OPT2022: 14th Annual Workshop on Optimization for Machine Learning

How Does Sharpness-Aware Minimization Minimizes Sharpness?

Kaiyue Wen WENKY20@MAILS.TSINGHUA.EDU.CN *Tsinghua University*
Tengyu Ma TENGYUMA@STANFORD.EDU *Stanford University*
Zhiyuan Li ZHIYUANLI@STANFORD.EDU *Stanford University*

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 another notion of sharpness, which is most likely to be the preferred notion for practical performance. The key mechanism behind this intriguing phenomenon is the implicit alignment between the gradient and the top eigenvector of Hessian when running SAM.

1. Introduction

Modern deep nets are often overparametrized and have the capacity to fit even randomly labeled data [24]. 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 [6, 7, 13, 14, 21]. Partly motivated by these studies, Foret et al. [9], Wu et al. [23], Zheng et al. [26] 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. [9] 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 $R_\rho^{\max}(x) = L_\rho^{\max}(x) - L(x)$ the *worst-direction sharpness* at x and thus SAM is intended to minimize the original training loss plus the worst-direction sharpness at x.

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

Type of Sharpness	Symbol	Definition	Limiting Regularizers Among Minimizers
Worst-direction	$L_\rho^{\max}$	$\max_{\|v\|_2\le 1} L(x+\rho v)$	$\lambda_1(\nabla^2 L(x))/2$ (Theorem 12)
Ascent-direction	L_ρ^{asc}	$L\left(x+\rho\frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right)$	$\lambda_{\min}(\nabla^2 L(x))/2$(Theorem 13)
Average-direction	L_ρ^{avg}	$\mathbb{E}_{v\sim N(0,I)} L(x+\rho\frac{v}{\|v\|_2})$	$\mathrm{Tr}(\nabla^2 L(x))/2D$ (Theorem 14)

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

However, even evaluation of $L_\rho^{\max}(x)$ is computationally expensive, not to mention optimization. Thus [9, 26] proposed to approximate the worst perturbation direction by the direction of the gradient and implement the second notion of sharpness, which corresponds to (2). We call $R_\rho^{\mathrm{asc}}(x) = L_\rho^{\mathrm{asc}}(x) - L(x)$ the *ascent-direction sharpness* at x.

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

Intriguingly, in the generalization analysis of SAM by [9, 23], 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 [20]. We call the third notion $R_\rho^{\mathrm{avg}}(x) = L_\rho^{\mathrm{avg}}(x) - L(x)$ the *average-direction sharpness* at x, where $L_\rho^{\mathrm{avg}}(x) = \mathbb{E}_{g\sim N(0,I)} L\left(x+\rho g/\|g\|\right)$.

For further acceleration, Foret et al. [9], Zheng et al. [26] 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.*, $\nabla L_\rho^{\mathrm{asc}}(x) \approx \nabla L\left(x+\rho\frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right)$ 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 [8, 16] In particular, we make the following theoretical contributions.

1. We prove that full-batch SAM does minimize worst-direction sharpness. (Theorem 8)
2. Surprisingly, when batch size is 1, SAM minimizes average-direction sharpness. (Theorem 11)
3. We characterize the explicit biases of three notions of sharpness among minimizers when perturbation radius ρ goes to zero. (Theorems 12 to 14, 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 an alignment phenomenon between the gradient and the top eigenvector of Hessian when running SAM.

2. 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 $P_{x,\Gamma}^\top = I_D - P_{x,\Gamma}$. 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 Arora et al. [2], Fehrman et al. [8], Li et al. [16], we make the below assumption.

Assumption 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 met 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. Let $U = \{x \in \mathbb{R}^D | \Phi(x) \text{ exists and } \Phi(x) \in \Gamma\}$. Assumption 1 implies that U is open and Φ is in $\mathcal{C}^3$ on U (from Lemma B.15 [2]).

3. Explicit and Implicit Bias in the Full-batch Setting

Section 3.1 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 3.2 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 local minimizers. In Appendix C.1 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.

3.1. Worst- and ascent-direction sharpness have different explicit bias

The intuition of approximating $R_\rho^{\max}$ by R_ρ^{asc} comes from the following Taylor expansions [9, 23].

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

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

For most of points x with non-zero gradient, their leading terms are both the first order term and are the same, since $\sup_{\|v\|_2\leq 1} v^\top \nabla L(x) = \|\nabla L(x)\|_2$. Unfortunately, the first order term vanishes when we actually try to minimize the regularized objective, *i.e.*, the sharpness-aware loss L_ρ^{asc} or $L_\rho^{\max}$, because every minimizer of the original loss has zero gradient. When one attempts to optimize the regularized loss, the original loss must first be optimized, meaning the first order term goes away and the first-order approximation becomes trivial. A quick way to see this is that any global minimizer of the original loss L will kill the first order term and is a $O(\rho^2)$-approximate minimizer of the sharpness-aware loss. In order to allow the regularizer to actually "regularize" the learning algorithm, the goal must be at least reaching $O(\rho^2)$ error, and what really matters is indeed the second order term.

In this section, we aim to understand under Assumption 1 what the explicit biases of various notions of sharpness among different minimizers are. Theorem 4 will be our main theoretic tool to analyze the explicit bias for small perturbation radius ρ.

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 $R^{\mathrm{asc}}_\rho = L^{\mathrm{asc}}_\rho - L$ 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.*, $R^{\mathrm{avg}}_\rho(x) \approx \frac{\rho^2}{2D}\mathrm{Tr}[\nabla^2 L(x)]$. Therefore we make the following assumption to allow the regularizer to be mildly negative.

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

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

Definition 3 (Limiting Regularizer) *We define the* limiting regularizer *of $\{R_\rho\}$ as the function*

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

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 $\forall x \in \Gamma \cap V, 0 < \rho \leq \rho_{x^*}, \left| S(x) - \inf_{\|x'-x\|_2 \leq \epsilon \cdot \rho} R_\rho(x')/\rho^2 \right| < \epsilon$. 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 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 *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 2 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*

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

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 also admits the following alternative form which doesn't involve R_ρ. Corollary 5 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 5 *Under the setting of Theorem 4, 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 5 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.

Theorem 6 (Summary of Theorem 13,12 and 14) *$R^{\text{asc}}_\rho, R^{\max}_\rho, R^{\text{avg}}_\rho$ satisfy Assumption 2 and admit good limiting regularizers on Γ. (see Table 1)*

Using Theorem 6, we can apply Corollary 5 to characterize their explicit biases, which are all different.

3.2. 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.

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

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 8, which is the main theorem of this section.

Assumption 7 *For $x \in \Gamma$, there exists a positive eigengap,i.e., $\lambda_1(\nabla^2 L(x)) > \lambda_2(\nabla^2 L(x))$.*

Theorem 8 (Main, Theorems 46 and 47 stated informally) *Let $\{x(t)\}$ be the iterates of SAM (3) with $x(0) = x_{init} \in U$, then under Assumptions 1 and 7, 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 $O(\rho)$.*

Theorem 8 shows that SAM decreases the largest eigenvalue of Hessian of loss locally around the manifold of local minimizers.

4. 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 [9]. Towards explaining this phenomenon, Foret et al. [9] argues intuitively that stochastic SAM minimizes stochastic worst-direction sharpness.

In this section, we focus on SGD with batch size 1. We still need Assumption 1 in this section. We first by analyzing the explicit bias of the stochastic ascent- and worst-direction sharpness in

Section 4.1 via the tools developed in Section 3.1. It turns out they are all proportional to the trace of hessian as $\rho \to 0$. In Section 4.2, 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 $l = 1, \dots, M$. 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 $y \in R$.

4.1. Stochastic worst- and ascent-direction sharpness have same explicit bias

Below we specify *stochastic worst-direction sharpness* and *stochastic ascent-direction sharpness* as $\mathbb{E}_k[R^{\max}_{k,\rho}] = \mathbb{E}_k[L^{\max}_{k,\rho}] - L$ and $\mathbb{E}_k[R^{\text{asc}}_{k,\rho}] = \mathbb{E}_k[L^{\text{asc}}_{k,\rho}] - L$. 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 9 *Stochastic worst-direction sharpness $\mathbb{E}_k[R^{\max}_{k,\rho}]$ satisfies Assumption 2 and admits $\text{Tr}(\nabla^2 L(\cdot))/2$ as a good limiting regularizer on Γ.*

Theorem 10 *Stochastic ascent-direction sharpness $\mathbb{E}_k[R^{\text{asc}}_{k,\rho}]$ satisfies Assumption 2 and admits $\text{Tr}(\nabla^2 L(\cdot))/2$ as a good limiting regularizer on Γ.*

4.2. 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 $[M]$.

$$x(t+1) = x(t) - \eta \nabla L_{k_t}\left(x + \rho \frac{\nabla L_{k_t}(x)}{\|\nabla L_{k_t}(x)\|_2}\right), \tag{7}$$

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

$$X(\tau) = X(0) - \frac{1}{2}\int_{s=0}^{\tau} P^{\top}_{X(s),\Gamma} \nabla \text{Tr}(X(s))ds, X(0) = \Phi(x_{\text{init}}). \tag{8}$$

Theorem 11 *Let $\{x(t)\}$ be the iterates defined by SAM (7) and $x(0) = x_{init} \in U$, then under Assumption 1, for all η and ρ such that $(\eta + \rho)\log(1/\eta\rho)$ is sufficiently small, 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, with probability at least $1 - O(\sqrt{\rho})$;*
- ***Phase II:*** *Stochastic SAM tracks the solution X of (8), the Riemannian Gradient Flow with respect to $\text{Tr}(\nabla^2 L(\cdot))$, with time scaling as $\|\Phi(x(\lceil T_3/(\eta\rho^2)\rceil)) - X(T_3)\| = O((\eta+\rho)\log(1/\eta\rho)$.*

5. 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.

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

References

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

[2] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 948–1024. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/arora22a.html.

[3] 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.

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

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

[6] 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*, pages 1019–1028. JMLR. org, 2017.

[7] 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.

[8] 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.

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

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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*, pages 5905–5914. PMLR, 2021.

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

[17] 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*, pages 12360–12370, 2022.

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

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

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

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

[22] 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.

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

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

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

[26] 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*, pages 8156–8165, 2021.

Contents

Appendix A. Related Works

Sharpness and Generalization. The study on the connection between sharpness and generalization can be traced back to [10]. Keskar et al. [14] 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. [12] extends this by finding a correlation of the sharpness and the ratio between learning rate to batch size. Dinh et al. [6] shows that one can easily construct networks with good generalization but with arbitrary large sharpness by reparametrization. Dziugaite and Roy [7], Neyshabur et al. [21], Wei and Ma [22] give theoretical guarantees on the generalization error using sharpness-related measures. [13] performs 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. [9], Zheng et al. [26] concurrently proposed to minimize the loss at the perturbed from current parameter towards the worst direction to improve generalization. Wu et al. [23] proposed the almost identical method for a different purpose, robust generalization of adversarial training. [15] proposed a different metric for SAM to fix the rescaling problem pointed out by [6]. Liu et al. [17] proposed an more computationally efficient version of SAM. Zhao et al. [25] proposed to improve generalization by penalizing gradient norm. Their proposed algorithm can be viewed as a generalization of SAM. Andriushchenko and Flammarion [1] 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.

Sharpness Minimization as Implicit Bias. Recent theoretical works Blanc et al. [3], Damian et al. [4], Li et al. [16] showed that SGD with label noise is implicitly biased to local minimizers with a smaller trace of Hessian. Arora et al. [2] showed that normalized GD implicitly decreases the largest eigenvalue of the Hessian. Lyu et al. [18] 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.

Appendix B. Discussion and Implication of main result

B.1. Limiting Regularizer

Theorem 12 *Worst-direction sharpness $R_\rho^{\max}$ satisfies Assumption 2 and admits $\lambda_1(\nabla^2 L(\cdot))/2$ as a good limiting regularizer on Γ.*

Theorem 13 *Ascent-direction sharpness R_ρ^{asc} satisfies Assumption 2 and admits $\lambda_M(\nabla^2 L(\cdot))/2$ as a good limiting regularizer on Γ.*

Theorem 14 *Average-direction sharpness R_ρ^{avg} satisfies Assumption 2 and admits $\mathrm{Tr}(\nabla^2 L(\cdot))/(2D)$ 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.

The analysis for ascent-direction sharpness is more tricky as $R_\rho^{\mathrm{asc}}(x) = \infty$ 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 $R_\rho^{\mathrm{asc}}(x') \approx \rho^2/2 \cdot \nabla L(x')^\top \nabla^2 L(x) \nabla L(x')/\|\nabla L(x')\|^2$. Here the crucial step of the proof is that because of Assumption 1, $\nabla L(x)/\left\|\nabla L(x)\right\|_2$ 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 1 ensures this property extends to general losses.

Unlike in the full-batch setting where the implicit regularizer of ascent-direction sharpness and worst-direction sharpness have different explicit bias, in the stochastic case 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 regularizer of the average regularizers by definition.

B.2. Full-batch setting

As a corollary of Theorem 8, we can also show that the largest eigenvalue of the limiting flow closely tracks the regularized training loss.

Corollary 15 *For all $T_3' > 0$, for all ρ, η such that $\eta \ln(1/\rho)$ and ρ/η are sufficiently small we have $\forall T_3' < \eta\rho^2 t \le T_3$, $\|R_\rho^{\max}(x(t)) - \rho^2 \lambda_1(X(\eta\rho^2 t))/2\| = \tilde{O}(\eta\rho^2)$*

Recall Theorem 13 shows that the largest eigenvalue of Hessian is the limiting regularizer of the *worst-direction sharpness*, leveraging the equivalence relationship in Theorem 4, 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 16 *Define U' as in Theorem 4, 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 $L_\rho^{\max}(x(\lceil T/(\eta\rho^2)\rceil)) \le \epsilon\rho^2 + \inf_{x\in U'} L_\rho^{\max}(x)$*

B.3. Stochastic Setting

Corollary 17 and 18 below are stochastic counterparts of Corollary 15 and 16, 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 17 *For all $T_3' > 0$, for all ρ, η such that $(\eta+\rho)\log(1/\eta\rho)$ are sufficiently small we have $\forall T_3' \le \eta\rho^2 t \le T_3$, $\|\mathbb{E}_k[R^{\max}_{k,\rho}](x(t)) - \rho^2 \mathrm{Tr}(\nabla^2 L(X(\eta\rho^2 t)))/2\| = \tilde{O}((\eta+\rho)\rho^2)$.*

Corollary 18 *Define U' as in Theorem 4, suppose $X(\infty) = \lim_{t\to\infty} X(t)$ exists and is a minimizer of $\mathrm{Tr}(\nabla^2 L(x))$ in $U' \cap \Gamma$, for all $\epsilon > 0$, there exists a constant $T > 0$, then for all ρ, η such that $(\eta + \rho)\log(1/\eta\rho)$ are sufficiently small we have $\mathbb{E}_k[L^{\max}_{k,\rho}](\lceil T/(\eta\rho^2)\rceil)) \le \epsilon\rho^2 + \inf_{x\in U'} \mathbb{E}_k[L^{\max}_{k,\rho}](x)$*

Appendix C. Proof overview

C.1. Proof Sketch for Phase II in Theorem 8

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. [2], Lyu et al. [18], 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. [2], Lyu et al. [18], 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 35) Therefore we have that

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

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

$$\begin{aligned}&\nabla L\left(x + \rho\frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right)\\ =&\nabla L(x) - \rho\nabla^2 L(x)\frac{\nabla L(x)}{\|\nabla L(x)\|_2} + \frac{\rho^2}{2}\partial^2(\nabla L)(x)\left[\frac{\nabla L(x)}{\|\nabla L(x)\|_2}, \frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right] + O(\rho^3).\end{aligned} \tag{10}$$

Lemma 19 (Lemma B.16 [2]) *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 (10) into (9) and simplify the expression using Lemma 19. We have that

$$\begin{aligned}\Phi(x(t+1)) - \Phi(x(t)) =& 0 + O(\rho\left\|\partial\Phi(x(t))\nabla^2 L(x(t))\right\|\|\nabla L(x(t))\|_2) \\ &-\frac{\eta\rho^2}{2}\partial\Phi(x)\partial^2(\nabla L)(x)\left[\frac{\nabla L(x)}{\|\nabla L(x)\|_2}, \frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right] + O(\eta^3\rho^2 + \eta\rho^3)\end{aligned} \tag{11}$$

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 30).

Lemma 20 *For $x \in \Gamma$, $\partial\Phi(x) = P_{x,\Gamma}^{\top}$, $\partial\Phi(x)\nabla^2 L(x) = 0$, where $P_{x,\Gamma}^{\top}$ is the orthogonal projection matrix of the tangent space of Γ at x.*

Next we use Lemma 20 to further simplify (11). 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

$$\begin{aligned}
\Phi(x(t+1)) - \Phi(x(t)) &= -\frac{\eta\rho^2}{2}\partial\Phi(x)\partial^2(\nabla L)(x)\left[\frac{\nabla L(x)}{\|\nabla L(x)\|_2}, \frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right] + O(\eta^3\rho^2 + \eta\rho^3) \\
&= -\frac{\eta\rho^2}{2}\partial\Phi(x)\partial^2(\nabla L)(x)\left[v_1(\nabla^2 L(x)), v_1(\nabla^2 L(x))\right] + O(\eta^3\rho^2 + \eta\rho^3) \\
&= -\frac{\eta\rho^2}{2}\partial\Phi(x(t))\nabla\lambda_1(\nabla^2 L(x(t))) + O(\eta^3\rho^2 + \eta\rho^3) \\
&= -\frac{\eta\rho^2}{2}\partial\Phi(\Phi(x(t)))\cdot\nabla\lambda_1(\nabla^2 L(x))|_{x=\Phi(x(t))} + O(\eta^3\rho^2 + \eta\rho^3), \quad (12)
\end{aligned}$$

where the second to last step we use the following property about the derivative of eigenvalue (Lemma 57) 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 1. Below we first give the formal statement of the implicit alignment on quadratic loss Theorem 21. 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_ρ^{asc}, because optimizing L_ρ^{asc} would rather make the gradient align to the smallest eigenvector!

Theorem 21 *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 (13), 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 $\eta\lambda_1(A) < 1$.*

$$x(t+1) = x(t) - \eta A\left(x(t) + \rho\frac{Ax(t)}{\|Ax(t)\|}\right), \tag{13}$$

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

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

and the result of Theorem 21 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):

$$x(t+1) = x(t) - \eta\nabla L(x(t)) - \eta\rho\nabla^2 L(x(t))\frac{\nabla L(x(t))}{\|\nabla L(x(t))\|_2} + O(\eta\rho^2). \tag{15}$$

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 (15) and Taylor expansion on $\nabla L(x(t+1))$ at $x(t)$, we have that

$$\nabla L(x(t+1)) = \nabla L(x(t)) - \nabla^2 L(x(t))\,(x(t+1)-x(t)) + O(\eta^2\rho^2) \tag{16}$$

$$= \nabla L(x(t)) - \eta(\nabla^2 L(x(t)))^2 \left(\nabla L(x(t)) + \rho\frac{\nabla L(x(t))}{\|\nabla L(x(t)))\|}\right) + O(\eta\rho^2) \tag{17}$$

Equation (16) 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.

C.2. Proof Sketch for Stochastic Case

Given our results in Section 3, it's natural to ask if we can justify 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 4.2).

The fundamental difference between stochastic SAM (7) 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 22 formally states this property. For convenience, we denote $\frac{d^2\ell(y',y_k)}{d^2y'}|_{y'=f_k(x)}\nabla f_k(x)/\|\nabla f_k(x)\|$ as $\Lambda_k(x), w_k(x)$.

Lemma 22 *Under Assumption 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 $\exists s \in \{1,-1\}\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = sw_k(p) + O(\|x-p\|)$*

With Lemma 22, we can show $\Phi(x(t+1)) - \Phi(x(t)) = -\eta\rho^2 P_{t,\Gamma}^\top \nabla(\Lambda_k(x))/2 + \tilde{O}(\eta^2\rho^2)$ 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 3.2. As we have $\mathbb{E}[\Lambda_k(x)] = (\sum_k \Lambda_k(x))/M = \mathrm{Tr}(\nabla^2 L(x))$, 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 51.

Appendix D. Proof Details

D.1. Setup

We will first restate our main assumptions in Section 2.

Assumption 23 *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 $1 \le M \le D$, where for all $x \in \Gamma$, x is a local minimizer of L, $L(x) = 0$ and rank$(\nabla^2 L(x)) = M$.*

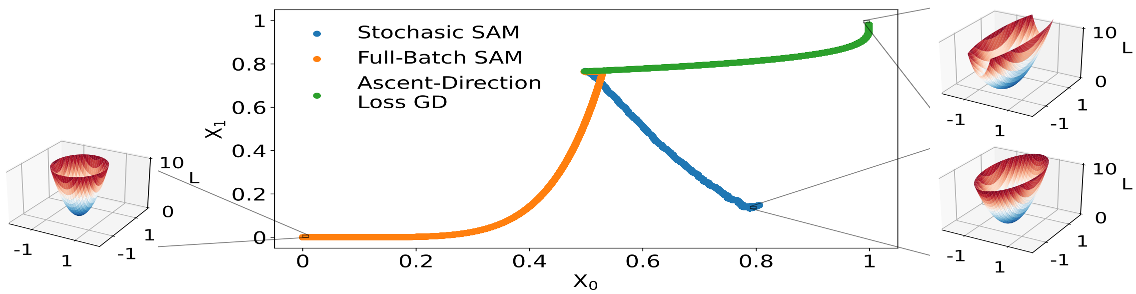

Figure 1: Visualization of the different biases of different sharpness notions on a 4D-toy example. For $x, y \in \mathbb{R}^2$, consider loss $L(x,y) = F_0(x)y_0^2 + F_1(x)y_1^2$ with $F_0(x) = x_0^2 + 6x_1^2 + 8$ and $F_1(x) = 4(1-x_0)^2 + (1-x_1)^2 + 1$. L has a zero loss manifold $\{y = 0\}$ and 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 (7) (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)

Assumption 24 *For $x \in \Gamma$, there exists a positive eigengap,i.e., $\lambda_1(x) > \lambda_2(x)$.*

For stochastic loss, we use notation in Assumption 49, a general assumption containing our setup in Theorem 11. We additionally define Φ_k as the gradient flow with respect to L_k.

We abuse the notation and 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 $\lambda_i(t) := \lambda_i(x(t)), v_i(t) := v_i(x(t)), P_{t,\Gamma}^\top := P_{\Phi(x(t)),\Gamma}^\top, P_{t,\Gamma} := P_{\Phi(x(t)),\Gamma}$.

In our proof, we repeatedly discuss compact set in Γ and their neighborhoods. We will use $K \subset \Gamma$ to denote a compact set. This notation may have different meanings in different proof and will be clearly stated. We further define $K^d = \{x | \mathrm{Dist}(x, K) \leq d\}$.

We first present some lemmas regularizing the behavior of L and Γ near C.

Definition 25 *A function L is $\mu-$PL in a set U iff $\forall x \in U$, $\|\nabla L(x)\|^2 \geq 2\mu(L(x) - \inf_{x \in U} L(x))$.*

Definition 26 *The spectral 2-norm of a k-order tensor $\Gamma_{i_1,\ldots,i_k} \in R^{d_1 \times \ldots \times d_k}$ is defined as*

$$\|\Gamma\| = \max_{\|x_i\| \in R^{d_i}, \|x_i\|=1} \Gamma[x_1, \ldots, x_k].$$

Lemma 27 *Given C, there is a sufficiently small $r(C) > 0$ such that*

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

We further assume

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

and all these constants are greater than 1. We also abuse the notation and use notations like ζ_k to denote the same norm for stochastic loss.

Lemma 28 *Given C, there is a sufficiently small $h(C) > 0$ such that*

1. $\sup_{x\in K^h} L(x) - \inf_{x\in K^h} L(x) \leq \frac{\mu\rho^2}{8}$
2. $\forall x \in K^h, \Phi(x) \in K^{r/2}$

Then we have

Lemma 29 *For any $x \in K^h$, we have*

1. *The entire gradient flow trajectory with respect to every stochastic loss L_k and L lies in K^r.*
2. *The whole segment $\overline{x\Phi(x)}$ and $\overline{x\Phi_k(x)}$ lies in K^r.*

The proof of these lemmas can be found in [2].

D.2. Lemma about Φ

In this section, we will introduce some geometric lemma about SAM, which will be heavily used in the analysis below.

Lemma 30 *For $x \in K^h$, we have*

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

Lemma 31 *For $x \in K^h$, we have*

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

Lemma 32 *At any point $x \in K^h$, we have*

$$\|\nabla L\left(x\right) - \nabla^2 L\left(\Phi(x)\right)\left(x - \Phi(x)\right)\| \leq \frac{\nu}{2}\|x - \Phi(x)\|^2$$
$$\left|\frac{\|\nabla L\left(x\right)\|}{\|\nabla^2 L\left(\Phi(x)\right)\left(x - \Phi(x)\right)\|} - 1\right| \leq \frac{2\nu}{\mu}\|x - \Phi(x)\|$$
$$\frac{\nabla L\left(x\right)}{\|\nabla L\left(x\right)\|} = \frac{\nabla^2 L\left(\Phi(x)\right)\left(x - \Phi(x)\right)}{\|\nabla^2 L\left(\Phi(x)\right)\left(x - \Phi(x)\right)\|} + O(\frac{\nu}{\mu}\|x - \Phi(x)\|)$$

The proof of above lemmas can be found in [2].

Lemma 33 *For $x \in K^h$,*

$$\|\partial\Phi(x)\nabla L_k(x)\| \leq (\nu_k + \zeta_k\xi)\|x - \Phi(x)\|^2$$
$$\|\partial\Phi(x)\nabla^2 L_k(x)\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|}\| \leq (\nu_k + \zeta_k\xi)\|x - \Phi(x)\|$$

Proof Consider doing Taylor expansion,

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

and

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

Here we use Lemma 28 to ensure the approximation is correct. ∎

Lemma 34 *There exists $h_1 > 0$, for $x \in K^h$ and $p \in C, \nabla^2 L_k(p) = \Lambda_k(p)w_k(p)w_k(p)^T$, suppose $\|x - p\| < h_1$, there exists $s \in \{1, -1\}$,*

$$\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = sw_k(p) + O(\|x - p\|)$$

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

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

Proof There are two ways we may use to estimate the direction $\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|}$.
First Way According to Lemma 32,

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

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

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

Define $\nabla^2 L_k(O) = v_i v_i^T$, using Davis-Kahan Theorem 55, we would have $\exists s \in \{-1, 1\}$, such that $\|w_k(\Phi_k(x)) - s w_k(p)\| \leq \zeta \|\Phi_k(x) - p\|$

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

According to Lemma 30, we have $\|x - \Phi_k(x)\| \leq \frac{\|\nabla L_k(x)\|}{\mu} \leq \frac{\zeta\|x-p\|}{\mu}$. This implies,

$$\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = s w_k(p) + O(\|x - p\|) \tag{18}$$

Second Way There is another direct way to consider the direction of $\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|}$. Doing a Taylor expansion at O,

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

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

$$\begin{aligned} \big| \|\nabla L_k(x)\| - \|w_k w_k^T(x - p)\| \big| &\leq O(\|x - p\|^2) \\ \frac{\big| \|\nabla L_k(x)\| - \|w_k w_k^T(x - p)\| \big|}{\|w_k w_k^T(x - p)\|} &\leq O(\frac{\|x - p\|^2}{\|w_k w_k^T(x - p)\|}) = O(\|x - p\|^{1/2}) \end{aligned}$$

Hence we have

$$\frac{\nabla L_k(x)}{\|\nabla L_k(x)\|} = sign(w_k^T(x - p)) w_k + O(\|x - p\|^{1/2}) \tag{19}$$

Comparing (18) and (19), we have there exists h_1, such that if $\|x-p\| \leq h_1$, $s = sign(w_k^T(x-p))$ when $|w_k^T(x - p)| \geq \|x - p\|^{3/2}$. The final inequality is self-explanatory. ■

We will abuse notation slightly and suppose h_1 in Lemma 34 satisfies $h_1 \geq h$.

Lemma 35 *Suppose $x \in K^h$ and $y = x - \eta \nabla L\left(x + \rho \frac{\nabla L(x)}{\|\nabla L(x)\|}\right)$,*

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

Proof Using Taylor Expansion and Lemma 28,

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

Further

$$\begin{aligned} y - x &= -\eta \nabla L\left(x + \rho \frac{\nabla L(x)}{\|\nabla L(x)\|}\right) \\ &= -\eta \nabla L(x) - \eta \rho \nabla^2 L(x) \frac{\nabla L(x)}{\|\nabla L(x)\|} - \eta \rho^2 \partial \nabla^2 L(x)[\frac{\nabla L(x)}{\|\nabla L(x)\|}, \frac{\nabla L(x)}{\|\nabla L(x)\|}]/2 + O(\eta \rho^3 \Upsilon) \\ \Rightarrow & \|y - x + \eta \nabla L(x) + \eta \rho \nabla^2 L(x) \frac{\nabla L(x)}{\|\nabla L(x)\|} + \eta \rho^2 \partial \nabla^2 L(x)[\frac{\nabla L(x)}{\|\nabla L(x)\|}, \frac{\nabla L(x)}{\|\nabla L(x)\|}]/2\| \leq \eta \rho^3 \Upsilon \end{aligned}$$

This implies

$$\|y - x\| = \eta \|\nabla L \left(x + \rho \frac{\nabla L\left(x\right)}{\|\nabla L\left(x\right)\|} \right) \| \leq \eta \|\nabla L\left(x\right)\| + \eta \zeta \rho$$
$$\|\partial\Phi(x)(y-x)\| \leq \eta \|\partial\Phi(x)\nabla L\left(x\right) + \rho \partial\Phi(x)\nabla^2 L(x) \frac{\nabla L\left(x\right)}{\|\nabla L\left(x\right)\|}\| + \eta\rho^2\xi$$

Using Lemma 31,

$$\|\partial\Phi(x)(y-x)\| \leq \eta\rho\|\nabla L\left(x\right)\| \|\partial^2\Phi(x)\left[\frac{\nabla L\left(x\right)}{\|\nabla L\left(x\right)\|}, \frac{\nabla L\left(x\right)}{\|\nabla L\left(x\right)\|}\right]\| + \eta\rho^2\xi$$
$$\leq \xi\eta\rho\|\nabla L\left(x\right)\| + \eta\rho^2\xi$$

Putting together we have

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

■

Lemmas in this section will be repeatedly used in our proofs.

D.3. Explicit Bias

We will first prove the generic Theorem 4 and then apply this theorems to characterize a variety of regularizers.

Lemma 36 *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 $x \in U', \mathrm{dist}(x, \Gamma) \leq \rho_0 \Rightarrow \mathrm{dist}(x, \overline{U' \cap \Gamma}) \leq h_2$*

Proof 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 $\mathrm{dist}(x_k, \Gamma) \leq \rho_k$ and $\mathrm{dist}(x_k, \overline{U' \cap \Gamma}) \geq h_2$. Then consider accumulation point of x_k x^* in $\overline{U'}$. Then we would have $x^* \in \Gamma \cap \overline{U'} = \overline{U' \cap \Gamma}$. ■

Lemma 37 *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 $x \in U', L(x) \leq \rho_1 \Rightarrow \mathrm{dist}(x, \overline{U' \cap \Gamma}) \leq h_2$*

Proof 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 $L(x_k) \leq \rho_k$ and $\mathrm{dist}(x_k, \overline{U' \cap \Gamma}) \geq h_2$. Then consider accumulation point of x_k x^* in $\overline{U'}$. Then we would have $x^* \in \Gamma \cap \overline{U'} = \overline{U' \cap \Gamma}$. ■

Proof [Proof of Theorem 4] We will first prove $\Rightarrow$ side. The proof consists of four steps. Define h as the constant in Lemma 29 with $K = \overline{U' \cap H}$.

Step 1 Consider $x_0 \in U' \cap \Gamma$, satisfying

$$S(x_0) \le \inf_{x \in U' \cap \Gamma} S(x) + O(\rho).$$

Then using definition 3, we have there exists sufficiently small $\epsilon_\rho < \rho^2$, such that exists $\|x_1 - x_0\| \le \epsilon_\rho < \rho^2$ and $\|S(x_0) - R_\rho(x_1)/\rho^2\| = O(\rho^2)$. Now that as $x_0 \in U'$, we have for sufficiently small ρ, $x_1 \in U'$. By Lemma 36, we have for sufficiently small ρ, $\overline{x_0 x_1} \in \overline{U' \cap H}^h$, then consider doing Taylor Expansion, we would have $L(x_1) + R_\rho(x_1) \le \rho^2 S(x_0) + o(\rho^2)$.

Step 2 Now consider $L(u) + R_\rho(u) \le \rho^2 \inf_{x \in U' \cap \Gamma} S(x) + \epsilon\rho^2 + o(\rho^2)$. We easily have $L(u) \le O(\rho^2)$ by Assumption 2. By Lemma 37, we have $\overline{u\Phi(u)} \in K^h$ and that $\|u - \Phi(u)\| = O(\rho)$ by Lemma 30. This implies $\mathrm{dist}(\overline{U' \cap H}, \Phi(u)) = o(1)$. Notice now, we have shown $R_\rho(u) \le \rho^2 \inf S(x) + \rho^2\epsilon + o(\rho^2)$

Step 3 Now applying definition of a good limiting regularizer and Finite covering theorem, we have this further implies $S(\Phi(u))\rho^2 \le R_\rho(u) + o(\rho^2)$. We also have $S(\Phi(u)) \ge \inf_{x \in U' \cap \Gamma} S(x) - o(1)$ as $S \in \mathcal{C}^0$. Finally $R_\rho(u) \ge S(\Phi(u))\rho^2 - o(\rho^2) \ge \rho^2 \inf S(x) - o(\rho^2)$. We also have $L(u) \le o(\rho^2)$, proving our claim.

For the $\Leftarrow$ side, we need to provide a lower bound for $L(x) + R_\rho(x)$. Define constant C_1 such that for all x, $\|x - \Phi(x)\| \ge C_1\rho$, we have $L(x) \ge (C_{U'} + \inf_{x \in U' \cap \Gamma} S_(x) + 1)\rho^2$ where $C_{U'}$ is the constant in Assumption 2. Then we have for x such that $\|x - \Phi(x)\| \ge C_1\rho$, we have $L(x) + R_\rho(x) \ge (\inf_{x \in U' \cap \Gamma} S_(x) + 1)\rho^2$. For $\|x - \Phi(x)\| \le C_1\rho$, we have by definition of good limiting regularizer and Finite covering theorem, $R_\rho(x) \ge \rho^2 S(\Phi(x)) - o(\rho^2)$. As $\mathrm{dist}\Phi(x), \overline{U' \cap \Gamma} = o(1)$, we have $R_\rho(x) \ge \inf_{x \in U' \cap \Gamma} S(x) - o(\rho^2)$, hence $L(x) + R_\rho(x) \ge \inf_{x \in U' \cap \Gamma} S_(x)\rho^2$. ■

Theorem 4 implies Corollary 38 saying that the minimum of the regularized loss is approximately the sum of the minimum of loss and the minimum of limiting regularizer.

Corollary 38 *Under the setting of Theorem 4,*

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

Proof [Proof of Corollary 38] In Theorem 4, choose $\epsilon = 0$, and choose u such that $L(u) + R_\rho(u) \le \inf_{x \in U'}(L(x) + R_\rho(x)) + o(\rho^2)$, the result is then clear. ■

We will then prove the Corollary 5.

Proof [Proof of Corollary 5] By Theorem 4, 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 $\rho^2 S(\Phi(x)) - o(\rho^2) \le R_\rho(u) \le \rho^2 \inf_{x \in U' \cap H} S(x)$. 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 [Proof of Theorem 12]

Step 1 For assumption 2

$$\begin{aligned} R_\rho^{\max}(p) &= \max_{\|v\|_2 \le 1} L(p + \rho v) - L(p) \ge \max_{\|v\|_2 \le 1} \left(\rho\langle \nabla L(p), v\rangle + \rho^2 v^T \nabla^2 L(p) v/2\right) - \Upsilon\rho^3 \\ &\ge -\Upsilon\rho^3 \ge -C\rho^2 \end{aligned}$$

Step 2 For definition of good limiting regularizer, by Davis-Kahan Theorem and assumption 1, $S(x) = \lambda_1(x)/2$ is non-negative and continuous on Γ.

We also have

$$\lim_{\rho\to 0}\lim_{r\to 0}\inf_{\|x'-x\|\le r}\frac{R_\rho^{\max}(x')}{\rho^2} = \lim_{\rho\to 0}\frac{R_\rho^{\max}(x)}{\rho^2} = \lambda_1(\nabla^2 L(x))/2$$

Further for $x^* \in \Gamma$, consider a sufficiently small open set V containing x^* in which $\|\nabla^3 L\|$ is bounded, for $x \in V\cap\Gamma$, for $\|x'-x\| \le C\rho$, we have

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

On the other hand

$$R_\rho^{\max}(x) = \max_{\|v\|_2\le 1} L(x+\rho v) - L(x) \ge \rho^2\lambda_1(\nabla^2 L(x))/2 - \Upsilon\rho^3$$

■

Proof [Proof of Theorem 13]

Step 1 For assumption 2

$$\begin{aligned}R_\rho^{\max}(p) &= L(p+\rho\frac{\nabla L(p)}{\|\nabla L(p)\|}) - L(p)\\ &\ge \left(\rho\|\nabla L(p)\| + \rho^2(\frac{\nabla L(p)}{\|\nabla L(p)\|})^T\nabla^2 L(p)\frac{\nabla L(p)}{\|\nabla L(p)\|}/2\right) - \Upsilon\rho^3\\ &\ge -C\rho^2\end{aligned}$$

This constant is by taking minimizer of $\lambda_1(\nabla^2 L)$ over $\overline{U'}$.

Step 2 For definition of good limiting regularizer, by Davis-Kahan Theoremm 55 and assumption 1, $S(x) = \lambda_M(x)/2$ is non-negative and continuous on Γ.

Further for $x^* \in \Gamma$, consider a sufficiently small open set V containing x^* in which $\|\nabla^3 L\|$ is bounded, for $x \in V\cap\Gamma$, for $\|x'-x\| \le C\rho$, then we easily have $\|x' - \Phi(x)\| \le O(\rho)$, we have

$$\begin{aligned}R_\rho^{\max}(x') &= L(x'+\rho\frac{\nabla L(x')}{\|\nabla L(x')\|}) - L(x')\\ &\ge \rho\|\nabla L(x)\| + \rho^2(\frac{\nabla L(x')}{\|\nabla L(x')\|})^T\nabla^2 L(x)\frac{\nabla L(x)}{\|\nabla L(x)\|}/2 - O(\rho^3)\\ &\ge \rho^2(\frac{\nabla L(x)}{\|\nabla L(x)\|})^T\nabla^2 L(\Phi(x))\frac{\nabla L(x)}{\|\nabla L(x)\|}/2 - O(\rho^3)\end{aligned}$$

Using Lemma 32 and Theorem 55, we have

$$\begin{aligned}R_\rho^{\max}(x') &\ge \rho^2\lambda_M(\Phi(x)) - O(\rho^3)\\ &\ge \rho^2\lambda_M(u) - O(\rho^3)\end{aligned}$$

We also have

$$\lim_{\rho\to 0}\lim_{r\to 0}\inf_{\|x'-x\|\le r}\frac{R^{\mathrm{asc}}_\rho(x')}{\rho^2}=\lim_{\rho\to 0}\lim_{r\to 0}\inf_{\|x'-x\|\le r}\frac{\rho^2\lambda_M(\Phi(x'))}{2\rho^2}=\lambda_M(\nabla^2 L(x))/2$$

Combining we have $S(x)$ is a good limiting regularizer of $R(x)$.

■

Proof [Proof of Theorem 14]

Step 1 For assumption 2

$$\begin{aligned}R^{\mathrm{avg}}_\rho(p)&=\mathbb{E}_{g\sim N(0,I)}L(p+g/\|g\|)-L(p)\\&\ge\left(\rho^2(g/\|g\|)^T\nabla^2 L(p)g/\|g\|\right)/2-\Upsilon\rho^3\\&\ge -C\rho^2\end{aligned}$$

This constant is by taking minimizer of $\mathrm{Tr}(\nabla^2 L)$ over $\overline{U'}$.

Step 2 For definition of good limiting regularizer, by Davis-Kahan Theoremm 55 and assumption 1, $S(x)=\mathrm{Tr}(x)/(2D)$ is non-negative and continuous on Γ.

Further for $x^*\in\Gamma$, consider a sufficiently small open set V containing x^* in which $\|\nabla^3 L\|$ is bounded, for $x\in V\cap\Gamma$, for $\|x'-x\|\le C\rho$, we have

$$\begin{aligned}R^{\max}_\rho(x')&=\mathbb{E}_{g\sim N(0,I)}L(x'+g/\|g\|)-L(x')\\&\ge\left(\rho^2(g/\|g\|)^T\nabla^2 L(x')g/\|g\|\right)/2-O(\rho^3)\\&\ge\rho^2\mathrm{Tr}(\nabla^2 L(x'))/2D-O(\rho^3)\\&\ge\rho^2\mathrm{Tr}(\nabla^2 L(x))/2D-O(\rho^3)\end{aligned}$$

We also have

$$\lim_{\rho\to 0}\lim_{r\to 0}\inf_{\|x'-x\|\le r}\frac{R^{\mathrm{avg}}_\rho(x')}{\rho^2}=\lim_{\rho\to 0}\frac{R^{\mathrm{avg}}_\rho(x)}{\rho^2}=\mathrm{Tr}(\nabla^2 L(x))/2D$$

Combining we have $S(x)$ is a good limiting regularizer of $R(x)$.

■

Proof [Proof of Theorem 9]

By Theorem 50, Assumption 49 holds.

Easily deducted from Theorem 12 $\Lambda_k(x)$ is a good limiting regularizer for $R^{\max}_{k,\rho}$ on Γ_k. Then as $\Gamma\subset\Gamma_k$, $\Lambda_k(x)$ is a good limiting regularizer for $R^{\max}_{k,\rho}$ on Γ. Hence $S(x)=\sum_k\Lambda_k(x)/2M=\mathrm{Tr}(\nabla^2\mathrm{L(x)})/2$ is a good limiting regularizer of $\mathbb{E}_k[R^{\max}_{k,\rho}](x)$ on Γ.

■

Proof [Proof of Theorem 10]

By Theorem 50, Assumption 49 holds.

Easily deducted from Theorem 13 $\Lambda_k(x)$ is a good limiting regularizer for $R^{\mathrm{asc}}_{k,\rho}$ 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 $R^{\max}_{k,\rho}$ on Γ.Hence $S(x)=\sum_k\Lambda_k(x)/2M=\mathrm{Tr}(\nabla^2\mathrm{L(x)})/2$ is a good limiting regularizer of $\mathbb{E}_k[R^{\mathrm{asc}}_{k,\rho}](x)$ on Γ. ■

D.4. Full-batch SAM on Quadratic Loss: Proof of Theorem 21

We first simplify the iterate as

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

Define $\tilde{x}(t) = \frac{Ax(t)}{\rho}$. We have

$$\tilde{x}(t+1) = \tilde{x}(t) - \eta A \tilde{x}(t) - \eta \frac{A^2 \tilde{x}(t)}{\|\tilde{x}(t)\|} \tag{20}$$

Our proof consists of three steps

(1) *Preparation Phase* $\exists T_1, \forall t > T_1, \|P^{(j:D)}\tilde{x}(t)\| \le \eta\lambda_j^2$

(2) *Alignment Phase* Define $S_1 = \{t | \|\tilde{x}(t)\| \le \frac{\eta\lambda_1^2}{2-\eta\lambda_1}, t > T_1\}$, suppose $t, t' \in S_1, t \le t'$, then $|\tilde{x}_1(t)| < |\tilde{x}_1(t')|$ (Lemma 45) and we have $t \in S_1$ or $t+1 \in S_1$ for $t \ge T_1$(Lemma 43).

(3) *Length Convergence* $\|\tilde{x}(t)\|$ will converge to $\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$

D.4.1. PREPARATION PHASE

We define $\mathbb{I}_j = \{\tilde{x} | \|P^{(j:D)}\tilde{x}\| \le \eta\lambda_j^2\}$ and we will prove the following two lemmas. Lemma 39 will show this is an invariant set for update rule 20 and Lemma 40 will show that all vectors not in this set will shrink exponentially in norm.

Lemma 39 *If $\tilde{x}(t) \in \mathbb{I}_j$, using update rule 20, $\tilde{x}(t+1) \in \mathbb{I}_j$*

Proof We have by update rule 20,

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

Hence

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

.

We have $\|\tilde{x}(t)\| \le \frac{\eta\lambda_j^2}{1-\eta\lambda_j}$ by assumption,

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

Then we have $\|I - P^{(j:D)}\eta A - \eta \frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|}\| \le \max(1, \eta\lambda_j + \eta \frac{\lambda_j^2}{\|P^{(j:D)}\tilde{x}(t)\|} - 1)$
Hence

$$\begin{aligned}
\|P^{(j:D)}\tilde{x}(t+1)\| &\le \max(\|P^{(j:D)}\tilde{x}(t)\|, \eta\lambda_j^2 - (1-\eta\lambda_j)\|P^{(j:D)}\tilde{x}(t)\|) \\
&\le \eta\lambda_j^2
\end{aligned}$$

.

Here the last equation use $1 - \eta\lambda_j \geq 0$. We have by definition $\tilde{x}(t+1) \in \mathbb{I}_j$

∎

Lemma 40 *If $\tilde{x}(t) \notin \mathbb{I}_j$, then $\|P^{(j:D)}\tilde{x}(t+1)\| \leq \max(1 - \eta\lambda_D, \eta\lambda_j)\|P^{(j:D)}\tilde{x}(t)\|$*

Proof

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

.

We have $\|\tilde{x}(t)\| \geq \|P^{(j:D)}\tilde{x}(t)\| > \eta\lambda_j^2$, hence $\eta\frac{P^{(j:D)}A^2}{\|\tilde{x}(t)\|} \prec \eta\frac{P^{(j:D)}A^2}{\eta\lambda_j^2} \prec I$
This implies

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

Hence we have

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

. ∎

Lemma 41 *Choosing $T_1 = \max_j\left(-\log_{\max(1-\eta\lambda_D, \eta\lambda_j)}\max(\frac{\|\tilde{x}(0)\|}{\eta\lambda_j^2}, 1)\right)$, then $\forall t \geq T_1, D > j \geq 1, \tilde{x}(t) \in \mathbb{I}_j$*

Proof Proof by contradiction, suppose $\exists j, T > T_1, \tilde{x}(T) \notin \mathbb{I}_j$.
By Lemma 39, $\forall t < T, \tilde{x}(t) \notin \mathbb{I}_j$.
Then by Lemma 40,

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

which is a contradiction.

∎

D.4.2. ALIGNMENT PHASE

Define $\theta(t) = \arccos(\frac{|\langle\tilde{x}(t), e_1\rangle|}{\|\tilde{x}(t)\|})$, $\tilde{x}_i(t) = \langle\tilde{x}, e_i\rangle$
We will first show the following lemma.

Lemma 42 *When $\|\tilde{x}(t)\| \leq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, using update rule 20*

$$|\tilde{x}_1(t+1)| > |\tilde{x}_1(t)| \ \& \ \cos(\tilde{x}(t+1)) \geq \cos(\tilde{x}(t))$$

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

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

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

On the other hand, $\|P^{(2:D)}\tilde{x}(t+1)\| \le \max\left(|1-\eta\lambda_2-\eta\frac{\lambda_2^2}{\|\tilde{x}(t)\|}|, |1-\eta\lambda_D-\eta\frac{\lambda_D^2}{\|\tilde{x}(t)\|}|\right)\|P^{(2:D)}\tilde{x}(t)\|$

Notice that

$$1-\eta\lambda_1-\eta\frac{\lambda_1^2}{\|\tilde{x}(t)\|} < 1-\eta\lambda_2-\eta\frac{\lambda_2^2}{\|\tilde{x}(t)\|} \le 1-\eta\lambda_D-\eta\frac{\lambda_D^2}{\|\tilde{x}(t)\|} \le 1-\eta\lambda_D < 1 < \eta\lambda_1+\eta\frac{\lambda_1^2}{\|\tilde{x}(t)\|} - 1$$

Hence $\max\left(|1-\eta\lambda_2-\eta\frac{\lambda_2^2}{\|\tilde{x}(t)\|}|, |1-\eta\lambda_D-\eta\frac{\lambda_D^2}{\|\tilde{x}(t)\|}|\right) < |1-\eta\lambda_1-\eta\frac{\lambda_1^2}{\|\tilde{x}(t)\|}|$. ■

Lemma 43 *$\|\tilde{x}(t)\| > \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, $\tilde{x}(t) \in \cap\mathbb{I}_j$, then using update rule 20,*

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

Proof

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

Consider the following three cases.

Case 1 $\forall i, \left|(1-\eta\lambda_1)\|\tilde{x}(t)\|-\eta\lambda_1^2\right| \ge \left|(1-\eta\lambda_i)\|\tilde{x}(t)\|-\eta\lambda_i^2\right|$

In this case, we have $\|\tilde{x}(t+1)\| \le \left|(1-\eta\lambda_1)\|\tilde{x}(t)\|-\eta\lambda_1^2\right| = \eta\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|$

A more detailed analysis would show $\|\tilde{x}(t+1)\|$ is upper bounded by

$$\sqrt{\left(\eta\lambda_1^2-(1-\eta\lambda_1)\|\tilde{x}(t)\|\right)^2\cos^2(\theta(t)) + \max\{|\eta\lambda_2^2-(1-\eta\lambda_2)\|\tilde{x}(t)\||, |\eta\lambda_D^2-(1-\eta\lambda_D)\|\tilde{x}(t)\||\}\sin^2(\theta(t))}$$

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

As

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

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

$$\|\tilde{x}(t)\| > \frac{\eta\lambda_1^2+\eta\lambda_i^2}{2-\eta\lambda_1-\eta\lambda_i} \tag{21}$$

Combining with $\tilde{x}(t)\in\mathbb{I}_1 \Rightarrow \|\tilde{x}(t)\| \le \eta\lambda_1^2$, we have $\eta < \frac{\lambda_1-\lambda_i}{\lambda_1^2}$.

Now consider the following vertors,

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

.
Then we have

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

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

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

Using AM-GM inequality, we have

$$
\begin{aligned}
\lambda_{i+j-1}\lambda_{i+j}^2 &\le \frac{\lambda_{i+j-1}^3 + 2\lambda_{i+j}^3}{3} \\
\lambda_{i+j-1}^2\lambda_{i+j}^2 &\le \frac{\lambda_{i+j-1}^4 + \lambda_{i+j}^4}{2}
\end{aligned}
$$

Hence

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

So,

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

We further discuss three cases

Case 2.1 $\eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}} < \frac{\eta\lambda_1^2+\eta\lambda_i^2}{2-\eta\lambda_1-\eta\lambda_i}$.

In this case we have $\|\tilde{x}(t)\| > \frac{\eta\lambda_1^2+\eta\lambda_i^2}{2-\eta\lambda_1-\eta\lambda_i} > \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$,then

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

The second line is because $(1-\eta\lambda_1)\|\tilde{x}(t)\| + \eta^2\lambda_i^2(\frac{1}{2}\lambda_i^2+\lambda_1^2)\frac{1}{\|\tilde{x}(t)\|}$ monotonously increase w.r.t $\|\tilde{x}(t)\|$ when $\|\tilde{x}(t)\| > \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$. The last line is due to technical lemma Lemma 59.

Case 2.2 $\eta\lambda_1^2 \ge \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}} \ge \frac{\eta\lambda_1^2+\eta\lambda_i^2}{2-\eta\lambda_1-\eta\lambda_i}$.

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

The second line is because of AM-GM inequality. The last line is due to technical lemma Lemma 61.

Case 2.3 $\eta\lambda_1^2 < \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$.

In this case we have $\|\tilde{x}(t)\| < \eta\lambda_1^2 < \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$, then

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

The second line is because $(1-\eta\lambda_1)\|\tilde{x}(t)\| + \eta^2\lambda_i^2(\frac{1}{2}\lambda_i^2+\lambda_1^2)\frac{1}{\|\tilde{x}(t)\|}$ monotonously decrease w.r.t $\|\tilde{x}(t)\|$ when $\|\tilde{x}(t)\| < \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$. The last line is due to technical lemma Lemma 60.

■

Lemma 44 *If $\|\tilde{x}(t)\| \le \frac{\eta\lambda_1^2}{2-\eta\lambda_1}, \tilde{x}(t) \in \cap\mathbb{I}_j$, then $\|\tilde{x}(t+1)\| \le \eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|$*

Proof This can be directly inferred from the proof of lemma 42 ■

Lemma 45 *Define $S_1 = \{t \mid \|\tilde{x}(t)\| \le \frac{\eta\lambda_1^2}{2-\eta\lambda_1}, t > T_1\}$, suppose $t, t' \in S_1, t \le t'$, then $|\tilde{x}_1(t)| < |\tilde{x}_1(t')|$*

Proof For $t \in S_1$, by Lemma 43, $t+1 \in S_1$ or $t+1 \notin S_1, t+2 \in S_1$.

Case 1 $t+1 \in S_1$, we can use Lemma 42 to show $|\tilde{x}_1(t)| < |\tilde{x}_1(t+1)|$.

Case 2 $t+1 \notin S_1, t+2 \in S_1$.

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

We only need to prove

$$\begin{aligned}
&(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|)(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t+1)\|) > \|\tilde{x}(t)\|\|\tilde{x}(t+1)\| \\
\iff &\eta^2\lambda_1^4 - \eta\lambda_1^2(1-\eta\lambda_1)(\|\tilde{x}(t)\| + \|\tilde{x}(t+1)\|) + (-2\eta\lambda_1 + \eta^2\lambda_1^2)\|\tilde{x}(t)\|\|\tilde{x}(t+1)\| \ge 0 \\
\iff &\eta^2\lambda_1^4 - \eta\lambda_1^2(1-\eta\lambda_1)\|\tilde{x}(t)\| \ge \left((2\eta\lambda_1 - \eta^2\lambda_1^2)\|\tilde{x}(t)\| + \eta\lambda_1^2(1-\eta\lambda_1)\right)\|\tilde{x}(t+1)\|
\end{aligned}$$

Now using Lemma 44, we only need to prove,

$$\eta^2\lambda_1^4 - \eta\lambda_1^2(1-\eta\lambda_1)\|\tilde{x}(t)\| \ge \left((2\eta\lambda_1 - \eta^2\lambda_1^2)\|\tilde{x}(t)\| + \eta\lambda_1^2(1-\eta\lambda_1)\right)\left(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|\right)$$

Through some calculation, this is equivalent to

$$((2-\eta\lambda_1)\|\tilde{x}(t)\| - \eta\lambda_1^2)((1-\eta\lambda_1)\|\tilde{x}(t)\| - \eta\lambda_1^2) \ge 0$$

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

Concluding the two cases and use induction, we can get the desired result.

■

D.4.3. LENGTH CONVERGENCE

As Lemma 45 show, $\|\tilde{x}_1(t)\|$ increase monotonously for $t \in S_1$. We can inferred from Lemma 43, S_1 is infinite.

$\forall \epsilon > 0, \exists T_\epsilon$ satisfies $\forall t, t' \in S_1, t' > t > T_\epsilon$, $\frac{\|\tilde{x}_1(t')\|}{\|\tilde{x}_1(t)\|} < 1 + \epsilon$.

Then $\forall t > T_\epsilon$, we have

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

or

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

Hence $\|\tilde{x}(t)\| \geq \min\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1^2+\epsilon}, \frac{\eta^2\lambda_1^3}{(2-\lambda_1\eta)\lambda_1\eta+\epsilon}\right), \forall t > T_\epsilon, t \in S_1$.

As $\forall t \notin S_1, t > T_\epsilon$ we have $\|\tilde{x}(t)\| \geq \eta\lambda_i\sqrt{\frac{\frac{1}{2}\lambda_i^2+\lambda_1^2}{1-\eta\lambda_1}}$.

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

Further by Lemma 44, we can prove $\forall t > T_\epsilon+1$, $\|\tilde{x}(t)\| \leq \eta\lambda_1^2 - (1-\eta\lambda_1)\min\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1^2+\epsilon}, \frac{\eta^2\lambda_1^3}{(2-\lambda_1\eta)\lambda_1\eta+\epsilon}\right)$.

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

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

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

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

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

D.5. Full-batch SAM on General Loss: Proof of Theorem 8

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$-k$ eigenspace of $\nabla^2 L(\Phi(x))$. We will provide a fine-grained convergence bound on $R_j(x)$.

Theorem 46 (Phase I) *Let $\{x(t)\}$ be the iterates defined by SAM (3) and $x(0) = x_{init} \in U$, then under Assumption 1 there exists a constant T_1, such that for any $T_1' > T_1$, it holds for all η, ρ such that $(\eta + \rho)\log(1/\eta\rho)$ is sufficiently small, we have*

$$\max_{T_1 \log(1/\eta\rho) \le \eta t \le T_1' \log(1/\eta\rho)} \max_{j \in [M]} R_j(x) = O(\eta\rho^2)$$

$$\max_{T_1 \log(1/\eta\rho) \le \eta t \le T_1' \log(1/\eta\rho)} \|\Phi(x(t)) - \Phi(x_{init})\| = O((\eta + \rho)\log(1/\eta\rho))$$

Theorem 46 implies SAM will converge to an $O(\eta\rho)$ neighbor of Γ. Notice in the time frame defined by Theorem 46, $x(t)$ effectively operates at a local regime around $\Phi(\lceil -T_1 \log \eta\rho/\eta \rceil)$, this allows us to approximate L with the quadratic Taylor expansion of L at $\Phi(\lceil -T_1 \log \eta\rho/\eta \rceil)$ and give us the following theorem.

Theorem 47 (Phase II) *Let $\{x(t)\}$ be the iterates defined by SAM (3) under Assumptions 1 and 7, further assuming that (1) $\max_j R_j(x(0)) = O(\eta\rho^2)$, (2) $\|\Phi(x(0)) - \Phi(x_{init})\| = O((\eta+\rho)\log(1/\eta\rho))$ and (3) $|\langle x(0) - \Phi(x(0)), v_1(x(0))\rangle| \ge \Omega(\rho^2)$, then there exists constant $T_2 > 0$, for any $T_3 > 0$ till which solution of (6) exists, for all η, ρ such that $\eta \ln(1/\rho)$ and ρ/η is sufficiently small,*

$$\max_{t \le T_3/\eta\rho^2} \|\Phi(x(t)) - X(\eta\rho^2 t)\| = O((\eta + \rho)\log(1/\eta\rho))$$

$$\min_{T_2 \log(1/\rho)/\eta \le t \le T_3/\eta\rho^2} |\langle x(t) - \Phi(x(t)), v_1(x(t))\rangle| = \Theta(\eta\rho)$$

$$\max_{T_2 \log(1/\rho)/\eta \le t \le T_3/\eta\rho^2} \max_{j \in [2:M]} |\langle x(t) - \Phi(x(t)), v_j(x(t))\rangle| = O(\eta\rho^2)$$

In this section we will define K as $\{X(t)\}$ where X is the solution of (6).

D.5.1. PHASE I: PROOF OF THEOREM 46

The proof of Theorem 46 is further split into three subphases.

In **Subphase A**, we will show that the trajectory of SAM will track gradient flow to the working zone K^h. This subphase will take time $O(\frac{1}{\eta})$ for sufficiently small η and ρ. At the end of this subphase $x(t) - \Phi(x(t)) = O(1), \Phi(x(t)) - \Phi(x(0)) = O(\eta + \rho)$.

In **Subphase B**, we will show that in the working zone, the loss will continue to decrease until $\|\nabla L\| = O(\rho)$. This will take time $O(\frac{-\log\rho}{\eta})$ for sufficiently small η and ρ. At the end of this subphase $x(t) - \Phi(x(t)) = O(\rho), \Phi(x(t)) - \Phi(x(0)) = O(-(\eta + \rho)\log\rho)$.

In **Subphase C**, we will show that $R_j(x)$ will shrink exponentially to $O(\eta\rho^{3/2})$ and $x(t)$ will stay in the invariant sets $I_j = \{R_j(x) \le O(\eta\rho^{3/2})\}$.

Subphase A We know $\exists T_0, \|\Phi(x_{\text{init}}, T_0) - \Phi(x_{\text{init}})\| \le \frac{h}{4}, L(\Phi(x_{\text{init}}, T_0)) < \frac{h^2\mu}{16}$. Using standard approximation theory, $\exists \eta_0, \rho_0 > 0$, such that

$$\eta < \eta_0, \rho < \rho_0 \Rightarrow \|x(\frac{T_0}{\eta}) - \Phi(x_{\text{init}}, T_0)\| \le O(\eta + \rho),$$

$$L(x(\frac{T_0}{\eta})) \le \frac{h^2\mu}{8}$$

This further implies

$$\begin{aligned}\|\Phi(x(\frac{T_0}{\eta})) - \Phi(x_{\text{init}})\| &= \|\Phi(x(\frac{T_0}{\eta})) - \Phi(\Phi(x_{\text{init}}, T_0))\| \\ &\le O(\|x(\frac{T_0}{\eta}) - \Phi(x_{\text{init}}, T_0)\|) \\ &\le O(\eta + \rho)\end{aligned}$$

Subphase B Define

$$D(x) = \|\Phi(x) - \Phi(x(\frac{T_0}{\eta}))\|$$

In this subphase, we will track the descent of loss to show that $\exists t$, such that

$$\begin{aligned}\|\nabla(L(x(t)))\| \le 4\zeta\rho \\ \|\Phi(x(t)) - \Phi(x_{\text{init}})\| \le O(-(\eta+\rho)\log\rho)\end{aligned}$$

for sufficiently small η. We require $\eta \le \frac{1}{2\zeta}$. We assume $\inf_{x\in U} L(x) = 0$.

We also requires $-\eta\log\rho$ being sufficiently small, i.e ρ is not too small compared to η. So that

$$\begin{aligned}\log_{1-\frac{\eta\mu}{8}}(\frac{64\zeta^2\rho^2}{h^2})(2\eta\rho\zeta^2 h + 2\eta^2\zeta^4 h^2) &\le \frac{-\log\frac{64\zeta^2\rho^2}{h^2}}{\frac{\eta\mu}{8}}(2\eta\rho\zeta^2 h + 2\eta^2\zeta^4 h^2) \\ &\le -1024\log\rho(2\rho\zeta^2 h + 2\eta\zeta^4 h^2) \\ &\le \frac{h}{8}\end{aligned}$$

We will prove the following proposition,

$$\begin{aligned}&\|\nabla L(x)\| \ge 4\zeta\rho, t \le \frac{T_0}{\eta} + \log_{1-\frac{\eta\mu}{8}}(\frac{64\zeta^2\rho^2}{h^2}) \\ &L(x(t)) \le \frac{h^2\mu}{8}, D(x(t)) \le (2\eta\rho\zeta^2 h + 2\eta^2\zeta^4 h^2)(t - \frac{T_0}{\eta}) \\ \Rightarrow &L(x(t+1)) \le (1 - \frac{\eta\mu}{8})L(x(t)), D(x(t+1)) \le (2\eta\rho\zeta^2 h + 2\eta^2\zeta^4 h^2)(t + 1 - \frac{T_0}{\eta})\end{aligned}$$

Proof

By lemma 31, we have

$$\|x(t) - \Phi(x(t))\| \le \sqrt{\frac{2L(x(t))}{\mu}} \le \frac{h}{2}$$

Further, given the choice of η, ρ, $\|\Phi(x(t)) - \Phi(x_{\text{init}})\| \le \frac{h}{4}$.

Hence we have $x(t) \in K^{\frac{3h}{4}}$. By Lemma 35. we have $\overline{x(t)x(t+1)} \subset K^h$

Under update rule (3), using the smoothness of L, we have

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

■

We have that

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

Hence

$$L(x(t+1))\leq L(x(t))-\eta\|\nabla L\left(x(t)\right)\|^2+\eta\zeta\rho\|\nabla L\left(x(t)\right)\|+\zeta\eta^2\|\nabla L\left(x(t)\right)\|^2+\zeta^3\eta^2\rho^2 33$$

By induction hypothesis, we have

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

Hence as $\overline{x(t+1)x(t)}\in K^h$

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

This implies

$$L(x(t+1))\leq(1-\frac{\eta\mu}{8})L(x(t))$$

Using Lemma 30

$$\begin{aligned}\|\Phi(x(t+1))-\Phi(x(t))\|&\leq\zeta\eta\rho\|\nabla L\left(x\right)\|+\eta\rho^2\nu+\zeta^2\eta^2\|\nabla L\left(x\right)\|^2+\zeta^3\eta^2\rho^2\\ &\leq\eta\rho\zeta^2h+\eta\rho^2\nu+\eta^2\zeta^4h^2+\eta^2\rho^2\zeta^3\\ &\leq 2\eta\rho\zeta^2h+2\eta^2\zeta^4h^2\end{aligned}$$

The induction is complete.
Now define t_1 the minimal $t\geq\frac{T_0}{\eta}$, such that $\|\nabla L(x(t))\|\leq 4\zeta\rho$.
If $t_1>\frac{T_0}{\eta}+\log_{1-\frac{\eta\mu}{8}}(\frac{64\zeta^2\rho^2}{h^2})$, then by the induction,

$$\begin{aligned}L(\frac{T_0}{\eta}+\log_{1-\frac{\eta\mu}{8}}(\frac{64\zeta^2\rho^2}{h^2}))&\leq(1-\frac{\eta\mu}{8})^{\log_{1-\frac{\eta\mu}{8}}(\frac{64\zeta^2\rho^2}{h^2}}L(\frac{T_0}{\eta})\\ &\leq\frac{64\zeta^2\rho^2}{h^2}L(\frac{T_0}{\eta})\\ &\leq 8\zeta^2\rho^2\mu\\ \Rightarrow\nabla L\left(\frac{T_0}{\eta}+\log_{1-\frac{\eta\mu}{8}}(\frac{64\zeta^2\rho^2}{h^2})\right)&\leq 4\zeta\rho.\end{aligned}$$

This is a contradiction.

Subphase C Recall the definition of $R_j(x)$,

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

In this subphase, we will show that $R_j(x)$ will shrink exponentially to $O(\eta\rho^2)$ and $x(t)$ will stay in the invariant sets $I_j = \{R_j(x) \le O(\eta\rho^2 + (\eta\rho)^{3/2})\}$.

Define $\hat{x}(t) = x(t) - \Phi(x(t)), A(t) = \nabla^2 L\left(\Phi(x(t))\right), \tilde{x}(t) = A(t)\hat{x}(t)$

We will prove the induction hypothesis for $t_1 \le t \le t_1 + 10\log_{1-\eta\mu}\frac{\eta\mu^3}{4\zeta^2}$,

$$\|\tilde{x}(t)\| \ge \eta\lambda_1(t)^2 \Rightarrow \|\tilde{x}(t+1)\| \le (1-\eta\mu)\|\tilde{x}(t)\|$$
$$\|\tilde{x}(t)\| \le \eta\lambda_1(t)^2\mu \Rightarrow \|\tilde{x}(t+1)\| \le \eta\rho\lambda_1^2(t) + 2c_1\eta\rho^2$$

As we have $\|\tilde{x}(t_1)\| = \|A(t_1)\hat{x}(t_1)\| \le \frac{\zeta}{\mu}\|\nabla L\left(x(t_1)\right)\| \le \frac{4\zeta^2\rho}{\mu}$.

Combining with the induction hypothesis, we have $\|\tilde{x}(t)\| \le \frac{4\zeta^2\rho}{\mu}$.

Then we have,

$$\begin{aligned}
\|\Phi(x(t+1)) - \Phi(x(t))\| &\le \zeta\eta\rho\|\nabla L\left(x(t)\right)\| + \eta\rho^2\nu + \zeta^2\eta^2\|\nabla L\left(x(t)\right)\|^2 + \zeta^3\eta^2\rho^2 \\
&\le \zeta^2\eta\rho\|x(t) - \Phi(x(t))\| + \zeta^4\eta^2\|x(t) - \Phi(x(t))\|^2 + \eta\rho^2\nu + \zeta^3\eta^2\rho^2 \\
&\le \frac{\zeta\eta\rho}{\mu}\|\tilde{x}(t)\| + \frac{\zeta^4\eta^2}{\mu^2}\|\tilde{x}(t)\|^2 + \eta\rho^2\nu + \zeta^3\eta^2\rho^2 \\
&\le c_0\eta\rho^2
\end{aligned}$$

As $t_1 \le t \le t_1 + 10\log_{1-\eta\mu}\frac{\eta\mu^3}{4\zeta^2} \le t_1 + 10\frac{-\log\frac{\eta\mu^3}{4\zeta^2}}{\eta\mu}$, this implies

$$\|\Phi(x(t)) - \Phi(x(t_1))\| \le O(-\rho^2\log\eta)$$
$$\|\Phi(x(t)) - \Phi(x_{\text{init}})\| \le O(-(\eta+\rho)\log\rho)$$

We have $x(t) \in K^{\frac{h}{2}}$ Using Lemma 35, we conclude that $\overline{x(t)x(t+1)} \subset K^h$.

$$\|\left(x(t+1) - x(t)\right) + \left(\eta\nabla L\left(x(t)\right) + \eta\rho\nabla^2 L\left(x(t)\right)\frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|}\right)\| \le \nu\rho^2\eta$$

Now Using Lemma 32, we have

$$\|(x(t+1) - x(t)) + \eta\nabla^2 L\left(\Phi(x(t))\right)\left(x(t) - \Phi(x(t))\right) + \eta\rho\nabla^2 L\left(x(t)\right)\frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|}\| \le \nu\rho^2\eta + \nu\eta\|x(t) - \Phi(x(t))$$

Further we have

$$\|\frac{\nabla L\left(x(t)\right)}{\|\nabla L\left(x(t)\right)\|} - \frac{\nabla^2 L\left(\Phi(x(t))\right)\left(x(t) - \Phi(x(t))\right)}{\|\nabla L\left(x(t)\right)\|}\| \le \frac{\nu\|x(t) - \Phi(x(t))\|^2}{2\|\nabla L\left(x(t)\right)\|} \le \frac{\nu\|x(t) - \Phi(x(t))\|}{2\mu}$$
$$\|\frac{\nabla^2 L\left(\Phi(x(t))\right)\left(x(t) - \Phi(x(t))\right)}{\|\nabla L\left(x(t)\right)\|} - \frac{\nabla^2 L\left(\Phi(x(t))\right)\left(x(t) - \Phi(x(t))\right)}{\|\nabla^2 L\left(\Phi(x(t))\right)\left(x(t) - \Phi(x(t))\right)\|}\| \le \frac{4v}{3\mu}\|x(t) - \Phi(x(t))\|$$

Hence

$$
\begin{aligned}
&\|(x(t+1)-x(t))+\eta\nabla^2 L\left(\Phi(x(t))\right)(x(t)-\Phi(x(t)))+\eta\rho\nabla^2 L\left(x(t)\right)\frac{\nabla^2 L\left(\Phi(x(t))\right)(x(t)-\Phi(x(t)))}{\|\nabla^2 L\left(\Phi(x(t))\right)(x(t)-\Phi(x(t)))\|}\| \\
\leq&\nu\rho^2\eta+\nu\eta\|x(t)-\Phi(x(t))\|^2+\eta\rho\frac{2\zeta\nu}{\mu}\|x(t)-\Phi(x(t))\|
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\|(x(t+1)-x(t))+\eta\nabla^2 L\left(\Phi(x(t))\right)(x(t)-\Phi(x(t)))+\eta\rho\nabla^2 L\left(\Phi(x(t))\right)\frac{\nabla^2 L\left(\Phi(x(t))\right)(x(t)-\Phi(x(t)))}{\|\nabla^2 L\left(\Phi(x(t))\right)(x(t)-\Phi(x(t)))\|}\| \\
\leq&\nu\rho^2\eta+\nu\eta\|x(t)-\Phi(x(t))\|^2+\eta\rho\frac{2\zeta\nu}{\mu}\|x(t)-\Phi(x(t))\|+\nu\eta\rho\|x(t)-\Phi(x(t))\|
\end{aligned}
$$

This implies,

$$
\begin{aligned}
&\|A(t)\left((x(t+1)-x(t))+\eta\tilde{x}(t)+\eta\rho A(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\right)\| \\
\leq&\zeta\nu\rho^2\eta+\zeta\nu\eta\|x(t)-\Phi(x(t))\|^2+\zeta\eta\rho\frac{2\zeta\nu}{\mu}\|x(t)-\Phi(x(t))\|+\zeta\nu\eta\rho\|x(t)-\Phi(x(t))\|
\end{aligned}
$$

Also by Lemma 35

$$
\begin{aligned}
&\|A(t)(x(t+1)-x(t))-\tilde{x}(t+1)+\tilde{x}(t)\| \\
=&\|A(t)(x(t+1)-x(t))-A(t+1)x(t+1)+A(t)x(t)+A(t+1)\Phi(x(t+1))-A(t)\Phi(x(t))\| \\
=&\|(A(t)-A(t+1))(x(t+1)-\Phi(x(t+1)))+A(t)(\Phi(x(t+1))-\Phi(x(t)))\| \\
\leq&\nu\|\Phi(x(t+1))-\Phi(x(t))\|\|x(t+1)-\Phi(x(t+1))\|+\zeta\|\Phi(x(t+1))-\Phi(x(t))\| \\
\leq&(\nu h+\zeta)\|\Phi(x(t+1))-\Phi(x(t))\| \\
\leq&(\nu h+\zeta)(\zeta\eta\rho\|\nabla L\left(x\right)\|+\eta\rho^2\nu+\zeta^2\eta^2\|\nabla L\left(x\right)\|^2+\zeta^3\eta^2\rho^2)
\end{aligned}
$$

Combining with induction hypothesis, we know exists constant c_1, such that

$$
\|\tilde{x}(t+1)-\tilde{x}(t)+\eta A(t)\tilde{x}(t)+\eta\rho A^2(t)\frac{\tilde{x}(t)}{\|\tilde{x}(t)\|}\|\leq c_1\eta\rho^2
$$

We first bound $\|\tilde{x}(t)\|$, as in quadratic case, if $\|\tilde{x}(t)\|>\eta\rho\lambda_1^2(t)$, we would have

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

Choosing ρ small enough, we have

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

If$\|\tilde{x}(t)\| \le \eta\rho\lambda_1^2(t)$

$$\begin{aligned}\|\tilde{x}(t+1)\| &\le \eta\rho\lambda_1^2(t) + c_1\eta\rho^2 \\ &\le \eta\rho\lambda_1^2(t+1) + 2c_1\eta\rho^2\end{aligned}$$

Here we use

$$\max_i \|\lambda_i(t) - \lambda_i(t+1)\| \le \|A(t+1) - A(t)\|_2 \le c_0\kappa\eta\rho^2$$

Now define t_2 the minimal $t \ge t_1$, such that $\|\tilde{x}(t)\| \le \eta\rho\lambda_1^2(t)$.
If $t_2 > t_1 + \log_{1-\eta\mu}(\frac{\eta\mu^3}{4\zeta^2})$, then by the induction,

$$\begin{aligned}\|\tilde{x}(t_1 + \log_{1-\eta\mu}(\frac{\eta\mu^3}{4\zeta^2}) + 1)\| &\le (1-\eta\mu)^{\log_{1-\eta\mu}(\frac{\eta\mu^3}{4\zeta^2})}\|\tilde{x}(t_1)\| \\ &\le \frac{\eta\mu^3}{4\zeta^2}\|\tilde{x}(t_1)\| \\ &\le \frac{\eta\mu^3}{4\zeta^2}\zeta\|x(t_1) - \Phi(x(t_1)\| \\ &\le \frac{\eta\mu^2}{4\zeta}\|\nabla L(t_1)\| \\ &\le \mu^2\eta\rho \\ &\le \lambda_1^2(t_1 + \log_{1-\eta\mu}(\frac{\eta\mu^3}{4\zeta^2}) + 1)\eta\rho\end{aligned}$$

This is a contradiction.
Following the induction, we further have for $t_2 \le t \le t_1 + T_1' \log_{1-\eta\mu}(\frac{\eta\mu^3}{4\zeta^2})$,

$$\|x(t)\| \le \eta\rho\lambda_1^2(t+1) + 2c_1\eta\rho^2$$

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

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

We would have

$$\min_{k\in s_k, l\in S_j} |\lambda_k(t+1) - \lambda_l(t+1)| \ge \rho - 8\nu\eta\rho^2 \ge 0.99\rho$$

We would then have for $t \ge t_2$ (analogous to proof in Section D.4.1)
If $\sqrt{\sum_{i=j}^p \|P_{S^{(i)}}^{(t)}\tilde{x}(t)\|^2} > \max_{k\in S_j}\lambda_k^2(t)\eta\rho$

$$\sqrt{\sum_{i=j}^p \|P_{S^{(i)}}^{(t)}\tilde{x}(t+1)\|^2} \le \max\{(1-\eta\lambda_D(t+1)\|\sum_{i=j}^p P_{S^{(i)}}^{(t)}\tilde{x}(t)\| - \eta\rho\lambda_D(t+1)^2, \eta\max_{k\in S_j}\lambda_k(t+1)\|\sum_{i=j}^p P_{S^{(i)}}^{(t)}\tilde{x}(t)\|\} \cdot$$

If $\sqrt{\sum_{i=j}^{p} \|P_{S^{(i)}}^{(t)}\tilde{x}(t)\|^2} \le \max_{k\in S_j} \lambda_k^2(t)\eta\rho$

$$\begin{aligned}\sqrt{\sum_{i=j}^{p} \|P_{S^{(i)}}^{(t)}\tilde{x}(t+1)\|^2} &\le \max_{k\in S_j} \lambda_k^2(t)\eta\rho + c_1\eta\rho^2 \\ &\le \max_{k\in S_j} \lambda_k^2(t+1)\eta\rho + 2c_1\eta\rho^2\end{aligned}$$

Further we have $\|P_{s_k}^{(t)} - P_{s_k}^{(t+1)}\| \le O(\nu\eta\rho)$ by the Lemma 53

So we have

If $\sqrt{\sum_{i=j}^{p} \|P_{S^{(i)}}^{(t)}\tilde{x}(t)\|^2} > \max_{k\in S_j} \lambda_k^2(t)\eta\rho$

$$\sqrt{\sum_{i=j}^{p} \|P_{S^{(i)}}^{(t+1)}\tilde{x}(t+1)\|^2} \le \max\{(1-\eta\lambda_D)\|\sum_{i=j}^{p} P_{S^{(i)}}^{(t)}\tilde{x}(t)\| - \eta\rho\lambda_D^2, \eta\max_{k\in S_j}\lambda_k\|\sum_{i=j}^{p} P_{S^{(i)}}^{(t)}\tilde{x}(t)\|\} + c_1\eta\rho^2$$

If $\sqrt{\sum_{i=j}^{p} \|P_{S^{(i)}}^{(t)}\tilde{x}(t)\|^2} \le \max_{k\in S_j} \lambda_k^2(t)\eta\rho$

$$\begin{aligned}\sqrt{\sum_{i=j}^{p} \|P_{S^{(i)}}^{(t+1)}\tilde{x}(t+1)\|^2} &\le \max_{k\in S_j} \lambda_k^2(t)\eta\rho + c_1\eta\rho^2 + O(\eta\rho^2) \\ &\le \max_{k\in S_j} \lambda_k^2(t+1)\eta\rho + 2c_1\eta\rho^2 + O(\eta\rho^2)\end{aligned}$$

Finally taking into quantization error, as all the eigenvalue in the same group at most differ $D\rho$, we would have

If $R_k(x(t)) \ge 0$

$$\begin{aligned}R_k(x(t+1)) + \lambda_k^2(t+1)\eta\rho &\le (1-\eta\lambda_D)R_k(x(t)) \\ &\le (1-\eta\mu)(R_k(x(t)) + \lambda_k^2(t)\eta\rho)\end{aligned}$$

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

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

Similar to the proof of existence of t_2, we can show the existence of $t_3 \le t_2 + \log_{1-\eta\mu}(\frac{\eta\mu^3}{4\zeta^2})$, such that for $t_3 \le t \le t_3 + T_1' \log_{1-\eta\mu}(\frac{\eta\mu^3}{4\zeta^2})$,

$$\max_j R_j(t) \le O(\eta\rho^2)$$

D.5.2. PHASE II: PROOF OF THEOREM 47

The proof consists of two subphase.

In **subphase A**, we will show $x(t) - \Phi(x(t))$ aligns with $v_1(t)$ in $O(\log(1/\rho)/\eta)$ steps.

In **subphase B**, we will show that the alignment continues to hold and $x(t)$ moves as a time-rescaled version of solution of Equation (6).

Subphase A This proof is analogous to Section D.4.2. To maintain consistency with previous section, we abuse notation and change the starting step to t_3. The time frame we are discussing is $t_3 \le t \le t_3 + T_1' \log_{1-\eta\mu}(\frac{\eta\mu^3}{4\zeta^2})$.

First Induction We will inductively prove $\|x_1(t)\|$ is $\Omega(\rho^2)$ and that there exists step t_4 for $t_4 \le t \le t_3 + T_1' \log_{1-\eta\mu}(\frac{\eta\mu^3}{4\zeta^2})$ that $x_1(t) \ge \frac{1}{4}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + 3\frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\rho - O(\eta\rho^2)$

For step t, we fixed the quadratic function as $\langle x - \Phi(x(t)), \nabla^2 L\left(\Phi(x(t))\right)(x - \Phi(x(t))\rangle$,

Define $\bar{x} = \frac{\nabla^2 L(\Phi(x(t)))(x-\Phi(x(t))}{\rho}$, $A = \nabla^2 L(t)$

We have for $t \le t_4 - 1$, $\|\bar{x}_1(t)\| \le \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)$.

By assumption, we have $\|\bar{x}_1(t)\| \ge \Omega(\rho)$.

We have

$$
\begin{aligned}
&\|(x(t+1) - x(t)) + \eta\nabla^2 L\left(\Phi(x(t))\right)(x(t) - \Phi(x(t))) + \eta\rho\nabla^2 L\left(\Phi(x(t))\right)\frac{\nabla^2 L\left(\Phi(x(t))\right)(x(t) - \Phi(x(t)))}{\|\nabla^2 L\left(\Phi(x(t))\right)(x(t) - \Phi(x(t)))\|}\| \\
\le&\nu\rho^2\eta + \nu\eta\|x(t) - \Phi(x(t))\|^2 + \eta\rho\frac{2\zeta\nu}{\mu}\|x(t) - \Phi(x(t))\| + \nu\eta\rho\|x(t) - \Phi(x(t))\| \\
\le&c_2\eta\rho^2
\end{aligned}
$$

Further we have, there exists constant c_4 such that

$$
\begin{aligned}
&\|(x(t+1) - x(t)) + \eta\rho\bar{x}(t) + \eta\rho A\frac{\bar{x}(t)}{\|\bar{x}(t)\|}\| \\
\le&c_2\eta\rho^2 + c_3\|\Phi(x(t)) - \Phi(x(t_3))\| \\
\le&-\frac{c_4}{\zeta}\eta\rho^2\log\rho
\end{aligned}
$$

Now we have

$$
\|\bar{x}(t+1) - \bar{x}(t) + \eta A\bar{x}(t) + \eta A^2\frac{\bar{x}(t)}{\|\bar{x}(t)\|}\| \le -c_4\eta\rho\log\rho
$$

So here our goal is to discuss the dynamics of the following perturbed version of quadratic SAM.

$$
\begin{aligned}
\|\bar{x}(t+1) - \bar{x}(t) + \eta A\bar{x}(t) + \eta A^2\frac{\bar{x}(t)}{\|\bar{x}(t)\|}\| &\le -c_4\eta\rho\log\rho \\
\|P^{(j:D)}\bar{x}(t)\| - \lambda_j^2\eta &\le c_5(\eta\rho + \eta^{3/2}\rho^{1/2})
\end{aligned}
$$

Define $\hat{x}(t+1)$,

$$
\hat{x}(t+1) = \bar{x}(t) - \eta A\bar{x}(t) - \eta A^2\frac{\bar{x}(t)}{\|\bar{x}(t)\|}
$$

In the quadratic case, we have Lemma 43 to show $\|\tilde{x}(t)\|$ can't stay greater $\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$ for two steps. Hence here we have if $\|\bar{x}(t)\| \ge \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, then $\|\hat{x}(t+1)\| \le \frac{\eta\lambda_1^2}{2-\eta\lambda_1} - O(\eta)$, which leads $\|\bar{x}(t+1)\| \le \|\hat{(}x(t+1))\| + \tilde{O}(\eta\rho) \le \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$. Here we can in fact prove a more subtle version of this lemma showing,

Lemma 48

$$\exists C_1(\lambda_1,\lambda_2,\lambda_D), C_2(\lambda_1,\lambda_2,\lambda_D) < 1, \|\bar{x}(t)\| \geq C_1 \frac{\eta\lambda_1^2}{2-\eta\lambda_1} \Rightarrow \|\bar{x}(t+1)\| \leq C_2 \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$$

Proof

If

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

Using Lemma 43, we have

$$\|\hat{x}(t)\| \leq \max(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} - \eta\frac{\lambda_D^4}{2\lambda_1^2}, \eta\lambda_1^2 - (1-\eta\lambda_1)\|\tilde{x}(t)\|) \leq c_6(\lambda_1,\lambda_D)\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$$
$$c_6(\lambda_1,\lambda_D) < 1$$

If

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

Then we have

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

This implies,

$$\|\hat{x}(t+1)\| \leq (\eta\lambda_1^2 - (1-\eta\lambda_1)\|\bar{x}(t)\|)\sqrt{\frac{\|\bar{x}_1^2(t)\|}{\|\bar{x}(t)\|^2} + (1 - \frac{\|\bar{x}_1^2(t)\|}{\|\bar{x}(t)\|^2})\max\{\frac{\lambda_1^2-\lambda_D^2}{\lambda_1^2}, \frac{\lambda_2^2}{\lambda_1^2}\}}$$

As we suppose

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

This implies

$$\frac{\|\bar{x}_1(t)\|}{\|\bar{x}(t)\|} \leq \frac{\lambda_1^2 + \lambda_2^2}{2\lambda_1^2}$$

Combining we have

$$\|\hat{x}(t+1)\| \leq c_7(\lambda_1,\lambda_2,\lambda_D)(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\bar{x}(t)\|)$$
$$c_7(\lambda_1,\lambda_2,\lambda_D) < 1$$

This implies

$$\|\bar{x}(t)\| \geq \frac{c_6\eta\lambda_1^2 - \frac{c_5+1}{\frac{\eta\lambda_1^2}{2-\eta\lambda_1}}}{c_6(1-\lambda_1)}$$
$$\Rightarrow \|\hat{x}(t+1)\| \leq \max\{c_7, \frac{c_6+1}{2}\}\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$$

As $\exists c_8$,

$$\|\hat{x}(t+1) - \bar{x}(t+1)\| \leq c_8\eta\rho$$

. We can conclude that

$$\exists C_1(\lambda_1, \lambda_2, \lambda_D), C_2(\lambda_1, \lambda_2, \lambda_D) < 1, \|\bar{x}(t)\| \geq C_1\frac{\eta\lambda_1^2}{2-\eta\lambda_1} \Rightarrow \|\bar{x}(t+1)\| \leq C_2\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$$

■

Define $T = \{t | \|\bar{x}(t)\| \leq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)\}$, $S = \{t | \|\bar{x}(t)\| \leq \frac{\eta\lambda_1^2}{2-\eta\lambda_1}\}$,
For $s \in S, n(s) := \min_{j>s}\{j \in S\}$.
We will show when $\|\bar{x}_1(t)\| \leq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)$,

$$\exists c_4(\lambda_1, \lambda_2, \lambda_D) > 1, \|\bar{x}_1(n(s))\| \geq c_4\|\bar{x}_1(s)\| \text{ or } \|\bar{x}_1(n(n(s)))\| \geq c_4\|\bar{x}_1(s)\|$$

Define $\underline{x}(t+1) = \hat{x}(t) - \eta A\hat{x}(t) - \eta A^2\frac{\hat{x}(t)}{\|\hat{x}(t)\|}$
Consider two cases

Case 1 $\|\bar{x}(s)\| \leq C_1\frac{\eta\lambda_1^2}{2-\eta\lambda_1}$

$$\begin{aligned}
\frac{\|\hat{x}_1(s+1)\|}{\|\bar{x}_1(t)\|} &= \frac{\eta\lambda_1^2 - (1-\eta\lambda_1)\|\bar{x}(t)\|}{\|\bar{x}(t)\|} \geq \frac{(2-C_1) - \eta\lambda_1 + C_1\eta\lambda_1}{C_1} \geq \frac{1}{C_1} \\
\frac{\|\underline{x}_1(s+2)\|}{\|\bar{x}_1(s)\|} &= \frac{(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\bar{x}(s)\|)(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\hat{x}(s+1)\|)}{\|\bar{x}(s)\|\|\hat{x}(s+1)\|} \\
&\geq \frac{(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\bar{x}(s)\|)\left(\eta\lambda_1^2 - (1-\eta\lambda_1)\left(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\bar{x}(s)\|\right)\right)}{\|\bar{x}(s)\|\left(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\bar{x}(s)\|\right)} \\
&= \frac{\eta\lambda_1^2 - (1-\eta\lambda_1)\left(\eta\lambda_1^2 - (1-\eta\lambda_1)\|\bar{x}(s)\|\right)}{\|\bar{x}(s)\|} \geq (1-\eta\lambda_1)^2 + \frac{(2-\eta\lambda_1)}{C} \\
&\geq 1 + C_3\eta
\end{aligned}$$

Here we require $C_3 \geq 0$ As $n(s) = s+1$ or $n(s) = s+2$ and we have

$$\|\underline{x} - \bar{x}\| \leq c_9\eta\rho$$
$$\|\hat{x} - \bar{x}\| \leq c_8\eta\rho$$

Also

$$|\bar{x}_1(s)| \geq \Omega(\rho)$$

We have

$$\|\bar{x}_1(n(s))\| \geq (1 + C_3\eta/2)\|\bar{x}_1(s)\|$$

Case 2 $\|\bar{x}(s)\| > C_1 \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, then $\|\bar{x}(s+1)\| \leq C_2 \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, $n(s) = s+1$
Similar to case 1 We have $\|\hat{x}_1(s+1)\| \geq \|\tilde{x}_1(s)\|$
So in fact we have $\bar{x}_1(n(s)) \geq \bar{x}_1(s) - O(\eta\rho)$, suppose $\frac{\bar{x}_1(t_3)}{\rho}$ is a sufficiently large constant, then we can assume $|\bar{x}_1(n(s))| \geq (1 - C_3\eta/8)|\bar{x}_1(s)|$.
As $\|\bar{x}(n(s))\| \leq C_2 \frac{\eta\lambda_1^2}{2-\eta\lambda_1}$, similar to case 1, $\|\bar{x}_1(n(n(s)))\| \geq (1 + C_3\eta/2)\|\bar{x}_1(n(s))\| \geq (1 + C_3\eta/2)\|\bar{x}_1(s)\|$

In conclusion, if $\|\bar{x}_1(s)\| \leq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)$, we would have

$$\exists c_4(\lambda_1, \lambda_2, \lambda_D) > 0, \|\bar{x}_1(n(s))\| \geq (1 + c_4\eta)\|\bar{x}_1(s)\| \text{ or } \|\bar{x}_1(n(n(s)))\| \geq (1 + c_4\eta)\|\bar{x}_1(s)\|$$

This implies $\exists t_4 \leq t_3 - 4(\log\rho)/(c_4\eta)$, such that

$$\|\bar{x}_1(t_4)\| \geq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)$$

As we have $\|\bar{x}_1(n(t))\| \geq \|\bar{x}_1(t)\| - \max\{c_8, c_9\}(\eta\rho + \eta^{3/2}\rho^{1/2})$ for $\|\bar{x}_1(t)\| \geq \frac{1}{2}\left(\frac{\eta\lambda_1^2}{2-\eta\lambda_1} + \frac{\eta\lambda_2^2}{2-\eta\lambda_2}\right)$, we would have

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

for $t_3 + T_1' \log_{1-\eta\mu}(\frac{\eta\mu^3}{4\zeta^2}) \geq t \geq t_4$ and the first induction is complete.

Second Induction Define $\hat{x}$ and $\bar{x}$ as before.

This implies for $t \geq t_4$,

$$-1 + c_{10}(\lambda_1, \lambda_2) \leq 1 - \eta\lambda_2 - \eta\frac{\lambda_2^2}{\|\tilde{x}(t)\|} \leq 1 - \eta\lambda_D - \eta\frac{\lambda_D^2}{\|\tilde{x}(t)\|} \leq 1 - \frac{\lambda_D^2}{2\lambda_1^2}$$
$$\|P^{(2:D)}\hat{x}(t+1)\| \leq \max(1 - \frac{\lambda_D^2}{2\lambda_1^2}, 1 - c_{10}(\lambda_1, \lambda_2))\|P^{(2:D)}\bar{x}(t)\|$$

As $\|P^{(2:D)}\bar{x}(t)\| \leq 2\lambda_2^2\eta$ We can inductively show that for $t \geq t_4 + \log_{\max(1-\frac{\lambda_D^2}{2\lambda_1^2}, 1-c_{10}(\lambda_1,\lambda_2))} \frac{\rho^2}{\zeta^2}$ iteration, $\|P^{(2:D)}\bar{x}(t+1)\| \leq 2c_5(\eta\rho)$

Now as we have

$$\|P^{(2:D)}(t) - P^{(2:D)}(t+1)\| \leq O(\nu\rho^2)$$
$$\|v_1(t) - v_1(t+1)\| \leq O(\nu\rho^2)$$
$$\|\lambda_1(t) - \lambda_1(t+1)\| \leq O(\nu\rho^2)$$

This implies we have for $t \geq t_4 + O(\log_{\max(1-\frac{\lambda_D^2}{2\lambda_1^2}, 1-c_{10}(\lambda_1,\lambda_2))} \frac{\rho^2}{\zeta^2})$

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

Subphase B We are now ready to show that $\Phi(x(t))$ will track the solution of (6). The main principal of this proof have been introduced in Section C.1.

To simplify our writing define $\theta(t) = \arccos(\langle v_1, \frac{\nabla L(x)}{\|\nabla L(x)\|}\rangle)$

We can inductively prove the following statement

$$
\begin{aligned}
\|\Phi(x(t)) - X(\eta\rho^2 t)\| &\le O(-(\eta+\rho)\log\rho) \\
\|x(t) - \Phi(x(t))\| &= \Theta(\frac{\eta\rho\lambda_1(t)}{2-\eta\lambda_1(t)}) \\
\|\bar{x}_1(t)\| &\ge \frac{1}{2}\left(\frac{\eta\lambda_1^2(t)}{2-\eta\lambda_1(t} + \frac{\eta\lambda_2^2(t)}{2-\eta\lambda_2(t)}\right)\rho - O(\eta\rho^2) \\
\|P^{(2:D)}(t)\bar{x}(t)\| &\le O(\eta\rho^2)
\end{aligned}
$$

The initial condition is satisfied by assumption.

We have

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

Using Lemma 31, we have

$$
\eta\rho\partial\Phi(x)\nabla^2 L(x)\frac{\nabla L(x)}{\|\nabla L(x)\|} = \eta\rho\|\nabla L(x)\|\|\partial^2\Phi(x)\left[\frac{\nabla L(x)}{\|\nabla L(x)\|}, \frac{\nabla L(x)}{\|\nabla L(x)\|}\right]/2\| \le \eta\rho\kappa\|\nabla L(x)\|
$$

This implies,

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

Further

$$
\begin{aligned}
&\|\Phi(x(t+1)) - \Phi(x(t)) - \eta\rho^2\partial\Phi(x)\partial\nabla^2 L(\Phi(x))[\frac{\nabla L(x)}{\|\nabla L(x)\|}, \frac{\nabla L(x)}{\|\nabla L(x)\|}]/2\| \\
\le&\eta\rho^3\Gamma + \eta\rho\kappa\|\nabla L(x)\| + \eta\rho^2\nu\|x-\Phi(x)\| + \frac{1}{2}\zeta\|x(t+1)-x(t)\|^2
\end{aligned}
$$

By induction we have $0 \le t \le \frac{T_3}{\eta\rho^2}$, we have $\|x - \Phi(x)\| = \Theta(\eta\rho), \theta = O(\rho)$

So we have

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

Further we have

$$
\begin{aligned}
&\|\eta\rho^2\partial\Phi(x)\partial\nabla^2 L(\Phi(x))[\frac{\nabla L(x)}{\|\nabla L(x)\|}, \frac{\nabla L(x)}{\|\nabla L(x)\|}]/2 - \eta\rho^2\partial\Phi(x)\partial\nabla^2 L(\Phi(x))[v_1(t), v_1(t)]/2 \\
\le&\eta\rho^2(O(\zeta\theta) + O(\frac{\nu\zeta\|x-\Phi(x(t))\|}{\mu})) \le O(\eta\rho^3)
\end{aligned}
$$

We have

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

This implies

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

Hence we can perform the induction and the accumulated approximation error will be of order $O(\rho + \eta)$.

D.5.3. PROOF OF COROLLARY

Proof [Proof of Corollary 15] We will do a Taylor expansion on $L^{\max}_\rho(x)$. By Theorem 46 and 47, for $t > T_3'/\eta\rho^2$, we have $\|X(\eta\rho^2 t) - x(t)\| = \tilde{O}(\eta + \rho)$ and $\|x(t) - \Phi(x(t))\| = O(\eta\rho)$

$$R^{\max}_\rho(x) = \max_v \rho v^T \nabla L(x) + \rho^2 v^T \nabla^2 L(x) v/2 + O(\rho^3)$$

Then as $\|v^T \nabla L(x)\| = O(\eta\rho)$, this implies

$$\begin{aligned}
R^{\max}_\rho(x) &= \rho^2 \max_v v^T \nabla^2 L(x) v/2 + O(\eta^2\rho^2 + \rho^3) \\
&= \rho^2 \max_v v^T \nabla^2 L(X(\eta\rho^2 t)) v/2 + \tilde{O}(\eta\rho^2) \\
&= \rho^2 \lambda_1(X(\eta\rho^2 t))/2 + \tilde{O}(\eta\rho^2)
\end{aligned}$$

■

Proof [Proof of Corollary 16]

Choose T such that $X(T)$ is sufficiently close to $X(\infty)$, such that $\lambda_1(X(T)) \le \lambda_1(X(\infty)) + 2\epsilon$

By corollary 15, we have $\|R^{\max}_\rho(x(\lceil T/(\eta\rho^2)\rceil)) - \rho^2\lambda_1(X(T))/2\| \le \tilde{O}(\eta\rho^2)$. This further implies $\|R^{\max}_\rho(x(\lceil T/(\eta\rho^2)\rceil)) - \rho^2\lambda_1(X(\infty))/2\| \le \epsilon\rho^2 + \tilde{O}(\eta\rho^2)$. We also have $\|L(x(\lceil T/(\eta\rho^2)\rceil))\| = O(\eta^2\rho^2)$. Then we can leverage Theorem 4 and Theorem 12 to get the desired bound. ■

D.6. Stochastic SAM: Proof of Theorem 11

We will first prove Lemma 22,

Proof [Proof of Lemma 22] Note that $\frac{d\ell(y',y_k)}{dy'}|_{y'=f_k(p)} = 0$, we have $\nabla^2 L_k(p) = \Lambda_k(p) w_k(p) w_k(p)^T$. Also note $\nabla L_k(x)/\|\nabla L_k(x)\| = \mathrm{sign}(\frac{d\ell(y',y_k)}{dy'}|_{y'=f_k(x)})\nabla f_k(x)/\|\nabla f_k(x)\|$. By Assumption 1, $\nabla^2 L(p) = \sum_k \nabla^2 L_k(p)/M = \sum_k \Lambda_k(p) w_k(p) w_k(p)^T/M$ has rank M , this implies $\forall k, \nabla f_k(p) \neq 0$, hence $\nabla f_k(x)/\|\nabla f_k(x)\|$ is in C^1 near p and we have proved our claim. ■

We will prove under a more general assumption.

Assumption 49 *Assume loss $L = \sum_k L_k/M$ and L_k belongs to $\mathcal{C}^4$, and there exists a manifold Γ_k that is $D-1$ dimensional $\mathcal{C}^2-$submanifold of $\mathbb{R}^D$, where for all $x \in \Gamma$, x is a global minimizer of L_k, $L_k(x) = 0$ and rank$(\nabla^2 L_k(x)) = 1$. Let $U_k = \{x \in \mathbb{R}^D | \Phi(x)$ exists and $\Phi_k(x) \in \Gamma_k\}$.*

We have U_k is open and Φ_k is in $\mathcal{C}^3$ on U_k.(from Lemma B.15 [2])

Theorem 50 shows that Setting in Theorem 11 satisfies assumption 49.

Theorem 50 *Suppose $L(x) = \sum_{k=1}^{M} L_k(x)/M$ and manifold Γ satisfy Assumption 1, and that $f_k(x) = y_k$ for every $x \in \Gamma, k \in [M]$. Then there exists $(D-1)$-dimensional $\mathcal{C}^2$ submanifolds of $\mathbb{R}^D$, such that $\cap_{k=1}^{M} \Gamma_k = \Gamma$ and for every $x \in \Gamma_k$ and $k \in [M]$, $f_k(x) = y_k$, that is, Γ_k is a manifold of global minimizers of L_k.*

Proof [Proof of Theorem 50]

By standard calculus, for $x \in \Gamma$, we have $\nabla^2 L(x) = \sum_{k=1}^{M} w_k w_k^T / M$. By Assumption 1,$\nabla^2 L(x)$ is full rank, this implies $w_k \neq 0$. Then we have in an open set $V(x)$ containing x, $\nabla f_k(x) \neq 0$. Then consider $V = \cup_{x \in \Gamma} V(x)$, which is an open set and in which $\nabla f_k \neq 0$. Now apply preimage theorem, we would have $\Gamma_k = \{x \in V | f_k(x) \neq y_k\}$ forms C^2 dimensional sub-manifolds. We also easily have $\cap \Gamma_k = \Gamma$ from definitions. ■

The following theorems shows that stochastic SAM (7) essentially minimize trace of Hessian of loss. Analogous to the full-batch setting, we will split the trajectory into two phase.

Theorem 51 (Phase I) *Let $\{x(t)\}$ be the iterates defined by SAM (3) and $x(0) = x_{init} \in U$, then under Assumption 1 and 49 there exists a constant T_1, it holds for sufficiently small $-(\eta + \rho)\log(\eta\rho)$, we have with probability $1 - O(\sqrt{\rho})$, $\min_{-T_1 \log \rho/\eta \geq t} \|x(t) - \Phi(x(t))\| = O(-(\eta\rho + \rho^2)\log \eta\rho)$ and $\max_{-T_1 \log \rho/\eta \geq t} \|\Phi(x_{init}) - \Phi(x(t))\| = O(-(\eta + \rho)\log(\eta\rho))$.*

Theorem 51 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 1 and 49, we have $\mathrm{Tr}(\nabla^2 L_k(x)) = \lambda_1(\nabla^2 L_k(x))$ is differentiable for $x \in \Gamma_i$. Hence $\mathrm{Tr}(\nabla^2 L(x)) = \sum_i \mathrm{Tr}(\nabla^2 L_k(x))$ is also differentiable and we have (8) is well defined for some finite time T_2.

Theorem 52 (Phase II) *Let $\{x(t)\}$ be the iterates defined by SAM (7) under Assumptions 1 and 49, assuming (1) $\forall t, k, L_k(x(t)) \neq 0$, (2) $\|x(0) - \Phi(x(0))\| = O(-(\eta\rho + \rho^2)\log \eta\rho)$ and (3) $\|\Phi(x_{init}) - \Phi(x(0))\| = O(-(\eta + \rho)\log(\eta\rho))$, then for any $T_2 > 0$ till which solution of (8) exists, for sufficiently small $-(\eta + \rho)\log(\eta\rho)$, we have with probability $1 - O(\eta\rho)$, for all $\eta\rho^2 t < T_2$, $\|\Phi(x(t)) - X(\eta\rho^2 t)\| = O(-(\eta + \rho)\log \eta\rho)$ and $\|x(t) - \Phi(x(t))\| = O(-(\eta\rho + \rho^2)\log \eta\rho)$.*

In this section we will define K as $\{X(t)\}$ where X is the solution of (8).

We will prove these theorems respectively in the following sections.

D.6.1. PHASE I: PROOF OF THEOREM 51

We will now discuss the convergence of 1-SAM to the manifold of minimizer. We will separate the dynamics into the following phase. Define ϕ as the gradient flow projection as in deterministic case.

Subphase A Gradient flow approximation, using standard approximation, we can show that $x(t)$ with high probability falls into a region K^h where the loss satisfies PL condition.

Subphase B After reaching the region, a detailed analysis will show that loss continue to decrease until $\|x(t) - \Phi(x(t))\| = \tilde{O}(\rho)$.

Subphase C Consider a quadratic approximation and we will get with high probability $x(t)$ will falls into $O(\eta\rho + \rho^2)$ neighbor of $\Phi(x(t))$.

Subphase A This is analogous to Subphase A in Section D.5.1 and we can suppose $\exists t_1, x(t_1) \in K^h$

Subphase B Define event $A(t)$ as $\{\nabla L\left(x(\tau)\right) \geq 4\zeta\rho, \forall \tau \leq t\}$.

We have if $\|\nabla L\left(x(t)\right)\| \geq 4\zeta\rho$

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

We have

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

We can then conclude that with $t_2 = \frac{2\log\frac{h^2}{16\rho^2\mu M\eta}}{M\eta\mu} + t_1$

$$16\zeta^2\rho^2\mu\mathbb{P}\left(A(t_2+1)\right) \leq \mathbb{E}[L(x(t_2+1))\mathbf{1}A(t_2+1)] \leq (1 - \frac{M\eta\mu}{2})^{t_2-t_1}L(x(t_1)) \leq \zeta^2 h^2(1 - \frac{M\eta\mu}{2})^{t_2-t_1}$$

We have

$$\mathbb{P}\left(A(t_2+1)\right) \leq M\eta$$

With an abuse of notation, suppose $\nabla L\left(x(t_2)\right) \leq 4\zeta\rho$, which implies $\|x(t_2) - \Phi(x(t_2))\| = O(\rho)$.

Subphase C After $\|\nabla L\left(x(t_2)\right)\| \leq 4\zeta\rho$, it becomes difficult to prove the loss continue to decrease. We proceed by consider a quadratic approximation. Now consider

$$\begin{aligned}
x(t+1) &= x(t) - M\eta\nabla L_k\left(x(t) + \rho\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|}\right) \\
&= x(t) - M\eta\nabla L_k\left(x(t)\right) - M\eta\rho\nabla^2 L_k(x(t))\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|} + O(M\eta\rho^2) \\
&= x(t) - M\eta\nabla L_k\left(x(t)\right) - M\eta\rho\Lambda_k w_k w_k^T\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|} + O(M\eta\rho^2)
\end{aligned}$$

Iteratively define $t_{2,j}$, for $1 \leq j \leq 3$, $t_{2,1} = t_2$.

For $j \leq 2$, inductively suppose $\|x(t_{2,j}) - \Phi(x(t_{2,j}))\| \leq O(\rho^{(j+1)/2} + M\eta\rho)$, let $p_j = \Phi(x(t_{2,j}))$, further assume $\nabla L_k(\Phi(x(t_{2,j}))) = v_i v_i^T$. Further suppose N as the normal space of Γ at p_j and T as the tangent space. Define P_N and P_T as projection to the space.

We have $\|P_T(x(t_{2,j}) - p_j)\| = O((x(t_{2,j}) - p_j)^2)$.

Consider $t_{2,j} \leq t \leq t_{2,j} + \frac{\rho^{\frac{j-2}{2}}}{M\eta}$
Define event $A_j(t) = \{\|x(\tau) - p_j\| \geq \rho^{(j+2)/2} | \forall \tau \leq t\}$
We have $\|P_N(x(t) - p_j)\| = O(\rho^{k+1} + M\eta\rho^2(t - t_{2,j})) \leq O(\rho^{(j+2)/2})$
Under $A_j(t)$, we would have $\|P_T(x(t) - p_j)\| = O(\|x(t) - p_j\|)$
By Lemma 34, we have

$$\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|} = s_k(t)w_k + O(\|x(t) - p_j\|)$$

We also have

$$s_k(t) \neq sign(w_k^T(x(t) - p_j)) \Rightarrow \|w_k^T(x(t) - p_j)\| \leq \|x(t) - p_j\|^{3/2}$$

Now by Taylor Expansion,

$$\begin{aligned}
x(t+1) - p_j &= (x(t) - p_j) - M\eta\Lambda_k w_k w_k^T(x(t) - p_j) + O(M\eta\|x(t) - p_j\|^2) \\
&\quad - M\eta\rho\Lambda_k s_k(t) w_k w_k^T w_k + O(M\eta\rho\|x(t) - p_j\|) + O(M\eta\rho^2) \\
&= (x(t) - p_j) - M\eta\Lambda_k w_k w_k^T(x(t) - p_j) - M\eta\rho\Lambda_k s_k(t) w_k w_k^T w_k + O(M\eta\rho^2)
\end{aligned}$$

We then have

$$\begin{aligned}
\mathbb{E}[\|x(t+1) - p_j\|^2 | x(t)] = \|x(t) - p_j\|^2 + 2\eta^2 \sum_k \Lambda_k^2 |w_k^T(x(t) - p_j)|^2 + \eta^2\rho^2 \sum_k \Lambda_k^2 + O(M\eta\rho^2\|x(t) - p_j\|) \\
- \eta \sum_k \Lambda_k |w_k^T(x(t) - p_j)|^2 - \eta\rho \sum_k \Lambda_k s_k(t) w_k^T(x(t) - p_j)
\end{aligned}$$

We lower bound $\sum_k \Lambda_k s_k(t) v_i^T(x(t) - p_j)$ again by Lemma 34, there exists C such that

$$\begin{aligned}
\sum_k \|v_i\| s_k(t) v_i^T(x(t) - p_j) &\geq \sum_k \|v_i^T(x(t) - p_j)\| - 2\kappa \sum_k \|v_i\| \|x(t) - p_j\|^{3/2} \\
&\geq C\|x(t) - p_j\|.
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\|x(t+1) - p_j\|^2 | x(t)] &\leq \|x(t) - p_j\|^2 + 2\eta^2 C_1 \|x(t) - p_j\|^2 + \eta^2\rho^2 C_1 + \eta\rho^2 C_1 \|x(t) - p_j\| \\
&\quad - \eta C\|x(t) - p_j\|^2 - 3\eta\rho C\|x(t) - p_j\| \\
&\leq \|x(t) - p_j\|^2 - 2\eta\rho C\|x(t) - p_j\| + \eta^2\rho^2 C_1
\end{aligned}$$

Hence we have if $\|x(t) - p_j\| \geq O(\eta\rho)$

$$\mathbb{E}[\|x(t+1) - p_j\| | x(t)] \leq \|x(t) - p_j\| - \alpha\eta\rho.$$

which implies,

$$\mathbb{E}[\|x(t+1) - p_j\| \mathbf{1} A_j(t+1)] \leq \mathbb{E}[\|x(t+1) - p_j\| \mathbf{1} A_j(t)] \leq \mathbb{E}[\|x(t) - p_j\| \mathbf{1} A_j(t)] - \alpha\eta\rho P(A_j(t))$$

Hence

$$\alpha\rho^{j/2}P(A_j(t_{2,j}+\frac{\rho^{(j-2)/2}}{\eta})\leq\alpha\eta\rho\sum_{t=t_{2,j}}^{t=t_{2,j}+\frac{\rho^{(j-2)/2}}{\eta}}P(A_j(t))\leq\|x(t_{2,j})-\Phi(x(t_{2,j}))\|=O(\rho^{(j+1)/2})$$

Hence with probability $O(\sqrt{\rho})$, there exists $t_{2,j+1}$, such that $\|x(t_{2,j+1})-p_j\|\leq O(\rho^{(j+2)/2})$, which further implies $\|x(t_{2,j+1})-\Phi(x(t_{2,j+1}))\|\leq O(\rho^{(j+2)/2})$ using Lemma 30.

Define $t_3=t_{2,3}$

D.6.2. PHASE II: PROOF OF THEOREM 52

We will inductively prove the following claim holds with probability $1-O(\eta\rho)$,

$$\|\Phi(x(t))-X(\eta\rho^2t)\|\leq O((\eta+\rho)\log(1/\rho))$$
$$\|x(t)-\Phi(x(t))\|=O(\rho(\eta+\rho)\log(1/\rho))$$

To be more precise, the induction mainly consists of two parts. The first part shows that $x(t)$ will stay close to the manifold with large probability and the second part shows the direction $\Phi(x(t))$ moves.

To be more succinct with previous section, we abuse notation and suppose the iteration starts at t_3.

Part I: Convergence Near Manifold We have $\|x(t_3)-\Phi(x(t_3))\|=O(\eta\rho+\rho^2)$. By induction hypothesis, we have $x(t)\in K^h$.

According to Lemma 33

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

Using the same argument in previous section, we have for sufficiently large constant A, if $A(\eta\rho+\rho^2)\log(1/\eta\rho)\geq\|x(t)-\Phi(x(t))\|\geq A(\eta\rho+\rho^2)$, then there exists constant α,B independent of A,

$$\mathbb{E}[\|x(t+1)-\Phi(x(t+1))\||x(t)]\leq\|x(t)-\Phi(x(t))\|-\alpha\eta\rho$$
$$\|(x(t+1)-\Phi(x(t+1)))-(x(t)-\Phi(x(t)))\|\leq B\eta\rho$$

.

We then have

$$\begin{aligned}Pr(y(t+1)\geq A(\eta\rho+\rho^2)\log(1/\eta\rho))=\sum_{\tau=t_3}^{t}&Pr(y(t+1)\geq A(\eta\rho+\rho^2)\log(1/\eta\rho)\\&\text{and } y(\tau)<A(\eta\rho+\rho^2)\text{ and}\\&\forall t+1\geq\tau'\geq\tau+1,y(\tau')>A(\eta\rho+\rho^2)\end{aligned}$$

We then consider each term,

$$\begin{aligned}&Pr(y(t+1)\geq A(\eta\rho+\rho^2)\log(1/\eta\rho)\text{ and }y(\tau)<A(\eta\rho+\rho^2)\text{ and }\forall t+1\geq\tau'\geq\tau+1,y(\tau')>A(\eta\rho+\rho^2))\\\leq&Pr(y(t+1)\geq A(\eta\rho+\rho^2)\log(1/\eta\rho)\text{ and }\forall t+1\geq\tau'\geq\tau+1,y(\tau')>A(\eta\rho+\rho^2)\\&|A(\eta\rho+\rho^2)<y(\tau+1)<(A+B)(\eta\rho+\rho^2))\end{aligned}$$

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

$$\tilde{y}(\tau') = \begin{cases} \|x(\tau') - \Phi(x(\tau'))\|, & \text{if } \tilde{y}(\tau'-1) = \|x(\tau'-1) - \Phi(x(\tau'-1))\| > A(\eta\rho + \rho^2) \\ \tilde{y}(\tau'-1) - \alpha\eta\rho, & \text{if otherwise} \end{cases}$$

Then clearly

$$Pr(y(t+1) \geq A(\eta\rho+\rho^2)\log(1/\eta\rho) \text{ and } \forall t+1 \geq \tau' \geq \tau+1, y(\tau') > A(\eta\rho+\rho^2)$$
$$|A(\eta\rho+\rho^2) < y(\tau+1) < (A+B)(\eta\rho+\rho^2)) \leq Pr(\tilde{y}(t+1) \geq A(\eta\rho+\rho^2)\log(1/\eta\rho))$$

We have

$$|\tilde{y}(t+1) - \tilde{y}(t)| \leq B\eta\rho$$
$$\mathbb{E}[\tilde{y}(t+1)] - \mathbb{E}[\tilde{y}(t)] \leq -\alpha\eta\rho$$

Now applying Azuma-Hoeffding bound(Lemma 56), we have

$$\begin{aligned} P(\tilde{y}(t+1) \geq \tilde{y}(\tau+1) - \alpha\eta\rho(t-\tau) + h) &\leq P(\tilde{y}(t+1) \geq \mathbb{E}[\tilde{y}(t+1)] + h) \\ &\leq exp(-\frac{2h^2}{(t-\tau)B^2\eta^2\rho^2}) \end{aligned}$$

Choosing $h = \alpha\eta\rho(t-\tau) - y(\tau+1) + A(\eta\rho+\rho^2)\log(1/\eta\rho)$

$$\begin{aligned} P(\tilde{y}(t+1) \geq A(\eta\rho+\rho^2)\log(1/\eta\rho)) &\leq exp(-2\frac{(\alpha\eta\rho(t-\tau) - y(\tau+1) + A(\eta\rho+\rho^2)\log(1/\eta\rho))^2}{(t-\tau)B^2\eta^2\rho^2}) \\ &\leq exp(-2\frac{(\alpha\eta\rho(t-\tau) - A(\eta\rho+\rho^2) + A(\eta\rho+\rho^2)\log(1/\eta\rho))^2}{(t-\tau)B^2\eta^2\rho^2}) \\ &\leq exp(-2\frac{(\alpha\eta\rho(t-\tau) - A(\eta\rho+\rho^2)\log(\eta\rho)/2)^2}{(t-\tau)B^2\eta^2\rho^2}) \\ &= exp(-2\frac{(\alpha(t-\tau) - A\log(\eta\rho)/2)^2}{(t-\tau)B^2}) \\ &\leq exp(-2\frac{\sqrt{A\alpha}}{B^2}\log(\eta\rho)) = \eta^{10}\rho^{10} \end{aligned}$$

We then have

$$Pr(y(t+1) \geq A(\eta\rho+\rho^2)\log(1/\eta\rho)) \leq \eta^{10}\rho^{10}(t-t_3) \leq \eta^8\rho^8$$

Part II: Direction of $\Phi(x(t+1)) - \Phi(x(t))$ We shall do a Taylor expansion and show that

$$\begin{aligned} x(t+1) &= x(t) - \eta\nabla L_k\left(x(t) + \rho\frac{\nabla L_k(x(t))}{\|\nabla L_k(x(t))\|}\right) \\ &= x(t) - \eta\nabla L_k(x(t)) - \eta\rho\nabla^2 L_k(x(t))\frac{\nabla L_k(x(t))}{\|\nabla L_k(x(t))\|} \\ &\quad - \eta\rho^2\partial^2(\nabla L_k)[\frac{\nabla L_k(x(t))}{\|\nabla L_k(x(t))\|}, \frac{\nabla L_k(x(t))}{\|\nabla L_k(x(t))\|}]/2 + O(\eta\rho^3) \end{aligned}$$

Now by induction we have,

$$\|x(t) - \Phi(x(t))\| = \tilde{O}(\eta\rho + \rho^2)$$

, then by Lemma 35, it implies

$$\|x(t+1) - x(t)\| = O(\eta\rho)$$

Then we have

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

Using Lemma 33, we have

$$\|\eta\partial\Phi(x(t))\nabla L_k\left(x(t)\right)\| = O(\eta\|x(t) - \Phi(x(t))\|^2) = O(\eta^3\rho^2 + \eta\rho^4)$$
$$\|\eta\rho\partial\Phi(x(t))\nabla^2 L_k\left(x(t)\right)\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|}\| = O(\eta\rho\|x(t) - \Phi(x(t))\|) = \tilde{O}(\eta^2\rho^2 + \eta\rho^3)$$

Hence

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

Notice finally that by Lemma 34

$$\begin{aligned}\partial\Phi(x(t))\partial^2(\nabla L_k)[\frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|}, \frac{\nabla L_k\left(x(t)\right)}{\|\nabla L_k\left(x(t)\right)\|}] &= \partial\Phi(\Phi(x(t)))\partial^2(\nabla L_k)[w_k, w_k] + O(\|x(t) - \Phi(x(t))\|) \\ &= P_{x,\Gamma}^{\top}\Phi(x(t))\nabla(\lambda_1(\nabla^2 L_k(\Phi(x(t))))) + O(\|x(t) - \Phi(x(t))\|)\end{aligned}$$

Hence we have

$$\Phi(x(t+1)) - \Phi(x(t)) = -\eta\rho^2 P_{x,\Gamma}^{\top}\Phi(x(t))\nabla(\lambda_1(\nabla^2 L_k(\Phi(x(t)))))/2 + \tilde{O}(\eta^2\rho^2 + \eta\rho^3)$$

Notice finally,

$$\mathbb{E}_k[MP_{x,\Gamma}^{\top}\Phi(x(t))\nabla(\lambda_1(\nabla^2 L_k(\Phi(x(t)))))] = P_{x,\Gamma}^{\top}\Phi(x(t))\nabla(\mathrm{Tr}(\nabla^2 L(\Phi(x(t)))))$$

Together with standard concentration bound, we would have $x(t)$ follows the Riemannian gradient flow on Γ for loss

$$\mathbb{E}_i\lambda_1\left(\nabla^2 L_k\left(\Phi(x(t))\right)\right) = \mathbb{E}_i\mathrm{Tr}\left(\nabla^2 L_k\left(\Phi(x(t))\right)\right) = \mathrm{Tr}\nabla^2 L\left(\Phi(x(t))\right).$$

D.6.3. PROOF OF COROLLARY

Proof [Proof of Corollary 17] We will do a Taylor expansion on $\mathbb{E}_k[L_{k,\rho}^{\max}](x)$. By Theorem 51 and 52, for $t > \eta\rho^2 T_3'$, we have $\|X(\eta\rho^2 t) - x(t)\| = \tilde{O}(\eta + \rho)$ and $\|x(t) - \Phi(x(t))\| = \tilde{O}(\eta\rho + \rho^2)$

$$\mathbb{E}_k[R_{k,\rho}^{\max}](x) = \max_v \mathbb{E}_k[\rho v^T\nabla L_k(x) + \rho^2 v^T\nabla^2 L_k(x)v/2] + O(\rho^3)$$

Then as $L_k(x) = \tilde{O}(\eta^2\rho^2 + \rho^4)$ and $\|v^T \nabla L_k(x)\| = \tilde{O}(\eta\rho + \rho^2)$, this implies

$$\begin{aligned}\mathbb{E}_k[R^{\max}_{k,\rho}](x) &= \rho^2 \mathbb{E}_k \max_v v^T \nabla^2 L(x) v/2 + \tilde{O}(\eta^2\rho^2 + \rho^3) \\ &= \rho^2 \mathbb{E}_k \max_v v^T \nabla^2 L(X(\eta\rho^2 t)) v/2 + \tilde{O}(\eta\rho^2 + \rho^3) \\ &= \rho^2 \mathrm{Tr}(X(\eta\rho^2 t))/2 + \tilde{O}(\eta\rho^2)\end{aligned}$$

■

Proof [Proof of Corollary 18]

Choose T such that $X(T)$ is sufficiently close to $X(\infty)$, such that $\mathrm{Tr}(X(T)) \le \mathrm{Tr}(X(\infty)) + 2\epsilon$

By corollary 17, we have $\|\mathbb{E}_k[R^{\max}_{k,\rho}](x(\lceil T/(\eta\rho^2)\rceil)) - \rho^2 \mathrm{Tr}(X(T))/2\| \le \tilde{O}(\eta\rho^2)$. This further implies $\|\mathbb{E}_k[R^{\max}_{k,\rho}](x(\lceil T/(\eta\rho^2)\rceil)) - \rho^2 \mathrm{Tr}(X(\infty))/2\| \le \epsilon\rho^2 + \tilde{O}(\eta\rho^2)$. We also have $\|L(x(\lceil T/(\eta\rho^2)\rceil))\| = O(\eta^2\rho^2)$. Then we can leverage Theorem 4 and Theorem 9 to get the desired bound. ■

D.7. Technical Lemmas

Lemma 53 (Cor. 4.3.15 in [11]) *Let $\Sigma, \hat{\Sigma} \in R^{D\times D}$ be symmetric and non-negative with eigenvalues $\lambda_1 \ge ... \ge \lambda_D$ and $\hat{\lambda}_1 \ge ... \ge \hat{\lambda}_D$, then for any i,*

$$|\hat{\lambda}_i - \lambda_i| \le \|\Sigma - \hat{\Sigma}\|_2$$

Definition 54 (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 55 *[Davis-Kahan $\sin(\theta)$ theorem [5]] Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \ge \ldots \ge \lambda_p$ and $\hat{\lambda}_1 \ge \ldots \ge \hat{\lambda}_p$ respectively. Fix $1 \le r \le s \le p$, let $d := s - r + 1$ 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 $\Delta := \min\left\{\max\{0, \lambda_s - \hat{\lambda}_{s+1}\}, \max\{0, \hat{\lambda}_{r-1} - \lambda_r\}\right\}$, where $\hat{\lambda}_0 := \infty$ and $\hat{\lambda}_{p+1} := -\infty$, we have for any unitary invariant norm $\|\cdot\|_*$,*

$$\Delta \cdot \|\sin\Theta(\hat{V}, V)\|_* \le \|\hat{\Sigma} - \Sigma\|_*.$$

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 $i \ne j \in [d]$. $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_d$ denotes the singular values of $\hat{V}^\top V$. $[\sin\Theta]_{ij}$ is defined as $\sin(\Theta_{ij})$.

Lemma 56 (Azuma-Hoeffding Bound) *Suppose Z_n is a super-martingale, suppose $-\alpha \le Z_{i+1} - Z_i \le \beta$, then for all $n > 0, a > 0$, we have*

$$P(|Z_n - Z_0| \ge a) \le 2exp(-a^2/(2N(\alpha+\beta)^2))$$

Lemma 57 ([19]) *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 43.

Lemma 58 *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 59
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,

$$(1-cb)b > a$$

$$\frac{1-cb}{2-ca-cb} = \frac{1}{\frac{a+b}{b} + \frac{b-a}{1-cb}} \geq \frac{1}{\frac{a+b}{b} + \frac{b^2-ab}{a}} = \frac{ab}{a^2+b^2} \geq \frac{a^2}{a^2+b^2}$$

Using the above anxiliary 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 60

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,

$$
\begin{aligned}
&cb^3 + a^2(2-cb-\frac{2}{3}ca) \leq \frac{b^2}{2-cb} + \frac{a^2(\frac{1}{2}a^2+b^2)}{b^2} \\
\iff &cb^3 + a^2(1-cb-\frac{2}{3}ca) \leq \frac{b^2}{2-cb} + \frac{a^4}{2b^2}
\end{aligned}
$$

We have the auxiliary inequality $\frac{1}{2-cb} > \frac{1}{2} + \frac{cb}{4}$.
Hence

$$
\begin{aligned}
&cb^3 + a^2(1-cb-\frac{2}{3}ca) \leq \frac{b^2}{2-cb} + \frac{a^4}{2b^2} \\
\Leftarrow&cb^3 + a^2(1-cb-\frac{2}{3}ca) \leq \frac{b^2}{2} + \frac{a^4}{2b^2} + \frac{cb^3}{4} \\
\Leftarrow&c(\frac{3b^3}{4} - ba^2 - \frac{2}{3}a^3) \leq \frac{b^2}{2} + \frac{a^4}{2b^2} - a^2
\end{aligned}
$$

Case 1 If $\frac{3b^3}{4} - ba^2 - \frac{2}{3}a^3 \leq 0$, then

$$
c(\frac{3b^3}{4} - ba^2 - \frac{2}{3}a^3) \leq 0 \leq \frac{b^2}{2} + \frac{a^4}{2b^2} - a^2
$$

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 58,$a > \frac{b}{2}$,$(b-a)^2(b+a) = (b^2-a^2)(b-a) \leq b^2(b-a) \leq \frac{b^3}{2}$

■

Lemma 61 *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

$$\begin{aligned}
F(a) &:= a^2(2 - cb - \frac{2}{3}ca) - 2a\sqrt{(b^2 + \frac{1}{2}a^2)(1-cb)} \\
S_a(c,b) &:= \{a | 0 \leq a \leq b, 0 < c \leq \frac{b-a}{b^2}, b^2 \geq a\sqrt{\frac{a^2+2b^2}{2(1-cb)}} \geq \frac{a^2+b^2}{2-ca-cb}\} \\
a_{min}(c,b) &:= \inf S_a(c,b) \\
a_{max}(c,b) &:= \sup S_a(c,b) \leq b - cb^2
\end{aligned}$$

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 $u := cb, v := \frac{a}{b}$, 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 $\sqrt{\frac{(1-u)^2}{2} + 1} \geq \sqrt{\frac{(1-u)^2+1}{2}} \geq (1-u)$,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 43,

$$\begin{aligned}F(a_{min}) &= a_{min}^2(2 - cb - \frac{2}{3}ca_{min}) - (1-cb)\frac{c(a_{min}^2+b^2)}{2-ca_{min}-cb} - ca_{min}^2(\frac{1}{2}a_{min}^2 + b^2)\frac{2-ca_{min}-cb}{(a_{min}^2+b^2)} \\ &\leq \frac{1}{c}(\frac{cb^2}{2-cb} - cb^2)\end{aligned}$$

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.

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.

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 60,

$$\begin{aligned}F(a_{max}) &= a_{max}^2(2 - cb - \frac{2}{3}ca_{max}) - (1-cb)cb^2 - ca_{max}^2(\frac{1}{2}a_{max}^2 + b^2)\frac{1}{b^2} \\ &\leq \frac{1}{c}(\frac{cb^2}{2-cb} - cb^2)\end{aligned}$$

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 $v := \frac{a_{max}}{b}, cb = 1 - v$
Then by Lemma 58, $\frac{1}{2} \leq v \leq \frac{9}{10}$

$$\begin{aligned}
F(a_{max}) &= a_{max}^2(2 - cb - \frac{2}{3}ca_{max}) - 2a_{max}\sqrt{(b^2 + \frac{1}{2}a_{max}^2)(1 - cb)} \\
&= b^2\left(v^2(2 - (1 - v) - \frac{2}{3}(1 - v)v) - 2v\sqrt{(1 + \frac{v^2}{2})v}\right)
\end{aligned}$$

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

$$\begin{aligned}
F(a) &\leq \max\left(F(a_{min}(c, b)), F(a_{max}(c, b))\right) \\
&\leq \frac{1}{c}(\frac{cb^2}{2 - cb} - cb^2)
\end{aligned}$$

$\blacksquare$