Understanding the Generalization Benefit of Normalization Layers: Sharpness Reduction

Kaifeng Lyu **Zhiyuan Li** **Sanjeev Arora**
Department of Computer Science
Princeton University
{klyu,zhiyuanli,arora}@cs.princeton.edu

Abstract

Normalization layers (e.g., Batch Normalization, Layer Normalization) were introduced to help with optimization difficulties in very deep nets, but they clearly also help generalization, even in not-so-deep nets. Motivated by the long-held belief that flatter minima lead to better generalization, this paper gives mathematical analysis and supporting experiments suggesting that normalization (together with accompanying weight-decay) encourages GD to reduce the sharpness of loss surface. Here "sharpness" is carefully defined given that the loss is scale-invariant, a known consequence of normalization. Specifically, for a fairly broad class of neural nets with normalization, our theory explains how GD with a finite learning rate enters the so-called Edge of Stability (EoS) regime, and characterizes the trajectory of GD in this regime via a continuous sharpness-reduction flow.

1 Introduction

Training modern deep neural nets crucially relies on normalization layers to make the training process less sensitive to hyperparameters and initialization. The two of the most popular normalization layers are Batch Normalization (BN) [55] for vision tasks and Layer Normalization (LN) [9] for language tasks. Recent works also proposed other normalization layers aiming for better performance, most notably including Group Normalization (GN) [120], Weight Normalization (WN) [102], Scaled Weight Standardization (SWS) [97, 53, 14], etc. Most normalization layers amount to a reparametrization of the neural net so that the loss becomes invariant to the scale of most parameters, and with a minor change, to *all* parameters: $\mathcal{L}(c\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w})$ for all scalings $c > 0$ [55, 7, 77]. The current paper assumes this scale-invariance for all parameters and analyzes the trajectory of gradient descent with *weight decay* (WD):

$$\boldsymbol{w}_{t+1} \leftarrow (1 - \hat{\eta}\hat{\lambda})\boldsymbol{w}_t - \hat{\eta}\nabla\mathcal{L}(\boldsymbol{w}_t). \tag{1}$$

The use of WD is a common practice that has been adopted in training state-of-the-art neural nets, such as ResNets [46, 47] and Transformers [29, 15]. Previous ablation studies showed that adding WD to normalized nets indeed leads to better generalization [126, 72, 125]. More notably, Liu et al. [83] conducted experiments of training ResNets initialized from global minima with poor test accuracy, and showed that SGD with WD escapes from those bad global minima and attains good test accuracy. In contrast, training with vanilla SGD yields significant generalization degradation.

In the traditional view, WD regularizes the model by penalizing the parameter norm, but this may appear nonsensical for scale-invariant loss because one can scale down the norm arbitrarily without changing the loss value. However, the scale of the parameter *does* matter in backward propagation, and thus WD can affect the training dynamics. In particular, simple calculus shows $\nabla\mathcal{L}(\boldsymbol{w}) = \frac{1}{\|\boldsymbol{w}\|_2}\nabla\mathcal{L}(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}) \propto \frac{1}{\|\boldsymbol{w}\|_2}$ and $\nabla^2\mathcal{L}(\boldsymbol{w}) = \frac{1}{\|\boldsymbol{w}\|_2^2}\nabla^2\mathcal{L}(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}) \propto \frac{1}{\|\boldsymbol{w}\|_2^2}$, so WD is in effect trying to enlarge the gradient and Hessian in training. This makes the training dynamics very different from unnormalized nets and requires revisiting classical convergence analyses [77, 78, 84, 80].

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

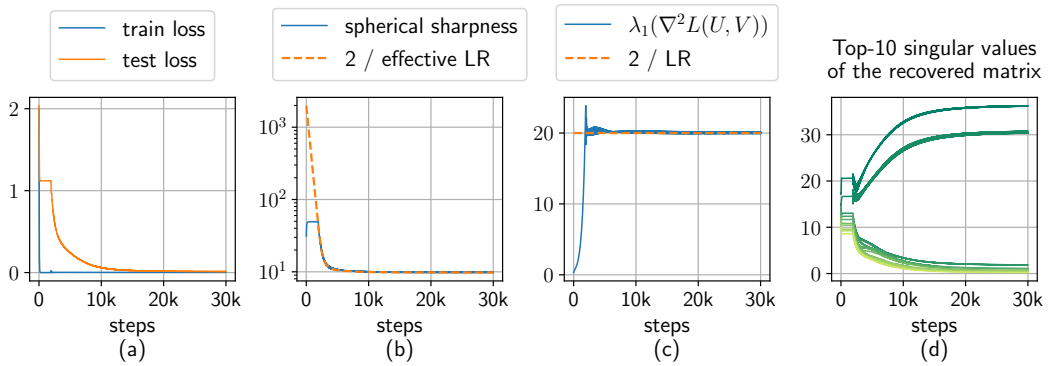


Figure 1: Experiment on overparameterized matrix completion with Batch Normalization. Given 800 (32%) entries Ω of a rank-2 matrix $\boldsymbol{M} \in \mathbb{R}^{50\times 50}$, use GD+WD to optimize the loss $\mathcal{L}(\boldsymbol{U},\boldsymbol{V}) := \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}(\mathrm{BN}([\boldsymbol{U}\boldsymbol{V}^\top]_{i,j}) - M_{i,j})^2$, where $\boldsymbol{U},\boldsymbol{V} \in \mathbb{R}^{50\times 50}$ (thus no explicit constraint on rank). Starting from step $\sim$ 2k, spherical sharpness drops significantly (**b**), which encourages low-rank (**d**) and causes the test loss (MSE of all entries) to decrease from 1.12 to 0.013 (**a**). See also Appendix P.1.

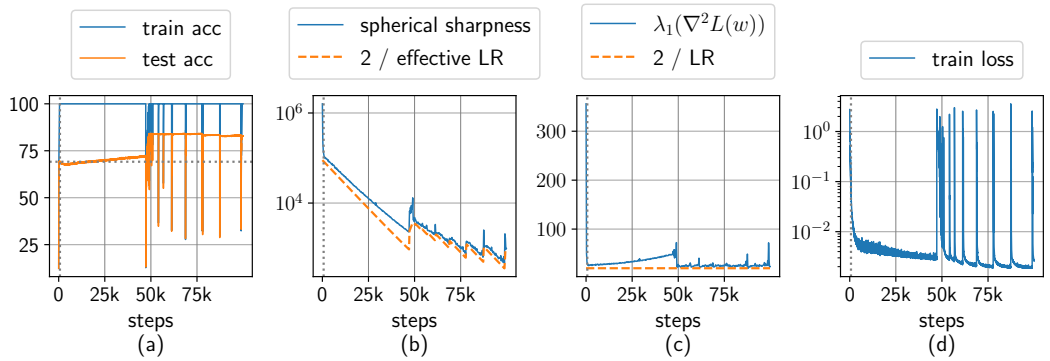


Figure 2: In training a smooth and scale-invariant VGG-11 on CIFAR-10 with (full-batch) GD+WD, the spherical sharpness keeps decreasing and the test accuracy keeps increasing. BN is added after every linear layer to ensure scale-invariance. 100% training accuracy is achieved after $\sim$ 680 steps (dotted line), but as the training continues for $\sim$ 47k steps, the spherical sharpness keeps decreasing (**b**) and the test accuracy increases from 69.1% to 72.0% (**a**). Then the training exhibits destabilization but the test accuracy is further boosted to 84.3%. Removing either of BN or WD eliminates this phenomenon; see Appendices P.4 and P.5.

The current paper aims to improve mathematical understanding of how normalization improves generalization. While this may arise from many places, we focus on the training dynamics of (full-batch) GD (1) as a necessary first step towards understanding SGD, and show that the interplay between normalization and WD provably induces an implicit bias to persistently reduce the *sharpness* of the local loss landscape during the training process, which we call the *sharpness-reduction bias*.

It is long believed that flatter minima generalize better [50, 63, 95], but the notion of sharpness/flatness makes sense only if it is carefully defined in consideration of various symmetries in neural nets. One of the most straightforward measures of sharpness is the maximum eigenvalue of Hessian, namely $\lambda_1(\nabla^2\mathcal{L}(\boldsymbol{w}_t))$. But for normalized nets, this sharpness measure is vulnerable to weight rescaling, because one can scale the weight norm to make a minimizer arbitrarily flat [31]. Also, this sharpness measure may not decrease with the number of training steps: an empirical study by Cohen et al. [24] shows that for various neural nets (including normalized nets), GD has an overwhelming tendency to persistently increase $\lambda_1(\nabla^2\mathcal{L}(\boldsymbol{w}_t))$ until it reaches the *Edge of Stability (EoS) regime*, a regime where $\lambda_1(\nabla^2\mathcal{L}(\boldsymbol{w}_t))$ stays around $2/\hat{\eta}$ ($\hat{\eta}$ is the learning rate). See also Figure 2c.

1.1 Our Contributions

The sharpness measure we use in this paper takes care of the scale-invariance in normalized nets. We are motivated by our experiments on matrix completion (with BN) and CIFAR-10, where our sharpness measure decreases as the training proceeds, and the generalization improves accordingly; see Figures 1 and 2. We note that techniques from previous works [92, 95, 37] can be easily adopted here to establish a PAC-Bayes bound on the test error, where our sharpness measure appears as an additive term (see Appendix C).

Definition 1.1 (Spherical Sharpness)**.** For a scale-invariant loss $\mathcal{L}(\boldsymbol{w})$ (i.e., $\mathcal{L}(c\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w})$ for all $c > 0$), the spherical sharpness at $\boldsymbol{w} \in \mathbb{R}^D$ is defined by $\lambda_1(\nabla^2 \mathcal{L}(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}))$, the maximum eigenvalue of the Hessian matrix after projecting $\boldsymbol{w}$ onto the unit sphere.

Based on Definition 1.1, we study the aforementioned sharpness-reduction bias in training normalized nets with GD+WD (defined in (1)). For constant learning rate $\hat{\eta}$ and weight decay $\hat{\lambda}$, we can rewrite this rule equivalently as Projected Gradient Descent (PGD) on the unit sphere with *adaptive* learning rates, $\boldsymbol{\theta}_{t+1} \leftarrow \Pi(\boldsymbol{\theta}_t - \tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t))$, where $\boldsymbol{\theta}_t := \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_2}$ is the direction of $\boldsymbol{w}_t$, and $\tilde{\eta}_t$ is the "effective" learning rate at step t (see Lemma 3.1). We call $\tilde{\eta}_t$ adaptive because it can be shown to resemble the behaviors of adaptive gradient methods (e.g., RMSprop [49]): $\tilde{\eta}_t$ increases when gradient is small and decreases when gradient is large (Figure 3). Our main contributions are as follows:

1. After $\boldsymbol{\theta}_t$ reaches a point near the manifold of minimizers of $\mathcal{L}$, we theoretically show that the effective learning rate $\tilde{\eta}_t$ increases until GD enters a regime where $2/\tilde{\eta}_t$ roughly equals to the spherical sharpness (or equivalently $2/\hat{\eta} \approx \lambda_1(\nabla^2 \mathcal{L}(\boldsymbol{w}_t))$), namely the EoS regime (Section 4.1).
2. In the EoS regime, we show that for GD with a small (but finite) learning rate, $\boldsymbol{\theta}_t$ oscillates around the manifold and moves approximately along a sharpness-reduction flow, which is a gradient flow for minimizing spherical sharpness on the manifold (with gradient-dependent learning rate) (Section 4.2).
3. As an application of our theory, we show that for linear regression with BN, GD+WD finds the minimizer that corresponds to the linear model with minimum weight norm, which looks surprisingly the same as the conventional effect of WD but is achieved through the completely different sharpness-reduction mechanism (Section 5).
4. We experimentally verified the sharpness-reduction phenomenon predicted by our theorem and its benefits to generalization on CIFAR-10 with VGG-11 and ResNet-20, as well as matrix completion with BN (Appendix P).
5. We generalize our theoretical results of sharpness-reduction bias to a broader class of adaptive gradient methods, most notably a variant of RMSprop with scalar learning rate (Appendix B).

Technical Contribution. Our proof technique is novel and may have independent interest to the ML community. The main challenge is that we need to analyze the implicit bias of GD in the EoS regime which crucially relies on step size being finite — this is in sharp contrast to many previous works on implicit bias of GD [107, 106, 87, 59, 43, 42, 76, 100, 4, 22, 79, 88, 101, 108, 38] where the same bias exists at infinitesimal LR. Our analysis is inspired by a previous line of works [13, 25, 81] showing that label noise can drive SGD to move on the minimizer manifold along the direction of minimizing the trace of Hessian. We borrow a few lemmas from those analyses, but the overall proof strategy is very different because our setting does not even have any stochastic gradient noise. Instead, we connect the dynamics in the EoS regime to power methods and show that GD oscillates around the minimizer manifold. This oscillation then becomes a driving power that pushes the parameter to move on the manifold. Finally, we analyze the speed of this movement by modeling two key parameters of the dynamics as a 1-dimensional Hamiltonian system (Figure 6). To the best of our knowledge, we are the first to provide theoretical proof for a sharpness measure to decrease during the standard GD training, without any additional regularization (e.g., label noise [13, 25, 81]) and without involving uncommon variants of GD (e.g., normalized GD or non-smooth wrappings on the loss function [8]).

2 Related Works

Sharpness and Generalization. It has been long believed that flat minima generalize better [50]. Several empirical studies [63, 74, 117, 57] verified the positive correlation between flatness and generalization. Neyshabur et al. [95] justified this via PAC-Bayes theory [92]. Several other theoretical papers explored the generalization properties of flat minima specifically for two-layer nets [13, 94, 44, 81, 30] and deep linear nets [93]. Jiang et al. [60] conducted extensive experiments for all existing generalization measures to evaluate their correlation and causal relationships with generalization error, concluding that sharpness-based measures perform the best overall. In light of this, Foret et al. [37] proposed SAM algorithm to improve the generalization by minimizing the sharpness. Despite so many positive results on sharpness-based measures, a common issue of many works is that the measures may suffer from sensitivity to rescaling of parameters in deep nets [31]. Another issue is that the minima could lie in asymmetric valleys that are flat on one side and sharp on the other [45].

Understanding Normalization Layers. The benefits of normalization layers can be shown in various aspects. A series of works studied the forward propagation of deep nets at random initialization, showing that normalization layers stabilize the growth of intermediate layer outputs with depth [14, 10, 28], provably avoid rank collapse [26] and orthogonalize representations [27]. Although these works mainly focused on BN [55], Lubana et al. [85], Labatie et al. [67] provided thorough discussions on the applicability of these arguments to other normalization layers. It is also believed that BN has a unique regularization effect through the noise in batch statistics [86, 111, 104]. Several other works argued that normalization layers lead to a smoothening or preconditioning effect of the loss landscape [103, 12, 39, 61, 82, 68], which may help optimization. By analyzing the training dynamics, Arora et al. [7] rigorously proved that normalization yields an auto-tuning effect of the effective learning rate $\tilde{\eta}_t$, which makes the asymptotic speed of optimization much less sensitive to the learning rate and initialization. In linear regression settings, Cai et al. [16], Kohler et al. [65] showed that training with BN leads to a faster convergence rate; Wu et al. [119] studied the implicit regularization effect of WN [102]. For two-layer nets with normalization, Ma and Ying [90] derived a mean-field formulation of the training dynamics; Dukler et al. [33] proved a convergence rate via NTK-based analysis. The current paper focuses on the interplay between normalization and WD during training, whereas all the above works either do not analyze the dynamics or assume no WD.

Interplay Between Normalization and WD. A common feature of normalization layers (including but not limited to BN, WN, LN, GN, SWS) is that they make the loss invariant to the scale of layer weights. In presence of both scale-invariance and WD, training dynamics can go out of the scope of the classical optimization theory, e.g., one can train the net to small loss even with learning rates exponentially increasing [77]. A series of works investigated into the interplay between normalization and WD and argued that the training dynamic with SGD eventually reaches an "equilibrium" state, where the parameter norm [78, 113, 21] and the size of angular update [114] become stable. Li et al. [78], Wang and Wang [115] provided empirical and theoretical evidence that the function represented by the net also equilibrates to a stationary distribution that is independent of initialization. This could be related to Liu et al. [83]'s experiments on the ability of SGD with WD to escape from bad initialization, but it remains unclear why the generalization should be good at the equilibrium state. In this paper, we focus on (full-batch) GD, which is the most basic and important special case of SGD.

3 Preliminaries

Let $\mathbb{S}^{D-1} := \{\boldsymbol{\theta} \in \mathbb{R}^D : \|\boldsymbol{\theta}\|_2 = 1\}$ be the unit sphere equipped with subspace topology. We say a loss function $\mathcal{L}(\boldsymbol{w})$ defined on $\mathbb{R}^D \setminus \{\mathbf{0}\}$ is *scale-invariant* if $\mathcal{L}(c\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w})$ for all $c > 0$. In other words, the loss value does not change with the parameter norm. For a differentiable scale-invariant function $\mathcal{L}(\boldsymbol{w})$, the gradient is (-1)-homogeneous and it is always perpendicular to $\boldsymbol{w}$, i.e., $\nabla\mathcal{L}(c\boldsymbol{w}) = c^{-1}\nabla\mathcal{L}(\boldsymbol{w})$ for all $c > 0$ and $\langle \nabla\mathcal{L}(\boldsymbol{w}), \boldsymbol{w}\rangle = 0$ (see Lemma D.1).

The focus of this paper is the dynamics of GD+WD on scale-invariant loss. (1) gives the update rule for learning rate (LR) $\hat{\eta}$ and weight decay (WD) $\hat{\lambda}$. We use $\boldsymbol{\theta}_t := \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_2}$ to denote the projection of $\boldsymbol{w}_t$ onto $\mathbb{S}^{D-1}$ at step t. We write GD+WD on scale-invariant loss as a specific kind of Projected Gradient Descent (PGD) and define the *effective learning rate* to be the LR $\tilde{\eta}_t := \frac{\hat{\eta}}{(1-\hat{\eta}\hat{\lambda})\|\boldsymbol{w}_t\|_2^2}$ that appears in the update rule of PGD. This notion is slightly different from the effective learning rate $\frac{\hat{\eta}}{\|\boldsymbol{w}_t\|_2^2}$ defined in previous works [113, 52, 7], but ours is more convenient for our analysis.

Lemma 3.1. *When the parameters $\boldsymbol{w}_t$ are updated as (1), $\boldsymbol{\theta}_t$ satisfies the following equation:*

$$\boldsymbol{\theta}_{t+1} = \Pi(\boldsymbol{\theta}_t - \tilde{\eta}_t \nabla\mathcal{L}(\boldsymbol{\theta}_t)), \tag{2}$$

*where $\tilde{\eta}_t := \frac{\hat{\eta}}{(1-\hat{\eta}\hat{\lambda})\|\boldsymbol{w}_t\|_2^2}$ is called the **effective learning rate** at step t, and $\Pi : \boldsymbol{w} \mapsto \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$ is the projection operator that projects any vector onto the unit sphere.*

4 GD+WD on Scale-Invariant Loss Functions

This section analyzes GD+WD (1) on a scale-invariant loss $\mathcal{L}(\boldsymbol{w})$, in particular what happens after approaching a manifold of local minimizers. Section 4.1 analyzes the dynamics in the stable regime, where loss is guaranteed to decrease monotonically, and Theorem 4.2 suggests $\boldsymbol{w}_t$ can get close to a local minimizer at some time t_0. We show that the effective LR keeps increasing after t_0, causing

GD+WD to eventually leave this stable regime and enter a new regime which we call the Edge of Stability (EoS). In Section 4.2, we establish our main theorem, which connects the dynamics of $\boldsymbol{w}_t$ in the EoS regime to a sharpness-reduction flow.

4.1 GD+WD Eventually Leaves the Stable Regime

A standard step of analyzing optimization methods is to do Taylor expansion locally for the loss function, and show that how the optimization method decreases the loss using a *descent lemma*. In our case of scale-invariant loss functions, we use $\boldsymbol{H}(\boldsymbol{w}) := \nabla^2 \mathcal{L}(\boldsymbol{w}) \in \mathbb{R}^{D\times D}$ to denote the Hessian matrix of $\mathcal{L}$ at $\boldsymbol{w} \in \mathbb{R}^D$, and $\lambda_1^{\mathrm{H}}(\boldsymbol{w}) := \lambda_1(\boldsymbol{H}(\boldsymbol{w}))$ to denote the top eigenvalue of $\boldsymbol{H}(\boldsymbol{w})$.

Lemma 4.1 (Descent Lemma)**.** *For scale-invariant loss $\mathcal{L}(\boldsymbol{w})$, at step t of GD+WD we have*

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \le \mathcal{L}(\boldsymbol{\theta}_t) - \tilde{\eta}_t(1 - \tilde{\eta}_t \lambda_{\max}^{(t)}/2)\|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2.$$

where $\lambda_{\max}^{(t)} := \sup_{\alpha\in[0,\tilde{\eta}_t]} \left\{\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}_t - \alpha\nabla\mathcal{L}(\boldsymbol{\theta}_t))\right\}$ is an upper bound of spherical sharpness locally.

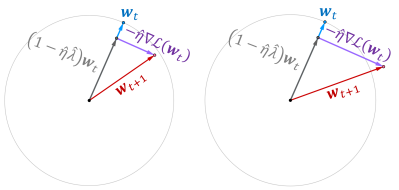
This descent lemma shows that the training loss $\mathcal{L}(\boldsymbol{\theta}_t)$ keeps decreasing as long as the effective LR $\tilde{\eta}_t$ is smaller than $2/\lambda_{\max}^{(t)}$, We call the regime of $\tilde{\eta}_t < 2/\lambda_{\max}^{(t)}$ as the *stable regime* of GD+WD. If $\tilde{\eta}_t \approx 2/\lambda_{\max}^{(t)}$ with a small difference, then we call it as the *Edge of Stability (EoS) regime*. We remark that this condition for EoS regime is essentially the *same* as $\hat{\eta} \approx 2/\lambda_1^{\mathrm{H}}(\boldsymbol{w})$ in Cohen et al. [24]'s definition because $\tilde{\eta}_t \cdot \lambda_{\max}^{(t)} \approx \hat{\eta} \cdot \lambda_1^{\mathrm{H}}(\boldsymbol{w})$; see Appendix G.3.

Fix an initial point $\boldsymbol{w}_0 \in \mathbb{R}^D \setminus \{\boldsymbol{0}\}$. Now we aim to characterize the dynamics of GD+WD when LR $\hat{\eta}$ and WD $\hat{\lambda}$ are small enough. The convergence rate of GD+WD has been analyzed by Li et al. [80]. Here we present a variant of their theorem that bounds both the gradient and effective LR.

Theorem 4.2 (Variant of Theorem D.2, Li et al. [80])**.** *Let $\mathcal{L}(\boldsymbol{w})$ be a scale-invariant loss function and $\rho_2 := \sup\{\|\nabla^2\mathcal{L}(\boldsymbol{w})\|_2 : \boldsymbol{w} \in \mathbb{S}^{D-1}\}$ be the smoothness constant of $\mathcal{L}$ restricted on the unit sphere. For GD+WD (1) with $\hat{\eta}\hat{\lambda} \le 1/2$ and $\tilde{\eta}_0 \le \frac{1}{\pi^2\rho_2(1-\hat{\eta}\hat{\lambda})}$, let $T_0 := \left\lceil \frac{1}{2\hat{\eta}\hat{\lambda}} \ln \frac{\|\boldsymbol{w}_0\|_2^2}{\rho_2\pi^2\hat{\eta}} \right\rceil$ steps, there must exist $0 \le t \le T_0$ such that $\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2^2 \le 8\pi^4\rho_2^2\hat{\lambda}\hat{\eta}$ and $\tilde{\eta}_t \le \frac{2}{\pi^2\rho_2(1-\hat{\eta}\hat{\lambda})}$.*

Theorem 4.2 shows that for some $t_0 \le T_0$, $\|\nabla\mathcal{L}(\boldsymbol{\theta}_{t_0})\|_2^2 \le O(\hat{\lambda}\hat{\eta})$ and $\tilde{\eta}_{t_0} \le \frac{1}{\pi^2\rho_2} < \frac{2}{\rho_2}$, which means $\boldsymbol{\theta}_{t_0}$ is an approximate first-order stationary point of $\mathcal{L}$ on the unit sphere. This does not guarantee that $\boldsymbol{\theta}_{t_0}$ is close to any global minimizer, but in practice the training loss rarely gets stuck at a non-optimal value when the model is overparameterized [70, 96, 71, 125]. We are thus motivated to study the case where $\boldsymbol{\theta}_{t_0}$ not only has small gradient $\|\nabla\mathcal{L}(\boldsymbol{\theta}_{t_0})\|_2^2 \le O(\hat{\lambda}\hat{\eta})$ but also is close to a local minimizer $\boldsymbol{\theta}^* \in \mathbb{S}^{D-1}$ of $\mathcal{L}$ in the sense that $\|\boldsymbol{\theta}_{t_0} - \boldsymbol{\theta}^*\|_2 \le O((\hat{\lambda}\hat{\eta})^{1/2})$ (assuming smoothness, the latter implies the former).

As the gradient is small near the local minimizer $\boldsymbol{\theta}^*$, starting from step t_0, the norm of $\boldsymbol{w}_t$ decreases due to the effect of WD. See Figure 3a. Since the effective LR is inversely proportional to $\|\boldsymbol{w}_t\|_2^2$, this leads to the effective LR to increase. Then Theorem 4.4 will show that the GD+WD dynamic eventually leaves the stable regime at some time $t_1 > t_0$, and enters the EoS regime where $\tilde{\eta}_t \approx 2/\lambda_{\max}^{(t)}$.



(a) (b)

Figure 3: The norm of $\boldsymbol{w}_t$ decreases when gradient is small and increases when gradient is large.

To establish Theorem 4.4, we need to assume that $\mathcal{L}$ satisfies Polyak-Łojasiewicz (PL) condition locally, which is a standard regularity condition in the optimization literature to ease theoretical analysis around a minimizer. Intuitively, PL condition guarantees that the gradient grows faster than a quadratic function as we move a parameter $\boldsymbol{\theta}$ away from $\boldsymbol{\theta}^*$. Note that PL condition is strictly weaker than convexity as the function can still be non-convex under PL condition (see, e.g., [62]).

Definition 4.3 (Polyak-Łojasiewicz Condition)**.** For a scale-invariant loss $\mathcal{L}(\boldsymbol{w})$ and $\mu > 0$, we say that $\mathcal{L}$ satisfies μ-Polyak-Łojasiewicz condition (or μ-PL) locally around a local minimizer $\boldsymbol{\theta}^*$ on $\mathbb{S}^{D-1}$ if for some neighborhood $U \subseteq \mathbb{S}^{D-1}$ of $\boldsymbol{\theta}^*$, $\forall \boldsymbol{\theta} \in U : \frac{1}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2^2 \ge \mu \cdot (\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*))$.

Theorem 4.4. *Let $\mathcal{L}(\boldsymbol{w})$ be a $\mathcal{C}^2$-smooth scale-invariant loss that satisfies μ-PL around a local minimizer $\boldsymbol{\theta}^*$ on the unit sphere, and $\rho_2 := \sup\{\|\nabla^2\mathcal{L}(\boldsymbol{w})\|_2 : \boldsymbol{w} \in \mathbb{S}^{D-1}\}$. For GD+WD on $\mathcal{L}(\boldsymbol{w})$ with learning rate $\hat{\eta}$ and weight decay $\hat{\lambda}$, if at some step t_0, $\|\boldsymbol{\theta}_{t_0} - \boldsymbol{\theta}^*\|_2 \le O((\hat{\lambda}\hat{\eta})^{1/2})$ and $\tilde{\eta}_{t_0} \le \frac{2}{\rho_2} < \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}^*)}$, and if $\hat{\lambda}\hat{\eta}$ is small enough, then there exists a time $t_1 > t_0$ such that $\|\boldsymbol{\theta}_{t_1} - \boldsymbol{\theta}^*\|_2 = O((\hat{\lambda}\hat{\eta})^{1/2})$ and $\tilde{\eta}_{t_1} = \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}^*)} + O((\hat{\lambda}\hat{\eta})^{1/2})$.*

4.2 Dynamics at the Edge of Stability

From the analysis in the previous subsection, we know that $\boldsymbol{\theta}_t$ can get close to a local minimizer $\boldsymbol{\theta}^*$ and enter the EoS regime at some step t_1. But what happens after t_1?

Figure 4 gives a warm-up example on a 3D scale-invariant loss $\mathcal{L} : \mathbb{R}^3 \setminus \{\mathbf{0}\} \to \mathbb{R}$, where the black line is a manifold Γ consisting of all the minimizers. In training with GD+WD, $\boldsymbol{\theta}_t$ first goes close to a local minimizer $\boldsymbol{\zeta}_0$, then Theorem 4.4 suggests that WD causes the effective LR to steadily increase until the dynamic enters the EoS regime. Now something interesting happens — $\boldsymbol{\theta}_t$ moves a bit away from $\boldsymbol{\zeta}_0$ and starts to oscillate around the manifold Γ. This oscillation is not completely perpendicular to Γ but actually forms a small angle that pushes $\boldsymbol{\theta}_t$ to move downward persistently until $\boldsymbol{\theta}_t$ approaches the minimizer $\boldsymbol{\zeta}_*$ denoted in the plot.

For a general scale-invariant loss $\mathcal{L} : \mathbb{R}^D \setminus \{\mathbf{0}\} \to \mathbb{R}$, which minimizer does $\boldsymbol{\theta}_t$ move towards? In this work, we consider the setting where there is a manifold Γ consisting only of local minimizers (but not necessarily all of them). We show that $\boldsymbol{\theta}_t$ always oscillates around the manifold once it approaches the manifold and enters the EoS regime, and meanwhile $\boldsymbol{\theta}_t$ keeps moving in a direction of reducing spherical sharpness.

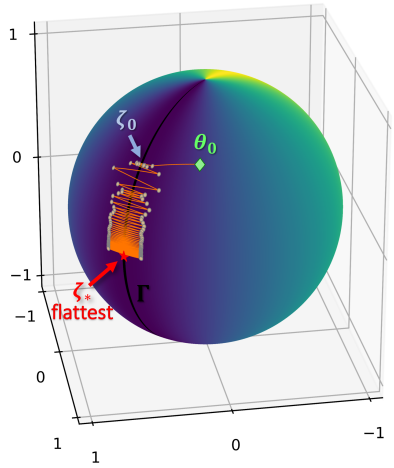


Figure 4: The trajectory of $\boldsymbol{\theta}_t$ on a 3D scale-invariant loss function. Darker color means lower loss on the unit sphere, and points in the black line are minimizers (see Appendix F). In the end, $\boldsymbol{\theta}_t$ approaches the flattest one (red star).

4.2.1 Assumptions

Now we formally introduce our main assumption on the local minimizer manifold Γ.

Assumption 4.5. The loss function $\mathcal{L} : \mathbb{R}^D \setminus \{\mathbf{0}\} \to \mathbb{R}$ is $\mathcal{C}^4$-smooth and scale-invariant. Γ is a $\mathcal{C}^2$-smooth, $(D_\Gamma - 1)$-dimensional submanifold of $\mathbb{S}^{D-1}$ for some $0 \le D_\Gamma < D$, where every $\boldsymbol{\theta} \in \Gamma$ is a local minimizer of $\mathcal{L}$ on $\mathbb{S}^{D-1}$ and $\mathrm{rank}(\boldsymbol{H}(\boldsymbol{\theta})) = D - D_\Gamma$.

Scale-invariance has become a standard assumption in studying neural nets with normalization layers [77, 78, 84]. For VGG and ResNet, the scale-invariance can be ensured after making minor changes to the architectures (see Appendix Q.1). The training loss $\mathcal{L}$ may not be smooth if the activation is ReLU, but lately it has become clear that differentiable activations such as Swish [98], GeLU [48] can perform equally well. Swish is indeed used in our VGG-11 experiments (Figure 2), but ResNet with ReLU activation also exhibits a sharpness-reduction bias empirically (see Appendix P.2).

For any local minimizer $\boldsymbol{\theta} \in \Gamma$, the eigenvalues $\lambda_k^{\mathrm{H}}(\boldsymbol{\theta})$ must be non-negative. And $\lambda_k^{\mathrm{H}}(\boldsymbol{\theta}) = 0$ for all $D - D_\Gamma < k \le D$, since Γ is of dimension $D_\Gamma - 1$. The condition $\mathrm{rank}(\boldsymbol{H}(\boldsymbol{\theta})) = D - D_\Gamma$ ensures that the Hessian is maximally non-degenerate on Γ, which also appears as a key assumption in previous works [81, 8, 35]. This condition simplifies the calculus on Γ in our analysis as it ensures that the null space of the matrix $\boldsymbol{H}(\boldsymbol{\theta})$ equals to the tangent space of Γ at $\boldsymbol{\theta} \in \Gamma$. It is also closely related to PL condition (Definition 4.3) as Assumption 4.5 implies that $\mathcal{L}(\boldsymbol{\theta})$ satisfies μ-PL (for some $\mu > 0$) locally around every $\boldsymbol{\theta} \in \Gamma$ on the unit sphere (Arora et al. [8], Lemma B.3).

To ease our analysis, we also need the following regularity condition to ensure that the largest eigenvalue is unique. In our experiments, sharpness reduction happens even when the multiplicity of the top eigenvalue is more than 1, but we leave the analysis of that case to future work.

Assumption 4.6. For all $\boldsymbol{\theta} \in \Gamma$, $\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}) > \lambda_2^{\mathrm{H}}(\boldsymbol{\theta})$. That is, the top eigenvalue of $\boldsymbol{H}(\boldsymbol{\theta})$ is unique.

4.2.2 Main Theorem

First, we define $\eta_{\text{in}} := \hat{\eta}\hat{\lambda}$ as the intrinsic learning rate (name from Li et al. [78]) for convenience. As suggested in Theorems 4.2 and 4.4, $\boldsymbol{\theta}_t$ can get close to a local minimizer and be in the EoS regime at some step t_1: if $\boldsymbol{\zeta}_0$ is the local minimizer, then $\|\boldsymbol{\theta}_{t_1} - \boldsymbol{\zeta}_0\|_2 = O(\eta_{\text{in}}^{1/2})$ and $\tilde{\eta}_{t_1} = \frac{2}{\lambda_1^{\text{H}}(\boldsymbol{\zeta}_0)} + O(\eta_{\text{in}}^{1/2})$. In our main theorem, we start our analysis from step t_1 while setting $t_1 = 0$ WLOG (otherwise we can shift the step numbers). We connect GD+WD in the EoS regime to the following gradient flow (3) on the manifold Γ minimizing spherical sharpness (with gradient-dependent learning rate), and show that one step of GD+WD tracks a time interval of length η_{in} in the gradient flow.

$$\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0 \in \Gamma, \qquad \frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{\zeta}(\tau) = -\frac{2\nabla_\Gamma \log \lambda_1^{\text{H}}(\boldsymbol{\zeta}(\tau))}{4 + \|\nabla_\Gamma \log \lambda_1^{\text{H}}(\boldsymbol{\zeta}(\tau))\|_2^2}. \tag{3}$$

Here we use the notation $\nabla_\Gamma R(\boldsymbol{\theta})$ for any $R : \mathbb{R}^D \to \mathbb{R}$ to denote the projection of $\nabla R(\boldsymbol{\theta})$ onto the tangent space $\mathsf{T}_{\boldsymbol{\theta}}(\Gamma)$ at $\boldsymbol{\theta} \in \Gamma$. $\boldsymbol{\zeta}(\tau)$ reduces sharpness as it moves in direction of the negative gradient of $\log \lambda_1^{\text{H}}(\boldsymbol{\zeta}(\tau))$ on Γ. A simple chain rule shows how fast the spherical sharpness decreases:

$$\frac{\mathrm{d}}{\mathrm{d}t}\log \lambda_1^{\text{H}}(\boldsymbol{\zeta}(\tau)) = -\frac{2\|\nabla_\Gamma \log \lambda_1^{\text{H}}(\boldsymbol{\zeta}(\tau))\|_2^2}{4 + \|\nabla_\Gamma \log \lambda_1^{\text{H}}(\boldsymbol{\zeta}(\tau))\|_2^2} \approx \begin{cases} -\frac{1}{2}\|\nabla_\Gamma \log \lambda_1^{\text{H}}(\boldsymbol{\zeta}(\tau))\|_2^2 & \text{for small gradient;} \\ -2 & \text{for large gradient.} \end{cases}$$

Note that it is not enough to just assume that $\boldsymbol{\theta}_0$ is close to $\boldsymbol{\zeta}_0$. If $\boldsymbol{\theta}_0 = \boldsymbol{\zeta}_0$ holds exactly, then the subsequent dynamic of $\boldsymbol{w}_t$ is described by $\boldsymbol{w}_t = (1 - \hat{\eta}\hat{\lambda})^t \boldsymbol{w}_0$ with direction unchanged. There are also some other bad initial directions of $\boldsymbol{w}_0$ that may not lead to the sharpness-reduction bias. This motivates us to do a smoothed analysis for the initial direction: the initial direction is $\boldsymbol{\zeta}$ with tiny random perturbation, where the perturbation scale is allowed to vary from $\exp(-\eta_{\text{in}}^{-o(1)})$ to $\eta_{\text{in}}^{1/2-o(1)}$, and we show that a good initial direction is met with high probability as $\eta_{\text{in}} \to 0$.[1] Alternatively, one can regard it as a modeling of the tiny random noise in GD+WD due to the precision errors in floating-point operations. See Figure 5b; the training loss can never be exactly zero in practice.

Initialization Scheme. Given a local minimizer $\boldsymbol{\zeta}_0 \in \Gamma$, we initialize $\boldsymbol{w}_0 \in \mathbb{R}^D \setminus \{\boldsymbol{0}\}$ as follows: draw $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma_0^2 \boldsymbol{I}/D)$ from Gaussian and set the direction of $\boldsymbol{w}_0$ to $\frac{\boldsymbol{\zeta}_0 + \boldsymbol{\xi}}{\|\boldsymbol{\zeta}_0 + \boldsymbol{\xi}\|_2}$, where σ_0 can take any value in $[\exp(-\eta_{\text{in}}^{-o(1)}), \eta_{\text{in}}^{1/2-o(1)}]$; then set the parameter norm $\|\boldsymbol{w}_0\|_2$ to be any value that satisfies $\left|\tilde{\eta}_0 - \frac{2}{\lambda_1^{\text{H}}(\boldsymbol{\zeta}_0)}\right| \le \eta_{\text{in}}^{1/2-o(1)}$, where $\tilde{\eta}_0 := \frac{\hat{\eta}}{(1-\hat{\eta}\hat{\lambda})\|\boldsymbol{w}_0\|_2^2}$ is the effective LR for the first step.

Theorem 4.7. *Under Assumptions 4.5 and 4.6, for GD+WD (1) with sufficiently small intrinsic learning rate $\eta_{\text{in}} := \hat{\eta}\hat{\lambda}$, if we follow the above initialization scheme for some $\boldsymbol{\zeta}_0 \in \Gamma$, then with probability $1 - O(\eta_{\text{in}}^{1/2-o(1)})$, the trajectory of $\boldsymbol{\theta}_t := \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_2}$ approximately tracks a sharpness-reduction flow $\boldsymbol{\zeta} : [0, T] \to \Gamma$ that starts from $\boldsymbol{\zeta}_0$ and evolves as the ODE (3) up to time T (if solution exists), in the sense that $\|\boldsymbol{\theta}_t - \boldsymbol{\zeta}(t\eta_{\text{in}})\|_2 = O(\eta_{\text{in}}^{1/4-o(1)})$ for all $0 \le t \le T/\eta_{\text{in}}$.*

Remark 4.8 (Magnitude of Oscillation)**.** As suggested by Figure 4, $\boldsymbol{\theta}_t$ actually oscillates around the manifold. But according to our analysis, the magnitude of oscillation is as small as $O(\eta_{\text{in}}^{1/2-o(1)})$, so it is absorbed into our final bound $O(\eta_{\text{in}}^{1/4-o(1)})$ for the distance between $\boldsymbol{\theta}_t$ and $\boldsymbol{\zeta}(t\eta_{\text{in}})$.

4.2.3 Proof Idea

Throughout our proof, we view GD+WD for $\boldsymbol{w}_t$ as a PGD for $\boldsymbol{\theta}_t$ with effective LR $\tilde{\eta}_t$ (Lemma 3.1). To track $\boldsymbol{\theta}_t$ with $\boldsymbol{\zeta}(t\eta_{\text{in}})$, for each step t, we construct a local minimizer $\boldsymbol{\phi}_t \in \Gamma$ that serves as the "projection" of $\boldsymbol{\theta}_t$ onto the manifold Γ, in the sense that the displacement $\boldsymbol{x}_t := \boldsymbol{\theta}_t - \boldsymbol{\phi}_t$ is approximately perpendicular to the tangent space of Γ at $\boldsymbol{\phi}_t$. Our entire proof works through induction. According to the initial conditions, the dynamic is initially in the EoS regime: $\|\boldsymbol{x}_t\|_2 \le \eta_{\text{in}}^{1/2-o(1)}$ and $|\tilde{\eta}_t - 2/\lambda_1^{\text{H}}(\boldsymbol{\phi}_t)| \le \eta_{\text{in}}^{1/2-o(1)}$ at $t = 0$. In our induction, we maintain the induction hypothesis that these two EoS conditions continue to hold for all $t \ge 0$.

[1] Here $\eta_{\text{in}}^{-o(1)}$ can be constant, $O(\log(1/\eta_{\text{in}}))$, or $O(\text{polylog}(1/\eta_{\text{in}}))$, but not $\eta_{\text{in}}^{-\epsilon}$ if $\epsilon > 0$ is a constant. As mentioned later, this need for random initialization is very similar to the one needed in power method for computing eigenvalues.

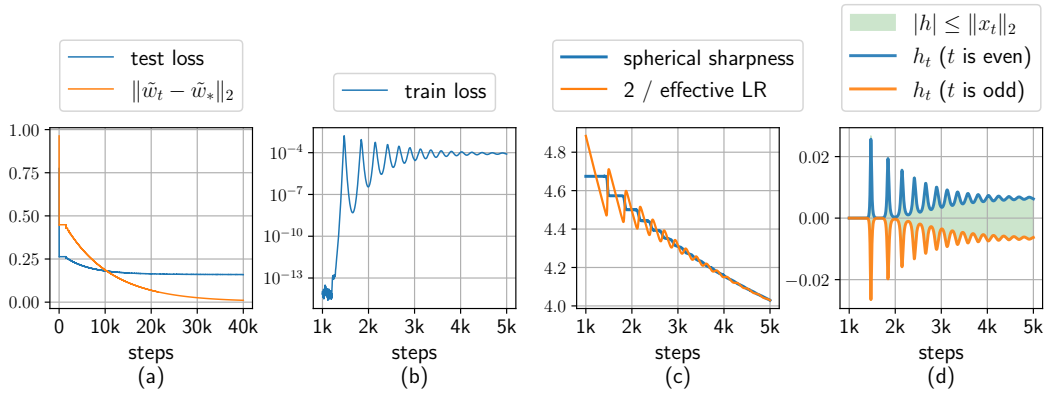


Figure 5: Illustration of the oscillation and periodic behaviors of GD+WD on linear regression with BN (see Sections 4.2.3 and 5). The training loss decreases to $\approx 10^{-14}$ in the first 1k steps and achieves test loss 0.26. Starting from step $\sim$ 1k, the dynamic enters the EoS regime. **(a).** The test loss decreases to 0.16 as a distance measure to the flattest solution (M) decreases towards 0; **(b).** The training loss oscillates around $\sim 10^{-4}$ in the EoS regime; **(c).** $2/\tilde{\eta}_t$ switches back and forth between being smaller and larger than $\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$; **(d).** The parameter oscillates around the minimizer manifold along the top eigenvector direction, and the magnitude of oscillation $|h_t|$ rises and falls periodically.

Period-Two Oscillation. A key insight in our proof is that after a few initial steps, $\boldsymbol{\theta}_t$ is oscillating around $\boldsymbol{\phi}_t$ along the $\pm \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\theta})$ directions, where $\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\theta})$ is a unit top eigenvector of $\boldsymbol{H}(\boldsymbol{\theta})$ and is chosen in a way that $\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\theta})$ is continuous on Γ. More specifically, $\boldsymbol{x}_t = h_t \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(\|\boldsymbol{x}_t\|_2^2)$ for $h_t := \langle \boldsymbol{x}_t, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) \rangle$. The oscillation is of period 2: $h_t > 0$ when t is even and $h_t < 0$ when t is odd. See Figure 5d for an example.

This oscillation can be connected to a power method for the matrix $\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}_t)$. In the EoS regime, we can approximate $\boldsymbol{\theta}_{t+1}$ (when $\boldsymbol{x}_t$ is small) as $\boldsymbol{\theta}_{t+1} = \Pi(\boldsymbol{\theta}_t - \tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t)) \approx \Pi(\boldsymbol{\theta}_t - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}_t)\boldsymbol{x}_t) \approx \boldsymbol{\theta}_t - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}_t)\boldsymbol{x}_t$ by Taylor expansions of $\nabla \mathcal{L}$ and $\Pi : \mathbb{R}^D \setminus \{\boldsymbol{0}\} \to \mathbb{S}^{D-1}$. We can further show that $\boldsymbol{\phi}_{t+1} \approx \boldsymbol{\phi}_t$ due to our choice of projections. Then the connection to power method is shown below:

$$\boldsymbol{x}_{t+1} \approx \boldsymbol{\theta}_{t+1} - \boldsymbol{\phi}_t \approx (\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}_t))\boldsymbol{x}_t.$$

By simple linear algebra, $\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$ is an eigenvector of $\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}_t)$, associated with eigenvalue $1 - \tilde{\eta}_t \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) \approx -1$. The remaining eigenvalues are $\{1 - \tilde{\eta}_t \lambda_i^{\mathrm{H}}(\boldsymbol{\phi}_t)\}_{i=2}^D$, where $\lambda_i^{\mathrm{H}}(\boldsymbol{\phi}_t)$ is the i-th largest eigenvalue of $\boldsymbol{H}(\boldsymbol{\theta}_t)$, and they lie in the range $(-1, 1]$ since $\lambda_i^{\mathrm{H}}(\boldsymbol{\phi}_t) \in [0, \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t))$. Using a similar analysis to power method, we show that $\boldsymbol{x}_t$ quickly aligns to the direction of $\pm \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$ after a few initial steps, as the corresponding eigenvalue has approximately the largest absolute value.[2]

To formally establish the above result, we need a tiny initial alignment between $\boldsymbol{x}_0$ and $\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_0)$, just as the initial condition in power method. This is where we need the initial random perturbation.

Oscillation Drives $\boldsymbol{\phi}_t$ to Move. This period-two oscillation is the driving power to push $\boldsymbol{\phi}_t$ to move on the manifold. The main idea here is to realize that the oscillation direction deviates slightly from the direction of $\pm \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$ by using a higher-order approximation. We specifically use the Taylor approximation to show that this deviation leads $\boldsymbol{\phi}_t$ to move slightly on Γ: after each cycle of oscillation, $\boldsymbol{\phi}_{t+2} \approx \boldsymbol{\phi}_t - 4h_t^2 \nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(\eta_{\mathrm{in}}^{1.5-o(1)})$, which resembles two steps of gradient descent on Γ to minimize the logarithm of spherical sharpness with learning rate $2h_t^2$,

Periodic Behavior of h_t and $\tilde{\eta}_t$. It remains to analyze the dynamics of h_t so that we can know how fast the sharpness reduction is. Our analysis is inspired by an empirical study from Lobacheva et al. [84], which reveals a periodic behavior of gradients and effective learning rates in training normalized nets with weight decay. In our theoretical setting, we capture this periodic behavior by showing that h_t and $\tilde{\eta}_t$ *do* evolve periodically. See Figures 5c and 5d for an example.

The key is that $\tilde{\eta}_t$ changes as an adaptive gradient method: $\tilde{\eta}_t$ increases when gradient is small and decreases when gradient is large (due to the effect of WD; see Figures 3a and 3b), and in our case the gradient norm scales as $|h_t|$ since $\nabla \mathcal{L}(\boldsymbol{\theta}_t) \approx h_t \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$. By the power method approximation,

[2] Our construction of $\boldsymbol{\phi}_t$ ensures that $\boldsymbol{x}_t$ only has a small overlap with the 1-eigenspace of $\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}_t)$, so $\boldsymbol{x}_t$ can only align to $\pm \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$.

$h_{t+2} \approx (1-\tilde{\eta}_t \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t))^2 h_t$, so $|h_t|$ decreases when $\tilde{\eta}_t < 2/\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$. But $|h_t|$ cannot decrease forever, since $\tilde{\eta}_t$ increases when $|h_t|$ is sufficiently small. When $\tilde{\eta}_t$ rises to over $2/\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$, $|h_t|$ changes from decreasing to increasing according to our approximation. But h_t cannot increase indefinitely either, since $\tilde{\eta}_t$ decreases when $|h_t|$ is sufficiently large. A period finishes when $\tilde{\eta}_t < 2/\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$ holds again.

In our theoretical analysis, we connect this periodic behavior with a 1-dimensional Hamiltonian system (see Appendix H.2), and show that $2h_t^2$ in each step can be approximated by its average value in the period without incurring a large error. Further calculations show that this average value is approximately $\frac{2\eta_{\mathrm{in}}}{4+\|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\zeta}(t\eta_{\mathrm{in}}))\|_2}$, the learning rate in the flow (3) multiplied with η_{in}. We can therefore conclude that each step of $\boldsymbol{\phi}_t$ (or $\boldsymbol{\theta}_t$) tracks a time interval of η_{in} in the flow.

Extensions. We note that this periodic behavior is not limited to GD+WD on scale-invariant loss, since the above intuitive argument holds as long as the effective LR changes adaptively with respect to gradient change. Based on this intuition, an important notion called *Quasi-RMSprop scheduler* is proposed. For a PGD method, a learning rate scheduler is a rule for changing the effective LR in each step, and Quasi-RMSprop is a specific class of schedulers we define, including the way that the effective LR changes in GD+WD on scale-invariant loss (if viewed as PGD). Our proof is done in a unified way that works as long as the effective LR changes in each step according to a Quasi-RMSprop scheduler. As a by-product, a similar theorem can be proved for GD (without projection) on non-scale-invariant loss if the LR changes as a Quasi-RMSprop in each step. For example, we can extend our analysis to RMSprop with a scalar learning rate. See Appendix B.

5 Case Study: Linear Regression with Batch Normalization

In this section, we analyze the GD+WD dynamics on linear regression with Batch Normalization (BN), as a simple application of our theory. Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a dataset, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ are inputs and regression targets. We study the over-parameterized case where $d \gg n$, and we assume that the regression targets are generated by an unknown linear model.

A classic linear model is parameterized by $(\boldsymbol{w}, b) \in \mathbb{R}^d \times \mathbb{R}$ and outputs $\boldsymbol{w}^\top \boldsymbol{x} + b$ given input $\boldsymbol{x}$, but now we add a BN to the output. More specifically, we consider a batch-normalized linear model $\Phi(\boldsymbol{x}; \boldsymbol{w}, \gamma, \beta) := \gamma \cdot \frac{\boldsymbol{w}^\top \boldsymbol{x} - \mu_1}{\sigma_1} + \beta$, where μ_1, σ_1 are the mean and variance of $\{\boldsymbol{w}^\top \boldsymbol{x}_i\}_{i=1}^n$ over the whole dataset[3], and the bias term b is cancelled out due to BN. Note that $\Phi(\boldsymbol{x}; \boldsymbol{w}, \gamma, \beta)$ is still a linear function with respect to $\boldsymbol{x}$. Let $\boldsymbol{\mu}_{\mathrm{x}} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_{\mathrm{x}} \in \mathbb{R}^{d \times d}$ be the mean and covariance of the input data $\{\boldsymbol{x}_i\}_{i=1}^n$. Then $\Phi(\boldsymbol{x}; \boldsymbol{w}, \gamma, \beta)$ can be rewritten as:

$$\Phi(\boldsymbol{x}; \boldsymbol{w}, \gamma, \beta) = \tilde{\boldsymbol{w}}^\top \boldsymbol{x} + \tilde{b}, \qquad \text{where} \quad \tilde{\boldsymbol{w}} := \gamma \boldsymbol{w} / \|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}, \quad \tilde{b} := \beta - \tilde{\boldsymbol{w}}^\top \boldsymbol{\mu}_{\mathrm{x}}. \tag{4}$$

No matter how $\boldsymbol{w}$ is set, the output mean and variance of Φ are always β and γ^2. To simplify our analysis, we fix β, γ to be non-trainable constants so that the mean and variance of Φ's output match with those of $\{y_i\}_{i=1}^n$, that is, we set $\beta = \mu_{\mathrm{y}}$ and $\gamma = \sigma_{\mathrm{y}}$ to be the mean and standard deviation of y_i over the whole dataset. Then the training loss is $\mathcal{L}(\boldsymbol{w}) := \frac{1}{n} \sum_{i \in [n]} (\Phi(\boldsymbol{x}_i; \boldsymbol{w}, \gamma, \beta) - y_i)^2$.

Theorem 5.1. *In our setting of linear regression with BN, the sharpness-reduction flow $\boldsymbol{\zeta}$ defined in* (3) *converges to the solution $\boldsymbol{w}^* \in \mathbb{S}^{d-1}$ that minimizes sharpness $\lambda_1^{\mathrm{H}}(\boldsymbol{w}^*)$ on Γ, regardless of the initialization. Moreover, the coefficients $(\tilde{\boldsymbol{w}}, \tilde{b})$ associated with $\boldsymbol{w}^*$ (defined in* (4)*) are the optimal solution of the following constrained optimization problem* (M)*:*

$$\min \quad \|\boldsymbol{w}\|_2^2 \quad \textit{s.t.} \quad \boldsymbol{w}^\top \boldsymbol{x}_i + b = y_i, \quad \forall i \in [n]. \tag{M}$$

At first sight the result may appear trivial because the intent of WD is to regularize L^2-norm. But this is deceptive because in scale-invariant nets the regularization effect of WD is not explicit. This result also challenges conventional view of optimization. GD is usually viewed as a discretization of its continuous counterpart, gradient flow (GF), and theoretical insight for the discrete update including convergence rate and implicit bias is achieved by analyzing the continuous counterpart (See Appendix A for a list). However, GF does not have the same sharpness-reduction bias as GD. As discussed in [77], adding WD only performs a time-rescaling on the GF trajectory on scale-invariant loss, but does not change the point that GF converge to if we project the trajectory onto the unit sphere. One can easily show that GF may converge to any zero-loss solution, but no matter how small

[3] Note that the batch size is n here as we are running full-batch GD

LR is, GD exhibits the sharpness-reduction bias towards the optimal solution of (M). To our best knowledge, this result is the first concrete example where even with arbitrarily small LR, GD can still generalize better than GF under natural settings.

6 Discussion

Experimental Verification of Sharpness Reduction. Besides Figures 1 and 2, Appendix P.1 provides additional matrix completion experiments with different data size, and Appendix P.2 provides CIFAR-10 experiments with ResNet-20. In all these experiments, we observed that GD continues to improve the test accuracy even after fitting the training set, and this phenomenon is correlated with the decreasing trend of spherical sharpness. See also Appendix P.3 for the validation for the periodic behavior we analyze in theory.

Ablation Studies on Normalization and Weight Decay. Our theoretical analysis crucially relies on the interplay between normalization and WD to establish the sharpness-reduction flow. We also conducted ablation studies on normalization and WD to highlight the importance of this interplay. First, if normalization is removed, the spherical sharpness becomes undefined, and we do not know if GD implicitly minimizes any sharpness measure. But even if a similar measure does exist, it cannot be strongly related to generalization, because we can verify that the test accuracy becomes very bad without normalization (56.8% on CIFAR-10, Figure 14), and continuing training after fitting the training set no longer improves test accuracy. Second, if WD is removed, the analysis in Arora et al. [7] guarantees convergence in the stable regime, and we can verify that the spherical sharpness and test accuracy stop changing when the loss is small. The final test accuracy is stuck at 66.4% (Figure 15), whereas training with WD leads to 84.3%.

Explaining the Progressive Sharpening and EoS Phenomena. Cohen et al. [24] conducted extensive empirical studies on the dynamics of GD in deep learning (without weight decay), formally $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \hat{\eta}\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)$. They observed the *progressive sharpening* phenomenon: $\lambda_1(\nabla^2\tilde{\mathcal{L}}(\boldsymbol{w}_t))$ tends to increase so long as it is less than $2/\hat{\eta}$. Then they observed that the training typically enters the EoS regime, which they define as a regime that (1) $\lambda_1(\nabla^2\tilde{\mathcal{L}}(\boldsymbol{w}_t))$ hovers right at, or just above $2/\hat{\eta}$; and (2) the training loss $\tilde{\mathcal{L}}(\boldsymbol{w}_t)$ goes up and down over short timescales, yet still decreases in the long-term run. A recent research trend focuses on explaining the progressive sharpening and EoS phenomena [1, 91, 8, 18]. Our work corresponds to an important special case where $\tilde{\mathcal{L}}(\boldsymbol{w})$ is a scale-invariant loss with L^2-regularization, namely $\mathcal{L}(\boldsymbol{w}) + \frac{\hat{\lambda}}{2}\|\boldsymbol{w}\|_2^2$. By analyzing the interplay between normalization and WD, the first part of our results (Section 4.1) attributes progressive sharpening to norm change, and the second part (Section 4.2) justifies in theory that the training can make progress in the EoS regime. See Appendix G.3 for more discussion.

7 Conclusions and Future Work

We exhibited settings where gradient descent has an implicit bias to reduce spherical sharpness in training neural nets with normalization layers and weight decay, and we verified experimentally this sharpness-reduction bias predicted by our theorem as well as its generalization benefit on CIFAR-10.

Our theoretical analysis applies to dynamics around a minimizer manifold and requires a small (but finite) learning rate so that we can show that the parameter oscillates locally and approximately tracks a sharpness-reduction flow. We note that in practice a decrease in spherical sharpness is observed even with moderate LR and even before getting close to a minimizer manifold. Explaining these phenomena is left for future work. Now we list some other future directions. The first is to generalize our results to SGD, where the sharpness measure may not be the spherical sharpness and could depend on the structure of gradient noise. Second, to understand the benefit of reducing spherical sharpness on specific tasks, e.g., why does reducing spherical sharpness encourage low-rank on matrix completion with BN (Figure 1)? Third, to study sharpness-reduction bias for neural net architectures that are not scale-invariant on all parameters (e.g., with certain unnormalized layers).

Acknowledgements

This work is funded by NSF, ONR, Simons Foundation, DARPA and SRC. ZL is also supported by Microsoft Research PhD Fellowship.

References

[1] Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. 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 247–257. PMLR, 17–23 Jul 2022.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019.

[4] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7411–7422. Curran Associates, Inc., 2019.

[5] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[6] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8139–8148. Curran Associates, Inc., 2019.

[7] Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. In *International Conference on Learning Representations*, 2019.

[8] 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.

[9] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[10] David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 342–350. PMLR, 06–11 Aug 2017.

[11] David Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.

[12] Johan Bjorck, Carla Gomes, and Bart Selman. Understanding batch normalization. *arXiv preprint arXiv:1806.02375*, 2018.

[13] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 483–513. PMLR, 09–12 Jul 2020.

[14] Andrew Brock, Soham De, and Samuel L Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. In *International Conference on Learning Representations*, 2021.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[16] Yongqiang Cai, Qianxiao Li, and Zuowei Shen. A quantitative analysis of the effect of batch normalization on gradient descent. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 882–890. PMLR, 09–15 Jun 2019.

[17] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[18] Lei Chen and Joan Bruna. On gradient descent convergence beyond the edge of stability. *arXiv preprint arXiv:2206.04172*, 2022.

[19] Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep re{lu} networks? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=fgd7we_uZa6.

[20] Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019. doi: 10.1109/TSP.2019.2937282.

[21] Vitaliy Chiley, Ilya Sharapov, Atli Kosson, Urs Koster, Ryan Reece, Sofia Samaniego de la Fuente, Vishal Subbiah, and Michael James. Online normalization for training neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[22] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

[23] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 2937–2947. Curran Associates, Inc., 2019.

[24] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

[25] Alex Damian, Tengyu Ma, and Jason D Lee. Label noise SGD provably prefers flat global minimizers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27449–27461. Curran Associates, Inc., 2021.

[26] Hadi Daneshmand, Jonas Kohler, Francis Bach, Thomas Hofmann, and Aurelien Lucchi. Batch normalization provably avoids ranks collapse for randomly initialised deep networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18387–18398. Curran Associates, Inc., 2020.

[27] Hadi Daneshmand, Amir Joudaki, and Francis Bach. Batch normalization orthogonalizes representations in deep random networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[28] Soham De and Sam Smith. Batch normalization biases residual blocks towards the identity function in deep networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19964–19975. Curran Associates, Inc., 2020.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

[30] Lijun Ding, Dmitriy Drusvyatskiy, and Maryam Fazel. Flat minima generalize for low-rank matrix recovery. *arXiv preprint arXiv:2203.03756*, 2022.

[31] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 06–11 Aug 2017.

[32] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[33] Yonatan Dukler, Quanquan Gu, and Guido Montufar. Optimization theory for ReLU neural networks trained with normalization layers. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2751–2760. PMLR, 13–18 Jul 2020.

[34] K. J. Falconer. Differentiation of the Limit Mapping in a Dynamical System. *Journal of the London Mathematical Society*, s2-27(2):356–372, 04 1983. ISSN 0024-6107. doi: 10.1112/jlms/s2-27.2.356.

[35] 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):1–48, 2020.

[36] Robert L. Foote. Shorter notes: Regularity of the distance function. *Proceedings of the American Mathematical Society*, 92(1):153–155, 1984. ISSN 00029939, 10886826.

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

[38] Rong Ge, Yunwei Ren, Xiang Wang, and Mo Zhou. Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems*, 34, 2021.

[39] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241. PMLR, 09–15 Jun 2019.

[40] Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Edward Dahl, Zachary Nado, and Orhan Firat. A loss curvature perspective on training instabilities of deep learning models. In *International Conference on Learning Representations*, 2022.

[41] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6151–6159. Curran Associates, Inc., 2017.

[42] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9482–9491. Curran Associates, Inc., 2018.

[43] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[44] Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2315–2357. PMLR, 15–19 Aug 2021.

[45] Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *Advances in neural information processing systems*, 32, 2019.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 630–645, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

[48] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

[49] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a: Overview of mini-batch gradient descent. Technical report, 2012. URL https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

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

[51] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[52] Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[53] Lei Huang, Xianglong Liu, Yang Liu, Bo Lang, and Dacheng Tao. Centered weight normalization in accelerating training of deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[54] Hikaru Ibayashi and Masaaki Imaizumi. Exponential escape efficiency of SGD from sharp minima in non-stationary regime. *arXiv preprint arXiv:2111.04004*, 2021.

[55] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[56] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8571–8580. Curran Associates, Inc., 2018.

[57] Stanisław Jastrzębski, 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.

[58] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho*, and Krzysztof Geras*. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2020.

[59] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17176–17186. Curran Associates, Inc., 2020.

[60] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

[61] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. The normalization method for alleviating pathological sharpness in wide neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[62] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851*, ECML PKDD 2016, pages 795–811, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 9783319461274. doi: 10.1007/978-3-319-46128-1_50.

[63] 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. In *International Conference on Learning Representations*, 2017.

[64] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2698–2707. PMLR, 10–15 Jul 2018.

[65] Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann, Ming Zhou, and Klaus Neymeyr. Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 806–815. PMLR, 16–18 Apr 2019.

[66] Lingkai Kong and Molei Tao. Stochasticity of deterministic gradient descent: Large learning rate for multiscale objective function. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2625–2638. Curran Associates, Inc., 2020.

[67] Antoine Labatie, Dominic Masters, Zach Eaton-Rosen, and Carlo Luschi. Proxy-normalizing activations to match batch normalization while removing batch dependence. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16990–17006. Curran Associates, Inc., 2021.

[68] Susanna Lange, Kyle Helfrich, and Qiang Ye. Batch normalization preconditioning for neural network training. *Journal of Machine Learning Research*, 23(72):1–41, 2022.

[69] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[70] Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[71] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

[72] Aitor Lewkowycz and Guy Gur-Ari. On the training dynamics of deep networks with L_2 regularization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4790–4799. Curran Associates, Inc., 2020.

[73] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

[74] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

[75] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[76] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2–47. PMLR, 06–09 Jul 2018.

[77] Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2020.

[78] Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14544–14555. Curran Associates, Inc., 2020.

[79] Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

[80] Zhiyuan Li, Srinadh Bhojanapalli, Manzil Zaheer, Sashank Reddi, and Sanjiv Kumar. Robust training of neural networks using scale invariant architectures. 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 12656–12684. PMLR, 17–23 Jul 2022.

[81] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022.

[82] Zinan Lin, Vyas Sekar, and Giulia Fanti. Why spectral normalization stabilizes GANs: Analysis and improvements. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9625–9638. Curran Associates, Inc., 2021.

[83] Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and sgd can reach them. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8543–8552. Curran Associates, Inc., 2020.

[84] Ekaterina Lobacheva, Maxim Kodryan, Nadezhda Chirkova, Andrey Malinin, and Dmitry P Vetrov. On the periodic behavior of neural network training with batch normalization and weight decay. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21545–21556. Curran Associates, Inc., 2021.

[85] Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. Beyond batchnorm: Towards a unified understanding of normalization in deep learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4778–4791. Curran Associates, Inc., 2021.

[86] Ping Luo, Xinjiang Wang, Wenqi Shao, and Zhanglin Peng. Towards understanding regularization in batch normalization. In *International Conference on Learning Representations*, 2019.

[87] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

[88] Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34, 2021.

[89] Chao Ma and Lexing Ying. On linear stability of SGD and input-smoothness of neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[90] Chao Ma and Lexing Ying. A Riemannian mean field formulation for two-layer neural networks with batch normalization. *Research in the Mathematical Sciences*, 9(3):47, July 2022. ISSN 2197-9847.

[91] Chao Ma, Lei Wu, and Lexing Ying. The multiscale structure of neural network loss functions: The effect on optimization and origin. *arXiv preprint arXiv:2204.11326*, 2022.

[92] David McAllester. Simplified PAC-Bayesian margin bounds. In *Learning theory and Kernel machines*, pages 203–215. Springer, 2003.

[93] Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7108–7118. PMLR, 13–18 Jul 2020.

[94] Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17749–17761. Curran Associates, Inc., 2021.

[95] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[96] Ioannis Panageas and Georgios Piliouras. Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 2:1–2:12, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-029-3. doi: 10.4230/LIPIcs.ITCS.2017.2.

[97] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization. *arXiv preprint arXiv:1903.10520*, 2019.

[98] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[99] Akshay Rangamani, Nam H. Nguyen, Abhishek Kumar, Dzung Phan, Sang Peter Chin, and Trac D. Tran. A scale invariant measure of flatness for deep network minima. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1680–1684, 2021.

[100] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21174–21187. Curran Associates, Inc., 2020.

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

[102] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[103] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2483–2493. Curran Associates, Inc., 2018.

[104] Alexander Shekhovtsov and Boris Flach. Stochastic normalizations as bayesian learning. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 463–479, Cham, 2019. Springer International Publishing. ISBN 978-3-030-20890-5.

[105] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[106] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

[107] Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *International Conference on Learning Representations*, 2018.

[108] Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34, 2021.

[109] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[110] Hidenori Tanaka and Daniel Kunin. Noether's learning dynamics: Role of symmetry breaking in neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25646–25660. Curran Associates, Inc., 2021.

[111] Mattias Teye, Hossein Azizpour, and Kevin Smith. Bayesian uncertainty estimation for batch normalized deep networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4907–4916. PMLR, 10–15 Jul 2018.

[112] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9636–9647. PMLR, 13–18 Jul 2020.

[113] Twan van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

[114] Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using sgd and weight decay. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6380–6391. Curran Associates, Inc., 2021.

[115] Yi Wang and Zhiren Wang. Three-stage evolution and fast equilibrium for SGD with non-degerate critical points. 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 23092–23113. PMLR, 17–23 Jul 2022.

[116] Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022.

[117] Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

[118] Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[119] Xiaoxia Wu, Edgar Dobriban, Tongzheng Ren, Shanshan Wu, Zhiyuan Li, Suriya Gunasekar, Rachel Ward, and Qiang Liu. Implicit regularization and convergence for weight normalization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2835–2847. Curran Associates, Inc., 2020.

[120] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[121] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.

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

[123] Mingyang Yi, Qi Meng, Wei Chen, Zhi-ming Ma, and Tie-Yan Liu. Positively scale-invariant flatness of relu neural networks. *arXiv preprint arXiv:1903.02237*, 2019.

[124] Mingyang Yi, Huishuai Zhang, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Bn-invariant sharpness regularizes the training model to better generalization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4164–4170. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[125] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[126] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2019.

[127] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7654–7663. PMLR, 09–15 Jun 2019.

[128] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sections 4.2.1 and 7.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We are basically a theoretical work studying the generalization myestery in deep learning. We do not see any negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions of our main theorem for GD+WD on scale-invariant loss are stated in Section 4.2.1. For GD/PGD with Quasi-RMSprop schedulers in general, see Appendix B.3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendices G and H for the proofs for our main theorems, Appendix O for the proof for the linear example.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See our supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix Q.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Most of our experiments have only run once due to computational constraints, but we verified the sharpness-reduction bias across various settings.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix Q.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents

1 Introduction **1**

1.1 Our Contributions 2

2 Related Works **3**

3 Preliminaries **4**

4 GD+WD on Scale-Invariant Loss Functions **4**

4.1 GD+WD Eventually Leaves the Stable Regime 5

4.2 Dynamics at the Edge of Stability 6

4.2.1 Assumptions 6

4.2.2 Main Theorem 7

4.2.3 Proof Idea 7

5 Case Study: Linear Regression with Batch Normalization **9**

6 Discussion **10**

7 Conclusions and Future Work **10**

A Additional Related Works **24**

B A General Theory for a Broader Class of Adaptive Gradient Methods **26**

B.1 Scalar RMSprop and Quasi-RMSprop Scheduler 26

B.2 Reformulation of GD+WD on Scale-Invariant Loss via Quasi-RMSprop Scheduler 27

B.3 Main Results for GD/PGD with Quasi-RMSprop Scheduler 28

B.3.1 Spherical Optimization 28

B.3.2 Full Space Optimization 29

C PAC-Bayes Bounds Based on Spherical Sharpness **31**

D Additional Preliminaries **33**

D.1 Additional Notations 33

D.2 Scale-Invariant Functions 33

D.3 Polyak-Łojasiewicz Condition 34

D.3.1 Full Space Optimization 34

D.3.2 Spherical Optimization 35

E Supplementary Material for Appendix B.2 **35**

E.1 Proof for Theorem B.7 35

E.2 Proof for Theorem B.9 36

F Details of the 3D Example **36**

G Supplementary Material for Section 4.1 **37**
G.1 Proof for Descent Lemma . . . 37
G.2 Proof for Theorem 4.4: GD Eventually Enters the EoS Regime . . . 37
G.3 Connection to the EoS Regime in Cohen et al.'s Definition . . . 38

H Proof Outlines of Our Theorems on Sharpness Reduction **40**
H.1 Additional Notations . . . 40
H.2 RMS-drift Process: Introduction . . . 41
H.3 Reduction to RMS-drift Process: The Case of Full Space Optimization . . . 42
H.3.1 Construction of Working Zones . . . 42
H.3.2 Good Initialization . . . 43
H.3.3 Alignment Phase . . . 43
H.3.4 Drifting Phase . . . 43
H.4 Reduction to RMS-drift Process: The Case of Spherical Optimization . . . 43
H.4.1 Construction of Working Zones . . . 44
H.4.2 Good Initialization . . . 44
H.4.3 Alignment Phase . . . 44
H.4.4 Drifting Phase . . . 44
H.5 RMS-drift Process: Analysis . . . 44
H.6 Finalizing Proofs . . . 45

I Lemmas for Working Zones **45**
I.1 Construction of Working Zones . . . 45
I.2 Gradient Flow Projection . . . 45
I.2.1 Full Space Optimization . . . 45
I.2.2 Spherical Optimization . . . 46

J Lemmas for Gradient Descent **46**
J.1 Full Space Optimization . . . 46
J.2 Spherical Optimization . . . 48

K Lemmas for Quasi-RMSprop Schedulers **51**

L Reduction to RMS-drift Process: The Case of Full Space Optimization **52**
L.1 Good Initialization . . . 52
L.2 Alignment Phase . . . 52
L.3 Drifting Phase . . . 54

M Reduction to RMS-drift Process: The Case of Spherical Optimization **57**
M.1 Good Initialization . . . 58
M.2 Alignment Phase . . . 58
M.3 Drifting Phase . . . 58

N Analysis of RMS-drift Process **58**
N.1 Conservation of Energy 58
N.2 Flow Approximation 60

O Proofs for Linear Regression with Batch Normalization **62**

P Experiments **64**
P.1 Validation of Sharpness Reduction on Matrix Completion 64
P.2 Validation of Sharpness Reduction on CIFAR-10 66
P.3 Periodic Behaviors 68
P.4 Ablation Study: Normalization 69
P.5 Ablation Study: Weight Decay 69
P.6 Ablation Study: Initial Effective Learning Rate 71

Q Experiment Details **72**
Q.1 Additional Details of CIFAR-10 Experiments 72
Q.2 Additional Details of Linear Regression Experiments 75
Q.3 Computing Spherical Sharpness 75

R Discussion on the Affine Parameters of the Final BN **75**
R.1 Squared Loss 75
R.2 Crossentropy Loss 76

A Additional Related Works

Sharpness Measures and Parameter Rescaling. To best capture the generalization performance, the measure of sharpness should give the same value whenever the function represented by the neural net is the same. As mentioned in the introduction, $\lambda_1(\nabla^2 \mathcal{L}(\boldsymbol{w}))$ does not satisfy this property for normalized nets because it is sensitive to weight rescaling. The spherical sharpness takes care of scale-invariance w.r.t. all parameters, but it is certainly not the only rescaling symmetry in deep nets. However, to the best of our knowledge, the spherical sharpness is the only measure that provably decreases in training normalized nets. Although many other sharpness measures may take care of more symmetries [123, 124, 112, 99, 37], it is unclear whether GD/SGD is implicitly reducing them.

SGD Noise Helps to Escape Sharp Minima. It has been a folklore that noise in stochastic gradient helps escapes sharp local minima. With the simplification of assuming the loss is quadratic and treating SGD as its canonical continuous SDE approximation, Zhu et al. [127], Xie et al. [121] showed that anisotropic noise (*e.g.*, the noise covariance is equal to Hessian) has a better escape efficiency of SGD out of a sharp minimizer, in comparison with isotropic noise. Under the same assumptions, Ibayashi and Imaizumi [54] proved the exponential escape efficiency without assuming SDE reaches the stationary distribution. Kleinberg et al. [64] showed that SGD can escape sharp local minima assuming one-point convexity. While all previous escaping analysis of SGD are based on the continuous approximation, another approach called stability analysis is able to show that SGD cannot converge to sharp local minima when learning rate is larger than some threshold [117, 118, 89].

Implicit Bias of GD. There are mainly two types of implicit bias result for GD, where the first type of result applies essentially to the continuous limit of gradient descent, namely gradient flow, and tolerates error discretization and stochasticity when the learning rate is sufficiently small [107, 106, 87, 59, 43, 42, 76, 100, 4, 22, 79, 88, 101, 108, 38]. The analyses of GD/SGD based on Neural Tangent Kernel (NTK) also essentially belong to this type, because though the analysis in NTK regime tolerates stochasticity and finite learning rate, GD/ SGD do not learn different solutions compared to gradient flow. Such works includes (but are not limited to) [56, 75, 32, 6, 5, 3, 2, 128, 23, 122, 17, 19]. This type of result typically relates the generalization quality to the initialization of GD, and GD in such regimes cannot escape from bad local minima once reaching there. In contrast, the second type of results, to which the current paper belongs, fundamentally relies on the discrete nature of gradient descent. For example, Barrett and Dherin [11] showed that for small LR, gradient descent is approximately equal to gradient flow minimizing a new objective, *i.e.*, the original objective plus η times squared norm of gradients. Kong and Tao [66] studied a special class of "multiscale" loss functions, and they showed that large learning rate introduces chaos to the dynamics of GD and provides a mechanism to escape local minima. Wang et al. [116] proved that for the matrix factorization problem, GD with a large learning rate has an implicit bias towards a solution with balanced matrix norms. Stability analysis [117, 118, 94, 89] also belongs to this type.

Comparison with Arora et al. [8]. The paper by Arora et al. [8] is probably the most related work among the second type of the implicit bias results. They assume that there is a smooth function L which satisfies certain regularity conditions around the minimizer manifold (including that Hessian is maximally non-degenerate on the manifold), and show that running normalized GD on L or GD on $\sqrt{L}$ with sufficiently small LR tracks a deterministic flow on the minimizer manifold and decreases the largest eigenvalue of $\nabla^2 L$, which has a similar flavor to our result. However, our setting (scale invariant loss + WD) is more natural and their result and technique do not apply to our setting because there is no minimizer under their definition, not to mention manifolds. To show the spherical sharpness decreases, we have to develop new proof techniques, including connecting the dynamics to a 1-dimensional Hamiltonian system. Another difference is that our analysis applies to gradient descent directly, but the analysis by Arora et al. [8] requires injecting stochastic noise to gradients.

Edge of Stability. Cohen et al. [24] provided an extensive empirical study showing that GD typically occurs at the Edge of Stability (EoS), where the top eigenvalue of Hessian is approximately 2 / LR and the descent lemma does not guarantee the loss to decrease. Ahn et al. [1] explored the dynamics of GD in the EoS regime through insightful experiments, and they attribute the EoS phenomenon to the lack of flat stationary points near GD trajectory and the existence of a subset near minima that is forward invariant under GD update. Ma et al. [91] proved the EoS phenomenon for a class of loss functions that are decomposable as a sum of 1-dimensional functions with subquadratic growth. Chen and Bruna [18] provided detailed analyses for the EoS phenomenon on two-layer single-neuron net and matrix factorization. The aforementioned work by Arora et al. [8] analyzes

normalized GD on L or GD on $\sqrt{L}$ also in the EoS regime, where the latter case corresponds to a class of loss functions that grow approximately linearly near minima (e.g., the absolute value function $|x|$). The evolution of the eigenvalues of Hessian has also been studied for SGD [40, 58, 73]. Our work can be seen as a theoretical explanation of the EoS phenomenon for scale-invariant loss functions with L^2-regularization (see also Appendix G.3), which are a more broad and natural class of training loss in deep learning compared with those being studied in previous theoretical analyses.

B A General Theory for a Broader Class of Adaptive Gradient Methods

Our main theorem for GD+WD on scale-invariant loss, Theorem 4.7, is actually a corollary of a more general theorem that holds for any PGD on $\mathbb{S}^{D-1}$ with effective learning rates changing adaptively according a specific kind of update rules. We name this kind of update rules as *Quasi-RMSprop Scheduler*, where the name is due to its similarity to RMSprop [49]. One of the main reasons that we define this concept is that proving a theorem for quasi-RMSprop schedulers in general is no harder than proving that only for GD+WD on scale-invariant loss, and sometimes the math involved is more simple and elegant when we analyze the dynamics at a higher level through quasi-RMSprop.

In this section, we introduce this key notion, Quasi-RMSprop scheduler. To motivate it, we first recall the update rule of RMSprop and present one of its variants with a single scalar learning rate. Then we formally introduce the notion of quasi-RMSprop schedulers, which is a class of rules for setting effective learning rates similarly as RMSprop. Then we prove a recursive formula for the effective learning rates when GD+WD optimizes a scale-invariant loss, and we categorize it as an instance of quasi-RMSprop scheduler.

After introducing this key notion, we then present our general theorem that holds for any PGD, or even GD, as long as the effective learning rate is set by a quasi-RMSprop scheduler in each step. Examples include Scalar RMSprop (Corollary B.14) and GD+WD on scale-invariant loss (Theorem 4.7).

B.1 Scalar RMSprop and Quasi-RMSprop Scheduler

The usual RMSprop algorithm maintains a vector $\boldsymbol{v}_t$ storing the moving average of the squared gradients for every coordinate, i.e., $\boldsymbol{v}_{t+1} \leftarrow \beta \boldsymbol{v}_t + (1-\beta)(\nabla \mathcal{L}(\boldsymbol{\theta}_t))^{\odot 2}$, where $\boldsymbol{g}^{\odot 2}$ stands for the vector obtained by squaring $\boldsymbol{g}$ coordinatewise. When updating the training parameter, RMSprop divides the usual GD update by the square root of the moving average coordinatewise, i.e., $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \frac{\eta}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}} \odot \nabla \mathcal{L}(\boldsymbol{\theta}_t)$, where ϵ is a small constant to avoid division by zero. Here all the addition, division, square root operations are coordinatewise, and $\odot$ stands for coordinatewise multiplication.

Now we consider a variant of RMSprop, which we call *Scalar RMSprop*, where the moving average v_t is maintained as a scalar storing the moving average of the squared norm of gradients, rather than a vector that stores the moving averages separately for each coordinate.

Definition B.1 (Scalar RMSprop, Standard Form)**.** *Scalar RMSprop* is an iterative method with the following update rule:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \frac{\eta}{\sqrt{v_t}} \nabla \mathcal{L}(\boldsymbol{\theta}_t), \tag{5}$$

$$v_{t+1} \leftarrow \beta v_t + (1-\beta) \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2. \tag{6}$$

Besides that v_t is changed from a vector to a scalar, another difference is that the gradient is divided by $\sqrt{v_t}$ in (5), while in the usual RMSprop it is $\sqrt{\boldsymbol{v}_{t+1}}$. In fact, our later analysis also applies if $\sqrt{v_t}$ is changed to $\sqrt{v_{t+1}}$, but the version with $\sqrt{v_t}$ leads to simpler math.

An alternative view of Scalar RMSprop is to regard it as GD with time-varying learning rate $\tilde{\eta}_t$, that is, $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t)$, where $\tilde{\eta}_t$ is the learning rate being used at the t-th step, which we call the *effective learning rate* at step t.

We view the effective learning rate here as the output of a *learning rate scheduler*.[4] In this view, we call the learning rate scheduler for Scalar RMSprop as *RMSprop scheduler*.

Definition B.2 (Gradient-Based Learning Rate Scheduler)**.** A *gradient-based learning rate scheduler* is an algorithm $\mathcal{H}$ that reads from a stream of vectors $\boldsymbol{g}_0, \boldsymbol{g}_1, \boldsymbol{g}_2, \ldots$, For all $t \geq 0$, $\mathcal{H}$ outputs a real number $\tilde{\eta}_t := \mathcal{H}(\boldsymbol{g}_0, \boldsymbol{g}_1, \ldots, \boldsymbol{g}_t)$ as soon as $\mathcal{H}$ finishes reading the first $t+1$ vectors. At each step of gradient descent equipped with a gradient-based learning rate scheduler $\mathcal{H}$, the gradients in training are revealed one by one as an input stream to $\mathcal{H}$, and $\tilde{\eta}_t$ produced by $\mathcal{H}$ is used as the effective learning rate at step t.

Definition B.3 (RMSprop Scheduler)**.** Given a constant $\tilde{v}_0 > 0$, a base learning rate η and a decay rate β as hyperparameters, an RMSprop scheduler $\mathcal{H}_{\text{RMS}}$ is a gradient-based learning rate

[4] The scheduler here is similar to `torch.optim.lr_scheduler` in PyTorch, see https://pytorch.org/docs/stable/optim.html#how-to-adjust-learning-rate

scheduler that reads from a stream of vectors $\boldsymbol{g}_0, \boldsymbol{g}_1, \ldots$, and generates effective learning rates $\tilde{\eta}_t := \mathcal{H}_{\text{RMS}}(\boldsymbol{g}_0, \ldots, \boldsymbol{g}_t)$ for all $t \geq 0$ according to the following recursion:

$$\tilde{\eta}_t \leftarrow \frac{1}{\sqrt{\tilde{v}_t}}, \qquad \tilde{v}_{t+1} \leftarrow \beta \tilde{v}_t + (1-\beta)\bar{g}_t^2, \qquad \text{where} \quad \bar{g}_t := \|\boldsymbol{g}_t\|_2/\eta.$$

Definition B.4 (Scalar RMSprop, Alternative Form)**.** *Scalar RMSprop* is gradient descent with a RMSprop scheduler $\mathcal{H}_{\text{RMS}}$.

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t), \qquad \text{where} \qquad \tilde{\eta}_t := \mathcal{H}_{\text{RMS}}(\nabla \mathcal{L}(\boldsymbol{\theta}_0), \ldots, \nabla \mathcal{L}(\boldsymbol{\theta}_t)). \tag{7}$$

Note that we use $\tilde{v}_t$ as an internal state of $\mathcal{H}_{\text{RMS}}$ in Definition B.3. It is easy to verify that $\tilde{v}_t$ is nothing but a reparameterization of v_t: setting $\tilde{v}_t = v_t/\eta^2$ in Definition B.3 recovers the update rule in Definition B.1.

Now we introduce the notion of quasi-RMSprop scheduler, which is a class of gradient-based learning rate schedulers that have update rules similar to RMSprop scheduler. At first reading, one can just ignore the details and regard quasi-RMSprop scheduler as an RMSprop scheduler with negligible perturbations when $\beta \to 1$.

Definition B.5 (Quasi-RMSprop Scheduler)**.** A *quasi-RMSprop scheduler* $\mathcal{H}$ is a gradient-based learning rate scheduler parameterized by a base learning rate η and a decay rate β, satisfying the following properties:

1. (η, β) is allowed to take value from a hyperparameter space $\mathcal{P}_{\mathcal{H}} \subseteq (0, +\infty) \times (0, 1)$.
2. There exist thresholds $\eta_{\max}, \beta_{\min}$, a continuous function $\delta : (0, +\infty) \to \mathbb{R}$, and a polynomial $P : \mathbb{R} \to \mathbb{R}$ such that the following holds. If $(\eta, \beta) \in \mathcal{P}_{\mathcal{H}}$ and $\eta < \eta_{\max}, \beta > \beta_{\min}$, for any input stream $\boldsymbol{g}_0, \boldsymbol{g}_1, \boldsymbol{g}_2, \ldots$, there exists a sequence of positive real numbers $\tilde{v}_0, \tilde{v}_1, \tilde{v}_2, \ldots$ such that $\tilde{\eta}_t := \mathcal{H}(\boldsymbol{g}_0, \ldots, \boldsymbol{g}_t)$ satisfies the following two inequalities for all $t \geq 0$:

$$\begin{aligned} \left| \tilde{\eta}_t - \frac{1}{\sqrt{\tilde{v}_t}} \right| &\leq \delta(\tilde{v}_t) \cdot (1-\beta) \cdot (1 + \bar{g}_t^2) \\ \left| \tilde{v}_{t+1} - \left(\beta \tilde{v}_t + (1-\beta)\bar{g}_t^2\right) \right| &\leq \delta(\tilde{v}_t) \cdot (1-\beta)^2 \cdot P(\bar{g}_t) \end{aligned} \qquad \text{where} \quad \bar{g}_t := \|\boldsymbol{g}_t\|_2/\eta.$$

   The real number $\tilde{v}_t$ is called the *moment estimate* at step t associated with the input stream and effective learning rates.

It is clear that a RMSprop scheduler can be seen as a quasi-RMSprop scheduler with the same hyperparameters η, β, and $\delta \equiv 0$.

B.2 Reformulation of GD+WD on Scale-Invariant Loss via Quasi-RMSprop Scheduler

Now we reformulate GD+WD on scale-invariant loss as PGD with a quasi-RMSprop scheduler. Recall that we say a loss function $\mathcal{L}(\boldsymbol{w})$ is scale-invariant if $\mathcal{L}(c\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w})$ for all $c > 0$, and Lemma 3.1 converts GD+WD on scale-invariant functions to Projected Gradient Descent (PGD) on unit sphere, i.e., $\boldsymbol{\theta}_{t+1} \leftarrow \Pi(\boldsymbol{\theta}_t - \tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t))$, where $\tilde{\eta}_t := \frac{\hat{\eta}}{(1-\hat{\eta}\hat{\lambda})\|\boldsymbol{w}_t\|_2^2}$ is the effective learning rate at step t. However, the evolution of $\tilde{\eta}_t$ over time is unclear unless we know how the parameter norm $\|\boldsymbol{w}_t\|_2$ changes. Similar as the above analysis for Scalar RMSprop, where we decompose Scalar RMSprop as the GD method and an RMSprop scheduler, here we abstract the evolution of $\tilde{\eta}_t$ as a learning rate scheduler, which we call *GWSI scheduler* (name picked from the initials of **G**D+**W**D on **S**cale-**I**nvariant loss) and view GD+WD on scale-invariant loss as PGD on $\mathbb{S}^{D-1}$ with effective learning rates being set by a GWSI scheduler.

Definition B.6 (GWSI Scheduler)**.** Given a constant $\tilde{v}_0 > 0$, a base learning rate η and a decay rate β as hyperparameters, a GWSI scheduler $\mathcal{H}_{\text{GWSI}}$ is a gradient-based learning rate scheduler that reads from a stream of vectors $\boldsymbol{g}_0, \boldsymbol{g}_1, \ldots$, and generates effective learning rates $\tilde{\eta}_t := \mathcal{H}_{\text{GWSI}}(\boldsymbol{g}_0, \ldots, \boldsymbol{g}_t)$ for all $t \geq 0$ according to the following recursion:

$$\tilde{\eta}_t \leftarrow \frac{1}{\sqrt{\tilde{v}_t}}, \qquad \tilde{v}_{t+1} \leftarrow \beta \tilde{v}_t + (1-\beta)\bar{g}_t^2 + \frac{1}{4\beta \tilde{v}_t}(1-\beta)^2 \bar{g}_t^4, \qquad \text{where} \quad \bar{g}_t := \|\boldsymbol{g}_t\|_2/\eta.$$

Theorem B.7. *For gradient descent (1) with learning rate $\hat{\eta} > 0$ and weight decay $\hat{\lambda} > 0$ on scale-invariant function $\mathcal{L}(\boldsymbol{w})$, let $\boldsymbol{\theta}_t := \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|_2}$ be the direction of $\boldsymbol{w}_t$ and $\tilde{\eta}_t := \frac{\hat{\eta}}{(1-\hat{\eta}\hat{\lambda})\|\boldsymbol{w}_t\|_2^2}$ be the effective learning rate at time t. Then $\tilde{\eta}_t$ evolves exactly the same as the GWSI scheduler with hyperparameters $\tilde{v}_0 := \frac{(1-\hat{\eta}\hat{\lambda})^2\|\boldsymbol{w}_0\|_2^4}{\hat{\eta}^2}$, $\beta := (1-\hat{\eta}\hat{\lambda})^4$, $\eta := \sqrt{(\beta^{-1}-1)/2}$, and we can write the dynamics of $\boldsymbol{\theta}_t$ as:*

$$\boldsymbol{\theta}_{t+1} \leftarrow \Pi(\boldsymbol{\theta}_t - \tilde{\eta}_t \nabla\mathcal{L}(\boldsymbol{\theta}_t)), \qquad \textit{where} \qquad \tilde{\eta}_t := \mathcal{H}_{\textit{GWSI}}(\nabla\mathcal{L}(\boldsymbol{\theta}_0), \ldots, \nabla\mathcal{L}(\boldsymbol{\theta}_t)).$$

Remark B.8. To the best of our knowledge, this particular form of the recursion formula (Definition B.6) for the effective learning rates of GD+WD on scale-invariant loss does not appear in prior works, but some variants have been studied before. Arora et al. [7] derived a similar formula when the weight decay is zero. Li et al. [78] obtained a Stochastic Differential Equation (SDE) for SGD+WD on scale-invariant functions, which is essentially a continuous approximation of our formulation when $\hat{\eta}\hat{\lambda} \to 0$. Tanaka and Kunin [110] studied the continuous-time approximation of the momentum method with WD on scale-invariant loss and establish a connection to adaptive gradient methods.

We defer the proof for Theorem B.7 to Appendix E.1. The formula of GWSI scheduler clearly resembles RMSprop scheduler: the only difference is that GWSI scheduler has an extra $\frac{1}{4\beta\tilde{v}_t}(1-\beta)^2\bar{g}_t^4$ term, which is negligible when $(1-\beta)^2$ is small. This intuition is formalized through the definition of quasi-RMSprop. It can be easily seen that GWSI scheduler is a quasi-RMSprop scheduler with the same hyperparameters η, β.

In fact, we can use Theorem B.7 as a basis to obtain a better way to write GD+WD on scale-invariant loss as PGD with a quasi-RMSprop scheduler, in which η, β are expressed more simply in terms of the intrinsic learning rate $\eta_{\text{in}} := \hat{\eta}\hat{\lambda}$. The main idea is that $\beta = 1 - 4\eta_{\text{in}} + O(\eta_{\text{in}}^2)$ and $\eta = \sqrt{2\eta_{\text{in}}} \cdot (1 + O(\eta_{\text{in}}))$ when η_{in} is small, and we can absorb these approximation errors into the error bounds in Definition B.5. We defer the details to Appendix E.2.

Theorem B.9. *There exists a quasi-RMSprop scheduler $\mathcal{H}_{\textit{QRMS}}$ with hyperparameter space $\{(\eta, \beta) : \beta = 1 - 2\eta^2, \eta \in (0, \frac{1}{\sqrt{2}})\}$ such that the following holds for GD+WD on scale-invariant loss. If the intrinsic learning rate $\eta_{\text{in}} := \hat{\eta}\hat{\lambda}$ lies in the range $(0, 1/4)$, then we can set the hyperparameters of $\mathcal{H}_{\textit{QRMS}}$ to be $\eta = \sqrt{2\eta_{\text{in}}}, \beta = 1 - 4\eta_{\text{in}}$ so that the dynamics of $\boldsymbol{\theta}_t$ can be written as*

$$\boldsymbol{\theta}_{t+1} \leftarrow \Pi(\boldsymbol{\theta}_t - \tilde{\eta}_t \nabla\mathcal{L}(\boldsymbol{\theta}_t)), \qquad \textit{where} \qquad \tilde{\eta}_t := \mathcal{H}_{\textit{QRMS}}(\nabla\mathcal{L}(\boldsymbol{\theta}_0), \ldots, \nabla\mathcal{L}(\boldsymbol{\theta}_t)).$$

B.3 Main Results for GD/PGD with Quasi-RMSprop Scheduler

As mentioned in Section 4.2.3, a key step in our analysis is to show the oscillation and periodic behaviors of the effective learning rates. But these behaviors can show up in other algorithms besides GD+WD on scale-invariant loss.

In the following, we first present Theorem 4.7, which holds for any PGD on $\mathbb{S}^{D-1}$ with effective learning rates being set by a quasi-RMSprop scheduler. We focus on the case where $\eta > 0$ is small and $\beta = C_{\text{b}}\eta^2 + O(\eta^4)$ for some constant $C_{\text{b}} > 0$. Our theorem shows that if $\boldsymbol{\theta}_0$ is initially near a local minimizer manifold Γ and is in the EoS regime, then $\boldsymbol{\theta}_t$ approximately tracks a sharpness-reduction flow defined as follows:

$$\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0 \in \Gamma, \qquad \frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{\zeta}(\tau) = -\frac{\nabla_\Gamma \log \lambda_1^{\text{H}}(\boldsymbol{\zeta}(\tau))}{4 + \frac{2}{C_{\text{b}}}\|\nabla_\Gamma \log \lambda_1^{\text{H}}(\boldsymbol{\zeta}(\tau))\|_2^2}, \tag{8}$$

Besides, we also generalize our theorem to any GD with quasi-RMSprop scheduler, where Scalar RMSprop serves as an important example.

The proof outlines of these theorems are given in Appendix H, and the proof details are spread over Appendices I to N.

B.3.1 Spherical Optimization

Consider Projected Gradient Descent (PGD) on the unit sphere $\mathbb{S}^{D-1}$:

$$\boldsymbol{\theta}_{t+1} \leftarrow \Pi(\boldsymbol{\theta}_t - \tilde{\eta}_t \nabla\mathcal{L}(\boldsymbol{\theta}_t)), \tag{9}$$

where $\Pi : \boldsymbol{w} \mapsto \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$ is the projection operator, and the effective learning rate $\tilde{\eta}_t$ is set by a quasi-RMSprop scheduler with base learning rate $\eta > 0$ and decay rate $\beta = 1 - C_{\mathrm{b}}\eta^2 + O(\eta^4)$.

Recall that the dynamic is in the EoS regime if $\tilde{\eta}_t \approx 2/\lambda_{\max}^{(t)}$ (see Lemma 4.1). When $\boldsymbol{\theta}_0$ is around a local minimizer $\boldsymbol{\zeta}_0$ and $\tilde{\eta}_0$ is generated by a quasi-RMSprop scheduler, the condition of being in EoS is essentially $\frac{1}{\sqrt{\tilde{v}_0}} \approx \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\zeta}_0)}$. In the following theorem, we show that PGD with quasi-RMSprop scheduler evolves as (8) if it is initially in the EoS regime.

Initialization Scheme. Given a local minimizer $\boldsymbol{\zeta}_0 \in \Gamma$ and a hyperparameter α_0, we initialize the initial direction $\boldsymbol{\theta}_0$ and initial moment estimate $\tilde{v}_0$ as follows: draw $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma_0^2 \boldsymbol{I}/D)$ from Gaussian and set the direction of $\boldsymbol{w}_0$ to $\frac{\boldsymbol{\zeta}_0+\boldsymbol{\xi}}{\|\boldsymbol{\zeta}_0+\boldsymbol{\xi}\|_2}$, where σ_0 can take any value in $[\exp(-\alpha_0^2)\eta, \alpha_0\eta]$; then set $\tilde{v}_0$ to be any positive value that satisfies $\left|\frac{1}{\sqrt{\tilde{v}_0}} - \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\zeta}_0)}\right| \le \alpha_0\eta$.

Theorem B.10. *Under Assumptions 4.5 and 4.6, for PGD described as (9) and initialized as above scheme for some $\boldsymbol{\zeta}_0 \in \Gamma$ and some $1 \le \alpha_0 \le \eta^{-o(1)}$, with probability $1 - O(\alpha_0\eta\sqrt{\log(1/\eta)})$, the trajectory of $\boldsymbol{\theta}_t$ approximately tracks a sharpness-reduction flow $\boldsymbol{\zeta} : [0, T] \to \Gamma$ that starts from $\boldsymbol{\zeta}_0$ and evolves as the ODE (8) (if solution exists), in the sense that $\|\boldsymbol{\theta}_t - \boldsymbol{\zeta}(t\eta^2)\|_2 = O(\alpha_0^2\eta^{1/2}\log(1/\eta))$ for all $0 \le t \le T/\eta^2$.*

Theorem 4.7 is a direct corollary of Theorem B.10, as GD+WD on scale-invariant loss can be seen as PGD with GWSI scheduler.

Proof for Theorem 4.7. Theorem B.9 implies that GD+WD on scale-invariant loss with LR $\hat{\eta}$ and WD $\hat{\lambda}$ can be seen as a PGD on $\mathbb{S}^{D-1}$ with quasi-RMSprop scheduler, where the base learning rate and decay rate of this scheduler is $\eta = \sqrt{2\eta_{\mathrm{in}}}$ and $\beta = 1 - 4\eta_{\mathrm{in}}$. Therefore, $\beta = 1 - C_{\mathrm{b}}\eta^2$ for $C_{\mathrm{b}} = 2$.

Now we apply Theorem B.10 to prove Theorem 4.7. It is easy to translate the initial conditions of Theorem 4.7 to Theorem B.10 with $\alpha_0 = \eta^{-o(1)} = \eta_{\mathrm{in}}^{-o(1)}$. Let $\hat{\boldsymbol{\zeta}}(t)$ be the sharpness-reduction flow defined as in (8) with $C_{\mathrm{b}} = 2$ and horizon $\hat{T} := 2T$.

$$\hat{\boldsymbol{\zeta}}(0) = \boldsymbol{\zeta}_0 \in \Gamma, \qquad \frac{\mathrm{d}}{\mathrm{d}\tau}\hat{\boldsymbol{\zeta}}(\tau) = -\frac{\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\hat{\boldsymbol{\zeta}}(\tau))}{4 + \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\hat{\boldsymbol{\zeta}}(\tau))\|_2^2}. \tag{10}$$

Then with probability $1 - O(\alpha_0\eta\sqrt{1/\delta}) = 1 - O(\eta^{1-o(1)})$, $\boldsymbol{\theta}_t$ approximately tracks $\hat{\boldsymbol{\zeta}}(\tau)$ in the sense that $\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\zeta}}(t\eta^2)\|_2 = O(\eta^{1/2-o(1)})$ for all $0 \le t \le \hat{T}/\eta^2$. Replacing η with $\sqrt{2\eta_{\mathrm{in}}}$ gives $\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\zeta}}(2t\eta_{\mathrm{in}})\|_2 = O(\eta_{\mathrm{in}}^{1/4-o(1)})$ for all $0 \le t \le T/\eta_{\mathrm{in}}$, and the success probability becomes $1 - O(\eta_{\mathrm{in}}^{1/2-o(1)})$. We can finish the proof by noting that $\hat{\boldsymbol{\zeta}}(2\tau)$ is just $\boldsymbol{\zeta}(\tau)$ defined in (3). □

B.3.2 Full Space Optimization

Now we present our general theorem for GD on $\mathbb{R}^D$ with quasi-RMSprop scheduler. Here GD with quasi-RMSprop scheduler can be written as:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t), \tag{11}$$

where the effective learning rate $\tilde{\eta}_t$ is set by a quasi-RMSprop scheduler with base learning rate $\eta > 0$ and decay rate $\beta = 1 - C_{\mathrm{b}}\eta^2 + O(\eta^4)$.

Similar to the spherical case, we assume that there is a manifold Γ consisting of local minimizers, but now we are not assuming scale-invariance.

Assumption B.11. The loss function $\mathcal{L} : \mathbb{R}^D \to \mathbb{R}$ is $\mathcal{C}^4$-smooth. Γ is a $\mathcal{C}^2$-smooth, D_Γ-dimensional submanifold of $\mathbb{S}^{D-1}$ for some $0 \le D_\Gamma \le D$, where every $\boldsymbol{\theta} \in \Gamma$ is a local minimizer of $\mathcal{L}$ on $\mathbb{R}^D$ and $\mathrm{rank}(\boldsymbol{H}(\boldsymbol{\theta})) = D - D_\Gamma$.

The following assumption is essentially the same as Assumption 4.6 except that now Γ is defined differently.

Assumption B.12. For all $\boldsymbol{\theta} \in \Gamma$, $\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}) > \lambda_2^{\mathrm{H}}(\boldsymbol{\theta})$. That is, the top eigenvalue of $\boldsymbol{H}(\boldsymbol{\theta})$ is unique.

Based on Assumptions B.11 and B.12 above, we study GD starting in the EoS regime. Similar to the spherical case, we focus on the case where $\boldsymbol{\theta}_0$ is around a local minimizer $\boldsymbol{\zeta} \in \Gamma$, and EoS is the regime in which $\tilde{\eta}_t \approx \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\zeta}_0)}$, which essentially means $\frac{1}{\sqrt{\tilde{v}_0}}$ if the effective LR is set by a quasi-RMSprop scheduler. In the theorem below, we show that GD with quasi-RMSprop scheduler tracks the sharpness-reduction flow defined in (8). Note that the ODE here is the same as the spherical case, but Γ is defined differently.

Initialization Scheme. Given a local minimizer $\boldsymbol{\zeta}_0 \in \Gamma$ and a hyperparameter α_0, we initialize the initial parameter $\boldsymbol{\theta}_0$ and initial moment estimate $\tilde{v}_0$ as follows: draw $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{I}/D)$ from Gaussian and set $\boldsymbol{w}_0 \leftarrow \boldsymbol{\zeta}_0 + \boldsymbol{\xi}$, where σ_0 can take any value in $[\exp(-\alpha_0^2)\eta, \alpha_0\eta]$; then set $\tilde{v}_0$ to be any positive value that satisfies $\left|\frac{1}{\sqrt{\tilde{v}_0}} - \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\zeta}_0)}\right| \leq \alpha_0\eta$.

Theorem B.13. *Under Assumptions B.11 and B.12, for GD described as* (11) *and initialized as above scheme for some $\boldsymbol{\zeta}_0 \in \Gamma$ and some $1 \leq \alpha_0 \leq \eta^{-o(1)}$, with probability $1 - O(\alpha_0\eta\sqrt{\log(1/\eta)})$, the trajectory of $\boldsymbol{\theta}_t$ approximately tracks a sharpness-reduction flow $\boldsymbol{\zeta} : [0, T] \to \Gamma$ that starts from $\boldsymbol{\zeta}_0$ and evolves as the ODE* (8) *(if solution exists), in the sense that $\|\boldsymbol{\theta}_t - \boldsymbol{\zeta}(t\eta^2)\|_2 = O(\alpha_0^2\eta^{1/2}\log(1/\eta))$ for all $0 \leq t \leq T/\eta^2$.*

A direct corollary is that Scalar RMSprop follows the sharpness-reduction flow (8), since the RMSprop scheduler is a quasi-RMSprop scheduler.

Corollary B.14. *The statement of Theorem B.13 holds for Scalar RMSprop if the decay rate is set to $\beta = 1 - C_{\mathrm{b}}\eta^2 + O(\eta^2)$ for some constant $C_{\mathrm{b}} > 0$.*

C PAC-Bayes Bounds Based on Spherical Sharpness

In this section we give the PAC-Bayes bound for generalization error using spherical sharpness (Definition 1.1). We will start with our setting and then recap the classic PAC-Bayes theorem in [92]. The main result in this section is Theorem C.2.

Setting. Let $\ell(\boldsymbol{w}, \boldsymbol{z})$ be the loss of parameter $\boldsymbol{w}$ on data point $\boldsymbol{z}$ and assume $\ell_{\max} = \sup_{z\in\mathcal{Z}, \boldsymbol{w}\in\mathbb{S}^{D-1}} \ell(\boldsymbol{w}, \boldsymbol{z}) < \infty$. Let $\mathcal{S} := \{\boldsymbol{z}_i\}_{i=1}^n$ where $\boldsymbol{z}_i$ are sampled independently. Different to the previous notation, we use $\mathcal{L}_\mathcal{S}$ to denote the empirical loss on training dataset $\mathcal{S}$ where we run optimization algorithms and $\mathcal{L}$ to denote the population loss. Let $\rho_3(\mathcal{L}_\mathcal{S}) := \sup_{\boldsymbol{\theta}\in\mathbb{S}^{D-1}} \|\nabla^3 \mathcal{L}_\mathcal{S}(\boldsymbol{\theta})\|_2$. Since $\mathbb{S}^{D-1}$ is compact and $\|\nabla^3 \mathcal{L}_\mathcal{S}(\cdot)\|_2$ is continuous, $\rho_3(\mathcal{L}_\mathcal{S})$ is finite.

Theorem C.1 (PAC-Bayes theorem [92])**.** *Given any distribution P on $\mathbb{R}^D$, with at least $1-\delta$ probability over the randomness of the dataset $\mathcal{S}$, for all distribution Q on $\mathbb{R}^D$, it holds that*

$$\mathbb{E}_{\boldsymbol{w}\sim Q}\mathcal{L}_S(\boldsymbol{w}) - \mathbb{E}_{\boldsymbol{w}\sim Q}\mathcal{L}(\boldsymbol{w}) \le \ell_{\max}\sqrt{\frac{D_{\mathrm{KL}}(Q\|P) + \ln(n/\delta)}{2(n-1)}}.$$

Now we are ready to state the main theorem in this section, Theorem C.2, which shows that small spherical sharpness $\lambda_1(\nabla^2 \mathcal{L}_S(\boldsymbol{\theta}))$ leads to small generalization error.

Theorem C.2. *For any $\sigma \le \frac{1}{2+2\sqrt{(\ln n)/D}}$, with at least $1-\delta$ probability over the randomness of the dataset $\mathcal{S}$, where $\mathcal{S} := \{\boldsymbol{z}_i\}_{i=1}^n$ and every $\boldsymbol{z}_i$ is sampled independently, for any $\boldsymbol{\theta} \in \mathbb{S}^{D-1}$,*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I}_D/D)}\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) - \mathcal{L}_\mathcal{S}(\boldsymbol{\theta})\\ &\quad\le \frac{\sigma^2}{2}\lambda_1(\nabla^2\mathcal{L}_\mathcal{S}(\boldsymbol{\theta})) + \frac{16\sigma^3}{3}\rho_3(\mathcal{L}_\mathcal{S})(1+((\ln n)/D)^{1.5}) + \ell_{\max}\sqrt{\frac{D/\sigma^2 + 2\ln(n/\delta)}{n-1}},\end{aligned}$$

Thus with the standard assumption in [37] that $\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I}_D/D)}\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) \ge \mathcal{L}(\boldsymbol{\theta})$, we have the same upper bound for $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}_\mathcal{S}(\boldsymbol{\theta})$.

To prove Theorem C.2, we will first need the following lemma. The proof is standard so it is omitted.

Lemma C.3. *Let $Q = \mathcal{N}(\boldsymbol{\mu}_Q, \sigma_Q^2\boldsymbol{I}_D)$ and $P = \mathcal{N}(\boldsymbol{\mu}_P, \sigma_P^2\boldsymbol{I}_D)$, we have that*

$$D_{KL}(Q\|P) = \frac{1}{2}\left[\frac{D\sigma_Q^2 + \|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|_2^2}{\sigma_P^2} - D + D\log\left(\frac{\sigma_P^2}{\sigma_Q^2}\right)\right]$$

Proof of Theorem C.2. Let $Q := \mathcal{N}(\boldsymbol{\theta}, \sigma^2\boldsymbol{I}_D/D)$, $P := \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}_D/D)$. Then $D_{KL}(Q\|P) = \frac{D}{2\sigma^2}$. By Theorem C.1 we have

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I}_D/D)}\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) - \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I}_D/D)}\mathcal{L}_S(\boldsymbol{\theta}+\boldsymbol{\epsilon}) \le \ell_{\max}\sqrt{\frac{\frac{D}{2\sigma^2} + \ln(n/\delta)}{2(n-1)}}.$$

Let $h(\sigma) := (1+\sqrt{(\ln n)/D})\sigma$. By assumption it holds that $h(\sigma) \le \frac{1}{2}$. Thus by Lemma 1 in [69], we have for any positive t:

$$\mathbb{P}[\|\boldsymbol{\epsilon}\|_2^2 - \sigma^2 \ge 2\sigma^2\sqrt{t/D} + 2t\sigma^2/D] \le \exp(-t). \tag{12}$$

Therefore, with probability $1 - 1/\sqrt{n}$, we have

$$\|\boldsymbol{\epsilon}\|_2^2 \le \sigma^2\left(1 + 2\sqrt{(\ln\sqrt{n})/D} + 2\cdot(\ln\sqrt{n})/D\right) \le \sigma^2\left(1+\sqrt{(\ln n)/D}\right)^2 \le h(\sigma)^2.$$

Thus for any $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\|_2 = 1$, we have

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I}_D/D)}\mathcal{L}_S(\boldsymbol{\theta}+\boldsymbol{\epsilon}) \le \ell_{\max}/\sqrt{n} + \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I}_D/D)}\left[\mathcal{L}_S(\boldsymbol{\theta}+\boldsymbol{\epsilon})\mathbb{1}_{[\|\boldsymbol{\epsilon}\|_2\le h(\sigma)]}\right].$$

By Taylor expansion,

$$\mathcal{L}_\mathcal{S}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) \le \mathcal{L}_\mathcal{S}(\boldsymbol{\theta}) + \langle\boldsymbol{\epsilon}, \nabla\mathcal{L}_\mathcal{S}(\boldsymbol{\theta})\rangle + \frac{1}{2}\left\langle\boldsymbol{\epsilon}, \nabla^2\mathcal{L}_\mathcal{S}(\boldsymbol{\theta})\boldsymbol{\epsilon}\right\rangle + \frac{1}{6}\sup_{\lambda\in[0,1]}\left\langle\nabla^3\mathcal{L}_\mathcal{S}(\boldsymbol{\theta}+\lambda\boldsymbol{\epsilon}), \boldsymbol{\epsilon}^{\otimes 3}\right\rangle.$$

Note that $\|\boldsymbol{\theta}+\lambda\boldsymbol{\epsilon}\|_2 \geq \|\boldsymbol{\theta}\|_2 - \lambda\|\boldsymbol{\epsilon}\|_2 \geq \frac{1}{2}$ for all $\lambda \in [0,1]$, it holds that $\left\|\nabla^3\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}+\lambda\boldsymbol{\epsilon})\right\| = \|\boldsymbol{\theta}+\lambda\boldsymbol{\epsilon}\|_2^{-3}\left\|\nabla^3\mathcal{L}_{\mathcal{S}}(\frac{\boldsymbol{\theta}+\lambda\boldsymbol{\epsilon}}{\|\boldsymbol{\theta}+\lambda\boldsymbol{\epsilon}\|_2})\right\|_2 \leq 8\rho_3(\mathcal{L}_{\mathcal{S}})$, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I}_D/D)}\left[\mathcal{L}_S(\boldsymbol{\theta}+\boldsymbol{\epsilon})\mathbb{1}_{[\|\boldsymbol{\epsilon}\|\leq h(\sigma)]}\right] &\leq \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) + \frac{\sigma^2}{2D}\operatorname{Tr}[\nabla^2\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})] + \frac{4}{3}\rho_3(\mathcal{L}_{\mathcal{S}})h(\sigma)^3 \\ &\leq \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) + \frac{\sigma^2}{2}\lambda_1(\nabla^2\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})) + \frac{4}{3}\rho_3(\mathcal{L}_{\mathcal{S}})h(\sigma)^3.\end{aligned}$$

Thus we conclude that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I}_D/D)}\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) - \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) \\ \leq\ &\frac{\sigma^2}{2}\lambda_1(\nabla^2\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})) + \frac{4}{3}\rho_3(\mathcal{L}_{\mathcal{S}})h(\sigma)^3 + \ell_{\max}\left(\sqrt{\frac{\frac{D}{2\sigma^2}+\ln(n/\delta)}{2(n-1)}}+1\right) \\ \leq\ &\frac{\sigma^2}{2}\lambda_1(\nabla^2\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})) + \frac{16\sigma^3}{3}\rho_3(\mathcal{L}_{\mathcal{S}})(1+(\ln n/D)^{1.5}) + \ell_{\max}\sqrt{\frac{D/\sigma^2+2\ln(n/\delta)}{n-1}},\end{aligned}$$

which completes the proof. □

D Additional Preliminaries

D.1 Additional Notations

We use $\mathrm{cl}(\mathcal{M})$ to denote the closure of a set $\mathcal{M}$. For $\boldsymbol{\theta} \in \mathbb{R}^D$ and $\mathcal{M} \subseteq \mathbb{R}^D$, we use $d_2(\boldsymbol{\theta}, \mathcal{M}) := \inf\{\|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2 : \boldsymbol{\phi} \in \mathcal{M}\}$ to denote the L^2-distance from $\boldsymbol{\theta}$ to $\mathcal{M}$. For $\boldsymbol{\theta} \in \mathbb{R}^D$ and $\epsilon \geq 0$, $B^\epsilon(\boldsymbol{\theta}) := \{\boldsymbol{\theta}' \in \mathbb{R}^D : \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 < \epsilon\}$ is the open ϵ-ball centered at $\boldsymbol{\theta}$. For a set $\mathcal{M} \subseteq \mathbb{R}^D$, $\mathcal{M}^\epsilon := \bigcup_{\boldsymbol{\theta} \in \mathcal{M}} B^\epsilon(\boldsymbol{\theta})$ is the (open) ϵ-neighborhood of $\mathcal{M}$. All the manifolds in our paper refer to manifolds without boundary. For a manifold Γ, we use $\mathsf{T}_p(\Gamma)$ and $\mathsf{N}_p(\Gamma)$ to denote the tangent and normal space of Γ at a point $p \in \Gamma$.

Given a function $f : \mathbb{R}^D \to \mathbb{R}^n$ and vectors $\boldsymbol{x}, \boldsymbol{v} \in \mathbb{R}^D$, we use $\partial f_{\boldsymbol{x}}[\boldsymbol{v}]$ to denote the directional derivative $\partial f_{\boldsymbol{x}}[\boldsymbol{v}] := \lim_{t \to 0} \frac{1}{t} f(\boldsymbol{x} + t\boldsymbol{v})$, which also equals to the Jacobian of f at $\boldsymbol{x}$ multiplied with $\boldsymbol{v}$. We use $\partial^2 f_{\boldsymbol{x}}[\boldsymbol{v}, \boldsymbol{u}]$ to denote the second-order derivative $\partial(\partial f_{\boldsymbol{x}}[\boldsymbol{v}])_{\boldsymbol{x}}[\boldsymbol{u}]$. For a real-valued function $\mathcal{L} : \mathbb{R}^D \to \mathbb{R}$, we use $\nabla \mathcal{L}(\boldsymbol{\theta})$ for gradient, $\nabla^2 \mathcal{L}(\boldsymbol{\theta})$ for Hessian. For third-order derivatives of $\mathcal{L}$, we define $\partial^3 \mathcal{L}_{\boldsymbol{\theta}}[\boldsymbol{v}, \boldsymbol{u}] := \partial^2 (\nabla \mathcal{L})_{\boldsymbol{\theta}}[\boldsymbol{v}, \boldsymbol{u}] \in \mathbb{R}^D$. If $\mathcal{L}$ is $\mathcal{C}^4$-smooth, then the following Taylor expansion holds for $\nabla \mathcal{L}$:

$$\nabla \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{x}) = \nabla \mathcal{L}(\boldsymbol{\theta}) + \nabla^2 \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{x} + \frac{1}{2} \partial^3 \mathcal{L}_{\boldsymbol{\theta}}[\boldsymbol{x}, \boldsymbol{x}] + O(\|\boldsymbol{x}\|_2^3).$$

D.2 Scale-Invariant Functions

The following lemma summarizes a few important properties of scale-invariant loss that have been exploited in previous works [7, 77]. For completeness, we include a proof here.

Lemma D.1. *The following hold for a twice-differentiable scale-invariant function $\mathcal{L}(\boldsymbol{w})$:*

1. *The gradient is (-1)-homogeneous and it is always perpendicular to $\boldsymbol{w}$, i.e., $\nabla \mathcal{L}(c\boldsymbol{w}) = c^{-1} \nabla \mathcal{L}(\boldsymbol{w})$ for all $c > 0$ and $\langle \nabla \mathcal{L}(\boldsymbol{w}), \boldsymbol{w} \rangle = 0$;*
2. *The Hessian matrix is (-2)-homogeneous, i.e., $\nabla^2 \mathcal{L}(c\boldsymbol{w}) = c^{-2} \nabla^2 \mathcal{L}(\boldsymbol{w})$ for all $c > 0$.*
3. *$\nabla^2 \mathcal{L}(\boldsymbol{w}) \boldsymbol{w} = -\nabla \mathcal{L}(\boldsymbol{w})$.*

Proof. Taking gradients with respect to $\boldsymbol{w}$ on both sides of $\mathcal{L}(c\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w})$ gives $\nabla \mathcal{L}(c\boldsymbol{w}) = c^{-1} \nabla \mathcal{L}(\boldsymbol{w})$. Taking gradients again proves that $\nabla^2 \mathcal{L}(c\boldsymbol{w}) = c^{-2} \nabla^2 \mathcal{L}(\boldsymbol{w})$.

Taking derivative with respect to c on both sides of $\mathcal{L}(c\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w})$ gives $\langle \nabla \mathcal{L}(c\boldsymbol{w}), \boldsymbol{w} \rangle = 0$. Taking $c = 1$ gives $\langle \nabla \mathcal{L}(\boldsymbol{w}), \boldsymbol{w} \rangle = 0$. Finally, we take gradients with respect to $\boldsymbol{w}$, then $\nabla^2 \mathcal{L}(\boldsymbol{w}) \boldsymbol{w} + \nabla \mathcal{L}(\boldsymbol{w}) = \mathbf{0}$. □

Now we provide proofs for Lemma 3.1 and Theorem 4.2 in Section 3.

Proof for Lemma 3.1. By (1) and definition of $\boldsymbol{\theta}_t$,

$$\boldsymbol{\theta}_{t+1} = \Pi(\boldsymbol{w}_{t+1}) = \Pi\left((1 - \hat{\eta}\hat{\lambda})\boldsymbol{w}_t - \hat{\eta} \nabla \mathcal{L}(\boldsymbol{w}_t)\right).$$

Since $\Pi(\boldsymbol{x}) = \Pi(c\boldsymbol{x})$ for all $c > 0$, we can divide $\boldsymbol{w}_{t+1}$ by $(1 - \hat{\eta}\hat{\lambda})\|\boldsymbol{w}_t\|_2$ and obtain

$$\boldsymbol{\theta}_{t+1} = \Pi\left(\boldsymbol{\theta}_t - \frac{\hat{\eta}}{(1 - \hat{\eta}\hat{\lambda})\|\boldsymbol{w}_t\|_2} \nabla \mathcal{L}(\boldsymbol{w}_t)\right).$$

Note that $\nabla \mathcal{L}(\boldsymbol{w}_t) = \frac{1}{\|\boldsymbol{w}_t\|_2} \mathcal{L}(\boldsymbol{\theta}_t)$ by Lemma D.1. So we can further rewrite the above formula:

$$\boldsymbol{\theta}_{t+1} = \Pi\left(\boldsymbol{\theta}_t - \frac{\hat{\eta}}{(1 - \hat{\eta}\hat{\lambda})\|\boldsymbol{w}_t\|_2^2} \nabla \mathcal{L}(\boldsymbol{\theta}_t)\right),$$

which proves the lemma by definition of $\tilde{\eta}_t$. □

Proof of Theorem 4.2. The first claim is directly from the Lemma D.2 of [80]. The second claim about $\hat{\eta}$ holds by scrutinizing their proof. □

D.3 Polyak-Łojasiewicz Condition

Definition D.2 (Polyak-Łojasiewicz)**.** For a loss function $\mathcal{L}(\boldsymbol{\theta})$ and a constant $\mu > 0$, we say that $\mathcal{L}$ satisfies μ-Polyak-Łojasiewicz condition (or μ-PL for brevity) on a set U if

$$\frac{1}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2^2 \geq \mu \cdot \left(\mathcal{L}(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in U} \mathcal{L}(\boldsymbol{\theta}')\right),$$

for all $\boldsymbol{\theta} \in U$.

D.3.1 Full Space Optimization

Theorem D.3. *Let $\mathcal{L} : \mathbb{R}^D \to \mathbb{R}$ be a $\mathcal{C}^3$-smooth function, and Γ be a $\mathcal{C}^1$-smooth, D_Γ-dimensional submanifold of $\mathbb{R}^D$, where every $\boldsymbol{\theta} \in \Gamma$ is a local minimizer of $\mathcal{L}$ and $\operatorname{rank}(\nabla^2\mathcal{L}(\boldsymbol{\theta})) = D - D_\Gamma$. If $\mathcal{Z}$ is a compact subset of Γ, then there exist $\epsilon > 0, \mu > 0$ such that $\operatorname{cl}(\mathcal{Z}^\epsilon) \cap \Gamma$ is compact and $\mathcal{L}$ satisfies μ-PL on $\mathcal{Z}^\epsilon$.*

Proof. Since $\mathcal{Z}$ is compact and Γ is a submanifold of $\mathbb{R}^D$, we can choose a small $\delta > 0$ such that $\mathcal{N} := \Gamma \cap \operatorname{cl}(\mathcal{Z}^\delta)$ is compact.

It can be shown that there exists an open neighborhood U of the compact submanifold $\mathcal{N}$ such that for every $\boldsymbol{\theta} \in U$, the nearest point on $\mathcal{N}$, $P(\boldsymbol{\theta}) := \arg\min\{\|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2 : \boldsymbol{\phi} \in \mathcal{N}\}$, exists and is unique [36].

We choose $\epsilon < \delta/2$ to be small enough so that $\epsilon < \delta/2$ and $\mathcal{Z}^\epsilon \subseteq U$. For $\boldsymbol{\theta} \in \mathcal{Z}^\epsilon$, $P(\boldsymbol{\theta})$ lies in the interior of the manifold $\mathcal{N}$, since $d_2(P(\boldsymbol{\theta}), \mathcal{Z}) \leq \|P(\boldsymbol{\theta}) - \boldsymbol{\theta}\|_2 + d(\boldsymbol{\theta}, \mathcal{Z}) < \delta$. Then it must hold that $\boldsymbol{\theta} - P(\boldsymbol{\theta}) \in \mathsf{N}_{P(\boldsymbol{\theta})}(\mathcal{N})$; otherwise the differential of $\|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2^2$ on $\mathcal{N}$ is non-zero at $\boldsymbol{\phi} = P(\boldsymbol{\theta})$, which contradicts to the fact that $P(\boldsymbol{\theta})$ is the nearest point to $\boldsymbol{\theta}$ on $\mathcal{N}$.

Since $\mathcal{L} \in \mathcal{C}^3$, and $\nabla^2\mathcal{L}(\boldsymbol{\theta})$ is of constant rank $D - D_\Gamma$ on the compact manifold $\mathcal{N}$, there exist $\lambda_{\min} > 0, \lambda_{\max} > 0$ such that $\lambda_{D-D_\Gamma}(\nabla^2\mathcal{L}(\boldsymbol{\theta})) \geq \lambda_{\min}$ and $\lambda_1(\nabla^2\mathcal{L}(\boldsymbol{\theta})) \leq \lambda_{\max}$ for all $\boldsymbol{\theta} \in \mathcal{N}$. Also by $\mathcal{L} \in \mathcal{C}^3$ and compactness of $\mathcal{N}$, there exists $C_3 > 0$ such that the following Taylor expansions hold for all $\boldsymbol{\theta} \in \mathcal{Z}^\epsilon$,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(P(\boldsymbol{\theta})) &\leq (\boldsymbol{\theta} - P(\boldsymbol{\theta}))^\top \nabla^2\mathcal{L}(P(\boldsymbol{\theta}))(\boldsymbol{\theta} - P(\boldsymbol{\theta})) + C_3\|\boldsymbol{\theta} - P(\boldsymbol{\theta})\|_2^3, \\
&\leq (\lambda_{\max} + C_3\epsilon) \cdot \|\boldsymbol{\theta} - P(\boldsymbol{\theta})\|_2^2. \\
\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2^2 &\geq (\boldsymbol{\theta} - P(\boldsymbol{\theta}))^\top \left(\nabla^2\mathcal{L}(P(\boldsymbol{\theta}))\right)^2 (\boldsymbol{\theta} - P(\boldsymbol{\theta})) - C_3\|\boldsymbol{\theta} - P(\boldsymbol{\theta})\|_2^3 \\
&\geq (\lambda_{\min}^2 - C_3\epsilon) \cdot \|\boldsymbol{\theta} - P(\boldsymbol{\theta})\|_2^2.
\end{aligned}$$

Then $\mathcal{L}$ satisfies μ-PL on $\mathcal{Z}^\epsilon$ for $\mu := \frac{\lambda_{\min}^2 - C_3\epsilon}{2(\lambda_{\max} + C_3\epsilon)}$, which is positive if we choose ϵ to be small enough in the beginning. □

Theorem D.4. *If $\boldsymbol{\theta}^*$ is a local minimizer of a $\mathcal{C}^2$-smooth function $\mathcal{L} : \mathbb{R}^D \to \mathbb{R}$ and $\mathcal{L}$ is μ-PL on an open neighborhood U of $\boldsymbol{\theta}^*$, then for any $\boldsymbol{\theta}_0$ sufficiently close to $\boldsymbol{\theta}^*$, a gradient flow $\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = -\nabla\mathcal{L}(\boldsymbol{\theta})$ starting with $\boldsymbol{\theta}_0$ converges to a point $\boldsymbol{\theta}_\infty$ as $t \to +\infty$ and $\|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}^*\|_2 = O(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2)$.*

Proof. Let $T := \inf\{t : \boldsymbol{\theta}(t) \notin U\}$. For all $t < T$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*))^{1/2} &= \frac{1}{2}(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*))^{-1/2} \cdot \left\langle \nabla\mathcal{L}(\boldsymbol{\theta}), \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} \right\rangle \\
&= -\frac{1}{2}(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*))^{-1/2} \cdot \|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2 \cdot \left\|\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}\right\|_2.
\end{aligned}$$

Since $(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*))^{1/2} \leq \frac{1}{\sqrt{2\mu}}\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*))^{1/2} \leq -\frac{\sqrt{2\mu}}{2}\left\|\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}\right\|_2.$$

Integrating on both sides proves the following

$$\frac{\sqrt{2\mu}}{2}\int_0^T \left\|\frac{\mathrm{d}\boldsymbol{\theta}(\tau)}{\mathrm{d}\tau}\right\|_2 \mathrm{d}\tau \leq \sqrt{\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}(\boldsymbol{\theta}^*)} = O(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2).$$

So if $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2$ is small enough, then $T = +\infty$ and $\boldsymbol{\theta}(t)$ converges to a point in U as $t \to +\infty$. Moreover, $\|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}^*\|_2 \leq \|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}_0\|_2 + \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2 = O(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2)$. □

D.3.2 Spherical Optimization

Theorem D.5. *Let $\mathcal{L} : \mathbb{R}^D \to \mathbb{R}$ be a $\mathcal{C}^3$-smooth scale-invariant function, and Γ be a $\mathcal{C}^1$-smooth, $(D_\Gamma - 1)$-dimensional submanifold of $\mathbb{S}^{D-1}$, where every $\boldsymbol{\theta} \in \Gamma$ is a local minimizer of $\mathcal{L}$ on $\mathbb{S}^{D-1}$ and $\operatorname{rank}(\nabla^2 \mathcal{L}(\boldsymbol{\theta})) = D - D_\Gamma$. If $\mathcal{Z}$ is a compact subset of Γ, then there exist $\epsilon > 0, \mu > 0$ such that $\operatorname{cl}(\mathcal{Z}^\epsilon) \cap \Gamma$ is compact and $\mathcal{L}$ satisfies μ-PL on $\mathcal{Z}^\epsilon \cap \mathbb{S}^{D-1}$.*

Proof. Let $\Gamma' := \{\nu\boldsymbol{\theta} : \boldsymbol{\theta} \in \Gamma, \nu > 0\}$. Then Γ' is a $\mathcal{C}^1$-smooth, D_Γ-dimensional submanifold of $\mathbb{R}^D$, where every $\boldsymbol{\theta} \in \Gamma$ is a local minimizer of $\mathcal{L}$ on $\mathbb{R}^D$ and $\operatorname{rank}(\nabla^2 \mathcal{L}(\boldsymbol{\theta})) = D - D_\Gamma$. By Theorem D.5, there exist $\epsilon > 0, \mu > 0$ such that $\mathcal{L}$ satisfies μ-PL on $\mathcal{Z}^\epsilon$, so it satisfies μ-PL on $\mathcal{Z}^\epsilon \cap \mathbb{S}^{D-1}$. □

Theorem D.6. *If $\boldsymbol{\theta}^*$ is a local minimizer of a $\mathcal{C}^2$-smooth and scale-invariant function $\mathcal{L} : \mathbb{R}^D \setminus \{\mathbf{0}\} \to \mathbb{R}$ and $\mathcal{L}$ is μ-PL on an open neighborhood U of $\boldsymbol{\theta}^*$ on $\mathbb{S}^{D-1}$, then for any $\boldsymbol{\theta}_0 \in \mathbb{S}^{D-1}$ sufficiently close to $\boldsymbol{\theta}^*$, a gradient flow $\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = -\nabla \mathcal{L}(\boldsymbol{\theta})$ starting with $\boldsymbol{\theta}_0$ converges to a point $\boldsymbol{\theta}_\infty$ as $t \to +\infty$ and $\|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}^*\|_2 = O(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2)$.*

Proof. Since $\mathcal{L}$ is μ-PL on $U \subseteq \mathbb{S}^{D-1}$ and scale-invariant, we know $\mathcal{L}$ is $\frac{\mu}{2}$-PL on an open set in $\mathbb{R}^D$, $U' = \{\boldsymbol{w} : \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \in U, \|\boldsymbol{w}\|_2 \in [\frac{1}{\sqrt{2}}, \sqrt{2}]\}$. The proof is completed by applying Theorem D.4. □

E Supplementary Material for Appendix B.2

E.1 Proof for Theorem B.7

Lemma E.1. *In the setting of Theorem B.7,*

$$\|\boldsymbol{w}_{t+1}\|_2^2 = (1 - \hat{\eta}\hat{\lambda})^2 \|\boldsymbol{w}_t\|_2^2 + \frac{\hat{\eta}^2}{\|\boldsymbol{w}_t\|_2^2} \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2. \tag{13}$$

Proof. Recall that $\boldsymbol{w}_{t+1} = (1 - \hat{\eta}\hat{\lambda})\boldsymbol{w}_t - \hat{\eta}\nabla \mathcal{L}(\boldsymbol{w}_t)$. By scale-invariance, $\langle \nabla \mathcal{L}(\boldsymbol{w}_t), \boldsymbol{w}_t \rangle = 0$. Then by Pythagorean theorem (or Gougu Theorem),

$$\|\boldsymbol{w}_{t+1}\|_2^2 = (1 - \hat{\eta}\hat{\lambda})^2 \|\boldsymbol{w}_t\|_2^2 + \hat{\eta}^2 \|\nabla \mathcal{L}(\boldsymbol{w}_t)\|_2^2.$$

Since $\nabla \mathcal{L}(\boldsymbol{w}_t) = \frac{1}{\|\boldsymbol{w}_t\|_2} \nabla \mathcal{L}(\boldsymbol{\theta}_t)$ by scale-invariance, we can rewrite the last term $\hat{\eta}^2 \|\nabla \mathcal{L}(\boldsymbol{w}_t)\|_2^2$ as $\frac{\hat{\eta}^2}{\|\boldsymbol{w}_t\|_2^2} \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2$, which implies (13). □

Proof for Theorem B.7. Squaring both sides of (13), we have

$$\|\boldsymbol{w}_{t+1}\|_2^4 = (1 - \hat{\eta}\hat{\lambda})^4 \|\boldsymbol{w}_t\|_2^4 + 2(1 - \hat{\eta}\hat{\lambda})^2 \hat{\eta}^2 \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2 + \frac{\hat{\eta}^4}{\|\boldsymbol{w}_t\|_2^4} \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^4.$$

Let $\beta := (1 - \hat{\eta}\hat{\lambda})^4$, $\tilde{v}_t := \frac{1}{\hat{\eta}_t^2} = \frac{(1-\hat{\eta}\hat{\lambda})^2}{\hat{\eta}^2} \|\boldsymbol{w}_t\|_2^4$. Then

$$\begin{aligned}
\tilde{v}_{t+1} &= \frac{(1 - \hat{\eta}\hat{\lambda})^2}{\hat{\eta}^2} \left((1 - \hat{\eta}\hat{\lambda})^4 \|\boldsymbol{w}_t\|_2^4 + 2(1 - \hat{\eta}\hat{\lambda})^2 \hat{\eta}^2 \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2 + \frac{\hat{\eta}^4}{\|\boldsymbol{w}_t\|_2^4} \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^4 \right) \\
&= (1 - \hat{\eta}\hat{\lambda})^4 \tilde{v}_t + 2(1 - \hat{\eta}\hat{\lambda})^4 \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2 + \frac{(1 - \hat{\eta}\hat{\lambda})^2 \hat{\eta}^2}{\|\boldsymbol{w}_t\|_2^4} \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^4 \\
&= \beta \tilde{v}_t + 2\beta \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2 + \frac{\beta}{\tilde{v}_t} \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^4,
\end{aligned}$$

where the last equality uses the definition of β and $\tilde{v}_t$.

Let $\eta := \sqrt{\frac{1-\beta}{2\beta}}$ and $\bar{g}_t := \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2 / \eta$. Then

$$\begin{aligned}
\tilde{v}_{t+1} &= \beta \tilde{v}_t + 2\beta \cdot \eta^2 \bar{g}_t^2 + \frac{\beta}{\tilde{v}_t} \cdot \eta^4 \bar{g}_t^4 \\
&= \beta \tilde{v}_t + (1 - \beta) \bar{g}_t^2 + \frac{1}{4\beta \tilde{v}_t} (1 - \beta)^2 \bar{g}_t^4,
\end{aligned}$$

which is exactly the update rule of GWSI scheduler. □

E.2 Proof for Theorem B.9

Proof. We specify the quasi-RMSprop scheduler $\mathcal{H}_{\text{QRMS}}$ as follows. Let $\mathcal{P}_{\mathcal{H}_{\text{QRMS}}} := \{(\eta, \beta) : \beta = 1 - 2\eta^2, \eta \in (0, \frac{1}{\sqrt{2}})\}$. Given hyperparameters η, β, we define $\beta' := (1 - \frac{1}{4}(1-\beta))^4$ and $\eta' := \sqrt{(\beta'-1)/2}$. Then $\mathcal{H}_{\text{QRMS}}$ produces the effective LRs as a GWSI scheduler with (η', β'):

$$\tilde{\eta}_t \leftarrow \frac{1}{\sqrt{\tilde{v}_t}}, \quad \tilde{v}_{t+1} \leftarrow \beta' \tilde{v}_t + (1-\beta')\hat{g}_t^2 + \frac{1}{4\beta'\tilde{v}_t}(1-\beta')^2 \hat{g}_t^4, \quad \text{where} \quad \hat{g}_t := \|\boldsymbol{g}_t\|_2/\eta'.$$

When $\eta = \sqrt{2\eta_{\text{in}}}$ and $\beta = 1 - 4\eta_{\text{in}}$, it is easy to see that $\mathcal{H}_{\text{QRMS}}$ produces the same effective LRs as GD+WD on scale-invariant functions (Theorem B.7). Now we only need to verify that $\mathcal{H}_{\text{QRMS}}$ is indeed a quasi-RMSprop scheduler.

When β is close enough to 1, we have $\beta' = \beta + O((1-\beta)^2)$, $\beta' \geq 1/2$, $\eta' = \eta \cdot (1 + O(1-\beta))$. Let C_0 be a constant such that $\eta/\eta' \leq C_0, |1 - (\eta/\eta')^2| \leq C_0(1-\beta), 1 - \beta' \leq C_0(1-\beta), |\beta' - \beta| \leq C_0(1-\beta)^2$. Let $\bar{g}_t := \|\boldsymbol{g}_t\|_2/\eta$. Then $\hat{g}_t = (\eta/\eta')\bar{g}_t$, and thus

$$\hat{g}_t \leq C_0 \bar{g}_t, \qquad |\hat{g}_t^2 - \bar{g}_t^2| = |1 - (\eta/\eta')^2| \cdot \bar{g}_t^2 \leq C_0(1-\beta)\bar{g}_t^2.$$

We only need to verify that $\left|\tilde{v}_{t+1} - \left(\beta\tilde{v}_t + (1-\beta)\bar{g}_t^2\right)\right| \leq \delta(\tilde{v}_t) \cdot (1-\beta)^2 \cdot P(\bar{g}_t)$ for some continuous function δ and some polynomial P.

$$\begin{aligned}
&\left|\tilde{v}_{t+1} - \left(\beta\tilde{v}_t + (1-\beta)\bar{g}_t^2\right)\right| \\
&\quad \leq \left|\tilde{v}_{t+1} - \left(\beta'\tilde{v}_t + (1-\beta')\hat{g}_t^2\right)\right| + (1-\beta')\left|\hat{g}_t^2 - \bar{g}_t^2\right| + |\beta' - \beta| \cdot (\tilde{v}_t + \bar{g}_t^2).
\end{aligned}$$

For the first term, we have

$$\left|\tilde{v}_{t+1} - \left(\beta'\tilde{v}_t + (1-\beta')\hat{g}_t^2\right)\right| = \frac{1}{4\beta'\tilde{v}_t}(1-\beta')^2\hat{g}_t^4 \leq \frac{C_0^6}{2\tilde{v}_t} \cdot (1-\beta)^2 \cdot \bar{g}_t^4.$$

For the second and third terms, we have

$$(1-\beta')\left|\hat{g}_t^2 - \bar{g}_t^2\right| \leq C_0^2(1-\beta)^2\bar{g}_t^2, \qquad |\beta' - \beta| \cdot (\tilde{v}_t + \bar{g}_t^2) \leq C_0(1-\beta)^2 \cdot (\tilde{v}_t + \bar{g}_t^2).$$

Finally we can conclude

$$\begin{aligned}
\left|\tilde{v}_{t+1} - \left(\beta\tilde{v}_t + (1-\beta)\bar{g}_t^2\right)\right| &\leq \frac{C_0^6}{2\tilde{v}_t} \cdot (1-\beta)^2 \cdot \bar{g}_t^4 + C_0^2(1-\beta)^2\bar{g}_t^2 + C_0(1-\beta)^2 \cdot (\tilde{v}_t + \bar{g}_t^2) \\
&\leq \left(\frac{C_0^6}{2\tilde{v}_t} + C_0^2 + C_0(1+\tilde{v}_t)\right) \cdot (1-\beta)^2 \cdot (\bar{g}_t^4 + \bar{g}_t^2),
\end{aligned}$$

which verifies that $\mathcal{H}_{\text{QRMS}}$ is indeed a quasi-RMSprop scheduler. □

F Details of the 3D Example

In this section, we give more details for Figure 4.

The loss $\mathcal{L}(\boldsymbol{w})$ is constructed as follows. First, we define the following scale-invariant function:

$$F(x, y, z) := 2 - \frac{x+y}{\sqrt{x^2 - xy + y^2}}.$$

By taking gradient on $\mathbb{S}^2$, one can easily see that the minimum is attained when (x, y) points to $(1, 1)$ in direction. In other words, the minimizer manifold of F is $\Gamma := \{(x, y, z) \in \mathbb{S}^2 : x = y > 0\}$.

Then we fix an orthogonal matrix $\boldsymbol{U}$ (generated randomly) and define $\mathcal{L} : \mathbb{R}^3 \setminus \{\boldsymbol{0}\} \to \mathbb{R}, \boldsymbol{w} \mapsto F(\boldsymbol{U}\boldsymbol{w})$, i.e., the function F after an orthogonal transformation. For plotting the figure, we transform the coordinates back to the domain of F.

The initial point is $\boldsymbol{w}_0 = (0.3, 1.3, 1.2)$ in the domain of F. We run gradient descent on $\mathcal{L}$ with LR $\hat{\eta} = 0.5$ and WD $\hat{\lambda} = 0.08$. It can be seen from the figure that $\boldsymbol{\theta}_t$ does not stop moving after reaching $\boldsymbol{\zeta}_0$. The point that $\boldsymbol{\theta}_t$ eventually oscillate around is $\boldsymbol{\zeta}_* = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$.

One can check that the Hessian matrix of F at $(x, y, z) \in \Gamma$ is

$$H(x,y,z) = \frac{3}{x^2+y^2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \frac{3}{1-z^2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Therefore, the spherical sharpness is controlled by $|z|$. The smaller the absolute value of z, the flatter the minimizer. And the flattest one is $\boldsymbol{\zeta}_*$, which has z-coordinate being zero. This matches with our theory of sharpness-reduction bias as GD+WD moves along Γ and oscillates near $\boldsymbol{\zeta}_*$ in the end.

G Supplementary Material for Section 4.1

G.1 Proof for Descent Lemma

Proof for Lemma 4.1. By Taylor expansion,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_{t+1}) = \mathcal{L}(\boldsymbol{\theta}_t - \tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t)) &\le \mathcal{L}(\boldsymbol{\theta}_t) - \langle \nabla \mathcal{L}(\boldsymbol{\theta}_t), \tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t)\rangle + \frac{1}{2}\lambda_{\max}^{(t)} \|\tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2 \\ &= \mathcal{L}(\boldsymbol{\theta}_t) - \tilde{\eta}_t(1 - \tilde{\eta}_t \lambda_{\max}^{(t)}/2)\|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2,\end{aligned}$$

which proves the lemma. □

G.2 Proof for Theorem 4.4: GD Eventually Enters the EoS Regime

Let $R := C_0(\hat{\lambda}\hat{\eta})^{1/2}$ be a radius so that $\boldsymbol{\theta}^*$ is a minimizer of $\mathcal{L}$ on $U := B^R(\boldsymbol{\theta}) \cap \mathbb{S}^{D-1}$ and μ-PL holds within U, where C_0 is a large constant to be specified later. Let $\lambda_{\max} := \sup\{\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in U\}$. Then we know that $\lambda_{\max} = \lambda_1^{\mathrm{H}}(\boldsymbol{\theta}^*) + O(\hat{\lambda}\hat{\eta})$. Let T_0 be the largest number so that $\boldsymbol{\theta}_t \in U$ for all $t_0 \le t \le T_0$, We define a potential function $\Psi(\boldsymbol{\theta}) := \sqrt{\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)}$.

Lemma G.1. *If $\boldsymbol{\theta}_t \in U$ and $\tilde{\eta}_t < \frac{2}{\lambda_{\max}}$ for some $t_0 \le t < T_0$, then*

$$\Psi(\boldsymbol{\theta}_t) - \Psi(\boldsymbol{\theta}_{t+1}) \ge \frac{\sqrt{2\mu}}{2}(1 - \tilde{\eta}_t \lambda_{\max}/2)\tilde{\eta}_t \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2.$$

Proof. By descent lemma,

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \le \mathcal{L}(\boldsymbol{\theta}_t) - \tilde{\eta}_t(1 - \tilde{\eta}_t \lambda_{\max}/2)\|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2.$$

Then

$$\Psi(\boldsymbol{\theta}_t) - \Psi(\boldsymbol{\theta}_{t+1}) = \frac{\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1})}{\Psi(\boldsymbol{\theta}_t) + \Psi(\boldsymbol{\theta}_{t+1})} \ge \frac{(1 - \tilde{\eta}_t \lambda_{\max}/2)\tilde{\eta}_t \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2^2}{2\Psi(\boldsymbol{\theta}_t)}.$$

By μ-PL, $\|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2 \ge \sqrt{2\mu} \cdot \Psi(\boldsymbol{\theta}_t)$. Combining these together proves the lemma. □

Lemma G.2. *There exists $C_2 = O(1)$ such that $\|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 \le C_2(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*))$ for all $\boldsymbol{\theta} \in U$.*

Proof. It is equivalent to give an upper bound for $\sup_{\boldsymbol{\theta} \in U}\{G(\boldsymbol{\theta})\}$, where $G(\boldsymbol{\theta}) := \frac{\|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2}{\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)}$. Since $\mathcal{L} \in \mathcal{C}^2$, G is continuous in its domain. So it suffices to upper bound $G(\boldsymbol{\theta})$ around every singular point, i.e., around every minimizer of $\mathcal{L}$ on U. And for every minimizer $\boldsymbol{\theta}' \in U$, we can do Taylor expansions for $\mathcal{L}$ and $\nabla \mathcal{L}$ to show that $G(\boldsymbol{\theta})$ is indeed bounded by $O(1)$ around $\boldsymbol{\theta}'$. □

Proof for Theorem 4.4. Recall that $\tilde{\eta}_{t_0} \le \frac{2}{\rho_2} < \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}^*)}$. Let $\delta := 2 - \tilde{\eta}_{t_0}\lambda_{\max} \in (0, 2)$ and T_1 be the largest number so that $\tilde{\eta}_t \le \frac{2-\delta/4}{\lambda_{\max}}$ for all $t_0 \le t \le T_1$. By Lemma G.1, for all $t_0 \le t < \min\{T_0, T_1\}$,

$$\Psi(\boldsymbol{\theta}_t) - \Psi(\boldsymbol{\theta}_{t+1}) \ge \frac{\sqrt{2\mu}}{16}\delta \tilde{\eta}_t \|\nabla \mathcal{L}(\boldsymbol{\theta}_t)\|_2.$$

Telescoping the sum we have $\frac{\sqrt{2\mu}}{16}\delta \sum_{\tau=t_0}^{t-1} \tilde{\eta}_\tau \|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_2 \le \Psi(\boldsymbol{\theta}_0)$. By smoothness of $\mathcal{L}$, $\Psi(\boldsymbol{\theta}_{t_0}) = O(\|\boldsymbol{\theta}_{t_0} - \boldsymbol{\theta}^*\|_2) = O((\hat{\lambda}\hat{\eta})^{1/2})$. So for all $t_0 \le t \le \min\{T_0, T_1\}$,

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t_0}\|_2 \le \sum_{\tau=t_0}^{t-1} \tilde{\eta}_\tau \|\nabla \mathcal{L}(\boldsymbol{\theta}_\tau)\|_2 = O((\hat{\lambda}\hat{\eta})^{1/2}),$$

which implies that $T_0 > T_1$ or $T_0 = T_1 = +\infty$ if we choose C_0 to be large enough.

By Theorem B.7, $\tilde{\eta}_t$ can be seen as the output of a GWSI scheduler with $\beta = (1-\hat{\lambda}\hat{\eta})^4 = 1 - \Theta(\hat{\lambda}\hat{\eta})$ and $\eta = \sqrt{(\beta^{-1}-1)/2} = O((\hat{\lambda}\hat{\eta})^{1/2})$. Then by the update rule,

$$\tilde{\eta}_{t+1}^{-2} = \beta\tilde{\eta}_t^{-2} + (1-\beta)\bar{g}_t^2 + \frac{\tilde{\eta}_t^2}{4\beta}(1-\beta)^2\bar{g}_t^4, \qquad \text{where} \qquad \bar{g}_t := \|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2/\eta. \tag{14}$$

For $\boldsymbol{\theta} \in U$, we have $\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2^2 \le C_2(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*))$ by Lemma G.2. So for all $t_0 \le t \le T_0$, we have

$$\bar{g}_t = \frac{1}{\eta}\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2 \le \frac{C_2}{\eta}(\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}^*)) \le \frac{C_2}{\eta}(\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}(\boldsymbol{\theta}^*)) = O(1).$$

Then $\tilde{\eta}_{t+1}^{-2} = \beta\tilde{\eta}_t^{-2} + (1-\beta)\cdot O(1)$, which implies $\tilde{\eta}_t \ge \Omega(1)$ for some $t = t_0 + O(\frac{1}{1-\beta}\log(\tilde{\eta}_0^{-2}))$. Therefore, we can infer that it must hold for some steps t that $\tilde{\eta}_t \in \left[c_{\min}, \frac{2-\delta/4}{\lambda_{\max}}\right]$, where $c_{\min}$ is some constant.

As $\bar{g}_t \ge 0$, the update rule (14) also implies $\tilde{\eta}_{t+1}^{-2} \ge \beta\tilde{\eta}_t^{-2}$, or equivalently $\tilde{\eta}_{t+1} \le (1-\hat{\eta}\hat{\lambda})^{-2}\tilde{\eta}_t$. This suggests that the number of steps such that $\tilde{\eta}_t \in \left[\frac{2-c}{\lambda_{\max}}, \frac{2-\delta/4}{\lambda_{\max}}\right]$, is at least $\Omega(1/(\hat{\eta}\hat{\lambda}))$. When $\tilde{\eta}_t$ does lie in this range, by Lemma 4.1 we have

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \le \mathcal{L}(\boldsymbol{\theta}_t) - \frac{1}{8}\delta\tilde{\eta}_t\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2^2$$

Combining with μ-PL gives

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_{t+1}) - \mathcal{L}(\boldsymbol{\theta}^*) &\le (1-\mu\delta\tilde{\eta}_t/4)\cdot(\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}^*)) \\ &\le (1-\mu\delta c_{\min}/4)\cdot(\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}^*)).\end{aligned}$$

Thus the loss decays by a constant factor in every step (the factor is in $(0,1)$ as we can choose $c_{\min}$ as small as we want). As this process lasts for at least $\Omega(1/(\hat{\eta}\hat{\lambda}))$ steps, the loss first decreases to $\mathcal{L}(\boldsymbol{\theta}^*) + O((\hat{\eta}\hat{\lambda})^{10})$ after $O(\log\frac{1}{\hat{\eta}\hat{\lambda}})$ steps, then it stays small until T_1.

Now we show that T_1 is finite. By Lemma G.2 and (14), when the loss is $\mathcal{L}(\boldsymbol{\theta}^*) + O((\hat{\eta}\hat{\lambda})^{10})$ the effective LR steadily grows as $\tilde{\eta}_{t+1} = (1-\hat{\eta}\hat{\lambda})^{-2}\tilde{\eta}_t + o(1)$. So at some step t, it must hold that $\tilde{\eta}_t > \frac{2-\delta/4}{\lambda_{\max}}$, which proves $T_1 < +\infty$.

Let C_1 be a large constant, and T_2 be the largest number so that $\tilde{\eta}_t < \frac{2-2C_1(\hat{\eta}\hat{\lambda})^{1/2}}{\lambda_{\max}}$ for all $t_0 \le t \le T_2$. By Lemma G.1, for all $T_1 \le t < \min\{T_2, T_0\}$,

$$\Psi(\boldsymbol{\theta}_t) - \Psi(\boldsymbol{\theta}_{t+1}) \ge \frac{\sqrt{2\mu}}{2}C_1(\hat{\eta}\hat{\lambda})^{1/2}\tilde{\eta}_t\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2.$$

Telescoping the sum gives $(\hat{\eta}\hat{\lambda})^{1/2}\sum_{\tau=T_1}^{t-1}\tilde{\eta}_\tau\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_2 \le O(\Psi(\boldsymbol{\theta}_{T_1})) \le O((\hat{\eta}\hat{\lambda})^5)$, where the last inequality is due to $\mathcal{L}(\boldsymbol{\theta}_{T_1}) = \mathcal{L}(\boldsymbol{\theta}^*) + O((\hat{\eta}\hat{\lambda})^{10})$. This shows that $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{T_1}\|_2 = O((\hat{\eta}\hat{\lambda})^{4.5})$, and thus $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 = O((\hat{\eta}\hat{\lambda})^{1/2})$ by triangle inequality. Now we have $T_0 > T_2$ or $T_0 = T_2 = +\infty$ when C_0 is chosen to be large enough. We can finish the proof with a similar argument as for T_1 to show that T_2 cannot be infinite either. □

G.3 Connection to the EoS Regime in Cohen et al.'s Definition

Now we elaborate how our definition of EoS $\tilde{\eta}_t \approx 2/\lambda_{\max}^{(t)}$ is related to the original definition of EoS in Cohen et al. [24]. In their work, they studied the dynamics of GD (without weight decay). When the loss is $\tilde{\mathcal{L}}$, the update rule is given by $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \hat{\eta}\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)$. They define the EoS regime as a regime in which (1) $\lambda_1(\nabla^2\tilde{\mathcal{L}}(\boldsymbol{w}_t))$ hovers right at, or just above $2/\hat{\eta}$; and (2) the training loss $\tilde{\mathcal{L}}(\boldsymbol{w}_t)$ goes up and down over short timescales, yet still decreases in the long-term run.

View I: Rewriting as GD. We can write GD+WD on scale-invariant loss as GD on scale-invariant loss with L^2-regularization, i.e., GD on $\tilde{\mathcal{L}}(\boldsymbol{w}) := \mathcal{L}(\boldsymbol{w}) + \frac{\hat{\lambda}}{2}\|\boldsymbol{w}\|_2^2$. Now we show that $\tilde{\eta}_t \approx \frac{2}{\lambda_{\max}^{(t)}}$ is essentially the same as $\hat{\eta} \approx \frac{2}{\lambda_1(\nabla^2\tilde{\mathcal{L}}(\boldsymbol{w}_t))}$.

It suffices to show $\tilde{\eta}_t \cdot \lambda_{\max}^{(t)} \approx \hat{\eta} \cdot \lambda_1(\nabla^2\tilde{\mathcal{L}}(\boldsymbol{w}_t))$. When the gradient is small, $\boldsymbol{\theta}_t$ does not move far in one step, then $\lambda_{\max}^{(t)} \approx \lambda_1(\nabla^2\mathcal{L}(\boldsymbol{\theta}_t))$. By scale-invariance, $\lambda_1(\nabla^2\mathcal{L}(\boldsymbol{\theta}_t)) = \|\boldsymbol{w}_t\|_2^2 \cdot \lambda_1(\nabla^2\mathcal{L}(\boldsymbol{w}_t))$ (Lemma D.1). Recall that $\tilde{\eta}_t := \frac{\hat{\eta}}{(1-\hat{\eta}\hat{\lambda})\|\boldsymbol{w}_t\|_2^2}$. Then we have

$$\tilde{\eta}_t \cdot \lambda_{\max}^{(t)} \approx \frac{\hat{\eta}}{(1-\hat{\eta}\hat{\lambda})\|\boldsymbol{w}_t\|_2^2} \cdot \|\boldsymbol{w}_t\|_2^2 \cdot \lambda_1(\nabla^2\mathcal{L}(\boldsymbol{w}_t)) \approx \frac{\hat{\eta}}{(1-\hat{\eta}\hat{\lambda})} \cdot \lambda_1(\nabla^2\mathcal{L}(\boldsymbol{w}_t)) \approx \hat{\eta} \cdot \lambda_1(\nabla^2\mathcal{L}(\boldsymbol{w}_t)).$$

Note that $\hat{\eta} \cdot \lambda_1(\nabla^2\tilde{\mathcal{L}}(\boldsymbol{w}_t)) = \hat{\eta} \cdot \lambda_1(\nabla^2\mathcal{L}(\boldsymbol{w}_t)) + \hat{\eta}\hat{\lambda}$. When $\hat{\eta}\hat{\lambda}$ is small, we can then conclude that $\tilde{\eta}_t \cdot \lambda_{\max}^{(t)} \approx \hat{\eta} \cdot \lambda_1(\nabla^2\tilde{\mathcal{L}}(\boldsymbol{w}_t))$.

Now we show below that our main theorem on sharpness-reduction bias implies the second condition in Cohen et al. [24]'s definition, that is, $\tilde{\mathcal{L}}$ decreases in the long-term run.

For the regularizer $\frac{\hat{\lambda}}{2}\|\boldsymbol{w}_t\|_2^2$, note that $\|\boldsymbol{w}_t\|_2^2 \approx \hat{\eta}/\tilde{\eta}_t \approx \frac{1}{2}\hat{\eta}\lambda_1(\nabla^2\mathcal{L}(\boldsymbol{\theta}_t))$ in the EoS regime. So $\|\boldsymbol{w}_t\|_2^2$ as well as the regularizer is decreasing due to the sharpness-reduction bias (Theorem 4.7).

The scale-invariant part $\mathcal{L}(\boldsymbol{\theta}_t)$ is not always decreasing, but now we show that its time average can be upper bounded by the time average of norm squared. By Lemma E.1 and Lemma D.1, we have

$$\|\boldsymbol{w}_{t+1}\|_2^2 - \|\boldsymbol{w}_t\|_2^2 = (2-\hat{\eta}\hat{\lambda})\hat{\eta}\hat{\lambda}\|\boldsymbol{w}_t\|_2^2 + \hat{\eta}^2\|\nabla\mathcal{L}(\boldsymbol{w}_t)\|_2^2.$$

Since $\|\boldsymbol{w}_t\|_2^2$ decreases in the long run, we know that for any long enough time window T_0 to $T_1 - 1$, $\sum_{t=T_0}^{T_1-1}\|\nabla\mathcal{L}(\boldsymbol{w}_t)\|_2^2 \lesssim \sum_{t=T_0}^{T_1-1}\frac{2\hat{\lambda}}{\hat{\eta}}\|\boldsymbol{w}_t\|_2^2$. Further due to the alignment between the gradient and the top eigenvalue of the Hessian in the EoS regime, we have $\mathcal{L}(\boldsymbol{w}_t) \approx \frac{\|\nabla\mathcal{L}(\boldsymbol{w}_t)\|_2^2}{2\lambda_1(\nabla^2\mathcal{L}(\boldsymbol{w}_t))} \approx \frac{\hat{\eta}}{4}\|\nabla\mathcal{L}(\boldsymbol{w}_t)\|_2^2$. Therefore, we conclude that the average loss over a long enough time window is always upper bounded by the average of squared weight norm, that is,

$$\frac{1}{T_1-T_0}\sum_{t=T_0}^{T_1-1}\mathcal{L}(\boldsymbol{w}_t) \lesssim \frac{1}{T_1-T_0}\sum_{t=T_0}^{T_1-1}\frac{\hat{\eta}}{4}\cdot\frac{2\hat{\lambda}}{\hat{\eta}}\|\boldsymbol{w}_t\|_2^2 \approx \frac{1}{T_1-T_0}\sum_{t=T_0}^{T_1-1}\frac{\hat{\lambda}}{2}\|\boldsymbol{w}_t\|_2^2,$$

where the last step uses the fact that GD operates in EoS. So $\mathcal{L}(\boldsymbol{w}_t)$ decreases in the long-term run.

Combining the above two parts, we can conclude that the regularized loss $\tilde{\mathcal{L}}(\boldsymbol{w}_t)$ has a tendency to decrease in the long-term run.

View II: Generalizing EoS to PGD. For a gradient-based method in general, $2/\hat{\eta}$ should be replaced to the maximum sharpness bound that the loss function is guaranteed to decrease through Taylor expansions, e.g., Cohen et al. [24] derived the bounds exactly for Polyak and Nesterov momentum in Appendix B of their paper. In our definition, we view GD+WD on scale-invariant loss as PGD on $\mathbb{S}^{D-1}$, and thus we define the EoS regime for PGD as the regime where $\tilde{\eta}_t \approx 2/\lambda_{\max}^{(t)}$, where $\lambda_{\max}^{(t)}$ is the local upper bound of spherical sharpness (Lemma 4.1). This captures the first condition of Cohen et al. [24]'s definition. Repeating our argument in View I, we can show the second condition, namely the condition that the loss $\mathcal{L}(\boldsymbol{\theta}_t)$ decreases in the long-term run.

Progressive Sharpening. All the above discussion is about the EoS phenomenon. Another phenomenon identified by Cohen et al. [24] is progressive sharpening, which is the phenomenon that $\lambda_1(\nabla^2\tilde{\mathcal{L}}(\boldsymbol{w}_t))$ tends to increase so long as it is less than $2/\hat{\eta}$. Theorem 4.4 in our paper justifies this phenomenon in View II, i.e., if $\tilde{\eta}_t$ is less than $\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}^*)$, then $\tilde{\eta}_t$ increases until it reaches $\frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}^*)}$. The key insight in our analysis is that WD decreases the norm when gradient is small, and smaller norm leads larger $\lambda_1(\nabla^2\tilde{\mathcal{L}}(\boldsymbol{w}_t))$. This shows that the progressive sharpening phenomenon in our case can be well explained by the interplay between normalization and WD.

H Proof Outlines of Our Theorems on Sharpness Reduction

In this section, we give proof outlines of Theorems B.10 and B.13. The main proof idea for the spherical case is stated in Section 4.2.3, but technically it is easier to state and prove the lemmas for full space optimization. Therefore, we present the full details for the full space case and omit certain details for the spherical case if they are similar to the full space case.

As mentioned in Section 4.2.3, a key ingredient in our proof is to show that the period-2 oscillation drives the parameter to move along the manifold. For both full space and spherical cases, we project $\boldsymbol{\theta}_t$ onto the manifold Γ with a carefully-defined projection function, $\boldsymbol{\phi}_t := \Phi(\boldsymbol{\theta}_t)$. Then we show that after a period of oscillation (2 steps), the projection drifts from $\boldsymbol{\phi}_t$ to a new position $\boldsymbol{\phi}_{t+2}$ approximately along the direction of $\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$. To analyze the speed of each drift, we show that the oscillation can be tracked with two variables $h \in \mathbb{R}$, $u \in \mathbb{R}$, where h is related to the displacement of $\boldsymbol{\theta}_t$ from the manifold, u is related to the closeness of the current dynamic to the edge of stability. We formally define a discrete process called *RMS-drift process* and show that the oscillations in both full space and spherical cases can be regarded as RMS-drift processes.

The rest of the section is organized as follows. In Appendix H.1 we introduce some additional notations. In Appendix H.2 we formally introduce the concept of RMS-drift process. In Appendix H.3 we show how to reduce our analysis in the full space case into studying an RMS-drift process. In Appendix H.4 we show how to reduce our analysis in the spherical case into studying an RMS-drift process. Finally, in Appendix H.5 we analyze the RMS-drift process and show that the projections of parameters in training can be tracked with the sharpness-reduction flow defined in (8).

H.1 Additional Notations

We follow the notations in Appendix D. We also need some additional notations in this section. It is implied by the uniqueness of the top eigenvalue (Assumption B.12 or Assumption 4.6) and $\mathcal{L} \in \mathcal{C}^4$ that $\lambda_1^{\mathrm{H}}(\boldsymbol{\theta})$ is $\mathcal{C}^2$-smooth on Γ, and we can construct a $\mathcal{C}^2$-smooth function $\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\theta})$ such that $\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\theta})$ is a unit top eigenvector of $\boldsymbol{H}(\boldsymbol{\theta})$ on Γ. Let $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\theta})$ be the projection matrix onto the null space of $\boldsymbol{H}_t$, and $\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\theta})$ be the projection matrix onto the space spanned by the eigenvectors of $\boldsymbol{H}_t$ corresponding to non-zero and non-top eigenvalues ($\lambda \neq 0, \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$). For any local minimizer $\boldsymbol{\theta} \in \Gamma$, we define $\gamma(\boldsymbol{\theta}) := \frac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\theta})} \min\left\{\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}) - \lambda_2^{\mathrm{H}}(\boldsymbol{\theta}), \lambda_{D-D_\Gamma}^{\mathrm{H}}(\boldsymbol{\theta})\right\}$ to be the relative eigenvalue gap (between λ_1^{H} and λ_2^{H}, or between $\lambda_{D-D_\Gamma}^{\mathrm{H}}$ and 0). Then $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\theta}), \boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\theta}), \gamma(\boldsymbol{\theta})$ are all $\mathcal{C}^2$-smooth on Γ. Finally, we define $\mu(\boldsymbol{\theta}) := \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\theta})}$.

We define $\Phi(\boldsymbol{\theta})$ to be the convergence point of a gradient flow of $\mathcal{L}$ starting from $\boldsymbol{\theta}$. That is, the limit of $\tilde{\boldsymbol{\theta}}(t)$ as $t \to +\infty$ when $\tilde{\boldsymbol{\theta}}(t)$ is described by the following ODE:

$$\frac{\mathrm{d}\tilde{\boldsymbol{\theta}}}{\mathrm{d}t} = -\nabla \mathcal{L}(\tilde{\boldsymbol{\theta}}), \quad \text{where} \quad \tilde{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}.$$

We leave $\Phi(\boldsymbol{\theta})$ undefined if the ODE does not converge to any point.

For GD/PGD with quasi-RMSprop scheduler (Definition B.5), the state at step t can be written as $(\boldsymbol{\theta}_t, \tilde{v}_t)$, where $\boldsymbol{\theta}_t$ is the trainable parameter in GD/PGD and $\tilde{v}_t$ is the moment estimate. Whenever $\Phi(\boldsymbol{\theta}_t)$ exists and is in Γ, we define

$$\begin{aligned} \boldsymbol{\phi}_t &:= \Phi(\boldsymbol{\theta}_t) & \boldsymbol{H}_t &:= \boldsymbol{H}(\boldsymbol{\phi}_t) \\ \mu_t &:= \mu(\boldsymbol{\phi}_t) & \boldsymbol{U}_t &:= \boldsymbol{I} - \mu_t \boldsymbol{H}_t \\ \boldsymbol{x}_t &:= \boldsymbol{\theta}_t - \boldsymbol{\phi}_t \end{aligned}$$

Most importantly, we define two hidden variables (h_t, u_t) as follows:

$$h_t := \tfrac{1}{\eta}\langle \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t), \boldsymbol{x}_t \rangle, \qquad u_t := \tfrac{1}{\eta}(\mu_t^2 \tilde{v}_t - 1). \tag{15}$$

Note that the h_t defined here *differs* with that in Section 4.2.3 by a factor of $1/\eta$. We introduce this factor for the sake of convenience, and we will use the definition with the factor $1/\eta$ only in our theoretical analysis.

H.2 RMS-drift Process: Introduction

Definition H.1. A *drift state* is described by a tuple $S = (h, u, \boldsymbol{\phi})$ in the drift state space $\mathcal{S} := \mathbb{R} \times \mathbb{R} \times \Gamma$. We say that S is *α-bounded* if $\max\{|h|, |u|\} \le \alpha$.

Definition H.2. Given two drift states $S_t = (h_t, u_t, \boldsymbol{\phi}_t)$, $S_{t+2} = (h_{t+2}, u_{t+2}, \boldsymbol{\phi}_{t+2})$ in the drift state space $\mathcal{S}$, for learning rate $\eta > 0$ and hyperparameter $C_{\mathrm{b}} > 0$, we say that the transition $S_t \to S_{t+2}$ is a *C_0-RMS-drift transition* if for all $\alpha \ge 1$, as long as S_t is α-bounded, S_{t+2} is close to an auxiliary state $S'_{t+2} := (h'_{t+2}, u'_{t+2}, \boldsymbol{\phi}'_{t+2})$ in the following sense:

$$
\begin{aligned}
h'_{t+2} &:= (1 - 2\eta u_t) h_t, & |h_{t+2} - h'_{t+2}| &\le C_0 \alpha^2 |h_t| \eta^2, \\
u'_{t+2} &:= u_t + 4\eta h_t^2 (2C_{\mathrm{b}} + \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2^2) - 2\eta C_{\mathrm{b}}, & |u_{t+2} - u'_{t+2}| &\le C_0 \alpha (1 + h_t^2) \eta^2, \\
\boldsymbol{\phi}'_{t+2} &:= \boldsymbol{\phi}_t - 2\eta^2 h_t^2 \nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t), & \|\boldsymbol{\phi}_{t+2} - \boldsymbol{\phi}'_{t+2}\|_2 &\le C_0 \alpha h_t^2 \eta^3.
\end{aligned}
$$

For a sequence of states $S_0, S_2, S_4, \ldots, S_{2M}$, we say it is a *$C_0$-RMS-drift process* if $S_t \to S_{t+2}$ is a C_0-RMS-drift transition for all even numbers $0 \le t < 2M$.

In our analysis of GD/PGD with quasi-RMSprop scheduler, we can rewrite the dynamics as RMS-drift processes (after a few warm-up steps), where $\boldsymbol{\phi}_t := \Phi(\boldsymbol{\theta}_t)$ is the gradient flow projection of the parameter at step t onto Γ, and h_t, u_t are two hidden variables defined in (15). This RMS-drift process serves as an abstraction of the original dynamics that contains the minimal but sufficient amount of information so that we can compute the continuous approximation for the trajectory of $\boldsymbol{\phi}_t$.

In RMS-drift process, $\boldsymbol{\phi}_t$ evolves as gradient descent on Γ for minimizing $\log \lambda_1^{\mathrm{H}}(\boldsymbol{\theta})$, and the corresponding learning rate is changing with h_t per step. To obtain the final flow approximation (8), we need to sum up the learning rates over time.

An intuitive way to understand RMS-drift process is to use the following first-order continuous approximation with time scaling $(h(\tau), u(\tau), \boldsymbol{\phi}(\tau)) \approx (h_{\tau/\eta}, u_{\tau/\eta}, \boldsymbol{\phi}_{\tau/\eta})$ and ignore all the second order terms $O(\eta^2)$:

$$
\frac{\mathrm{d}h}{\mathrm{d}\tau} = -uh, \qquad \frac{\mathrm{d}u}{\mathrm{d}\tau} = 2h^2(2C_{\mathrm{b}} + \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi})\|_2^2) - C_{\mathrm{b}}, \qquad \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\tau} = \mathbf{0}.
$$

This approximation gives an important insight: h_t and u_t are changing much faster than $\boldsymbol{\phi}_t$ when η is small. Therefore, we can analyze this first-order approximation to obtain an average value of $2\eta^2 h_t^2$, and use this average value as the "effective" learning rate in the flow approximation.

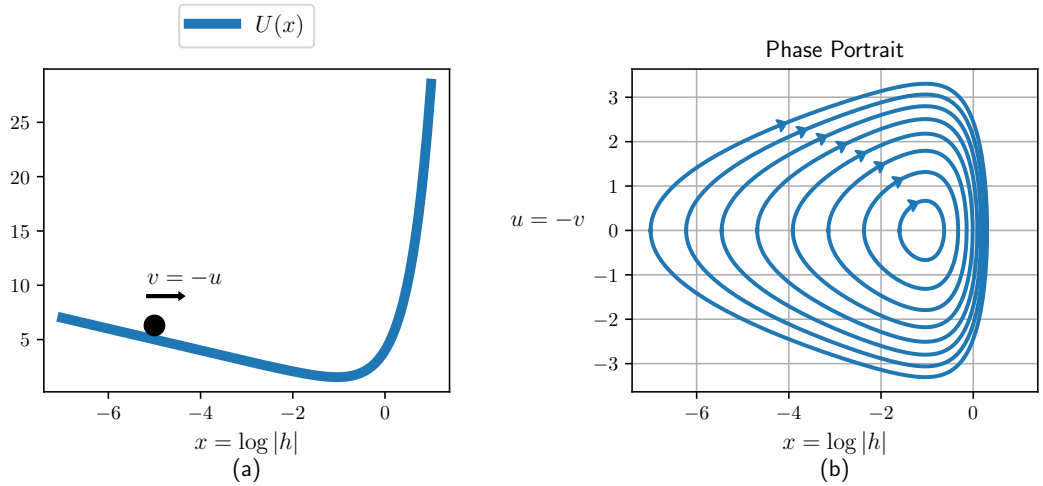


Figure 6: A visualization of the dynamical system described by (16) with $K = 2, C_{\mathrm{b}} = 1$. This system can be associated with a physical system in which a unit-mass particle moves in a potential well $U(x) = K^2 e^{2x} - C_{\mathrm{b}} x$ without any energy loss; see (a). $\log|h|$ can be seen as the position of the particle, and $-u$ can be seen as the velocity. This system must be periodic because of the conservation of energy, and it can be seen clearly from its phase portrait (b).

In fact, (h, u) forms a 1-dimensional Hamiltonian system after a coordinate transformation and evolves periodically in the above ODE. To see this, we can compute the time derivatives of $(\log|h|, -u)$ while letting $K := \sqrt{2C_{\mathrm{b}} + \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi})\|_2^2}$:

$$\frac{\mathrm{d}\log|h|}{\mathrm{d}\tau} = -u, \qquad \frac{\mathrm{d}(-u)}{\mathrm{d}\tau} = -\left(2K^2h^2 - C_{\mathrm{b}}\right). \tag{16}$$

Consider a unit-mass particle in the system with position $x(t)$, velocity $v(t) = x'(t)$ and acceleration $a(t) = v'(t)$. Suppose that this system has the potential energy $U(x) := K^2e^{2x} - C_{\mathrm{b}}x$. Then we can see that the position and velocity of this particle evolve exactly the same as $(\log|h|, -u)$! This also shows that $(\log|h|, -u)$ evolves periodically because 1-dimensional Hamiltonian system with unimodal potential energy must be periodic. See also Figure 6.

In Appendix H.5 below, we will use this observation to obtain the flow approximation nicely. But now we first outline how to reduce the original dynamics to an RMS-drift process.

H.3 Reduction to RMS-drift Process: The Case of Full Space Optimization

Now we outline how to reduce the dynamics to an RMS-drift process in the setting of Theorem B.13, the main theorem for the case of full-space optimization.

H.3.1 Construction of Working Zones

Let $\mathcal{Z} := \{\boldsymbol{\zeta}(t) : t \in [0, T]\} \subseteq \Gamma$ be the set of points passed by the sharpness-reduction flow (8). Inspired by Arora et al. [8], we construct a two-level nested working zone $(\mathcal{Z}^{\epsilon_0}, \mathcal{Z}^{\epsilon_1})$, where $\mathcal{Z}^{\epsilon_0}, \mathcal{Z}^{\epsilon_1}$ are essentially the ϵ_0- and ϵ_1-neighborhoods of $\mathcal{Z}$ with ϵ_0, ϵ_1 carefully chosen. In our later analysis, we will ensure that $\boldsymbol{\theta}_t$ is always in $\mathcal{Z}^{\epsilon_0}$, and its gradient flow projection $\boldsymbol{\phi}_t$ is always well-defined and lies in $\mathcal{Z}^{\epsilon_1} \cap \Gamma$.

The following lemma shows some important properties of the working zone in our construction. We defer the proof to Appendix I.1.

Lemma H.3 (Working Zone Lemma)**.** *There exist $0 < \epsilon_0 < \epsilon_1$ such that $\mathcal{Z}^{\epsilon_0}, \mathcal{Z}^{\epsilon_1}$ satisfy the following:*

1. $\mathrm{cl}(\mathcal{Z}^{\epsilon_1}) \cap \Gamma$ *is compact;*
2. $\mathcal{L}$ *satisfies $\mu_{\mathtt{PL}}$-PL on $\mathcal{Z}^{\epsilon_1}$ for some $\mu_{\mathtt{PL}} > 0$;*
3. Φ *is well-defined on $\mathcal{Z}^{\epsilon_0}$, and $\Phi(\boldsymbol{\theta}) \in \mathcal{Z}^{\epsilon_1} \cap \Gamma$ for all $\boldsymbol{\theta} \in \mathcal{Z}^{\epsilon_0}$;*
4. Φ *is $\mathcal{C}^3$-smooth on $\mathcal{Z}^{\epsilon_0}$;*
5. $\gamma(\boldsymbol{\theta}) \geq \gamma_{\min}$ *holds uniformly on $\mathcal{Z}^{\epsilon_1} \cap \Gamma$ for some $\gamma_{\min} > 0$.*

It can be seen from Lemma H.3 that $\boldsymbol{\phi}_t, \boldsymbol{H}_t, \mu_t, \boldsymbol{U}_t, \boldsymbol{x}_t, h_t, u_t$ are all well-defined as long as $\boldsymbol{\theta}_t \in \mathcal{Z}^{\epsilon_0}$. Now we define some useful notions in the EoS regime. A state is α-bounded if $\boldsymbol{\theta}_t$ is close to Γ and the dynamic is in the EoS regime due to $\tilde{v}_t \approx \mu_t^{-2}$. A state is α-deviated if $\boldsymbol{\theta}_t$ is not too close to Γ.

Definition H.4 (α-Bounded State)**.** We say that the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ at some step t is α-bounded if $\boldsymbol{\theta}_t \in \mathcal{Z}^{\epsilon_0}$, $\|\boldsymbol{x}_t\|_2 \leq \alpha\eta$ and $|u_t| \leq \alpha$.

Definition H.5 (α-Deviated State)**.** We say that the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ at some step t is α-deviated (from the manifold Γ) if $\|\boldsymbol{x}_t\|_2 \geq \eta\exp(-\alpha^2)$ or $\boldsymbol{\theta}_t \notin \mathcal{Z}^{\epsilon_0}$.

Next, we define a quantitative measurement for how much $\boldsymbol{x}_t$ aligns to the top eigenvector of $\boldsymbol{H}_t$. Note that $\boldsymbol{x}_t$ can be decomposed into the projections onto the top eigenspace, the null space, and the space spanned by non-zero and non-top eigenvectors; we can write this decomposition as $\boldsymbol{x}_t = h_t\eta\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + \boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t + \boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t$. The following lemma shows that $\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)$ is always negligible, so we characterize the alignment to the top eigenvector only through comparing $\eta h_t\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$ and $\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t$.

Lemma H.6. *At any step t, if $\boldsymbol{\theta}_t \in \mathcal{Z}^{\epsilon_0}$, then $\|\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 = O(\|\boldsymbol{x}_t\|_2^2)$.*

Proof. Direct consequence of Lemma I.3. □

Definition H.7 (p-Misaligned State)**.** We say that the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ at some step t is at most p-misaligned (to the top eigenvector) if $\boldsymbol{\theta}_t \in \mathcal{Z}^{\epsilon_0}$, $\|\boldsymbol{P}^{\mathrm{H}}_{\mathtt{NZT}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 \le p \cdot h_t\eta$.

H.3.2 Good Initialization

First, we show that the initialization satisfies some desirable properties with high probability. The proof is deferred to Appendix L.1.

Lemma H.8. *There exists $\delta = O(\alpha_0\eta\sqrt{\log(1/\eta)})$ such that the following holds. With probability $1-\delta$, the initial state is $O(\alpha_0\sqrt{\log(1/\delta)})$-bounded, $O(\alpha_0 + \sqrt{\log(1/\delta)})$-deviated, at most $O(1/\delta)$-misaligned, and satisfies $\|\boldsymbol{\phi}_0 - \boldsymbol{\zeta}_0\|_2 \le O(\alpha_0\eta\sqrt{\log(1/\delta)})$.*

H.3.3 Alignment Phase

At initialization, the state is not well-aligned to the top eigenvector. But in the following, we show that it becomes at most $O(|h_t|\eta)$-misaligned after only $\eta^{-o(1)}$ steps. We defer the proofs to Appendix L.2.

The key lemma is the following, which gives good approximations for various important variables.

Lemma H.9. *For small enough base learning rate η, at any step t, if the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ is α-bounded for some $1 \le \alpha \le \eta^{-o(1)}$, then $\boldsymbol{\theta}_{t+1} \in \mathcal{Z}^{\epsilon_0}$, and*

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + O(\|\boldsymbol{x}_t\|_2^2) \tag{17}$$

$$\boldsymbol{x}_{t+1} = (\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{x}_t + O(\|\boldsymbol{x}_t\|_2^2) \tag{18}$$

$$h_{t+1} = -h_t + O(\alpha\|\boldsymbol{x}_t\|_2) \tag{19}$$

$$\|\boldsymbol{P}^{\mathrm{H}}_{\mathtt{NZT}}(\boldsymbol{\phi}_{t+1})\boldsymbol{x}_{t+1}\|_2 \le (1 - 1.9\gamma_{\min})\|\boldsymbol{P}^{\mathrm{H}}_{\mathtt{NZT}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 + O(\|\boldsymbol{x}_t\|_2^2) \tag{20}$$

$$u_{t+1} = u_t + O(\alpha^2\eta) \tag{21}$$

Applying the above lemma through an induction proves the following theorem.

Theorem H.10. *There exists $\delta = O(\alpha_0\eta\sqrt{\log(1/\eta)})$ and $T_1 = O(\log\frac{1}{\eta} + \alpha_0\sqrt{\log(1/\delta)})$ such that the following holds. Let $\alpha_{\max} := \alpha_0\sqrt{\log(1/\delta)}$. If the initial state is $O(\alpha_{\max})$-bounded, $O(\alpha_{\max})$-deviated, at most $O(1/\delta)$-misaligned, and satisfies $\|\boldsymbol{\phi}_0 - \boldsymbol{\zeta}_0\|_2 \le O(\alpha_{\max}\eta)$, then at step $t = T_1$, the state is at most $O(|h_t|\eta)$-misaligned while still being $O(\alpha_{\max})$-bounded, $O(\alpha_{\max})$-deviated, and satisfying $\|\boldsymbol{\phi}_t - \boldsymbol{\zeta}_0\|_2 \le O(\alpha_{\max}\eta)$.*

H.3.4 Drifting Phase

After the alignment phase, the state is now at most $O(|h_t|\eta)$-misaligned. Then we have the following lemma showing that $(h_t, u_t, \boldsymbol{\phi}_t)$ evolves as an $O(1)$-RMS-drift process. We defer the proof to Appendix L.3.

Lemma H.11. *For small enough base learning rate η, at any step t, if for some $1 \le \alpha \le \eta^{-o(1)}$, the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ is α-bounded and at most $O(|h_t|\eta)$-misaligned, then $\boldsymbol{\theta}_{t+2} \in \mathcal{Z}^{\epsilon_0}$, and*

$$h_{t+2} = (1 - 2\eta u_t)h_t + O(\alpha^2|h_t|\eta^2), \tag{22}$$

$$\|\boldsymbol{P}^{\mathrm{H}}_{\mathtt{NZT}}(\boldsymbol{\phi}_{t+2})\boldsymbol{x}_{t+2}\|_2 \le (1 - 1.9\gamma_{\min})^2\|\boldsymbol{P}^{\mathrm{H}}_{\mathtt{NZT}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 + O(h_t^2\eta^2), \tag{23}$$

$$u_{t+2} = u_t + 4\eta h_t^2(2C_{\mathrm{b}} + \|\nabla_\Gamma \log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2^2) - 2\eta C_{\mathrm{b}} + O(\alpha(1+h_t^2)\eta^2), \tag{24}$$

$$\boldsymbol{\phi}_{t+2} = \boldsymbol{\phi}_t - 2\eta^2 h_t^2\nabla_\Gamma \log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(\alpha h_t^2\eta^3). \tag{25}$$

In other words, $(h_t, u_t) \to (h_{t+2}, u_{t+2})$ is an $O(1)$-RMS-drift transition if the state at step t is α-bounded and at most $O(|h_t|\eta)$-misaligned.

H.4 Reduction to RMS-drift Process: The Case of Spherical Optimization

Now we outline how to reduce the dynamics to an RMS-drift process in the setting of Theorem B.13, the main theorem for the case of spherical optimization. The basic logic is the same as the case of full space optimization, so we only list the new lemma and theorem statements here. The proofs in this section are deferred to Appendices I.1 and M.

H.4.1 Construction of Working Zones

We still define $\mathcal{Z} := \{\boldsymbol{\zeta}(t) : t \in [0, T]\} \subseteq \Gamma$. The construction of working zone becomes the following. We defer the proof to Appendix I.1.

Lemma H.12 (Working Zone Lemma)**.** *There exist $0 < \epsilon_0 < \epsilon_1$ such that $\mathcal{Z}^{\epsilon_0}, \mathcal{Z}^{\epsilon_1}$ satisfy the following:*

1. $\mathrm{cl}(\mathcal{Z}^{\epsilon_1}) \cap \Gamma$ *is compact;*
2. $\mathcal{L}$ *satisfies $\mu_{\mathtt{PL}}$-PL on $\mathcal{Z}^{\epsilon_1}$ for some $\mu_{\mathtt{PL}} > 0$;*
3. Φ *is well-defined on $\mathcal{Z}^{\epsilon_0} \cap \mathbb{S}^{D-1}$ and $\Phi(\boldsymbol{\theta}) \in \mathcal{Z}^{\epsilon_1} \cap \Gamma$ for all $\boldsymbol{\theta} \in \mathcal{Z}^{\epsilon_0} \cap \mathbb{S}^{D-1}$;*
4. Φ *is $\mathcal{C}^3$-smooth on $\mathcal{Z}^{\epsilon_1} \cap \mathbb{S}^{D-1}$;*
5. $\gamma(\boldsymbol{\theta}) \geq \gamma_{\min}$ *holds uniformly on $\mathcal{Z}^{\epsilon_1} \cap \Gamma$ for some $\gamma_{\min} > 0$.*

In the working zone, we continue to define α-bounded states, α-deviated states, at most p-misaligned states following the same definitions as the full space case except that the definition of working zone is changed.

Definition H.13 (α-Bounded State)**.** We say that the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ at some step t is α-bounded if $\boldsymbol{\theta}_t \in \mathcal{Z}^{\epsilon_0} \cap \mathbb{S}^{D-1}$, $\|\boldsymbol{x}_t\|_2 \leq \alpha\eta$ and $|u_t| \leq \alpha$.

Definition H.14 (α-Deviated State)**.** We say that the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ at some step t is α-deviated (from the manifold Γ) if $\|\boldsymbol{x}_t\|_2 \geq \eta \exp(-\alpha^2)$ or $\boldsymbol{\theta}_t \notin \mathcal{Z}^{\epsilon_0} \cap \mathbb{S}^{D-1}$.

Lemma H.15. *At any step t, if $\boldsymbol{\theta}_t \in \mathcal{Z}^{\epsilon_0} \cap \mathbb{S}^{D-1}$, then $\|\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 = O(\|\boldsymbol{x}_t\|_2^2)$.*

Proof. Direct consequence of Lemma I.6. □

Definition H.16 (p-Misaligned State)**.** We say that the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ at some step t is at most p-misaligned (to the top eigenvector) if $\boldsymbol{\theta}_t \in \mathcal{Z}^{\epsilon_0} \cap \mathbb{S}^{D-1}$, $\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 \leq p \cdot h_t\eta$.

H.4.2 Good Initialization

Lemma H.17. *The same statement as Lemma H.8 holds for the spherical case.*

H.4.3 Alignment Phase

Lemma H.18. *For small enough base learning rate η, at any step t, if the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ is α-bounded for some $1 \leq \alpha \leq \eta^{-o(1)}$, then $\boldsymbol{\theta}_{t+1} \in \mathcal{Z}^{\epsilon_0} \cap \mathbb{S}^{D-1}$, and (17) to (21) in the full space case continue to hold in the spherical case.*

Theorem H.19. *The same statement as Theorem H.10 holds for the spherical case.*

H.4.4 Drifting Phase

Lemma H.20. *For small enough base learning rate η, at any step t, if for some $1 \leq \alpha \leq \eta^{-o(1)}$, the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ is α-bounded and at most $O(|h_t|\eta)$-misaligned, then $\boldsymbol{\theta}_{t+2} \in \mathcal{Z}^{\epsilon_0} \cap \mathbb{S}^{D-1}$, and (22) to (25) in the full space case continue to hold in the spherical case.*

H.5 RMS-drift Process: Analysis

Now we outline how to obtain the final flow approximation (8) from the RMS-drift process. We say that an RMS-drift process $S_0, \ldots, S_{2M}$ is *in the working zone* if $\boldsymbol{\phi}_t \in \mathcal{Z}^{\epsilon_1} \cap \Gamma$ for all even numbers $0 \leq t \leq 2M$. We focus on RMS-drift processes in the working zone, and later we will show that the RMS-drift processes of interest are indeed in the working zone.

First, we define a potential function that resembles the total energy (or Hamiltonian) in physics.

Definition H.21 (Energy)**.** For a drift state $S = (h, u, \boldsymbol{\phi})$, we define the energy $E(S)$ as follows:

$$E(S) := \frac{1}{2}u^2 + (2C_{\mathrm{b}} + \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi})\|_2^2)h^2 + C_{\mathrm{b}} \log \frac{1}{|h|}.$$

If the energy is bounded by α^2 at some step t, then it is easy to see that the state at step t is $O(\alpha)$-bounded and $O(\alpha)$-deviated.

The first key lemma is the conservation of energy in RMS-drift process, which shows that the energy is preserved for $O(1/\eta^2)$ steps. The proof is deferred to Appendix N.1.

Theorem H.22. *For an $O(1)$-RMSdrift process $S_0, \ldots, S_{2M}$ in the working zone, if $E(S_0) \le \alpha^2$ for some parameter $1 \le \alpha \le \eta^{-o(1)}$ and $M = O(1/\eta^2)$, then $E(S_t) = O(\alpha^2)$ for all even numbers $0 \le t \le 2M$.*

Next, we show that as long as the states are $\eta^{-o(1)}$-bounded, the RMS-drift process tracks the sharpness-reduction flow nicely. We defer the proof to Appendix N.2.

Theorem H.23. *Let $\boldsymbol{\zeta} : [0, T] \mapsto \mathcal{Z}^{\epsilon_1} \cap \Gamma$ be a sharpness-reduction flow defined in (8). For an $O(1)$-RMS-drift process $S_0, \ldots, S_{2M}$ in the working zone, where $M := \lfloor \frac{T}{2\eta^2} \rfloor$, if $\|\boldsymbol{\phi}_0 - \boldsymbol{\zeta}(0)\|_2 \le O(\alpha^2 \eta^{1/2})$, and S_t is $O(\alpha)$-bounded for all even numbers $0 \le t \le 2M$, then $\|\boldsymbol{\phi}_t - \boldsymbol{\zeta}(t\eta^2)\|_2 \le O(\alpha^2 \eta^{1/2})$ for all even numbers $0 \le t \le 2M$.*

H.6 Finalizing Proofs

Proof for Theorem B.13. First, we use Lemma H.8 to ensure a good initialization. Then we apply Theorem H.10 to show that the state at $t = T_1$ is at most $O(|h_t|\eta)$-misaligned, $O(\alpha_{\max})$-bounded, $O(\alpha_{\max})$-deviated, and satisfies $\|\boldsymbol{\phi}_t - \boldsymbol{\zeta}_0\|_2 \le O(\alpha_{\max}\eta)$. By smoothness of $\boldsymbol{\zeta}$, we also have $\|\boldsymbol{\phi}_t - \boldsymbol{\zeta}(t\eta^2)\|_2$. Then we consider the even-indexed and odd-indexed steps separately and do an induction for each of them to show that the later dynamics (1) has bounded energy $O(\alpha_{\max}^2)$ (Theorem H.22); (2) stays at most $O(|h_t|\eta)$-misaligned (Lemma H.11); (3) follows the flow (Theorem H.23). The proof is done when the induction proceeds to T/η^2. □

Proof for Theorem B.10. Same as above but we invoke the spherical version of lemmas and theorems. □

I Lemmas for Working Zones

I.1 Construction of Working Zones

Proof for Lemma H.3. By Assumption B.11 and Theorem D.3, there exists ϵ_1 such that Items 1 and 2 hold. By Theorem D.4, we can choose $\epsilon_0' > 0$ small enough so that any gradient flow starting in $\mathcal{Z}^{\epsilon_0'}$ moves at most $O(\epsilon_0')$ in distance and converges in $\mathcal{Z}^{\epsilon_1}$. We can further combine this with the results by Falconer [34] to show that Φ is $\mathcal{C}^3$-smooth in a neighborhood of $\mathcal{Z}$. Then we can choose $\epsilon_0 > 0$ small enough so that $\epsilon_0 < \epsilon_0'$ and $\mathcal{Z}^{\epsilon_0}$ is a subset of that neighborhood, which ensures Items 3 and 4. Finally, Item 5 is directly implied by the compactness of $\mathrm{cl}(\mathcal{Z}^{\epsilon_1}) \cap \Gamma$. □

Proof for Lemma H.12. We can use a similar argument as above but with Theorems D.5 and D.6. □

I.2 Gradient Flow Projection

In the working zone, $\Phi(\boldsymbol{\theta})$ is well-defined and $\mathcal{C}^3$-smooth. Now we highlight some useful properties.

I.2.1 Full Space Optimization

The following two lemmas are from Li et al. [81].

Lemma I.1. *For $\boldsymbol{\phi} \in \mathcal{Z}^{\epsilon_0} \cap \Gamma$, $\partial\Phi_{\boldsymbol{\phi}}[\boldsymbol{x}] = \boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\boldsymbol{x}$, which also equals to the projection of $\boldsymbol{x}$ onto the tangent space $\mathsf{T}_{\boldsymbol{\phi}}(\Gamma)$ at $\boldsymbol{\phi}$.*

Lemma I.2. *For $\boldsymbol{\phi} \in \mathcal{Z}^{\epsilon_0} \cap \Gamma$ and $\boldsymbol{x} \in \mathsf{N}_{\boldsymbol{\phi}}(\Gamma)$, $\partial^2\Phi_{\boldsymbol{\phi}}[\boldsymbol{x}, \boldsymbol{x}] = \mathbf{0}$.*

The following lemma can be proved by Taylor expansion.

Lemma I.3. *For $\boldsymbol{\theta} \in \mathcal{Z}^{\epsilon_0}$ and $\boldsymbol{\phi} = \Phi(\boldsymbol{\theta})$, then $\|\partial\Phi_{\boldsymbol{\phi}}[\boldsymbol{\theta} - \boldsymbol{\phi}]\|_2 \le O(\|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2^2)$.*

Proof. When $\|\boldsymbol{\theta}-\boldsymbol{\phi}\|_2$ is small enough, the linear interpolation of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ lies in $\mathcal{Z}^{\epsilon_1}$. By Taylor expansion, $\Phi(\boldsymbol{\theta})=\Phi(\boldsymbol{\phi})+\partial\Phi_{\boldsymbol{\phi}}[\boldsymbol{\theta}-\boldsymbol{\phi}]+O(\|\boldsymbol{\theta}-\boldsymbol{\phi}\|_2^2)$. In fact, $\Phi(\boldsymbol{\theta})=\boldsymbol{\phi}=\Phi(\boldsymbol{\phi})$. So we can conclude $\partial\Phi_{\boldsymbol{\phi}}[\boldsymbol{\theta}-\boldsymbol{\phi}]=O(\|\boldsymbol{\theta}-\boldsymbol{\phi}\|_2^2)$. □

I.2.2 Spherical Optimization

In the spherical case, we have the following lemmas that highly resemble those in the full space case. All these lemmas can be proved in the same manner: we can first define $\Gamma':=\{\nu\boldsymbol{\theta}:\boldsymbol{\theta}\in\Gamma,\nu\in(1/2,2)\}$ and apply the counterparts of these lemmas in the full space case; then translate the results back to the spherical case.

Lemma I.4. *For $\boldsymbol{\phi}\in\mathcal{Z}^{\epsilon_0}\cap\Gamma$, $(\boldsymbol{I}-\boldsymbol{\phi}\boldsymbol{\phi}^\top)\partial\Phi_{\boldsymbol{\phi}}[\boldsymbol{x}]=(\boldsymbol{I}-\boldsymbol{\phi}\boldsymbol{\phi}^\top)\boldsymbol{P}_0^{\mathrm{H}}\boldsymbol{x}$, which also equals to the projection of $\boldsymbol{x}$ onto the tangent space $\mathsf{T}_{\boldsymbol{\phi}}(\Gamma)$ at $\boldsymbol{\phi}$.*

Lemma I.5. *For $\boldsymbol{\phi}\in\mathcal{Z}^{\epsilon_0}\cap\Gamma$ and $\boldsymbol{x}\in\mathsf{N}_{\boldsymbol{\phi}}(\Gamma)$, if $\langle\boldsymbol{x},\boldsymbol{\phi}\rangle=0$, then $\partial^2\Phi_{\boldsymbol{\phi}}[\boldsymbol{x},\boldsymbol{x}]=\mathbf{0}$.*

Lemma I.6. *For $\boldsymbol{\theta}\in\mathcal{Z}^{\epsilon_0}\cap\mathbb{S}^{D-1}$ and $\boldsymbol{\phi}=\Phi(\boldsymbol{\theta})$, then $\|\partial\Phi_{\boldsymbol{\phi}}[\boldsymbol{\theta}-\boldsymbol{\phi}]\|_2\le O(\|\boldsymbol{\theta}-\boldsymbol{\phi}\|_2^2)$.*

J Lemmas for Gradient Descent

J.1 Full Space Optimization

Lemma J.1. *Under Assumption B.11, consider one step of gradient descent $\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t-\tilde{\eta}_t\nabla\mathcal{L}(\boldsymbol{\theta}_t)$ with effective learning rate $\tilde{\eta}_t$. Let $\boldsymbol{\phi}\in\mathcal{Z}^{\epsilon_0}\cap\Gamma$ be a local minimizer. If for some parameters $\alpha=\eta^{-o(1)}$ and $r=O(\alpha)$, $\|\boldsymbol{\theta}_t-\boldsymbol{\phi}\|_2=O(r\eta)$ and $\tilde{\eta}_t=\frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}+O(\alpha\eta)$, then after one step, we have the following approximations for $\boldsymbol{\theta}_{t+1}$:*

1. *Zeroth-order approximation:*
$$\boldsymbol{\theta}_{t+1}-\boldsymbol{\phi}=O(r\eta).$$
2. *First-order approximation:*
$$\begin{aligned}\boldsymbol{\theta}_{t+1}-\boldsymbol{\phi}&=(\boldsymbol{I}-\tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t-\boldsymbol{\phi})+O(r^2\eta^2)\\&=\left(\boldsymbol{I}-\tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}(\boldsymbol{\phi})\right)(\boldsymbol{\theta}_t-\boldsymbol{\phi})+O(\alpha r\eta^2).\end{aligned}$$
3. *Second-order approximation:*
$$\boldsymbol{\theta}_{t+1}-\boldsymbol{\phi}=(\boldsymbol{I}-\tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t-\boldsymbol{\phi})-\tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{\theta}_t-\boldsymbol{\phi},\boldsymbol{\theta}_t-\boldsymbol{\phi}]+O(\alpha r^2\eta^3).$$

Proof. For the zeroth-order approximation, we expand $\mathcal{L}(\boldsymbol{\theta}_t)$ around $\boldsymbol{\phi}$ using Taylor expansion:
$$\nabla\mathcal{L}(\boldsymbol{\theta}_t)=\nabla\mathcal{L}(\boldsymbol{\phi})+O(r\eta)=O(r\eta).$$
So $\boldsymbol{\theta}_{t+1}-\boldsymbol{\phi}=\boldsymbol{\theta}_t-\boldsymbol{\phi}-\tilde{\eta}_t\nabla\mathcal{L}(\boldsymbol{\theta}_t)=O(r\eta)$.

For the first-order approximation, we expand $\mathcal{L}(\boldsymbol{\theta}_t)$ around $\boldsymbol{\phi}$ using Taylor expansion:
$$\begin{aligned}\nabla\mathcal{L}(\boldsymbol{\theta}_t)&=\nabla\mathcal{L}(\boldsymbol{\phi})+\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t-\boldsymbol{\phi})+O((r\eta)^2)\\&=\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t-\boldsymbol{\phi})+O(r^2\eta^2).\end{aligned}$$
So $\boldsymbol{\theta}_{t+1}-\boldsymbol{\phi}=\boldsymbol{\theta}_t-\boldsymbol{\phi}-\tilde{\eta}_t\nabla\mathcal{L}(\boldsymbol{\theta}_t)=(\boldsymbol{I}-\tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t-\boldsymbol{\phi})+O(r^2\eta^2)$.

For the second-order approximation, we again use Taylor expansion to expand $\mathcal{L}(\boldsymbol{\theta}_t)$:
$$\begin{aligned}\nabla\mathcal{L}(\boldsymbol{\theta}_t)&=\nabla\mathcal{L}(\boldsymbol{\phi})+\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t-\boldsymbol{\phi})+\frac{1}{2}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{\theta}_t-\boldsymbol{\phi},\boldsymbol{\theta}_t-\boldsymbol{\phi}]+O((r\eta)^3)\\&=\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t-\boldsymbol{\phi})+\frac{1}{2}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{\theta}_t-\boldsymbol{\phi},\boldsymbol{\theta}_t-\boldsymbol{\phi}]+O(r^3\eta^3).\end{aligned}$$

So we have

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} - \boldsymbol{\phi} &= \boldsymbol{\theta}_t - \boldsymbol{\phi} - \tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t) \\
&= (\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t - \boldsymbol{\phi}) - \frac{\tilde{\eta}_t}{2} \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{\theta}_t - \boldsymbol{\phi}, \boldsymbol{\theta}_t - \boldsymbol{\phi}] + O(r^3\eta^3) \\
&= (\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t - \boldsymbol{\phi}) - \left(\tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} + O(\alpha\eta)\right) \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{\theta}_t - \boldsymbol{\phi}, \boldsymbol{\theta}_t - \boldsymbol{\phi}] + O(r^3\eta^3) \\
&= (\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t - \boldsymbol{\phi}) - \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{\theta}_t - \boldsymbol{\phi}, \boldsymbol{\theta}_t - \boldsymbol{\phi}] + O(\alpha r^2\eta^3),
\end{aligned}$$

where the last equality uses $r = O(\alpha)$. □

Lemma J.2. *Under Assumption B.11, consider two steps of gradient descent with effective learning rates $\tilde{\eta}_t$ and $\tilde{\eta}_{t+1}$:*

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \tilde{\eta}_t \nabla \mathcal{L}(\boldsymbol{\theta}_t), \\
\boldsymbol{\theta}_{t+2} &= \boldsymbol{\theta}_{t+1} - \tilde{\eta}_{t+1} \nabla \mathcal{L}(\boldsymbol{\theta}_{t+1}).
\end{aligned}$$

Let $\boldsymbol{\phi} \in \mathcal{Z}^{\epsilon_0} \cap \Gamma$ be a local minimizer. If for some parameters $\alpha = \eta^{-o(1)}$ and $r = O(\alpha)$, $\|\boldsymbol{\theta}_t - \boldsymbol{\phi}\|_2 = O(r\eta)$, $\tilde{\eta}_t = \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} + O(\alpha\eta)$, and $\tilde{\eta}_{t+1} = \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} + O(\alpha\eta)$, then after two steps,

$$\boldsymbol{\theta}_{t+2} - \boldsymbol{\phi} = (\boldsymbol{I} - \tilde{\eta}_{t+1} \boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t - \boldsymbol{\phi}) - \boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_t - \boldsymbol{\phi}) + O(\alpha r^2\eta^3),$$

where $\boldsymbol{\psi}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}) := \frac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} \left(\boldsymbol{U}_{\boldsymbol{\phi}} \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}] + \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}, \boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}]\right)$, $\boldsymbol{U}_{\boldsymbol{\phi}} := \boldsymbol{I} - \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} \boldsymbol{H}(\boldsymbol{\phi})$.

Proof. Let $\hat{\boldsymbol{x}}_\tau := \boldsymbol{\theta}_\tau - \boldsymbol{\phi}$ for all $\tau \in \{t, t+1, t+2\}$. By Lemma J.1, we have $\hat{\boldsymbol{x}}_{t+1} = O(r\eta)$ and the following first-order and second-order approximations for $\hat{\boldsymbol{x}}_{t+1}$:

$$\begin{aligned}
\hat{\boldsymbol{x}}_{t+1} &= \boldsymbol{U}_{\boldsymbol{\phi}} \hat{\boldsymbol{x}}_t + O(\alpha r\eta^2). \\
\hat{\boldsymbol{x}}_{t+1} &= (\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}))\hat{\boldsymbol{x}}_t - \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}_t] + O(\alpha r^2\eta^3).
\end{aligned}$$

Note that $\tilde{\eta}_{t+1} = \tilde{\eta}_t + O((1+r^2)\eta^2) = \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} + O(\alpha\eta)$. By Lemma J.1 again, we have the following second-order approximation for $\hat{\boldsymbol{x}}_{t+2}$:

$$\hat{\boldsymbol{x}}_{t+2} = (\boldsymbol{I} - \tilde{\eta}_{t+1} \boldsymbol{H}(\boldsymbol{\phi}))\hat{\boldsymbol{x}}_{t+1} - \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\hat{\boldsymbol{x}}_{t+1}, \hat{\boldsymbol{x}}_{t+1}] + O(\alpha r^2\eta^3).$$

Now we combine the two steps together.

$$\begin{aligned}
\hat{\boldsymbol{x}}_{t+2} &= (\boldsymbol{I} - \tilde{\eta}_{t+1} \boldsymbol{H}(\boldsymbol{\phi})) \left((\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}))\hat{\boldsymbol{x}}_t - \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}_t] + O(\alpha r^2\eta^3)\right) \\
&\quad - \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}_t + O(\alpha r\eta^2), \boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}_t + O(\alpha r\eta^2)] + O(\alpha r^2\eta^3) \\
&= (\boldsymbol{I} - \tilde{\eta}_{t+1} \boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}))\hat{\boldsymbol{x}}_t - \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} \boldsymbol{U}_{\boldsymbol{\phi}} \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{x}}_t] \\
&\quad - \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} \partial^3 \mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}_t, \boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}_t] + O(\alpha r^2\eta^3) \\
&= (\boldsymbol{I} - \tilde{\eta}_{t+1} \boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}))\hat{\boldsymbol{x}}_t - \boldsymbol{\psi}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_t) + O(\alpha r^2\eta^3),
\end{aligned}$$

where the second equality uses $\boldsymbol{I} - \tilde{\eta}_{t+1} \boldsymbol{H}(\boldsymbol{\phi}) = \boldsymbol{U}_{\boldsymbol{\phi}} + O(\alpha\eta)$. □

The following lemma characterizes the function $\boldsymbol{\psi}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}})$ in Lemma J.2 when $\hat{\boldsymbol{x}}$ is the top eigenvector of $\boldsymbol{H}(\boldsymbol{\phi})$. We will use this property later in Appendix L.3.

Lemma J.3. *Under Assumptions B.11 and B.12, for $\boldsymbol{\phi} \in \Gamma$,*

$$\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})) = (2\boldsymbol{I} - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} \boldsymbol{H}(\boldsymbol{\phi})) \nabla \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}),$$

where $\boldsymbol{\psi}_{\boldsymbol{\phi}}$ is defined as in Lemma J.2. Moreover,

$$\begin{aligned}
\langle \boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}) \rangle &= 0, \\
\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}) \boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})) &= 2\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}).
\end{aligned}$$

Proof. Note that $\boldsymbol{U}_{\boldsymbol{\phi}}\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}) = -\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})$. Then we can rewrite $\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}))$ as follows:

$$\begin{aligned}
\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})) &= \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\left(\boldsymbol{U}_{\boldsymbol{\phi}}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})] + \partial^3\mathcal{L}_{\boldsymbol{\phi}}[-\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}), -\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})]\right) \\
&= \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}(2\boldsymbol{I} - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}(\boldsymbol{\phi}))\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})] \\
&= \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}(2\boldsymbol{I} - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}(\boldsymbol{\phi}))\nabla\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}) \\
&= (2\boldsymbol{I} - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}(\boldsymbol{\phi}))\nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}).
\end{aligned}$$

Since $(2\boldsymbol{I} - \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}(\boldsymbol{\phi}))\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}) = 2\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}) - 2\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}) = \boldsymbol{0}$, we have $\langle\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})\rangle = 0$.

The projection matrix onto the tangent space of Γ at $\boldsymbol{\phi}$ equals to $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})$ by Lemma I.1. Then

$$\begin{aligned}
\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})) = \boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})(2\boldsymbol{I} - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}(\boldsymbol{\phi}))\nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}) &= (2\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}) - \boldsymbol{0})\nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}) \\
&= 2\nabla_{\Gamma}\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}),
\end{aligned}$$

which proves the lemma. □

J.2 Spherical Optimization

Lemma J.4. *Under Assumption 4.5, consider one step of projected gradient descent on $\mathbb{S}^{D-1}$, $\boldsymbol{\theta}_{t+1} = \Pi(\boldsymbol{\theta}_t - \tilde{\eta}_t\nabla\mathcal{L}(\boldsymbol{\theta}_t))$ with effective learning rate $\tilde{\eta}_t$. Let $\boldsymbol{\phi} \in \mathcal{Z}^{\epsilon_0} \cap \Gamma$ be a local minimizer. If for some parameters $\alpha = \eta^{-o(1)}$ and $r = O(\alpha)$, $\|\boldsymbol{\theta}_t - \boldsymbol{\phi}\|_2 = O(r\eta)$ and $\tilde{\eta}_t = \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} + O(\alpha\eta)$, then after one step, we have the following approximations for $\boldsymbol{\theta}_{t+1}$:*

1. *Zeroth-order approximation:*

$$\boldsymbol{\theta}_{t+1} - \boldsymbol{\phi} = O(r\eta).$$

2. *First-order approximation:*

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} - \boldsymbol{\phi} &= (\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t - \boldsymbol{\phi}) + O(r^2\eta^2) \\
&= \left(\boldsymbol{I} - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}(\boldsymbol{\phi})\right)(\boldsymbol{\theta}_t - \boldsymbol{\phi}) + O(\alpha r\eta^2).
\end{aligned}$$

3. *Second-order approximation:*

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} - \boldsymbol{\phi} = (\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t - \boldsymbol{\phi}) &- \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{\theta}_t - \boldsymbol{\phi}, \boldsymbol{\theta}_t - \boldsymbol{\phi}] \\
&- \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2\boldsymbol{\phi} + O(\alpha r^2\eta^3).
\end{aligned}$$

Proof. Let $\hat{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_t - \tilde{\eta}_t\nabla\mathcal{L}(\boldsymbol{\theta}_t)$. Then $\boldsymbol{\theta}_{t+1} = \frac{\hat{\boldsymbol{\theta}}_{t+1}}{\|\hat{\boldsymbol{\theta}}_{t+1}\|_2}$.

Since $\nabla\mathcal{L}(\boldsymbol{\theta}_t)$ is perpendicular to $\boldsymbol{\theta}_t$, $\|\hat{\boldsymbol{\theta}}_{t+1}\|_2 = \sqrt{\|\boldsymbol{\theta}_t\|_2^2 + \tilde{\eta}_t^2\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2^2}$. By Taylor expansion, $\nabla\mathcal{L}(\boldsymbol{\theta}_t) = O(r\eta)$. Then the norm of $\hat{\boldsymbol{\theta}}_{t+1}$ can be estimated by

$$\|\hat{\boldsymbol{\theta}}_{t+1}\|_2 = \sqrt{\|\boldsymbol{\theta}_t\|_2^2 + \tilde{\eta}_t^2\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2^2} = \sqrt{1 + O(r^2\eta^2)} = 1 + O(r^2\eta^2).$$

Then $\boldsymbol{\theta}_{t+1} = \frac{\hat{\boldsymbol{\theta}}_{t+1}}{\|\hat{\boldsymbol{\theta}}_{t+1}\|_2} = \hat{\boldsymbol{\theta}}_{t+1} + O(r^2\eta^2)$. Combining with Lemma J.1 proves the zeroth- and first-order approximations.

To prove the second-order approximation, we need a tighter estimate for the norm of $\hat{\boldsymbol{\theta}}_{t+1}$. By Taylor expansion, we have $\nabla\mathcal{L}(\boldsymbol{\theta}_t) = \boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi}) + O(r^2\eta^2)$. Squaring both sides gives

$\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2^2 = \|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2 + O(r^3\eta^3)$. Then we have

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_{t+1}\|_2 &= \sqrt{\|\boldsymbol{\theta}_t\|_2^2 + \tilde{\eta}_t^2\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2^2} \\
&= \sqrt{1 + \left(\tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} + O(\alpha\eta)\right)^2 (\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2 + O(r^3\eta^3))} \\
&= \sqrt{1 + \tfrac{4}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2 + O(\alpha r^2\eta^3)} \\
&= \sqrt{1 + \tfrac{4}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2} + O(\alpha r^2\eta^3) \\
&= 1 + \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2 + O(\alpha r^2\eta^3).
\end{aligned}$$

So $\boldsymbol{\theta}_{t+1}$ can be estimated by

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} = \frac{\hat{\boldsymbol{\theta}}_{t+1}}{\|\hat{\boldsymbol{\theta}}_{t+1}\|_2} &= \frac{\hat{\boldsymbol{\theta}}_t}{1 + \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2 + O(\alpha r^2\eta^3)} \\
&= \left(1 - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2\right)\hat{\boldsymbol{\theta}}_t + O(\alpha r^2\eta^3).
\end{aligned}$$

Then $\boldsymbol{\theta}_{t+1} - \boldsymbol{\phi}$ can be estimated by

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} - \boldsymbol{\phi} &= \left(1 - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2\right)\hat{\boldsymbol{\theta}}_t - \boldsymbol{\phi} + O(\alpha r^2\eta^3) \\
&= \left(1 - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2\right)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}) \\
&\quad - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2\boldsymbol{\phi} + O(\alpha r^2\eta^3) \\
&= (\hat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}) + O(r^3\eta^3) \\
&\quad - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2\boldsymbol{\phi} + O(\alpha r^2\eta^3) \\
&= (\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t - \boldsymbol{\phi}) - \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{\theta}_t - \boldsymbol{\phi}, \boldsymbol{\theta}_t - \boldsymbol{\phi}] \\
&\quad - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})(\boldsymbol{\theta}_t - \boldsymbol{\phi})\|_2^2\boldsymbol{\phi} + O(\alpha r^2\eta^3),
\end{aligned}$$

where the last equality uses the second-order approximation in Lemma J.1. □

Lemma J.5. *Under Assumption 4.5, consider two steps of projectioned gradient descent on $\mathbb{S}^{D-1}$ with effective learning rates $\tilde{\eta}_t$ and $\tilde{\eta}_{t+1}$:*

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} &= \Pi(\boldsymbol{\theta}_t - \tilde{\eta}_t\nabla\mathcal{L}(\boldsymbol{\theta}_t)), \\
\boldsymbol{\theta}_{t+2} &= \Pi(\boldsymbol{\theta}_{t+1} - \tilde{\eta}_{t+1}\nabla\mathcal{L}(\boldsymbol{\theta}_{t+1})).
\end{aligned}$$

Let $\boldsymbol{\phi} \in \mathcal{Z}^{\epsilon_0} \cap \Gamma$ be a local minimizer. If for some parameters $\alpha = \eta^{-o(1)}$ and $r = O(\alpha)$, $\|\boldsymbol{\theta}_t - \boldsymbol{\phi}\|_2 = O(r\eta)$, $\tilde{\eta}_t = \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} + O(\alpha\eta)$, and $\tilde{\eta}_{t+1} = \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} + O(\alpha\eta)$, then after two steps,

$$\boldsymbol{\theta}_{t+2} - \boldsymbol{\phi} = (\boldsymbol{I} - \tilde{\eta}_{t+1}\boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{\theta}_t - \boldsymbol{\phi}) - \boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_t - \boldsymbol{\phi}) + O(\alpha r^2\eta^3),$$

where

$$\begin{aligned}
\boldsymbol{\psi}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}) &:= \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\left(\boldsymbol{U}_{\boldsymbol{\phi}}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}] + \partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}, \boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}]\right) \\
&\qquad + \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\left(\|\boldsymbol{H}(\boldsymbol{\phi})\hat{\boldsymbol{x}}\|_2^2 + \|\boldsymbol{H}(\boldsymbol{\phi})\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}\|_2^2\right)\boldsymbol{\phi}, \\
\boldsymbol{U}_{\boldsymbol{\phi}} &:= \boldsymbol{I} - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}(\boldsymbol{\phi}).
\end{aligned}$$

Proof. Let $\hat{\boldsymbol{x}}_\tau := \boldsymbol{\theta}_\tau - \boldsymbol{\phi}$ for all $\tau \in \{t, t+1, t+2\}$. Let $\widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\boldsymbol{x})$ be the following function:

$$\widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}) := \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}] + \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})\hat{\boldsymbol{x}}\|_2^2\boldsymbol{\phi}.$$

By Lemma J.4, we have $\hat{\boldsymbol{x}}_{t+1} = O(r\eta)$ and the following first-order and second-order approximations for $\hat{\boldsymbol{x}}_{t+1}$:

$$\begin{aligned}
\hat{\boldsymbol{x}}_{t+1} &= \boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}_t + O(\alpha r\eta^2). \\
\hat{\boldsymbol{x}}_{t+1} &= (\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi}))\hat{\boldsymbol{x}}_t - \widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_t) + O(\alpha r^2\eta^3).
\end{aligned}$$

Note that $\tilde{\eta}_{t+1} = \tilde{\eta}_t + O((1+r^2)\eta^2) = \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})} + O(\alpha\eta)$. By Lemma J.4 again, we have the following second-order approximation for $\hat{\boldsymbol{x}}_{t+2}$:

$$\hat{\boldsymbol{x}}_{t+2} = (\boldsymbol{I} - \tilde{\eta}_{t+1}\boldsymbol{H}(\boldsymbol{\phi}))\hat{\boldsymbol{x}}_{t+1} - \widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_{t+1}) + O(\alpha r^2\eta^3).$$

Now we combine the two steps together.

$$\begin{aligned}
\hat{\boldsymbol{x}}_{t+2} &= (\boldsymbol{I} - \tilde{\eta}_{t+1}\boldsymbol{H}(\boldsymbol{\phi}))\left((\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi}))\hat{\boldsymbol{x}}_t - \widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_t) + O(\alpha r^2\eta^3)\right) \\
&\quad - \widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}_t + O(\alpha r\eta^2)) + O(\alpha r^2\eta^3) \\
&= (\boldsymbol{I} - \tilde{\eta}_{t+1}\boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi})\hat{\boldsymbol{x}}_t - (\boldsymbol{U}_{\boldsymbol{\phi}} + O(\alpha\eta))\widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_t) + O(\alpha r^2\eta^3) \\
&\quad - \left(\widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}_t) + O(\alpha r^2\eta^3)\right) + O(\alpha r^2\eta^3) \\
&= (\boldsymbol{I} - \tilde{\eta}_{t+1}\boldsymbol{H}(\boldsymbol{\phi}))(\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}(\boldsymbol{\phi})\hat{\boldsymbol{x}}_t - \left(\boldsymbol{U}_{\boldsymbol{\phi}}\widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_t) + \widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}_t)\right) + O(\alpha r^2\eta^3).
\end{aligned}$$

where the second equality uses $\boldsymbol{I} - \tilde{\eta}_{t+1}\boldsymbol{H}(\boldsymbol{\phi}) = \boldsymbol{U}_{\boldsymbol{\phi}} + O(\alpha\eta)$. Finally, we note that

$$\begin{aligned}
\boldsymbol{U}_{\boldsymbol{\phi}}\widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}) + \widehat{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}) &= \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{U}_{\boldsymbol{\phi}}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\hat{\boldsymbol{x}},\hat{\boldsymbol{x}}] + \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})\hat{\boldsymbol{x}}\|_2^2\boldsymbol{U}_{\boldsymbol{\phi}}\boldsymbol{\phi} \\
&\quad + \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}},\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}] + \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\|\boldsymbol{H}(\boldsymbol{\phi})\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}\|_2^2\boldsymbol{\phi} \\
&= \tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\left(\boldsymbol{U}_{\boldsymbol{\phi}}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\hat{\boldsymbol{x}},\hat{\boldsymbol{x}}] + \partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}},\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}]\right) \\
&\quad + \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\left(\|\boldsymbol{H}(\boldsymbol{\phi})\hat{\boldsymbol{x}}\|_2^2 + \|\boldsymbol{H}(\boldsymbol{\phi})\boldsymbol{U}_{\boldsymbol{\phi}}\hat{\boldsymbol{x}}\|_2^2\right)\boldsymbol{\phi} \\
&= \boldsymbol{\psi}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}),
\end{aligned}$$

where the second equality uses the fact that $\boldsymbol{U}_{\boldsymbol{\phi}} = (\boldsymbol{I} - \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}(\boldsymbol{\phi}))\boldsymbol{\phi} = \boldsymbol{\phi} + \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\nabla\mathcal{L}(\boldsymbol{\phi}) = \boldsymbol{\phi}$ (Lemma D.1). □

Lemma J.6. *Under Assumptions 4.5 and 4.6, for $\boldsymbol{\phi} \in \Gamma$,*

$$\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})) = (2\boldsymbol{I} - \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}_t)\nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}) + 4\boldsymbol{\phi},$$

where $\boldsymbol{\psi}_{\boldsymbol{\phi}}$ is defined as in Lemma J.5. Moreover,

$$\langle\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})\rangle = 0, \tag{26}$$

$$\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})) = 2\nabla_\Gamma\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}). \tag{27}$$

Proof. Let $\boldsymbol{V}_{\boldsymbol{\phi}} := 2\boldsymbol{I} - \frac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\boldsymbol{H}(\boldsymbol{\phi})$. Using a similar argument as in Lemma J.3,

$$\tfrac{1}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}\left(\boldsymbol{U}_{\boldsymbol{\phi}}\partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}),\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})] + \partial^3\mathcal{L}_{\boldsymbol{\phi}}[-\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}),-\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})]\right) = \boldsymbol{V}_{\boldsymbol{\phi}}\nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}).$$

Also notice that

$$\tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2}\left(\|\boldsymbol{H}(\boldsymbol{\phi})\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})\|_2^2 + \|\boldsymbol{H}(\boldsymbol{\phi})\boldsymbol{U}_{\boldsymbol{\phi}}\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})\|_2^2\right)\boldsymbol{\phi} = \tfrac{2}{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})}(\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2 + \lambda_1^{\mathrm{H}}(\boldsymbol{\phi})^2)\boldsymbol{\phi} = 4\boldsymbol{\phi}.$$

Combining these together proves $\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})) = \boldsymbol{V}_{\boldsymbol{\phi}}\nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}) + 4\boldsymbol{\phi}$.

To obtain the last two equations (26), (27), we first note that $\boldsymbol{\phi}$ is a 0-eigenvector of $\boldsymbol{H}(\boldsymbol{\phi})$ since $\boldsymbol{H}(\boldsymbol{\phi})\boldsymbol{\phi} = -\nabla\mathcal{L}(\boldsymbol{\phi}) = \boldsymbol{0}$ by Lemma D.1. So we have $\langle\boldsymbol{\phi}, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})\rangle = 0$ and $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\boldsymbol{\phi} = \boldsymbol{\phi}$.

To prove (26), we note that $\boldsymbol{V}_{\boldsymbol{\phi}}\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}) = \boldsymbol{0}$. Then

$$\langle\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})\rangle = \langle\nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}), \boldsymbol{V}_{\boldsymbol{\phi}}\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})\rangle + 4\langle\boldsymbol{\phi}, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})\rangle = 0.$$

To prove (27), first we note that $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\boldsymbol{V}_{\boldsymbol{\phi}} = 2\boldsymbol{I}$ and $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\boldsymbol{\phi} = \boldsymbol{\phi}$, which implies

$$\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})) = 2\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}) + 4\boldsymbol{\phi}.$$

Then by Lemma I.4, we can decompose $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\nabla \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi})$ as a component parallel to $\boldsymbol{\phi}$ and a component perpendicular to $\boldsymbol{\phi}$:

$$\begin{aligned}\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\nabla \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}) &= (\boldsymbol{I} - \boldsymbol{\phi}\boldsymbol{\phi}^\top)\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\nabla \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}) + \langle \nabla \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}), \boldsymbol{\phi}\rangle \boldsymbol{\phi} \\ &= \nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}) + \frac{\partial}{\partial c} \log \lambda_1^{\mathrm{H}}(c\boldsymbol{\phi})\bigg|_{c=1}.\end{aligned}$$

For the second term, we note that $\lambda_1^{\mathrm{H}}(c\boldsymbol{\phi}) = c^{-2}\lambda_1^{\mathrm{H}}(\boldsymbol{\phi})$ by scale-invariance, and thus we have $\frac{\partial}{\partial c} \log \lambda_1^{\mathrm{H}}(c\boldsymbol{\phi}) = -2/c$ and $\frac{\partial}{\partial c} \log \lambda_1^{\mathrm{H}}(c\boldsymbol{\phi})\big|_{c=1} = -2$. Combining all these together gives

$$\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi})\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi})) = 2\left(\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}) - 2\boldsymbol{\phi}\right) + 4\boldsymbol{\phi} = 2\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}),$$

which proves (27). □

K Lemmas for Quasi-RMSprop Schedulers

Lemma K.1. *Given gradients $\{\boldsymbol{g}_t\}_{t\geq 0}$, let $\tilde{\eta}_0, \tilde{\eta}_1, \tilde{\eta}_2, \dots$ be the effective learning rates produced by a quasi-RMSprop scheduler with base learning rate η and decay rate β, and let $\{\tilde{v}_t\}_{t\geq 0}$ be the corresponding moment estimating sequence. Consider the case of $\eta = o(1)$, $\beta = 1 - C_{\mathrm{b}}\eta^2 + O(\eta^4)$ for some $C_{\mathrm{b}} = \Theta(1)$. For $\mu = \Theta(1)$ and some $t \geq 0$, define $\hat{u}_\tau := \frac{1}{\eta}(\mu^2\tilde{v}_\tau - 1)$ for $\tau \in \{t, t+1\}$. If $\|\boldsymbol{g}_t\|_2 \leq \alpha\eta$ and $|\mu^2\tilde{v}_t - 1| \leq \alpha\eta$ for some $\alpha = (\frac{1}{\eta})^{o(1)}$ at step t, then the following holds*

$$\tilde{\eta}_t = \mu \cdot \left(1 - \tfrac{1}{2}\eta\hat{u}_t\right) + O(\alpha^2\eta^2). \tag{28}$$

$$\hat{u}_{t+1} = \hat{u}_t + C_{\mathrm{b}}\eta(\mu^2\bar{g}_t^2 - 1) + O(\alpha\eta^2). \tag{29}$$

Proof. By definition of RMSprop-like learning rate scheduler,

$$\tilde{\eta}_t = \frac{1}{\sqrt{\tilde{v}_t}} + O(\alpha^2\eta^2) \tag{30}$$

$$\tilde{v}_{t+1} = \beta\tilde{v}_t + (1-\beta)\bar{g}_t^2 + O(\mathrm{poly}(\alpha)\eta^4). \tag{31}$$

By definition of $\hat{u}_t$, we can express $\tilde{v}_t$ as $\tilde{v}_t = \frac{1}{\mu^2}(1 + \eta\hat{u}_t)$. Combining this with (30) proves (28):

$$\begin{aligned}\tilde{\eta}_t = \frac{1}{\sqrt{\frac{1}{\mu^2}(1+\eta\hat{u}_t)}} = \frac{\mu}{\sqrt{1+\eta\hat{u}_t}} &= \mu \cdot \left(1 - \tfrac{1}{2}\eta\hat{u}_t + O(\alpha^2\eta^2)\right) \\ &= \mu \cdot \left(1 - \tfrac{1}{2}\eta\hat{u}_t\right) + O(\alpha^2\eta^2).\end{aligned}$$

By substituting $\frac{1}{\mu^2}(1+\eta\hat{u}_t)$ for $\tilde{v}_t$ in (31) we have

$$\frac{1}{\mu^2}(1+\eta\hat{u}_{t+1}) = \beta \cdot \frac{1}{\mu^2}(1+\eta\hat{u}_t) + (1-\beta)\bar{g}_t^2 + O(\mathrm{poly}(\alpha)\eta^4).$$

Multiplying μ^2 and subtracting 1 on both sides gives

$$\begin{aligned}\eta\hat{u}_{t+1} &= -(1-\beta) + \beta\eta\hat{u}_t + (1-\beta)\mu^2\bar{g}_t^2 + O(\mathrm{poly}(\alpha)\eta^4) \\ &= \eta\hat{u}_t + (1-\beta)\left(\mu^2\bar{g}_t^2 - \eta\hat{u}_t - 1\right) + O(\mathrm{poly}(\alpha)\eta^4).\end{aligned}$$

Now we divide η on both sides. Then we have

$$\begin{aligned}\hat{u}_{t+1} &= \hat{u}_t + \left(C_{\mathrm{b}}\eta + O(\eta^3)\right)\left(\mu^2\bar{g}_t^2 - \eta\hat{u}_t - 1\right) + O(\mathrm{poly}(\alpha)\eta^3) \\ &= \hat{u}_t + C_{\mathrm{b}}\eta\left(\mu^2\bar{g}_t^2 - \eta\hat{u}_t - 1\right) + O(\mathrm{poly}(\alpha)\eta^3) \\ &= \hat{u}_t + C_{\mathrm{b}}\eta(\mu^2\bar{g}_t^2 - 1) + O(\alpha\eta^2),\end{aligned}$$

which proves (29). □

Lemma K.2. *In the setting of Lemma K.1, for $\mu = \Theta(1)$ and some $t \geq 0$, define $\hat{u}_\tau := \frac{1}{\eta}(\mu^2\tilde{v}_\tau - 1)$ for $\tau \in \{t, t+1, t+2\}$. If $\|\boldsymbol{g}_t\|_2 \leq \alpha\eta$, $\|\boldsymbol{g}_{t+1}\|_2 \leq \alpha\eta$, and $|\mu^2\tilde{v}_t - 1| \leq \alpha\eta$ for some $\alpha = (\frac{1}{\eta})^{o(1)}$ at step t, then the following holds*

$$\tilde{\eta}_{t+1} = \mu \cdot \left(1 - \tfrac{1}{2}\eta\hat{u}_t\right) + O(\alpha^2\eta^2). \tag{32}$$

$$\hat{u}_{t+2} = \hat{u}_t + C_{\mathrm{b}}\eta(\mu^2\bar{g}_t^2 + \mu^2\bar{g}_{t+1}^2 - 2) + O(\alpha\eta^2). \tag{33}$$

Proof. Note that (29) implies $\hat{u}_{t+1} = O(\alpha)$. Applying the inequalities (28) and (29) to step $t+1$ proves (32) and (33). □

L Reduction to RMS-drift Process: The Case of Full Space Optimization

In this section, we let $\{(\boldsymbol{\theta}_t, v_t)\}_{t\geq 0}$ be a trajectory of gradient descent with quasi-RMSprop scheduler, and let η, β be the base learning rate and decay rate. We follow the notations and terminologies in Appendix H.

L.1 Good Initialization

Proof for Lemma H.8. Let $\hat{\boldsymbol{x}}_0 := \boldsymbol{\theta}_0 - \boldsymbol{\zeta}_0$ and $r = \|\hat{\boldsymbol{x}}_0\|_2$ for short. By Gaussian concentration and anti-concentration, with probability $1-\delta$, the following holds:

$$\Omega(\delta\sigma_0) \leq r \leq O(\sigma_0\sqrt{\log(1/\delta)}),$$
$$|\langle \hat{\boldsymbol{x}}_0, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\zeta}_0)\rangle| \geq \Omega(\delta r).$$

By Taylor expansion of Φ and Lemma I.1,

$$\begin{aligned}\boldsymbol{\phi}_0 &= \boldsymbol{\zeta}_0 + \partial\Phi_{\boldsymbol{\zeta}_0}[\hat{\boldsymbol{x}}_0] + O(r^2)\\ &= \boldsymbol{\zeta}_0 + \boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\zeta}_0)\hat{\boldsymbol{x}}_0 + O(r^2).\end{aligned}$$

Thus we can approximate $\boldsymbol{x}_0$ by

$$\boldsymbol{x}_0 = \hat{\boldsymbol{x}}_0 + \boldsymbol{\zeta}_0 - \boldsymbol{\phi}_0 = (\boldsymbol{I} - \boldsymbol{P}_0^{\mathrm{H}})\hat{\boldsymbol{x}}_0 + O(r^2).$$

Now we give a lower bound for $|h_0|\eta$:

$$\begin{aligned}|h_0|\eta = |\langle \boldsymbol{x}_0, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_0)\rangle| &\geq |\langle \hat{\boldsymbol{x}}_0, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\zeta}_0)\rangle| - |\langle \hat{\boldsymbol{x}}_0 - \boldsymbol{x}_0, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\zeta}_0)\rangle| - \|\hat{\boldsymbol{x}}_0\|_2 \cdot \|\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\zeta}_0) - \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_0)\|_2\\ &\geq \Omega(\delta r) - O(r^2) - O(r^2)\\ &\geq \Omega(r)\cdot(\Omega(\delta) - O(r)).\end{aligned}$$

Since $r = O(\sigma_0\sqrt{\log(1/\delta)})$, we can choose $\delta := C_0\alpha_0\eta\sqrt{\log(1/\eta)}$ with a large enough C_0 such that the above inequality gives $|h_0|\eta \geq \Omega(\delta r)$.

Now we verify the conditions claimed in the lemma statement. First, we can see from the following that the initial state is $O(\alpha_0\sqrt{\log(1/\delta)})$-bounded:

$$\begin{aligned}\|\boldsymbol{x}_0\|_2 &\leq \|(\boldsymbol{I} - \boldsymbol{P}_0^{\mathrm{H}})\hat{\boldsymbol{x}}_0\|_2 + O(r^2) \leq O(r) \leq O(\alpha_0\eta\sqrt{\log(1/\delta)})\\ |u_0| &= \tfrac{1}{\eta}|\mu_t^2\tilde{v}_0 - 1| \leq O(\alpha_0).\end{aligned}$$

It is also $O(\alpha_0 + \sqrt{\log(1/\delta)})$-deviated since

$$\|\boldsymbol{x}_0\|_2 \geq |h_0|\eta \geq \Omega(\delta r) \geq \Omega(\delta^2\exp(-\alpha_0^2)) \geq \eta\exp\left(-O\left(\alpha_0 + \sqrt{\log(1/\delta)}\right)^2\right).$$

Next, we verify that the initial state is $O(1/\delta)$-misaligned:

$$\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_0)\boldsymbol{x}_0\|_2 \leq \|\boldsymbol{x}_0\|_2 \leq O(r) \leq O(1/\delta)\cdot|h_0|\eta.$$

Finally, $\|\boldsymbol{\phi}_0 - \boldsymbol{\zeta}_0\|_2 \leq O(r) \leq O(\alpha_0\eta\sqrt{\log(1/\delta)})$. □

L.2 Alignment Phase

We define $r_t = \|\boldsymbol{x}_t\|_2/\eta$ and the following notations for this subsection.

$$\hat{\boldsymbol{x}}_{t+1} := \boldsymbol{\theta}_{t+1} - \boldsymbol{\phi}_t \qquad \hat{u}_{t+1} := \tfrac{1}{\eta}(\mu_t^2\tilde{v}_{t+1} - 1)$$
$$\hat{h}_{t+1} := \tfrac{1}{\eta}\langle \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t), \hat{\boldsymbol{x}}_{t+1}\rangle$$

Lemma L.1. *If $\boldsymbol{\theta}_t \in \mathcal{Z}^{\epsilon_0}$ at step t, then $\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)(\boldsymbol{I} - \mu_t\boldsymbol{H}_t)\|_2 \leq 1 - 2\gamma_{\min}$.*

Proof. If $\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t), \ldots, \lambda_D^{\mathrm{H}}(\boldsymbol{\phi}_t)$ are the eigenvalues of $\boldsymbol{H}(\boldsymbol{\phi}_t)$, then $\{1 - {}^{2\lambda_i^{\mathrm{H}}(\boldsymbol{\phi}_t)}/_{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)}\}_{i=1}^D$ are the eigenvalues of $\boldsymbol{I} - \mu_t\boldsymbol{H}_t$. By definition of $\gamma_{\min}$, we have $1 - {}^{2\lambda_i^{\mathrm{H}}(\boldsymbol{\phi}_t)}/_{\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)} \in [-1+2\gamma_{\min}, 1-2\gamma_{\min}]$ as long as $\lambda_i^{\mathrm{H}}(\boldsymbol{\phi}_t) \neq 0, \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$. Therefore we have $\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)(\boldsymbol{I} - \mu_t\boldsymbol{H}_t)\|_2 \leq 1 - 2\gamma_{\min}$. □

Lemma L.2. *In the setting of Lemma H.9,*

$$\hat{\boldsymbol{x}}_{t+1} = (\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}_t)\boldsymbol{x}_t + O(r_t^2\eta^2) \tag{34}$$

$$\hat{h}_{t+1} = -h_t + O(\alpha r_t \eta) \tag{35}$$

$$\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+1}\|_2 = (1 - 1.9\gamma_{\min})\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 + O(r_t^2\eta^2) \tag{36}$$

$$\hat{u}_{t+1} = u_t + O(\alpha^2\eta) \tag{37}$$

Proof for Lemma L.2. By Lemma J.1, we have the following first-order approximation for $\hat{\boldsymbol{x}}_{t+1}$:

$$\hat{\boldsymbol{x}}_{t+1} = (\boldsymbol{I} - \tilde{\eta}_t \boldsymbol{H}(\boldsymbol{\phi}_t))\,\boldsymbol{x}_t + O(r_t^2\eta^2),$$

which proves (34). Then we can prove (35) as follows:

$$\begin{aligned}\hat{h}_{t+1} = \langle \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t), \hat{\boldsymbol{x}}_{t+1}\rangle &= (1 - \tilde{\eta}_t\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t))h_t + O(r_t^2\eta^2)\\ &= (1 - \mu_t\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t))h_t + O(\alpha r_t\eta^2)\\ &= -h_t + O(\alpha r_t\eta^2).\end{aligned}$$

To prove (36) from (34), we note that if η is sufficiently small, then $\tilde{\eta}_t$ is sufficiently close to μ_t. In this case, by Lemma L.1 we have $\|(\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{x}_t\|_2 \le (1 - 1.9\gamma_{\min})\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2$. Combining this with (34) proves (36).

Finally, we prove (37) By Taylor expansion of $\mathcal{L}$ around $\boldsymbol{\phi}_t$, $\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|_2 = O(\|\boldsymbol{x}_t\|_2) = O(r_t\eta)$. By Lemma K.1, we can approximate $\hat{u}_{t+1}$ by

$$\begin{aligned}\hat{u}_{t+1} &= u_t + C_{\mathrm{b}}\eta(\mu_t^2\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)/\eta\|_2^2 - 1) + O(\alpha\eta^2)\\ &= u_t + C_{\mathrm{b}}\eta(\mu_t^2\cdot O(r_t^2) - 1) + O(\alpha\eta^2)\\ &= u_t + O((r_t^2+1)\eta) + O(\alpha\eta^2)\\ &= u_t + O(\alpha^2\eta),\end{aligned}$$

which proves (37). □

Proof for Lemma H.9. Combining (34) with a Taylor expansion of Φ gives

$$\begin{aligned}\boldsymbol{\phi}_{t+1} = \Phi(\boldsymbol{\phi}_t + \hat{\boldsymbol{x}}_{t+1}) &= \boldsymbol{\phi}_t + \partial\Phi_{\boldsymbol{\phi}_t}[\hat{\boldsymbol{x}}_{t+1}] + O(r_t^2\eta^2)\\ &= \boldsymbol{\phi}_t + \partial\Phi_{\boldsymbol{\phi}_t}[(\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{x}_t] + O(r_t^2\eta^2).\end{aligned}$$

By Lemma I.1 and Lemma I.3, $\partial\Phi_{\boldsymbol{\phi}_t}[(\boldsymbol{I} - \mu_t\boldsymbol{H}_t)\boldsymbol{x}_t] = \partial\Phi_{\boldsymbol{\phi}_t}[\boldsymbol{x}_t] = O(r_t^2\eta^2)$. So we have $\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + O(r_t^2\eta^2)$, which proves (17). Then $\boldsymbol{x}_{t+1} = \hat{\boldsymbol{x}}_{t+1} + \boldsymbol{\phi}_t - \boldsymbol{\phi}_{t+1} = \hat{\boldsymbol{x}}_{t+1} + O(r_t^2\eta^2)$, which proves (18). Finally, by $\mathcal{C}^1$-smoothness of $\mu(\boldsymbol{\phi}), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}), \boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi})$ on Γ, (35), (36), (37) imply (19), (20), (21) respectively. □

Proof for Theorem H.10. Let $\delta := C_0\alpha_0\eta\sqrt{\log(1/\eta)}$, where $C_0 = O(1)$ is a sufficiently large constant so that our proof can work. Let $\alpha_{\max} := \alpha_0\sqrt{\log(1/\delta)}$.

Note that the initial state is $O(1/\delta)$-misaligned. So $\|\boldsymbol{x}_0\|_2 = O(h_0\eta/\delta)$. By Lemma H.9, for all $t \le T_0 := \left\lceil \frac{1}{\log(1-1.9\gamma_{\min})}\log\frac{1}{\delta}\right\rceil$, we can prove by induction that the states are $O(\alpha_{\max})$-bounded, and

$$\begin{aligned}\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 &= O\big((1 - 1.9\gamma_{\min})^t\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_0)\boldsymbol{x}_0\|_2\big)\\ \|\boldsymbol{x}_t\|_2 &= O((1 - 1.9\gamma_{\min})^t\|\boldsymbol{x}_0\|_2),\\ \boldsymbol{\phi}_t &= \boldsymbol{\phi}_0 + O(\|\boldsymbol{x}_0\|_2^2)\\ h_t &= (-1)^t h_0 + O(\alpha_{\max}\|\boldsymbol{x}_0\|_2)\end{aligned}$$

At $t = T_0$, $\|\boldsymbol{x}_t\|_2 = O(h_0\eta)$, $\boldsymbol{\phi}_t = \boldsymbol{\phi}_0 + O(\|\boldsymbol{x}_0\|_2^2)$, $|h_t - (-1)^t h_0| \le O(\alpha_{\max}h_0\eta/\delta)$, $\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 = O(h_0\eta)$. When C_0 is sufficiently large, it also holds that $|h_t - (-1)^t h_0| \le |h_0|/4$.

After that, for all $T_0 \le t \le T_1 := \left\lceil \frac{1}{\log(1-1.9\gamma_{\min})} \log \frac{1}{\delta h_0 \eta} \right\rceil = O(\alpha_{\max})$, we can prove by induction that the states are $O(\alpha_{\max})$-bounded and

$$\begin{aligned}
\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 &= O\big((1-1.9\gamma_{\min})^t \|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_0)\boldsymbol{x}_0\|_2 + h_0^2\eta^2\big) \\
\|\boldsymbol{x}_t\|_2 &= O(h_0\eta), \\
\boldsymbol{\phi}_t &= \boldsymbol{\phi}_0 + O(\|\boldsymbol{x}_0\|_2^2 + h_0^2\eta^2 t) \\
h_t &= (-1)^{t-T_0} h_{T_0} + O(\alpha_{\max} h_0 \eta t)
\end{aligned}$$

At $t = T_1$, $\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 = O(h_0^2\eta^2)$, $\|\boldsymbol{x}_t\|_2 = O(h_0\eta)$, $\boldsymbol{\phi}_t = \boldsymbol{\phi}_0 + O(\alpha_{\max}^3\eta^2) = \boldsymbol{\phi}_0 + O(\alpha_{\max}\eta)$, $|h_t - (-1)^{t-T_1}| \le O(\alpha_{\max}^2 h_0 \eta)$. When η is sufficiently small, it also holds that $|h_t - (-1)^{t-T_1} h_{T_1}| \le |h_0|/4$. So $|h_{T_1} - (-1)^t h_0| \le |h_0|/2$. Putting all these together proves the theorem. □

L.3 Drifting Phase

We define $\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{x})$ as in Lemma J.2. We abuse the notation to write $\boldsymbol{\psi}_t(\boldsymbol{x}) = \boldsymbol{\psi}_{\boldsymbol{\phi}_t}(\boldsymbol{x})$, that is,

$$\boldsymbol{\psi}_t(\boldsymbol{x}) = \frac{\mu_t}{2}\left(\boldsymbol{U}_t \partial^3 \mathcal{L}_{\boldsymbol{\phi}_t}[\boldsymbol{x},\boldsymbol{x}] + \partial^3 \mathcal{L}_{\boldsymbol{\phi}_t}[\boldsymbol{U}_t\boldsymbol{x}, \boldsymbol{U}_t\boldsymbol{x}]\right).$$

We also define the following notations for this subsection.

$$\hat{\boldsymbol{x}}_{t+2} := \boldsymbol{\theta}_{t+2} - \boldsymbol{\phi}_t, \qquad \hat{u}_{t+2} := \frac{1}{\eta}(\mu_t^2 \tilde{v}_{t+2} - 1), \qquad \hat{h}_{t+2} := \frac{1}{\eta}\langle \boldsymbol{x}_{t+2}, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\rangle.$$

Lemma L.3. *If the state $(\boldsymbol{\theta}_t, \tilde{v}_t)$ at step t is $\eta^{-o(1)}$-bounded and $O(|h_t|\eta)$-misaligned, then*

$$\boldsymbol{x}_t = h_t\eta \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(h_t^2\eta^2).$$

Proof. Note that we have the decomposition $\boldsymbol{x}_t = h_t\eta\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + \boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t + \boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t$. By definition of $O(|h_t|\eta)$-misaligned state, $\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t = O(h_t^2\eta^2)$. By Lemma H.6, $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t = O(\|\boldsymbol{x}_t\|_2^2)$.

$$\|\boldsymbol{x}_t\|_2 = h_t\eta + O(h_t^2\eta^2) + O(\|\boldsymbol{x}_t\|_2^2).$$

Solving this equation gives $\|\boldsymbol{x}_t\|_2 = O(|h_t|\eta)$. Then $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t = O(h_t^2\eta^2)$, and therefore we have $\boldsymbol{x}_t = h_t\eta\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(h_t^2\eta^2)$. □

Lemma L.4. *In the setting of Lemma H.11,*

$$\hat{\boldsymbol{x}}_{t+2} = (\boldsymbol{I} - \tilde{\eta}_{t+1}\boldsymbol{H}_t)(\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{x}_t - \eta^2 h_t^2 \boldsymbol{\psi}_t(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)) + O(\alpha h_t^2\eta^3), \tag{38}$$

$$\tilde{\eta}_t = \mu_t \cdot (1 - \tfrac{1}{2}\eta u_t) + O(\alpha^2\eta^2), \tag{39}$$

$$\tilde{\eta}_{t+1} = \mu_t \cdot (1 - \tfrac{1}{2}\eta u_t) + O(\alpha^2\eta^2). \tag{40}$$

$$\hat{u}_{t+2} = u_t + 8C_{\mathrm{b}}\eta h_t^2 - 2C_{\mathrm{b}}\eta + O(\alpha(1+h_t^2)\eta^2). \tag{41}$$

Proof. Define $\bar{g}_t = \|\nabla\mathcal{L}(\boldsymbol{\theta}_t)/\eta\|_2$ and $\bar{g}_{t+1} = \|\nabla\mathcal{L}(\boldsymbol{\theta}_{t+1})/\eta\|_2$. Since the state at step t is $O(|h_t|\eta)$-misaligned, $\|\boldsymbol{x}_t\|_2 \le O(|h_t|\eta)$. By Taylor expansion of $\nabla\mathcal{L}$ around $\boldsymbol{\phi}_t$,

$$\nabla\mathcal{L}(\boldsymbol{\theta}_t) = \nabla\mathcal{L}(\boldsymbol{\phi}_t) + \nabla^2\mathcal{L}(\boldsymbol{\phi}_t)\boldsymbol{x}_t + O(h_t^2\eta^2) = \boldsymbol{H}_t\boldsymbol{x}_t + O(h_t^2\eta^2).$$

So $\bar{g}_t = \|\boldsymbol{H}_t\boldsymbol{x}_t/\eta\|_2 + O(h_t^2\eta) = O(h_t)$. Then Lemma K.1 implies (39) and the following approximation for $\hat{u}_{t+1}$:

$$\begin{aligned}
\hat{u}_{t+1} &= u_t + C_{\mathrm{b}}\eta(\mu_t^2\bar{g}_t^2 - 1) + O(\alpha\eta^2) \\
&= u_t + C_{\mathrm{b}}\eta(\mu_t^2\|\boldsymbol{H}_t\boldsymbol{x}_t/\eta\|_2^2 - 1) + O((\alpha + |h_t|^3)\eta^2).
\end{aligned}$$

As (39) verifies $\tilde{\eta}_t = \mu_t + O(\alpha\eta)$, we can use Lemma J.1 to derive the zeroth-order and first-order approximations for $\boldsymbol{\theta}_{t+1}$: $\boldsymbol{\theta}_{t+1} - \boldsymbol{\phi}_t = O(|h_t|\eta)$ and $\boldsymbol{\theta}_{t+1} - \boldsymbol{\phi}_t = \boldsymbol{U}_t\boldsymbol{x}_t + O(\alpha|h_t|\eta^2)$. Then by Taylor expansion of $\nabla\mathcal{L}$ around $\boldsymbol{\phi}_t$ again,

$$\nabla\mathcal{L}(\boldsymbol{\theta}_{t+1}) = \nabla\mathcal{L}(\boldsymbol{\phi}_t) + \nabla^2\mathcal{L}(\boldsymbol{\phi}_t)(\boldsymbol{U}_t\boldsymbol{x}_t + O(\alpha|h_t|\eta^2)) + O(h_t^2\eta^2) = \boldsymbol{H}_t\boldsymbol{U}_t\boldsymbol{x}_t + O(\alpha h_t\eta^2).$$

So $\bar{g}_{t+1} = \|\boldsymbol{H}_t\boldsymbol{U}_t\boldsymbol{x}_t/\eta\|_2 + O(\alpha|h_t|\eta) = O(|h_t|)$. Then Lemma K.2 implies (40). We can further apply Lemma J.2 to obtain the following:

$$\hat{\boldsymbol{x}}_{t+2} = (\boldsymbol{I} - \tilde{\eta}_{t+1}\boldsymbol{H}_t)(\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{x}_t - \boldsymbol{\psi}_t(\boldsymbol{x}_t) + O(\alpha h_t^2\eta^3). \tag{42}$$

Note that Lemma L.3 implies that $\boldsymbol{x}_t = h_t\eta\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(h_t^2\eta^2)$. Then by Lemma J.3

$$\boldsymbol{\psi}_t(\boldsymbol{x}_t) = \boldsymbol{\psi}_t(\eta h_t\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)) + O(h_t^2\eta^2 \cdot |h_t|\eta) = \eta^2 h_t^2\boldsymbol{\psi}_t(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)) + O(|h_t|^3\eta^3).$$

Combining this with (42) gives (38).

Finally, we derive the approximation for $\hat{u}_{t+2}$. By Lemma K.2,

$$\hat{u}_{t+2} = u_t + C_{\mathrm{b}}\eta(\mu_t^2\bar{g}_t^2 + \mu_t^2\bar{g}_{t+1}^2 - 2) + O(\alpha\eta^2). \tag{43}$$

Since $\boldsymbol{x}_t = h_t\eta\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(h_t^2\eta^2)$, for $\bar{g}_t$ we have

$$\begin{aligned}\bar{g}_t = \|\boldsymbol{H}_t(\eta h_t\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(h_t^2\eta^2))/\eta\|_2 + O(h_t^2\eta) &= \|h_t\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2 + O(h_t^2\eta)\\ &= \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)|h_t| + O(h_t^2\eta).\end{aligned}$$

Similarly, for $\bar{g}_{t+1}$ we have

$$\begin{aligned}\bar{g}_{t+1} = \|\boldsymbol{H}_t\boldsymbol{U}_t(\eta h_t\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(h_t^2\eta^2))/\eta\|_2 &= \left\|-h_t\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\right\|_2 + O(h_t^2\eta)\\ &= \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)|h_t| + O(h_t^2\eta).\end{aligned}$$

So both $\bar{g}_t^2$ and $\bar{g}_{t+1}^2$ can be approximated by $\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)^2h_t^2 + O(|h_t|^3\eta)$. Combining this with (43) gives

$$\begin{aligned}\hat{u}_{t+2} &= u_t + C_{\mathrm{b}}\eta\mu_t^2(\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)^2h_t^2 + O(|h_t|^3\eta)) + C_{\mathrm{b}}\eta\mu_t^2(\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)^2h_t^2 + O(|h_t|^3\eta))\\ &\quad - 2C_{\mathrm{b}}\eta + O\big(\alpha(1+h_t^2)\eta^2\big)\\ &= u_t + 8C_{\mathrm{b}}\eta h_t^2 - 2C_{\mathrm{b}}\eta + O(\alpha(1+h_t^2)\eta^2),\end{aligned}$$

which implies (41). □

Lemma L.5. *In the setting of Lemma H.11,*

$$\hat{h}_{t+2} = (1 - 2\eta u_t)h_t + O(\alpha^2|h_t|\eta^2) \tag{44}$$

$$\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+2} = \boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t - 2\eta^2h_t^2\nabla_\Gamma\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(\alpha h_t^2\eta^3) \tag{45}$$

$$\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+2}\|_2 \le (1 - 1.9\gamma)^2\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2 + O(h_t^2\eta^2) \tag{46}$$

Proof. In the following, we derive the approximations from (38),

$$\hat{\boldsymbol{x}}_{t+2} = (\boldsymbol{I} - \tilde{\eta}_{t+1}\boldsymbol{H}_t)(\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{x}_t - \eta^2h_t^2\boldsymbol{\psi}_t(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)) + O(\alpha h_t^2\eta^3).$$

Approximation for $\hat{h}_{t+2}$. Note that $\boldsymbol{x}_t = \eta h_t\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(h_t^2\eta^2)$ since the state at step t is $O(|h_t|\eta)$-misaligned. For $\hat{h}_{t+2}$, we have

$$\begin{aligned}\hat{h}_{t+2} &= \tfrac{1}{\eta}\left(\langle\boldsymbol{x}_t, (\boldsymbol{I} - \tilde{\eta}_{t+1}\boldsymbol{H}_t)(\boldsymbol{I} - \tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\rangle - \eta^2h_t^2\langle\boldsymbol{\psi}_t(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\rangle + O(\alpha h_t^2\eta^3)\right)\\ &= (1 - \tilde{\eta}_{t+1}\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t))(1 - \tilde{\eta}_t\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t))h_t + O(\alpha h_t^2\eta^2),\end{aligned}$$

where we use the fact that $\langle\boldsymbol{\psi}_t(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\rangle = 0$ by Lemma J.3.

By (39), (40), $\tilde{\eta}_\tau = \mu_t \cdot (1 - \frac{1}{2}\eta u_t) + O(\alpha^2\eta^2)$ for $\tau \in \{t, t+1\}$. Note that $\mu_t \cdot \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) = 2$. Then for $\tau \in \{t, t+1\}$,

$$\begin{aligned}1 - \tilde{\eta}_\tau\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) &= 1 - \mu_t \cdot (1 - \tfrac{1}{2}\eta u_t) \cdot \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(\alpha^2\eta^2)\\ &= 1 - 2 \cdot (1 - \tfrac{1}{2}\eta u_t)) + O(\alpha^2\eta^2)\\ &= -1 + \eta u_t + O(\alpha^2\eta^2).\end{aligned}$$

Then $(1 - \tilde{\eta}_{t+1}\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t))(1 - \tilde{\eta}_t\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t))$ can be approximated by

$$(1 - \tilde{\eta}_{t+1}\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t))(1 - \tilde{\eta}_t\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)) = (-1 + \eta u_t + O(\alpha^2\eta^2))^2 = 1 - 2\eta u_t + O(\alpha^2\eta^2).$$

Therefore, we have $\hat{h}_{t+2} = \big(1 - 2\eta u_t + O(\alpha^2\eta^2)\big)\, h_t + O(\alpha h_t^2\eta^2)$, which implies (44).

Approximation for $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+2}$. For $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+2}$, we have

$$\begin{aligned}\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+2} &= \boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)(\boldsymbol{I}-\tilde{\eta}_{t+1}\boldsymbol{H}_t)(\boldsymbol{I}-\tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{x}_t - \eta^2 h_t^2 \boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{\psi}_t(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)) + O(\alpha h_t^2\eta^3)\\ &= \boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)(\boldsymbol{I}-\tilde{\eta}_{t+1}\boldsymbol{H}_t)(\boldsymbol{I}-\tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{x}_t - 2\eta^2 h_t^2\nabla_\Gamma \log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(\alpha h_t^2\eta^3),\end{aligned}$$

where we use the fact that $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{\psi}_t(\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)) = 2\nabla_\Gamma \log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$ by Lemma J.3. To obtain (45), we only need to note that $\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)(\boldsymbol{I}-\tilde{\eta}_{t+1}\boldsymbol{H}_t)(\boldsymbol{I}-\tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{x}_t = \boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t$.

Approximation for $\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+2}\|_2$. To approximate $\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+2}\|_2$, we note that if η is sufficiently small, then $\tilde{\eta}_t, \tilde{\eta}_{t+1}$ are sufficiently close to μ_t. In this case, by Lemma L.1 we have $\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)(\boldsymbol{I}-\tilde{\eta}_{t+1}\boldsymbol{H}_t)(\boldsymbol{I}-\tilde{\eta}_t\boldsymbol{H}_t)\boldsymbol{x}_t\|_2 \le (1-1.9\gamma_{\min})^2\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\boldsymbol{x}_t\|_2$. Combining this with (38) proves (46). □

Lemma L.6. *In the setting of Lemma H.11, the approximation (25) holds for $\boldsymbol{\phi}_{t+2}$ and*

$$\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_{t+2}) = \log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) - 2\eta^2 h_t^2\|\nabla_\Gamma \log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2^2 + O(\alpha h_t^2\eta^3). \tag{47}$$

Proof. By Taylor expansion,

$$\boldsymbol{\phi}_t = \Phi(\boldsymbol{\phi}_t) + \partial\Phi_{\boldsymbol{\phi}_t}[\boldsymbol{x}_t] + \frac{1}{2}\partial^2\Phi_{\boldsymbol{\phi}_t}[\boldsymbol{x}_t,\boldsymbol{x}_t] + O(|h_t|^3\eta^3). \tag{48}$$

$$\boldsymbol{\phi}_{t+2} = \Phi(\boldsymbol{\phi}_t) + \partial\Phi_{\boldsymbol{\phi}_t}[\hat{\boldsymbol{x}}_{t+2}] + \frac{1}{2}\partial^2\Phi_{\boldsymbol{\phi}_t}[\hat{\boldsymbol{x}}_{t+2},\hat{\boldsymbol{x}}_{t+2}] + O(|h_t|^3\eta^3). \tag{49}$$

By definition, $\Phi(\boldsymbol{\phi}_t) = \boldsymbol{\phi}_t$. By (45) and Lemma I.1,

$$\partial\Phi_{\boldsymbol{\phi}_t}[\hat{\boldsymbol{x}}_{t+2}] = \partial\Phi_{\boldsymbol{\phi}_t}[\boldsymbol{x}_t] - 2\eta^2 h_t^2\nabla_\Gamma \log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(\alpha h_t^2\eta^3).$$

Note that $\boldsymbol{x}_t = \eta h_t \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(h_t^2\eta^2)$ since the state at step t is $O(|h_t|\eta)$-misaligned. Also note that the identity $\partial^2\Phi_{\boldsymbol{\phi}_t}[\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t),\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)] = \boldsymbol{0}$ holds by Lemma I.2. Then

$$\begin{aligned}\partial^2\Phi_{\boldsymbol{\phi}_t}[\boldsymbol{x}_t,\boldsymbol{x}_t] &= \partial^2\Phi_{\boldsymbol{\phi}_t}[\eta h_t\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t), \eta h_t\boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t)] + O(|h_t|^3\eta^3)\\ &= O(|h_t|^3\eta^3).\end{aligned}$$

Similarly, we have $\partial^2\Phi_{\boldsymbol{\phi}_t}[\hat{\boldsymbol{x}}_{t+2},\hat{\boldsymbol{x}}_{t+2}] = O(|h_t|^3\eta^3)$ since (45) implies $\|\boldsymbol{P}_0^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+2}\|_2 = O(h_t^2\eta^2)$ and (46) implies $\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+2}\|_2 = O(h_t^2\eta^2)$.

Now we can prove (25) by subtracting (49) with (48):

$$\begin{aligned}\boldsymbol{\phi}_{t+2} - \boldsymbol{\phi}_t &= (\partial\Phi_{\boldsymbol{\phi}_t}[\hat{\boldsymbol{x}}_{t+2}] - \partial\Phi_{\boldsymbol{\phi}_t}[\boldsymbol{x}_t]) + \tfrac{1}{2}\left(\partial^2\Phi_{\boldsymbol{\phi}_t}[\hat{\boldsymbol{x}}_{t+2},\hat{\boldsymbol{x}}_{t+2}] - \partial^2\Phi_{\boldsymbol{\phi}_t}[\boldsymbol{x}_t,\boldsymbol{x}_t]\right) + O(|h_t|^3\eta^3)\\ &= \left(-2\eta^2 h_t^2\nabla_\Gamma \log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(\alpha h_t^2\eta^3)\right) + O(|h_t|^3\eta^3) + O(|h_t|^3\eta^3)\\ &= -2\eta^2 h_t^2\nabla_\Gamma \log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(\alpha h_t^2\eta^3).\end{aligned}$$

Finally, for $\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_{t+2})$ we have

$$\begin{aligned}\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_{t+2}) - \log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) &= \langle\nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t), \boldsymbol{\phi}_{t+2}-\boldsymbol{\phi}_t\rangle + O((h_t^2\eta^2)^2)\\ &= \langle\nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t), -2\eta^2 h_t^2\nabla_\Gamma\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\rangle + O(\alpha h_t^2\eta^3 + h_t^4\eta^4)\\ &= -2\eta^2 h_t^2\|\nabla_\Gamma\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2^2 + O(\alpha h_t^2\eta^3),\end{aligned}$$

which proves (47). □

Proof for Lemma H.11. We have verified (25) in Lemma L.6. By (25) and definitions of $\hat{\boldsymbol{x}}_{t+2}$ and $\boldsymbol{x}_{t+2}$, we have

$$\boldsymbol{x}_{t+2} - \hat{\boldsymbol{x}}_{t+2} = \boldsymbol{\phi}_t - \boldsymbol{\phi}_{t+2} = 2\eta^2 h_t^2\nabla_\Gamma\log\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) + O(\alpha h_t^2\eta^3).$$

And we can write $\boldsymbol{x}_{t+2} - \hat{\boldsymbol{x}}_{t+2} = \boldsymbol{\phi}_t - \boldsymbol{\phi}_{t+2} = O(h_t^2\eta^2)$ as a loose approximation.

Approximation for h_{t+2}. For h_{t+2}, we have

$$\begin{aligned}
h_{t+2} - \hat{h}_{t+2} &= \langle \boldsymbol{x}_{t+2}, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_{t+2}) \rangle - \langle \hat{\boldsymbol{x}}_{t+2}, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) \rangle \\
&= \langle \boldsymbol{x}_{t+2}, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_{t+2}) - \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) \rangle + \langle \boldsymbol{x}_{t+2} - \hat{\boldsymbol{x}}_{t+2}, \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) \rangle \\
&= O(|h_t|\eta) \cdot O(\|\boldsymbol{\phi}_{t+2} - \boldsymbol{\phi}_t\|_2) + 2\eta^2 h_t^2 \langle \nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) \rangle + O(\alpha h_t^2 \eta^3) \\
&= O(|h_t|^3 \eta^3) + 0 + O(\alpha h_t^2 \eta^3) \\
&= O(\alpha h_t^2 \eta^3),
\end{aligned}$$

where the fourth equality is due to $\langle \nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t), \boldsymbol{v}_1^{\mathrm{H}}(\boldsymbol{\phi}_t) \rangle = 0$ and $\|\boldsymbol{\phi}_{t+2} - \boldsymbol{\phi}_t\|_2 = O(h_t^2 \eta^2)$. Combining this with (44) proves the claimed approximation (22).

Approximation for $\|\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_{t+2})\boldsymbol{x}_{t+2}\|_2$. For $\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_{t+2})\boldsymbol{x}_{t+2}$, we have

$$\begin{aligned}
\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_{t+2})\boldsymbol{x}_{t+2} - \boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)\hat{\boldsymbol{x}}_{t+2} &= (\boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_{t+2}) - \boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t))\boldsymbol{x}_{t+2} + \boldsymbol{P}_{\mathtt{NZT}}^{\mathrm{H}}(\boldsymbol{\phi}_t)(\boldsymbol{x}_{t+2} - \hat{\boldsymbol{x}}_{t+2}) \\
&= O(\|\boldsymbol{\phi}_{t+2} - \boldsymbol{\phi}_t\|_2) \cdot O(|h_t|\eta) + O(\|\boldsymbol{x}_{t+2} - \hat{\boldsymbol{x}}_{t+2}\|_2) \\
&= O(h_t^2 \eta^2) \cdot O(|h_t|\eta) + O(h_t^2 \eta^2) \\
&= O(h_t^2 \eta^2),
\end{aligned}$$

where the third equality is due to $\boldsymbol{x}_{t+2} - \hat{\boldsymbol{x}}_{t+2} = \boldsymbol{\phi}_t - \boldsymbol{\phi}_{t+2} = O(h_t^2 \eta^2)$. Combining this with (46) proves the claimed approximation (23).

Approximation for u_{t+2}. Now we prove the formula for u_{t+2}. Note that $\boldsymbol{\phi}_{t+2} - \boldsymbol{\phi}_t = O(h_t^2 \eta^2)$ implies that $\mu_{t+2} - \mu_t = O(h_t^2 \eta^2)$. Then we have

$$\begin{aligned}
\eta\,(u_{t+2} - \hat{u}_{t+2}) &= (\mu_{t+2}^2 - \mu_t^2)\tilde{v}_{t+2} \\
&= (\mu_{t+2} - \mu_t)(2\mu_t + O(h_t^2 \eta^2)) \cdot (\tfrac{1}{\mu_t^2} + O(\alpha\eta)) \\
&= (\mu_{t+2} - \mu_t)\left(2\mu_t \cdot \tfrac{1}{\mu_t^2} + O(\alpha\eta)\right) \\
&= \tfrac{2}{\mu_t}(\mu_{t+2} - \mu_t) + O(\alpha h_t^2 \eta^3).
\end{aligned}$$

Note that $\log \mu_{t+2} - \log \mu_t = \log\left(1 + \frac{1}{\mu_t}(\mu_{t+2} - \mu_t)\right) = \frac{1}{\mu_t}(\mu_{t+2} - \mu_t) + O((h_t^2 \eta^2)^2)$. By (47),

$$\log \mu_{t+2} - \log \mu_t = \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t) - \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_{t+2}) = 2\eta^2 h_t^2 \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2^2 + O(\alpha h_t^2 \eta^3).$$

Combining these together gives the following approximation for $\frac{1}{\mu_t}(\mu_{t+2} - \mu_t)$:

$$\begin{aligned}
\tfrac{1}{\mu_t}(\mu_{t+2} - \mu_t) &= \log \mu_{t+2} - \log \mu_t + O(h_t^4 \eta^4) \\
&= 2\eta^2 h_t^2 \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2^2 + O(\alpha h_t^2 \eta^3).
\end{aligned}$$

So $\eta\,(u_{t+2} - \hat{u}_{t+2}) = 4\eta^2 h_t^2 \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2^2 + O(\alpha h_t^2 \eta^3)$. Then by (41), we have

$$\begin{aligned}
u_{t+2} &= \hat{u}_{t+2} + 4\eta h_t^2 \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2^2 + O(\alpha h_t^2 \eta^2) \\
&= u_t + 8\eta C_{\mathrm{b}} h_t^2 + 4\eta h_t^2 \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2^2 - 2\eta C_{\mathrm{b}} + O\big(\alpha(1 + h_t^2)\eta^2\big) \\
&= u_t + 4\eta h_t^2 (2C_{\mathrm{b}} + \|\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)\|_2^2) - 2\eta C_{\mathrm{b}} + O(\alpha(1 + h_t^2)\eta^2),
\end{aligned}$$

which proves the claimed approximation (24). □

M Reduction to RMS-drift Process: The Case of Spherical Optimization

In this section, we let $\{(\boldsymbol{\theta}_t, v_t)\}_{t \geq 0}$ be a trajectory of projected gradient descent with quasi-RMSprop scheduler, and let η, β be the base learning rate and decay rate. We follow the notations and terminologies in Appendix H.

As the analysis in the spherical case is nearly the same as the full space case, we only discuss the difference here.

M.1 Good Initialization

Proof for Lemma H.17. Using a similar argument as in Lemma H.8, we know that the lemma holds if there is no projection in the random initialization. But the projection only leads to an error of order $O(r^2)$, so the lemma holds. □

M.2 Alignment Phase

Proofs for Lemma H.18 and Theorem H.19. The proof is essentially the same as Lemma H.9 and Theorem H.10. To see this, we only need to note that we only have used a linear approximation of the update rule with error $O(r_t^2)$, and the linear approximation remains unchanged if we add a projection operator (Lemma J.4). □

M.3 Drifting Phase

We define $\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{x})$ as in Lemma J.5. We abuse the notation to write $\boldsymbol{\psi}_t(\boldsymbol{x}) = \boldsymbol{\psi}_{\boldsymbol{\phi}_t}(\boldsymbol{x})$, that is,

$$\boldsymbol{\psi}_t(\boldsymbol{x}) := \frac{\mu_t}{2}\left(\boldsymbol{U}_t\partial^3\mathcal{L}_{\boldsymbol{\phi}_t}[\boldsymbol{x},\boldsymbol{x}] + \partial^3\mathcal{L}_{\boldsymbol{\phi}}[\boldsymbol{U}_t\boldsymbol{x},\boldsymbol{U}_t\boldsymbol{x}]\right) + \frac{\mu_t^2}{2}\left(\|\boldsymbol{H}_t\boldsymbol{x}\|_2^2 + \|\boldsymbol{H}_t\boldsymbol{U}_t\boldsymbol{x}\|_2^2\right)\boldsymbol{\phi}_t. \tag{50}$$

Lemma M.1. *In the setting of Lemma H.20, the same statement as Lemmas L.4 and L.5 holds, where $\boldsymbol{\psi}_t$ is interpreted as (50).*

Proof. We can follow the argument in the proof for Lemmas L.4 and L.5, but now we are using Lemmas J.5 and J.6 to establish the proof. □

Lemma M.2. *In the setting of Lemma H.20, the same statement as Lemma L.6 holds.*

Proof. The argument is the same as Lemma L.6, but we apply Lemmas I.4 and I.5 in doing Taylor expansion for $\Phi(\boldsymbol{\theta})$. □

Proof for Lemma H.20. Same as Lemma H.11 but we invoke the spherical version of lemmas. □

N Analysis of RMS-drift Process

In this section, we provide proofs for theorems in Appendix H.5. For convenience, we define $R(\boldsymbol{\theta}) = \nabla\log\lambda_1^{\mathrm{H}}(\boldsymbol{\theta})$, $K_t := \sqrt{2C_{\mathrm{b}} + \|\nabla_\Gamma R(\boldsymbol{\phi}_t)\|_2^2}$. Then a C_0-RMS-drift transition $S_t \to S_{t+2}$ can be written as

$$\begin{aligned}
h'_{t+2} &:= (1-2\eta u_t)h_t, & |h_{t+2} - h'_{t+2}| &\le C_0\alpha^2|h_t|\eta^2,\\
u'_{t+2} &:= u_t + 4\eta K_t^2 h_t^2 - 2\eta C_{\mathrm{b}}, & |u_{t+2} - u'_{t+2}| &\le C_0\alpha(1+h_t^2)\eta^2,\\
\boldsymbol{\phi}'_{t+2} &:= \boldsymbol{\phi}_t - 2\eta^2 h_t^2\nabla_\Gamma R(\boldsymbol{\phi}_t), & \|\boldsymbol{\phi}_{t+2} - \boldsymbol{\phi}'_{t+2}\|_2 &\le C_0\alpha h_t^2\eta^3.
\end{aligned}$$

N.1 Conservation of Energy

To establish the conservation of energy, we first compute the change in energy after one transition.

Lemma N.1. *Given two drift states $S_0 = (h_0, u_0, \boldsymbol{\phi}_0)$ and $S_2 = (h_2, u_2, \boldsymbol{\phi}_2)$ in the working zone, for learning rate η and hyperparameter $C_{\mathrm{b}} > 0$, if S_0 is α-bounded for some $1 \le \alpha \le \eta^{-o(1)}$, and $S_0 \to S_2$ is a C_0-RMSdrift transition, then*

$$E(S_2) - E(S_0) = O\big(\alpha^2(1+h_0^2)\eta^2\big) = \begin{cases} O(\alpha^2\eta^2) & |h_0| \le 2,\\ O(\alpha^2 h_0^2\eta^2) & |h_0| > 2.\end{cases}$$

Proof for Lemma N.1. h_0 and h_2 have the same sign when η is small enough. We can decompose $E(S_2) - E(S_0)$ as follows:

$$E(S_2) - E(S_0) = \underbrace{\frac{1}{2}(u_2^2 - u_0^2)}_{=:\delta_1} + \underbrace{K_0^2(h_2^2 - h_0^2)}_{=:\delta_2} + \underbrace{(\|\nabla_\Gamma R(\boldsymbol{\phi}_2)\|_2^2 - \|\nabla_\Gamma R(\boldsymbol{\phi}_0)\|_2^2)h_2^2}_{=:\delta_3} + \underbrace{C_{\mathrm{b}}\log\frac{h_2}{h_0}}_{=:\delta_4}.$$

Now we bound each error term. For δ_1 and δ_2, we use the formula $a^2 - b^2 = 2b(a-b) + (a-b)^2$:

$$\begin{aligned}
\delta_1 &= u_0(u_2 - u_0) + \tfrac{1}{2}(u_2 - u_0)^2 \\
&= u_0\left(4\eta h_0^2 K_0^2 - 2\eta C_{\mathrm{b}} + O(\alpha(1+h_0^2)\eta^2)\right) + O\left((h_0^2\eta)^2\right). \\
&= 4\eta u_0 h_0^2 K_0^2 - 2\eta u_0 C_{\mathrm{b}} + O(\alpha^2(1+h_0^2)\eta^2). \\
\delta_2 &= K_0^2\left(2h_0(h_2 - h_0) + (h_2 - h_0)^2\right) \\
&= K_0^2\left(\left(-4\eta u_0 h_0^2 + O(\alpha^2 h_0^2\eta^2)\right) + O\left((\alpha h_0\eta)^2\right)\right) \\
&= -4\eta u_0 h_0^2 K_0^2 + O(\alpha^2 h_0^2 \eta^2).
\end{aligned}$$

For δ_3, we use the Lipschitzness of $\nabla_\Gamma R(\boldsymbol{\theta})$:

$$\delta_3 = O(\|\boldsymbol{\phi}_2 - \boldsymbol{\phi}_0\|_2) \cdot h_2^2 = O(h_0^2\eta^2) \cdot O(h_0^2) = O(h_0^4\eta^2).$$

For δ_4, note that $\log(1+z) \le z + O(z^2)$ when $z = o(1)$. Then

$$\begin{aligned}
\delta_4 = -C_{\mathrm{b}} \log\left(1 + \frac{h_2 - h_0}{h_0}\right) &= -C_{\mathrm{b}} \cdot \frac{h_2 - h_0}{h_0} + O(\alpha^2\eta^2) \\
&= 2\eta u_0 C_{\mathrm{b}} + O(\alpha^2\eta^2).
\end{aligned}$$

Adding $\delta_1, \delta_2, \delta_3, \delta_4$ together gives

$$\begin{aligned}
E(S_2) - E(S_0) = &+ 4\eta u_0 h_0^2 K_0^2 && - 2\eta u_0 C_{\mathrm{b}} && + O(\alpha^2(1+h_0^2)\eta^2) \\
&- 4\eta u_0 h_0^2 K_0^2 && && + O(\alpha^2 h_0^2\eta^2) \\
& && && + O(h_0^4\eta^2) \\
& && + 2\eta C_{\mathrm{b}} u_0 && + O(\alpha^2\eta^2).
\end{aligned}$$

So $E(S_2) - E(S_0) = O(\alpha^2(1+h_0^2)\eta^2)$. □

To sum up the energy change over time, we need the following lemma.

Lemma N.2. *For any $M = o((\alpha\eta)^{-2})$, if $S_0, \ldots, S_{2M}$ is an $O(1)$-RMS-drift process in the working zone, and S_t is α-bounded for all even numbers $t \le 2M$, then*

$$\sum_{m=0}^{M-1} h_{2m}^2 = \frac{C_{\mathrm{b}}}{2K_0^2} M + O(\alpha^2\eta^2 M^2 + \alpha/\eta).$$

Proof for Lemma N.2. By the update rule of u_t, we have

$$u_{2M} - u_0 = \sum_{m=0}^{M-1} u_{2m+2} - u_{2m} = \sum_{m=0}^{M-1} \left(4\eta K_{2m}^2 h_{2m}^2 - 2\eta C_{\mathrm{b}} + O(\alpha^3\eta^2)\right).$$

Since S_{2m} is α-bounded for all $m \le M$, $\boldsymbol{\phi}_{2m} - \boldsymbol{\phi}_0 = O(\alpha^2\eta^2 M)$. By smoothness of R, $\|\nabla_\Gamma R(\boldsymbol{\phi}_{2m})\|_2^2 = \|\nabla_\Gamma R(\boldsymbol{\phi}_0)\|_2^2 + O(\alpha^2\eta^2 M)$, then $K_{2m} = K_0 + O(\alpha^2\eta^2 M)$. Since S_0 and S_{2M} are α-bounded, $u_T - u_0 = O(\alpha)$. Combining all these together,

$$O(\alpha) = \sum_{m=0}^{M-1} \left(4\eta\left(K_0^2 + O(\alpha^2\eta^2 M)\right) h_{2m}^2 - 2\eta C_{\mathrm{b}} + O(\alpha^3\eta^2)\right)$$

Let $Q := \sum_{m=0}^{M-1} h_{2m}^2$. Then we have

$$O(\alpha) = 4\eta\left(K_0^2 + O(\alpha^2\eta^2 M)\right) Q - 2\eta C_{\mathrm{b}} M + O(\alpha^3\eta^2 M).$$

Rearranging the terms while noting that $\alpha^3\eta^2 M = o(\alpha)$, we have

$$4\eta\left(K_0^2 + O(\alpha^2\eta^2 M)\right) Q = 2\eta C_{\mathrm{b}} M + O(\alpha).$$

So we can estimate Q by

$$Q = \frac{2C_{\mathrm{b}} M + O(\alpha/\eta)}{4K_0^2 + O(\alpha^2\eta^2 M)} = \frac{C_{\mathrm{b}}}{2K_0^2} M + O(\alpha^2\eta^2 M^2 + \alpha/\eta),$$

which completes the proof. □

Lemma N.3. *For an $O(1)$-RMSdrift process $S_0, \ldots, S_{2M}$ in the working zone, if $E(S_0) \le \alpha^2$ for some parameter $1 \le \alpha \le \eta^{-o(1)}$ and $M = O(1/\eta^{1.5})$, then $E(S_t) = E(S_0) + O(\alpha^2\eta^{0.5})$ for all even numbers $0 \le t \le 2M$.*

Proof for Lemma N.3. We do a bootstrap. As $E(S_0) \le O(\alpha^2)$, we can leverage Lemma N.1 to prove by induction that $E(S_t) \le E(S_0) + O(\alpha^4\eta^2 M)$ for all $0 \le t \le 2M$. Then we apply Lemma N.1 again. For all $N \le M$,

$$\begin{aligned} E(S_{2N}) - E(S_0) &= \sum_{m=0}^{N-1} O(\alpha^2(1+h_{2m}^2)\eta^2) \\ &= O(\alpha^2\eta^2)\left(N + \sum_{m=0}^{N-1} h_{2m}^2\right) \\ &\le O(\alpha^2\eta^2)\left(N + O(N + \alpha^2\eta^2N^2 + \alpha/\eta)\right) \\ &\le O(\alpha^2\eta^{0.5}). \end{aligned}$$

where the third line uses Lemma N.2. □

Proof for Theorem H.22. We group the M transitions into $O(1/\eta^{0.5})$ segments of length $O(1/\eta^{1.5})$. We can do an induction with Lemma N.3 applied on each segment to show that for all t in the k-th segment, $E(S_t) \le (1 + O(\eta^{0.5}))^k E(S_0)$. Noting that $k = O(1/\eta^{0.5})$ and $(1 + O(\eta^{0.5}))^{O(1/\eta^{0.5})} = O(1)$ finishes the proof. □

N.2 Flow Approximation

Lemma N.4. *In the setting of Theorem H.23 but with $M = \Theta(1/\eta^{1.5})$, if S_t is α-bounded for all even numbers $0 \le t \le 2M$, then*

$$\|\boldsymbol{\phi}_t - \boldsymbol{\zeta}(t\eta^2)\|_2 \le O(\alpha^2\eta).$$

Proof for Lemma N.4. When S_t is α-bounded, $\|\boldsymbol{\phi}_{t+2} - \boldsymbol{\phi}_t\|_2 = O(\alpha^2\eta^2)$, so $\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_0\|_2 = O(\alpha^2\eta^2 M) = O(\alpha^2\eta^{0.5})$. For every $N \le M$ we have

$$\begin{aligned} \boldsymbol{\phi}_{2N} - \boldsymbol{\phi}_0 = \sum_{m=0}^{N-1} \boldsymbol{\phi}_{2m+2} - \boldsymbol{\phi}_{2m} &= \sum_{m=0}^{N-1} \left(-2\eta^2 h_{2m}^2 \nabla_\Gamma R(\boldsymbol{\phi}_{2m}) + O(\alpha^3\eta^3)\right) \\ &= \sum_{m=0}^{N-1} \left(-2\eta^2 h_{2m}^2 (\nabla_\Gamma R(\boldsymbol{\phi}_0) + O(\alpha^2\eta^{0.5})) + O(\alpha^3\eta^3)\right) \\ &= \sum_{m=0}^{N-1} \left(-2\eta^2 h_{2m}^2 \nabla_\Gamma R(\boldsymbol{\phi}_0) + O(\alpha^2\eta^{2.5})\right) \\ &= -\underbrace{2\eta^2\left(\sum_{m=0}^{N-1} h_{2m}^2\right)}_{=:\delta} \nabla_\Gamma R(\boldsymbol{\phi}_0) + O(\alpha^2\eta). \end{aligned}$$

By Lemma N.2, we have

$$\begin{aligned} \delta = 2\eta^2 \cdot \left(\frac{C_{\mathrm{b}}}{2K_0^2} N + O(\alpha^2\eta^2N^2 + \alpha/\eta)\right) &= \frac{C_{\mathrm{b}}}{K_0^2}\eta^2 N + 2\eta^2 \cdot O(\alpha^2/\eta) \\ &= \frac{C_{\mathrm{b}}}{K_0^2}\eta^2 N + O(\alpha^2\eta). \end{aligned}$$

Note that $\frac{C_{\mathrm{b}}^2}{2K_0^2}\nabla_\Gamma R(\boldsymbol{\phi}_0) = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\zeta}(0) + O(\alpha^2\eta^{1/2})$. Then we have

$$\begin{aligned}
\boldsymbol{\phi}_{2N} &= \boldsymbol{\phi}_0 - \delta\nabla_\Gamma R(\boldsymbol{\phi}_0) + O(\alpha^2\eta) \\
&= \boldsymbol{\phi}_0 - (2\eta^2 N)\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\zeta}(\boldsymbol{\phi}_0) + O(\alpha^2\eta^{1/2})\right) + O(\alpha^2\eta) \\
&= \boldsymbol{\zeta}(2N\eta^2) + O\big((\eta^2 N)^2\big) + O(\alpha^2\eta) \\
&= \boldsymbol{\zeta}(2N\eta^2) + O(\alpha^2\eta),
\end{aligned}$$

where the third equality uses the smoothness of $\boldsymbol{\zeta}(t)$. □

Proof for Theorem H.23. We group the M transitions into $O(1/\eta^{1/2})$ segments of length $O(1/\eta^{1.5})$. Then we leverage Lemma N.4 and do an induction to show that $\|\boldsymbol{\phi}_t - \boldsymbol{\zeta}(t\eta^2)\|_2 \le O(\alpha^2\eta^{1/2})$ for all even numbers $0 \le t \le 2M$. □

O Proofs for Linear Regression with Batch Normalization

Lemma O.1. *Assume that the regression targets y_i are generated by a linear model. For linear regression with BN, the global minimizer manifold of $\mathcal{L}(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n}(\Phi(\boldsymbol{w}_i;\boldsymbol{w},\sigma_{\mathrm{y}},\mu_{\mathrm{y}}) - y_i)^2$ on the unit sphere is*

$$\Gamma := \left\{ \boldsymbol{w} \in \mathbb{S}^{d-1} : \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}}, \boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{x}} \right\rangle = \frac{y_i - \mu_{\mathrm{y}}}{\sigma_{\mathrm{y}}} \right\}.$$

For any global minimizer $\boldsymbol{w} \in \Gamma$, the Hessian matrix $\boldsymbol{H}(\boldsymbol{w})$ of the loss is given by

$$\boldsymbol{H}(\boldsymbol{w}) = 2\|\tilde{\boldsymbol{w}}\|_2^2 \left(\boldsymbol{\Sigma}_{\mathrm{x}} - \boldsymbol{z}\boldsymbol{z}^\top\right),$$

where $\tilde{\boldsymbol{w}} := \frac{\sigma_{\mathrm{y}}\boldsymbol{w}}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}}$ as defined in (4), and $\boldsymbol{z} := \frac{1}{n}\sum_{i=1}^{n}\frac{y_i-\mu_{\mathrm{y}}}{\sigma_{\mathrm{y}}}(\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{x}})$.

Proof. The model output $\Phi(\boldsymbol{x};\boldsymbol{w},\mu_{\mathrm{y}},\sigma_{\mathrm{y}})$ can be written as

$$\Phi(\boldsymbol{x};\boldsymbol{w},\mu_{\mathrm{y}},\sigma_{\mathrm{y}}) = \sigma_{\mathrm{y}}\left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}}, \boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{x}} \right\rangle + \mu_{\mathrm{y}}.$$

Then it is easy to verify that the global minimizer manifold is Γ.

Now we compute the Hessian. Let $\tilde{\boldsymbol{x}}_i := \boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{x}}$ and $q_i := \frac{y_i-\mu_{\mathrm{y}}}{\sigma_{\mathrm{y}}}$. Because of the use of squared loss, on Γ the Hessian can be written as the sum of outer products of gradients:

$$\boldsymbol{H}(\boldsymbol{w}) = \frac{2}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{w}}\Phi(\boldsymbol{x}_i;\boldsymbol{w},\sigma_{\mathrm{y}},\mu_{\mathrm{y}})\nabla_{\boldsymbol{w}}\Phi(\boldsymbol{x}_i;\boldsymbol{w},\sigma_{\mathrm{y}},\mu_{\mathrm{y}})^\top.$$

For each gradient we have

$$\nabla_{\boldsymbol{w}}\Phi(\boldsymbol{x}_i;\boldsymbol{w},\sigma_{\mathrm{y}},\mu_{\mathrm{y}}) = \frac{\sigma_{\mathrm{y}}}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}}\left(\boldsymbol{I} - \frac{\boldsymbol{\Sigma}_{\mathrm{x}}\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}^2}\right)\tilde{\boldsymbol{x}}_i = \frac{\sigma_{\mathrm{y}}}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}}\left(\boldsymbol{I} - \boldsymbol{\Sigma}_{\mathrm{x}}\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top\right)\tilde{\boldsymbol{x}}_i$$

Note that $\boldsymbol{\Sigma}_{\mathrm{x}}\tilde{\boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\top\tilde{\boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^{n}q_i\tilde{\boldsymbol{x}}_i =: \boldsymbol{z}$. Then we can simplify the gradient by

$$\nabla_{\boldsymbol{w}}\Phi(\boldsymbol{x}_i;\boldsymbol{w},\sigma_{\mathrm{y}},\mu_{\mathrm{y}}) = \frac{\sigma_{\mathrm{y}}}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}}\left(\tilde{\boldsymbol{x}}_i - q_i\boldsymbol{z}\right)$$

Now we simplify the Hessian.

$$\begin{aligned}
\boldsymbol{H}(\boldsymbol{w}) &= \frac{2}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{w}}\Phi(\boldsymbol{x}_i;\boldsymbol{w},\sigma_{\mathrm{y}},\mu_{\mathrm{y}})\nabla_{\boldsymbol{w}}\Phi(\boldsymbol{x}_i;\boldsymbol{w},\sigma_{\mathrm{y}},\mu_{\mathrm{y}})^\top \\
&= \frac{2\sigma_{\mathrm{y}}^2}{n\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}^2}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{x}}_i - q_i\boldsymbol{z}\right)\left(\tilde{\boldsymbol{x}}_i - q_i\boldsymbol{z}\right)^\top \\
&= \frac{2\sigma_{\mathrm{y}}^2}{n\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}^2}\left(\sum_{i=1}^{n}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\top - \sum_{i=1}^{n}q_i\tilde{\boldsymbol{x}}_i\boldsymbol{z}^\top - \sum_{i=1}^{n}q_i\boldsymbol{z}\tilde{\boldsymbol{x}}_i^\top + \sum_{i=1}^{n}q_i^2\boldsymbol{z}\boldsymbol{z}^\top\right) \\
&= \frac{2\sigma_{\mathrm{y}}^2}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}^2}\left(\boldsymbol{\Sigma}_{\mathrm{x}} - \boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{z}\boldsymbol{z}^\top + \boldsymbol{z}\boldsymbol{z}^\top\right) \\
&= \frac{2\sigma_{\mathrm{y}}^2}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}^2}\left(\boldsymbol{\Sigma}_{\mathrm{x}} - \boldsymbol{z}\boldsymbol{z}^\top\right).
\end{aligned}$$

We complete the proof by noting that $\frac{\sigma_{\mathrm{y}}^2}{\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_{\mathrm{x}}}^2} = \|\tilde{\boldsymbol{w}}\|_2^2$. □

Proof for Theorem 5.1. By Lemma O.1, for $\boldsymbol{w} \in \mathbb{S}^{D-1}$,

$$\nabla_\Gamma \log \lambda_1^{\mathrm{H}}(\boldsymbol{\theta}) = \nabla_\Gamma \log \|\tilde{\boldsymbol{w}}\|_2^2 = \frac{1}{\|\tilde{\boldsymbol{w}}\|_2^2}\nabla_\Gamma\|\tilde{\boldsymbol{w}}\|_2^2.$$

By simple calculation, it can be verified that the only point that has $\nabla_\Gamma \|\tilde{\boldsymbol{w}}\|_2^2 = \mathbf{0}$ is the unique point $\boldsymbol{w}^*$ on Γ that is a linear combination of $\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{x}}$. As the shperical sharpness is bounded from below and there is only one stationary point, the sharpness-reduction flow must converge on Γ and the convergence point must be $\boldsymbol{w}^*$.

Since $\boldsymbol{w}^*$ is a linear combination of $\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{x}}$, the associated $\tilde{\boldsymbol{w}}^*$ should be the least square solution (without bias) of this "shifted" dataset: $\{(\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{x}}, y_i - \mu_{\mathrm{y}})\}$. In other words, $\tilde{\boldsymbol{w}}^*$ is the optimal solution of the following constrained optimization problem:

$$\min \quad \|\boldsymbol{w}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{w}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathrm{x}}) = y_i - \mu_{\mathrm{y}}, \quad \forall i \in [n].$$

When $\boldsymbol{w}$ is given in the above optimization problem, there is only a unique b such that $\boldsymbol{w}^\top \boldsymbol{x}_i + b = y_i$. So we can introduce a bias to this problem without changing the optimal solution:

$$\min \quad \|\boldsymbol{w}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{w}^\top \boldsymbol{x}_i + b = y_i, \quad \forall i \in [n].$$

One can also easily see that this b must match with $\tilde{b}$. Therefore, we can conclude that the sharpness-reduction flow (3) finds to the optimal solution of (M) at convergence. □

P Experiments

In this section, we provide experiments on matrix completion and CIFAR-10 to validate the main claim in our theory: GD+WD on scale-invariant loss persistently reduces spherical sharpness in the EoS regime (the regime where $2/\tilde{\eta}_t$ roughly equals to the spherical sharpness). In addition, we validate that the generalization performance continues to improve as the spherical sharpness decreases. See Appendices P.1 and P.2. Then in Appendix P.3, we validate a key proof insight: the magnitude of oscillation and effective LR evolve periodically.

We also provide a series of ablation studies. In Appendices P.4 and P.5, we demonstrate that the two key components in our theoretical setup, normalization and WD, are crucial. In Appendix P.6, we show that the initial effective LR does not affect the final performance when the intrinsic LR is fixed.

P.1 Validation of Sharpness Reduction on Matrix Completion

First, we conduct experiments on matrix completion: there is an unknown low-rank ground-truth matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, and only N entries of $\boldsymbol{M}$ are known. The goal is to recover $\boldsymbol{M}$ given the known entries. Matrix completion has been studied via deep learning techniques (see, e.g., [20] for a survey). Matrix completion can be connected to supervised learning as follows. Each entry of the matrix can

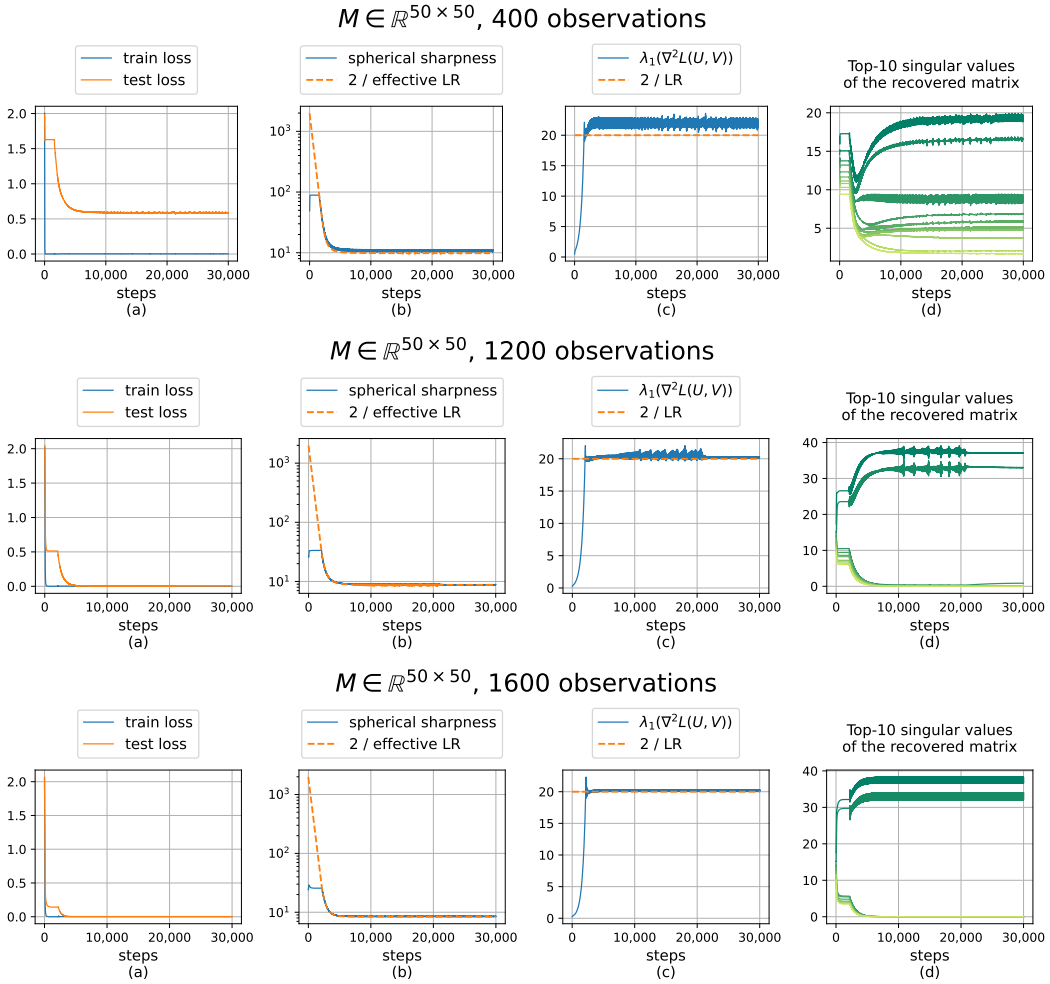


Figure 7: Overparameterized matrix completion with BN, where the ground-truth matrix $\boldsymbol{M} \in \mathbb{R}^{50\times 50}$ is of rank 2, and the number of observations varies from 400 to 1600. See Figure 1 for the case of 800 observations. The test loss starts to decrease significantly as soon as the spherical sharpness starts to decrease.

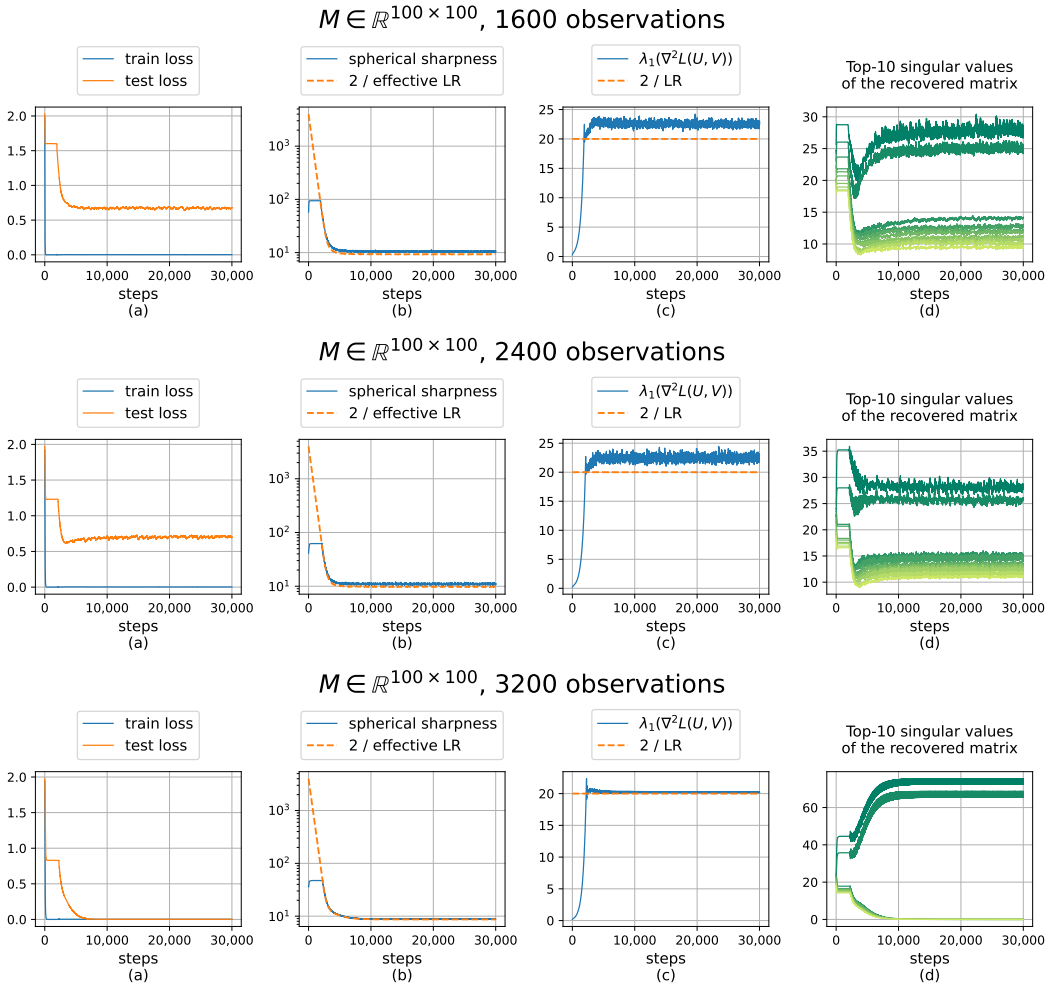


Figure 8: Overparameterized matrix completion with BN, where the ground-truth matrix $\boldsymbol{M} \in \mathbb{R}^{50\times 50}$ is of rank 2, and the number of observations vary from 1600 to 3200. The test loss starts to decrease significantly as soon as the spherical sharpness starts to decrease.

be seen as a data point, where the N observed entries constitute the training set. As in supervised learning, given a subset of data points, the goal of matrix completion is to build a model to predict the rest of the data points.

Here we empirically study solving matrix completion using an overparameterized scale-invariant model. We can observe that spherical sharpness is indeed decreasing as soon as the process enters the EoS regime, and the reduction of spherical sharpness encourages low-rank.

In our experiments, we generate the ground-truth matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ as follows. First, we set $\widetilde{\boldsymbol{M}} \leftarrow \boldsymbol{U}_*\boldsymbol{V}_*^\top$ for two random matrices $\boldsymbol{U}_*, \boldsymbol{V}_* \in \mathbb{R}^{d\times 2}$, where every entry is sampled uniformly from $[-1, 1]$. Then we obtain $\boldsymbol{M}$ by normalizing $\widetilde{\boldsymbol{M}}$ so that the second moment of the entries is 1, i.e., $\boldsymbol{M} \leftarrow (d \cdot \|\widetilde{\boldsymbol{M}}\|_{\mathrm{F}}^{-1})\widetilde{\boldsymbol{M}}$. We uniformly sample N entries of $\boldsymbol{M}$ to serve as the observations, and use $\Omega \subseteq [d] \times [d]$ to denote the index set of observed entries.

Matrix completion has been studied by a line of works [41, 76, 4, 100, 79, 108] as a test-bed for the implicit regularization of gradient descent. More specifically, they parameterize the target matrix as $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^\top$ where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{d\times d}$ are two trainable matrices, and run GD to minimize the squared loss $\mathcal{L}(\boldsymbol{U}, \boldsymbol{V}) := \frac{1}{N}\sum_{(i,j)\in\Omega}(W_{i,j} - M_{i,j})^2$. Although there is no explicit constraint on rank, GD with small random initialization can still exhibit an implicit bias towards low-rank solutions.

Inspired by this line of works, we conduct matrix completion experiments to test if GD+WD exhibits the same low-rank bias in training overparameterized models with BN. More specifically, we parameterize the target matrix as $\boldsymbol{W} = \mathrm{BN}(\boldsymbol{U}\boldsymbol{V}^\top)$, where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{d\times d}$ are two trainable matrices. Given the observed positions Ω, the output of the model for a single position $(i,j) \in \Omega$ is $\frac{\gamma}{\sigma}[\boldsymbol{U}\boldsymbol{V}^\top]_{i,j}$, where $\sigma^2 := \frac{1}{N}\sum_{(i,j)\in\Omega}[\boldsymbol{U}\boldsymbol{V}^\top]_{i,j}^2$ is the second moment of $[\boldsymbol{U}\boldsymbol{V}^\top]_{i,j}$ over all observed positions, and γ is a rescaling factor. Multiplying the factor $\frac{\gamma}{\sigma}$ can be seen as doing BN over observed entries because it rescales the output in the same manner as BN. But the difference is that, for the sake of simplicity, we do not subtract the mean. To ensure the loss to be scale-invariant, we also fix γ to match the second moment of observed entries, i.e., $\gamma^2 := \frac{1}{N}\sum_{(i,j)\in\Omega} M_{i,j}^2$.

To train this model, we use the standard squared loss, and we run gradient descent with LR $\hat{\eta} = 0.1$ and WD $\hat{\lambda} = 0.01$ to optimize the loss. It is obvious that this loss is scale-invariant due to BN.

$$\mathcal{L}(\boldsymbol{U}, \boldsymbol{V}) := \frac{1}{N}\sum_{(i,j)\in\Omega}\left(\frac{\gamma}{\sigma}[\boldsymbol{U}\boldsymbol{V}^\top]_{i,j} - M_{i,j}\right)^2.$$

For an unobserved entry (i,j), the model uses the same batch statistics as for observed entries, and predicts $\frac{\gamma}{\sigma}[\boldsymbol{U}\boldsymbol{V}^\top]_{i,j}$, where γ, σ are the same as above. So we measure the test loss as

$$\bar{\mathcal{L}}(\boldsymbol{U}, \boldsymbol{V}) := \frac{1}{d^2}\sum_{i,j\in[d]}\left(\frac{\gamma}{\sigma}[\boldsymbol{U}\boldsymbol{V}^\top]_{i,j} - M_{i,j}\right)^2.$$

We note that the loss function has no explicit constraint on rank. But surprisingly, in our experiments, GD+WD tends to prefer low-rank solution as soon as the sharpness-reduction bias starts to occur. See Figures 1 and 7 for experiments on reconstructing a rank-2 matrix of size 50×50, and Figure 8 for experiments on reconstructing a rank-2 matrix of size 100×100. In all these experiments, we can observe that the train loss first decreases to near zero in a short time, but the test loss remains at a high level. Then after effective LR increases for some more steps, the dynamics enters the EoS regime, and as predicted by our theory, the spherical sharpness starts to decrease. Meanwhile, the test loss also starts to decrease significantly. We notice that the gap between the second and the third largest singular values of the recovered matrix is enlarged at the same time, suggesting that this reduction of spherical sharpness encourages the recovered matrix to be low-rank.

P.2 Validation of Sharpness Reduction on CIFAR-10

Now we present experiments on CIFAR-10 with crossentropy loss to validate the sharpness-reduction bias in a more realistic setting. We run full-batch GD with accompanying WD on three different architectures: a scale-invariant variant of VGG-11, a scale-invariant variant of pre-activation ResNet-20, and the standard pre-activation ResNet-20. We fix LR $\hat{\eta} = 0.1$ and WD $\hat{\lambda} = 5 \times 10^{-4}$. See Appendix Q.1 for more details on the architectures and training procedures.

The experiment for the scale-invariant variants of VGG-11 and ResNet-20 are presented in Figures 2 and 9. Here our ResNet-20 is non-smooth due to the use of ReLU, but our VGG-11 is smooth because we choose to use Swish activation and mean pooling in this network. For both the smooth VGG-11 and non-smooth ResNet-20, it can be seen from the plots that in both cases the spherical sharpness has an overall tendency to decrease over time, and the test accuracy is increasing accordingly.

We note that this sharpness-reduction bias actually goes beyond the setting that our theory can directly apply: the dynamic enters the EoS regime in the very beginning of training, but our theory only analyzes the dynamic near a local minimizer manifold. This is because that the LR here is not small enough so that $\boldsymbol{\theta}_t$ can approach to a local minimizer before the dynamic enters the EoS regime. In fact, Figure 11 shows that in the scale-invariant VGG-11 experiment, the initial spherical sharpness is much larger than $2/\tilde{\eta}_0$, then after a few steps, the spherical sharpness decreases to a level that is close to $2/\tilde{\eta}_t$ and the dynamic enters the EoS regime.

Besides the scale-invariant models, we also validate the sharpness-reduction bias on the standard pre-activation ResNet-20, which is not (fully) scale-invariant but only scale-invariant to a part of its parameters. For evaluating spherical sharpness, we compute the partial Hessian only with respect to that part of parameters. See Figure 10 for the plot and Appendix Q.1 for more experimental details. Although our theory can only cover scale-invariant models, we can still observe the sharpness-reduction bias in experiments, and the test accuracy increases as spherical sharpness decreases.

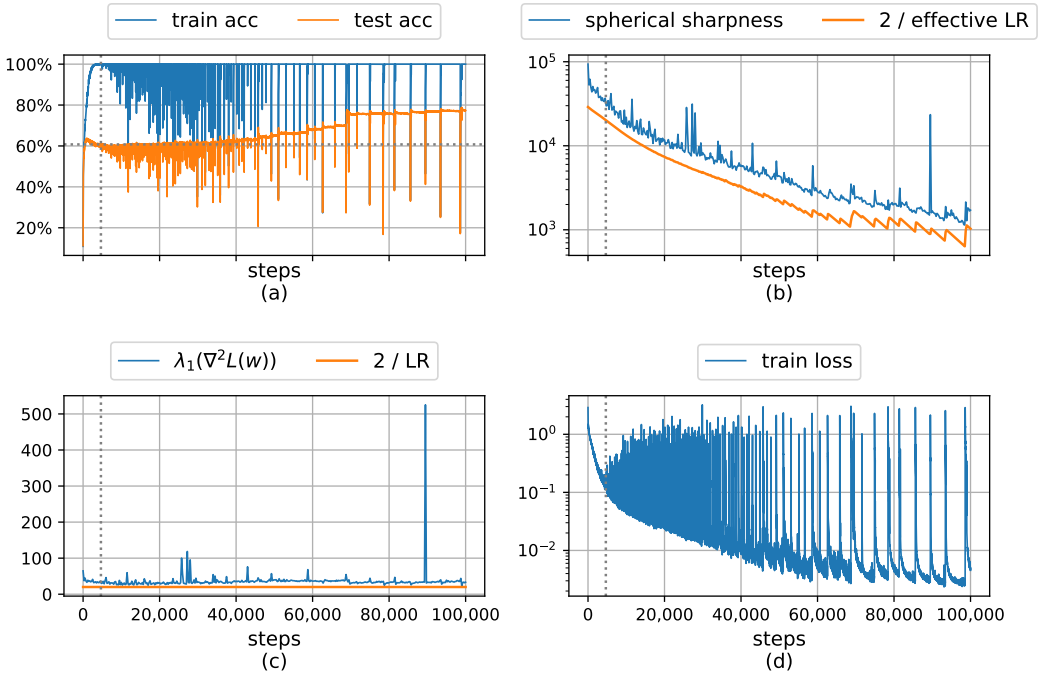


Figure 9: In training a scale-invariant ResNet-20 on CIFAR-10 with (full-batch) GD+WD, the spherical sharpness decreases over time. 100% training accuracy is achieved after ~ 4700 steps (dotted line), but as the training continues, the test accuracy increases from 60.8% to 77.4%.

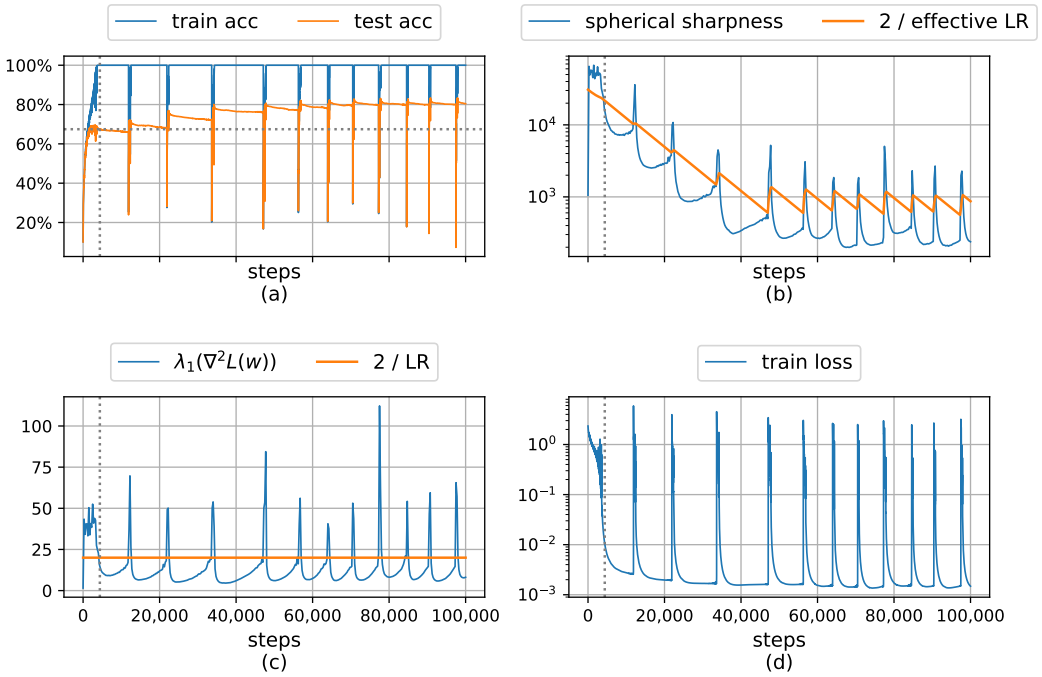


Figure 10: In training the standard pre-activation ResNet-20 on CIFAR-10 with (full-batch) GD+WD, the spherical sharpness is decreasing in the long term run. 100% training accuracy is achieved after ~ 4400 steps (dotted line), but as the training continues, the test accuracy increases from 67.5% to 80.4%. Spherical sharpness is only evaluated for a scale-invariant part of the trainable parameters.

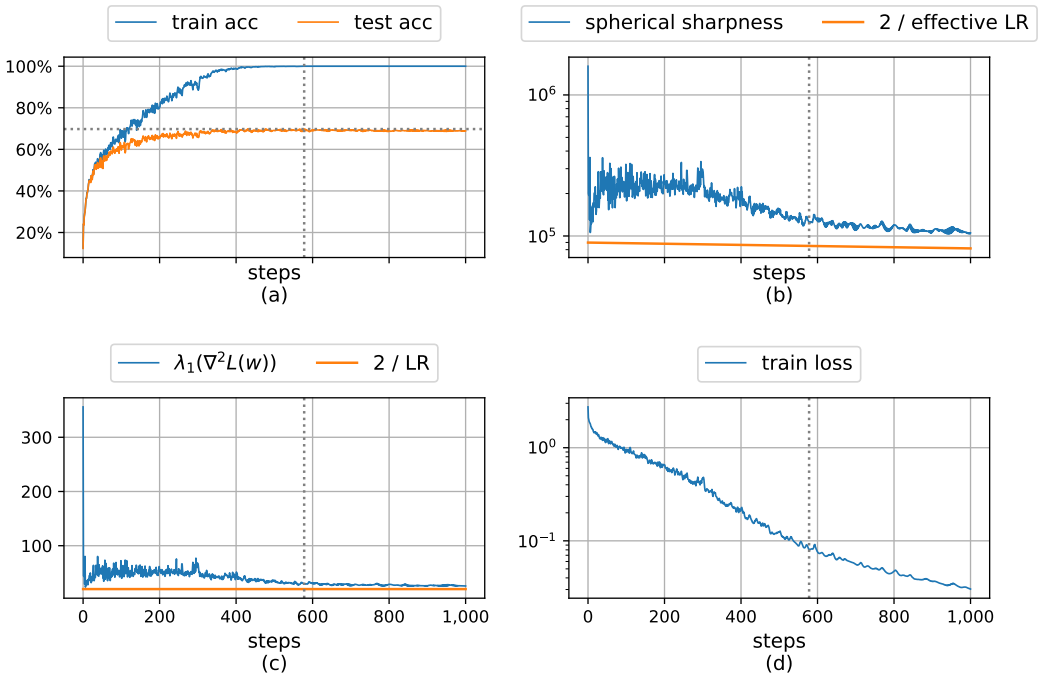


Figure 11: The first 1500 steps in training a smooth and scale-invariant VGG-11 on CIFAR-10 with GD+WD (see also Figure 2). The spherical sharpness is much larger than $2/\tilde{\eta}_0$ initially but is then reduced to the same level as $2/\tilde{\eta}_t$ after a few steps.

P.3 Periodic Behaviors

A key insight in our proof is that the magnitude of oscillation and effective LR evolve periodically in the EoS regime. Following Section 4.2.3, we can divide the EoS regime into two sub-regimes that occur alternatively in training: (1) the sub-regime where $2/\tilde{\eta}_t$ is smaller than $\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$ and the magnitude of oscillation $|h_t|$ keeps increasing; and (2) the sub-regime where $2/\tilde{\eta}_t$ is bigger than $\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$ and the magnitude of oscillation $|h_t|$ keeps decreasing. Here $\boldsymbol{\phi}_t$ is a projection of $\boldsymbol{\theta}_t$ onto the local minimizer manifold, $\lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$ is the spherical sharpness computed at $\boldsymbol{\phi}_t$, and h_t is the inner product between $\boldsymbol{\theta}_t - \boldsymbol{\phi}_t$ and the top eigenvector of $\boldsymbol{H}(\boldsymbol{\phi}_t)$.

Figure 5 provides a nice validation of this periodic behavior in linear regression with BN. We further validate the periodic behavior in matrix completion by visualizing $\tilde{\eta}_t$ in Figure 12 around the moment that the dynamic enters the EoS. We do not plot $|h_t|$ because computing $\boldsymbol{\phi}_t$ is inefficient, but we can observe that the training loss evolves periodically, which signals that $|h_t|$ evolves periodically as well.

In CIFAR-10 experiments with realistic LR and WD, the periodic behavior still exists, but it may deviate from the regime that our theory can capture. In most of our CIFAR-10 experiments, only the sub-regime where $2/\tilde{\eta}_t < \lambda_1^{\mathrm{H}}(\boldsymbol{\phi}_t)$ can be observed, and the phenomenon that $2/\tilde{\eta}_t$ switches back and forth between being smaller and larger than the spherical sharpness does not occur anymore. However, the change of $\tilde{\eta}_t$ and spherical sharpness still cause to the loss and gradient norm to fluctuate periodically. See Figure 13 for details. This difference is because that the loss function of a neural net on CIFAR-10 is much less smooth than that of linear regression and matrix completion. To make the dynamic happen in the regime that our theory describes, the LR and WD need to be very small so that $\boldsymbol{\theta}_t$ is sufficiently close to the minimizer manifold Γ, but they cannot be that small in practice due to computational inefficiency.

The periodic behavior can also happen in a macroscopic scale on CIFAR-10. E.g., in Figure 2, the training loss spikes to a high value around step 50,000, but then recovers to a near-zero value after a few hundred steps. Such a macroscopic periodic behavior may not always lead to sharpness reduction, in contrast to the microscopic periodic behavior of our interest. But as noted in the extensive empirical study by Lobacheva et al. [84], this macroscopic periodic behavior sometimes leads to better generalization. They also provide a theoretical analysis for the cause assuming the

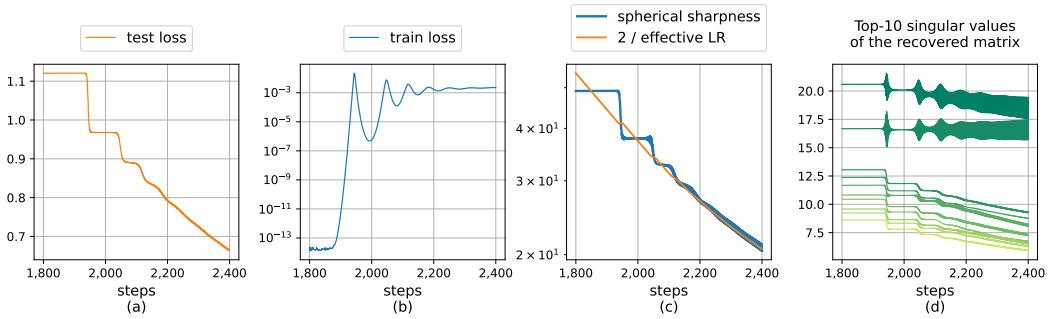


Figure 12: A closer look at the dynamic in the matrix completion experiment (Figure 1) around the moment that it enters the EoS regime. $2/\tilde{\eta}_t$ switches back and forth between being smaller and larger than the spherical sharpness (computed as $\lambda_1^{\mathrm{H}}(\boldsymbol{\theta}_t)$ for efficiency), which causes the training loss to oscillate periodically.

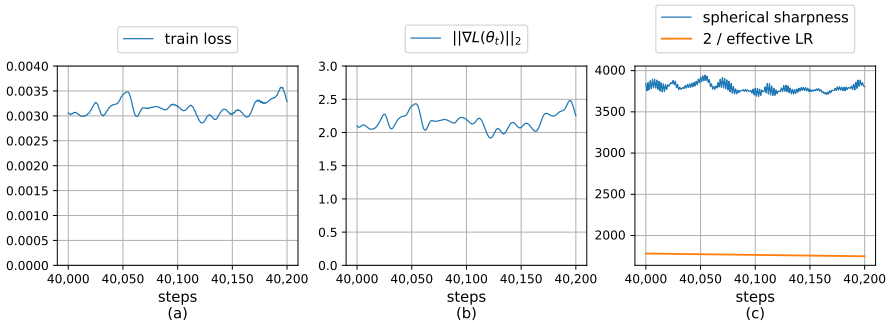


Figure 13: A closer look at a sample of 200 steps in the middle of the VGG-11 training on CIFAR-10 (Figure 2). $2/\tilde{\eta}_t$ is always smaller than the spherical sharpness, but the training loss and gradient norm still fluctuate periodically.

gradient norm is both lower and upper bounded. However, their explanation is not completely satisfactory because the gradient norm does not admit a lower bound when the parameter is close to the minimizer manifold Γ. We leave it a future work to look further into the cause of macroscopic periodic behavior and how it helps generalization.

P.4 Ablation Study: Normalization

Figure 2 and many other experiments in Appendices P.1 and P.2 have shown that GD+WD on normalized nets can continue to improve test accuracy even after reaching 100% test accuracy. In our theoretical analysis, we connect this phenomenon with the reduction of spherical sharpness during training.

Now we empirically validate that this phenomenon is indeed linked the presence of normalization layers in neural nets. We train a VGG-11 in the same setting as Figure 2, but now we remove all the normalization layers. See Figure 14. The test accuracy does not increase anymore after the training accuracy reaches 100%.

Note that the spherical sharpness is meaningful only for normalized nets. For unnormalized nets, even if GD is implicitly reducing a similar sharpness measure, such a measure cannot be strongly related to generalization because the test accuracy is not increasing accordingly.

P.5 Ablation Study: Weight Decay

Now we conduct experiments to study the effects of weight decay (WD) on sharpness reduction. It is crucial to have non-zero WD in our theoretical analysis. Otherwise, it is implied by Lemma E.1 that the effective LR is monotone decreasing, but our theory only applies to the effective LRs that can

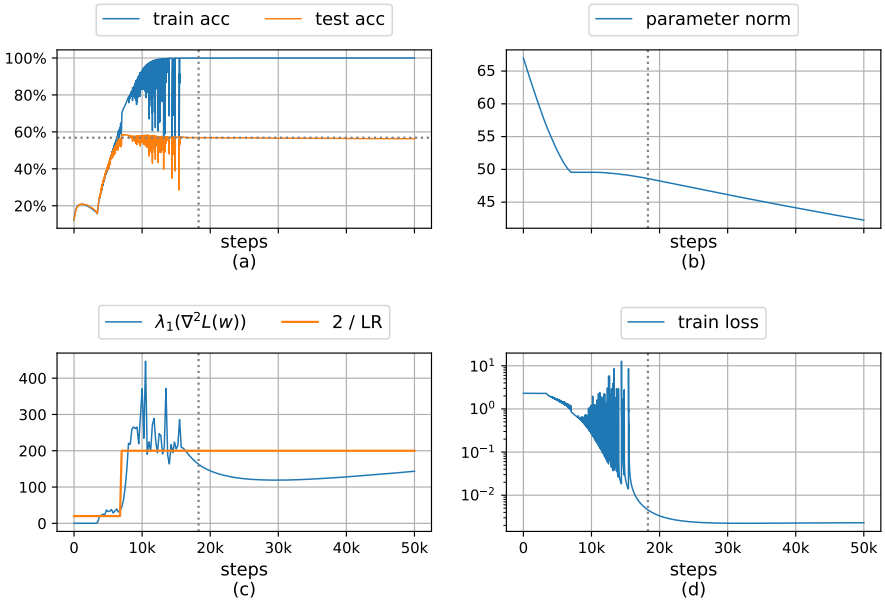


Figure 14: A VGG-11 is trained on CIFAR-10 in the same setting as Figure 2 but now the normalization layers are all removed. The test accuracy slowly decreases from 56.8% to 56.3% after the training accuracy reaches 100%. The LR is set to 0.1 initially and is decayed to 0.01 at step $7k$ to avoid instability.

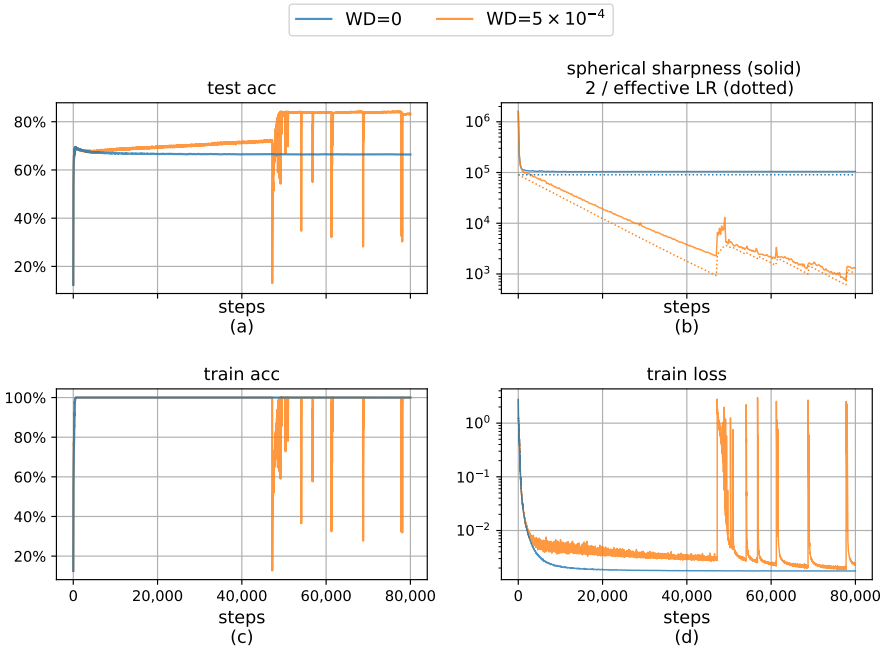


Figure 15: In training scale-invariant VGG-11 on CIFAR-10 with full-batch GD (LR $\tilde{\eta} = 0.1$), weight decay plays an important role in sharpness reduction. When weight decay $\hat{\lambda} = 5 \times 10^{-4}$, the spherical sharpness persistently decreases, and the test accuracy increases from 69.1% to 84.3% (see also Figure 2). But if there is no weight decay, the spherical sharpness does not change much after a few initial steps, and the test accuracy is stuck at 66.4%.

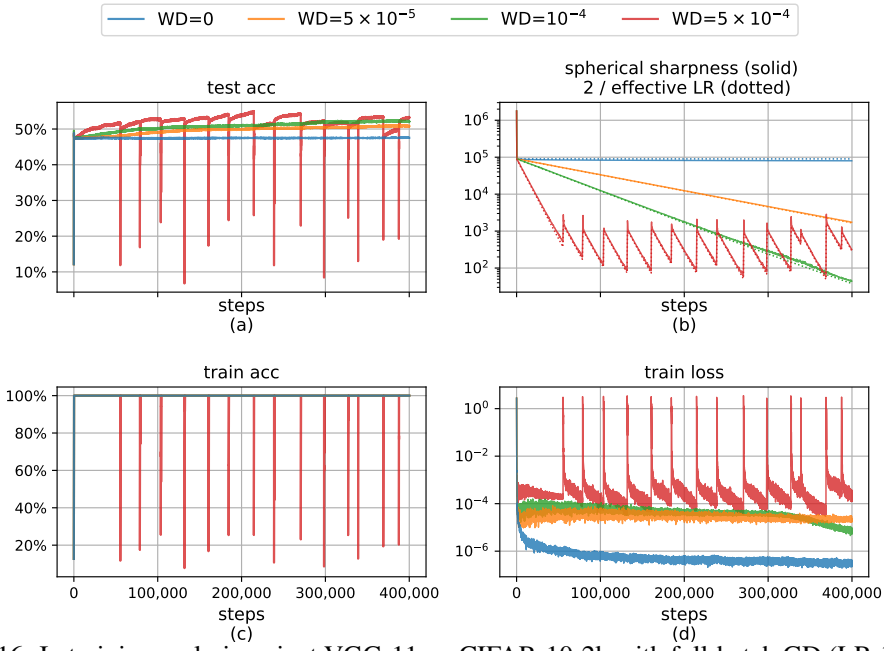


Figure 16: In training scale-invariant VGG-11 on CIFAR-10-2k with full-batch GD (LR $\tilde{\eta} = 0.1$), the sharpness-reduction bias occurs as long as the weight decay is non-zero, but smaller WD leads to longer training time to reduce the spherical sharpness to the same level of that with larger WD. The best test accuracy achieved within 400,000 steps is 48.9% when $\hat{\lambda} = 0$, 51.1% when $\hat{\lambda} = 5 \times 10^{-5}$, 52.5% when $\hat{\lambda} = 10^{-4}$, 55.1% when $\hat{\lambda} = 5 \times 10^{-4}$.

be viewed as quasi-RMSprop (see Definition B.5 and Theorem B.9), in which the effective LR can either increase or decrease as gradient norm changes. A previous work by Arora et al. [7] provides a detailed theoretical analysis in this case showing that the dynamic always stays in the stable regime after a few warm-up steps, and the parameter eventually converges to a stationary point on the unit sphere under standard assumptions in optimization.

We can verify through experiments that the requirement of WD to be non-zero is not a technical artifact, but indeed a necessity in practice to exhibit the sharpness-reduction bias. We train a scale-invariant VGG-11 following exactly the same hyperparameters and initialization as Figure 2, except that we set WD $\hat{\lambda}$ to zero. The result is presented in Figure 15. The spherical sharpness is no longer decreasing with time when WD is zero, and the final test accuracy is much lower than the VGG-11 with WD 5×10^{-4}.

We further conduct experiments with smaller (but non-zero) WD than our default value 5×10^{-4}. For the sake of computational efficiency, our experiments are conducted on a subset of CIFAR-10 images consisting of 2K images, which we call CIFAR-10-2k (see Appendix Q.1). The result is presented in Figure 16, from which we can observe that GD exhibits a sharpness-reduction bias as long as WD is non-zero. But note that smaller WD leads to a slower speed of sharpness reduction. This is an expected phenomenon because the speed of tracking the sharpness-reduction flow (3) is controlled by the intrinsic LR $\eta_{\text{in}} := \hat{\eta}\hat{\lambda}$. The smaller the WD, the smaller the intrinsic LR (when LR $\hat{\eta}$ is fixed).

P.6 Ablation Study: Initial Effective Learning Rate

As we have seen in Figure 11, in training the scale-invariant VGG-11 on CIFAR-10, if $\hat{\eta} = 0.1$ and $\hat{\lambda} = 5 \times 10^{-4}$, then initially the spherical sharpness is much larger than $2/\tilde{\eta}_0$, and after a few steps, it decreases to the same level as $2/\tilde{\eta}_t$, thus bringing the dynamic into the EoS regime.

Now we conduct an ablation study to see how the dynamics change with different initial effective LR. For computational efficiency, we run experiments on CIFAR-10-2k (see Appendix Q.1) to train our scale-invariant variant of VGG-11. We first set LR $\hat{\eta} = 0.1$ and WD $\hat{\lambda} = 5 \times 10^{-4}$, then we change the LR $\hat{\eta}$ to $10, 10^{-3}, 10^{-5}$ while rescaling the WD $\hat{\lambda}$ inverse proportionally. In this way, the intrinsic

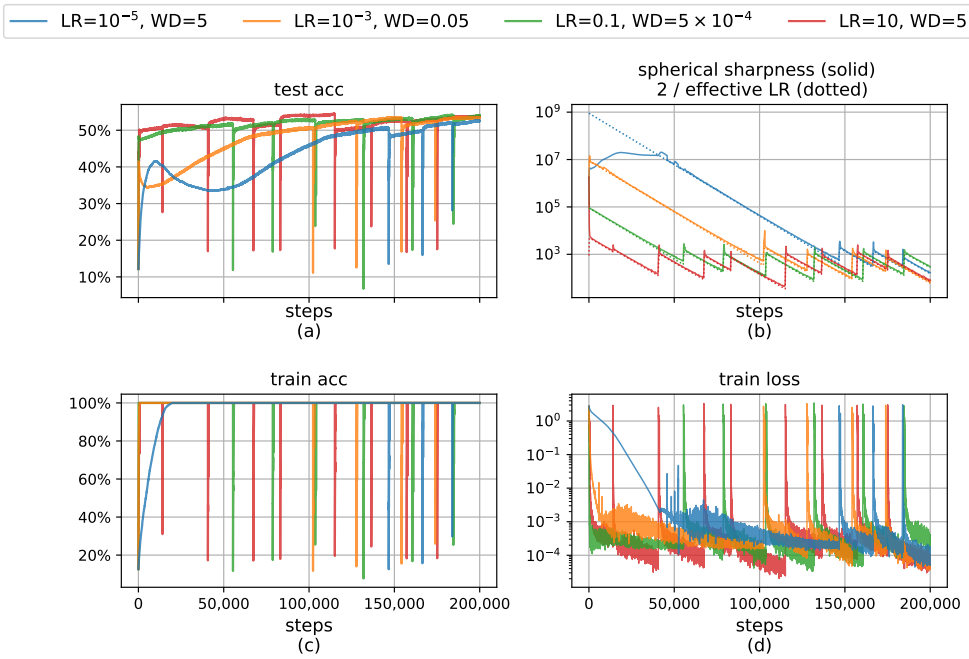


Figure 17: In training scale-invariant VGG-11 on CIFAR-10-2k with full-batch GD and various different initial effective learning rates $\tilde{\eta}_0$. We scale learning rate $\hat{\eta}$ and weight decay $\hat{\lambda}$ inverse proportionally while fixing the intrinsic LR $\eta_{\text{in}} = 5 \times 10^{-5}$. The best test accuracies achieved within 200,000 steps are similar in all runs, with 52.7% for $\hat{\eta} = 10^{-5}$, 53.7% for $\hat{\eta} = 10^{-3}$, 54.3% for $\hat{\eta} = 0.1$, 54.6% for $\hat{\eta} = 10$.

LR remains unchanged. See Figure 17 for the results. For large initial effective LR, the dynamic quickly enters the EoS regime; but when the initial effective LR is as small as 10^{-5}, the dynamic is in the stable regime initially, and $2/\tilde{\eta}_t$ falls so slowly that the dynamic enters the EoS regime only after reaching 100% training accuracy. In all experiments, the spherical sharpness persistently decreases in the EoS regime, and the test accuracy increases accordingly. In the end, all the experiment runs reach nearly the same spherical sharpness and test accuracy, regardless the initial effective LR.

Q Experiment Details

All our experiments were conducted on NVIDIA RTX A5000 GPUs. The longest experiments are the full-batch training experiments on CIFAR-10, each of which took 7 days to run.

Q.1 Additional Details of CIFAR-10 Experiments

Our implementation of full-batch GD is based on the code of Cohen et al. [24][5]. Our implementations of VGGNets [105] and ResNets [46, 47] on CIFAR-10 are based on a high-starred GitHub repository of Wei Yang (bearpaw)[6]. We do not use any data augmentation. As it is not feasible to do BN over the full dataset under our GPU memory constraints, we use ghost batch normalization [51] instead, where we split the dataset into 50 ghost batches of size 1000.

Some of our experiments are conducted on only a subset of CIFAR-10 consisting of 2000 images, which we call CIFAR-10-2k. We construct the dataset by scanning the full CIFAR-10 dataset and taking the first 200 images for each of the 10 classes. We use the standard batch normalization with full batch for all CIFAR-10-2k experiments.

[5] https://github.com/locuslab/edge-of-stability
[6] https://github.com/bearpaw/pytorch-classification

Three different neural network architectures are tested in our experiments, and we refer to them as scale-invariant VGG-11, standard ResNet-20, and scale-invariant ResNet-20 respectively.

Scale-Invariant VGG-11. Our scale-invariant VGG-11 architecture is similar to the configuration A of the original VGGNet [105], but we make the following changes. We add a BN layer (without affine parameters) between every convolution and activation to introduce scale-invariance, We use mean pooling instead of max pooling, Swish instead of ReLU to make the training loss smooth. We replace the final 3 fully-connected layers with only one fully-connected layer as in Yang's implementation. All convolutional and fully-connected layers have no bias terms. We add a BN layer after the last fully-connect layer, in which the affine parameters are fixed to $\gamma = \frac{3}{10}\ln 91 \approx 1.353$, $\beta = 0$ (see Appendix R for discussion). A visualization of the full architecture is given in Figure 18, whose output function can be shown to be scale-invariant with respect to the trainable parameters.

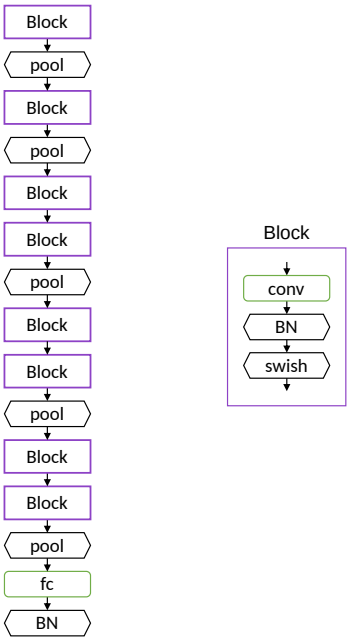
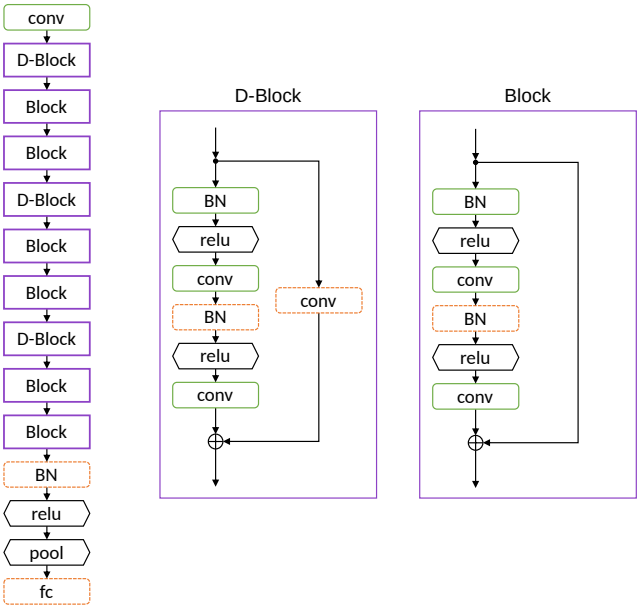


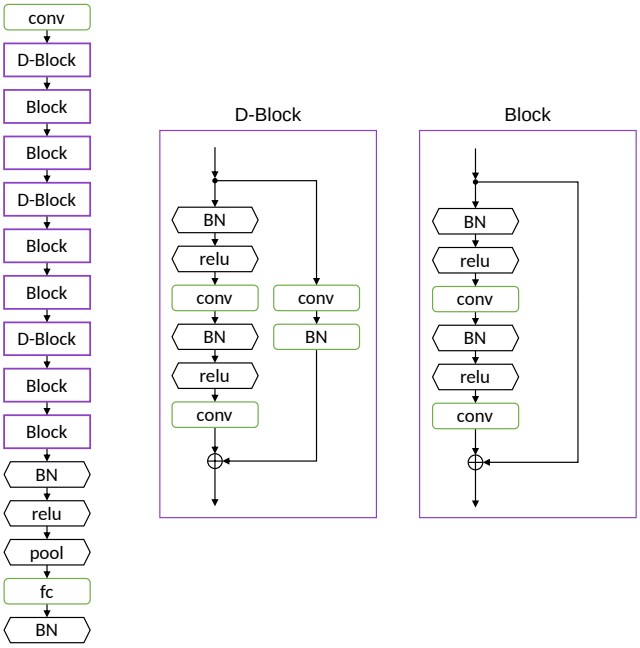
Figure 18: The architecture of a scale-invariant variant of VGG-11. Each rounded rectangle stands for a layer with trainable parameters, including convolutional and fully-connected layers. Each hexagon stands for a layer with no trainable parameters, including mean pooling, swish, BN without affine parameters, and BN with affine parameters fixed.

Standard ResNet-20. We also verify the sharpness-reduction bias on a more realistic setting with ResNet-20 and ReLU. In some of our experiments, we use Yang's implementation for the pre-activation variant of ResNet-20. All convolutional layers have no bias, but the final fully-connected layer does have bias terms. We note that the output function of this architecture is not scale-invariant to all its trainable parameters. However, it can be shown that the output function is indeed scale-invariant to a large part of its parameters. More specifically, the trainable parameters can be split into two parts, $\boldsymbol{w}$ and $\boldsymbol{z}$, and the output function remains unchanged if we replace $\boldsymbol{w}$ with $c\boldsymbol{w}$ for all $c > 0$. In our experiments, we measure the spherical sharpness by $\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z})$ while letting $\boldsymbol{\theta} := \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$. See Figure 19a for a visualization of the architecture.

Scale-Invariant ResNet-20. The standard ResNet-20 can be scale-invariant after making minor changes. First, we remove the affine parameters in all existing BNs. Then, we also add an extra BN (with trainable affine parameters) to each shortcut for blocks with downsampling, following Li and Arora [77]. Finally, we add a BN layer after the last fully-connect layer, in which the affine parameters are fixed to $\gamma = \frac{3}{10}\ln 91 \approx 1.353$, $\beta = 0$ (see Appendix R for discussion). A visualization of the full architecture is given in Figure 19b, whose output function can be shown to be scale-invariant with respect to the trainable parameters.



(a) The architecture of ResNet-20, following Yang's implementation.



(b) The architecture of a scale-invariant variant of ResNet-20.

Figure 19: The ResNet architectures used in our experiments. Each rounded rectangle stands for a layer with trainable parameters, including convolutional and fully-connected layers, and BN with trainable affine parameters. Each hexagon stands for a layer with no trainable parameters, including mean pooling, ReLU, BN without affine parameters, and BN with affine parameters fixed. In Yang's implementation of ResNet-20, the output function is scale-invariant to the parameters in layers marked as rounded green rectangles; in the scale-invariant variant, the output function is scale-invariant to all trainable parameters.

Q.2 Additional Details of Linear Regression Experiments

Our experiments on linear regression (Figure 5) follows strictly as Section 5. For generating data, we first sample the ground-truth weight $\boldsymbol{w}_{\mathrm{GT}} \in \mathbb{S}^{d-1}$ uniformly, where $d = 40$, and we sample the ground-truth bias as $b_{\mathrm{GT}} \sim \mathcal{N}(0, 0.01)$, Then we sample $n = 20$ points from Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathrm{diag}(1/d, 2/d, 3/d, \ldots, 1)$.

To generate Figure 5, we run gradient descent with LR $\hat{\eta} = 0.5$ and WD $\hat{\lambda} = 2 \times 10^{-4}$. This particular choice of hyperparameters is for better visualization of the periodic behavior described in Section 4.2.3. If we enlarge LR or WD, the period will be shortened accordingly.

Q.3 Computing Spherical Sharpness

To compute the spherical sharpness at a point $\boldsymbol{w} \in \mathbb{R}^D \setminus \{\mathbf{0}\}$ (Definition 1.1), we utilize the formula $\lambda_1(\nabla^2 \mathcal{L}(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2})) = \|\boldsymbol{w}\|_2^2 \cdot \lambda_1(\nabla^2 \mathcal{L}(\boldsymbol{w}))$, which can be proved for all scale-invariant loss by simple calculus (see also Lemma D.1). For computing the eigenvalues of $\nabla^2 \mathcal{L}(\boldsymbol{w})$, we invoke the Lanczos algorithm from SciPy library (`scipy.sparse.linalg.eigsh`), where a Hessian-vector product oracle $\boldsymbol{x} \mapsto \nabla^2 \mathcal{L}(\boldsymbol{w})\boldsymbol{x}$ is implemented with PyTorch and passed to the Lanczos algorithm as a linear operator to find its eigenvalues.

In full-batch training of the full CIFAR-10 dataset, it is time-consuming to compute even a single Hessian-vector product. Following Cohen et al. [24], we compute Hessian based on only the first 5000 training data points.

Note that in our theory we focus on the spherical sharpness of minimizers. However, in practice, it is usually time-consuming to compute the minimizers exactly, so in matrix completion and CIFAR-10 experiments, we compute the spherical sharpness directly at the current parameter $\boldsymbol{w}_t$. But for linear regression experiments (Figure 5), the spherical sharpness is indeed computed at minimizers in each step because our computational power is sufficient in this setting. Specifically, we compute the minimizer by doing projected gradient descent on $\mathbb{S}^{D-1}$ with fixed learning rate 0.005 until the loss decreases to 10^{-8}.

R Discussion on the Affine Parameters of the Final BN

In our experiments for scale-invariant models, we add a BN to the final linear layer and fix the affine parameters to be constants. Now we analyze its effects on regression tasks with squared loss and classification tasks with crossentropy loss, and discuss how to set the affine parameters reasonably.

We consider a scale-invariant neural net in general while assuming that a BN is put as the last layer. For input $\boldsymbol{x}$ and parameter $\boldsymbol{w}$, let $(F_1(\boldsymbol{x}; \boldsymbol{w}), \ldots, F_C(\boldsymbol{x}; \boldsymbol{w})) \in \mathbb{R}^C$ be the output before the final BN. After the final BN, the output is given by the following function $\Phi_k(\boldsymbol{w}; \boldsymbol{w})$:

$$\bar{F}_k(\boldsymbol{x}; \boldsymbol{w}) := \frac{F_k(\boldsymbol{x}; \boldsymbol{w}) - \mu_k}{\sigma_k},$$
$$\Phi_k(\boldsymbol{x}; \boldsymbol{w}) := \gamma_k \bar{F}_k(\boldsymbol{x}; \boldsymbol{w}) + \beta_k,$$

where μ_k and σ_k^2 are mean and variance of $\{F_k(\boldsymbol{x}_i; \boldsymbol{w})\}_{i=1}^n$ over the training set.

R.1 Squared Loss

For linear regression experiments, we use squared loss as the loss function, and we have $C = 1$ because the output is a scalar. Then as discussed in Section 5, the mean and variance of $\Phi_1(\boldsymbol{x}; \boldsymbol{w})$ are always β_1 and γ_1^2 regardless of $\boldsymbol{w}$. Therefore, to ensure that Φ has sufficient representation power to express the regression targets $\{y_i\}_{i=1}^n$, we should fix the affine parameters in a way that β_1 and γ_1^2 match with the mean and variance of $\{y_i\}_{i=1}^n$.

In the general case when $C \geq 1$, the k-th output unit has mean and variance β_k and γ_k^2. We can argue similarly that we should fix the affine parameters in a way that β_k and γ_k^2 match with the mean and variance of $\{y_{i,k}\}_{i=1}^n$, where $y_{i,k}$ stands for the k-th coordinate of the regression target for the i-th

data point. In other words,

$$\beta_k \leftarrow \mu_{\mathrm{y}}^{(k)} := \frac{1}{n}\sum_{i=1}^{n} y_{i,k}, \qquad \gamma_k \leftarrow \sigma_{\mathrm{y}}^{(k)} := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{i,k} - \mu_{\mathrm{y}}^{(k)})^2}.$$

R.2 Crossentropy Loss

In our CIFAR-10 experiments, we fix all γ_k to a constant value γ, and fix β_k to zero. We note that fixing these affine parameters in this way for an overparameterized model is equivalent to adding label smoothing to the training loss in some sense.

The basic idea is to assume the model has sufficient express any function F, then Φ_k can be any function of mean zero and covariance γ^2 on the training set. If the dataset has C classes and each class has equal numbers of samples, with some efforts one can show that the minimum loss is attained when the pre-softmax logit for an input is $\gamma(C-1)^{1/2}$ for the correct class and $-\gamma(C-1)^{-1/2}$ for every wrong class. In this case, the output probability $P_y(\boldsymbol{x}_i)$ for a class y is

$$P_y(\boldsymbol{x}_i) = \begin{cases} \frac{1}{1+(C-1)\exp(-\gamma((C-1)^{1/2}+(C-1)^{-1/2}))} & \text{if } y = y_i; \\ \frac{1}{(C-1)+\exp(\gamma((C-1)^{1/2}+(C-1)^{-1/2}))} & \text{otherwise.} \end{cases} \tag{51}$$

Label smoothing is a regularization technique proposed by Szegedy et al. [109] to improve the generalization of neural nets by replacing the one-hot hard labels with soft labels that assign a probability of ϵ/C to each wrong class. To see the connection to label smoothing, we can let $\epsilon := \frac{C}{(C-1)+\exp(\gamma((C-1)^{1/2}+(C-1)^{-1/2}))}$, then $P_y(\boldsymbol{x}_i) = \epsilon/C$ if $y \neq y_i$ and $P_y(\boldsymbol{x}_i) = 1-(1-1/C)\epsilon$ if $y = y_i$. In our CIFAR-10 experiments, $\gamma = \frac{3}{10}\ln 91 \approx 1.353$. Then by simple calculation $\epsilon = 0.1$, and therefore the neural net trained with γ_k and β_k fixed to γ and 0 is encouraged to produce the output probabilities to match with the soft labels.

The minimum loss is non-zero when the affine parameters are fixed. We can see from (51) that the loss attained by any minimizer is

$$\mathcal{L}_{\min} = \ln(1 + (C-1)\exp(-\gamma((C-1)^{1/2} + (C-1)^{-1/2}))). \tag{52}$$

In making plots for our experiments, we always subtract the original loss with its theoretical minimum value $\mathcal{L}_{\min}$. When $\gamma = \frac{3}{10}\ln 91$, this minimum value is $\mathcal{L}_{\min} = \ln\frac{100}{91} \approx 0.0943$.

Choosing γ to be $\frac{3}{10}\ln 91$ is mainly because of its connection to label smoothing with $\epsilon = 0.1$, which is the value chosen in [109]. But other choices of γ can also lead to the same sharpness-reduction bias in practice; see Figure 20.

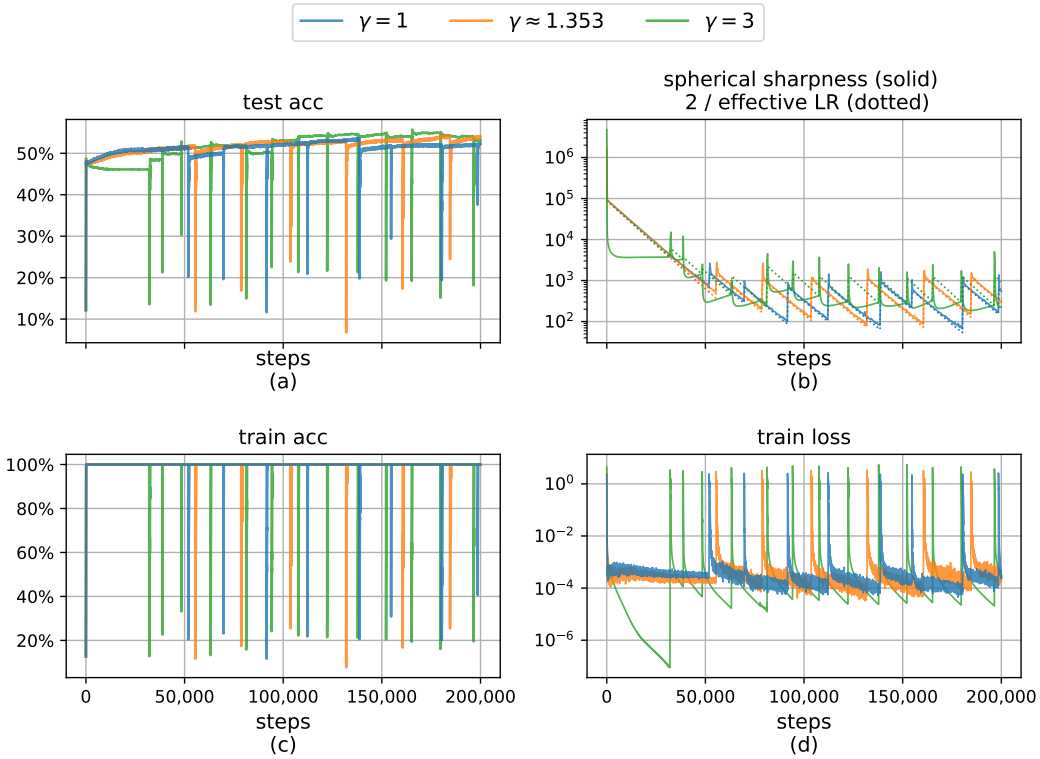


Figure 20: Training scale-invariant VGG-11 on CIFAR-10-2k with full-batch GD (LR $\hat{\eta} = 0.1$, WD $\hat{\lambda} = 5 \times 10^{-4}$) while fixing the rescaling parameter γ in the final BN to different values. The sharpness-reduction bias can be observed in all cases, but for larger γ the periodic behavior is more likely to happen at a macroscopic scale.