HOW DOES SHARPNESS-AWARE MINIMIZATION MINIMIZE SHARPNESS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Sharpness-Aware Minimization (SAM) is a highly effective regularization technique for improving the generalization of deep neural networks for various settings. However, the underlying working of SAM remains elusive because of various intriguing approximations in the theoretical characterizations. SAM intends to penalize a notion of sharpness of the model but implements a computationally efficient variant; moreover, a third notion of sharpness was used for proving generalization guarantees. The subtle differences in these notions of sharpness can indeed lead to significantly different empirical results. This paper rigorously nails down the exact sharpness notion that SAM regularizes and clarifies the underlying mechanism. We also show that the two steps of approximations in the original motivation of SAM individually lead to inaccurate local conclusions, but their combination accidentally reveals the correct effect, when full-batch gradients are applied. Furthermore, we also prove that the stochastic version of SAM in fact regularizes the third notion of sharpness mentioned above, which is most likely to be the preferred notion for practical performance. The key mechanism behind this intriguing phenomenon is the alignment between the gradient and the top eigenvector of Hessian when SAM is applied.

1 INTRODUCTION

Modern deep nets are often overparametrized and have the capacity to fit even randomly labeled data (Zhang et al., 2016). Thus, a small training loss does not necessarily imply good generalization. Yet, standard gradient-based training algorithms such as SGD are able to find generalizable models. Recent empirical and theoretical studies suggest that generalization is well-correlated with the sharpness of the loss landscape at the learned parameter (Keskar et al., 2016; Dinh et al., 2017; Dziugaite & Roy, 2017; Neyshabur et al., 2017; Jiang et al., 2019). Partly motivated by these studies, Foret et al. (2021); Wu et al. (2020); Zheng et al. (2021) propose to penalize the sharpness of the landscape to improve the generalization. We refer this method to *Sharpness-Aware Minimization* (SAM) and focus on the version of Foret et al. (2021) in this paper.

Despite its empirical success, the underlying working of SAM remains elusive because of the various intriguing approximations made in its derivation and analysis. There are three different notions of sharpness involved – SAM intends to optimize the first notion, the sharpness along the worst direction but actually implements a computationally efficient notion, the sharpness along the direction of the gradient. But in the analysis, a third notion of sharpness is actually used to prove generalization guarantees, which admits the first notion as an upper bound. The subtle difference between the three notions can lead to very different explicit biases. (see Figure 1 for demonstraion)

More concretely, let L be the training loss, x be the parameter and ρ be the *perturbation radius*, a hyperparameter requiring tuning. The first notion corresponds to the following optimization problem (1), where we call $R_\rho^{\text{Max}}(x) = L_\rho^{\text{Max}}(x) - L(x)$ the *worst-direction sharpness* at x and thus SAM is intended to minimize the original training loss plus the worst-direction sharpness at x.

$$\min_x L_\rho^{\text{Max}}(x), \quad \text{where} \quad L_\rho^{\text{Max}}(x) = \max_{\|v\|_2 \le 1} L(x + \rho v), \tag{1}$$

However, even evaluation of $L_\rho^{\text{Max}}(x)$ is computationally expensive, not to mention optimization. Thus Foret et al. (2021); Zheng et al. (2021) propose to approximate the worst perturbation direction by the direction of the gradient and implement the second notion of sharpness, which corresponds

Type of Sharpness	Symbol	Definition	Explicit Bias Among Minimizers
Worst-direction	L_ρ^{Max}	$\max_{\|v\|_2 \le 1} L(x+\rho v)$	$\min_x \lambda_1(\nabla^2 L(x))$
Ascent-direction	L_ρ^{Asc}	$L\left(x+\rho \frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right)$	$\min_x \lambda_{\text{min}}(\nabla^2 L(x))$
Average-direction	L_ρ^{Avg}	$\mathbb{E}_{v\sim N(0,I)} L(x+\rho \frac{v}{\|v\|_2})$	$\min_x \text{Tr}(\nabla^2 L(x))$

Table 1: Definitions and explicit biases of different notions of sharpness. Here λ_{min} denotes to the smallest *non-zero* eigenvalue.

to (2). We call $R_\rho^{\text{Asc}}(x) = L_\rho^{\text{Asc}}(x) - L(x)$ the *ascent-direction sharpness* at x.

$$\min_x L_\rho^{\text{Asc}}(x), \quad \text{where} \quad L_\rho^{\text{Asc}}(x) = L\left(x+\rho\frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right). \tag{2}$$

Intriguingly, in the generalization analysis of SAM by Foret et al. (2021); Wu et al. (2020), the first notion of sharpness, *i.e.*, the worst-direction sharpness, is only used for upper bounding the third notion of sharpness via the PAC Bayesian theory (McAllester, 2003). We call the third notion $R_\rho^{\text{Avg}}(x) = L_\rho^{\text{Avg}}(x) - L(x)$ the *average-direction sharpness* at x, where $L_\rho^{\text{Avg}}(x) = \mathbb{E}_{g\sim N(0,I)} L\left(x+\rho g/\|g\|\right)$.

For further acceleration, Foret et al. (2021); Zheng et al. (2021) omit the gradient through other occurrence of x and approximate the gradient of ascent-direction sharpness by gradient taken after one-step ascent, *i.e.*, $\nabla L_\rho^{\text{Asc}}(x) \approx \nabla L\left(x+\rho\frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right)$ and derive the update rule of SAM, where η is the learning rate.

$$\text{Sharpness-Aware Minimization (SAM):} \quad x(t+1) = x(t) - \eta \nabla L\left(x+\rho\frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right) \tag{3}$$

In this paper, we analyze the explicit bias of various notions of sharpness and the optimization trajectory of SAM. Our analysis is performed for small perturbation radius ρ and learning rate η under the setting where the minimizers of loss form a manifold following Fehrman et al. (2020); Li et al. (2021) In particular, we make the following theoretical contributions.

1. We prove that full-batch SAM does minimize worst-direction sharpness. (Theorem 4.9)
2. Surprisingly, when batch size is 1, SAM minimizes average-direction sharpness. (Theorem 4.9)
3. We characterize the explicit biases of three notions of sharpness among minimizers when perturbation radius ρ goes to zero. (Theorems 4.5 to 4.7, also see Table 1) Surprisingly, both heuristic approximations made for the update rule of SAM lead to inaccurate solutions, that is, (1) minimizing worst-direction sharpness and ascent-direction sharpness induce different biases among minimizers, and (2) SAM doesn't minimize ascent-direction sharpness.

The key mechanism behind this implicit bias of SAM is the implicit alignment between the gradient and the top eigenvector of Hessian when running SAM.

2 Related Works

Sharpness and Generalization. The study on the connection between sharpness and generalization can be traced back to Hochreiter & Schmidhuber (1997). Keskar et al. (2016) observed a positive correlation between the batch size, the generalization error, and the sharpness of the loss landscape when changing the batch size. Jastrzebski et al. (2017) extends this by finding a correlation of the sharpness and the ratio between learning rate to batch size. Dinh et al. (2017) show that one can easily construct networks with good generalization but with arbitrary large sharpness by reparametrization. Dziugaite & Roy (2017); Neyshabur et al. (2017); Wei & Ma (2019) give theoretical guarantees on the generalization error using sharpness-related measures. Jiang et al. (2019) perform a large-scale empirical study on various generalization measures and showed that sharpness-based measures have the highest correlation with generalization.

Background on Sharpness Aware-Minimization. Foret et al. (2021); Zheng et al. (2021) concurrently proposed to minimize the loss at the perturbed from current parameter towards the worst direction to improve generalization. Wu et al. (2020) propose the almost identical method for a different purpose, robust generalization of adversarial training. Kwon et al. (2021) proposed a different metric for SAM to fix the rescaling problem pointed out by Dinh et al. (2017). Liu et al. (2022) propose an more computationally efficient version of SAM. Zhao et al. (2022) proposed to improve generalization by penalizing gradient norm. Their proposed algorithm can be viewed as a

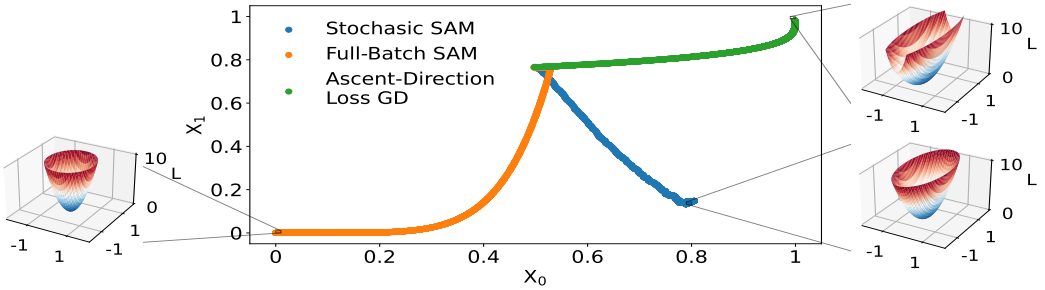

Figure 1: Visualization of the different biases of different sharpness notions on a 4D-toy example. For $x, y \in \mathbb{R}^2$, consider loss $L(x, y) = F_0(x)y_0^2 + F_1(x)y_1^2$ with $F_0(x) = x_0^2 + 6x_1^2 + 8$ and $F_1(x) = 4(1 - x_0)^2 + (1 - x_1)^2 + 1$. L has a zero loss manifold $\{y = 0\}$ and the eigenvalues of its Hessian on manifold are $F_0(x)$ and $F_1(x)$ with $F_0(x) \geq 8 > 6 \geq F_1(x)$ on $[0, 1]^2$. As our theory predicts, (1). full-batch SAM (3) finds the minimizer with the smallest top eigenvalue, $F_0(x)$; (2). GD on Ascent-direction Loss L_ρ^{Asc} (2) finds the minimizer with the smallest bottom eigenvalue, $F_0(x)$. (3). Stochastic SAM (16) (with $L_0(x, y) = F_0(x)x_0^2, L_1(x, y) = F_1(x)y_1^2$) finds the minimizer with smallest trace of Hessian. Loss landscape $L(x, \cdot)$ are visualized as 3D plots at converged x to illustrate the different biases. (*cf.*Table 1)

generalization of SAM. Andriushchenko & Flammarion (2022) studied a variant of SAM where the step size of ascent step is ρ instead of $\frac{\rho}{\|\nabla L(x)\|_2}$. They showed that for a simple model this variant of SAM has a stronger regularization effect when batch size is 1 compared to the full-batch case and argued that this might be the explanation that SAM generalizes better with small batch sizes.

3 Notations and Assumptions

For any integer k, we define $\mathcal{C}^k$ as the set of k times continuously differentiable functions. For any mapping F, we define $\partial F(x)[u]$ and $\partial^2 F(x)[u, v]$ as the first and second order directional derivative of $F(x)$ along the direction u (and v). Given a differential submanifold Γ of $\mathbb{R}^D$ and a point $x \in \Gamma$, define $P_{x,\Gamma}$ as the projection operator onto the manifold of the normal space of Γ at x and $P_{x,\Gamma}^{\perp} = I_D - P_{x,\Gamma}$. We fix our initialization as x_{init} and our loss function as $L : \mathbb{R}^D \to \mathbb{R}$. Given the loss function, the gradient flow can be defined as mapping $\phi : \mathbb{R}^D \times [0, \infty) \to \mathbb{R}^D$ satisfying $\phi(x, \tau) = x - \int_0^\tau \nabla L(\phi(x, t))dt$. We further define the limiting map $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ as $\Phi(x) = \lim_{\tau \to \infty} \phi(x, \tau)$. For any positive definite symmetry matrix $A \in \mathbb{R}^{D \times D}$, define $\{\lambda_i(A), v_i(A)\}_{i \in [D]}$ as all its eigenvalues and eigenvectors satisfying $\lambda_1(A) \geq \lambda_2(A) ... \geq \lambda_3(A)$ and $\|v_i(A)\|_2 = 1$. Our analysis assumes sufficiently small η and ρ and uses $O(\cdot)$ to hide constant.

Following Fehrman et al. (2020); Li et al. (2021); Arora et al. (2022), we make the below assumption.

Assumption 3.1. *Assume loss $L : \mathbb{R}^D \to \mathbb{R}$ belongs to $\mathcal{C}^4$, and there exists a manifold Γ that is $D - M$ 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 global minimizer of L, $L(x) = 0$ and rank$(\nabla^2 L(x)) = M$.*

The smoothness assumption is meet with networks with smooth activation functions and the existence of the manifold is due to the vast overparameterization of the modern neural network. The full rank assumption is necessary for the analysis to guarantee the differentiability of Φ.

Definition 3.2 (Attraction Set)**.** *Let U be the attraction set of Γ under gradient flow, that is, a neighborhood of Γ containing all points starting from which gradient flow w.r.t. loss L converges to some point in Γ, or mathematically, $U \triangleq \{x \in \mathbb{R}^D | \Phi(x)$ exists and $\Phi(x) \in \Gamma\}$.*

Assumption 3.1 implies that U is open and Φ is in $\mathcal{C}^3$ on U (from Lemma B.15 Arora et al. (2022)).

4 Explicit and Implicit Bias in the Full-batch Setting

Section 4.1 provides a general theorem to properly analyze the explicit bias of various notions of sharpness among different minimizers. We then apply our machinery on *ascent-direction sharpness* and *worst-direction sharpness* and show that they have different explicit biases. In Section 4.2 we provide our main theorem in the full-batch setting, that SAM implicitly minimizes the worst-direction sharpness, via characterizing its limiting dynamics as learning rate ρ and η goes to 0 with

a Riemmanian gradient flow with respect to the top eigenvalue of the Hessian of the loss on the manifold of global minimizers. In Section 4.3 we sketch the proof of the implicit bias of SAM and identified a key property behind the implicit bias, which is the implicit alignment between the gradient and the top eigenvector of the Hessian throughout the training.

4.1 Worst- and Ascent-direction Sharpness have Different Explicit Bias

The intuition of approximating R_ρ^{Max} by R_ρ^{Asc} comes from the following Taylor expansions (Foret et al., 2021; Wu et al., 2020), which works for small ρ:

$$R_\rho^{\text{Max}}(x) = \sup_{\|v\|_2 \le 1} L(x+\rho v) - L(x) = \sup_{\|v\|_2 \le 1} \left(\rho v^\top \nabla L(x) + \frac{\rho^2}{2} v^\top \nabla^2 L(x) v + O(\rho^3) \right) \quad (4)$$

$$R_\rho^{\text{Asc}}(x) = L\left(x + \rho \frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right) - L(x) = \rho \|\nabla L(x)\|_2 + \frac{\rho^2}{2} \frac{\nabla L(x)^\top \nabla^2 L(x) \nabla L(x)}{\|\nabla L(x)\|_2^2} + O(\rho^3) \quad (5)$$

Here, the preference among the global minima is what we are mainly concerned with. Since $\sup_{\|v\|_2 \le 1} v^\top \nabla L(x) = \|\nabla L(x)\|_2$ when $\|\nabla L(x)\|_2 > 0$, the leading terms in Equations (4) and (5) are both the first order term, $\rho \|\nabla L(x)\|_2$, and are the same. However, it is erroneous to think that the first order term decides the explicit bias, as the first order term $\|\nabla L(x)\|_2$ vanishes at the global minimizers of the loss L and thus the second order term becomes the leading term. Any global minimizer x of the original loss L is an $O(\rho^2)$-approximate minimizer of the sharpness-aware loss because $\nabla L(x) = 0$. Therefore, the sharpness-aware loss needs to be of order ρ^2 so that we can guarantee the second-order terms in Equation (4) and/or Equation (5) to be non-trivially small.

In this section, we aim to understand under Assumption 3.1 what the explicit biases of various notions of sharpness among different minimizers are. Theorem 4.3 will be our main theoretic tool to analyze the explicit bias for small perturbation radius ρ. Below we first introduce some notations and assumption.

Notation for Regularizers. Let $R_\rho : \mathbb{R}^D \to \mathbb{R} \cup \{\infty\}$ be a family of regularizers parameterized by ρ. If R_ρ is not well-defined at some x, then we let $R_\rho(x) = \infty$. This convention will be useful when analyzing ascent-direction sharpness $R_\rho^{\text{Asc}} = L_\rho^{\text{Asc}} - L$ which is not defined when $\nabla L(x) = 0$. This convention will not change the minimizers of the regularized loss. Intuitively, a regularizer should always be non-negative, but however, when far away from manifold, regularizer $R_\rho(x)$ can actually be negative, *e.g.*, $R_\rho^{\text{Avg}}(x) \approx \frac{\rho^2}{2D} \text{Tr}[\nabla^2 L(x)]$. Therefore we make the following assumption to allow the regularizer to be mildly negative.

Assumption 4.1. *Suppose for any bounded closed set $B \subset U$, there exists $C > 0$, such that for sufficiently small ρ, $\forall x \in B, R_\rho(x) \ge -C\rho^2$.*

The following concept of limiting regularizer is of crucial role in our analysis.

Definition 4.2 (Limiting Regularizer)**.** *We define the* limiting regularizer *of $\{R_\rho\}$ as the function*

$$S : \Gamma \to \mathbb{R}, \quad S(x) = \lim_{\rho \to 0} \lim_{r \to 0} \inf_{\|x' - x\|_2 \le r} R_\rho(x')/\rho^2.$$

We say the limiting regularizer S of $\{R_\rho\}$ is a good *around some $x^* \in \Gamma$, if there is an open set V containing x^*, such that S is a non-negative continuous function in V and for any $\epsilon > 0$, there is some $\rho_{x^*} > 0$, it holds that $\forall x \in \Gamma \cap V, \rho \le \rho_{x^*}, \left| S(x) - \inf_{\|x'-x\|_2 \le \epsilon \cdot \rho} R_\rho(x')/\rho^2 \right| < \epsilon$. We say the limiting regularizer S is* good *on Γ, if S is good around every point $x \in \Gamma$.*

The high-level intuition behind the definition of limiting regularizer is that we want to use it to capture the second order term in the Taylor expansion of regularizer R_ρ when $\rho \to 0$. When the second order term is continuous in x, the definition of $S(x)$ can also be simplified as $R_\rho(x)/\rho^2$. The intuition of the concept of a good limiting regularizer is that, the regularizer should not change very fast, especially in an $O(\rho)$ neighborhood of the minimizer. If so, the minimizer of the regularized loss may be $\Omega(\rho)$ away from any minimizer to reduce the regularizer at the cost of increasing the original loss, which makes the limiting regularizer unable to capture the explicit bias of the regularizer.

Theorem 4.3. *Let U' be any bounded open set such that its closure $\overline{U'}$ is contained in U and that $\overline{U'} \cap \Gamma = \overline{U' \cap \Gamma}$. Then for any family of parametrized regularizers $\{R_\rho\}$ admitting a good limiting*

regularizer on Γ and satisfying Assumption 4.1 and any $\epsilon \geq 0$, there is a $\rho_0 > 0$, such that for all $u \in U'$ and $\rho < \rho_0$, it holds that

$$L(u) + R_\rho(u) \leq \inf_{x \in U'} (L(x) + R_\rho(x)) + \epsilon\rho^2 + o(\rho^2)$$
$$\iff \left(L(u) - \inf_{x \in U'} L(x)\right) + \left|R_\rho(u) - \rho^2 \inf_{x \in U' \cap \Gamma} S(x)\right| \leq \epsilon\rho^2 + o(\rho^2)$$

For the applications we are interested in in this paper, the good limiting regularizer S can be continuously extended to the entire space $\mathbb{R}^D$. In such a case, the implication of "$\Longrightarrow$" of Theorem 4.3 also admits the following alternative form which doesn't involve R_ρ. Corollary 4.4 implies minimizing regularized loss $L(x) + R_\rho(x)$ is equivalent to minimizing the limiting regularizer of $\{R_\rho\}_\rho$, $S(x)$ on the global minimizer manifold Γ.

Corollary 4.4. *Under the setting of Theorem 4.3, let $\overline{S}$ be an continuous extension of S to $\mathbb{R}^d$, if $L(u) + R_\rho(u) \leq \inf_{x \in U'} (L(x) + R_\rho(x)) + \epsilon\rho^2 + o(\rho^2)$, then we have that $L(u) - \inf_{x \in U'} L(x) = O(\rho^2)$ and that $\left|\overline{S}(u) - \inf_{x \in U' \cap \Gamma} \overline{S}(x)\right| = \epsilon + o(1)$.*

Corollary 4.4 suggests a sharp phase transition of the property of the solution of $\min_x L(x) + R_\rho(x)$ when the optimization error drops from $\omega(\rho^2)$ to $O(\rho^2)$. when the optimization error is larger than $\omega(\rho^2)$, no regularization effect happens and any minimizer satisfies the requirement. When the error becomes $O(\rho^2)$, there is a non-trivial restriction on the (extended) limiting regularizer.

Below we show all of worst-, ascent- and average-direction sharpness admit good limiting regularizers. Thus we can apply Corollary 4.4 to characterize their explicit biases, which are all different.

Theorem 4.5. *Worst-direction sharpness R_ρ^{Max} satisfies Assumption 4.1 and admits $\lambda_1(\nabla^2 L(\cdot))/2$ as a good limiting regularizer on Γ.*

Theorem 4.6. *Ascent-direction sharpness R_ρ^{Asc} satisfies Assumption 4.1 and admits $\lambda_M(\nabla^2 L(\cdot))/2$ as a good limiting regularizer on Γ.*

Theorem 4.7. *Average-direction sharpness R_ρ^{Avg} satisfies Assumption 4.1 and admits $\operatorname{Tr}(\nabla^2 L(\cdot))/(2D)$ as a good limiting regularizer on Γ.*

When R_ρ is continuous at some $x \in \Gamma$, the definition of $S(x)$ can be simplified as $\lim_{\rho \to 0} R_\rho(x)/\rho^2$. Worst- and average- direction sharpness fall into this type and the limiting regularizer can be solved straightforwardly by Taylor expansion, which is well-known in literature (Keskar et al., 2016; Kaur et al., 2022; Zhuang et al., 2022; Orvieto et al., 2022). Theorems 4.5 and 4.7 can be viewed as a restatement of the existing results in our manifold setting.

Our major contribution in this subsection is the analysis for ascent-direction sharpness, which is more tricky as $R_\rho^{\text{Asc}}(x) = \infty$ and thus is not continuous for any $x \in \Gamma$. To minimize R_ρ^{Asc} around x, we can pick $x' \to x$ to make $\|\nabla L(x)\|_2 \to 0$ but not equal to 0. By (5), we have $R_\rho^{\text{Asc}}(x') \approx \rho^2/2 \cdot \nabla L(x')^\top \nabla^2 L(x) \nabla L(x')/\|\nabla L(x')\|^2$. Here the crucial step of the proof is that because of Assumption 3.1, $\nabla L(x)/\|\nabla L(x)\|_2$ must almost lie in the column span of $\nabla^2 L(x)$, and thus implies $\inf_{x'} \nabla L(x')^\top \nabla^2 L(x) \nabla L(x')/\|\nabla L(x')\|^2 \overset{\rho \to 0}{\to} \lambda_M(\nabla^2 L(x))$. The above alignment property between the gradient and the column space of Hessian can be checked directly for any non-negative quadratic function and the maximal Hessian rank assumption in Assumption 3.1 ensures this property extends to general losses.

4.2 SAM Provably Decreases Worst-direction Sharpness Locally

Though ascent-direction sharpness has different explicit bias from worst-direction sharpness, in this subsection we will show that surprisingly, SAM, an heuristic method designed to minimize ascent-direction sharpness, provably decreases worst-direction sharpness. The main result here is an exact characterization of the trajectory of SAM (3) via the following ordinary differential equation (ODE) (6), when learning rate η and perturbation radius ρ are small and the initialization $x(0) = x_{\text{init}}$ is in U. We call the solution of (6) the *limiting flow* of SAM, which is exactly the Riemannian Gradient Flow on the manifold Γ with respect to $\lambda_1(\nabla^2 L(\cdot))$. In other words, the ODE (6) is essentially

a projected gradient descent algorithm with loss $\lambda_1(\nabla^2 L(\cdot))$ on the constraint set Γ and an infinitesimal learning rate.

$$X(\tau) = X(0) - \frac{1}{2}\int_{s=0}^{\tau} P^{\perp}_{X(s),\Gamma}\nabla\lambda_1(\nabla^2 L(X(s)))ds, X(0) = \Phi(x_{\text{init}}). \tag{6}$$

Note $\lambda_1(\nabla^2 L(x))$ may not be differentiable at x if $\lambda_1(\nabla^2 L(x)) = \lambda_2(\nabla^2 L(x))$, thus to ensure the (6) is well-defined, we assume there is a positive eigengap for L on Γ. Assuming ODE (6) has a solution till time T_3, we have Theorem 4.9, which is the main theorem of this section.

Assumption 4.8. *For $x \in \Gamma$, there exists a positive eigengap,i.e., $\lambda_1(\nabla^2 L(x)) > \lambda_2(\nabla^2 L(x))$.*

Theorem 4.9 (Main, Theorems C.31 and C.32 stated informally)**.** *Let $\{x(t)\}$ be the iterates of SAM (3) with $x(0) = x_{init} \in U$, then under Assumptions 3.1 and 4.8, for all η, ρ such that $\eta\ln(1/\rho)$ and ρ/η are sufficiently small, the dynamics of SAM can be split into two phases:*

- ***Phase I:*** *SAM follows Gradient Flow with respect to L until entering an $O(\eta\rho)$ neighborhood of the manifold Γ in $\tilde{O}(\frac{1}{\eta})$ steps;*
- ***Phase II:*** *SAM tracks the solution X of (6), the Riemannian Gradient Flow with respect to $\lambda_1(\nabla^2 L(\cdot))$ in the $O(\eta\rho)$ neighborhood in the sense that $\max_{0\le T\le T_3}\|\Phi(x(\lceil T/(\eta\rho^2)\rceil)) - X(T)\| = O((\eta+\rho)\log(1/\eta\rho))$. Moreover, the angle between $\nabla L(x(t))$ and $v_1(\nabla^2 L(x(t)))$ is at most $O(\rho)$.*

Theorem 4.9 shows that SAM decreases the largest eigenvalue of Hessian of loss locally around the manifold of global minimizers. As a corollary of Theorem 4.9, we can also show that the largest eigenvalue of the limiting flow closely tracks the regularized training loss.

Corollary 4.10. *For all $T_3' > 0$, for all ρ, η such that $\eta\ln(1/\rho)$ and ρ/η are sufficiently small we have $\forall T_3' < \eta\rho^2 t \le T_3$, $\|R_\rho^{\text{Max}}(x(t)) - \rho^2\lambda_1(\nabla^2 L(X(\eta\rho^2 t))/2\| = \tilde{O}(\eta\rho^2)$*

Recall Theorem 4.6 shows that the largest eigenvalue of Hessian is the limiting regularizer of the *worst-direction sharpness*, leveraging the equivalence relationship in Theorem 4.3, below we show that full-batch SAM provably minimizes *worst-direction sharpness* if we additionally assume the limiting flow converges to a local minimizer of the top eigenvalue of Hessian.

Corollary 4.11. *Define U' as in Theorem 4.3, suppose $X(\infty) = \lim_{t\to\infty} X(t)$ exists and is a minimizer of $\lambda_1(\nabla^2 L(x))$ in $U' \cap \Gamma$, for all $\epsilon > 0$, there exists a constant $T > 0$, then for all ρ, η such that $\eta\ln(1/\rho)$ and ρ/η are sufficiently small we have $L_\rho^{\text{Max}}(x(\lceil T/(\eta\rho^2)\rceil)) \le \epsilon\rho^2 + \inf_{x\in U'} L_\rho^{\text{Max}}(x)$*

4.3 Proof Sketch for Phase II in Theorem 4.9

Now we sketch the proof for the ODE-based characterization of the trajectory of SAM in Phase II. The framework of the analysis is similar to Arora et al. (2022); Lyu et al. (2022), where the high-level idea is to use $\Phi(x(t))$ as a proxy for $x(t)$ and study the dynamics of $\Phi(x(t))$ via Taylor expansion, which turns out to be dependent on the alignment between the gradient and the eigenvectors of the Hessian. In particular, like Arora et al. (2022); Lyu et al. (2022), we show that the gradient aligns to the top eigenvector of Hessian, and thus encourages the SAM dynamics to reduce the top eigenvalue of the Hessian.

Taylor Expansion on Φ. In Phase II, it can be shown that $\|x(t) - \Phi(x(t))\| = O(\eta\rho)$ holds for every step, this implies $\|x(t+1) - x(t)\|_2 = O(\rho\eta)$. (See Lemma C.12) Therefore we have that

$$\begin{aligned}\Phi(x(t+1)) - \Phi(x(t)) =& -\eta\partial\Phi(x(t))(x(t+1) - x(t)) + O(\eta\|x(t+1) - x(t)\|^2)\\ =& -\eta\partial\Phi(x(t))\nabla L\left(x + \rho\frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right) + O(\eta^3\rho^2)\end{aligned} \tag{7}$$

Now we apply Taylor expansion on $\nabla L\left(x + \rho\frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right)$ around x and get

$$\begin{aligned}&\nabla L\left(x + \rho\frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right)\\ =&\nabla L(x) - \rho\nabla^2 L(x)\frac{\nabla L(x)}{\|\nabla L(x)\|_2} + \frac{\rho^2}{2}\partial^2(\nabla L)(x)\left[\frac{\nabla L(x)}{\|\nabla L(x)\|_2}, \frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right] + O(\rho^3).\end{aligned} \tag{8}$$

Lemma 4.12 (Lemma B.16 Arora et al. (2022)). *For $x \in U$, $\partial\Phi(x)\nabla L(x) = 0$, $\partial\Phi(x)\nabla^2 L(x)\nabla L(x) = -\partial^2\Phi(x)[\nabla L(x), \nabla L(x)]$.*

Now we plug (8) into (7) and simplify the expression using Lemma 4.12. We have that

$$\begin{aligned}\Phi(x(t+1)) - \Phi(x(t)) =& 0 + O(\rho \left\|\partial\Phi(x(t))\nabla^2 L(x(t))\right\| \|\nabla L(x(t))\|_2) \\ &- \frac{\eta\rho^2}{2}\partial\Phi(x)\partial^2(\nabla L)(x)\left[\frac{\nabla L(x)}{\|\nabla L(x)\|_2}, \frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right] + O(\eta^3\rho^2 + \eta\rho^3)\end{aligned} \tag{9}$$

where in the last step we use the property that $\|x(t) - \Phi(x(t))\| = O(\eta\rho)$, which further implies $\|\nabla L(x(t))\|_2 = O(\eta\rho)$ in Phase II (See Lemma C.7).

Lemma 4.13. *For $x \in \Gamma$, $\partial\Phi(x) = P^{\perp}_{x,\Gamma}$, $\partial\Phi(x)\nabla^2 L(x) = 0$, where $P^{\perp}_{x,\Gamma}$ is the orthogonal projection matrix of the tangent space of Γ at x.*

Next we use Lemma 4.13 to further simplify (9). Since the entire phase II happens in an $O(\eta\rho)$-neighborhood of manifold Γ, we have that $\|x(t) - \Phi(x(t))\|_2 = O(\eta\rho)$, thus both $\left\|\partial\Phi(x(t))\nabla^2 L(x(t))\right\|$ and $\|\nabla L(x(t))\|_2$ are $O(\eta\rho)$. So far, the proof is almost completed, with the implicit alignment between gradient and top eigenvector of Hessian being the last missing piece. Suppose we have $\left\|\frac{\nabla L(x)}{\|\nabla L(x)\|_2} - v_1(\nabla^2 L(x))\right\| = O(\rho)$, then it holds that

$$\begin{aligned}\Phi(x(t+1)) - \Phi(x(t)) &= -\frac{\eta\rho^2}{2}\partial\Phi(x)\partial^2(\nabla L)(x)\left[\frac{\nabla L(x)}{\|\nabla L(x)\|_2}, \frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right] + O(\eta^3\rho^2 + \eta\rho^3) \\ &= -\frac{\eta\rho^2}{2}\partial\Phi(x)\partial^2(\nabla L)(x)\left[v_1(\nabla^2 L(x)), v_1(\nabla^2 L(x))\right] + O(\eta^3\rho^2 + \eta\rho^3) \\ &= -\frac{\eta\rho^2}{2}\partial\Phi(x(t))\nabla\lambda_1(\nabla^2 L(x(t))) + O(\eta^3\rho^2 + \eta\rho^3) \\ &= -\frac{\eta\rho^2}{2}\partial\Phi(\Phi(x(t)))\cdot\nabla\lambda_1(\nabla^2 L(x))|_{x=\Phi(x(t))} + O(\eta^3\rho^2 + \eta\rho^3),\end{aligned} \tag{10}$$

where the second to last step we use the following property about the derivative of eigenvalue (Lemma C.69) and the last step is due to Taylor expansion.

Implicit Hessian-gradient Alignment. It remains to explain why the gradient implicitly aligns to the top eigenvector of Hessian, which is the key component of the analysis in Phase II. The proof strategy here is to first show alignment for a quadratic loss function, and then generalize its proof to general loss functions satisfying Assumption 3.1. Below we first give the formal statement of the implicit alignment on quadratic loss Theorem 4.14. Note this alignment property is an implicit property of the SAM algorithm – it is not due to the fact that SAM is an approximation of GD on L^{Asc}_{ρ}, because optimizing L^{Asc}_{ρ} would rather make the gradient align to the smallest eigenvector!

Theorem 4.14. *Suppose A is a positive definite symmetric matrix with unique top eigenvalue. Consider running SAM on loss $L(x) := \frac{1}{2}x^T Ax$ as (11), then for almost every $x(0)$, we have $x(t)$ converges in direction to $v_1(A)$ and $\lim_{t\to\infty}\|x(t)\| = \frac{\eta\rho\lambda_1(A)}{2-\eta\lambda_1(A)}$ with $\eta\lambda_1(A) < 1$.*

$$x(t+1) = x(t) - \eta A\left(x(t) + \rho\frac{Ax(t)}{\|Ax(t)\|}\right), \tag{11}$$

Equivalently, we can reformulate the update rule through the lens of the gradient $\nabla L(x) = Ax$:

$$\nabla L(x(t+1)) = \nabla L(x(t)) - \eta(\nabla^2 L(x(t)))^2\left(\nabla L(x(t)) + \rho\frac{\nabla L(x(t))}{\|\nabla L(x(t)))\|}\right), \tag{12}$$

and the result of Theorem 4.14 becomes the alignment between gradient and the top eigenvector of Hessian. This property can be generalized to the analysis of general loss as the dynamic of SAM near a fixed point on Γ is similar to the quadratic case with small perturbation. To see this, we apply Taylor expansion on the update rule of SAM (3):

$$x(t+1) = x(t) - \eta\nabla L(x(t)) - \eta\rho\nabla^2 L(x(t))\frac{\nabla L(x(t))}{\|\nabla L(x(t))\|_2} + O(\eta\rho^2). \tag{13}$$

Since phase II happens in an $O(\eta\rho)$-neighborhood of manifold Γ, we have $\|x(t+1)-x(t)\|_2 = O(\eta\rho)$. Then by (13) and Taylor expansion on $\nabla L(x(t+1))$ at $x(t)$, we have that

$$\nabla L(x(t+1)) = \nabla L(x(t)) - \nabla^2 L(x(t))\,(x(t+1)-x(t)) + O(\eta^2\rho^2) \tag{14}$$

$$= \nabla L(x(t)) - \eta(\nabla^2 L(x(t)))^2 \left(\nabla L(x(t)) + \rho\frac{\nabla L(x(t))}{\|\nabla L(x(t)))\|}\right) + O(\eta\rho^2) \tag{15}$$

Equation (14) is a $O(\eta\rho^2)$-perturbed version of the update rule in the quadratic case. Note this is a higher order term comparing to the other two terms, which have orders $\Theta(\eta^2\rho)$ and $\Theta(\eta\rho)$ respectively, the implicit alignment between Hessian and gradient happens for the same reason as in the quadratic case. We further show once this alignment happens, it will be kept until the end of our analysis, which is $\Theta(\eta^{-1}\rho^{-2})$ steps.

5 Explicit and Implicit bias in the Stochastic Setting

In practice, people usually use SAM in the stochastic mini-batch setting, and the test accuracy improves as the batch size decreases (Foret et al., 2021). Towards explaining this phenomenon, Foret et al. (2021) argue intuitively that stochastic SAM minimizes stochastic worst-direction sharpness. Given our results in Section 4, it's natural to ask if we can justify the above intuition by showing the implicit Hessian-gradient alignment in the stochastic setting. Unfortunately, such alignment is not possible in the most general setting. For example, take a simple quadratic loss $L(x) = \frac{L_1(x)+L_2(x)}{2}$, where $L_k(x) = 0.5x^\top A_k x$ and A_k is a positive definite matrix for $k = 1, 2$. If A_1 and A_2 have different top eigendirection, then no x can simultaneously satisfy that $\nabla L_k(x) = A_k x$ aligns to the top eigenvector of $\nabla^2 L_k(x) = A_k$. Yet when the batch size is 1, we can prove rigorously that stochastic SAM minimizes stochastic worst-direction sharpness (Section 5.2), which is the main result in this section.

Below we start by introducing our setting for SGD with batch size 1. We still need Assumption 3.1 in this section. We first by analyzing the explicit bias of the stochastic ascent- and worst-direction sharpness in Section 5.1 via the tools developed in Section 4.1. It turns out they are all proportional to the trace of hessian as $\rho \to 0$. In Section 5.2, we show stochastic SAM locally decreases trace of Hessian.

Setting Let $f_k(x)$ be the model output on kth data where f_k is a $\mathcal{C}^4$-smooth function and y_k be the ith label for $l = 1, \ldots, M$. We define the loss on kth data as $L_k(x) = \ell(f_k(x), y_k)$ and the total loss $L = \sum_{k=1}^M L_k/M$, where $\ell(y', y)$ is a $\mathcal{C}^4$-smooth function satisfying the following properties[1]: (1). $\arg\min_{y'\in\mathbb{R}} \ell(y', y) = y$, for any $y \in \mathbb{R}$; (2). $\frac{d^2\ell(y',y)}{d^2y'}|_{y'=y} > 0$, for any $y \in R$; (3). There exists x_0, such that for any $k \in [M]$, it holds that $f_k(x_0) = y_k$.

5.1 Stochastic Worst- and Ascent-direction Sharpness have Same Explicit Bias

Below we specify *stochastic worst-direction sharpness* and *stochastic ascent-direction sharpness* as $\mathbb{E}_k[R^{\text{Max}}_{k,\rho}] = \mathbb{E}_k[L^{\text{Max}}_{k,\rho}] - L$ and $\mathbb{E}_k[R^{\text{Asc}}_{k,\rho}] = \mathbb{E}_k[L^{\text{Asc}}_{k,\rho}] - L$. Here $L^{\text{max}}_{k,\rho}$ and $L^{\text{asc}}_{k,\rho}$ is defined as in Equations (1) and (2) with L replaced by L_k. Unlike the full-batch setting, these two sharpness have same explicit bias, or more precisely, they have the same limiting regularizers. We omit the result on the stochastic average-direction sharpness as it is the same as its counterpart in the full-batch case.

Theorem 5.1. *Stochastic worst-direction sharpness* $\mathbb{E}_k[R^{\text{Max}}_{k,\rho}]$ *satisfies Assumption 4.1 and admits* $\text{Tr}(\nabla^2 L(\cdot))/2$ *as a good limiting regularizer on* Γ.

Theorem 5.2. *Stochastic ascent-direction sharpness* $\mathbb{E}_k[R^{\text{Asc}}_{k,\rho}]$ *satisfies Assumption 4.1 and admits* $\text{Tr}(\nabla^2 L(\cdot))/2$ *as a good limiting regularizer on* Γ.

Unlike in the full-batch setting where the implicit regularizer of ascent-direction sharpness and worst-direction sharpness have different explicit bias, here they are the same because there is no difference between the maximum and minimum of its non-zero eigenvalue for rank-1 Hessian of each individual loss, and that the average of limiting regularizers is equal to the limiting regualrizer of the average regularizers by definition.

[1]Examples include ℓ_2 loss: $\ell(y, y') = 0.5(y - y')^2$.

5.2 STOCHASTIC SAM MINIMIZES STOCHASTIC WORST-DIRECTION SHARPNESS

Stochastic SAM: Recall L_k is the loss on kth data, we use *stochastic SAM* to denote the following update rule, where k_t is sampled i.i.d from uniform distribution on $[M]$.

$$x(t+1) = x(t) - \eta \nabla L_{k_t}\left(x + \rho \frac{\nabla L_{k_t}(x)}{\|\nabla L_{k_t}(x)\|_2}\right), \tag{16}$$

The goal of this section is to show stochastic SAM tracks the following Riemannian gradient flow with respect to $\mathrm{Tr}(\nabla^2 L(\cdot))$ on the manifold for sufficiently small η and ρ,

$$X(\tau) = X(0) - \frac{1}{2}\int_{s=0}^{\tau} P^{\perp}_{X(s),\Gamma} \nabla \mathrm{Tr}(X(s)) ds, X(0) = \Phi(x_{\text{init}}). \tag{17}$$

The fundamental difference between stochastic SAM (16) with batch size 1 and that in the general setting is that the rank of the Hessian of each stochastic loss at minimizers is only rank-1, which enforces the gradient $\nabla L_k(x) \approx \nabla^2 L_k(x)(x - \Phi(x))$ to (almost) lie in the direction of the top eigenvalue of the Hessian. Lemma 5.3 formally states this property. For convenience, we denote $\frac{d^2\ell(y',y_k)}{d^2 y'}|_{y'=f_k(x)}\|\nabla f_k(x)\|^2, \nabla f_k(x)/\|\nabla f_k(x)\|$ as $\Lambda_k(x), w_k(x)$.

Lemma 5.3. *Under Assumption 3.1, for any $x \in U$ and $p \in \Gamma, \nabla^2 L_k(p) = \Lambda_k(p) w_k(p) w_k(p)^T$ and $\exists s \in \{1, -1\} \frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = s w_k(p) + O(\|x - p\|)$*

With Lemma 5.3, we can show $\Phi(x(t+1)) - \Phi(x(t)) = -\eta\rho^2 P^{\perp}_{t,\Gamma}\nabla(\Lambda_k(x))/2 + \tilde{O}(\eta^2\rho^2)$ when $\|x(t+1) - x(t)\| = \tilde{O}(\eta\rho)$, via a similar but slightly more complicated analysis on $\Phi(x(t+1)) - \Phi(x(t))$ as in Section 4.2. As we have $\mathbb{E}[\Lambda_k(x)] = (\sum_k \Lambda_k(x))/M = \mathrm{Tr}(\nabla^2 L(x)), x \in \Gamma$, we have that locally SAM is essentially performing a stochastic projected gradient descent with respect to $\mathrm{Tr}(\nabla^2 L(\cdot))$. It still remains to prove stochastic SAM gets $\tilde{O}(\eta\rho)$ close of the manifold, which is addressed by Theorem C.50. Put everything together, we have the following main theorem for this section.

Theorem 5.4. *Let $\{x(t)\}$ be the iterates defined by SAM (16) and $x(0) = x_{init} \in U$, then under Assumption 3.1, for all η and ρ such that $(\eta + \rho)\log(1/\eta\rho)$ is sufficiently small, with probability at least $1 - O(\sqrt{\rho} + \sqrt{\eta})$, the dynamics of SAM can be split into two phases:*

- ***Phase I:*** *Stochastic SAM follows Gradient Flow with respect to L until entering an $O(\eta\rho)$ neighborhood of the manifold Γ in $\tilde{O}(\frac{1}{\eta})$ steps;*
- ***Phase II:*** *Stochastic SAM tracks the solution X of (17), the Riemannian Gradient Flow with respect to $\mathrm{Tr}(\nabla^2 L(\cdot))$, with time scaling as $\|\Phi(x(\lceil T_3/(\eta\rho^2)\rceil)) - X(T_3)\| = O(\eta^{1/4} + \rho^{1/4})$.*

Corollaries 5.5 and 5.6 below are stochastic counterparts of Corollaries 4.10 and 4.11, which says that the trace of Hessian of the limiting flow of well-tracks the stochastic worst-direction sharpness, and therefore when the limiting flow converges to a local minimizer of trace of Hessian, stochastic SAM minimizes the stochastic worst-direction sharpness.

Corollary 5.5. *Under the condition of Theorem 5.4, we have that with probability $1 - O(\sqrt{\eta} + \sqrt{\rho})$, the difference between the stochastic worst-direction sharpness of the iterates and the corresponding scaled trace of Hessian along the limiting flow is at most $O\big((\eta^{1/4} + \rho^{1/4})\rho^2\big)$, that is,*

$$\left|\mathbb{E}_k[R^{\text{Max}}_{k,\rho}(x(\lceil T_3/(\eta\rho^2)\rceil))] - \rho^2 \mathrm{Tr}(\nabla^2 L(X(T_3)))/2\right| = O\big((\eta^{1/4} + \rho^{1/4})\rho^2\big).$$

Corollary 5.6. *Define U' as in Theorem 4.3, suppose $X(\infty) = \lim_{t\to\infty} X(t)$ exists and is a minimizer of $\mathrm{Tr}(\nabla^2 L(x))$ in $U' \cap \Gamma$. Then for all $\epsilon > 0$, there exists a constant $T_\epsilon > 0$, such that for all ρ, η such that $(\eta + \rho)\ln(1/\eta\rho)$ are sufficiently small, we have that with probability $1 - O(\sqrt{\eta} + \sqrt{\rho})$,*

$$\mathbb{E}_k[L^{\text{Max}}_{k,\rho}(x(\lceil T_\epsilon/(\eta\rho^2)\rceil))] \le \epsilon\rho^2 + \inf_{x\in U'} \mathbb{E}_k[L^{\text{Max}}_{k,\rho}(x)].$$

6 CONCLUSION

In this work, we have performed a rigorous mathematical analysis of the explicit bias of various notions of sharpness when used as regularizers and the implicit bias of the SAM algorithm. In particular, we show the explicit biases of worst-, ascent- and average-direction sharpness around the manifold of minimizers are minimizing the largest eigenvalue, the smallest nonzero eigenvalue, and the trace of Hessian of the loss function. We show that in the full-batch setting, SAM provably decreases the largest eigenvalue of Hessian, while in the stochastic setting when batch size is 1, SAM provably decreases the trace of Hessian.

REFERENCES

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, pp. 639–668. PMLR, 2022.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on edge of stability in deep learning. *arXiv preprint arXiv:2205.09745*, 2022.

Peter L Bartlett, Philip M Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *arXiv preprint arXiv:2210.01513*, 2022.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. *arXiv preprint arXiv:1904.09080*, 2019.

Yaim Cooper. The loss landscape of overparameterized neural networks. *arXiv preprint arXiv:1804.10200*, 2018.

Alex Damian, Tengyu Ma, and Jason Lee. Label noise sgd provably prefers flat global minimizers, 2021.

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

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1019–1028. JMLR. org, 2017.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21:136, 2020.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

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

Stanisław Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Simran Kaur, Jeremy Cohen, and Zachary C Lipton. On the maximum hessian eigenvalue and generalization. *arXiv preprint arXiv:2206.10654*, 2022.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

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

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. In *International Conference on Learning Representations*, 2021.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *arXiv preprint arXiv:2206.07085*, 2022.

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

David McAllester. Simplified pac-bayesian margin bounds. In *Learning theory and Kernel machines*, pp. 203–215. Springer, 2003.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Antonio Orvieto, Anant Raj, Hans Kersting, and Francis Bach. Explicit regularization in overparametrized models via noise injection. *arXiv preprint arXiv:2206.04613*, 2022.

Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. *arXiv preprint arXiv:1910.04284*, 2019.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. *arXiv preprint arXiv:2202.03599*, 2022.

Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8156–8165, 2021.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. *arXiv preprint arXiv:2203.08065*, 2022.

CONTENTS

1 Introduction **1**

2 Related Works **2**

3 Notations and Assumptions **3**

4 Explicit and Implicit Bias in the Full-batch Setting **3**

4.1 Worst- and Ascent-direction Sharpness have Different Explicit Bias 4

4.2 SAM Provably Decreases Worst-direction Sharpness Locally 5

4.3 Proof Sketch for Phase II in Theorem 4.9 6

5 Explicit and Implicit bias in the Stochastic Setting **8**

5.1 Stochastic Worst- and Ascent-direction Sharpness have Same Explicit Bias 8

5.2 Stochastic SAM Minimizes Stochastic Worst-direction Sharpness 9

6 Conclusion **9**

A Additional Related Works **13**

B More Discussions of Manifold Assumption **13**

C Proof **14**

C.1 Proof Setups 14

C.2 Properties of Limiting Map of Gradient Flow, Φ 15

C.3 Explicit Bias 18

C.4 Full-batch SAM on Quadratic Loss (Proof of Theorem 4.14) 23

C.4.1 Entering Invariant Set 24

C.4.2 Alignment to Top Eigenvector 25

C.5 Full-batch SAM on General Loss (Proof of Theorem 4.9) 31

C.5.1 Phase I (Proof of Theorem C.31) 31

C.5.2 Phase II (Proof of Theorem C.32) 42

C.5.3 Proofs of Corollaries 4.10 and 4.11 48

C.6 Analysis for 1-SAM (Proof of Theorem 5.4) 48

C.6.1 Phase I (Proof of Theorem C.50) 49

C.6.2 Phase II (Proof of Theorem C.51) 58

C.6.3 Proofs of Corollaries 5.5 and 5.6 59

C.6.4 Other Omitted Proofs for 1-SAM 60

C.7 Technical Lemmas 60

D Omitted Proofs on Continuous Approximation **65**

A Additional Related Works

Sharpness Minimization as Implicit Bias. Recent theoretical works Blanc et al. (2019); Damian et al. (2021); Li et al. (2021) showed that SGD with label noise is implicitly biased to local minimizers with a smaller trace of Hessian. Arora et al. (2022) showed that normalized GD implicitly decreases the largest eigenvalue of the Hessian. Lyu et al. (2022) showed that GD with weight decay on a scale invariant loss function implicitly decreases the spherical sharpness, *i.e.*, the largest eigenvalue of the Hessian evaluated at the normalized parameter.

In a concurrent work, Bartlett et al. (2022) prove that on quadratic loss, the iterate of SAM (Equation (11)) and its gradient converges to the top eigenvector of Hessian, which is almost the same as our Theorem 4.14. Assuming such alignment for a general loss, the work of Bartlett et al. (2022) shows that the largest eigenvalue of Hessian decreases in the next step. This paper also proves such a Hessian-gradient alignment for general loss functions and an end-to-end theorem showing that the largest eigenvalue of Hessian and worst-direction sharpness decrease along the trajectory of SAM (Theorem 4.9), which are not shown in Bartlett et al. (2022). Moreover, this paper also characterizes implicit bias of stochastic SAM with batch size 1, which is minimizing the average-direction sharpness, while Bartlett et al. (2022) only considers the deterministic case.

Comparison with Arora et al. (2022). Our proof uses similar framework as Arora et al. (2022). However our analysis has its own difficulty for the following reasons. First, Arora et al. (2022) only deal with deterministic case, while our analysis extends to stochastic SAM as well (Section 5). Second, our analysis for the deterministic case is different from that of Arora et al. (2022) in the following two aspects:

1. The alignment analysis is more complicated because we have two hyperparameters,learning rate η and perturbation radius ρ, while Arora et al. (2022) only needs to deal with one hyperparameter, learning rate η.
2. The mechanism of penalizing worst-direction sharpness is different, which can be seen from the dependency of the sharpness-reduction rate over learning rate η. In Arora et al. (2022), normalized GD reduces the sharpness via a second-order effect of GD and thus the sharpness is reduced by $O(\eta^2)$ per step. In our analysis, for fixed small perturbation radius ρ, the sharpness is reduced by $O(\rho^2\eta)$ per step, which is linear in η.

B More Discussions of Manifold Assumption

Validity. Recent empirical studies have shown that there are essentially no barriers in loss landscape between different minimizers, that is, the set of minimizers are path-connected (Draxler et al., 2018; Garipov et al., 2018). Motivated by this empirical discovery, we make the assumption below following Fehrman et al. (2020); Damian et al. (2021); Li et al. (2021), which is theoretically justified by Cooper (2018) under a generic setting.

Importance. The connectivity of the set of local minimizers implied by the manifold assumption above allows us to take limits of perturbation radius $\rho \to 0$ while still yield interesting and insightful implicit bias results. So far almost all analysis of implicit bias for general model parameterizations relies on Taylor expansion, e.g. Blanc et al. (2019); Damian et al. (2021); Li et al. (2021); Arora et al. (2022), so does the derivation of the SAM algorithm (Foret et al., 2020; Wu et al., 2020). Thus it's crucial to consider small perturbation size ρ. On the contrary, if the set of global minimizers are a set of discrete points, then with small perturbation radius ρ, implicit bias of optimizers is not sufficient to drive the iterate from global minimum to the other one.

Limitation. As proved by Cooper (2018), the manifold setting holds generically for regression problems with smooth loss. The manifold assumption doesn't hold for networks with ReLU (which is non-smooth) and classification problems with softmax loss (where an overparametrized network may achieve arbitrarily small but non-zero loss). In the latter case, there may not even exist a global minimizer, not to mention a manifold of minimizers.

C PROOF

C.1 PROOF SETUPS

In this section we provide details of our proof setups, including notations and assumptions/settings.

We first introduce some additional notations that will be used in the proofs. For any subset $S \in \mathbb{R}^D$, we define $\text{Dist}(x, S) \triangleq \inf_{y \in S} \|x - y\|_2$. For any $d > 0$ and any subset $S \in \mathbb{R}^D$, we define $S^d \triangleq \{x \in \mathbb{R}^D \mid \text{Dist}(x, S) \leq d\}$. Our convention is to use K to denote a compact set and U to denote an open set.

Below we restate our main assumption in the full-batch case and related notations in Section 3. Throughout the analysis, we fix our initialization as x_{init}, our loss function as $L : \mathbb{R}^D \to \mathbb{R}$.

Assumption 3.1. *Assume loss $L : \mathbb{R}^D \to \mathbb{R}$ belongs to $\mathcal{C}^4$, and there exists a manifold Γ that is $D - M$ 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 global minimizer of L, $L(x) = 0$ and rank$(\nabla^2 L(x)) = M$.*

Notations for Full-Batch Setting: Given any point $x \in \Gamma$, define $P_{x,\Gamma}$ as the projection operator onto the manifold of the normal space of Γ at x and $P^{\perp}_{x,\Gamma} = I_D - P_{x,\Gamma}$. Given the loss function L, its gradient flow is denoted by mapping $\phi : \mathbb{R}^D \times [0, \infty) \to \mathbb{R}^D$. Here, $\phi(x, \tau)$ denotes the iterate at time τ of a gradient flow starting at x and is defined as the unique solution of $\phi(x, \tau) = x - \int_0^\tau \nabla L(\phi(x, t))dt$, $\forall x \in \mathbb{R}^D$. We further define the limiting map of $\phi(x, \cdot)$ as $\Phi(x) = \lim_{\tau \to \infty} \phi(x, \tau)$, that is, $\Phi(x)$ denotes the convergent point of the gradient flow starting from x. For convenience, we define $\lambda_i(x), v_i(x)$ as $\lambda_i(\nabla^2 L(\Phi(x))), v_i(\nabla^2 L(\Phi(x)))$ whenever the latter is well defined. When $x(t)$ and Γ is clear from context, we also use $\lambda_i(t) := \lambda_i(x(t)), v_i(t) := v_i(x(t)), P^{\perp}_{t,\Gamma} := P^{\perp}_{\Phi(x(t)),\Gamma}, P_{t,\Gamma} := P_{\Phi(x(t)),\Gamma}$.

Definition 3.2 (Attraction Set)**.** *Let U be the attraction set of Γ under gradient flow, that is, a neighborhood of Γ containing all points starting from which gradient flow w.r.t. loss L converges to some point in Γ, or mathematically, $U \triangleq \{x \in \mathbb{R}^D | \Phi(x)$ exists and $\Phi(x) \in \Gamma\}$.*

Below we restate the setting for stochastic loss of batch size one in Section 5.

Setting C.1. *Let $f_k(x)$ be the model output on kth data where f_k is a $\mathcal{C}^4$-smooth function and y_k be the ith label for $l = 1, \ldots, M$. We define the loss on kth data as $L_k(x) = \ell(f_k(x), y_k)$ and the total loss $L = \sum_{k=1}^M L_k/M$, where $\ell(y', y)$ is a $\mathcal{C}^4$-smooth function satisfying the following properties[2]: (1). $\arg\min_{y' \in \mathbb{R}} \ell(y', y) = y$, for any $y \in \mathbb{R}$; (2). $\frac{d^2 \ell(y', y)}{d^2 y'}|_{y'=y} > 0$, for any $y \in R$; (3). There exists x_0, such that for any $k \in [M]$, it holds that $f_k(x_0) = y_k$.*

In our analysis, we prove our main theorems in the stochastic setting under a more general condition than Setting C.1, which is Condition C.2 (on top of Assumption 3.1). The only usage of Setting C.1 in the proof is Lemma 5.3 and theorem C.49.

Condition C.2. *Total loss $L = \frac{1}{M} \sum_{k=1}^M L_k$. For each $k \in [M]$, L_k is $\mathcal{C}^4$, and there exists a $(D-1)$-dimensional $\mathcal{C}^2$-submanifold of $\mathbb{R}^D$, Γ_k, where for all $x \in \Gamma_k$, x is a global minimizer of L_k, $L_k(x) = 0$ and rank$(\nabla^2 L_k(x)) = 1$. Moreover, $\Gamma \subset \cap_{k=1}^M \Gamma_k$ for Γ defined in Assumption 3.1.*

Notations for Stochastic Setting: Since L_k is rank-1 on Γ_k for each $k \in [M]$, we can write it as $L_k(x) = \Lambda_k(x) w_k(x) w_k^\top(x)$ for any $x \in \Gamma$, where w_k is a continuous function on Γ with pointwise unit norm. Given the loss function L_k, its gradient flow is denoted by mapping $\phi_k : \mathbb{R}^D \times [0, \infty) \to \mathbb{R}^D$. Here, $\phi_k(x, \tau)$ denotes the iterate at time τ of a gradient flow starting at x and is defined as the unique solution of $\phi_k(x, \tau) = x - \int_0^\tau \nabla L_k(\phi_k(x, t))dt$, $\forall x \in \mathbb{R}^D$. We further define the limiting map Φ_k as $\Phi_k(x) = \lim_{\tau \to \infty} \phi_k(x, \tau)$, that is, $\Phi_k(x)$ denotes the convergent point of the gradient flow starting from x. Similar to Definition 3.2, we define $U_k = \{x \in \mathbb{R}^D | \Phi(x)$ exists and $\Phi_k(x) \in \Gamma_k\}$ be the attraction set of Γ_i. We have that each U_k is open and Φ_k is $\mathcal{C}^2$ on U_k by Lemma B.15 in Arora et al. (2022).

Definition C.3. *A function L is μ-PL in a set U iff $\forall x \in U$, $\|\nabla L(x)\|_2^2 \geq 2\mu(L(x) - \inf_{x \in U} L(x))$.*

[2]Examples include ℓ_2 loss: $\ell(y, y') = 0.5(y - y')^2$.

Definition C.4. *The spectral 2-norm of a k-order tensor $X_{i_1,\dots,i_k} \in R^{d_1\times\dots\times d_k}$ is defined as*

$$\|X\|_2 = \max_{x_i\in R^{d_i}, \|x_i\|_2=1} X[x_1, \dots, x_k].$$

Lemma C.5 (Arora et al. (2022) Lemma B.2)**.** *Given any compact set $K \subseteq \Gamma$, there exist $r(K), \mu(K), \Delta(K) \in \mathbb{R}^+$ such that*

1. *$K^{r(K)} \cap \Gamma$ is compact.*
2. *$K^{r(K)} \subset U \cap (\cap_k U_k)$.*
3. *L is $\mu(K)$-PL on $K^{r(K)}$.*
4. $\inf_{x\in K^{r(K)}}(\lambda_1(\nabla^2 L(x)) - \lambda_2(\nabla^2 L(x))) \geq \Delta(K) > 0$.
5. $\inf_{x\in K^{r(K)}} \lambda_M(\nabla^2 L(x)) \geq \mu(K) > 0$.
6. $\inf_{x\in K^{r(K)}} \lambda_1(\nabla^2 L_k(x)) \geq \mu(K) > 0$.

Given compact set $K \subset \Gamma$, we further define

$$\zeta(K) = \sup_{x\in K^{r(K)}} \|\nabla^2 L(x)\|_2,\ \nu(K) = \sup_{x\in K^{r(K)}} \|\nabla^3 L(x)\|_2,\ \Upsilon(K) = \sup_{x\in K^{r(K)}} \|\nabla^4 L(x)\|_2,$$
$$\xi(K) = \sup_{x\in K^{r(K)}} \|\nabla^2 \Phi(x)\|_2,\ \chi(K) = \sup_{x\in K^{r(K)}} \|\nabla^3 \Phi(x)\|_2.$$

Similarly, we use notations like $\zeta_k(K), \nu_k(K), \Upsilon_k(K), \xi_k(K), \chi_k(K)$ to denote the counterpart of the above quantities defined for stochastic loss L_k and its limiting map Φ_k for $k \in [M]$.

Lemma C.6 (Arora et al. (2022), Lemma B.5 and B.7)**.** *Given any compact subset $K \subset \Gamma$, let $r(K)$ be defined in Lemma C.5, there exist $0 < h(K) < r(K)$ such that*

1. $\sup_{x\in K^{h(K)}} L(x) - \inf_{x\in K^{h(K)}} L(x) \leq \frac{\mu(K)\rho^2(K)}{8}$.
2. $\forall x \in K^{h(K)}, \Phi(x) \in K^{r(K)/2}$.
3. $\forall x \in K^{h(K)}, \|x - \Phi(x)\|_2 \leq \frac{8\mu(K)^2}{\zeta(K)\nu(K)}$.
4. *The whole segment $\overline{x\Phi(x)}$ lies in $K^{r(K)}$, so does $\overline{x\Phi_k(x)}$, for any $k \in [D]$.*

The proof of the lemmas above can be found in Arora et al. (2022). Readers should note that although Arora et al. (2022) only prove these lemmas when K is a special compact set (the trajectory of an ODE), all the proof does not use any property of K other than it is a compact subset of Γ, and thus our Lemmas C.5 and C.6 hold for general compact subsets of Γ.

In the rest part of the appendix, for convenience we will drop the dependency on K in various constants when there is no ambiguity.

C.2 Properties of Limiting Map of Gradient Flow, Φ

In our analysis, the property of Φ will be heavily used. In this section, we will recap some related lemmas from Arora et al. (2022), and then introduce some new lemmas for the stochastic setting with batch size one.

Lemma C.7 (Arora et al. (2022) Lemma B.6)**.** *Given any compact set $K \subset \Gamma$, for any $x \in K^h$,*

$$\|x - \Phi(x)\|_2 \leq \int_0^\infty \|\frac{d\phi(x,t)}{dt}\|_2 \leq \sqrt{\frac{2(L(x) - L(\Phi(x)))}{\mu}} \leq \frac{\|\nabla L(x)\|_2}{\mu}.$$

Lemma C.8. *Given any compact set $K \subset \Gamma$, for any $x \in K^h$,*

$$\|\nabla L(x)\|_2 \leq \zeta\|x - \Phi(x)\|_2 \leq \zeta\sqrt{\frac{2(L(x) - L(\Phi(x)))}{\mu}}.$$

Proof. The first inequality is by Lemma C.5 and Taylor Expansion. The second inequality is by Lemma C.7. □

Lemma C.9 (Arora et al. (2022) Lemmas B.16 and B.22)**.** *Given any compact set $K \subset \Gamma$, for any $x \in K^h$,*

$$\begin{aligned}\partial\Phi(x)\nabla L\left(x\right) &= 0, x \in U;\\ \partial\Phi\left(x\right)\nabla^2 L\left(x\right)\nabla L\left(x\right) &= -\partial^2\Phi\left(x\right)\left[\nabla L\left(x\right),\nabla L\left(x\right)\right], x \in U;\\ \partial\Phi\left(x\right)\partial^2(\nabla L)(x)[v_1,v_1] &= P^{\perp}_{X,\Gamma}\nabla(\lambda_1(\nabla^2(L(x)))), x \in \Gamma.\end{aligned}$$

Lemma C.10 (Arora et al. (2022) Lemmas B.8 and B.9)**.** *Given any compact set $K \subset \Gamma$, for any $x \in K^h$,*

$$\begin{aligned}\|P^{\perp}_{\Phi(x),\Gamma}(x-\Phi(x))\|_2 &\leq \frac{\zeta\nu}{4\mu^2}\|x-\Phi(x)\|_2^2;\\ \|\nabla L\left(x\right)-\nabla^2 L\left(\Phi(x)\right)\left(x-\Phi(x)\right)\|_2 &\leq \frac{\nu}{2}\|x-\Phi(x)\|_2^2;\\ \left|\frac{\|\nabla L\left(x\right)\|_2}{\|\nabla^2 L\left(\Phi(x)\right)\left(x-\Phi(x)\right)\|_2}-1\right| &\leq \frac{2\nu}{\mu}\|x-\Phi(x)\|_2;\\ \frac{\nabla L\left(x\right)}{\|\nabla L\left(x\right)\|} &= \frac{\nabla^2 L\left(\Phi(x)\right)\left(x-\Phi(x)\right)}{\|\nabla^2 L\left(\Phi(x)\right)\left(x-\Phi(x)\right)\|_2}+O(\frac{\nu}{\mu}\|x-\Phi(x)\|_2).\end{aligned}$$

The proof of above lemmas can be found in Arora et al. (2022).

Lemma C.11. *Given any compact set $K \subset \Gamma$, for any $x \in K^h$,*

$$\begin{aligned}\|\partial\Phi(x)\nabla L_k(x)\|_2 &\leq (\nu_k+\zeta_k\xi)\|x-\Phi(x)\|_2^2\\ \|\partial\Phi(x)\nabla^2 L_k(x)\frac{\nabla L_k\left(x\right)}{\|\nabla L_k\left(x\right)\|}\|_2 &\leq (\nu_k+\zeta_k\xi)\|x-\Phi(x)\|_2\end{aligned}$$

Proof. By Lemma C.6 and Taylor Expansion,

$$\begin{aligned}\|\partial\Phi(x)\nabla L_k(x)\|_2 &\leq \|\partial\Phi(x)\nabla^2 L_k(\Phi(x))(x-\Phi(x))\|_2+\nu_k\|x-\Phi(x)\|_2^2\\ &\leq \|\partial\Phi(\Phi(x))\nabla^2 L_k(\Phi(x))(x-\Phi(x))\|_2+\nu_k\|x-\Phi(x)\|_2^2+\zeta_k\xi\|x-\Phi(x)\|_2^2\\ &= \|P^{\perp}_{x,\Gamma}\partial\Phi(\Phi(x))\nabla^2 L_k(\Phi(x))(x-\Phi(x))\|_2+\nu_k\|x-\Phi(x)\|_2^2+\zeta_k\xi\|x-\Phi(x)\|_2^2\\ &= (\nu_k+\zeta_k\xi)\|x-\Phi(x)\|_2^2,\end{aligned}$$

this proves the first claim.

Again by Lemma C.5 and Taylor Expansion,

$$\begin{aligned}\|\partial\Phi(x)\nabla^2 L_k(x)\frac{\nabla L_k\left(x\right)}{\|\nabla L_k\left(x\right)\|}\|_2 &\leq \|\partial\Phi(x)\nabla^2 L_k(\Phi(x))\frac{\nabla L_k\left(x\right)}{\|\nabla L_k\left(x\right)\|}\|_2+\nu_k\|x-\Phi(x)\|_2\\ &\leq \|\partial\Phi(\Phi(x))\nabla^2 L_k(\Phi(x))\frac{\nabla L_k\left(x\right)}{\|\nabla L_k\left(x\right)\|}\|_2+(\nu_k+\zeta_k\xi)\|x-\Phi(x)\|_2\\ &= (\nu_k+\zeta_k\xi)\|x-\Phi(x)\|_2,\end{aligned}$$

this proves the second claim. □

Lemma C.12. *Given any compact set $K \subset \Gamma$, for any $x \in K^h$ and $y = x-\eta\nabla L\left(x+\rho\frac{\nabla L(x)}{\|\nabla L(x)\|}\right)$,*

$$\begin{aligned}\|y-x\|_2 &\leq \eta\|\nabla L\left(x\right)\|_2+\eta\zeta\rho\\ \|\Phi(x)-\Phi(y)\|_2 &\leq \xi\eta\rho\|\nabla L\left(x\right)\|_2+\nu\eta\rho^2+\xi\eta^2\|\nabla L\left(x\right)\|_2^2+\xi\zeta^2\eta^2\rho^2\\ &\leq \zeta\xi\eta\rho\|x-\Phi(x)\|_2+\zeta^2\xi\eta^2\|x-\Phi(x)\|_2^2+\nu\eta\rho^2+\xi\zeta^2\eta^2\rho^2\end{aligned}$$

Proof. For sufficient small ρ, $x+\rho\frac{\nabla L(x)}{\|\nabla L(x)\|} \in K^r$. By Taylor Expansion,

$$\|y-x\|_2 = \eta\|\nabla L\left(x+\rho\frac{\nabla L\left(x\right)}{\|\nabla L\left(x\right)\|}\right)\|_2 \leq \eta\|\nabla L\left(x\right)\|_2+\eta\zeta\rho$$

This further implies that for sufficiently small η and ρ, $\overline{xy} \in K^r$.

Again by Taylor Expansion,

$$\|\partial\Phi(x)(y-x)\|_2 \leq \eta\|\partial\Phi(x)\nabla L\left(x\right) + \rho\partial\Phi(x)\nabla^2 L(x)\frac{\nabla L\left(x\right)}{\|\nabla L\left(x\right)\|}\|_2 + \eta\rho^2\nu/2\,.$$

By Lemma C.9, $\partial\Phi(x)\nabla L\left(x\right) = 0$ and $\partial\Phi\left(x\right)\nabla^2 L\left(x\right)\nabla L\left(x\right) = -\partial^2\Phi\left(x\right)\left[\nabla L\left(x\right), \nabla L\left(x\right)\right]$. Hence,

$$\begin{aligned}\|\partial\Phi(x)(y-x)\|_2 &\leq \eta\rho\|\nabla L\left(x\right)\|_2\|\partial^2\Phi(x)\left[\frac{\nabla L\left(x\right)}{\|\nabla L\left(x\right)\|}, \frac{\nabla L\left(x\right)}{\|\nabla L\left(x\right)\|}\right]\|_2 + \eta\rho^2\nu/2 \\ &\leq \xi\eta\rho\|\nabla L\left(x\right)\|_2 + \eta\rho^2\nu/2\,.\end{aligned}$$

As $\overline{xy} \in K^r$, by Taylor Expansion,

$$\|\Phi(y) - \Phi(x)\|_2 \leq \|\partial\Phi(x)(y-x)\|_2 + \xi\|y-x\|_2^2/2$$

Putting together we have

$$\|\Phi(x) - \Phi(y)\|_2 \leq \xi\eta\rho\|\nabla L\left(x\right)\|_2 + \eta\rho^2\nu + \xi\eta^2\|\nabla L\left(x\right)\|_2^2 + \xi\zeta^2\eta^2\rho^2\,.$$

Finally, by Lemma C.8, we have

$$\begin{aligned}\|\Phi(x) - \Phi(y)\|_2 &\leq \xi\eta\rho\|\nabla L\left(x\right)\|_2 + \nu\eta\rho^2 + \xi\eta^2\|\nabla L\left(x\right)\|_2^2 + \xi\zeta^2\eta^2\rho^2 \\ &\leq \zeta\xi\eta\rho\|x - \Phi(x)\|_2 + \zeta^2\xi\eta^2\|x - \Phi(x)\|_2^2 + \nu\eta\rho^2 + \xi\zeta^2\eta^2\rho^2\,.\end{aligned}$$

□

Lemma C.13. *Given any compact set $K \subset \Gamma$, for any $x \in K^h$ and $y = x - \eta\nabla L_k\left(x + \rho\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|}\right)$,*

$$\begin{aligned}\|y - x\|_2 &\leq \eta\|\nabla L_k\left(x\right)\|_2 + \eta\zeta\rho \\ \|\Phi(x) - \Phi(y)\|_2 &\leq O(\eta\|\nabla L(x)\|_2^2 + \eta\rho\|\nabla L(x)\|_2 + \eta\rho^2)\end{aligned}$$

Proof. For sufficient small ρ, $x + \rho\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} \in K^r$. By Taylor Expansion,

$$\|y - x\|_2 = \eta\|\nabla L_k\left(x + \rho\frac{\nabla L_k\left(x\right)}{\|\nabla L_k\left(x\right)\|}\right)\|_2 \leq \eta\|\nabla L_k\left(x\right)\|_2 + \eta\zeta\rho\,.$$

This further implies that for sufficiently small η and ρ, $\overline{xy} \in K^r$.

Again by Taylor Expansion,

$$\|\partial\Phi(x)(y-x)\|_2 \leq \eta\|\partial\Phi(x)\nabla L_k\left(x\right) + \rho\partial\Phi(x)\nabla^2 L_k(x)\frac{\nabla L_k\left(x\right)}{\|\nabla L_k\left(x\right)\|}\|_2 + \eta\rho^2\nu/2\,.$$

We further have by Lemma C.7,

$$\begin{aligned}&\|\partial\Phi(x)\nabla L_k\left(x\right)\| \\ \leq&\|\partial\Phi(\Phi(x))\nabla L_k\left(x\right)\| + \xi\|\nabla L_k\left(x\right)\|_2\|x - \Phi(x)\| \\ \leq&\|\partial\Phi(\Phi(x))\nabla^2 L_k(\Phi(x))(x - \Phi(x))\| + \nu\|x - \Phi(x)\|_2^2 + \zeta\xi\|x - \Phi(x)\|_2^2 \\ \leq&\frac{\nu}{\mu}\|\nabla L(x)\|_2^2 + \frac{\zeta\xi}{\mu^2}\|\nabla L(x)\|_2^2\,.\end{aligned}$$

Similarly

$$
\begin{aligned}
&\|\rho\partial\Phi(x)\nabla^2 L_k(x)\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|}\|_2 \\
\leq&\|\rho\partial\Phi(\Phi(x))\nabla^2 L_k(x)\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|}\|_2 + \rho\zeta\xi\|x-\Phi(x)\|_2 \\
\leq&\|\rho\partial\Phi(\Phi(x))\nabla^2 L_k(\Phi(x))\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|}\|_2 + \rho\zeta\xi\|x-\Phi(x)\|_2 + \rho\nu\|x-\Phi(x)\|_2 \\
\leq&\rho\frac{\zeta\xi}{\mu^2}\|\nabla L(x)\|_2 + \rho\frac{\nu}{\mu}\|\nabla L(x)\|_2 .
\end{aligned}
$$

□

C.3 Explicit Bias

We will first prove the generic Theorem 4.3 and then apply this theorems to characterize a variety of regularizers.

Lemma C.14. $\Gamma = U \cap \overline{\Gamma}$.

Proof of Lemma C.14. For the first claim, for any point $x \in \overline{U'} \cap \overline{\Gamma}$, there exists $\{x_k\}_{k=1}^{\infty} \in \Gamma$ such that $\lim_{k\to\infty} x_k = x$. Since $x \in \overline{U'} \subseteq U$ and Φ is continuous in U, it holds that Φ is continuous at x, thus $\lim_{k\to\infty}\Phi(x_k) = \Phi(x) \in \Gamma$. However $\Phi(x_k) = x_k$ because $x_k \in \Gamma, \forall k$. Thus we know $x = \Phi(x) \in \Gamma$. □

Lemma C.15. *Let U' be any bounded open set such that its closure $\overline{U'} \subseteq U$. If $\overline{U'} \cap \Gamma \subseteq \overline{U' \cap \Gamma}$, then $\overline{U'} \cap \Gamma = \overline{U' \cap \Gamma}$.*

Proof of Lemma C.14. By Lemma C.14, it holds that $\overline{U'} \cap \Gamma = \overline{U'} \cap U \cap \overline{\Gamma} = \overline{U'} \cap \overline{\Gamma}$. Note that $\overline{U' \cap \Gamma} \subseteq \overline{U'}, \overline{U' \cap \Gamma} \subseteq \overline{\Gamma}$, we have that $\overline{U' \cap \Gamma} \subseteq \overline{U'} \cap \overline{\Gamma} = \overline{U'} \cap \Gamma$, which completes the proof. □

Lemma C.16. *Let U' be any bounded open set such that its closure $\overline{U'} \subseteq U$. Further assume $\overline{U'} \cap \Gamma = \overline{U' \cap \Gamma}$, then for all $h_2 > 0, \exists \rho_0 > 0$ if $x \in U', \operatorname{dist}(x, \Gamma) \leq \rho_0 \Rightarrow \operatorname{dist}(x, \overline{U' \cap \Gamma}) \leq h_2$*

Proof. We will prove by contradiction. Suppose there exists $h_2 > 0$ and $\{x_k\}_{k=1}^{\infty} \in U'$, such that $\lim_{k\to\infty}\operatorname{dist}(x_k, \Gamma) = 0$ but $\forall k > 0, \operatorname{dist}(x_k, \overline{U' \cap \Gamma}) \geq h_2$. Since U' is bounded, $\overline{U'}$ is compact and thus $\{x_k\}_{k=1}^{\infty}$ has at least one accumulate point x^* in $\overline{U'} \subseteq U$. Since U is the attraction set of Γ under gradient flow, we know that $\Phi(x^*) \in \Gamma$. Now we claim $x^* \in \Gamma$. This is because $\lim_{k\to\infty}\operatorname{dist}(x_k, \Gamma) = 0$ and thus there exists a sequence of points on Γ, $\{y_k\}_{k=1}^{\infty}$, where $\lim_{k\to\infty}\|x_k - y_k\| = 0$. Thus we have that $x^* = \lim_{k\to\infty} y_k = \lim_{k\to\infty}\Phi(y_k) = \Phi(x^*)$, where the last step we used that $x^* \in U$ and Φ is continuous on U. By the definition of U, $x^* \in U \iff \Phi(x^*) \in \Gamma$, thus $x^* \in \Gamma$. Then we would have $x^* \in \overline{U'} \cap \Gamma$, which is contradictory to $\operatorname{dist}(x_k, \overline{U'} \cap \Gamma) \geq \operatorname{dist}(x_k, \overline{U' \cap \Gamma}) \geq h_2, \forall k > 0$. This completes the proof. □

Lemma C.17. *Let U' be any bounded open set such that its closure $\overline{U'} \subseteq U$. then for all $h_2 > 0, \exists \rho_1 > 0$ if $x \in U', L(x) \leq \rho_1 \Rightarrow \operatorname{dist}(x, \overline{U' \cap \Gamma}) \leq h_2$*

Proof. We will prove by contradiction. If there exists a list of $\rho_1, ..., \rho_k, ...$, such that $\rho_k \to 0$ and there exists $x_k \in U'$, such that $L(x_k) \leq \inf_{x\in U'} L(x) + \rho_k$ and $\operatorname{dist}(x_k, \overline{U' \cap \Gamma}) \geq h_2$. Since U' is bounded, $\overline{U'}$ is compact and thus $\{x_k\}_{k=1}^{\infty}$ has at least one accumulate point x^* in $\overline{U'} \subseteq U$. Since L is continuous in U, $L(x^*) = \lim_{k\to\infty} L(x_k) = \inf_{x\in U'} L(x)$. Thus x^* is a local minimizer of L and thus has zero gradient, which further implies that $x^* = \Phi(x^*)$. Thus $x^* \in \overline{U'} \cap \Gamma$, which is contradictory to $\operatorname{dist}(x_k, \overline{U'} \cap \Gamma) \geq \operatorname{dist}(x_k, \overline{U' \cap \Gamma}) \geq h_2, \forall k > 0$. This completes the proof. □

Lemma C.18. *Let U' be any bounded open set such that its closure $\overline{U'} \subseteq U$ and $\overline{U'} \cap \Gamma = \overline{U' \cap \Gamma}$. Suppose regularizers $\{R_\rho\}$ admits a limiting regularizer S on Γ, then*

$$
\inf_{x\in U'}(L(x) + R_\rho(x)) \leq \rho^2 \inf_{x\in U'\cap\Gamma} S(x) + \inf_{x\in U'} L(x) + o(\rho^2).
$$

Proof of Lemma C.18. First choose sufficiently small ρ, such that $\rho < h(\overline{U' \cap \Gamma})$. Choose an approximate minimizer of $S(x)$, $x_0 \in U' \cap \Gamma$, such that $S(x_0) \leq \inf_{x \in U' \cap \Gamma} S(x) + \rho^2$. Then by the definition of limiting regularizers (Definition 4.2) and the assumption that U' is open, there exists $x_1 \in U'$ satisfying that $\|x_1 - x_0\|_2 \leq r_\rho < \rho^2$ and $R_\rho(x_1)/\rho^2 - S(x_0) \leq \rho^2$. Thus, $R_\rho(x_1) \leq \rho^2 S(x_0) + \rho^4$.

As $\|x_1 - x_0\|_2 \leq \rho^2 < h$ and $x_0 \in \overline{U' \cap \Gamma}$. This further leads to $\overline{x_0 x_1} \in \overline{U' \cap \Gamma}^h$. By Taylor expansion on L at x_0, we would have $L(x_1) \leq L(x_0) + O(\|x_0 - x_1\|_2^2) = \inf_{x \in U' \cap \Gamma} L(x) + O(\rho^4)$. Thus it holds that

$$\inf_{x \in U'} (L(x) + R_\rho(x)) \leq L(x_1) + R_\rho(x_1) \leq \rho^2 \inf_{x \in U' \cap \Gamma} S(x) + \inf_{x \in U'} L(x) + O(\rho^4).$$

This completes the proof. □

Lemma C.19. *Let U' be any bounded open set such that its closure $\overline{U'} \subseteq U$ and $\overline{U'} \cap \Gamma = \overline{U' \cap \Gamma}$. Suppose regularizers $\{R_\rho\}$ admits a good limiting regularizer S on Γ, then for all $u \in U'$,*

$$\|u - \Phi(u)\|_2 = O(\rho) \implies R_\rho(u) \geq \rho^2 \inf_{x \in U' \cap \Gamma} S(x) - o(\rho^2)\,.$$

Proof of Lemma C.19. Define $K = \mathrm{dist}(u, \overline{U' \cap \Gamma})$. Define $r = r(K), h = h(K)$ as the constant in Lemma C.5 with $K = \overline{U' \cap \Gamma}$. Note K is compact and by Lemma C.15, $K = \overline{U'} \cap \Gamma \subset \Gamma$. By Lemma C.5, we have $K^r \cap \Gamma$ is a compact set, so is $K^h \cap \Gamma$. Since S is a good limiting regularizer for $\{R_\rho\}$, by Definition 4.2, for any $x^* \in K^h \cap \Gamma$, there exists open neighborhood of x^*, V_{x^*} such that for any $C, \epsilon_1 > 0$, there is a ρ_{x^*} such that

$$\forall x \in V_{x^*} \text{ and } \rho \leq \rho_{x^*}, \quad \left| S(x) - \inf_{\|x' - x\|_2 \leq C \cdot \rho} R_\rho(x')/\rho^2 \right| < \epsilon_1.$$

Note that $K^h \cap \Gamma$ is compact, there exists a finite subset of $K^h \cap \Gamma$, $\{x_k\}_k$, such that $K^h \cap \Gamma \subset \cup_k V_{x_k}$. Hence for any $C, \epsilon_1 > 0$, there is some $\rho_K = \min_k \rho_{x_k} > 0$, it holds that,

$$\forall x \in K^h \cap \Gamma \text{ and } \rho \leq \rho_K, \quad \left| S(x) - \inf_{\|x' - x\|_2 \leq C \cdot \rho} R_\rho(x')/\rho^2 \right| < \epsilon_1. \tag{18}$$

We can rewrite Equation (18) as for any $C > 0$,

$$\sup_{x \in K^h \cap \Gamma} \left| S(x) - \inf_{\|x' - x\|_2 \leq C \cdot \rho} R_\rho(x')/\rho^2 \right| = o(1), \quad \text{as } \rho \to 0. \tag{19}$$

As $u \in U' \subseteq U$, we have that $\Phi(u) \in \Gamma$. If $\|u - \Phi(u)\|_2 = O(\rho)$, then $\mathrm{dist}(u, \Gamma) \leq O(\rho)$. By Lemma C.16, we have that $\mathrm{dist}(u, K) = o(1)$. This further implies $\mathrm{dist}(\Phi(u), K) \leq \mathrm{dist}(u, K) + \mathrm{dist}(\Phi(u), u) = o(1)$. Hence we have that $\Phi(u) \in K^h \cap \Gamma$ for sufficiently small ρ. Thus we can pick $x = \Phi(u)$ in Equation (19) and C sufficiently large, which yields that

$$\rho^2 S(\Phi(u)) \leq \inf_{\|u' - \Phi(u)\|_2 \leq O(\rho)} R_\rho(u') + o(\rho^2) \leq R_\rho(u) + o(\rho^2), \tag{20}$$

where the last step is because $\|u - \Phi(u)\|_2 = O(\rho)$. On the other hand, we have that

$$S(\Phi(u)) \geq \inf_{x \in U' \cap \Gamma} S(x) - o(1)\,. \tag{21}$$

as S is continuous on Γ and $\mathrm{dist}(\overline{U' \cap \Gamma}, \Phi(u)) = o(1)$. Combining Equations (20) and (21), we have $R_\rho(u) \geq \rho^2 \inf_{x \in U' \cap \Gamma} S(x) - o(\rho^2)$. □

Proof of Theorem 4.3 . We will first prove the $\Rightarrow$ direction. By Lemma C.18, it holds that

$$L(u) + R_\rho(u) \leq \rho^2 \inf_{x \in U' \cap \Gamma} S(x) + \inf_{x \in U' \cap \Gamma} L(x) + \epsilon \rho^2 + o(\rho^2), \tag{22}$$

which implies $R_\rho(u) \leq \rho^2 \inf_{x \in U' \cap \Gamma} S(x) + \rho^2 \epsilon + o(\rho^2)$.

Moreover, combining Equation (22) with Assumption 4.1, we have that

$$L(u) \leq \inf_{x \in U' \cap \Gamma} L(x) + O(\rho^2). \tag{23}$$

By Equation (23) and lemma C.7, $\|u - \Phi(u)\|_2 = O(\rho)$. Then by Lemma C.19, we have $R_\rho(u) \geq \rho^2 \inf_{x \in U' \cap \Gamma} S(x) - o(\rho^2)$.

Putting this back into Equation (22), we have $L(u) \leq \inf_{x \in U' \cap \Gamma} L(x) + \epsilon\rho^2 + o(\rho^2)$. This implies,

$$L(u) - \inf_{x \in U' \cap \Gamma} L(x) - (R_\rho(u) - \rho^2 \inf_{x \in U' \cap \Gamma} S(x)) \leq \epsilon\rho^2 + o(\rho^2),$$

Finally notice that Equation (22) is equivalent to

$$L(u) - \inf_{x \in U' \cap \Gamma} L(x) + (R_\rho(u) - \rho^2 \inf_{x \in U' \cap \Gamma} S(x)) \leq \epsilon\rho^2 + o(\rho^2),$$

Putting together the above claims we have our results.

For the $\Leftarrow$ side, we will first lower bound $L(x) + R_\rho(x)$ for $x \in U'$. Suppose $C_{U'}$ is the constant in Assumption 4.1 Define $C_1 = \sqrt{2\frac{C_{U'} + \inf_{x \in \overline{U'} \cap \Gamma} S(x) + 1}{\mu}}$. We discuss by cases. For sufficiently small ρ,

1. If $x \notin K^h$, then by Lemma C.17, $L(x)$ is lower bounded by a positive constant.
2. If $x \in K^h$ and $\|x - \Phi(x)\|_2 \geq C_1\rho$, then by Lemma C.7,

$$L(x) \geq \frac{\mu\|x - \Phi(x)\|_2^2}{2} \geq (C_{U'} + \inf_{x \in U' \cap \Gamma} S(x) + 1)\rho^2 .$$

   This implies $L(x) + R_\rho(x) \geq (\inf_{x \in U' \cap \Gamma} S(x) + 1)\rho^2 + \inf_{x \in U'} L(x)$.
3. If $\|x - \Phi(x)\|_2 \leq C_1\rho$, by Lemma C.19, $R_\rho(x) \geq \rho^2 \inf_{x \in U' \cap \Gamma} S(x) - o(\rho^2)$, hence

$$L(x) + R_\rho(x) + o(\rho^2) \geq \inf_{x \in U' \cap \Gamma} S(x)\rho^2 + \inf_{x \in U'} L(x) .$$

The result then follows directly from this lower bound.

□

Theorem 4.3 implies Corollary C.20 saying that the minimum of the regularized loss is approximately the sum of the minimum of loss and the minimum of limiting regularizer.

Corollary C.20. *Under the setting of Theorem 4.3,*

$$\left| \inf_{x \in U'} (L(x) + R_\rho(x)) - \inf_{x \in U'} L(x) - \rho^2 \inf_{x \in U' \cap \Gamma} S(x) \right| \leq o(\rho^2)$$

Proof of Corollary C.20. In Theorem 4.3, choose $\epsilon = 0$, and choose u such that $L(u) + R_\rho(u) \leq \inf_{x \in U'}(L(x) + R_\rho(x)) + o(\rho^2)$, the result is then clear. □

We will then prove the Corollary 4.4.

Proof of Corollary 4.4. By Theorem 4.3, we have $\|L(u) - \inf_{x \in U'} L(x)\| = O(\rho^2)$. This implies $\|u - \Phi(u)\| = O(\rho)$ and also $\mathrm{dist}(\overline{U' \cap H}, \Phi(u)) = o(1)$.

We also have $\rho^2 S(\Phi(x)) - o(\rho^2) \leq R_\rho(u) \leq \rho^2 \inf_{x \in U' \cap H} S(x)$. We have $\|S(\Phi(x)) - \inf_{x \in U' \cap H} S(x)\| = o(1)$. This further implies $\|\overline{S}(\Phi(x)) - \inf_{x \in U' \cap H} S(x)\| = o(1)$

□

Proof of Theorem 4.5.

1. We will first verify Assumption 4.1. For fixed compact set $B \subset U$, as $\|\nabla^3 L(x)\|_2$ is continuous, there exists constant ν, such that $\forall x \in B^1$, $\|\nabla^3 L(x)\|_2 \leq \nu$. Then by Taylor Expansion,

$$\begin{aligned} R_\rho^{\text{Max}}(x) &= \max_{\|v\|_2 \leq 1} L(x+\rho v) - L(x) \\ &\geq \max_{\|v\|_2 \leq 1} \left(\rho \langle \nabla L(x), v\rangle + \rho^2 v^T \nabla^2 L(x) v/2\right) - \nu\rho^3/6 \\ &\geq -\nu\rho^3/6\,. \end{aligned}$$

2. Now we verify $S^{\text{Max}}(x) = \lambda_1(\nabla^2 L(\cdot))/2$ is the limiting regularizer of R_ρ^{Max}. Let x be any point in Γ, by continuity of R_ρ^{Max},

$$\lim_{\rho\to 0}\lim_{r\to 0}\inf_{\|x'-x\|_2\leq r} \frac{R_\rho^{\text{Max}}(x')}{\rho^2} = \lim_{\rho\to 0}\frac{R_\rho^{\text{Max}}(x)}{\rho^2} = \lambda_1(\nabla^2 L(x))/2\,.$$

3. Finally we verify definition of good limiting regularizer, by Assumption 3.1, $S^{\text{Max}}(x) = \lambda_1(x)/2$ is non-negative and continuous on Γ. For any $x^* \in \Gamma$, choose a sufficiently small open convex set V containing x^* such that $\forall x \in V^1$, $\|\nabla^3 L(x)\|_2 \leq \nu$. For any $x \in V \cap \Gamma$, for any x' satisfying that $\|x'-x\|_2 \leq C\rho$, by Theorem C.67,

$$\begin{aligned} R_\rho^{\text{Max}}(x') &= \max_{\|v\|_2\leq 1} L(x'+\rho v) - L(x') \geq \max_{\|v\|_2\leq 1}\left(\rho\langle \nabla L(x'), v\rangle + \rho^2 v^T\nabla^2 L(x')v/2\right) - \nu\rho^3/6 \\ &\geq \rho^2\lambda_1(\nabla^2 L(x'))/2 - \nu\rho^3/6 \geq \rho^2\lambda_1(\nabla^2 L(x))/2 - O(\rho^3)\,. \end{aligned}$$

This implies $\inf_{\|x'-x\|_2\leq C\rho} R_\rho^{\text{Max}}(x') \geq \rho^2\lambda_1(\nabla^2 L(x))/2 - O(\rho^3)$.

On the other hand, for any $x \in V\cap\Gamma$,

$$\begin{aligned} R_\rho^{\text{Max}}(x) &= \max_{\|v\|_2\leq 1} L(x+\rho v) - L(x) \leq \max_{\|v\|_2\leq 1}\left(\rho\langle \nabla L(x), v\rangle + \rho^2 v^T\nabla^2 L(x)v/2\right) + \nu\rho^3 \\ &= \max_{\|v\|_2\leq 1} \rho^2 v^T\nabla^2 L(x)v/2 + \nu\rho^3 = \rho^2\lambda_1(\nabla^2 L(x'))/2 + O(\rho^3)\,. \end{aligned}$$

This implies $\inf_{\|x'-x\|_2\leq C\rho} R_\rho^{\text{Max}}(x') \leq \rho^2\lambda_1(\nabla^2 L(x))/2 + O(\rho^3)$.

Thus, we conclude that $\left|\inf_{\|x'-x\|_2\leq C\rho} R_\rho^{\text{Max}}(x')/\rho^2 - \lambda_1(\nabla^2 L(x))/2\right| = O(\rho), \forall x \in V\cap\Gamma$, indicating S^{Max} is a good limiting regularizer of R_ρ^{Max} on Γ.

□

Proof of Theorem 4.6.

1. We will first prove Assumption 4.1 holds. For any fixed compact set $B \subset U$, as $\lambda_1(\nabla^2 L)$ and $\|\nabla^3 L\|$is continuous, there exists constant C, such that $\forall x \in B^2$, $\lambda_1(\nabla^2 L) > -\zeta$ and $\|\nabla^3 L(x)\| < \nu$. Then by Taylor Expansion,

$$\begin{aligned} R_\rho^{\text{Asc}}(x) &= L(x + \rho\frac{\nabla L(x)}{\|\nabla L(x)\|}) - L(x) \\ &\geq \left(\rho\|\nabla L(x)\|_2 + \rho^2(\frac{\nabla L(x)}{\|\nabla L(x)\|})^T\nabla^2 L(x)\frac{\nabla L(x)}{\|\nabla L(x)\|}/2\right) - \nu\rho^3/6 \\ &\geq -(\zeta+\nu/6)\rho^2. \end{aligned}$$

2. Now we verify $S^{\text{Asc}}(x) = \text{Tr}(\nabla^2 L(\cdot))/2$ is the limiting regularizer of R_ρ^{Asc}. Let x be any point in Γ. Let $K = \{x\}$ and choose $h = h(K)$ as in Lemma C.5. For any $x' \in K^h \cap U'$,

$$\begin{aligned} R_\rho^{\text{Asc}}(x') &= L(x' + \rho\frac{\nabla L(x')}{\|\nabla L(x')\|}) - L(x') \\ &\geq \rho\|\nabla L(x')\|_2 + \rho^2(\frac{\nabla L(x')}{\|\nabla L(x')\|})^T\nabla^2 L(x')\frac{\nabla L(x')}{\|\nabla L(x')\|}/2 - \nu\rho^3/6 \\ &\geq \rho^2(\frac{\nabla L(x')}{\|\nabla L(x')\|})^T\nabla^2 L(\Phi(x'))\frac{\nabla L(x')}{\|\nabla L(x')\|}/2 - \nu\rho^3/6\,. \end{aligned}$$

By Lemma C.10, we have $\frac{\nabla L(x')}{\|\nabla L(x')\|} = \frac{\nabla^2 L(\Phi(x'))(x'-\Phi(x'))}{\|\nabla^2 L(\Phi(x'))(x'-\Phi(x'))\|_2} + O(\frac{\nu}{\mu}\|x'-\Phi(x')\|_2)$. Hence

$$R_\rho^{\text{Asc}}(x') \geq \rho^2 \lambda_M(\nabla^2 L(\Phi(x')))/2 - \zeta\rho^2 O(\|x'-\Phi(x')\|_2) - \nu\rho^3/6\,.$$

This implies $\lim_{\rho\to 0}\lim_{r\to 0}\inf_{\|x'-x\|_2\leq r}\frac{R_\rho^{\text{Asc}}(x')}{\rho^2} \geq \lambda_M(\nabla^2 L(\Phi(x')))/2$.
We now show the above inequality is in fact equality. If we choose $x_r'' = x + rv_M$, then by Taylor Expansion

$$\begin{aligned}\nabla L(x_r'') &= \nabla L(x) + \nabla^2 L(x)(x_r''-x) + O(\|x_r''-x\|^2)\\ &= rv_M + O(r^2)\end{aligned}$$

This implies

$$\lim_{r\to 0}\frac{\nabla L\left(x_r''\right)}{\|\nabla L\left(x_r''\right)\|} = v_M.$$

We also have $\lim_{r\to 0}\nabla^2 L(x_r'') = \nabla^2 L(x)$ and $\lim_{r\to 0}\nabla^L(x_r'') = 0$. Putting together,

$$\begin{aligned}\lim_{r\to 0} R_\rho^{\text{Asc}}(x_r'') &= \lim_{r\to 0} L(x_r'' + \rho\frac{\nabla L\left(x_r''\right)}{\|\nabla L\left(x_r''\right)\|}) - L(x_r'')\\ &= \lim_{r\to 0}\Big(\rho\|\nabla L\left(x_r''\right)\|_2 + \rho^2(\frac{\nabla L\left(x_r''\right)}{\|\nabla L\left(x_r''\right)\|})^T\nabla^2 L(x_r'')\frac{\nabla L\left(x_r''\right)}{\|\nabla L\left(x_r''\right)\|}/2 + O(\nu\rho^3)\Big)\\ &= \rho^2\lambda_M(\nabla^2 L(x))/2 + O(\rho^3).\end{aligned}$$

This implies $\lim_{\rho\to 0}\lim_{r\to 0}\inf_{\|x'-x\|_2\leq r}\frac{R_\rho^{\text{Asc}}(x')}{\rho^2} \leq \lim_{\rho\to 0}\lim_{r\to 0}\frac{R_\rho^{\text{Asc}}(x_r'')}{\rho^2} = \lambda_M(\nabla^2 L(x))/2$.
Hence the limiting regularizer S is exactly $\lambda_M(\nabla^2 L(\cdot))/2$.

3. Finally we verify definition of good limiting regularizer, by Assumption 3.1, $S^{\text{Max}}(x) = \lambda_M(x)/2$ is non-negative and continuous on Γ. For any $x^* \in \Gamma$, choose a sufficiently small open convex set V containing x^* such that $\forall x \in V^1$, $\|\nabla^3 L(x)\|_2 \leq \nu$. For any $x \in V\cap\Gamma$, for any x' satisfying that $\|x'-x\|_2 \leq C\rho$,

$$\begin{aligned}R_\rho^{\text{Asc}}(x') &= L(x' + \rho\frac{\nabla L\left(x'\right)}{\|\nabla L\left(x'\right)\|}) - L(x')\\ &\geq \rho\|\nabla L\left(x'\right)\|_2 + \rho^2(\frac{\nabla L\left(x'\right)}{\|\nabla L\left(x'\right)\|})^T\nabla^2 L(x')\frac{\nabla L\left(x'\right)}{\|\nabla L\left(x'\right)\|}/2 - \nu\rho^3/6\\ &\geq \rho^2(\frac{\nabla L\left(x'\right)}{\|\nabla L\left(x'\right)\|})^T\nabla^2 L(\Phi(x'))\frac{\nabla L\left(x'\right)}{\|\nabla L\left(x'\right)\|}/2 - \nu\rho^3/6\,.\end{aligned}$$

By Lemma C.10, we have $\frac{\nabla L(x')}{\|\nabla L(x')\|} = \frac{\nabla^2 L(\Phi(x'))(x'-\Phi(x'))}{\|\nabla^2 L(\Phi(x'))(x'-\Phi(x'))\|_2} + O(\frac{\nu}{\mu}\|x'-\Phi(x')\|_2)$. This implies $\inf_{\|x'-x\|_2\leq C\rho} R_\rho^{\text{Asc}}(x') \geq \rho^2\lambda_M(\nabla^2 L(x))/2 - O(\rho^3)$.
On the other hand, simillar to the proof in the second part, we have $\inf_{\|x'-x\|_2\leq C\rho} R_\rho^{\text{Asc}}(x') \leq \rho^2\lambda_M(\nabla^2 L(x))/2 + O(\rho^3)$.
Thus, we conclude that $\left|\inf_{\|x'-x\|_2\leq C\rho} R_\rho^{\text{Max}}(x')/\rho^2 - \lambda_1(\nabla^2 L(x))/2\right| = O(\rho), \forall x \in V\cap\Gamma$, indicating S^{Max} is a good limiting regularizer of R_ρ^{Max} on Γ.

□

Proof of Theorem 4.7.

1. We will first verify Assumption 4.1. For fixed compact set $B \subset U$, as $\|\nabla^3 L(x)\|_2$ is continuous, there exists constant ν, such that $\forall x \in B^1$, $\|\nabla^3 L(x)\|_2 \leq \nu$. Then by Taylor Expansion,

$$\begin{aligned}R_\rho^{\text{Avg}}(x) &= \mathbb{E}_{g\sim N(0,I)} L(x + \rho\frac{g}{\|g\|}) - L(x)\\ &\geq \mathbb{E}_{g\sim N(0,I)}\left(\rho\langle\nabla L(x), \frac{g}{\|g\|}\rangle + \rho^2(\frac{g}{\|g\|})^T\nabla^2 L(x)\frac{g}{2\|g\|}\right) - \nu\rho^3/6\\ &\geq -\nu\rho^3/6\,.\end{aligned}$$

2. Now we verify $S^{\text{Max}}(x) = \text{Tr}(\nabla^2 L(\cdot))/2D$ is the limiting regularizer of R_ρ^{Avg}. Let x be any point in Γ, by continuity of R_ρ^{Avg},

$$\lim_{\rho\to 0}\lim_{r\to 0}\inf_{\|x'-x\|_2\le r}\frac{R_\rho^{\text{Avg}}(x')}{\rho^2} = \lim_{\rho\to 0}\frac{R_\rho^{\text{Avg}}(x)}{\rho^2} = \text{Tr}(\nabla^2 L(x))/2D\,.$$

3. Finally we verify definition of good limiting regularizer, by Assumption 3.1, $S^{\text{Avg}}(x) = \text{Tr}(x)/2D$ is non-negative and continuous on Γ. For any $x^* \in \Gamma$, choose a sufficiently small open convex set V containing x^* such that $\forall x \in V^1, \|\nabla^3 L(x)\|_2 \le \nu$. For any $x \in V \cap \Gamma$, for any x' satisfying that $\|x'-x\|_2 \le C\rho$, by Theorem C.67,

$$\begin{aligned}
R_\rho^{\text{Avg}}(x') &= \mathbb{E}_{g\sim N(0,I)} L(x' + \rho\frac{g}{\|g\|}) - L(x') \\
&\ge \mathbb{E}_{g\sim N(0,I)}\left(\rho\langle\nabla L(x'), \frac{g}{\|g\|}\rangle + \rho^2 \frac{g}{\|g\|}^T \nabla^2 L(x')\frac{g}{\|2g\|}\right) - \nu\rho^3/6 \\
&\ge \rho^2 \text{Tr}(\nabla^2 L(x'))/2D - \nu\rho^3/6 \ge \rho^2\text{Tr}(\nabla^2 L(x))/2D - O(\rho^3)\,.
\end{aligned}$$

This implies $\inf_{\|x'-x\|_2\le C\rho} R_\rho^{\text{Avg}}(x') \ge \rho^2\text{Tr}(\nabla^2 L(x))/2D - O(\rho^3)$.
On the other hand, for any $x \in V \cap \Gamma$,

$$\begin{aligned}
R_\rho^{\text{Avg}}(x) &= \mathbb{E}_{g\sim N(0,I)} L(x + \rho\frac{g}{\|g\|}) - L(x) \\
&\le \mathbb{E}_{g\sim N(0,I)}\left(\rho\langle\nabla L(x), \frac{g}{\|g\|}\rangle + \rho^2 \frac{g}{\|g\|}^T \nabla^2 L(x)\frac{g}{2\|g\|}\right) + \nu\rho^3 \\
&= \mathbb{E}_{g\sim N(0,I)}\rho^2 \frac{g}{\|g\|}^T \nabla^2 L(x)\frac{g}{2\|g\|} + \nu\rho^3 = \rho^2\text{Tr}(\nabla^2 L(x'))/2D + O(\rho^3)\,.
\end{aligned}$$

This implies $\inf_{\|x'-x\|_2\le C\rho} R_\rho^{\text{Avg}}(x') \le \rho^2\text{Tr}(\nabla^2 L(x))/2D + O(\rho^3)$.

Thus, we conclude that $\left|\inf_{\|x'-x\|_2\le C\rho} R_\rho^{\text{Avg}}(x')/\rho^2 - \text{Tr}(\nabla^2 L(x))/2D\right| = O(\rho), \forall x \in V \cap \Gamma$, indicating S^{Avg} is a good limiting regularizer of R_ρ^{Avg} on Γ.

□

Proof of Theorem 5.1. By Theorem C.49, Condition C.2 holds.

Easily deducted from Theorem 4.5 $\Lambda_k(x)$ is a good limiting regularizer for $R_{k,\rho}^{\max}$ on Γ_k. Then as $\Gamma \subset \Gamma_k$, $\Lambda_k(x)$ is a good limiting regularizer for $R_{k,\rho}^{\max}$ on Γ. Hence $S(x) = \sum_k \Lambda_k(x)/2M = \text{Tr}(\nabla^2\text{L(x)})/2$ is a good limiting regularizer of $\mathbb{E}_k[R_{k,\rho}^{\text{Max}}](x)$ on Γ. □

Proof of Theorem 5.2. By Theorem C.49, Condition C.2 holds.

Easily deducted from Theorem 4.6 $\Lambda_k(x)$ is a good limiting regularizer for $R_{k,\rho}^{\text{asc}}$ on Γ_k as the codimension of Γ_k is 1. Then as $\Gamma \subset \Gamma_k$, $\Lambda_k(x)$ is a good limiting regularizer for $R_{k,\rho}^{\max}$ on Γ.Hence $S(x) = \sum_k \Lambda_k(x)/2M = \text{Tr}(\nabla^2 L(x))/2$ is a good limiting regularizer of $\mathbb{E}_k[R_{k,\rho}^{\text{Asc}}](x)$ on Γ. □

C.4 Full-batch SAM on Quadratic Loss (Proof of Theorem 4.14)

The goal of this section is to prove Theorem 4.14. In this section, we use $A \prec B$ to indicate $B - A$ is positive semi-definite.

Theorem 4.14. *Suppose A is a positive definite symmetric matrix with unique top eigenvalue. Consider running SAM on loss $L(x) := \frac{1}{2}x^T Ax$ as (11), then for almost every $x(0)$, we have $x(t)$ converges in direction to $v_1(A)$ and $\lim_{t\to\infty}\|x(t)\| = \frac{\eta\rho\lambda_1(A)}{2-\eta\lambda_1(A)}$ with $\eta\lambda_1(A) < 1$.*

$$x(t+1) = x(t) - \eta A\left(x(t) + \rho\frac{Ax(t)}{\|Ax(t)\|}\right), \tag{11}$$

Proof of Theorem 4.14. We first rewrite the iterate as

$$x(t+1) = x(t) - \eta A x(t) - \eta\rho \frac{A^2 x(t)}{\|Ax(t)\|_2}\,.$$

Define $\tilde{x}(t) \triangleq \frac{\nabla L(x(t))}{\rho} = \frac{Ax(t)}{\rho}$, and we have

$$\tilde{x}(t+1) = \tilde{x}(t) - \eta A \tilde{x}(t) - \eta \frac{A^2 \tilde{x}(t)}{\|\tilde{x}(t)\|_2}\,. \tag{24}$$

We suppose $A \in R^{D\times D}$ and use λ_i, v_i to denote $\lambda_i(A), v_i(A)$.

Further, we define that

$$\begin{aligned}
P^{(j:D)} &\triangleq \sum_{i=j}^{D} v_i(A) v_i(A)^T, \\
\mathbb{I}_j &\triangleq \{\tilde{x} \mid \|P^{(j:D)}\tilde{x}\|_2 \le \eta\lambda_j^2\}\,, \\
\tilde{x}_i(t) &\triangleq \langle \tilde{x}(t), v_i\rangle\,, \\
S &\triangleq \{t \mid \|\tilde{x}(t)\|_2 \le \frac{\eta\lambda_1^2}{2-\eta\lambda_1}, t > T_1\}\,.
\end{aligned}$$

By Lemma C.21, $\mathbb{I}_j$ is an invariant set for update rule Equation (24).

Our proof consists of two steps.

(1) *Entering Invariant Set.* Lemma C.22 implies that there exists constant $T_1 > 0$, such that $\forall t > T_1, \|P^{(j:D)}\tilde{x}(t)\|_2 \le \eta\lambda_j^2$

(2) *Alignment to Top Eigenvector.* Lemmas C.29 and C.30 show that $\|\tilde{x}(t)\|_2$ and $|\tilde{x}_1(t)|$ converge to $\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, which implies our final results.

□

C.4.1 Entering Invariant Set

In this subsection, we will prove the following three lemmas.

1. Lemma C.21 shows $\mathbb{I}_j$ is an invariant set for update rule (Equation (24)).
2. Lemma C.22 shows that under the update rule (Equation (24)), all iterates not in $\mathbb{I}_j$ will shrink exponentially in ℓ_2 norm.
3. Lemma C.23 combines Lemmas C.21 and C.22 to show that for sufficiently large t, $x(t) \in \cap_j \mathbb{I}_j$.

Lemma C.21. *For $t \ge 0$, if $\eta\lambda_1(A) < 1$ and $\tilde{x}(t) \in \mathbb{I}_j$, then $\tilde{x}(t+1) \in \mathbb{I}_j$.*

Proof of Lemma C.21. By (Equation (24)), we have that

$$P^{(j:D)}\tilde{x}(t+1) = (I - P^{(j:D)}\eta A - \eta\frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|_2})P^{(j:D)}\tilde{x}(t)\,.$$

Hence we have that

$$\begin{aligned}
\|P^{(j:D)}\tilde{x}(t+1)\|_2 &= \|(I - P^{(j:D)}\eta A - \eta\frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|_2})P^{(j:D)}\tilde{x}(t)\|_2 \\
&\le \|I - P^{(j:D)}\eta A - \eta\frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|_2}\|_2 \|P^{(j:D)}\tilde{x}(t)\|_2\,.
\end{aligned}$$

Because $\tilde{x}(t) \in \mathbb{I}_j$, $\|\tilde{x}(t)\|_2 \le \frac{\eta\lambda_j^2}{1-\eta\lambda_j}$. This implies,

$$I(1 - \eta\lambda_j - \eta\frac{\lambda_j^2}{\|P^{(j:D)}\tilde{x}(t)\|_2}) \prec I(1 - \eta\lambda_j - \eta\frac{\lambda_j^2}{\|\tilde{x}(t)\|_2}) \prec I - P^{(j:D)}\eta A - \eta\frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|_2} \prec I\,.$$

Furthermore, $\|I - P^{(j:D)}\eta A - \eta\frac{P^{(j:D)}A^2}{\|_2\tilde{x}(t)\|}\|_2 \leq \max(1, \eta\lambda_j + \eta\frac{\lambda_j^2}{\|P^{(j:D)}\tilde{x}(t)\|_2} - 1)$.

Hence it holds that

$$\|P^{(j:D)}\tilde{x}(t+1)\|_2 \leq \max(\|P^{(j:D)}\tilde{x}(t)\|_2, \eta\lambda_j^2 - (1-\eta\lambda_j)\|P^{(j:D)}\tilde{x}(t)\|_2) \leq \eta\lambda_j^2,$$

where the last equality is because $1 - \eta\lambda_j \geq 0$. This above inequality is exactly the definition of $\tilde{x}(t+1) \in \mathbb{I}_j$ and thus is proof is completed. □

Lemma C.22. *For $t \geq 0$, if $\eta\lambda_1(A) < 1$ and $\tilde{x}(t) \notin \mathbb{I}_j$, then*

$$\begin{aligned}\|P^{(j:D)}\tilde{x}(t+1)\|_2 &\leq \max\left(1 - \eta\lambda_D - \eta\frac{\lambda_D^2}{\|\tilde{x}(t)\|_2}, \eta\lambda_j\right)\|P^{(j:D)}\tilde{x}(t)\|_2 \\ &\leq \max\left(1-\eta\lambda_D, \eta\lambda_j\right)\|P^{(j:D)}\tilde{x}(t)\|_2 .\end{aligned} \tag{25}$$

Proof of Lemma C.22. Note that

$$\begin{aligned}\|P^{(j:D)}\tilde{x}(t+1)\|_2 &= \|(I - P^{(j:D)}\eta A - \eta\frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|_2})P^{(j:D)}\tilde{x}(t)\|_2 \\ &\leq \|P^{(j:D)} - P^{(j:D)}\eta A - \eta\frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|_2}\|_2\|P^{(j:D)}\tilde{x}(t)\|_2 .\end{aligned}$$

As $\tilde{x}(t) \notin \mathbb{I}_j$, We have $\|\tilde{x}(t)\|_2 \geq \|P^{(j:D)}\tilde{x}(t)\|_2 > \eta\lambda_j^2$, hence $\eta\frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|_2} \prec \eta\frac{P^{(j:D)}A^2}{\eta\lambda_j^2} \prec P^{(j:D)}$.

This implies that

$$-\eta\lambda_j P^{(j:D)} \prec -P^{(j:D)}\eta A \prec P^{(j:D)} - P^{(j:D)}\eta A - \eta\frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|_2},$$

and

$$P^{(j:D)} - P^{(j:D)}\eta A - \eta\frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|_2} \prec P^{(j:D)}(1-\eta\lambda_D) - \eta\frac{\lambda_D^2}{\|\tilde{x}(t)\|_2}.$$

Hence we have that

$$\begin{aligned}\|P^{(j:D)}\tilde{x}(t+1)\|_2 &\leq \max\left(1 - \eta\lambda_D - \eta\frac{\lambda_D^2}{\|\tilde{x}(t)\|_2}, \eta\lambda_j\right)\|P^{(j:D)}\tilde{x}(t)\|_2. \\ &\leq \max\left(1-\eta\lambda_D, \eta\lambda_j\right)\|P^{(j:D)}\tilde{x}(t)\|_2\end{aligned}$$

This completes the proof. □

Lemma C.23. *Choosing $T_1 = \max_j\left(-\log_{\max(1-\eta\lambda_D, \eta\lambda_j)}\max(\frac{\|\tilde{x}(0)\|_2}{\eta\lambda_j^2}, 1)\right)$, then $\forall t \geq T_1, D > j \geq 1, \tilde{x}(t) \in \mathbb{I}_j$*

Proof of Lemma C.23. We will prove by contradiction. Suppose $\exists j \in [D]$ and $T > T_1$, such that $\tilde{x}(T) \notin \mathbb{I}_j$. By Lemma C.21, it holds that $\forall t < T, \tilde{x}(t) \notin \mathbb{I}_j$. Then by Lemma C.22,

$$\|P^{(j:D)}\tilde{x}(T)\|_2 \leq \max\left(1-\eta\lambda_D, \eta\lambda_j\right)^T\|P^{(j:D)}\tilde{x}(0)\|_2 \leq \eta\lambda_j^2,$$

which leads to a contradiction. □

C.4.2 Alignment to Top Eigenvector

In this subsection, we prove the following lemmas towards showing that $\tilde{x}(t)$ converges in direction to $v_1(A)$ up to a proper sign flip.

1. Lemma C.24 shows that under update rule (Equation (24)), $t \notin S \Rightarrow t+1 \in S$ for sufficiently large t, where the definition of S is $\{t | \|\tilde{x}(t)\|_2 \leq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}, t > T_1\}$.

2. Lemma C.28, a combination of Lemmas C.25 and C.26, shows that following update rule (Equation (24)), $\tilde{x}_1(t)$ increases for $t \in S$.
3. Lemma C.29 shows that $\|\tilde{x}(t)\|$ converges to $\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$ under Equation (24).
4. Lemma C.30 shows that $\|\tilde{x}_1(t)\|_2$ converges to $\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$ under Equation (24).

Lemma C.24. *For $t \geq 0$, if $\|\tilde{x}(t)\|_2 > \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, $\tilde{x}(t) \in \cap\mathbb{I}_j$, then*

$$\|\tilde{x}(t+1)\|_2 \leq \max(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} - \eta\frac{\lambda_D^4}{2\lambda_1^2}, \eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2)$$

Proof of Lemma C.24. Note that

$$\begin{aligned}\tilde{x}(t+1) &= (I - \eta A - \eta\frac{A^2}{\|\tilde{x}(t)\|_2})\tilde{x}(t) \\ &= \frac{1}{\|\tilde{x}(t)\|_2}\sum_{j=1}^{D}\left((1-\eta\lambda_j)\|\tilde{x}(t)\|_2 - \eta\lambda_j^2\right)\tilde{x}_j(t)v_j\end{aligned}$$

Consider the following two cases.

1 If for any i, such that $\left|(1-\eta\lambda_1)\|\tilde{x}(t)\|_2 - \eta\lambda_1^2\right| \geq \left|(1-\eta\lambda_i)\|\tilde{x}(t)\|_2 - \eta\lambda_i^2\right|$, then we have

$$\|\tilde{x}(t+1)\|_2 \leq \left|(1-\eta\lambda_1)\|\tilde{x}(t)\|_2 - \eta\lambda_1^2\right| = \eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2\,.$$

2 If there exists i, such that $\left|(1-\eta\lambda_1)\|\tilde{x}(t)\|_2 - \eta\lambda_1^2\right| < \left|(1-\eta\lambda_i)\|\tilde{x}(t)\|_2 - \eta\lambda_i^2\right|$, then suppose WLOG, i is the smallest among such index.
As

$$\eta\lambda_i^2 - (1-\eta\lambda_i)\|\tilde{x}(t)\|_2 < \eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2 = \left|(1-\eta\lambda_1)\|\tilde{x}(t)\|_2 - \eta\lambda_1^2\right|$$

We have $-\eta\lambda_i^2 + (1-\eta\lambda_i)\|\tilde{x}(t)\|_2 > \eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2$. Equivalently,

$$\|\tilde{x}(t)\|_2 > \frac{\eta\lambda_1^2 + \eta\lambda_i^2}{2-\eta\lambda_1-\eta\lambda_i} \tag{26}$$

Combining with $\tilde{x}(t) \in \mathbb{I}_1 \Rightarrow \|\tilde{x}(t)\|_2 \leq \eta\lambda_1^2$, we have $\eta < \frac{\lambda_1-\lambda_i}{\lambda_1^2}$.
Now consider the following vertors,

$$\begin{aligned}v^{(1)}(t) &\triangleq (\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2)\tilde{x}(t)\,, \\ v^{(2)}(t) &\triangleq ((2-\eta\lambda_1-\eta\lambda_i)\|\tilde{x}(t)\|_2 - \eta\lambda_i^2 - \eta\lambda_1^2)P^{(i:D)}\tilde{x}(t)\,, \\ v^{(2+j)}(t) &\triangleq ((\eta\lambda_{i+j-1} - \eta\lambda_{i+j})\|\tilde{x}(t)\|_2 - \eta\lambda_{i+j}^2 + \eta\lambda_{i+j-1}^2)P^{(i+j:D)}\tilde{x}(t), 1 \leq j \leq D-i\,.\end{aligned}$$

Then we have

$$\begin{aligned}\|\tilde{x}(t+1)\|_2 =& \|\frac{1}{\|\tilde{x}(t)\|_2}\sum_{j=1}^{D}\left((1-\eta\lambda_j)\|\tilde{x}(t)\|_2 - \eta\lambda_j^2\right)\tilde{x}_j(t)v_j\|_2 \\ \leq& \|\frac{1}{\|\tilde{x}(t)\|_2}\sum_{j=1}^{i-1}\left(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2\right)\tilde{x}_j(t)v_j\| + \\ & \|\frac{1}{\|\tilde{x}(t)\|_2}\sum_{j=i}^{D}\left((1-\eta\lambda_j)\|\tilde{x}(t)\|_2 - \eta\lambda_j^2\right)\tilde{x}_j(t)v_j\|_2 \\ \leq& \frac{1}{\|\tilde{x}(t)\|_2}\sum_{j=1}^{D+1-i}\|v^{(j)}\|_2\end{aligned}$$

By assumption, we have $\tilde{x}(t) \in \cap\mathbb{I}_j$, hence we have

$$\begin{aligned}\|v^{(1)}(t)\|_2 &= (\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2)\|\tilde{x}(t)\|_2\,, \\ \|v^{(2)}(t)\|_2 &\leq \eta((2-\eta\lambda_1-\eta\lambda_i)\|\tilde{x}(t)\|_2 - \eta\lambda_i^2 - \eta\lambda_1^2)\lambda_i^2\,, \\ \|v^{(2+j)}(t)\|_2 &\leq \eta((\eta\lambda_{i+j-1} - \eta\lambda_{i+j})\|\tilde{x}(t)\|_2 - \eta\lambda_{i+j}^2 + \eta\lambda_{i+j-1}^2)\lambda_{i+j}^2, 1 \leq j \leq D-i\,.\end{aligned}$$

Using AM-GM inequality, we have

$$\lambda_{i+j-1}\lambda_{i+j}^2 \leq \frac{\lambda_{i+j-1}^3 + 2\lambda_{i+j}^3}{3}\,,$$
$$\lambda_{i+j-1}^2\lambda_{i+j}^2 \leq \frac{\lambda_{i+j-1}^4 + \lambda_{i+j}^4}{2}\,.$$

Hence

$$\begin{aligned}
\|v^{(2+j)}(t)\|_2 &\leq \eta((\eta\lambda_{i+j-1} - \eta\lambda_{i+j})\|\tilde{x}(t)\|_2 - \eta\lambda_{i+j}^2 + \eta\lambda_{i+j-1}^2)\lambda_{i+j}^2 \\
&\leq \eta^2\|\tilde{x}(t)\|_2\frac{\lambda_{i+j-1}^3 - \lambda_{i+j}^3}{3} + \eta^2\frac{\lambda_{i+j-1}^4 - \lambda_{i+j}^4}{2}, 1 \leq j \leq D-i \\
\sum_{j=1}^{D-i}\|v^{(2+j)}(t)\|_2 &\leq \eta^2\|\tilde{x}(t)\|_2\frac{\lambda_i^3 - \lambda_D^3}{3} + \eta^2\frac{\lambda_i^4 - \lambda_D^4}{2}\,.
\end{aligned}$$

Putting together,

$$\begin{aligned}
\|\tilde{x}(t+1)\|_2 \leq& \frac{1}{\|\tilde{x}(t)\|_2}\sum_{j=1}^{D+1-i}\|v^{(i)}\|_2 \\
\leq& \eta\lambda_1^2 + \eta\lambda_i^2(2 - \eta\lambda_1 - \eta\lambda_i) + \eta^2\frac{\lambda_i^3 - \lambda_D^3}{3} - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2 \\
&- \eta^2\lambda_i^2(\lambda_i^2 + \lambda_1^2)\frac{1}{\|\tilde{x}(t)\|_2} + \eta^2\frac{\lambda_i^4 - \lambda_D^4}{2}\frac{1}{\|\tilde{x}(t)\|_2} \\
\leq& \eta\lambda_1^2 + \eta\lambda_i^2(2 - \eta\lambda_1 - \frac{2}{3}\eta\lambda_i) - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2 \\
&- \eta^2\lambda_i^2(\frac{1}{2}\lambda_i^2 + \lambda_1^2)\frac{1}{\|\tilde{x}(t)\|_2} - \eta^2\frac{\lambda_D^4}{2\|\tilde{x}(t)\|_2} \\
\leq& \eta\lambda_1^2 + \eta\lambda_i^2(2 - \eta\lambda_1 - \frac{2}{3}\eta\lambda_i) - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2 \\
&- \eta^2\lambda_i^2(\frac{1}{2}\lambda_i^2 + \lambda_1^2)\frac{1}{\|\tilde{x}(t)\|_2} - \eta\frac{\lambda_D^4}{2\lambda_1^2}\,.
\end{aligned}$$

We further discuss three cases

1. If $\eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}} < \frac{\eta\lambda_1^2+\eta\lambda_i^2}{2-\eta\lambda_1-\eta\lambda_i}$, we have $\|\tilde{x}(t)\|_2 > \frac{\eta\lambda_1^2+\eta\lambda_i^2}{2-\eta\lambda_1-\eta\lambda_i} > \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$,then

$$\begin{aligned}
\|\tilde{x}(t+1)\|_2 \leq& \eta\lambda_1^2 + \eta\lambda_i^2(2 - \eta\lambda_1 - \frac{2}{3}\eta\lambda_i) - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2 \\
&- \eta^2\lambda_i^2(\frac{1}{2}\lambda_i^2 + \lambda_1^2)\frac{1}{\|\tilde{x}(t)\|_2} - \eta\frac{\lambda_D^4}{2\lambda_1^2} \\
\leq& \eta\lambda_1^2 + \eta\lambda_i^2(2 - \eta\lambda_1 - \frac{2}{3}\eta\lambda_i) - (1-\eta\lambda_1)\frac{\eta\lambda_1^2 + \eta\lambda_i^2}{2-\eta\lambda_1-\eta\lambda_i} \\
&- \eta^2\lambda_i^2(\frac{1}{2}\lambda_i^2 + \lambda_1^2)\frac{2-\eta\lambda_1-\eta\lambda_i}{\eta\lambda_1^2+\eta\lambda_i^2} - \eta\frac{\lambda_D^4}{2\lambda_1^2} \\
\leq& \frac{\eta\lambda_1^2}{2-\eta\lambda_1} - \eta\frac{\lambda_D^4}{2\lambda_1^2}\,.
\end{aligned}$$

The second line is because $(1-\eta\lambda_1)\|\tilde{x}(t)\|_2 + \eta^2\lambda_i^2(\frac{1}{2}\lambda_i^2 + \lambda_1^2)\frac{1}{\|\tilde{x}(t)\|_2}$ monotonously increase w.r.t $\|\tilde{x}(t)\|_2$ when $\|\tilde{x}(t)\|_2 > \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$. The last line is due to technical lemma Lemma C.71.

2. If $\eta\lambda_1^2 \geq \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}} \geq \frac{\eta\lambda_1^2+\eta\lambda_i^2}{2-\eta\lambda_1-\eta\lambda_i}$, then

$$\begin{aligned}\|\tilde{x}(t+1)\|_2 \leq & \eta\lambda_1^2 + \eta\lambda_i^2(2-\eta\lambda_1-\frac{2}{3}\eta\lambda_i) - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2 \\ & - \eta^2\lambda_i^2(\frac{1}{2}\lambda_i^2+\lambda_1^2)\frac{1}{\|\tilde{x}(t)\|_2} - \eta\frac{\lambda_D^4}{2\lambda_1^2} \\ \leq & \eta\lambda_1^2 + \eta\lambda_i^2(2-\eta\lambda_1-\frac{2}{3}\eta\lambda_i) - 2\eta\lambda_i\sqrt{(\lambda_1^2+\frac{1}{2}\lambda_i^2)(1-\eta\lambda_1)} - \eta\frac{\lambda_D^4}{2\lambda_1^2} \\ \leq & \frac{\eta\lambda_1^2}{2-\eta\lambda_1} - \eta\frac{\lambda_D^4}{2\lambda_1^2}\,.\end{aligned}$$

The second line is because of AM-GM inequality. The last line is due to technical lemma Lemma C.73.

3. If $\eta\lambda_1^2 < \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$, we have $\|\tilde{x}(t)\|_2 < \eta\lambda_1^2 < \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$, then

$$\begin{aligned}\|\tilde{x}(t+1)\|_2 \leq & \eta\lambda_1^2 + \eta\lambda_i^2(2-\eta\lambda_1-\frac{2}{3}\eta\lambda_i) - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2 \\ & - \eta^2\lambda_i^2(\frac{1}{2}\lambda_i^2+\lambda_1^2)\frac{1}{\|\tilde{x}(t)\|_2} - \eta\frac{\lambda_D^4}{2\lambda_1^2} \\ \leq & \eta\lambda_1^2 + \eta\lambda_i^2(2-\eta\lambda_1-\frac{2}{3}\eta\lambda_i) - (1-\eta\lambda_1)\eta\lambda_1^2 \\ & - \eta\lambda_i^2(\frac{1}{2}\lambda_i^2+\lambda_1^2)\frac{1}{\lambda_1^2} - \eta\frac{\lambda_D^4}{2\lambda_1^2} \\ \leq & \frac{\eta\lambda_1^2}{2-\eta\lambda_1} - \eta\frac{\lambda_D^4}{2\lambda_1^2}\,.\end{aligned}$$

The second line is because $(1-\eta\lambda_1)\|\tilde{x}(t)\|_2 + \eta^2\lambda_i^2(\frac{1}{2}\lambda_i^2+\lambda_1^2)\frac{1}{\|\tilde{x}(t)\|_2}$ monotonously decrease w.r.t $\|\tilde{x}(t)\|_2$ when $\|\tilde{x}(t)\|_2 < \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$. The last line is due to technical lemma Lemma C.72.

□

Lemma C.25. *if* $\|\tilde{x}(t)\|_2 \leq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, *it holds that*

$$|\tilde{x}_1(t+1)| \geq |\tilde{x}_1(t)|\,.$$

Proof. We have $|\tilde{x}_1(t+1)| = |1-\eta\lambda_1 - \eta\frac{\lambda_1^2}{\|\tilde{x}(t)\|_2}||\tilde{x}_1(t)|$.

We also have $\eta\frac{\lambda_1^2}{\|\tilde{x}(t)\|_2} > 2-\eta\lambda_1^2$, hence $1-\eta\lambda_1-\eta\frac{\lambda_1^2}{\|\tilde{x}(t)\|_2} < -1$.

Hence we have that $|\tilde{x}_1(t+1)| > |\tilde{x}_1(t)|$. □

Lemma C.26. *For any* $t \geq 0$, *if* $\|\tilde{x}(t)\|_2 \leq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, $\tilde{x}(t) \in \cap\mathbb{I}_j$, *it holds that*

$$\|\tilde{x}(t+1)\|_2 \leq \eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2\,.$$

Proof. The result holds as

$$\|I-\eta A-\eta\frac{A^2}{\|\tilde{x}(t)\|_2}\|_2 \leq \max_{1\leq j\leq D}\{|1-\eta\lambda_j-\eta\frac{\lambda_j^2}{\|\tilde{x}(t)\|}|\} = \eta\frac{\lambda_1^2}{\|\tilde{x}(t)\|} - (1-\eta\lambda_j)\,.$$

□

Lemma C.27. *For any $t \geq 0$, if $\|\tilde{x}(t)\|_2 \leq \frac{\eta\lambda_1^2}{1-\eta\lambda_1}$, it holds that*

$$\begin{aligned}&\|\tilde{x}(t+1)\|_2\\ \leq&(\lambda_1^2-(1+\eta\lambda_1)\|\tilde{x}(t)\|_2)\times\\ &\sqrt{\frac{|\tilde{x}_1(t)|^2}{\|\tilde{x}(t)\|^2}+\left(\max\{\frac{|(1-\eta\lambda_2)\|\tilde{x}(t)\|_2-\eta\lambda_2^2|}{\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2},\frac{|(1-\eta\lambda_D)\|\tilde{x}(t)\|_2-\eta\lambda_D^2|}{\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2}\}\right)^2(1-\frac{|\tilde{x}_1(t)|^2}{\|\tilde{x}(t)\|^2})}.\end{aligned}$$

Proof. We will discuss the movement along v_1 and orthogonal to v_1. First,

$$\begin{aligned}\|P^{(2:D)}\tilde{x}(t+1)\|_2 &= \|(I-P^{(2:D)}\eta A-\eta\frac{P^{(2:D)}A^2}{\|\tilde{x}(t)\|_2})P^{(2:D)}\tilde{x}(t)\|_2\\ &\leq \|P^{(2:D)}-P^{(2:D)}\eta A-\eta\frac{P^{(2:D)}A^2}{\|\tilde{x}(t)\|_2}\|_2\|P^{(2:D)}\tilde{x}(t)\|_2\\ &\leq \max\{|1-\eta\lambda_2-\frac{\eta\lambda_2^2}{\|\tilde{x}(t)\|_2}|,|1-\eta\lambda_D-\frac{\eta\lambda_D^2}{\|\tilde{x}(t)\|_2}|\}\|P^{(2:D)}\tilde{x}(t)\|_2\,.\end{aligned}$$

Second

$$|\tilde{x}_1(t+1)| = (\frac{\lambda_1^2}{\|\tilde{x}(t)\|_2}-1+\eta\lambda_1)|\tilde{x}_1(t)|.$$

Hence we have that

$$\begin{aligned}&\|\tilde{x}(t+1)\|_2\\ \leq&(\lambda_1^2-(1+\eta\lambda_1)\|\tilde{x}(t)\|_2)\times\\ &\sqrt{\frac{|\tilde{x}_1(t)|^2}{\|\tilde{x}(t)\|^2}+\left(\max\{\frac{|(1-\eta\lambda_2)\|\tilde{x}(t)\|_2-\eta\lambda_2^2|}{\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2},\frac{|(1-\eta\lambda_D)\|\tilde{x}(t)\|_2-\eta\lambda_D^2|}{\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2}\}\right)^2(1-\frac{|\tilde{x}_1(t)|^2}{\|\tilde{x}(t)\|^2})}.\end{aligned}$$

□

Lemma C.28. *For $t, t' \in S, 0 \leq t \leq t'$, then $|\tilde{x}_1(t)| \leq |\tilde{x}_1(t')|$.*

Proof. For $t \in S$, by Lemma C.24, $t+1 \in S$ or $t+1 \notin S, t+2 \in S$. We will discuss by case.

1. If $t+1 \in S$, we can use Lemma C.25 to show $|\tilde{x}_1(t)| \leq |\tilde{x}_1(t+1)|$.
2. If $t+1 \notin S, t+2 \in S$, then

$$|\tilde{x}_1(t+2)| = \frac{(\eta\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2)(\eta\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t+1)\|_2)}{\|\tilde{x}(t)\|_2\|\tilde{x}(t+1)\|_2}|\tilde{x}_1(t)|\,.$$

As

$$\begin{aligned}&(\eta\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2)(\eta\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t+1)\|_2) \geq \|\tilde{x}(t)\|_2\|\tilde{x}(t+1)\|_2\\ \iff&\eta^2\lambda_1^4-\eta\lambda_1^2(1-\eta\lambda_1)(\|\tilde{x}(t)\|_2+\|\tilde{x}(t+1)\|_2)\\ &\geq(2\eta\lambda_1-\eta^2\lambda_1^2)\|\tilde{x}(t)\|_2\|\tilde{x}(t+1)\|_2\\ \iff&\eta^2\lambda_1^4-\eta\lambda_1^2(1-\eta\lambda_1)\|\tilde{x}(t)\|_2\\ &\geq\left((2\eta\lambda_1-\eta^2\lambda_1^2)\|\tilde{x}(t)\|_2+\eta\lambda_1^2(1-\eta\lambda_1)\right)\|\tilde{x}(t+1)\|_2\,,\end{aligned}$$

combining with Lemma C.26, we only need to prove,

$$\begin{aligned}&\eta^2\lambda_1^4-\eta\lambda_1^2(1-\eta\lambda_1)\|\tilde{x}(t)\|_2\\ &\geq\left((2\eta\lambda_1-\eta^2\lambda_1^2)\|\tilde{x}(t)\|_2+\eta\lambda_1^2(1-\eta\lambda_1)\right)\left(\eta\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2\right)\,.\end{aligned}$$

Through some calculation, this is equivalent to

$$((2-\eta\lambda_1)\|\tilde{x}(t)\|_2-\eta\lambda_1^2)((1-\eta\lambda_1)\|\tilde{x}(t)\|_2-\eta\lambda_1^2) \geq 0\,.$$

which holds for $\|\tilde{x}(t)\|_2 \leq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$.

Combining the two cases and using induction, we can get the desired result.

□

Lemma C.29. *$\|\tilde{x}(t)\|$ converges to $\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$ when $t \to \infty$.*

Proof. By Lemma C.28, $|\tilde{x}_1(t)|$ increases monotonously for $t \in S$. By Lemma C.24, S is infinite. By Lemma C.22, for sufficiently large t, $|\tilde{x}_1(t)|$ is bounded. Combining the three facts, we know $\tilde{x}_1(t)$ for $t \in S$ converges.

Formally $\forall \epsilon > 0$, there exists $T_\epsilon > 0$ such that $\forall t, t' \in S, t' > t > T_\epsilon$, $\frac{\|\tilde{x}_1(t')\|_2}{\|\tilde{x}_1(t)\|_2} < 1 + \epsilon$.

Then by Lemma C.24, $\forall t \in S, t+1 \in S$ or $t+2 \in S$, we will discuss by case. For $t \geq T_\epsilon$,

1. If $t+1 \in S$, then

$$1 + \epsilon \geq \frac{\|\tilde{x}_1(t+1)\|_2}{\|\tilde{x}_1(t)\|_2} = \frac{\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2}{\|\tilde{x}(t)\|_2}.$$

2. If $t+1 \notin S$ and $t+2 \in S$, then

$$\begin{aligned}
1 + \epsilon &\geq \frac{\|\tilde{x}_1(t+2)\|_2}{\|\tilde{x}_1(t)\|_2} \\
&= \frac{(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2)(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t+1)\|_2)}{\|\tilde{x}(t)\|_2\|\tilde{x}(t+1)\|_2} \\
&\geq \frac{(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2)\left(\eta\lambda_1^2 - (1-\eta\lambda_1)\left(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2\right)\right)}{\|\tilde{x}(t)\|_2\left(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2\right)} \\
&= \frac{\eta\lambda_1^2 - (1-\eta\lambda_1)\left(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|_2\right)}{\|\tilde{x}(t)\|_2}.
\end{aligned}$$

   Here in the last inequality, we apply Lemma C.26.

Concluding, $\|\tilde{x}(t)\|_2 \geq \min\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1^2+\epsilon}, \frac{\eta^2\lambda_1^3}{(2-\lambda_1\eta)\lambda_1\eta+\epsilon}\right), \forall t > T_\epsilon, t \in S$. As $\forall t \notin S, t > T_\epsilon$, we have $\|\tilde{x}(t)\|_2 \geq \frac{\eta\lambda_1^2}{2-\eta\lambda_1^2}$. Hence we have $\forall t > T_\epsilon, \|\tilde{x}(t)\|_2 \geq \min\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1^2+\epsilon}, \frac{\eta^2\lambda_1^3}{(2-\lambda_1\eta)\lambda_1\eta+\epsilon}\right)$.

Further by Lemma C.26, $\forall t > T_\epsilon + 1, \|\tilde{x}(t)\|_2 \leq \eta\lambda_1^2 - (1-\eta\lambda_1)\min\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1^2+\epsilon}, \frac{\eta^2\lambda_1^3}{(2-\lambda_1\eta)\lambda_1\eta+\epsilon}\right)$.

Combining both bound, we have $\lim_{t\to\infty} \|\tilde{x}(t)\|_2 = \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$. □

Lemma C.30. *$\|\tilde{x}_1(t)\|_2$ converges to $\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, when $t \to \infty$.*

Proof. Notice that

$$\|P^{(2:D)}\tilde{x}(t+1)\|_2 \leq \max\left(|1 - \eta\lambda_2 - \eta\frac{\lambda_2^2}{\|\tilde{x}(t)\|_2}|, |1 - \eta\lambda_D - \eta\frac{\lambda_D^2}{\|\tilde{x}(t)\|_2}|\right)\|P^{(2:D)}\tilde{x}(t)\|_2.$$

When $\|\tilde{x}(t)\|_2 > \frac{\eta\lambda_2^2}{2-\eta\lambda_2-\delta}$,

$$\begin{aligned}
-1 + \delta \leq 1 - \eta\lambda_2 - \eta\frac{\lambda_2^2}{\|\tilde{x}(t)\|_2} \leq 1 - \eta\lambda_D - \eta\frac{\lambda_D^2}{\|\tilde{x}(t)\|_2} \leq 1 - \eta\lambda_D \\
\|P^{(2:D)}\tilde{x}(t+1)\|_2 \leq \max(1-\eta\lambda_D, 1-\delta)\|P^{(2:D)}\tilde{x}(t)\|_2
\end{aligned}$$

Hence for sufficiently large t, $\|P^{(2:D)}\tilde{x}(t)\|_2$ shrinks exponentially, showing that $\lim_{t\to\infty} \|\tilde{x}_1(t)\|_2 = \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$. □

C.5 FULL-BATCH SAM ON GENERAL LOSS (PROOF OF THEOREM 4.9)

The goal of this section is to prove the following theorem.

Theorem 4.9 (Main, Theorems C.31 and C.32 stated informally)**.** *Let $\{x(t)\}$ be the iterates of SAM (3) with $x(0) = x_{init} \in U$, then under Assumptions 3.1 and 4.8, for all η, ρ such that $\eta \ln(1/\rho)$ and ρ/η are sufficiently small, the dynamics of SAM can be split into two phases:*

- ***Phase I:*** *SAM follows Gradient Flow with respect to L until entering an $O(\eta\rho)$ neighborhood of the manifold Γ in $\tilde{O}(\frac{1}{\eta})$ steps;*
- ***Phase II:*** *SAM tracks the solution X of (6), the Riemannian Gradient Flow with respect to $\lambda_1(\nabla^2 L(\cdot))$ in the $O(\eta\rho)$ neighborhood in the sense that $\max_{0\le T\le T_3} \|\Phi(x(\lceil T/(\eta\rho^2)\rceil)) - X(T)\| = O((\eta+\rho)\log(1/\eta\rho))$. Moreover, the angle between $\nabla L(x(t))$ and $v_1(\nabla^2 L(x(t)))$ is at most $O(\rho)$.*

To prove the theorem, we will separate the dynamic of SAM on general loss L to two phases.

Define

$$R_j(x) = \sqrt{\sum_{i=j}^{M} \lambda_i^2(x)\langle v_i(x), x - \Phi(x)\rangle^2 - \eta\rho\lambda_j(x)}, j \in [M], x \in U,$$

which is the length projection of $x - \Phi(x)$ on button$-k$ non-zero eigenspace of $\nabla^2 L(\Phi(x))$. We will provide a fine-grained convergence bound on $R_j(x)$.

Theorem C.31 (Phase I)**.** *Let $\{x(t)\}$ be the iterates defined by SAM (Equation (3)) and $x(t) = x_{init} \in U$, then under Assumption 3.1 there exists a constant T_1, such that for any $T_1' > T_1$, it holds for all η, ρ such that $(\eta+\rho)\ln(1/\eta\rho)$ is sufficiently small, we have*

$$\max_{T_1 \ln(1/\eta\rho)\le \eta t\le T_1' \ln(1/\eta\rho)} \max_{j\in[M]} R_j(x) = O(\eta\rho^2)$$
$$\max_{T_1 \ln(1/\eta\rho)\le \eta t\le T_1' \ln(1/\eta\rho)} \|\Phi(x(t)) - \Phi(x_{init})\| = O((\eta+\rho)\ln(1/\eta\rho))$$

Theorem C.31 implies SAM will converge to an $O(\eta\rho)$ neighbor of Γ. Notice in the time frame defined by Theorem C.31, $x(t)$ effectively operates at a local regime around $\Phi(\lceil T_1 \ln(1/\eta\rho)/\eta\rceil)$, this allows us to approximate L with the quadratic Taylor expansion of L at $\Phi(\lceil T_1 \ln(1/\eta\rho)/\eta\rceil)$ and give us the following theorem.

Theorem C.32 (Phase II)**.** *Let $\{x(t)\}$ be the iterates defined by SAM (Equation (3)) under Assumptions 3.1 and 4.8, for all η, ρ such that $\eta\ln(1/\rho)$ and ρ/η is sufficiently small, further assuming that (1) $\max_j R_j(x(0)) = O(\eta\rho^2)$, (2) $\|\Phi(x(0)) - \Phi(x_{init})\| = O((\eta+\rho)\ln(1/\eta\rho))$ and (3) $|\langle x(0) - \Phi(x(0)), v_1(x(0))\rangle| \ge \Omega(\rho^2)$, the iterates $x(t)$ tracks the solution X of Equation (6). Quantitatively for $t = \lceil T_3/\eta\rho^2\rceil$, we have that*

$$\|\Phi(x(t)) - X(\eta\rho^2 t)\| = O(\eta\ln(1/\rho)).$$

Moreover the angle between $\nabla L(x(t))$ and the top eigenspace of $\nabla^2 L(\Phi(x(t)))$ is at most $O(\rho)$. Quantitatively,

$$|\langle x(t) - \Phi(x(t)), v_1(x(t))\rangle| = \Theta(\eta\rho).$$
$$\max_{j\in[2:M]} |\langle x(t) - \Phi(x(t)), v_j(x(t))\rangle| = O(\eta\rho^2).$$

In this section we will define K as $\{X(t) \mid 0 \le t \le T_3\}$ where X is the solution of Equation (6). To simplify our proof, we assume WLOG $L(x) = 0$ for $x \in \Gamma$.

C.5.1 PHASE I (PROOF OF THEOREM C.31)

The goal of this section is to prove Theorem C.31.

Proof. The proof consists of three major steps.

1. *Tracking Gradient Flow* Lemma C.33 shows the existence of step $t_{\mathrm{GF}} = O(1/\eta)$ such that $x(t_{\mathrm{GF}})$ is in a subset of K^h and $\Phi(x(t_{\mathrm{GF}}))$ is $O(\eta+\rho)$ close to $\Phi(x_{\text{init}})$.
2. *Decreasing Loss* Lemma C.35 shows the existence of step $t_{\mathrm{DEC}} = O(\ln(1/\rho)/\eta)$ such that $x(t_{\mathrm{DEC}})$ is in $O(\rho)$ neighbor of Γ and $\Phi(x(t_{\mathrm{DEC}}))$ is $O((\eta+\rho)\ln(1/\rho))$ close to $\Phi(x_{\text{init}})$.
3. *Entering Invariant Set* Lemmas C.40 and C.42 shows the existence of step $t_{\mathrm{INV}} = O(\ln(1/\rho\eta)/\eta)$ such that for any t satisfying $t_{\mathrm{INV}} \le t \le t_{\mathrm{INV}} + \Theta(\ln(1/\eta)/\eta)$, we have that $x(t) \in \cap_{k\in[M]}\mathbb{I}_k$ and $\Phi(x(t))$ is $O((\eta+\rho)\ln(1/\eta\rho))$ close to $\Phi(x_{\text{init}})$.

□

Tracking Gradient Flow Lemma C.33 shows that the iterates $x(t)$ tracks gradient flow to an $O(1)$ neighbor of Γ.

Lemma C.33. *Under condition of Theorem C.31, there exists $t_{\mathrm{GF}} = O(1/\eta)$, such that the iterate $x(t_{\mathrm{GF}})$ is $O(1)$ close to the manifold Γ and $\Phi(x(t_{\mathrm{GF}}))$ is $O(\eta+\rho)$ is close to $\Phi(x_{init})$. Quantitatively,*

$$\begin{aligned} L(x(t_{\mathrm{GF}})) &\le \frac{\mu h^2}{32} \\ \|x(t_{\mathrm{GF}}) - \Phi(x(t_{\mathrm{GF}}))\| &\le h/4\,, \\ \|\Phi(x(t_{\mathrm{GF}})) - \Phi(x_{init})\| &= O(\eta+\rho)\,. \end{aligned}$$

Proof. Choose $C = \frac{1}{4}\sqrt{\frac{\mu}{\zeta}}$.

There exists $T > 0$, such that

$$\|\phi(x_{\text{init}}, T) - \Phi(x_{\text{init}})\|_2 \le Ch/2\,.$$

Consider

$$\begin{aligned} x(t+1) &= x(t) - \eta\nabla L(x(t) + \rho\frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|}) \\ &= x(t) - \eta\nabla L(x(t)) + O(\eta\rho)\,. \end{aligned}$$

By Theorem D.5, by choosing $p = \eta$ and $q = \eta+\rho$, the iterates $x(t)$ tracks gradient flow $\phi(x_{\text{init}}, T)$ in $O(1/\eta)$ steps. Quantitatively for $t_{\mathrm{GF}} = \lceil\frac{T}{\eta}\rceil$, we have that

$$\|x(t_{\mathrm{GF}}) - \phi(x_{\text{init}}, T)\|_2 \le O(\eta+\rho)\,.$$

This implies $x(t_{\mathrm{GF}}) \in K^h$, hence by Taylor Expansion on Φ,

$$\begin{aligned} \|\Phi(x(t_{\mathrm{GF}})) - \Phi(x_{\text{init}})\|_2 &= \|\Phi(x(t_{\mathrm{GF}})) - \Phi(\phi(x_{\text{init}}, T))\|_2 \\ &\le O(\|x(t_{\mathrm{GF}}) - \phi(x_{\text{init}}, T)\|_2) \\ &\le O(\eta+\rho)\,. \end{aligned}$$

This implies

$$\begin{aligned} \|x(t_{\mathrm{GF}}) - \Phi(x(t_{\mathrm{GF}}))\|_2 \le& \|x(t_{\mathrm{GF}}) - \phi(x_{\text{init}}, T_0)\|_2 + \|\phi(x_{\text{init}}, T_0) - \Phi(x_{\text{init}})\|_2 \\ &+ \|\Phi(x_{\text{init}}) - \Phi(x(t_{\mathrm{GF}}))\|_2 \\ \le& Ch/2 + O(\eta+\rho) \le Ch \le h/4\,. \end{aligned}$$

By Taylor Expansion,

$$L(x(t_{\mathrm{GF}})) \le \zeta\|x(t_{\mathrm{GF}}) - \Phi(x(t_{\mathrm{GF}}))\|_2^2/2 \le \frac{\mu h^2}{32}$$

□

Decreasing Loss Lemma C.35 shows that the iterates $x(t)$ converges to an $O(\rho)$ neighbor of Γ in $O(\ln(1/\rho)/\eta)$ steps.

Lemma C.34. *Under condition of Theorem C.31, if $x(t) \in K^h$ and $\|\nabla L(x(t))\| \geq 4\zeta\rho$, then we have that $L(x(t+1))$ decreases with respect to $L(x(t))$, quantitatively, we have that*

$$L(x(t+1)) \leq L(x(t))(1-\eta\mu/8)\,.$$

Moreover the movement of the projection of the iterates on the manifold is bounded, quantitatively, we have that

$$\|\Phi(x(t+1)) - \Phi(x(t))\| \leq O(\eta^2)\,.$$

Proof. As $x(t) \in K^h$ and L is μ-PL in K^h, we have $L(x(t)) \geq 0$.

As $x(t) \in K^h$, by Lemma C.12 and Taylor Expansion, we have $\|\overline{x(t)x(t+1)}\| = O(\eta)$. hence for sufficiently small η, $\overline{x(t)x(t+1)} \subset K^r$. Using similar argument, the segment from $x(t)$ to $x(t) + \rho\frac{\nabla L(x(t))}{\|\nabla L(x(t))\|}$ is in K^r.

Then by Taylor Expansion on L,

$$\begin{aligned} L(x(t+1)) &= L(x(t) - \eta\nabla L\left(x(t) + \rho\frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|}\right)) \\ &\leq L(x(t)) - \eta\left\langle \nabla L\left(x(t)\right), \nabla L\left(x(t) + \rho\frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|}\right)\right\rangle \\ &+ \frac{\zeta\eta^2\|\nabla L\left(x(t) + \rho\frac{\nabla L(x(t))}{\|\nabla L(x(t))\|}\right)\|^2}{2}\,. \end{aligned} \tag{27}$$

By Taylor Expansion on ∇L, we have that

$$\|\nabla L\left(x(t) + \rho\frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|}\right) - \nabla L\left(x(t)\right)\| \leq \zeta\rho\,.$$

After plugging in Equation (27), we have that

$$L(x(t+1)) \leq L(x(t)) - \eta\|\nabla L\left(x(t)\right)\|^2 + \eta\zeta\rho\|\nabla L\left(x(t)\right)\| + \zeta\eta^2\|\nabla L\left(x(t)\right)\|^2 + \zeta^3\eta^2\rho^2\,. \tag{28}$$

As $\|\nabla L(x(t))\| \geq 4\zeta\rho$, we have that the following term is bounded.

$$\begin{aligned} \zeta\eta^2\|\nabla L\left(x(t)\right)\|^2 &\leq \frac{1}{2}\eta\|\nabla L\left(x(t)\right)\|^2\,, \\ \eta\zeta\rho\|\nabla L\left(x(t)\right)\| &\leq \frac{1}{4}\eta\|\nabla L\left(x(t)\right)\|^2\,, \\ \zeta^3\eta^2\rho \leq \zeta^2\eta\rho^2 &\leq \frac{1}{16}\eta\|\nabla L\left(x(t)\right)\|^2\,. \end{aligned}$$

After plugging in Equation (28), by Lemma C.8,

$$\begin{aligned} L(x(t+1)) &\leq L(x(t)) - \frac{1}{16}\eta\|\nabla L\left(x(t)\right)\|^2 \\ &\leq L(x(t))(1-\eta\mu/8)\,. \end{aligned}$$

As $x(t) \in K^h$, by Taylor Expansion, we have

$$\|\nabla L\left(x(t)\right)\| \leq \zeta h\,.$$

Hence by Lemma C.12 and Taylor Expansion,

$$\begin{aligned} \|\Phi(x(t+1)) - \Phi(x(t))\| &\leq \xi\eta\rho\|\nabla L\left(x\right)\|_2 + \nu\eta\rho^2 + \xi\eta^2\|\nabla L\left(x\right)\|_2^2 + \xi\zeta^2\eta^2\rho^2 \\ &= O(\eta^2)\,. \end{aligned}$$

□

Lemma C.35. *Under condition of Theorem C.31, assuming there exists t_{GF} such that $L(x(t_{\mathrm{GF}})) \leq \frac{\mu h^2}{32}$ and $x(t_{\mathrm{GF}}) \in K^{h/4}$, then there exists $t_{\mathrm{DEC}} = t_{\mathrm{GF}} + O(\ln(1/\rho)/\eta)$, such that $x(t_{\mathrm{DEC}})$ is in $O(\rho)$ neighbor of Γ, quantitatively, we have that*

$$\|\nabla L(x(t_{\mathrm{DEC}}))\|_2 \leq 4\zeta\rho\,.$$

Moreover the movement of the projection of $\Phi(x(\cdot))$ on the manifold is bounded,

$$\|\Phi(x(t_{\mathrm{GF}})) - \Phi(x(t_{\mathrm{DEC}}))\|_2 = O(\eta \ln(1/\rho))\,.$$

Proof. Choose t_{DEC} as the minimal $t \geq t_{\mathrm{GF}}$ such that $\|\nabla L(x(t_{\mathrm{DEC}}))\|_2 \leq 4\zeta\rho$. Define $C = \lceil \ln_{1-\frac{\eta\mu}{8}}(64\rho^2/h^2)\rceil = O(\ln(1/\rho)/\eta)$.

We will first perform an induction on $t \leq \min\{t_{\mathrm{DEC}}, t_{\mathrm{GF}} + C\} = t_{\mathrm{GF}} + O(\ln(1/\rho)/\eta)$ to show that

$$\begin{aligned} L(x(t)) &\leq (1-\eta\mu/8)^{t-t_{\mathrm{GF}}} L(x(t_{\mathrm{GF}})) \\ \|\Phi(x(t)) - \Phi(x(t_{\mathrm{GF}}))\| &= O(\eta^2(t - t_{\mathrm{GF}})) \end{aligned}$$

For $t = t_{\mathrm{GF}}$, the result holds trivially. Suppose the induction hypothesis holds for t. Then by C.7 and Taylor Expansion,

$$\begin{aligned} \|\Phi(x(t)) - x(t)\| &\leq \sqrt{\frac{2L(x(t_{\mathrm{GF}}))}{\mu}} \\ &\leq h/4\,. \end{aligned}$$

Then we have that

$$\begin{aligned} \mathrm{dist}(K, x(t)) \leq& \mathrm{dist}(K, x(t_{\mathrm{GF}})) + \|x(t_{\mathrm{GF}}) - \Phi(x(t_{\mathrm{GF}}))\|_2 \\ &+ \|\Phi(x(t_{\mathrm{GF}})) - \Phi(x(t))\| + \|\Phi(x(t)) - x(t)\| \\ \leq& 3h/4 + O(\eta^2(t - t_{\mathrm{GF}})) = 3h/4 + O(\eta\ln(1/\rho)) \leq h\,. \end{aligned}$$

That is $x(t) \in K^h$. Then as $t \leq t_{\mathrm{DEC}}$, $\|\nabla L(x(t))\|_2 \geq 4\zeta\rho$. Then by Lemma C.34, we have that

$$\begin{aligned} L(x(t+1)) &\leq (1-\eta\mu/8)L(x(t)) \leq (1-\eta\mu/8)^{t+1-t_{\mathrm{GF}}} L(x(t_{\mathrm{GF}}))\,, \\ \|\Phi(x(t+1)) - \Phi(x(t_{\mathrm{GF}}))\| &\leq \|\Phi(x(t+1)) - \Phi(x(t))\| + \|\Phi(x(t)) - \Phi(x(t_{\mathrm{GF}}))\| \\ &\leq O(\eta^2(t - t_{\mathrm{GF}}))\,. \end{aligned}$$

The induction is complete.

Now if $t_{\mathrm{DEC}} \geq t_{\mathrm{GF}} + C = t_{\mathrm{GF}} + O(\ln(1/\rho)/\eta)$, As the result of the induction, we have that

$$\begin{aligned} L(x(t_{\mathrm{GF}} + C)) &\leq (1 - \frac{\eta\mu}{8})^C L(t_{\mathrm{GF}}) \\ &\leq \frac{64\rho^2}{h^2} L(t_{\mathrm{GF}}) \\ &\leq 8\rho^2\mu\,. \end{aligned}$$

By Lemma C.8, we have that

$$\|\nabla L(x(t_{\mathrm{GF}} + C))\|_2 \leq \zeta\sqrt{\frac{2L(x(t_{\mathrm{GF}} + C))}{\mu}} = 4\zeta\rho\,.$$

This is a contradiction.

Hence we have that $t_{\mathrm{DEC}} \leq t_{\mathrm{GF}} + C = t_{\mathrm{GF}} + O(\ln(1/\rho)/\eta)$. By induction, we have that

$$\|\Phi(x(t_{\mathrm{DEC}})) - \Phi(x(t_{\mathrm{GF}}))\| = O(\eta^2(t_{\mathrm{DEC}} - t_{\mathrm{GF}})) = O(\eta\ln(1/\rho))\,.$$

□

Entering Invariant Set We will introduce some notations that is required for the proof in this and following subsection.

Define

$$\begin{aligned}\hat{x} &= x - \Phi(x)\,,\\ A(x) &= \nabla^2 L\left(\Phi(x)\right)\,,\\ \tilde{x} &= A(x)\hat{x}\,,\\ \tilde{x}_j &= \langle \tilde{x}, v_j(x)\rangle\,,\\ P^{(j:D)}(x) &= \sum_{i=j}^{M} v_i(x) v_i^T(x)\,.\end{aligned}$$

Note $\tilde{x} \approx \nabla L(x)$ for x near the manifold Γ. We also use $\tilde{x}(t),\ A(t)$ and $\hat{x}(t)$ to denote $\tilde{x(t)},\ A(x(t))$ and $\hat{x(t)}$.

Recall the original definition of $R_j(x)$ is

$$R_j(x) = \sqrt{\sum_{i=j}^{M} \lambda_i^2(x)\langle v_i(x), x - \Phi(x)\rangle^2} - \eta\rho\lambda_j^2(x)\,,$$

Based on the above notions, we can rephrase the notion R as

$$R_j(x) = \|P^{(j:D)}(x)\tilde{x}\|^2 - \eta\rho\lambda_j^2(x)\,.$$

We additionally define the approximate invariant set $\mathbb{I}_j$ as

$$\mathbb{I}_j = \{R_j(x) \le O(\eta\rho^2)\}\,.$$

Lemma C.36. *Under condition of Theorem C.31, assuming t satisfy that $x(t) \in K^h$, then we have that*

$$\frac{\mu}{2}\|x(t) - \Phi(x(t))\| \le \|\tilde{x}(t)\| \le \zeta\|x(t) - \Phi(x(t))\|$$

Proof. First by Lemma C.10, $\Phi(x(t)) \in K^r$, hence

$$\begin{aligned}\|\tilde{x}(t)\| &= \|\nabla^2 L(\Phi(x(t)))(x(t) - \Phi(x(t)))\|\\ &\le \zeta\|x(t) - \Phi(x(t))\|\,.\end{aligned}$$

Also

$$\begin{aligned}\|\tilde{x}(t)\| &= \|\nabla^2 L(\Phi(x(t)))(x(t) - \Phi(x(t)))\|\\ &\ge \mu\|P^{\perp}_{\Phi(x(t)),\Gamma}(x(t) - \Phi(x(t)))\|\,.\end{aligned}$$

By Lemma C.10 and Lemma C.6, we have

$$\begin{aligned}\|x(t) - \Phi(x(t))\| &\le \|P^{\perp}_{\Phi(x(t)),\Gamma}(x(t) - \Phi(x(t)))\| + \|P_{\Phi(x(t)),\Gamma}(x(t) - \Phi(x(t)))\|\\ &\le \frac{\zeta\nu}{4\mu^2}\|x(t) - \Phi(x(t))\|^2 + \frac{1}{\mu}\|\tilde{x}(t)\|\\ &\le \frac{1}{2}\|x(t) - \Phi(x(t))\| + \frac{1}{\mu}\|\tilde{x}(t)\|.\end{aligned}$$

Hence $\|x(t) - \Phi(x(t))\| \le \frac{2}{\mu}\|\tilde{x}(t)\|$. □

Lemma C.37. *Under condition of Theorem C.31, assuming t satisfy that $x(t) \in K^h$ and $\|\tilde{x}(t)\|_2 = O(\rho)$, then we have that*

$$\|\Phi(x(t+1)) - \Phi(x(t))\| = O(\eta\rho^2)\,.$$

Proof. By Lemma C.36, we have $\|x(t) - \Phi(x(t))\| = O(\rho)$. By Lemma C.12, we have that

$$\begin{aligned}\|\Phi(x(t+1)) - \Phi(x(t))\| &\leq \zeta\xi\eta\rho\|x - \Phi(x)\|_2 + \zeta^2\xi\eta^2\|x - \Phi(x)\|_2^2 + \nu\eta\rho^2 + \xi\zeta^2\eta^2\rho^2 \\ &\leq O(\eta\rho^2)\end{aligned}$$

□

Lemma C.38. *Under condition of Theorem C.31, assuming t satisfy $x(t) \in K^{h/2}$ and $\|x(t) - \Phi(x(t))\|_2 = O(\rho)$, define x' as $x'(t) = x(t)$ and*

$$\begin{aligned}x'(t+1) = x'(t) &- \eta\nabla^2 L(\Phi(x(t)))(x'(t) - \Phi(x)) \\ &-\eta\rho\nabla^2 L(\Phi(x(t)))\frac{\nabla^2 L(\Phi(x(t)))(x'(t) - \Phi(x))}{\|\nabla^2 L(\Phi(x(t)))(x'(t) - \Phi(x))\|_2}.\end{aligned}$$

Then

$$\|x'(t+1) - x(t+1)\|_2 = O(\eta\rho^2)$$

and further if $\|x(t+1) - \Phi(x(t+1))\|_2 = \Omega(\eta\rho)$, then

$$\|x'(t+2) - x(t+2)\|_2 = O(\eta\rho^2).$$

Proof. For simplicity, we will sometime omit the t in $x(t)$.

By $\|x(t) - \Phi(x(t))\| = O(\rho)$, $x(t) \in K^{h/2}$, and Lemma C.12, we have that $\|x(t+1) - x(t)\| = O(\eta\rho)$ and hence $x(t+1) \in K^{3h/4}$. This also implies $\|x(t+1) - \Phi(x(t+1))\|_2 = O(\rho)$. Similarly we have $x(t+2) \in K^{3h/4}$.

For $k \in \{1, 2\}$, by Taylor Expansion,

$$\begin{aligned}x(t+k+1) =& x(t+k) - \eta\nabla L(x(t+k)) - \eta\rho\nabla^2 L(x(t+k))\frac{\nabla L\left(x(t+k)\right)}{\|\nabla L\left(x(t+k)\right)\|} + O(\eta\rho^2) \\ =& x(t+k) - \eta\nabla^2 L(\Phi(x(t+k)))(x(t+k) - \Phi(x(t+k))) + O(\eta\rho^2) \\ &- \eta\rho\nabla^2 L(\Phi(x(t+k)))\frac{\nabla L\left(x(t+k)\right)}{\|\nabla L\left(x(t+k)\right)\|} + O(\eta\rho^2) \\ =& x(t+k) - \eta\nabla^2 L(\Phi(x(t+k)))(x(t+k) - \Phi(x(t+k))) \\ &- \eta\rho\nabla^2 L(\Phi(x(t+k)))\frac{\nabla L\left(x(t+k)\right)}{\|\nabla L\left(x(t+k)\right)\|} + O(\eta\rho^2).\end{aligned}$$

Now by Lemmas C.36 and C.37, $\|\Phi(x(t+k)) - \Phi(x(t))\|_2 = O(\eta\rho^2)$,

$$\begin{aligned}x(t+k+1) =& x(t+k) - \eta\nabla^2 L(\Phi(x))(x(t+k) - \Phi(x(t))) \\ &- \eta\rho\nabla^2 L(\Phi(x))\frac{\nabla L\left(x(t+k)\right)}{\|\nabla L\left(x(t+k)\right)\|} + O(\eta\rho^2).\end{aligned} \tag{29}$$

Now we first prove the first claim, we have for $k = 0$, $\|x(t+k) - x'(t+k)\|_2 = 0$, by Lemma C.10 and eq. (29),

$$\begin{aligned}x(t+1) =& x(t) - \eta\nabla^2 L(\Phi(x))(x(t) - \Phi(x)) - \eta\rho\frac{\nabla^2 L(\Phi(x))(x(t) - \Phi(x))}{\|\nabla^2 L(\Phi(x))(x(t) - \Phi(x))\|_2} + O(\eta\rho^2) \\ =& x'(t+1) + O(\eta\rho^2).\end{aligned}$$

The second claim is slightly more complex. By the first claim and Lemma C.10, we have that

$$\begin{aligned}\frac{\nabla L\left(x(t+1)\right)}{\|\nabla L\left(x(t+1)\right)\|} =& \frac{\nabla^2 L(\Phi(x(t+1)))(x(t+1) - \Phi(x(t+1)))}{\|\nabla^2 L(\Phi(x(t+1)))(x(t+1) - \Phi(x(t+1)))\|_2} \\ &+ O(\|x(t+1) - \Phi(x(t+1))\|_2).\end{aligned} \tag{30}$$

We first show $\|\nabla^2 L(\Phi(x(t+1)))(x(t+1)-\Phi(x(t+1)))\|_2$ is of order $\|x(t+1)-\Phi(x(t+1))\|_2 = \Omega(\rho^2)$ to show that the normalized gradient term is stable with respect to small perturbation,

$$\begin{aligned}
&\|\nabla^2 L(\Phi(x(t+1)))(x(t+1)-\Phi(x(t+1)))\|_2\\
\geq&\|P_{\Phi(x(t+1)),\Gamma}\nabla^2 L(\Phi(x(t+1)))(x(t+1)-\Phi(x(t+1)))\|_2\\
\geq&\|\nabla^2 L(\Phi(x(t+1)))P_{\Phi(x(t+1)),\Gamma}(x(t+1)-\Phi(x(t+1)))\|_2\\
\geq&\mu\|P_{\Phi(x(t+1)),\Gamma}(x(t+1)-\Phi(x(t+1)))\|_2\\
\geq&\mu(\|(x(t+1)-\Phi(x(t+1)))\|_2-\|P^{\perp}_{\Phi(x(t+1)),\Gamma}(x(t+1)-\Phi(x(t+1)))\|_2)\\
\geq&\mu(\|(x(t+1)-\Phi(x(t+1)))\|_2-\frac{\nu\zeta}{4\mu^2}\|x(t+1)-\Phi(x(t+1))\|_2^2)\\
\geq&\frac{\mu}{2}\|(x(t+1)-\Phi(x(t+1)))\|_2=\Omega(\eta\rho).
\end{aligned}$$

Based on Lemma C.12, we have

$$\Phi(x(t+1))-\Phi(x(t))=O(\eta\rho^2).$$

We further have by the first claim and Lemma C.37,

$$\begin{aligned}
&\nabla^2 L(\Phi(x(t+1)))(x(t+1)-\Phi(x(t+1)))-\nabla^2 L(\Phi(x))(x'(t+1)-\Phi(x(t)))\\
=&\nabla^2 L(\Phi(x))(x(t+1)-\Phi(x(t+1)))-\nabla^2 L(\Phi(x))(x'(t+1)-\Phi(x(t))\\
&+O(\|x(t+1)-\Phi(x(t+1))\|_2\|\Phi(x(t+1))-\Phi(x)\|_2)\\
=&\nabla^2 L(\Phi(x))(x(t+1)-\Phi(x(t+1)))-\nabla^2 L(\Phi(x))(x'(t+1)-\Phi(x(t)))+O(\eta\rho^3)\\
=&\nabla^2 L(\Phi(x))(x(t+1)-x'(t+1))+\nabla^2 L(\Phi(x))(\Phi(x(t+1))-\Phi(x(t)))+O(\eta\rho^3)\\
=&O(\eta\rho^2)
\end{aligned}$$

This implies

$$\frac{\nabla^2 L(\Phi(x(t+1)))(x(t+1)-\Phi(x(t+1)))}{\|\nabla^2 L(\Phi(x(t+1)))(x(t+1)-\Phi(x(t+1)))\|_2}=\frac{\nabla^2 L(\Phi(x))(x'(t+1)-\Phi(x))}{\|\nabla^2 L(\Phi(x))(x'(t+1)-\Phi(x))\|_2}+O(\rho)$$

Combining with Equation (30), we have

$$\frac{\nabla L\left(x(t+1)\right)}{\|\nabla L\left(x(t+1)\right)\|}=\frac{\nabla^2 L(\Phi(x))(x'(t+1)-\Phi(x))}{\|\nabla^2 L(\Phi(x))(x'(t+1)-\Phi(x))\|_2}+O(\rho)$$

By the above approximation and Equation (29),

$$x(t+2)=x'(t+2)+O(\eta\rho^2)\,.$$

□

Lemma C.39. *Under condition of Theorem C.31, assuming assuming t satisfy that $x(t)\in K^{3h/4}$ and $\|\tilde{x}(t)\|_2=O(\rho)$, then we have that*

$$\|\tilde{x}(t+1)-\tilde{x}(t)+\eta A(t)\tilde{x}(t)+\eta\rho A^2(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\|_2=O(\eta\rho^2)\,.$$

Proof. By Lemma C.38, we know

$$\|x(t+1)-x(t)+\eta\tilde{x}(t)+\eta\rho A(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\|\leq O(\eta\rho^2)\,.$$

This implies

$$\begin{aligned}
&\|A(t)(x(t+1)-\Phi(x(t)))-\tilde{x}(t)+\eta A(t)\tilde{x}(t)+\eta\rho A^2(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\|\\
=&\|A(t)(x(t+1)-x(t)+\eta\tilde{x}(t)+\eta\rho A(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|})\|\\
\leq&\zeta\|x(t+1)-x(t)+\eta\tilde{x}(t)+\eta\rho A(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\|=O(\eta\rho^2)\,.
\end{aligned}\tag{31}$$

We also have

$$\begin{aligned}
&\tilde{x}(t+1) - A(t)(x(t+1) - \Phi(x(t))) \\
=&(A(t+1) - A(t))(x(t+1) - \Phi(x(t+1))) - A(t)(\Phi(x(t)) - \Phi(x(t+1))) \\
=&O(\eta\rho^2)\,.
\end{aligned}$$

Plugging in Equation (31), we have that

$$\|\tilde{x}(t+1) - \tilde{x}(t) + \eta A(t)\tilde{x}(t) + \eta\rho A^2(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\|_2 = O(\eta\rho^2)\,.$$

□

Lemma C.40. *Under condition of Theorem C.31, assuming there exists t_{DEC} such that $x(t_{\mathrm{DEC}}) \in K^{h/2}$ and $\|\nabla L(x(t_{\mathrm{DEC}}))\| \leq 4\zeta\rho$, then there exists $t_{\mathrm{DEC2}} = t_{\mathrm{DEC}} + O(\ln(1/\eta)/\eta)$, such that $x(t_{\mathrm{DEC2}})$ is in $\mathbb{I}_1 \cap K^{3h/4}$.*

Furthermore, for any t satisfying $t_{\mathrm{DEC2}} \leq t \leq t_{\mathrm{DEC2}} + \Theta(\ln(1/\eta)/\eta)$, we have that $x(t) \in \mathbb{I}_1 \cap K^{3h/4}$ and $\|\Phi(x(t)) - \Phi(x(t_{\mathrm{DEC}}))\| = O(\rho^2 \ln(1/\eta))$.

Proof. For simplicity, denote $C = \lceil \ln_{1-\eta\mu} \frac{\eta\mu^3}{4\zeta^2} \rceil + \Theta(\ln(1/\rho)/\eta) = O(\ln(1/\eta)/\eta)$. Here the quantity $\Theta(\ln(1/\rho)/\eta)$ is the same quantity in the statement of the lemma.

We will prove the induction hypothesis for $t_{\mathrm{DEC}} \leq t \leq t_{\mathrm{DEC}} + 2C$,

$$\begin{cases}
\|\tilde{x}(t-1)\| \geq \eta\rho\lambda_1^2(t), t > t_{\mathrm{DEC}} \Rightarrow \|\tilde{x}(t)\| \leq (1-\eta\mu)\|\tilde{x}(t-1)\|, \\
\|\tilde{x}(t-1)\| \leq \eta\rho\lambda_1^2(t-1), t > t_{\mathrm{DEC}} \Rightarrow \|\tilde{x}(t)\| \leq \eta\rho\lambda_1^2(t) + O(\eta\rho^2), \\
\|\Phi(x(t)) - \Phi(x(t_{\mathrm{DEC}}))\| \leq O(\eta\rho^2(t - t_{\mathrm{DEC}})), \\
x(t) \in K^{3h/4}.
\end{cases}$$

The induction hypothesis holds trivially for $t = t_{\mathrm{DEC}}$.

Assume the induction hypothesis holds for $t' \leq t$. By Lemmas C.7 and C.36, $\|\tilde{x}(t_{\mathrm{DEC}})\|_2 \leq \zeta\|x(t_{\mathrm{DEC}}) - \Phi(x(t_{\mathrm{DEC}}))\| \leq \frac{\zeta}{\mu}\|\nabla L(x(t_{\mathrm{DEC}}))\| \leq \frac{4\zeta^2}{\mu}\rho$. Combining with the induction hypothesis, we have $\|\tilde{x}(t)\| \leq \frac{4\zeta^2}{\mu}\rho$.

By $x(t) \in K^{3h/4}$ and Lemma C.37, we have that

$$\|\Phi(x(t+1)) - \Phi(x(t))\| \leq O(\eta\rho^2)\,.$$

Hence we have that

$$\begin{aligned}
\|\Phi(x(t+1)) - \Phi(x(t_{\mathrm{DEC}}))\| &\leq \|\Phi(x(t+1)) - \Phi(x(t))\| + \|\Phi(x(t)) - \Phi(x(t_{\mathrm{DEC}}))\| \\
&\leq O(\eta\rho^2(t+1-t_{\mathrm{DEC}})). \qquad (32)
\end{aligned}$$

This proves the third statement of the induction hypothesis.

By $\|\tilde{x}(t)\| = O(\rho)$ and Lemma C.39, we have that

$$\|\tilde{x}(t+1) - \tilde{x}(t) + \eta A(t)\tilde{x}(t) + \eta\rho A^2(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\|_2 = O(\eta\rho^2)\,.$$

Analogous to the proof of Lemmas C.21 and C.22, we have

1. If $\|\tilde{x}(t)\| > \eta\rho\lambda_1^2(t)$, we would have

$$\begin{aligned}
&\|\tilde{x}(t) - \eta A(t)\tilde{x}(t) - \eta\rho A^2(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\| \\
\leq &\|\tilde{x}(t)\|\|I - \eta A(t) - \eta\rho A^2(t)\frac{1}{\|\tilde{x}(t)\|}\| \\
\leq &\|\tilde{x}(t)\| \max\{\eta\lambda_1, 1 - \eta\lambda_D - \eta\rho\lambda_D^2\frac{1}{\|x(t)\|}\} \\
\leq &\max\{(1-\eta\lambda_D)\|\tilde{x}(t)\| - \eta\rho\lambda_D^2, \eta\lambda_1\|\tilde{x}(t)\|\} \\
\leq &\max\{(1-\eta\mu)\|\tilde{x}(t)\| - \eta\rho\mu^2, \eta\zeta\|\tilde{x}(t)\|\}
\end{aligned}$$

   Hence we have

$$\|\tilde{x}(t+1)\| \leq \max\{(1-\eta\mu)\|\tilde{x}(t)\| - \eta\rho\mu^2, \eta\zeta\|\tilde{x}(t)\|\} + O(\eta\rho^2) \leq (1-\eta\mu)\|\tilde{x}(t)\|.$$

2. If $\|\tilde{x}(t)\|_2 \leq \eta\rho\lambda_1^2(t)$, then by Lemma C.21, we have that

$$\|\tilde{x}(t) - \eta A(t)\tilde{x}(t) - \eta\rho A^2(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\|_2 \leq \eta\rho\lambda_1^2(t)\,.$$

   Hence by Lemma C.65

$$\begin{aligned}
\|\tilde{x}(t+1)\| &\leq \eta\rho\lambda_1^2(t) + O(\eta\rho^2) \\
&\leq \eta\rho\lambda_1^2(t+1) + O(\eta\rho^2)\,.
\end{aligned}$$

Concluding the two cases, we have shown the first and second claim of the induction hypothesis holds. Hence we can show that $\|\tilde{x}(t+1)\| \leq \frac{4\zeta^2}{\mu}\rho$. Then by Lemma C.36, we have that $\|x(t+1) - \Phi(x(t+1))\| \leq \frac{8\zeta^2}{\mu^2}\rho$.

As $t \leq t_{\mathrm{DEC}} + 2C = t_{\mathrm{DEC}} + O(\ln(1/\eta)/\eta)$, by Equation (32),

$$\|\Phi(x(t+1)) - \Phi(x(t_{\mathrm{DEC}}))\| \leq O(-\rho^2\ln\eta)\,.$$

This implies

$$\begin{aligned}
\mathrm{dist}(x(t+1), K) \leq &\mathrm{dist}(x(t_{\mathrm{DEC}}), K) + \|x(t_{\mathrm{DEC}}) - \Phi(x(t_{\mathrm{DEC}}))\| \\
&+ \|\Phi(x(t_{\mathrm{DEC}})) - \Phi(x(t+1))\| + \|x(t+1) - \Phi(x(t+1))\| \\
= &h/2 + O(\rho^2\ln(1/\eta)) + O(\rho) \leq 3h/4.
\end{aligned}$$

This proves the fourth claim of the inductive hypothesis.

The induction is complete.

Now define t_{DEC2} the minimal $t \geq t_{\mathrm{DEC}}$, such that $\|\tilde{x}(t)\| \leq \eta\rho\lambda_1^2(t)$.

If $t_{\mathrm{DEC2}} > t_{\mathrm{DEC}} + C$, then by the induction, Lemmas C.7 and C.36,

$$\begin{aligned}
\|\tilde{x}(t_{\mathrm{DEC}} + C)\| &\leq (1-\eta\mu)^C\|\tilde{x}(t_{\mathrm{DEC}})\| \\
&\leq \frac{\eta\mu^3}{4\zeta^2}\|\tilde{x}(t_{\mathrm{DEC}})\| \\
&\leq \frac{\eta\mu^3}{4\zeta^2}\zeta\|x(t_{\mathrm{DEC}}) - \Phi(x(t_{\mathrm{DEC}})\| \\
&\leq \frac{\eta\mu^2}{4\zeta}\|\nabla L(t_{\mathrm{DEC}})\| \\
&\leq \mu^2\eta\rho \\
&\leq \lambda_1^2(t_{\mathrm{DEC}} + C)\eta\rho\,.
\end{aligned}$$

This is a contradiction. Hence we have $t_{\mathrm{DEC2}} \leq t_{\mathrm{DEC}} + C$. By the induction hypothesis $x(t_{\mathrm{DEC2}}) \in \mathbb{I}_1 \cap K^{3h/4}$.

Furthermore by induction, for any t satisfying $t_{\mathrm{DEC2}} \le t \le t_{\mathrm{DEC}} + 2C$, we have that

$$\|\tilde{x}(t)\| \le \eta\rho\lambda_1^2(t) + O(\eta\rho^2)\,.$$

By the induction hypothesis $x(t) \in \mathbb{I}_1 \cap K^{3h/4}$ and $\|\Phi(x(t)) - \Phi(x(t_{\mathrm{DEC}}))\| = O(\rho^2 \ln(1/\rho))$. □

Lemma C.41. *Under condition of Theorem C.31, assuming assuming t satisfy that $x(t) \in \mathbb{I}_1 \cap K^{3h/4}$, then we have that*

$$\begin{cases} R_k(x(t)) \ge 0 \Rightarrow R_k(x(t+1)) + \lambda_k^2(t+1)\eta\rho \le (1-\eta\mu)(R_k(x(t)) + \lambda_k^2(t)\eta\rho), \\ R_k(x(t)) \le 0 \Rightarrow R_k(x(t+1)) \le O(\eta\rho^2). \end{cases}$$

Proof. As $x(t) \in \mathbb{I}_1$, $\|\tilde{x}(t)\|_2 \le \zeta\eta\rho + O(\eta\rho^2)$.

As $\|\tilde{x}(t)\|_2 = O(\rho)$, we have $\overline{x(t)x(t+1)} \subset K^h$ and $\overline{\Phi(x(t))\Phi(x(t+1))} \subset K^r$.

We will begin with a quantization technique separating $[M]$ into disjoint continuous subset $S_1, ..., S_p$ such that $\forall i \ne j$,

$$\min_{k\in S_i, l\in S_j} |\lambda_k(t) - \lambda_l(t)| \ge \rho\,.$$

By Lemmas C.37 and C.65, we have that for any $n \in [M]$,

$$\begin{aligned} |\lambda_k(t) - \lambda_k(t+1)| &= O(\|\nabla^2 L(\Phi(x(t))) - \nabla^2 L(\Phi(x(t+1)))\|) \\ &= O(\|\Phi(x(t)) - \Phi(x(t+1))\|) \\ &= O(\eta\rho^2). \end{aligned}$$

This implies

$$\min_{k\in S_i, l\in S_j} |\lambda_k(t+1) - \lambda_l(t+1)| \ge \rho - O(\eta\rho^2) \ge 0.99\rho\,.$$

Define

$$P_{S^{(i)}}^{(t)} \triangleq \sum_{k\in S_i} v_n(t)v_n(t)^T\,.$$

By Theorem C.67, for any k,

$$\|P_{S_k}^{(t)} - P_{S_k}^{(t+1)}\| \le O(\frac{\|\nabla^2 L(\Phi(x(t))) - \nabla^2 L(\Phi(x(t+1)))\|}{\rho}) = O(\eta\rho)\,.$$

By Lemma C.39, we have that

$$\|\tilde{x}(t+1) - \tilde{x}(t) + \eta A(t)\tilde{x}(t) + \eta\rho A^2(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\|_2 = O(\eta\rho^2)\,.$$

We will write $x'(t+1)$ as shorthand of $\tilde{x}(t) - \eta A(t)\tilde{x}(t) - \eta\rho A^2(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}$.

Now we discuss by cases,

1. If $\sqrt{\sum_{i=j}^p \|P_{S^{(i)}}^{(t)}\tilde{x}(t)\|^2} > \max_{k\in S_j} \lambda_k^2(t)\eta\rho > \mu^2\eta\rho$, by Lemma C.23,

$$\begin{aligned} &\sqrt{\sum_{i=j}^p \|P_{S^{(i)}}^{(t)}\tilde{x}(t+1)\|^2} \le \sqrt{\sum_{i=j}^p \|P_{S^{(i)}}^{(t)}x'(t+1)\|^2} + O(\eta\rho^2) \\ &\le \max\{\big(1-\eta\lambda_D(t+1)\big)\|\sum_{i=j}^p P_{S^{(i)}}^{(t)}\tilde{x}(t)\| - \eta\rho\lambda_D(t+1)^2\frac{\|\sum_{i=j}^p P_{S^{(i)}}^{(t)}\tilde{x}(t)\|}{\|\tilde{x}(t)\|}, \\ &\quad \eta\max_{k\in S_j}\lambda_k(t+1)\|\sum_{i=j}^p P_{S^{(i)}}^{(t)}\tilde{x}(t)\|\} + O(\eta\rho^2) \\ &\le \max\{\big(1-\eta\mu\big)\|\sum_{i=j}^p P_{S^{(i)}}^{(t)}\tilde{x}(t)\| - \eta\rho\frac{\mu^3}{2\zeta}, \eta\zeta\|\sum_{i=j}^p P_{S^{(i)}}^{(t)}\tilde{x}(t)\|\} + O(\eta\rho^2)\,. \end{aligned}$$

This further implies

$$\sqrt{\sum_{i=j}^{p}\|P_{S^{(i)}}^{(t+1)}\tilde{x}(t+1)\|^2} \leq \sqrt{\sum_{i=j}^{p}\|P_{S^{(i)}}^{(t)}\tilde{x}(t+1)\|^2} + O(\eta\rho\|\tilde{x}(t+1)\|)$$
$$\leq \max\{(1-\eta\mu)\|\sum_{i=j}^{p}P_{S^{(i)}}^{(t)}\tilde{x}(t)\| - \eta\rho\frac{\mu^3}{2\zeta}, \eta\zeta\|\sum_{i=j}^{p}P_{S^{(i)}}^{(t)}\tilde{x}(t)\|\} + O(\eta\rho^2)$$
$$\leq (1-\eta\mu)\|\sum_{i=j}^{p}P_{S^{(i)}}^{(t)}\tilde{x}(t)\|\,.$$

2. If $\sqrt{\sum_{i=j}^{p}\|P_{S^{(i)}}^{(t)}\tilde{x}(t)\|^2} \leq \max_{k\in S_j}\lambda_k^2(t)\eta\rho$, then by Lemma C.21, we have that

$$\|\sum_{i=j}^{p}P_{S^{(i)}}^{(t)}x'(t+1)\|_2 \leq \eta\rho\max_{k\in S_j}\lambda_k^2(t)\,.$$

Hence we have that

$$\sqrt{\sum_{i=j}^{p}\|P_{S^{(i)}}^{(t)}\tilde{x}(t+1)\|^2} \leq \sqrt{\sum_{i=j}^{p}\|P_{S^{(i)}}^{(t)}x'(t+1)\|^2} + O(\eta\rho^2)$$
$$\leq \max_{k\in S_j}\lambda_k^2(t)\eta\rho + O(\eta\rho^2)$$
$$\leq \max_{k\in S_j}\lambda_k^2(t+1)\eta\rho + O(\eta\rho^2)\,.$$

This further implies

$$\sqrt{\sum_{i=j}^{p}\|P_{S^{(i)}}^{(t+1)}\tilde{x}(t+1)\|^2} \leq \sqrt{\sum_{i=j}^{p}\|P_{S^{(i)}}^{(t)}\tilde{x}(t+1)\|^2} + O(\eta\rho\|\tilde{x}(t+1)\|)$$
$$\leq \max_{k\in S_j}\lambda_k^2(t)\eta\rho + O(\eta\rho^2)$$
$$\leq \max_{k\in S_j}\lambda_k^2(t+1)\eta\rho + O(\eta\rho^2)\,.$$

Finally taking into quantization error, as all the eigenvalue in the same group at most differ $D\rho$, for any $i \in S_j$, we have that $-\lambda_i^2(t+1) + \max_{k\in S_j}\lambda_k^2(t+1) \leq 2D\zeta\rho + D^2\rho^2$.

Hence the previous discussion concludes as

1. If $R_k(x(t)) \geq 0$

$$R_k(x(t+1)) + \lambda_k^2(t+1)\eta\rho \leq (1-\eta\mu)(R_k(x(t)) + \lambda_k^2(t).\eta\rho)$$

2. If $R_k(x(t)) < 0$

$$R_k(x(t+1)) \leq O(\eta\rho^2).$$

□

Lemma C.42. *Under condition of Theorem C.31, assuming there exists t_{DEC2} such that for any t satisfying $t_{\mathrm{DEC2}} \leq t \leq t_{\mathrm{DEC2}} + \Theta(\ln(1/\eta)/\eta)$, we have that $x(t) \in \mathbb{I}_1 \cap K^{3h/4}$. Then there exists $t_{\mathrm{INV}} = t_{\mathrm{DEC2}} + O(\ln(1/\eta)/\eta))$ such that for any t satisfying $t_{\mathrm{INV}} \leq t \leq t_{\mathrm{INV}} + \Theta(\ln(1/\eta)/\eta)$, we have that*

$$x(t) \in \cap_{k\in[M]}\mathbb{I}_k \cap K^{7h/8}\,.$$
$$\|\Phi(x(t)) - \Phi(x(t_{\mathrm{DEC2}}))\| = O(\rho^2\ln(1/\eta))\,.$$

Proof. The proof is almost identical with Lemma C.40 replacing the first two iterative hypothesis to Lemma C.41 and is omitted here. □

C.5.2 PHASE II (PROOF OF THEOREM C.32)

The goal of this section is to prove Theorem C.32.

Proof. First by Lemma C.47, there exists t_{ALIGN} such that $\nabla L(x(t_{\text{ALIGN}}))$ aligns with $v_1(t_{\text{ALIGN}})$ with $O(\rho)$ angle.

Then we can inductively use Lemma C.48 and theorem D.5 with $p = \eta\rho^2$ and $q = \eta + \rho$ to show that $x(t)$ tracks the Riemmannian gradient flow with approximation error $O(\eta \ln(1/\rho))$ and hence maintains in K^h and Lemma C.47 to show that $x(t)$ keeps alignment with $v_1(t)$ for $T_3/\eta\rho^2$ step. □

Alignment to Top Eigenvector We will continue to use the notations introduced in Appendix C.5.1.

We further define

$$S = \{t | \|\tilde{x}(t)\| \leq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho + O(\eta\rho^2)\},$$
$$T = \{t | \|\tilde{x}(t)\| \leq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho\},$$
$$U = \{t | \Omega(\rho^2) \leq \|\tilde{x}_1(t)\| \leq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho\}.$$

Here the constant in O depends on the constant in $\mathbb{I}_j$ and will be made clear in Lemma C.44.

For $s \in S$, define $\text{next}(t)$ as the smallest integer greater than s in S.

Lemma C.43. *Under the condition of Theorem C.32, there exist constants $C_1, C_2 < 1$ independent of η and ρ, if $\|\tilde{x}_1(t)\|_2 \leq \frac{1}{2}\left(\frac{\eta\lambda_1(t)^2}{2-\eta\lambda_1(t)} + \frac{\eta\lambda_2(t)^2}{2-\eta\lambda_2(t)}\right)\rho$ and $x(t) \in \cap_j \mathbb{I}_j \cap K^{7h/8}$, then*

$$\|\tilde{x}(t)\|_2 \geq C_1\frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho \Rightarrow \|\tilde{x}(t+1)\|_2 \leq C_2\frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho$$

Proof. By Lemma C.39, if we write $x'(t+1)$ as shorthand of $\tilde{x}(t) - \eta A(t)\tilde{x}(t) - \eta\rho A^2(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}$, then $\|\tilde{x}(t+1) - x'(t+1)\| = O(\eta\rho^2)$.

Define $\mathbb{I}_j^{\text{quad}}$ as $\{x | R_j(x) \leq 0\}$. Then we can find a surrogate $x_{\text{sur}}(t)$ such that $x_{\text{sur}}(t) \in \cap_j \mathbb{I}_j^{\text{quad}} \cap K^h$ and $\|x_{\text{sur}}(t) - \tilde{x}(t)\|_2 = O(\eta\rho^2)$. We will write $x'_{\text{sur}}(t+1)$ as shorthand of $x_{\text{sur}}(t) - \eta A(t)x_{\text{sur}}(t) - \eta\rho A^2(t)\frac{x_{\text{sur}}(t)}{\|x_{\text{sur}}(t)\|}$.

Let

$$h(t) \triangleq (2-t)\sqrt{\frac{1}{2t}\left(\frac{(\zeta-\Delta)^2}{\zeta^2}+1\right) + (1-\frac{1}{2t}\left(\frac{(\zeta-\Delta)^2}{\zeta^2}+1\right))\max\{\frac{\zeta^2-\mu^2}{\zeta^2}, \frac{(\zeta-\Delta)^2}{\zeta^2}\}}.$$

As $h(1) < 1$, we can choose $C_1 < 1$, such that $h(C_1) < 1$.

We can further choose $C_2 = \max\{(h(C_1)+1)/2, 1 - \frac{\mu^2}{3\zeta^2}\} < 1$.

We will discuss by cases

1. If

$$\|\tilde{x}(t)\|_2 \geq \frac{\eta\lambda_1^4}{\lambda_1^2(1-\eta\lambda_D) + (\lambda_1^2 - \lambda_D^2)(1-\eta\lambda_1)}\rho$$

Then

$$\frac{\|\tilde{x}(t)\|_2}{\frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho} = \frac{\lambda_1^2(2-\eta\lambda_1)}{\lambda_1^2(1-\eta\lambda_D)+(\lambda_1^2-\lambda_D^2)(1-\eta\lambda_1)}$$
$$= \frac{\lambda_1^2(2-\eta\lambda_1)}{\lambda_1^2(2-\eta\lambda_1-\eta\lambda_D)-\lambda_D^2(1-\eta\lambda_1)}$$
$$\geq \frac{1}{1-\frac{\lambda_D^2}{\lambda_1^2}\frac{1-\eta\lambda_1}{2-\eta\lambda_1}} \geq 1+\frac{\lambda_D^2}{\lambda_1^2}\frac{1-\eta\lambda_1}{2-\eta\lambda_1} \geq 1+\frac{\mu^2}{3\zeta^2}\,.$$

In such case we have

$$\|\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}-\frac{x_{\text{sur}}(t)}{\|x_{\text{sur}}(t)\|}\| = O(\rho)\,.$$

Then we have $\|x'_{\text{sur}}(t+1)-x'(t+1)\| = O(\eta\rho^2)$. By Lemma C.24, we have that

$$\|\tilde{x}(t+1)\|_2 \leq \|\tilde{x}(t+1)-x'(t+1)\|_2+\|\tilde{x}(t+1)-x'_{\text{sur}}(t+1)\|+\|x'_{\text{sur}}(t+1)\|$$
$$\leq \max(\frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho-\eta\rho\frac{\lambda_D^4}{2\lambda_1^2},\eta\rho\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2)+O(\eta\rho^2)$$
$$\leq \max(1-\frac{\lambda_D^4(2-\eta\lambda_1)}{2\lambda_1^4},(2-\eta\lambda_1)-(1-\eta\lambda_1)(1+\frac{\mu^2}{3\zeta^2}))\frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho$$
$$\leq (1-\frac{\mu^2}{3\zeta^2})\frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho \leq C_2\frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho$$

2. If

$$\|\tilde{x}(t)\|_2 \leq \frac{\eta\lambda_1^4}{\lambda_1^2(1-\eta\lambda_D)+(\lambda_1^2-\lambda_D^2)(1-\eta\lambda_1)}\rho \leq \frac{\eta\lambda_1^2}{1-\eta\lambda_1}\rho.$$

Then we have

$$\frac{|-\eta\rho\lambda_D^2+(1-\eta\lambda_D)\|\tilde{x}(t)\|_2|}{\eta\rho\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2} \leq \frac{\lambda_1^2-\lambda_D^2}{\lambda_1^2}.$$
$$\frac{|\eta\rho\lambda_2^2-(1-\eta\lambda_2)\|\tilde{x}(t)\|_2|}{\eta\rho\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2} \leq \frac{\lambda_2^2}{\lambda_1^2}.$$

By Lemma C.27,

$$\|x'(t+1)\|_2$$
$$\leq(\eta\rho\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2)\sqrt{\frac{\|\tilde{x}_1^2(t)\|_2}{\|\tilde{x}(t)\|_2^2}+(1-\frac{\|\tilde{x}_1^2(t)\|_2}{\|\tilde{x}(t)\|_2^2})\max\{\frac{\lambda_1^2-\lambda_D^2}{\lambda_1^2},\frac{\lambda_2^2}{\lambda_1^2}\}}$$
$$\leq(\eta\rho\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|_2)\sqrt{\frac{\|\tilde{x}_1^2(t)\|_2}{\|\tilde{x}(t)\|_2^2}+(1-\frac{\|\tilde{x}_1^2(t)\|_2}{\|\tilde{x}(t)\|_2^2})\max\{\frac{\zeta^2-\mu^2}{\zeta^2},\frac{(\zeta-\Delta)^2}{\zeta^2}\}}\,.$$

.

As

$$\|\tilde{x}_1(t)\|_2 \leq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1}+\frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho$$

For $\|\tilde{x}(t)\|_2 \geq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho C_1$,

$$\frac{\|\tilde{x}_1(t)\|_2}{\|\tilde{x}(t)\|_2} \leq \frac{1}{2}\left(\frac{\lambda_2^2(2-\eta\lambda_1)}{\lambda_1^2(2-\eta\lambda_2)}+1\right)/C_1 \leq \frac{1}{2}\left(\frac{\lambda_2^2}{\lambda_1^2}+1\right)/C_1 \leq \frac{1}{2C_1}\left(\frac{(\zeta-\Delta)^2}{\zeta^2}+1\right).$$

After plugging in, we have that

$$\|\tilde{x}(t+1)\|_2 \leq \|x'(t+1)\|_2+O(\eta\rho^2) \leq h(C_1)\frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho+O(\eta\rho^2) \leq C_2\frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho.$$

This concludes the proof. □

Lemma C.44. *Under the condition of Theorem C.32, assuming that (1) $x(t) \in \cap_j \mathbb{I}_j$, (2) $t \in S \cap K^{15h/16}$ then $\mathrm{next}(t)$ is always well defined and $\mathrm{next}(t) \leq t+2$.*

Proof. If $t+1 \notin S$, then

$$\|\tilde{x}(t+1)\| \geq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho + \Theta(\eta\rho^2).$$

Following similar argument in Lemma C.21, we have that $x(t+1) \in \cap_j \mathbb{I}_j$.

By Lemma C.39, if we write $x'(t+2)$ as shorthand of $\tilde{x}(t+1) - \eta A(t+1)\tilde{x}(t+1) - \eta\rho A^2(t+1)\frac{\tilde{x}(t+1)}{\|\tilde{x}(t+1)\|}$, then $\|\tilde{x}(t+2) - x'(t+2)\| = O(\eta\rho^2)$.

Define $\mathbb{I}_j^{\mathrm{quad}}$ as $\{x | R_j(x) \leq 0\}$. Then we can find a surrogate $x_{\mathrm{sur}}(t+1)$ such that $x_{\mathrm{sur}}(t+1) \in \cap_j \mathbb{I}_j^{\mathrm{quad}} \cap K^h$, and $\|x_{\mathrm{sur}}(t+1) - \tilde{x}(t+1)\|_2 = O(\eta\rho^2)$. We will write $x'_{\mathrm{sur}}(t+2)$ as shorthand of $x_{\mathrm{sur}}(t+1) - \eta A(t+1)x_{\mathrm{sur}}(t+1) - \eta\rho A^2(t+1)\frac{x_{\mathrm{sur}}(t+1)}{\|x_{\mathrm{sur}}(t+1)\|}$.

As $\|\tilde{x}(t+1)\| = \Omega(\eta\rho)$, we have

$$\|\frac{x_{\mathrm{sur}}(t+1)}{\|x_{\mathrm{sur}}(t+1)\|} - \frac{\tilde{x}(t+1)}{\|\tilde{x}(t+1)\|}\|_2 = O(\rho)\,.$$

Hence we have that $\|\tilde{x}(t+2) - x'_{\mathrm{sur}}(t+2)\| = \|\tilde{x}(t+2) - x'(t+2)\| + \|x'(t+2) - x'_{\mathrm{sur}}(t+2)\| = O(\eta\rho^2)$

Notice we have $\|x_{\mathrm{sur}}(t)\|_2 \geq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho$ for properly chosen function in the definition S, hence, by Lemma C.24

$$\|x'_{\mathrm{sur}}(t+1)\|_2 \leq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho.$$

This further implies $t+2 \in S$. □

Lemma C.45. *Under the condition of Theorem C.32, there exists constant $C > 0$ independent of η and ρ, assuming that (1) $x(t) \in \cap_j \mathbb{I}_j$, (2) $t \in S \cap K^{7h/8}$ (3) $\Omega(\rho^2) \leq \|\tilde{x}_1(t)\|$, then*

$$\|\tilde{x}_1(\mathrm{next}(t))\| \geq \|\tilde{x}_1(t)\| - O(\eta\rho^2)\,.$$

Proof. This is by standard approximation as in previous proof and Lemma C.28. □

Lemma C.46. *Under the condition of Theorem C.32, there exists constant $C > 0$ independent of η and ρ, assuming that (1) $x(t) \in \cap_j \mathbb{I}_j$, (2) $t \in S \cap K^{7h/8}$ (3) $\Omega(\rho^2) \leq \|\tilde{x}_1(t)\| \leq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho$, then*

$$\|\tilde{x}_1(\mathrm{next}(t))\| \geq \min\{(1+C\eta)\|\tilde{x}_1(t)\|, \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho\}\,,$$

$$\textit{or } \|\tilde{x}_1(\mathrm{next}(\mathrm{next}(t)))\| \geq \min\{(1+C\eta)\|\tilde{x}_1(t)\|, \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho\}.$$

Proof. In this proof, we will sometime drop the t in $\lambda_k(t)$ or $A(t)$. Applying Lemma C.44, we have $\mathrm{next}(t)$ and $\mathrm{next}(\mathrm{next}(t))$ are well-defined. We can suppose $\|\tilde{x}_1(\mathrm{next}(t))\|_2 \leq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)$, else the result holds already.

By assumption, we have $\|\tilde{x}_1(t)\| \geq \Omega(\rho^2)$.

Using Lemma C.39,

$$\|\tilde{x}(t+1) - \tilde{x}(t) + \eta A\tilde{x}(t) + \eta\rho A^2\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\| \leq O(\eta\rho^2)\,.$$

Denote

$$x'(t+1) = \tilde{x}(t) + \eta A\tilde{x}(t) + \eta\rho A^2 \frac{\tilde{x}(t)}{\|\tilde{x}(t)\|},$$

as the one step update of SAM on the quadratic approximation of the general loss.

Now using Lemma C.43 and the induction hypothesis, we have for some C_1 and C_2 smaller than 1, $\|\tilde{x}(t)\| \geq C_1 \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho \Rightarrow \|x'(t+1)\| \leq C_2 \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho$.

We will discuss by cases,

1 If $\|\tilde{x}(t)\| \leq C_1 \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho$
If $\text{next}(t) = t+1$, then

$$\frac{\|x_1'(t+1)\|}{\|\tilde{x}_1(t)\|} = \frac{\eta\rho\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|}{\|\tilde{x}(t)\|} \geq \frac{(2-C_1) - \eta\lambda_1 + C_1\eta\lambda_1}{C_1} \geq \frac{1}{C_1}$$

As we have $\tilde{x}_1(t) = \Omega(\rho^2)$, we have $x_1'(t+1) = \Omega(\rho^2)$, then as $\|\tilde{x}_1(t+1) - x_1'(t+1)\| = O(\eta\rho^2)$, this implies

$$\|\tilde{x}_1(t+1)\| \geq \|x_1'(t+1)\| - O(\eta\rho^2) \geq \frac{1}{C_1}\|\tilde{x}_1(t+1)\| - O(\eta\rho^2) \geq \frac{1}{2}(\frac{1}{C_1}+1)\|\tilde{x}_1(t)\|.$$

If $\text{next}(t) = t+2$, define $x'(t+2) = x'(t+1) - \eta A x'(t+1) - \eta\rho A^2 \frac{x'(t+1)}{\|x'(t+1)\|}$, as $\|\tilde{x}_1(t+1)\| = \Omega(\eta\rho)$, by Lemma C.38, we have $\|x'(t+2) - \tilde{x}(t+2)\| = O(\eta\rho^2)$.

$$\begin{aligned}
\frac{\|x'(t+2)\|}{\|\tilde{x}_1(t)\|} &= \frac{(\eta\rho\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|)(\eta\rho\lambda_1^2 - (1-\eta\lambda_1)\|x'(t+1)\|)}{\|\tilde{x}(t)\|\|x'(t+1)\|} \\
&\geq \frac{(\eta\rho\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|)\left(\eta\rho\lambda_1^2 - (1-\eta\lambda_1)\left(\eta\rho\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|\right)\right)}{\|\tilde{x}(t)\|\left(\eta\rho\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|\right)} \\
&= \frac{\eta\rho\lambda_1^2 - (1-\eta\lambda_1)\left(\eta\rho\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|\right)}{\|\tilde{x}(t)\|} \\
&\geq (1-\eta\lambda_1)^2 + \frac{\eta\lambda_1}{C_1}(2-\eta\lambda_1) \geq 1 + 4C\eta.
\end{aligned}$$

Combining with $|\tilde{x}_1(t)| \geq \Omega(\rho^2)$, we have that

$$\|\tilde{x}_1(\text{next}(t))\| \geq (1+C\eta)\|\tilde{x}_1(t)\|$$

2 *Case 2* $\|\tilde{x}(t)\| > C_1 \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho$, then $\|\tilde{x}(t+1)\| \leq C_2 \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho$, $\text{next}(t) = t+1$
By Lemma C.45, $\|\tilde{x}_1(t+1)\| \geq (1-C\eta)\|\tilde{x}_1(t)\|$.
As $\|\tilde{x}(\text{next}(t))\| \leq C_2 \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, similar to the first case,

$$\|\tilde{x}_1(\text{next}(\text{next}(t)))\| \geq (1+4C\eta)\|\tilde{x}_1(\text{next}(t))\| \geq (1+C\eta)\|\tilde{x}_1(t)\|.$$

In conclusion, if $\|\tilde{x}_1(t)\| \leq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho$, we would have there exists $C > 0$

$$\|\tilde{x}_1(\text{next}(t))\| \geq (1+C\eta)\|\tilde{x}_1(t)\| \text{ or } \|\tilde{x}_1(\text{next}(\text{next}(t)))\| \geq (1+C\eta)\|\tilde{x}_1(t)\|.$$

□

Lemma C.47. *Under the condition of Theorem C.32, there exists constant $T_2 > 0$ independent of η and ρ, we would have that when $t = t_{\text{ALIGN}} = \lceil T_2 \ln(1/\rho)/\eta \rceil$,*

$$\begin{aligned}
|\langle x(t) - \Phi(x(t)), v_1(x(t))\rangle| &= \Theta(\eta\rho)\,, \\
\max_{j\in[2:M]} |\langle x(t) - \Phi(x(t)), v_j(x(t))\rangle| &= O(\eta\rho^2)\,.
\end{aligned}$$

Further if $x(t') \in K^h$ holds for $t' = 0, 1, ..., t_{\text{LOCAL}}$, then for t satisfying $t_{\text{ALIGN}} \leq t \leq t_{\text{LOCAL}}$

$$\begin{aligned}
|\langle x(t) - \Phi(x(t)), v_1(x(t))\rangle| &= \Theta(\eta\rho)\,, \\
\max_{j\in[2:M]} |\langle x(t) - \Phi(x(t)), v_j(x(t))\rangle| &= O(\eta\rho^2)\,.
\end{aligned}$$

Proof. Let C be the constant defined in Lemma C.46.

Suppose WLOG $\|\tilde{x}_1(0)\| \geq \rho^2$. Define

$$C_1 \triangleq \lceil \log_{1+C\eta}(\frac{\eta\lambda_1^2}{2-\eta\lambda_1}/\rho)\rceil$$

$$C_2 \triangleq C_1 + \lceil \ln_{\max\{1-\frac{\mu^2}{2\zeta^2}, 1-\frac{\Delta^2}{4\zeta^2}\}} \frac{\rho^2}{\zeta^2}\rceil = O(\log(1/\rho)/\eta).$$

We will choose t_{ALIGNMID} as the minimal $t \in S$, such that $\|\tilde{x}_1(t)\| \geq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho$.

Then by induction and Lemmas C.45 and C.46, we easily have that for $t \leq \min\{C_2+1, t_{\text{ALIGNMID}}\}$ and $t \in S$, we have that

$$x(t) \in K^{7h/8} \cap \cap_j \mathbb{I}_j\,,$$

$$\|\tilde{x}_1(t)\| \geq \min\{(1+C\eta)^{t/4}\|\tilde{x}_1(0)\|, \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho\}$$

$$\text{or} \quad \|\tilde{x}_1(\text{next}(t))\| \geq \min\{(1+C\eta)^{t/4}\|\tilde{x}_1(0)\|, \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho\}\,.$$

The detailed induction is analogous to previous inductive argument and is omitted. If $t_{\text{ALIGNMID}} \geq C_1$, then we have for the minimal $t \geq C_1$ and $t \in S$

$$\|\tilde{x}_1(t)\| \geq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\rho\,.$$

This is a contradiction and we have that $t_{\text{ALIGNMID}} \leq C_1$.

As we have $\|\tilde{x}_1(\text{next}(t))\| \geq \|\tilde{x}_1(t)\| - O(\eta\rho^2)$ for $\|\tilde{x}_1(t)\| \geq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho$,

$$\|\tilde{x}(t)\| \geq \|\tilde{x}_1(t))\| \geq \frac{1}{4}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + 3\frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho.$$

for $C_2 \geq t \geq t_{\text{ALIGNMID}}$.

We will then show that for $t \geq t_{\text{ALIGNMID}} + C_1$ iteration, $\|P^{(2:D)}\bar{x}(t+1)\| \leq O(\eta\rho^2)$.

For $C_1 \geq t \geq t_{\text{ALIGNMID}}$,

$$1-\eta\lambda_2 - \eta\rho\frac{\lambda_2^2}{\|\tilde{x}(t)\|} \leq 1-\eta\lambda_D - \eta\rho\frac{\lambda_D^2}{\|\tilde{x}(t)\|} \leq 1 - \frac{\lambda_D^2}{2\lambda_1^2} \leq 1 - \frac{\mu^2}{\zeta^2}$$

Notice that,

$$\begin{aligned}
&1-\eta\lambda_2 - \eta\rho\frac{\lambda_2^2}{\|\tilde{x}(t)\|} \\
\geq &1-\eta\lambda_2 - \eta\rho\frac{\lambda_2^2}{\|\tilde{x}(t)\|} \\
\geq &1-\eta\lambda_2 - \frac{4\lambda_2^2}{\lambda_1^2+3\lambda_2^2}(2-\eta\lambda_2) \\
\geq &-1 + \frac{2(\lambda_1^2-\lambda_2^2)}{\lambda_1^2+3\lambda_2^2} \geq -1 + \frac{\Delta^2}{2\zeta^2}
\end{aligned}$$

Hence,

$$\|P^{(2:D)}(t)x'(t+1)\|_2 \leq \max\{1-\frac{\mu^2}{\zeta^2}, 1-\frac{\Delta^2}{2\zeta^2}\}\|P^{(2:D)}(t)\tilde{x}(t)\|_2$$

Now by Lemma C.65 and Theorem C.67,

$$\begin{aligned}
\|P^{(2:D)}(t) - P^{(2:D)}(t+1)\| &\leq O(\eta\rho^2) \\
\|v_1(t) - v_1(t+1)\| &\leq O(\eta\rho^2) \\
\|\lambda_1(t) - \lambda_1(t+1)\| &\leq O(\eta\rho^2)
\end{aligned}$$

By Lemma C.39,

$$\|x'(t+1) - \tilde{x}(t+1)\| = O(\eta\rho^2)$$

Combining,

$$\|P^{(2:D)}(t+1)\tilde{x}(t+1)\| \leq \max\{1 - \frac{\mu^2}{2\zeta^2}, 1 - \frac{\Delta^2}{4\zeta^2}\}\|P^{(2:D)}(t)\tilde{x}(t)\| + O(\eta\rho^2)$$

Hence when $t = t_{\text{ALIGN}} = t_{\text{ALIGNMID}} + C_2$,

$$\|\tilde{x}(t)\| \geq \|\tilde{x}_1(t)\| \geq \Omega(\eta\rho)$$
$$\|P^{(2:D)}(t)\tilde{x}(t)\| \leq O(\eta\rho^2)$$

By $x(t) \in \mathbb{I}_1$, we easily have $\|\tilde{x}_1(t)\| = O(\eta\rho)$. Hence we conclude that

$$\|\tilde{x}_1(t)\| = \Theta(\eta\rho)$$
$$\|P^{(2:D)}(t)\tilde{x}(t)\| = O(\eta\rho^2)\,.$$

The second claim is just another induction similar to previous steps and is omitted as well.

□

Tracking Riemannian Gradient Flow We are now ready to show that $\Phi(x(t))$ will track the solution of Equation (6). The main principal of this proof has been introduced in Section 4.3.

Lemma C.48. *Under the condition of Theorem C.32, for any t satisfying that*

$$x(t) \in K^h,$$
$$\|\tilde{x}_1(t)\| = \Theta(\eta\rho),$$
$$\|P^{(2:D)}(t)\tilde{x}(t)\| = O(\eta\rho^2)\,,$$

it holds that

$$\|\Phi(x(t+1)) - \Phi(x(t)) + \eta\rho^2 P^{\perp}_{X,\Gamma}\nabla\lambda_1(t)/2\| \leq O(\eta\rho^3 + \eta^2\rho^2)\,.$$

Proof. To begin with, we can approximate $\Phi(x(t+1)) - \Phi(x(t))$ by its first order Taylor Expansion, by Lemma C.12,

$$\|\Phi(x(t+1)) - \Phi(x(t)) - \partial\Phi(x(t))(x(t+1) - x(t))\| = O(\|x(t+1) - x(t)\|^2) = O(\eta^2\rho^2)\,.$$

Then by plugging in the update rule and another Taylor Expansion,

$$\|\partial\Phi(x(t))(x(t+1) - x(t)) - \eta\rho\partial\Phi(x(t))\nabla^2 L(x)\frac{\nabla L(x)}{\|\nabla L(x)\|}$$
$$-\eta\rho^2\partial\Phi(x(t))\partial\nabla^2 L(x)[\frac{\nabla L(x)}{\|\nabla L(x)\|}, \frac{\nabla L(x)}{\|\nabla L(x)\|}]/2\|_2 = O(\eta\rho^3).$$

Using Lemma C.9, we have

$$\|\eta\rho\partial\Phi(x(t))\nabla^2 L(x)\frac{\nabla L(x)}{\|\nabla L(x)\|}\| = \eta\rho\|\nabla L(x)\|\|\partial^2\Phi(x(t))\left[\frac{\nabla L(x)}{\|\nabla L(x)\|}, \frac{\nabla L(x)}{\|\nabla L(x)\|}\right]\| = O(\eta\rho\|\nabla L(x)\|)\,.$$

Putting together, we have that

$$\|\Phi(x(t+1)) - \Phi(x(t)) - \eta\rho^2\partial\Phi(x(t))\partial\nabla^2 L(\Phi(x(t)))[\frac{\nabla L(x)}{\|\nabla L(x)\|}, \frac{\nabla L(x)}{\|\nabla L(x)\|}]/2\|$$
$$\leq O(\eta^2\rho^2 + \eta\rho^3) + O(\eta\rho\|\nabla L(x)\|)\,.$$

As we have $\|\tilde{x}(t)\| = \Theta(\eta\rho)$, hence by Lemmas C.8 and C.36,

$$\|\Phi(x(t+1)) - \Phi(x(t)) - \eta\rho^2 \partial\Phi(x(t))\partial\nabla^2 L\left(\Phi(x(t))\right)[\frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|}, \frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|}]/2\|$$
$$\leq O(\eta\rho^3 + \eta^2\rho^2)$$

Finally, we have that

$$\|\eta\rho^2 \partial\Phi(x(t))\partial\nabla^2 L\left(\Phi(x(t))\right)[\frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|}, \frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|}]/2$$
$$- \eta\rho^2 \partial\Phi(x(t))\partial\nabla^2 L\left(\Phi(x(t))\right)[v_1(t), v_1(t)]/2\| \leq O(\eta\rho^3)$$

as the angle between $\frac{\nabla L(x)}{\|\nabla L(x)\|}$ and $v_1(t)$ is $O(\rho)$.

By Lemma C.9, it holds that

$$\partial\Phi(x(t))\partial\nabla^2 L\left(\Phi(x(t))\right)[v_1(t), v_1(t)] = P^{\perp}_{X,\Gamma}\nabla(\lambda_1(t))$$

Putting together we have that,

$$\|\Phi(x(t+1)) - \Phi(x(t)) + \eta\rho^2 P^{\perp}_{X,\Gamma}\nabla\lambda_1(t)/2\| \leq O(\eta\rho^3 + \eta^2\rho^2).$$

It completes the proof. □

C.5.3 Proofs of Corollaries 4.10 and 4.11

Proof of Corollary 4.10. We will do a Taylor expansion on L^{Max}_{ρ}. By Theorem C.31 and C.32, we have $\|x(\lceil T_3/\eta\rho^2\rceil)) - X(T_3)\| = \tilde{O}(\eta + \rho)$ and $\|x(\lceil T_3/\eta\rho^2\rceil)) - \Phi(x(\lceil T_3/\eta\rho^2\rceil)))\|_2 = O(\eta\rho)$. For convenience, we denote $x(\lceil T_3/\eta\rho^2\rceil)$ by x.

$$R^{\text{Max}}_{\rho}(x) = \max_{\|v\|_2\leq 1} \rho v^T\nabla L(x) + \rho^2 v^T\nabla^2 L(x)v/2 + O(\rho^3)$$

Since $\max_{\|v\|_2\leq 1}\|v^T\nabla L(x)\|_2 = O(\|x - \Phi(x)\|_2) = O(\eta\rho)$, it holds that

$$\begin{aligned} R^{\text{Max}}_{\rho}(x) &= \rho^2 \max_{\|v\|_2\leq 1} v^T\nabla^2 L(x)v/2 + O(\eta^2\rho^2 + \rho^3) \\ &= \rho^2\lambda_1(\nabla^2 L(x)) + O(\eta^2\rho^2 + \rho^3) \\ &= \rho^2\lambda_1(\nabla^2 L(X(T_3))) + \tilde{O}(\eta\rho^2), \end{aligned}$$

which completes the proof. □

Proof of Corollary 4.11. We choose T such that $X(T_\epsilon)$ is sufficiently close to $X(\infty)$, such that $\lambda_1(X(T_\epsilon)) \leq \lambda_1(X(\infty)) + \epsilon/2$. By Corollary 4.10 (let $T_3 = T_\epsilon$), we have that for all ρ, η such that $\eta\ln(1/\rho)$ and ρ/η are sufficiently small, $\|R^{\text{Max}}_{\rho}(x(\lceil T_\epsilon/(\eta\rho^2)\rceil)) - \rho^2\lambda_1(X(T_\epsilon))/2\| \leq \tilde{o}(1)$. This further implies $\|R^{\text{Max}}_{\rho}(x(\lceil T_\epsilon/(\eta\rho^2)\rceil)) - \rho^2\lambda_1(X(\infty))/2\| \leq \epsilon\rho^2 + o(1)$. We also have $L(x(\lceil T_\epsilon/(\eta\rho^2)\rceil)) - \inf_{x\in U'} L(x) = o(1)$. Then we can leverage Theorem 4.3 and Theorem 4.5 to get the desired bound. □

C.6 Analysis for 1-SAM (Proof of Theorem 5.4)

The goal of this section is to prove the following theorem.

Theorem 5.4. *Let $\{x(t)\}$ be the iterates defined by SAM (16) and $x(0) = x_{init} \in U$, then under Assumption 3.1, for all η and ρ such that $(\eta + \rho)\log(1/\eta\rho)$ is sufficiently small, with probability at least $1 - O(\sqrt{\rho} + \sqrt{\eta})$, the dynamics of SAM can be split into two phases:*

- ***Phase I:*** *Stochastic SAM follows Gradient Flow with respect to L until entering an $O(\eta\rho)$ neighborhood of the manifold Γ in $\tilde{O}(\frac{1}{\eta})$ steps;*
- ***Phase II:*** *Stochastic SAM tracks the solution X of (17), the Riemannian Gradient Flow with respect to $\text{Tr}(\nabla^2 L(\cdot))$, with time scaling as $\|\Phi(x(\lceil T_3/(\eta\rho^2)\rceil)) - X(T_3)\| = O(\eta^{1/4} + \rho^{1/4})$.*

As mentioned in our proof setups in Appendix C, we will prove Theorem 5.4 under a more general (and weaker) condition, namely Condition C.2 and Assumption 3.1. The only usage of Setting C.1 in the proof is Theorem C.49 and lemma 5.3.

Theorem C.49. *Condition C.2 holds under Setting C.1 and assumption 3.1.*

Proof of Theorem C.49. 1. $L = \frac{1}{M}\sum_{k=1}^{M} L_k$ by definition.
2. $\forall k \in [M]$, $L_k(x) = \ell(f_k(x), y_k)$ is $\mathcal{C}^4$ as ℓ and f_k are both $\mathcal{C}^4$.
3. For any $x \in \Gamma$, by Lemma 5.3, we have $\nabla f_k(x) \neq 0$. Then there exists an open neighborhood V_k such that $\Gamma \subset V_k$ and $\nabla f_k(x) \neq 0$ for any $x \in V_k, k \in [M]$. Then applying implicit function theorem, for any $k \in M$ there exists a $(D-1)$-dimensional $\mathcal{C}^4$-manifold $\Gamma'_k \subset V_k$, such that for any $x' \in V$, $f_k(x') = y_k$ if and only if $x' \in \Gamma'_k$. As for any $x \in \Gamma \subset V_k$, $f_k(x') = y_k$, we can infer that $\Gamma \subset \Gamma'_k$. Then $\Gamma \subset \cup_{k=1}^{M} \Gamma_k$.
4. For any $x \in \Gamma_k$, we have $f_k(x) = y_k$, which implies $L_k(x) = 0$. Also as $x \in V$,$\nabla f_k(x) \neq 0$. By Lemma C.64, we have $\mathrm{rank}(\nabla^2 L(x)) = 1$.

□

Analogous to the full-batch setting, we will split the trajectory into two phase.

Theorem C.50 (Phase I)**.** *Let $\{x(t)\}$ be the iterates defined by SAM (Equation (16)) and $x(0) = x_{init} \in U$, then under Assumption 3.1 and C.2, for almost every x_{init}, there exists a constant T_1, it holds for sufficiently small $(\eta+\rho)\ln 1/\eta\rho$, we have with probability $1 - O(\sqrt{\rho})$, $\min_{T_1 \ln(1/\eta\rho)/\eta \geq t} \|x(t) - \Phi(x(t))\|_2 = O(\eta\rho)$ and $\max_{T_1 \ln(1/\eta\rho)/\eta \geq t} \|\Phi(x_{init}) - \Phi(x(t))\|_2 = O((\eta+\rho)\ln 1/\eta\rho)$.*

Theorem C.50 shows that SAM will converges to an $\tilde{O}(\eta\rho)$ neighborhood of the manifold without getting far away from $\Phi(x(0))$, where we can perform a local analysis on the trajectory of $\Phi(x(t))$.

Under Assumptions 3.1 and C.2, we have $\mathrm{Tr}(\nabla^2 L_k(x)) = \lambda_1(\nabla^2 L_k(x))$ is differentiable for $x \in \Gamma_i$. Hence $\mathrm{Tr}(\nabla^2 L(x)) = \sum_{k=1}^{M} \mathrm{Tr}(\nabla^2 L_k(x))$ is also differentiable and we have (17) is well defined for some finite time T_2.

Theorem C.51 (Phase II)**.** *Let $\{x(t)\}$ be the iterates defined by SAM (Equation (16)) under Assumptions 3.1 and C.2, assuming (1) $\|x(0) - \Phi(x(0))\|_2 = O(\eta\rho)$ and (2) $\|\Phi(x_{init}) - \Phi(x(0))\|_2 = O((\eta+\rho)\ln 1/(\eta\rho))$, then for almost every $x(0)$, for any $T_2 > 0$ till which solution of (17) X exists, for sufficiently small $(\eta+\rho)\ln 1/(\eta\rho)$, we have with probability $1 - O(\eta\rho)$, for all $\eta\rho^2 t < T_2$, $\|\Phi(x(t)) - X(\eta\rho^2 t)\|_2 = O((\eta+\rho)\ln 1/\eta\rho)$ and $\|x(t) - \Phi(x(t))\|_2 = O(\eta\rho \ln 1/\eta\rho)$.*

Combining Theorems C.49 to C.51, the proof of Theorem 5.4 is clear.

Now we recall our notations for stochastic setting with batch size one.

Notations for Stochastic Setting: Since L_k is rank-1 on Γ_k for each $k \in [M]$, we can write it as $L_k(x) = \Lambda_k(x) w_k(x) w_k^\top(x)$ for any $x \in \Gamma_k$, where w_k is a continuous function on Γ_k with pointwise unit norm. Given the loss function L_k, its gradient flow is denoted by mapping $\phi_k : \mathbb{R}^D \times [0, \infty) \to \mathbb{R}^D$. Here, $\phi_k(x, \tau)$ denotes the iterate at time τ of a gradient flow starting at x and is defined as the unique solution of $\phi_k(x, \tau) = x - \int_0^\tau \nabla L_k(\phi_k(x, t))dt$, $\forall x \in \mathbb{R}^D$. We further define the limiting map Φ_k as $\Phi_k(x) = \lim_{\tau\to\infty} \phi_k(x, \tau)$, that is, $\Phi_k(x)$ denotes the convergent point of the gradient flow starting from x. Similar to Definition 3.2, we define $U_k = \{x \in \mathbb{R}^D | \Phi(x) \text{ exists and } \Phi_k(x) \in \Gamma_k\}$ be the attraction set of Γ_i. We have that each U_k is open and Φ_k is $\overline{\mathcal{C}}^2$ on U_k by Lemma B.15 in Arora et al. (2022).

In this section we will define K as $\{X(t) \mid t \in [0, T_3]\}$ where X is the solution of (17).

We will assume the update is always well defined.

C.6.1 Phase I (Proof of Theorem C.50)

The goal of this section is to prove Theorem C.50.

Proof. The proof consists of two major steps.

1. *Tracking Gradient Flow* Lemma C.52 shows the existence of step $t_{\mathrm{GF}} = O(1/\eta)$ such that $x(t_{\mathrm{GF}})$ is in a subset of K^h and $\Phi(x(t_{\mathrm{GF}}))$ is $O(\eta+\rho)$ close to $\Phi(x_{\text{init}})$.
2. *Decreasing Loss* Lemmas C.56 and C.61 shows existence of step $t_{\mathrm{DEC2}} = O(\ln(1/\eta\rho)/\eta)$ such that $x(t_{\mathrm{DEC2}})$ is in a $O(\eta\rho)$ neighbor of K^h and $\Phi(x(t_{\mathrm{DEC2}}))$ is $O((\eta+\rho)\ln(1/\eta\rho))$ close to $\Phi(x_{\text{init}})$.

□

Tracking Gradient Flow

Lemma C.52. *Under condition of Theorem C.50, with probability $1 - O(\sqrt{\rho} + \sqrt{\eta})$, there exists $t_{\mathrm{GF}} = O(1/\eta)$, such that the iterate $x(t_{\mathrm{GF}})$ is $O(1)$ close to the manifold Γ and $\Phi(x(t_{\mathrm{GF}}))$ is $O(\eta+\rho)$ is close to $\Phi(x_{init})$. Quantitatively,*

$$
\begin{aligned}
L(x(t_{\mathrm{GF}})) &\leq \frac{\mu h^2}{32} \\
\|x(t_{\mathrm{GF}}) - \Phi(x(t_{\mathrm{GF}}))\| &\leq h/4\,, \\
\|\Phi(x(t_{\mathrm{GF}})) - \Phi(x_{init})\| &= O(\eta^{1/4} + \rho^{1/4})\,.
\end{aligned}
$$

Proof. Choose $C = \frac{1}{4}\sqrt{\frac{\mu}{\zeta}}$.

There exists $T > 0$, such that

$$
\|\phi(x_{\text{init}}, T) - \Phi(x_{\text{init}})\|_2 \leq Ch/2\,.
$$

Consider

$$
\begin{aligned}
x(t+1) &= x(t) - \eta\nabla L_k(x(t) + \rho\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|}) \\
&= x(t) - \eta\nabla L_k(x(t)) + O(\eta\rho)\,.
\end{aligned}
$$

By Theorem D.5, by choosing $p = \eta$ and $q = \eta + \rho$, for sufficiently small η and ρ, the iterates $x(t)$ tracks gradient flow $\phi(x_{\text{init}}, T)$ in $O(1/\eta)$ steps in expectation, Quantitatively, with probability $1 - \sqrt{\eta} - \sqrt{\rho}$, for $t_{\mathrm{GF}} = \lceil \frac{T_0}{\eta} \rceil$, we have that

$$
\|x(t_{\mathrm{GF}}) - \mathbb{E}[\phi(x_{\text{init}}, T)]\|_2 \leq O(\eta^{1/4} + \rho^{1/4})\,.
$$

This implies $x(t_{\mathrm{GF}}) \in K^h$, hence by Taylor Expansion on Φ,

$$
\begin{aligned}
\|\Phi(x(t_{\mathrm{GF}})) - \Phi(x_{\text{init}})\|_2 &= \|\Phi(x(t_{\mathrm{GF}})) - \Phi(\phi(x_{\text{init}}, T))\|_2 \\
&\leq O(\|x(t_{\mathrm{GF}}) - \phi(x_{\text{init}}, T)\|_2) \\
&\leq O(\eta^{1/4} + \rho^{1/4})\,.
\end{aligned}
$$

This implies

$$
\begin{aligned}
\|x(t_{\mathrm{GF}}) - \Phi(x(t_{\mathrm{GF}}))\|_2 \leq& \|x(t_{\mathrm{GF}}) - \phi(x_{\text{init}}, T_0)\|_2 + \|\phi(x_{\text{init}}, T_0) - \Phi(x_{\text{init}})\|_2 \\
&+ \|\Phi(x_{\text{init}}) - \Phi(x(t_{\mathrm{GF}}))\|_2 \\
\leq& Ch/2 + O(\eta^{1/4} + \rho^{1/4}) \leq Ch \leq h/4\,.
\end{aligned}
$$

By Taylor Expansion,

$$
L(x(t_{\mathrm{GF}})) \leq \zeta\|x(t_{\mathrm{GF}}) - \Phi(x(t_{\mathrm{GF}}))\|_2^2/2 \leq \frac{\mu h^2}{32}\,.
$$

□

Decreasing Loss

Lemma C.53. *Under condition of Theorem C.50, assuming $x(t_0) \in K^{h/4}$ and for any t satisfying $t_0 \le t \le t_0 + O(\ln(1/\eta\rho)/\eta)$, $\max_{t_0 \le \tau \le t_0 + O(\ln(1/\eta\rho)/\eta)} L(x(\tau)) \le \frac{\mu h^2}{16}$, it holds that*

$$x(\tau) \in K^h, \forall t_0 \le \tau \le t.$$

Moreover, we have that

$$\|\Phi(x(t)) - \Phi(x(t_0))\| = O((\eta + \rho)\ln(1/\eta\rho)).$$

Proof. We will prove by induction. For $\tau = t_0$, the result holds trivially. Suppose the result holds for $t-1$, then for any τ satisfying $t_0 \le \tau \le t-1$, by Lemmas C.7 and C.13,

$$\begin{aligned}\|\Phi(x(\tau+1)) - \Phi(x(\tau))\| &\le \xi\eta\rho\|\nabla L\left(x(\tau)\right)\|_2 + \nu\eta\rho^2 + \xi\eta^2\|\nabla L\left(x(\tau)\right)\|_2^2 + \xi\zeta^2\eta^2\rho^2 \\ &= O(\eta^2 + \eta\rho)\,.\end{aligned}$$

Also by Lemma C.7, $\|x(t) - \Phi(x(t))\|_2 \le h/2\sqrt{2}$, this implies,

$$\begin{aligned}\mathrm{dist}(K, x(t)) \le& \mathrm{dist}(K, x(t_0)) + \|x(t_0) - \Phi(x(t_0))\|_2 \\ &+ \|\Phi(x(t_0)) - \Phi(x(t))\| + \|\Phi(x(t)) - x(t)\| \\ \le& 0.99h + O(\eta^2(t - t_{\mathrm{GF}})) = 0.99h + O(\eta\ln(1/\eta\rho)) \le h\,.\end{aligned}$$

□

Lemma C.54. *Under condition of Theorem C.50, if $x(\tau) \in K^h$, then we have that*

$$\mathbb{E}[L(x(\tau+1))|x(\tau)] \le L(x(\tau)) - \frac{\eta\mu}{2}L(x(\tau))\,.$$

Moreover it holds that,

$$\mathbb{E}[\ln L(x(\tau+1))|x(\tau)] \le \ln\mathbb{E}[L(x(\tau+1))|x(\tau)] \le \ln L(x(\tau)) - \frac{\eta\mu}{2}\,.$$

Proof. By Lemma C.13 and Taylor Expansion,

$$\begin{aligned}&\mathbb{E}[L(x(\tau+1))|x(\tau)] \\ =&\mathbb{E}\left[L\left(x(\tau) - \eta\nabla L_k[x(\tau) + \rho\frac{\nabla L_k\left(x(\tau)\right)}{\|\nabla L_k\left(x(\tau)\right)\|}]\right)|x(\tau)\right] \\ \le&\mathbb{E}\left[L(x(\tau)) - \eta\left\langle\nabla L\left(x(\tau)\right), \nabla L_k\left(x(\tau) + \rho\frac{\nabla L_k\left(x(\tau)\right)}{\|\nabla L_k\left(x(\tau)\right)\|}\right)\right\rangle\right] \\ &+ \mathbb{E}\left[\frac{\zeta\eta^2}{2}\|\nabla L_k[x(\tau) + \rho\frac{\nabla L_k\left(x(\tau)\right)}{\|\nabla L_k\left(x(\tau)\right)\|}]\|_2^2\right] \\ \le& L(x(\tau)) - \eta\|\nabla L\left(x(\tau)\right)\|_2^2 + \eta\rho\zeta\|\nabla L\left(x(\tau)\right)\|_2 + \zeta\eta^2\mathbb{E}[\|\nabla L_k(x(\tau))\|_2^2] + \zeta^3\eta^2\rho^2 \\ \le& L(x(\tau)) - \frac{\eta}{2}\|\nabla L\left(x(\tau)\right)\|_2^2 \\ \le& L(x(\tau)) - \frac{\eta\mu}{2}L(x(\tau))\,.\end{aligned}$$

□

Lemma C.55. *Under condition of Theorem C.50, assuming $x(t_0) \in K^{h/4}$ and $L(x(t_0)) \le \frac{\mu h^2}{32}$, then with probability $1 - O(\rho)$, for any t satisfying $t_0 \le t \le t_0 + O(\ln(1/\eta\rho)/\eta)$, it holds that $x(t) \in K^h$. Moreover, we have that*

$$\|\Phi(x(t)) - \Phi(x(t_0))\| = O((\eta + \rho)\ln(1/\eta\rho)).$$

Proof. By Uniform Bound and Lemma C.53,

$$\begin{aligned}
&\mathbb{P}(\exists t_0 \le t \le t_0 + O(\ln(1/\eta\rho)/\eta), L(x(t)) \ge \frac{\mu h^2}{16}) \\
\le &\sum_{t=t_0}^{t_0+O(\ln(1/\eta\rho)/\eta)} \mathbb{P}(L(x(t)) \ge \frac{\mu h^2}{16} \quad \text{and} \quad L(x(\tau)) \le \frac{\mu h^2}{16}, \forall t_0 \le \tau \le t-1) \\
\le &\sum_{t=t_0}^{t_0+O(\ln(1/\eta\rho)/\eta)} \mathbb{P}(L(x(t)) \ge \frac{\mu h^2}{16} \quad \text{and} \quad x(\tau) \in K^h, \forall t_0 \le \tau \le t-1)
\end{aligned}$$

Consider each term, and applying uniform bound again,

$$\begin{aligned}
&P(L(x(t)) \ge \frac{\mu h^2}{16} \quad \text{and} \quad x(\tau) \in K^h, \forall t_0 \le \tau \le t-1) \\
\le &\sum_{\tau=t_0}^{t} P(L(x(t)) \ge \frac{\mu h^2}{16} \quad \text{and} \quad L(x(\tau)) \le \frac{\mu h^2}{32} \\
&\qquad \text{and} \quad \forall t-1 \ge \tau' \ge \tau+1, \frac{\mu h^2}{16} > L(x(\tau')) > \frac{\mu h^2}{32} \\
&\qquad \text{and} \quad \forall t-1 \ge \tau'' \ge \tau, x(\tau'') \in K^h) \,.
\end{aligned}$$

Then if we consider each term, we have that it is bounded by

$$\begin{aligned}
&P(L(x(t)) \ge \frac{\mu h^2}{16} \quad \text{and} \quad \forall t-1 \ge \tau' \ge \tau+1, L(x(\tau')) > \frac{\mu h^2}{32} \\
&\quad \text{and} \quad \forall t-1 \ge \tau'' \ge \tau, x(\tau'') \in K^h \mid L(x(\tau)) \le \frac{\mu h^2}{32}) \,.
\end{aligned}$$

Define a coupled process $\tilde{L}(\tau+1) = \ln L(x(\tau+1))$ and

$$\tilde{L}(\tau') = \begin{cases} \ln L(x(\tau')), & \text{if } \tilde{L}(\tau'-1) = \ln L(x(\tau'-1)) \ge \ln(\frac{\mu h^2}{32}), \\ \tilde{L}(\tau'-1) - \eta\mu/2, & \text{if otherwise.} \end{cases}$$

Then clearly

$$\begin{aligned}
&P(L(x(t)) \ge \frac{\mu h^2}{16} \quad \text{and} \quad \forall t \ge \tau' \ge \tau+1, L(x(\tau')) > \frac{\mu h^2}{32} \\
&\quad \text{and} \quad \forall t \ge \tau'' \ge \tau, x(\tau'') \in K^h \mid L(x(\tau)) \le \frac{\mu h^2}{32}) \\
\le &P(\tilde{L}(t) \ge \ln(\frac{\mu h^2}{16})) \,.
\end{aligned}$$

Consider a fixed τ' satisfying $\tau+1 \le \tau' \le t$. By Lemma C.54, we have that

$$\tilde{L}(x(\tau'+1)) - \tilde{L}(x(\tau')) \le -\eta\mu/2.$$

Hence $\tilde{L}(t) + \eta\mu t/2$ is a super martingale.

Further it holds that if $L(x(\tau'-1)) \ge (\frac{\mu h^2}{32})$, then

$$L(x(\tau'-1)) - L(x(\tau')) = O(\|x(\tau'-1) - x(\tau')\|) = O(\eta) \,.$$

Using the smoothness at $\log(x)$ at $\frac{\mu h^2}{32}$ which is a positive constant,

$$\|\tilde{L}(\tau'+1) - \tilde{L}(\tau')\| \le O(\eta) \le C\eta \,.$$

Here C is a constant independent of η. This implies $\tilde{L}(x(\tau+1)) \le \frac{\mu h^2}{16\sqrt{2}}$

Now by Azuma-Hoeffding bound (Lemma C.68), we have that

$$P(\tilde{L}(t) - \tilde{L}(\tau+1) + (t-\tau-1)\eta\mu/2 > a) \leq 2\exp(-\frac{a^2}{8(t-\tau-1)(C+\mu)^2\eta^2}).$$

With $a = \ln(\frac{\mu h^2}{16\tilde{L}(\tau+1)}) + (t-\tau-1)\eta\mu/2 \geq (\ln 2 + (t-\tau-1)\eta\mu)/2$, we have that

$$\begin{aligned} P(\tilde{L}(t) > \ln(\frac{\mu h^2}{16})) &\leq 2\exp(-\frac{(\ln 2 + (t-\tau-1)\eta\mu)^2}{32(C+\mu)^2\eta^2}) \\ &\leq 2\exp(-\frac{\ln 2(t-\tau-1)\mu}{8(C+\mu)^2\eta}) \end{aligned}$$

Hence we have

$$\begin{aligned} &\mathbb{P}(\exists t_0 \leq t \leq t_0 + O(\ln(1/\eta\rho)/\eta), L(x(t)) \geq \frac{\mu h^2}{16}) \\ \leq &O(2\exp(-\frac{\ln 2(t-\tau-1)\mu}{8(C+\mu)^2\eta})\ln^2(1/\eta\rho)/\eta^2) \leq \rho. \end{aligned}$$

Hence with probability $1-\rho$, $L(x(t)) \leq \frac{\mu h^2}{16}, \forall t_0 \leq t \leq t_0 + O(\ln(1/\eta\rho)/\eta)$, combining with Lemma C.53, we have completed our proof.

□

Lemma C.56. *Under condition of Theorem C.50, assuming there exists t_{GF} such that $L(x(t_{\mathrm{GF}})) \leq \frac{\mu h^2}{32}$ and $x(t_{\mathrm{GF}}) \in K^{h/4}$, then with probability $1-O(\rho)$, there exists $t_{\mathrm{DEC}} = t_{\mathrm{GF}} + O(\ln(1/\rho)/\eta)$, such that $x(t_{\mathrm{DEC}})$ is in $O(\rho)$ neighbor of Γ, quantitatively, we have that*

$$\|\nabla L(x(t_{\mathrm{DEC}}))\|_2 \leq 4\zeta\rho\,.$$

Moreover the movement of the projection of $\Phi(x(\cdot))$ on the manifold is bounded,

$$\|\Phi(x(t_{\mathrm{GF}})) - \Phi(x(t_{\mathrm{DEC}}))\|_2 = O((\eta+\rho)\ln(1/\rho))\,.$$

Proof. For simplicity of writing, define $T_1 \triangleq \lceil \frac{2\ln\frac{h^2}{256\rho^3\mu}}{\eta\mu} \rceil = O(\ln(1/\rho)/\eta)$.

By Lemma C.55, we may assume $x(t) \in K^h$ for $t_{\mathrm{GF}} \leq t \leq T_1 + t_{\mathrm{GF}}$.

Define indicator function as

$$\mathcal{A}(t) = \mathbf{1}[\nabla L\left(x(\tau)\right) \geq 4\zeta\rho, \forall t \geq \tau \geq t_{\mathrm{GF}}]\,.$$

By Lemma C.54, we have that,

$$\mathbb{E}[L(x(t+1))\mathcal{A}(t+1)] \leq \mathbb{E}[L(x(t+1))\mathcal{A}(t)] \leq (1-\frac{\eta\mu}{2})\mathbb{E}[L(x(t))\mathcal{A}(t)].$$

We can then conclude that with $T_2 = T_1 + t_{\mathrm{GF}}$, using Lemma C.8,

$$8\mu\rho^2\mathbb{E}\mathcal{A}(T_2+1) \leq \mathbb{E}[L(x(T_2+1))\mathcal{A}(T_2+1)] \leq (1-\frac{\eta\mu}{2})^{T_1}L(x(t_{\mathrm{GF}})) \leq 8\mu\rho^3.$$

We have

$$\mathbb{E}\mathcal{A}(T_2+1) \leq \rho.$$

This implies $\mathcal{A}(T_2+1) = 0$ with probability $1-O(\rho)$, which indicates the existence of t_{DEC}. The second claim is a direct application of Lemma C.55. □

Lemma C.57 (A general version of Lemma 5.3)**.** *Under Assumption 3.1 and Condition C.2, for $x \in K^h$ and $p \in C$, $\nabla^2 L_k(p) = \Lambda_k(p) w_k(p) w_k(p)^\top$, there exists $s \in \{1, -1\}$,*

$$\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = s w_k(p) + O(\|x-p\|_2).$$

Further if $|w_k^\top (x-p)| \geq \|x-p\|_2^{3/2}$, then $s = \operatorname{sign}(w_k^\top (x-p))$. This implies

$$\begin{aligned}\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|}^\top (x-p) &\geq s w_k^\top (x-p) - O(\|x-p\|_2^2) \\ &\geq \|w_k^\top (x-p)\|_2 - O(\|x-p\|_2^{3/2}).\end{aligned}$$

Proof. We will calculate the direction of $\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|}$ using two different approximations and compare them to get our result.

1. According to Lemma C.10,

$$\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = \frac{\nabla^2 L_k(\Phi_k(x))(x-\Phi_k(x))}{\|\nabla^2 L_k(\Phi_k(x))(x-\Phi_k(x))\|_2} + O(\|x-\Phi_k(x)\|_2).$$

   Suppose $\nabla^2 L_k(\Phi_k(x)) = \Lambda_k(\Phi_k(x)) w_k(\Phi_k(x)) w_k(\Phi_k(x))^\top$, then

$$\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = w_k(\Phi_k(x)) + O(\|x-\Phi_k(x)\|_2)$$

   As $\nabla^2 L_k(p) = \Lambda_k(p) w_k(p) w_k(p)^\top$, using Davis-Kahan Theorem C.67, we would have $\exists s \in \{-1, 1\}$, such that $\|w_k(\Phi_k(x)) - s w_k(p)\|_2 \leq \zeta \|\Phi_k(x) - p\|_2$.

$$\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = s w_k(p) + O(\|\Phi_k(x)-p\|_2 + \|x-p\|_2).$$

   According to Lemma C.7, we have $\|x-\Phi_k(x)\|_2 \leq \frac{\|\nabla L_k(x)\|_2}{\mu} \leq \frac{\zeta \|x-p\|_2}{\mu}$. This implies,

$$\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = s w_k(p) + O(\|x-p\|_2). \tag{33}$$

   Equation (33) is our first statement.
2. By Taylor expansion at p,

$$\nabla L_k(x) = \Lambda_k(x) w_k(p) w_k(p)^\top (x-p) + O(\nu \|x-p\|_2^2).$$

   That being said, when $|w_k^\top (x-p)| \geq \|x-p\|_2^{3/2}$, we have

$$\begin{aligned}&\|\nabla L_k(x) - \Lambda_k w_k w_k^\top (x-p)\|_2 \leq O(\|x-p\|_2^2). \\ &\|\nabla L_k(x)\| \geq \|\Lambda_k w_k w_k^\top (x-p)\|_2 - O(\|x-p\|_2^2) \geq \Omega(\|x-p\|_2^{3/2}).\end{aligned}$$

   Concluding,

$$\left\|\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} - \frac{\Lambda_k w_k w_k^\top (x-p)}{\|\Lambda_k w_k w_k^\top (x-p)\|}\right\|_2 \leq O(\|x-p\|_2^{1/2})$$

   Hence we have

$$\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = \operatorname{sign}(w_k^\top (x-p)) w_k + O(\|x-p\|_2^{1/2}). \tag{34}$$

   Comparing (33) and (34), we have $s = \operatorname{sign}(w_k(p)^\top (x-p))$ when $|w_k^\top (x-p)| \geq \|x-p\|_2^{3/2}$.

□

Lemma C.58. *Under condition of Theorem C.50, for any constant $C > 0$ independent of η, ρ, there exists constant $C_1 > C_2 > 0$ independent of η, ρ, if $x(t) \in K^h$ and $C_1 \eta \rho \leq \|x(t) - \Phi(x(t))\| \leq C\rho$, then we have that*

$$\mathbb{E}_k[\|x(t+1) - \Phi(x(t+1))\|_2 \mid x(t)] \leq \|x(t) - \Phi(x(t))\|_2 - C_2 \eta \rho.$$

Proof. By Lemma C.8, $\|x(t) - \Phi(x(t))\| = O(\rho)$. Hence we have that by Taylor Expansion,

$$\begin{aligned} x(t+1) &= x(t) - \eta \nabla L_k \left(x(t) + \rho \frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|} \right) \\ &= x(t) - \eta \nabla L_k\left(x(t)\right) - \eta\rho \nabla^2 L_k(x(t)) \frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|} + O(\eta\rho^2) \\ &= x(t) - \eta \nabla L_k\left(x(t)\right) - \eta\rho \Lambda_k w_k w_k^\top \frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|} + O(\eta\rho^2)\,. \end{aligned}$$

Here Λ_k, w_k indicates $\Lambda_k(\Phi(x(t))), w_k(\Phi(x(t)))$.

Notice that given $\|x(t) - \Phi(x(t))\| = O(\rho)$, by Lemma C.13, we have that

$$\begin{aligned} \|\Phi(x(t+1)) - \Phi(x(t))\|_2 &= O(\eta\rho^2), \\ \|x(t+1) - x(t)\|_2 &= O(\eta\rho). \end{aligned}$$

This implies $x(t+1) \in K^r$.

Further by Taylor Expansion, $\nabla L_k(x(t)) = \Lambda_k w_k w_k^\top (x(t) - \Phi(x(t))) + O(\rho^2)$.

By Lemma C.57, we have for some $s_k(t) \in \{-1, 1\}$.

$$w_k^\top \frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|} = s_k(t) w_k + O(\|x(t) - \Phi(x(t))\|_2)\,.$$

We also have

$$s_k(t) \neq \text{sign}(w_k^\top (x(t) - \Phi(x(t)))) \Rightarrow \|w_k^\top (x(t) - \Phi(x(t)))\|_2 \leq \|x(t) - \Phi(x(t))\|_2^{3/2}. \tag{35}$$

Concluding,

$$\begin{aligned} &x(t+1) - \Phi(x(t+1)) \\ =&(x(t) - \Phi(x(t))) - \eta \Lambda_k w_k w_k^\top (x(t) - \Phi(x(t))) - \eta\rho \Lambda_k s_k(t) w_k w_k^\top w_k + O(\eta\rho^2). \end{aligned}$$

After we take square and expectation,

$$\begin{aligned} &\mathbb{E}[\|x(t+1) - \Phi(x(t+1))\|_2^2 \mid x(t)] \\ \leq&\|x(t) - \Phi(x(t))\|_2^2 + \frac{2\eta^2}{M} \sum_{k=1}^M \Lambda_k^2 |w_k^\top (x(t) - \Phi(x(t)))|^2 + \frac{2\eta^2\rho^2}{M} \sum_{k=1}^M \Lambda_k^2 \\ &- 2\frac{\eta}{M} \sum_{k=1}^M \Lambda_k |w_k^\top (x(t) - \Phi(x(t)))|^2 - 2\frac{\eta\rho}{M} \sum_{k=1}^M \Lambda_k s_k(t) w_k^\top (x(t) - \Phi(x(t))) \\ &+ O(\eta\rho^2 \|x(t) - \Phi(x(t))\| + \eta^2\rho^3)\,. \end{aligned}$$

We will then carefully examine each positive term,

$$\begin{aligned} \frac{2\eta^2}{M} \sum_{k=1}^M \Lambda_k^2 |w_k^\top (x(t) - \Phi(x(t)))|^2 &= 2M\eta^2 (x(t) - \Phi(x(t)))^\top \nabla^2 L(x(t))^2 (x(t) - \Phi(x(t))) \\ &\leq 2M\zeta\eta^2 \|x(t) - \Phi(x(t))\|^2 = O(\eta^2\rho^2)\,. \end{aligned}$$

$$\frac{2\eta^2\rho^2}{M} \sum_{k=1}^M \Lambda_k^2 \leq 2\zeta^2\eta^2\rho^2 = O(\eta^2\rho^2)\,.$$

This implies,

$$\begin{aligned} &\mathbb{E}[\|x(t+1) - \Phi(x(t+1))\|_2^2 \mid x(t)] \\ \leq&\|x(t) - \Phi(x(t))\|_2^2 - 2\frac{\eta\rho}{M} \sum_{k=1}^M \Lambda_k s_k(t) w_k^\top (x(t) - \Phi(x(t))) \\ &+ O(\eta\rho^2 \|x(t) - \Phi(x(t))\| + \eta^2\rho^2)\,. \end{aligned}$$

We will now lower bound $\sum_{k=1}^{M} \Lambda_k s_k(t) w_k^\top (x(t) - \Phi(x(t)))$. By Equation (35),

$$\begin{aligned}\sum_{k=1}^{M} \Lambda_k s_k(t) w_k^\top (x(t) - \Phi(x(t))) &\geq \sum_{k=1}^{M} \Lambda_k \|w_k^\top (x(t) - \Phi(x(t)))\|_2 - 2\sum_{k=1}^{M} \Lambda_k \|x(t) - \Phi(x(t))\|_2^{3/2} \\ &\geq \sum_{k=1}^{M} \Lambda_k \|w_k^\top (x(t) - \Phi(x(t)))\|_2 - O(\|x(t) - \Phi(x(t))\|_2^{3/2}) \,.\end{aligned}$$

For $\sum_{k=1}^{M} \Lambda_k \|w_k^\top (x(t) - \Phi(x(t)))\|_2$, by Lemma Lemma C.10,

$$\begin{aligned}\sum_{k=1}^{M} \Lambda_k \|w_k^\top (x(t) - \Phi(x(t)))\|_2 &\geq \sqrt{\sum_{k=1}^{M} \Lambda_k^2 \|w_k^\top (x(t) - \Phi(x(t)))\|_2^2} \\ &= \sqrt{(x(t) - \Phi(x(t)))^\top \nabla^2 L(\Phi(x(t)))^2 (x(t) - \Phi(x(t)))} \\ &= \|\nabla^2 L(\Phi(x(t)))(x(t) - \Phi(x(t)))\|_2 \\ &\geq \mu \|\partial\Phi(\Phi(x(t)))(x(t) - \Phi(x(t)))\|_2 \\ &\geq \mu \|x(t) - \Phi(x(t))\|_2 - O(\|x(t) - \Phi(x(t))\|_2^2) \,.\end{aligned}$$

Concluding, we have that

$$\sum_{k=1}^{M} \Lambda_k s_k(t) w_k^\top (x(t) - \Phi(x(t))) \geq \mu \|x(t) - \Phi(x(t))\|_2 / 2 \,.$$

So

$$\begin{aligned}&\mathbb{E}[\|x(t+1) - \Phi(x(t+1))\|_2^2 \mid x(t)] \\ \leq &\|x(t) - \Phi(x(t))\|_2^2 - \frac{\mu\eta\rho}{M} \|x(t) - \Phi(x(t))\|_2 \\ &+ O(\eta\rho^2 \|x(t) - \Phi(x(t))\| + \eta^2\rho^2) \\ \leq &(\|x(t) - \Phi(x(t))\|_2 - C_2\eta\rho)^2 \,.\end{aligned}$$

The inequality holds if $\|x(t) - \Phi(x(t))\|_2 > C_1\eta\rho$.

Finally by Jenson's Inequality,

$$\mathbb{E}[\|x(t+1) - \Phi(x(t+1))\|_2 | x(t)] \leq \|x(t) - \Phi(x(t))\|_2 - C_2\eta\rho.$$

□

Lemma C.59. *Under condition of Theorem C.50, for any constant $C > 0$ independent of η, ρ, there exists constant $C_3 > 0$ independent of η, ρ, if $x(t) \in K^h$ and $\|x(t) - \Phi(x(t))\| \leq C\rho$, then we have that*

$$|\|x(t+1) - \Phi(x(t+1))\|_2 - \|x(t) - \Phi(x(t))\|_2| \leq C_3\eta\rho \,.$$

Proof. This is a direct application of Lemma C.13. □

Lemma C.60. *Under condition of Theorem C.50, assuming $x(t_0) \in K^{h/2}$ and $\|x(t_0) - \Phi(x(t_0))\| \leq f(\eta, \rho)$ for some fixed function f and $f(\eta, \rho) \in \Omega(\eta\rho \ln^2(1/\eta\rho)) \cap O(\rho)$, then with probability $1 - O(\rho)$, for any t satisfying $t_0 \leq t \leq t_0 + O(\ln(1/\eta\rho)/\eta)$, it holds that $\|x(t) - \Phi(x(t)\| \leq 2f(\eta, \rho)$. Moreover, we have that*

$$\|\Phi(x(t)) - \Phi(x(t_0))\| = O((\eta + \rho)\ln(1/\eta\rho)).$$

Proof. By Lemma C.55, we have that $x(t) \in K^h$ for any t satisfying that $t_0 \leq t \leq t_0 + O(\ln(1/\eta\rho)/\eta)$ and with probability $1 - O(\rho)$ we will suppose this hold for the following deduction.

By Uniform Bound,

$$\begin{aligned}
&\mathbb{P}(\exists t_0 \le t \le t_0 + O(\ln(1/\eta\rho)/\eta), \|x(t) - \Phi(x(t)\| \ge 2f(\eta,\rho)) \\
\le &\sum_{t=t_0}^{t_0+O(\ln(1/\eta\rho)/\eta)} \mathbb{P}(\|x(t) - \Phi(x(t))\| \ge 2f(\eta,\rho)) \quad \text{and} \quad \|x(\tau) - \Phi(x(\tau))\| \le 2f(\eta,\rho), \forall t_0 \le \tau \le t-1).
\end{aligned}$$

Consider each term and apply Uniform bound again,

$$\begin{aligned}
&\mathbb{P}(\|x(t) - \Phi(x(t))\| \ge 2f(\eta,\rho)) \quad \text{and} \quad \|x(\tau) - \Phi(x(\tau))\| \le 2f(\eta,\rho), \forall t_0 \le \tau \le t-1) \\
\le &\sum_{\tau=t_0}^{t} \mathbb{P}(\|x(t) - \Phi(x(t))\| \ge 2f(\eta,\rho)) \quad \text{and} \quad \|x(\tau) - \Phi(x(\tau))\| \le f(\eta,\rho), \\
&\text{and} \quad f(\eta,\rho) \le \|x(\tau') - \Phi(x(\tau'))\| \le 2f(\eta,\rho), \forall \tau+1 \le \tau' \le t-1).
\end{aligned}$$

Then if we consider each term, it is bounded by

$$\begin{aligned}
&\mathbb{P}(\|x(t) - \Phi(x(t))\| \ge 2f(\eta,\rho)) \\
&\quad \text{and} \quad f(\eta,\rho) \le \|x(\tau') - \Phi(x(\tau'))\| \le 2f(\eta,\rho), \forall \tau+1 \le \tau' \le t-1 \\
&\quad | \ \|x(\tau) - \Phi(x(\tau))\| \le f(\eta,\rho)).
\end{aligned} \tag{36}$$

Now let C be the positive constant satisfying $2f(\eta,\rho) \le C\rho$, suppose C_1, C_2 are the constants corresponds to C in Lemma C.58 and C_3 is the constant correspond to C in Lemma C.59. By definition $C_3 > C_2$.

Define a coupled process $\tilde{y}(\tau+1) = y(\tau+1)$ and

$$\tilde{y}(\tau') = \begin{cases} \|x(\tau') - \Phi(x(\tau'))\|_2, & \text{if } \tilde{y}(\tau'-1) = \|x(\tau'-1) - \Phi(x(\tau'-1))\|_2 > f(\eta,\rho) \\ \tilde{y}(\tau'-1) - C_2\eta\rho, & \text{if otherwise.} \end{cases}$$

Now clearly Equation (36) is bounded by $\mathbb{P}(\tilde{y}(t) \ge 2f(\eta,\rho))$.

As $\mathbb{E}[\tilde{y}(\tau')] \le \tilde{y}(\tau'-1) - C_2\eta\rho$ by Lemma C.58 and $\|\tilde{y}(\tau') - \tilde{y}(\tau'-1)\| \le C_3\eta\rho$ by Lemma C.59. This implies $\|\tilde{y}(\tau')\| - C_2\eta\rho\tau'$ is a super martingale. By Azuma-Hoeffding bound(Lemma C.68), we have

$$P(\tilde{y}(t) \ge \tilde{y}(\tau+1) - C_2\eta\rho(t-\tau-1) + h) \le 2\exp(-\frac{h^2}{4(t-\tau-1)(C_3+C_2)^2\eta^2\rho^2}).$$

Choosing $h = C_2\eta\rho(t-\tau-1) - \|x(\tau+1) - \Phi(x(\tau+1))\| + 2f(\eta,\rho)$

$$\begin{aligned}
&P(\tilde{y}(t+1) \ge 2f(\eta,\rho)) \\
\le &2\exp(-\frac{(C_2\eta\rho(t-\tau) - \|x(\tau+1) - \Phi(x(\tau+1))\| + 2f(\eta,\rho))^2}{8(t-\tau)(C_3+C_2)^2\eta^2\rho^2}) \\
\le &2\exp(-\frac{(C_2\eta\rho(t-\tau) + f(\eta,\rho)/2)^2}{4(t-\tau)(C_3+C_2)^2\eta^2\rho^2}) \\
\le &2\exp(-\frac{C_2 f(\eta,\rho)}{2(C_3+C_2)^2\eta\rho}) \le \eta^{10}\rho^{10}.
\end{aligned}$$

We then have

$$\mathbb{P}(\exists t_0 \le t \le t_0 + O(\ln(1/\eta\rho)/\eta), \|x(t) - \Phi(x(t)\| \ge 2f(\eta,\rho)) \le \rho.$$

□

Lemma C.61. *Under condition of Theorem C.50, assuming there exists t_{DEC} such that $x(t_{\mathrm{DEC}}) \in K^{h/2}$ and $\|\nabla L(x(t_{\mathrm{DEC}}))\| \le 4\zeta\rho$, then there exists $t_{\mathrm{DEC2}} = t_{\mathrm{DEC}} + O(\ln(1/\eta\rho)/\eta)$, such that $\|x(t_{\mathrm{DEC2}}) - \Phi(t_{\mathrm{DEC2}})\| \le O(\eta\rho)$.*

Furthermore, for any t satisfying $t_{\mathrm{DEC2}} \le t \le t_{\mathrm{DEC2}} + \Theta(\ln(1/\eta\rho)/\eta)$, we have that $\|\Phi(x(t)) - \Phi(x(t_{\mathrm{DEC}}))\| = O(\rho^2 \ln(1/\eta\rho))$.

Proof. We have that $x(t) \in K^h$ (Lemma C.55) and $\|x(t) - \Phi(x(t))\| \leq C\rho$ for some constant C (Lemma C.60) for any t satisfying that $t_{\mathrm{DEC}} \leq t \leq t_{\mathrm{DEC}} + O(\ln(1/\eta\rho)/\eta)$ with probability $1 - O(\rho)$ and we will suppose this holds for the following deduction. The second statement then follows directly from Lemma C.13.

Let C_1, C_2 be the constant in Lemma C.58 corresponding to C, For simplicity of writing, define $T_1 \triangleq \lceil \frac{C\ln(\frac{C}{C_1\eta\rho^2})}{C_2\eta} \rceil = O(\ln(1/\eta\rho)/\eta)$. Define indicator function as

$$\mathcal{A}(t) = \mathbf{1}[\|x(t) - \Phi(x(t))\| \geq C_1\eta\rho, \forall t \geq \tau \geq t_{\mathrm{GF}}]\,.$$

By Lemma C.58, we have that,

$$\begin{aligned}
\mathbb{E}[\|x(t+1) - \Phi(x(t+1))\|\mathcal{A}(t+1)] &\leq \mathbb{E}[\|x(t+1) - \Phi(x(t+1))\|\mathcal{A}(t)] \\
&\leq \mathbb{E}[\|x(t) - \Phi(x(t))\|\mathcal{A}(t)] - C_2\eta\rho\mathbb{E}[\mathcal{A}(t)] \\
&\leq \mathbb{E}[\|x(t) - \Phi(x(t))\|\mathcal{A}(t)](1 - \frac{C_2\eta}{C})
\end{aligned}$$

We can then conclude that with $T_2 = T_1 + t_{\mathrm{DEC}}$, using Lemma C.8,

$$\begin{aligned}
C_1\eta\rho\mathbb{E}\mathcal{A}(T_2+1) &\leq \mathbb{E}[\|x(T_2+1) - \Phi(x(T_2+1))\|_2\mathcal{A}(T_2+1)] \\
&\leq (1 - \frac{C_2\eta}{C})^{T_1}\|x(t_{\mathrm{DEC}}) - \Phi(x(t_{\mathrm{DEC}}))\| \leq C_1\eta\rho^3.
\end{aligned}$$

This implies $\mathcal{A}(T_2+1) = 0$ with probability $1 - O(\rho)$, which indicates the existence of t_{DEC2}. □

C.6.2 Phase II (Proof of Theorem C.51)

The goal of this section is to prove Theorem C.50.

Proof. We can inductively use Lemma C.63 and theorem D.5 to show that $x(t)$ tracks the Riemannian gradient flow and maintains in K^h with probability $1 - O(\sqrt{\eta} + \sqrt{\rho})$ and Lemma C.62 to show that $x(t)$ is in an $\tilde{O}(\eta\rho)$ neighbor of Γ for $T_3/\eta\rho^2$ step. □

Convergence Near Manifold

Lemma C.62. *Under condition of Theorem C.51, assuming $x(t) \in K^h, \forall t_0 \leq t \leq t_0 + O(1/\eta\rho^2)$ and $\|x(t_0) - \Phi(x(t_0))\| \leq f(\eta,\rho)$ for some fixed function f and $f(\eta,\rho) \in \Omega(\eta\rho\ln^2(1/\eta\rho)) \cap O(\rho)$, then with probability $1 - O(\eta\rho)$, for any t satisfying $t_0 \leq t \leq t_0 + O(1/\eta\rho^2)$, it holds that $\|x(t) - \Phi(x(t)\| \leq 2f(\eta,\rho)$.*

Proof. The proof is almost identical to Lemma C.60 and is omitted. □

Tracking Riemannian Gradient Flow

Lemma C.63. *Under the condition of Theorem C.51, for any t satisfying that $x(t) \in K^h$ and*

$$\|x(t) - \Phi(x(t))\| = O(\eta\rho\ln^2(1/\eta\rho)).$$

It holds that

$$\|\Phi(x(t+1)) - \Phi(x(t)) + \eta\rho^2 P^{\perp}_{\Phi(x(t)),\Gamma}\nabla\mathrm{Tr}\nabla^2 L(\Phi(x(t)))/2\| \leq \tilde{O}(\eta\rho^3 + \eta^2\rho^2)\,.$$

Proof. By Taylor Expansion,

$$\begin{aligned}
x(t+1) &= x(t) - \eta\nabla L_k\left(x(t) + \rho\frac{\nabla L_k(x(t))}{\|\nabla L_k(x(t))\|}\right) \\
&= x(t) - \eta\nabla L_k(x(t)) - \eta\rho\nabla^2 L_k(x(t))\frac{\nabla L_k(x(t))}{\|\nabla L_k(x(t))\|} \\
&\quad - \eta\rho^2\partial^2(\nabla L_k)[\frac{\nabla L_k(x(t))}{\|\nabla L_k(x(t))\|}, \frac{\nabla L_k(x(t))}{\|\nabla L_k(x(t))\|}]/2 + O(\eta\rho^3).
\end{aligned}$$

Now as$\|x(t)-\Phi(x(t))\|_2=\tilde{O}(\eta\rho)$, by Lemma C.13, it implies

$$\|x(t+1)-x(t)\|_2=O(\eta\rho)\,.$$

Then we have

$$\|\Phi(x(t+1))-\Phi(x(t))-\partial\Phi(x(t))(x(t+1)-x(t))\|_2\leq\xi\|x(t+1)-x(t)\|_2^2=O(\eta^2\rho^2).$$

Using Lemma C.11, we have

$$\|\eta\partial\Phi(x(t))\nabla L_k\left(x(t)\right)\|_2=O(\eta\|x(t)-\Phi(x(t))\|_2^2)=O(\eta^3\rho^2+\eta\rho^4),$$
$$\|\eta\rho\partial\Phi(x(t))\nabla^2L_k\left(x(t)\right)\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|}\|_2=O(\eta\rho\|x(t)-\Phi(x(t))\|_2)=\tilde{O}(\eta^2\rho^2+\eta\rho^3).$$

Hence

$$\|\Phi(x(t+1))-\Phi(x(t))+\eta\rho^2\partial\Phi(x(t))\partial^2(\nabla L_k)[\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|},\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|}]/2\|_2=\tilde{O}(\eta^2\rho^2+\eta\rho^3).$$

Notice finally that by Lemma C.57

$$\begin{aligned}&\partial\Phi(x(t))\partial^2(\nabla L_k)[\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|},\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|}]\\=&\partial\Phi(\Phi(x(t)))\partial^2(\nabla L_k)[w_k,w_k]+O(\|x(t)-\Phi(x(t))\|_2)\\=&P^{\perp}_{\Phi(x(t)),\Gamma}\nabla(\lambda_1(\nabla^2L_k(\Phi(x(t)))))+O(\|x(t)-\Phi(x(t))\|_2).\end{aligned}$$

Hence we have

$$\Phi(x(t+1))-\Phi(x(t))=-\eta\rho^2P^{\perp}_{\Phi(x(t)),\Gamma}\nabla(\lambda_1(\nabla^2L_k(\Phi(x(t)))))/2+\tilde{O}(\eta^2\rho^2+\eta\rho^3)$$

This completes the proof. □

C.6.3 Proofs of Corollaries 5.5 and 5.6

Proof of Corollary 5.5. We will do Taylor expansion on $\mathbb{E}_k[L^{\text{Max}}_{k,\rho}](x)$. By Theorem C.50 and C.51, we have $\|x(\lceil T_3/\eta\rho^2\rceil)-X(T_3)\|_2=O(\eta^{1/4}+\rho^{1/4})$ and $\|\Phi(x(\lceil T_3/\eta\rho^2\rceil))-x(\lceil T_3/\eta\rho^2\rceil)\|_2=O(\eta^{1/4}+\rho^{1/4})$. For convenience, we denote $x(\lceil T_3/\eta\rho^2\rceil)$ by x.

$$\mathbb{E}_k[R^{\text{Max}}_{k,\rho}](x)=\max_{\|v\|\leq1}\mathbb{E}_k[\rho v^\top\nabla L_k(x)+\rho^2v^\top\nabla^2L_k(x)v/2]+O(\rho^3)$$

Since $\max_{\|v\|\leq1}|v^\top\nabla L_k(x)|=O(\|x-\Phi(x)\|)=O(\eta^{1/4}+\rho^{1/4})$, it holds that,

$$\begin{aligned}\mathbb{E}_k[R^{\text{Max}}_{k,\rho}](x)&=\rho^2\mathbb{E}_k[\max_{\|v\|\leq1}v^\top\nabla^2L(x)v/2]+O\big((\eta^{1/4}+\rho^{1/4})\rho^2\big)\\&=\rho^2\mathbb{E}_k\max_{\|v\|\leq1}[v^\top\nabla^2L(X(T_3))v/2]+O\big((\eta^{1/4}+\rho^{1/4})\rho^2\big)\\&=\rho^2\text{Tr}(X(T_3))/2+O\big((\eta^{1/4}+\rho^{1/4})\rho^2\big)\end{aligned}$$

□

Proof of Corollary 5.6. We choose T_ϵ such that $X(T_\epsilon)$ is sufficiently close to $X(\infty)$, such that $\text{Tr}(X(T_\epsilon))\leq\text{Tr}(X(\infty))+\epsilon/2$. By corollary 5.5 (let $T_3=T_\epsilon$), we have for all ρ,η such that $(\eta+\rho)\ln(1/\eta\rho)$ is sufficiently small, $\|\mathbb{E}_k[R^{\text{Max}}_{k,\rho}](x(\lceil T_\epsilon/(\eta\rho^2)\rceil))-\rho^2\text{Tr}(X(T))/2\|_2\leq o(1)$. This further implies $\|\mathbb{E}_k[R^{\text{Max}}_{k,\rho}](x(\lceil T_\epsilon/(\eta\rho^2)\rceil))-\rho^2\text{Tr}(X(\infty))/2\|_2\leq\epsilon\rho^2/2+o(1)$. We also have $L(x(\lceil T_\epsilon/(\eta\rho^2)\rceil))-\inf_{x\in U'}L(x)=o(1)$. Then we can leverage Theorems 4.3 and 5.1 to get the desired bound. □

C.6.4 OTHER OMITTED PROOFS FOR 1-SAM

We will use $\ell'(y, y_k)$ and $\ell''(y, y_k)$ to denote $\frac{d\ell(y', y_k)}{dy'}|_{y'=y}$ and $\frac{d^2\ell(y', y_k)}{dy'^2}|_{y'=y}$.

Lemma C.64. *Under Setting C.1, fix $k \in [M]$, for any p satisfying $\ell(f_k(p), y_k) = 0$, we have that*

$$\nabla^2 L_k(p) = \ell''(f_k(p), y_k)\nabla f_k(p)(\nabla f_k(p))^\top .$$

Proof. $\ell(f_k(p), y_k) = 0$ implies $\ell'(f_k(p), y_k) = 0$. Then by Taylor Expansion,

$$\begin{aligned}\nabla^2 L_k(p) &= \nabla_p^2 \ell(f_k(p), y_k)\\ &= \partial_p[\ell'(f_k(p), y_k)\nabla f_k(p)]\\ &= \ell''(f_k(p), y_k)\nabla f_k(p)\nabla f_k(p)^\top + \ell'(f_k(p), y_k)\nabla^2 f_k(p)\\ &= \ell''(f_k(p), y_k)\nabla f_k(p)\nabla f_k(p)^\top .\end{aligned}$$

This concludes the proof. □

Proof of Lemma 5.3. By Lemma C.64, as $L(p) = \frac{1}{M}\sum_{k=1}^M L_k(p)$, we have

$$\nabla^2 L(p) = \frac{1}{M}\sum_{k=1}^M \frac{\partial^2 \ell(y', y_k)}{(\partial y')^2}|_{y'=f_k(x)}\nabla f_k(p)\nabla f_k(p)^\top .$$

By definition of Γ in Setting C.1, we have for any $p \in \Gamma$, $\{\nabla f_k(p)\}_{k=1}^n$ are linearly independent, which implies that $\nabla f_k(p) \neq 0$ for any $p \in \Gamma$.

For any $p \in \Gamma$, as $\frac{\nabla f_k(p)}{\|\nabla f_k(p)\|}$ is well defined and continuous at p, there exists a open ball V containing p such that for any $x \in V$, $\|\nabla f_k(x)\|_2 \geq C_1 > 0$ and $\|\nabla[\frac{\nabla f_k(x)}{\|\nabla f_k(x)\|}]\|_2 \leq C_2$ for some constants C_1 and C_2.

Suppose $\nabla L_k(x) \neq 0$, then as by Taylor Expansion,

$$\nabla L_k(x) = \ell'(f_k(x), y_k)\nabla f_k(x) .$$

We have $\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = \frac{\nabla f_k(x)}{\|\nabla f_k(x)\|} = \frac{\nabla f_k(p)}{\|\nabla f_k(p)\|} + C_2\|x - p\|$, which completes the proof. □

We note that the alignment result in Lemma 5.3 is not directly used in our proof. Instead, we use its generalized version Lemma C.57 which holds under holds under a more general condition than Setting C.1, namely Condition C.2.

C.7 TECHNICAL LEMMAS

Lemma C.65 (Corollary 4.3.15 in Horn & Johnson (2012))**.** *Let $\Sigma, \hat{\Sigma} \in R^{D\times D}$ be symmetric and non-negative with eigenvalues $\lambda_1 \geq ... \geq \lambda_D$ and $\hat{\lambda}_1 \geq ... \geq \hat{\lambda}_D$, then for any i,*

$$|\hat{\lambda}_i - \lambda_i| \leq \|\Sigma - \hat{\Sigma}\|_2$$

Definition C.66 (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 C.67. *[Davis-Kahan $\sin(\theta)$ theorem (Davis & Kahan, 1970)] Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ respectively. Fix $1 \leq r \leq s \leq p$, let $d \triangleq s - r + 1$ and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\Sigma v_j = \lambda_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\lambda}_j\hat{v}_j$ for $j = r, r+1, \ldots, s$. Define $\Delta \triangleq \min\left\{\max\{0, \lambda_s - \hat{\lambda}_{s+1}\}, \max\{0, \hat{\lambda}_{r-1} - \lambda_r\}\right\}$, where $\hat{\lambda}_0 \triangleq \infty$ and $\hat{\lambda}_{p+1} \triangleq -\infty$, we have for any unitary invariant norm $\|\cdot\|_*$,*

$$\Delta \cdot \|\sin\Theta(\hat{V}, V)\|_* \leq \|\hat{\Sigma} - \Sigma\|_* .$$

Here $\Theta(\hat{V}, V) \in \mathbb{R}^{d\times d}$*, with* $\Theta(\hat{V}, V)_{j,j} = \arccos \sigma_j$ *for any* $j \in [d]$ *and* $\Theta(\hat{V}, V)_{i,j} = 0$ *for all* $i \neq j \in [d]$*.* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ *denotes the singular values of* $\hat{V}^\top V$*.* $[\sin\Theta]_{ij}$ *is defined as* $\sin(\Theta_{ij})$*.*

Lemma C.68 (Azuma-Hoeffding Bound)**.** *Suppose* Z_n *is a super-martingale, suppose* $-\alpha \leq Z_{i+1} - Z_i \leq \beta$*, then for all* $n > 0, a > 0$*, we have*

$$P(|Z_n - Z_0| \geq a) \leq 2exp(-a^2/(2n(\alpha+\beta)^2))$$

Lemma C.69 (Magnus (1985))**.** *Let* $A : \mathbb{R}^D \to \mathbb{R}^{D\times D}$ *be any* $\mathcal{C}^1$ *symmetric matrix function and* $x^* \in \mathbb{R}^D$ *satisfying* $\lambda_1(A(x^*)) > \lambda_2(A(x^*))$ *and* v_1 *be the top eigenvector of* $A(x^*)$*. It holds that* $\nabla\lambda_1(A(x))|_{x=x^*} = \nabla(v_1^\top A(x)v_1)|_{x=x^*}$*.*

We then present some of the technical lemmas we required to prove Lemma C.24.

Lemma C.70. *If* $0 < c < \frac{b-a}{b^2}, a\sqrt{\frac{a^2+2b^2}{2(1-cb)}} \geq \frac{a^2+b^2}{2-ca-cb}$*, then* $a > \frac{1}{2}b, cb \leq \frac{1}{2}$

Proof. Notice that

$$ca\sqrt{\frac{a^2+b^2}{1-cb}} \geq ca\sqrt{\frac{a^2+2b^2}{2(1-cb)}} \geq \frac{cb^2+ca^2}{2-cb-ca} \geq \frac{cb^2+ca^2}{2-cb}.$$

So

$$\sqrt{1-cb} + \frac{1}{\sqrt{1-cb}} \geq \sqrt{1+\frac{b^2}{a^2}}$$

As $c < \frac{b-a}{b^2}$, we have $1 > 1 - cb > \frac{a}{b}$.

So

$$\sqrt{\frac{a}{b}} + \sqrt{\frac{b}{a}} \geq \sqrt{1+\frac{b^2}{a^2}}$$

The above inequality implies $a \geq \frac{1}{2}b$. As $c < \frac{b-a}{b^2}$,$cb \leq \frac{1}{2}$

□

Lemma C.71. *When* $0 < a < b, 0 < c < \frac{b-a}{b^2}$*, we have*

$$cb^2 + ca^2(2-cb-\frac{2}{3}ca) - (1-cb)\frac{c(a^2+b^2)}{2-ca-cb} - ca^2(\frac{1}{2}a^2+b^2)\frac{2-ca-cb}{(a^2+b^2)} \leq \frac{cb^2}{2-cb}$$

Proof. Equivalently, we are going to prove

$$(1-cb)b^2\left(\frac{1}{2-ca-cb} - \frac{1}{2-cb}\right) + a^2\frac{1-cb}{2-ca-cb} + a^2(\frac{1}{2}a^2+b^2)\frac{2-ca-cb}{(a^2+b^2)} \geq a^2(2-cb-\frac{2}{3}ca)$$

Further simplifying, we only need to prove

$$\frac{(1-cb)cab^2}{(2-cb)(2-ca-cb)} + a^2\frac{1-cb}{2-ca-cb} \geq \frac{1}{3}ca^3 + \frac{a^4}{2(a^2+b^2)}(2-ca-cb)$$

We have the following auxiliary inequalities,

$$(1-cb)b > a$$

$$\frac{1-cb}{2-ca-cb} \geq \frac{1}{\frac{a+b}{b} + \frac{b-a}{(1-cb)b}} \geq \frac{1}{\frac{a+b}{b} + \frac{b-a}{a}} = \frac{ab}{a^2+b^2} \geq \frac{a^2}{a^2+b^2}$$

$$\frac{1-cb}{a^2} \geq \frac{2-ca-cb}{a^2+b^2}$$

Using the above auxiliary inequalities we have

$$\begin{aligned}
&\frac{(1-cb)cab^2}{(2-cb)(2-ca-cb)}+a^2\frac{1-cb}{2-ca-cb}\geq\frac{1}{3}ca^3+\frac{a^4}{2(a^2+b^2)}(2-ca-cb)\\
\Leftarrow&\frac{ca^2b}{(2-cb)(2-ca-cb)}+\left(1-\frac{1}{2}(2-ca-cb)\right)\frac{a^2(1-cb)}{2-ca-cb}\geq\frac{1}{3}ca^3\\
\Leftarrow&\frac{ca^2b}{(2-cb)(2-ca-cb)}+\frac{ca^2(a+b)(1-cb)}{2(2-ca-cb)}\geq\frac{1}{3}ca^3\\
\Leftarrow&\frac{ca^2b}{(2-cb)(2-ca-cb)}+\frac{ca^2b(1-cb)}{2(2-ca-cb)}\geq\frac{1}{3}ca^2b\\
\Leftarrow&\frac{1}{(2-cb)^2}+\frac{1-cb}{2(2-cb)}\geq\frac{1}{3}\\
\Leftarrow&3(1-cb)(2-cb)+6\geq 2(2-cb)^2\\
\Leftarrow&(cb)^2-cb+4\geq 0
\end{aligned}$$

□

Lemma C.72. *When* $0<a<b, 0<c<\frac{b-a}{b^2}, a\sqrt{\frac{a^2+2b^2}{2(1-cb)}}\geq\frac{a^2+b^2}{2-ca-cb}$, *we have*

$$cb^2+ca^2(2-cb-\frac{2}{3}ca)-(1-cb)cb^2-ca^2(\frac{1}{2}a^2+b^2)\frac{1}{b^2}\leq\frac{cb^2}{2-cb}$$

Proof. Equivalently, we are going to prove,

$$\begin{aligned}
&cb^3+a^2(2-cb-\frac{2}{3}ca)\leq\frac{b^2}{2-cb}+\frac{a^2(\frac{1}{2}a^2+b^2)}{b^2}\\
\Longleftrightarrow\ &cb^3+a^2(1-cb-\frac{2}{3}ca)\leq\frac{b^2}{2-cb}+\frac{a^4}{2b^2}
\end{aligned}$$

We have the auxiliary inequality $\frac{1}{2-cb}>\frac{1}{2}+\frac{cb}{4}$.

Hence

$$\begin{aligned}
&cb^3+a^2(1-cb-\frac{2}{3}ca)\leq\frac{b^2}{2-cb}+\frac{a^4}{2b^2}\\
\Leftarrow&cb^3+a^2(1-cb-\frac{2}{3}ca)\leq\frac{b^2}{2}+\frac{a^4}{2b^2}+\frac{cb^3}{4}\\
\Leftarrow&c(\frac{3b^3}{4}-ba^2-\frac{2}{3}a^3)\leq\frac{b^2}{2}+\frac{a^4}{2b^2}-a^2
\end{aligned}$$

1 *Case 1,* If $\frac{3b^3}{4}-ba^2-\frac{2}{3}a^3\leq 0$, then

$$c(\frac{3b^3}{4}-ba^2-\frac{2}{3}a^3)\leq 0\leq\frac{b^2}{2}+\frac{a^4}{2b^2}-a^2$$

2 *Case 2,* If $\frac{3b^3}{4}-ba^2-\frac{2}{3}a^3>0$, then

$$\begin{aligned}
&c(\frac{3b^3}{4}-ba^2-\frac{2}{3}a^3)\leq\frac{b^2}{2}+\frac{a^4}{2b^2}-a^2\\
\Leftarrow&\frac{b-a}{b^2}(\frac{3b^3}{4}-ba^2-\frac{2}{3}a^3)\leq\frac{(b^2-a^2)^2}{2b^2}\\
\Leftarrow&2(\frac{3b^3}{4}-ba^2-\frac{2}{3}a^3)\leq(b-a)(b+a)^2\\
\Leftarrow&2(b^3-ba^2)-(b-a)(b+a)^2\leq\frac{b^3}{2}+\frac{4a^3}{3}\\
\Leftarrow&(b-a)(2b(a+b)-(a+b)^2)\leq\frac{b^3}{2}+\frac{4a^3}{3}\\
\Leftarrow&(b-a)^2(b+a)\leq\frac{b^3}{2}+\frac{4a^3}{3}
\end{aligned}$$

Using Lemma C.70,$a > \frac{b}{2}$,$(b-a)^2(b+a) = (b^2-a^2)(b-a) \leq b^2(b-a) \leq \frac{b^3}{2}$

□

Lemma C.73. *When* $0 \leq a \leq b, 0 \leq c \leq \frac{b-a}{b^2}, b^2 \geq a\sqrt{\frac{a^2+2b^2}{2(1-cb)}} \geq \frac{a^2+b^2}{2-ca-cb}$, *we have*

$$cb^2 + ca^2(2 - cb - \frac{2}{3}ca) - 2ca\sqrt{(b^2 + \frac{1}{2}a^2)(1-cb)} \leq \frac{cb^2}{2-cb}$$

Proof. Define

$$\begin{aligned}
F(a) &\triangleq a^2(2 - cb - \frac{2}{3}ca) - 2a\sqrt{(b^2 + \frac{1}{2}a^2)(1-cb)} \\
S_a(c,b) &\triangleq \{a | 0 \leq a \leq b, 0 < c \leq \frac{b-a}{b^2}, b^2 \geq a\sqrt{\frac{a^2+2b^2}{2(1-cb)}} \geq \frac{a^2+b^2}{2-ca-cb}\} \\
a_{min}(c,b) &\triangleq \inf S_a(c,b) \\
a_{max}(c,b) &\triangleq \sup S_a(c,b) \leq b - cb^2
\end{aligned}$$

Here we suppose WLOG $S_a(c,b) \neq \phi$.

Consider

$$\begin{aligned}
\frac{dF(a)}{da} &= 2a(2 - cb - \frac{2}{3}ca) - \frac{2}{3}ca^2 - 2\sqrt{(b^2 + \frac{1}{2}a^2)(1-cb)} - a^2\sqrt{\frac{1-cb}{b^2+\frac{1}{2}a^2}} \\
\frac{d^2F(a)}{da^2} &= 2(2 - cb - \frac{2}{3}ca) - \frac{4}{3}ca - \frac{4}{3}ca - a\sqrt{\frac{1-cb}{b^2+\frac{1}{2}a^2}} - 2a\sqrt{\frac{1-cb}{b^2+\frac{1}{2}a^2}} + \frac{a^3}{2(b^2+\frac{1}{2}a^2)^{\frac{3}{2}}}\sqrt{1-cb} \\
&\geq 4 - 2cb - 4ca - 3a\sqrt{\frac{1-cb}{b^2+\frac{1}{2}a^2}}
\end{aligned}$$

Define $u \triangleq cb, v \triangleq \frac{a}{b}$, then $u + v \leq 1$.

$$\begin{aligned}
\frac{d^2F(a)}{da^2} &\geq 4 - 2u - 4uv - 3\sqrt{1-u}\frac{1}{\sqrt{\frac{1}{2}+\frac{1}{v^2}}} \\
&\geq 4 - 2u - 4u(1-u) - 3\sqrt{1-u}\frac{1}{\sqrt{\frac{1}{2}+\frac{1}{(1-u)^2}}} \\
&\geq 4u^2 - 6u + 4 - 3\sqrt{1-u}\frac{(1-u)}{\sqrt{\frac{(1-u)^2}{2}+1}}
\end{aligned}$$

As $\sqrt{\frac{(1-u)^2}{2}+1} \geq \sqrt{\frac{(1-u)^2+1}{2}} \geq \sqrt{1-u}$,we have

$$\frac{d^2F(a)}{da^2} \geq 4u^2 - 6u + 4 - 3(1-u) = 4u^2 + 1 - 3u > 0$$

The above inequality shows that $F(a)$ is convex w.r.t to a for $a_{min}(c,b) \leq a \leq a_{max}(c,b)$.

Hence $F(a) \leq \max\left(F(a_{min}(c,b)), F(a_{max}(c,b))\right)$

Part 1 We abuse the notation and use a_{min} a shorthand for $a_{min}(c,b)$.

We have $a_{min}\sqrt{\frac{a_{min}^2+2b^2}{2(1-cb)}} = \frac{a_{min}^2+b^2}{2-ca_{min}-cb}$. This implies

$$2a_{min}\sqrt{(b^2+\frac{1}{2}a_{min}^2)(1-cb)} = (1-cb)\frac{(a_{min}^2+b^2)}{2-ca_{min}-cb} + a_{min}^2(\frac{1}{2}a_{min}^2+b^2)\frac{2-ca_{min}-cb}{(a_{min}^2+b^2)}$$

Hence using Lemma C.71,

$$\begin{aligned}F(a_{min}) &= a_{min}^2(2-cb-\frac{2}{3}ca_{min}) - (1-cb)\frac{c(a_{min}^2+b^2)}{2-ca_{min}-cb} - ca_{min}^2(\frac{1}{2}a_{min}^2+b^2)\frac{2-ca_{min}-cb}{(a_{min}^2+b^2)}\\ &\leq \frac{1}{c}(\frac{cb^2}{2-cb}-cb^2)\end{aligned}$$

Part 2 We abuse the notation and use a_{max} a shorthand for $a_{max}(c,b)$.
It's not easy to see which boundary condition a_{max} satisfy, hence we will discuss by cases.

1 *Case 1*,$a_{max}\sqrt{\frac{a_{max}^2+2b^2}{2(1-cb)}} = \frac{a_{max}^2+b^2}{2-ca_{max}-cb}$, in this case we simply redo the calculation in Part 1.

2 *Case 2*, $b^2 = a_{max}\sqrt{\frac{a_{max}^2+2b^2}{2(1-cb)}}$. This implies

$$2a_{max}\sqrt{(b^2+\frac{1}{2}a_{max}^2)(1-cb)} = (1-cb)b^2 + a_{max}^2(\frac{1}{2}a_{max}^2+b^2)\frac{1}{b^2}$$

Hence using Lemma C.72,

$$\begin{aligned}F(a_{max}) &= a_{max}^2(2-cb-\frac{2}{3}ca_{max}) - (1-cb)cb^2 - ca_{max}^2(\frac{1}{2}a_{max}^2+b^2)\frac{1}{b^2}\\ &\leq \frac{1}{c}(\frac{cb^2}{2-cb}-cb^2)\end{aligned}$$

Case 3 $cb^2 = b - a_{max}$ As $1-cb = \frac{a_{max}}{b}$ and $b^2 \geq a_{max}\sqrt{\frac{b(a_{max}^2+2b^2)}{2a_{max}}}$.
This implies $a_{max}^3 + 2a_{max}b^2 - 2b^3 \leq 0 \Rightarrow a_{max} < \frac{9}{10}b$.
Define $v \triangleq \frac{a_{max}}{b}$, $cb = 1-v$
Then by Lemma C.70, $\frac{1}{2} \leq v \leq \frac{9}{10}$

$$\begin{aligned}F(a_{max}) &= a_{max}^2(2-cb-\frac{2}{3}ca_{max}) - 2a_{max}\sqrt{(b^2+\frac{1}{2}a_{max}^2)(1-cb)}\\ &= b^2\left(v^2(2-(1-v)-\frac{2}{3}(1-v)v) - 2v\sqrt{(1+\frac{v^2}{2})v}\right)\end{aligned}$$

We will prove the following inequality,

$$v^2(2-(1-v)-\frac{2}{3}(1-v)v) - 2v\sqrt{(1+\frac{v^2}{2})v} \leq \frac{1}{2-cb} - 1 = \frac{-v}{1+v}$$

In fact we can directly show

$$\begin{aligned}& v^2(1+v) + \frac{v}{1+v} \leq 2v\sqrt{(1+\frac{v^2}{2})v}\\ \iff\ & v(1+v) + \frac{1}{1+v} \leq 2\sqrt{(1+\frac{v^2}{2})v}\end{aligned}$$

for $v \in [0.5, 0.9]$.

Hence we have

$$\begin{aligned}F(a) &\leq \max\left(F(a_{min}(c,b)), F(a_{max}(c,b))\right)\\ &\leq \frac{1}{c}(\frac{cb^2}{2-cb}-cb^2)\end{aligned}$$

□

D OMITTED PROOFS ON CONTINUOUS APPROXIMATION

Our main results Theorems 4.9 and 5.4 characterize the dynamics of full-batch SAM and 1-SAM using continuous approximations for both the first phase and second phase. In this section we provide the omitted details for such continuous approximation. Our analysis is mainly based on Li et al. (2019). Below we first restate their main theorem adapted to our setting.

Assumption D.1. *The function $b : \mathbb{R}^D \to \mathbb{R}^D$ is uniform Lipschitz, that is,* $\sup_{x,y\in\mathbb{R}^D} \frac{\|b(x)-b(y)\|_2}{\|x-y\|_2}$.

Definition D.2. *Let G denote the set of continuous functions $\mathbb{R}^D \to \mathbb{R}$ of at most polynomial growth, i.e. $g \in G$ if there exists positive integers $\kappa_1, \kappa_2 > 0$ such that*

$$|g(x)| \le \kappa_1(1 + |x|^{2\kappa_2}),$$

for all $x \in \mathbb{R}^d$. Moreover, for each integer $\alpha \ge 1$ we denote by G_α the set of α-times continuously differentiable functions $\mathbb{R}^D \to \mathbb{R}$ which, together with its partial derivatives up to and including order α, belong to G. Moreover, if g depends on additional parameters, we say $g \in G_\alpha$ if the constants κ_1, κ_2 are independent of these parameters, i.e., $g \in G_\alpha$ uniformly. Finally, the definition generalizes to vector-valued functions coordinate-wise.

Let $\psi_b(x,t), t \in \mathbb{R}, t \ge 0$ be the unique solution of

$$\psi_b(x,t) = x + \int_{s=0}^{t} b(\psi_b(x,t))ds, \tag{37}$$

and let $\overline{\psi}(x)$ be a random variable depending on x and $\overline{\psi}(x,k), k \in \mathbb{N} \triangleq (\overline{\psi})^k(x)$. Note here we abuse the notation and assume the randomness in application of $\overline{\psi}$ is independent. For example, $\overline{\phi}(x)$ could be used to denote the one-step update of 1-SAM, $\overline{\psi}(x,1) \equiv \overline{\psi}(x) = x - \eta\nabla L_k(x + \rho\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|_2})$ where k is independently sampled, and $\overline{\psi}(x,2)$ denote the (random) iterate of 1-SAM achieved after two steps of updates. In this section (including the following theorem), $\overline{\psi}$ depends on η and ρ, but we omit the explicit dependency on η and ρ for notational convenience.

Below is a modified version of Theorem 3 in Li et al. (2019), adapted to our setting. The main difference between the original version and our adaption is that our discrete update $\overline{\psi}(x)$ have two hyperparameters η and ρ, where the proper continuous time scaling depends on both hyperparameters. In detail, in Phase I, the target continuous dynamics is the gradient flow and in Phase II, the target continuous dynamics is the Riemmannien gradient flow on manifold. (See Table 2 for a summary.) Therefore, we use $p(\eta,\rho)$ to denote the amount of continuous time that one-step discrete update corresponds to, which is the η in Li et al. (2019). We further use $q(\eta,\rho)$ to denote the total error. Though the original theorem statement in Li et al. (2019) only applies to the case $q(\eta,\rho) = p^\alpha(\eta,\rho)$, their proof indeed extends to the more general case, $p^\alpha(\eta,\rho) = O(q(\eta,\rho))$, as stated below.

Theorem D.3 (Adapted version of Theorem 3, Li et al. (2019))**.** *Let $T > 0$, $\alpha \ge 1$ be an integer,$p, q : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ be two functions satisfying that $p(\eta,\rho) = O(\eta+\rho)$, $\Omega(q(\eta,\rho)) = p^\alpha(\eta,\rho) = o(1)$, and $N = \lfloor T/p(\eta,\rho)\rfloor$. Let $\Delta(x) \triangleq \overline{\psi}(x) - x$ and $\widetilde{\Delta}(x) \triangleq \psi_b(x, p(\eta,\rho)) - x$. Under Assumption D.1, suppose further that the following conditions holds:*

(1) There exists a $K_1 \in G$ independent of η, ρ such that

$$\left|\mathbb{E}\prod_{j=1}^{s}\Delta_{(i_j)}(x) - \prod_{j=1}^{s}\widetilde{\Delta}_{(i_j)}(x)\right| \le K_1(x)p(\eta,\rho)q(\eta,\rho),$$

for $s = 1, 2, \ldots, \alpha$ and

$$\mathbb{E}\prod_{j=1}^{\alpha+1}\left|\Delta_{(i_j)}(x)\right| \le K_1(x)p^{\alpha+1}(\eta,\rho),$$

for all $i_j \in \{1, \ldots, D\}$.

(2) For each $m \geq 1$, the $2m$-moment of $\overline{\psi}(x,k)$ is uniformly bounded with respect to k, η and ρ, i.e., there exists $K_2 \in G$, independent of η, ρ, k, such that

$$\mathbb{E}|\overline{\psi}(x,k)|^{2m} \leq K_2(x), \tag{38}$$

for all $k = 0, \ldots, N \equiv \lfloor T/p(\eta,\rho) \rfloor$.

Then for each $g \in G^{\alpha+1}$, there exists a constant $C > 0$, independent of η, ρ, such that

$$\max_{k=0,\ldots,\lfloor T/p(\eta,\rho)\rfloor} |\mathbb{E}(g(\overline{\psi}(x,k))) - g(\psi_b(x, k \cdot p(\eta,\rho)))| \leq Cq(\eta,\rho).$$

	p	q	b
Lemma C.33	η	$\rho+\eta$	$-\nabla L(\cdot)$
Theorem C.32	$\eta\rho^2$	$\rho+\eta$	$-\partial\Phi(\cdot)\nabla\lambda_1(\nabla^2 L(\cdot))/2$
Lemma C.52	η	$\rho+\eta$	$-\nabla L(\cdot)$
Theorem C.51	$\eta\rho^2$	$\rho+\eta$	$-\partial\Phi(\cdot)\nabla\mathrm{Tr}(\nabla^2 L(\cdot))/2$

Table 2: Summary of applications of Theorem D.5 in our analysis

Theorem D.4 (Lemma 29, Li et al. (2019))**.** *Suppose there is a function $p : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ and a constant $c > 0$ such that $\left\|\overline{\psi}(x) - x\right\|_2 \leq C(1+|x|)p(\eta,\rho)$ for all $x \in \mathbb{R}^D$, then for any fixed $T > 0$ and any $m \geq 1$, $\mathbb{E}|x_k|^m$ exists and is uniformly bounded in η and $k = 0, \ldots, N = \lfloor N/p(\eta,\rho) \rfloor$.*

Combining Theorems D.3 and D.4, we have the following theorem:

Theorem D.5. *Let $T > 0$, $p, q : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ be two functions satisfying that $p(\eta,\rho) = O(\eta+\rho)$, $q(\eta,\rho) = \Omega(p(\eta,\rho)) = o(1)$, and $N = \lfloor T/p(\eta,\rho) \rfloor$. Let $\Delta(x) \triangleq \overline{\psi}(x) - x$ and $\widetilde{\Delta}(x) \triangleq \psi_b(x, p(\eta,\rho)) - x$. Suppose there is an open neighborhood V of $\{\psi_b(x,t)\}_{t=0}^T$ that b is uniform Lipschitz in V and $\frac{\|\Delta(x)\|_2}{p(\eta,\rho)}$ is bounded in V. Further suppose that the following conditions holds: There exists a constant K independent of η, ρ such that for all $x \in V$,*

$$\left\|\mathbb{E}\Delta(x) - \widetilde{\Delta}(x)\right\| \leq Kp(\eta,\rho)q(\eta,\rho).$$

Then there exists a constant $C > 0$, independent of η, ρ, such that for each $k = 0, \ldots, \lfloor T/p(\eta,\rho) \rfloor$,

$$\mathbb{P}\left[\left\|\overline{\psi}(x,k) - \psi_b(x, k \cdot p(\eta,\rho))\right\|_2 \geq C\sqrt{q(\eta,\rho)}\right] \leq \sqrt{p(\eta,\rho)}.$$

Moreover, when $\overline{\psi}(x)$ is deterministic, there exists a constant $C > 0$, independent of η, ρ, such that

$$\max_{k=0,\ldots,\lfloor T/p(\eta,\rho)\rfloor} \left\|\overline{\psi}(x,k) - \psi_b(x, k \cdot p(\eta,\rho))\right\|_2 \leq Cq(\eta,\rho).$$

Proof of Theorem D.5. The proof invokes Theorem D.3 with $\alpha = 1$ and $g_t(x) = \|x - \psi_b(x,t)\|_2^2$ for each $0 \leq t \leq T$and $g^i(x) = x_i$ for each $i \in [M]$.

Since there is an open neighborhood V of $\{\psi_b(x,t)\}_{t=0}^T$ that b is uniform Lipschitz in V, without loss of generality, we can assume that b is uniform Lipschitz in $\mathbb{R}^D$, because by classical results approximation results, $\left\|\overline{\psi}(x,k) - \psi_b(x, kp(\eta,\rho))\right\|_2$ will be larger than constants for at most exponential small probability and thus we can just modify b outside V such that it is uniform Liptshitz. This will not affect our final high probability bounds. By Theorem D.4, we know that the condition (2) of Theorem D.3 holds.

For the same reason,, we can assume $\left\|\mathbb{E}\Delta(x) - \widetilde{\Delta}(x)\right\| \leq Kp(\eta,\rho)q(\eta,\rho)$ holds for each $x \in \mathbb{R}^D$, otherwise we can modify $\overline{\psi}(x)$ outside V. We can also assume $\frac{\|\Delta(x)\|_2}{p(\eta,\rho)}$ is not only bounded in V, but in the entire $\mathbb{R}^D$. Since we set $\alpha = 1$, it holds that

$$\mathbb{E}\prod_{j=1}^{\alpha+1} \left|\Delta_{(i_j)}(x)\right| \leq O(p^{\alpha+1}(\eta,\rho)),$$

for all $i_j \in \{1, \ldots, D\}$.

Now by Theorem D.3, choose $g(x) = \|x - \psi_b(x, kp(\eta, \rho))\|_2^2$ for each k, we know Then there exists a constant $C > 0$, independent of η, ρ, such that

$$\mathbb{E} \left\| \overline{\psi}(x, k) - \psi_b(x, k \cdot p(\eta, \rho)) \right\|_2^2 \leq Cq(\eta, \rho).$$

By Markov inequality, we know that the probability of $\left\| \overline{\psi}(x, k) - \psi_b(x, k \cdot p(\eta, \rho)) \right\|_2 \geq C\sqrt{q(\eta, \rho)}$ is at most $\sqrt{q(\eta, \rho)}$.

When $\overline{\psi}(x)$ is deterministic, setting $g^i = x_i$ for each $i \in [D]$, we know that there exists a constant $C > 0$, independent of η, ρ, such that

$$\max_{k=0,\ldots,\lfloor T/p(\eta,\rho) \rfloor} \left\| \overline{\psi}(x, k) - \psi_b(x, k \cdot p(\eta, \rho)) \right\|_2 \leq Cq(\eta, \rho).$$

□