{\|\widetilde{x}(0)\|-\lambda_1}{\lambda_D}, 0\}$, $\widetilde{x}(t) \in \mathcal{I}_1$. Finally once $\widetilde{x}(t) \in \mathcal{I}_1$, we can upper bound $\|\widetilde{x}(t)\|$ by λ_1, and thus $\left\|P^{(j:D)}\widetilde{x}(t)\right\|$ shrinks at least by a factor of $\frac{\lambda_D}{\lambda_1}$ per step, which implies $\widetilde{x}(t)$ will be in $\mathcal{I}_j$ in another $\frac{\lambda_1}{\lambda_j}\ln\frac{\lambda_1}{\lambda_j}$ steps.(Corollary A.4) □

Once the component of $\widetilde{x}(t)$ on an eigenvector becomes 0, it stays 0. So without loss of generality we can assume that after the preparation phase, the projection of $\widetilde{x}(t)$ along the top eigenvector v_1 is non-zero, otherwise we can study the problem in the subspace excluding the top eigenvector.

Lemma 3.4 (Alignment Phase)**.** *If* $\widetilde{x}(T) \in \cap_{j=1}^D \mathcal{I}_j$ *holds for some* T, *then for any* t', t *such that* $T \leq t \leq t'$ *and* $\|\widetilde{x}(t)\| \leq 0.5\lambda_1$, *it holds* $|\langle v_1, \widetilde{x}(t)\rangle| \leq |\langle v_1, \widetilde{x}(t')\rangle|$.

Proof of Lemma 3.4. First, Lemma 3.5 (proved in Appendix A) shows that the norm of the iterate $\widetilde{x}(t)$ remains above $0.5\lambda_1$ for only one time-step.

Lemma 3.5. *For any* t *with* $\widetilde{x}(t) \in \cap_{j=1}^D \mathcal{I}_j$, *if* $\|\widetilde{x}(t)\| > \lambda_1/2$, *then* $\|\widetilde{x}(t+1)\| \leq \max\left(\frac{\lambda_1}{2} - \frac{\lambda_D^2}{2\lambda_1}, \lambda_1 - \|\widetilde{x}(t)\|\right)$.

Thus, for any t with $\widetilde{x}(t) \in \cap_{j=1}^D \mathcal{I}_j$ and $\|\widetilde{x}(t)\| \leq \frac{\lambda_1}{2}$, either $\|\widetilde{x}(t+1)\| \leq \frac{\lambda_1}{2}$, or $\|\widetilde{x}(t+1)\| > \frac{\lambda_1}{2}$, which in turn implies that $\|\widetilde{x}(t+2)\| \leq \frac{\lambda_1}{2}$ by Lemma 3.5. The proof of Lemma 3.4 is completed by induction on Lemma 3.6.

Lemma 3.6. *For any step* t *with* $\|\widetilde{x}(t)\| \leq \lambda_1/2$, *for any* $k \in \{1, 2\}$, $|\langle v_1, \widetilde{x}(t+k)\rangle| \geq |\langle v_1, \widetilde{x}(t)\rangle|$.

For simplicity, we defer the proof of Lemma 3.6 into Appendix A. Proof of case $k = 1$ in Lemma 3.6 follows directly from plugging the assumption $\|\widetilde{x}(t)\| \leq \frac{\lambda_1}{2}$ into (2) (See Lemma A.5). The case of $k = 2$ in Lemma 3.6 follows from Lemma A.7. □

To complete the proof for Theorem 3.1, we relate the increase in the projection along v_1 at any step t, $|\langle v_1, \widetilde{x}(t)\rangle|$, to the magnitude of the angle between $\widetilde{x}(t)$ and the top eigenspace, θ_t. Briefly speaking, we show that if $\|\widetilde{x}(t)\| \leq \frac{\lambda_1}{2}$, $|\langle v_1, \widetilde{x}(t)\rangle|$ has to increase by a factor of $\Theta(\theta_t^2)$ in two steps. Since $|\langle v_1, \widetilde{x}(t)\rangle|$ is bounded and monotone increases among $\{t \mid \|\widetilde{x}(t)\| \leq \frac{\lambda_1}{2}\}$ by Lemma 3.4, we conclude that θ_t gets arbitrarily small for sufficiently large t with $\|\widetilde{x}(t)\| \leq \frac{\lambda_1}{2}, \|\widetilde{x}(t+2)\| \leq \frac{\lambda_1}{2}$ satisfied. Since the one-step normalized GD update ((2)) is continuous when bounded

away from origin, with a careful analysis, we conclude $\theta_t \to 0$ for all iterates. Please see Appendix A.3 for details.

Equivalence to GD on $\sqrt{\frac{1}{2}x^\top Ax}$: We can show that GD on loss $\sqrt{L}(x) = \sqrt{\frac{1}{2}x^\top Ax}$, follows the same update rule as Normalized GD on $L(x) = \frac{1}{2}x^\top Ax$, up to a linear transformation, because $x(t+1) = x(t) - \eta\nabla\sqrt{L}(x(t)) = x(t) - \eta\frac{Ax(t)}{\sqrt{2x(t)^\top Ax(t)}}$. Denoting $\widetilde{x}(t) = \frac{1}{\eta}(2A)^{1/2}x(t)$, we can easily check $\widetilde{x}(t)$ also satisfies update rule (2).

4. Main Results

In this section we present the main results of this paper. Section 4.1 is for preliminary and notations. In Section 4.2, we make two key assumptions for our analysis.

4.1. Preliminary and Notations

For any integer k, we denote $\mathcal{C}^k$ as the set of the k times continuously differentiable functions. For any mapping F, we use $\partial F(x)[u]$ and $\partial^2 F(x)[u,v]$ to denote the first and second order directional derivative of F at x along the derivation of u (and v). Given the loss function L, the gradient flow (GF) governed by L can be described through a mapping $\phi : \mathbb{R}^D \times [0,\infty) \to \mathbb{R}^D$ satisfying $\phi(x,\tau) = x - \int_0^\tau \nabla L(\phi(x,s))\mathrm{d}s$. We further define the limiting map of gradient flow as $\Phi : \mathbb{R}^D \to \mathbb{R}^D$, $\Phi(x) = \lim_{\tau\to\infty}\phi(x,\tau)$. We define $B_x(r)$ for any $r \in \mathbb{R}$ and $x \in \mathbb{R}^D$ as the uniform distribution over the set $\{y \in \mathbb{R}^D \mid \|x-y\|_2 \le r\}$.

For a matrix $A \in \mathbb{R}^{D\times D}$, we denote its eigenvalue-eigenvector pairs by $\{\lambda_i(A), v_i(A))\}_{i\in[D]}$. For simplicity, whenever Φ is defined at point x, we use $\{(\lambda_i(x), v_i(x))\}_{i=1}^D$ to denote the eigenvector-eigenvalue pairs of $\nabla^2 L(\Phi(x))$, with $\lambda_1(x) > \lambda_2(x) \ge \lambda_3(x) \ldots \ge \lambda_D(x)$. When the iterates $x(t)$ is clear in the context, we also use shorthand $\lambda_i(t) := \lambda_i(x(t))$, $v_i(t) := v_i(x(t))$, and $\theta_t \in [0, \frac{\pi}{2}]$ to denote the angle between $\nabla^2 L(\Phi(x(t)))(x(t) - \Phi(x(t)))$ and top eigenspace of $\nabla^2 L(\Phi(x(t)))$. Given a differentiable submanifold Γ of $\mathbb{R}^D$ and point $x \in \Gamma$, we use $P_{x,\Gamma} : \mathbb{R}^D \to \mathbb{R}^D$ to denote the projection operator onto the normal space of Γ at x, and $P^\perp_{x,\Gamma} := I_D - P_{x,\Gamma}$. As before, for convenience, we use the shorthand $P_{t,\Gamma} := P_{\Phi(x(t)),\Gamma}$ and $P^\perp_{t,\Gamma} := P^\perp_{\Phi(x(t)),\Gamma}$.

In this section, we focus on the setting where LR η goes to 0 and we fix the initialization x_{init} and the loss function L throughout this paper. We use $\mathcal{O}(\cdot)$ to hide constants about x_{init} and L.

4.2. Key Assumptions on Manifold of Local Minimizers

Following Fehrman et al. (2020); Li et al. (2022), we make the following two assumptions throughout the paper.

Assumption 4.1. Assume that the loss $L : \mathbb{R}^D \to \mathbb{R}$ is a $\mathcal{C}^4$ function, and that Γ is a $(D-M)$ dimensional $\mathcal{C}^2$-submanifold of $\mathbb{R}^D$ for some integer $1 \le M \le D$, where for all $x \in \Gamma, x$ is a local minimizer of L with $L(x) = 0$ and $\mathrm{rank}\left(\nabla^2 L(x)\right) = M$.

Assumption 4.2. Assume that U is an open neighborhood of Γ satisfying that gradient flow w.r.t. L starting in U converges to some point in Γ, i.e. for all $x \in U$, $\Phi(x) \in \Gamma$. (Then Φ is $\mathcal{C}^3$ on U (Falconer, 1983)).

The smoothness assumption is satisfied for networks with smooth activation functions like tanh and GeLU. The existence of manifold is due to the vast overparametrization in modern deep networks and preimage theorem. (See a discussion in section 3.1 of Li et al. (2022)) We also assume U is an open neighborhood of Γ such that gradient flow starting from every point in U converges to Γ.

4.3. Results for Normalized GD

We first denote the iterates of Normalized GD with LR η by $x_\eta(t)$, with $x_\eta(0) \equiv x_{\text{init}}$ for all η:

$$\text{Normalized GD: } x_\eta(t+1) = x_\eta(t) - \eta\frac{\nabla L(x_\eta(t))}{\|\nabla L(x_\eta(t))\|}. \quad (3)$$

The first theorem demonstrates the movement in the manifold, when the iterate travels from x_{init} to a position that is $\mathcal{O}(\eta)$ distance closer to the manifold (more specifically, $\Phi(x_{\text{init}})$). Moreover, just like the result in the quadratic case, we have more fine-grained bounds on the projection of $x_\eta(t) - \Phi(x_\eta(t))$ into the bottom-k eigenspace of $\nabla^2 L(\Phi(x_\eta(t)))$ for every $k \in [D]$. For convenience, we will denote the quantity $\sqrt{\sum_{i=j}^M \lambda_i^2(x)\langle v_i(x), x - \Phi(x)\rangle^2} - \lambda_j(x)\eta$ by $R_j(x)$ for all $j \in [M]$ and $x \in U$. In the quadratic case, Lemma 3.3 shows that $R_j(x)$ will eventually become non-positive for normalized GD iterates. Similarly, for the general loss, the following theorem shows that $R_j(x_\eta(t))$ eventually becomes approximately non-positive (smaller than $\mathcal{O}(\eta^2)$) in $\mathcal{O}(\frac{1}{\eta})$ steps.

Theorem 4.3 (Phase I). *Let $\{x_\eta(t)\}_{t\in\mathbb{N}}$ be the iterates of Normalized GD (3) with LR η and $x_\eta(0) = x_{init} \in U$. There is $T_1 > 0$ such that for any $T_1' > T_1$, it holds that for sufficiently small η that (1) $\max_{T_1 \le \eta t \le T_1'} \|x_\eta(t) - \Phi(x_{init})\| \le O(\eta)$ and (2) $\max_{T_1 \le \eta t \le T_1', j\in[D]} R_j(x_\eta(t)) \le O(\eta^2)$.*

Our main contribution is the analysis for the second phase (Theorem 4.4), which says just like the quadratic case, the angle between $\nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))$ and the top eigenspace of $\nabla^2 L(\Phi(x_\eta(t)))$, will be $\mathcal{O}(\eta)$ on average. As a result, the dynamics of Normalized GD tracks the riemannian gradient flow with respect to $\log(\lambda_1(\nabla^2 L(\cdot)))$

Algorithm 1 Perturbed Normalized Gradient Descent

Input: loss function $L : \mathbb{R}^D \to \mathbb{R}$, initial point x_{init}, maximum number of iteration T, LR η, frequency parameter $T_{\text{freq}} = \Theta(\eta^{-0.1})$, noise parameter $r = \Theta(\eta^{100})$.
for $t = 1$ **to** T **do**
  Generate $n(t) \sim B_0(r)$ if $t \mod T_{\text{freq}} = 0$, else set $n(t) = 0$.
  $x(t) \leftarrow x(t-1) - \eta \frac{\nabla L(x(t))}{\|\nabla L(x(t))\|} + n(t)$.
end for

on manifold, that is, the unique solution of (4).

$$X(\tau) = \Phi(x_{\text{init}}) - \frac{1}{4}\int_{s=0}^{\tau} P^{\perp}_{X(s),\Gamma} \nabla \log \lambda_1(X(s)) \mathrm{d}s \quad (4)$$

Note (4) is not guaranteed to have a well-defined solution for all $t \geq 0$, for the following two reasons: (1) if the multiplicity of top eigenvalue is larger than 1, $\lambda_1(\nabla^2 L(\cdot))$ may not be differentiable and (2) the projection matrix is only defined on Γ and the equation becomes undefined when the solution leaves Γ. We define T_2^{fav} as the set of all time $T_2 > 0$ such that for any $T_2 \in T_2^{\text{fav}}$, (4) satisfies: for any $0 \leq \tau \leq T_2$, we have (1) $X(\tau) \in U$, and (2) $\lambda_1(\nabla^2 L(X(\tau))) - \lambda_2(\nabla^2 L(X(\tau))) > 0$.

For a rigorous characterization of the dynamics in the second phase, we need to make the following modifications: (1). we add negligible noise of magnitude $\mathcal{O}(\eta^{100})$ every $\eta^{-0.1}$ steps; (2). we assume for each $\eta > 0$, there exist some step $t = \Theta(1/\eta)$ in phase I, except the guaranteed condition (1) and (2) (by Theorem 4.3), the additional condition (3) also holds. This assumption is mild because we only require (3) to hold for one step among $\Theta(1/\eta)$ steps from $\frac{T_1}{\eta}$ to $\frac{T_1'}{\eta}$, where T_1 is the constant given by Theorem 4.3 and T_1' is arbitrary constant larger than T_1. This assumption also holds empirically for all our experiments in Section 6.

Theorem 4.4 (Phase II). *Let $\{x_\eta(t)\}_{t\in\mathbb{N}}$ be the iterates of perturbed Normalized GD (Algorithm 1) with LR η. If the initialization $x_\eta(0)$ satisfy that*
(1) $\|x_\eta(0) - \Phi(x_{init})\| \leq O(\eta)$,
(2) $\max_{j\in[D]} R_j(x_\eta(0)) \leq O(\eta^2)$, and additionally
(3) $\min\{|\langle v_1(x_\eta(0)), x_\eta(0) - \Phi(x_\eta(0))\rangle|, -R_1(x_\eta(0))\} \geq \Omega(\eta)$, then for any time $T_2 \in T_2^{\text{fav}}$, it holds for sufficiently small η, with probability at least $1 - O(\eta^{10})$, that $\left\|\Phi(x_\eta(\lfloor T_2/\eta^2 \rfloor)) - X(T_2)\right\| = O(\eta)$ and $\frac{1}{\lfloor T_2/\eta^2 \rfloor}\sum_{t=0}^{\lfloor T_2/\eta^2 \rfloor} \theta_t \leq \mathcal{O}(\eta)$.

4.4. Results for GD on $\sqrt{L}$

In this subsection, we denote the iterates of GD on $\sqrt{L}$ with LR η by $x_\eta(t)$, with $x_\eta(0) \equiv x_{\text{init}}$ for all η:

$$\text{GD on } \sqrt{L}: \quad x_\eta(t+1) = x_\eta(t) - \eta \nabla \sqrt{L}(x_\eta(t)) \quad (5)$$

Similar to Normalized GD, we will have two phases. The first theorem demonstrates the movement in the manifold, when the iterate travels from x_{init} to a position that is $\mathcal{O}(\eta)$ distance closer to the manifold. For convenience, we will denote the quantity $\sqrt{\sum_{i=j}^{M} \lambda_i(x)\langle v_i(x), x - \Phi(x)\rangle^2} - \eta\sqrt{1/2}\lambda_j(x)$ by $\overline{R}_j(x)$ for all $j \in [d]$ and $x \in U$.

Theorem 4.5 (Phase I). *Let $\{x_\eta(t)\}_{t\in\mathbb{N}}$ be the iterates of Normalized GD (5) with LR η and $x_\eta(0) = x_{init} \in U$. There is $T_1 \in \mathbb{R}^+$ such that for any $T_1' \in \mathbb{R}^+$, it holds for sufficiently small η that (1) $\max_{T_1 \leq \eta t \leq T_1'} \|x_\eta(t) - \Phi(x_{init})\| \leq O(\eta)$ and (2) $\max_{T_1 \leq \eta t \leq T_1', j\in[D]} \overline{R}_j(x_\eta(t)) \leq O(\eta^2)$.*

The next result demonstrates that close to the manifold, the trajectory implicitly minimizes sharpness. We have an equivalent definition of T_2^{fav} for (6).

Theorem 4.6 (Phase II). *Let $\{x_\eta(t)\}_{t\in\mathbb{N}}$ be the iterates of perturbed GD on $\sqrt{L}$ (Algorithm 3) . If the initialization $x_\eta(0)$ satisfy that (1) $\|x_\eta(0) - \Phi(x_{init})\| \leq O(\eta)$, (2) $\max_{j\in[D]} \overline{R}_j(x_\eta(t)) \leq O(\eta^2)$, and additionally (3) $\min\{|\langle v_1(x_\eta(0)), x_\eta(0) - \Phi(x_\eta(0))\rangle|, -R_1(x_\eta(t))\} \geq \Omega(\eta)$, then for any time $T_2 \in T_2^{\text{fav}}$, it holds for sufficiently small η, with probability at least $1 - O(\eta^{10})$, that $\left\|\Phi(x_\eta(\lfloor T_2/\eta^2 \rfloor)) - X(T_2)\right\| = O(\eta^{1/2})$ and $\frac{1}{\lfloor T_2/\eta^2 \rfloor}\sum_{t=0}^{\lfloor T_2/\eta^2 \rfloor} \theta_t \leq \mathcal{O}(\eta^{1/2})$.*

$$X(\tau) = \Phi(x_{init}) - \frac{1}{8}\int_{s=0}^{\tau} P^{\perp}_{X(s),\Gamma} \nabla \lambda_1(X(s)) \mathrm{d}s. \quad (6)$$

4.5. Operating on the Edge of Stability

We can show that both Normalized GD on L and GD on $\sqrt{L}$ is on Edge of Stability in their phase II, that is, at least in one of every two consecutive steps, the stableness is at least 2 and the loss oscillates in every two consecutive steps. Interestingly, the average loss over two steps still monotonically decreases, even when operating on the edge of Stability (see Figure 1 for illustration), as indicated by the following theorems. Note that Theorems 4.4 and 4.6 ensures that the average of θ_t are $\mathcal{O}(\eta)$ and $\mathcal{O}(\sqrt{\eta})$.

Theorem 4.7 (Stableness, Normalized GD). *Under the setting of Theorem 4.4, by viewing Normalized GD as GD with time-varying LR $\eta_t := \frac{\eta}{\|\nabla L(x_\eta(t))\|}$, we have $[S_L(x_\eta(t), \eta_t)]^{-1} + [S_L(x_\eta(t+1), \eta_{t+1})]^{-1} = 1 + \mathcal{O}(\theta_t + \eta)$. Moreover, we have $\sqrt{L(x_\eta(t))} + \sqrt{L(x_\eta(t+1))} = \eta\sqrt{\frac{\lambda_1(\nabla^2 L(x_\eta(t)))}{2}} + \mathcal{O}(\eta\theta_t)$.*

Theorem 4.8 (Stableness, GD on $\sqrt{L}$). *Under the setting of Theorem 4.6, we have $[S_{\sqrt{L}}(x_\eta(t), \eta_t)] \geq \Omega(\frac{1}{\eta\theta_t})$. Moreover, we have $\sqrt{L(x_\eta(t))} + \sqrt{L(x_\eta(t+1))} = \eta\lambda_1(\nabla^2 L(x_\eta(t))) + \mathcal{O}(\eta\theta_t)$.*

5. Proof Overview

We sketch the proof of the Normalized GD in phase I and II respectively in Section 5.2. Then we briefly discuss how to prove the results for GD with $\sqrt{L}$ with same analysis in Section 5.3. We start by introducing the properties of limit map of gradient flow Φ in Section 5.1, which plays a very important role in the analysis.

5.1. Properties of Φ

The limit map of gradient flow Φ lies at the core of our analysis. When LR η is small, one can show $x_\eta(t)$ will be $\mathcal{O}(\eta)$ close to manifold and $\Phi(x_\eta(t))$. Therefore, $\Phi(x_\eta(t))$ captures the essential part of the implicit regularization of Normalized GD and characterization of the trajectory of $\Phi(x_\eta(t))$ immediately gives us that of $\Phi(x_\eta(t))$ up to $\mathcal{O}(\eta)$. Below we first recap a few important properties of Φ that will be used later this section, which makes the analysis of $\Phi(x_\eta(t))$ convenient.

Lemma 5.1. *[Lemmas B.15, B.17 and B.19] Under Assumptions 4.1 and 4.2, Φ satisfies the following two properties: (1) $\partial\Phi(x)\nabla L(x) = 0$ for any $x \in U$, and (2) for any $x \in \Gamma$, if $\lambda_1(x) > \lambda_2(x)$, $\partial^2\Phi(x)[v_1(x), v_1(x)] = -\frac{1}{2}P^{\perp}_{x,\Gamma}\nabla \log \lambda_1(x)$.*

With the Normalized GD update (3) for $x_\eta(t+1) - x_\eta(t)$, using a second order taylor expansion of Φ, we have

$$\Phi(x_\eta(t+1)) - \Phi(x_\eta(t)) \tag{7}$$
$$= \frac{\eta^2}{2}\partial^2\Phi(x_\eta(t))\left[\frac{\nabla L(x_\eta(t))}{\|\nabla L(x_\eta(t))\|}, \frac{\nabla L(x_\eta(t))}{\|\nabla L(x_\eta(t))\|}\right] + \mathcal{O}(\eta^3),$$

where we use the first claim of Lemma 5.1 in the final step. Therefore, we have $\Phi(x_\eta(t+1)) - \Phi(x_\eta(t)) = \mathcal{O}(\eta^2)$, which means $\Phi(x_\eta(t))$ moves slowly along the manifold, at a rate of at most $\mathcal{O}(\eta^2)$ step. The Taylor expansion of Φ, (7), plays a crucial role in our analysis for both Phase I and II and will be used repeatedly.

5.2. Analysis for Normalized GD

Analysis for Phase I, Theorem 4.3: The Phase I itself can be divided into two subphases: (A). Normalized GD iterate $x_\eta(t)$ gets $O(\eta)$ close to manifold; (B). counterpart of preparation phase in the quadratic case: local movement in the $\mathcal{O}(\eta)$-neighborhood of the manifold which decreases $R_j(x_\eta(t))$ to $\mathcal{O}(\eta^2)$. Below we sketch their proofs:

Subphase (A): First, with a very classical result in ODE approximation theory, normalized GD with small LR will track the normalized gradient flow, which is a time-rescaled version of standard gradient flow, with $\mathcal{O}(\eta)$ error, and enter a small neighborhoods of the manifold where Polyak-Łojasiewicz (PL) condition holds. Since then, Normalized GD decreases the fast loss with PL condition and the gradient has to be $O(\eta)$ small in $O(\frac{1}{\eta})$ steps. (See details in Appendix C.1).

Subphase (B): The result in subphase (B) can be viewed as a generalization of Lemma 3.3 when the loss function is $\mathcal{O}(\eta)$-approximately quadratic, in both space and time. More specifically, it means $\left\|\nabla^2 L(\Phi(x_\eta(t))) - \nabla^2 L(x)\right\| \le \mathcal{O}(\eta)$ for all x which is $\mathcal{O}(\eta)$-close to some $\Phi(x_\eta(t'))$ with $t' - t \le \mathcal{O}(1/\eta)$. This is because by Taylor expansion (7), $\|\Phi(x_\eta(t)) - \Phi(x_\eta(t'))\| = \mathcal{O}(\eta^2(t'-t)) = \mathcal{O}(\eta)$, and again by Taylor expansion of $\nabla^2 L$, we know $\left\|\nabla^2 L(x) - \nabla^2 L(\Phi(x_\eta(t)))\right\| = \mathcal{O}(\|x - \Phi(x_\eta(t))\|) = \mathcal{O}(\eta)$.

With a similar proof technique, we show $x_\eta(t)$ enters an invariant set around the manifold Γ, that is, $\{x \in U \mid R_j(x) \le O(\eta^2), \forall j \in [D]\}$. Formally, we show the following analog of Lemma 3.3:

Lemma 5.2 (Preparation Phase, Informal version of Lemma C.1)**.** *Let $\{x_\eta(t)\}_{t\ge 0}$ be the iterates of Normalized GD (3) with LR η. If for some step t_0, $\|x_\eta(t_0) - \Phi(x_\eta(t_0))\| = \mathcal{O}(\eta)$, then for sufficiently small LR η and all steps $t \in [t_0 + \Theta(1), \Theta(\eta^{-2})]$ steps, the iterate $x_\eta(t)$ satisfy $\max_{j\in[M]} R_j(x_\eta(t)) \le O(\eta^2)$.*

Analysis for Phase II, Theorem 4.4: Similar to the subphase (B) in the Phase I, the high-level idea here is again that $x_\eta(t)$ locally evolves like normalized GD with quadratic loss around $\Phi(x_\eta(t))$ and with an argument similar to the alignment phase of quadratic case (though technically more complicated), we show $x_\eta(t) - \Phi(x_\eta(t))$ approximately aligns to the top eigenvector of $\nabla^2 L(\Phi(x_\eta(t)))$, denoted by $v_1(t)$ and so does $\nabla L(x_\eta(t))$. Plugging the second claim of Lemma 5.1 into the Taylor expansion of Φ (7), we immediately get that $\Phi(x_\eta(t+1)) - \Phi(x_\eta(t)) \approx -\frac{\eta^2}{4}P^{\perp}_{\Phi(x_\eta(t)),\Gamma}\nabla \log \lambda_1(t)$.

We now have a more detailed look at the movement in Φ. Since $\Phi(x_\eta(t))$ belongs to the manifold, we have $\nabla L(\Phi(x_\eta(t))) = 0$ and so $\nabla L(x_\eta(t)) = \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t))) + \mathcal{O}(\eta^2)$ using a Taylor expansion. This helps us derive a relation between the Normalized GD update and the top eigenvector of the hessian (simplified version of Lemma B.10):

$$\exists s \in \{\pm 1\}, \quad \frac{\nabla L(x_\eta(t))}{\|\nabla L(x_\eta(t))\|} = s v_1(t) + \mathcal{O}(\theta_t + \eta). \tag{8}$$

Incorporating the above into the movement in $\Phi(x_\eta(t))$ from (7) gives: $\Phi(x_\eta(t+1)) - \Phi(x_\eta(t)) = \frac{\eta^2}{2}\partial^2\Phi(x_\eta(t))[v_1(t), v_1(t)] + \mathcal{O}(\eta^2\theta_t + \eta^3)$. Using the second property of Lemma 5.1 yields Lemma 5.3.

Lemma 5.3 (Movement in the manifold, Informal version of Lemma B.13)**.** *Under the setting in Theorem 4.4, for sufficiently small η, we have at any step $t \le \lfloor T_2/\eta^2 \rfloor$, $\Phi(x_\eta(t+1)) - \Phi(x_\eta(t)) = -\frac{\eta^2}{4}P^{\perp}_{t,\Gamma}\nabla \log \lambda_1(t) + \mathcal{O}(\eta^3 + \eta^2\theta_t)$.*

To complete the proof of Theorem 4.4, we show that for small enough η, the trajectory of $\Phi(x_\eta(\tau/\eta^2))$ is $O(\eta^3\lfloor T_2/\eta^2\rfloor + \eta^2\sum_{t=0}^{\lfloor T_2/\eta^2\rfloor}\theta_t)$-close to $X(\tau)$ for any $\tau \le T_2$, where $X(\cdot)$ is the flow given by (4). This error is $\mathcal{O}(\eta)$, since $\sum_{t=0}^{\lfloor T_2/\eta^2\rfloor}\theta_t = \mathcal{O}(\lfloor T_2/\eta^2\rfloor\eta)$. One technical difficulty towards showing the average of θ_t is only $\mathcal{O}(\eta)$ is that our current analysis requires $|\langle v_1(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))\rangle|$ doesn't vanish, that is, remains $\Omega(\eta)$ large throughout the entire training process. This is guaranteed by Lemma 3.4 in quadratic case, but the analysis breaks when the loss is only approximately quadratic and the alignment $|\langle v_1(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))\rangle|$could decrease by $\mathcal{O}(\theta_t\eta^2)$ per step. Once the alignment becomes too small, even if the angle θ_t is small, the normalized GD dynamics become chaotic and super sensitive to any perturbation. Our current proof technique cannot deal with this case, which is why we have to make the additional assumptions in Theorem 4.4.

Role of η^{100} noise. With the additional assumption that the initial alignment is $\Omega(\eta)$, we can show adding any poly(η) perturbation (even as small as $\Omega(\eta^{100})$) suffices to prevent the aforementioned bad case, that is, $|\langle v_1(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))\rangle|$ stays $\Omega(\eta)$ large. The intuition why $\Omega(\eta^{100})$ perturbation works again comes from quadratic case – it's clear that $\widetilde{x} = cv_1$ for any $|c| \le 1$ is a stationary point for two-step normalized GD updates for quadratic loss under the setting of Section 3. But if c is smaller than critical value determined by the eigenvalues of the hessian, the stationary point is unstable, meaning any deviation away from the top eigenspace will be amplified until the alignment increases above the critical threshold. Based on this intuition, the formal argument, Lemma E.11 uses the techniques from the 'escaping saddle point' analysis (Jin et al., 2017). Adding noise is not necessary in experiments to observe the predicted behavior (see 'Alignment' in Figure 5 where no noise is added). On one hand, it might be because the floating point errors served the role of noise. On the other hand, we suspect it's not necessary even for theory, just like GD gets stuck at saddle point only when initialized from a zero measure set even without noise (Lee et al., 2016; 2017).

5.3. Analysis for GD on $\sqrt{L}$

In this subsection we will make an additional assumption that $L(x) = 0$ for all $x \in \Gamma$. The analysis then will follow a very similar strategy as the analysis for Normalized GD. However, the major difference from the analysis for Normalized GD comes from the update rule for $x_\eta(t)$ when it is $\mathcal{O}(\eta)$-close to the manifold:

$$\exists s \in \{\pm 1\}, \quad \nabla\sqrt{L}(x_\eta(t)) = s\sqrt{\lambda_1(t)}v_1(t) + \mathcal{O}(\eta + \theta_t).$$

Thus, the effective learning rate is $\sqrt{\lambda_1(t)}\eta$ at any step t. This shows up, when we compute the movement in Φ.

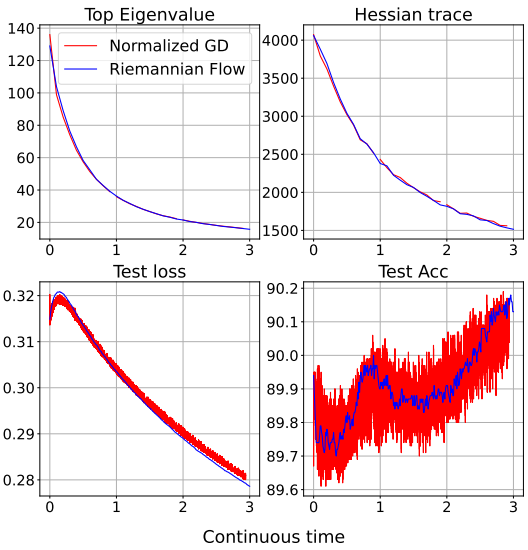

Figure 3: Normalized GD and Riemannian flow have almost the same behavior under proper time scalings, for a 2-layer network on MNIST initialized with tiny loss.

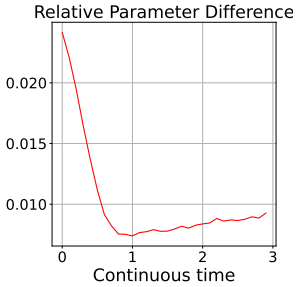

Figure 4: The trajectory of Normalized GD is close to that of the limiting flow minimizing the sharpness on manifold, as predicted by our theory. Relative parameter difference is the ratio of the norm of the difference between the parameters of the two trajectories to the norm of parameters of Normalized GD trajectory at the same continuous time.

Lemma 5.4 (Movement in the manifold, Informal version of Lemma G.1)**.** *Under the setting in Theorem 4.6, for sufficiently small η, we have at any step $t \le \lfloor T_2/\eta^2\rfloor$, $\Phi(x_\eta(t+1)) - \Phi(x_\eta(t)) = -\frac{\eta^2}{8}P^\perp_{t,\Gamma}\nabla\lambda_1(t) + \mathcal{O}(\eta^3 + \eta^2\theta_t)$.*

6. Experiments

Verifying convergence to limiting flow on MNIST: We first verify the closeness between the Riemannian gradient flow w.r.t. the top eigenvalue and Normalized GD, as predicted by Theorem 4.4, on a 1 hidden-layer fully connected network on MNIST (LeCun & Cortes, 2010). The network had 784 hidden units, with GeLU activation function (Hendrycks & Gimpel, 2016). We used the loss function L as the mean squared loss to ensure the existence of

minimizers and thus the manifold. For efficient training on a single GPU, we consider a sample of 1000 randomly selected points from the training data.

We first trained the model with Gradient Descent to reach loss of order 10^{-3}. Starting from this checkpoint, we make two different runs, one for Normalized GD and another for Riemannian gradient flow w.r.t. the top eigenvalue (see Appendix H for details). We plot the behavior of the network w.r.t. continuous time defined for Normalized GD as $\#\text{GradientSteps} \times \eta^2/4$, and for Riemannian flow as $\#\text{GradientSteps} \times \eta$, where η is the learning rate. We track the behavior of Test Loss, Test accuracy, the top eigenvalue of the Hessian and also the trace of the Hessian in Figure 3. We see that there is an exact match between the behavior of the four functions, which supports our theory. Moreover, Figure 4 computes the norm of the difference in the parameters between the two runs, and shows that the runs stay close to each other in the parameter space throughout training.

Verification for Predicted Phenomena on Real-life Models: Details in Appendix H show that it is very inefficient to simulate the Riemannian gradient flows for Real-life Models. Hence, we observe the behavior of different test functions throughout the training to verify our theoretical findings. We perform our experiments on a VGG-16 model (Simonyan & Zisserman, 2014) trained on CIFAR-10 dataset (Krizhevsky et al.) with Normalized GD and GD with $\sqrt{L}$. For efficient full-batch training, we trained the model on a sample of randomly chosen 5000 examples from the training dataset. To meet the smoothness requirement by our theory, we modified our network in two ways, (a) we used GeLU activation in place of the non-smooth ReLU activation, and (b) we used average pooling in place of the non-smooth max-pooling (Boureau et al., 2010). We used ℓ_2 loss instead of softmax loss to ensure the existence of minimizers and thus the manifold. We plot the behavior of the following four functions in Figure 5: Top eigenvalue of the Hessian, Alignment, Stableness, and Test accuracy. Alignment is defined as $\frac{1}{\lambda_1 \|g\|^2} g^\top (\nabla^2 L) g$, where $\nabla^2 L$ is the Hessian, g is the gradient and λ_1 is the top eigenvalue of the Hessian. To check the behavior for Stableness, we plot $\frac{\eta}{\|g\|} \times \lambda_1$ for Normalized GD and $\frac{\eta}{2\sqrt{L}} \times \lambda_1$ for GD with $\sqrt{L}$, which are lower bounds on the Stableness of the Hessian (1.1). We observe that the alignment function reaches close to 1, towards the end of training. The top eigenvalue decreases over time (as predicted by Theorems 4.4 and 4.6), and the stableness hovers around 2 at the end of training.

7. Conclusion

The recent discovery of Edge of Stability phenomenon in Cohen et al. (2021) calls for a reexamination of how we understand optimization in deep learning. The current paper gives two concrete settings with fairly general loss functions,

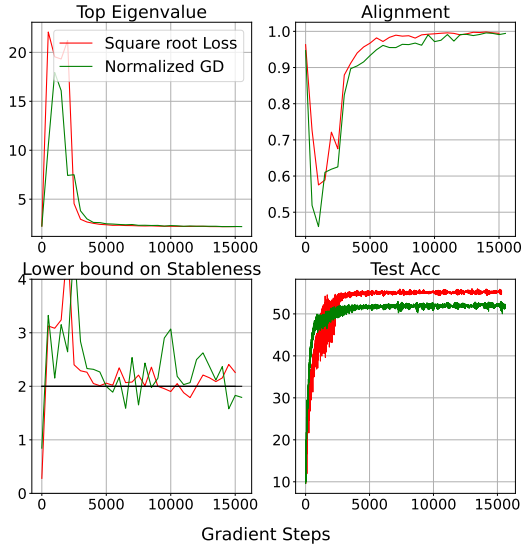

Figure 5: We verify our theoretical claims in the second phase —(a) the sharpness decreases; (b) gradient aligns with the top eigenvector of Hessian; (c) stableness will be higher than 2 — under the setting of training VGG-16 on CIFAR-10 dataset with Normalized GD on L and GD with $\sqrt{L}$ loss respectively.

where gradient updates can be shown to decrease loss over many iterations even after stableness is lost. Furthermore, in one setting the trajectory is shown to amount to reduce the sharpness (i.e., the maximum eigenvalue of the Hessian of the loss), thus rigorously establishing an effect that has been conjectured for decades in deep learning literature and was definitively documented for GD in Cohen et al. (2021). Our analysis crucially relies upon learning rate η being finite, in contrast to many recent results on implicit bias that required an infinitesimal LR. Even the alignment analysis of Normalized GD to the top eigenvector for quadratic loss in Section 3 appears to be new.

One limitation of our analysis is that it only applies close to the manifold of local minimizers. By contrast, in experiments, the EoS phenomenon, including the control of sharpness, begins much sooner. Addressing this gap, as well as analysing the EoS for the loss L itself (as opposed to $\sqrt{L}$ as done here) is left for future work. Very likely this will require novel understanding of properties of deep learning losses, which we were able to circumvent by looking at $\sqrt{L}$ instead. Exploration of EoS-like effects in SGD setting would also be interesting, although we first need definitive experiments analogous to Cohen et al. (2021).

Acknowledgements

The authors are supported by NSF, ONR, Simons Foundation, DARPA, and SRC. ZL is also supported by Microsoft Research Ph.D. Fellowship.

References

Ahn, K., Zhang, J., and Sra, S. Understanding the unstable convergence of gradient descent. *arXiv preprint arXiv:2204.01050*, 2022.

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 2019a.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019b.

Arora, S., Cohen, N., and Hazan, E. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pp. 244–253. PMLR, 2018.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019a.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019b.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. On exact computation with an infinitely wide neural net. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 8141–8150, 2019c.

Azulay, S., Moroshko, E., Nacson, M. S., Woodworth, B. E., Srebro, N., Globerson, A., and Soudry, D. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pp. 468–477. PMLR, 2021.

Barrett, D. and Dherin, B. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.

Bihari, I. A generalization of a lemma of bellman and its application to uniqueness problems of differential equations. *Acta Mathematica Hungarica*, 7(1):81–94, 1956.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pp. 483–513. PMLR, 2020.

Boureau, Y.-L., Ponce, J., and LeCun, Y. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 111–118, 2010.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Damian, A., Ma, T., and Lee, J. D. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34, 2021.

Davis, C. and Kahan, W. M. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Falconer, K. J. Differentiation of the limit mapping in a dynamical system. *Journal of the London Mathematical Society*, s2-27(2):356–372, 1983. ISSN 0024-6107. doi: 10.1112/jlms/s2-27.2.356.

Fehrman, B., Gess, B., and Jentzen, A. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21, 2020.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018a.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. *In Advances in Neural Information Processing Systems*, 2018b.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018c.

Gunasekar, S., Woodworth, B., and Srebro, N. Mirrorless mirror descent: A natural derivation of mirror descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 2305–2313. PMLR, 2021.

He, H., Huang, G., and Yuan, Y. Asymmetric valleys: Beyond sharp and flat local minima. *Advances in neural information processing systems*, 32, 2019.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Computation*, 1997.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jiang*, Y., Neyshabur*, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pp. 1724–1732. PMLR, 2017.

Katzenberger, G. S. Solutions of a stochastic differential equation forced onto a manifold by a large drift. *The Annals of Probability*, pp. 1587–1628, 1991.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Kwon, J., Kim, J., Park, H., and Choi, I. K. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5905–5914. PMLR, 18–24 Jul 2021.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257. PMLR, 2016.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

Li, Q., Tai, C., and Weinan, E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pp. 2101–2110. PMLR, 2017.

Li, Q., Tai, C., and Weinan, E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520, 2019.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in Neural Information Processing Systems*, 31, 2018.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pp. 2–47. PMLR, 2018.

Li, Z., Luo, Y., and Lyu, K. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2020a.

Li, Z., Lyu, K., and Arora, S. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33, 2020b.

Li, Z., Wang, T., and Arora, S. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022.

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeLIgBKPS.

Lyu, K., Li, Z., Wang, R., and Arora, S. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34, 2021.

Ma, C. and Ying, L. On linear stability of SGD and input-smoothness of neural networks. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Magnus, J. R. On differentiating eigenvalues and eigenvectors. *Econometric theory*, 1(2):179–191, 1985.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Rangamani, A., Nguyen, N. H., Kumar, A., Phan, D., Chin, S. P., and Tran, T. D. A scale invariant measure of flatness for deep network minima. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1680–1684, 2021.

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

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Tsuzuku, Y., Sato, I., and Sugiyama, M. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9636–9647. PMLR, 13–18 Jul 2020.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Wu, L., Zhu, Z., et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

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

Yi, M., Meng, Q., Chen, W., Ma, Z.-m., and Liu, T.-Y. Positively scale-invariant flatness of relu neural networks. *arXiv preprint arXiv:1903.02237*, 2019a.

Yi, M., Zhang, H., Chen, W., Ma, Z.-M., and Liu, T.-Y. Bn-invariant sharpness regularizes the training model to better generalization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 4164–4170. International Joint Conferences on Artificial Intelligence Organization, 7 2019b.

Yi, M., Meng, Q., Chen, W., and Ma, Z.-M. Towards accelerating training of batch normalization: A manifold perspective. *arXiv preprint arXiv:2101.02916*, 2021.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

A. Omitted Proofs for Results for Quadratic Loss Functions

We first recall the settings and notations. Let A be a positive definite matrix. Without loss of generality, we can assume A is diagonal, *i.e.*, $A = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_D) \in \mathbb{R}^{D\times D}$, where $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_D > 0$ and the eigenvectors are the standard basis vectors $e_1, \cdots, e_D$ of the D-dimensional space. We will denote $P^{(j:D)} = \sum_{i=j}^{D} e_i e_i^\top$ as the projection matrix onto the subspace spanned by $e_j, \ldots, e_D$.

Recall the loss function L is defined as $L(x) = \frac{1}{2}x^\top Ax$. The Normalized GD update (LR= η)is given by $x(t+1) = x(t) - \eta\frac{Ax(t)}{\|Ax(t)\|}$. A substitution $\widetilde{x}(t) := \frac{Ax(t)}{\eta}$ gives the following update rule:

$$\widetilde{x}(t+1) = \widetilde{x}(t) - A\frac{\widetilde{x}(t)}{\|\widetilde{x}(t)\|}. \tag{2}$$

Note Normalized GD (2) is not defined at $\|\widetilde{x}(t)\| = 0$. Moreover, it's easy to check that if at some time step t $|\langle v_1, \widetilde{x}(t)\rangle| = 0$, $|\langle v_1, \widetilde{x}(t')\rangle| = 0$ holds for any $t' \geq t$. Thus it's necessary to assume $|\langle v_1, \widetilde{x}(t)\rangle| \neq 0$ for all $t \in \mathbb{N}$ in order to prove alignment to the top eigenvector of A for Normalized GD (2).

Now we recall the main theorem for Normalized GD on quadratic loss functions:

Theorem 3.1. *If $|\langle v_1, \widetilde{x}(t)\rangle| \neq 0$, $\forall t \geq 0$, then there exists $0 < C < 1$ and $s \in \{\pm 1\}$ such that $\lim_{t\to\infty} \widetilde{x}(2t) = Cs\lambda_1 v_1$ and $\lim_{t\to\infty} \widetilde{x}(2t+1) = (C-1)s\lambda_1 v_1$.*

We also note that GD on $\sqrt{L}$ with any LR η can also be reduced to update rule (2), as shown in the discussion at the end of Section 3.

A.1. Proofs for Preparation Phase

In this subsection, we show (1). $\mathcal{I}_j$ is indeed an invariant set for normalized GD $\forall j \in [D]$ and (2). from any initialization, normalized GD will eventually go into their intersection $\cap_{j=1}^{D}\mathcal{I}_j$.

Lemma A.1. *For any $t \in \mathbb{N}$ and $j \in [D]$, $\left\|P^{(j:D)}\widetilde{x}(t)\right\| \leq \lambda_j \implies \left\|P^{(j:D)}\widetilde{x}(t+1)\right\| \leq \lambda_j$. In other words, $\{\mathcal{I}_j\}_{j=1}^{D}$ are invariant sets of update rule Equation (2).*

Proof of Lemma A.1. Note $P^{(j:D)}A = P^{(j:D)}AP^{(j:D)}$, by definition of Normalized GD (2), we have

$$P^{(j:D)}\widetilde{x}(t+1) = P^{(j:D)}\widetilde{x}(t) - P^{(j:D)}A\frac{\widetilde{x}(t)}{\|\widetilde{x}(t)\|} = \left(I - \frac{P^{(j:D)}A}{\|\widetilde{x}(t)\|}\right)P^{(j:D)}\widetilde{x}(t),$$

which implies

$$\left\|P^{(j:D)}\widetilde{x}(t+1)\right\| \leq \left\|I - \frac{P^{(j:D)}A}{\|\widetilde{x}(t)\|}\right\|\left\|P^{(j:D)}\widetilde{x}(t)\right\|. \tag{9}$$

Note that $P^{(j:D)}A \preccurlyeq \lambda_j I$, $\left\|P^{(j:D)}\widetilde{x}(t)\right\| \leq \|\widetilde{x}(t)\|$ and $\left\|P^{(j:D)}\widetilde{x}(t)\right\| \leq \lambda_j$ by assumption, we have

$$-\frac{\lambda_j}{\left\|P^{(j:D)}\widetilde{x}(t)\right\|}I \preccurlyeq -\frac{P^{(j:D)}A}{\|\widetilde{x}(t)\|} \preccurlyeq I - \frac{P^{(j:D)}A}{\|\widetilde{x}(t)\|} \preccurlyeq I \preccurlyeq \frac{\lambda_j}{\left\|P^{(j:D)}\widetilde{x}(t)\right\|}I.$$

Therefore $\left\|I - \frac{P^{(j:D)}A}{\|\widetilde{x}(t)\|}\right\| \leq \frac{\lambda_j}{\left\|P^{(j:D)}\widetilde{x}(t)\right\|}$ and thus we conclude $\left\|P^{(j:D)}\widetilde{x}(t+1)\right\| \leq \lambda_j$. □

Lemma A.2. *For any $t \in \mathbb{N}$ and $j \in [D]$, if $\left\|P^{(j:D)}\widetilde{x}(t)\right\| \geq \lambda_j$, then $\left\|P^{(j:D)}\widetilde{x}(t+1)\right\| \leq (1 - \frac{\lambda_D}{\|\widetilde{x}(t)\|})\left\|P^{(j:D)}\widetilde{x}(t)\right\|$.*

Proof of Lemma A.2. Since $\lambda_j \leq \left\|P^{(j:D)}\widetilde{x}(t)\right\| \leq \|\widetilde{x}(t)\|$, we have $0 \preccurlyeq I - \frac{P^{(j:D)}A}{\|\widetilde{x}(t)\|} \preccurlyeq 1 - \frac{\lambda_D}{\|\widetilde{x}(t)\|}$. Therefore $\left\|I - \frac{P^{(j:D)}A}{\|\widetilde{x}(t)\|}\right\| \leq 1 - \frac{\lambda_D}{\|\widetilde{x}(t)\|}$. The proof is completed by plugging this into Equation (9). □

Lemma A.2 has the following two direct corollaries.

Corollary A.3. *For any initialization $\widetilde{x}(0)$ and $t \geq \frac{\|\widetilde{x}(0)\|-\lambda_1}{\lambda_D}$, $\|\widetilde{x}(t)\| \leq \lambda_1$, that is, $\widetilde{x}(t) \in \mathcal{I}_1$.*

Proof of Corollary A.3. Set $j = 1$ in Lemma A.2, it holds that $\|\widetilde{x}(t+1)\| \leq \|\widetilde{x}(t)\| - \lambda_D$ whenever $\|\widetilde{x}(t)\| \geq \lambda_1$. Thus $\left\|\widetilde{x}(\lceil \frac{\|\widetilde{x}(0)\|-\lambda_1}{\lambda_D} \rceil)\right\| \leq \lambda_1$. The proof is completed as $\mathcal{I}_1$ is an invariant set by Lemma A.1. □

Corollary A.4. *For any coordinate $j \in [D]$ and initial point $\widetilde{x}(0) \in \mathcal{I}_1$, if $t \geq \frac{\lambda_1}{\lambda_D} \ln \frac{\lambda_1}{\lambda_j}$ then $\left\|P^{(j:D)}\widetilde{x}(t)\right\| \leq \lambda_j$.*

Proof of Corollary A.4. Since $\mathcal{I}_1$ is an invariant set, we have $\|\widetilde{x}(t)\| \leq \lambda_1$ for all $t \geq 0$. Thus let $T = \lfloor \frac{\lambda_1}{\lambda_D} \ln \frac{\lambda_1}{\lambda_j} \rfloor$, we have

$$\left\|P^{(j:D)}\widetilde{x}(T)\right\| \leq e^{-T\frac{\lambda_D}{\lambda_1}} \left\|P^{(j:D)}\widetilde{x}(0)\right\| \leq \frac{\lambda_j}{\lambda_1} \|\widetilde{x}(0)\| \leq \lambda_j.$$

The proof is completed since I_j is a invariant set for any $j \in [D]$ by Lemma A.1. □

A.2. Proofs for Alignment Phase

In this subsection, we analyze how normalized GD align to the top eigenvector once it goes through the preparation phase, meaning $\widetilde{x}(t) \in \cap_{j=1}^{D} \mathcal{I}_j$ for all t in alignment phase.

Lemma 3.5. *For any t with $\widetilde{x}(t) \in \cap_{j=1}^{D} \mathcal{I}_j$, if $\|\widetilde{x}(t)\| > \lambda_1/2$, then $\|\widetilde{x}(t+1)\| \leq \max\left(\frac{\lambda_1}{2} - \frac{\lambda_D^2}{2\lambda_1}, \lambda_1 - \|\widetilde{x}(t)\|\right)$.*

Proof. The update at step t as:

$$\widetilde{x}(t+1) = \frac{1}{\|\widetilde{x}(t)\|}\left(\|\widetilde{x}(t)\| I - A\right)\widetilde{x}(t) = \frac{1}{\|\widetilde{x}(t)\|}\begin{bmatrix} (\|\widetilde{x}(t)\| - \lambda_1)\widetilde{x}_1(t) \\ (\|\widetilde{x}(t)\| - \lambda_2)\widetilde{x}_2(t) \\ \vdots \\ (\|\widetilde{x}(t)\| - \lambda_D)\widetilde{x}_D(t) \end{bmatrix}.$$

Let the index k be the smallest integer such that $\lambda_{k+1} < 2\|\widetilde{x}(t)\| - \lambda_1$. If no such index exists, then one can observe that $\|\widetilde{x}(t+1)\| \leq \lambda_1 - \|\widetilde{x}(t)\|$. Assuming that such an index exists in $[D]$, we have $\lambda_k \geq 2\|\widetilde{x}(t)\| - \lambda_1$ and $\|\widetilde{x}(t)\| - \lambda_j \leq \lambda_1 - \|\widetilde{x}(t)\|, \forall j \leq k$. Now consider the following vectors:

$$\begin{aligned}
v^{(1)}(t) &:= (\lambda_1 - \|\widetilde{x}(t)\|)\widetilde{x}(t), \\
v^{(2)}(t) &:= (2\|\widetilde{x}(t)\| - \lambda_1 - \lambda_k)P^{(k:D)}\widetilde{x}(t), \\
v^{(2+j)}(t) &:= (\lambda_{k+j-1} - \lambda_{k+j})P^{(k+j:D)}\widetilde{x}(t), \forall 1 \leq j \leq D-k.
\end{aligned}$$

By definition of k, $|\,\|\widetilde{x}(t)\| - \lambda_j| \leq |\,\|\widetilde{x}(t)\| - \lambda_1|$. Thus

$$\begin{aligned}
\|\widetilde{x}(t+1)\| &\leq \frac{1}{\|\widetilde{x}(t)\|}\left\|\begin{bmatrix} (\|\widetilde{x}(t)\| - \lambda_1)\widetilde{x}_1(t) \\ \vdots \\ (\|\widetilde{x}(t)\| - \lambda_1)\widetilde{x}_k(t) \\ (\|\widetilde{x}(t)\| - \lambda_{k+1})\widetilde{x}_{k+1}(t) \\ \vdots \\ (\|\widetilde{x}(t)\| - \lambda_D)\widetilde{x}_D(t) \end{bmatrix}\right\| \\
&= \frac{1}{\|\widetilde{x}(t)\|}\left\|v^{(1)}(t) + v^{(2)}(t) + \ldots + v^{(D-k+2)}(t)\right\| \\
&\leq \frac{1}{\|\widetilde{x}(t)\|}\left(\left\|v^{(1)}(t)\right\| + \left\|v^{(2)}(t)\right\| + \ldots + \left\|v^{(D-k+2)}(t)\right\|\right).
\end{aligned}$$

By assumption, we have $\widetilde{x}(t) \in \cap_{j=1}^{D} \mathcal{I}_j$. Thus

$$\begin{aligned}
\left\| v^{(1)}(t) \right\| &= (\lambda_1 - \|\widetilde{x}(t)\|) \|\widetilde{x}(t)\| \\
\left\| v^{(2)}(t) \right\| &\le (2\|\widetilde{x}(t)\| - \lambda_1 - \lambda_k)\lambda_k \\
\left\| v^{(2+j)}(t) \right\| &\le (\lambda_{k-1+j} - \lambda_{k+j})\lambda_{k+j}, \text{ for all } j \ge 1.
\end{aligned}$$

Hence,

$$\begin{aligned}
\sum_{j\ge 2} \left\| v^{(j)}(t) \right\| &= (2\|\widetilde{x}(t)\| - \lambda_1 - \lambda_k)\lambda_k + \sum_{j\ge k} (\lambda_j - \lambda_{j+1})\lambda_{j+1} \\
&= (2\|\widetilde{x}(t)\| - \lambda_1)\lambda_k + \sum_{j\ge k} \lambda_j\lambda_{j+1} - \sum_{j\ge k} \lambda_j^2 \\
&\le \frac{(2\|\widetilde{x}(t)\| - \lambda_1)^2 + \lambda_k^2}{2} + \sum_{j\ge k} \frac{\lambda_j^2 + \lambda_{j+1}^2}{2} - \sum_{j\ge k} \lambda_j^2 \\
&\le \frac{(2\|\widetilde{x}(t)\| - \lambda_1)^2}{2} - \frac{\lambda_D^2}{2},
\end{aligned}$$

where we applied AM-GM inequality multiple times in the pre-final step.

Thus,

$$\begin{aligned}
\|\widetilde{x}(t+1)\| &\le \frac{1}{\|\widetilde{x}(t)\|} \left(\left\| v^{(1)}(t) \right\| + \left\| v^{(2)}(t) \right\| + \ldots + \left\| v^{(D-k+1)}(t) \right\| \right) \\
&\le \frac{(2\|\widetilde{x}(t)\| - \lambda_1)^2}{2\|\widetilde{x}(t)\|} - \frac{\lambda_D^2}{2\|\widetilde{x}(t)\|} + \lambda_1 - \|\widetilde{x}(t)\| \\
&= \|\widetilde{x}(t)\| + \frac{\lambda_1^2 - \lambda_D^2}{2\|\widetilde{x}(t)\|} - \lambda_1 \\
&\le \frac{\lambda_1}{2} - \frac{\lambda_D^2}{2\lambda_1},
\end{aligned}$$

where the final step is because $\frac{\lambda_1}{2} \le \|\widetilde{x}(t)\| \le \lambda_1$ and that the maximal value of a convex function is attained at the boundary of an interval.

□

Lemma A.5. *At any step t and $i \in [D]$, if $\|\widetilde{x}(t)\| \gtreqless \frac{\lambda_i}{2}$, then $|\widetilde{x}_i(t+1)| \lesseqgtr |\widetilde{x}_i(t)|$, where $\gtreqless$ denotes larger than, equal to and smaller than respectively. (Same for $\lesseqgtr$, but in the reverse order)*

Proof. From the Normalized GD update rule, we have $\widetilde{x}_i(t+1) = \widetilde{x}_i(t)\left(1 - \frac{\lambda_i}{\|\widetilde{x}(t)\|}\right)$, for all $i \in [D]$. Thus

$$\frac{\lambda_1}{\|\widetilde{x}(t)\|} \lesseqgtr 2 \iff \left| 1 - \frac{\lambda_1}{\|\widetilde{x}(t)\|} \right| \lesseqgtr 1 \iff |\widetilde{x}_i(t+1)| \lesseqgtr |\widetilde{x}_i(t)|,$$

which completes the proof. □

Lemma A.6. *At any step t, if $\|\widetilde{x}(t)\| \le \frac{\lambda_1}{2}$, then*

$$(\lambda_1 - \|\widetilde{x}(t)\|)\cos\theta_t \le \|\widetilde{x}(t+1)\| \le \lambda_1 - \|\widetilde{x}(t)\| - \frac{\lambda}{2\lambda_1}\left(1 - \frac{\lambda}{\lambda_1}\right)\lambda_1 \sin^2\theta_t,$$

where $\theta_t = \arctan \frac{\|P^{(2:D)}\widetilde{x}(t)\|}{|e_1^\top \widetilde{x}(t)|}$ and $\lambda = \min(\lambda_1 - \lambda_2, \lambda_D)$.

Proof. We first show that the left side inequality holds by the following update rule for $\langle e_1, \widetilde{x}(t)\rangle$:

$$\langle e_1, \widetilde{x}(t+1)\rangle = (\|\widetilde{x}(t)\| - \lambda_1)\frac{\langle e_1, \widetilde{x}(t)\rangle}{\|\widetilde{x}(t)\|}.$$

Since $\|\widetilde{x}(t+1)\| \geq |\langle e_1, \widetilde{x}(t+1)\rangle|$ and θ_t denotes the angle between e_1 and $\widetilde{x}(t+1)$, we get the left side inequality.

Now, we focus on the right hand side inequality. First of all, the update in the coordinate $j \in [2, D]$ is given by

$$\langle e_j, \widetilde{x}(t+1)\rangle = (\|\widetilde{x}(t)\| - \lambda_j)\frac{\langle e_j, \widetilde{x}(t)\rangle}{\|\widetilde{x}(t)\|}.$$

Then, we have

$$\begin{aligned}
\|\widetilde{x}(t+1)\|^2 &= \sum_{j=1}^{D}\langle e_j, \widetilde{x}(t+1)\rangle^2 \\
&= \sum_{j=1}^{D}(\|\widetilde{x}(t)\| - \lambda_j)^2\left(\frac{\langle e_j, \widetilde{x}(t)\rangle}{\|\widetilde{x}(t)\|}\right)^2 \\
&= (\|\widetilde{x}(t)\| - \lambda_1)^2\cos^2\theta_t + \sum_{j=2}^{D}(\|\widetilde{x}(t)\| - \lambda_j)^2\left(\frac{\langle e_j, \widetilde{x}(t)\rangle}{\|\widetilde{x}(t)\|}\right)^2 \\
&\leq (\|\widetilde{x}(t)\| - \lambda_1)^2\cos^2\theta_t + (\|\widetilde{x}(t)\| - \overline{\lambda})^2\sum_{j=2}^{D}\left(\frac{\langle e_j, \widetilde{x}(t)\rangle}{\|\widetilde{x}(t)\|}\right)^2 \\
&= (\|\widetilde{x}(t)\| - \lambda_1)^2\cos^2\theta_t + (\|\widetilde{x}(t)\| - \overline{\lambda})^2\sin^2\theta_t \\
&= (\|\widetilde{x}(t)\| - \lambda_1)^2 + (\lambda_1 - \overline{\lambda})(2\|\widetilde{x}(t)\| - \overline{\lambda} - \lambda_1)\sin^2\theta_t \\
&\leq (\|\widetilde{x}(t)\| - \lambda_1)^2 - \overline{\lambda}(\lambda_1 - \overline{\lambda})\sin^2\theta_t,
\end{aligned}$$

where in the fourth step, we have used $\overline{\lambda} = \mathrm{argmax}_{\lambda_i | 2\leq i\leq D}\left|\|\widetilde{x}(t)\| - \lambda_i\right|$. The final step uses $\|\widetilde{x}(t)\| < \frac{\lambda_1}{2}$. Hence, using the fact that $\sqrt{1-y} \leq 1 - y/2$ for any $y \leq 1$, we have

$$\begin{aligned}
\|\widetilde{x}(t+1)\| &\leq \lambda_1 - \|\widetilde{x}(t)\| - \frac{1}{2(\lambda_1 - \|\widetilde{x}(t)\|)}\overline{\lambda}(\lambda_1 - \lambda)\sin^2\theta_t \\
&\leq \lambda_1 - \|\widetilde{x}(t)\| - \frac{\overline{\lambda}}{2\lambda_1}\left(-\frac{\overline{\lambda}}{\lambda_1}\right)\lambda_1\sin^2\theta_t,
\end{aligned}$$

where again in the final step, we have used $\|\widetilde{x}(t)\| < \frac{\lambda_1}{2}$. The above bound can be further bounded by

$$\begin{aligned}
\|\widetilde{x}(t+1)\| &\leq \lambda_1 - \|\widetilde{x}(t)\| - \frac{\overline{\lambda}}{2\lambda_1}\left(1 - \frac{\overline{\lambda}}{\lambda_1}\right)\lambda_1\sin^2\theta_t \\
&\leq \lambda_1 - \|\widetilde{x}(t)\| - \frac{1}{2}\left(\min_{\lambda'\in\{\lambda_2,\lambda_D\}}\frac{\lambda'}{\lambda_1}\left(1 - \frac{\lambda'}{\lambda_1}\right)\right)\lambda_1\sin^2\theta_t \\
&= \lambda_1 - \|\widetilde{x}(t)\| - \frac{1}{2}\left(\frac{\lambda}{\lambda_1}\left(1 - \frac{\lambda}{\lambda_1}\right)\right)\lambda_1\sin^2\theta_t,
\end{aligned}$$

where we have used $\lambda = \min(\lambda_1 - \lambda_2, \lambda_D)$.

□

Lemma A.7. *If at some step t, $\|\widetilde{x}(t+1)\| + \|\widetilde{x}(t)\| \leq \lambda_1$, then $|\widetilde{x}_1(t+2)| \geq |\widetilde{x}_1(t)|$, where the equality holds only when $\|\widetilde{x}(t+1)\| + \|\widetilde{x}(t)\| = \lambda_1$. Therefore, by Lemma A.6, we have :*

$$\|\widetilde{x}(t)\| \leq \frac{\lambda_1}{2} \implies |\widetilde{x}_1(t+2)| \geq |\widetilde{x}_1(t)|\,(1 + 2\frac{\lambda}{\lambda_1}(1 - \frac{\lambda}{\lambda_1})\sin^2\theta_t),$$

where $\theta_t = \arctan\frac{\|P^{(2:D)}\widetilde{x}(t)\|}{|e_1^\top\widetilde{x}(t)|}$, and $\lambda = \min(\lambda_1 - \lambda_2, \lambda_D)$.

Proof of Lemma A.7. Using the Normalized GD update rule, we have

$$\widetilde{x}_1(t+1) = \left(1 - \frac{\lambda_1}{\|\widetilde{x}(t)\|}\right)\widetilde{x}_1(t), \quad \widetilde{x}_1(t+2) = \left(1 - \frac{\lambda_1}{\|\widetilde{x}(t+1)\|}\right)\widetilde{x}_1(t+1).$$

Combining the two updates, we have

$$\begin{aligned}
|\widetilde{x}_1(t+2)| &= \left|\left(1 - \frac{\lambda_1}{\|\widetilde{x}(t)\|}\right)\left(1 - \frac{\lambda_1}{\|\widetilde{x}(t+1)\|}\right)\right| |\widetilde{x}_1(t)| \\
&= \left|1 + \frac{\lambda_1^2 - \lambda_1(\|\widetilde{x}(t)\| - \|\widetilde{x}(t+1)\|)}{\|\widetilde{x}(t)\|\,\|\widetilde{x}(t+1)\|}\right| |\widetilde{x}_1(t)| \\
&\geq |\widetilde{x}_1(t)|\,,
\end{aligned}$$

where the equality holds only when $\|\widetilde{x}(t+1)\| + \|\widetilde{x}(t)\| = \lambda_1$.

Moreover, with the additional condition that $\|\widetilde{x}(t)\| < \frac{\lambda_1}{2}$, we have from Lemma A.6, $\|\widetilde{x}(t+1)\| \leq \lambda_1 - \|\widetilde{x}(t)\| - \lambda(\lambda_1 - \lambda)\sin^2\theta_t$, where $\lambda = \min(\lambda_1 - \lambda_2, \lambda_D)$.

Hence, retracing the steps we followed before, we have

$$\begin{aligned}
|\widetilde{x}_1(t+2)| &= \left|1 + \frac{\lambda_1^2\lambda_1(-\|\widetilde{x}(t)\| - \|\widetilde{x}(t+1)\|)}{\|\widetilde{x}(t)\|\,\|\widetilde{x}(t+1)\|}\right| |\widetilde{x}_1(t)| \\
&\geq \left|1 + \frac{\lambda(\lambda_1 - \lambda)\sin^2\theta_t}{\|\widetilde{x}(t)\|\,\|\widetilde{x}(t+1)\|}\right| |\widetilde{x}_1(t)| \\
&\geq \left|1 + 2\frac{\lambda}{\lambda_1}(1 - \frac{\lambda}{\lambda_1})\sin^2\theta_t\right| |\widetilde{x}_1(t)|\,,
\end{aligned}$$

where the final step follows from using $\|\widetilde{x}(t)\| \leq \frac{\lambda_1}{2}$ and $\|\widetilde{x}(t+1)\| \leq \lambda_1 - \|\widetilde{x}(t)\| \leq \lambda_1$.

□

A.3. Proof of Main theorems for Quadratic Loss

Proof of Theorem 3.1. The analysis will follow in two phases:

1. **Preparation phase:** $\widetilde{x}(t)$ enters and stays in an invariant set around the origin, that is, $\cap_{j=1}^{D}\mathcal{I}_j$, where $\mathcal{I}_j := \{\widetilde{x} \mid \sum_{i=j}^{D}\langle e_i, \widetilde{x}(t)\rangle^2 \leq \lambda_j^2\}$. (See Lemma 3.3, which is a direct consequence of Lemma A.1 and Corollary A.3.)
2. **Alignment phase:** The projection of $\widetilde{x}(t)$ on the top eigenvector, $|\langle \widetilde{x}(t), e_1\rangle|$, is shown to increase monotonically among the steps among the steps $\{t \mid \|\widetilde{x}(t)\| \leq 0.5\}$, up until convergence, since it's bounded. (Lemma 3.4)

   By Lemma A.7, the convergence of $|\langle \widetilde{x}(t), e_1\rangle|$ would imply the convergence of $\widetilde{x}(t)$ to e_1 in direction.

Below we elaborate the convergence argument in the alignment phase. For convenience, we will use θ_t to denote the angle between e_1 and $\widetilde{x}(t)$ and we assume $\widetilde{(0)} \in \cap_{j=1}^{D}\mathcal{I}_j$ without loss of generality. We first define $S := \{t \in \mathbb{N} \mid \|\widetilde{x}(t)\| \leq \frac{\lambda_1}{2}\}$ and $S' := \{t \in S \mid t+2 \in S\}$. The result in alignment phase says that $\frac{1}{\lambda_1}|\widetilde{x}_1(t)|$ monotone increases and converges to some constant $C \in (0, \frac{1}{2}]$ among all $t \in S$, thus $\lim_{t\to\infty, t\in S'} \frac{|\widetilde{x}_1(t+2)|}{|\widetilde{x}_1(t)|} = 1$. By Lemma A.7, we have $\lim_{t\to\infty, t\in S'} \theta_t = 0$. Since the one-step update function $F(\widetilde{x}) = \widetilde{x} - A\frac{\widetilde{x}}{\|\widetilde{x}\|}$ is uniformly lipschitz when $\|\widetilde{x}\|$ is bounded away from zero, we know $\lim_{t\to\infty, t\in S'} \theta_{t+k} = 0,\ \forall k \in \mathbb{N}$.

Now we claim $\forall t \geq 3$, there is some $k \in \{0, 1, 3\}$ such that $t - k \in S'$. This is because Lemma 3.5 says that if $t \notin S$, then both $t-1, t+1 \in S$. Thus for any $t \notin S$, $t-1 \in S'$. Therefore, for any $t \in S/S'$, if $t-2 \notin S$, then $t-3 \in S'$. Thus we conclude that $\forall t \geq 3$, there is some $k \in \{0, 1, 3\}$ such that $t - k \in S'$, which implies $\lim_{t\to\infty} \theta_t = 0$. Hence $\lim_{t\to\infty} \|\widetilde{x}(t+1) - \widetilde{x}(t)\| = \lambda_1$, meaning for sufficiently large t, $\widetilde{x}_1(t)$ flips its sign per step and thus $\lim_{t\to\infty} \widetilde{x}(t+2) - \widetilde{x}(t) = 0$, $\lim_{t\to\infty} \|\widetilde{x}(t+1)\| + \|\widetilde{x}(t)\| = \lambda_1$.

If $C = \frac{1}{2}$, then we must have $\lim_{t\to\infty} \|\widetilde{x}(t)\| = \frac{\lambda_1}{2}$ and we are done in this case. If $C < \frac{1}{2}$, note that $\lim_{t\to\infty, t\in S'} |\widetilde{x}_1(t)| = C\lambda_1$, it must hold that $\lim_{t\to\infty, t\in S'} \|\widetilde{x}(t+1)\| = (1-C)\lambda_1$, thus there is some large $T \in S$ such that for all $t \in S, t \geq T$, $t+1 \notin S$. By Lemma 3.5, $t+2 \in S$. Thus we conclude $\lim_{t\to\infty} \widetilde{x}(T+2t) = C\lambda s e_1$ for some $s \in \{-1, 1\}$ and thus $\lim_{t\to\infty} \widetilde{x}(T+2t+1) = (C-1)\lambda s e_1$. This completes the proof. □

A.4. Some Extra Lemmas (only used in the general loss case)

For a general loss function L satisfying Assumption 4.1, the loss landscape looks like a strongly convex quadratic function locally around its minimizer. When sufficient small learning rate, the dynamics will be sufficiently close to the manifold and behaves like that in quadratic case with small perturbations. Thus it will be very useful to have more refined analysis for the quadratic case, as they allow us to bound the error in the approximate quadratic case quantitatively. Lemmas A.8 to A.11 are such examples. Note that they are only used in the proof of the general loss case, but not in the quadratic loss case.

Lemma A.8 is a slightly generalized version of Lemma 3.5.

Lemma A.8. *Suppose at time t, $\left\|P^{(j:D)}\widetilde{x}(t)\right\| \leq \lambda_j(1+\frac{\lambda_D^2}{\lambda_1^2})$, for all $j \in [D]$, if $\|\widetilde{x}(t)\| > \frac{\lambda_1}{2}$, then $\|\widetilde{x}(t+1)\| \leq \frac{\lambda_1}{2}$.*

Proof of Lemma A.8. The proof is similar to the proof of Lemma 3.5. Let the index k be the smallest integer such that $\lambda_{k+1} < 2\|\widetilde{x}(t)\| - \lambda_1$. If no such index exists, then one can observe that $\|\widetilde{x}(t+1)\| \leq \lambda_1 - \|\widetilde{x}(t)\|$. Assuming that such an index exists in $[D]$, we have $\lambda_k \geq 2\|\widetilde{x}(t)\| - \lambda_1$ and $\|\widetilde{x}(t)\| - \lambda_j \leq \lambda_1 - \|\widetilde{x}(t)\|, \forall j \leq k$. With the same decomposition and estimation, since $\widetilde{x}(t) \in \cap_{j=1}^D (1+\frac{\lambda_D^2}{\lambda_1^2})\mathcal{I}_j$, we have

$$\begin{aligned}
&\left\|v^{(1)}(t)\right\| = (\lambda_1 - \|\widetilde{x}(t)\|)\|\widetilde{x}(t)\| \\
&\left\|v^{(2)}(t)\right\| \leq (1+\frac{\lambda_D^2}{\lambda_1^2})(2\|\widetilde{x}(t)\| - \lambda_1 - \lambda_k)\lambda_k \\
&\left\|v^{(2+j)}(t)\right\| \leq (1+\frac{\lambda_D^2}{\lambda_1^2})(\lambda_{k-1+j} - \lambda_{k+j})\lambda_{k+j}, \text{ for all } j \geq 1.
\end{aligned}$$

Thus we conclude

$$\begin{aligned}
\|\widetilde{x}(t+1)\| \leq& \frac{1}{\|\widetilde{x}(t)\|}\left(\left\|v^{(1)}(t)\right\| + \left\|v^{(2)}(t)\right\| + \ldots + \left\|v^{(D-k+1)}(t)\right\|\right) \\
\leq& \frac{\lambda_1}{2}(1-\frac{\lambda_D^2}{\lambda_1^2})(1+\frac{\lambda_1^2}{\lambda_D^2}) \leq \frac{\lambda_1}{2},
\end{aligned}$$

which completes the proof. □

Lemma A.9. *Consider the function $g : \mathbb{R} \to \mathbb{R}$, with $g(\lambda) = \frac{\lambda_1}{2}\left(1 - \sqrt{1 - 2\frac{\lambda}{\lambda_1}\left(1-\frac{\lambda}{\lambda_1}\right)}\right)$. For any small constant $c > 0$, consider any t with $\widetilde{x}(t) \in \cap_{j=1}^D \mathcal{I}_j$, with $\widetilde{x}(t)$ satisfying*

- $|\langle e_1, \widetilde{x}(t)\rangle| \leq (1-2c)g(\lambda_k)$.
- $\theta_t \leq \sqrt{c\,|\langle e_1, \widetilde{x}(t)\rangle|}$,

where $\theta_t = \arctan \frac{\|P^{(2:D)}(\widetilde{x}(t))\|}{|\langle e_1, \widetilde{x}(t)\rangle|}$.

Then, for any coordinate $1 \leq k \leq D$,

$$\left|\frac{\langle e_k, \widetilde{x}(t+2)\rangle}{\langle e_1, \widetilde{x}(t+2)\rangle}\right| \geq (1+c)\left|\frac{\langle e_k, \widetilde{x}(t)\rangle}{\langle e_1, \widetilde{x}(t)\rangle}\right|.$$

Proof of Lemma A.9. From the quadratic update, we have the update rule as:

$$\widetilde{x}_k(t+1) = \widetilde{x}_k(t)\left(1 - \frac{\lambda_k}{\|\widetilde{x}(t)\|}\right), \text{ for all } k \in \{1, \dots, D\}.$$

Thus, we have for any $1 \le k \le d$,

$$\begin{aligned}\left|\frac{\langle e_k, \widetilde{x}(t+2)\rangle}{\langle e_1, \widetilde{x}(t+2)\rangle}\right| &= \left|\left(1 - \frac{\lambda_1 - \lambda_k}{\lambda_1 - \|\widetilde{x}(t)\|}\right)\left(1 - \frac{\lambda_1 - \lambda_k}{\lambda_1 - \|\widetilde{x}(t+1)\|}\right)\frac{\langle e_k, \widetilde{x}(t)\rangle}{\langle e_1, \widetilde{x}(t)\rangle}\right| \\ &= \left|\left(1 - \frac{(\lambda_1 - \lambda_k)(\lambda_1 + \lambda_k - \|\widetilde{x}(t)\| - \|\widetilde{x}(t+1)\|)}{(\lambda_1 - \|\widetilde{x}(t+1)\|)(\lambda_1 - \|\widetilde{x}(t)\|)}\right)\frac{\langle e_k, \widetilde{x}(t)\rangle}{\langle e_1, \widetilde{x}(t)\rangle}\right|.\end{aligned}$$

Thus, as long as, the following holds true:

$$\frac{(\lambda_1 - \lambda_k)(\lambda_1 + \lambda_k - \|\widetilde{x}(t)\| - \|\widetilde{x}(t+1)\|)}{(\lambda_1 - \|\widetilde{x}(t+1)\|)(\lambda_1 - \|\widetilde{x}(t)\|)} \ge 2 + c,$$

we must have

$$\left|\frac{\langle e_k, \widetilde{x}(t+2)\rangle}{\langle e_1, \widetilde{x}(t+2)\rangle}\right| \ge (1+c)\left|\frac{\langle e_k, \widetilde{x}(t)\rangle}{\langle e_1, \widetilde{x}(t)\rangle}\right|.$$

We can use $(\lambda_1 - \|\widetilde{x}(t)\|)\cos\theta_t \le \|\widetilde{x}(t+1)\| \le \lambda_1 - \|\widetilde{x}(t)\| - \frac{\lambda}{2\lambda_1}\left(1 - \frac{\lambda}{\lambda_1}\right)\lambda_1 \sin^2\theta_t$, where $\lambda = \min(\lambda_1 - \lambda_2, \lambda_D)$ from Lemma A.6 to show the following with additional algebraic manipulation:

$$\frac{(\lambda_1 - \lambda_k)(\lambda_1 + \lambda_k - \|\widetilde{x}(t)\| - \|\widetilde{x}(t+1)\|)}{(\lambda_1 - \|\widetilde{x}(t+1)\|)(\lambda_1 - \|\widetilde{x}(t)\|)} \ge \frac{(\lambda_1 - \lambda_k)\lambda_k}{(\lambda_1 - (\lambda_1 - \|\widetilde{x}(t)\|)\cos\theta_t)(\lambda_1 - \|\widetilde{x}(t)\|)}.$$

Hence, it suffices to show that

$$\frac{(\lambda_1 - \lambda_k)\lambda_k}{(\lambda_1 - (\lambda_1 - \|\widetilde{x}(t)\|)\cos\theta_t)(\lambda_1 - \|\widetilde{x}(t)\|)} \ge 2 + c.$$

The left hand side can be simplified as

$$\begin{aligned}\frac{(\lambda_1 - \lambda_k)\lambda_k}{(\lambda_1 - (\lambda_1 - \|\widetilde{x}(t)\|)\cos\theta_t)(\lambda_1 - \|\widetilde{x}(t)\|)} &= \frac{(\lambda_1 - \lambda_k)\lambda_k}{(2\lambda_1 \sin^2(\theta_t/2) + |\langle e_1, \widetilde{x}(t)\rangle|)(\lambda_1 - \|\widetilde{x}(t)\|)} \\ &\ge \frac{(\lambda_1 - \lambda_k)\lambda_k}{\lambda_1\theta_t^2/2 + |\langle e_1, \widetilde{x}(t)\rangle|)(\lambda_1 - |\langle e_1, \widetilde{x}(t)\rangle|)} \\ &\ge \frac{(\lambda_1 - \lambda_k)\lambda_k}{|\langle e_1, \widetilde{x}(t)\rangle|\,(\lambda_1 + \frac{c}{2}\lambda_1 - |\langle e_1, \widetilde{x}(t)\rangle|)},\end{aligned}$$

where the last step we use that $|\theta_t| \le \sqrt{c\,|\langle e_1, \widetilde{x}(t)\rangle|}$, we only need

$$(2+c)\,|\langle e_1, \widetilde{x}(t)\rangle|^2 - 2\lambda_1(1 + c/2)(2+c)\,|\langle e_1, \widetilde{x}(t)\rangle| + (\lambda_1 - \lambda_k)\lambda_k \ge 0.$$

The above inequality is true when $|\langle e_1, \widetilde{x}(t)\rangle| \le (1 - 2c)\, g(\lambda_k)$.

□

Lemma A.10. *Consider the function $g : \mathbb{R} \to \mathbb{R}$, with $g(\lambda) = \frac{\lambda_1}{2}\left(1 - \sqrt{1 - 2\frac{\lambda}{\lambda_1}\left(1 - \frac{\lambda}{\lambda_1}\right)}\right)$. Consider any coordinate $2 \le k \le D$. For any constant $0 < c < 4\frac{\lambda_k}{\lambda_1}(1 - \frac{\lambda_k}{\lambda_1})$, consider any t with $\widetilde{x}(t) \in \cap_{j=1}^{D}\mathcal{I}_j$, with $\widetilde{x}(t)$ satisfying*

$$0.5\lambda_1 \ge \|\widetilde{x}(t)\| \ge (1+c)g(\lambda_k).$$

Then, the following must hold true at time t.

$$\left|\frac{\langle e_k, \widetilde{x}(t+2)\rangle}{\langle e_1, \widetilde{x}(t+2)\rangle}\right| \le (1 - 0.5c)\left|\frac{\langle e_k, \widetilde{x}(t)\rangle}{\langle e_1, \widetilde{x}(t)\rangle}\right|,$$

Proof. By the Normalized GD update, we have:

$$\left|\frac{\langle e_k, \widetilde{x}(t+2)\rangle}{\langle e_1, \widetilde{x}(t+2)\rangle}\right| = \left|\left(\frac{1-\frac{\lambda_k}{\|\widetilde{x}(t+1)\|}}{1-\frac{\lambda_k}{\|\widetilde{x}(t+1)\|}}\right)\left(\frac{1-\frac{\lambda_1}{\|\widetilde{x}(t)\|}}{1-\frac{\lambda_1}{\|\widetilde{x}(t)\|}}\right)\right|\left|\frac{\langle e_k, \widetilde{x}(t)\rangle}{\langle e_1, \widetilde{x}(t)\rangle}\right|$$
$$= \left|\left(1-\frac{(\lambda_1-\lambda_k)(\lambda_1+\lambda_k-\|\widetilde{x}(t)\|-\|\widetilde{x}(t+1)\|)}{(\lambda_1-\|\widetilde{x}(t+1)\|)(\lambda_1-\|\widetilde{x}(t)\|)}\right)\frac{\langle e_k, \widetilde{x}(t)\rangle}{\langle e_1, \widetilde{x}(t)\rangle}\right|. \tag{10}$$

Now, we focus on the term $\frac{(\lambda_1-\lambda_k)(\lambda_1+\lambda_k-\|\widetilde{x}(t)\|-\|\widetilde{x}(t+1)\|)}{(\lambda_1-\|\widetilde{x}(t+1)\|)(\lambda_1-\|\widetilde{x}(t)\|)}$. For simplicity, we will denote the term as $\mathrm{ratio}(\lambda_1, \lambda_k, \|\widetilde{x}(t)\|, \|\widetilde{x}(t+1)\|)$. The term behaves differently, depending on whether $\|\widetilde{x}(t)\| \geq \lambda_k$ or $\|\widetilde{x}(t)\| \leq \lambda_k$:

1. If $\|\widetilde{x}(t)\| \geq \lambda_k$, which is only possible when $\lambda_k \leq \frac{\lambda_1}{2}$, we find that $\mathrm{ratio}(\lambda_1, \lambda_k, \|\widetilde{x}(t)\|, \|\widetilde{x}(t+1)\|)$ is a monotonically decreasing function w.r.t. $\|\widetilde{x}(t+1)\|$, keeping other terms fixed. Using the fact that $\|\widetilde{x}(t+1)\| \leq \lambda_1 - \|\widetilde{x}(t)\|$ from Lemma A.6, we can bound the term as:

$$\min_{\lambda_k \leq a \leq 0.5\lambda_1} \mathrm{ratio}(\lambda_1, \lambda_k, a, \lambda_1 - a) \leq \mathrm{ratio}(\lambda_1, \lambda_k, \|\widetilde{x}(t)\|, \|\widetilde{x}(t+1)\|)$$
$$\leq \max_{\lambda_k \leq a \leq 0.5\lambda_1} \mathrm{ratio}(\lambda_1, \lambda_k, a, 0).$$

   We can simplify $\mathrm{ratio}(\lambda_1, \lambda_k, a, 0)$ as $\frac{(\lambda_1+\lambda_k-a)(\lambda_1-\lambda_k)}{\lambda_1(\lambda_1-a)}$ for any a, and can be shown to be atmost $1+\frac{\lambda_k}{\lambda_1}$ ($\leq 3/2$) for any a in the range $(\lambda_k, 0.5\lambda_1)$. Furthermore, $\mathrm{ratio}(\lambda_1, \lambda_k, a, \lambda_1 - a)$ simplifies as $\frac{\lambda_k(\lambda_1-\lambda_k)}{a(\lambda_1-a)}$ for any a, and can be shown to be atleast $4\frac{\lambda_k}{\lambda_1}(1-\lambda_k/\lambda_1)$ in the range $(\lambda_k, 0.5\lambda_1)$, which it attains at $a = \lambda_k$.

2. If $\|\widetilde{x}(t)\| \leq \lambda_k$, we find that $\mathrm{ratio}(\lambda_1, \lambda_k, \|\widetilde{x}(t)\|, \|\widetilde{x}(t+1)\|)$ is a monotonically increasing function w.r.t. $\|\widetilde{x}(t+1)\|$, keeping other terms fixed. Using the fact that $\|\widetilde{x}(t+1)\| \leq \lambda_1 - \|\widetilde{x}(t)\|$ from Lemma A.6, we can bound the term as:

$$\min_{(1+c)g(\lambda_k) \leq a \leq \min(0.5\lambda_1, \lambda_k)} \mathrm{ratio}(\lambda_1, \lambda_k, a, 0) \leq \mathrm{ratio}(\lambda_1, \lambda_k, \|\widetilde{x}(t)\|, \|\widetilde{x}(t+1)\|)$$
$$\leq \max_{(1+c)g(\lambda_k) \leq a \leq \min(0.5\lambda_1, \lambda_k)} \mathrm{ratio}(\lambda_1, \lambda_k, a, \lambda_1 - a).$$

   Continuing in the similar way as the previous case, we show that $\mathrm{ratio}(\lambda_1, \lambda_k, a, 0)$ is at least $1-(\lambda_k/\lambda_1)^2$ in the range $((1+c)g(\lambda_k), \min(0.5\lambda_1, \lambda_k))$. $\mathrm{ratio}(\lambda_1, \lambda_k, a, \lambda_1 - a)$ is maximized in the range $((1+c)g(\lambda_k), \min(0.5\lambda_1, \lambda_k))$ at $a = (1+c)g(\lambda_k)$ and is atmost $(2-0.5c)g(\lambda_k)$.

Thus, we have shown that

$$2\frac{\lambda_k}{\lambda_1}(1-\frac{\lambda_k}{\lambda_1}) \leq \min\left(4\frac{\lambda_k}{\lambda_1}(1-\frac{\lambda_k}{\lambda_1}), 1-(\frac{\lambda_k}{\lambda_1})^2\right) \leq \frac{(\lambda_1-\lambda_k)(\lambda_1+\lambda_k-\|\widetilde{x}(t)\|-\|\widetilde{x}(t+1)\|)}{(\lambda_1-\|\widetilde{x}(t+1)\|)(\lambda_1-\|\widetilde{x}(t)\|)} \leq 2-0.5c.$$

The result follows after substituting this bound in Equation (10).

□

Lemma A.11. *At any step t, if* $\|\widetilde{x}(t)\| \leq \frac{\lambda_1}{2}$, $|\tan(\angle(\widetilde{x}(t+1), e_1))| \leq \max(\frac{\lambda_2}{\lambda_1}, 1-2\frac{\lambda_D}{\lambda_1})\,|\tan(\angle(\widetilde{x}(t), e_1))|$.

Proof of Lemma A.11. From the Normalized GD update rule, we have

$$\widetilde{x}_i(t+1) = \widetilde{x}_i(t)\left(1-\frac{\lambda_i}{\|\widetilde{x}(t)\|}\right), \text{ for all } i \in [D],$$

implying $|\widetilde{x}_i(t+1)| < \left|\left(1-\frac{1}{\|\widetilde{x}(t)\|}\right)\right||\widetilde{x}_i(t)|$ for all $i \in [2, D]$, since $\lambda_i < 1$.

Since $\lambda_i < \lambda_1$ and $\|\widetilde{x}(t)\| \leq \frac{\lambda_1}{2}$, it holds that

$$\frac{|\widetilde{x}_i(t+1)|}{|\widetilde{x}_1(t+1)|} = \left|\frac{1-\frac{\lambda_i}{\|\widetilde{x}(t)\|}}{1-\frac{\lambda_1}{\|\widetilde{x}(t)\|}}\right|\frac{|\widetilde{x}_i(t)|}{|\widetilde{x}_1(t)|} = \left|1-\frac{\lambda_1-\lambda_i}{\lambda_1-\|\widetilde{x}(t)\|}\right|\frac{|\widetilde{x}_i(t)|}{|\widetilde{x}_1(t)|} \leq \max(\frac{\lambda_i}{\lambda_1}, 1-2\frac{\lambda_i}{\lambda_1})\frac{|\widetilde{x}_i(t)|}{|\widetilde{x}_1(t)|}.$$

Finally we conclude

$$\frac{\left\|P^{(2:D)}\widetilde{x}(t+1)\right\|}{|\widetilde{x}_1(t+1)|} \leq \max(\frac{\lambda_2}{\lambda_1}, 1-2\frac{\lambda_D}{\lambda_1})\frac{\left\|P^{(2:D)}\widetilde{x}(t)\right\|}{|\widetilde{x}_1(t)|}.$$

Recall $|\tan(\angle(v,e_1))| = \frac{\left\|P^{(2:D)}v\right\|}{|\langle e_1,v\rangle|}$ for any vector v, the claim follows from re-arranging the terms.

□

B. Setups for General Loss Functions

Before we start the analysis for Normalized GD for general loss functions in Appendix C, we need to introduce some new notations and terminologies to complete the formal setup. We will start by first recapping some core assumptions and definitions in the main paper.

Assumption 4.1. Assume that the loss $L : \mathbb{R}^D \to \mathbb{R}$ is a $\mathcal{C}^4$ function, and that Γ is a $(D-M)$ dimensional $\mathcal{C}^2$-submanifold of $\mathbb{R}^D$ for some integer $1 \leq M \leq D$, where for all $x \in \Gamma, x$ is a local minimizer of L with $L(x) = 0$ and $\operatorname{rank}\left(\nabla^2 L(x)\right) = M$.

Assumption 4.2. Assume that U is an open neighborhood of Γ satisfying that gradient flow w.r.t. L starting in U converges to some point in Γ, i.e. for all $x \in U$, $\Phi(x) \in \Gamma$. (Then Φ is $\mathcal{C}^3$ on U (Falconer, 1983)).

Notations: We define $\Phi : U \to \Gamma$ as the limit map of gradient flow below. We summarize various properties of Φ from (Li et al., 2022) in Appendix B.2.

$$\Phi(x) = \lim_{\tau\to\infty} \phi(x,\tau), \quad \text{where} \quad \phi(x,\tau) = x - \int_0^\tau \nabla L(\phi(x,s))ds. \tag{11}$$

For a matrix $A \in \mathbb{R}^{D\times D}$, we denote its eigenvalue-eigenvector pairs by $\{\lambda_i(A), v_i(A))\}_{i\in[D]}$. For simplicity, whenever Φ is defined and $\mathcal{C}^2$ at point x, we use $\{(\lambda_i(x), v_i(x))\}_{i=1}^D$ to denote the eigenvector-eigenvalue pairs of $\nabla^2 L(\Phi(x))$, with $\lambda_1(x) > \lambda_2(x) \geq \lambda_3(x) \ldots \geq \lambda_D(x)$. Given a differentiable submanifold Γ of $\mathbb{R}^D$ and point $x \in \Gamma$, we use $\mathrm{N}_x\Gamma$ and $\mathrm{T}_x\Gamma$ to denote the normal space and the tangent space of the manifold Γ for any point $x \in \Gamma$. We use $P_{x,\Gamma} : \mathbb{R}^D \to \mathbb{R}^D$ to denote the projection operator onto the normal space of Γ at x, and $P^\perp_{x,\Gamma} := I_D - P_{x,\Gamma}$. Similar to quadratic case, for any $x \in U$, we use $\widetilde{x}$ to denote $\nabla^2 L(\Phi(x))(x - \Phi(x))$ for notational convenience. Additionally, for any $x \in U$, we use $\theta(x)$ to denote the angle between $\widetilde{x}$ and the top eigenspace of the hessian at $\Phi(x)$, i.e. $\theta(x) = \arctan \frac{\left\|P^{(2:M)}_{\Phi(x),\Gamma}\widetilde{x}\right\|}{|\langle v_1(x),\widetilde{x}\rangle|}$. Furthermore, when the iterates $x(t)$ is clear in the context, we use shorthand $\lambda_i(t) := \lambda_i(x(t))$, $v_i(t) := v_i(x(t))$, $P_{t,\Gamma} := P_{\Phi(x(t)),\Gamma}$, $P^\perp_{t,\Gamma} := P^\perp_{\Phi(x(t)),\Gamma}$ and θ_t to denote $\theta(x(t))$ when $x(t)$ is clear in the context. We define the function $g_t : \mathbb{R} \to \mathbb{R}$ for every $t \in \mathbb{N}$ as

$$g_t(\lambda) = \frac{1}{2}\left(1 - \sqrt{1 - 2\frac{\lambda}{\lambda_1(t)}\left(1 - \frac{\lambda}{\lambda_1(t)}\right)}\right).$$

Given any two points x, y, we use $\overline{xy}$ to denote the line segment between x and y, *i.e.*, $\{z \mid \exists\lambda \in [0,1], z = (1-\lambda)x + \lambda y\}$.

The main result of this paper focuses on the trajectory of Normalized GD from fixed initialization x_{init} with LR η converges to 0, which can be roughly split into two phases. In the first phase, Theorem 4.3 shows that the normalized GD trajectory converges to the gradient flow trajectory, $\phi(x_{\text{init}}, \cdot)$. In second phase, Theorem 4.4 shows that the normalized GD trajectory converges to the limiting flow which decreases sharpness on Γ, (4). Therefore, for sufficiently small η, the entire trajectory of normalized GD will be contained in a small neighbourhood of gradient flow trajectory Z and limiting flow trajectory Y. The convergence rate given by our proof depends on the various local constants like smoothness of L and Φ in this small neighbourhood, which intuitively can be viewed as the actual "working zone" of the algorithm. The constants are upper bounded or lower bounded from zero because this "working zone" is compact after fixing the stopping time of (4), which is denoted by T_2.

$$X(\tau) = \Phi(x_{\text{init}}) - \frac{1}{4}\int_{s=0}^{\tau} P^\perp_{X(s),\Gamma}\nabla \log\lambda_1(X(s))\mathrm{d}s \tag{4}$$

Below we give formal definitions of the "working zones" and the corresponding properties. For any point $y \in \mathbb{R}^D$ and positive number r, we define $B_r(y) := \{x \in \mathbb{R}^D \mid \|y - x\| < r\}$ as the open ℓ_2 norm ball centered at y and $\overline{B}_r(y)$ as its closure. For any set S and positive number r, we define $S^r := \cup_{y \in S} B_r(y)$ and $\overline{B}_r(S) := \cup_{y \in S} \overline{B}_r(y)$. Given any stopping time $T_2 > 0$, we denote the trajectory of limiting flow Equation (4) $\{X(\tau)\}_{\tau=0}^{T_2}$ by Y and we define $Y^{\boldsymbol{\epsilon}} := \cup_{y \in Y} \overline{B}_y(\boldsymbol{\epsilon})$ where $\boldsymbol{\epsilon}$ is some sufficiently small constant determined later in Lemma B.3. $Y^{\boldsymbol{\epsilon}}$ will be the "working zone" of Normalized GD in the second phase. By definition, $Y^{\boldsymbol{\epsilon}}$ are compact. To ensure Equation (4) is well-defined, we have to make T_2 small enough such that (1) $Y \in U$, with $\Phi(\cdot)$ being well-defined along Y, and (2) $\lambda_1(\nabla^2 L(\cdot))$ is differentiable, which yields the following definition of T_2^{fav}.

Definition B.1. T_2^{fav} is the set of all time T_2 such that for any $T_2 \in T_2^{\text{fav}}$, (4) is well-defined up to time T_2, *i.e.*, for any $0 \le \tau \le T_2$, we have

1. $X(\tau) \in U$;
2. $\lambda_1(\nabla^2 L(X(\tau))) - \lambda_2(\nabla^2 L(X(\tau))) > 0$.

For convenience, we define $\boldsymbol{\Delta} := \frac{1}{2} \inf_{x \in Y} \left(\lambda_1(\nabla^2 L(x)) - \lambda_2(\nabla^2 L(x))\right)$ and $\boldsymbol{\mu} := \frac{1}{4} \inf_{x \in Y^{\boldsymbol{\rho}}} \lambda_M(\nabla^2 L(x))$. By Assumption 4.1 and the first bullet of Definition B.1, we have $\boldsymbol{\mu} > 0$. By the second bullet of Definition B.1, $\boldsymbol{\Delta} > 0$.

Below we construct the "working zone" of the second phase, $Y^{\boldsymbol{\rho}}$ and $Y^{\boldsymbol{\epsilon}}$, where $0 < \boldsymbol{\epsilon} < \boldsymbol{\rho}$, implying $Y^{\boldsymbol{\epsilon}} \subset Y^{\boldsymbol{\rho}}$. The reason that we need the two-level nested "working zones" is that even though we can ensure all the points in $Y^{\boldsymbol{\rho}}$ have nice properties as listed in Lemma B.3, we cannot ensure the trajectory of gradient flow from $x \in Y^{\boldsymbol{\rho}}$ to $\Phi(x)$ or the line segment $\overline{x\Phi(x)}$ is in $Y^{\boldsymbol{\rho}}$, which will be crucial for the geometric lemmas (in Appendix B.1) that we will heavily use in the trajectory analysis around the manifold. For this reason we further define $Y^{\boldsymbol{\epsilon}}$ and Lemma B.6 guarantees the trajectory of gradient flow from x to $\Phi(x)$ or the line segment $\overline{x\Phi(x)}$ whenever $x \in Y^{\boldsymbol{\rho}}$.

Definition B.2 (PL condition)**.** A function L is said to be $\boldsymbol{\mu}$-PL in a set U iff for all $x \in U$,

$$\|\nabla L(x)\|^2 \ge 2\boldsymbol{\mu}(L(x) - \inf_{x \in U} L(x)).$$

Lemma B.3. *Given Y, there are sufficiently small $\boldsymbol{\rho} > 0$ such that*

1. *$Y^{\boldsymbol{\rho}} \cap \Gamma$ is compact;*
2. *$Y^{\boldsymbol{\rho}} \subset U$;*
3. *L is $\boldsymbol{\mu}$-PL on $Y^{\boldsymbol{\rho}}$; (see Definition B.2)*
4. *$\inf_{x \in Y^{\boldsymbol{\rho}}} \left(\lambda_1(\nabla^2 L(x)) - \lambda_2(\nabla^2 L(x))\right) \ge \boldsymbol{\Delta} > 0$;*
5. *$\inf_{x \in Y^{\boldsymbol{\rho}}} \lambda_M(\nabla^2 L(x)) \ge \boldsymbol{\mu} > 0$.*

Proof of Lemma B.3. We first claim for every $y \in Y$, for all sufficiently small $\boldsymbol{\rho}_y > 0$ (*i.e.* for all $\boldsymbol{\rho}_y$ smaller than some threshold depending on y), the following three properties hold (1) $\overline{B}_y(\boldsymbol{\rho}_y) \cap \Gamma$ is compact; (2) $\overline{B}_y(\boldsymbol{\rho}_y) \cap \Gamma \subset U$ and (3) L is $\boldsymbol{\mu}$-PL on $\overline{B}_y(\boldsymbol{\rho}_y \cap \Gamma)$.

Among the above three claims, (2) is immediate. (1) holds because $\overline{B}_y(\boldsymbol{\rho}_y) \cap \Gamma$ is bounded and we can make $\boldsymbol{\rho}_y$ small enough to ensure $\overline{B}_y(\boldsymbol{\rho}_y) \cap \Gamma$ is closed. For (3), by Proposition 7 of (Fehrman et al., 2020), we define $p(y) := \operatorname{argmin}_{x \in \Gamma} \|x - y\|$ which is uniquely defined and $\mathcal{C}^1$ in $\overline{B}_y(\boldsymbol{\rho}_y)$ for sufficiently small $\boldsymbol{\rho}_y$. Moreover, Lemma 14 in (Fehrman et al., 2020) shows that $\left\|\nabla L(x) - \nabla^2 L(p(x))(x - p(x))\right\| \le c \|x - p(x)\|_2^2$ for all x in $B_y(\boldsymbol{\rho}_y)$ uniformly and some constant c. Thus for small enough $\boldsymbol{\rho}_y$,

$$\|\nabla L(x)\|^2 \ge (x - p(x))^\top (\nabla^2 L(p(x)))^2 (x - p(x)) - O(\|x - p(x)\|^3) \tag{12}$$

Furthermore, by Lemma 10 in (Fehrman et al., 2020), it holds that $x - p(x) \in \mathrm{N}_{p(x)}\Gamma = \operatorname{span}(\{v_i(p(x))\}_{i=1}^M)$, which implies

$$(x - p(x))^\top (\nabla^2 L(p(x)))^2 (x - p(x)) \ge \lambda_M(\nabla^2 L(p(x)))(x - p(x))^\top \nabla^2 L(p(x))(x - p(x)), \tag{13}$$

and that

$$(x - p(x))^\top \nabla^2 L(p(x))(x - p(x)) \ge \lambda_M(\nabla^2 L(p(x))) \|x - p(x)\|_2^2 .$$

Thus for any $c' > 0$, for sufficiently small $\boldsymbol{\rho}_y$, $(x - p(x))^\top \nabla^2 L(p(x))(x - p(x)) \geq c' \|x - p(x)\|^3$. Combining Equations (12) and (13), we conclude that for sufficiently small $\boldsymbol{\rho}_y$,

$$\|\nabla L(x)\|^2 \geq \lambda_M(\nabla^2 L(p(x)))(x - p(x))^\top \nabla^2 L(p(x))(x - p(x)) - O(\|x - p(x)\|_2^3)$$

Again for sufficiently small $\boldsymbol{\rho}_y$, by Taylor expansion of L at $p(x)$, we have

$$\frac{1}{2}(x - p(x))^\top \nabla^2 L(p(x))(x - p(x)) \geq L(x) - O(\|x - p(x)\|^3).$$

Thus we conclude

$$\|\nabla L(x)\|^2 \geq 2\lambda_M(\nabla^2 L(p(x)))L(x) - O(\|x - p(x)\|^3) \geq \lambda_M(\nabla^2 L(p(x)))L(x) \geq 2\boldsymbol{\mu} L(x).$$

Meanwhile, since $\lambda_M(\nabla^2 L(p(x)))$ and $\lambda_1(\nabla^2 L(p(x))) - \lambda_2(\nabla^2 L(p(x)))$ are continuous functions in x, we can also choose a sufficiently small $\boldsymbol{\rho}_y$ such that for all $x \in \overline{B}_y(\boldsymbol{\rho}_y)$, $\lambda_M(\nabla^2 L(p(x))) \geq \frac{1}{2}\lambda_M(\nabla^2 L(p(y))) = \frac{1}{2}\lambda_M(\nabla^2 L(y)) > \boldsymbol{\Delta}$ and $\lambda_1(\nabla^2 L(p(x))) - \lambda_2(\nabla^2 L(p(x))) \geq \frac{1}{2}\big(\lambda_1(\nabla^2 L(p(y))) - \lambda_2(\nabla^2 L(p(y)))\big) = \frac{1}{2}\big(\lambda_1(\nabla^2 L(y)) - \lambda_2(\nabla^2 L(y))\big) \geq \boldsymbol{\mu}$. Further note $Y \subset \cup_{y \in Y} B_y(\boldsymbol{\rho}_y)$ and Y is a compact set, we can take a finite subset of Y, Y', such that $Y \subset \cup_{y \in Y'} B_y(\boldsymbol{\rho}_y)$. Taking $\boldsymbol{\rho} := \min_{y \in Y'} \frac{\boldsymbol{\rho}_y}{2}$ completes the proof. □

Definition B.4. The spectral 2-norm of a k-order tensor $\mathcal{T} = (t_{i_1 i_2 \cdots i_k}) \in \mathbb{R}^{d_1 \times d_2 \times \cdots d_k}$ is defined as the maximum of the following constrained multilinear optimization problem:

$$\|\mathcal{T}\| = \max\left\{\mathcal{T}\left(x^{(1)}, \cdots, x^{(k)}\right) : \left\|x^{(i)}\right\|_2 = 1, x^{(i)} \in \mathbb{R}^{d_i}, i = 1, 2, \ldots, k\right\}.$$

Here, $\mathcal{T}\left(x^1, \cdots, x^{(k)}\right) = \sum_{i_1=1}^{d_1} \sum_{i_2=1}^{d_2} \cdots \sum_{i_k=1}^{d_k} t_{i_1 i_2 \ldots i_d} x_{i_1}^{(1)} x_{i_2}^{(2)} \ldots x_{i_k}^{(k)}$.

Definition B.5. We define the following constants regarding smoothness of L and Φ of various orders over $Y^{\boldsymbol{\rho}}$.

$$\boldsymbol{\zeta} = \sup_{x \in Y^{\boldsymbol{\rho}}} \left\|\nabla^2 L(x)\right\|, \quad \boldsymbol{\nu} = \sup_{x \in Y^{\boldsymbol{\rho}}} \left\|\nabla^3 L(x)\right\|, \quad \boldsymbol{\Upsilon} = \sup_{x \in Y^{\boldsymbol{\rho}}} \left\|\nabla^4 L(x)\right\|,$$
$$\boldsymbol{\xi} = \sup_{x \in Y^{\boldsymbol{\rho}}} \left\|\nabla^2 \Phi(x)\right\|, \quad \boldsymbol{\chi} = \sup_{x \in Y^{\boldsymbol{\rho}}} \left\|\nabla^3 \Phi(x)\right\|,$$

We assume each of the constants $\boldsymbol{\zeta}, \boldsymbol{\nu}, \boldsymbol{\Upsilon}, \boldsymbol{\xi}, \boldsymbol{\chi}$ are at least 1 for simplicity (otherwise we can set them to be 1 and our bound still holds)

Lemma B.6. *Given $\boldsymbol{\rho}$ as defined in Lemma B.3, there is an $\boldsymbol{\epsilon} \in (0, \boldsymbol{\rho})$ such that*

1. $\sup_{x \in Y^{\boldsymbol{\epsilon}}} L(x) - \inf_{x \in Y^{\boldsymbol{\epsilon}}} L(x) < \frac{\boldsymbol{\mu}\boldsymbol{\rho}^2}{8}$;
2. $\forall x \in Y^{\boldsymbol{\epsilon}}$, $\Phi(x) \in Y^{\frac{\boldsymbol{\rho}}{2}}$;
3. $\boldsymbol{\epsilon} \leq \frac{2\boldsymbol{\mu}^2}{\boldsymbol{\nu}\boldsymbol{\zeta}}$.

Proof of Lemma B.6. For every $y \in Y$, there is an $\boldsymbol{\epsilon}_y$, such that $\forall x \in B_y(\boldsymbol{\epsilon}_y)$, it holds that $L(x) < \frac{\boldsymbol{\mu}\boldsymbol{\rho}^2}{8}$, $\Phi(x) \in Y^{\frac{\boldsymbol{\rho}}{2}}$ and $\lambda_1(\nabla^2 L(x)) - \lambda_2(\nabla^2 L(x))) \geq \boldsymbol{\Delta}$, as both $L(x)$, $\Phi(x)$ and eigenvalue functions are continuous. Further note $Y \subset \cup_{y \in Y} B_y(\boldsymbol{\epsilon}_y)$ and Y is a compact set, we can take a finite subset of Y, Y', such that $Y \subset \cup_{y \in Y'} B_y(\boldsymbol{\epsilon}_y)$. Taking $\boldsymbol{\epsilon} := \min\{\min_{y \in Y'} \frac{\boldsymbol{\epsilon}_y}{2}, \frac{2\boldsymbol{\mu}^2}{\boldsymbol{\nu}\boldsymbol{\zeta}}\}$ completes the proof. □

Summary for Setups: The initial point x_{init} is chosen from an open neighborhood of manifold Γ, U, where the infinite-time limit of gradient flow Φ is well-defined and for any $x \in U$, $\Phi(x) \in \Gamma$. We consider normalized GD with sufficiently small LR η such that the trajectory enters a small neighborhood of limiting flow trajectory, $Y^{\boldsymbol{\rho}}$. Moreover, L is $\boldsymbol{\mu}$-PL on $Y^{\boldsymbol{\rho}}$ and the eigengaps and smallest eigenvalues are uniformly lower bounded by positive $\boldsymbol{\Delta}$, $\boldsymbol{\mu}$ respectively on $Y^{\boldsymbol{\rho}}$. Finally, we consider a proper subset of $Y^{\boldsymbol{\rho}}$, $Y^{\boldsymbol{\epsilon}}$, as the final "working zone" in the second phase (defined in Lemma B.6), which enjoys more properties than $Y^{\boldsymbol{\rho}}$, including Lemmas B.8 to B.11.

B.1. Geometric Lemmas

In this subsection we present several geometric lemmas which are frequently used in the trajectory analysis of normalized GD. Below is a brief summary:

- Lemma B.7: Inequalities connecting various terms: the distance between x and $\Phi(x)$, the length of GF trajectory from x to $\Phi(x)$, square root of loss and gradient norm;
- Lemma B.8: For any $x \in Y^{\boldsymbol{\epsilon}}$, the gradient flow trajectory from x to $\Phi(x)$ and the line segment between x and $\Phi(x)$ are all contained in $Y^{\boldsymbol{\rho}}$, so it's "safe" to use Taylor expansions along GF trajectory or $\overline{x\Phi(x)}$ to derive properties;
- Lemmas B.9 to B.11: for any $x \in Y^{\boldsymbol{\epsilon}}$, the normalized GD dynamics at x can be roughly viewed as approximately quadratic around $\Phi(x)$ with positive definite matrix $\nabla^2 L(\Phi(x))$.
- Lemma B.12: In the "working zone", $Y^{\boldsymbol{\rho}}$, one-step normalized GD update with LR η only changes $\Phi(x_t)$ by $O(\eta^2)$.
- Lemma B.14: In the "working zone", $Y^{\boldsymbol{\rho}}$, one-step normalized GD update with LR η decreases $\sqrt{L(x) - \min_{y\in Y} L(y)}$ by $\eta\frac{\sqrt{2\boldsymbol{\mu}}}{4}$ if $\|\nabla L(x)\| \geq \frac{\zeta}{\eta}$.

Lemma B.7. *If the trajectory of gradient flow starting from x, $\phi(x,t)$, stays in $Y^{\boldsymbol{\rho}}$ for all $t \geq 0$, then we have*

$$\|x - \Phi(x)\| \leq \int_{t=0}^{\infty} \left\| \frac{\mathrm{d}\phi(x,t)}{\mathrm{d}t} \right\| \mathrm{d}t \leq \sqrt{\frac{2(L(x) - L(\Phi(x)))}{\boldsymbol{\mu}}} \leq \frac{\|\nabla L(x)\|}{\boldsymbol{\mu}}.$$

Proof of Lemma B.7. Since $\Phi(x)$ is defined as $\lim_{t\to\infty}\phi(x,t)$ and $\phi(x,0) = x$, the left-side inequality follows immediately from triangle inequality. The right-side inequality is by the definition of PL condition. Below we prove the middle inequality.

Since $\forall t \geq 0$, $\phi(x,t) \in Y^{\boldsymbol{\rho}}$, it holds that $\|\nabla L(\phi(x,t))\|^2 \geq 2\boldsymbol{\mu}(L(\phi(x,t)) - L(\Phi(x)))$ by the choice of $\boldsymbol{\rho}$ in Lemma B.3. Without loss of generality, we assume $L(y) = 0, \forall y \in \Gamma$. Thus we have

$$\int_{t=0}^{\infty} \|\nabla L(\phi(x,t))\| \,\mathrm{d}t \leq \int_{t=0}^{\infty} \frac{\|\nabla L(\phi(x,t))\|^2}{\sqrt{2\boldsymbol{\mu}L(\phi(x,t))}} \mathrm{d}t.$$

Since $\mathrm{d}\phi(x,t) = -\nabla L(\phi(x,t))\mathrm{d}t$, if holds that

$$\int_{t=0}^{\infty} \frac{\|\nabla L(\phi(x,t))\|^2}{\sqrt{2\boldsymbol{\mu}L(\phi(x,t))}} \mathrm{d}t \leq \int_{t=0}^{\infty} \frac{-\mathrm{d}L(\phi(x,t))}{\sqrt{2\boldsymbol{\mu}L(\phi(x,t))}} = \int_{t=0}^{\infty} \sqrt{\frac{2}{\boldsymbol{\mu}}} \mathrm{d}\sqrt{L(\phi(x,t))} = \sqrt{\frac{2L(\phi(x,0))}{\boldsymbol{\mu}}}.$$

The proof is complete since $\phi(x,0) = x$ and we assume $L(\Phi(x))$ is 0. □

Lemma B.8. *Let $\boldsymbol{\rho}, \boldsymbol{\epsilon}$ be defined in Lemmas B.3 and B.6. For any $x \in Y^{\boldsymbol{\epsilon}}$, we have*

1. *The entire trajectory of gradient flow starting from x is contained in $Y^{\boldsymbol{\rho}}$,* i.e., *$\phi(x,t) \in Y^{\boldsymbol{\rho}}$, $\forall t \geq 0$;*
2. *For any $t \geq 0$, line segment $\overline{\Phi(x)\phi(x,t)}$ is contained in Y^{ρ},* i.e., *$\|\Phi(x) - \phi(x,t)\| \leq \boldsymbol{\rho}$, $\forall t \geq 0$.*

Proof of Lemma B.8. Let time $\tau^* > 0$ be the smallest time after which the trajectory of GF is completely contained in $Y^{\boldsymbol{\rho}}$, that is, $\tau^* := \inf\{t \geq 0 \mid \forall t' \geq t, \phi(x,t') \in Y^{\boldsymbol{\rho}}\}$. Since $Y^{\boldsymbol{\rho}}$ is closed and $\phi(x,\cdot)$ is continuous, we have $\phi(x,\tau^*) \in Y^{\boldsymbol{\rho}}$.

Since $\forall \tau \geq \tau^*$, $\phi(x,\tau) \in Y^{\boldsymbol{\rho}}$, by Lemma B.7, it holds that $\|\phi(x,\tau^*) - \Phi(x)\| \leq \sqrt{\frac{2(L(\phi(x,\tau^*)) - L(\Phi(x)))}{\boldsymbol{\mu}}}$.

Note that loss doesn't increase along GF, we have $L(\phi(x,\tau^*)) - L(\Phi(x)) \leq L(x) - L(\Phi(x)) \leq \frac{\boldsymbol{\mu}\boldsymbol{\rho}^2}{8}$, which implies that $\|\phi(x,\tau^*) - \Phi(x)\| \leq \frac{\boldsymbol{\rho}}{2}$.

Now we prove the first claim. Suppose GF trajectory starting from x leaves $Y^{\boldsymbol{\rho}}$, or equivalently $\tau^* > 0$, since $\lim_{\tau\to\infty}\phi(x,\tau) = \Phi(x) \in Y^{\frac{\boldsymbol{\rho}}{2}}$, there must exist a time $\tau^* > 0$ such that $\phi(x,\tau^*)$ is on the boundary of $Y^{\boldsymbol{\rho}}$, that is, $\inf_{y\in Y}\|y - \phi(x,\tau^*)\| = \boldsymbol{\rho}$. By triangle inequality, this implies that $\|\Phi(x) - \phi(x,\tau^*)\| \geq \frac{\boldsymbol{\rho}}{2}$. Contradiction!

The second claim also follows from the same estimation. Since $\tau^* = 0$, for any $t \geq 0$, $\|\phi(x,t) - \Phi(x)\| \leq \int_{\tau=t}^{\infty} \|\nabla L(\phi(x,\tau))\| \,\mathrm{d}\tau \leq \int_{\tau=\tau^*}^{\infty} \|\nabla L(\phi(x,\tau))\| \,\mathrm{d}\tau \leq \frac{\boldsymbol{\rho}}{2}$.

□

The following theorem shows that the projection of x in the tangent space of $\Phi(x)$ is small when x is close to the manifold. In particular if we can show that in a discrete trajectory with a vanishing learning rate η, the iterates $\{x_\eta(t)\}$ stay in $Y^{\boldsymbol{\epsilon}}$, we can interchangeably use $\|x_\eta(t) - \Phi(x_\eta(t))\|$ with $\|P_{t,\Gamma}(x_\eta(t) - \Phi(x_\eta(t)))\|$, with an additional error of $\mathcal{O}(\eta^3)$, when $\|P_{t,\Gamma}(x_\eta(t) - \Phi(x_\eta(t)))\| \leq \mathcal{O}(\eta)$.

Lemma B.9. *For all $x \in Y^{\boldsymbol{\epsilon}}$, we have*

$$\left\|P^{\perp}_{\Phi(x),\Gamma}(x - \Phi(x))\right\| \leq \frac{\boldsymbol{\nu\zeta}}{4\boldsymbol{\mu}^2} \|x - \Phi(x)\|^2,$$

and

$$\left\|P_{\Phi(x),\Gamma}(x - \Phi(x))\right\| \geq \|x - \Phi(x)\| \left(1 - \frac{\boldsymbol{\nu}^2\boldsymbol{\zeta}^2}{16\boldsymbol{\mu}^4} \|x - \Phi(x)\|^2\right) \geq \frac{3}{4} \|x - \Phi(x)\|.$$

Proof. First of all, we can track the decrease in loss along the Gradient flow trajectory starting from x. At any time τ, we have

$$\frac{d}{d\tau} L(\phi(x,\tau)) = \langle \nabla L(\phi(x,\tau)), \frac{d}{d\tau}\phi(x,\tau)\rangle = -\|\nabla L(\phi(x,\tau))\|^2,$$

where $\phi(x,0) = x$. Without loss of generality, we assume $L(y) = 0, \forall y \in \Gamma$. Using the fact that L is $\boldsymbol{\mu}$-PL on $Y^{\boldsymbol{\rho}}$ and the GF trajectory starting from any point in $Y^{\boldsymbol{\epsilon}}$ stays inside $Y^{\boldsymbol{\rho}}$ (from Lemma B.8), we have

$$\frac{d}{d\tau} L(\phi(x,\tau)) \leq -2\boldsymbol{\mu} L(\phi(x,\tau)),$$

which implies

$$L(\phi(x,\tau)) \leq L(\phi(x,0)) e^{-2\boldsymbol{\mu}\tau}$$

By Lemma B.7, we have

$$\|\phi(x,\tau) - \Phi(x)\| \leq \sqrt{\frac{2}{\boldsymbol{\mu}}}\sqrt{L}(\phi(x,\tau)) \leq \sqrt{\frac{2L(\phi(x,0))e^{-2\boldsymbol{\mu}\tau}}{\boldsymbol{\mu}}}. \tag{14}$$

Moreover, we can relate $L(\phi(x,0)$ with $\|\Phi(x) - x\|$ with a second order taylor expansion:

$$L(x) = L(\Phi(x)) + \langle \nabla L(\Phi(x)), x - \Phi(x)\rangle + \int_{s=0}^{1} (1-s)(x - \Phi(x))^\top \nabla^2 L(sx + (1-s)\Phi(x))(x - \Phi(x)) ds$$

where in the final step, we have used the fact that $L(\Phi(x)) = 0$ and $\nabla L(\Phi(x)) = 0$. By Lemma B.8, we have $\overline{x\Phi(x)} \subset Y^{\boldsymbol{\rho}}$. Thus $\max_{s\in[0,1]} \left\|\nabla^2 L(sx + (1-s)\Phi(x))\right\| \leq \boldsymbol{\zeta}$ from Definition B.5 and it follows that

$$L(x) \leq \int_{s=0}^{1} (1-s)\boldsymbol{\zeta} \|x - \Phi(x)\|^2 ds = \frac{\boldsymbol{\zeta}}{2} \|\Phi(x) - x\|^2, \tag{15}$$

Finally we focus on the movement in the tangent space. It holds that

$$\left\|P^{\perp}_{\Phi(x),\Gamma}(\phi(x,\infty) - \phi(x,0))\right\| \leq \left\|P^{\perp}_{\Phi(x),\Gamma} \int_0^{\infty} \nabla L(\phi(x,\tau))\right\| d\tau \leq \int_0^{\infty} \left\|P^{\perp}_{\Phi(x),\Gamma} \nabla L(\phi(x,\tau))\right\| d\tau. \tag{16}$$

By Lemma B.8, we have $\overline{\phi(x,\tau)\Phi(x)} \subset Y^{\rho}$ for all $\tau \geq 0$ and thus

$$\left\|\nabla L(\phi(x,\tau)) - \nabla^2 L(\Phi(x))\big(\Phi(x)\big)\big(\phi(x,\tau) - \Phi(x)\big)\right\| \leq \frac{\boldsymbol{\nu}}{2}\left\|\phi(x,\tau) - \Phi(x)\right\|^2.$$

Since $P^{\perp}_{\Phi(x),\Gamma}$ is the projection matrix for the tangent space, $P^{\perp}_{\Phi(x),\Gamma}\nabla^2 L(\Phi(x)) = 0$ and thus by Equation (14)

$$\left\|P^{\perp}_{\Phi(x),\Gamma}\nabla L(\phi(x,\tau))\right\| \leq \frac{\boldsymbol{\nu}}{2}\left\|\phi(x,\tau) - \Phi(x)\right\|^2 \leq \frac{\boldsymbol{\nu} L(\phi(x,0))e^{-2\boldsymbol{\mu}\tau}}{\boldsymbol{\mu}} \tag{17}$$

Plug Equation (17) into Equation (16), we conclude that

$$\left\|P^{\perp}_{\Phi(x),\Gamma}(\phi(x,\infty) - x)\right\| \leq \int_{\tau=0}^{\infty} \frac{\boldsymbol{\nu} L(\phi(x,0))e^{-2\boldsymbol{\mu}\tau}}{\boldsymbol{\mu}} = \frac{\boldsymbol{\nu} L(x)}{2\boldsymbol{\mu}^2} \leq \frac{\boldsymbol{\nu}\boldsymbol{\zeta}\left\|x - \Phi(x)\right\|^2}{4\boldsymbol{\mu}^2} \tag{18}$$

For the second claim, simply note that

$$\left\|P^{\perp}_{\Phi(x),\Gamma}(x - \Phi(x))\right\| = \sqrt{\left\|x - \Phi(x)\right\|^2 - \left\|P_{\Phi(x),\Gamma}(x - \Phi(x))\right\|^2} \geq \left\|x - \Phi(x)\right\| - \frac{\left\|P_{\Phi(x),\Gamma}(x - \Phi(x))\right\|^2}{\left\|x - \Phi(x)\right\|}$$

The left-side inequality of the second inequality is proved by plugging the first claim into the above inequality. Note by the property (3) in Lemma B.6, $\frac{\boldsymbol{\nu}^2\boldsymbol{\zeta}^2}{16\boldsymbol{\mu}^4}\left\|x - \Phi(x)\right\|^2 \leq \frac{1}{4}$, the right-side inequality is also proved. □

Lemma B.10. *At any point $x \in Y^{\boldsymbol{\epsilon}}$, we have*

$$\left\|\nabla L(x) - \nabla^2 L(\Phi(x))(x - \Phi(x))\right\| \leq \frac{1}{2}\boldsymbol{\nu}\left\|x - \Phi(x)\right\|^2.$$

and

$$\left|\frac{\|\nabla L(x)\|}{\|\nabla^2 L(\Phi(x))(x - \Phi(x))\|} - 1\right| \leq \frac{4\boldsymbol{\nu}}{3\boldsymbol{\mu}}\left\|x - \Phi(x)\right\|,$$

Moreover, the normalized gradient of L can be written as

$$\frac{\nabla L(x)}{\|\nabla L(x)\|} = \frac{\nabla^2 L(\Phi(x))(x - \Phi(x))}{\|\nabla^2 L(\Phi(x))(x - \Phi(x))\|} + \mathcal{O}(\frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}\left\|x - \Phi(x)\right\|). \tag{19}$$

Proof of Lemma B.10. Using taylor expansion at x, we have using $\nabla L(\Phi(x)) = 0$:

$$\begin{aligned}
\left\|\nabla L(x) - \nabla^2 L(\Phi(x))(x - \Phi(x))\right\| &= \left\|\int_0^1 (1-s)\partial^2(\nabla L)(sx + (1-s)\Phi(x))[x - \Phi(x), x - \Phi(x)]ds\right\| \\
&\leq \left\|\int_0^1 (1-s)ds\right\| \max_{0\leq s\leq 1}\left\|\partial^2(\nabla L)(sx + (1-s)\Phi(x))\right\|\left\|x - \Phi(x)\right\|^2 \\
&\leq \frac{1}{2}\boldsymbol{\nu}\left\|x - \Phi(x)\right\|^2.
\end{aligned}$$

Further note $\left\|\nabla^2 L(\Phi(x))(x - \Phi(x))\right\| \geq \left\|P_{\Phi(x),\Gamma}\nabla^2 L(\Phi(x))(x - \Phi(x))\right\| = \left\|\nabla^2 L(\Phi(x))P_{\Phi(x),\Gamma}(x - \Phi(x))\right\| \geq \boldsymbol{\mu}\left\|P_{\Phi(x),\Gamma}(x - \Phi(x))\right\|$, we have

$$\left|\frac{\|\nabla L(x)\|}{\|\nabla^2 L(\Phi(x))(x - \Phi(x))\|} - 1\right| \leq \frac{\boldsymbol{\nu}\left\|x - \Phi(x)\right\|^2}{\boldsymbol{\mu}\left\|P_{\Phi(x),\Gamma}(x - \Phi(x))\right\|} \leq \frac{4\boldsymbol{\nu}}{3\boldsymbol{\mu}}\left\|x - \Phi(x)\right\|,$$

where we use Lemma B.9 since $x \in Y^{\boldsymbol{\epsilon}}$. Thus, the normalized gradient at any step t can be written as

$$\frac{\nabla L(x)}{\|\nabla L(x)\|} = \frac{\nabla^2 L(\Phi(x))[x - \Phi(x)] + \mathcal{O}(\boldsymbol{\nu}\left\|x - \Phi(x)\right\|^2)}{\|\nabla^2 L(\Phi(x))(x - \Phi(x))\|\left(1 + \mathcal{O}\left(\frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}\left\|x - \Phi(x)\right\|\right)\right)}.$$

$$= -\frac{\nabla^2 L(\Phi(x))[x-\Phi(x)]}{\|\nabla^2 L(\Phi(x))[x-\Phi(x)]\|} + \mathcal{O}(\frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}\|x-\Phi(x)\|),$$

which completes the proof. □

Lemma B.11. *Consider any point $x \in Y^{\boldsymbol{\epsilon}}$. Then,*

$$\left|\left\langle v_1(x), \frac{\nabla L(x)}{\|\nabla L(x)\|}\right\rangle\right| \geq \cos\theta - \mathcal{O}(\frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}\|x-\Phi(x)\|),$$

where $\theta = \arctan \frac{\left\|P^{(2:M)}_{\Phi(x),\Gamma}\widetilde{x}\right\|}{|\langle v_1(x),\widetilde{x}\rangle|}$, with $\widetilde{x} = \nabla^2 L(\Phi(x))(x-\Phi(x))$.

Proof of Lemma B.11. From Lemma B.10, we have that

$$\frac{\nabla L(x)}{\|\nabla L(x)\|} = \frac{\nabla^2 L(\Phi(x))(x-\Phi(x))}{\|\nabla^2 L(\Phi(x))(x-\Phi(x))\|} + \mathcal{O}(\frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}\|x-\Phi(x)\|).$$

Hence,

$$\begin{aligned}\frac{|\langle v_1(x), \nabla L(x)\rangle|}{\|\nabla L(x)\|} &= \frac{\left|\langle v_1(x), \nabla^2 L(\Phi(x))(x-\Phi(x))\rangle\right|}{\|\nabla^2 L(\Phi(x))(x-\Phi(x))\|} + \mathcal{O}(\frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}\|x-\Phi(x)\|)\\ &\geq \cos\theta - \mathcal{O}(\frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}\|x-\Phi(x)\|).\end{aligned}$$

□

Lemma B.12. *For any $\overline{xy} \in Y^{\boldsymbol{\epsilon}}$ where $y = x - \eta\frac{\nabla L(x)}{\|\nabla L(x)\|}$ is the one step Normalized GD update from x, we have*

$$\|\Phi(y) - \Phi(x)\| \leq \frac{1}{2}\boldsymbol{\xi}\eta^2.$$

Moreover, we must have for every $1 \leq k \leq M$,

$$\left|\lambda_k(\nabla^2 L(\Phi(x))) - \lambda_k(\nabla^2 L(\Phi(y)))\right| \leq \frac{1}{4}\boldsymbol{\nu}\boldsymbol{\xi}\eta^2,$$

and

$$\left\|v_1(\nabla^2 L(\Phi(x))) - v_1(\nabla^2 L(\Phi(y)))\right\| \leq \frac{1}{2}\frac{\boldsymbol{\nu}\boldsymbol{\xi}\eta^2}{\boldsymbol{\Delta} - \frac{1}{4}\boldsymbol{\nu}\boldsymbol{\xi}\eta^2} = \frac{\boldsymbol{\nu}\boldsymbol{\xi}\eta^2}{2\boldsymbol{\Delta}} + \mathcal{O}(\frac{\boldsymbol{\nu}^2\boldsymbol{\xi}^2\eta^4}{\boldsymbol{\Delta}}).$$

Proof. By Lemma B.15, we have $\partial\Phi(x)\nabla L(x) = 0$ for all $x \in U$. Thus we have

$$\begin{aligned}\|\Phi(y) - \Phi(x)\| =& \eta\left\|\int_{s=0}^{1} \partial\Phi\left(x - s\eta\frac{\nabla L(x)}{\|\nabla L(x)\|}\right)\frac{\nabla L(x)}{\|\nabla L(x)\|}\mathrm{d}s\right\|\\ =& \eta\left\|\int_{s=0}^{1}\left(\partial\Phi\left(x - s\eta\frac{\nabla L(x)}{\|\nabla L(x)\|}\right) - \partial\Phi(x)\right)\frac{\nabla L(x)}{\|\nabla L(x)\|}\mathrm{d}s\right\|\\ \leq& \eta\int_{s=0}^{1}\left\|\partial\Phi\left(x - s\eta\frac{\nabla L(x)}{\|\nabla L(x)\|}\right) - \partial\Phi(x)\right\|\mathrm{d}s\\ \leq& \eta^2\int_{s=0}^{1} s\sup_{s'\in[0,s]}\left\|\nabla^2\Phi((1-s')x + s'y)\right\|\mathrm{d}s\\ =& \frac{\eta^2}{2}\sup_{s'\in[0,1]}\left\|\nabla^2\Phi((1-s')x + s'y)\right\|\\ \leq& \frac{1}{2}\boldsymbol{\xi}\eta^2,\end{aligned}$$

where the final step follows from using Definition B.5.

For the second claim, we have for every $1 \le k \le M$,

$$\begin{aligned}
\left|\lambda_k(\nabla^2 L(\Phi(x))) - \lambda_k(\nabla^2 L(\Phi(y)))\right| &\le \left\|\nabla^2 L(\Phi(x)) - \nabla^2 L(\Phi(y))\right\| \\
&= \left\|\int_{s=0}^{1} (1-s)\partial^2(\nabla L)(\Phi(sx+(1-s)y))(\Phi(x)-\Phi(y))ds\right\| \\
&\le \left|\int_{s=0}^{1}(1-s)ds\right| \max_{s\in[0,1]} \left\|\partial^2(\nabla L)(\Phi(sx+(1-s)y))\right\| \|\Phi(x)-\Phi(y))\| \\
&\le \frac{1}{4}\boldsymbol{\nu\xi}\eta^2,
\end{aligned}$$

where the first step involves Theorem F.2.

The third claim follows from using Theorem F.4. Again,

$$\begin{aligned}
\left\|v_1(\nabla^2 L(\Phi(x))) - v_1(\nabla^2 L(\Phi(y)))\right\| &\le \frac{\left\|\nabla^2 L(\Phi(x)) - \nabla^2 L(\Phi(y))\right\|}{\lambda_1(\nabla^2 L\Phi(x)) - \lambda_2(\nabla^2 L(\Phi(y)))} \\
&\le \frac{1}{2}\frac{\boldsymbol{\nu\xi}\eta^2}{\lambda_1(\nabla^2 L(\Phi(x))) - \lambda_2(\nabla^2 L(\Phi(y))} \\
&\le \frac{1}{2}\frac{\boldsymbol{\nu\xi}\eta^2}{\lambda_1(\nabla^2 L(\Phi(x))) - \lambda_2(\nabla^2 L(\Phi(x)) - \frac{1}{4}\boldsymbol{\nu\xi}\eta^2} \\
&\le \frac{1}{2}\frac{\boldsymbol{\nu\xi}\eta^2}{\boldsymbol{\Delta} - \frac{1}{4}\boldsymbol{\nu\xi}\eta^2},
\end{aligned}$$

where we borrow the bound on $\left\|\nabla^2 L(\Phi(x)) - \nabla^2 L(\Phi(y))\right\|$ from our previous calculations. The final step follows from the constants defined in Definition B.5. □

Lemma B.13. *For any $\overline{xy} \in Y^{\boldsymbol{\epsilon}}$ where $y = x - \eta\frac{\nabla L(x)}{\|\nabla L(x)\|}$ is the one step Normalized GD update from x, we have that*

$$\begin{aligned}
\Phi(y) - \Phi(x) = &-\frac{\eta^2}{4} P^{\perp}_{\Phi(x),\Gamma}\nabla(\log\lambda_1(\nabla^2 L(\Phi(x)))) \\
&+ \mathcal{O}(\eta^2\boldsymbol{\xi}\theta) + \mathcal{O}(\frac{\boldsymbol{\nu\xi}\left\|x-\Phi(x)\right\|\eta^2}{\boldsymbol{\mu}}) + \mathcal{O}(\boldsymbol{\chi}\left\|x-\Phi(x)\right\|\eta^2) + \mathcal{O}(\boldsymbol{\chi}\eta^3).
\end{aligned}$$

Here $\theta = \arctan\frac{\left\|P^{(2:M)}_{\Phi(x),\Gamma}\widetilde{x}\right\|}{|\langle v_1(x),\widetilde{x}\rangle|}$, with $\widetilde{x} = \nabla^2 L(\Phi(x))(x-\Phi(x))$. Additionally, we have that

$$\left\|P_{\Phi(x),\Gamma}(\Phi(y)-\Phi(x))\right\| \le \mathcal{O}(\boldsymbol{\chi}\left\|x-\Phi(x)\right\|\eta^2) + \mathcal{O}(\boldsymbol{\chi}\eta^3) + \mathcal{O}(\frac{\boldsymbol{\nu\xi}}{\boldsymbol{\mu}}\left\|x-\Phi(x)\right\|\eta^2).$$

Proof of Lemma B.13. By Taylor expansion for Φ at x, we have

$$\begin{aligned}
\Phi(y) - \Phi(x) =&\partial\Phi(x)\,(y-x) + \frac{1}{2}\partial^2\Phi(x)[y-x, y-x] + \mathcal{O}(\boldsymbol{\chi}\left\|y-x\right\|^3) \\
=&\partial\Phi(x)\left(-\eta\frac{\nabla L(x)}{\|\nabla L(x)\|}\right) + \frac{\eta^2}{2}\partial^2\Phi(x)\left[\frac{\nabla L(x)}{\|\nabla L(x)\|}, \frac{\nabla L(x)}{\|\nabla L(x)\|}\right] + \mathcal{O}(\boldsymbol{\chi}\eta^3) \\
=&\frac{\eta^2}{2}\partial^2\Phi(x)\left[\frac{\nabla L(x)}{\|\nabla L(x)\|}, \frac{\nabla L(x)}{\|\nabla L(x)\|}\right] + \mathcal{O}(\boldsymbol{\chi}\eta^3),
\end{aligned}$$

where in the pre-final step, we used the property of Φ from Lemma B.15. In the final step, we have used a second order taylor expansion to bound the difference between $\partial^2\Phi(x)$ and $\partial^2\Phi(\Phi(x))$. Additionally, we have used $y - x = \eta\frac{\nabla L(x)}{\|\nabla L(x)\|}$ from the Normalized GD update rule.

Applying Taylor expansion on Φ again but at $\Phi(x)$, we have that

$$\Phi(y)-\Phi(x)=\frac{\eta^2}{2}\partial^2\Phi(\Phi(x))\left[\frac{\nabla L(x)}{\|\nabla L(x)\|},\frac{\nabla L(x)}{\|\nabla L(x)\|}\right]+\mathcal{O}(\boldsymbol{\chi}\|x-\Phi(x)\|\eta^2)+\mathcal{O}(\boldsymbol{\chi}\eta^3) \tag{20}$$

Also, at $\Phi(x)$, since $v_1(x)$ is the top eigenvector of the hessian $\nabla^2 L$, we have that from Corollary B.22,

$$\partial^2\Phi(\Phi(x))\left[v_1(x)v_1(x)^\top\right]=-\frac{1}{2\lambda_1(x)}\partial\Phi(\Phi(x))\partial^2(\nabla L)(\Phi(x))[v_1(x),v_1(x)]. \tag{21}$$

By Lemma B.11, it holds that

$$\left\|\operatorname{sign}\left(\left\langle\frac{\nabla L(x)}{\|\nabla L(x)\|},v_1(x)\right\rangle\right)\frac{\nabla L(x)}{\|\nabla L(x)\|}-v_1(x)\right\|\le 2\sin\frac{\theta}{2}+\mathcal{O}(\frac{\boldsymbol{\nu}\|x-\Phi(x)\|}{\boldsymbol{\mu}})\le\theta+\mathcal{O}(\frac{\boldsymbol{\nu}\|x-\Phi(x)\|}{\boldsymbol{\mu}}). \tag{22}$$

Plug Equations (21) and (22) into Equation (20), we have that

$$\begin{aligned}\Phi(y)-\Phi(x)=&-\frac{\eta^2}{2}\frac{1}{2\lambda_1(x)}\partial\Phi(\Phi(x))\partial^2(\nabla L)(\Phi(x))[v_1(x),v_1(x)]\\&+\mathcal{O}(\eta^2\boldsymbol{\xi}\theta)+\mathcal{O}(\frac{\boldsymbol{\nu\xi}\|x-\Phi(x)\|\eta^2}{\boldsymbol{\mu}})+\mathcal{O}(\boldsymbol{\chi}\|x-\Phi(x)\|\eta^2)+\mathcal{O}(\boldsymbol{\chi}\eta^3).\end{aligned}$$

By Lemma B.17, for any $x\in\Gamma$, $\partial\Phi(x)$ is the projection matrix onto the tangent space $\mathrm{T}_{\Phi(x)}\Gamma$. Thus, $\partial\Phi(\Phi(x))=P^{\perp}_{\Phi(x),\Gamma}$. Thus the proof of the first claim is completed by noting that $\partial\Phi(\Phi(x))\partial^2(\nabla L)(\Phi(x))[v_1(x),v_1(x)]=P^{\perp}_{\Phi(x),\Gamma}\nabla(\log\lambda_1(\nabla^2L(\Phi(x))))$ by Corollary B.23.

For the second claim, continuing from Equation (20), we have that

$$\begin{aligned}\Phi(y)-\Phi(x)=&\frac{\eta^2}{2}\partial^2\Phi(\Phi(x))\left[\frac{\nabla L(x)}{\|\nabla L(x)\|},\frac{\nabla L(x)}{\|\nabla L(x)\|}\right]+\mathcal{O}(\boldsymbol{\chi}\|x-\Phi(x)\|\eta^2)+\mathcal{O}(\boldsymbol{\chi}\eta^3)\\=&\frac{\eta^2}{2}\partial^2\Phi(\Phi(x))\left[\Sigma\right]+\mathcal{O}(\boldsymbol{\chi}\|x-\Phi(x)\|\eta^2)+\mathcal{O}(\boldsymbol{\chi}\eta^3)+\mathcal{O}(\frac{\boldsymbol{\nu\xi}}{\boldsymbol{\mu}}\|x-\Phi(x)\|\eta^2),\end{aligned}$$

where $\Sigma=P_{\Phi(x),\Gamma}\frac{\nabla L(x)}{\|\nabla L(x)\|}\left(P_{\Phi(x),\Gamma}\frac{\nabla L(x)}{\|\nabla L(\Phi(x))\|}\right)^\top$ and the last step is by Lemma B.10. Here $P_{\Phi(x),\Gamma}$ denotes the projection matrix of the subspace spanned by $v_1(x),\dots,v_M(x)$.

By Lemmas B.17, B.18 and B.21, we have that $P_{\Phi(x),\Gamma}\partial^2\Phi(\Phi(x))\left[\Sigma\right]=-P_{\Phi(x),\Gamma}\partial\Phi(x)\partial^2(\nabla L)(x)[\mathcal{L}^{-1}_{\nabla^2L(x)}\Sigma]=0$, we conclude that

$$\left\|P_{\Phi(x),\Gamma}(\Phi(y)-\Phi(x))\right\|\le\mathcal{O}(\boldsymbol{\chi}\|x-\Phi(x)\|\eta^2)+\mathcal{O}(\boldsymbol{\chi}\eta^3)+\mathcal{O}(\frac{\boldsymbol{\nu\xi}}{\boldsymbol{\mu}}\|x-\Phi(x)\|\eta^2),$$

which completes the proof. □

Lemma B.14. *Let $L_{min}=\min_{y\in U}L(y)$. For any $\overline{xy}\in Y^{\boldsymbol{\epsilon}}$ where $y=x-\eta\frac{\nabla L(x)}{\|\nabla L(x)\|}$ is the one step Normalized GD update from x, if $\|\nabla L(x_\eta(t))\|\ge\boldsymbol{\zeta}\eta$, we have that*

$$\sqrt{L(y)-L_{min}}\le\sqrt{L(x)-L_{min}}-\eta\frac{\sqrt{2\boldsymbol{\mu}}}{4}.$$

Proof of Lemma B.14. By Taylor expansion, we have that

$$L(y)\le L(x)-\eta\|\nabla L(x)\|+\frac{\boldsymbol{\zeta}\eta^2}{2}.$$

Thus for $\|\nabla L(x_\eta(t))\| \geq \boldsymbol{\zeta}\eta$, we have that

$$L(y) - L(x) \leq -\frac{\eta}{2}\|\nabla L(x)\| \leq -\eta\frac{\sqrt{2\boldsymbol{\mu}}}{2}\sqrt{L(x) - L_{min}} \leq 0,$$

where the last step is because L is μ-PL on $Y^{\boldsymbol{\epsilon}}$. In other words, we have that

$$\sqrt{L(y) - L_{min}} - \sqrt{L(x) - L_{min}} \leq -\eta\frac{\sqrt{L(x) - L_{min}}}{\sqrt{L(y) - L_{min}} + \sqrt{L(x) - L_{min}}}\frac{\sqrt{2\boldsymbol{\mu}}}{2} \leq -\eta\frac{\sqrt{2\boldsymbol{\mu}}}{4},$$

where in the last step we use $L(y) - L(x) \leq 0$. This completes the proof. □

B.2. Properties of limiting map of gradient flow, Φ

All the following Lemmas B.15 to B.19 and B.21 and Definition B.20 have been taken from (Li et al., 2022).

Lemma B.15. *For any $x \in U$, it holds that (1). $\partial\Phi(x)\nabla L(x) = 0$ and (2). $\partial^2\Phi(x)[\nabla L(x), \nabla L(x)] = -\partial\Phi(x)\nabla^2 L(x)\nabla L(x)$.*

Lemma B.16. *For any $x \in \Gamma$ and any $v \in \mathrm{T}_x\Gamma$, it holds that $\nabla^2 L(x)v = 0$.*

Lemma B.17. *For any $x \in \Gamma$, $\partial\Phi(x) \in \mathbb{R}^{D\times D}$ is the projection matrix onto the tangent space $\mathrm{T}_x\Gamma$, i.e. $\partial\Phi(x) = P^{\perp}_{x,\Gamma}$.*

Lemma B.18. *For any $x \in \Gamma$, if $v_1, \ldots, v_M$ denotes the non-zero eigenvectors of the hessian $\nabla^2 L(\Phi(x))$, then $v_1, \ldots, v_M \in \mathrm{N}_x\Gamma$.*

Lemma B.19. *For any $x \in \Gamma$ and $u \in \mathrm{N}_x\Gamma$, it holds that*

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

Definition B.20 (Lyapunov Operator)**.** For a symmetric matrix H, we define $W_H = \left\{\Sigma \in \mathbb{R}^{D\times D} \mid \Sigma = \Sigma^\top, HH^\dagger\Sigma = \Sigma = \Sigma HH^\dagger\right\}$ and Lyapunov Operator $\mathcal{L}_H : W_H \to W_H$ as $\mathcal{L}_H(\Sigma) = H^\top\Sigma + \Sigma H$. It's easy to verify $\mathcal{L}_H^{-1}$ is well-defined on W_H.

Lemma B.21. *For any $x \in \Gamma$ and $\Sigma = \mathrm{span}\{uu^\top \mid u \in \mathrm{N}_x\Gamma\}$,*

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

We will also use the following two corollaries of Lemma B.21.

Corollary B.22. *For any $x \in \Gamma$, if u denotes the top eigenvector of $\nabla^2 L(x)$, then*

$$\partial^2\Phi(x)[uu^\top] = -\frac{1}{2\lambda_1(\nabla^2 L(x))}\partial\Phi(x)\partial^2(\nabla L)(x)[u, u]$$

Corollary B.23. *For any $x \in \Gamma$ and any eigenvector u of $\nabla^2 L(x)$, then*

$$\partial^2\Phi(x)[uu^\top] = -\frac{1}{2}P^{\perp}_{x,\Gamma}\nabla\log(\lambda_1(\nabla^2 L(x))).$$

Proof of Corollary B.23. The proof follows from using Corollary B.22 and the derivative of λ_1 from Theorem F.1. □

C. Analysis of Normalized GD on General Loss Functions

C.1. Phase I, Convergence

We restate the theorem concerning Phase I for the Normalized GD algorithm. Recall the following notation for each $1 \leq j \leq M$:

$$R_j(x) := \sqrt{\sum_{i=j}^{M}\lambda_i^2(x)\langle v_i(x), x - \Phi(x)\rangle^2} - \lambda_j(x)\eta, \quad \text{for all } x \in U.$$

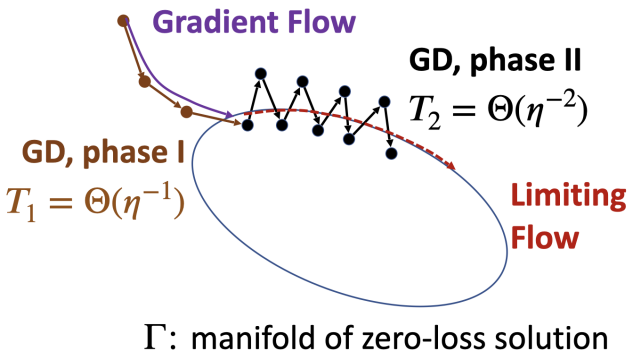

Figure 6: Illustration for two-phase dynamics of Normalized GD and GD on $\sqrt{L}$ on a 1D zero loss manifold Γ. For sufficiently small LR η, Phase I is close to Gradient Flow and lasts for $\Theta(\eta^{-1})$ steps, while Phase II is close to the limiting flow which decreases the sharpness of the loss and lasts for $\Theta(\eta^{-2})$ steps. GD iterate oscillates along the top eigenvector of the Hessian with the period equal to two steps. (cf. Figure 2 in (Li et al., 2022))

Theorem 4.3 (Phase I). *Let $\{x_\eta(t)\}_{t\in\mathbb{N}}$ be the iterates of Normalized GD (3) with LR η and $x_\eta(0) = x_{init} \in U$. There is $T_1 > 0$ such that for any $T_1' > T_1$, it holds that for sufficiently small η that (1) $\max_{T_1 \le \eta t \le T_1'} \|x_\eta(t) - \Phi(x_{init})\| \le O(\eta)$ and (2) $\max_{T_1 \le \eta t \le T_1', j\in[D]} R_j(x_\eta(t)) \le O(\eta^2)$.*

The intuition behind the above theorem is that for sufficiently small LR η, $x_\eta(t)$ will track the normalized gradient flow starting from x_{init}, which is a time-rescaled version of the standard gradient flow. Thus the normalized GF will enter Y^ϵ and so does normalized GD. Since L satisfies PL condition in Y^ϵ, the loss converges quickly and the iterate $x_\eta(t)$ gets η to manifold. To finish, we need the following theorem, which is the approximately-quadratic version of Lemma 3.3 when the iterate is $\mathcal{O}(\eta)$ close to the manifold.

Lemma C.1. *Suppose $\{x_\eta(t)\}_{t\ge 0}$ are iterates of Normalized GD (3) with a learning rate η and $x_\eta(0) = x_{init}$. There are constants $C > 0$, such that for any constant $\varsigma > 0$, if at some time t', $x_\eta(t') \in Y^\epsilon$ and satisfies $\frac{\|x_\eta(t') - \Phi(x_\eta(t'))\|}{\eta} \le \varsigma$, then for all $\bar{t} \ge t' + C\frac{\zeta\varsigma}{\mu}\log\frac{\varsigma\zeta}{\mu}$, the following must hold true for all $1 \le j \le M$:*

$$\sqrt{\sum_{i=j}^{M} \langle v_i(\bar{t}), \widetilde{x}(\bar{t})\rangle^2} \le \eta\lambda_j(\bar{t}) + \mathcal{O}(\nu\xi\eta^2) + \mathcal{O}(\frac{\nu\zeta^2\varsigma}{\mu^2}\eta^2) + \mathcal{O}(\sqrt{D}\xi\zeta\varsigma\nu\eta^2) + \mathcal{O}(\eta^2 D), \tag{23}$$

provided $\eta \le \mathcal{O}(\frac{\mu^3}{\zeta^3\varsigma^2\xi\nu\sqrt{D}})$ and that for all steps $t \in \{t, \dots, \bar{t}-1\}$, $\overline{x_\eta(t)x_\eta(t+1)} \subset Y^\epsilon$.

The proof of the above theorem is in Appendix D.1.

Proof of Theorem 4.3. We define the Normalized gradient flow as $\overline{\phi}(x,\tau) = x - \int_0^\tau \frac{\nabla L(\overline{\phi}(x,s))}{\|\nabla L(\overline{\phi}(x,s))\|}\mathrm{d}s$. Since $\overline{\phi}(x,\cdot)$ is only a time rescaling of $\phi(x,\cdot)$, they have the same limiting mapping, i.e., $\Phi(x) = \lim_{\tau\to T_x}\overline{\phi}(x,\tau)$, where T_x is the length of the trajectory of the gradient flow starting from x.

Let T_x be the length of the GF trajectory starting from x, and we know $\lim_{\tau\to T_x}\overline{\phi}(x,\tau) = \Phi(x)$, where $\overline{\phi}(x,\tau)$ is defined as the Normalized gradient flow starting from x. In Lemmas B.3 and B.6 we show there is a small neighbourhood around $\Phi(x_{\text{init}})$, Y^ϵ such that L is μ-PL in Y^ϵ. Thus we can take some time $T_0 < T_{x_{\text{init}}}$ such that $\overline{\phi}(x_{\text{init}}, T_0) \in Y^{\epsilon/2}$ and $L(\overline{\phi}(x_{\text{init}}), T_0) \le \frac{1}{2}L_{\text{critical}}$, where $L_{\text{critical}} := \frac{\epsilon^2\mu}{8}$. (Without loss of generality, we assume $\min_{y\in Y} L(y) = 0$) By standard ODE approximation theory, we know there is some small η_0, such that for all $\eta \le \eta_0$, $\left\|x_\eta(\lceil T_0/\eta\rceil) - \overline{\phi}(x_{\text{init}}, T_0)\right\| = \mathcal{O}(\eta)$, where $\mathcal{O}(\cdot)$ hides constants depending on the initialization x_{init} and the loss function L.

Without loss of generality, we can assume η_0 is small enough such that $x_\eta(\lceil T_0/\eta\rceil) \in Y^\epsilon$ and $L(x_\eta(\lceil T_0/\eta\rceil)) \le L_{\text{critical}}$. Now let t_η be the smallest integer (yet still larger than $\lceil T_0/\eta\rceil$) such that $\overline{x_\eta(t_\eta)x_\eta(t_\eta - 1)} \not\subset Y^\epsilon$ and we claim that there

is $t \in \{\lceil T_0/\eta \rceil, \dots, t_\eta\}$, $\|\nabla L(x_\eta(t))\| < \boldsymbol{\zeta}\eta$. By the definition of t_η, we know for any $t \in \{\lceil T_0/\eta \rceil + 1, \dots, t_\eta - 1\}$, by Lemma B.12 we have $\|\Phi(x_\eta(t)) - \Phi(x_\eta(t-1))\| \le \boldsymbol{\xi}\eta^2$, and by Lemma B.14, $\sqrt{L(x_\eta(t))} - \sqrt{x_\eta(t-1)} \le -\eta\frac{\sqrt{2\boldsymbol{\mu}}}{4}$ if $\|\nabla L(x_\eta(t))\| \ge \boldsymbol{\zeta}\eta$. If the claim is not true, since $\sqrt{L(x_\eta(t))}$ decreases $\eta\frac{\sqrt{2\boldsymbol{\mu}}}{4}$ per step, we have

$$0 \le \sqrt{L(x_\eta(t_\eta - 1))} \le \sqrt{L(x_\eta(\lceil T_0/\eta \rceil))} - (t_\eta - \lceil T_0/\eta \rceil - 1)\eta\frac{\sqrt{2\boldsymbol{\mu}}}{4},$$

which implies that $t_\eta - \lceil T_0/\eta \rceil - 1 \le \frac{\boldsymbol{\epsilon}}{\eta}$, and therefore by Lemma B.12,

$$\|\Phi(x_\eta(t_\eta - 1)) - \Phi(x_\eta(\lceil T_0/\eta \rceil))\| \le (t_\eta - \lceil T_0/\eta \rceil - 1)\frac{\boldsymbol{\xi}\eta^2}{2} = \frac{\boldsymbol{\xi}\eta\boldsymbol{\epsilon}}{2}$$

Thus we have

$$\|\Phi(x_\eta(t_\eta - 1)) - \Phi(x_{\text{init}})\| \le \|\Phi(x_\eta(t_\eta - 1)) - \Phi(x_\eta(\lceil T_0/\eta \rceil))\| + \left\|\Phi(x_\eta(\lceil T_0/\eta \rceil) - \Phi(\overline{\phi}(x_{\text{init}}, T_0)))\right\| = \mathcal{O}(\eta).$$

Meanwhile, by Lemma B.7, we have $\|\Phi(x_\eta(t_\eta - 1)) - x_\eta(t_\eta - 1)\| \le \sqrt{\frac{2L(x_\eta(t_\eta - 1))}{\boldsymbol{\mu}}} \le \sqrt{\frac{2L(x_\eta(\lceil T_0/\eta \rceil))}{\boldsymbol{\mu}}} = \frac{\boldsymbol{\epsilon}}{2}$. Thus for any $\kappa \in [0, 1]$, we have $\|\kappa x_\eta(t_\eta) + (1-\kappa)x_\eta(t_\eta - 1) - \Phi(x_{\text{init}})\|$ is upper bounded by

$$\kappa\|x_\eta(t_\eta) - x_\eta(t_\eta - 1)\| + \|\Phi(x_\eta(t_\eta - 1)) - x_\eta(t_\eta - 1)\| + \|\Phi(x_\eta(t_\eta - 1)) - \Phi(x_{\text{init}})\| = \kappa\eta + \frac{\boldsymbol{\epsilon}}{2} + \mathcal{O}(\eta),$$

which is smaller than $\boldsymbol{\epsilon}$ since we can set η_0 sufficiently small. In other words, $\overline{\Phi(x_\eta(t_\eta))\Phi(x_\eta(t_\eta - 1))} \subset Y^{\boldsymbol{\epsilon}}$, which contradicts with the definition of t_η. So far we have proved our claim that there is $t'_\eta \in \{\lceil T_0/\eta \rceil, \dots, t_\eta\}$, $\left\|\nabla L(x_\eta(t'_\eta))\right\| < \boldsymbol{\zeta}\eta$. Moreover, since $\sqrt{L(x_\eta(t))}$ decreases $\eta\frac{\sqrt{2\boldsymbol{\mu}}}{4}$ per step before t'_η, we know $t'_\eta - \lceil T_0/\eta \rceil \le \frac{\boldsymbol{\epsilon}}{\eta}$. By Lemma B.7, we know $\left\|x_\eta(t'_\eta) - \Phi(x_\eta(t'_\eta))\right\| \le \frac{\boldsymbol{\zeta}\eta}{\boldsymbol{\mu}}$.

Now we claim that for any T'_1, $t_\eta \ge \frac{T'_1}{\eta} + 1$ for sufficiently small threshold η_0 and $\eta \le \eta_0$. Below we prove this claim by contradiction. If the claim is not true, that is, $t_\eta < \frac{T'_1}{\eta} + 1$, if $t_\eta \le C\frac{\boldsymbol{\zeta}\boldsymbol{\varsigma}}{\boldsymbol{\mu}}\log\frac{\boldsymbol{\varsigma}\boldsymbol{\zeta}}{\boldsymbol{\mu}} + t'_\eta$ with $\boldsymbol{\varsigma} = \frac{\boldsymbol{\zeta}}{\boldsymbol{\mu}}$, we know $\|x_\eta(t_\eta) - \Phi(x_{\text{init}})\| \le \left\|x_\eta(t_\eta) - x_\eta(t'_\eta)\right\| + \left\|x_\eta(t'_\eta) - \Phi(x_\eta(t'_\eta))\right\| + \left\|\Phi(x_\eta(t'_\eta)) - \Phi(x_{\text{init}})\right\| = \mathcal{O}(\eta)$, which implies that $\overline{x_\eta(t_\eta)x_\eta(t_\eta - 1)} \in Y$. If $t_\eta \ge C\frac{\boldsymbol{\zeta}\boldsymbol{\varsigma}}{\boldsymbol{\mu}}\log\frac{\boldsymbol{\varsigma}\boldsymbol{\zeta}}{\boldsymbol{\mu}} + t'_\eta$, by Lemma C.1, we have $\|x_\eta(t_\eta) - \Phi(x_\eta(t_\eta))\| = \mathcal{O}(\eta)$. By Lemma B.12, we have $\|\Phi(x_\eta(t_\eta)) - \Phi(x_\eta(\lceil T_0/\eta \rceil))\| \le \mathcal{O}(\eta)$. Thus again $\|x_\eta(t_\eta) - \Phi(x_{\text{init}})\| \le \|x_\eta(t_\eta) - \Phi(x_\eta(t_\eta))\| + \|\Phi(x_\eta(t_\eta)) - \Phi(x_\eta(\lceil T_0/\eta \rceil))\| + \|\Phi(x_\eta(\lceil T_0/\eta \rceil)) - \Phi(x_{\text{init}})\| = \mathcal{O}(\eta)$, which implies that $\overline{x_\eta(t_\eta)x_\eta(t_\eta - 1)} \in Y$.

Thus for any T'_1, $t_\eta \ge \frac{T'_1}{\eta} + 1$ for sufficiently small threshold η_0 and $\eta \le \eta_0$. To complete the proof of Theorem 4.3, we pick T_1 to be any real number strictly larger than $\boldsymbol{\epsilon} + T_0$, as $\frac{T_1}{\eta} > C\frac{\boldsymbol{\zeta}\boldsymbol{\varsigma}}{\boldsymbol{\mu}}\log\frac{\boldsymbol{\varsigma}\boldsymbol{\zeta}}{\boldsymbol{\mu}} + \frac{\boldsymbol{\epsilon}}{\eta} + \lceil T_0/\eta \rceil \ge C\frac{\boldsymbol{\zeta}\boldsymbol{\varsigma}}{\boldsymbol{\mu}}\log\frac{\boldsymbol{\varsigma}\boldsymbol{\zeta}}{\boldsymbol{\mu}} + t'_\eta$ when η is sufficiently small with $\boldsymbol{\varsigma} = \frac{\boldsymbol{\zeta}}{\boldsymbol{\mu}}$. By Lemma C.1 the second claim of Theorem 4.3 is proved. Using the same argument again, we know $\forall \frac{T_1}{\eta} \le t \le \frac{T'_1}{\eta}$, it holds that $\|\Phi(x_\eta(t)) - \Phi(x_{\text{init}})\| \le \mathcal{O}(\eta)$.

□

C.2. Phase II, Limiting Flow

We first restate the main theorem that demonstrates that the trajectory implicitly minimizes sharpness.

Theorem 4.4 (Phase II). *Let $\{x_\eta(t)\}_{t\in\mathbb{N}}$ be the iterates of perturbed Normalized GD (Algorithm 1) with LR η. If the initialization $x_\eta(0)$ satisfy that*
(1) $\|x_\eta(0) - \Phi(x_{init})\| \le O(\eta)$,
(2) $\max_{j\in[D]} R_j(x_\eta(0)) \le O(\eta^2)$, and additionally
(3) $\min\{|\langle v_1(x_\eta(0)), x_\eta(0) - \Phi(x_\eta(0))\rangle|, -R_1(x_\eta(0))\} \ge \Omega(\eta)$, then for any time $T_2 \in T_2^{\text{fav}}$, it holds for sufficiently small η, with probability at least $1 - O(\eta^{10})$, that $\left\|\Phi(x_\eta(\lfloor T_2/\eta^2 \rfloor)) - X(T_2)\right\| = O(\eta)$ and $\frac{1}{\lfloor T_2/\eta^2 \rfloor}\sum_{t=0}^{\lfloor T_2/\eta^2 \rfloor}\theta_t \le \mathcal{O}(\eta)$.

To prove the above lemma, we first show the movement in the manifold for the discrete trajectory for Algorithm 1 with some learning rate η.

To simplify presentation in the upcoming lemmas, we will introduce few novel notations and their desired values/upper bounds.

Definition C.2. For ease of notation in the upcoming lemmas and proofs, we define the following notations and their desired values or upper bounds. Here, $\mathcal{O}$, Ω and Θ hide numerical constants.

$$
\begin{aligned}
t_{\text{escape}} &= \Theta(\frac{M\boldsymbol{\zeta}^2}{\boldsymbol{\mu}\boldsymbol{\Delta}}\log\frac{1}{\eta}) \\
c_{\text{escape}} &= \Theta(\frac{\boldsymbol{\beta}^6\boldsymbol{\mu}^6}{M\boldsymbol{\zeta}^6\boldsymbol{\nu}^3}) \\
\boldsymbol{\varrho} &= \Theta(\boldsymbol{\zeta}^2+\boldsymbol{\nu}) \\
\boldsymbol{\alpha} &= \Theta(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3\boldsymbol{\beta}}) \\
\boldsymbol{\Psi} &= \Theta\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\Delta}+\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\frac{\boldsymbol{\alpha}}{\boldsymbol{\beta}^2}\right) \\
r &= \eta^{100} \\
\boldsymbol{\Psi}_{\text{norm}} &= \Theta\left(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}+\sqrt{D}\boldsymbol{\xi}\boldsymbol{\zeta}\boldsymbol{\varsigma}\boldsymbol{\nu}+D\right) \\
\boldsymbol{\Psi}_{\text{G}} &= \Theta\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^3\boldsymbol{\xi}\boldsymbol{\nu}^2\boldsymbol{\chi}}{\boldsymbol{\mu}^5\boldsymbol{\beta}^3\Delta}\right).
\end{aligned}
$$

To recall, the limiting flow is given by

$$
X(\tau) = \Phi(x_{\text{init}}) - \frac{1}{4}\int_{s=0}^{\tau} P^{\perp}_{X(s),\Gamma}\nabla\log\lambda_1(X(s))\mathrm{d}s \tag{4}
$$

Let T_2 be the time up until which solution to the limiting flow exists.

Lemma B.13 shows the movement in Φ, which can be informally given as follows: in each step t,

$$
\Phi(x_\eta(t+1)) - \Phi(x_\eta(t)) = -\frac{\eta^2}{4}P^{\perp}_{\Phi(x_\eta(t)),\Gamma}\nabla\log\lambda_1(x_\eta(t)) + \mathcal{O}((\theta_t + \|x_\eta(t)-\Phi(x_\eta(t))\|)\eta^2), \tag{24}
$$

provided $\overline{\Phi(x_\eta(t))\Phi(x_\eta(t+1))} \in Y^{\boldsymbol{\epsilon}}$.

Motivated by this update step, we show that the trajectory of $\Phi(x_\eta(\cdot))$ is close to the limiting flow, for a small enough learning rate η. This isn't trivial, because even though the trajectories look similar, we introduce a noise in Equation (24) at each step, which can exponentially blow up over time. One of the major facts that helped us to bound the error between the two trajectories is that the error introduced in Equation (24) is at most $\mathcal{O}(\eta^2)$ at each step. Furthermore, the total error across the trajectory is given by $\sum_{t=0}^{t_2}\mathcal{O}(\eta^2\theta_t+\eta^3)$, which is of the order $\mathcal{O}(\eta)$ using the result from Lemma E.1.

Proof of Theorem 4.4. Without loss of generality, we can change assumption (3) in the theorem statement into $\|\widetilde{x}_\eta(0)\| \le \eta\lambda_1(0)/2 + \mathcal{O}(\boldsymbol{\Psi}_{\text{norm}}\eta^2)$ and $|\langle v_1(x_\eta(0)), x_\eta(0)-\Phi(x_\eta(0))\rangle| \ge \Omega(\eta)$. This is because we know from Lemma D.1, that the norm can't stay above $\frac{\lambda_1(\cdot)}{2}\eta + \Omega(\boldsymbol{\Psi}_{\text{norm}}\eta^2)$ for two consecutive steps. Moreover, if $|v_1(0), x_\eta(0)-\Phi(x_\eta(0))| \ge \Omega(\eta)$ but $\eta\lambda_1(0)/2 + \Omega(\boldsymbol{\Psi}_{\text{norm}}\eta^2) \le \|\widetilde{x}_\eta(0)\| \le \eta\lambda_1(0) - \Omega(\eta)$, we can further show that $|v_1(1), x_\eta(1)-\Phi(x_\eta(1))| \ge \Omega(\eta)$ from the update rule of Normalized GD (Lemma B.10). Thus, we can shift our analysis by one time-step if our assumption isn't true at step 0. This simplification of assumption helps us to prove the second claim using Lemma E.1.

We will follow an inductive analysis to prove two major claims. Suppose we denote $\mathrm{diff}(t_2)$ as the quantity $\left\|\Phi(x_\eta(t_2)) - X(t_2\eta^2)\right\|$ at any step t_2. At $t_2=0$, we have $\mathrm{diff}(0) = \|\Phi(x_\eta(0)) - X(0)\| = \|\Phi(x_\eta(0)) - \Phi(x_{\text{init}})\| \le \mathcal{O}(\eta)$, using the fact that we start from $x_\eta(0)$ which is $\mathcal{O}(\eta)$-close to $\Phi(x_{\text{init}})$ (i.e. $\|x_\eta(0)-\Phi(x_{\text{init}})\| \le \mathcal{O}(\eta)$), and is also $\mathcal{O}(\eta)$-close to the manifold (i.e. $\|x_\eta(0)-\Phi(x_\eta(0))\| \le \mathcal{O}(\eta)$).

Our inductive hypothesis is then for all step $1 \le t_2 \le \lfloor T_2/\eta^2 \rfloor$, the following holds true with probability at least $1-\eta^{10}$:

1. $\mathrm{diff}(t_2) \le (1+\boldsymbol{\beta}_{\mathrm{lip}}\eta^2)\mathrm{diff}(t_2-1) + \boldsymbol{\beta}_{\mathrm{lip}}\boldsymbol{\gamma}_{\mathrm{ub}}\eta^4 + \left(\mathcal{O}(\eta^2\boldsymbol{\xi}\theta_t) + \mathcal{O}(\frac{\boldsymbol{\zeta\nu\xi}\eta^3}{\boldsymbol{\mu}^2}) + \mathcal{O}(\frac{\boldsymbol{\chi\zeta}\eta^3}{\boldsymbol{\mu}}) + \mathcal{O}(\boldsymbol{\Upsilon}\eta^3)\right)$.

2. $\overline{x_\eta(t)x_\eta(t+1)} \in Y^{\boldsymbol{\epsilon}}$.

Here is our inductive argument, suppose the hypothesis is true till some step t_2. We can extend the hypothesis to step t_2+1 as follows:

1. First, we focus on $\mathrm{diff}(\cdot)$. Using Lemma B.13 to quantify the movement in $\Phi(\cdot)$, and Equation (4) to quantify the movement in $X(\cdot)$, we have

$$\begin{aligned}
\mathrm{diff}(t_2+1) &= \left\|\Phi(x_\eta(t_2+1)) - X((t_2+1)\eta^2)\right\| \\
&\le \left\|\Phi(x_\eta(t_2)) - X(t_2\eta^2)\right\| \\
&\quad + \left\| \frac{\eta^2}{4} P^{\perp}_{\Phi(x_\eta(t_2)),\Gamma}\nabla\log\lambda_1(x_\eta(t_2)) - \frac{1}{4}\int_{\tau=(t_2)\eta^2}^{(t_2+1)\eta^2} P^{\perp}_{X(\tau),\Gamma}\nabla\log\lambda_1(X(\tau))d\tau \right\| \\
&\quad + \mathcal{O}(\eta^2\boldsymbol{\xi}\theta_{t_2}) + \mathcal{O}(\frac{\boldsymbol{\nu\xi}\left\|x_\eta(t_2)-\Phi(x_\eta(t_2))\right\|\eta^2}{\boldsymbol{\mu}}) + \mathcal{O}(\boldsymbol{\chi}\left\|x_\eta(t_2)-\Phi(x_\eta(t_2))\right\|\eta^2) + \mathcal{O}(\boldsymbol{\chi}\eta^3) \\
&\le \left\|\Phi(x_\eta(t_2)) - X(t_2\eta^2)\right\| \\
&\quad + \left\| \frac{\eta^2}{4} P^{\perp}_{\Phi(x_\eta(t_2)),\Gamma}\nabla\log\lambda_1(x_\eta(t_2)) - \frac{1}{4}\int_{\tau=(t_2)\eta^2}^{(t_2+1)\eta^2} P^{\perp}_{X(\tau),\Gamma}\nabla\log\lambda_1(X(\tau))d\tau \right\| \\
&\quad + \left(\mathcal{O}(\eta^2\boldsymbol{\xi}\theta_{t_2}) + \mathcal{O}(\frac{\boldsymbol{\zeta\nu\xi}\eta^3}{\boldsymbol{\mu}^2}) + \mathcal{O}(\frac{\boldsymbol{\chi\zeta}\eta^3}{\boldsymbol{\mu}}) + \mathcal{O}(\boldsymbol{\Upsilon}\eta^3)\right). \qquad (25)
\end{aligned}$$

Under the assumption that we have started from a point that has $\max_{1\le j\le M} R_j(x_\eta(0)) \le \mathcal{O}(\eta^2)$, we have from Lemma C.1, that the iterate should satisfy the condition $\max_{1\le j\le M} R_j(x_\eta(t_2)) \le \mathcal{O}(\eta^2)$ at step t_2 as well. This helps us bound $\|x_\eta(t_2)-\Phi(x_\eta(t_2))\| \le \mathcal{O}(\boldsymbol{\zeta}\eta/\boldsymbol{\mu})$ in the second step.

We now focus on the second term in the R.H.S. of the above inequality. First of all, we can simplify $\int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} P^{\perp}_{X(\tau),\Gamma}\nabla\log\lambda_1(X(\tau))d\tau$ using

$$\begin{aligned}
&\left\| \int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} P^{\perp}_{X(\tau),\Gamma}\nabla\log\lambda_1(X(\tau))d\tau - \int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} P^{\perp}_{X(t_2\eta^2),\Gamma}\nabla\log\lambda_1(X(t_2\eta^2))d\tau \right\| \\
\le& \int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} \left\| P^{\perp}_{X(\tau),\Gamma}\nabla\log\lambda_1(X(\tau)) - P^{\perp}_{X(t_2\eta^2),\Gamma}\nabla\log\lambda_1(X(t_2\eta^2)) \right\| d\tau \\
\le& \int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} \boldsymbol{\beta}_{\mathrm{lip}} \left\| X(\tau) - X(t_2\eta^2) \right\| d\tau \\
=& \int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} \boldsymbol{\beta}_{\mathrm{lip}} \left\| \int_{s=t_2\eta^2}^{\tau} P^{\perp}_{X(s),\Gamma}\nabla\log\lambda_1(X(s)ds \right\| d\tau \\
\le& \int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} \int_{s=t_2\eta^2}^{\tau} \boldsymbol{\beta}_{\mathrm{lip}} \left\| P^{\perp}_{X(s),\Gamma}\nabla\log\lambda_1(X(s) \right\| dsd\tau \\
\le& \int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} \int_{s=t_2\eta^2}^{\tau} \boldsymbol{\beta}_{\mathrm{lip}}\boldsymbol{\gamma}_{\mathrm{ub}} dsd\tau \\
\le& \boldsymbol{\beta}_{\mathrm{lip}}\boldsymbol{\gamma}_{\mathrm{ub}}\eta^4.
\end{aligned}$$

Here, we use the upper bounds on the magnitude ($\boldsymbol{\gamma}_{\mathrm{ub}}$) and the lipschitz constant ($\boldsymbol{\beta}_{\mathrm{lip}}$) of the function $P^{\perp}_{\Phi(x),\Gamma}\nabla\log\lambda_1(x)$ in the domain $x \in Y^{\boldsymbol{\epsilon}}$ from Lemma F.5.

Moreover using the same strategy, we have

$$\begin{aligned}
&\left\|\frac{\eta^2}{4}P^{\perp}_{\Phi(x_\eta(t_2)),\Gamma}\nabla\log\lambda_1(x_\eta(t_2)) - \frac{1}{4}\int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} P^{\perp}_{X(t_2\eta^2),\Gamma}\nabla\log\lambda_1(X(t_2\eta^2))d\tau\right\| \\
=&\frac{\eta^2}{4}\left\|P^{\perp}_{\Phi(x_\eta(t_2)),\Gamma}\nabla\log\lambda_1(x_\eta(t_2)) - P^{\perp}_{X(t_2\eta^2),\Gamma}\nabla\log\lambda_1(X(t_2\eta^2))\right\| \\
=&\frac{\eta^2}{4}\left\|P^{\perp}_{\Phi(\Phi(x_\eta(t_2))),\Gamma}\nabla\log\lambda_1(\Phi(x_\eta(t_2))) - P^{\perp}_{X(t_2\eta^2),\Gamma}\nabla\log\lambda_1(X(t_2\eta^2))\right\| \\
\leq&\frac{\eta^2}{4}\boldsymbol{\beta}_{\text{lip}}\left\|\Phi(x_\eta(t_2)) - X(t_2\eta^2)\right\|.
\end{aligned}$$

In the pre-final step, we have used the fact that $\Phi(\Phi(x)) = \Phi(x)$ for any $x \in \Gamma$, which follows from the definition of Φ itself. Moreover, from the notations that we use, $\lambda_1(x) = \lambda_1(\Phi(x))$.

Thus, we have

$$\begin{aligned}
&\left\|\frac{\eta^2}{4}P^{\perp}_{\Phi(x_\eta(t_2)),\Gamma}\nabla\log\lambda_1(x_\eta(t_2)) - \frac{1}{4}\int_{\tau=(t_2)\eta^2}^{(t_2+1)\eta^2} P^{\perp}_{X(\tau),\Gamma}\nabla\log\lambda_1(X(\tau))d\tau\right\| \\
\leq&\left\|\int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} P^{\perp}_{X(\tau),\Gamma}\nabla\log\lambda_1(X(\tau))d\tau - \int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} P^{\perp}_{X(t_2\eta^2),\Gamma}\nabla\log\lambda_1(X(t_2\eta^2))d\tau\right\| \\
&+\frac{1}{4}\left\|\frac{\eta^2}{4}P^{\perp}_{\Phi(x_\eta(t_2)),\Gamma}\nabla\log\lambda_1(x_\eta(t_2)) - \frac{1}{4}\int_{\tau=t_2\eta^2}^{(t_2+1)\eta^2} P^{\perp}_{X(t_2\eta^2),\Gamma}\nabla\log\lambda_1(X(t_2\eta^2))d\tau\right\| \\
\leq&\frac{\eta^2}{4}\boldsymbol{\beta}_{\text{lip}}\left\|\Phi(x_\eta(t_2)) - X(t_2\eta^2)\right\| + \boldsymbol{\beta}_{\text{lip}}\boldsymbol{\gamma}_{\text{ub}}\eta^4.
\end{aligned}$$

Continuing from Equation (25), we have

$$\begin{aligned}
\text{diff}(t_2+1) \leq&(1+\frac{\eta^2}{4}\boldsymbol{\beta}_{\text{lip}})\left\|\Phi(x_\eta(t_2)) - X(t_2\eta^2)\right\| + \boldsymbol{\beta}_{\text{lip}}\boldsymbol{\gamma}_{\text{ub}}\eta^4 \\
&+\left(\mathcal{O}(\eta^2\boldsymbol{\xi}\theta_{t_2}) + \mathcal{O}(\frac{\boldsymbol{\zeta}\boldsymbol{\nu}\boldsymbol{\xi}\eta^3}{\boldsymbol{\mu}^2}) + \mathcal{O}(\frac{\boldsymbol{\chi}\boldsymbol{\zeta}\eta^3}{\boldsymbol{\mu}}) + \mathcal{O}(\boldsymbol{\Upsilon}\eta^3)\right) \\
=&(1+\frac{\eta^2}{4}\boldsymbol{\beta}_{\text{lip}})\text{diff}(t_2) + \boldsymbol{\beta}_{\text{lip}}\boldsymbol{\gamma}_{\text{ub}}\eta^4 \\
&+\left(\mathcal{O}(\eta^2\boldsymbol{\xi}\theta_{t_2}) + \mathcal{O}(\frac{\boldsymbol{\zeta}\boldsymbol{\nu}\boldsymbol{\xi}\eta^3}{\boldsymbol{\mu}^2}) + \mathcal{O}(\frac{\boldsymbol{\chi}\boldsymbol{\zeta}\eta^3}{\boldsymbol{\mu}}) + \mathcal{O}(\boldsymbol{\Upsilon}\eta^3)\right).
\end{aligned}$$

Thus, we have shown that we can extend the first hypothesis claim to step $t_2 + 1$.

2. To show the second hypothesis claim, note that we have assumed that both the claims hold true till time t_2. We can then bound $\text{diff}(\text{t}_2)$ as follows:

$$\begin{aligned}
\text{diff}(t_2) \leq&(1+\boldsymbol{\beta}_{\text{lip}}\eta^2)^{t_2}\text{diff}(0) \\
&+\sum_{t=0}^{t_2}(1+\boldsymbol{\beta}_{\text{lip}}\eta^2)^t\left(\boldsymbol{\beta}_{\text{lip}}\boldsymbol{\gamma}_{\text{ub}}\eta^4 + \left(\mathcal{O}(\eta^2\boldsymbol{\xi}\theta_t) + \mathcal{O}(\frac{\boldsymbol{\zeta}\boldsymbol{\nu}\boldsymbol{\xi}\eta^3}{\boldsymbol{\mu}^2}) + \mathcal{O}(\frac{\boldsymbol{\chi}\boldsymbol{\zeta}\eta^3}{\boldsymbol{\mu}}) + \mathcal{O}(\boldsymbol{\Upsilon}\eta^3)\right)\right) \\
\leq&(1+\boldsymbol{\beta}_{\text{lip}}\eta^2)^{t_2}\text{diff}(0) \\
&+(1+\boldsymbol{\beta}_{\text{lip}}\eta^2)^{t_2}\sum_{t=0}^{t_2}\left(\boldsymbol{\beta}_{\text{lip}}\boldsymbol{\gamma}_{\text{ub}}\eta^4 + \mathcal{O}(\eta^2\boldsymbol{\xi}\theta_t) + \mathcal{O}(\frac{\boldsymbol{\zeta}\boldsymbol{\nu}\boldsymbol{\xi}\eta^3}{\boldsymbol{\mu}^2}) + \mathcal{O}(\frac{\boldsymbol{\chi}\boldsymbol{\zeta}\eta^3}{\boldsymbol{\mu}}) + \mathcal{O}(\boldsymbol{\Upsilon}\eta^3)\right) \\
\leq&\mathcal{O}(e^{T_2\boldsymbol{\beta}_{\text{lip}}}\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}^3\boldsymbol{\nu}^2\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}T_2\eta),
\end{aligned}$$

In the final step, we have used the sum of the angles θ_t from Lemma E.1, which requires $\|x_\eta(t+1) - x_\eta(t)\| \in Y^\epsilon$ for all $1 \le t \le t_2 - 1$ and is true by the inductive hypothesis. To give a rough upper bound, we have also used $t_2 \le \frac{T_2}{\eta^2}$. Since, $\text{diff}(t_2) \le \mathcal{O}(\eta)$, we have for sufficiently small learning rate η, $\Phi(x_\eta(t_2)) \in Y^\epsilon$. Moreover, under the assumption that we have started from a point that has $\max_{1\le j\le M} R_j(x_\eta(0)) \le \mathcal{O}(\eta^2)$, we have from Lemma C.1, that the iterate should satisfy the condition $\max_{1\le j\le M} R_j(x_\eta(t_2)) \le \mathcal{O}(\eta^2)$ at step t_2 as well. This helps us bound $\|x_\eta(t_2) - \Phi(x_\eta(t_2))\| \le \mathcal{O}(\boldsymbol{\zeta}\eta/\boldsymbol{\mu})$. Furthermore, $\|x_\eta(t_2+1) - x_\eta(t_2)\| \le \mathcal{O}(\eta)$ from the update step of (perturbed) Normalized GD (Algorithm 1). Thus, we must have for sufficiently small learning rate η, $\|x_\eta(t_2+1) - x_\eta(t_2)\| \in Y^\epsilon$.

Thus, we have shown that our inductive hypothesis is true. The first claim follows from using $\text{diff}(\lfloor T_2/\eta^2 \rfloor)$. The second claim will follow from using Lemma E.1.

□

D. Phase I, OmittedProof of the main lemmas

D.1. Proof of Lemma C.1

Proof of Lemma C.1. The Normalized GD update at any step t can be written as (from Lemma B.10)

$$x_\eta(t+1) - x_\eta(t) = -\eta \frac{\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]}{\|\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\|} + \mathcal{O}(\frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}\eta \|x_\eta(t) - \Phi(x_\eta(t))\|). \tag{26}$$

Thus, using the notation $\widetilde{x} = \nabla^2 L(\Phi(x))(x - \Phi(x))$, we have

$$\begin{aligned}
\widetilde{x}_\eta(t+1) - \widetilde{x}_\eta(t) &= \widetilde{x}_\eta(t+1) - \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t+1) - \Phi(x_\eta(t))) \\
&\quad + \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t+1) - \Phi(x_\eta(t))) - \widetilde{x}_\eta(t) \\
&= \nabla^2 L(\Phi(x_\eta(t+1)))(\Phi(x_\eta(t)) - \Phi(x_\eta(t+1))) \\
&\quad + (\nabla^2 L(\Phi(x_\eta(t+1))) - \nabla^2 L(\Phi(x_\eta(t))))(x_\eta(t+1) - \Phi(x_\eta(t))) \\
&\quad + \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t+1) - x_\eta(t)) \\
&= \nabla^2 L(\Phi(x_\eta(t+1)))(\Phi(x_\eta(t)) - \Phi(x_\eta(t+1))) \\
&\quad + (\nabla^2 L(\Phi(x_\eta(t+1))) - \nabla^2 L(\Phi(x_\eta(t))))(x_\eta(t+1) - \Phi(x_\eta(t))) \\
&\quad + \nabla^2 L(\Phi(x_\eta(t)))\Big[- \eta \frac{\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]}{\|\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\|} \\
&\quad + \mathcal{O}(\frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}\eta \|x_\eta(t) - \Phi(x_\eta(t))\|)\Big] \\
&= -\eta \nabla^2 L(\Phi(x_\eta(t))) \frac{\widetilde{x}_\eta(t)}{\|\widetilde{x}_\eta(t)\|} + \text{err} + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}}{\boldsymbol{\mu}}\eta \|x_\eta(t) - \Phi(x_\eta(t))\|),
\end{aligned}$$

where err denotes $\nabla^2 L(\Phi(x_\eta(t+1)))(\Phi(x_\eta(t)) - \Phi(x_\eta(t+1))) + (\nabla^2 L(\Phi(x_\eta(t+1))) - \nabla^2 L(\Phi(x_\eta(t))))(x_\eta(t+1) - \Phi(x_\eta(t)))$. From Lemma B.12, we have $\|\Phi(x_\eta(t)) - \Phi(x_\eta(t+1))\| \le \mathcal{O}(\boldsymbol{\xi}\eta^2)$, which further implies, $\left\|\nabla^2 L(\Phi(x_\eta(t+1))) - \nabla^2 L(\Phi(x_\eta(t)))\right\| \le \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2)$. Hence,

$$\begin{aligned}
\|\text{err}\| &\le \left\|\nabla^2 L(\Phi(x_\eta(t+1)))\right\| \|\Phi(x_\eta(t)) - \Phi(x_\eta(t+1))\| \\
&\quad + \left\|\nabla^2 L(\Phi(x_\eta(t+1))) - \nabla^2 L(\Phi(x_\eta(t)))\right\| \|x_\eta(t+1) - \Phi(x_\eta(t))\| \\
&\le \left\|\nabla^2 L(\Phi(x_\eta(t+1)))\right\| \|\Phi(x_\eta(t)) - \Phi(x_\eta(t+1))\| \\
&\quad + \left\|\nabla^2 L(\Phi(x_\eta(t+1))) - \nabla^2 L(\Phi(x_\eta(t)))\right\| (\|x_\eta(t+1) - \Phi(x_\eta(t+1))\| + \|\Phi(x_\eta(t+1)) - \Phi(x_\eta(t))\|) \\
&\le \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2).
\end{aligned}$$

Hence,

$$\widetilde{x}_\eta(t+1) = \left(I - \eta \frac{\nabla^2 L(\Phi(x_\eta(t)))}{\|\widetilde{x}_\eta(t)\|}\right) \widetilde{x}_\eta(t) + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}}{\boldsymbol{\mu}}\eta \|x_\eta(t) - \Phi(x_\eta(t))\|). \tag{27}$$

Since $\|x_\eta(t) - \Phi(x_\eta(t))\| \le \mathcal{O}(\eta)$, the trajectory is similar to the trajectory in the qudratic model with an $\mathcal{O}(\eta^2)$ error, with the hessian fixed at $\nabla^2 L(\Phi(x_\eta(t)))$, and hence we can apply the same techniques from Corollary A.4 and Lemma A.1.

First, we consider the norm of the vector $\widetilde{x}_\eta(t)$ for $t' + 1 \le t \le \bar{t}$. We will show the following induction hypothesis:

$$\|\widetilde{x}_\eta(t)\| \le 1.01\eta\boldsymbol{\zeta}\boldsymbol{\varsigma}.$$

1. **Base case:** $(t = t')$. We have $\|\widetilde{x}_\eta(t')\| = \left\|\nabla^2 L(\Phi(x_\eta(t')))[x_\eta(t') - \Phi(x_\eta(t'))]\right\| \le \eta\lambda_1(t)\boldsymbol{\varsigma} \le \eta\boldsymbol{\zeta}\boldsymbol{\varsigma}$.
2. **Induction case:**$(t > t')$. Suppose the hypothesis holds true for $t-1$. Then,

$$\|x_\eta(t-1) - \Phi(x_\eta(t-1))\| \le \frac{1}{\lambda_M(t)} \|\widetilde{x}_\eta(t-1)\| \le \frac{1.01\eta\boldsymbol{\varsigma}\boldsymbol{\zeta}}{\boldsymbol{\mu}}.$$

We consider the following two cases:

(a) If $\|\widetilde{x}_\eta(t-1)\| \ge \eta\lambda_1(t)$. We can directly apply Corollary A.3 on (27) to show that

$$\begin{aligned}
\|\widetilde{x}_\eta(t)\| &\le \left(1 - \frac{\eta\lambda_M(t-1)}{2\|\widetilde{x}_\eta(t-1)\|}\right)\|\widetilde{x}_\eta(t-1)\| + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}}{\boldsymbol{\mu}}\eta\|x_\eta(t-1) - \Phi(x_\eta(t-1))\|) \\
&\le \|\widetilde{x}_\eta(t-1)\| - \frac{\eta\lambda_M(t-1)}{2} + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}}{\boldsymbol{\mu}^2}\boldsymbol{\varsigma}\eta^2) \\
&\le \|\widetilde{x}_\eta(t-1)\| - \frac{\eta\lambda_M(t-1)}{4},
\end{aligned}$$

where the final step follows if $\eta \le \mathcal{O}(\frac{\boldsymbol{\mu}^3}{\boldsymbol{\zeta}\boldsymbol{\varsigma}\boldsymbol{\nu}\boldsymbol{\xi}})$. Hence, $\|\widetilde{x}_\eta(t)\| < \|\widetilde{x}_\eta(t-1)\| \le \eta\boldsymbol{\zeta}\boldsymbol{\varsigma}$.

(b) If $\|\widetilde{x}_\eta(t-1)\| \le \eta\lambda_1(t)$. Then, we can directly apply Lemma A.1 on (27) to show that

$$\begin{aligned}
\|\widetilde{x}_\eta(t)\| &\le \eta\lambda_1(t) + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}}{\boldsymbol{\mu}}\eta\|x_\eta(t-1) - \Phi(x_\eta(t-1))\|) \\
&\le \eta\lambda_1(t) + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) \\
&\le 1.01\eta\lambda_1(t),
\end{aligned}$$

where the last step follows from using $\eta \le \mathcal{O}(\frac{\lambda_1(t)\boldsymbol{\mu}^2}{\boldsymbol{\nu}\boldsymbol{\xi}\boldsymbol{\zeta}\boldsymbol{\varsigma}})$.

Hence, we have shown that, $\|x_\eta(t) - \Phi(x_\eta(t))\| \le \frac{1}{\lambda_M(t)}\|\widetilde{x}_\eta(t)\| \le \frac{1.01\eta\boldsymbol{\varsigma}\boldsymbol{\zeta}}{\boldsymbol{\mu}}$ for all time $t' \le t \le \bar{t}$.

We complete the proof of Lemma C.1 with a similar argument as that for the quadratic model (see Corollary A.4 and Lemma A.1). The major difference from the quadratic model is that here the hessian changes over time, along with its eigenvectors and eigenvalues. Hence, we need to take care of the errors introduced in each step by the change of hessian.

We will divide the eigenvalues at time t' into groups such that eigenvalues in different groups are differed by at least η. *i.e.*, we divide $[M]$ into disjoint subsets $S_1, \cdots, S_p$ (with $1 \le p \le M$) such that for any $i, j \in [p]$ with $i \ne j$,

$$\min_{k \in S_i, \ell \in S_j} |\lambda_k(t') - \lambda_\ell(t')| \ge \eta.$$

From Lemma B.12, we have $\|\Phi(x_\eta(t)) - \Phi(x_\eta(t+1))\| \le \boldsymbol{\xi}\eta^2$, which further implies, $\left\|\nabla^2 L(\Phi(x_\eta(t+1))) - \nabla^2 L(\Phi(x_\eta(t)))\right\| \le \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2)$. That implies, using Theorem F.2, $|\lambda_j(t) - \lambda_j(t+1)| \le \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2)$ for any $j \in [M]$. Hence, after time t, we must have for any $i, j \in [p]$ with $i \ne j$,

$$\min_{k \in S_i, \ell \in S_j} |\lambda_k(t) - \lambda_\ell(t)| \ge \eta - \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2(t - t')).$$

Thus, for time $t' + 1 \le t \le t' + \mathcal{O}(\frac{\boldsymbol{\varsigma}\boldsymbol{\zeta}}{\boldsymbol{\mu}} \log \frac{\boldsymbol{\zeta}\boldsymbol{\varsigma}}{\boldsymbol{\mu}})$, we have

$$\min_{k \in S_i, \ell \in S_j} |\lambda_k(t) - \lambda_\ell(t)| \ge 0.99\eta,$$

provided $\eta \le \mathcal{O}(\frac{\boldsymbol{\mu}}{\boldsymbol{\zeta\varsigma\nu\xi}\log\frac{\boldsymbol{\varsigma\zeta}}{\boldsymbol{\mu}}})$.

For all $1 \le j \le M$, we consider the following two cases for any time $t'+1 \le t \le t' + \mathcal{O}(\frac{\boldsymbol{\varsigma\zeta}}{\boldsymbol{\mu}}\log\frac{\boldsymbol{\zeta\varsigma}}{\boldsymbol{\mu}})$:

1. If $\sqrt{\sum_{i=j}^{M}\langle v_i(t), \widetilde{x}_\eta(t)\rangle^2} > \eta\lambda_j(t)$, then we can apply Corollary A.4 on (27) to show that

$$\begin{aligned}
&\sqrt{\sum_{i=j}^{M}\langle v_i(t), \widetilde{x}_\eta(t+1)\rangle^2}\\
&\le \left(1 - \frac{\eta\lambda_M(t)}{2\left\|\widetilde{x}_\eta(t)\right\|}\right)\sqrt{\sum_{i=j}^{M}\langle v_i(t), \widetilde{x}_\eta(t)\rangle^2} + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}}\eta\left\|x_\eta(t) - \Phi(x_\eta(t))\right\|)\\
&\le \left(1 - \frac{\lambda_M(t)}{2\boldsymbol{\zeta\varsigma}}\right)\sqrt{\sum_{i=j}^{M}\langle v_i(t), \widetilde{x}_\eta(t)\rangle^2} + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}}\eta\left\|x_\eta(t) - \Phi(x_\eta(t))\right\|)\\
&\le \left(1 - \frac{\lambda_M(t)}{2\boldsymbol{\zeta\varsigma}}\right)\sqrt{\sum_{i=j}^{M}\langle v_i(t), \widetilde{x}_\eta(t)\rangle^2} + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2).
\end{aligned}$$

2. If $\sqrt{\sum_{i=j}^{M}\langle v_i(t), \widetilde{x}_\eta(t)\rangle^2} \le \eta\lambda_j(t)$, then we can apply Lemma A.1 on (27) to show that

$$\begin{aligned}
\sqrt{\sum_{i=j}^{M}\langle v_i(t), \widetilde{x}_\eta(t+1)\rangle^2} &\le \eta\lambda_j(t) + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}}\eta\left\|x_\eta(t) - \Phi(x_\eta(t))\right\|)\\
&\le \eta\lambda_j(t) + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta\varsigma\zeta}}{\boldsymbol{\mu}^2}\eta^2).
\end{aligned}$$

Denote by $P_{S_i}^{(t)}$ the projection matrix at time t onto the subspace spanned by $\{v_k(t)\}_{k\in S_i}$. Reconciling with the eigen subspaces, we have for any $j \in [p]$,

1. If $\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t)\right\|^2} > \eta\max_{k\in S_j}\lambda_k(t)$, then

$$\begin{aligned}
&\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t+1)\right\|^2}\\
&\le \left(1 - \frac{\lambda_M(t)}{2\boldsymbol{\zeta\varsigma}}\right)\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t)\right\|^2} + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2).
\end{aligned}$$

2. $\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t)\right\|^2} \le \eta\max_{k\in S_j}\lambda_k(t)$,

$$\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t+1)\right\|^2} \le \eta\max_{k\in S_j}\lambda_k + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta\varsigma\zeta}}{\boldsymbol{\mu}^2}\eta^2).$$

From Lemma B.12, we have $\left\|\Phi(x_\eta(t)) - \Phi(x_\eta(t+1))\right\| \le \mathcal{O}(\boldsymbol{\xi}\eta^2)$, which further implies,

$$\left\|\nabla^2 L(\Phi(x_\eta(t+1))) - \nabla^2 L(\Phi(x_\eta(t)))\right\| \le \mathcal{O}(\boldsymbol{\nu\xi}\eta^2).$$

That implies, $|\lambda_j(t) - \lambda_j(t+1)| \leq \mathcal{O}(\boldsymbol{\nu\xi}\eta^2)$ for any $j \in [M]$. Furthermore, we can use Theorem F.4 to have for any $i \in [p]$ $\left\|P_{S_i}^{(t)} - P_{S_i}^{(t+1)}\right\| \leq \mathcal{O}(\boldsymbol{\nu\xi}\eta)$, since we have created the eigen subspaces such that the eigenvalue gap between any two distinct eigen subspaces is at least 0.99η in the desired interval.

Reconciling the additional error terms due to the movement in the hessian, we have

1. For any subspace $\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t)\right\|^2} > \eta\max_{k\in S_j}\lambda_k(t)$, we have

$$\begin{aligned}
&\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t+1)}\widetilde{x}_\eta(t+1)\right\|^2}\\
&\leq \left(1-\frac{\lambda_M(t)}{2\boldsymbol{\zeta\varsigma}}\right)\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t)\right\|^2} + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi\zeta\varsigma\nu}\eta^2)\\
&\leq \left(1-\frac{\lambda_M(t)}{4\boldsymbol{\zeta\varsigma}}\right)\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t)\right\|^2},
\end{aligned}$$

   where the final step follows if $\eta \leq \mathcal{O}(\frac{\boldsymbol{\mu}^3}{\boldsymbol{\zeta}^3\boldsymbol{\varsigma}^2\boldsymbol{\xi\nu}\sqrt{D}})$.

2. If $\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t)\right\|^2} \leq \eta\max_{k\in S_j}\lambda_k(t)$, we have

$$\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t+1)}\widetilde{x}_\eta(t+1)\right\|^2} \leq \eta\max_{k\in S_j}\lambda_k + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi\zeta\varsigma\nu}\eta^2).$$

Hence, if $\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t)\right\|^2} > \eta\max_{k\in S_j}\lambda_k$, its value drops by a factor $\left(1-\frac{\lambda_M(t)}{4\boldsymbol{\zeta\varsigma}}\right)$. And if it is already below $\eta\max_{k\in S_j}\lambda_k$, it doesn't go $\mathcal{O}(\eta^2)$ beyond $\eta\max_{k\in S_j}\lambda_k$.

Since, at any time t, any two eigenvalues that belong to the same group can't be farther than ηD, we have: if for all $j \in [p]$,

$$\sqrt{\sum_{i=j}^{p}\left\|P_{S_i}^{(t)}\widetilde{x}_\eta(t)\right\|^2} \leq \eta\max_{k\in S_j}\lambda_k + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi\zeta\varsigma\nu}\eta^2),$$

then we must have for all $j \in [M]$,

$$\sqrt{\sum_{i=j}^{M}|\langle v_i(t), \widetilde{x}_\eta(t)\rangle|^2} \leq \eta\lambda_j(t) + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi\zeta\varsigma\nu}\eta^2) + \mathcal{O}(\eta^2 D).$$

Thus, at $\bar{t} = t' + \mathcal{O}(\frac{\boldsymbol{\zeta\varsigma}}{\boldsymbol{\mu}}\log\frac{\boldsymbol{\varsigma\zeta}}{\boldsymbol{\mu}})$, we must have for all $1 \leq j \leq M$,

$$\sqrt{\sum_{i=j}^{M}\langle v_i(\bar{t}), \widetilde{x}_\eta(\bar{t})\rangle^2} \leq \eta\lambda_j(\bar{t}) + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi\zeta\varsigma\nu}\eta^2) + \mathcal{O}(\eta^2 D).$$

To finish the argument, we need to show that the above condition continues to hold true for any $\bar{t} \geq t' + \mathcal{O}(\frac{\boldsymbol{\zeta\varsigma}}{\boldsymbol{\mu}}\log\frac{\boldsymbol{\varsigma\zeta}}{\boldsymbol{\mu}})$. We give a proof sketch here (we aren't rigorous here, since the argument is very much the same). Suppose the hypothesis is true at some time $\bar{t}$. We can repeat the above argument inductively at steps $\bar{t}$ to get the condition at step $\bar{t}+1$. First, we define

the groups $S_1(\bar{t}), \cdots, S_p(\bar{t})$ (for some $1 \leq p \leq M$) on the basis of the eigenvalues at the current step $\bar{t}$. Then, we show that for any $j \in [p]$,

$$\sqrt{\sum_{i=j}^{p} \left\| P_{S_i(\bar{t})} \widetilde{x}_\eta(\bar{t}+1) \right\|^2} \leq \eta \max_{k \in S_j(\bar{t})} \lambda_k(\bar{t}) + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi}\boldsymbol{\zeta}\boldsymbol{\varsigma}\boldsymbol{\nu}\eta^2).$$

Taking, the movement in hessian into account, we have for any $j \in [p]$,

$$\sqrt{\sum_{i=j}^{p} \left\| P_{S_i(\bar{t})} \widetilde{x}_\eta(\bar{t}+1) \right\|^2} \leq \eta \max_{k \in S_j(\bar{t})} \lambda_k(\bar{t}+1) + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi}\boldsymbol{\zeta}\boldsymbol{\varsigma}\boldsymbol{\nu}\eta^2).$$

Finally, we can get back to the projection on the eigenvectors by using our construction that any two eigenvalues that belong to the same group $S_j(\bar{t})$ can't be farther than ηD: for any $j \in [M]$,

$$\sqrt{\sum_{i=j}^{M} \langle v_i(\bar{t}+1), \widetilde{x}_\eta(\bar{t}+1)\rangle^2} \leq \eta\lambda_j(\bar{t}+1) + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi}\boldsymbol{\zeta}\boldsymbol{\varsigma}\boldsymbol{\nu}\eta^2) + \mathcal{O}(\eta^2 D).$$

□

D.2. Some interesting properties of the condition in Equation (23)

Thus, we can claim that after the initial phase, the following condition will continue to hold true for all $1 \leq j \leq M$:

$$\sqrt{\sum_{i=j}^{M} \langle v_i(t), \widetilde{x}_\eta(t)\rangle^2} \leq \lambda_j(t)\eta + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi}\boldsymbol{\zeta}\boldsymbol{\varsigma}\boldsymbol{\nu}\eta^2) + \mathcal{O}(\eta^2 D), \tag{28}$$

where $\widetilde{x}_\eta(t) = \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))$. We will call the above condition as the alignment condition from now onwards. If $\eta \leq \mathcal{O}(\frac{\boldsymbol{\mu}^3}{\boldsymbol{\nu}\boldsymbol{\xi}\boldsymbol{\varsigma}\boldsymbol{\zeta}^2 D})$, we can have

$$\sqrt{\sum_{i=j}^{M} \langle v_i(t), \widetilde{x}_\eta(t)\rangle^2} \leq 2\lambda_j(t)\eta,$$

for all $1 \leq j \leq M$. We will be using this bound, when the error bound can be allowed to stay loose.

From the alignment condition (28), we can derive the following property that continues to hold true throughout the trajectory, once the condition is satisfied:

Lemma D.1. *If at time t, $x_\eta(t) \in Y^{\boldsymbol{\epsilon}}$, and the condition (28) holds true, then if $\|\widetilde{x}_\eta(t)\| > \frac{\eta\lambda_1(t)}{2}$, we have:*

$$\|\widetilde{x}_\eta(t+1)\| \leq \frac{\eta\lambda_1(t)}{2} + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi}\boldsymbol{\zeta}\boldsymbol{\varsigma}\boldsymbol{\nu}\eta^2) + \mathcal{O}(\eta^2 D),$$

provided $\overline{x_\eta(t)x_\eta(t+1)} \in Y^{\boldsymbol{\epsilon}}$.

We will continue to denote the error term by $\boldsymbol{\Psi}_{\text{norm}}\eta^2$, where

$$\boldsymbol{\Psi}_{\text{norm}} = \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2} + \sqrt{D}\boldsymbol{\xi}\boldsymbol{\zeta}\boldsymbol{\varsigma}\boldsymbol{\nu} + D).$$

The proof follows from applying Lemma A.8 using the alignment condition Equation (28). Hence, the iterate $\widetilde{x}_\eta(t)$ can't stay at norm larger than $0.5\eta\lambda_1(t) + \boldsymbol{\Psi}_{\text{norm}}\eta^2$ for time larger than 1. Thus, we will state the remaining lemmas for time t, s.t. $\|\widetilde{x}_\eta(t)\| \leq 0.5\eta\lambda_1(t) + \boldsymbol{\Psi}_{\text{norm}}\eta^2$.

Another useful lemma is to show that the magnitude along the top eigenvector increases when $\|\widetilde{x}_\eta(t)\| \leq \frac{\eta\lambda_1(t)}{2} + \mathcal{O}(\boldsymbol{\Psi}_{\text{norm}}\eta^2)$.

Algorithm 2 Grouping into 1-cycles and 2-cycles

Input: Interval $(\widetilde{t}, \bar{t})$, Iterates of Algorithm 1 in the interval: $\{\widetilde{x}_\eta(t)\}_{t\in(\widetilde{t},\bar{t})}$, Top eigenvalue of the hessian $\nabla^2 L(\Phi(x_\eta(t)))$ for all t in the interval.
Requires: $\left\|\widetilde{x}_\eta(\widetilde{t})\right\| \le 0.5\lambda_1(\widetilde{t})\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$.
Initialize: $N_0, N_1, N_2 \leftarrow \emptyset$, $t \leftarrow \widetilde{t}$.
while $t \le \bar{t}$ **do**
  if $\|\widetilde{x}_\eta(t+2)\| > 0.5\lambda_1(t+2)\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$ **then**
    $N_0 \leftarrow N_0 \cup \{t\}$
    $t \leftarrow t+1$
  else if $\|\widetilde{x}_\eta(t+2)\| \le 0.5\lambda_1(t+2)\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$ **then**
    $N_1 \leftarrow N_1 \cup \{t\}$
    $N_2 \leftarrow N_2 \cup \{t+1\}$
    $t \leftarrow t+1$
  end if
end while
Return: N_0, N_1, N_2

Lemma D.2. *For any constant $\boldsymbol{\beta} > 0$, consider any time t such that $x_\eta(t) \in Y^{\boldsymbol{\epsilon}}$, $\|\widetilde{x}_\eta(t)\| \le \frac{\eta\lambda_1(t)}{2} + \mathbf{\Psi}_{\text{norm}}\eta^2$, $\left|v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\right| \ge \boldsymbol{\beta}\eta$, and the alignment condition (Equation (28)) holds true, then:*

$$|v_1(t+1), \widetilde{x}_\eta(t+1)| \ge (1 + \mathcal{O}(\mathbf{\Psi}_{\text{norm}}\eta))\,|v_1(t), \widetilde{x}_\eta(t)| + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}^2}\eta^2),$$

Moreover,

$$\theta_{t+1} \le \left(1 - \frac{\boldsymbol{\Delta\mu}}{\boldsymbol{\zeta}} + \mathcal{O}(\frac{\mathbf{\Psi}_{\text{norm}}}{\boldsymbol{\beta}}\eta) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}^2\boldsymbol{\beta}}\eta)\right)\theta_t + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}^2\boldsymbol{\beta}}\eta).$$

For both the results to hold true, we must have $\overline{x_\eta(t)x_\eta(t+1)} \in Y^{\boldsymbol{\epsilon}}$.

The proof follows from using the noisy quadratic update rule for Normalized GD from Equation (19) (from Lemma B.10) and using the result for the increase in the projection along the top eigenvector for a quadratic model from Lemma A.5. For the second claim, we use the result of the drop in angle for a quadratic model from Lemma A.11.

Corollary D.3. *If for $2 \le k \le M$, $\|\widetilde{x}_\eta(t)\| \le \frac{\eta\lambda_k(t)}{2} + \mathbf{\Psi}_{\text{norm}}\eta^2$ and alignment condition (Equation (28)) holds true, the following must hold true:*

$$\left|v_k(t+1)^\top \widetilde{x}_\eta(t+1)\right| \ge (1 + \mathcal{O}(\mathbf{\Psi}_{\text{norm}}\eta))\left|v_k(t)^\top \widetilde{x}_\eta(t)\right| + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}^2}\eta^2).$$

The proof follows from using the noisy quadratic update for Normalized GD in Lemma B.10 (Equation (19)) and the behavior in a quadratic model along the non-top eigenvectors in Lemma A.5.

E. Phase II, Proof of the main lemmas

E.1. Average of angle across time (Phase II)

In Phase II, we start from a point $x_\eta(0)$, such that (1) $\|x_\eta(0) - \Phi(x_{\text{init}})\| \le O(\eta)$, (2) $\max_{j\in[D]} R_j(x_\eta(t)) \le O(\eta^2)$, and additionally (3) $|\langle v_1(x_\eta(0)), x_\eta(0) - \Phi(x_\eta(0))\rangle| = \Omega(\eta)$.

Formally, recall our notation on θ_t as $\theta_t = \arctan \frac{\left\|P^{(2:M)}_{t,\Gamma}\widetilde{x}_\eta(t)\right\|}{|\langle v_1(t), \widetilde{x}_\eta(t)\rangle|}$, with our notation of $\widetilde{x}_\eta(t)$ as $\nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))$. Moreover, recall the definition of the function $g_t : \mathbb{R} \to \mathbb{R}$ as

$$g_t(\lambda) = \frac{1}{2}\left(1 - \sqrt{1 - 2\frac{\lambda}{\lambda_1(t)}\left(1 - \frac{\lambda}{\lambda_1(t)}\right)}\right).$$

The condition (Equation (28)) that was shown to hold true in Phase II is:

$$\sqrt{\sum_{i=j}^{M}\langle v_i(t), \widetilde{x}_\eta(t)\rangle^2} \leq \lambda_j(t)\eta + \mathcal{O}(\boldsymbol{\nu\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2) + \mathcal{O}(\sqrt{D}\boldsymbol{\xi\zeta\varsigma\nu}\eta^2) + \mathcal{O}(\eta^2 D).$$

Further, we had proved in Lemma D.1 that if $\|\widetilde{x}_\eta(t)\| > \frac{\eta\lambda_1(t)}{2}$, the following must hold true:

$$\|\widetilde{x}_\eta(t+1)\| \leq \frac{\eta\lambda_1(t)}{2} + \boldsymbol{\Psi}_{\text{norm}}\eta^2.$$

Thus, the iterate can't stay greater than $\frac{\eta\lambda_1(t)}{2} + \Omega(\eta^2)$ for more than 1 timestep. We will heavily use this property of the iterate in this section.

In the first lemma, we show that the sum of the angles across the entire trajectory in any interval $[0, t_2]$ with $t_2 = \Omega(1/\eta^2)$, is at most $\mathcal{O}(\eta t_2)$.

Lemma E.1 (Sum of the angles)**.** *For any $T_2 > 0$ for which solution of Equation (4) exists, consider an interval $[0, t_2]$, with $\Omega(1/\eta^2) \leq t_2 \leq \lfloor T_2/\eta^2 \rfloor$. Suppose Algorithm 1 is run with learning rate η for t_2 steps, starting from a point $x_\eta(0)$ that satisfies (1) $\max_{j\in[D]} R_j(x_\eta(0)) \leq O(\eta^2)$, and (2) $|\langle v_1(0), x_\eta(0) - \Phi(x_\eta(0))\rangle| \geq \boldsymbol{\beta}\eta$, $\|\widetilde{x}_\eta(0)\| \leq \frac{\eta\lambda_1(0)}{2} + \boldsymbol{\Psi}_{\text{norm}}\eta^2$ for some constant $0 < \boldsymbol{\beta} \leq \frac{\boldsymbol{\mu\Delta}}{8\boldsymbol{\zeta}^2}$ independent of η. The following holds true with probability at least $1 - \eta^{10}$:*

$$\sum_{\ell=0}^{t_2}\theta_\ell \leq \mathcal{O}\left(\frac{\boldsymbol{\Upsilon\zeta}^2\boldsymbol{\xi\nu\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}t_2\eta\right),$$

provided η has been set sufficiently small, and for all time $0 \leq t \leq t_2 - 1$, $\overline{x_\eta(t)x_\eta(t+1)} \subset Y^{\boldsymbol{\epsilon}}$.

Proof. **Split analysis into blocks:** We split the analysis of the entire trajectory along $[0, t_2)$ into different blocks in the following inductive way. We use $0 = t^{(1)} < t^{(2)} < t^{(3)} < \cdots < t^{(k)} = t_2$ to denote the starting points of each of these blocks. The definition of $t^{(d)}$ depends on $t^{(d-1)}$ for all $d > 0$.

For each of the blocks $[t^{(d-1)}, t^{(d)})$, with $d > 0$, we will show the following results:

1. The average angle inside the block is $\mathcal{O}(\eta)$.
2. If $G_{t^{(d-1)}} \geq \boldsymbol{\beta}\eta$, then $G_{t^{(d)}} \geq \boldsymbol{\beta}\eta$.

We will show the analysis for a general block $[t^{(d-1)}, t^{(d)})$ for some $d > 0$. We define $t^{(d)}$ from $t^{(d-1)}$ as follows: at $t^{(d-1)}$, we can divide $[M] \setminus \{1\}$ into disjoint subsets $S_1, \cdots, S_p$ (with $1 \leq p \leq M$) such that for any $i, j \in [p]$ with $i \neq j$,

$$\min_{k\in S_i, \ell\in S_j}\Big|\Big|0.5\lambda_1(t^{(d-1)}) - \lambda_k(t^{(d-1)})\Big| - \Big|0.5\lambda_1(t^{(d-1)}) - \lambda_\ell(t^{(d-1)})\Big|\Big| \geq 10^{-3}\lambda_1(t^{(d-1)})$$

Then, we define

$$t^{(d)} = \min_{t > t^{(d-1)}}\Big\{t \mid \min_{i,j\in[p]}\min_{k\in S_i,\ell\in S_j}||0.5\lambda_1(t) - \lambda_k(t)| - |0.5\lambda_1(t) - \lambda_\ell(t)|| \leq \frac{1}{2}\times 10^{-3}\lambda_1(t),$$
$$\|\widetilde{x}_\eta(t)\| \leq 0.5\lambda_1(t)\eta + \boldsymbol{\Psi}_{\text{norm}}\eta^2\Big\}.$$

Moreover, we have the following two properties from the above definition of $S_1, \cdots, S_p$:

1. We must have from the definition of g, for any $i, j \in [p]$, $\min_{k\in S_i,\ell\in S_j}\left|g_{t^{(d-1)}}(\lambda_k(t^{(d-1)})) - g_{t^{(d-1)}}(\lambda_\ell(t^{(d-1)}))\right| \geq \frac{2}{3}\times 10^{-3}$. Thus, we sort them as follows: for any $i, j \in [p]$, $\min_{\ell\in S_i} g_{t^{(d-1)}}(\lambda_\ell(t^{(d-1)})) > \max_{\ell\in S_j} g_{t^{(d-1)}}(\lambda_\ell(t^{(d-1)}))$, if $i > j$.

2. Each eigenvalue changes by at most $\frac{1}{2}\boldsymbol{\nu}\boldsymbol{\xi}\eta^2$ in each step using Lemma B.12. Then, we have $t^{(d)} \geq t^{(d-1)} + \Omega(\frac{1}{\boldsymbol{\zeta}\boldsymbol{\xi}\boldsymbol{\nu}\eta^2})$. Moreover, the order among $S_1, \cdots, S_p$ w.r.t. the function $g_\cdot$ remains the same, i.e. for any $i, j \in [p]$, $\min_{\ell \in S_i} g_t(\lambda_\ell(t)) > \max_{\ell \in S_j} g_t(\lambda_\ell(t))$, if $i > j$.

At each step t, we first define different strips as $I_k(t) = [(1 - \frac{1}{100M}) \min_{\ell \in S_k} g_t(\lambda_\ell(t)) - \mathcal{O}(\boldsymbol{\Psi}_{\mathrm{G}}\eta^{3-0.1}), (1 + \frac{1}{100M}) \max_{\ell \in S_k} g_t(\lambda_\ell(t))]$ for all $1 \leq k \leq p$. We further define $I_{p+1}(t) = \{\boldsymbol{\beta}\eta\}$.

Claim 1: We argue the average of the angles in $[t^{(d-1)}, t^{(d)})$ is of order $\mathcal{O}(\eta)$. We split the entire interval $[t^{(d-1)}, t^{(d)})$ into different smaller trunks in the following way. We use $t^{(d-1)} = \widetilde{t}_0 < \widetilde{t}_1 < \widetilde{t}_2 \ldots \widetilde{t}_\ell = t^{(d)}$ to denote the starting step of each trunk. Each $\widetilde{t}_i$ is defined from $\widetilde{t}_{i-1}$ for $i > 0$. The behavior of each trunk depends on the magnitude of the iterate along the top eigenvector of hessian. We classify the trunks on the basis of $2p + 1$ possibilities: Consider a general $\widetilde{t}_i$,

A. If $G_{\widetilde{t}_i} \geq \max\{y \in I_1(\widetilde{t}_i)\}$, then we define $\widetilde{t}_{i+1}$ as

$$\widetilde{t}_{i+1} = \min_{t > \widetilde{t}_i}\{t \mid G_t \leq (1 + \frac{1}{200M}) \max_{\ell \in S_1} g_t(\lambda_\ell(t)), \|\widetilde{x}(t)\| \leq 0.5\lambda_1(t)\eta + \boldsymbol{\Psi}_{\mathrm{norm}}\eta^2\}.$$

$B(k)$. For any $1 \leq k \leq p$, if $G_{\widetilde{t}_i} \in I_k(\widetilde{t}_i)$, then we define $\widetilde{t}_{i+1}$ as

$$\widetilde{t}_{i+1} = \min_{t > \widetilde{t}_i}\{t \mid G_t \geq \max y \in I_k(t), \|\widetilde{x}(t)\| \leq 0.5\lambda_1(t)\eta + \boldsymbol{\Psi}_{\mathrm{norm}}\eta^2\}.$$

$C(k)$. For any $1 \leq k \leq p$, if $\max\{y \in I_{k+1}(\widetilde{t}_i)\} - \mathcal{O}(\boldsymbol{\Psi}_{\mathrm{G}}\eta^{3-0.1}) \leq G_{\widetilde{t}_i} \leq \min\{y \in I_k(\widetilde{t}_i)\}$, then we define $\widetilde{t}_{i+1}$ as

$$\widetilde{t}_{i+1} = \min_{t > \widetilde{t}_i}\{t \mid G_t \geq \min\{y \in I_k(t)\}, \|\widetilde{x}(t)\| \leq 0.5\lambda_1(t)\eta + \boldsymbol{\Psi}_{\mathrm{norm}}\eta^2\}.$$

We analyse the behavior of a general $\widetilde{t}_i$ when it falls in any of the above cases:

A. First of all, since $G_t \geq (1 + \frac{1}{200M}) \max_{k \in [M]} g_t(\lambda_k(t))$ for all $\widetilde{t}_i \leq t < \widetilde{t}_{i+1}$ we can show from Lemma E.2 that the angle with the top eigenvector quickly drops to $\mathcal{O}(\eta)$ in at most t_{escape} time-steps. Moreover, the iterate's magnitude can only drop along the top eigenvector when the angle with the top eigenvector is smaller than $\mathcal{O}(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta)$, and the drop is at most $\mathcal{O}(\boldsymbol{\Psi}_{\mathrm{G}}\eta^3)$ (Corollary E.9). Thus, during alignment of the iterate to the top eigenvector, G_t never drops. Moreover, after the alignment, it takes $\Omega(\frac{1}{\eta^2})$ steps for the iterate's magnitude along the top eignvector to drop below $(1 + \frac{1}{200M}) \max_{k \in [M]} g_\cdot(\lambda_k(\cdot))$. Hence,

$$\left|\widetilde{t}_{i+1} - \widetilde{t}_i\right| \geq \Omega\left(\frac{1}{\boldsymbol{\Psi}_{\mathrm{G}}\eta^2}\right), \quad \sum_{t=\widetilde{t}_i}^{\widetilde{t}_{i+1}} \theta_t \leq \mathcal{O}\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}(\widetilde{t}_{i+1} - \widetilde{t}_i)}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta\right).$$

After G_t drops out of $I_1(t)$, it moves to case $B(1)$.

$B(k)$. For any $1 \leq k \leq p$, if $G_{\widetilde{t}_i} \in I_k(\widetilde{t}_i)$, then $\widetilde{t}_{i+1}$ is defined as the time at which it escapes the strip I_k. From Lemma E.3 we have that the sum of angle over this time is

$$\sum_{t=\widetilde{t}_i}^{\widetilde{t}_{i+1}} \theta_t = \mathcal{O}\left(\frac{\boldsymbol{\zeta}^5}{\boldsymbol{\mu}^2\boldsymbol{\Delta}^2} + \frac{M\boldsymbol{\zeta}}{\boldsymbol{\beta}^3}t_{\mathrm{escape}} + \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta(\widetilde{t}_{i+1} - \widetilde{t}_i)\right).$$

Moreover, from Lemma E.4, we have the following two major claims about the regions to which the iterate escapes:

i. If $k > 1$ and $G_{\widetilde{t}_{i+1}} \geq \max_{y \in I_k(\widetilde{t}_{i+1})}$ i.e. it goes above the strip $I_k(\widetilde{t}_{i+1})$, then it never returns back to this strip in the future.

ii. Moreover, G_t never goes to any region below $I_k(t)$, i.e. there exists no time $t > \widetilde{t}_i$, where $G_t \leq \min\{y \in I_k(t)\}$.

Hence, at time $\widetilde{t}_{i+1}$, we move to cases $C(j)$ or $B(j)$ or A, where $j < k$, and never return back to case $B(k)$.

$C(k)$. For any $1 \le k \le p$, if $G_{\widetilde{t}_i}$, if $\max\{y \in I_{k+1}(\widetilde{t}_i)\} - \mathcal{O}(\boldsymbol{\Psi}_{\mathrm{G}}\eta^{3-0.1}) \le G_{\widetilde{t}_i} \le \min\{y \in I_k(\widetilde{t}_i)\}$, then we have from Lemma E.4 that the iterate quickly moves beyond $\min\{y \in I_k(\widetilde{t}_i)\}$ within $\mathcal{O}(\frac{\boldsymbol{\zeta}^2}{\boldsymbol{\mu}\boldsymbol{\Delta}} c_{\mathrm{escape}}^{-2}\eta^{-0.1}\log\frac{1}{\eta})$ steps. Thus,

$$\sum_{t=\widetilde{t}_i}^{\widetilde{t}_{i+1}} \theta_t = \mathcal{O}(\frac{\boldsymbol{\zeta}^2}{\boldsymbol{\mu}\boldsymbol{\Delta}} c_{\mathrm{escape}}^{-2}\eta^{-0.1}\log\frac{1}{\eta}).$$

Moreover, from the discussion on cases $B(j)$ and $C(j)$ for $j < k$, we never return back to case $C(k)$ once we escape it.

Thus to summarize, for each $C(1), \cdots, C(M), B(2), \cdots, B(M)$, there can be at most one trunk that represents the behavior. There are only a constant number of trunks that represent cases A and $B(1)$, since $B(1)$ is followed by A, where the iterate is provably stuck for $\Omega\left(\frac{1}{\eta^2}\right)$ steps.

All in all, we must have

$$\begin{aligned}\sum_{t=t^{(d-1)}}^{t^{(d)}} \theta_t &= \mathcal{O}\left(\frac{\boldsymbol{\zeta}^5 M}{\boldsymbol{\mu}^2\boldsymbol{\Delta}^2} + \frac{M^2\boldsymbol{\zeta}}{\boldsymbol{\beta}^3} t_{\mathrm{escape}}\right) + \mathcal{O}(\frac{\boldsymbol{\zeta}^2 M}{\boldsymbol{\mu}\boldsymbol{\Delta}} c_{\mathrm{escape}}^{-2}\eta^{-0.1}\log\frac{1}{\eta}) + \mathcal{O}\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}(t^{(d)} - t^{(d-1)})}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta\right) \\ &= \mathcal{O}\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}(t^{(d)} - t^{(d-1)})}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta\right),\end{aligned}$$

where the final step follows from setting η sufficiently small and the fact that $t^{(d)} - t^{(d-1)} \ge \Omega(\frac{1}{\eta^2})$.

Claim 2: For the second claim, note that from our definition of the cases, if $\widetilde{t}_0 := t^{(d-1)}$ represents case $C(p)$, i.e. $\boldsymbol{\beta}\eta \le G_{\widetilde{t}_0} \le \min\{y \in I_p(\widetilde{t}_0)\}$, then at time $\widetilde{t}_1$, $G_{\widetilde{t}_1}$ must be inside or above the strip $I_p(\widetilde{t}_1)$ and never returns back. Hence, at the end of the block, we know $G_{t^{(d)}}$ must be at least $\left(1 - \frac{1}{100M}\right)\min_{k \in g_{t^{(d)}}(\lambda_k(t^{(d)}))}\eta \ge \frac{\boldsymbol{\Delta}\boldsymbol{\mu}}{8\boldsymbol{\zeta}^2}\eta$, where the simplification follows from the constants in Definition B.5 and the assumption that the iterate never leaves $Y^{\boldsymbol{\epsilon}}$. Since, $\boldsymbol{\beta}$ was chosen a constant smaller than $\frac{\boldsymbol{\Delta}\boldsymbol{\mu}}{8\boldsymbol{\zeta}^2}$, we have $G_{t^{(d)}} \ge \boldsymbol{\beta}\eta$.

Combining the blocks: Thus, in summary, we have

(a) The average angle in each of the blocks is $\mathcal{O}(\eta)$.
(b) The iterate never drops the magnitude along the top eigenvector below $\boldsymbol{\beta}\eta$.

Combining the angles over all the blocks, we will have

$$\sum_{\ell=0}^{t_2} \theta_\ell \le \mathcal{O}\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}} t_2\eta\right).$$

□

Lemma E.2. *Consider the setting of Lemma E.1. For any time t, where $x_\eta(t) \in Y^{\boldsymbol{\epsilon}}$, if $\|\widetilde{x}_\eta(t)\| \le 0.5\eta\lambda_1(t) + \boldsymbol{\Psi}_{\mathrm{norm}}\eta^2$, and $G_t \ge \left(1 + \frac{1}{100M}\right)c_{\mathrm{thres}}(t)\eta$, then*

$$\begin{aligned}\tan\theta_{t+1} &\le \max\left(1 - \frac{\lambda_M(t)}{\lambda_1(t)}, \frac{\lambda_2(t)}{\lambda_1(t)}\right)\tan\theta_t \\ \tan\theta_{t+2} &\le \left(1 - 2\min\left(\frac{1}{200M}, \min_{i \le M}\frac{\lambda_i(t)}{2\lambda_1(t)}(1 - \frac{\lambda_i(t)}{\lambda_1(t)})\right)\right)\tan\theta_t,\end{aligned}$$

where G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$, provided $\eta \le \mathcal{O}(\frac{\boldsymbol{\mu}^2}{M\boldsymbol{\nu}\boldsymbol{\zeta}^2})$ and $\overline{x_\eta(t)x_\eta(t+1)}, \overline{x_\eta(t+1)x_\eta(t+2)} \subset Y^{\boldsymbol{\epsilon}}$. Here, $c_{\mathrm{thres}}(t) = \max_{k \in [M]} g_t(\lambda_k(t))$.

Proof. The proof follows from using the noisy update rule for Normalized GD, as derived in Equation (27). Which says that the Normalized GD update is very close to the update in a quadratic model with an additional $\mathcal{O}(\eta^2)$ error. The result then follows from using Lemma A.11 and Lemma A.10, that computes the convergence rate towards the top eigenvector for a quadratic model. □

Lemma E.3. *Consider the setting of Lemma E.1. Consider any time $\bar{t}$, where $x_\eta(\bar{t}) \in Y^\epsilon$, and $\left\|\widetilde{x}_\eta(\bar{t})\right\| \leq 0.5\eta\lambda_1(\bar{t}) + \boldsymbol{\Psi}_{\text{norm}}\eta^2$. Suppose we are also given p disjoint subsets of $[M]$, $S_1, \cdots, S_p$ (with $1 \leq p \leq M$) and a step $t_{\text{stop}} \geq \bar{t}$, such that for any $i, j \in [p]$ with $i \neq j$, and for any $\bar{t} \leq t \leq t_{\text{stop}}$, we can guarantee*

$$\min_{k \in S_i, \ell \in S_j} \left|\left|\frac{\lambda_1}{2} - \lambda_k(t)\right| - \left|\frac{\lambda_1}{2} - \lambda_\ell(t)\right|\right| \geq \frac{1}{2} \times 10^{-3}\lambda_1(t),$$

and the subsets are arranged such that $\min_{\ell \in S_i} g_t(\lambda_\ell(t)) > \max_{\ell \in S_j} g_t(\lambda_\ell(t)), \quad$ if $i > j$.

Consider any subset S_k for $1 \leq k \leq p$. If $(1 - \frac{1}{100M}) \min_{\ell \in S_k} g_{\bar{t}}(\lambda_\ell(\bar{t})) \leq G_{\bar{t}} \leq (1 + \frac{1}{100M}) \max_{\ell \in S_k} g_{\bar{t}}(\lambda_\ell(\bar{t}))$ and suppose there exists some time $\bar{t} \leq t' \leq t_{\text{stop}}$ such that the iterate is stuck inside this region in the interval $(\bar{t}, t')$. I.e. for all $t \in (\bar{t}, t')$, whenever $\|\widetilde{x}_\eta(t)\| \leq 0.5\eta\lambda_1(t) + \boldsymbol{\Psi}_{\text{norm}}\eta^2$, we must have $(1 - \frac{1}{100M}) \min_{\ell \in S_k} g_t(\lambda_\ell(t)) \leq G_t \leq (1 + \frac{1}{100M}) \max_{\ell \in S_k} g_t(\lambda_\ell(t))$. Then,

$$\sum_{\ell=\bar{t}}^{t'} \theta_\ell \leq \mathcal{O}\left(\frac{\boldsymbol{\zeta}^5}{\boldsymbol{\mu}^2\boldsymbol{\Delta}^2} + \frac{M\boldsymbol{\zeta}}{\boldsymbol{\beta}^3} t_{\text{escape}} + \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta(t' - \bar{t})\right),$$

where G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$, provided for all $\bar{t} \leq \ell < t'$, $\overline{x_\eta(\ell)x_\eta(\ell+1)} \subset Y^\epsilon$.

Proof. We will sketch the outline of the proof here. First of all, we use the noisy update rule for Normalized GD, as derived in Lemma B.10. Which says that the Normalized GD update is very close to the update in a quadratic model with an additional $\mathcal{O}(\eta^2)$ error. Keeping this in mind, we then divide our trajectory in the interval $(\bar{t}, t')$ as per Algorithm 2 into three subsets N_0, N_1, N_2. Please see Appendix E.2 for a summary on the properties of these 3 sets. To recall, they are

1. All the time steps t in N_0 and N_1 have the norm of the iterate $\widetilde{x}_\eta(t)$ at most $0.5\lambda_1(t)\eta + \boldsymbol{\Psi}_{\text{norm}}\eta^2$.
2. For any step t in N_0, we must have the norm of the iterate $\widetilde{x}_\eta(t+2)$ at least $0.5\lambda_1(t+2)\eta + \boldsymbol{\Psi}_{\text{norm}}\eta^2$.
3. For any step t in N_1, we must have the norm of the iterate $\widetilde{x}_\eta(t+2)$ at most $0.5\lambda_1(t+2)\eta + \boldsymbol{\Psi}_{\text{norm}}\eta^2$.
4. For any step t in N_1, we have $t+1$ in N_2. Moreover, for any time t in N_2, we must have $t - 1 \in N_1$.

Consider the following arguments:

1. First of all, the magnitude along all eigenvectors $v_i(\cdot)$ for $i \in \cup_{j>k} S_j$ can't be greater than $\boldsymbol{\alpha}\eta^2$ for more than t_{escape} number of steps from Lemma E.19.
2. Furthermore, the magnitude along all eigenvectors $v_i(\cdot)$ for $i \in \cup_{j<k} S_j$ can't be greater than $\boldsymbol{\alpha}\eta^2$ for more than $\mathcal{O}\left(\frac{M\boldsymbol{\zeta}}{\boldsymbol{\beta}^3} t_{\text{escape}}\right)$ number of steps from Lemma E.17. Thus, we will only consider the steps at which the magnitude along the eigenvectors $v_i(\cdot)$ for $i \in \cup_{j \neq k} S_j$ is small.
3. Consider any $t \in N_1$. Using the behavior of $|\langle v_1(t), \widetilde{x}_\eta(t)\rangle|$ from Lemma D.2 and the behavior of G_t from Lemma E.8, we can show that in each of the time-frames, if $\theta_t > \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta$, $G_{t+2} \geq (1 + \frac{\boldsymbol{\mu}\boldsymbol{\Delta}}{\boldsymbol{\zeta}^2}\sin^2\theta_t)G_t$.

   Suppose, we divide N_1 into groups, $N_1^{(1)}$ and $N_1^{(2)}$, such that $G_{t+2} > G_t$ if $t \in N_1^{(1)}$ and $G_{t+2} \leq G_t$ if $t \in N_1^{(2)}$. Since, the increase in G_t during this interval can be at most from $\left(1 - \frac{1}{100M}\right) \min_{i \in S_k} g_t(\lambda_i(t))$ to $\boldsymbol{\zeta}\eta$ (using our alignment condition from Equation (28)), we must have

$$\sum_{t \in N_1^{(1)}} \theta_t \leq \mathcal{O}\left(\frac{\boldsymbol{\zeta}^3}{0.99\boldsymbol{\mu}\boldsymbol{\Delta} c_{\text{thres}}(\bar{t})}\right).$$

   Moreover, if $G_{t+2} \leq G_t$ at any step t, then we have $\theta_t \leq \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta$. That implies,

$$\sum_{t \in N_1^{(2)}} \theta_t \leq \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta \left|N_1^{(2)}\right| \leq \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}(t' - \bar{t})\eta.$$

Thus,

$$\sum_{t\in N_1}\theta_t = \sum_{t\in N_1^{(1)}}\theta_t + \sum_{t\in N_1^{(2)}}\theta_t \le \mathcal{O}\left(\frac{\boldsymbol{\zeta}^3}{0.99\boldsymbol{\mu}\boldsymbol{\Delta}c_{\text{thres}}(\bar{t})} + \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}(t'-\bar{t})\eta\right).$$

Using a very rough bound of $\frac{\boldsymbol{\mu}\boldsymbol{\Delta}}{\boldsymbol{\zeta}^2}$ for $c_{\text{thres}}(\bar{t})$, we have the first term of the above bound as $\mathcal{O}\left(\frac{\boldsymbol{\zeta}^5}{0.99\boldsymbol{\mu}^2\boldsymbol{\Delta}^2}\right)$.

4. Moreover, using the fact that the angle drops whenever the norm of the iterate is less than $\frac{\eta\lambda_1(t)}{2}$ (Lemma E.6), we must have $\theta_t < \theta_{t-1}$. Furthermore, as listed before, for any time $t \in N_2$, $t-1 \in N_1$. That implies,

$$\sum_{t\in N_1\cup N_2}\theta_t \le \mathcal{O}\left(\frac{\boldsymbol{\zeta}^5}{\boldsymbol{\mu}^2\boldsymbol{\Delta}^2} + \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta(t'-\bar{t})\right).$$

5. We look at the angles of the remaining time-steps, which is N_0. Recall that we are only looking at steps t, where the magnitude along the eigenvectors $v_i(\cdot)$ for $i \in \cup_{j\neq k}S_j$ is less than $\boldsymbol{\alpha}\eta^2$. Furthermore, the iterate is stuck inside this thin region, where the magnitude along the top eigenvector is inside this thin strip of $(1-\frac{1}{100M})\min_{\ell\in S_k} g_t(\lambda_\ell(t)) \le G_t \le (1+\frac{1}{100M})\max_{\ell\in S_k} g_t(\lambda_\ell(t))$. Using the difference between the subspaces as 10^{-3}, we have that for any $i \in S_k$, the magnitude along i^{th} eigenvector is inside $0.99g_t(\lambda_i(t)) \le G_t \le 1.01g_t(\lambda_i(t))$. Thus, we can use the result from the quadratic model (Lemma A.10) that θ_t can be at most $1.01\theta_{t-2}$. This implies, the sum of the remaining angles can be at most 1.02 times the above bound. That can be incorporated into the above bound to get

$$\sum_{t\in N_1\cup N_2\cup N_0}\theta_t \le \mathcal{O}\left(\frac{\boldsymbol{\zeta}^5}{\boldsymbol{\mu}^2\boldsymbol{\Delta}^2} + \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta(t'-\bar{t})\right).$$

□

Lemma E.4. *Consider the same setting as Lemma E.3. For any subset S_k with $1 \le k \le p$, if at time $\bar{t}$, $(1-\frac{1}{100M})\min_{\ell\in S_k} g_{\bar{t}}(\lambda_\ell(\bar{t})) \le G_{\bar{t}} \le (1+\frac{1}{100M})\max_{\ell\in S_k} g_{\bar{t}}(\lambda_\ell(\bar{t}))$, then we have the following two claims that hold with probability at least $1-\eta^{10}$:*

1. *If $k > 1$, and there exists some time $t' \le t_{\text{stop}}$ such that $G_{t'} \ge (1+\frac{1}{100M})\max_{\ell\in S_k} g_{t'}(\lambda_\ell(t'))$, then for all time $t' \le \widetilde{t} \le t_{\text{stop}}$, $G_{\widetilde{t}} \ge (1+\frac{1}{100M})\max_{\ell\in S_k} g_{\widetilde{t}}(\lambda_\ell(\widetilde{t}))$.*

   In addition, there must exist a time $\widetilde{t} = t' + \mathcal{O}(\frac{\boldsymbol{\zeta}^2}{\boldsymbol{\mu}\boldsymbol{\Delta}}c_{\text{escape}}^{-2}\eta^{-0.1}t_{\text{escape}})$, such that

$$G_{\widetilde{t}} \ge (1-\frac{1}{100M})\min_{\ell\in S_{k-1}} g_{\widetilde{t}}(\lambda_\ell(\widetilde{t})),$$

   provided $\left\|\widetilde{x}_\eta(\widetilde{t})\right\| \le 0.5\eta\lambda_1(\widetilde{t}) + \boldsymbol{\Psi}_{\text{norm}}\eta^2$.
2. *There doesn't exist a time $t' \le t_{\text{stop}}$ such that $G_{t'} \le (1-\frac{1}{100M})\min_{\ell\in S_k} g_{t'}(\lambda_\ell(t')) + \mathcal{O}(\boldsymbol{\Psi}_{\text{G}}\eta^{3-0.1})$.*

The result holds true when $\eta \le \widetilde{\mathcal{O}}(\frac{\boldsymbol{\mu}^{10}\boldsymbol{\beta}^9\boldsymbol{\Delta}}{\boldsymbol{\zeta}^{10}\boldsymbol{\nu}^6 D\boldsymbol{\xi}})$, and for all time $t \le \bar{t} < t'$, $\overline{x_\eta(\bar{t})x_\eta(\bar{t}+1)} \subset Y^{\boldsymbol{\epsilon}}$. Here G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$, $c_{\text{escape}} = \Theta(\frac{\boldsymbol{\beta}^6\boldsymbol{\mu}^6}{\boldsymbol{\zeta}^6\boldsymbol{\nu}^3})$, and $t_{\text{escape}} = \mathcal{O}(\frac{M\boldsymbol{\zeta}^2}{\boldsymbol{\mu}\boldsymbol{\Delta}}\log 1/\eta)$.

Proof. We outline the proof sketch here. For the first claim, we show that whenever the magnitude of the iterate along the top eigenvector crosses above the strip, it never re-enters the strip. Moreover, in $\mathcal{O}(\eta^{-0.1}t_{\text{escape}})$ time, the magnitude of the iterate along the top eigenvector jumps to the higher strip. The second claim shows that the magnitude of the iterate along the top eigenvector never drops below the strip. Both the claims will follow from Corollary E.9 and Corollary E.10 that shows that the drop in the magnitude along the top eigenvector can only be of order $\mathcal{O}(\boldsymbol{\Psi}_{\text{G}}\eta^3)$ when the angle of the iterate with the top eigenvector is below $\mathcal{O}\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta\right)$. In that case, we can wait for $\eta^{-0.1}$ steps to apply the result of Lemma E.11, which shows that a minor injection of η^{100} noise can guarantee the increase in the magnitude along the top eigenvector by a constant factor.

□

E.2. Properties of Algorithm 2

The properties of the three sets N_0, N_1, N_2 in an interval $(\widetilde{t}, \bar{t})$ given by the algorithm in Algorithm 2 are:

1. All the time steps t in N_0 and N_1 have the norm of the iterate $\widetilde{x}_\eta(t)$ at most $0.5\lambda_1(t)\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$.
2. For any step t in N_0, we must have the norm of the iterate $\widetilde{x}_\eta(t+2)$ at least $0.5\lambda_1(t+2)\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$.
3. For any step t in N_1, we must have the norm of the iterate $\widetilde{x}_\eta(t+2)$ at most $0.5\lambda_1(t+2)\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$.
4. For any step t in N_1, we have $t+1$ in N_2. Moreover, for any time t in N_2, we must have $t-1 \in N_1$.

Lemma E.5. *For any step t in N_0, $t+1$ can't be in N_0.*

Proof. Suppose there exists a time t, such that t and $t+1$ are both in N_0. Then, from the properties of N_0 outlined before, we must have $\|\widetilde{x}_\eta(t+2)\| \geq 0.5\lambda_1(t+2)\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$ and $\|\widetilde{x}_\eta(t+3)\| \geq 0.5\lambda_1(t+3)\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$. However, this contradicts the result of Lemma D.1 which shows that the norm of the iterate can't be over $0.5\lambda_1(\cdot)\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$ for more than one steps. □

Lemma E.6. *For any step t in N_0 and N_1, $\theta_{t+1} \leq \left(1 - \frac{\boldsymbol{\Delta\mu}}{\boldsymbol{\zeta}} + \mathcal{O}(\frac{\mathbf{\Psi}_{\text{norm}}}{\boldsymbol{\beta}}\eta) + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu^2\beta}}\eta)\right)\theta_t + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu^2\beta}}\eta)$.*

Proof. Using the property of N_0 and N_1 outlined above, we have the norm of $\widetilde{x}_\eta(t)$ at most $0.5\lambda_1(t)\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$. Then, we can directly use the result from Lemma D.2 to show that the angle has to drop, albeit an error of $\mathcal{O}(\eta)$. □

Lemma E.7. *For any step t in N_0 and N_1,*

$$\begin{aligned}&|\langle v_1(t+1), x_\eta(t+1) - \Phi(x_\eta(t+1))\rangle| \\ &\geq (1 + \mathcal{O}(\mathbf{\Psi}_{\text{norm}}\eta))\,|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}^2}\eta^2).\end{aligned}$$

Proof. Using the property of N_0 and N_1 outlined above, we have the norm of $\widetilde{x}_\eta(t)$ at most $0.5\lambda_1(t)\eta + \mathbf{\Psi}_{\text{norm}}\eta^2$. Then, we can directly use the result from Lemma D.2 to show that the magnitude along the top eignvector has to increase, albeit an error of $\mathcal{O}(\eta^2)$. □

E.3. Main Helping Lemmas for Phase II

Here, we will state the important three lemmas that we used for the proof of Lemma E.1. We have implicitly assumed in all the lemmas, that Equation (28) holds true for the time under consideration, which we showed in Lemma C.1, and also the fact that we start Phase II from a point where the alignment along the top eigenvector is non negligible.

The following lemma shows the behavior of the iterate along the top eigenvector.

Lemma E.8 (Behavior along the top eigenvector)**.** *Suppose $\eta \leq \mathcal{O}(\min(\frac{\boldsymbol{\mu}^3}{\boldsymbol{\zeta}^3\boldsymbol{\varsigma}^2\boldsymbol{\xi\nu}\sqrt{D}}, \frac{\boldsymbol{\mu}^3}{\boldsymbol{\nu\xi\varsigma\zeta}^2 D}, \frac{\boldsymbol{\mu}^3}{\boldsymbol{\nu\xi\varsigma\zeta}^2 D}, \frac{\boldsymbol{\Delta}}{\mathbf{\Psi}_{\text{norm}}}))$. Consider any time t, such that $x_\eta(t) \in Y^\epsilon$, where $\|\widetilde{x}_\eta(t)\| \leq \frac{1}{2}\eta\lambda_1(t) + \mathbf{\Psi}_{\text{norm}}\eta^2$ holds true. If G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$ and G_{t+2} denotes the quantity $|\langle v_1(t+2), x_\eta(t+2) - \Phi(x_\eta(t+2))\rangle|$, then the following holds true:*

$$\begin{aligned}G_{t+2} \geq (1 + \frac{1}{4}\min_{2\leq j\leq M}\frac{\lambda_j(t)(\lambda_1(t) - \lambda_j(t))}{\lambda_1^2(t)}\sin^2\theta_t)G_t \\ - \mathcal{O}(\frac{\boldsymbol{\Upsilon\zeta}^2\boldsymbol{\xi\nu\chi}}{\boldsymbol{\mu}^3\boldsymbol{\Delta}}\eta^3 + (1 + \eta/G_t)\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t}\sin\theta_t),\end{aligned}$$

provided $G_t \geq \Omega(\eta^{1.5})$ and $\overline{x_\eta(t)x_\eta(t+1)}, \overline{x_\eta(t+1)x_\eta(t+2)} \subset Y^\epsilon$. Here θ_t is given by $\arctan(\frac{\left\|P_{t,\Gamma}^{(2:M)}\widetilde{x}_\eta(t)\right\|}{|\langle v_1(t), \widetilde{x}_\eta(t)\rangle|})$, with $P_{t,\Gamma}^{(2:M)}$ denoting the projection matrix onto the subspace spanned by $v_2(t), \ldots, v_M(t)$.

The proof of the above lemma is given in Appendix E.3.1.

A corollary of the above lemma is that when the magnitude along the top eigenvalue is $\Omega(\eta)$, the magnitude drops, only when the angle of the iterate with the top eigenvector is $\mathcal{O}(\eta)$.

Corollary E.9. *Consider any time t, such that $x_\eta(t) \in Y^\epsilon$, where $\|\widetilde{x}_\eta(t)\| \le \frac{1}{2}\eta\lambda_1(t) + \boldsymbol{\Psi}_{\rm norm}\eta^2$ holds true. If G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$ and G_{t+2} denotes the quantity $|\langle v_1(t+2), x_\eta(t+2) - \Phi(x_\eta(t+2))\rangle|$, then $G_{t+2} \ge G_t$ for all*

$$|\theta_t| \ge \Omega\left(\max\left(\sqrt{\frac{\boldsymbol{\Upsilon\zeta^2\xi\nu\chi}}{\boldsymbol{\mu}^3\Delta}\frac{\eta^3}{G_t}}, (1+\eta/G_t)\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t^2}\right)\right),$$

provided $G_t \ge \Omega(\eta^{1.5})$, and $\overline{x_\eta(t)x_\eta(t+1)}, \overline{x_\eta(t+1)x_\eta(t+2)} \subset Y^\epsilon$. Moreover, if $G_t \ge \boldsymbol{\beta}\eta$ for some $\boldsymbol{\beta} \ge 0$, then the above bound can be simplified as

$$|\theta_t| \ge \Omega\left(\frac{\boldsymbol{\Upsilon\zeta^2\xi\nu\chi}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^3\boldsymbol{\Delta}}\eta\right).$$

The next corollary shows that if the magnitude along the top eigenvector drops, when it is $\Omega(\eta)$, it can only drop by a magnitude of $\mathcal{O}(\eta^3)$.

Corollary E.10. *Consider any time t, such that $x_\eta(t) \in Y^\epsilon$ and $\|\widetilde{x}_\eta(t)\| \le \frac{1}{2}\eta\lambda_1(t) + \boldsymbol{\Psi}_{\rm norm}\eta^2$. Let G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$, then*

$$G_{t+2} \ge G_t - \mathcal{O}\left(\frac{\boldsymbol{\Upsilon\zeta^2\xi\nu\chi}}{\boldsymbol{\mu}^3\Delta}\eta^3 + (1+\eta/G_t)\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t}\sqrt{\frac{\boldsymbol{\Upsilon\zeta^2\xi\nu\chi}}{\boldsymbol{\mu}^3\Delta}\frac{\eta^3}{G_t}}\right),$$

provided $G_t \ge \Omega(\eta^{1.5})$ and $\overline{x_\eta(t)x_\eta(t+1)}, \overline{x_\eta(t+1)x_\eta(t+2)} \subset Y^\epsilon$. Therefore, there is some $\boldsymbol{\beta} > 0$, such that whenever $G_t \ge \boldsymbol{\beta}\eta$, we have $G_{t+2} \ge G_t - \mathcal{O}(\boldsymbol{\Psi}_{\rm G}\eta^3)$, where $\boldsymbol{\Psi}_{\rm G} = \frac{\boldsymbol{\Upsilon\zeta^3\xi\nu^2\chi}}{\boldsymbol{\mu}^5\boldsymbol{\beta}^3\Delta}$.

E.3.1. Behavior along Top Eigenvalue

Lemma E.8 (Behavior along the top eigenvector)**.** *Suppose $\eta \le \mathcal{O}(\min(\frac{\boldsymbol{\mu}^3}{\boldsymbol{\zeta}^3\boldsymbol{\zeta}^2\boldsymbol{\xi\nu}\sqrt{D}}, \frac{\boldsymbol{\mu}^3}{\boldsymbol{\nu\xi\varsigma\zeta}^2 D}, \frac{\boldsymbol{\mu}^3}{\boldsymbol{\nu\xi\varsigma\zeta}^2 D}, \frac{\boldsymbol{\Delta}}{\boldsymbol{\Psi}_{\rm norm}}))$. Consider any time t, such that $x_\eta(t) \in Y^\epsilon$, where $\|\widetilde{x}_\eta(t)\| \le \frac{1}{2}\eta\lambda_1(t) + \boldsymbol{\Psi}_{\rm norm}\eta^2$ holds true. If G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$ and G_{t+2} denotes the quantity $|\langle v_1(t+2), x_\eta(t+2) - \Phi(x_\eta(t+2))\rangle|$, then the following holds true:*

$$\begin{aligned} G_{t+2} \ge (1 + \frac{1}{4}\min_{2\le j\le M}\frac{\lambda_j(t)(\lambda_1(t)-\lambda_j(t))}{\lambda_1^2(t)}\sin^2\theta_t)G_t \\ - \mathcal{O}(\frac{\boldsymbol{\Upsilon\zeta^2\xi\nu\chi}}{\boldsymbol{\mu}^3\Delta}\eta^3 + (1+\eta/G_t)\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t}\sin\theta_t), \end{aligned}$$

provided $G_t \ge \Omega(\eta^{1.5})$ and $\overline{x_\eta(t)x_\eta(t+1)}, \overline{x_\eta(t+1)x_\eta(t+2)} \subset Y^\epsilon$. Here θ_t is given by $\arctan(\frac{\left\|P_{t,\Gamma}^{(2:M)}\widetilde{x}_\eta(t)\right\|}{|\langle v_1(t), \widetilde{x}_\eta(t)\rangle|})$, with $P_{t,\Gamma}^{(2:M)}$ denoting the projection matrix onto the subspace spanned by $v_2(t), \ldots, v_M(t)$.

Proof. Using the Normalized GD update, we have

$$\begin{aligned} &\langle v_1(t), x_\eta(t+1) - \Phi(x_\eta(t))\rangle - \langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle \\ &= -\eta\frac{\langle v_1(t), \nabla L(x_\eta(t))\rangle}{\|\nabla L(x_\eta(t))\|} \\ &= -\eta\frac{\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \frac{1}{2}\partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]\rangle}{\left\|\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \frac{1}{2}\partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]\right\|} \\ &+ \mathcal{O}(\frac{\boldsymbol{\Upsilon\zeta}^2}{\boldsymbol{\mu}^3}\eta^3) \end{aligned} \tag{29}$$

$$= -\eta \frac{v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \frac{1}{2} v_1(t)^\top \partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]}{\left\| P_{t,\Gamma} \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \frac{1}{2} P_{t,\Gamma} \partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))] \right\|}$$
$$+ \mathcal{O}(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\eta^3) \tag{30}$$
$$= -\eta \frac{v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]}{\| P_{t,\Gamma} \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] \|} + \text{err} + \mathcal{O}(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\eta^3). \tag{31}$$

The steps followed above are as follows:

1. In Equation (29), we use taylor expansion to expand $\nabla L(x_\eta(t))$ around $\Phi(x_\eta(t))$, with an error of $\mathcal{O}(\boldsymbol{\Upsilon} \|x_\eta(t) - \Phi(x_\eta(t))\|^3)$. Since by alignment condition Equation (28),we have $\|\widetilde{x}_\eta(t)\| \leq \mathcal{O}(\lambda_1(t)\eta)$, we have $\|x_\eta(t) - \Phi(x_\eta(t))\| \leq \mathcal{O}(\frac{\lambda_1(t)}{\lambda_M(t)})$. This adds an error of magnitude $\mathcal{O}(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\eta^3)$ to the entire term.

2. In Equation (30), we divide the vector $\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \frac{1}{2}\partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]$ using its projection onto the subspace S_1 spanned by $v_1(t), \ldots, v_M(t)$ and the subspace S_2 spanned by the rest of the eigenvectors $v_{M+1}(t), \ldots, v_D(t)$. Since $\nabla^2 L(\Phi(x_\eta(t)))$ only projects onto the subspace S_1, S_2 can only get its component from $\partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]$, which is of norm $\mathcal{O}(\boldsymbol{\nu} \|x_\eta(t) - \Phi(x_\eta(t))\|^2)$. Thus, we can only consider the vector in S_1, with an error vector $\mathcal{O}(\boldsymbol{\nu} \|x_\eta(t) - \Phi(x_\eta(t))\|^2)$ orthogonal to S_1. Taking the norm of this vector, we get an additional error term of magnitude $\mathcal{O}(\eta^4)$.

3. In the final step (Equation (31)), we only consider $\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]$, which gives out an error term err.

Now, we show that $|\text{err}| \leq \mathcal{O}(\eta^2 \theta_t)$.

First of all, $\left\|\partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]\right\| \leq \mathcal{O}(\boldsymbol{\nu} \|x_\eta(t) - \Phi(x_\eta(t))\|^2) = \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^2}\eta^2)$. Suppose $P_{t,\Gamma}^{(2:M)}$ denotes the projection matrix onto the subspace spanned by $v_2(t), \ldots, v_M(t)$, then denoting by $\text{res}^{(2:M)}$ the vector $P_{t,\Gamma}^{(2:M)} \partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]$ and by res the element $v_1(t)^\top \partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]$, we have

$$|\text{res}|, \left\|\text{res}^{(2:M)}\right\| \leq \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^2}\eta^2). \tag{32}$$

Thus,

$$-\frac{1}{\eta}\text{err} + \frac{v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]}{\|P_{t,\Gamma} \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\|}$$
$$= \frac{v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + v_1(t)^\top \partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]}{\left\|P_{t,\Gamma} \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \frac{1}{2} P_{t,\Gamma} \partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]\right\|} \tag{33}$$
$$= \frac{v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \text{res}}{\left\|P_{t,\Gamma} \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \text{res} \cdot v_1(t) + \text{res}^{(2:M)}\right\|} \tag{34}$$
$$= \frac{v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \text{res}}{\sqrt{(v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \text{res})^2 + \left\|P_{t,\Gamma}^{(2:M)} \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))] + \text{res}^{(2:M)}\right\|^2}} \tag{35}$$
$$= \frac{v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]}{\|P_{t,\Gamma} \nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\|} \left[\sqrt{1 + \text{err}'}\right]$$

where

$$\mathrm{err}' = \frac{\left[\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}\right]^2 - \left[\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]+\mathrm{res}^{(2:M)}\right\|}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]+\mathrm{res}}\right]^2}{1+\left[\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}\right]^2}.$$

We followed the following steps in the above set of equations:

1. In Equation (33), we just copied the term from Equation (30), which was divided into terms err and $\frac{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}{\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\|}$.

2. In Equation (34), we introduced the terms res and $\mathrm{res}^{(2:M)}$ to represent $\partial^2(\nabla L)(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t)), x_\eta(t)-\Phi(x_\eta(t))]$.

3. In Equation (35), we expanded the norm in the denominator using the projection along the vector $v_1(t)$ and the subspace spanned by $v_2(t),\ldots,v_M(t)$.

The magnitude of err' can now be bounded as follows:

$$\begin{aligned}
|\mathrm{err}'| &= \left|\frac{\left[\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}\right]^2 - \left[\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]+\mathrm{res}^{(2:M)}\right\|}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]+\mathrm{res}}\right]^2}{1+\left[\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}\right]^2}\right| \\
&= \frac{1}{\sec^2\theta_t}\left|\left[\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}\right]^2\right. \\
&\quad \left. - \left[\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]+\mathrm{res}^{(2:M)}\right\|}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]+\mathrm{res}}\right]^2\right| \qquad (36) \\
&\le \mathcal{O}\left(\frac{1}{G_t\lambda_1(t)\sec^2\theta_t}\left(\left\|\mathrm{res}^{(2:M)}\right\| + |\mathrm{res}|\tan\theta_t\right)\cdot\left|\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|}{v_1^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}\right|\right) \qquad (37) \\
&\le \mathcal{O}\left(\frac{1}{G_t\lambda_1(t)}\left(\left\|\mathrm{res}^{(2:M)}\right\| + |\mathrm{res}|\tan\theta_t\right)\cdot|\sin 2\theta_t|\right). \qquad (38)
\end{aligned}$$

We followed the following steps in the above set of equations:

1. In Equation (36), we used the definition of θ_t to represent $\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|}{v_1^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}$.

2. In Equation (37), we bound the magnitude of $\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|+\text{res}}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]+\text{res}^{(2:M)}}$ by

$$\begin{aligned}
&\left|\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|+\text{res}^{(2:M)}}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]+\text{res}}\right| \\
&\le\left|\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|}{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}\right| \\
&+\mathcal{O}\Big(\frac{1}{|v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]|} \\
&\cdot\Big(\left\|\text{res}^{(2:M)}\right\|+|\text{res}|\,\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|}{|v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]|}\Big)\Big) \\
&=\frac{\left\|P_{t,\Gamma}^{(2:M)}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right\|}{|v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]|}+\mathcal{O}\left(\frac{1}{G_t\lambda_1(t)}\left(\left\|\text{res}^{(2:M)}\right\|+|\text{res}|\tan\theta_t\right)\right).
\end{aligned}$$

In the final step, we have used the following steps:

$$\left|v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\right|=\left|\lambda_1(t)v_1(t)^\top[x_\eta(t)-\Phi(x_\eta(t))]\right|=\lambda_1(t)G_t.$$

The pre-final step is true iff $G_t>\Omega(\left\|\text{res}^{(2:M)}\right\|+|\text{res}|)$. Since we have a bound of $\mathcal{O}(\frac{\boldsymbol{\nu\zeta}^2}{\boldsymbol{\mu}^2}\eta^2)$ on $|\text{res}|,\left\|\text{res}^{(2:M)}\right\|$ from Equation (32), having a lower bound of $\Omega(\eta^{1.5})$ on G_t suffices, provided $\eta\le\mathcal{O}(\frac{\boldsymbol{\mu}^4}{\boldsymbol{\nu}^2\boldsymbol{\zeta}^4})$.

Hence, combining everything, we have

$$\begin{aligned}
|\text{err}|&\le\eta\left|\sqrt{1+\text{err}'}-1\right|\left|\frac{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}{\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\|}\right| \\
&\le\eta\mathcal{O}(|err'|) \\
&\le\mathcal{O}(\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2G_t\lambda_1(t)}\sin\theta_t),
\end{aligned}$$

where in the second step, we have used the fact that $\left|\frac{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}{\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\|}\right|=|\cos\theta_t|<1$.

Thus, continuing from Equation (31), we have

$$\begin{aligned}
\langle v_1(t),x_\eta(t+1)-\Phi(x_\eta(t))\rangle-\langle v_1(t),x_\eta(t)-\Phi(x_\eta(t))\rangle&=-\eta\frac{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]}{\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t)-\Phi(x_\eta(t))]\|} \\
&+\mathcal{O}(\frac{\boldsymbol{\Upsilon\zeta}^2}{\boldsymbol{\mu}^3}\eta^3+\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t}\sin\theta_t). \qquad (39)
\end{aligned}$$

Similarily, we can show

$$\begin{aligned}
&\langle v_1(t),x_\eta(t+2)-\Phi(x_\eta(t))\rangle-\langle v_1(t),x_\eta(t+1)-\Phi(x_\eta(t))\rangle \\
&=-\eta\frac{v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1)-\Phi(x_\eta(t))]}{\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1)-\Phi(x_\eta(t))]\|} \\
&+\mathcal{O}\left(\frac{\boldsymbol{\Upsilon\zeta}^2}{\boldsymbol{\mu}^3}\eta^3+\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\left|v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1)-\Phi(x_\eta(t))]\right|}\sin\widetilde{\theta}_t\right),
\end{aligned}$$

where $\cos\widetilde{\theta}_t=\frac{\left|v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1)-\Phi(x_\eta(t))]\right|}{\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1)-\Phi(x_\eta(t))]\|}$. Hence, we can combine the two equations to have

$$\langle v_1(t),x_\eta(t+2)-\Phi(x_\eta(t))\rangle=a_{t+1}a_t\langle v_1(t),x_\eta(t+2)-\Phi(x_\eta(t))\rangle+\overline{\text{err}}, \qquad (40)$$

where

$$
\begin{aligned}
a_{t+1} &= 1 - \frac{\eta\lambda_1(t)}{\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1) - \Phi(x_\eta(t))]\|},\\
a_t &= 1 - \frac{\eta\lambda_1(t)}{\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\|},\\
|\overline{\text{err}}| &\le |a_{t+1}|\,\mathcal{O}(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\eta^3 + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\eta^3}{\boldsymbol{\mu}^2 G_t\lambda_1(t)}\sin\theta_t)\\
&\quad + \mathcal{O}\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\eta^3 + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\eta^3}{\boldsymbol{\mu}^2\,|v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1) - \Phi(x_\eta(t))]|}\sin\widetilde{\theta}_t\right).
\end{aligned}
$$

Using Lemma D.2 that predicts an increase (albeit an error of $\mathcal{O}(\eta^2)$) on the projection along the top eigenvector, we have

$$
|\overline{\text{err}}| \le |a_{t+1}|\,\mathcal{O}(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\eta^3 + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\eta^3}{\boldsymbol{\mu}^2 G_t\lambda_1(t)}\sin\theta_t) + \mathcal{O}\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\eta^3 + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\eta^3}{\boldsymbol{\mu}^2 G_t\lambda_1(t)}\sin\theta_t\right).
$$

Since,

$$
\begin{aligned}
\left\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1) - \Phi(x_\eta(t))]\right\| &\ge \left|v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1) - \Phi(x_\eta(t))]\right|\\
&\ge \frac{1}{2}\left|v_1(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\right| = \frac{1}{2}\lambda_1(t)G_t,
\end{aligned}
$$

using Lemma D.2, that predicts an increase (albeit an error of $\mathcal{O}(\eta^2)$) on the projection along the top eigenvector, in the pre-final step, we have $|a_{t+1}| \le \frac{2\eta}{G_t}$. Thus, overall,

$$
|\overline{\text{err}}| \le \mathcal{O}\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\eta^4}{\boldsymbol{\mu}^3 G_t} + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\eta^4}{\boldsymbol{\mu}^2 G_t^2\lambda_1(t)}\sin\theta_t + \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\zeta}}{\boldsymbol{\mu}^3}\eta^3 + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\eta^3}{\boldsymbol{\mu}^2 G_t\lambda_1(t)}\sin\theta_t\right).
$$

Using the steps used for finding the noisy quadratic update rule of Normalized GD in Lemma B.10 (Equation (19)), we can show that

$$
\begin{aligned}
&\langle v_j(t), x_\eta(t+1) - \Phi(x_\eta(t))\rangle - \langle v_j(t), x_\eta(t) - \Phi(x_\eta(t))\rangle\\
&= -\eta\frac{v_j(t)^\top\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]}{\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\|} + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}}{\boldsymbol{\mu}}\eta^2),
\end{aligned}
\tag{41}
$$

for all $2 \le j \le M$. Hence, with Equation (39) and Equation (41), we can further show that

$$
\begin{aligned}
\left\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1) - \Phi(x_\eta(t))]\right\| &\le \eta\lambda_1(t) - \left\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\right\|\\
&\quad - 2\eta\min_{2\le j\le M}\frac{\lambda_j(t)(\lambda_1(t) - \lambda_j(t))}{\lambda_1(t)}\sin^2\theta_t \qquad (42)\\
&\quad + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}}{\boldsymbol{\mu}}\eta^2\sin\theta_t + \frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\eta^3 + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t}\sin\theta_t). \qquad (43)
\end{aligned}
$$

Further, from alignment condition(Equation (28)), we have

$$
\begin{aligned}
\left\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1) - \Phi(x_\eta(t))]\right\|, \left\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\right\|&\\
\le \lambda_1(t)\eta + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2\boldsymbol{\varsigma}}{\boldsymbol{\mu}^2}\eta^2)&\\
+ \mathcal{O}(\sqrt{D}\boldsymbol{\xi}\boldsymbol{\zeta}\boldsymbol{\varsigma}\boldsymbol{\nu}\eta^2) + \mathcal{O}(\eta^2 D)&\\
\le 2\lambda_1(t)\eta,&
\end{aligned}
$$

if $\eta \le \mathcal{O}(\frac{\boldsymbol{\mu}^3}{\boldsymbol{\nu}\boldsymbol{\xi}\boldsymbol{\varsigma}\boldsymbol{\zeta}^2 D})$.

Continuing from Equation (40), we have

$$|\langle v_1(t), x_\eta(t+2) - \Phi(x_\eta(t))\rangle| = |a_{t+1}a_t\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| + \overline{\text{err}}$$
$$\geq (1 + \frac{1}{2}\min_{2\leq j\leq M} \frac{\lambda_j(t)(\lambda_1(t) - \lambda_j(t))}{\lambda_1^2(t)} \sin^2\theta_t)G_t + \widetilde{\text{err}}, \tag{44}$$

where we use Equation (43) and the bounds on $\left\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1) - \Phi(x_\eta(t))]\right\|$, $\left\|P_{t,\Gamma}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\right\|$ in the final step, with

$$|\widetilde{\text{err}}| \leq \overline{\text{err}} + \frac{G_t}{2\eta\lambda_1(t)}\mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}}\eta^2\sin\theta_t + \frac{\boldsymbol{\Upsilon\zeta}^2}{\boldsymbol{\mu}^3}\eta^3 + \frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t}\sin\theta_t)$$
$$\leq \mathcal{O}(\frac{\boldsymbol{\Upsilon\zeta}^2}{\boldsymbol{\mu}^3}\eta^3 + (1+\eta/G_t)\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t}\sin\theta_t).$$

The proof now follows from Equation (44), after taking into account $\|v_1(t) - v_1(t+2)\|$ and $\|\Phi(x_\eta(t+2)) - \Phi(x_\eta(t)\|$. From Lemma B.12, we have $\|\Phi(x_\eta(t)) - \Phi(x_\eta(t+1))\| \leq \mathcal{O}(\boldsymbol{\xi}\eta^2)$, which further implies, $\left\|\nabla^2 L(\Phi(x_\eta(t+1))) - \nabla^2 L(\Phi(x_\eta(t)))\right\| \leq \mathcal{O}(\boldsymbol{\nu\xi}\eta^2)$. Thus, we can use Theorem F.4 to have $\|v_1(t) - v_1(t+1)\| \leq \mathcal{O}(\frac{\boldsymbol{\nu\xi}\eta^2}{\lambda_1(t)-\lambda_2(t)}) = \mathcal{O}(\frac{\boldsymbol{\nu\xi}\eta^2}{\boldsymbol{\Delta}})$. From Lemma B.13, we have $|\langle v_1(t), \Phi(x_\eta(t+1)) - \Phi(x_\eta(t))\rangle| \leq \mathcal{O}(\frac{\boldsymbol{\xi\zeta\nu\chi}\eta^3}{\boldsymbol{\mu}^2})$. Thus,

$$\begin{aligned}
G_{t+2} &= |\langle v_1(t+2), x_\eta(t+2) - \Phi(x_\eta(t+2))\rangle| \\
&= |\langle v_1(t), x_\eta(t+2) - \Phi(x_\eta(t))\rangle| + \mathcal{O}(\frac{\boldsymbol{\nu\xi}\eta^3}{\boldsymbol{\Delta}} + \frac{\boldsymbol{\xi\zeta\nu\chi}\eta^3}{\boldsymbol{\mu}^2\boldsymbol{\Delta}}) \\
&\geq (1 + \frac{1}{2}\min_{2\leq j\leq M} \frac{\lambda_j(t)(\lambda_1(t) + \mathcal{O}(\boldsymbol{\Psi}_{\text{norm}}\eta) - \lambda_j(t))}{\lambda_1^2(t)} \sin^2\theta_t)G_t \\
&+ \mathcal{O}(\frac{\boldsymbol{\Upsilon\zeta}^2}{\boldsymbol{\mu}^3}\eta^3 + (1+\eta/G_t)\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t}\sin\theta_t) + \mathcal{O}(\frac{\boldsymbol{\nu\xi}\eta^3}{\boldsymbol{\Delta}} + \frac{\boldsymbol{\xi\zeta\nu\chi}\eta^3}{\boldsymbol{\mu}^2}) \\
&\geq (1 + \frac{1}{4}\min_{2\leq j\leq M} \frac{\lambda_j(t)(\lambda_1(t) + \mathcal{O}(\boldsymbol{\Psi}_{\text{norm}}\eta) - \lambda_j(t))}{\lambda_1^2(t)} \sin^2\theta_t)G_t \\
&+ \mathcal{O}(\frac{\boldsymbol{\Upsilon\zeta}^2}{\boldsymbol{\mu}^3}\eta^3 + (1+\eta/G_t)\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t}\sin\theta_t) + \mathcal{O}(\frac{\boldsymbol{\nu\xi}\eta^3}{\boldsymbol{\Delta}} + \frac{\boldsymbol{\xi\zeta\nu\chi}\eta^3}{\boldsymbol{\mu}^2}).
\end{aligned}$$

The error term in the pre-final step can be further upper bounded as as $\mathcal{O}(\frac{\boldsymbol{\Upsilon\zeta}^2\boldsymbol{\xi\nu\chi}}{\boldsymbol{\mu}^3\boldsymbol{\Delta}}\eta^3 + (1+\eta/G_t)\frac{\boldsymbol{\nu\zeta}^2\eta^3}{\boldsymbol{\mu}^2\lambda_1(t)G_t}\sin\theta_t)$. The final step follows if $\eta \leq \mathcal{O}(\frac{\boldsymbol{\Delta}}{\boldsymbol{\Psi}_{\text{norm}}})$.

□

E.3.2. Movement along top eigenvector when iterate drops below threshold

In this section, we will show that the projection along the top eigenvector cannot drop below a certain threshold. Formally, we will show the following lemma that predicts the increase in the projection along the top eigenvector in $\mathcal{O}(\log 1/\eta)$ steps, whenever the projection drops below a certain threshold $c_{\text{thres}}(t)$.

Lemma E.11. *Denote $r = \eta^{100}$. For any constant $0 < \boldsymbol{\beta} < \frac{\boldsymbol{\mu\Delta}}{8\boldsymbol{\zeta}^2}$, consider any time step t, with $x_\eta(t) \in Y^{\boldsymbol{\epsilon}}$ and $x_\eta(t)$ satisfying the following:*

1. $\boldsymbol{\beta}\eta \leq |\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| \leq \left(1 - \frac{1}{100M}\right)c_{\text{thres}}(t)\eta$.
2. $|\langle v_i(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| \leq \boldsymbol{\alpha}\eta^2$, *for all* $2 \leq i \leq M$.

Here $c_{\text{thres}}(t)$ is equal to $\max_{k\in[M]} g_t(\lambda_k(t))$ and $\boldsymbol{\alpha} = \Theta(\frac{\boldsymbol{\nu\zeta}^2}{\boldsymbol{\mu}^2\boldsymbol{\beta}})$.

Then, with probability at least $1-\eta^{12}$, the following holds true, after one step of noise perturbation with noise generated from $B_0(r)$ followed by $t_{\text{escape}}+2=\Theta(\log 1/\eta)$ steps of normalized gradient descent ($\bar{t}=t+t_{\text{escape}}+2$):

$$G_{\bar{t}} \geq \left(1+\frac{\boldsymbol{\mu\Delta}}{256\boldsymbol{\zeta}^2}\frac{c_{\text{escape}}^2}{c_{\text{thres}}^2(t)}\right)G_t$$

with $c_{\text{escape}}=\Theta(\frac{\boldsymbol{\beta}^6\boldsymbol{\mu}^6}{\boldsymbol{\zeta}^6\boldsymbol{\nu}^3})$, provided $\eta \leq \widetilde{\mathcal{O}}(\frac{\boldsymbol{\mu}^{10}\boldsymbol{\beta}^9\boldsymbol{\Delta}^{1/2}}{\boldsymbol{\zeta}^{10}\boldsymbol{\nu}^6})$ and $\overline{x_\eta(t')x_\eta(t'+1)} \subset Y^{\boldsymbol{\epsilon}}$ for all time $t \leq t' \leq \bar{t}$. Here $G_t=|\langle v_1(t), x_\eta(t)-\Phi(x_\eta(t))\rangle|$.

Proof. We argue as follows: With a small perturbation to $x_\eta(t)$, the magnitude along the eigenvector $v_k(t)$ is destined to grow to at least $\frac{1}{4}c_{\text{escape}}\eta$ after t_{escape} steps of Normalized GD with high probability (see Lemma E.12). Here, $v_k(t)$ is the eigenvector corresponding to the eigenvalue $\lambda_k(t)=\text{argmax}_{\lambda_i(t)|1\leq i\leq M} g_t(\lambda_i(t))$.

After we have shown that $\left\|P_{t,\Gamma}^{(2:M)}(x_\eta(t)-\Phi(x_\eta(t)))\right\|$ is at least $\Omega(\eta)$ after at most $\mathcal{O}(\log 1/\eta)$ number of steps, we use the behavior of G_t derived in Lemma E.8 to show that

$$G_{t+t_{\text{escape}}+2} \geq \left(1+\frac{\boldsymbol{\mu\Delta}}{128\boldsymbol{\zeta}^2}\frac{c_{\text{escape}}^2}{c_{\text{thres}}^2(t)}\right)G_{t+t_{\text{escape}}},$$

provided $\eta=\mathcal{O}\left(\min\left(\frac{\boldsymbol{\mu}^3\boldsymbol{\Delta}\boldsymbol{\beta}^2 c_{\text{escape}}}{\boldsymbol{\zeta}^4\boldsymbol{\nu}c_{\text{thres}}(t)}, \left(\frac{\boldsymbol{\mu}^4\Delta^2 c_{\text{escape}}^2\boldsymbol{\beta}}{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^4\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}c_{\text{thres}}^2}\right)^{1/2}\right)\right)$.

We will further require the behavior of the projection along the top eigenvector from Corollary E.10 to show that the projection along the top eigenvector can drop only when the angle is $\mathcal{O}(\eta)$, and hence, the drop in the magnitude along the top eigenvector is at-most t_{escape} times $\mathcal{O}(\boldsymbol{\Psi}_{\text{G}}\eta^3)$, where $\boldsymbol{\Psi}_{\text{G}}=\max\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\Delta}, \frac{\boldsymbol{\nu}^2\boldsymbol{\zeta}^4}{\boldsymbol{\mu}^6\boldsymbol{\beta}^5}\right)$. That is

$$G_{t+t_{\text{escape}}} \geq G_t + \mathcal{O}(\boldsymbol{\Psi}_{\text{G}}t_{\text{escape}}\eta^3).$$

The final bound will then follow using $\eta \leq \widetilde{\mathcal{O}}\left(\left(\frac{\boldsymbol{\mu\Delta}}{\boldsymbol{\Psi}_{\text{G}}\boldsymbol{\zeta}^2}\frac{c_{\text{escape}}^2}{c_{\text{thres}}^2(t)}\right)^{1/2}\right)$.

□

Lemma E.12. *Consider any time t, with $x_\eta(t)\in Y^{\boldsymbol{\epsilon}}$. Suppose $x_\eta(t)$ satisfies the conditions in Lemma E.11. The constants $c_{\text{escape}}, c_{\text{thres}}(t), r, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ have been taken from Lemma E.11. Define $\mathcal{X}_{\text{stuck}}$ as the region in $B_{x_\eta(t)}(r)$ such that starting from any point $u\in\mathcal{X}_{\text{stuck}}$, the points $\{u(\widetilde{t})\}_{\widetilde{t}\in[t_{\text{escape}}]}$, with $u(0):=u$, obtained using t_{escape} steps of normalized gd satisfy:*

$$\left\|P_{t,\Gamma}^{(2:M)}(u(\widetilde{t})-\Phi(x_\eta(t)))\right\| \leq \boldsymbol{\alpha}\eta^2, \quad \text{for all } \widetilde{t}\in[t_{\text{escape}}], \tag{45}$$

where $P_{t,\Gamma}^{(2:M)}$ denotes the subspace spanned by $v_2(t),\ldots,v_M(t)$.

Consider two points u and w in $B_{x_\eta(t)}(r)$, with the property $w=u+\eta^{12}rv_k(t)$, where $v_k(t)$ denotes the eigenvector corresponding to the eigenvalue $\lambda_k(t)=\text{argmax}_{\lambda_i(t)|1\leq i\leq M} g_t(\lambda_i(t))$. Then, at least one of u and w is not present in the region $\mathcal{X}_{\text{stuck}}$. Moreover,

$$\left\|P_{t,\Gamma}^{(2:M)}(u(\widetilde{t})-w(\widetilde{t}))\right\| \geq \frac{1}{4}c_{\text{escape}}\eta.$$

Proof. W.l.o.g. we assume u lies in the region $\mathcal{X}_{\text{stuck}}$. Then, consider the two sequences obtained with t_{escape} steps of normalized gd, $\{u(\widetilde{t}), w(\widetilde{t})\}_{\widetilde{t}\in[t_{\text{escape}}]}$:

$$u(0)=u, \quad w(0)=w, \quad u(\widetilde{t})=u(\widetilde{t}-1)-\eta\frac{\nabla L(u(\widetilde{t}))}{\|\nabla L(u(\widetilde{t}))\|}, \quad w(\widetilde{t})=w(\widetilde{t}-1)-\eta\frac{\nabla L(w(\widetilde{t}))}{\|\nabla L(w(\widetilde{t}))\|}.$$

We will show the following:

$$\left\| P_{t,\Gamma}^{(2:M)}(u(t_{\text{escape}}) - w(t_{\text{escape}})) \right\| \geq \Omega(\eta).$$

An important claim to note is the following:

Lemma E.13. *Both the trajectories $\{u(\widetilde{t}), w(\widetilde{t})\}_{\widetilde{t} \leq t_{\text{escape}}}$ satisfy a modified version of the alignment condition (Equation (28)), i.e. for all $1 \leq j \leq M$:*

$$\sqrt{\sum_{i=j}^{M} \langle v_i(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))] \rangle^2} \leq \lambda_j(t)\eta + \mathcal{O}(\boldsymbol{\Psi}_{\text{norm}}\eta^2),$$

$$\sqrt{\sum_{i=j}^{M} \langle v_i(t), \nabla^2 L(\Phi(x_\eta(t)))[w(\widetilde{t}) - \Phi(x_\eta(t))] \rangle^2} \leq \lambda_j(t)\eta + \mathcal{O}(\boldsymbol{\Psi}_{\text{norm}}\eta^2).$$

Note that the condition has been slightly changed to use $\{v_i(t)\}$ as reference coordinate system and $\Phi(x_\eta(t))$ as reference point. The above lemma follows from the fact that both $u(0)$ and $w(0)$ are r-close to $x_\eta(t)$, which itself satisfies the alignment condition (Equation (28)). Thus, both $u(0)$ and $w(0)$ initially follow the desired condition. Since, both the trajectories follow Normalized GD updates, the proof will follow from applying the same technique used in the proof of Lemma C.1. Another result to keep in mind is the following modified version of Lemma D.2.

Lemma E.14. *If $\left\|\nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\right\| \leq \eta\frac{\lambda_1(t)}{2} + \boldsymbol{\Psi}_{\text{norm}}\eta^2$,*

$$\begin{aligned}
&\left|v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))(u(\widetilde{t}+1) - \Phi(x_\eta(t)))\right| \\
&\geq (1 + \mathcal{O}(\boldsymbol{\Psi}_{\text{norm}}\eta)) \left|v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))(u(\widetilde{t}) - \Phi(x_\eta(t)))\right| + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}^2}\eta^2).
\end{aligned}$$

Similarly, if $\left\|\nabla^2 L(\Phi(x_\eta(t)))[w(\widetilde{t}) - \Phi(x_\eta(t))]\right\| \leq \eta\frac{\lambda_1(t)}{2} + \boldsymbol{\Psi}_{\text{norm}}\eta^2$,

$$\begin{aligned}
&\left|v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))(w(\widetilde{t}+1) - \Phi(x_\eta(t)))\right| \\
&\geq (1 + \mathcal{O}(\boldsymbol{\Psi}_{\text{norm}}\eta)) \left|v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))(w(\widetilde{t}) - \Phi(x_\eta(t)))\right| + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}^2}\eta^2).
\end{aligned}$$

If $\left\|\nabla^2 L(\Phi(x_\eta(t)))[w(\widetilde{t}) - \Phi(x_\eta(t))]\right\| \leq \eta\frac{\lambda_1(t)}{2} + \boldsymbol{\Psi}_{\text{norm}}\eta^2$, $\left\|\nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\right\| \leq \eta\frac{\lambda_1(t)}{2} + \boldsymbol{\Psi}_{\text{norm}}\eta^2$, and $u(\boldsymbol{\gamma})$ denotes $\boldsymbol{\gamma}u(0) + (1-\boldsymbol{\gamma})w(0)$ for any $\boldsymbol{\gamma} \in [0,1]$,

$$\begin{aligned}
&\left|v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))(F(u(\boldsymbol{\gamma})) - \Phi(x_\eta(t)))\right| \\
&\geq (1 + \mathcal{O}(\boldsymbol{\Psi}_{\text{norm}}\eta)) \left|v_1(t)^\top \nabla^2 L(\Phi(x_\eta(t)))(u(\boldsymbol{\gamma}) - \Phi(x_\eta(t)))\right| + \mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}^2}\eta^2),
\end{aligned}$$

where $F(x) = x - \eta\frac{\nabla L(x)}{\|\nabla L(x)\|}$.

The above lemma uses $\{v_i(t)\}$ as reference coordinate system and $\Phi(x_\eta(t))$ as reference point. The above lemma follows from showcasing Normalized GD updates of $u(\widetilde{t})$ and $w(\widetilde{t})$ as equivalent to the update in a quadratic model, with an additional noise of $\mathcal{O}(\frac{\boldsymbol{\nu\zeta}}{\boldsymbol{\mu}}\eta^2)$, similar to Equation (27).

Continuing with the proof of Lemma E.12, we first consider the behavior of u. Since u stays in the region $\mathcal{X}_{\text{stuck}}$ for t_{escape} steps of normalized gd, we have for any time-step $\widetilde{t}$:

$$\begin{aligned}
&\left\| \frac{u(\widetilde{t}) - \Phi(x_\eta(t))}{\left\|u(\widetilde{t}) - \Phi(x_\eta(t))\right\|} - v_1(t) \right\| \leq \frac{\boldsymbol{\alpha}}{\boldsymbol{\beta}}\eta, \\
\text{or } &\left\| \frac{u(\widetilde{t}) - \Phi(x_\eta(t))}{\left\|u(\widetilde{t}) - \Phi(x_\eta(t))\right\|} + v_1(t) \right\| \leq \frac{\boldsymbol{\alpha}}{\boldsymbol{\beta}}\eta
\end{aligned} \tag{46}$$

Further, applying the same technique from Lemma E.8, we can show that

$$\begin{aligned}&\left|\langle v_1(t), u(\widetilde{t}+2) - \Phi(x_\eta(t))\rangle - \langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right| \\ &\le \mathcal{O}\left(\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\Delta}} + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\frac{\boldsymbol{\alpha}}{\boldsymbol{\beta}^2}\right)\eta^3\right) = \mathcal{O}(\boldsymbol{\Psi}\eta^3),\end{aligned} \tag{47}$$

where we replace $\left(\frac{\boldsymbol{\Upsilon}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}\boldsymbol{\chi}}{\boldsymbol{\mu}^3\boldsymbol{\Delta}} + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3}\frac{\boldsymbol{\alpha}}{\boldsymbol{\beta}^2}\right)$ by $\boldsymbol{\Psi}$ for simplicity of presentation.

Initially, because u was initialized close to $x_\eta(t)$, we must have $|\langle v_1(t), u(0) - \Phi(x_\eta(t))\rangle - \langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| \le r$. Hence, $\left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle - \langle v_1(t), u(0) - \Phi(x_\eta(t))\rangle\right| \le \mathcal{O}(\boldsymbol{\Psi}\eta^3 t_{\text{escape}})$ for all $\widetilde{t} \in [t_{\text{escape}}]$. With $t_{\text{escape}} \sim \mathcal{O}(\log 1/\eta)$ and with $\eta \le \widetilde{\mathcal{O}}((\boldsymbol{\beta}/M\boldsymbol{\Psi})^{1/2})$, we must have

$$\left(1 - \frac{1}{200M}\right) c_{\text{thres}}(t) \ge \left(1 + \frac{1}{200M}\right)|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| \tag{48}$$

$$\begin{aligned}&\ge \left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right| \\ &\ge 0.999\,|\langle v_1(t), u(0) - \Phi(x_\eta(t))\rangle| \ge 0.999(|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| - r) \ge 0.998\boldsymbol{\beta}\eta,\end{aligned} \tag{49}$$

for any $t \le \widetilde{t} \le t + t_{\text{escape}}$.

Now, we consider the behavior of $w(\cdot)$ and $u(\cdot)$. Consider an even time step $0 \le \widetilde{t} \le t_{\text{escape}}$. From the update rule of w and u, we have

$$w(\widetilde{t}+2) - u(\widetilde{t}+2) = F(F(w(\widetilde{t}))) - F(F(u(\widetilde{t}))),$$

where the function $F : \mathbb{R}^D \to \mathbb{R}^D$, $F(v) = v - \eta\frac{\nabla L(v)}{\nabla\|L(v)\|}$ is the one-step update rule of Normalized GD.

Now, we use taylor expansion of F around $u(\widetilde{t})$ to get

$$w(\widetilde{t}+2) - u(\widetilde{t}+2) = F(F(w(\widetilde{t}))) - F(F(u(\widetilde{t}))) = \nabla_{u(\widetilde{t})}F(F(u(\widetilde{t})))(w(\widetilde{t}) - u(\widetilde{t})) + \text{err},$$

where $\|\text{err}\|$ can be bounded as follows:

$$\begin{aligned}&\max_{\boldsymbol{\gamma}\in[0,1]:u(\boldsymbol{\gamma})=\boldsymbol{\gamma}u(\widetilde{t})+(1-\boldsymbol{\gamma})w(\widetilde{t})} \frac{1}{2}\left\|\nabla^2_{u(\boldsymbol{\gamma})}F(F(u(\boldsymbol{\gamma})))\right\|\left\|w(\widetilde{t}) - u(\widetilde{t})\right\|^2 \\ &= \max_{\boldsymbol{\gamma}\in[0,1]:u(\boldsymbol{\gamma})=\boldsymbol{\gamma}u(\widetilde{t})+(1-\boldsymbol{\gamma})w(\widetilde{t})} \frac{1}{2}\left\|\nabla_{u(\boldsymbol{\gamma})}[\nabla_{F(u(\boldsymbol{\gamma}))}F(F(u(\boldsymbol{\gamma})))\nabla_{u(\boldsymbol{\gamma})}F(u(\boldsymbol{\gamma}))]\right\|\left\|w(\widetilde{t}) - u(\widetilde{t})\right\|^2 \\ &\le \max_{\boldsymbol{\gamma}\in[0,1]:u(\boldsymbol{\gamma})=\boldsymbol{\gamma}u(\widetilde{t})+(1-\boldsymbol{\gamma})w(\widetilde{t})} \eta\mathcal{O}\left(\frac{1}{\|\nabla L(u(\boldsymbol{\gamma}))\|^2} + \frac{1}{\|\nabla L(F(u(\boldsymbol{\gamma})))\|^2}\right)\left\|w(\widetilde{t}) - u(\widetilde{t})\right\|^2 \\ &\cdot \max\left(\left\|\partial^2(\nabla L)(u(\boldsymbol{\gamma}))\right\|, \left\|\nabla^2 L(u(\boldsymbol{\gamma}))\right\|^2, \left\|\partial^2(\nabla L)(F(u(\boldsymbol{\gamma})))\right\|, \left\|\nabla^2 L(F(u(\boldsymbol{\gamma})))\right\|^2\right) \\ &\le \max_{\boldsymbol{\gamma}\in[0,1]:u(\boldsymbol{\gamma})=\boldsymbol{\gamma}u(\widetilde{t})+(1-\boldsymbol{\gamma})w(\widetilde{t})} \left(\frac{1}{\|\nabla L(u(\boldsymbol{\gamma}))\|^2} + \frac{1}{\|\nabla L(F(u(\boldsymbol{\gamma})))\|^2}\right)\cdot\boldsymbol{\varrho}\eta\left\|w(\widetilde{t}) - u(\widetilde{t})\right\|^2,\end{aligned}$$

where the constant $\boldsymbol{\varrho} = \mathcal{O}(\boldsymbol{\zeta}^2 + \boldsymbol{\nu})$. Thus we conclude

$$\left\|w(\widetilde{t}+2) - u(\widetilde{t}+2) - H(u(\widetilde{t}))(w(\widetilde{t}) - u(\widetilde{t}))\right\| \le \boldsymbol{\varrho}\eta\mu(\widetilde{t})\left\|w(\widetilde{t}) - u(\widetilde{t})\right\|^2, \tag{50}$$

where $H(u(\widetilde{t}))$ is given by

$$\begin{aligned}&H(u(\widetilde{t})) := \partial F \circ F(u(\widetilde{t})) = \partial F(u(\widetilde{t}+1)))\partial F(u(\widetilde{t})) = A_{\widetilde{t}+1}A_{\widetilde{t}}, \\ &A_{\widetilde{t}} := \partial F(u(\widetilde{t})) = I - \eta\left[I - \frac{\nabla L(u(\widetilde{t}))\nabla L(u(\widetilde{t}))^\top}{\left\|\nabla L(u(\widetilde{t}))\right\|^2}\right]\frac{\nabla^2 L(u(\widetilde{t}))}{\left\|\nabla L(u(\widetilde{t}))\right\|},\end{aligned}$$

and $\mu(\widetilde{t})$ is given by

$$\mu(\widetilde{t}) = \max_{\gamma\in[0,1]:u(\gamma)=\gamma u(\widetilde{t})+(1-\gamma)w(\widetilde{t})} \left(\frac{1}{\|\nabla L(u(\gamma))\|^2} + \frac{1}{\|\nabla L(F(u(\gamma)))\|^2}\right), \tag{51}$$

Now we claim $A_{\widetilde{t}}$ can be approximated as below with $\|B_{\widetilde{t}}\| \le \mathcal{O}(\frac{\nu\zeta^2}{\mu^2\beta^2}\eta)$. Furthermore, $\|A_{\widetilde{t}}\| \le \mathcal{O}(\frac{1}{\beta})$.

$$A_{\widetilde{t}} = I - \eta\left[I - v_1(t)v_1(t)^\top\right]\frac{\nabla^2 L(\Phi(x_\eta(t)))}{\left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\rangle\right|} + B_{\widetilde{t}},$$

The following strategies have been used to obtain the above approximation:

1. $\left\|\nabla^2 L(u(\widetilde{t})) - \nabla^2 L(\Phi(x_\eta(t)))\right\| \le \mathcal{O}(\nu\left\|u(\widetilde{t}) - \Phi(x_\eta(t))\right\|) = \mathcal{O}(c_{\text{thres}}(t)\nu\eta)$. Here, we use $\left\|u(\widetilde{t}) - \Phi(x_\eta(t))\right\| \le 2c_{\text{thres}}(t)\eta$, since $\|u(0) - \Phi(x_\eta(t))\| \le c_{\text{thres}}(t) + r$, and the conditions from Equation (47) and Equation (45) imply that the norm stays below $2c_{\text{thres}}(t)$, if $\eta \le \mathcal{O}(\frac{c_{\text{thres}}(t)}{\beta})$. We can further bound the error by using $c_{\text{thres}}(t) \le \lambda_1(t)\eta$.
2. Using taylor expansion and the bound on $\left\|u(\widetilde{t}) - \Phi(x_\eta(t))\right\|$, $\nabla L(u(\widetilde{t})) = \nabla^2 L(\Phi(x_\eta(\widetilde{t})))(u(\widetilde{t}) - \Phi(x_\eta(t))) + \mathcal{O}(\nu c_{\text{thres}}(t)^2\eta^2)$. Also, this further implies $\left\|\nabla L(u(\widetilde{t}))\right\| \ge \lambda_1(t)\left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right| + \mathcal{O}(\nu c_{\text{thres}}(t)^2\eta^2)$. Using the update from Equation (47) and the bound on t_{escape}, we must have $\left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right| \ge \beta\eta - \mathcal{O}(\Psi\eta^3 t_{\text{escape}}) \ge \frac{1}{2}\beta\eta$ for $\eta \le \widetilde{\mathcal{O}}((\frac{\beta}{\Psi})^{1/2})$. Thus, $\left\|\nabla L(u(\widetilde{t}))\right\| \ge \lambda_1(t)\frac{1}{2}\beta\eta + \mathcal{O}(\nu c_{\text{thres}}(t)^2\eta^2) \ge \Omega(\mu\beta\eta)$, if $\eta \le \mathcal{O}(\frac{\mu\beta}{\nu c_{\text{thres}}(t)^2})$.
3. We use the condition from Equation (46) to show that $\frac{\nabla L(u(\widetilde{t}))}{\|\nabla L(u(\widetilde{t}))\|}\left(\frac{\nabla L(u(\widetilde{t}))}{\|\nabla L(u(\widetilde{t}))\|}\right)^\top = v_1(t)v_1(t)^\top + \mathcal{O}(\frac{\alpha}{\beta}\eta)$.

Similarly, we can show that:

$$A_{\widetilde{t}+1} = I - \eta\left[I - v_1(t)v_1(t)^\top\right]\frac{\nabla^2 L(\Phi(x_\eta(t)))}{\eta\lambda_1(t) - \left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\rangle\right|} + B_{\widetilde{t}+1},$$

with $\left\|A_{\widetilde{t}+1}\right\| \le \mathcal{O}(\frac{1}{\beta})$ and $\left\|B_{\widetilde{t}+1}\right\| \le \mathcal{O}(\frac{\nu\zeta^2}{\mu^2\beta^2}\eta)$.

Consider the following error term,

$$\text{err}(\widetilde{t}) := w(\widetilde{t}+2) - u(\widetilde{t}+2) - \prod_{0\le i\le\widetilde{t}:i\%2=0} H(u(i))(w(0) - u(0)), \tag{52}$$

By Equation (50), the following property holds with function μ defined in Equation (51):

$$\begin{aligned}&\left\|\text{err}(\widetilde{t}) - H(u(\widetilde{t}))\text{err}(\widetilde{t}-2)\right\| \le \varrho\eta\mu(\widetilde{t})\left\|w(\widetilde{t}) - u(\widetilde{t})\right\|^2 \text{ for all } \widetilde{t} \ge 0,\\ &\text{err}(-2) = 0,\end{aligned}$$

Finally, we use Lemma E.15 and Lemma E.16 to handle the main and error terms in Equation (52),

$$\begin{aligned}|\langle v_k(t), w(t_{\text{escape}}) - u(t_{\text{escape}})\rangle| &= \left|v_k(t)^\top \prod_{0\le\widetilde{t}\le t_{\text{escape}}:\widetilde{t}\%2=0} H(u(\widetilde{t}))(w(0) - u(0)) + v_k(t)^\top\text{err}(t_{\text{escape}})\right|\\ &\ge \left|v_k(t)^\top \prod_{0\le\widetilde{t}\le t_{\text{escape}}:\widetilde{t}\%2=0} H(u(\widetilde{t}))(w(0) - u(0))\right| - \|\text{err}(t_{\text{escape}})\|\\ &\ge \frac{1}{4}c_{\text{escape}}\eta.\end{aligned}$$

which completes the proof of Lemma E.12. □

Lemma E.15.

$$\left| v_k(t)^\top \prod_{0 \le \widetilde{t} \le t_{\text{escape}} : \widetilde{t}\%2=0} H(u(\widetilde{t}))(w(0)-u(0)) \right| \ge \frac{1}{2} c_{\text{escape}} \eta.$$

Lemma E.16.

$$\|\text{err}(t_{\text{escape}})\| \le \frac{1}{4} c_{\text{escape}} \eta.$$

Proof of Lemma E.15. Recall that from Lemma E.12 where we left off,

1. $x_\eta(t)$ satisfies (a) $\boldsymbol{\beta}\eta \le |\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| \le \left(1 - \frac{1}{100M}\right) c_{\text{thres}}(t)\eta$, and (b) $|\langle v_i(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| \le \boldsymbol{\alpha}\eta^2$, for all $2 \le i \le M$.
2. $v_k(t)$ denotes the eigenvector corresponding to the eigenvalue $\lambda_k(t) = \text{argmax}_{\lambda_i(t) | 1 \le i \le M} g_t(\lambda_i(t))$.
3. Initial condition between w and u is given by $w(0) - u(0) = \eta^3 r v_k(t)$.
4. From Equation (47), $\left|\langle v_1(t), u(\widetilde{t}+2) - u(\widetilde{t})\rangle\right| \le \mathcal{O}(\boldsymbol{\Psi}\eta^3) \quad$ for all $\widetilde{t} \in [t_{\text{escape}}]$.
5. The function H is defined by:

$$H(u(\widetilde{t})) = A_{\widetilde{t}+1} A_{\widetilde{t}}$$
$$A_{\widetilde{t}} = I - \eta \left[I - \frac{\nabla L(u(\widetilde{t})) \nabla L(u(\widetilde{t}))^\top}{\left\|\nabla L(u(\widetilde{t}))\right\|^2} \right] \frac{\nabla^2 L(u(\widetilde{t}))}{\left\|\nabla L(u(\widetilde{t}))\right\|}.$$

6. $A_{\widetilde{t}}$ was further simplified as,

$$A_{\widetilde{t}} = I - \eta \left[I - v_1(t) v_1(t)^\top \right] \frac{\nabla^2 L(\Phi(x_\eta(t)))}{\left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\rangle\right|} + B_{\widetilde{t}}$$
$$A_{\widetilde{t}+1} = I - \eta \left[I - v_1(t) v_1(t)^\top \right] \frac{\nabla^2 L(\Phi(x_\eta(t)))}{\eta\lambda_1(t) - \left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\rangle\right|} + B_{\widetilde{t}+1},$$

with $\|B_{\widetilde{t}}\|, \left\|B_{\widetilde{t}+1}\right\| \le \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^2\boldsymbol{\beta}^2}\eta)$. Further, we showed that $\|A_{\widetilde{t}}\|, \left\|A_{\widetilde{t}+1}\right\| \le \mathcal{O}(\frac{1}{\boldsymbol{\beta}})$.

Thus, the term under consideration can be simplified as follows,

$$\begin{aligned}
&\prod_{0 \le \widetilde{t} \le t_{\text{escape}} : \widetilde{t}\%2=0} H(u(\widetilde{t}))(w(0)-u(0)) \\
&= \eta^3 r \prod_{0 \le \widetilde{t} \le t_{\text{escape}} : \widetilde{t}\%2=0} H(u(\widetilde{t})) v_k(t) \\
&= \eta^3 r \prod_{0 \le \widetilde{t} \le t_{\text{escape}} : \widetilde{t}\%2=0} A_{\widetilde{t}+1} A_{\widetilde{t}} v_k(t) \\
&= \eta^3 r \\
&\prod_{0 \le \widetilde{t} \le t_{\text{escape}} : \widetilde{t}\%2=0} \left[I - \eta \left[I - v_1(t) v_1(t)^\top \right] \frac{\nabla^2 L(\Phi(x_\eta(t)))}{\eta\lambda_1(t) - \left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\rangle\right|} + B_{\widetilde{t}+1} \right] \\
&\qquad \cdot \left[I - \eta \left[I - v_1(t) v_1(t)^\top \right] \frac{\nabla^2 L(\Phi(x_\eta(t)))}{\left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\rangle\right|} + B_{\widetilde{t}} \right] v_k(t) \\
&= \eta^3 r \prod_{0 \le \widetilde{t} \le t_{\text{escape}} : \widetilde{t}\%2=0} M_{\widetilde{t}} v_k(t) + \text{rem},
\end{aligned}$$

where using the bounds on $\{A_{\widetilde{t}}, A_{\widetilde{t}+1}, B_{\widetilde{t}}, B_{\widetilde{t}+1}\}_{0\le\widetilde{t}\le t_{\text{escape}}}$, we have

$$\begin{aligned}
&\|\text{rem}\| \\
&\le \max_{\widetilde{t}\le t_{\text{escape}}} \left(\|B_{\widetilde{t}}\| + \left\|B_{\widetilde{t}+1}\right\|\right) \cdot \max_{\widetilde{t}\le t_{\text{escape}}} \left(\|A_{\widetilde{t}}\| + \left\|A_{\widetilde{t}+1}\right\|\right) \cdot \Big\| \sum_{0\le\widetilde{t}\le t_{\text{escape}}:\widetilde{t}\%2=0} \quad \prod_{0\le j\le t_{\text{escape}}:j\%2=0, j\neq\widetilde{t}} M_j \Big\| \\
&\le \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta^4 r) \cdot \Big\| \sum_{0\le\widetilde{t}\le t_{\text{escape}}:\widetilde{t}\%2=0} \quad \prod_{0\le j\le t_{\text{escape}}:j\%2=0, j\neq\widetilde{t}} M_j \Big\|. \qquad (53)
\end{aligned}$$

For simplicity of presentation, we have used $M_{\widetilde{t}}$ to define

$$\left[I - \eta\left[I - v_1(t)v_1(t)^\top\right] \frac{\nabla^2 L(\Phi(x_\eta(t)))}{\eta\lambda_1(t) - \left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\rangle\right|}\right]\left[I - \eta\left[I - v_1(T)v_1(t)^\top\right] \frac{\nabla^2 L(\Phi(x_\eta(t)))}{\left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\rangle\right|}\right].$$

Initially, $|\langle v_1(t), u(0) - \Phi(x_\eta(t))\rangle - \langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| \le r$. Also, we have

$$\left|\langle v_1(t), u(\widetilde{t}+2) - u(\widetilde{t})\rangle\right| \le \mathcal{O}(\boldsymbol{\Psi}\eta^3) \quad \text{for all } \widetilde{t} \in [t_{\text{escape}}].$$

From the behavior of $u(\widetilde{t})$ from Equation (49), we have $\left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right| \le \left(1 - \frac{1}{200M}\right) c_{\text{thres}}(t)\eta$. Recall that $c_{\text{thres}}(t)$ was chosen as $\max_{1\le k\le M} g_t(\lambda_k(t))$. We will showcase below that for the chosen upper bound of $c_{\text{thres}}(t)$, $v_k(t)$ acts as the top eigenvector of $M_{\widetilde{t}}$ for any $\widetilde{t} \le t_{\text{escape}}$. For all $j \in [2, M]$ and $\widetilde{t} \in [t_{\text{escape}}]$, we have:

$$M_{\widetilde{t}}v_j(t) = \left[1 - \eta\frac{\lambda_j(t)/\lambda_1(t)}{\eta - \left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right|}\right]\left[1 - \eta\frac{\lambda_j(t)/\lambda_1(t)}{\left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right|}\right] v_j(t),$$

with $M_{\widetilde{t}}v_1(t) = v_1(t)$. When $\left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right| \le c_{\text{thres}}(t)$, $M_{\widetilde{t}}v_j(t) \ge M_{\widetilde{t}}v_1(t)$, for all $j \ge 2$. Furthermore, $M_{\widetilde{t}}v_j(t)$ maximizes when $j = k$. Thus, the value of $c_{\text{thres}}(t)$ has been strategically chosen to make $v_k(t)$ the maximum eigenvector for the matrices $\{M_{\widetilde{t}}\}_{0\le\widetilde{t}\le t_{\text{escape}}}$.

Furthermore, with $v_k(t)$ the top eigenvector of $M_{\widetilde{t}}$, we can show

$$\begin{aligned}
\|M_{\widetilde{t}}\| &= \left|\left[1 - \eta\frac{\lambda_k(t)/\lambda_1(t)}{\eta - \left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right|}\right]\left[1 - \eta\frac{\lambda_k(t)/\lambda_1(t)}{\left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right|}\right]\right| \\
&\ge \left|\left[1 - \frac{\lambda_k(t)/\lambda_1(t)}{1 - \left(1 - \frac{1}{200M}\right) c_{\text{thres}}(t)}\right]\left[1 - \frac{\lambda_k(t)/\lambda_1(t)}{\left(1 - \frac{1}{200M}\right) c_{\text{thres}}(t)}\right]\right|, \quad \text{for all } \widetilde{t} \in [t_{\text{escape}}],
\end{aligned}$$

since we showed before that $\left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right| \le \left(1 - \frac{1}{200M}\right) c_{\text{thres}}(t)\eta$, .

Now, we explain our choice of t_{escape}. We select t_{escape} s.t.

$$\left|\left\langle v_k(t), \eta^3 r \prod_{0\le\widetilde{t}\le t_{\text{escape}}:\widetilde{t}\%2=0} M_{\widetilde{t}}v_k(t)\right\rangle\right| = c_{\text{escape}}\eta.$$

That is, we select the time step $\widetilde{t}$, where the magnitude of the useful term $\prod_{0\le\widetilde{t}\le t_{\text{escape}}:\widetilde{t}\%2=0} M_{\widetilde{t}}v_k(t)$ along the eigenvector $v_k(t)$ reaches $c_{\text{escape}}\eta$. With $c_{\text{thres}}(t) = g_t(\lambda_k(t))$, we have $\left|\left[1 - \frac{\lambda_k(t)/\lambda_1(t)}{1-\left(1-\frac{1}{200M}\right)c_{\text{thres}}(t)}\right]\left[1 - \frac{\lambda_k(t)/\lambda_1(t)}{\left(1-\frac{1}{200M}\right)c_{\text{thres}}(t)}\right]\right| \ge 1.001$ and so, we just need $t_{\text{escape}} \le \mathcal{O}(\log(c_{\text{escape}}/\eta))$.

With this choice of t_{escape}, we must have from Equation (53),

$$\|\text{rem}\| \le \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta^4 r t_{\text{escape}}) \cdot \frac{c_{\text{escape}}}{\eta^2 r} \le \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3} t_{\text{escape}} c_{\text{escape}}\eta^2),$$

with

$$\left|\left\langle v_k(t), \prod_{0\le\widetilde{t}\le t_{\text{escape}}:\widetilde{t}\%2=0} H(u(\widetilde{t}))(w(0)-u(0))\right\rangle\right|$$
$$\ge c_{\text{escape}}\eta + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3} t_{\text{escape}} c_{\text{escape}}\eta^2) \ge \frac{1}{2}c_{\text{escape}}\eta,$$

where the last step follows if $\eta \le \widetilde{\mathcal{O}}(\frac{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}{\boldsymbol{\nu}\boldsymbol{\zeta}^2})$. Thus, we have shown that with the appropriate choice of t_{escape}, the magnitude of $\prod_{0\le\widetilde{t}\le t_{\text{escape}}:\widetilde{t}\%2=0} H(u(\widetilde{t}))(w(0)-u(0))$ can reach at least $\frac{1}{2}c_{\text{escape}}\eta$ along the eigenvector $v_k(t)$.

□

Proof of Lemma E.16. Recall that from Lemma E.12 where we left off,

1. $x_\eta(t)$ satisfies (a) $\boldsymbol{\beta}\eta \le |\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| \le \left(1 - \frac{1}{100M}\right) c_{\text{thres}}(t)\eta$, and (b) $|\langle v_i(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| \le \boldsymbol{\alpha}\eta^2$, for all $2 \le i \le M$.
2. $v_k(t)$ denotes the eigenvector corresponding to the eigenvalue $\lambda_k(t) = \text{argmax}_{\lambda_i(t)|1\le i\le M}\, g_t(\lambda_i(t))$.
3. Initial condition between w and u is given by $w(0) - u(0) = \eta^3 r v_k(t)$, where r denotes the magnitude of the noise.
4. The difference between $w(\widetilde{t})$ and $u(\widetilde{t})$ changes as follows (Equation (50)):
$$\left\|w(\widetilde{t}+2) - u(\widetilde{t}+2) - H(u(\widetilde{t}))(w(\widetilde{t}) - u(\widetilde{t}))\right\| \le \boldsymbol{\varrho}\eta\mu(\widetilde{t}) \left\|w(\widetilde{t}) - u(\widetilde{t})\right\|^2. \tag{54}$$
5. The difference between $\text{err}(\widetilde{t})$ and $H(u(\widetilde{t}))\text{err}(\widetilde{t}-2)$ is given by :
$$\left\|\text{err}(\widetilde{t}) - H(u(\widetilde{t}))\text{err}(\widetilde{t}-2)\right\| \le \boldsymbol{\varrho}\eta\mu(\widetilde{t}) \left\|w(\widetilde{t}) - u(\widetilde{t})\right\|^2 \text{ for all } \widetilde{t} \ge 0,$$
$$\text{err}(-2) = 0,$$
with function μ defined in Equation (51):
$$\mu(\widetilde{t}) = \max_{\boldsymbol{\gamma}\in[0,1]:u(\boldsymbol{\gamma})=\boldsymbol{\gamma}u(\widetilde{t})+(1-\boldsymbol{\gamma})w(\widetilde{t})} \left(\frac{1}{\|\nabla L(u(\boldsymbol{\gamma}))\|^2} + \frac{1}{\|\nabla L(F(u(\boldsymbol{\gamma})))\|^2}\right),$$
where the function F was defined as $F(x) = x - \eta\frac{\nabla L(x)}{\|\nabla L(x)\|}$.
6. From Equation (47), the update of $u(\widetilde{t})$ along the top eigenvector $v_1(t)$ is given by:
$$\left|\langle v_1(t), u(\widetilde{t}+2) - u(\widetilde{t})\rangle\right| \le \mathcal{O}(\boldsymbol{\Psi}\eta^3) \quad \text{for all } \widetilde{t} \in [t_{\text{escape}}],$$
for $\widetilde{t} \le t_{\text{escape}}$.
7. The function H is defined by:
$$H(u(\widetilde{t})) = A_{\widetilde{t}+1}A_{\widetilde{t}}$$
$$A_{\widetilde{t}} = I - \eta\left[I - \frac{\nabla L(u(\widetilde{t}))\nabla L(u(\widetilde{t}))^\top}{\left\|\nabla L(u(\widetilde{t}))\right\|^2}\right]\frac{\nabla^2 L(u(\widetilde{t}))}{\left\|\nabla L(u(\widetilde{t}))\right\|}.$$
8. $A_{\widetilde{t}}$ was further simplified as,
$$A_{\widetilde{t}} = I - \eta\left[I - v_1(t)v_1(t)^\top\right]\frac{\nabla^2 L(\Phi(x_\eta(t)))}{\left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\rangle\right|} + B_{\widetilde{t}}$$
$$A_{\widetilde{t}+1} = I - \eta\left[I - v_1(t)v_1(t)^\top\right]\frac{\nabla^2 L(\Phi(x_\eta(t)))}{\eta\lambda_1(t) - \left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\widetilde{t}) - \Phi(x_\eta(t))]\rangle\right|} + B_{\widetilde{t}+1},$$
with $\|B_{\widetilde{t}}\|, \left\|B_{\widetilde{t}+1}\right\| \le \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^2\boldsymbol{\beta}^2}\eta)$. Further, we showed that $\|A_{\widetilde{t}}\|, \left\|A_{\widetilde{t}+1}\right\| \le \mathcal{O}(\frac{1}{\boldsymbol{\beta}})$.

To proceed ahead, we will need few results from the proof of Lemma E.15.

1. The function H can be further simplified for all $u(\widetilde{t})$ as $H(u(\widetilde{t})) = M_{\widetilde{t}} + C_{\widetilde{t}}$, where $\|C_{\widetilde{t}}\| = \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^3}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta)$ and $\|M_{\widetilde{t}}\| = \left|\left[1-\eta\frac{\lambda_k(t)/\lambda_1(t)}{\eta-\left|\langle v_1(t),u(\widetilde{t})-\Phi(x_\eta(t))\rangle\right|}\right]\left[1-\eta\frac{\lambda_k(t)/\lambda_1(t)}{\left|\langle v_1(t),u(\widetilde{t})-\Phi(x_\eta(t))\rangle\right|}\right]\right|$. Further, $\|M_{\widetilde{t}}\| = \max_{j\le t_{\text{escape}}}\|M_j\| + \mathcal{O}(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log 1/\eta)$ for all $\widetilde{t}\le t_{\text{escape}}$.
2. $v_k(t)$ is the top eigenvector of $M_{\widetilde{t}}$.
3. t_{escape} is set such that
$$\left|\left\langle v_k(t), \eta^3 r \prod_{0\le\widetilde{t}\le t_{\text{escape}}:\widetilde{t}\%2=0} M_{\widetilde{t}} v_k(t)\right\rangle\right| = c_{\text{escape}}\eta.$$
Also, we only need $t_{\text{escape}} = \mathcal{O}(\log 1/\eta)$.

We will use an induction procedure to bound $\text{err}(t_{\text{escape}})$. Denote by $\boldsymbol{\varphi} = \max_{\widetilde{t}\le t_{\text{escape}}}\|M_{\widetilde{t}}\|$. Then, from the results listed above, we can show that

$$\left\|\prod_{0\le i\le\widetilde{t}:i\%2=0} M_i\right\| = \prod_{0\le i\le\widetilde{t}:i\%2=0}\|M_i\|$$
$$= \prod_{0\le i\le\widetilde{t}:i\%2=0}\max_{j\le t_{\text{escape}}}\|M_i\| + \mathcal{O}(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}t_{\text{escape}}\eta^2\log 1/\eta) = \boldsymbol{\varphi}^{\widetilde{t}/2} + \mathcal{O}(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log^2 1/\eta).$$

With t_{escape} set such that $\left\|\prod_{0\le i\le t_{\text{escape}}:i\%2=0} M_i\right\| = \frac{c_{\text{escape}}}{\eta^2 r}$, we must have

$$\boldsymbol{\varphi}^{t_{\text{escape}}/2} \le \frac{c_{\text{escape}}}{\eta^2 r} + \mathcal{O}(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log^2 1/\eta). \tag{55}$$

Hence,

$$\left\|H(u(\widetilde{t}))\right\| = \boldsymbol{\varphi}^{\widetilde{t}/2} + \mathcal{O}(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log^2 1/\eta) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^3}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta). \tag{56}$$

The induction hypothesis is as follows for all $\widetilde{t}\le t_{\text{escape}}$:

1. $\mu(\widetilde{t}) \le \frac{4}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2\eta^2}$.
2. $\left\|\text{err}(\widetilde{t})\right\| \le \boldsymbol{\varepsilon}^{\widetilde{t}}\left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\right)^3\frac{1}{\eta}\left\|w(0)-u(0)\right\|^2$.
3. $\frac{1}{\eta}\left\|w(\widetilde{t}+2)-u(\widetilde{t}+2)\right\| \le \frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\boldsymbol{\varepsilon}^{\widetilde{t}/2}\frac{1}{\eta}\left\|w(0)-u(0)\right\|$, where $\boldsymbol{\varepsilon}$ is given by $\boldsymbol{\varphi} + \mathcal{O}(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log^2 1/\eta) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^3}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta)$.

Before proving the above hypothesis, we will first look at the behavior of the term $\left|\langle v_1(t), u(\widetilde{t})-\Phi(x_\eta(t))\rangle\right|$. From Equation (49), we have

$$\left(1-\frac{1}{200M}\right)c_{\text{thres}}(t)\eta \ge \left|\langle v_1(t), u(\widetilde{t})-\Phi(x_\eta(t))\rangle\right| \ge 0.998\boldsymbol{\beta}\eta.$$

Since, u is in $\mathcal{X}_{\text{stuck}}$, we have $\left\|P^{(2:M)}_{t,\Gamma}(u(\widetilde{t})-\Phi(x_\eta(t)))\right\| \le \boldsymbol{\alpha}\eta^2$. Carefully choosing $\eta \le \mathcal{O}(\frac{\boldsymbol{\beta}}{\boldsymbol{\alpha}})$, we can bound $\left\|u(\widetilde{t})-\Phi(x_\eta(t))\right\| \le 1.005c_{\text{thres}}(t)\eta$.

We can verify that the claims hold true for $\widetilde{t} = -2$. Now, we prove the induction hypothesis as follows:

1. Suppose the conditions hold true for all $i \le \widetilde{t}$, where $\widetilde{t} \le t_{\text{escape}}$. Since, $\left\|\text{err}(\widetilde{t})\right\| \le \boldsymbol{\varepsilon}^{\widetilde{t}} \left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\right)^3 \frac{1}{\eta} \left\|w(0) - u(0)\right\|^2$, with the following set of computation, we can show that $\left\|w(\widetilde{t}) - u(\widetilde{t})\right\| \le \frac{1}{2}\beta\eta$.

$$\begin{aligned}
\left\|w(\widetilde{t}) - u(\widetilde{t})\right\| &\le \frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\boldsymbol{\varepsilon}^{\widetilde{t}/2} \left\|w(0) - u(0)\right\| \\
&\le \frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\boldsymbol{\varphi}^{\widetilde{t}/2} \left(1 + \widetilde{t}\mathcal{O}\left(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log^2 1/\eta + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^3}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta\right)\right) \left\|w(0) - u(0)\right\| \qquad (57)\\
&\le \frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\left(\frac{c_{\text{escape}}}{\eta^2 r} + \mathcal{O}(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log^2 1/\eta)\right) \\
&\quad \left(1 + \widetilde{t}\mathcal{O}\left(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log^2 1/\eta + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^3}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta\right)\right) \left\|w(0) - u(0)\right\| \qquad (58)\\
&\le \frac{16\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2} c_{\text{escape}}\eta \qquad (59)\\
&\le \frac{1}{10}\boldsymbol{\beta}\eta. \qquad (60)
\end{aligned}$$

Here, in Equation (57), we have used the relation derived before: $\boldsymbol{\varepsilon} = \boldsymbol{\varphi} + \mathcal{O}\left(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log^2 1/\eta + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^3}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta\right)$. In Equation (58), we have used $\widetilde{t} \le t_{\text{escape}}$, and the bound on $\boldsymbol{\varphi}$ from Equation (55). In Equation (59), we simplify using $\eta \le \widetilde{\mathcal{O}}(\min((\frac{\boldsymbol{\beta}^3}{\boldsymbol{\Psi}})^{0.5}, \frac{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}{\boldsymbol{\nu}\boldsymbol{\zeta}^3}))$. The final step (Equation (60)) justifies the bound on $c_{\text{escape}} = \mathcal{O}\left(\frac{\boldsymbol{\beta}^3\boldsymbol{\mu}^2}{\boldsymbol{\varrho}}\right)$.

This further implies,

$$\left|\langle u(\widetilde{t}) - w(\widetilde{t}), v_1(t)\rangle\right| \le \left\|u(\widetilde{t}) - w(\widetilde{t})\right\| \le 0.1\boldsymbol{\beta}\eta.$$

From the update rule for $u(\widetilde{t})$ mentioned in Equation (49), we have $\left|\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right| \ge 0.998\boldsymbol{\beta}\eta$. Thus, we can deduce that

$$\begin{aligned}
&\left|\langle v_1(t), w(\widetilde{t}) - \Phi(x_\eta(t))\rangle\right| \ge 0.5\boldsymbol{\beta}\eta, \qquad (61)\\
&\text{sign}(\langle v_1(t), u(\widetilde{t}) - \Phi(x_\eta(t))\rangle) = \text{sign}(\langle v_1(t), w(\widetilde{t}) - \Phi(x_\eta(t))\rangle). \qquad (62)
\end{aligned}$$

If $u(\boldsymbol{\gamma})$ denotes $\lambda u(\widetilde{t}) + (1-\lambda)w(\widetilde{t})$ for any $\lambda \in [0,1]$, then we must have $|\langle v_1(t), u(\lambda) - \Phi(x_\eta(t))\rangle| \ge 0.5\boldsymbol{\beta}\eta$. Using taylor expansion: $\nabla L(u(\boldsymbol{\gamma})) = \nabla^2 L(\Phi(x_\eta(t)))(u(\boldsymbol{\gamma}) - \Phi(x_\eta(t))) + \mathcal{O}(\boldsymbol{\nu}\left\|u(\boldsymbol{\gamma}) - \Phi(x_\eta(t))\right\|^2)$ and hence, we must have $\left\|\nabla L(u(\boldsymbol{\gamma}))\right\| \ge \lambda_1(t)\left|\langle v_1(t), u(\boldsymbol{\gamma}) - \Phi(x_\eta(t))\rangle\right| + \mathcal{O}(\boldsymbol{\nu}\left\|u(\boldsymbol{\gamma}) - \Phi(x_\eta(t))\right\|^2) \ge \frac{1}{2}\lambda_1(t)\boldsymbol{\beta}\eta$.

With $\left\|u(\widetilde{t}) - \Phi(x_\eta(t))\right\| \le 1.05c_{\text{thres}}(t)\eta$ and $\left\|w(0) - \Phi(x_\eta(t))\right\| \le \left\|u(\widetilde{t}) - \Phi(x_\eta(t))\right\| + \left\|u(\widetilde{t}) - w(\widetilde{t})\right\| \le 1.15c_{\text{thres}}(t)\eta$, we can apply Lemma E.14 to show $\left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[F(u(\boldsymbol{\gamma})) - \Phi(x_\eta(t))]\rangle\right| \ge \left|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))[u(\boldsymbol{\gamma}) - \Phi(x_\eta(t))]\rangle\right| + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}}{\boldsymbol{\mu}}\eta^2)$. That implies, $\left\|\nabla L(F(u(\boldsymbol{\gamma})))\right\| \ge \frac{1}{2}\lambda_1(t)\boldsymbol{\beta}\eta$.

Hence, $\mu(\widetilde{t}) = \max_{\boldsymbol{\gamma}} \frac{1}{\left\|\nabla L(u(\boldsymbol{\gamma}))\right\|^2} + \frac{1}{\left\|\nabla L(F(u(\boldsymbol{\gamma})))\right\|^2} \le \frac{4}{\lambda_1^2(t)\boldsymbol{\beta}^2\eta^2} \le \frac{4}{\boldsymbol{\mu}^2\boldsymbol{\beta}^2\eta^2}$.

2. The error term $\text{err}(\widetilde{t})$ is recursively defined as

$$\begin{aligned}
&\left\|\text{err}(\widetilde{t}) - H(u(\widetilde{t}))\text{err}(\widetilde{t} - 2)\right\| \le \boldsymbol{\varrho}\eta\mu(\widetilde{t})\left\|w(\widetilde{t}) - u(\widetilde{t})\right\|^2 \text{ for all } \widetilde{t} \ge 0,\\
&\text{err}(-2) = 0.
\end{aligned}$$

The norm on $\text{err}(\widetilde{t})$ can be recursively bounded as :

$$\left\|\text{err}(\widetilde{t})\right\| \le \sum_{j=0}^{\widetilde{t}} \prod_{j < i \le \widetilde{t}: i\%2=0} \left\|H(u(i))\right\| \boldsymbol{\varrho}\eta\mu(j)\left\|w(j) - u(j)\right\|^2 \qquad (63)$$

$$\leq \sum_{j=0}^{\widetilde{t}} \prod_{j<i\leq\widetilde{t}:i\%2=0} \|H(u(i))\| \frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2\eta} \|w(j)-u(j)\|^2 \tag{64}$$

$$\leq \sum_{j=0}^{\widetilde{t}} \prod_{j<i\leq\widetilde{t}:i\%2=0} \|H(u(i))\| \frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2\eta} \cdot \left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\boldsymbol{\varepsilon}^{j/2}\right)^2 \|w(0)-u(0)\|^2 \tag{65}$$

$$\leq \sum_{j=0}^{\widetilde{t}} \boldsymbol{\varepsilon}^{\widetilde{t}/2-j/2} \frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2\eta} \cdot \left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\boldsymbol{\varepsilon}^{j/2}\right)^2 \|w(0)-u(0)\|^2 \tag{66}$$

$$\leq \sum_{j=0}^{\widetilde{t}} \boldsymbol{\varepsilon}^{\widetilde{t}/2+j/2} \left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\right)^3 \frac{1}{\eta} \|w(0)-u(0)\|^2 \tag{67}$$

$$\leq \boldsymbol{\varepsilon}^{\widetilde{t}} \left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\right)^3 \frac{1}{\eta} \|w(0)-u(0)\|^2 \tag{68}$$

The following steps have been followed in the previous set of computations:

(a) In Equation (64) and Equation (65), we use the values of $\mu(j)$ and $\|w(j)-u(j)\|$ derived using induction hypothesis for $j \leq \widetilde{t}$.

(b) In Equation (66), we replace $\|H(u(i))\|$ by $\boldsymbol{\varepsilon}$, which is equal to $\boldsymbol{\varphi} + \mathcal{O}(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log^2 1/\eta) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^3}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta)$.

3. We have from Equation (54),

$$\begin{aligned}\frac{1}{\eta}\left\|w(\widetilde{t}+2)-u(\widetilde{t}+2)\right\| &\leq \frac{1}{\eta}\left\|H(u(\widetilde{t}))\right\|\left\|w(\widetilde{t})-u(\widetilde{t})\right\| + \boldsymbol{\varrho}\mu(\widetilde{t})\left\|w(\widetilde{t})-u(\widetilde{t})\right\|^2 \\ &\leq \frac{1}{\eta}\boldsymbol{\varepsilon}\|w(0)-u(0)\| + \frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2\eta^2}\left\|w(\widetilde{t})-u(\widetilde{t})\right\|.\end{aligned}$$

Since, the above inequality holds true for all time-steps $i \leq \widetilde{t}$, we can use gronwall's inequality (Bihari, 1956) and show that

$$\begin{aligned}\frac{1}{\eta}\left\|w(\widetilde{t}+2)-u(\widetilde{t}+2)\right\| &\leq \frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\frac{(\boldsymbol{\varepsilon}^{\widetilde{t}+2}-\boldsymbol{\varepsilon}^{\widetilde{t}/2+1})\frac{1}{\eta}\|w(0)-u(0)\|}{(\boldsymbol{\varepsilon}^{\widetilde{t}+2}-\boldsymbol{\varepsilon}^{\widetilde{t}/2+1})+\boldsymbol{\varepsilon}^{\widetilde{t}/2+1}\frac{1}{\eta}\|w(0)-u(0)\|} \\ &\leq \frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\boldsymbol{\varepsilon}^{\widetilde{t}/2+1}\frac{1}{\eta}\|w(0)-u(0)\|.\end{aligned}$$

Here, $\boldsymbol{\varepsilon}$ is given by $\boldsymbol{\varphi} + \mathcal{O}(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2\log^2 1/\eta) + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^3}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta)$.

The proof will follow by bounding $\|\mathrm{err}(t_{\mathrm{escape}})\|$. The steps are similar to the ones used in Equation (57)-Equation (60).

$$
\begin{aligned}
\|\mathrm{err}(t_{\mathrm{escape}})\| &\le \boldsymbol{\varepsilon}^{t_{\mathrm{escape}}} \left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\right)^3 \frac{1}{\eta} \|w(0)-u(0)\|^2 \\
&\le \boldsymbol{\varepsilon}^{t_{\mathrm{escape}}} \left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\right)^3 \frac{1}{\eta} \|w(0)-u(0)\|^2 \\
&\le \boldsymbol{\varphi}^{t_{\mathrm{escape}}} \left(1 + t_{\mathrm{escape}}\mathcal{O}\left(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2 \log^2 1/\eta + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^3}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta\right)\right) \left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\right)^3 \frac{1}{\eta}(\eta^3 r)^2 \\
&\le \left(\frac{c_{\mathrm{escape}}}{\eta^2 r} + \mathcal{O}(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2 \log^2 1/\eta)\right)^2 \\
&\left(1 + t_{\mathrm{escape}}\mathcal{O}\left(\frac{\boldsymbol{\Psi}}{\boldsymbol{\beta}^3}\eta^2 \log^2 1/\eta + \frac{\boldsymbol{\nu}\boldsymbol{\zeta}^3}{\boldsymbol{\mu}^2\boldsymbol{\beta}^3}\eta\right)\right) \left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\right)^3 \frac{1}{\eta}(\eta^3 r)^2 \\
&\le 8c_{\mathrm{escape}}^2 \left(\frac{4\boldsymbol{\varrho}}{\boldsymbol{\beta}^2\boldsymbol{\mu}^2}\right)^3 \eta. \\
&\le \frac{1}{4} c_{\mathrm{escape}} \eta,
\end{aligned}
$$

where in the final step, we use the bound on $c_{\mathrm{escape}} = \mathcal{O}(\frac{\boldsymbol{\beta}^6\boldsymbol{\mu}^6}{\boldsymbol{\varrho}^3})$.

□

Lemma E.17. *Consider any coordinate $2 \le k \le M$. For any constants $0 < \boldsymbol{\beta} < \frac{\boldsymbol{\mu}\boldsymbol{\Delta}}{8\boldsymbol{\zeta}^2}$, suppose at time step t, $x_\eta(t)$ is in $Y^{\boldsymbol{\epsilon}}$, and the following :*

1. $\boldsymbol{\beta}\eta \le |\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| < \left(1 - \frac{1}{100M}\right) g_t(\lambda_k(t))$.
2. $|\langle v_k(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| > \boldsymbol{\alpha}\eta^2$.

Here $\boldsymbol{\alpha} = \Theta(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3\boldsymbol{\beta}})$.

Then, we must have some time $\bar{t} \le t + \mathcal{O}\left(\frac{M\boldsymbol{\zeta}}{\boldsymbol{\beta}^3} t_{\mathrm{escape}}\right)$ such that either of the above two conditions breaks, i.e. we have either of the following two conditions hold true:

1. *The magnitude along the top eigenvector grows beyond $\left(1 - \frac{1}{100M}\right) g_t(\lambda_k(t))$, i.e.*

$$
\left|\langle v_1(\bar{t}), (x_\eta(\bar{t}) - \Phi(x_\eta(\bar{t})))\rangle\right| \ge \left(1 - \frac{1}{100M}\right) g_t(\lambda_k(\bar{t})), \tag{69}
$$

   when the norm of $\widetilde{x}_\eta(t)$ is at most $0.5\lambda_1(t)\eta + \mathcal{O}(\boldsymbol{\Psi}_{\mathrm{G}}\eta^2)$.

2. *The magnitude along the eigenvector $v_k(\cdot)$ drops below $\boldsymbol{\alpha}\eta^2$, i.e.*

$$
\left|\langle v_k(\bar{t}), x_\eta(\bar{t}) - \Phi(x_\eta(\bar{t}))\rangle\right| \le \boldsymbol{\alpha}\eta^2. \tag{70}
$$

   Moreover, in this case, we must have

$$
\left|\langle v_1(\bar{t}), (x_\eta(\bar{t}) - \Phi(x_\eta(\bar{t})))\rangle\right| \ge \left(1 + \Omega(\frac{\boldsymbol{\mu}\boldsymbol{\Delta}\boldsymbol{\beta}^3}{M\boldsymbol{\zeta}})\right) |\langle v_1(t), (x_\eta(t) - \Phi(x_\eta(t)))\rangle| .
$$

The result holds true provided $\eta \le \widetilde{\mathcal{O}}(\frac{\boldsymbol{\mu}^3\boldsymbol{\beta}^2}{\boldsymbol{\zeta}^2\boldsymbol{\nu}})$ and for all time $t \le t' < \bar{t}$, $\overline{x_\eta(t')x_\eta(t'+1)} \subset Y^{\boldsymbol{\epsilon}}$.

Proof. We divide the interval $(t, \bar{t})$ into sub-intervals using Algorithm 2 into three subsets N_0, N_1, and N_2. We refer the reader to Appendix E.2 for a brief on the properties of the three sets.

Suppose, we denote the time-steps that belong to N_0 or N_1 by $t = t_1 < t_2 < t_3 \cdots < t_g = \bar{t}$. Then, we present our induction argument as follows: suppose we have shown that any of the final conditions (Equation (70) and Equation (69)) aren't true till t_i. Then, if $t_i \in N_0$, we can directly use the property from Lemma E.5 that $t_{i+1} = t_i + 1 \in N_1$. Moreover, from Lemma E.6, we must have $|\langle v_1(t_{i+1}), (x_\eta(t_{i+1}) - \Phi(x_\eta(t_{i+1})))\rangle| \geq |\langle v_1(t_i), (x_\eta(t_i) - \Phi(x_\eta(t_i)))\rangle| + \mathcal{O}(\eta^2)$.

However, if $t_i \in N_1$, we will consider the following two cases, depending on how large θ_{t_i} is:

1. If $\theta_{t_i} \leq \mathcal{O}\left(\sqrt{\frac{\boldsymbol{\beta}}{M\boldsymbol{\zeta}}}\right)$, then from Lemma E.18 we must have some time $t_i < t_j \leq t_i + t_{\text{escape}}$ where either Equation (69) holds true or the angle θ_{t_j} becomes greater than $\Omega\left(\sqrt{\frac{\boldsymbol{\beta}}{M\boldsymbol{\zeta}}}\right)$, along with the increase in the magnitude along the top eigenvector as

$$|\langle v_1(t_j), (x_\eta(t_j) - \Phi(x_\eta(t_j)))\rangle| \geq \left(1 + \Omega(\frac{\boldsymbol{\mu\Delta\beta}^3}{M\boldsymbol{\zeta}})\right) |\langle v_1(t_i), (x_\eta(t_i) - \Phi(x_\eta(t_i)))\rangle| .$$

2. If $\theta_{t_i} \geq \Omega\left(\sqrt{\frac{\boldsymbol{\beta}}{M\boldsymbol{\zeta}}}\right)$, we can use Lemma E.8 to have

$$|\langle v_1(t_i + 2), (x_\eta(t_i + 2) - \Phi(x_\eta(t_i + 2)))\rangle| \geq \left(1 + \Omega(\frac{\boldsymbol{\beta\mu\Delta}}{M\boldsymbol{\zeta}})\right) |\langle v_1(t_i), (x_\eta(t_i) - \Phi(x_\eta(t_i)))\rangle| .$$

Thus, the magnitude along the top eigenvector will keep on increasing monotonically, unless one the two conditions (Equation (69) and Equation (70)) hold true. □

Lemma E.18. *Consider any coordinate $2 \leq k \leq M$. For any constant $0 < \boldsymbol{\beta} < \frac{\boldsymbol{\mu\Delta}}{8\boldsymbol{\zeta}^2}$, suppose at time step t, $x_\eta(t)$ is in $Y^{\boldsymbol{\epsilon}}$, satisfies the alignment condition (Equation (28)) and the following :*

1. $\boldsymbol{\beta}\eta \leq |\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| < \left(1 - \frac{1}{100M}\right) g_t(\lambda_k(t))$.
2. $|\langle v_k(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| > \boldsymbol{\alpha}\eta^2$.
3. $\theta_t \leq \mathcal{O}\left(\sqrt{\frac{\boldsymbol{\beta}}{M\boldsymbol{\zeta}}}\right)$.

Here $\boldsymbol{\alpha} = \Theta(\frac{\boldsymbol{\nu\zeta}^2}{\boldsymbol{\mu}^3\boldsymbol{\beta}})$. Then, we must have some time $\bar{t} \leq t + t_{\text{escape}}$ such that either of the two above conditions breaks, i.e. we must have one of the two conditions hold true:

1. $\left|\langle v_1(\bar{t}), (x_\eta(\bar{t}) - \Phi(x_\eta(\bar{t})))\rangle\right| \geq \left(1 - \frac{1}{100M}\right) g_t(\lambda_k(\bar{t}))$.
2. $\theta_{\bar{t}} \geq \Omega(\sqrt{\frac{\boldsymbol{\beta}}{M\boldsymbol{\zeta}}})$. *Moreover, in this case, we must have*

$$\left|\langle v_1(\bar{t}), (x_\eta(\bar{t}) - \Phi(x_\eta(\bar{t})))\rangle\right| \geq \left(1 + \Omega(\frac{\boldsymbol{\mu\Delta\beta}^3}{M\boldsymbol{\zeta}})\right) |\langle v_1(t), (x_\eta(t) - \Phi(x_\eta(t)))\rangle| .$$

The result requires $\eta \leq \widetilde{\mathcal{O}}(\frac{\boldsymbol{\mu}^3\boldsymbol{\Delta\beta}^2}{\boldsymbol{\zeta}^2\boldsymbol{\nu\xi}})$ and for all time $t \leq t' < \bar{t}$, $\overline{x_\eta(t')x_\eta(t'+1)} \subset Y^{\boldsymbol{\epsilon}}$.

Proof. Here, $\boldsymbol{\alpha}$ has been selected such that if $x_\eta(t)$ satisfies $|\langle v_k(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| \geq \boldsymbol{\alpha}\eta^2$, then $|\langle v_k(t+1), x_\eta(t+1) - \Phi(x_\eta(t+1))\rangle| > |\langle v_k(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$. The proof strategy will be very similar to the strategy used in the previous lemmas, and we outline the sketch here: Consider the eigenvector $v_k(t)$,

1. First of all, Lemma E.8 can be slightly modified to show that at any step $t' \in [t, t + t_{\text{escape}}]$,

$$|v_1(t), x_\eta(t' + 2) - \Phi(x_\eta(t))| \geq |v_1(t), x_\eta(t') - \Phi(x_\eta(t))| + \mathcal{O}(\frac{\boldsymbol{\zeta}^2\boldsymbol{\nu}}{\boldsymbol{\mu}^3\boldsymbol{\beta}^2}\eta^2),$$

provided $|v_1(t), x_\eta(t') - \Phi(x_\eta(t))| \geq \mathcal{O}(\boldsymbol{\beta})$. Note the difference from Lemma E.8 is that the reference point has been changed to $\Phi(x_\eta(t))$ and the reference top eigenvetor is $v_1(t)$.

Thus, with $t_{\text{escape}} = \mathcal{O}(\log 1/\eta)$, we can apply mathematical induction to have

$$|v_1(t), x_\eta(t + t_{\text{escape}}) - \Phi(x_\eta(t))| \geq \frac{1}{2}\boldsymbol{\beta}\eta,$$

for $\eta \leq \widetilde{\mathcal{O}}(\frac{\boldsymbol{\mu}^3\boldsymbol{\beta}^2}{\boldsymbol{\zeta}^2\boldsymbol{\nu}})$.

2. Now, we can use Equation (19) (Lemma B.10) to show that the Normalized GD update is equivalent to update in quadratic model, with an additional $\mathcal{O}(\eta^2)$ error.

$$x_\eta(t+1) - x_\eta(t) = -\eta\frac{\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]}{\|\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t))]\|} + \mathcal{O}(\frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}\eta\,\|x_\eta(t) - \Phi(x_\eta(t))\|).$$

We can then consider the updates of $\langle v_i(t), x_\eta(t+2) - \Phi(x_\eta(t))\rangle - \langle v_i(t), x_\eta(t) - \Phi(x_\eta(t))\rangle$ using the update in a quadratic model outlined in Lemma A.9. That is,

$$\begin{aligned}
&\frac{\left|\langle v_k(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t+2) - \Phi(x_\eta(t)))\rangle\right|}{|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t+2) - \Phi(x_\eta(t)))\rangle|} \\
&= \left(1 - \eta\frac{\lambda_1(t) - \lambda_k(t)}{\lambda_1(t)\eta - \|\nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t+1) - \Phi(x_\eta(t)))\|}\right) \\
&\cdot \left(1 - \eta\frac{\lambda_1(t) - \lambda_k(t)}{\lambda_1(t)\eta - \|\nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\|}\right)\frac{\left|\langle v_k(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\rangle\right|}{|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\rangle|} \\
&+ \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3\boldsymbol{\beta}}\eta) \qquad (71) \\
&\geq \left(1 + \frac{1}{200M}\right)\frac{\left|\langle v_k(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\rangle\right|}{|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\rangle|} + \mathcal{O}(\frac{\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3\boldsymbol{\beta}}\eta) \\
&\geq \left(1 + \frac{1}{400M}\right)\frac{\left|\langle v_k(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\rangle\right|}{|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\rangle|},
\end{aligned}$$

where we bound $\|x_\eta(t) - \Phi(x_\eta(t))\|$ by $\mathcal{O}(\frac{\boldsymbol{\zeta}}{\boldsymbol{\mu}})$ and lower-bound $\left|v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\right|$ by $\boldsymbol{\mu}\boldsymbol{\beta}$ to get the second step. The third step follows from using the same argument as the one used for the quadratic update in Lemma A.9. The final step holds true if $\boldsymbol{\alpha} \geq \Omega(\frac{c_{\text{thres}}(t)\boldsymbol{\nu}\boldsymbol{\zeta}^2}{\boldsymbol{\mu}^3\boldsymbol{\beta}})$, since $\frac{\left|\langle v_k(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\rangle\right|}{|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))\rangle|} \geq \frac{\boldsymbol{\alpha}\eta}{c_{\text{thres}}(t)}$.

Thus, we can continue the above argument for all $t' \in [t, t + t_{\text{escape}}]$ to get

$$\begin{aligned}
&\frac{\left|\langle v_k(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t' + 2) - \Phi(x_\eta(t)))\rangle\right|}{|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t' + 2) - \Phi(x_\eta(t)))\rangle|} \\
&\geq \left(1 + \frac{1}{400M}\right)\frac{\left|\langle v_k(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t') - \Phi(x_\eta(t)))\rangle\right|}{|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t') - \Phi(x_\eta(t)))\rangle|},
\end{aligned}$$

up until we satisfy one of the two conditions:

(a) $\frac{\left\|P^{(2:M)}_{t,\Gamma}\widetilde{x}_\eta(t')\right\|}{|\langle v_1(t), \widetilde{x}_\eta(t')\rangle|} \geq \Omega\left(\sqrt{\frac{|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t') - \Phi(x_\eta(t)))\rangle|}{M\eta\lambda_1(t)}}\right) \geq \Omega(\sqrt{\frac{\boldsymbol{\beta}}{M\boldsymbol{\zeta}}})$.

(b) $|\langle v_1(t), x_\eta(t') - \Phi(x_\eta(t))\rangle| \geq \left(1 - \frac{1}{101M}\right) g_t(\lambda_k(t))$.

3. We then bound $\|v_k(t + t_{\text{escape}}) - v_k(t)\|$ and $\|\Phi(x_\eta(t + t_{\text{escape}}) - \Phi(x_\eta(t))\|$ by $\mathcal{O}(\boldsymbol{\xi}\eta^2 t_{\text{escape}})$ using Lemma B.12. Combining everything, we have the final bound, provided $\eta \leq \mathcal{O}(\frac{\boldsymbol{\mu}\boldsymbol{\Delta}\boldsymbol{\beta}}{\boldsymbol{\zeta}^2\boldsymbol{\xi}})$.

4. Moreover, we can show from observing Equation (71) that the value of $\frac{\left|\langle v_k(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t'+2) - \Phi(x_\eta(t)))\rangle\right|}{|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t'+2) - \Phi(x_\eta(t)))\rangle|}$ is at most $\frac{1}{\boldsymbol{\beta}}$ times $\frac{\left|\langle v_k(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t') - \Phi(x_\eta(t)))\rangle\right|}{|\langle v_1(t), \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t') - \Phi(x_\eta(t)))\rangle|}$. Thus, we can argue that there exists some step t', where

$\theta_{t'} = \Theta(\sqrt{\frac{\beta^3}{M\zeta}})$, if the initial angle is strictly less than $\mathcal{O}(\sqrt{\frac{\beta^3}{M\zeta}})$. We can then use the result from Lemma E.8 to show that the magnitude along the top eigenvector has to increase by a factor $\left(1 + \frac{\mu\Delta\beta^3}{M\zeta}\right)$, when $\theta_{t'} = \Theta(\sqrt{\frac{\beta^3}{M\zeta}})$. Moreover, since the angle $\theta_{t'}$ is $\Omega(\eta)$ in all the steps, the magnitude along the top eigenvector never drops in any other step. Overall, we must have a increase in the magnitude along the top eigenvector by a factor $\left(1 + \frac{\mu\Delta\beta^3}{M\zeta}\right)$.

□

Lemma E.19. *Consider any coordinate $1 \le k \le M$. Suppose at time step t, $x_\eta(t)$ is in Y^ϵ, satisfies the alignment condition (Equation (28)) and the following :*

1. $\left(1 + \frac{1}{100M}\right) g_t(\lambda_k(t))\eta \le |\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| < 0.5\eta$.
2. $|\langle v_k(t), x_\eta(t) - \Phi(x_\eta(t))\rangle| > \alpha\eta^2$.

Here $\alpha = \Theta(\frac{\nu\zeta^2}{\mu^2\beta})$.

Then, we have have some time $\bar{t} \le t + t_{\text{escape}}$ such that the magnitude along $v_k(t)$ drops below $\alpha\eta^2$, i.e.

$$\left|\langle v_k(\bar{t}), x_\eta(\bar{t}) - \Phi(x_\eta(\bar{t}))\rangle\right| \le \alpha\eta^2,$$

when $\left\|\nabla^2 L(\Phi(x_\eta(\bar{t})))(x_\eta(\bar{t}) - \Phi(x_\eta(\bar{t})))\right\| \le 0.5\lambda_1(t)\eta + \Psi_{\text{norm}}\eta^2$. The results hold true when $\eta \le \widetilde{\mathcal{O}}(\frac{\mu^3\beta^2}{\zeta^2\nu})$ and for all time $t \le t' < \bar{t}$, $\overline{x_\eta(t')x_\eta(t'+1)} \subset Y^\epsilon$.

The proof is going to be very similar to the proof of Lemma E.18, where the only difference will be that we need to use Lemma A.10 in place of Lemma A.9, when we use the result for the quadratic model.

E.4. Ommited Proof for operating on Edge of Stability

Proof of Theorem 4.7. According to the proof of Theorem 4.4, we know for all t, it holds that $R_j(x_\eta(t)) \le \mathcal{O}(\eta^2)$. Thus $S_L(x_\eta(t), \eta_t) = \eta_t \cdot \sup_{0\le s\le\eta_t} \lambda_1(\nabla^2 L(x_\eta(t) - s\nabla L(x_\eta(t)))) = \eta_t(\lambda_1(t) + \mathcal{O}(\eta))$, which implies that $[S_L(x_\eta(t), \eta_t)]^{-1} = \frac{\|\nabla L(x_\eta(t))\|}{\eta\lambda_1(t)} + \mathcal{O}(\eta) = \frac{\|\widetilde{x}_\eta(t)\|}{\eta\lambda_1(t)} + \mathcal{O}(\eta)$. The proof for the first claim is completed by noting that $\frac{1}{\eta}(\|\widetilde{x}_\eta(t)\| + \|\widetilde{x}_\eta(t+1)\|) = \lambda_1(t) + \mathcal{O}(\eta + \theta_t)$ as an analog of the quadratic case.

For the second claim, it's easy to check that $\sqrt{L(x_\eta(t))} = \frac{\|\widetilde{x}_\eta(t)\|}{\sqrt{2\lambda_1(t)}} + \mathcal{O}(\eta\theta_t)$. Thus have $\sqrt{L(x_\eta(t))} + \sqrt{L(x_\eta(t+1))} = \frac{\|\widetilde{x}_\eta(t)\|}{\sqrt{2\lambda_1(t)}} + \frac{\|\widetilde{x}_\eta(t+1)\|}{\sqrt{2\lambda_1(t+1)}} + \mathcal{O}(\eta(\theta_t + \theta_{t+1}))$. Note that $\lambda_1(t) - \lambda_1(t+1) = \mathcal{O}(\eta^2)$ and $\theta_{t+1} = \mathcal{O}(\theta_t)$, we conclude that $\sqrt{L(x_\eta(t))} + \sqrt{L(x_\eta(t+1))} = \eta\sqrt{\frac{\lambda_1(\nabla^2 L(x_\eta(t)))}{2}} + \mathcal{O}(\eta\theta_t)$. □

F. Some Useful Lemmas About Eigenvalues and Eigenvectors

Theorem F.1 (Derivative of eigenvalues and eigenvectors of a matrix, Theorem 1 in (Magnus, 1985))**.** *Let x_0 be a real symmetric $n \times n$ matrix. Let u_0 be a normalized eigenvector associated with a simple eigenvalue λ_0 of X_0. Then a real valued function λ and a vector valued function u are defined for all X in some neighborhood $N(x_0) \subset \mathbb{R}^{n\times n}$ of X_0, such that*

$$\lambda(X_0) = \lambda_0, \quad u(X_0) = u_0,$$

and

$$Xu = \lambda u, \quad u^\top u = 1, \quad X \in N(X_0).$$

Moreover, the functions λ and u are ∞ times differentiable on $N(X_0)$ and the differentials at X_0 are

$$d\lambda = u_0^\top (dX)u_0, \quad du = (\lambda_0 I_n - X_0)^\dagger (dX)u_0.$$

In the above theorem, a simple eigenvalue is defined as an eigenvalue with multiplicity 1.

Theorem F.2. *[Eigenvalue perturbation for symmetric matrices, Cor. 4.3.15 in (Horn & Johnson, 2012)] Let $\Sigma, \widehat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ and $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_p$ respectively. Then, for any $i \leq p$, we have*

$$\left|\lambda_i - \widehat{\lambda}_i\right| \leq \left\|\Sigma - \widehat{\Sigma}\right\|_2.$$

The next theorem is the Davis-Kahan $\sin(\theta)$ theorem, that bounds the change in the eigenvectors of a matrix on perturbation. Before presenting the theorem, we need to define the notion of unitary invariant norms. Examples of such norms include the frobenius norm and the spectral norm.

Definition F.3 (Unitary invariant norms)**.** A matrix norm $\|\cdot\|_*$ on the space of matrices in $\mathbb{R}^{p\times d}$ is unitary invariant if for any matrix $K \in \mathbb{R}^{p\times d}$, $\|UKW\|_* = \|K\|_*$ for any unitary matrices $U \in \mathbb{R}^{p\times p}, W \in \mathbb{R}^{d\times d}$.

Theorem F.4. *[Davis-Kahan $\sin(\theta)$ theorem (Davis & Kahan, 1970)] Let $\Sigma, \widehat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ and $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_p$ respectively. Fix $1 \leq r \leq s \leq p$, let $d := s - r + 1$ and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$ and $\widehat{V} = (\widehat{v}_r, \widehat{v}_{r+1}, \ldots, \widehat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\Sigma v_j = \lambda_j v_j$ and $\widehat{\Sigma}\widehat{v}_j = \widehat{\lambda}_j \widehat{v}_j$ for $j = r, r+1, \ldots, s$. Define $\boldsymbol{\Delta} := \min\left\{\max\{0, \lambda_s - \widehat{\lambda}_{s+1}\}, \max\{0, \widehat{\lambda}_{r-1} - \lambda_r\}\right\}$, where $\widehat{\lambda}_0 := \infty$ and $\widehat{\lambda}_{p+1} := -\infty$, we have for any unitary invariant norm $\|\cdot\|_*$,*

$$\boldsymbol{\Delta} \cdot \|\sin\Theta(\widehat{V}, V)\|_* \leq \|\widehat{\Sigma} - \Sigma\|_*.$$

Here $\Theta(\widehat{V}, V) \in \mathbb{R}^{d\times d}$, with $\Theta(\widehat{V}, V)_{j,j} = \arccos\sigma_j$ for any $j \in [d]$ and $\Theta(\widehat{V}, V)_{i,j} = 0$ for all $i \neq j \in [d]$. $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ denotes the singular values of $\widehat{V}^\top V$. $[\sin\Theta]_{ij}$ is defined as $\sin(\Theta_{ij})$.

Lemma F.5 (Parameter bounds)**.** *The upper bound $\boldsymbol{\gamma}_{\text{ub}}$ and lipschitz-constant $\boldsymbol{\beta}_{\text{lip}}$ for the function $P^{\perp}_{\Phi(x),\Gamma}\nabla\log\lambda_1(x)$ for any point $x \in Y^{\boldsymbol{\epsilon}}$ can be given as*

$$\boldsymbol{\beta}_{\text{lip}} = \frac{\boldsymbol{\nu}\boldsymbol{\xi}}{\boldsymbol{\mu}} + \frac{\boldsymbol{\nu}\boldsymbol{\xi}^2\boldsymbol{\epsilon}}{\boldsymbol{\mu}} + \frac{\boldsymbol{\Upsilon} + \boldsymbol{\nu}^2\boldsymbol{\Delta}^{-1} + \boldsymbol{\nu}^2}{\boldsymbol{\mu}^2}$$
$$\boldsymbol{\gamma}_{\text{ub}} = \frac{\boldsymbol{\nu}}{\boldsymbol{\mu}}.$$

Proof. First, we focus on the bound of the function $P^{\perp}_{\Phi(x),\Gamma}\nabla\log\lambda_1(x)$. $P^{\perp}_{\Phi(x),\Gamma}$ is just a projection matrix, while $\nabla\log\lambda_1(x) = \frac{\nabla^3 L(\Phi(x))[v_1(x), v_1(x)]}{\lambda_1(x)}$, using Theorem F.4. Each term can then be bounded using the definition from Definition B.5.

Now, we look at the lipschitz constant of the function $P^{\perp}_{\Phi(x),\Gamma}\nabla\log\lambda_1(x)$. Using the derivative of the function $P^{\perp}_{\Phi(x),\Gamma}\nabla\log\lambda_1(x)$ at any point x, we have

$$\begin{aligned}
&\left\|\nabla_x P^{\perp}_{\Phi(x),\Gamma}\nabla\log\lambda_1(x)\right\| \\
&\leq \left\|\nabla_x P^{\perp}_{\Phi(x),\Gamma}\right\| \|\nabla\log\lambda_1(x)\| + \left\|P^{\perp}_{\Phi(x),\Gamma}\right\| \|\nabla^2\log\lambda_1(x\| \\
&\leq \left\|\nabla_x P^{\perp}_{\Phi(x),\Gamma}\right\| \|\nabla\log\lambda_1(x)\| + \left\|P^{\perp}_{\Phi(x),\Gamma}\right\| \left\|\frac{\nabla^2\lambda_1(x)}{\lambda_1(x)^2}\right\| + \left\|P^{\perp}_{\Phi(x),\Gamma}\right\| \frac{\|\nabla\lambda_1(x)\|^2}{\lambda_1(x)^2}.
\end{aligned}$$

We can bound $\left\|\nabla_x P^{\perp}_{\Phi(x),\Gamma}\right\|$ by $\left\|\partial^2\Phi(x)\right\| \|\partial\Phi(x)\|$ using the equivalence between $P^{\perp}_{\Phi(x),\Gamma}$ and $\partial\Phi(\Phi(x))$ from Lemma B.17. Moreover, using taylor expansion, we can bound $\|\partial\Phi(x)\| \leq \|\partial\Phi(\Phi(x))\| + \boldsymbol{\xi}\boldsymbol{\epsilon}$, for any $x \in Y^{\boldsymbol{\epsilon}}$. Moreover, since $P^{\perp}_{\Phi(x),\Gamma} = \partial\Phi(\Phi(x))$, we must have $\|\partial\Phi(\Phi(x))\| = 1$. Using the bound on the second derivative of Φ from Definition B.5, we have $\left\|\nabla_x P^{\perp}_{\Phi(x),\Gamma}\right\| \leq \boldsymbol{\xi} + \boldsymbol{\xi}^2\boldsymbol{\epsilon}$.

We can further use Theorem F.4 to get the desired derivatives:

$$\begin{aligned}
\left\|\nabla^2\lambda_1(x)\right\| &\leq \left\|\nabla^3 L(\Phi(x))\right\| + \|x\| \|\nabla v_1(x)\| \\
&\leq \left\|\nabla^3 L(\Phi(x))\right\| + \|x\|^2 \frac{1}{\lambda_1(x) - \lambda_2(x)}.
\end{aligned}$$

Algorithm 3 Perturbed Gradient Descent on $\sqrt{L}$

Input: loss function $L : \mathbb{R}^D \to \mathbb{R}$, initial point x_{init}, maximum number of iteration T, LR η, Frequency parameter $T_{\text{freq}} = \Theta(\eta^{-0.1})$, noise parameter $r = \Theta(\eta^{100})$.
for $t = 1$ **to** T **do**
  Generate $n(t) \sim B_0(r)$ if $t \mod T_{\text{freq}} = 0$, else set $n(t) = 0$.
  $x(t) \leftarrow x(t-1) - \eta \nabla \sqrt{L}(x(t)) + n(t)$.
end for

We can finally bound each of the terms using Definition B.5.

□

G. Analysis of $\sqrt{L}$

The analysis will follow the same line of proof used for the analysis of Normalized GD. Hence, we write down the main lemmas that are different from the analysis of Normalized GD. Rest of the lemmas are nearly the same and hence, we have omitted them.

The major difference between the results of Normalized GD and GD with $\sqrt{L}$ is in the behavior along the manifold Γ (for comparison, see Lemma B.13 for Normalized GD and Lemma G.11 for GD with $\sqrt{L}$). Another difference between the results of Normalized GD and GD with $\sqrt{L}$ is in the error rates mentioned in Theorem 4.4 and Theorem 4.6. The difference comes from the stronger behavior of the projection along the top eigenvector that we showed for Normalized GD in Lemma E.8, but doesn't hold for GD with $\sqrt{L}$ (see Lemma G.6). This difference shows up in the sum of angles across the trajectory (for comparison, see Lemma E.1 for Normalized GD and Lemma G.4 for GD with $\sqrt{L}$), and is finally reflected in the error rates.

G.1. Notations

The notations will be the same as Appendix B . However, here we will use $\widetilde{x}_\eta(t)$ to denote $\left(\nabla^2 L(\Phi(x_\eta(t)))\right)^{1/2}(x_\eta(t) - \Phi(x_\eta(t)))$. We will now denote Y as the limiting flow given by Equation (6).

$$X(\tau) = \Phi(x_{\text{init}}) - \frac{1}{8}\int_{s=0}^{\tau} P^{\perp}_{X(s),\Gamma} \nabla \lambda_1(X(s)) \mathrm{d}s. \tag{6}$$

G.2. Phase I, convergence

Here, we will show a very similar stability condition for the GD update on $\sqrt{L}$ as the one (Lemma C.1) derived for Normalized GD. Recall our notation $\widetilde{x}_\eta(t) = \sqrt{\nabla^2 L(\Phi(x_\eta(t)))}(x_\eta(t) - \Phi(x_\eta(t)))$.

Lemma G.1. *Suppose $\{x_\eta(t)\}_{t\geq 0}$ are iterates of GD with $\sqrt{L}$ (5) with a learning rate η and $x_\eta(0) = x_{init}$. There are constants $C > 0$, such that for any constant $\boldsymbol{\varsigma} > 0$, if at some time t', $x_\eta(t') \in Y^{\boldsymbol{\epsilon}}$ and satisfies $\frac{\left\|x_\eta(t') - \Phi(x_\eta(t'))\right\|}{\eta} \leq \boldsymbol{\varsigma}$, then for all $\bar{t} \geq t' + C\frac{\boldsymbol{\zeta}\boldsymbol{\varsigma}}{\boldsymbol{\mu}} \log \frac{\boldsymbol{\varsigma}\boldsymbol{\zeta}}{\boldsymbol{\mu}}$, the following must hold true for all $1 \leq j \leq M$:*

$$\sqrt{\sum_{i=j}^{M} \langle v_i(\bar{t}), \widetilde{x}_\eta(\bar{t})\rangle^2} \leq \eta\sqrt{\frac{1}{2}\lambda_j^2(\bar{t})} + \mathcal{O}(\boldsymbol{\nu}\boldsymbol{\xi}\boldsymbol{\zeta}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\zeta}\boldsymbol{\nu}\boldsymbol{\varsigma}\eta^2}{\boldsymbol{\mu}}) + \mathcal{O}(\boldsymbol{\xi}\boldsymbol{\zeta}^2\boldsymbol{\varsigma}\boldsymbol{\nu}\sqrt{D}\eta^2) + \mathcal{O}(\eta^2 D), \tag{72}$$

provided $\eta \leq \mathcal{O}(\frac{\boldsymbol{\mu}^3}{\boldsymbol{\zeta}^3\boldsymbol{\varsigma}^2\boldsymbol{\xi}\boldsymbol{\nu}\sqrt{D}})$ and that for all steps $t \in \{t, \ldots, \bar{t}-1\}$, $\overline{x_\eta(t)x_\eta(t+1)} \subset Y^{\boldsymbol{\epsilon}}$.

Proof. The proof exactly follows the strategy used in Lemma C.1. We will outline the major milestones in the proof to help the interested readers.

1. We can use the noisy update formulation from Lemma G.9 and the bound on the movement in Φ from Lemma G.11 to get for any time t with $\bar{t} \geq t \geq t_0$ (similar to Equation (27)):

$$\widetilde{x}_\eta(t+1) = \left(I - \eta \frac{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x_\eta(t)))}}{\|\widetilde{x}_\eta(t)\|}\right)\widetilde{x}_\eta(t) + \mathcal{O}(\boldsymbol{\nu\xi\zeta}\eta^2) + \mathcal{O}(\frac{\boldsymbol{\zeta}^{1/2}\boldsymbol{\nu}\,\|x_\eta(t) - \Phi(x_\eta(t))\|}{\boldsymbol{\mu}}\eta).$$

2. Secondly, we show for all $t_0 + 1 \leq t \leq \bar{t}$, $\|\widetilde{x}_\eta(t)\| \leq 1.01\eta\sqrt{\boldsymbol{\zeta}}\boldsymbol{\varsigma}$ using an induction argument. We require $\eta \leq \mathcal{O}(\frac{\boldsymbol{\mu}^3}{\boldsymbol{\zeta}^2\boldsymbol{\nu\varsigma\xi}})$.

3. Finally, we show the desired bound by coupling the trajectory with the quadratic model, where the quadratic update is governed by $\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x_\eta(t)))}$ at each step t. Here, we need $\eta \leq \mathcal{O}(\frac{\boldsymbol{\mu}^3}{\boldsymbol{\zeta}^3\boldsymbol{\varsigma}^2\boldsymbol{\xi\nu}\sqrt{D}})$.

□

A simple version of the above condition can be given as:

$$\sqrt{\sum_{i=j}^{M}\langle v_i(\bar{t}), \widetilde{x}_\eta(\bar{t})\rangle^2} \leq \eta\sqrt{2\lambda_j^2(\bar{t})}, \tag{73}$$

provided $\eta \leq \mathcal{O}(\frac{\boldsymbol{\mu}^2}{\boldsymbol{\nu\zeta}^2\boldsymbol{\xi\varsigma}D})$. For simplicity of presentation, we have used

$$\boldsymbol{\Psi}_{\text{norm}} = \mathcal{O}(\boldsymbol{\nu\xi\zeta} + \frac{\boldsymbol{\zeta\nu\varsigma}\eta^2}{\boldsymbol{\mu}}\boldsymbol{\xi\zeta}^2\boldsymbol{\varsigma\nu}\sqrt{D} + D).$$

Hence, we can derive the following property that continues to hold true throughout the trajectory, once the condition Equation (72) is satisfied:

Lemma G.2. *If $\eta \leq \mathcal{O}(\frac{\boldsymbol{\mu}^4}{D\boldsymbol{\varsigma\zeta}^2\boldsymbol{\xi\nu}})$ and condition Equation (72) holds true, then if $\|\widetilde{x}_\eta(t)\| > \frac{\eta\sqrt{0.5\lambda_1^2(t)}}{2} + \boldsymbol{\Psi}_{\text{norm}}\eta^2$, the following must hold true:*

$$\|\widetilde{x}_\eta(t+1)\| \leq \frac{\eta\sqrt{0.5\lambda_1^2(t)}}{2} + \boldsymbol{\Psi}_{\text{norm}}\eta^2.$$

The proof follows from applying Lemma A.8 for a quadratic update with $\sqrt{\nabla^2 L(\Phi(x_\eta(t)))}$, using the stability condition Equation (72).

Thus, we will consider the trajectory in cycles of length 2, with the norm of $\widetilde{x}_\eta(t) \leq \frac{\eta\sqrt{0.5\lambda_1(t)}}{2}$ at the start of the cycle.

Another useful lemma is to show that the magnitude along the top eigenvector increases when $\|\widetilde{x}_\eta(t)\| \leq \frac{\eta\sqrt{0.5\lambda_1^2(t)}}{2}$.

Lemma G.3. *If at time t, $\|\widetilde{x}_\eta(t)\| \leq \frac{\eta\sqrt{0.5\lambda_1(t)}}{2} + \boldsymbol{\Psi}_{\text{norm}}\eta^2$ and stability condition (Equation (72)) holds true, the following must hold true:*

$$\left|v_1(t+1)^\top \widetilde{x}_\eta(t+1)\right| \geq \left|v_1(t)^\top \widetilde{x}_\eta(t)\right| - \mathcal{O}(\frac{\boldsymbol{\zeta}^{3/2}\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}}\eta^2) - \boldsymbol{\Psi}_{\text{norm}}\eta^2.$$

The proof follows from using the noisy update of GD on $\sqrt{L}$ from Lemma G.9 and using the quadratic update result from Lemma A.5.

G.3. Phase II, limiting flow

To recall, the limiting flow given by

$$X(\tau) = \Phi(x_{\text{init}}) - \frac{1}{8}\int_{s=0}^{\tau} P^{\perp}_{X(s),\Gamma}\nabla\lambda_1(X(s))\mathrm{d}s. \tag{6}$$

Let T_2 be the time up until which solution to the limiting flow exists.

Lemma G.11 shows the movement in Φ, which can be informally given as follows: in each step t,

$$\Phi(x_\eta(t+1)) - \Phi(x_\eta(t)) = -\frac{\eta^2}{8}P^{\perp}_{t,\Gamma}\nabla\lambda_1(\nabla^2 L(\Phi(x_\eta(t)))) + \mathcal{O}(\eta^2(\theta_t + \left\|x_\eta(t) - \Phi(x_{\eta(t)})\right\|)), \tag{74}$$

provided $\overline{\Phi(x_\eta(t))\Phi(x_\eta(t+1))} \in Y^{\epsilon}$.

Motivated by this update rule, we show that the trajectory of $\Phi(x_\eta(\cdot))$ is close to the limiting flow in Equation (6), for a small enough learning rate η. The major difference from Theorem 4.4 comes from the fact that the total error introduced in Equation (74) over an interval $[0, t_2]$ is $\sum_{t=0}^{t_2}\mathcal{O}(\eta^2\theta_t + \eta^3)$, which is of the order $\mathcal{O}(\eta^{1/2})$ using the result of Lemma G.4.

G.3.1. Average of the angles

The first lemma shows that the sum of the angles in an interval $[0, t_2]$ of length $\Omega(1/\eta^2)$ is atmost $\mathcal{O}(t_2\eta^{1/2})$.

Lemma G.4. *For any $T_2 > 0$ for which solution of Equation (6) exists, consider an interval $[0, t_2]$, with $t_2 \le \lfloor T_2/\eta^2 \rfloor$. Suppose Algorithm 3 is run with learning rate η for t_2 steps, starting from a point $x_\eta(0)$ that satisfies (1) $\max_{j\in[D]}\overline{R}_j(x_\eta(0)) \le O(\eta^2)$, and (2) $|v_1(0), x_\eta(0) - \Phi(x_\eta(0))| \ge \boldsymbol{\beta}\eta$ for some constant $0 < \boldsymbol{\beta} \le \frac{\boldsymbol{\mu}\boldsymbol{\Delta}}{8\boldsymbol{\zeta}^2}$ independent of η, with $\|\widetilde{x}_\eta(0)\| \le \frac{\sqrt{0.5\lambda_1^2(0)}}{2}\eta + \boldsymbol{\Psi}_{\text{norm}}\eta^2$. For any $T_2 > 0$ for which solution to Equation (6) exists, and any integer $t_2 \le \frac{T_2}{\eta^2}$, the following holds true with probability at least $1 - \eta^{10}$:*

$$\sum_{\ell=0}^{t_2}\theta_\ell \le \mathcal{O}\left(\sqrt{\frac{\boldsymbol{\zeta}^4\boldsymbol{\nu}\boldsymbol{\xi}}{\boldsymbol{\mu}^{5/2}\boldsymbol{\Delta}\boldsymbol{\beta}^2}\eta}\right),$$

provided η is sufficiently small and for all time $0 \le t \le t_2 - 1$, $\overline{x_\eta(t)x_\eta(t+1)} \subset Y^{\epsilon}$.

Proof. The proof is very similar to the proof of Lemma E.1. The only difference shows up from the result of Lemma G.5, and hence we show that the sum of the angles when the iterate is stuck in any of the regions satisfying conditions $B(k)/C(k)$ for the interval $(\widetilde{t}_i, \widetilde{t}_{i+1})$ is $\mathcal{O}\left(\frac{\boldsymbol{\zeta}^5}{\boldsymbol{\mu}^2\boldsymbol{\Delta}^2} + \frac{M\boldsymbol{\zeta}}{\boldsymbol{\beta}^3}t_{\text{escape}} + (\widetilde{t}_{i+1} - \widetilde{t}_i)\sqrt{\frac{\boldsymbol{\zeta}^4\boldsymbol{\nu}\boldsymbol{\xi}}{\boldsymbol{\mu}^{5/2}\boldsymbol{\Delta}\boldsymbol{\beta}^2}\eta}\right)$. □

Lemma G.5. *Consider any time $\bar{t}$, where $x_\eta(\bar{t}) \in Y^{\epsilon}$, where $\left\|\widetilde{x}_\eta(\bar{t})\right\| \le 0.5\eta\lambda_1(\bar{t}) + \boldsymbol{\Psi}_{\text{norm}}\eta^2$. Suppose we are also given p disjoint subsets of $[M]$, $S_1, \cdots, S_p$ (with $1 \le p \le M$) and a step $t_{\text{stop}} \ge \bar{t}$, such that for any $i, j \in [p]$ with $i \ne j$, and for any $\bar{t} \le t \le t_{\text{stop}}$, we can guarantee*

$$\min_{k\in S_i,\ell\in S_j}\left|\left|\frac{1}{2} - \lambda_k(t)\right| - \left|\frac{1}{2} - \lambda_\ell(t)\right|\right| \ge \frac{1}{2}\times 10^{-3}\lambda_1(t),$$
$$\min_{\ell\in S_i} g_t(\lambda_\ell(t)) > \max_{\ell\in S_j} g_t(\lambda_\ell(t)), \quad \text{if } i > j.$$

Consider any subset S_k for $1 \le k \le p$. If $(1 - \frac{1}{100M})\min_{\ell\in S_k} g_{\bar{t}}(\lambda_\ell(\bar{t})) \le G_{\bar{t}} \le (1 + \frac{1}{100M})\max_{\ell\in S_k} g_{\bar{t}}(\lambda_\ell(\bar{t}))$ and suppose there exists some time $\bar{t} \le t' \le t_{\text{stop}}$ such that the iterate is stuck inside this region in the interval $(\bar{t}, t')$. I.e. for all $t \in (\bar{t}, t')$, whenever $\|\widetilde{x}_\eta(t)\| \le 0.5\eta\lambda_1(t) + \boldsymbol{\Psi}_{\text{norm}}\eta^2$, we must have $(1 - \frac{1}{100M})\min_{\ell\in S_k} g_t(\lambda_\ell(t)) \le G_t \le (1 + \frac{1}{100M})\max_{\ell\in S_k} g_t(\lambda_\ell(t))$. Then,

$$\sum_{\ell=\bar{t}}^{t'}\theta_\ell \le \mathcal{O}\left(\frac{\boldsymbol{\zeta}^5}{\boldsymbol{\mu}^2\boldsymbol{\Delta}^2} + \frac{M\boldsymbol{\zeta}}{\boldsymbol{\beta}^3}t_{\text{escape}} + (t' - \bar{t})\sqrt{\frac{\boldsymbol{\zeta}^4\boldsymbol{\nu}\boldsymbol{\xi}}{\boldsymbol{\mu}^{5/2}\boldsymbol{\Delta}\boldsymbol{\beta}^2}\eta}\right),$$

where G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$, provided for all $\bar{t} \le \ell < t'$, $\overline{x_\eta(\ell)x_\eta(\ell+1)} \subset Y^{\epsilon}$.

Proof. The proof will follow exactly as Lemma E.3. The only difference is that G_t is bound to increase when $\theta_t \geq \Omega\left(\sqrt{\frac{\boldsymbol{\zeta}^4\boldsymbol{\nu}\boldsymbol{\xi}}{\boldsymbol{\mu}^{5/2}\boldsymbol{\Delta}\boldsymbol{\beta}^2}\eta}\right)$ from Corollary G.7. Thus, at step 3, when we take the sum of the angle over the steps where G_t is bound to decrease, we must have $\sum_{t\in N_2^{(2)}}\theta_t \leq \mathcal{O}\left(\sqrt{\frac{\boldsymbol{\zeta}^4\boldsymbol{\nu}\boldsymbol{\xi}}{\boldsymbol{\mu}^{5/2}\boldsymbol{\Delta}\boldsymbol{\beta}^2}\eta}(t'-\bar{t})\right)$. □

G.3.2. Behavior along the top eigenvector

Lemma G.6. *Suppose $\eta \leq \mathcal{O}(\frac{\boldsymbol{\mu}^4}{D\boldsymbol{\varsigma}\boldsymbol{\zeta}^2\boldsymbol{\xi}\boldsymbol{\nu}})$. Consider any time t, such that $x_\eta(t) \in Y^\epsilon$, where $\|\widetilde{x}_\eta(t)\| \leq \frac{1}{2}\eta\sqrt{0.5\lambda_1^2(t)} + \boldsymbol{\Psi}_{\text{norm}}\eta^2$ holds true. If G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$ and G_{t+2} denotes the quantity $|\langle v_1(t+2), x_\eta(t+2) - \Phi(x_\eta(t+2))\rangle|$, then the following holds true:*

$$G_{t+2} \geq (1 + \frac{1}{2}\min_{2\leq j\leq M}\frac{\lambda_j(t)(\lambda_1(t)-\lambda_j(t))}{\lambda_1^2(t)}\sin^2\theta_t)G_t - \mathcal{O}(\frac{\eta}{G_t}\frac{\boldsymbol{\zeta}^2\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}}\eta^2) - \mathcal{O}(\boldsymbol{\xi}\boldsymbol{\zeta}\eta^2),$$

provided $G_t \geq \Omega(\eta^{1.5})$ and $\overline{x_\eta(t)x_\eta(t+1)}, \overline{x_\eta(t+1)x_\eta(t+2)} \subset Y^\epsilon$. Here θ_t is given by $\cos^{-1}(\frac{|\langle v_1(t), \widetilde{x}_\eta(t)\rangle|}{\|\widetilde{x}_\eta(t)\|})$, with $P_{t,\Gamma}$ denoting the projection matrix onto the subspace spanned by $v_1(t), \ldots, v_M(t)$, and $\widetilde{x}_\eta(t) = \nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))$.

Proof. Here, we will follow a much simpler approach than Lemma E.8 to have a weaker error bound. The stronger error bounds in Lemma E.8 were due to the very specific update rule of Normalized GD.

From Lemma G.9, we have

$$\frac{\nabla L(x_\eta(t))}{\sqrt{L(x_\eta(t))}} = \frac{\nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]}} + \mathcal{O}(\frac{\boldsymbol{\zeta}^{3/2}\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}}\eta),$$

where we have used the fact that $x_\eta(t) - \Phi(x_\eta(t))$ satisfies the stability condition from Equation (72) and hence, $\|x_\eta(t) - \Phi(x_\eta(t))\| \leq \mathcal{O}(\frac{\sqrt{\boldsymbol{\zeta}^2}}{\sqrt{\boldsymbol{\mu}}})$.

Thus, we have

$$\begin{aligned}
&x_\eta(t+1) - \Phi(x_\eta(t)) \\
&= x_\eta(t) - \Phi(x_\eta(t)) - \eta\frac{\nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t) - \Phi(x_\eta(t)))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]}} + \mathcal{O}(\frac{\boldsymbol{\zeta}^{3/2}\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}}\eta^2).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&x_\eta(t+2) - \Phi(x_\eta(t)) \\
&= x_\eta(t+1) - \Phi(x_\eta(t)) - \eta\frac{\nabla^2 L(\Phi(x_\eta(t)))(x_\eta(t+1) - \Phi(x_\eta(t)))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1) - \Phi(x_\eta(t)), x_\eta(t+1) - \Phi(x_\eta(t))]}} + \mathcal{O}(\frac{\boldsymbol{\zeta}^{3/2}\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}}\eta^2).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sqrt{\nabla^2 L(\Phi(x_\eta(t)))}(x_\eta(t+2) - \Phi(x_\eta(t))) \\
&= \left(I - \eta\frac{2^{-1/2}\nabla^2 L(\Phi(x_\eta(t)))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1) - \Phi(x_\eta(t)), x_\eta(t+1) - \Phi(x_\eta(t))]}}\right) \\
&\cdot \left(I - \eta\frac{2^{-1/2}\nabla^2 L(\Phi(x_\eta(t)))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t) - \Phi(x_\eta(t)), x_\eta(t) - \Phi(x_\eta(t))]}}\right)\sqrt{\nabla^2 L(\Phi(x_\eta(t)))}(x_\eta(t) - \Phi(x_\eta(t))) \\
&+ \mathcal{O}(\frac{\eta}{G_t}\frac{\boldsymbol{\zeta}^2\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}}\eta^2).
\end{aligned}$$

The final error appears due to the possible blow-up from $\left(I - \eta \frac{2^{-1/2}\nabla^2 L(\Phi(x_\eta(t)))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x_\eta(t)))[x_\eta(t+1)-\Phi(x_\eta(t)),x_\eta(t+1)-\Phi(x_\eta(t))]}}\right)$, which has been bounded by $\frac{\eta\sqrt{\zeta}}{G_{t+1}} \leq \mathcal{O}(\frac{\eta\sqrt{\zeta}}{G_t})$, using the value of G_{t+1} from Lemma G.3.

Hence, the update is similar to the update in a quadratic model, with $\nabla^2 L(\Phi(x_\eta(t)))$ guiding the updates. The final bound comes from using Lemma A.7, and then reconciling the errors introduced by the changes in the top eigenvector and the function Φ from Lemma G.11. □

A corollary of the above lemma is that when the magnitude along the top eigenvalue is $\Omega(\eta)$, the magnitude drops, only when the angle of the iterate with the top eigenvector is $\mathcal{O}(\eta^{1/2})$.

Corollary G.7. *Consider any time t, such that $x_\eta(t) \in Y^\epsilon$, where $\|\widetilde{x}_\eta(t)\| \leq \frac{1}{2}\eta\sqrt{0.5\lambda_1^2(t)} + \boldsymbol{\Psi}_{\text{norm}}\eta^2$ holds true. If G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$ and G_{t+2} denotes the quantity $|\langle v_1(t+2), x_\eta(t+2) - \Phi(x_\eta(t+2))\rangle|$, then $G_{t+2} \geq G_t$ for all*

$$|\theta_t| \geq \Omega\left(\sqrt{\frac{\boldsymbol{\zeta}^2}{\boldsymbol{\mu}\boldsymbol{\Delta}}\frac{\eta}{G_t^2}\frac{\boldsymbol{\zeta}^2\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}}\eta^2 + \boldsymbol{\xi}\boldsymbol{\zeta}\frac{\eta^2}{G_t}}\right),$$

provided $G_t \geq \Omega(\eta^{1.5})$, and $\overline{x_\eta(t)x_\eta(t+1)}, \overline{x_\eta(t+1)x_\eta(t+2)} \subset Y^\epsilon$. Moreover, if $G_t \geq \boldsymbol{\beta}\eta$ for some $\boldsymbol{\beta} \geq 0$, then the above bound can be simplified as

$$|\theta_t| \geq \Omega\left(\sqrt{\frac{\boldsymbol{\zeta}^4\boldsymbol{\nu}\boldsymbol{\xi}}{\boldsymbol{\mu}^{5/2}\boldsymbol{\Delta}\boldsymbol{\beta}^2}\eta}\right).$$

The next corollary shows that if the magnitude along the top eigenvector drops, when it is $\Omega(\eta)$, it can only drop by a magnitude of $\mathcal{O}(\eta^2)$ at any step.

Corollary G.8. *Consider any time t, such that $x_\eta(t) \in Y^\epsilon$, where $\|\widetilde{x}_\eta(t)\| \leq \frac{1}{2}\eta\sqrt{0.5\lambda_1^2(t)} + \boldsymbol{\Psi}_{\text{norm}}\eta^2$ holds true. If G_t denotes the quantity $|\langle v_1(t), x_\eta(t) - \Phi(x_\eta(t))\rangle|$ and G_{t+2} denotes the quantity $|\langle v_1(t+2), x_\eta(t+2) - \Phi(x_\eta(t+2))\rangle|$, then*

$$G_{t+2} \geq G_t - \mathcal{O}(\frac{\eta}{G_t}\frac{\boldsymbol{\zeta}^2\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}}\eta^2) - \mathcal{O}(\boldsymbol{\xi}\boldsymbol{\zeta}\eta^2),$$

provided $G_t \geq \Omega(\eta^{1.5})$ and $\overline{x_\eta(t)x_\eta(t+1)}, \overline{x_\eta(t+1)x_\eta(t+2)} \subset Y^\epsilon$. Moreover, if $G_t \geq \boldsymbol{\beta}\eta$ for some $\boldsymbol{\beta} \geq 0$, then the above bound can be simplified as

$$G_{t+2} \geq G_t - \mathcal{O}(\boldsymbol{\Psi}_{\text{G}}\eta^2),$$

where $\boldsymbol{\Psi}_{\text{G}} = \frac{\boldsymbol{\zeta}^2\boldsymbol{\nu}\boldsymbol{\xi}}{\boldsymbol{\beta}\boldsymbol{\mu}^{3/2}}$.

G.4. Ommited Proof for operating on Edge of Stability

This proof is similar to that of Theorem 4.7.

Proof of Theorem 4.8. If $M = 1$, that is, the dimension of manifold Γ is $D-1$, we know $\overline{x_\eta(t)x_\eta(t+1)}$ will cross Γ, making the $\nabla^2\sqrt{L}$ diverges at the intersection and the first claim becomes trivial. If $M \geq 2$, we have $\nabla^2\sqrt{L} = \frac{2L\nabla^2 L - \nabla L\nabla L^\top}{4\sqrt{L}^3}$ diverges at the rate of $\frac{1}{\|\nabla L\|}$. It turns out that using basic geometry, one can show that the distance from $\Phi(x_\eta(t))$ to $\overline{x_\eta(t)x_\eta(t+1)}$ is $\mathcal{O}(\eta(\theta_t + \theta_{t+1}))$, thus $\sup_{0\leq s\leq\eta}\lambda_1(\nabla^2\sqrt{L}(x_\eta(t) - s\nabla\sqrt{L}(x_\eta(t)))) = \Omega(\frac{1}{\eta(\theta_t+\theta_{t+1})})$. The proof of the first claim is completed by noting that $\theta_{t+1} = \mathcal{O}(\theta_t)$.

For the second claim, it's easy to check that $\sqrt{L(x_\eta(t))} = \|\widetilde{x}_\eta(t)\| + \mathcal{O}(\eta)$. The proof for the first claim is completed by noting that $\|\widetilde{x}_\eta(t)\| + \|\widetilde{x}_\eta(t+1)\| = \eta\lambda_1(t) + \mathcal{O}(\eta + \theta_t)$ as an analog of the quadratic case. □

G.5. Geometric Lemmas for $\sqrt{L}$

Lemma G.9. *At any point $x \in Y^{\epsilon}$, we have*

$$\frac{\nabla L(x)}{\sqrt{L(x)}} = \frac{\nabla^2 L(\Phi(x))(x-\Phi(x))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x))[x-\Phi(x), x-\Phi(x)]}} + \mathcal{O}(\frac{\boldsymbol{\zeta}^{1/2}\boldsymbol{\nu}}{\boldsymbol{\mu}} \|x-\Phi(x)\|).$$

Also,

$$\left\| \frac{\nabla L(x)}{\sqrt{L(x)}} \right\| \le \sqrt{2\lambda_1(x)} + \mathcal{O}(\frac{\boldsymbol{\zeta}^{1/2}\boldsymbol{\nu}\sqrt{D}}{\boldsymbol{\mu}} \|x-\Phi(x)\|).$$

Proof. Using taylor expansion around $\Phi(x)$, we have

$$\nabla L(x) = \nabla L(\Phi(x)) + \nabla^2 L(\Phi(x))(x-\Phi(x)) + \mathrm{err},$$

where $\|\mathrm{err}\| = \mathcal{O}(\boldsymbol{\nu}\|x-\Phi(x)\|^2)$ Similarly

$$L(x) = L(\Phi(x)) + \langle \nabla L(\Phi(x)), (x-\Phi(x))\rangle + \frac{1}{2}\nabla^2 L(\Phi(x))[x-\Phi(x), x-\Phi(x)] + \mathrm{err}',$$

where $\|\mathrm{err}'\| = \mathcal{O}(\boldsymbol{\nu}\|x-\Phi(x)\|^3)$. Since $\Phi(x)$ is a local minimizer, we have $L(\Phi(x)) = 0$ and $\nabla L(\Phi(x)) = 0$.

$$\frac{\nabla L(x)}{\sqrt{L(x)}} = \frac{\nabla^2 L(\Phi(x))(x-\Phi(x))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x))[x-\Phi(x), x-\Phi(x)]}} + \mathcal{O}(\frac{\boldsymbol{\zeta}^{1/2}\boldsymbol{\nu}}{\boldsymbol{\mu}} \|x-\Phi(x)\|).$$

In the above result, we have bounded the following terms:

1. We can bound $\left\| \frac{\nabla^2 L(\Phi(x))(x-\Phi(x))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x))[x-\Phi(x),x-\Phi(x)]}} \right\|$, as

$$\left\| \frac{\nabla^2 L(\Phi(x))(x-\Phi(x))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x))[x-\Phi(x), x-\Phi(x)]}} \right\| = \left\| \sqrt{2\nabla^2 L\Phi(x)} \frac{\widetilde{x}}{\|\widetilde{x}\|} \right\| \le \sqrt{2\lambda_1(x)},$$

   where $\widetilde{x} = \sqrt{\nabla^2 L(\Phi(x))}(x-\Phi(x))$.

2. Also, using the assumptions on the eignvalues of $\nabla^2 L$, $\boldsymbol{\mu}\|x-\Phi(x)\| \le \left\|\nabla^2 L(\Phi(x))(x-\Phi(x))\right\| \le \boldsymbol{\zeta}\|x-\Phi(x)\|$.

□

Lemma G.10. *Consider any point $x \in Y^{\epsilon}$. Then,*

$$\left| \left\langle v_1(x), \frac{\nabla L(x)}{\sqrt{L(x)}} \right\rangle \right| \ge \cos\theta - \mathcal{O}(\frac{\boldsymbol{\zeta}^{1/2}\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}} \|x-\Phi(x)\|),$$

where $\theta = \arctan \frac{\left\|P^{(2:M)}_{\Phi(x),\Gamma}\widetilde{x}\right\|}{|\langle v_1(x), \widetilde{x}\rangle|}$, with $\widetilde{x} = \nabla^2 L(\Phi(x))(x-\Phi(x))$.

Proof. From Lemma G.9, we have

$$\frac{\nabla L(x)}{\sqrt{L(x)}} = \frac{\nabla^2 L(\Phi(x))(x-\Phi(x))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x))[x-\Phi(x), x-\Phi(x)]}} + \mathcal{O}(\frac{\boldsymbol{\zeta}^{1/2}\boldsymbol{\nu}}{\boldsymbol{\mu}} \|x-\Phi(x)\|).$$

Also, with $\widetilde{x} = \sqrt{\nabla^2 L\Phi(x)}(x - \Phi(x))$.

$$\left\| \frac{\nabla^2 L(\Phi(x))(x-\Phi(x))}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x))[x-\Phi(x), x-\Phi(x)]}} \right\| = \left\| \sqrt{\frac{1}{2}\nabla^2 L(\Phi(x))} \frac{\widetilde{x}}{\|\widetilde{x}\|} \right\| \le \sqrt{2\lambda_1(x)}.$$

Since, $P_{\Phi(x),\Gamma}$ is the projection matrix onto the non-zero eigenvector subspace of $\nabla^2 L(\Phi(x))$, we have

$$\begin{aligned}\frac{|\langle v_1(x), \nabla L(x)\rangle|}{\sqrt{L(x)}} &= \frac{\langle v_1(x), \nabla^2 L(\Phi(x))(x-\Phi(x))\rangle}{\sqrt{\frac{1}{2}\nabla^2 L(\Phi(x))[P_{\Phi(x),\Gamma}(x-\Phi(x)), P_{\Phi(x),\Gamma}(x-\Phi(x))]}} + \mathcal{O}(\frac{\boldsymbol{\zeta}^{1/2}\boldsymbol{\nu}}{\boldsymbol{\mu}} \|x-\Phi(x)\|) \\ &\ge \sqrt{2\lambda_1(\Phi(x))}\cos\theta - \mathcal{O}(\frac{\boldsymbol{\zeta}^{1/2}\boldsymbol{\nu}}{\boldsymbol{\mu}} \|x-\Phi(x)\|).\end{aligned}$$

Hence, combining the above two equations, we must have

$$\left\| v_1(x) - \sqrt{\frac{1}{2\lambda_1(\Phi(x))}} \operatorname{sign}\left(\left\langle v_1(x), \frac{\nabla L(x)}{\sqrt{L(x)}} \right\rangle \right) \frac{\nabla L(x)}{\sqrt{L(x)}} \right\| \le \theta + \mathcal{O}(\frac{\boldsymbol{\zeta}^{1/2}\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}} \|x-\Phi(x)\|).$$

□

Lemma G.11. *For any $\overline{xy} \in Y^{\boldsymbol{\epsilon}}$ where $y = x - \eta\nabla\sqrt{L(x)}$ is the one step update on $\sqrt{L}$ loss from x, we have*

$$\begin{aligned}\Phi(y) - \Phi(x) = -\frac{\eta^2}{8} P^{\perp}_{\Phi(x),\Gamma} \nabla\lambda_1(\nabla^2 L(\Phi(x))) + \mathcal{O}(\eta^2 \boldsymbol{\xi}\theta) + \mathcal{O}(\frac{\boldsymbol{\zeta}^{3/2}\boldsymbol{\nu}\boldsymbol{\xi}\|x-\Phi(x)\|}{\boldsymbol{\mu}^{3/2}}\eta^2) + \mathcal{O}(\boldsymbol{\chi}\|x-\Phi(x)\|\eta^2) \\ + \mathcal{O}(\boldsymbol{\Upsilon}\boldsymbol{\zeta}^{3/2}(1 + \frac{\boldsymbol{\nu}\sqrt{D}}{\boldsymbol{\mu}}\|x-\Phi(x)\| + \frac{\boldsymbol{\nu}^3 D^{3/2}}{\boldsymbol{\mu}^3}\|x-\Phi(x)\|^3)\eta^3).\end{aligned}$$

Here $\theta = \arctan \frac{\left\|P^{(2:M)}_{\Phi(x),\Gamma}\widetilde{x}\right\|}{|\langle v_1(x), \widetilde{x}\rangle|}$, with $\widetilde{x} = \nabla^2 L(\Phi(x))(x - \Phi(x))$. That implies, we have

$$\|\Phi(y) - \Phi(x)\| \le \mathcal{O}((\boldsymbol{\nu}+\boldsymbol{\xi})\eta^2) + \mathcal{O}(\frac{\boldsymbol{\zeta}^{3/2}\boldsymbol{\nu}\boldsymbol{\xi}\|x-\Phi(x)\|}{\boldsymbol{\mu}^{3/2}}\eta^2) + \mathcal{O}(\boldsymbol{\chi}\|x-\Phi(x)\|\eta^2),$$

for sufficiently small η.

Proof. We outline the major difference from the proof of Lemma B.13. Using taylor expansion for the function Φ, we have

$$\begin{aligned}&\Phi(y) - \Phi(x) \\ &= \partial\Phi(x)(y-x) + \frac{1}{2}\partial^2\Phi(x)[y-x, y-x] + \text{err} \\ &= \partial\Phi(x)\left(-\eta\frac{\nabla L(x)}{2\sqrt{L(x)}}\right) + \frac{\eta^2}{2}\partial^2\Phi(x)\left[\frac{\nabla L(x)}{2\sqrt{L(x)}}, \frac{\nabla L(x)}{\|\nabla L(x)\|}\right] + \text{err} \\ &= \frac{\eta^2}{2}\partial^2\Phi(x)\left[\frac{\nabla L(x)}{2\sqrt{L(x)}}, \frac{\nabla L(x)}{2\sqrt{L(x)}}\right] + \text{err},\end{aligned}$$

where in the final step, we used the property of Φ from Lemma B.15. We can bound the error term, using the bound on $\frac{\nabla L(x)}{\sqrt{L(x)}}$ from Lemma G.9:

$$\|\text{err}\| \le \mathcal{O}(\boldsymbol{\Upsilon}\boldsymbol{\zeta}^{3/2}(1 + \frac{\boldsymbol{\nu}\sqrt{D}}{\boldsymbol{\mu}}\|x-\Phi(x)\| + \frac{\boldsymbol{\nu}^3 D^{3/2}}{\boldsymbol{\mu}^3}\|x-\Phi(x)\|^3)\eta^3).$$

First of all, the function $\Phi \in \mathcal{C}^3$, hence:

$$\partial^2\Phi(x) = \partial^2\Phi(\Phi(x)) + \mathcal{O}(\boldsymbol{\chi}\,\|x - \Phi(x)\|) = \partial^2\Phi(\Phi(x)) + \mathcal{O}(\boldsymbol{\chi}\,\|x - \Phi(x)\|).$$

Also, at $\Phi(x)$, since $v_1(x)$ is the top eigenvector of the hessian $\nabla^2 L$, we have from Corollary B.22,

$$\partial^2\Phi(\Phi(x))\left[v_1(x)v_1(x)^\top\right] = -\frac{1}{2\lambda_1(x)}\partial\Phi(\Phi(x))\partial^2(\nabla L)(\Phi(x))[v_1(x), v_1(x)].$$

From Lemma G.10, we have

$$\left\| v_1(x) - \sqrt{\frac{1}{2\lambda_1(x)}}\,\mathrm{sign}\left(\left\langle v_1(x), \frac{\nabla L(x)}{\sqrt{L(x)}}\right\rangle\right)\frac{\nabla L(x)}{\sqrt{L(x)}}\right\| \le \theta + \mathcal{O}(\frac{\boldsymbol{\zeta}^{1/2}\boldsymbol{\nu}}{\boldsymbol{\mu}^{3/2}}\,\|x - \Phi(x)\|),$$

where recall our notation of $\theta = \arctan\frac{\left\|P^{(2:M)}_{\Phi(x),\Gamma}(x-\Phi(x))\right\|}{|\langle v_1(x), x-\Phi(x)\rangle|}$.

With further simplification, it turns out that

$$\Phi(y) - \Phi(x) = -\frac{\eta^2}{8}\partial\Phi(\Phi(x))\partial^2(\nabla L)(\Phi(x))[v_1(x), v_1(x)] + \mathrm{err}' + \mathrm{err},$$

with

$$\|\mathrm{err}'\| \le \mathcal{O}(\eta^2\boldsymbol{\zeta}\boldsymbol{\xi}\theta) + \mathcal{O}(\frac{\boldsymbol{\zeta}^{3/2}\boldsymbol{\nu}\boldsymbol{\xi}}{\boldsymbol{\mu}^{3/2}}\,\|x - \Phi(x)\|\,\eta^2) + \mathcal{O}(\boldsymbol{\zeta}\boldsymbol{\chi}\,\|x - \Phi(x)\|\,\eta^2).$$

The rest of the proof will follow the same strategy as Lemma B.13.

□

H. Additional Experimental Details

H.1. Experimental details

For Figure 1: For running GD on $\sqrt{L}$, we start from $(x, y) = (14.7, 3.)$, and use a learning rate $\eta = 0.5$. For running Normalized GD on L, we start from $(x, y) = (14.7, -3)$, and use a learning rate $\eta = 5$.

For Figure 2: We start Normalized GD from $\langle v_1, \widetilde{x}(0)\rangle = 10^{-4}$, $\langle v_2, \widetilde{x}(0)\rangle = 0.45$. We use a learning rate of 1 for the optimization updates.

H.2. Code for the riemannian flow corresponding to Normalized GD

We provide the code for running a single step of the riemannian flow (Equation (4)) corresponding to Normalized GD. Each update comprises of three major steps: a) computing $\nabla^3 L(x)[v_1(x), v_1(x)]$, b) a projection onto the tangent space of the manifold, and c) few steps of gradient descent with small learning rate to drop back to manifold (if error induced by the first 2 operations).

As we are running deterministic updates, all of the steps are pretty expensive, as they scale with the number of data points in the training set. Moreover, computing $\nabla^3 L(x)[v_1(x), v_1(x)]$ requires 3 backpropagations through the entire network. Finally, we need to run few steps of full batch GD with small learning rate, to make sure we fall back to the manifold, if we go out of the manifold with the discrete riemannian updates. Thus, running each step is several times more expensive than running a full batch gradient descent.

The pseudocodeAlgorithm 4. The loss L is equal to the average of n loss functions $\ell_i : \mathbb{R}^D \to \mathbb{R}$. We can alternatively think as having n training data and each ℓ_i computes the loss of the parameters x with respect to a sample from the training set.

Algorithm 4 Riemannian Update for Normalized GD

Input: n loss functions $\ell_i : \mathbb{R}^D \to \mathbb{R}$, initial point x_{init}, maximum number of iteration T, LR η, Projection LR η_{proj}, maximum number of projection iterations T_{proj}.
$x(0) \leftarrow x_{\text{init}}$.
for $t = 1$ **to** T **do**
  $L(x(t-1)) \leftarrow \frac{1}{n}\sum_{i=1}^{n} \ell_i(x(t-1))$.
  Compute v_1, the top eigenvector of $\nabla^2 L(x(t-1))$.
  Compute $\nabla\lambda_1(x(t-1)) = \nabla^3 L(x(t-1))[v_1, v_1]$.
  Compute $P_{x(t-1),\Gamma}\nabla\lambda_1(x(t-1))$ as the projection of $\nabla\lambda_1(x(t-1))$ on the space spanned by $\nabla\ell_1(x), \cdots, \nabla\ell_n(x)$.
  Compute $P^{\perp}_{x(t-1),\Gamma}\nabla\lambda_1(x(t-1)) = \nabla\lambda_1(x(t-1)) - P_{x(t-1),\Gamma}\nabla\lambda_1(x(t-1))$.
  $y(0) \leftarrow x(t-1) - \frac{\eta}{\lambda_1(x(t-1))} P^{\perp}_{x(t-1),\Gamma}\nabla\lambda_1(x(t-1))$.
  for $\widetilde{t} = 1$ **to** T_{proj} **do**
    {The next few steps involve GD to move back to the manifold.}
    $L(y(\widetilde{t}-1)) \leftarrow \frac{1}{n}\sum_{i=1}^{n} \ell_i(y(\widetilde{t}-1))$.
    $y(\widetilde{t}) = y(\widetilde{t}-1) - \eta_{\text{proj}}\nabla(L(y(\widetilde{t}-1)))$.
  end for
  $x(t) \leftarrow y(T_{\text{proj}})$.
end for