Adam Reduces a Unique Form of Sharpness: Theoretical Insights Near the Minimizer Manifold

Xinghan Li* **Haodong Wen*†** **Kaifeng Lyu‡**
Institute for Interdisciplinary Information Sciences
Tsinghua University
{xh-li22,whd25}@mails.tsinghua.edu.cn klyu@mail.tsinghua.edu.cn

Abstract

Despite the popularity of Adam optimizer in practice, most theoretical analyses study Stochastic Gradient Descent (SGD) as a proxy and little is known about how the solutions found by Adam differ. In this paper, we show that Adam reduces a specific form of sharpness measure shaped by its adaptive updates, leading to qualitatively different solutions from SGD. When the training loss is small, Adam wanders around the manifold of minimizers and takes semi-gradients to minimize this sharpness measure in an adaptive manner, a behavior we rigorously characterize via a continuous-time approximation using stochastic differential equations. We further illustrate how this behavior differs from that of SGD in a well-studied setting: When training overparameterized models with label noise, SGD has been shown to minimize the trace of the Hessian matrix, $\mathrm{tr}(\boldsymbol{H})$, whereas we prove that Adam minimizes $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{1/2})$ instead. In solving sparse linear regression with diagonal linear networks, Adam provably achieves better sparsity and generalization than SGD due to this difference. Finally, we note that our proof framework applies not only to Adam but also to many other adaptive gradient methods, including but not limited to RMSProp, Adam-mini, Adalayer and Shampoo. This provides a unified perspective for analyzing how adaptive optimizers reduce sharpness and may offer insights for future optimizer design.

1 Introduction

Due to the non-convexity of the loss landscape, neural networks trained in different ways can perform very differently on the test set, even if they achieve the same training loss or accuracy (Zhang et al., 2017; Keskar et al., 2017; Liu et al., 2023; Saunshi et al., 2024). To mathematically understand the generalization of neural networks, especially for over-parameterized models that admit many global minimizers, a key step is to understand the *implicit bias* of optimization methods (Neyshabur et al., 2014; Soudry et al., 2018). That is, beyond just minimizing the training loss, *what kinds of solutions are different optimizers implicitly biased towards?*

Many theoretical works on implicit bias focused on (full-batch) gradient descent or its continuous variant, gradient flow. This includes the works on the implicit bias towards max-margin classifiers (Soudry et al., 2018; Nacson et al., 2019; Lyu and Li, 2020; Ji and Telgarsky, 2020), implicit bias towards min-norm solutions (Lyu et al., 2024), and equivalence to kernel methods (Jacot et al., 2018; Chizat et al., 2019). However, these characterizations do not highlight the specific role of stochasticity in SGD, although it is more widely used in practice than gradient flow or full-batch gradient descent.

*Equal contribution; alphabet ordering

†Most work was done while Haodong was at Xi'an Jiaotong University.

‡Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Another line of works (Blanc et al., 2020; Damian et al., 2021; Li et al., 2021b) demonstrated that the gradient noise in SGD induces an additional form of implicit bias that reduces the *sharpness* of the solutions, a generalization measure that has been long observed to correlate with generalization (Hochreiter and Schmidhuber, 1997; Keskar et al., 2017; Jiang et al., 2020; Foret et al., 2021). More specifically, these works focus on the dynamics of SGD when the training loss is already small and the iterates are close to a manifold of minimizers. Li et al. (2021b) introduced a general framework to analyze the dynamics of SGD near the minimizer manifold, showing that SGD will not stop at arbitrary global minimizers, but drift and diffuse around the manifold, driving the iterates towards flatter regions of the loss landscape.

This behavior is mathematically characterized by a Stochastic Differential Equation (SDE), termed as *slow SDE* (Gu et al., 2023a), which accurately tracks the projected dynamics of SGD near the minimizer manifold over a timescale of $\mathcal{O}(\eta^{-2})$. The resulting dynamics reveal that SGD behaves like a gradient method on the manifold that takes semi-gradients to minimize a specific sharpness measure determined by the Hessian and gradient noise. See Section 3 for more details.

However, SGD is rarely used directly in modern deep learning. Instead, Adaptive Gradient Methods (AGMs) have become the de facto standard for training neural networks. Among them, Adam (Kingma and Ba, 2014) innovatively combines the moving average of the first and second moments of gradients to determine an adaptive learning rate for each parameter, and provides faster convergence and better stability than SGD across various domains (Ashish, 2017; Dosovitskiy et al., 2020; Schulman et al., 2017; Zhang et al., 2024c).

Despite the popularity of Adam, little is known about its implicit bias, especially how it is different from SGD in terms of reducing sharpness. In the literature, Ma et al. (2023) made attempts to generalize the slow SDE framework from SGD to Adam, but their analysis is specific to a two-dimensional loss function and involves a quasistatic approximation that lacks full mathematical rigor. Other works, such as Liu et al. (2023); Gu et al. (2024), leverage insights from the slow SDE developed for SGD to interpret empirical observations with Adam, but do not provide a theoretical analysis of Adam's own dynamics. A rigorous analysis of Adam's implicit bias in terms of sharpness remains an open problem.

Our Contributions. In this paper, we show that Adam implicitly reduces a unique form of sharpness and biases the iterates towards flatter regions in a way that is different from SGD, and provide separations between SGD and Adam in concrete theoretical cases.

1. In Section 4, we generalize the slow SDE for SGD to Adam. The slow SDE approximates the dynamics of Adam near the minimizer manifold, and reveals that Adam behaves like an adaptive gradient method that minimizes a unique form of sharpness by taking semi-gradients on the manifold.
2. In Section 5, we prove theoretically the generalization benefit of Adam under label noise settings. We show that under label noise setting, the implicit regularizer of Adam will reduce to $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{1/2})$ where $\boldsymbol{H}$ is the Hessian matrix. Compared to the $\mathrm{tr}(\boldsymbol{H})$ of SGD, this new kind of sharpness reduction usually aligns better with sparsity regularization, thus utilizing data more efficiently when the model is required to fit a sparse ground truth. We verify this anticipation experimentally through the diagonal net setting (Woodworth et al., 2020). We also demonstrated the discrepancy of the implicit biases of Adam and SGD through the matrix factorization setting in Appendix C.
3. Technically, our analysis holds for a general class of adaptive gradient methods (AGMs), including Adam, RMSProp, Adam-mini, Adalayer, and Shampoo (Gupta et al., 2018; Morwani et al., 2024). We develop several new tools that can be of independent interest, including a manifold projection operator tailored for AGMs, a high-probability convergence analysis for AGMs under PL conditions that directly gives a bound on $\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^*$.

2 Related Work

Implicit Bias of Adam. Despite Adam's widespread use, its implicit bias remains underexplored. Qian and Qian (2019) and Xie and Li (2024) analyzed AdaGrad and AdamW, but these techniques do not apply directly to Adam. Wang et al. (2021) showed Adam's regularizer matches SGD's under restrictive gradient-magnitude assumptions, and Zhang et al. (2024a) treated only linearly separable data, limiting practical relevance.

Also, another line of works on *implicit gradient regularization* (IGR) derive higher-order approximations for full-batch GD (Barrett and Dherin, 2020) and extend to Adam (Cattaneo et al., 2024; Cattaneo and Shigida, 2025). In particular, Cattaneo et al. (2024) argued that full-batch Adam with constant learning rate approximately follows an ODE that anti-regularizes sharpness when $\beta_1 < \beta_2$. Our work analyzes the dynamics of Adam for $\mathcal{O}(\eta^{-2})$ steps, a longer horizon than Cattaneo et al. (2024). Our analysis shows that with gradient noise, Adam can be characterized by an slow SDE that regularizes sharpness in the long term, offering a complementary perspective.

Slow SDE Approximation. To capture long-term behavior, we adopt the *slow SDE* technique of Li et al. (2021b) and Gu et al. (2023b). Standard SDE approximations (Li et al., 2018, 2021a; Cattaneo et al., 2024; Malladi et al., 2024) focus on the $\tilde{\mathcal{O}}(\eta^{-1})$ convergence phase and fail on the manifold. In contrast, slow SDEs peel off convergence to track the $\mathcal{O}(\eta^{-2})$ manifold dynamics accurately.

For more detailed discussion on the related work, please refer to Appendix B.

3 Preliminaries

Notations. Unless otherwise stated, for a square matrix $\boldsymbol{M}$, $\operatorname{diag}(\boldsymbol{M})$ denotes the vector consisting of its diagonal entries. The notation Diag has two usages: For a vector $\boldsymbol{v}$, $\operatorname{Diag}(\boldsymbol{v})$ denotes the diagonal matrix with $\boldsymbol{v}$ on its diagonal; and for a square matrix $\boldsymbol{M}$, $\operatorname{Diag}(\boldsymbol{M})$ denotes the diagonal matrix that only keeps $\boldsymbol{M}$'s diagonal entries and equals 0 elsewhere, i.e. $\operatorname{Diag}(\boldsymbol{M}) \stackrel{\text{def}}{=} \operatorname{Diag}(\operatorname{diag}(\boldsymbol{M}))$. For two vectors $\boldsymbol{u}$, $\boldsymbol{v}$ with the same dimension d, $\boldsymbol{u} \odot \boldsymbol{v}$ denotes element-wise multiplication $(u_1 v_1, \dots, u_d v_d)$. For any exponent p, $\boldsymbol{v}^{\odot p}$ denotes element-wise exponentiation, i.e. $\boldsymbol{v}^{\odot p} = (v_1^p, \dots, v_d^p)$, and $\sqrt{\boldsymbol{v}}$ means $\boldsymbol{v}^{\odot 1/2}$. We use $\mathbb{R}^d_{\geq 0}$ to denote the subset of $\mathbb{R}^d$ that has non-negative entries, and $\mathbb{S}^d_{++}$ to denote the space of $\mathbb{R}^{d\times d}$ positive definite matrices. We use the operator $\nabla_\Gamma(f)$ to represent the gradient of a function f projected to the tangent space of Γ.

For a mapping $F : \mathbb{R}^d \to \mathbb{R}^d$, we denote the Jacobian with respect to $\boldsymbol{\theta} \in \mathbb{R}^d$ as $\partial F(\boldsymbol{\theta}) \in \mathbb{R}^{d\times d}$, and $\partial^2 F(\boldsymbol{\theta})$ the second-order derivative at $\boldsymbol{\theta}$, which is a third-order tensor. Given a matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, we use the notation $\partial^2 F(\boldsymbol{\theta})[\boldsymbol{M}]$ to denote the second-order directional derivative of F at $\boldsymbol{\theta}$ in the direction $\boldsymbol{M}$, defined as $\partial^2 F(\boldsymbol{\theta})[\boldsymbol{M}] := \sum_{i\in[d]} \left\langle \frac{\partial^2 F_i}{\partial \boldsymbol{\theta}^2}, \boldsymbol{M} \right\rangle \boldsymbol{e}_i$, where F_i represents the i-th element in F, and $\boldsymbol{e}_i$ is the i-th vector of the standard basis. When the context is clear, we write $\partial^2(\nabla\mathcal{L})(\boldsymbol{\theta})[\boldsymbol{M}]$ as $\nabla^3\mathcal{L}(\boldsymbol{\theta})[\boldsymbol{M}]$ for brevity.

Loss Functions. Define $\ell(\boldsymbol{\theta}; \xi)$ as the loss function for a data sample ξ for a model with parameters $\boldsymbol{\theta}$. Define $\mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}_{\xi\sim\mathcal{S}}[\ell(\boldsymbol{\theta}; \xi)]$ as the training loss function, where $\mathcal{S}$ is the training dataset and $\xi \sim \mathcal{S}$ means the data sample ξ is drawn from $\mathcal{S}$ uniformly at random. Let $\mathcal{L}^* := \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta})$ be the minimum of training loss. Let $\mathcal{Z}(\boldsymbol{\theta})$ be the distribution of gradient noise $\nabla\ell(\boldsymbol{\theta}; \xi) - \nabla\mathcal{L}(\boldsymbol{\theta})$, which is a random variable that depends on $\boldsymbol{\theta}$. We define $\boldsymbol{\Sigma}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}(\boldsymbol{\theta})}[\boldsymbol{z}\boldsymbol{z}^\top]$ as the noise covariance matrix of gradients at $\boldsymbol{\theta}$. In this work, we make some regularity assumptions on the loss function and the gradient noise distribution. We begin with a smoothness condition on the loss function $\mathcal{L}$.

Assumption 3.1. *The loss function $\mathcal{L}$ and the matrix square root of the noise covariance $\boldsymbol{\Sigma}^{1/2}$ are $\mathcal{C}^5$-smooth on $\mathbb{R}^d$, i.e. all their partial derivatives up to order 5 exist and are continuous.*

Assuming smoothness on the loss function is a common practice in optimization analysis. Here, we specifically assume the $\mathcal{C}^5$-smoothness, which we found to be a minimal smoothness requirement for our proof to hold for all $\mathcal{C}^3$-smooth test functions in Theorem 4.1.

Moreover, we assume that the smoothness constant of $\mathcal{L}$ and the gradient noise are globally bounded:

Assumption 3.2. *$\mathcal{L}$ is ρ-smooth on $\mathbb{R}^d$, i.e. $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, $\|\nabla\mathcal{L}(\boldsymbol{\theta}_1) - \nabla\mathcal{L}(\boldsymbol{\theta}_2)\|_2 \leq \rho \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ and $\mathcal{L}$ is bounded from below, i.e. $\mathcal{L}^* = inf_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) > -\infty$.*

Assumption 3.3. *The noisy gradients are ℓ_2-bounded, i.e., there exists some constant R s.t. $\forall \boldsymbol{\theta} \in \mathbb{R}^d$, $\|\nabla\ell(\boldsymbol{\theta}; \xi)\|_2 \leq R$ almost surely for training data sample $\xi \sim \mathcal{S}_{\text{train}}$.*

SGD and Adam. SGD is an iterative method that starts from an initial point $\boldsymbol{\theta}_0$ and updates the parameters as $\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k - \eta\nabla\ell_k(\boldsymbol{\theta}_k)$ for all $k \geq 0$, where η is the learning rate, $\ell_k(\boldsymbol{\theta})$ is the loss function for the data sample ξ_k sampled at step k. Adam (Kingma and Ba, 2014) is a popular optimizer that updates the parameters as:

$$\begin{aligned}
\boldsymbol{m}_{k+1} &:= \beta_1 \boldsymbol{m}_k + (1-\beta_1)\nabla \ell_k(\boldsymbol{\theta}_k) \\
\boldsymbol{v}_{k+1} &:= \beta_2 \boldsymbol{v}_k + (1-\beta_2)\nabla \ell_k(\boldsymbol{\theta}_k)^{\odot 2} \\
\theta_{k+1,i} &:= \theta_{k,i} - \eta \frac{m_{k+1,i}}{\sqrt{v_{k+1,i}} + \epsilon} \quad \text{for all } i \in [d].
\end{aligned}$$

Note that in practice, it is common to normalize $\boldsymbol{m}_{k+1}$ and $\boldsymbol{v}_{k+1}$ by $1-\beta_1^{k+1}$ and $1-\beta_2^{k+1}$ respectively before the division. However, this normalization quickly becomes neglectable when k is large, so we ignore it for simplicity.

SDE First-Order Approximation For SGD. A *Stochastic Differential Equation* (SDE) is an extension of an ordinary differential equation that incorporates random perturbations, and is widely used to model systems under the influence of noise. An SDE on $\mathbb{R}^d$ takes the form $\mathrm{d}\boldsymbol{\theta}_t = b(\boldsymbol{\theta}_t)\mathrm{d}t + \sigma(\boldsymbol{\theta}_t)\mathrm{d}\boldsymbol{W}_t$ where $b : \mathbb{R}^d \to \mathbb{R}^d$ is the drift vector field, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ is the diffusion matrix, and $\{\boldsymbol{W}_t\}_{t\geq 0}$ is an m-dimensional Wiener process. A line of works (Li et al., 2015; Jastrzębski et al., 2017; Li et al., 2017; Smith et al., 2020; Li et al., 2019, 2021a) used the following SDE to serve as a first-order approximation of SGD, which we refer to as the *conventional SDE*:

$$\mathrm{d}\boldsymbol{\theta}_t = -\nabla\mathcal{L}(\boldsymbol{\theta}_t)\mathrm{d}t + \sqrt{\eta}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}_t)\mathrm{d}\boldsymbol{W}_t,$$

where the stochastic integral is taken in the Itô sense. For an introduction to Itô calculus, see Oksendal (2013). Later, Malladi et al. (2024) extended this type of SDE to Adam. Besides these conventional SDEs, below we introduce another type of SDE, slow SDE, that can more explicitly capture the implicit bias of SGD near a manifold of minimizers.

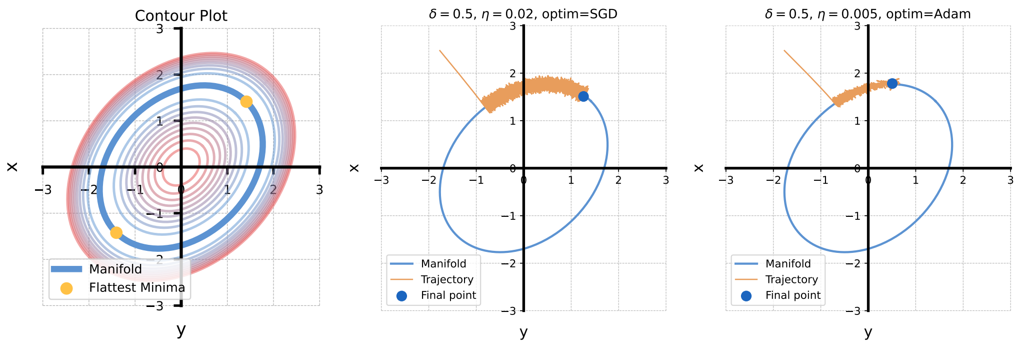
Manifold Assumption. Before going into the slow SDE, we introduce the *manifold assumption*. Previous studies (Garipov et al., 2018; Kuditipudi et al., 2019) have found that low-loss solutions are in fact connected to each other, a phenomenon known as mode connectivity. Wen et al. (2024) provided empirical evidence that the training dynamics of language model training usually happen in a structure similar to a river valley, where many low-loss solutions lie in the bottom of the valley. Motivated by these observations, many previous works (Li et al., 2021b; Fehrman et al., 2020; Lyu and Li, 2020; Gu et al., 2023a) assumed that the minimizers of the training loss function are not isolated points but connected and form a manifold Γ:

Assumption 3.4. *Γ is $\mathcal{C}^\infty$-smooth, $(d-m)$-dimensional compact submanifold of $\mathbb{R}^d$, where any $\zeta \in \Gamma$ is a local minimizer of $\mathcal{L}$. For all $\boldsymbol{\zeta} \in \Gamma$, rank$(\nabla^2\mathcal{L}(\boldsymbol{\zeta})) = m$. Additionally, there exists an open neighborhood of Γ, denoted as U, such that $\Gamma = \arg\min_{\boldsymbol{\theta}\in U} \mathcal{L}(\boldsymbol{\theta})$.*

With this assumption, if an optimization process converges and the learning rate η is sufficiently small, then the process will be trapped near some minimizer manifold which we denote by Γ.

Slow SDE. A line of works (Blanc et al., 2020; Damian et al., 2021; Li et al., 2021b) studied the dynamics of SGD near the manifold Γ and showed that SGD has an implicit bias towards flatter minimizers on Γ. This effect cannot be directly seen from conventional SDEs, so Li et al. (2021b) derived a new type of SDE approximation, called slow SDE, that can explicitly capture this effect. See Appendix A for an illustration of the difference between conventional SDEs and slow SDEs. Here we introduce the slow SDE for SGD following the formulation in Gu et al. (2024). For ease of presentation, we define the following projection operators $\Phi, P_{\boldsymbol{\zeta}}$ for points and differential forms respectively. Consider the gradient flow $\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = -\nabla\mathcal{L}(\boldsymbol{x}(t))$ with $\boldsymbol{x}(0) = \boldsymbol{x}$, and fix some point $\boldsymbol{\theta}_{\text{null}} \notin \Gamma$, we define the gradient flow projection of any $\boldsymbol{x}$, $\Phi(\boldsymbol{x})$, as $\lim_{t\to+\infty}\boldsymbol{x}(t)$ if the limit exists and belongs to Γ, and $\boldsymbol{\theta}_{\text{null}}$ otherwise. It can be shown by simple calculus (Li et al., 2021b) that $\partial\Phi(\boldsymbol{\zeta})$ equals the projection matrix onto the tangent space of Γ at $\boldsymbol{\zeta}$. We decompose the noise covariance $\boldsymbol{\Sigma}(\boldsymbol{\zeta})$ for $\boldsymbol{\zeta} \in \Gamma$ into two parts: the noise in the tangent space $\boldsymbol{\Sigma}_{\parallel}(\boldsymbol{\zeta}) := \partial\Phi(\boldsymbol{\zeta})\boldsymbol{\Sigma}(\boldsymbol{\zeta})\partial\Phi(\boldsymbol{\zeta})$ and the noise in the normal space $\boldsymbol{\Sigma}_{\Diamond}(\boldsymbol{\zeta}) := \boldsymbol{\Sigma}(\boldsymbol{\zeta}) - \boldsymbol{\Sigma}_{\parallel}(\boldsymbol{\zeta})$.

For any $\boldsymbol{\zeta} \in \Gamma$, matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$, we use $P_{\boldsymbol{\zeta}}(\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}\mathrm{d}t)$ to denote $\Phi(\boldsymbol{\zeta} + \boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}\mathrm{d}t) - \Phi(\boldsymbol{\zeta})$, which equals $\partial\Phi(\boldsymbol{\zeta})\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \left(\partial\Phi(\boldsymbol{\zeta})\boldsymbol{b} + \frac{1}{2}\partial^2\Phi(\boldsymbol{\zeta})[\boldsymbol{A}\boldsymbol{A}^\top]\right)\mathrm{d}t$ by Itô calculus. $P_{\boldsymbol{\zeta}}$ can be interpreted as projecting an infinitesimal step from $\boldsymbol{\zeta}$, so that $\boldsymbol{\zeta}$ after taking the projected step does not leave the manifold Γ. Now we are ready to state the SDE for Local SGD.



(a) (b) (c)

Figure 1: **(a)**: Coutour of the elliptical loss, from which we can see the two tips as the flattest minima. **(b)**: SGD implicitly minimizes $\mathrm{tr}(\boldsymbol{H})$ and converges to the flattest minima. **(c)**: Adam reduces sharpness too but converges to a different and sparser minimum.

Definition 3.1 (Slow SDE for SGD)**.** *Given $\eta > 0$ and $\boldsymbol{\zeta}_0 \in \Gamma$, define $\boldsymbol{\zeta}(t)$ as the solution of the following SDE with initial condition $\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0$:*

$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}}\Big(\underbrace{\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}_t}_{\text{(a) diffusion}} - \underbrace{\frac{1}{2}\nabla^3\mathcal{L}(\boldsymbol{\zeta})\big[\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})\big]\mathrm{d}t}_{\text{(b) drift}}\Big). \tag{1}$$

Here $\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})$ is defined as $\sum_{i,j:\,\lambda_i \neq 0 \vee \lambda_j \neq 0} \frac{1}{\lambda_i+\lambda_j}\langle \boldsymbol{\Sigma}_{\Diamond}(\boldsymbol{\zeta}),\, \boldsymbol{v}_i\boldsymbol{v}_j^\top\rangle\, \boldsymbol{v}_i\boldsymbol{v}_j^\top$, where $\{\boldsymbol{v}_i\}_{i=1}^d$ is an orthonormal eigenbasis of $\nabla^2\mathcal{L}(\zeta)$ with corresponding eigenvalues $\lambda_1, \dots, \lambda_d$.

Interpretation of the Slow SDE for SGD: Semi-gradient Descent This SDE on the minimizer manifold Γ splits naturally into a *diffusion* term $P_{\boldsymbol{\zeta}}\big(\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta})\,\mathrm{d}\boldsymbol{W}_t\big)$ injecting noise in the tangent space, and a *drift* term $-\frac{1}{2}\,P_{\boldsymbol{\zeta}}\big(\nabla^3\mathcal{L}(\boldsymbol{\zeta})\big[\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})\big]\mathrm{d}t\big)$ that can be seen as the negative *semi-gradient* of the following sharpness measure:

$$\mu(\boldsymbol{\zeta}) := \Big\langle \nabla^2\mathcal{L}(\boldsymbol{\zeta}), \widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})\Big\rangle.$$

Here we use the word "semi-gradient" (Mnih et al., 2015; Brandfonbrener and Bruna, 2019) because it is not exactly the gradient of $\mu(\boldsymbol{\zeta})$ but only the gradient with respect to the first argument of the inner product. More specifically, define $\mu(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) := \Big\langle \nabla^2\mathcal{L}(\boldsymbol{\zeta}_1), \widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta}_2)\Big\rangle$, then the drift term is essentially $-\frac{1}{2}\,\nabla_{\boldsymbol{\zeta}_1}\mu(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2)|_{\boldsymbol{\zeta}_1=\boldsymbol{\zeta},\boldsymbol{\zeta}_2=\boldsymbol{\zeta}}$ after projecting onto the tangent space of Γ at $\boldsymbol{\zeta}$. In other words, SGD near manifold takes semi-gradients to minimize the implicit regularizer $\langle \nabla^2\mathcal{L}(\boldsymbol{\zeta}), \widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})\rangle$ but pretend $\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})$ to be fixed, i.e. ignore the dependency of $\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})$ on $\boldsymbol{\zeta}$.

Example: Noisy Ellipse. We provide a toy example to illustrate the phenomenon described by the slow SDE for SGD: there are two parameters x, y and an elliptical loss with label noise $\mathcal{L}(x, y) = \frac{1}{2}\left(\frac{(x+y)^2}{2a^2} + \frac{(y-x)^2}{2b^2} - 1 - \delta\right)^2$. The label noise δ is sampled uniformly from $\{-0.5, 0.5\}$ at every step. As depicted in Fig. 1, SGD moves towards flatter minimizers after reaching the manifold. The same phenomenon can be observed for Adam, but Adam converges to a different minimizer that is closer to the axis (or, "sparser" in the parameter space). Understanding the difference between SGD and Adam is the main focus of this paper.

4 Theoretical Analysis of Adam

In this section, we generalize the slow SDE for SGD to a general class of adaptive gradient methods (AGMs), including Adam. We first present our novel slow SDE for a general class of AGMs, including Adam, and give an intuitive explanation for our results. Then, we discuss the difficulty of directly applying the slow SDE framework to Adam and other AGMs and how we resolve the problems.

A General Class of Adaptive Gradient Methods. We define a general class of AGMs as follows:

$$\begin{aligned}
\boldsymbol{m}_{k+1} &:= \beta_1\boldsymbol{m}_k + (1-\beta_1)\nabla\ell_k(\boldsymbol{\theta}_k)\\
\boldsymbol{v}_{k+1} &:= \beta_2\boldsymbol{v}_k + (1-\beta_2)V\big(\nabla\ell_k(\boldsymbol{\theta}_k)\nabla\ell_k(\boldsymbol{\theta}_k)^\top\big)\\
\boldsymbol{\theta}_{k+1} &:= \boldsymbol{\theta}_k - \eta S(\boldsymbol{v}_{k+1})\boldsymbol{m}_{k+1}.
\end{aligned}$$

where $S : \mathbb{R}^d_{\geq 0} \longrightarrow \mathbb{S}^d_{++}$ is a ρ_s-smooth function that maps a vector $\boldsymbol{v} \in \mathbb{R}^d_{\geq 0}$ (i.e. with non-negative entries) to a symmetric positive definite matrix $S(\boldsymbol{v}) \in \mathbb{S}^d_{++}$. In addition, we require S to satisfy $S(\boldsymbol{v}) \succeq \frac{1}{R_0}\boldsymbol{I}$ for some $R_0 > 0$ and any $\boldsymbol{v} \in \mathbb{R}^d$. We also require $V : \mathbb{R}^{d\times d} \longrightarrow \mathbb{R}^d$ to be a linear function "with positive semi-definite coefficients", specifically, $V(\boldsymbol{g}\boldsymbol{g}^\top) \in \mathbb{R}^d_{\geq 0}$ for all $\boldsymbol{g} \in \mathbb{R}^d$.

A number of currently used optimization algorithms, such as RMSProp, Adam, Adam-mini, Adafactor[1], Adalayer, AdaSGD, and Shampoo[2], all fit this framework. Note that we do not consider weight decays or bias corrections in these optimizers. Some examples of V and S functions are listed in Table 1, including the AdamE-λ optimizer that will be introduced in Section 5 as a tool to tune the implicit bias of Adam.

Prior to the results, we introduce two technical assumptions: S satisfies a mild smoothness condition, and $1 - \beta_1$ is of constant order.

Assumption 4.1. *The function S is $\mathcal{C}^4$-smooth on $\mathbb{R}^d_{\geq 0}$.*

Assumption 4.2. $\beta_1 \leq 0.9$.

Remark 4.1. *The threshold 0.9 in Assumption 4.2 can also be replaced by any constant below 1, and the approximation rate in our result will remain unaffected. So we actually consider the regime where $b_1 := 1 - \beta_1$ is of constant order. For real-world Natural Language Processing (NLP) models, BERT (Devlin et al., 2019), Transformer (Vaswani et al., 2017) and GPT (Radford et al., 2018) all use $\beta_1 = 0.9$. In computer vision (CV), pix2pix (Isola et al., 2017) uses $\beta_1 = 0.5$, while U-Net (Ronneberger et al., 2015) and ViT (Dosovitskiy et al., 2020) use $\beta_1 = 0.9$. Thus, assuming $\beta_1 \leq 0.9$ is consistent with standard practice across multiple aspects.*

4.1 Slow SDE Analysis for AGMs

Our SDE for AGMs characterizes the training dynamics near the manifold Γ. First we rigorously define the preconditioned projection mapping $\Phi_{\boldsymbol{S}}$ and the SDE projection formula as an extension to the Φ and $\boldsymbol{P}_{\boldsymbol{\zeta}}$ mentioned in Section 3, after which we present the SDE for AGMs we derived.

Definition 4.1 (Preconditioner Flow Projection)**.** *Fix a point $\theta_{null} \notin \Gamma$. Given a Positive Semi-Definite matrix $\boldsymbol{S}$, for $x \in \mathbb{R}^d$, consider the preconditioner flow $\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\boldsymbol{S}\nabla\mathcal{L}(x(t))$ with $x(0) = x$. We denote the preconditioner flow projection of x as $\Phi_{\boldsymbol{S}}(x)$, i.e. $\Phi_{\boldsymbol{S}}(x) := \lim_{t\to+\infty} x(t)$ if the limit exists and belongs to Γ, and $\Phi_{\boldsymbol{S}}(x) = \theta_{null}$ otherwise.*

Definition 4.2. *For any $\boldsymbol{\zeta} \in \Gamma$ and any differential form $\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}\mathrm{d}t$ in Itô calculus, where $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, and $b \in \mathbb{R}^d$. We use $\boldsymbol{P}_{\boldsymbol{\zeta},\boldsymbol{S}}(\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}\mathrm{d}t)$ as a shorthand for the differential form $\partial\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{S}\left(\partial\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})\boldsymbol{b} + \frac{1}{2}\partial^2\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})[\boldsymbol{A}\boldsymbol{A}^top]\right)\mathrm{d}t$.*

Definition 4.3 (Slow SDE for AGMs)**.** *Given learning rate η, $\frac{1-\beta_2}{\eta^2} = c$, $\boldsymbol{v}_0 \in \mathbb{R}^d$, $\boldsymbol{S}_t := S(\boldsymbol{v}(t))$, and $\boldsymbol{\zeta}_0 \in \Gamma$, $\boldsymbol{v}_0 \in \mathbb{R}^d$, we define $\boldsymbol{\zeta}(t)$ as the solution of the following SDE with initial point $(\boldsymbol{\zeta}(0), \boldsymbol{v}(0)) = (\boldsymbol{\zeta}_0, \boldsymbol{v}_0)$:*

$$
\begin{cases}
\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}(t),\boldsymbol{S}(t)}\left(\underbrace{\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta}(t);\boldsymbol{S}(t))\mathrm{d}\boldsymbol{W}_t}_{\textit{diffusion}} \underbrace{-\frac{1}{2}\boldsymbol{S}(t)\nabla^3\mathcal{L}(\boldsymbol{\zeta})\left[\boldsymbol{\Sigma}_{\diamond}(\boldsymbol{\zeta}(t);\boldsymbol{S}(t))\right]\mathrm{d}t}_{\textit{drift}}\right), \\
\mathrm{d}\boldsymbol{v}(t) = \underbrace{c\left(V(\boldsymbol{\Sigma}(\boldsymbol{\zeta})) - \boldsymbol{v}\right)\mathrm{d}t}_{\textit{preconditioner drift}}.
\end{cases}
\tag{2}
$$

$\boldsymbol{\Sigma}_{\diamond}(\boldsymbol{\zeta};\boldsymbol{S}) = \boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\zeta})\boldsymbol{S} - \boldsymbol{\Sigma}_{\parallel}(\boldsymbol{\zeta};\boldsymbol{S})$, $\boldsymbol{\Sigma}_{\parallel}(\boldsymbol{\zeta};\boldsymbol{S}) = \partial\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\zeta})\boldsymbol{S}\partial\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})$.

Note that the drift term in $\mathrm{d}\boldsymbol{\zeta}(t)$ can be interpreted as an *adaptive semi-gradient descent* process, in that this term drives the dynamics towards optimizing an adaptive loss function

$$\mu(\boldsymbol{\zeta}, \boldsymbol{v}) = \langle \nabla^2\mathcal{L}(\boldsymbol{\zeta}), \boldsymbol{\Sigma}_{\diamond}(\boldsymbol{\zeta}(t);\boldsymbol{S}(t))\rangle$$

as if $\boldsymbol{\Sigma}_{\diamond}(\boldsymbol{\zeta}(t);\boldsymbol{S}(t))$ has no dependence on $\boldsymbol{\zeta}$; also this gradient flow is preconditioned by a positive definite matrix $\boldsymbol{S}(t)$. Recall that the drift term in the slow SDE for SGD can be seen as a semi-gradient descent. In the AGM framework, it takes $\Theta(\eta^{-2})$ time for the preconditioner $\boldsymbol{S}(t)$ to make

[1]We ignore update clipping, i.e. we adopt the Algorithm 2 in Shazeer and Stern (2018).

[2]In practice, the Shampoo optimizer is often equipped with the exponential moving average (EMA) on the calculation of pre-conditioner (Morwani et al., 2024). Here we adopt this practical version of Shampoo instead of the original one (Gupta et al., 2018).

Table 1: Examples of optimizers in the AGM Framework.

Optimizer	Functions V/S	Regularizer under Label Noise with $\epsilon = 0$[3]	Remarks
SGD	**V:** $V(\boldsymbol{M}) = \mathbf{1_d}$ **S:** $S(\boldsymbol{v}) = \boldsymbol{I}_d$	tr$(\boldsymbol{H})$ (Blanc et al., 2020)	
Adam	**V:** $V(\boldsymbol{M}) = \mathrm{diag}(\boldsymbol{M})$ **S:** $S(\boldsymbol{v}) = \mathrm{Diag}\big(1/(\sqrt{\boldsymbol{v}} + \epsilon)\big)$	tr$\Big(\mathrm{Diag}(\boldsymbol{H})^{1/2}\Big)$	
Rmsprop	**V:** $V(\boldsymbol{M}) = \mathrm{diag}(\boldsymbol{M})$ **S:** $S(\boldsymbol{v}) = \mathrm{Diag}\big(1/(\sqrt{\boldsymbol{v}} + \epsilon)\big)$	tr$\Big(\mathrm{Diag}(\boldsymbol{H})^{1/2}\Big)$	
Adam-mini	**V:** $V(\boldsymbol{M})_i = \frac{1}{\lvert B_{\pi(i)}\rvert} \sum_{j \in B_{\pi(i)}} M_{jj}$ **S:** $S(\boldsymbol{v}) = \mathrm{Diag}\big(1/(\sqrt{\boldsymbol{v}} + \epsilon)\big)$	$\sum_{i\in[N]} \sqrt{\lvert B_i\rvert \cdot \mathrm{tr}\boldsymbol{H}_{B_i}}$	Blocks $\{B_1, B_2, \cdots, B_N\}$ partition $[d]$, and $i \in B_{\pi(i)}$.
Adalayer	**V:** $V(\boldsymbol{M})_i = \frac{1}{\lvert L_{\pi(i)}\rvert} \sum_{j \in L_{\pi(i)}} M_{jj}$ **S:** $S(\boldsymbol{v}) = \mathrm{Diag}\big(1/(\sqrt{\boldsymbol{v}} + \epsilon)\big)$	$\sum_{i\in[N]} \sqrt{\lvert L_i\rvert \cdot \mathrm{tr}\boldsymbol{H}_{L_i}}$	Layers $\{B_1, B_2, \cdots, B_N\}$ partition $[d]$, and $i \in L_{\pi(i)}$.
AdamE-λ	**V:** $V(\boldsymbol{M}) = \mathrm{diag}(\boldsymbol{M})$ **S:** $S(\boldsymbol{v}) = \mathrm{Diag}\big(1/(\boldsymbol{v}^{\odot\lambda} + \epsilon)\big)$	tr$\Big(\mathrm{Diag}(\boldsymbol{H})^{1-\lambda}\Big)$	
Shampoo[4]	**V:** $V(\boldsymbol{M}) = (V_L(\boldsymbol{M}),\, V_R(\boldsymbol{M}))$ $[V_L(\boldsymbol{M})]_{i,j} = \sum_k M_{i,k,k.j}$, $[V_R(\boldsymbol{M})]_{i,j} = \sum_k M_{k,i,j,k}$ **S:** $S(\boldsymbol{V}_L, \boldsymbol{V}_R)\left(\boldsymbol{V}_R^\top \otimes \boldsymbol{V}_L\right)^{-1/2}$	No explicit form	Details for vector version of shampoo and discussion of regularizer in Appendix J.

a significant (i.e. $\Theta(1)$) change, which coincides with the moving speed of the slow SDE of $\boldsymbol{\zeta}$. Therefore, compared to that of SGD, our SDE includes a new formula that tracks the motion of the preconditioner and injects adaptiveness accordingly in the semi-gradient descent process.

We prove that $\boldsymbol{\zeta}(t)$ always stays on the manifold Γ. And next, we present our main theorem showing that the above SDE in Equation (2) tracks the trajectory of Adam in a weak approximation sense.

Theorem 4.1. *Under Assumption 3.1–4.2, let $T > 0$ be a constant and let $\boldsymbol{X}(t) = (\boldsymbol{\zeta}(t), \boldsymbol{v}(t))$ be the solution to Equation* (2) *with initial condition:*

$$\boldsymbol{\zeta}(0) = \Phi\big(\boldsymbol{\theta}_0\big) \in \Gamma, \quad \boldsymbol{v}(0) = \boldsymbol{v}_0 \in \mathbb{R}^d,$$

and we define the projected state of Adam at time t as $\bar{\boldsymbol{X}}_t := (\Phi_{\boldsymbol{S}_t}(\boldsymbol{\theta}_t), \boldsymbol{v}_t)$. For any $\mathcal{C}^3$-smooth function $g(\boldsymbol{\theta})$,

$$\max_{0 \le t \le \frac{T}{\eta^2}} \Big| \mathbb{E}\big[g\big(\bar{\boldsymbol{X}}_t\big)\big] - \mathbb{E}\big[g\big(\boldsymbol{X}(t\eta^2)\big)\big] \Big| = \widetilde{\mathcal{O}}\big(\eta^{0.25}\big),$$

where $\widetilde{O}(\cdot)$ hides logarithmic factors and constants independent of η but may depend on $g(\boldsymbol{\theta})$.

Theorem 4.1 shows that with a small η, once Adam approaches the minimizer manifold, its long-horizon behavior within $\widetilde{\mathcal{O}}(\eta^{-2})$ steps is captured by the SDE in Equation (2).

4.2 Interpretation of The Slow SDEs for AGMs

Adaptive Projection Operator. Equation (1) employs a fixed projection operator $P_{\boldsymbol{\zeta}}$ to constrain the SDE to the manifold. As a comparison, the slow SDE for AGM uses an adaptive projection $P_{\boldsymbol{\zeta}, \boldsymbol{S}(t)}$ that depends on the current preconditioner $S(\boldsymbol{v}(t))$. In other words, SGD's projection is state-independent, but AGM's projection is state-dependent. This adaptive projection alters the way the stochastic trajectory evolves on the manifold, giving rise to a different implicit bias in AGMs versus SGD.

Effect of the Preconditioner on the Gradient Noise Covariance. Near the manifold, as the gradient of loss vanishes ($\nabla\mathcal{L}(\boldsymbol{\theta}) \to 0$), SGD's wandering around becomes noise-driven. For AGMs, the situation is more subtle.

First, one can show that the momentum term does not affect the implicit bias, consistent with prior theory (Wang et al., 2023). The reason why β_1 does not affect the implicit bias is that, after the iteration approaches the manifold, the difference between the current gradient g_t and momentum M_t becomes negligible in expectation.

Second, the AGM trajectory is influenced by its preconditioner. Concretely, the gradient-noise covariance matrix $\boldsymbol{\Sigma}$ is filtered through the preconditioner $\boldsymbol{S}(t)$ into $\boldsymbol{S}(t)\boldsymbol{\Sigma}\boldsymbol{S}(t)$ and then contributes to the SDE. Over a long time horizon, this modified noise term alters the deterministic drift direction, further distinguishing AGM's dynamics from those of vanilla SGD.

[3]The derivation of regularizers for each optimizer is discussed in Appendix I.

[4]Since the shampoo optimizer is designed to optimize matrix parameters such as $\Theta \in \mathbb{R}^{m\times n}$, here we slightly generalize the notation in the AGM framework, which is originally proposed for parameters with vector type. For details, see Appendix J.

4.3 Technical Difficulties and Proof Insights

4.3.1 Convergence Guarantee of AGMs

The core of our study is to consider the behavior of Adam's implicit bias around the minimizer manifold. However, to make our study self-contained, we first need to show that Adam can actually converge to the neighborhood of the minimizer manifold, which itself is already non-trivial. Unfortunately, without any constraint, Adam cannot provably converge to the minimizer manifold. In fact, the convergence issue of Adam has been debated from its birth. Reddi et al. (2018) show that Adam does not converge to the optimal solution even in some simple convex settings. Recent work (Dereich and Jentzen, 2024) gives Adam's ODE and shows that this ODE does not necessarily converge to the absorbing point of the gradient flow. So we present a statement of AGMs' convergence as a preparation for our subsequent study into Adam's behavior near the manifold. It is worth noting that, not only for Adam, the convergence holds for all AGMs under our framework.

Theorem 4.2 (Convergence Bound of AGMs, Stated Informally)**.** *Let Assumptions 3.2, 3.3 and 4.2 hold, and $\mathcal{L}$ satisfy the μ-PL condition. With a small learning rate η, it holds with high probability for some $K = \mathcal{O}(\frac{1}{\eta}\log\frac{1}{\eta})$ that $\mathcal{L}(\boldsymbol{\theta}_K) - \mathcal{L}^* = \tilde{\mathcal{O}}(\eta)$. See Theorem D.2 for a formal statement.*

4.3.2 Key Insights in the Derivation of Slow SDEs for AGMs

After the AGMs reach the neighborhood of the minimizer manifold, we can derive an analysis similar to the one in the local SGD paper (Gu et al., 2023a). Specifically, we use SDEs to approximate the AGMs after they reach the manifold neighborhood. However, unlike the usual SDE approximation, the SDEs we use here can track the AGMs for a much longer period of time, up to $\tilde{\mathcal{O}}(\eta^{-2})$ rather than the $\tilde{\mathcal{O}}(\eta^{-1})$, which is more common in the previous papers. This type of SDE is termed "*slow SDE*" by Gu et al. (2023a).

There are two obstacles preventing us from directly applying the analysis of slow SDEs from SGDs to AGMs. First, the obtaining of slow SDEs requires an accurate calculation of the variation of the first-order and second-order moments of the parameters over a relatively large number of steps (a "*giant step*" in the notation of Gu et al. (2023a)), and in the case of SGD, due to the nature of its rotational equivariance, we can always consider its Hessian matrix as a diagonal array, as well as its corresponding minimizer manifold as a space extended by some full-space standard bases, which greatly simplifies the computation. However, it is not the case for AGMs. Due to the effect of preconditioners $S(\boldsymbol{v}_k)$, the rotation equivariance is not satisfied here.

To resolve this, we generalize the gradient flow projection in Gu et al. (2023a); Li et al. (2021b) into a varying preconditioner flow projection. Based on this definition, reparameterizing to the original space lets us reuse the simple formulas employed previously (Gu et al., 2023a; Li et al., 2021b).

The second reason is that when β_2 is far from 1, the preconditioner changes too quickly, making the evolution of the moments hard to characterize. Conversely, when β_2 is extremely close to 1, the preconditioner changes so little as to be impractical. Accordingly, we focus on the regime $1 - \beta_2 = \mathcal{O}(\eta^2)$, which we call the "2-*scheme*." The key point is that this regime does not make the preconditioner's evolution negligible; rather, its slow but nontrivial drift shapes the SDE and can be tracked analytically.

5 Adam's Provable Generalization Benefit with Label Noise

In this section, we prove that with the label noise condition, the implicit regularizer of Adam reduces to a simpler form that aligns better with sparsity regularizations, and then verify experimentally.

5.1 Reduction of Adam's Implicit Regularizer with Label Noise

Label Noise. By *label noise* we refer to the condition that for all $\boldsymbol{\theta} \in \Gamma$, the covariance matrix $\boldsymbol{\Sigma}$ is a constant multiple of the Hessian: $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \alpha\,\nabla^2\mathcal{L}(\boldsymbol{\theta})$ for some constant α (Blanc et al., 2020). This condition was initially derived from the setting of adding an i.i.d. perturbation to the true label in each training step in a regression problem, but the term *label noise* can refer to any setting satisfying the proportionality condition. This proportionality greatly simplifies the analysis and has been widely used to study the implicit bias of SGD and related optimizers (Blanc et al., 2020; Damian et al., 2021; Li et al., 2021b; Gu et al., 2023a). (Li et al., 2021b) proved that Slow SDE for SGD reduces to an ODE with label noise, and we now show the same thing for AGMs:

Theorem 5.1 (Slow ODE for AGMs with Label Noise)**.** *Under the label noise condition, the Slow SDE for AGMs* (2) *becomes the following ODE:*

$$\begin{cases} \mathrm{d}\boldsymbol{\zeta}(t) = -\frac{\alpha}{2}\boldsymbol{S}_t\partial\Phi_{\boldsymbol{S}_t}(\boldsymbol{\zeta})\boldsymbol{S}_t\partial^2(\nabla\mathcal{L})(\boldsymbol{\zeta})[\boldsymbol{S}_t]\mathrm{d}t, \\ \mathrm{d}\boldsymbol{v}(t) = c\left(V(\boldsymbol{\Sigma}(\boldsymbol{\zeta})) - \boldsymbol{v}\right)\mathrm{d}t, \end{cases} \tag{3}$$

where $\boldsymbol{S}_t := \boldsymbol{S}(V_t)$*.*

See Appendix H for the proof. We then derive the implicit bias of Adam with label noise.

Lemma 5.1 (Adam's Implicit Bias with Label Noise)**.** *Let* $\lambda \in [0,1)$*. With the label noise condition and* $\epsilon = 0$*, the fixed point of* (3) *in the Adam case satisfies* $\nabla_\Gamma \mathit{tr}\left(\mathit{Diag}(\boldsymbol{H})^{1/2}\right) = 0$*.*

Proof Sketch. The fixed point of (3) satisfies $\boldsymbol{v} = V(\boldsymbol{\Sigma}(\boldsymbol{\zeta}))$ and $\nabla^3\mathcal{L}(\boldsymbol{\zeta})[S(\boldsymbol{v})] = 0$, since $S(\boldsymbol{v})$ is invertible. Let $\boldsymbol{H} = \nabla^2\mathcal{L}(\boldsymbol{\zeta}) = \boldsymbol{\Sigma}(\boldsymbol{\zeta})/\alpha$. In the Adam case, $\boldsymbol{v} = \mathrm{diag}(\boldsymbol{\Sigma}) = \alpha\cdot\mathrm{diag}(\boldsymbol{H})$, and $S(\boldsymbol{v}) = \mathrm{Diag}(1/(\sqrt{\boldsymbol{v}}+\epsilon))$. Integrating by parts yields $\nabla^3\mathcal{L}(\boldsymbol{\zeta})\left[S(\boldsymbol{v})\right] = \nabla\left[\langle \boldsymbol{H}, S(\boldsymbol{v})\rangle\right] - \nabla\left(S(\boldsymbol{v})\right)[\boldsymbol{H}]$. Then a straightforward simplification gives the result. Refer to Appendix H for the more detailed calculation. □

A Simple Way to Tune Adam's Implicit Bias: AdamE. The proof of Lemma 5.1 inspired the following simple variant of Adam: We define *AdamE* as an optimizer class that, is identical to Adam except that $S(\boldsymbol{v}) = \mathrm{Diag}(1/(\boldsymbol{v}^{\odot\lambda}+\epsilon))$ for a tunable parameter $\lambda \in [0,1)$. For a specified λ_0 we also use the term *AdamE with* $\lambda = \lambda_0$, or simply *AdamE-*λ_0. Note that AdamE with $\lambda = \frac{1}{2}$ coincides with Adam, and that all AdamE optimizers lie within the AGM framework. Applying the same method as in Lemma 5.1 yields the implicit bias of AdamE under label noise; the result is stated below.

Lemma 5.2 (AdamE's Implicit Bias with Label Noise)**.** *Let* $\lambda \in [0,1)$*. With the label noise condition and* $\epsilon = 0$*, the fixed point of* (3) *in the AdamE-*λ *case satisfies* $\nabla_\Gamma \mathit{tr}\left(\mathit{Diag}(\boldsymbol{H})^{1-\lambda}\right) = 0$*.*

Lemma 5.2 indicates that tuning the exponent of the second-order moment in Adam exactly results in tuning the exponent of $\mathrm{diag}(\nabla^2\mathcal{L}(\boldsymbol{\zeta}))$ in the implicit bias. When $\lambda = 0$, the implicit bias reduces to that of SGD, and AdamE also gets rid of the effect of second-order moments and reduces to SGD with momentum, which coincides perfectly. Next, we relate the implicit bias to sparsity and compare the performance of Adam, AdamE, and SGD in a simple experimental setup.

5.2 Example: Sparse Linear Regression with Diagonal Net

In this section, we adopt the *diagonal linear network* (diagonal net) setting proposed by Woodworth et al. (2020) as an experimental setting, which is also used by Li et al. (2021b) to study the implicit bias of SGD.

Setting (Diagonal Net with Label Noise): Let $\boldsymbol{w}^* \in \mathbb{R}^d$ be an unknown κ-sparse ground truth vector. Let $\{(\boldsymbol{z}_i, y_i)\}_{i\in[n]}$ be the training dataset where each $\boldsymbol{z}_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\,\{\pm 1\}^d$, and each y_i is generated by $\langle \boldsymbol{z}_i, \boldsymbol{w}^*\rangle$. Our parameter is defined as $\boldsymbol{\theta} = \binom{\boldsymbol{u}}{\boldsymbol{v}} \in \mathbb{R}^{2d}$. For any function g defined on $\mathbb{R}^{2d}$, we write $g(\boldsymbol{\theta})$ and $g(\boldsymbol{u},\boldsymbol{v})$ interchangeably. The loss function is defined as:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i(\boldsymbol{\theta}), \quad \text{where } \mathcal{L}_i(\boldsymbol{\theta}) = \frac{1}{2}\left(\left\langle \boldsymbol{z}_i, \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}\right\rangle - y_i\right)^2$$

where a label noise is added to the true label y during training. This setting can be viewed as using estimation $\widehat{\boldsymbol{w}} = \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}$ to approximate the ground truth vector $\boldsymbol{w}^*$ of a linear regression task. Note that $d \gg n$ here so the model is highly overparameterized: Theoretically, Li et al. (2021b) proved that $n = \mathcal{O}(\kappa \ln d)$ is enough for SGD to recover ground truth, and we will later show experimentally that less than 1000 training pairs is required for both Adam and SGD to achieve a low test loss when $d = 10000$. The manifold is defined as wherever zero train loss is achieved, i.e. $\Gamma = \left\{\boldsymbol{\theta} \middle| \langle \boldsymbol{z}_i, \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}\rangle = y_i, \forall i \in [n]\right\}$.

This setting allows us to relate the implicit bias directly to the sparsity of the estimated ground truth.

Lemma 5.3. *Let* $\boldsymbol{\theta}^*$ *be an optimal parameter minimizing the loss function* $\mathcal{L}$*, i.e.* $\boldsymbol{\theta}^* \in \Gamma$*. For each* $\boldsymbol{\theta} = \binom{\boldsymbol{u}}{\boldsymbol{v}} \in \Gamma$*, denote* $\widehat{\boldsymbol{w}} := \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}$ *and* $\boldsymbol{H} := \nabla^2\mathcal{L}(\boldsymbol{\theta})$*. We have the following:*

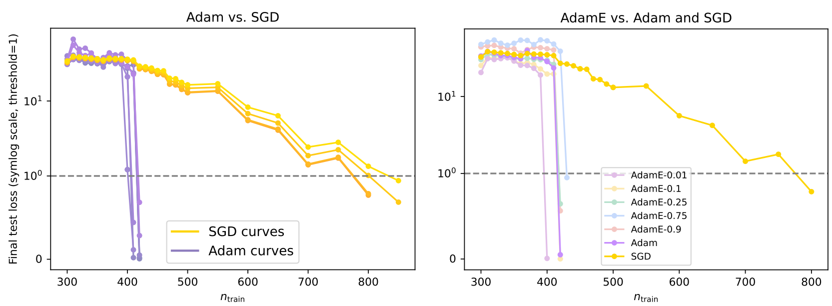
- *If* $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\Gamma} \mathit{tr}(\mathit{Diag}(\boldsymbol{H})^{0.5})$*, then we also have* $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\Gamma} \|\widehat{\boldsymbol{w}}\|_{0.5}$*.*
- *Furthermore, for any* $e_0 \in (0,1]$*, if* $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\Gamma} \mathit{tr}(\mathit{Diag}(\boldsymbol{H})^{e_0})$*, then we also have* $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\Gamma} \|\widehat{\boldsymbol{w}}\|_{e_0}$*.*

The main idea of the proof is that the training loss depends only on the combined quantity $\widehat{\boldsymbol{w}} = \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}$. Hence, if for some index i both u_i and v_i are nonzero, we can reduce the magnitudes of u_i and v_i while keeping $u_i^2 - v_i^2$ fixed, obtaining another minimizer with strictly smaller $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{e_0})$. Therefore, at any optimum we must have $u_i = 0$ or $v_i = 0$ for every i. Under this condition, $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{e_0})$ can be identified with $\|\widehat{\boldsymbol{w}}\|_{e_0}$. We provide the detailed derivation in Appendix H. $\square$

Lemma 5.3 gives the following insights: Implicitly regularizing $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{e_0})$ is equivalent to regularizing the ℓ_{e_0}-norm of the estimated ground truth $\widehat{\boldsymbol{w}} = \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}$: Adam corresponds to $\ell_{0.5}$, SGD to ℓ_1, and AdamE-λ to $\ell_{1-\lambda}$. Just as lasso (ℓ_1) is preferable to ridge (ℓ_2) for sparse ground-truth recovery, we therefore expect Adam and AdamE (with $\lambda > 0$) to recover sparse ground truth more efficiently than SGD. We verify this prediction below.

5.2.1 Result: Adam's Implicit Regularizer Facilitates Sparse Ground-truth Recovery

Fig. 2 shows the results of the experiment. We gradually increase the number of training points and train Adam, SGD, and AdamE under several configurations until convergence. We consider a configuration to have *recovered the ground truth* if the test loss falls below 1. As illustrated in Fig. 2a, Adam's test loss plunges towards zero at approximately $n_{\mathrm{train}} = 420$, whereas SGD's test loss decreases more gradually as the training set grows. To interpolate between different implicit biases, we evaluate AdamE for several values of λ. Fig. 2b indicates that AdamE, even with a small positive value of λ, exhibits the same sudden recovery behavior as Adam.



(a) (b)

Figure 2: Final test loss as a function of the training data size with $d = 10000$, $\kappa = 50$. Each plotted point is the final test loss after both the training and test losses have converged; its x-coordinate is the training data size and the curve denotes the optimizer and configuration. **(a)** Loss comparison between SGD with different learning rates, and Adam with different learning rates and β_2 values. **(b)** Loss comparison between AdamE with $\lambda = 0.01, 0.1, 0.25, 0.75, 0.9$, Adam, and SGD.

Takeaway. Adam's unique implicit bias aligns better with the fundamental target of reducing the sparsity of the model's output, which facilitates the recovery of the sparse ground truth compared to SGD, and this improvement mainly arises from the fact that Adam takes the second order moment into consideration. Starting from SGD, even if we introduce the second-order moment in the preconditioner for a little bit, it could result in significant assistance in sparse ground truth recovery.

However, we should also keep in mind that a clear interpretation of Adam's unique implicit bias, $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{1/2})$ relies heavily on the condition that $\boldsymbol{H}$ is diagonal. Only with this condition can we claim Adam as minimizing $\|\boldsymbol{h}\|_{0.5}$ instead of SGD's $\|\boldsymbol{h}\|_1$ where $\boldsymbol{h}$ is the vector consisting of all eigenvalues of $\boldsymbol{H}$. In other words, Adam's optimization on the implicit bias upon SGD only makes sense when $\boldsymbol{H}$ is diagonal. In the diagonal net setting this is indeed the case in expectation, but we will see in the next chapter that Adam's unique implicit bias may even lead to worse generalization when $\boldsymbol{H}$ is no longer diagonal.

6 Conclusions

We show that Adam implicitly minimizes the sharpness measure $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{1/2})$, leading to solutions and generalization behavior distinct from SGD. Our slow SDE framework rigorously captures Adam's adaptive semi-gradient drift near the minimizer manifold and recovers explicit separations in sparse linear regression and deep matrix factorization. Open directions include extending analysis beyond the "2-scheme" regime ($1 - \beta_2 = O(\eta^2)$) to intermediate regimes such as 1.5-scheme, characterizing Adam's implicit bias once iterates exit the local manifold neighborhood, and incorporating weight-decay (e.g., AdamW) to understand its effect on the effective sharpness regularizer.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions are accurately summarized in abstract and introduction, with no overclaims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed not to penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the proof for each theoretical result in either main part or the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we provide details and concrete settings of our experiments in the experiment sections.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide code and data for reproducing our experiments in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the optimizers and main settings in the experiment section, and other details in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report error bars across multiple random seeds for all main experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiments are tiny and can finish on CPU within an hour.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and complied with the NeurIPS ethics review guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the creators of all external datasets and assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

References

Anil, R., Gupta, V., Koren, T., Regan, K., and Singer, Y. (2020). Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*.

Ashish, V. (2017). Attention is all you need. *Advances in neural information processing systems*, 30:I.

Barrett, D. G. and Dherin, B. (2020). Implicit gradient regularization. *arXiv preprint arXiv:2009.11162*.

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

Brandfonbrener, D. and Bruna, J. (2019). Geometric insights into the convergence of nonlinear td learning. *arXiv preprint arXiv:1905.12185*.

Cattaneo, M. D., Klusowski, J. M., and Shigida, B. (2024). On the implicit bias of adam.

Cattaneo, M. D. and Shigida, B. (2025). How memory in optimization algorithms implicitly modifies the loss.

Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems*, 32.

Damian, A., Ma, T., and Lee, J. D. (2021). Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461.

Défossez, A., Bottou, L., Bach, F., and Usunier, N. (2020). A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*.

Dereich, S. and Jentzen, A. (2024). Convergence rates for the adam optimizer. *arXiv preprint arXiv:2407.21078*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *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. Association for Computational Linguistics.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, A. and Duan, J. (2006). Invariant manifold reduction for stochastic dynamical systems. *arXiv preprint math/0607366*.

Duistermaat, J. J. and Kolk, J. A. (2012). *Lie groups*. Springer Science & Business Media.

Falconer, K. (1983). Differentiation of the limit mapping in a dynamical system. *Journal of the London Mathematical Society*, 2(2):356–372.

Fehrman, B., Gess, B., and Jentzen, A. (2020). Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21(136):1–48.

Filipović, D. (2000). Invariant manifolds for weak solutions to stochastic equations. *Probability theory and related fields*, 118(3):323–341.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. (2021). Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of dnns.

Gatmiry, K., Li, Z., Ma, T., Reddi, S., Jegelka, S., and Chuang, C.-Y. (2023). What is the inductive bias of flatness regularization? a study of deep matrix factorization models. *Advances in Neural Information Processing Systems*, 36:28040–28052.

Gu, X., Lyu, K., Arora, S., Zhang, J., and Huang, L. (2024). A quadratic synchronization rule for distributed deep learning. In *The Twelfth International Conference on Learning Representations*.

Gu, X., Lyu, K., Huang, L., and Arora, S. (2023a). Why (and when) does local sgd generalize better than sgd? *arXiv preprint arXiv:2303.01215*.

Gu, X., Lyu, K., Huang, L., and Arora, S. (2023b). Why (and when) does local sgd generalize better than sgd?

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. (2025). Unified convergence analysis for adaptive optimization with moving average estimator.

Gupta, V., Koren, T., and Singer, Y. (2018). Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR.

HaoChen, J. Z., Wei, C., Lee, J., and Ma, T. (2021). Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357. PMLR.

Hinton, G., Srivastava, N., and Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2.

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

Hong, Y. and Lin, J. (2023). High probability convergence of adam under unbounded gradients and affine variance noise.

Iiduka, H. (2022). Theoretical analysis of adam using hyperparameters close to one without lipschitz smoothness.

Isola, P., Zhu, J., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134. IEEE.

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

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. (2017). Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*.

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

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. (2020). Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kuditipudi, R., Wang, X., Lee, H., Zhang, Y., Li, Z., Hu, W., Ge, R., and Arora, S. (2019). Explaining landscape connectivity of low-cost solutions for multilayer nets. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lee, J. (2012). *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer New York.

Li, Q., Tai, C., and E, W. (2018). Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations.

Li, Q., Tai, C., et al. (2015). Dynamics of stochastic gradient algorithms. *CoRR*.

Li, Q., Tai, C., et al. (2017). Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR.

Li, Q., Tai, C., et al. (2019). Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47.

Li, Z., Malladi, S., and Arora, S. (2021a). On the validity of modeling sgd with stochastic differential equations (sdes). In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12712–12725. Curran Associates, Inc.

Li, Z., Wang, T., and Arora, S. (2021b). What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*.

Lin, T., Stich, S. U., Patel, K. K., and Jaggi, M. (2018). Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*.

Lin, W., Lowe, S. C., Dangel, F., Eschenhagen, R., Xu, Z., and Grosse, R. B. (2025). Understanding and improving the shampoo optimizer via kullback-leibler minimization. *arXiv preprint arXiv:2509.03378*.

Liu, H., Xie, S. M., Li, Z., and Ma, T. (2023). Same pre-training loss, better downstream: Implicit bias matters for language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22188–22214. PMLR.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lyu, K., Jin, J., Li, Z., Du, S. S., Lee, J. D., and Hu, W. (2024). Dichotomy of early and late phase implicit biases can provably induce grokking. In *12th International Conference on Learning Representations, ICLR 2024*.

Lyu, K. and Li, J. (2020). Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*.

Lyu, K., Li, Z., and Arora, S. (2022). Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35:34689–34708.

Ma, C., Kunin, D., and Ying, L. (2023). A quasistatic derivation of optimization algorithms' exploration on minima manifolds.

Malladi, S., Lyu, K., Panigrahi, A., and Arora, S. (2022). On the sdes and scaling rules for adaptive gradient algorithms. *Advances in Neural Information Processing Systems*, 35:7697–7711.

Malladi, S., Lyu, K., Panigrahi, A., and Arora, S. (2024). On the sdes and scaling rules for adaptive gradient algorithms.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Morwani, D., Shapira, I., Vyas, N., Malach, E., Kakade, S., and Janson, L. (2024). A new perspective on shampoo's preconditioner. *arXiv preprint arXiv:2406.17748*.

Nacson, M. S., Gunasekar, S., Lee, J., Srebro, N., and Soudry, D. (2019). Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4683–4692. PMLR.

Neyshabur, B., Tomioka, R., and Srebro, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*.

Oksendal, B. (2013). *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media.

Qian, Q. and Qian, X. (2019). The implicit bias of adagrad on separable data. *Advances in Neural Information Processing Systems*, 32.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Reddi, S. J., Kale, S., and Kumar, S. (2018). On the convergence of adam and beyond. In *International Conference on Learning Representations*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer.

Saunshi, N., Karp, S., Krishnan, S., Miryoosefi, S., Reddi, S. J., and Kumar, S. (2024). On the inductive bias of stacking towards improving reasoning. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 71437–71464. Curran Associates, Inc.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shazeer, N. and Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Shi, H.-J. M., Lee, T.-H., Iwasaki, S., Gallego-Posada, J., Li, Z., Rangadurai, K., Mudigere, D., and Rabbat, M. (2023). A distributed data-parallel pytorch implementation of the distributed shampoo optimizer for training neural networks at-scale. *arXiv preprint arXiv:2309.06497*.

Shi, N. and Li, D. (2021). Rmsprop converges with proper hyperparameter. In *International conference on learning representation*.

Smith, S., Elsen, E., and De, S. (2020). On the generalization benefit of noise in stochastic gradient descent. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9058–9067. PMLR.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B., Meng, Q., Chen, W., and Liu, T.-Y. (2021). The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pages 10849–10858. PMLR.

Wang, B., Zhang, H., Meng, Q., Sun, R., Ma, Z.-M., and Chen, W. (2024a). On the convergence of adam under non-uniform smoothness: Separability from sgdm and beyond.

Wang, B., Zhang, Y., Zhang, H., Meng, Q., Sun, R., Ma, Z.-M., Liu, T.-Y., Luo, Z.-Q., and Chen, W. (2024b). Provable adaptivity of adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 2960–2969, New York, NY, USA. Association for Computing Machinery.

Wang, J. and Wiens, J. (2020). Adasgd: Bridging the gap between sgd and adam. *arXiv preprint arXiv:2006.16541*.

Wang, R., Malladi, S., Wang, T., Lyu, K., and Li, Z. (2023). The marginal value of momentum for small learning rate sgd. *arXiv preprint arXiv:2307.15196*.

Wen, K., Li, Z., Wang, J., Hall, D., Liang, P., and Ma, T. (2024). Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective.

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

Xie, S. and Li, Z. (2024). Implicit bias of AdamW: ℓ_∞-norm constrained optimization. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54488–54510. PMLR.

Zaheer, M., Reddi, S., Sachan, D., Kale, S., and Kumar, S. (2018). Adaptive methods for nonconvex optimization. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Zhang, C., Zou, D., and Cao, Y. (2024a). The implicit bias of adam on separable data.

Zhang, Q., Zhou, Y., and Zou, S. (2024b). Convergence guarantees for rmsprop and adam in generalized-smooth non-convex optimization with affine noise variance. *arXiv preprint arXiv:2404.01436*.

Zhang, Y., Chen, C., Ding, T., Li, Z., Sun, R., and Luo, Z.-Q. (2024c). Why transformers need adam: A hessian perspective.

Zhang, Y., Chen, C., Li, Z., Ding, T., Wu, C., Kingma, D. P., Ye, Y., Luo, Z.-Q., and Sun, R. (2025). Adam-mini: Use fewer learning rates to gain more.

Zhang, Y., Chen, C., Shi, N., Sun, R., and Luo, Z.-Q. (2022). Adam can converge without any modification on update rules. *Advances in neural information processing systems*, 35:28386–28399.

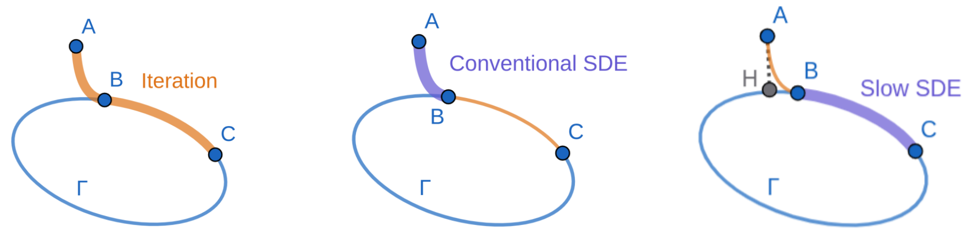
Zhao, R., Morwani, D., Brandfonbrener, D., Vyas, N., and Kakade, S. (2025). Deconstructing what makes a good optimizer for language models.

Zou, F., Shen, L., Jie, Z., Zhang, W., and Liu, W. (2019). A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11127–11135.

Contents

A Illustration of the Difference between Conventional SDE and Slow SDE

In this section, we illustrate the difference between conventional SDE and slow SDE. In Fig. 3, let Γ denotes a 1D manifold, then the discrete iteration of the optimization process can be seen as successive steps (orange, Fig. 3a) that starts from A, first converge to some point B in Γ and then move along Γ to C.



(a) (b) (c)

Figure 3: Comparison of conventional SDE and slow SDE.

The main intuition behind slow SDE is that the whole process $A \to B \to C$ can actually be decomposed into two motions: a convergence motion $A \to H$ (dashed, Fig. 3c) and an implicit regularization motion $H \to B \to C$. The convergence motion is fast and dominates the dymanics during the convergence phase, but it fades out as soon as convergence phase ends; meanwhile the slow, implicit regularization motion starts to dominate.

The conventional SDE approximates the convergence phase only, whose unit time corresponds to $\tilde{O}(\eta^{-1})$ steps (Fig. 3b). In contrast, slow SDE manages to separate the slow implicit regularization motion from the fast convergence, and approximate the implicit regularization near manifold only (Fig. 3c).

Remark A.1. *The projection method (which projects $A \to B \to C$ to $H \to B \to C$) varies in the analysis of different optimizers. Intuitively, the projection should reflect the converging direction driven by a clean (without noise) and continuous version of the optimizer. In SGD the projection is gradient flow; but in Adam we need to consider the preconditioning effect caused by $1/\sqrt{\boldsymbol{v}+\epsilon}$, so we add an SDE to track the preconditioner, and define a preconditioned gradient flow for projection.*

B Additional Related Work

Implicit Bias of SGD. The implicit bias of SGD has been studied over time; HaoChen et al. (2021) showed that SGD with label noise recovers the sparse ground-truth on a quadratically parameterized model. Blanc et al. (2020) proved that the fixed point of SGD is the minimizer of the trace of Hessian, in the zero-loss manifold, with MSE, label noise. Later on Damian et al. (2021) extended Blanc's results to global convergence and large learning rate. Li et al. (2021b) for the first time, gave a general framework for the implicit bias of SGD near the minimizer manifold, through an SDE manner. Gu et al. (2023b) generalized this framework to Local SGD (Lin et al., 2018), and developed a fine-grained analysis of the SDE proposed by Li et al. (2021b). Wang et al. (2023) showed that the Momentum does not affect the implicit bias.

Another line of work that also has limitations is *implicit gradient regularization (IGR)*. For full-batch GD, this method tries to find higher-order terms that can be added to the gradient flow ODE to approximate discrete GD iterations more accurately (Barrett and Dherin, 2020); This idea was later generalized to Adam's analysis by Cattaneo et al. (2024) and further to AdamW by Cattaneo and Shigida (2025). In particular, Cattaneo et al. (2024) argued that full-batch Adam with constant learning rate approximately follows an ODE that anti-regularizes sharpness when $\beta_1 < \beta_2$. Our work analyzes the dynamics of Adam for $\mathcal{O}(\eta^{-2})$ steps, a longer horizon than Cattaneo et al. (2024). Our analysis shows that with gradient noise, Adam can be characterized by an slow SDE that regularizes sharpness in the long term, offering a complementary perspective.

Implicit Bias of Adam. On the theoretical side, the current literature still lacks a rigorous understanding of the implicit bias of Adam, although it's more widely used than SGD in practical deep learning training, especially for large language models. However, many efforts have been made on this problem. Qian and Qian (2019) and Xie and Li (2024) characterized the implicit biases of AdaGrad and AdamW, respectively. However, their methods cannot be generalized to Adam. Wang et al. (2021) proved the implicit regularizer of Adam as identical to that of SGD, while their result requires that the gradient entries be lower than ϵ, which is typically not feasible in practice. Zhang et al. (2024a)'s analysis is also limited in its use, since it studies Adam's implicit bias on linear separable data, a condition generally not met by real-world applications.

Approximation of Stochastic First-Order Methods with Ito SDE. To fill in the gaps and provide a theoretical analysis that tracks iterations of Adam for a sufficiently long time, we use the same approximation tool as in the aforementioned Li et al. (2021b) and Gu et al. (2023a), namely *slow SDE* (termed by Gu et al. (2023a)). Specifically, optimizers such as SGD and Adam takes $\tilde{\mathcal{O}}(\eta^{-1})$ steps to converge onto the manifold, and then moves along the manifold for $\mathcal{O}(\eta^{-2})$ steps, during which the optimizer will be dominated by a slower implicit regularization dynamics, different from the mixing dynamics during the convergence phase. Conventional SDE papers such as Li et al. (2018, 2021a); Cattaneo et al. (2024); Malladi et al. (2024) approximate the iteration itself during the convergence phase, however, they struggle to bound the approximation error if extended to the manifold phase. Instead, a slow SDE peels the convergence dynamics off by only approximating the iteration's projection on the manifold. In this way, a slow SDE can track the optimizer's iteration during the whole manifold phase for $\mathcal{O}(\eta^{-2})$ time. This idea will be made explicit in Section 3.

Adaptive Gradient Methods. As a test-of-time optimizer that has revolutionized the field of deep learning (Kingma and Ba, 2014), Adam innovatively combined the moving average of the first and second moments of gradients to determine an adaptive learning rate. Adam has also spawned a family of derivative optimizers such as AdamW (Loshchilov and Hutter, 2017), AdaFactor (Shazeer and Stern, 2018), Adam-mini (Zhang et al., 2025), Adalayer (Zhao et al., 2025) and AdaSGD (Wang and Wiens, 2020), maintaining significant advantages over SGD in terms of empirical use. Under a more general framework of adaptive gradient methods, many optimizers also get huge success as adaptive gradient methods, such as RMSprop (Hinton et al., 2012), Adafactor (Shazeer and Stern, 2018).

C Matrix Factorization: Adam Implicitly Regularizes Sharpness Differently

The diagonal net experiments in Section 5 showed that Adam's implicit bias towards *sparsity* improves generalization relative to SGD. We now turn to supply the potentially negative impact of Adam's implicit bias in another controlled setting: **deep matrix factorization with label noise**, where the relevant implicit regularizers are analytically tractable. In this task, Adam is expected to minimize tr(Diag$(\boldsymbol{H})^{1/2}$) rather than tr($\boldsymbol{H}$). Leveraging existing theory, we therefore predict that (i) Adam will converge to a solution with tr($\boldsymbol{H}$) larger—but tr(Diag$(\boldsymbol{H})^{1/2}$) smaller—than SGD's solution, and (ii) once training reaches the interpolation regime, Adam will *generalize worse* than vanilla SGD in the presence of label noise. Our experiments confirm both predictions (Figure 4).

C.1 Problem setup

Consider an L-layer linear network with parameters $\boldsymbol{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L)$, where $\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$ and $d_i \geq \min\{d_0, d_L\}$ for all i. Let $\boldsymbol{M}^* \in \mathbb{R}^{d_L \times d_0}$ be a rank-r ground–truth matrix, and observe n i.i.d. linear measurements $\{(\boldsymbol{A}_i, b_i)\}_{i=1}^n$ generated by $b_i = \langle \boldsymbol{A}_i, \boldsymbol{M}^* \rangle$. With label noise and mini–batch size B the empirical loss at step t is

$$\mathcal{L}_t(\boldsymbol{W}) = \frac{1}{B} \sum_{i \in \mathcal{B}_t} \big(\langle \boldsymbol{A}_i, \boldsymbol{W}_L \cdots \boldsymbol{W}_1 \rangle - b_i + \xi_{t,i} \big)^2,$$

where $\mathcal{B}_t$ is a fresh batch of size B, and $\xi_{t,i} \sim \mathcal{N}(0, \sigma^2)$ are independent across (t, i).

Implicit regularization. It is known that vanilla SGD with label noise, with a small learning rate, implicitly minimizes the trace of the Hessian matrix of the loss after reaching the zero-loss manifold. In this matrix factorization task, Gatmiry et al. (2023) proved that the minimization of the trace of Hessian is roughly equivalent to the minimization of the nuclear norm of $\boldsymbol{W}^*$. Futher, when $\boldsymbol{M}^*$ itself has a low rank nature, the minimization of nuclear norm can induces better genralization. Hence, Adam, whose regularization term is $\mathrm{tr}(\mathrm{diag}(\boldsymbol{H})^{1/2})$ rather than $\mathrm{tr}(\boldsymbol{H})$, should converge to a solution different from that of SGD vonverges to. And the solution found by Adam should have a larger trace

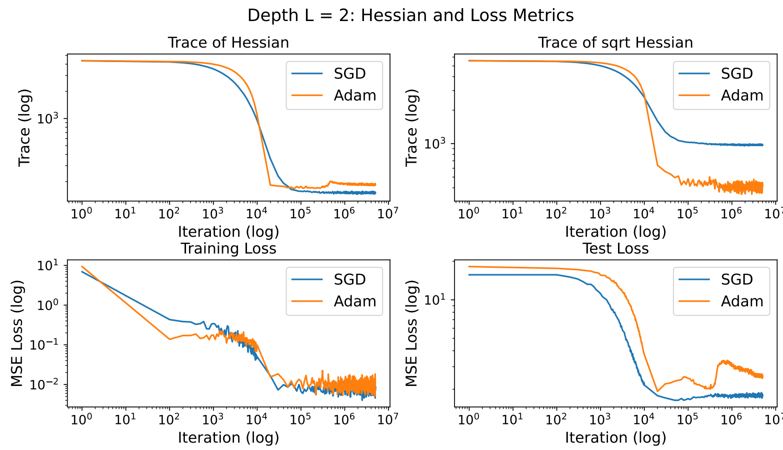


Figure 4: **Deep matrix factorization with label noise with deepth $L = 2$.** Adam and SGD are trained on identical data and noise realizations. *Top:* evolution of $\mathrm{tr}(\boldsymbol{H})$ and $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{1/2})$. *Bottom:* training and test MSE. Adam converges to a point with larger overall curvature but smaller diagonal curvature, and exhibits higher test error.

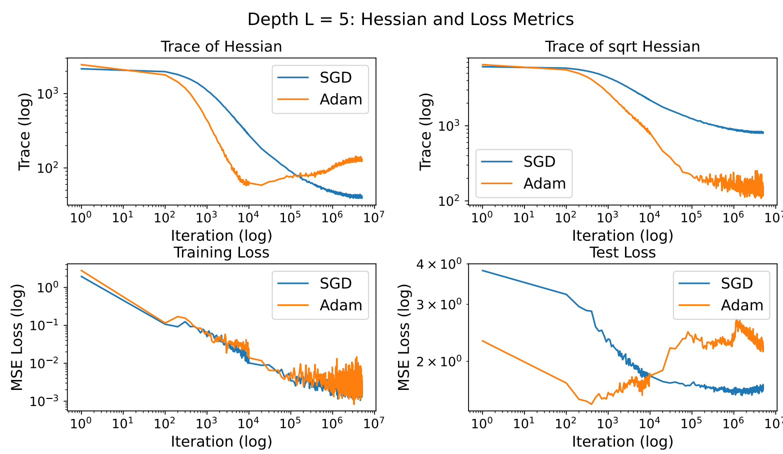


Figure 5: **Deep matrix factorization with label noise with deepth $L = 5$.**

of Hessian, but a smaller trace of the square root of the diagonal Hessian than the solution of SGD. We could also observe a generalization performance degradation of Adam.

C.2 Results

Our SGD setup follows Section 7 of Gatmiry et al. (2023). For Adam, we use the standard hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate 10^{-3}; all other settings are identical to SGD.

Figure 4 (top row) shows the evolution of curvature metrics. Adam drives $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{1/2})$ sharply downward while $\mathrm{tr}(\boldsymbol{H})$ remains high and even non-monotone, confirming that Adam does *not* target overall Hessian trace. Correspondingly, the bottom row shows that Adam attains a higher test MSE despite identical training error—evidence that its implicit bias is detrimental in this setting.

Takeaway. In deep matrix factorization with label noise, Adam's preference for minimizing the diagonal curvature leads it to sharper—and less generalizable—solutions than SGD, reinforcing that Adam's implicit regularization differs qualitatively from SGD's and can hurt performance when overall curvature matters.

D Formal Statements of the Main Results

In this section, we give the formal versions of the main results stated in Section 4, where we presented the two main theorems:

1. The AGM iterates converge to a neighborhood of the manifold (Theorem 4.2);
2. Moreover, once the iterates enter this neighborhood, their dynamics over $\mathcal{O}(\eta^{-2})$ discrete steps can be accurately tracked by a slow SDE (Theorem 4.1).

Recall that in the AGM framework, the transition from $\boldsymbol{\theta}_k$ to $\boldsymbol{\theta}_{k+1}$ is defined as:

$$\begin{aligned}\boldsymbol{m}_{k+1} &:= \beta_1 \boldsymbol{m}_k + (1-\beta_1)\nabla \ell_k(\boldsymbol{\theta}_k)\\ \boldsymbol{v}_{k+1} &:= \beta_2 \boldsymbol{v}_k + (1-\beta_2)V\big(\nabla \ell_k(\boldsymbol{\theta}_k)\nabla \ell_k(\boldsymbol{\theta}_k)^\top\big)\\ \boldsymbol{\theta}_{k+1} &:= \boldsymbol{\theta}_k - \eta S(\boldsymbol{v}_{k+1})\boldsymbol{m}_{k+1},\end{aligned}$$

under the following conditions:

1. $S: \mathbb{R}^d_{\geq 0} \longrightarrow \mathbb{S}^d_{++}$ is ρ_s-smooth, where $\mathbb{R}^d_{\geq 0}$ denotes the subset of $\mathbb{R}^d$ that has non-negative entries, and $\mathbb{S}^d_{++}$ denotes the space of $\mathbb{R}^{d\times d}$ positive definite matrices.
2. $S(\boldsymbol{v}) \succeq \frac{1}{R_0}\boldsymbol{I}$ for some $R_0 > 0$ and any $\boldsymbol{v} \in \mathbb{R}^d$.
3. $V: \mathbb{R}^{d\times d} \longrightarrow \mathbb{R}^d$ is linear, and $V(\boldsymbol{g}\boldsymbol{g}^\top) \in \mathbb{R}^d_{\geq 0}$ for all $\boldsymbol{g} \in \mathbb{R}^d$.

D.1 Slow SDE for AGMs

Theorem D.1. *Let Assumptions 3.2, 3.3 and 4.2 be satisfied. Let Γ denote a local minimizer manifold, and let η be a sufficiently small learning rate of an AGM. Then we have the following conclusions:*

1. *(Convergence to a near–manifold neighborhood) There exists a constant $\epsilon_0 > 0$ such that for any initial point $\boldsymbol{\theta}_0$ whose L2 distance from Γ does not exceed ϵ_0, and any $\delta \in (\eta^{200}, 1)$,*[5] *with probability at least $1-\delta$, the following holds for some $K_0 = \mathcal{O}(\frac{1}{\eta}\log\frac{1}{\eta})$:*

$$\mathcal{L}(\boldsymbol{\theta}_{K_0}) - \mathcal{L}^* = \mathcal{O}\left(\eta \log \frac{1}{\eta\delta}\right),$$

$$\|\boldsymbol{\theta}_{K_0} - \Phi_{\boldsymbol{S}_{K_0}}(\boldsymbol{\theta}_{K_0})\|_2 = \mathcal{O}\left(\sqrt{\eta \log \frac{1}{\eta\delta}}\right).$$

2. *(Formal restatement of Theorem 4.1: Slow SDE tracks AGM's trajectory in a weak approximation sense) Moreover, when Assumptions 3.4, 3.1 and 4.1 hold, we shift the timeline and redefine the final state $(\boldsymbol{\theta}_{K_0}, \boldsymbol{v}_{K_0})$ in conclusion 1 by $(\boldsymbol{\theta}_0, \boldsymbol{v}_0)$. Let $T > 0$ be a constant, $\boldsymbol{X}(t) = (\boldsymbol{\zeta}(t), \boldsymbol{v}(t))$ be the solution to Equation (2) with initial condition:*

$$\boldsymbol{\zeta}(0) = \Phi\big(\boldsymbol{\theta}_0\big) \in \Gamma, \quad \boldsymbol{v}(0) = \boldsymbol{v}_0 \in \mathbb{R}^d,$$

and define the parameters of Adam as $\bar{\boldsymbol{X}}_t := (\Phi_{\boldsymbol{S}_t}(\boldsymbol{\theta}_t), \boldsymbol{v}_t)$. For any C^3-smooth function $g(\boldsymbol{\theta})$,

$$\max_{0\leq t\leq \lfloor \frac{T}{\eta^2}\rfloor}\Big|\mathbb{E}\big[g\left(\bar{\boldsymbol{X}}_t\right)\big] - \mathbb{E}\big[g\left(\boldsymbol{X}(t\eta^2)\right)\big]\Big| = \widetilde{\mathcal{O}}\big(\eta^{0.25}\big),$$

where $\widetilde{O}(\cdot)$ hides logarithmic factors and constants that are independent of η but may depend on $g(\boldsymbol{\theta})$.

[5] The exponent here, along with the exponents related to the δ-goodness in Definition G.1, can be arbitrary large constant, which does not affect the order of following derivations.

D.2 Convergence Guarantee of AGMs

In the proof, the first part of Theorem D.1 is done by first proving a convergence result with global μ-PL condition, and then arguing that AGM starting near enough to the manifold will stick to the manifold with high probability. As mentioned in Section 4.3.1, the convergence under μ-PL condition can be seen as a separate technical contribution of our paper, which is stated below.

Definition D.1 (Polyak-Łojasiewicz Condition)**.** *For some $\mu, \bar{L} > 0$, we say some function $\mathcal{L} : \mathbb{R}^d \to d$ is $(\mu, \bar{L})$-Polyak-Łojasiewicz (abbreviated as $(\mu, \bar{L})$-PL), if and only if for all $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $\mathcal{L}(\boldsymbol{\theta}) \leq \bar{L}$:*

$$2\mu(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^*) \leq \|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 .$$

When $\bar{L} = +\infty$, we call this condition the μ-Polyak-Łojasiewicz (μ-PL) condition.

Theorem D.2 (Formal restatement of Theorem 4.2)**.** *Let Assumptions 3.2, 3.3 and 4.2 be satisfied, $\mathcal{L}$ be a function satisfying the μ-PL condition, and η be a sufficiently small learning rate of an AGM. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following holds for some $K = \mathcal{O}(\frac{1}{\eta} \log \frac{1}{\eta})$:*

$$\mathcal{L}(\boldsymbol{\theta}_K) - \mathcal{L}^* = \mathcal{O}\left(\eta \log \frac{1}{\eta\delta}\right),$$

$$\|\boldsymbol{\theta}_K - \Phi_{\boldsymbol{S}_K}(\boldsymbol{\theta}_K)\|_2 = \mathcal{O}\left(\sqrt{\eta \log \frac{1}{\eta\delta}}\right).$$

Remark. Note that Theorem D.2 is different from Part 1 of Theorem D.1 in the sense that Theorem D.2 requires the μ-PL condition to hold globally, while Part 1 of Theorem D.1 does not. Actually, the latter requires the iteration to start from some neighborhood of Γ. Later on, we will find from Lemma E.3 that μ-PL provably exists in a neighborhood of Γ, and we prove that an iteration of AGMs starting within that neighborhood stays within that neighborhood with high probability.

There have been many previous works discussing the convergence bound of Adam. For example, Reddi et al. (2018) and Dereich and Jentzen (2024) give convergence bounds under the convexity condition, Zou et al. (2019), Shi and Li (2021) and Zhang et al. (2022) focus on the cases where learning rates follow a $1/\sqrt{t}$ decay, and the bounds given by Zaheer et al. (2018), Zhang et al. (2022) and Wang et al. (2024b) do not decrease to 0 as $\eta \to 0$. Also, most works (Défossez et al., 2020; Guo et al., 2025; Iiduka, 2022; Wang et al., 2024a; Zhang et al., 2024b; Hong and Lin, 2023) only establish an upper bound on the average of gradient norms over the time of iteration. In contrast, we directly bound the loss term of the last step to $o(1)$. Going beyond convex loss functions, we establish the bound on μ-PL functions, and we focus on the constant learning rate schedule.

E Constructing the Working Zones

Note that it is generally hard to ensure some properties that are crucial to the feasibility of our analysis, such as the μ-PL condition or the well-definedness of preconditioned gradient projections. However, this becomes possible when we constrain the discussion inside some local neighborhood of a manifold. So in this subsection, we construct "working zones" around any local minimizer manifold Γ such that iterations inside the working zones will be captured by the manifold and obtain certain properties that support the analysis of slow SDE.

For any $\Gamma \subset \mathbb{R}^n$ being a nonempty set and $\boldsymbol{\theta} \in \mathbb{R}^n$, let $\|\cdot\|_2$ denote the Euclidean norm. We denote the distance from $\boldsymbol{\theta}$ to Γ as $\mathrm{dist}(\boldsymbol{\theta}, \Gamma) := \inf_{\boldsymbol{\zeta}\in\Gamma} \|\boldsymbol{\theta} - \boldsymbol{\zeta}\|_2$. Note that when Γ is closed, the infimum is attained (i.e., there exists $\boldsymbol{\zeta}^\star \in \Gamma$ with $\|\boldsymbol{\theta} - \boldsymbol{\zeta}^\star\|_2 = \mathrm{dist}(\boldsymbol{\theta}, \Gamma)$).

Definition E.1 (Neighborhood of a Manifold)**.** *For any manifold Γ and positive constant ϵ, the ϵ-neighborhood of Γ, denoted by Γ^ϵ is defined as the set of points $\boldsymbol{\theta}$ such that*

$$\mathrm{dist}(\boldsymbol{\theta}, \Gamma) \leq \epsilon.$$

Definition E.2 (Preconditioned gradient flow)**.** *For any differentiable function $\mathcal{L}$ and any matrix $\mathbf{S} \in \mathbb{R}^{d\times d}$, the* $\mathbf{S}$-preconditioned gradient flow *of $\mathcal{L}$ is the ordinary differential equation*

$$\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} = -\,\mathbf{S}\,\nabla\mathcal{L}\big(\boldsymbol{\theta}(t)\big).$$

*When the objective $\mathcal{L}$ (or the time dependence of $\boldsymbol{\theta}$) is clear from context, we may omit it in the notation and simply refer to the system as the $\boldsymbol{S}$-*preconditioned gradient flow *or the* gradient flow preconditioned by $\mathbf{S}$.

Lemma E.1. *Let $C_1, C_2 > 0$ with $C_1 \leq C_2$, and let $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ be ρ-smooth and satisfy the μ-PL condition. For any symmetric matrix $\boldsymbol{S}$ with $C_1 \boldsymbol{I} \preceq \boldsymbol{S} \preceq C_2 \boldsymbol{I}$, consider the $\boldsymbol{S}$-preconditioned gradient flow of $\mathcal{L}$ starting at $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$. Then for any $T > 0$,*

$$\|\boldsymbol{\theta}(T) - \boldsymbol{\theta}_0\|_2 \ \leq \ \frac{2C_2}{C_1\sqrt{2\mu}} \sqrt{\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*},$$

where $\mathcal{L}^* = \inf_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$.

Proof. Since $C_1 \boldsymbol{I} \preceq \boldsymbol{S} \preceq C_2 \boldsymbol{I}$, we have $\|\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{\theta})\|_2 \leq C_2 \|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2$ and $\langle \nabla\mathcal{L}(\boldsymbol{\theta}), \boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{\theta})\rangle \geq C_1 \|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2^2$ for any $\boldsymbol{\theta}$, which implies

$$\langle \nabla\mathcal{L}(\boldsymbol{\theta}), \boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{\theta})\rangle \geq \frac{C_1}{C_2} \|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2 \|\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{\theta})\|_2 .$$

Then plugging in the above equation gives that, for any $t < T$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}^*} &= \frac{1}{2}\left(\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}^*\right)^{-\frac{1}{2}} \cdot \left\langle \nabla\mathcal{L}(\boldsymbol{\theta}(t)), \frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} \right\rangle \\
&\leq -\frac{C_1}{2C_2}\left(\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}^*\right)^{-\frac{1}{2}} \cdot \|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\|_2 \left\|\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t}\right\|_2 \\
&\leq -\frac{C_1}{2C_2}\left(\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}^*\right)^{-\frac{1}{2}} \cdot \sqrt{2\mu(\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}^*)} \left\|\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t}\right\|_2 \\
&= -\frac{\sqrt{2\mu}C_1}{2C_2}\left\|\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t}\right\|_2 .
\end{aligned}$$

Integrating both sides gives us

$$\begin{aligned}
\sqrt{\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*} &\geq \frac{\sqrt{2\mu}C_1}{2C_2}\int_0^T \left\|\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t}\right\|_2 \\
&\geq \frac{\sqrt{2\mu}C_1}{2C_2}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}(T)\|_2 .
\end{aligned}$$

The above equations complete the proof. □

To avoid ambiguity, all comparisons between vectors and scalars are interpreted componentwise. Specifically, for $v \in \mathbb{R}^d$ and a scalar $c \in \mathbb{R}$ we write $v \leq c$ (resp. $v < c$) iff $v_i \leq c$ (resp. $v_i < c$) for every coordinate i. Equivalently $v \geq c$ (resp. $v > c$) means $v_i \geq c$ (resp. $v_i > c$) for all i. In particular the notation $0 \leq v \leq c$ means $0 \leq v_i \leq c$ for every i, which is the convention used in the sequel. Recall that $\mathbb{R}^d_{\geq 0}$ means $\left\{\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^d, \boldsymbol{v} \geq 0\right\}$. In the sequel, we slightly abuse this notation such that for any subset $I \subseteq \mathbb{R}$, $\mathbb{R}^d_I$ means $\left\{\boldsymbol{v} \mid \boldsymbol{v} \in R^d, v_i \in I \text{ for all } i \in [d]\right\}$.

Lemma E.2. *There exist constants $R_1, R_2 > 0$ such that for all $k \geq 0$, $0 \leq \boldsymbol{v}_k \leq R_1$ and $S(\boldsymbol{v}) \preceq R_2 \boldsymbol{I}$ almost surely. Moreover, S is Lipschitz on $\mathbb{R}^d_{[0,R_1]}$.*

Proof. From Assumption 3.3, all noisy gradients $\nabla\ell_k(\boldsymbol{\theta}_k)$ are uniformly bounded by a constant R. Hence $V\left(\nabla\ell_k(\boldsymbol{\theta}_k)\nabla\ell_k(\boldsymbol{\theta}_k)^\top\right)$ is also bounded. Combining with the condition that $V(\boldsymbol{g}\boldsymbol{g}^\top) \geq 0$ for all $\boldsymbol{g}$, we have $V\left(\nabla\ell_k(\boldsymbol{\theta}_k)\nabla\ell_k(\boldsymbol{\theta}_k)^\top\right) \in \mathbb{R}^d_{[0,R_1]}$ for some constant R_1. Since $\boldsymbol{v}_k$ is an exponential moving average of previous $V\left(\nabla\ell_t(\boldsymbol{\theta}_t)\nabla\ell_t(\boldsymbol{\theta}_t)^\top\right)$ terms and $\mathbb{R}^d_{[0,R_1]}$ is convex, we have $\boldsymbol{v}_k \in \mathbb{R}^d_{[0,R_1]}$ for all $k \geq 0$.

From Assumption 4.1, S is ρ_s-smooth, hence both S and ∇S are continuous, thus are bounded on the compact set $\mathbb{R}^d_{[0,R_1]}$. The boundedness of S gives the existence of R_2, while the boundedness of ∇S gives the Lipschitzness of S. □

We continue to use the notations R_1 and R_2 throughout the following part of the paper. Note that for all optimizers listed in Table 1, setting $R_1 = R^2$ is sufficient. Another thing to clarify is the relationship between the R_2 here and the stabilizing constant ϵ used by optimizers in Table 1. We will call it ϵ_{optim} here, so as to distinguish from the ϵ notations that represent a distance (for instance, the ϵ in Definition E.1 or Lemma E.3).

Remark E.1 (Relationship between R_2 and ϵ_{optim})**.** *Setting $R_2 := 1/\epsilon_{\text{optim}}$ here is theoretically enough for the requirement in Lemma E.2 to hold, but will introduce a large constant to the proof since ϵ_{optim} is very small in practice; However, in practice the gradient noise is also very likely to keep $\boldsymbol{v}$ away from zero, thus the operational R_0 that governs empirical convergence is usually much smaller than the worst-case $1/\epsilon_{\text{optim}}$.*

Now we are ready to construct working zones in which nice properties are ensured to benefit our analysis. For all $\epsilon > 0$, define $\mathcal{X}^\epsilon := \Gamma^\epsilon \times \mathbb{R}^d_{[0,R_1]}$ as the set of AGM states $(\boldsymbol{\theta}, \boldsymbol{v})$ where $\boldsymbol{\theta}$ lies in Γ^ϵ and $0 \le \boldsymbol{v} \le R_1$.

We construct nested working zones $(\Gamma^{\epsilon_1}, \Gamma^{\epsilon_2}, \Gamma^{\epsilon_3})$ in the following way:

Lemma E.3 (Working Zone Lemma)**.** *We denote the minimal distance of Γ and any other local minimizer manifold as ϵ_4. There exist positive constants $\epsilon_1, \epsilon_2, \epsilon_3$ such that $\epsilon_1 < \epsilon_2 < \epsilon_3 < \epsilon_4$ and $\Gamma^{\epsilon_1}, \Gamma^{\epsilon_2}, \Gamma^{\epsilon_3}$ satisfy the following properties:*

1. *$\mathcal{L}$ is μ-PL in Γ^{ϵ_3} for some constant $\mu > 0$.*
2. *For all matrices $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ with $\frac{1}{R_0}\boldsymbol{I} \preceq \boldsymbol{S} \preceq \frac{1}{\epsilon}\boldsymbol{I}$, starting from any initial point $\boldsymbol{\theta}_0 \in \Gamma^{\epsilon_2}$, the gradient flow preconditioned by $\boldsymbol{S}$ converges to a point in Γ.*
3. *Under Assumption 3.1 and Assumption 4.1 the function $F : \mathcal{X}^{\epsilon_2} \to \mathbb{R}^d, (\boldsymbol{\theta}, \boldsymbol{v}) \mapsto \Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})$ is $\mathcal{C}^4$ on $\mathcal{X}^{\epsilon_1}$.*

Proof. By Lemma H.3 in Lyu et al. (2022), there exists an ϵ_3-neighborhood of Γ where $\mathcal{L}$ is μ-PL for some $\mu > 0$. WLOG let $\epsilon_3 < \epsilon_4$.

We prove the second property by contradiction. Let $C_1 = 1/R_0$ and $C_2 = 1/\epsilon$. Let ϵ_2 be some constant such that $\epsilon_2 + \sqrt{\frac{\rho}{\mu}} \cdot \frac{C_2}{C_1}\epsilon_2 < \epsilon_3$. For any starting point $\boldsymbol{\theta}_0 \in \Gamma^{\epsilon_2}$, and any preconditioning matrix $\boldsymbol{S}$ satisfying $C_1\boldsymbol{I} \preceq \boldsymbol{S} \preceq C_2\boldsymbol{I}$, assume on the contrary that the preconditioned gradient flow starting from $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$ will leave Γ^{ϵ_3} at some finite time. Then let $T = \inf\{t : \boldsymbol{\theta}(t) \notin \Gamma^{\epsilon_3}\} < \infty$.

Using Lemma E.1 and combining the μ-PL condition, we conclude that

$$\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}(T)\|_2 \le \frac{2C_2}{\sqrt{2\mu}C_1}\sqrt{\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*} \le \frac{2C_2}{\sqrt{2\mu}C_1} \cdot \sqrt{\frac{\mu}{2}}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2 = \sqrt{\frac{\rho}{\mu}} \cdot \frac{C_2}{C_1}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2$$

for any $\boldsymbol{\theta}^* \in \Gamma$. Hence $\boldsymbol{\theta}(T) \in \Gamma^{\epsilon_3}$, which is a contradition.

Next we begin the construction of Γ^{ϵ_1} with Assumptions 3.1 and 4.1. Define a function $f(\boldsymbol{\theta}, \boldsymbol{v}) : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ as

$$f(\boldsymbol{\theta}, \boldsymbol{v}) := (-S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta}), \boldsymbol{v}),$$

then f is $\mathcal{C}^4$ on $\mathbb{R}^d \times \mathbb{R}^d_{[0,R_1]}$. Let $\tilde{r}$ be a constant such that $\tilde{r} > \epsilon_2$. Substituting $f_0 = f$, $r = \sqrt{\tilde{r}^2 + d \cdot R_1^2}$, $x_0 = (\boldsymbol{\theta}_0, \boldsymbol{v}_0)$ such that each entry of $\boldsymbol{v}_0$ is $R_1/2$ and $\boldsymbol{\theta}_0$ be arbitrary point in Γ, and $B = \mathcal{X}^{\tilde{r}}$ into Lemma B.4 in Duistermaat and Kolk (2012), we conclude that there exists some constant δ such that the mapping $\gamma_\delta(\boldsymbol{\theta}, \boldsymbol{v})$ defined by:

$$\boldsymbol{\theta}(0) = \boldsymbol{\theta}, \quad \frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} = -S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta}(t)), \quad \gamma_\delta(\boldsymbol{\theta}, \boldsymbol{v}) = \boldsymbol{\theta}(\delta)$$

is well-defined and $\mathcal{C}^4$ on $\mathcal{X}^{\tilde{r}}$. Note that we require a slight modification of the original proof since B is now a factorization of a ball and a hypercube instead of a ball, but the convexity of B is preserved, hence the modification is trivial.

Note that the constant δ can be independent with $\boldsymbol{\theta}_0$ to fulfill the requirements of Lemma B.4 in Duistermaat and Kolk (2012) since $\|\nabla\mathcal{L}\|_2$ and $\left\|\nabla^2\mathcal{L}\right\|_2$ can be uniformly bounded. Take $\epsilon_1 = 0.9\epsilon_2$, then for any $\boldsymbol{\theta} \in \Gamma^{\epsilon_1}$, a small open neighborbood of $\boldsymbol{\theta}$ stays in the ϵ_2-neighborhoods of two different points on Γ. Taking union of all $\boldsymbol{\theta}_0 \in \Gamma$, we conclude that γ_δ is $\mathcal{C}^4$ on $\mathcal{X}^{\epsilon_1}$. Finally, we use Theorem 6.4 in Falconer (1983) to conclude that $F(\boldsymbol{\theta}, \boldsymbol{v}) := \Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})$ is $\mathcal{C}^4$ on $\mathcal{X}^{\epsilon_1}$. $\square$

F Proof of the Convergence of AGMs

In this section, we aim to prove Theorem D.2 and the first part of Theorem D.1. Specifically, for some constant $\gamma = 1 - \Theta(\eta)$, we prove that the loss value of AGM converges to $\tilde{\mathcal{O}}(\gamma^K + \eta)$ within K steps with high probability. If we substitute $K = \mathcal{O}\left(\frac{1}{\eta}\log\frac{1}{\eta}\right)$, this will recover the first part of Theorem D.1; However, this convergence analysis works for any $K = \mathcal{O}(\mathrm{poly}(1/\eta))$, and substituting $K = \mathcal{O}(\eta^{-2})$ will give us a high probability guarantee that the iteration stays near manifold in the whole scope of our analysis, which helps the proof of the second part too.

First, we introduce some additional notations that will be used in our proof. In the AGM framework, an algorithm starts from initial state $\boldsymbol{\theta}_0$, and we set $\boldsymbol{m}_0 = \boldsymbol{v}_0 = \mathbf{0}$. For every $k \geq 0$, we use **step** $k+1$ to refer to the process of obtaining the noisy gradient $\nabla \ell_k(\boldsymbol{\theta}_k)$ and then $\boldsymbol{m}_{k+1}, \boldsymbol{v}_{k+1}$ and $\boldsymbol{\theta}_{k+1}$.

For any $k \geq 0$, to simplify the notation, we denote that

$$\boldsymbol{g}_k := \nabla \ell_k(\boldsymbol{\theta}_k), \quad \boldsymbol{z}_k := \ell_k(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_k) \sim \mathcal{Z}(\boldsymbol{\theta}_k), \quad \boldsymbol{S}_k := S(\boldsymbol{v}_k),$$
$$\boldsymbol{U}_{k+1} := S(\boldsymbol{v}_k)\boldsymbol{g}_k, \quad \boldsymbol{u}_{k+1} := S(\boldsymbol{v}_k)\boldsymbol{m}_{k+1}, \quad \boldsymbol{\phi}_k := \Phi_{\boldsymbol{S}_k}(\boldsymbol{\theta}_k).$$

We use **time** k to refer to the time right before step $k+1$ happens, i.e. the time right after we get $\boldsymbol{\theta}_k$. We also define $\{\mathcal{F}_k\}$ as the natural filteration generated by the history of optimization, where each $\mathcal{F}_k = \sigma\left(\boldsymbol{\theta}_0, \boldsymbol{z}_0, \cdots, \boldsymbol{z}_{k-1}\right)$ can be interpreted as "all the information available up to time k". We use the notation $\mathbb{E}_k$ to denote the expectation conditioned on $\mathcal{F}_k$.

To start with, we prove that the descent direction of each step does not veer off the direction of a preconditioned gradient descent, and the mismatch term can also be constrained by a list of martingales. After that, we can ensure a decay in the loss function every step, with some small perturbations that can be dealt with using Azuma-Hoeffding's inequality.

From Lemma F.1 throughout Lemma F.5, we will assume that the loss function $\mathcal{L}$ satisfies μ-PL condition at each iteration step, which is automatically satisfied in the setting of Theorem D.2. follows directly from the result. After that, we argue that if the loss function satisfies μ-PL only within some local neighborhood, an AGM starting near enough to the manifold will stick to the manifold with high probability, which leads to the first part of Theorem D.1.

Lemma F.1. *Let $\mathcal{L}$ satisfy Assumption 3.2. Define $\tilde{\boldsymbol{v}}_k := \beta_2 \boldsymbol{v}_{k-1} + (1-\beta_2)\mathbb{E}_{k-1}[V\left(\boldsymbol{g}_{k-1}\boldsymbol{g}_{k-1}^\top\right)]$. There exist a constant C_{1a} and a constant C_{1b} independent of $\mathcal{L}$, such that for any $k \geq 1$,*

$$\langle \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right), \boldsymbol{U}_k \rangle = \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right)^\top S(\tilde{\boldsymbol{v}}_k) \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right) - Y_k - X_k,$$

where Y_k and X_k are two $\mathcal{F}_k$-measurable random variables such that:

1. *$|Y_k| \leq C_{1a} \left\|\nabla \mathcal{L}(\boldsymbol{\theta}_{k-1})\right\|_2 \cdot \eta^2$ a.s.*
2. *$|X_k| \leq C_{1b} \left\|\nabla \mathcal{L}(\boldsymbol{\theta}_{k-1})\right\|_2$ a.s., and $\mathbb{E}_{k-1}[X_k] = 0$.*

Proof. We first peel the $S(\tilde{\boldsymbol{v}}_k)$ part off the $S(\boldsymbol{v}_k)$ term:

$$\begin{aligned}\langle \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right), \boldsymbol{U}_k \rangle &= \langle \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right), S(\boldsymbol{v}_k)\boldsymbol{g}_{k-1} \rangle \\ &= \langle \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right), S(\tilde{\boldsymbol{v}}_k)\boldsymbol{g}_{k-1} \rangle + \langle \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right), \left(S(\boldsymbol{v}_k) - S(\tilde{\boldsymbol{v}}_k)\right)\boldsymbol{g}_{k-1} \rangle.\end{aligned}$$

Define Y_k as $Y_k = -\langle \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right), \left(S(\boldsymbol{v}_k) - S(\tilde{\boldsymbol{v}}_k)\right)\boldsymbol{g}_{k-1} \rangle$, then it holds almost surely that

$$|Y_k| \leq \left\|\nabla \mathcal{L}(\boldsymbol{\theta}_{k-1})\right\|_2 \left\|\left(S(\boldsymbol{v}_k) - S(\tilde{\boldsymbol{v}}_k)\right)\boldsymbol{g}_{k-1}\right\|_2.$$

Since S is Lipscitz, V is linear and

$$\left\|\tilde{\boldsymbol{v}}_k - \boldsymbol{v}_k\right\|_2 = (1-\beta_2)\left\|\mathbb{E}_{k-1}\left[V\left(\boldsymbol{g}_{k-1}\boldsymbol{g}_{k-1}^\top\right)\right] - V\left(\boldsymbol{g}_{k-1}\boldsymbol{g}_{k-1}^\top\right)\right\|_2,$$

we conclude that $|Y_k| \leq C_{1a}\left\|\nabla \mathcal{L}(\boldsymbol{\theta}_{k-1})\right\|_2 \cdot \eta^2$ a.s. for some constant C_{1a}. The rest term $\langle \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right), S(\tilde{\boldsymbol{v}}_k)\boldsymbol{g}_{k-1} \rangle$ can also be decomposed into a deterministic part and a random part as:

$$\begin{aligned}\langle \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right), S(\tilde{\boldsymbol{v}}_k)\boldsymbol{g}_{k-1} \rangle &= \langle \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right), S(\tilde{\boldsymbol{v}}_k)\left(\nabla \mathcal{L}(\boldsymbol{\theta}_{k-1}) + \boldsymbol{z}_{k-1}\right) \rangle \\ &= \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right)^\top S(\tilde{\boldsymbol{v}}_k) \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right) + \left\langle \boldsymbol{z}_{k-1}, S(\tilde{\boldsymbol{v}}_k)^\top \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right) \right\rangle.\end{aligned}$$

Now we only need to let $X_k = \left\langle \boldsymbol{z}_{k-1}, S(\tilde{\boldsymbol{v}}_k)^\top \nabla\mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right)\right\rangle$. It's easy to see that $\mathbb{E}_{k-1}[X_k] = 0$ and $|X_k| \le C_{1b} \left\|\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1})\right\|_2$ a.s. for some constant C_{1b}. Finally, note that C_{1b} is the multiplication of a constant bounding the magnitude of $\boldsymbol{z}$ and R_2 which bounds $\|S\|_2$, which is independent of $\mathcal{L}$. This completes the proof. $\square$

Lemma F.2 (Descent Lemma of the AGM Framework)**.** *Let $\mathcal{L}$ satisfy Assumption 3.2. For any $k \ge 1$ it holds that*

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k-1}) \le C_2\eta^2 - \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_{i-1}), \boldsymbol{U}_i\right\rangle$$

for some constant C_2.

Proof. From the smoothness of $\mathcal{L}$ we have

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k-1}) \le -\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \eta\boldsymbol{u}_k\right\rangle + \frac{\rho\eta^2}{2}\left\|\boldsymbol{u}_k\right\|_2^2.$$

If $k = 1$, then $\boldsymbol{m}_k = (1-\beta_1)\boldsymbol{g}_{k-1}$, so $\boldsymbol{u}_k = (1-\beta_1)\boldsymbol{U}_k$, and the statement trivially holds as long as $C_2 \ge \frac{\rho}{2}\left\|\boldsymbol{u}_k\right\|_2^2$. If $k > 1$, then the $-\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \boldsymbol{u}_k\right\rangle$ term can be expanded as

$$\begin{aligned}
-\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \boldsymbol{u}_k\right\rangle &= -\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), S(\boldsymbol{v}_k)\boldsymbol{m}_k\right\rangle \\
&= -\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), S(\boldsymbol{v}_k)\left(\beta_1\boldsymbol{m}_{k-1} + (1-\beta_1)\boldsymbol{g}_{k-1}\right)\right\rangle \\
&= -\beta_1\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), S(\boldsymbol{v}_k)\boldsymbol{m}_{k-1}\right\rangle - (1-\beta_1)\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), S(\boldsymbol{v}_k)\boldsymbol{g}_{k-1}\right\rangle \\
&= -\beta_1\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-2}), S(\boldsymbol{v}_{k-1})\boldsymbol{m}_{k-1}\right\rangle - (1-\beta_1)\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \boldsymbol{U}_k\right\rangle \\
&\quad - \beta_1\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}) - \nabla\mathcal{L}(\boldsymbol{\theta}_{k-2}), S(\boldsymbol{v}_{k-1})\boldsymbol{m}_{k-1}\right\rangle \\
&\quad - \beta_1\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \left(S(\boldsymbol{v}_k) - S(\boldsymbol{v}_{k-1})\right)\boldsymbol{m}_{k-1}\right\rangle \\
&\le -\beta_1\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-2}), S(\boldsymbol{v}_{k-1})\boldsymbol{m}_{k-1}\right\rangle - (1-\beta_1)\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \boldsymbol{U}_k\right\rangle \\
&\quad + \beta_1\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}) - \nabla\mathcal{L}(\boldsymbol{\theta}_{k-2})\right\|_2\left\|S(\boldsymbol{v}_{k-1})\boldsymbol{m}_{k-1}\right\|_2 \\
&\quad + \beta_1\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1})\right\|_2\left\|\left(S(\boldsymbol{v}_k) - S(\boldsymbol{v}_{k-1})\right)\boldsymbol{m}_{k-1}\right\|_2.
\end{aligned}$$

Note that a single step of update on $\boldsymbol{\theta}$ and $\boldsymbol{v}$ is small since

$$\begin{aligned}
\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1} &= \eta\boldsymbol{u}_k, \\
\boldsymbol{v}_k - \boldsymbol{v}_{k-1} &= \beta_2\boldsymbol{v}_{k-1} + (1-\beta_2)V\left(\boldsymbol{g}_{k-1}\boldsymbol{g}_{k-1}^\top\right) - \boldsymbol{v}_{k-1} \\
&= (1-\beta_2)\left(V\left(\boldsymbol{g}_{k-1}\boldsymbol{g}_{k-1}^\top\right) - \boldsymbol{v}_{k-1}\right),
\end{aligned}$$

which implies that $\left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\right\|_2 = \mathcal{O}(\eta)$ and $\left\|\boldsymbol{v}_k - \boldsymbol{v}_{k-1}\right\|_2 = \mathcal{O}(\eta^2)$. We then leverage the smoothness of $\nabla\mathcal{L}$ and S to conclude that there exists some constant $\tilde{C}_2$ such that

$$-\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \boldsymbol{u}_k\right\rangle \le -\beta_1\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-2}), \boldsymbol{u}_{k-1}\right\rangle - (1-\beta_1)\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \boldsymbol{U}_k\right\rangle + \beta_1\tilde{C}_2\eta.$$

Giving that $\boldsymbol{u}_0 = \boldsymbol{0}$, we can expand this formula iteratively as

$$\begin{aligned}
-\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \boldsymbol{u}_k\right\rangle &\le -\beta_1\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-2}), \boldsymbol{u}_{k-1}\right\rangle - (1-\beta_1)\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \boldsymbol{U}_k\right\rangle + \beta_1\tilde{C}_2\eta \\
&\le -\beta_1^2\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-3}), \boldsymbol{u}_{k-2}\right\rangle - \beta_1(1-\beta_1)\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-2}), \boldsymbol{U}_{k-1}\right\rangle \\
&\quad - (1-\beta_1)\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1}), \boldsymbol{U}_k\right\rangle + \beta_1\tilde{C}_2\eta + \beta_1^2\tilde{C}_2\eta \\
&\le \cdots \\
&\le -(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{i-1}), \boldsymbol{U}_i\right\rangle + \beta_1^{k-i+1}\tilde{C}_2\eta \\
&\le \frac{\beta_1}{1-\beta_1}\tilde{C}_2\eta - (1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{i-1}), \boldsymbol{U}_i\right\rangle.
\end{aligned}$$

Plugging in, we get

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k-1}) &\le \frac{\beta_1}{1-\beta_1}\tilde{C}_2\eta^2 + \frac{\rho\eta^2}{2}\left\|\boldsymbol{u}_k\right\|_2^2 - \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{i-1}), \boldsymbol{U}_i\right\rangle \\
&\le C_2\eta^2 - \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{i-1}), \boldsymbol{U}_i\right\rangle,
\end{aligned}$$

for some constant C_2. $\square$

Lemma F.3. *Let $\mathcal{L}$ satisfy Assumption 3.2, and assume that μ-PL condition is satisfied at all $\boldsymbol{\theta}_k$ where $k \geq 0$. Define $\gamma := 1 - \frac{2\eta\mu(1-\beta_1)}{R_0}$. For any $k \geq 0$, we have*

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^* \leq \gamma^k \left(\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*\right) + \eta(1-\beta_1)\sum_{i=1}^{k} X_i \sum_{j=i}^{k} \gamma^{k-j}\beta_1^{j-i} + C_3\eta$$

for some constant C_3.

Proof. We start from Lemma F.2 and plug in Lemma F.1:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k-1}) &\leq C_2\eta^2 - \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_{i-1}), \boldsymbol{U}_i\right\rangle \\
&= C_2\eta^2 - \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}\left(\nabla\mathcal{L}\left(\boldsymbol{\theta}_{i-1}\right)^\top S(\tilde{\boldsymbol{v}}_i)\nabla\mathcal{L}\left(\boldsymbol{\theta}_{i-1}\right) - Y_i - X_i\right).
\end{aligned}$$

Since $|Y_i| \leq C_{1a}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_{i-1})\right\|_2 \cdot \eta^2$ for every i, the effect of Y is negligible:

$$\left|\eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}Y_i\right| \leq C_{1a}\eta^3 \cdot \max_{i=0}^{k}\left\{\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_{i-1})\right\|_2\right\} = o(\eta^2),$$

and we can absorb it into the $C_2\eta^2$ term to write out that

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k-1}) \leq \tilde{C}_3\eta^2 - \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}\left(\nabla\mathcal{L}\left(\boldsymbol{\theta}_{i-1}\right)^\top S(\tilde{\boldsymbol{v}}_i)\nabla\mathcal{L}\left(\boldsymbol{\theta}_{i-1}\right) - X_i\right)$$

for some constant $\tilde{C}_3$. Note that $S(\tilde{\boldsymbol{v}}_i) \succeq \frac{1}{R_0}$, so $\nabla\mathcal{L}\left(\boldsymbol{\theta}_{i-1}\right)^\top S(\tilde{\boldsymbol{v}}_i)\nabla\mathcal{L}\left(\boldsymbol{\theta}_{i-1}\right) \geq \frac{1}{R_0}\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_{i-1}\right)\right\|_2^2$ for any i, hence

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k-1}) \leq \tilde{C}_3\eta^2 - \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}\left(\frac{1}{R_0}\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_{i-1}\right)\right\|_2^2 - X_i\right). \tag{4}$$

Combining with the μ-PL property $\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_{i-1}\right)\right\|_2^2 \geq 2\mu\left(\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^*\right)$, we have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^* &\leq \tilde{C}_3\eta^2 + \mathcal{L}(\boldsymbol{\theta}_{k-1}) - \mathcal{L}^* - \frac{2\eta\mu(1-\beta_1)}{R_0}\sum_{i=1}^{k}\beta_1^{k-i}\left(\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^*\right) \\
&\quad + \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}X_i \\
&\leq \tilde{C}_3\eta^2 + \left(1 - \frac{2\eta\mu(1-\beta_1)}{R_0}\right)\left(\mathcal{L}(\boldsymbol{\theta}_{k-1}) - \mathcal{L}^*\right) + \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}X_i \\
&= \tilde{C}_3\eta^2 + \gamma\left(\mathcal{L}(\boldsymbol{\theta}_{k-1}) - \mathcal{L}^*\right) + \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}X_i.
\end{aligned}$$

Note that we can expand the $\mathcal{L}(\boldsymbol{\theta}_{k-1}) - \mathcal{L}^*$ term iteratively to obtain a generic formula for $\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^*$:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^* &\leq \gamma\left(\mathcal{L}(\boldsymbol{\theta}_{k-1}) - \mathcal{L}^*\right) + \eta(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}X_i + \tilde{C}_3\eta^2 \\
&\leq \gamma^k\left(\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*\right) + \eta(1-\beta_1)\sum_{j=1}^{k}\gamma^{k-j}\sum_{i=1}^{j}\beta_1^{j-i}X_i + \sum_{j=1}^{k}\gamma^{k-j}\tilde{C}_3\eta^2 \\
&\leq \gamma^k\left(\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*\right) + \eta(1-\beta_1)\sum_{i=1}^{k}X_i\sum_{j=i}^{k}\gamma^{k-j}\beta_1^{j-i} + C_3\eta,
\end{aligned}$$

where $C_3 = \tilde{C}_3 \cdot \frac{R_0}{2\mu(1-\beta_1)}$. $\square$

Corollary F.1. *Let $\mathcal{L}$ satisfy Assumption 3.2. There exists a constant $\bar{C}$ independent of $\mathcal{L}$, such that $\forall k > 0$, if*

$$\|\nabla \mathcal{L}(\boldsymbol{\theta}_{k-1})\|_2 > \bar{C},$$

then with sufficiently small η, we have

$$\mathcal{L}(\boldsymbol{\theta}_k) < \mathcal{L}(\boldsymbol{\theta}_{k-1}).$$

Proof. Note that (4) can be obtained without PL condition:

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k-1}) \le \tilde{C}_3 \eta^2 - \eta(1-\beta_1) \sum_{i=1}^{k} \beta_1^{k-i} \left(\frac{1}{R_0} \|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\|_2^2 - X_i \right).$$

From Lemma F.1, $|X_i| \le C_{1b} \|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\|_2, \forall i \le k$ where C_{1b} is a constant independent of $\mathcal{L}$. So

$$\frac{1}{R_0} \|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\|_2^2 - X_i \;\ge\; \frac{1}{R_0} \|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\|_2^2 - C_{1b} \|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\|_2 \;\ge\; -\frac{R_0 C_{1b}^2}{4},$$

where the last inequality uses $a^2 - 2ab \ge -b^2$ with $a = \|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\|_2 / \sqrt{R_0}$ and $b = C_{1b}\sqrt{R_0}/2$. Let $G_{i-1} := \|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\|_2$, then

$$\begin{aligned}
\sum_{i=1}^{k} \beta_1^{k-i} \left(\frac{1}{R_0} G_{i-1}^2 - X_i \right) &\ge \left(\frac{1}{R_0} G_{k-1}^2 - C_{1b} G_{k-1} \right) - \frac{R_0 C_{1b}^2}{4} \sum_{i=1}^{k-1} \beta_1^{k-i} \\
&\ge \frac{1}{R_0} G_{k-1}^2 - C_{1b} G_{k-1} - \frac{R_0 C_{1b}^2}{4} \cdot \frac{\beta_1}{1-\beta_1}.
\end{aligned}$$

As long as $\frac{1}{R_0} G_{k-1}^2 - C_{1b} G_{k-1} - \frac{R_0 C_{1b}^2}{4} \cdot \frac{\beta_1}{1-\beta_1} > 0$, a small η can ensure the loss strictly decreases at this step k. Set

$$\bar{C} := \frac{R_0 C_{1b}}{2} \left(1 + \sqrt{\frac{1}{1-\beta_1}} \right),$$

then any $G_{k-1} > \bar{C}$ meets this requirement. Moreover, $\bar{C}$ depends only on (R_0, C_{1b}, β_1) and is independent of $\mathcal{L}$. This proves the corollary. $\square$

Lemma F.4. *Let $\mathcal{L}$ satisfy Assumption 3.2, and assume that μ-PL condition is satisfied at all $\boldsymbol{\theta}_k$ where $k \ge 0$. Let $k \le K = \mathcal{O}(\mathrm{poly}(1/\eta))$ and let*

$$\psi : (\{0, 1, \cdots, k-1\} \times (0,1)) \longrightarrow \mathbb{R}^+$$

be a function. Let $\{X_i\}_{i=1}^k$ be any martingale difference sequence such that:

1. *X_i is $\mathcal{F}_i$-measurable and $\mathbb{E}_{i-1}[X_i] = 0$;*
2. *$|X_i| \le C_{1b} \|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\|_2$ a.s.*

for any $i \in [k]$. If for any $i \in [k]$ and $\delta \in (0,1)$, it holds with probability $1-\delta$ that

$$\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^* \le \psi(i, \delta),$$

then $\forall \delta \in (0,1)$, with probability $1-\delta$, we have $\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^* \le \psi\left(i, \frac{\delta}{2k}\right)$ for all $i \in [k]$, and that

$$\left| \sum_{i=1}^{k} \gamma^{k-i} X_i \right| \le C_4 \sqrt{\sum_{i=1}^{k} \gamma^{2k-2i} \psi\left(i, \frac{\delta}{2k}\right) \log \frac{4}{\delta}}$$

for some constant C_4.

Remark F.1. *The $\{X_i\}$ here may not necessarily equal the $\{X_i\}$ defined in Lemma F.1; we just make it general to benefit future steps. In fact, when we leverage this lemma later, we will multiply that of Lemma F.1 by some scalar $\in (0,1)$.*

Proof. Note that $\sum_{i=1}^{k} \gamma^{k-i} X_i$ is a sum of martingale differences. Moreover, since $\mathcal{L}$ is ρ-smooth and $\exists C_{1b}$ s.t. every $|X_i|$ is bounded by $C_{1b} \left\|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\right\|_2$ (Lemma F.1), we have

$$\begin{aligned}|X_i| &\leq C_{1b} \left\|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\right\|_2 \\ &\leq C_{1b} \sqrt{2\rho \left(\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^*\right)} \\ &\leq C_{1b} \sqrt{2\rho\psi(i, \delta')} \quad \text{ if } \mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^* \leq \psi(i, \delta').\end{aligned}$$

Since $\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^* \leq \psi(i, \delta')$ holds with probability $1-\delta'$ instead of probability 1, we create a new martingale difference sequence that masks out all the positions that exceed the bound. Specifically, we define $X'_{i,\delta'}$ as:

$$X'_{i,\delta'} = \begin{cases} X_i & \text{if } \mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^* \leq \psi(i, \delta'), \\ 0 & \text{else.} \end{cases}$$

This ensures that $\left|X'_{i,\delta'}\right| \leq C_{1b} \sqrt{2\rho\psi(i, \delta')}$ a.s. Then Azuma-Hoeffding's inequality gives us that for any ϵ',

$$\mathbb{P}\left[\left|\sum_{i=1}^{k} \gamma^{k-i} X'_{i,\delta'}\right| \geq \epsilon'\right] \leq 2 \exp\left(\frac{-\epsilon'^2}{4 \sum_{i=1}^{k} C_{1b}^2 \gamma^{2k-2i} \rho\psi(i, \delta')}\right),$$

denoting the right hand side as $\frac{\delta}{2}$ gives that for any δ, with probability $1 - \frac{\delta}{2}$,

$$\left|\sum_{i=1}^{k} \gamma^{k-i} X'_{i,\delta'}\right| \leq \sqrt{4 \sum_{i=1}^{k} C_{1b}^2 \gamma^{2k-2i} \rho\psi(i, \delta') \log \frac{4}{\delta}}.$$

Let $\delta' = \frac{\delta}{2k}$, by union bound, $\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^* \leq \psi\left(i, \frac{\delta}{2k}\right)$ for all $i \in [k]$ with probability $1 - \frac{\delta}{2}$, which also implies $X'_{i,\delta'} = X_i$ for all $i \in [k]$. So with probabilty $1 - \delta$, the following two statements hold simultaneously for all $i \in [k]$:

$$\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^* \leq \psi\left(i, \frac{\delta}{2k}\right)$$

and

$$\begin{aligned}\left|\sum_{i=1}^{k} \gamma^{k-i} X_i\right| &\leq \sqrt{4 \sum_{i=1}^{k} C_{1b}^2 \gamma^{2k-2i} \rho\psi(i, \delta') \log \frac{4}{\delta}} \\ &= C_4 \sqrt{\sum_{i=1}^{k} \gamma^{2k-2i} \psi\left(i, \frac{\delta}{2k}\right) \log \frac{4}{\delta}},\end{aligned}$$

where $C_4 = 2C_{1b}\sqrt{\rho}$. □

Lemma F.5 (Convergence Bound of the AGM Framework)**.** *Let $\mathcal{L}$ satisfy Assumption 3.2, and assume that μ-PL condition is satisfied at all $\boldsymbol{\theta}_k$ where $k \geq 0$. Let η be a small learning rate satisfying $\frac{\beta_1}{\gamma} = \beta_1/(1 - \frac{2\eta\mu(1-\beta_1)}{R_0}) \leq 0.95$. Let $K = \mathcal{O}(\mathrm{poly}(1/\eta))$. Under mild restrictions on K, for any $k \leq K$, $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that*

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^* \leq C_{5a} \cdot \gamma^k \left(\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*\right) + C_{5b} \cdot \eta \log \frac{K}{\delta}$$

for some constants C_{5a}, C_{5b}.

Proof. Denote $D_0 := \mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*$, and denote the bound with $1 - \delta$ probability as $\psi(k, \delta) := \gamma^k C_{5a} D_0 + C_{5b}\eta \log \frac{K}{\delta}$, where the constants C_{5a}, C_{5b} will be specified by us later. We prove by induction. When $k = 0$, the inequality

$$C_{5a} D_0 + C_{5b}\eta \log \frac{K}{\delta} \geq D_0$$

holds trivially as long as $C_{5a} \geq 1$. Now assume that the statement holds for $0, 1, \cdots, k-1$. From Lemma F.3, we have

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^* \leq \gamma^k \left(\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*\right) + \eta(1-\beta_1) \sum_{i=1}^{k} X_i \sum_{j=i}^{k} \gamma^{k-j} \beta_1^{j-i} + C_3 \eta.$$

We can bound the coefficients by

$$\begin{aligned}
\sum_{j=i}^{k} \gamma^{k-j} \beta_1^{j-i} &= \gamma^{k-i} \sum_{j=0}^{k-i} \left(\frac{\beta_1}{\gamma}\right)^j \\
&\leq \gamma^{k-i} \cdot \frac{1}{1 - \frac{\beta_1}{\gamma}} \\
&\leq 20 \gamma^{k-i},
\end{aligned}$$

where the last inequality is due to the assumption $\frac{\beta_1}{\gamma} \leq 0.95$ in the statement. Let $\tilde{X}_i := \frac{\sum_{j=i}^{k} \gamma^{k-j} \beta_1^{j-i}}{20 \gamma^{k-i}} X_i$, then $\left\{\tilde{X}_i\right\}_{i=1}^{k}$ is also a martingale difference sequence and

$$\left|\tilde{X}_i\right| \leq |X_i| \leq C_{1b} \left\|\nabla \mathcal{L}(\boldsymbol{\theta}_i)\right\|_2 \text{ a.s.}$$

From Lemma F.4, with probability $1-\delta$, $\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^* \leq \psi\left(i, \frac{\delta}{2k}\right)$ holds for all $i \in [k]$ and it also holds that

$$\left|\sum_{i=1}^{k} \gamma^{k-i} \tilde{X}_i\right| \leq C_4 \sqrt{\sum_{i=1}^{k} \gamma^{2k-2i} \psi\left(i, \frac{\delta}{2k}\right) \log \frac{4}{\delta}}.$$

The above arguments give

$$\begin{aligned}
&\eta(1-\beta_1) \sum_{i=1}^{k} X_i \sum_{j=i}^{k} \gamma^{k-j} \beta_1^{j-i} \\
&\leq 20 C_4 \eta (1-\beta_1) \sqrt{\sum_{i=1}^{k} \gamma^{2k-2i} \psi\left(i, \frac{\delta}{2k}\right) \log \frac{4}{\delta}} \\
&\leq 20 C_4 \eta (1-\beta_1) \sqrt{\sum_{i=1}^{k} \gamma^{2k-2i} \left(\gamma^i C_{5a} D_0 + C_{5b} \eta \log \frac{2kK}{\delta}\right) \log \frac{4}{\delta}} \\
&\leq 20 C_4 \eta (1-\beta_1) \sqrt{\sum_{i=1}^{k} \gamma^{2k-i} C_{5a} D_0 + \sum_{i=1}^{k} \gamma^{2k-2i} C_{5b} \eta \log \frac{2K^2}{\delta}} \cdot \sqrt{\log \frac{4}{\delta}} \\
&\leq 20 C_4 \eta (1-\beta_1) \sqrt{\frac{\gamma^k C_{5a} D_0}{1-\gamma} + \frac{C_{5b} \eta}{1-\gamma^2} \log \frac{2K^2}{\delta}} \cdot \sqrt{\log \frac{4}{\delta}} \\
&\leq 20 C_4 \eta (1-\beta_1) \left(\sqrt{\frac{\gamma^k C_{5a} D_0}{1-\gamma}} + \sqrt{\frac{C_{5b} \eta}{1-\gamma^2} \log \frac{2K^2}{\delta}}\right) \cdot \sqrt{\log \frac{4}{\delta}}.
\end{aligned}$$

As long as $K \geq \max\{2\delta^2, 4\}$ (which is a mild restriction on K), we have $\log \frac{2K^2}{\delta} \log \frac{4}{\delta} \leq 3 \log^2 \frac{K}{\delta}$ and $\log \frac{4}{\delta} \leq \log \frac{K}{\delta}$. Plugging in $\frac{1}{1-\gamma} = \frac{R_0}{2\mu(1-\beta_1)} \cdot \frac{1}{\eta}$, we have

$$\begin{aligned}
&\eta(1-\beta_1) \sum_{i=1}^{k} X_i \sum_{j=i}^{k} \gamma^{k-j} \beta_1^{j-i} \\
&\leq 20 C_4 (1-\beta_1) \left(\sqrt{\frac{C_{5a} R_0}{2\mu(1-\beta_1)}} \cdot \sqrt{\eta \gamma^k D_0 \log \frac{K}{\delta}} + \eta \sqrt{\frac{3 C_{5b} R_0}{2\mu(1-\beta_1)} \log^2 \frac{K}{\delta}}\right)
\end{aligned}$$

$$\leq 10C_4\sqrt{\frac{2C_{5a}R_0(1-\beta_1)}{\mu}}\cdot\sqrt{\eta\gamma^k D_0\log\frac{K}{\delta}}+10C_4\sqrt{\frac{6C_{5b}R_0(1-\beta_1)}{\mu}}\cdot\eta\log\frac{K}{\delta}$$
$$\leq C_{5c}\gamma^k D_0 + C_{5d}\eta\log\frac{K}{\delta},$$

where

$$C_{5c} = 5C_4\sqrt{2C_{5a}R_0(1-\beta_1)/\mu},$$
$$C_{5d} = 5C_4\sqrt{2C_{5a}R_0(1-\beta_1)/\mu} + 10C_4\sqrt{6C_{5b}R_0(1-\beta_1)/\mu}.$$

Now as long as $K \geq e\delta$ (so that $\log\frac{K}{\delta}\geq 1$), we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_k)-\mathcal{L}^* &\leq \gamma^k\left(\mathcal{L}(\boldsymbol{\theta}_0)-\mathcal{L}^*\right)+\eta(1-\beta_1)\sum_{i=1}^{k}X_i\sum_{j=i}^{k}\gamma^{k-j}\beta_1^{j-i}+C_3\eta \\ &\leq \gamma^k D_0 + C_{5c}\gamma^k D_0 + C_{5d}\eta\log\frac{K}{\delta}+C_3\eta \\ &\leq \left(C_{5c}+1\right)\gamma^k D_0 + \left(C_{5d}+C_3\right)\eta\log\frac{K}{\delta}.\end{aligned}$$

To complete the induction, we need C_{5a}, C_{5b} satisfy

$$\begin{cases} C_{5a} & \geq C_{5c}+1 \quad = 5C_4\sqrt{\frac{2C_{5a}D_0R_0(1-\beta_1)}{\mu}}+1, \\ C_{5b} & \geq C_{5d}+C_3 \quad = 5C_4\sqrt{\frac{2C_{5a}D_0R_0(1-\beta_1)}{\mu}}+10C_4\sqrt{\frac{6C_{5b}R_0(1-\beta_1)}{\mu}}+C_3.\end{cases}$$

Notice that the right-hand side grows at the rate of the square root of C_{5a} and C_{5b}, so there must exist some feasible constants C_{5a} and C_{5b}. Summarizing, under mild restrictions $K \geq \max\left\{2\delta^2, e\delta, 4\right\}$, the statement $\mathcal{L}(\boldsymbol{\theta}_k)-\mathcal{L}^* \leq \gamma^k C_{5a}D_0 + C_{5b}\eta\log\frac{K}{\delta}$ holds with probability $1-\delta$, completing the induction. □

Remark F.2. *The assumption in the statement, $\frac{\beta_1}{\gamma}\leq 0.95$, is very mild since $\beta_1 \leq 0.9$ (Assumption 4.2) and $1-\gamma$ equals a constant multiple of η, so with small η this condition is very easy to satisfy. Moreover, similar to Remark 4.1, the assumed threshold 0.95 can be replaced by any constant below 1, and the order of the convergence rate will remain unaffected.*

F.1 Proof of Convergence-Related Conclusions in Appendix D

Proof of Theorem D.2. This is a direct corollary following from Lemma F.5. The loss function $\mathcal{L}$ is global μ-PL in this case. Setting $k=K$, letting $\gamma^K = \mathcal{O}(\eta)$ gives $K = \mathcal{O}\left(\frac{1}{\eta}\log\frac{1}{\eta}\right)$, completing the proof. □

Now we move on to prove the first part of Theorem D.1. The main difficulty comes from the fact that $\mathcal{L}$ is only guaranteed to satisfy μ-PL condition within some neighborhood Γ^{ϵ_3}; The iteration, once getting out of that neighborhood, cannot be characterized. Hence we need to bound the probability of that event.

The trick here is to construct a proxy loss function $\tilde{\mathcal{L}}$ that, agrees with $\mathcal{L}$ near Γ but has a "wall of quadratic functions" upon $\mathcal{L}$ further away. $\tilde{\mathcal{L}}$ thus satisfies $(\mu,\bar{L})$-PL which allows us to use Lemma F.5. If the losses at all steps are small, this ensures that the iteration never leaves Γ^{ϵ_1}, where $\mathcal{L}$ and $\tilde{\mathcal{L}}$ are identical. We formalize this idea in the sequel.

Lemma F.6 (Tubular Neighborhood Theorem; Theorem 6.24 in Lee (2012))**.** *Let $\Gamma\subset\mathbb{R}^d$ satisfy Assumption 3.4 (in particular, Γ is a $\mathcal{C}^\infty$ compact embedded submanifold). Then there exists $\tau_\Gamma > 0$ such that, writing $N\Gamma$ for the normal bundle,*

$$V \;:=\; \{(p,\boldsymbol{u})\in N\Gamma : \|\boldsymbol{u}\|_2 < \tau_\Gamma\}, \qquad E: V\to\mathbb{R}^d, \quad E(p,\boldsymbol{u}) = p+\boldsymbol{u},$$

the map E is a diffeomorphism onto the open tube $U := \Gamma^{\tau_\Gamma}$. Consequently, the nearest–point projection $P: U\to\Gamma$ is well-defined and C^∞, and every $\boldsymbol{\theta}\in U$ can be written uniquely as

$$\boldsymbol{\theta} \;=\; P(\boldsymbol{\theta})+\nu(\boldsymbol{\theta}), \quad \nu(\boldsymbol{\theta})\in N_{P(\boldsymbol{\theta})}\Gamma.$$

Corollary F.2 (Smooth distance and unit normal on the tube)**.** *In the setting of Lemma F.6, let $U := \Gamma^{\tau_\Gamma}$ and write $E^{-1}(\boldsymbol{\theta}) = (P(\boldsymbol{\theta}), \nu(\boldsymbol{\theta}))$ on U. Define*

$$r(\boldsymbol{\theta}) := \operatorname{dist}(\boldsymbol{\theta}, \Gamma) = \|\nu(\boldsymbol{\theta})\|_2, \qquad \boldsymbol{n}(\boldsymbol{\theta}) := \frac{\nu(\boldsymbol{\theta})}{\|\nu(\boldsymbol{\theta})\|_2} \quad (\boldsymbol{\theta} \in U \setminus \Gamma).$$

Then r and $\boldsymbol{n}$ are $\mathcal{C}^\infty$ on $U \setminus \Gamma$, and

$$\nabla r(\boldsymbol{\theta}) = \boldsymbol{n}(\boldsymbol{\theta}), \qquad \forall \boldsymbol{\theta} \in U \setminus \Gamma. \tag{5}$$

Proof. By Lemma F.6, P and ν are $\mathcal{C}^\infty$ on U, hence so are r and $\boldsymbol{n}$ on $U \setminus \Gamma$. Set $g(\boldsymbol{\theta}) := r(\boldsymbol{\theta})^2 = \|\nu(\boldsymbol{\theta})\|_2^2$. By the chain rule,

$$\nabla g(\boldsymbol{\theta}) \;=\; 2\left(\nabla \nu(\boldsymbol{\theta})\right)^\top \nu(\boldsymbol{\theta}).$$

From the identity $\boldsymbol{\theta} = P(\boldsymbol{\theta}) + \nu(\boldsymbol{\theta})$ we have

$$\nabla \nu(\boldsymbol{\theta}) \;=\; I - \nabla P(\boldsymbol{\theta}).$$

Since $\nabla P(\boldsymbol{\theta})$ maps into $T_{P(\boldsymbol{\theta})}\Gamma$ and $\nu(\boldsymbol{\theta}) \in N_{P(\boldsymbol{\theta})}\Gamma$, it follows that

$$\left(\nabla P(\boldsymbol{\theta})\right)^\top \nu(\boldsymbol{\theta}) \;=\; \mathbf{0},$$

hence

$$\nabla g(\boldsymbol{\theta}) \;=\; 2\left(I - \nabla P(\boldsymbol{\theta})\right)^\top \nu(\boldsymbol{\theta}) \;=\; 2\,\nu(\boldsymbol{\theta}).$$

Therefore, for $\boldsymbol{\theta} \in U \setminus \Gamma$ (so $r(\boldsymbol{\theta}) > 0$),

$$\nabla r(\boldsymbol{\theta}) \;=\; \frac{1}{2\,r(\boldsymbol{\theta})}\,\nabla g(\boldsymbol{\theta}) \;=\; \frac{\nu(\boldsymbol{\theta})}{\|\nu(\boldsymbol{\theta})\|_2} \;=\; \boldsymbol{n}(\boldsymbol{\theta}),$$

which is (5).

□

Lemma F.7 (Nonobtuse angle between $\nabla\mathcal{L}$ and the outward normal)**.** *Let Γ satisfy Assumption 3.4 and let $\mathcal{L}$ satisfy Assumption 3.1 and Assumption 3.2. Then there exists a constant $\tau \in (0, \tau_\Gamma]$ such that, for all $\boldsymbol{\theta} \in \Gamma^\tau$ with nearest-point projection $P(\boldsymbol{\theta})$, distance $r(\boldsymbol{\theta}) := \|\boldsymbol{\theta} - P(\boldsymbol{\theta})\|_2$, and outward unit normal $\boldsymbol{n}(\boldsymbol{\theta}) := (\boldsymbol{\theta} - P(\boldsymbol{\theta}))/r(\boldsymbol{\theta})$, we have*

$$\big\langle \nabla \mathcal{L}(\boldsymbol{\theta}),\, \boldsymbol{n}(\boldsymbol{\theta}) \big\rangle \geq 0. \tag{6}$$

In other words, $\angle\big(\nabla\mathcal{L}(\boldsymbol{\theta}), \boldsymbol{n}(\boldsymbol{\theta})\big) \leq \pi/2$ on $\Gamma^\tau \setminus \Gamma$.

Remark F.3. *This lemma also implies that $t \mapsto \mathcal{L}\big(P(\boldsymbol{\theta}) + t\boldsymbol{n}(\boldsymbol{\theta})\big)$ is non-decreasing on $(0, \tau]$, since $\frac{d}{dt}\mathcal{L}\big(\boldsymbol{\phi} + t\boldsymbol{n}\big) = \langle \nabla\mathcal{L}(\boldsymbol{\phi} + t\boldsymbol{n}), \boldsymbol{n}\rangle \geq 0$ on $(0, \tau]$. To put it vividly, Γ^τ is a valley, with Γ being the floor at the center of it.*

Proof. By Assumption 3.4, each $\boldsymbol{\zeta} \in \Gamma$ is a local minimizer of $\mathcal{L}$, hence $\nabla\mathcal{L}(\boldsymbol{\zeta}) = \mathbf{0}$, and $\nabla^2\mathcal{L}(\boldsymbol{\zeta})$ is positive definite on $N_{\boldsymbol{\zeta}}\Gamma$, with a uniform lower-bound $m > 0$ of its eigenvalues on $N_{\boldsymbol{\zeta}}\Gamma$ (from the compactness of Γ). Let τ_Γ be as in Lemma F.6. For $\boldsymbol{\theta} \in \Gamma^{\tau_\Gamma}$ write $\boldsymbol{\phi} := P(\boldsymbol{\theta})$, $r := \|\boldsymbol{\theta} - \boldsymbol{\phi}\|_2$, and $\boldsymbol{n} := (\boldsymbol{\theta} - \boldsymbol{\phi})/r \in N_{\boldsymbol{\phi}}\Gamma$. Since $\mathcal{L}$ is $\mathcal{C}^5$-smooth, we can perform a third order Taylor expansion:

$$\nabla\mathcal{L}(\boldsymbol{\theta}) = \nabla\mathcal{L}(\boldsymbol{\phi}) + \nabla^2\mathcal{L}(\boldsymbol{\phi})(\boldsymbol{\theta} - \boldsymbol{\phi}) + \boldsymbol{\theta}^{(3)}, \qquad \|\boldsymbol{\theta}^{(3)}\|_2 \leq C^{(3)} r^2,$$

where $C^{(3)}$ is a constant independent of $\boldsymbol{\theta}$. Taking the inner product with $\boldsymbol{n}$ and using $\nabla\mathcal{L}(\boldsymbol{\phi}) = \mathbf{0}$ and $\boldsymbol{\theta} - \boldsymbol{\phi} = r\boldsymbol{n}$, we have

$$\big\langle \nabla\mathcal{L}(\boldsymbol{\theta}), \boldsymbol{n} \big\rangle = r\boldsymbol{n}^\top \nabla^2\mathcal{L}(\boldsymbol{\phi})\boldsymbol{n} + \big\langle \boldsymbol{\theta}^{(3)}, \boldsymbol{n} \big\rangle \geq mr - C^{(3)} r^2.$$

Choose $\tau := \min\{\tau_\Gamma,\, m/C^{(3)}\}$, then for all $r \leq \tau$, we obtain (6). □

Proof of the first part of Theorem D.1. First we construct a tubular neighborhood around Γ and introduce some notations. By Assumption 3.4 , Γ is the unique set of minimizers of $\mathcal{L}$ in some neighborhood U of Γ. By Lemma F.7, there is a $\tau \in (0, \tau_\Gamma]$ for which the nonobtuse condition $\langle \nabla \mathcal{L}(\boldsymbol{\theta}), \boldsymbol{n}(\boldsymbol{\theta}) \rangle \geq 0$ holds on $\Gamma^\tau \setminus \Gamma$. By Lemma E.3, there exists $\epsilon_3 > 0$ such that $\mathcal{L}$ is μ-PL on Γ^{ϵ_3}. Shrinking ϵ_3 if necessary, assume $\epsilon_3 \leq \tau$ and $\Gamma^{\epsilon_3} \subseteq U$.

Throughout this proof we work inside the tube Γ^{τ_Γ} given by Lemma F.6. In particular, for every $\boldsymbol{\theta} \in \Gamma^{\tau_\Gamma}$ we have the nearest–point projection $P(\boldsymbol{\theta}) \in \Gamma$, the normal offset $\nu(\boldsymbol{\theta}) \in N_{P(\boldsymbol{\theta})}\Gamma$, the distance and unit normal

$$r(\boldsymbol{\theta}) := \operatorname{dist}(\boldsymbol{\theta}, \Gamma) = \|\nu(\boldsymbol{\theta})\|_2, \qquad \boldsymbol{n}(\boldsymbol{\theta}) := \frac{\nu(\boldsymbol{\theta})}{\|\nu(\boldsymbol{\theta})\|_2},$$

and (on $U \setminus \Gamma$) the identity $\nabla r(\boldsymbol{\theta}) = \boldsymbol{n}(\boldsymbol{\theta})$ from Corollary F.2.

Next, recall the constant ϵ_1 constructed in Lemma E.3 such that $0 < \epsilon_1 < \epsilon_3$. Define the "gap level"

$$\mathcal{L}_m := \min\left\{\inf\left\{\mathcal{L}(\boldsymbol{\theta}) : r(\boldsymbol{\theta}) \in [\epsilon_1, \epsilon_3]\right\}, \bar{L}\right\}.$$

The set $\{\boldsymbol{\theta} : \ r(\boldsymbol{\theta}) \in [\epsilon_1, \epsilon_3]\}$ is compact and disjoint from Γ, hence $\mathcal{L}_m > \mathcal{L}^*$.

Then we define the proxy objective $\tilde{\mathcal{L}} : \mathbb{R}^d \to \mathbb{R}$ by

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) := \begin{cases} \mathcal{L}(\boldsymbol{\theta}), & \operatorname{dist}(\boldsymbol{\theta}, \Gamma) \leq \epsilon_1, \\ \mathcal{L}(\boldsymbol{\theta}) + \dfrac{C}{2}\big(\operatorname{dist}(\boldsymbol{\theta}, \Gamma) - \epsilon_1\big)^2, & \operatorname{dist}(\boldsymbol{\theta}, \Gamma) > \epsilon_1, \end{cases}$$

where C is a large constant satisfying $C \geq \mu$. Note that for $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) < \mathcal{L}_m \implies \operatorname{dist}(\boldsymbol{\theta}) < \epsilon_1,$$

so on the sublevel set $\{\tilde{\mathcal{L}} < \mathcal{L}_m\}$ we have $\tilde{\mathcal{L}} = \mathcal{L}$.

Now define

$$\bar{L} := \frac{C}{2}(\epsilon_3 - \epsilon_1)^2 + \mathcal{L}^*,$$

and we prove the core property of the proxy loss function: $\tilde{\mathcal{L}}$ is $(\mu, \bar{L})$-PL. First we consider the case $\boldsymbol{\theta} \in \Gamma^{\epsilon_3}$. On Γ^{ϵ_3}, the distance function $r(\boldsymbol{\theta}) = \operatorname{dist}(\boldsymbol{\theta}, \Gamma)$ is defined. Using $\nabla r(\boldsymbol{\theta}) = \boldsymbol{n}(\boldsymbol{\theta})$ from Corollary F.2 and the nonobtuse condition from Lemma F.7, for $r(\boldsymbol{\theta}) > \epsilon_1$,

$$\begin{aligned} \nabla \tilde{\mathcal{L}}(\boldsymbol{\theta}) &= \nabla \mathcal{L}(\boldsymbol{\theta}) + C\big(r(\boldsymbol{\theta}) - \epsilon_1\big)\, \boldsymbol{n}(\boldsymbol{\theta}), \\ \|\nabla \tilde{\mathcal{L}}(\boldsymbol{\theta})\|_2^2 &= \|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 + 2C\big(r(\boldsymbol{\theta}) - \epsilon_1\big)\, \langle \nabla \mathcal{L}(\boldsymbol{\theta}), \boldsymbol{n}(\boldsymbol{\theta}) \rangle + C^2\big(r(\boldsymbol{\theta}) - \epsilon_1\big)^2 \\ &\geq \|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 + C^2\big(r(\boldsymbol{\theta}) - \epsilon_1\big)^2. \end{aligned}$$

Since $\mathcal{L}$ is μ-PL on Γ^{ϵ_3},

$$\|\nabla \tilde{\mathcal{L}}(\boldsymbol{\theta})\|_2^2 \ \geq \ 2\mu\big(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^*\big) + 2C \cdot \frac{C}{2}\big(r(\boldsymbol{\theta}) - \epsilon_1\big)^2 \ \geq \ 2\min\{\mu, C\}\big(\tilde{\mathcal{L}}(\boldsymbol{\theta}) - \mathcal{L}^*\big).$$

For $r(\boldsymbol{\theta}) \leq \epsilon_1$, $\tilde{\mathcal{L}} = \mathcal{L}$ and the μ-PL inequality holds trivially. Our choice of C yields

$$\|\nabla \tilde{\mathcal{L}}(\boldsymbol{\theta})\|_2^2 \ \geq \ 2\mu\big(\tilde{\mathcal{L}}(\boldsymbol{\theta}) - \mathcal{L}^*\big) \qquad \text{for all } \ \boldsymbol{\theta} \in \Gamma^{\epsilon_3},$$

i.e., $\tilde{\mathcal{L}}$ is μ-PL on Γ^{ϵ_3}.

Combining with the fact that $\tilde{\mathcal{L}}(\boldsymbol{\theta}) > \bar{L}$ if $\boldsymbol{\theta} \notin \Gamma^{\epsilon_3}$, we conclude that $\tilde{\mathcal{L}}$ is $(\mu, \bar{L})$-PL.

Finally we come back to prove the main conclusion. Let $\bar{C}$ be the constant constructed in Corollary F.1. Note that for $\boldsymbol{\theta} \in \mathbb{R}^d \setminus \Gamma^{\epsilon_1}$,

$$\begin{aligned} \|\nabla \tilde{\mathcal{L}}(\boldsymbol{\theta})\|_2 \leq \bar{C} &\Rightarrow \|\nabla \mathcal{L}(\boldsymbol{\theta}) + C\big(r(\boldsymbol{\theta}) - \epsilon_1\big)\, \boldsymbol{n}(\boldsymbol{\theta})\|_2 \leq \bar{C} \\ &\Rightarrow \|C\big(r(\boldsymbol{\theta}) - \epsilon_1\big)\, \boldsymbol{n}(\boldsymbol{\theta})\|_2 \leq \bar{C} \quad \text{(from Lemma F.7)} \\ &\Rightarrow r(\boldsymbol{\theta}) \leq \frac{\bar{C}}{C} + \epsilon_1. \end{aligned}$$

Increasing C if necessary, we can let $\bar{C}/C < (\epsilon_3 - \epsilon_1)/2$, which implies

$$\epsilon_m := \frac{\bar{C}}{C} + \epsilon_1 < \frac{\epsilon_1 + \epsilon_3}{2}.$$

Take some $\mathcal{L}_0 > \mathcal{L}^*$ such that $\mathcal{L}_0 - \mathcal{L}^* < (\mathcal{L}_m - \mathcal{L}^*)/2C_{5a}$, where C_{5a} is the constant in Lemma F.5. Take a constant $\epsilon < \epsilon_1$ such that all loss values inside Γ^ϵ are bounded by $\mathcal{L}_0$. There exists a sufficiently small learning rate η such that (let $K = \lfloor (T+1)\eta^{-2} \rfloor$):

1. $C_{5a}\gamma^k (\mathcal{L}_0 - \mathcal{L}^*) + C_{5b}\eta \log \frac{K}{\delta} \le 0.99\mathcal{L}_m - \mathcal{L}^*, \forall k \le K, \delta \in (\eta^{200}, 1)$.
2. $\eta R/\epsilon < \epsilon_3 - \epsilon_m$.

The second property ensures that any single step of update cannot jump from the interior of Γ^{ϵ_m} to the exterior of Γ^{ϵ_3}. However when $\boldsymbol{\theta}_{k-1} \in \Gamma^{\epsilon_3} \setminus \Gamma^{\epsilon_m}$, it follows that $\|\nabla \tilde{\mathcal{L}}(\boldsymbol{\theta})\|_2 > \bar{C}$, so $\mathcal{L}(\boldsymbol{\theta}_k) < \mathcal{L}(\boldsymbol{\theta}_{k-1})$ from Corollary F.1. By induction, we conclude that for any $\boldsymbol{\theta}_0 \in \Gamma^\epsilon$, if we launch an AGM from $\boldsymbol{\theta}_0$ and train using $\tilde{\mathcal{L}}$ and η, all loss values $\tilde{\mathcal{L}}(\boldsymbol{\theta}_k), k \in [0, K-1]$ do not exceed $\bar{L}$, which means the μ-PL condition of $\tilde{\mathcal{L}}$ is satisfied at all $\boldsymbol{\theta}_k$ where $k \in [0, K-1]$. This meets the requirement of Lemma F.5.

By Lemma F.5 and the first property of η, we further conclude that all loss values $\tilde{\mathcal{L}}(\boldsymbol{\theta}_k)$ do not exceed $0.99\mathcal{L}_m$. Finally, by noting that

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) < \mathcal{L}_m \Rightarrow \boldsymbol{\theta} \in \Gamma^{\epsilon_1} \text{ and } \tilde{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}),$$

we conclude from Lemma F.5 that: For any $\boldsymbol{\theta}_0 \in \Gamma^\epsilon$, if we launch an AGM from $\boldsymbol{\theta}_0$ and train using $\mathcal{L}$ and η, for any $k \le K$, $\delta \in (\eta^{200}, 1)$, it holds almost surely that $\boldsymbol{\theta}_k \in \Gamma^{\epsilon_1}$, and it holds with probability at least $1 - \delta$ that

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^* \le C_{5a} \cdot \gamma^k (\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*) + C_{5b} \cdot \eta \log \frac{K}{\delta},$$

and it takes $K_0 = \mathcal{O}(\frac{1}{\eta} \log \frac{1}{\eta})$ time to reach $\mathcal{L}(K_0) - \mathcal{L}^* = \mathcal{O}\left(\eta \log \frac{1}{\eta\delta}\right)$, completing the proof.

□

G Proof of the SDE Approximation of AGMs

In this section, we present a detailed derivation of our slow SDE approximation of the AGM framework as shown in Theorem 4.1. Our slow SDE starts at the time of convergence, so for simplicity, we will "shift the timeline" in this section.

Remark G.1 (Time Shift)**.** *To simplify the notations, we redefine $\boldsymbol{\theta}_0$ and $\boldsymbol{v}_0$ as follows. Starting from Appendix G.1, $\boldsymbol{\theta}_0$ and $\boldsymbol{v}_0$ will no longer represent the parameters that are initialized at the actual beginning of training. Instead, they represent the $\boldsymbol{\theta}_{K_0}$ and $\boldsymbol{v}_{K_0}$ yielded by the first part of Theorem D.1, $(\boldsymbol{\theta}_1, \boldsymbol{v}_1)$ denoting $(\boldsymbol{\theta}_{K_0+1}, \boldsymbol{v}_{K_0+1})$, and so on. Our SDE approximation then describes the dynamics of AGMs after reaching the state $(\boldsymbol{\theta}_0, \boldsymbol{v}_0)$.*

Remark G.2. *Recall for any time step k that $\boldsymbol{\phi}_k := \Phi_{S(\boldsymbol{v}_k)}(\boldsymbol{\theta}_k)$. With the "time shift" described in Remark G.1, the time steps before K_0 will become negative. However in some parts of the following calculation, to deal with the first-order momentum we still need up to $\mathcal{O}(\log \frac{1}{\eta})$ past timesteps. Without loss of generality, we assume that at time K_0 the iteration has already converged for time $\mathcal{O}(\log \frac{1}{\eta})$, i.e., $\forall K_0 - \mathcal{O}(\log \frac{1}{\eta}) \le k \le K_0$,*

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^* = \mathcal{O}\left(\eta \log \frac{1}{\eta}\right),$$

$$\|\boldsymbol{\theta}_k - \boldsymbol{\phi}_k\|_2 = \mathcal{O}\left(\sqrt{\eta \log \frac{1}{\eta}}\right),$$

$$\|\nabla \mathcal{L}(\boldsymbol{\theta}_k)\|_2 = \mathcal{O}\left(\sqrt{\eta \log \frac{1}{\eta}}\right).$$

If not, we simply increase K_0 by $\mathcal{O}(\log \frac{1}{\eta})$ and the argument in the proof of the first part of Theorem D.1 still holds. After the time shift, the range of k above will become $-\mathcal{O}(\log \frac{1}{\eta}) \leq k \leq 0$. Therefore, negative timesteps may appear in the derivation below and they are not typos; We just need their three properties above to control the order of some terms.

G.1 Lemmas for Adaptive Manifold Projection

Before we characterize the projections, we introduce some properties of the preconditioned projection function in this part.

Lemma G.1 (Adaption of Lemma C.2 in Li et al. (2021b))**.** *For any $\boldsymbol{x} \in \mathbb{R}^d$, and any p.d matrix $\boldsymbol{S} \in \mathbb{R}^{d\times d}$, it holds that $\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{x}) = 0$, and*

$$\partial^2\Phi_{\boldsymbol{S}}(\boldsymbol{x})[\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{x}), \boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{x})] = -\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\nabla^2\mathcal{L}(\boldsymbol{x})\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{x}).$$

Proof. We consider a trajectory starting from $\boldsymbol{x}(0) = \boldsymbol{x}$, with an ODE $\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = -\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{x}(t))$, thus by the definition of $\Phi_{\boldsymbol{S}}$, we have $\Phi_{\boldsymbol{S}}(\boldsymbol{x}) = \Phi_{\boldsymbol{S}}(\boldsymbol{x}(t))$, then we have

$$\frac{\mathrm{d}\Phi_{\boldsymbol{S}}(\boldsymbol{x}(t))}{\mathrm{d}t} = -\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{x}) = 0.$$

Further, we take the second derivative of $\Phi_{\boldsymbol{S}}(\boldsymbol{x}(t))$ with repsect to t

$$\frac{\mathrm{d}^2\Phi_{\boldsymbol{S}}(\boldsymbol{x}(t))}{\mathrm{d}t^2} = \partial^2\Phi_{\boldsymbol{S}}(\boldsymbol{x})[\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{x}), \boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{x})] + \partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\nabla^2\mathcal{L}(\boldsymbol{x})\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{x}) = 0.$$

Taking $t = 0$ completes the proof. $\square$

Lemma G.2. *For any $\boldsymbol{x} \in \Gamma$, and a p.d matrix $\boldsymbol{S}$, it holds that $\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\nabla^2\mathcal{L}(\boldsymbol{x}) = \mathbf{0}$.*

Proof. From Lemma C.1 in Li et al. (2021b), we have for $u \in T_x(\Gamma)$, $\nabla^2\mathcal{L}(\boldsymbol{x})u = 0$, and for $u \in T_x^{\perp}(\Gamma)$, it is direct corollary of Lemma 4.3 in Li et al. (2021b) that

$$\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}u = 0.$$

The above identity completes the proof. $\square$

Lemma G.3. *For any $\boldsymbol{x} \in \Gamma$, $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, p.d matrix $\boldsymbol{S}$, and $v \in T_x(\Gamma)$, it holds that*

$$\partial^2\Phi_{\boldsymbol{S}}(\boldsymbol{x})[\boldsymbol{u}\boldsymbol{v}^\top] = -\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\partial^2(\nabla\mathcal{L})(\boldsymbol{x})[\nabla^2\mathcal{L}(\boldsymbol{x})^\dagger\boldsymbol{S}^{-1}\boldsymbol{u}\boldsymbol{v}^\top] - \boldsymbol{S}^{-1}\nabla^2\mathcal{L}(\boldsymbol{x})^\dagger\partial^2(\nabla\mathcal{L})(\boldsymbol{x})[\boldsymbol{S}\partial\Phi(\boldsymbol{x})\boldsymbol{u}\boldsymbol{v}^\top].$$

Proof. We define $\boldsymbol{P} := \boldsymbol{S}^{1/2}$. And we do a reparameterization as $\boldsymbol{x}' := \boldsymbol{P}^{-1}\boldsymbol{x}$, $\mathcal{L}'(\boldsymbol{x}) := \mathcal{L}(\boldsymbol{P}\boldsymbol{x})$, then we have

$$\begin{aligned}
\partial\Phi'(\boldsymbol{x}') &= \boldsymbol{P}\partial\Phi_{\boldsymbol{S}}(\boldsymbol{P}\boldsymbol{x})\boldsymbol{P} \\
\nabla^2 L'(\boldsymbol{x}') &= \boldsymbol{P}\nabla^2 L(\boldsymbol{P}\boldsymbol{x})\boldsymbol{P} \\
\partial^2(\nabla L')(\boldsymbol{x}')[\boldsymbol{M}] &= \boldsymbol{P}\partial^2(\nabla L)(\boldsymbol{P}\boldsymbol{x})[\boldsymbol{P}\boldsymbol{M}\boldsymbol{P}] \\
\partial^2\Phi'(\boldsymbol{x}')[\boldsymbol{M}] &= \boldsymbol{P}\partial^2\Phi(\boldsymbol{x})[\boldsymbol{P}\boldsymbol{M}\boldsymbol{P}].
\end{aligned}$$

Notice that in the space of $\boldsymbol{x}'$, the adaptive projection mapping $\Phi_{\boldsymbol{S}}$ turns into a fixed gradient flow projection. And this allows us to directly apply Lemma C.4 in Li et al. (2021b), which gives

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

A slight modification using the above transformations gives

$$\begin{aligned}
\partial^2\Phi_{\boldsymbol{S}}(\boldsymbol{x})[\boldsymbol{P}\boldsymbol{v}, \boldsymbol{P}\boldsymbol{u}] = &-\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\partial^2(\nabla\mathcal{L})(\boldsymbol{x})[\boldsymbol{P}\boldsymbol{v}, \nabla^2\mathcal{L}(\boldsymbol{x})^\dagger\boldsymbol{S}^{-1}\boldsymbol{P}\boldsymbol{u}] \\
&- \boldsymbol{S}^{-1}\nabla^2\mathcal{L}(\boldsymbol{x})^\dagger\partial^2(\nabla\mathcal{L})(\boldsymbol{x})[\boldsymbol{P}\boldsymbol{v}, \boldsymbol{S}\partial\Phi(\boldsymbol{x})\boldsymbol{P}\boldsymbol{u}].
\end{aligned}$$

We now redefine $\boldsymbol{u} = \boldsymbol{P}\boldsymbol{u}$, $\boldsymbol{v} = \boldsymbol{P}\boldsymbol{v}$, and we organize the above equation

$$\begin{aligned}
\partial^2\Phi_{\boldsymbol{S}}(\boldsymbol{x})[\boldsymbol{u}\boldsymbol{v}^\top] = &-\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\partial^2(\nabla\mathcal{L})(\boldsymbol{x})[\nabla^2\mathcal{L}(\boldsymbol{x})^\dagger\boldsymbol{S}^{-1}\boldsymbol{u}\boldsymbol{v}^\top] \\
&- \boldsymbol{S}^{-1}\nabla^2\mathcal{L}(\boldsymbol{x})^\dagger\partial^2(\nabla\mathcal{L})(\boldsymbol{x})[\boldsymbol{S}\partial\Phi(\boldsymbol{x})\boldsymbol{u}\boldsymbol{v}^\top].
\end{aligned}$$

We completes the proof. $\square$

G.2 Iteration Stays Near Manifold

Now we begin the final preparations before deriving the slow SDE near the manifold. Note that in the end of convergence analysis, the total steps equal $K = \lfloor (T+1)\eta^{-2} \rfloor$ and the converging step $K_0 = \mathcal{O}(\frac{1}{\eta}\log\frac{1}{\eta})$. So after time shifting, the high probability convergence of $\lfloor (T+1)\eta^{-2} \rfloor - \mathcal{O}(\frac{1}{\eta}\log\frac{1}{\eta}) > \lfloor T\eta^{-2} \rfloor$ steps are ensured in Lemma F.5. Now denote $K := \lfloor T\eta^{-2} \rfloor$ be the total number of steps in our analysis. Let β be some constant in $(0, 0.5)$, whose exact value will be specified later. First, we bound the movement of projected steps by showing that $\boldsymbol{\phi}$ shifts no more than $\tilde{\mathcal{O}}(\eta^{0.5-0.5\beta})$ within $\Delta K := \lfloor \eta^{-1-\beta} \rfloor$ steps, demonstrating the "slowness" of the dynamics of AGMs after the projection.

Lemma G.4. *If $\boldsymbol{\theta}_k$ stays inside Γ^{ϵ_1} for any $k \in [0, K]$, then for any $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability $1-\delta$, for any $k \in [0, K-\Delta K]$, $\Delta k \in [\Delta K]$,*

$$\|\boldsymbol{\phi}_{k+\Delta k} - \boldsymbol{\phi}_k\|_2 \le C_6 \eta^{0.5-0.5\beta}\sqrt{\log\frac{1}{\eta\delta}}$$

for some constant C_6.

Proof. Recall from Lemma E.3 that $\Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})$ is $\mathcal{C}^4$ on $\mathcal{X}^{\epsilon_1} := \Gamma^{\epsilon_1} \times \mathbb{R}^d_{[0,R_1]}$. Since $\mathcal{X}^{\epsilon_1}$ is compact, $\Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})$ is then bounded and Lipschitz on $\mathcal{X}^{\epsilon_1}$. Similarly, $\partial\Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})$ is bounded and Lipschitz on $\mathcal{X}^{\epsilon_1}$. For any $k \in [0, K)$, let $\bar{k} = k - 2\log_{\beta_1}\eta$, we have:

$$\begin{aligned}
\boldsymbol{\phi}_{k+1} - \boldsymbol{\phi}_k &= \Phi_{S(\boldsymbol{v}_{k+1})}(\boldsymbol{\theta}_{k+1}) - \Phi_{S(\boldsymbol{v}_k)}(\boldsymbol{\theta}_k) \\
&= \Phi_{S(\boldsymbol{v}_{\bar{k}})}(\boldsymbol{\theta}_{k+1}) - \Phi_{S(\boldsymbol{v}_{\bar{k}})}(\boldsymbol{\theta}_k) + \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right) \\
&= \partial\Phi_{S(\boldsymbol{v}_{\bar{k}})}(\boldsymbol{\theta}_k)(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k) + \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right) \\
&= \partial\Phi_{S(\boldsymbol{v}_{\bar{k}})}(\boldsymbol{\theta}_k)(\eta S(\boldsymbol{v}_{k+1})\boldsymbol{m}_{k+1}) + \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right) \\
&= \partial\Phi_{S(\boldsymbol{v}_{\bar{k}})}(\boldsymbol{\theta}_{\bar{k}})(\eta S(\boldsymbol{v}_{\bar{k}})\boldsymbol{m}_{k+1}) + \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right),
\end{aligned}$$

where the second equality comes from the fact that one step of update on $\boldsymbol{v}$ is of $\mathcal{O}(\eta^2)$ and the Lipschitzness of S and Φ, the third equality comes from $\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2 = \mathcal{O}(\eta)$, and the last equality follows from the boundedness and Lipschitzness of $\partial\Phi$. We can decompose $\boldsymbol{m}_k$ as:

$$\begin{aligned}
\boldsymbol{m}_{k+1} &= (1-\beta_1)\sum_{i=\bar{k}}^{k}\beta_1^{k-i}(\nabla\mathcal{L}(\boldsymbol{\theta}_i) + \boldsymbol{z}_i) + \mathcal{O}(\eta^2) \\
&= (1-\beta_1)\sum_{i=\bar{k}}^{k}\beta_1^{k-i}\left(\nabla\mathcal{L}(\boldsymbol{\theta}_{\bar{k}}) + \mathcal{O}\left(\eta\log\frac{1}{\eta}\right)\right) + (1-\beta_1)\sum_{i=\bar{k}}^{k}\beta_1^{k-i}\boldsymbol{z}_i + \mathcal{O}(\eta^2).
\end{aligned}$$

A key observation is that $\partial\Phi_{S(\boldsymbol{v}_{\bar{k}})}(\boldsymbol{\theta}_{\bar{k}})S(\boldsymbol{v}_{\bar{k}})\nabla\mathcal{L}(\boldsymbol{\theta}_{\bar{k}}) = 0$ from Lemma G.2, which allows us to view $\boldsymbol{\phi}_{k+1} - \boldsymbol{\phi}_k$ as $\sum_{i=\bar{k}}^{k}\tilde{\boldsymbol{z}}_{k,i} + \mathcal{O}(\eta^2\log\frac{1}{\eta})$ where $\tilde{\boldsymbol{z}}_{k,i} = \partial\Phi_{S(\boldsymbol{v}_{\bar{k}})}(\boldsymbol{\theta}_{\bar{k}})(\eta(1-\beta_1)\beta_1^{k-i}S(\boldsymbol{v}_{\bar{k}})\boldsymbol{z}_i)$. Note that $\tilde{\boldsymbol{z}}_{k,i}$ is $\mathcal{F}_{i+1}$-measurable and its mean is $\mathbf{0}$, since $\tilde{\boldsymbol{z}}_{k,i}$ just applies a linear tensor transformation to $\boldsymbol{z}_i$. If we define a constant $C_{6a} := \sup\left\{\left\|\partial\Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})\right\|_2 \mid (\boldsymbol{v}, \boldsymbol{\theta}) \in \mathcal{X}^{\epsilon_1}\right\} \cdot (1-\beta_1) \cdot \epsilon^{-1}$ that is independent of k and i, then $\|\tilde{\boldsymbol{z}}_{k,i}\|_2$ is almost surely bounded by $\eta\beta_1^{k-i}C_{6a}\|\boldsymbol{z}_i\|_2$.

For any $k \in [0, K-\Delta K]$ and $\Delta k \in [\Delta K]$, we have

$$\begin{aligned}
\boldsymbol{\phi}_{k+\Delta k} - \boldsymbol{\phi}_k &= \sum_{j=k}^{k+\Delta k-1}(\boldsymbol{\phi}_{j+1} - \boldsymbol{\phi}_j) \\
&= \sum_{j=k}^{k+\Delta k-1}\left(\sum_{i=j-2\log_{\beta_1}\eta}^{j}\tilde{\boldsymbol{z}}_{j,i} + O\left(\eta^2\log\frac{1}{\eta}\right)\right)
\end{aligned}$$

$$= \sum_{i=k-2\log_{\beta_1}\eta}^{k+\Delta k-1} \sum_{j=i}^{\min\left\{k+\Delta k-1, j+2\log_{\beta_1}\eta\right\}} \tilde{\boldsymbol{z}}_{j,i} + \tilde{\mathcal{O}}(\eta^{1-\beta})$$

Denote $\boldsymbol{Z}_i := \sum_{j=i}^{\min\left\{k+\Delta k-1, j+2\log_{\beta_1}\eta\right\}} \tilde{\boldsymbol{z}}_{j,i}$, then each $\boldsymbol{Z}_i$ is a linear transformation of $\boldsymbol{z}_i$ so it is with zero mean, and also $\|\boldsymbol{Z}_i\|_2 \le \eta \cdot \frac{C_{6a}}{1-\beta_1}\|\boldsymbol{z}_i\|_2 \le \eta \cdot \frac{C_{6a}R}{1-\beta_1}$ a.s. Azuma-Hoeffding's inequality then gives that for any $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability $1-\delta$,

$$\begin{aligned}\boldsymbol{\phi}_{k+\Delta k} - \boldsymbol{\phi}_k &\le \sqrt{2\eta^2\left(\frac{C_{6a}R}{1-\beta_1}\right)^2 \cdot \left(R_{\mathrm{grp}}H + 2\log_{\beta_1}\eta\right) \cdot \log\frac{2}{\delta}} \\ &\le C_{6b}\sqrt{\eta^{1-\beta}\log\frac{2}{\delta}}\end{aligned}$$

for some constant C_{6b}. Finally, plugging in $\delta' = \frac{\delta}{K\cdot\Delta K}$ and taking union bound over all $k \in [0, K-\Delta K]$ and $\Delta k \in [\Delta K]$ gives the theorem. □

With the concentration bounds so far, we can show that the dynamics behaves "well" during the whole iteration, and we formalize this idea below.

Definition G.1 (δ-good)**.** *For any $\delta = \mathcal{O}(\mathrm{poly}(\eta))$ and any step $\hat{K} \in [K]$, we define step $\hat{K}$ to be δ-good if and only if the simultaneous establishment of the following statements:*

1. *For any $k \in [0, \hat{K}]$, $\boldsymbol{\phi}_k \in \Gamma$ and $\|\boldsymbol{\theta}_k - \boldsymbol{\phi}_k\|_2 \le C_{8a}\sqrt{\eta\log\frac{1}{\eta\delta}}$.*
2. *For any $k \in [0, \hat{K}-\Delta K]$, $\Delta k \in [\Delta K]$, $\|\boldsymbol{\phi}_{k+\Delta k} - \boldsymbol{\phi}_k\|_2 \le C_{8b}\eta^{0.5-0.5\beta}\sqrt{\log\frac{1}{\eta\delta}}$.*

Here C_{8a} and $C_{8b} = C_6\sqrt{2}$ are two constants.

Lemma G.5. *When η is sufficiently small, with probability $1-\eta^{100}$, the event η^{100}-good holds for any step $\hat{K}$ in $[K]$.*

Proof. Denote $\delta := \eta^{100}$. From Lemma F.5, with probability $1-\delta/2$, all $k \in [0, K]$ satisfy $\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^* \le \mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^* + C_{5b}\eta\log\frac{2K^2}{\delta}$. Note that $D_0 := \mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*$ is of $\mathcal{O}(\eta\log\frac{1}{\eta\delta})$ since time 0 now refer to the time after convergence. Combining Lemma E.1, this implies $\|\boldsymbol{\theta}_k - \boldsymbol{\phi}_k\|_2 \le \frac{2C_2}{\sqrt{2\mu}C_1} \cdot \sqrt{C_{5a}D_0 + C_{5b}\eta\log\frac{2K^2}{\delta}}$ for any $k \in [0, K]$. When η is small enough such that $\|\boldsymbol{\theta}_k - \boldsymbol{\phi}_k\|_2 \le \frac{2C_2}{\sqrt{2\mu}C_1} \cdot \sqrt{C_{5a}D_0 + C_{5b}\eta\log\frac{2K^2}{\delta}} + \eta R/\epsilon < \epsilon_2$, any $\boldsymbol{\phi}_k \in \Gamma$ with $k \ge 0$ will imply $\boldsymbol{\phi}_{k+1} \in \Gamma$, since $\boldsymbol{\theta}_{k+1}$ cannot escape Γ^{ϵ_2}. Giving $\boldsymbol{\phi}_0 \in \Gamma$ and using induction, we conclude that all $\boldsymbol{\phi}_k \in \Gamma$ for $k \ge 0$.

When the above holds, the requirement of Lemma G.4 is met. Then with probability $1-\delta/2$, for any $k \in [0, K-\Delta K]$, $\Delta k \in [\Delta K]$, we have $\|\boldsymbol{\phi}_{k+\Delta k} - \boldsymbol{\phi}_k\|_2 \le C_6\eta^{0.5-0.5\beta}\sqrt{\log\frac{2}{\eta\delta}}$.

Finally, we just take the union of Lemma F.5 and Lemma G.4. With $\log\frac{2K^2}{\delta} \le 8\log\frac{1}{\eta\delta}$ and $\log\frac{2}{\eta\delta} \le 2\log\frac{1}{\eta\delta}$ (which are mild restrictions since η is small), we have the theorem. □

We have proved that our iteration will behave well with high probability, but chances still exist that the iteration is driven out of working zones and becomes intractable. We define a well-behaved sequence that manually redirects the iteration when extreme cases happen.

Definition G.2 (Well-behaved Sequence)**.** *Denote the event of step k being η^{100}-good as $\mathcal{E}_k$. Let $\boldsymbol{\phi}_{\mathrm{null}}$ be a fixed point on Γ. Starting from $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$ and $\hat{\boldsymbol{v}}_0 = \boldsymbol{v}_0$, we define a sequence of $(\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{v}}_k, \hat{\boldsymbol{m}}_k)$ as follows:*

$$\begin{aligned}\hat{\boldsymbol{m}}_{k+1} &:= \beta_1\hat{\boldsymbol{m}}_k + (1-\beta_1)(\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{z}_k) \\ \hat{\boldsymbol{v}}_{k+1} &:= \beta_2\hat{\boldsymbol{v}}_k + (1-\beta_2)V((\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{z}_k)(\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{z}_k)^\top) \\ \hat{\boldsymbol{\theta}}_{k+1} &:= \mathbf{1}_{\mathcal{E}_k}\boldsymbol{\theta}_{k+1} + \mathbf{1}_{\bar{\mathcal{E}}_k}\boldsymbol{\phi}_{\mathrm{null}},\end{aligned}$$

where $\mathbf{1}$ is the indicator function: $\mathbf{1}_{\mathcal{E}} = 1$ if event $\mathcal{E}$ happens and $\mathbf{1}_{\mathcal{E}} = 0$ otherwise.

Note that the update of $\hat{\boldsymbol{\theta}}_k$ can be written as

$$\hat{\boldsymbol{\theta}}_{k+1} := \hat{\boldsymbol{\theta}}_k - \eta S(\hat{\boldsymbol{v}}_{k+1})\hat{\boldsymbol{m}}_{k+1} \underbrace{-\mathbf{1}_{\bar{\mathcal{E}}_k}(\hat{\boldsymbol{\theta}}_k - \eta S(\hat{\boldsymbol{v}}_{k+1})\hat{\boldsymbol{m}}_{k+1}) + \mathbf{1}_{\bar{\mathcal{E}}_k}\boldsymbol{\phi}_{\mathrm{null}}}_{:=\boldsymbol{e}_k}$$

where $\boldsymbol{e}_k$ denotes the redirection under extreme cases. By definition, $\boldsymbol{e}_k$ equals zero in the vast majority of cases, and in other cases it's still bounded by a constant, so all moments of $\boldsymbol{e}_k$ are within $\mathcal{O}(\eta^{100})$ which is negligibly small.

G.3 Moment Calculation of AGMs Near Manifold

Additional Notations. To utilize the analysis framework in Gu et al. (2023b), we first introduce some notations needed. Consistent with Gu et al. (2023b), we pretend that AGMs proceed with $H = \frac{1}{\eta}$ local steps, as a single worker (without multiple workers). We denote every H steps as one round. Next, we define a "giant step", which encompasses $R_{\mathrm{grp}} = \frac{1}{\eta^{\beta}}$ rounds, corresponding to $R_{\mathrm{grp}} \cdot H$ steps. We consider a total timescope of $\frac{T}{\eta^2}$ steps, which corresponds to $\frac{T}{\eta^{1-\beta}}$ giant steps.

For any $0 \le s < R_{\mathrm{grp}}$ and $0 \le t \le H$, we use $\hat{\boldsymbol{\theta}}_t^{(s)}$ and $\hat{\boldsymbol{\theta}}_k$ (where $k = sH + t$) exchangeably to denote the parameter we get on the t-th local step of round s, which is also the k-th global step. Also note that for any $0 \le s < R_{\mathrm{grp}}$, $\hat{\boldsymbol{\theta}}_H^{(s)}$ and $\hat{\boldsymbol{\theta}}_0^{(s+1)}$ refer to the same thing. We define the notation $\hat{\boldsymbol{v}}_t^{(s)}$, $\hat{\boldsymbol{m}}_t^{(s)}$ and $\mathcal{E}_t^{(s)}$ in the same way as we did for $\boldsymbol{\theta}$. Furthermore, we define

$$\hat{\boldsymbol{g}}_t^{(s)} := \nabla \ell_t^{(s)}\left(\hat{\boldsymbol{\theta}}_t^{(s)}\right),\ \hat{\boldsymbol{S}}_k = S\left(\hat{\boldsymbol{v}}_k\right),\ \hat{\boldsymbol{S}}_t^{(s)} := S\left(\hat{\boldsymbol{v}}_t^{(s)}\right),\ \hat{\boldsymbol{S}}^{(s)} := \hat{\boldsymbol{S}}_0^{(s)},\ \hat{\boldsymbol{\phi}}^{(s)} := \boldsymbol{\Phi}_{\hat{\boldsymbol{S}}^{(s)}}\left(\hat{\boldsymbol{\theta}}_0^{(s)}\right),$$

$$\hat{\boldsymbol{x}}_t^{(s)} := \hat{\boldsymbol{\theta}}_t^{(s)} - \hat{\boldsymbol{\phi}}^{(s)},\ \Delta\hat{\boldsymbol{\phi}}^{(s)} := \hat{\boldsymbol{\phi}}^{(s)} - \hat{\boldsymbol{\phi}}^{(0)},\ \boldsymbol{\Sigma}_0 := \boldsymbol{\Sigma}(\hat{\boldsymbol{\phi}}^{(0)}),\ \boldsymbol{P}_{\parallel} := \partial\boldsymbol{\Phi}_{\hat{\boldsymbol{S}}^{(0)}}(\hat{\boldsymbol{\phi}}^{(0)}),\ \boldsymbol{P}_{\perp} := \boldsymbol{I} - \boldsymbol{P}_{\parallel},$$

$$\hat{\boldsymbol{q}}_t^{(s)} := \mathbb{E}\left[\hat{\boldsymbol{x}}_t^{(s)}\right],\ \hat{\boldsymbol{A}}_t^{(s)} := \mathbb{E}\left[\hat{\boldsymbol{x}}_t^{(s)}\hat{\boldsymbol{x}}_t^{(s)\top}\right],\ \hat{\boldsymbol{B}}_t^{(s)} := \mathbb{E}\left[\hat{\boldsymbol{x}}_t^{(s)}\Delta\hat{\boldsymbol{\phi}}^{(s)\top}\right].$$

Corollary G.1. *There exist constants C_{9a}, C_{9b}, C_{9c} such that for all $0 \le s < R_{\mathrm{grp}}$, $0 \le t \le H$,*

$$\left\|\hat{\boldsymbol{x}}_t^{(s)}\right\|_2 \le C_{9a}\sqrt{\eta \log \frac{1}{\eta}},$$

$$\left\|\hat{\boldsymbol{\theta}}_t^{(s)} - \hat{\boldsymbol{\theta}}_0^{(s)}\right\|_2 \le C_{9b}\sqrt{\eta \log \frac{1}{\eta}},$$

$$\left\|\hat{\boldsymbol{\phi}}^{(s)} - \hat{\boldsymbol{\phi}}^{(0)}\right\|_2 \le C_{9c}\eta^{0.5-0.5\beta}\sqrt{\log \frac{1}{\eta}}.$$

Proof. Substituting $\delta = \eta^{100}$. When $\mathcal{E}$ holds, this follows directly from the definition of δ-goodness; Otherwise, all $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$ are equal, and these quantities are equal to $\mathbf{0}$. □

Impact of Momentum. Our conclusion regrading to the impact of Momentum on the implicit bias is similar to the conclusion in Wang et al. (2023): It does not impact the implicit bias. Further, our analysis is based on moment methods and can give exact error bounds. First, we state some technical lemmas in order to show that introducing momentum will not cause the gradient to deviate too much from itself, i.e. $\mathbb{E}[\hat{\boldsymbol{m}}_t]$ is close to $\mathbb{E}[\hat{\boldsymbol{g}}_t]$. Once this guarantee is established, we can replace $\hat{\boldsymbol{m}}_t$ with $\hat{\boldsymbol{g}}_t$ in the moment calculation to simplify it. The general idea of the proof is to show that if i is close to t, then $\mathbb{E}[\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_{i-1})]$ will become close to $\mathbb{E}[\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_{t-1})]$, and if i is far from t, then the contribution of $\mathbb{E}[\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_{i-1})]$ would be negligible in $\mathbb{E}[\hat{\boldsymbol{m}}_t]$.

Lemma G.6. *For any $k \ge 0$, we have*

$$\left\|\mathbb{E}\left[\nabla\mathcal{L}\left(\hat{\boldsymbol{\theta}}_{k+1}\right) - \nabla\mathcal{L}\left(\hat{\boldsymbol{\theta}}_k\right)\right]\right\|_2 \le C_{10}\eta^{1.5}$$

for some constant C_{10}.

Proof. We have

$$\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_{k+1}) - \nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_k) = \nabla^2\mathcal{L}(\hat{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}}_{k+1} - \hat{\boldsymbol{\theta}}_k) + \mathcal{O}\left(\left\|\hat{\boldsymbol{\theta}}_{k+1} - \hat{\boldsymbol{\theta}}_k\right\|_2^2\right)$$
$$= \nabla^2\mathcal{L}(\hat{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}}_{k+1} - \hat{\boldsymbol{\theta}}_k) + \mathcal{O}(\eta^2) + \mathcal{O}(\|\boldsymbol{e}_k\|_2),$$

since $\left\|\hat{\boldsymbol{\theta}}_{k+1} - \hat{\boldsymbol{\theta}}_k\right\|_2 = \|\eta S(\hat{\boldsymbol{v}}_k)\hat{\boldsymbol{m}}_k - \boldsymbol{e}_k\|_2 = \mathcal{O}(\eta) + \mathcal{O}(\|\boldsymbol{e}_k\|_2)$. Let $\bar{k} = k - \log_{\beta_1}(\eta)$ be a threshold that is logarithmically close to k, then we have

$$\nabla^2\mathcal{L}(\hat{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}}_{k+1} - \hat{\boldsymbol{\theta}}_k) = \left(\nabla^2\mathcal{L}(\hat{\boldsymbol{\theta}}_{\bar{k}}) + \mathcal{O}\left(\left\|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{\bar{k}}\right\|_2\right)\right)\left(\hat{\boldsymbol{\theta}}_{k+1} - \hat{\boldsymbol{\theta}}_k\right)$$
$$= \nabla^2\mathcal{L}(\hat{\boldsymbol{\theta}}_{\bar{k}})\left(\hat{\boldsymbol{\theta}}_{k+1} - \hat{\boldsymbol{\theta}}_k\right) + \mathcal{O}\left(\eta\cdot\log_{\beta_1}(\eta)\cdot\eta\right) + \mathcal{O}(\|\boldsymbol{e}_k\|_2)$$
$$= \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{\theta}}_{\bar{k}})S(\hat{\boldsymbol{v}}_{k+1})\hat{\boldsymbol{m}}_{k+1} + \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right) + \mathcal{O}(\|\boldsymbol{e}_k\|_2).$$

Recentering the Hessian term to $\hat{\boldsymbol{\theta}}_{\bar{k}}$ allows us to take conditional expectation $\mathbb{E}_{\bar{k}}$ on $S(\hat{\boldsymbol{v}}_{k+1})\hat{\boldsymbol{m}}_{k+1}$:

$$\mathbb{E}\left[\nabla^2\mathcal{L}(\hat{\boldsymbol{\theta}}_{\bar{k}})S(\hat{\boldsymbol{v}}_{k+1})\hat{\boldsymbol{m}}_{k+1}\right] = \mathbb{E}\left[\nabla^2\mathcal{L}(\hat{\boldsymbol{\theta}}_{\bar{k}})\mathbb{E}_{\bar{k}}\left[S(\hat{\boldsymbol{v}}_{k+1})\hat{\boldsymbol{m}}_{k+1}\right]\right].$$

After that, notice that

$$\begin{aligned}\|\mathbb{E}_{\bar{k}}\left[S(\hat{\boldsymbol{v}}_{k+1})\hat{\boldsymbol{m}}_{k+1}\right]\|_2 &= \|\mathbb{E}_{\bar{k}}\left[S(\mathbb{E}_{\bar{k}}\left[\hat{\boldsymbol{v}}_{k+1}\right])\hat{\boldsymbol{m}}_{k+1}\right]\|_2 + \mathcal{O}(\|\hat{\boldsymbol{v}}_{k+1} - \mathbb{E}_{\bar{k}}\left[\hat{\boldsymbol{v}}_{k+1}\right]\|_2)\\ &= \|S(\mathbb{E}_{\bar{k}}\left[\hat{\boldsymbol{v}}_{k+1}\right])\mathbb{E}_{\bar{k}}[\hat{\boldsymbol{m}}_{k+1}]\|_2 + \mathcal{O}(\|\hat{\boldsymbol{v}}_{k+1} - \mathbb{E}_{\bar{k}}\left[\hat{\boldsymbol{v}}_{k+1}\right]\|_2)\\ &= \mathcal{O}(\underbrace{\|\mathbb{E}_{\bar{k}}[\hat{\boldsymbol{m}}_{k+1}\|_2}_{=:D_1}) + \mathcal{O}(\underbrace{\|\hat{\boldsymbol{v}}_{k+1} - \mathbb{E}_{\bar{k}}\left[\hat{\boldsymbol{v}}_{k+1}\right]\|_2}_{=:D_2})\end{aligned}$$

since S and $\hat{\boldsymbol{m}}$ are both bounded by constant scale. We figure out the orders of these two terms respectively:

$$\begin{aligned}D_1 &= \left\|\mathbb{E}_{\bar{k}}\left[\beta_1^{k-\bar{k}+1}\hat{\boldsymbol{m}}_{\bar{k}} + (1-\beta_1)\sum_{i=\bar{k}}^{k}\beta_1^{k-i}\hat{\boldsymbol{g}}_i\right]\right\|_2\\ &= \mathcal{O}\left(\beta_1^{\log_{\beta_1}(\eta)}\right) + \left\|\mathbb{E}_{\bar{k}}\left[(1-\beta_1)\sum_{i=\bar{k}}^{k}\beta_1^{k-i}\hat{\boldsymbol{g}}_i\right]\right\|_2\\ &= \mathcal{O}(\eta) + \left\|\mathbb{E}_{\bar{k}}\left[(1-\beta_1)\sum_{i=\bar{k}}^{k}\beta_1^{k-i}\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_i)\right]\right\|_2\\ &= \mathcal{O}(\eta) + \mathcal{O}(\eta^{0.5}) = \mathcal{O}(\eta^{0.5})\end{aligned}$$

since $\nabla\mathcal{L}$ is uniformly bounded by $\mathcal{O}(\eta^{0.5})$ after convergence (see Lemma F.5); And

$$\begin{aligned}D_2 &= (1-\beta_2)\sum_{i=\bar{k}}^{k}\beta_2^{k-i}\left(V(\hat{\boldsymbol{g}}_i\hat{\boldsymbol{g}}_i^\top) - \mathbb{E}_{\bar{k}}\left[V(\hat{\boldsymbol{g}}_i\hat{\boldsymbol{g}}_i^\top)\right]\right)\\ &= \mathcal{O}\left(b_2\cdot(k-\bar{k})\right)\\ &= \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right),\end{aligned}$$

since V is bounded by a constant scale. Now combining the above together, we have

$$\begin{aligned}\left\|\mathbb{E}[\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_{k+1}) - \nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_k)]\right\|_2 &= \eta\mathbb{E}\left[\nabla^2\mathcal{L}(\hat{\boldsymbol{\theta}}_{\bar{k}})\mathbb{E}_{\bar{k}}\left[S(\hat{\boldsymbol{v}}_{k+1})\hat{\boldsymbol{m}}_{k+1}\right]\right] + \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right)\\ &\quad + \mathcal{O}(\mathbb{E}[\|e_k\|_2])\\ &= \eta\cdot\mathcal{O}(D_1+D_2) + \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right) + \mathcal{O}(\eta^{100})\\ &= \mathcal{O}(\eta^{1.5}),\end{aligned}$$

which concludes the proof. ∎

With Lemma G.6, we are ready to deduce the closeness between $\mathbb{E}[\hat{\boldsymbol{m}}_k]$ and $\mathbb{E}[\hat{\boldsymbol{g}}_k]$.

Lemma G.7. *For any $k \geq 0$, let $\bar{k} = k - 2\log_{\beta_1}(\eta)$, we have*

$$\|\mathbb{E}_{\bar{k}}[\hat{\boldsymbol{m}}_{k+1} - \hat{\boldsymbol{g}}_{k+1}]\|_2 \leq C_{11}\eta^{1.5}\log\frac{1}{\eta}, \quad \textit{a.s.}$$

Note that this also implies that $\|\mathbb{E}[\hat{\boldsymbol{m}}_{k+1} - \hat{\boldsymbol{g}}_{k+1}]\|_2 \leq C_{11}\eta^{1.5}\log\frac{1}{\eta}$.

Proof. Expanding $\mathbb{E}_{\bar{k}}[\hat{\boldsymbol{m}}_{k+1}]$, we have

$$\begin{aligned}
\mathbb{E}_{\bar{k}}[\hat{\boldsymbol{m}}_{k+1}] &= \mathbb{E}_{\bar{k}}\left[(1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}\hat{\boldsymbol{g}}_i\right] \\
&= (1-\beta_1)\sum_{i=1}^{\bar{k}-1}\beta_1^{k-i}\hat{\boldsymbol{g}}_i + (1-\beta_1)\sum_{i=\bar{k}}^{k}\beta_1^{k-i}\mathbb{E}_{\bar{k}}[\hat{\boldsymbol{g}}_i] \\
&= \underbrace{(1-\beta_1)\sum_{i=1}^{\bar{k}-1}\beta_1^{k-i}\hat{\boldsymbol{g}}_i}_{:=E_1} + \underbrace{(1-\beta_1)\sum_{i=\bar{k}}^{k}\beta_1^{k-i}\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_i)}_{:=E_2}
\end{aligned}$$

Note that E_1 is neglegible:

$$\begin{aligned}
\|E_1\|_2 &= \left\|(1-\beta_1)\sum_{i=1}^{\bar{k}-1}\beta_1^{k-i}\hat{\boldsymbol{g}}_i\right\|_2 \\
&= (1-\beta_1)\sum_{i=1}^{\bar{k}-1}\beta_1^{k-i}\cdot\mathcal{O}(1) \\
&\leq (1-\beta_1)\sum_{i=2\log_{\beta_1}(\eta)}^{\infty}\beta_1^{i}\cdot\mathcal{O}(1) \\
&= \mathcal{O}\left(\beta_1^{2\log_{\beta_1}(\eta)}\right) = \mathcal{O}\left(\eta^2\right),
\end{aligned}$$

and that E_2 is close to $\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_k)$:

$$\begin{aligned}
\left\|E_2 - \mathbb{E}[\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_k)]\right\|_2 &= \left\|(1-\beta_1)\sum_{i=\bar{k}}^{k}\beta_1^{k-i}\mathbb{E}[\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_i)] - \mathbb{E}[\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_k)]\right\|_2 \\
&= \left\|(1-\beta_1)\sum_{i=\bar{k}}^{k}\beta_1^{k-i}\mathbb{E}\left[\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_i) - \nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_k)\right]\right\|_2 + \mathcal{O}(\eta^2) \\
&\leq (1-\beta_1)\cdot\left(k-\bar{k}\right)\cdot C_{10}\eta^{1.5} + \mathcal{O}(\eta^2). \quad \text{(by Lemma G.6)}
\end{aligned}$$

Combining the results of E_1 and E_2 gives

$$\begin{aligned}
\|\mathbb{E}_{\bar{k}}[\hat{\boldsymbol{m}}_k - \hat{\boldsymbol{g}}_k]\|_2 &\leq \|E_1\|_2 + \left\|E_2 - \mathbb{E}[\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_k)]\right\|_2 \\
&\leq (1-\beta_1)\cdot 2\log_{\beta_1}(\eta)\cdot C_{10}\eta^{1.5} + \mathcal{O}(\eta^2) \\
&\leq C_{11}\eta^{1.5}\log\frac{1}{\eta}
\end{aligned}$$

for some constant C_{11}, which completes the proof. □

G.3.1 Moment Calculation Within a Giant Step

In this part, we aim to give the change of first and second moments of ϕ and $\hat{\boldsymbol{v}}$, which is the basis of deriving the SDE for AGMs.

Now there are only a few preparations left before we get into the direct part of the moment calculation. For all $0 \le s < R_{\mathrm{grp}}$, $0 \le t \le H$. Note that $\|\hat{\boldsymbol{v}}_{k+1} - \hat{\boldsymbol{v}}_k\|_2 = (1-\beta_2)\left\|V\left(\hat{\boldsymbol{g}}_k\hat{\boldsymbol{g}}_k^\top\right) - \hat{\boldsymbol{v}}_k\right\|_2 = \mathcal{O}(1-\beta_2) = \mathcal{O}(\eta^2)$, so combining with the Lipschitzness of S gives

$$\left\|\hat{\boldsymbol{S}}_{k_2} - \hat{\boldsymbol{S}}_{k_1}\right\|_2 = \mathcal{O}\left((k_2-k_1)\eta^2\right)$$

for any $k_2 > k_1$ and $k_2 - k_1 = o\left(\eta^{-2}\right)$. Next, we begin our moment calculation analysis, starting from the update in a single step.

Lemma G.8. *For all $0 \le k \le R_{\mathrm{grp}}H$, it holds that*

$$\mathbb{E}\left[\hat{\boldsymbol{\theta}}_{k+1}\right] = \mathbb{E}\left[\hat{\boldsymbol{\theta}}_k - \eta\hat{\boldsymbol{S}}_0\hat{\boldsymbol{g}}_k\right] + \mathcal{O}\left(\eta^{2.5-\beta}\right).$$

Proof. We write the update rule of AGM under a single step as

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{k+1} &= \hat{\boldsymbol{\theta}}_k - \eta\hat{\boldsymbol{S}}_k\hat{\boldsymbol{m}}_{k+1} - \boldsymbol{e}_k \\
&= \hat{\boldsymbol{\theta}}_k - \eta\left[\hat{\boldsymbol{S}}_k\hat{\boldsymbol{g}}_k + \hat{\boldsymbol{S}}_k\left(\hat{\boldsymbol{m}}_{k+1} - \hat{\boldsymbol{g}}_k\right)\right] - \boldsymbol{e}_k \\
&= \hat{\boldsymbol{\theta}}_k - \eta\left[\hat{\boldsymbol{S}}_0\hat{\boldsymbol{g}}_k + \underbrace{\left(\hat{\boldsymbol{S}}_k - \hat{\boldsymbol{S}}_0\right)\hat{\boldsymbol{g}}_k}_{\Delta\hat{\boldsymbol{\theta}}_1} + \underbrace{\hat{\boldsymbol{S}}_k\left(\hat{\boldsymbol{m}}_{k+1} - \hat{\boldsymbol{g}}_k\right)}_{\Delta\hat{\boldsymbol{\theta}}_2}\right] - \boldsymbol{e}_k,
\end{aligned}$$

where we recall that $\boldsymbol{e}_k := -\mathbf{1}_{\bar{\mathcal{E}}_k}(\hat{\boldsymbol{\theta}}_k - \eta S(\hat{\boldsymbol{v}}_{k+1})\hat{\boldsymbol{m}}_{k+1}) + \mathbf{1}_{\bar{\mathcal{E}}_k}\boldsymbol{\phi}_{\mathrm{null}}$. We can prove that $\Delta\hat{\boldsymbol{\theta}}_1$ and $\Delta\hat{\boldsymbol{\theta}}_2$ are small enough to be negligible in expectation for our following calculation.

Specifically, if $k=0$ then $\Delta\hat{\boldsymbol{\theta}}_1 = \mathbf{0}$; and if $k > 0$, we can decompose $\mathbb{E}\left[\Delta\hat{\boldsymbol{\theta}}_1\right]$ as:

$$\begin{aligned}
\mathbb{E}\left[\Delta\hat{\boldsymbol{\theta}}_1\right] &= \mathbb{E}\left[\left(\hat{\boldsymbol{S}}_{k-1} - \hat{\boldsymbol{S}}_0\right)\hat{\boldsymbol{g}}_k + \left(\hat{\boldsymbol{S}}_k - \hat{\boldsymbol{S}}_{k-1}\right)\hat{\boldsymbol{g}}_k\right] \\
&= \mathbb{E}\left[\left(\hat{\boldsymbol{S}}_{k-1} - \hat{\boldsymbol{S}}_0\right)\nabla\mathcal{L}\left(\hat{\boldsymbol{\theta}}_k\right)\right] + \mathbb{E}\left[\left(\hat{\boldsymbol{S}}_k - \hat{\boldsymbol{S}}_{k-1}\right)\hat{\boldsymbol{g}}_k\right] \\
&= \mathcal{O}((k-1)\eta^2\cdot\eta^{0.5}) + \mathcal{O}(\eta^2) \\
&= \mathcal{O}(H\cdot R_{\mathrm{grp}}\cdot\eta^{2.5} + \eta^2) \\
&= \mathcal{O}(\eta^{1.5-\beta}).
\end{aligned}$$

Here, the second equality holds since the gradient noise term as step k $\boldsymbol{z}_k$ is conditioned on time k, when $\hat{\boldsymbol{S}}_{k-1}$ has already been determined, thus we can take the conditional expectation.

For $\Delta\hat{\boldsymbol{\theta}}_2$, let $\bar{k} = k - 2\log_{\beta_1}(\eta)$, we have

$$\begin{aligned}
\mathbb{E}\left[\Delta\hat{\boldsymbol{\theta}}_2\right] &= \mathbb{E}\left[\hat{\boldsymbol{S}}_{\bar{k}-1}\left(\hat{\boldsymbol{m}}_{k+1} - \hat{\boldsymbol{g}}_k\right) + \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right)\right] \\
&= \mathbb{E}\left[\hat{\boldsymbol{S}}_{\bar{k}-1}\mathbb{E}_{\bar{k}}\left[\left(\hat{\boldsymbol{m}}_{k+1} - \hat{\boldsymbol{g}}_k\right)\right]\right] + \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right) \\
&= \mathcal{O}\left(\eta^{1.5}\log\frac{1}{\eta}\right) + \mathcal{O}\left(\eta^2\log\frac{1}{\eta}\right) \\
&= \mathcal{O}\left(\eta^{1.5}\log\frac{1}{\eta}\right),
\end{aligned}$$

where the second-to-last equality follows from Lemma G.7. Finally, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{\theta}}_{k+1}\right] &= \mathbb{E}\left[\hat{\boldsymbol{\theta}}_k - \eta\hat{\boldsymbol{S}}_0\hat{\boldsymbol{g}}_k\right] + \mathcal{O}\left(\eta^{2.5-\beta}\right) + \mathcal{O}\left(\eta^{2.5}\log\frac{1}{\eta}\right) + \mathcal{O}(\eta^{100}) \\
&= \mathbb{E}\left[\hat{\boldsymbol{\theta}}_k - \eta\hat{\boldsymbol{S}}_0\hat{\boldsymbol{g}}_k\right] + \mathcal{O}\left(\eta^{2.5-\beta}\right),
\end{aligned}$$

which concludes the proof. $\square$

After getting the update rule of $\hat{\boldsymbol{\theta}}_k$, we then derive the moment change during the single round with $\boldsymbol{H}$ steps. To this end, we recall our modification of manifold projection from a "Gradient Flow" manner to a "Preconditioned Flow" manner in Definition 4.1.

Definition G.3 (Preconditioned Flow Projection)**.** *Fix a point $\theta_{null} \notin \Gamma$. Given a Positive Semi-Definite matrix $\boldsymbol{M}$. For $x \in \mathbb{R}^d$, consider the preconditioned flow $\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\boldsymbol{M}\nabla\mathcal{L}(x(t))$ with $x(0) = x$. We denote the preconditioned flow projection of x as $\Phi_{\boldsymbol{M}}(x)$, i.e. $\Phi_{\boldsymbol{M}}(x) := \lim_{t\to+\infty} x(t)$ if the limit exists and belongs to Γ, and $\Phi_{\boldsymbol{M}}(x) = \theta_{null}$ otherwise.*

We decompose the preconditioner matrix in the very begining of the giant step as $\hat{\boldsymbol{S}}_0 = \hat{\boldsymbol{S}}(\hat{\boldsymbol{v}}_0) = \boldsymbol{P}\boldsymbol{P}$, where $\boldsymbol{P} = \hat{\boldsymbol{S}}_0^{1/2}$. We then provide the first moment calculation of $\hat{\boldsymbol{\phi}}$ in the following lemma. Before that, we first introduce the operator $\mathcal{V}_{\boldsymbol{H}}$.

Definition G.4. *Given a Positive Semi-Definite matrix $\boldsymbol{H} \in \mathbb{R}^{d\times d}$, whose j-th eigenvalue and the corresponding orthonormal eigenvector are denoted by λ_j and $\boldsymbol{v}_j$. We then define the operator $\mathcal{V}_{\boldsymbol{H}}(\cdot) : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ as*

$$\mathcal{V}_{\boldsymbol{H}}(\cdot) = \sum_{i,j:\lambda_i\neq 0\vee\lambda_j\neq 0} \frac{1}{\lambda_i+\lambda_j} \left\langle \cdot, \boldsymbol{v}_i\boldsymbol{v}_j^\top \right\rangle \boldsymbol{v}_i\boldsymbol{v}_j^\top.$$

Intuitively, the above operator projects the one matrix into the basis of $\boldsymbol{H}$ and sums up the corresponding components with weights $\frac{1}{\lambda_i+\lambda_j}$. Then we present our moment calculation lemma.

Lemma G.9. *The expectation of the change of the manifold projection every round is*

$$\mathbb{E}\left[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}\right] = -\frac{H\eta^2}{2}\hat{\boldsymbol{S}}_0\partial\Phi_{\hat{\boldsymbol{S}}_0}(\hat{\boldsymbol{\phi}}^{(0)})\hat{\boldsymbol{S}}_0\partial^2\nabla\mathcal{L}(\hat{\boldsymbol{\phi}}_{(0)})\left[\boldsymbol{P}\mathcal{V}_{\nabla^2\mathcal{L}'(\hat{\boldsymbol{\phi}}'_{(0)})}(\boldsymbol{\Sigma}_{0,\boldsymbol{P}})\boldsymbol{P}\right] + \tilde{\mathcal{O}}(\eta^{1.5-\beta})$$

for $R_0 < s < R_{\mathrm{grp}}$, and

$$\mathbb{E}\left[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}\right] = \tilde{\mathcal{O}}(\eta)$$

for $s \leq R_0$, where $R_0 := \max\left\{\left\lceil \frac{10}{\lambda_{\max}\alpha}\log\frac{1}{\eta}\right\rceil, \left\lceil 2\log_{1/\beta}\frac{1}{\eta}\right\rceil\right\}$ and $\boldsymbol{\Sigma}_{0,\boldsymbol{P}} := \boldsymbol{P}\boldsymbol{\Sigma}_0\boldsymbol{P}$.

Proof. First, we consider the scenario when $R_0 < s < R_{\mathrm{grp}}$. Let $L'(\boldsymbol{x}) := L(\boldsymbol{P}\boldsymbol{x})$, then

$$\begin{aligned}
\nabla L'(\boldsymbol{x}) &= \boldsymbol{P}\nabla L(\boldsymbol{P}\boldsymbol{x})\\
\nabla^2 L'(\boldsymbol{x}) &= \boldsymbol{P}\nabla^2 L(\boldsymbol{P}\boldsymbol{x})\boldsymbol{P}\\
\boldsymbol{\Sigma}'(\boldsymbol{x}) &= \boldsymbol{P}\boldsymbol{\Sigma}(\boldsymbol{P}\boldsymbol{x})\boldsymbol{P}\\
\partial^2(\nabla L')(\boldsymbol{x})[\boldsymbol{M}] &= \boldsymbol{P}\partial^2(\nabla L)(\boldsymbol{P}\boldsymbol{x})[\boldsymbol{P}\boldsymbol{M}\boldsymbol{P}].
\end{aligned}$$

For a one-step GD update, we consider an auxiliary process $\{\hat{\boldsymbol{\theta}}'_t\}$

$$\begin{aligned}
\hat{\boldsymbol{\theta}}'_{t+1} &= \hat{\boldsymbol{\theta}}'_t - \eta\nabla\mathcal{L}'(\hat{\boldsymbol{\theta}}'_t) + \mathcal{O}\left(\eta^{2.5-\beta}\right)\\
&= \hat{\boldsymbol{\theta}}'_t - \eta\boldsymbol{P}\nabla\mathcal{L}(\boldsymbol{P}\hat{\boldsymbol{\theta}}'_t) + \mathcal{O}\left(\eta^{2.5-\beta}\right).
\end{aligned}$$

Similarly, we define $\hat{\boldsymbol{A}}_t^{'(s)} := \mathbb{E}[\hat{\boldsymbol{x}}_t^{'(s)}\hat{\boldsymbol{x}}_t^{'(s)\top}]$, $\hat{\boldsymbol{q}}_t^{'(s)} := \mathbb{E}[\hat{\boldsymbol{x}}_t^{'(s)}]$, and $\hat{\boldsymbol{B}}_t^{'(s)} := \mathbb{E}[\hat{\boldsymbol{x}}_t^{'(s)}\Delta\hat{\phi}'^{(s)\top}]$, and $\Phi(\boldsymbol{x})$ is the gradient flow projection of point $\boldsymbol{x}$. We further define $\hat{\boldsymbol{\phi}}'(s) := \Phi(\hat{\boldsymbol{\theta}}^{'(s)})$.

Now we are interested in the update of $\boldsymbol{P}\hat{\boldsymbol{\theta}}'$, which is

$$\boldsymbol{P}\hat{\boldsymbol{\theta}}'_{t+1} = \boldsymbol{P}\hat{\boldsymbol{\theta}}'_t - \eta\hat{\boldsymbol{S}}_0\nabla\mathcal{L}(\boldsymbol{P}\hat{\boldsymbol{\theta}}'_t) + \mathcal{O}\left(\eta^{2.5-\beta}\right). \tag{7}$$

One can obviously see the update rule of $\boldsymbol{P}\hat{\boldsymbol{\theta}}'$ resembles the update rule of $\hat{\boldsymbol{\theta}}$ in Lemma G.8. Now we set $\hat{\boldsymbol{\theta}}' = \boldsymbol{P}^{-1}\hat{\boldsymbol{\theta}}$, then Equation (7) is satisfied, and combining Equation (7) and Lemma G.8 gives

$$\boldsymbol{q}_{t+1}^{'(s)} = \boldsymbol{q}_{t+1}^{'(s)} - \eta\nabla\mathcal{L}'(\hat{\boldsymbol{\theta}}_t^{'(s)}) + \mathcal{O}\left(\eta^{2.5-\beta}\right).$$

Notice that the above equation resembles the single update for SGD, which allows us to apply Lemma I.36 from Gu et al. (2023b) for the update of $\hat{\boldsymbol{\theta}}'$, with loss function $\mathcal{L}'(\hat{\boldsymbol{\theta}})$, number of workers $k = 1$ and manifold projection $\Phi'(\hat{\boldsymbol{\theta}})$, which gives

$$\begin{aligned}\boldsymbol{P}\mathbb{E}\left[\hat{\boldsymbol{\phi}}'^{(s+1)} - \hat{\boldsymbol{\phi}}'^{(s)}\right] &= \mathbb{E}\left[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}\right] \\ &= -\frac{H\eta^2}{2}\boldsymbol{P}\boldsymbol{P}\partial\Phi_{\hat{\boldsymbol{S}}_0}(\hat{\boldsymbol{\phi}}^{(0)})\boldsymbol{P}\boldsymbol{P}\partial^2\nabla\mathcal{L}(\hat{\boldsymbol{\phi}}_{(0)})[\boldsymbol{P}\mathcal{V}_{\nabla^2\mathcal{L}'(\hat{\boldsymbol{\phi}}'_{(0)})}(\boldsymbol{P}\boldsymbol{\Sigma}_0\boldsymbol{P})\boldsymbol{P}] \\ &+ \tilde{O}(\eta^{1.5-\beta}),\end{aligned}$$

where the first equation uses the fact that $\boldsymbol{P}\hat{\boldsymbol{\phi}}(\hat{\boldsymbol{\theta}}') = \Phi_{\boldsymbol{S}}(\hat{\boldsymbol{\theta}})$, and it can be verified with the definitions of $\hat{\boldsymbol{\phi}}'$, $\Phi_{\boldsymbol{S}}$, and $\hat{\boldsymbol{\theta}}'$.

The proof when $s \le R_0$ is a direct conclusion of Lemma I.36 in Gu et al. (2023b) since the $R_0 \propto \log\frac{1}{\eta}$ in our case. $\square$

Notice the above equation for the moment of $\hat{\phi}$ contains ϕ'. The next corollary eliminates ϕ' from the formula.

Corollary G.2. *The expectation of the change of manifold projection every round is:*

$$\mathbb{E}\left[\phi^{(s+1)} - \phi^{(s)}\right] = \begin{cases} \frac{H\eta^2}{2}\hat{\boldsymbol{S}}_0\partial^2\Phi_{\hat{\boldsymbol{S}}_0}(\phi^{(0)})[\hat{\boldsymbol{S}}_0\boldsymbol{\Sigma}_0\hat{\boldsymbol{S}}_0] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), & R_0 < s < R_{\mathrm{grp}} \\ \tilde{\mathcal{O}}(\eta), & s \le R_0 \end{cases}$$

Proof. Notice that for the preconditioned projection, we also have the corresponding transformation

$$\begin{aligned}\partial\Phi'(\boldsymbol{x}') &= \boldsymbol{P}\partial\Phi_{\hat{\boldsymbol{S}}}(\boldsymbol{P}\boldsymbol{x}')\boldsymbol{P} \\ \partial^2\Phi'(\boldsymbol{x}')[\boldsymbol{M}] &= \boldsymbol{P}\partial^2\Phi(\boldsymbol{x}')[\boldsymbol{P}\boldsymbol{M}\boldsymbol{P}].\end{aligned}$$

The above two equations and Lemma I.36 in Gu et al. (2023b) complete the proof. $\square$

Lemma G.10. *The second moment of the change of manifold projection every round is*

$$\mathbb{E}\left[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})^\top\right] = \begin{cases} H\eta^2\hat{\boldsymbol{S}}_0\boldsymbol{P}_{\|,\hat{\boldsymbol{S}}}\hat{\boldsymbol{S}}_0\boldsymbol{\Sigma}_0\hat{\boldsymbol{S}}_0\boldsymbol{P}_{\|,\hat{\boldsymbol{S}}}\hat{\boldsymbol{S}}_0 + \tilde{O}(\eta^{1.5-\beta}), & R_0 < s < R_{\mathrm{grp}} \\ \tilde{O}(\eta), & s \le R_0 \end{cases}$$

where $R_0 := \max\left\{\left\lceil\frac{10}{\lambda_{\max}\alpha}\log\frac{1}{\eta}\right\rceil, \left\lceil 2\log_{1/\beta}\frac{1}{\eta}\right\rceil\right\}$ *and* $\boldsymbol{P}_{\|,\hat{\boldsymbol{S}}} := \partial\Phi_{\hat{\boldsymbol{S}}}(\hat{\boldsymbol{\phi}}^{(0)})$.

Proof. According to Lemma I.37 in Gu et al. (2023b), we could write the second moment for $\hat{\boldsymbol{\theta}}'$ as

$$\mathbb{E}\left[(\hat{\boldsymbol{\phi}}'^{(s+1)} - \hat{\boldsymbol{\phi}}'^{(s)})(\hat{\boldsymbol{\phi}}'^{(s+1)} - \hat{\boldsymbol{\phi}}'^{(s)})^\top\right] = \begin{cases} H\eta^2\boldsymbol{\Sigma}'_{0,\|} + \tilde{O}(\eta^{1.5-\beta}), & R_0 < s < R_{\mathrm{grp}} \\ \tilde{\mathcal{O}}(\eta), & s \le R_0. \end{cases}$$

Notice that

$$\begin{aligned}\Sigma'_{0,\|} &:= \partial\Phi(\hat{\boldsymbol{\phi}}'^{(0)})\Sigma'_0\partial\Phi(\hat{\boldsymbol{\phi}}'^{(0)}) \\ &= \boldsymbol{P}\partial\Phi_{\hat{\boldsymbol{S}}}(\hat{\boldsymbol{\phi}}^{(0)})\boldsymbol{P}\boldsymbol{P}\Sigma_0\boldsymbol{P}\boldsymbol{P}\partial\Phi_{\hat{\boldsymbol{S}}}(\hat{\boldsymbol{\phi}}^{(0)})\boldsymbol{P}.\end{aligned}$$

When $R_0 \le s < R_{\mathrm{grp}}$,

$$\begin{aligned}\mathbb{E}\left[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})^\top\right] &= \mathbb{E}\left[\boldsymbol{P}(\hat{\boldsymbol{\phi}}'^{(s+1)} - \hat{\boldsymbol{\phi}}'^{(s)})(\hat{\boldsymbol{\phi}}'^{(s+1)} - \hat{\boldsymbol{\phi}}'^{(s)})^\top\boldsymbol{P}\right] \\ &= \hat{\boldsymbol{S}}_0\boldsymbol{P}_{\|,\hat{\boldsymbol{S}}}\hat{\boldsymbol{S}}_0\boldsymbol{\Sigma}_0\hat{\boldsymbol{S}}_0\boldsymbol{P}_{\|,\hat{\boldsymbol{S}}}\hat{\boldsymbol{S}}_0.\end{aligned}$$

The proof when $s \le R_0$ is a direct conclusion of Lemma I.37 in Gu et al. (2023b) since the $R_0 \propto \log\frac{1}{\eta}$ in our case. $\square$

Then we give the moment change of $\hat{\boldsymbol{\phi}}$ within a single giant step.

Theorem G.1. *Given $\|\hat{\boldsymbol{\theta}}^{(0)} - \hat{\boldsymbol{\phi}}^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$, for $0 < \beta < 0.5$, the first and second moments of $\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})} := \hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})} - \hat{\boldsymbol{\phi}}^{(0)}$ are as follows:*

$$\mathbb{E}[\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})}] = \frac{\eta^{1-\beta}}{2}\hat{\boldsymbol{S}}_0\partial^2\Phi_{\hat{\boldsymbol{S}}_0}(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{S}}_0\boldsymbol{\Sigma}_0\hat{\boldsymbol{S}}_0] + \tilde{\mathcal{O}}(\eta^{1.5-2\beta}) + \tilde{\mathcal{O}}(\eta),$$

$$\mathbb{E}[\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})\top}] = \eta^{1-\beta}\hat{\boldsymbol{S}}_0\boldsymbol{\Sigma}_{\|}(\hat{\boldsymbol{\phi}}^{(0)}, \hat{\boldsymbol{S}}^{(0)})\hat{\boldsymbol{S}}_0 + \tilde{\mathcal{O}}(\eta^{1.5-1.5\beta}) + \tilde{\mathcal{O}}(\eta),$$

where $\boldsymbol{\Sigma}_{\|}(\phi^{(0)}, \hat{\boldsymbol{S}}^{(0)}) := \boldsymbol{P}_{\|,\hat{\boldsymbol{S}}}\hat{\boldsymbol{S}}_0\boldsymbol{\Sigma}_0\hat{\boldsymbol{S}}_0\boldsymbol{P}_{\|,\hat{\boldsymbol{S}}}$.

Proof. First we prove the first moment change as

$$\begin{aligned}
\mathbb{E}[\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})}] &= \mathbb{E}[\sum_{s=0}^{R_{\mathrm{grp}}-1} \hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}] \\
&= \sum_{s=0}^{R_0}\mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}] + \sum_{s=R_0+1}^{R_{\mathrm{grp}}-1}\mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}] \\
&= \frac{\eta^{1-\beta}}{2}\hat{\boldsymbol{S}}_0\partial^2\Phi_{\hat{\boldsymbol{S}}_0}(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{S}}_0\boldsymbol{\Sigma}_0\hat{\boldsymbol{S}}_0] + \tilde{\mathcal{O}}(\eta^{1.5-2\beta}) + \tilde{\mathcal{O}}(\eta).
\end{aligned}$$

The last equation is a direct conclusion of Corollary G.2.

And for the second moment, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{s=0}^{R_{\mathrm{grp}}-1} \hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}\right)\left(\sum_{s=0}^{R_{\mathrm{grp}}-1} \hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}\right)^\top\right] \\
&= \sum_{s=0}^{R_{\mathrm{grp}}-1} \mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})^\top] \\
&\quad + \sum_{s\neq s'}\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})]\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s'+1)} - \hat{\boldsymbol{\phi}}^{(s')})^\top] \\
&= \eta^{1-\beta}\hat{\boldsymbol{S}}_0\boldsymbol{\Sigma}_{\|}(\hat{\boldsymbol{\phi}}^{(0)}, \hat{\boldsymbol{S}}^{(0)})\hat{\boldsymbol{S}}_0 + \tilde{\mathcal{O}}(\eta^{1.5-1.5\beta}) + \tilde{\mathcal{O}}(\eta),
\end{aligned}$$

where the last equation uses $\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})]\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s'+1)} - \hat{\boldsymbol{\phi}}^{(s')})^\top] = \tilde{\mathcal{O}}(\eta^2)$. ☐

Next, we proceed with the updates of $\boldsymbol{v}$.

Lemma G.11. *Given $c := \frac{1-\beta_2}{\eta^2}$, and we have*

$$\mathbb{E}\left[\hat{\boldsymbol{v}}_0^{(R_{\mathrm{grp}})} - \hat{\boldsymbol{v}}_0^{(0)}\right] = c\eta^{1-\beta}\left(V\left(\boldsymbol{\Sigma}_0^{(0)}\right) - \hat{\boldsymbol{v}}_0^{(0)}\right) + \mathcal{O}\left(\eta^{1.5-1.5\beta}\right).$$

Proof. By the update rule of $\boldsymbol{v}$, we have

$$\begin{aligned}
\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)} &= \hat{\boldsymbol{v}}_H^{(s)} - \hat{\boldsymbol{v}}_0^{(s)} \\
&= \beta_2^H\hat{\boldsymbol{v}}_0^{(s)} + (1-\beta_2)\sum_{i=1}^{H}\beta_2^{H-i}\boldsymbol{V}\left(\hat{\boldsymbol{g}}_i^{(s)}\hat{\boldsymbol{g}}_i^{(s)^\top}\right) - \hat{\boldsymbol{v}}_0^{(s)} \\
&= \left(\beta_2^H - 1\right)\hat{\boldsymbol{v}}_0^{(0)} + (1-\beta_2)\sum_{i=1}^{H}\beta_2^{H-i}V\left(\hat{\boldsymbol{g}}_i^{(s)}\hat{\boldsymbol{g}}_i^{(s)\top}\right).
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{g}}_i^{(s)}\hat{\boldsymbol{g}}_i^{(s)^\top}\right] &= \mathbb{E}\left[\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_i^{(s)})\right] \\
&= \mathbb{E}\left[\boldsymbol{\Sigma}(\hat{\boldsymbol{\phi}}_0^{(0)} + \boldsymbol{x}_i^{(s)})\right] \\
&= \mathbb{E}\left[\boldsymbol{\Sigma}(\boldsymbol{\phi}_0^{(0)}) + \mathcal{O}\left(\eta^{0.5-0.5\beta}\right)\right] \\
&= \boldsymbol{\Sigma}_0^{(0)} + \mathcal{O}\left(\eta^{0.5-0.5\beta}\right).
\end{aligned}$$

Combining with the linearity of $\boldsymbol{V}$, we conclude that

$$\mathbb{E}\left[\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right] = \left(\beta_2^H - 1\right)\hat{\boldsymbol{v}}_0^{(0)} + \left(1-\beta_2^H\right)\boldsymbol{V}\left(\boldsymbol{\Sigma}_0^{(0)}\right) + \mathcal{O}\left(\eta^{1.5-0.5\beta}\right)$$

$$\mathbb{E}\left[\hat{\boldsymbol{v}}_0^{(s+1)}\right] = \beta_2^H\hat{\boldsymbol{v}}_0^{(s)} + \left(1-\beta_2^H\right)\boldsymbol{V}\left(\boldsymbol{\Sigma}_0^{(0)}\right) + \mathcal{O}\left(\eta^{1.5-0.5\beta}\right).$$

To transfer from $\hat{\boldsymbol{v}}_0^{(0)}$ to arbitrary $\hat{\boldsymbol{v}}_0^{(s)}$, we simply expand to get the result:

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{v}}_0^{(s)}\right] &= \beta_2^{sH}\hat{\boldsymbol{v}}_0^{(0)} + \left[\left(1-\beta_2^H\right)V\left(\boldsymbol{\Sigma}_0^{(0)}\right) + \mathcal{O}\left(\eta^{1.5-0.5\beta}\right)\right]\left(1+\beta_2^H+\beta_2^{2H}+\cdots+\beta_2^{(s-1)H}\right)\\
&= \beta_2^{sH}\hat{\boldsymbol{v}}_0^{(0)} + \left[\left(1-\beta_2^H\right)V\left(\boldsymbol{\Sigma}_0^{(0)}\right)\right]\left(\frac{1-\beta_2^{sH}}{1-\beta_2^H}\right) + \mathcal{O}\left(\eta^{1.5-0.5\beta}\right)\cdot\mathcal{O}\left(\eta^{-\beta}\right)\\
&= \beta_2^{sH}\hat{\boldsymbol{v}}_0^{(0)} + \left(1-\beta_2^{sH}\right)V\left(\boldsymbol{\Sigma}_0^{(0)}\right) + \mathcal{O}\left(\eta^{1.5-1.5\beta}\right).
\end{aligned}$$

Thus we have

$$\mathbb{E}\left[\hat{\boldsymbol{v}}_0^{(R_{\mathrm{grp}})} - \hat{\boldsymbol{v}}_0^{(0)}\right] = c\eta^{1-\beta}\left(V\left(\boldsymbol{\Sigma}_0^{(0)}\right) - \hat{\boldsymbol{v}}_0^{(0)}\right) + \mathcal{O}\left(\eta^{1.5-1.5\beta}\right).$$

where the last equation uses the fact that $1-\beta_2^{R_{\mathrm{grp}}H} = 1-(1-c\eta^{1-\beta})+O(\eta^{2-2\beta}) = c\eta+O(\eta^2)$. $\square$

Also, for the second moment change of $\hat{\boldsymbol{v}}$, we get the following lemma

Lemma G.12. *The second moment change of $\hat{\mathbf{v}}$ over a giant step is*

$$\mathbb{E}\left[\left(\hat{\boldsymbol{v}}_0^{(R_{\mathrm{grp}})} - \hat{\boldsymbol{v}}_0^{(0)}\right)\left(\hat{\boldsymbol{v}}_0^{(R_{\mathrm{grp}})} - \hat{\boldsymbol{v}}_0^{(0)}\right)^\top\right] = \mathcal{O}(\eta^{2-\beta}).$$

Proof.

$$\begin{aligned}
\mathbb{E}\Big[\left(\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right)\left(\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right)^\top\Big] &= \mathbb{E}\Big[\left((\beta_2^H-1)+(1-\beta_2)\sum_{i=1}^{H}\beta_2^{H-i}V\left(\hat{\boldsymbol{g}}_i^{(s)}\hat{\boldsymbol{g}}_i^{(s)\top}\right)\right)\\
&\qquad\left((\beta_2^H-1)+(1-\beta_2)\sum_{i=1}^{H}\beta_2^{H-i}V\left(\hat{\boldsymbol{g}}_i^{(s)}\hat{\boldsymbol{g}}_i^{(s)\top}\right)\right)^\top\Big]\\
&= \mathcal{O}\left((1-\beta_2^H)^2\right) = \mathcal{O}\left(\eta^2\right).
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\boldsymbol{v}}_0^{(R_{\mathrm{grp}})} - \hat{\boldsymbol{v}}_0^{(0)}\right)\left(\hat{\boldsymbol{v}}_0^{(R_{\mathrm{grp}})} - \hat{\boldsymbol{v}}_0^{(0)}\right)^\top\right] &= \mathbb{E}\left[\left(\sum_{s=0}^{R_{\mathrm{grp}}-1}\left(\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right)\right)\left(\sum_{s=0}^{R_{\mathrm{grp}}-1}\left(\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right)^\top\right)\right]\\
&= \sum_{s=0}^{R_{\mathrm{grp}}-1}\mathbb{E}\left[\left(\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right)\left(\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right)^\top\right]\\
&\quad + \sum_{s\neq s'}\mathbb{E}\left[\left(\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right)\right]\mathbb{E}\left[\left(\hat{\boldsymbol{v}}_0^{(s'+1)} - \hat{\boldsymbol{v}}_0^{(s')}\right)^\top\right]\\
&= \mathcal{O}(\eta^{2-\beta}).
\end{aligned}$$

The last equation uses

$$\mathbb{E}\Big[\left(\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right)\left(\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right)^\top\Big] = \mathcal{O}(\eta^2),$$

and

$$\mathbb{E}\left[\left(\hat{\boldsymbol{v}}_0^{(s+1)} - \hat{\boldsymbol{v}}_0^{(s)}\right)\right]\mathbb{E}\left[\left(\hat{\boldsymbol{v}}_0^{(s'+1)} - \hat{\boldsymbol{v}}_0^{(s')}\right)^\top\right] = \mathcal{O}(3-3\beta).$$

The above equation completes the proof. $\square$

G.4 Weak Approximation

After we get the first and second moment changes within a giant step, we now utilize the moment calculation to prove the SDE approximation part of Theorem D.1. First, we recall our slow SDE for AGMs

$$\begin{cases} \mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta},\boldsymbol{S}(t)}\left(\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta}(t);\boldsymbol{S}(t))\mathrm{d}\boldsymbol{W}_t - \frac{1}{2}\boldsymbol{S}(t)\nabla^3\mathcal{L}(\boldsymbol{\zeta})\left[\boldsymbol{\Sigma}_{\diamond}(\boldsymbol{\zeta}(t);\boldsymbol{S}(t))\right]\mathrm{d}t\right), \\ \mathrm{d}\boldsymbol{v}(t) = c\left(V(\boldsymbol{\Sigma}(\boldsymbol{\zeta})) - \boldsymbol{v}\right)\mathrm{d}t. \end{cases}$$

We then open the projection mapping $P_{\boldsymbol{\zeta},\boldsymbol{S}(t)}$ as

$$\begin{cases} \mathrm{d}\boldsymbol{\zeta} = \boldsymbol{S}(\boldsymbol{v})\partial\Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\zeta})\boldsymbol{S}(\boldsymbol{v})\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}_t + \frac{1}{2}\boldsymbol{S}(\boldsymbol{v})\partial^2\Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\zeta})\left[\boldsymbol{S}(\boldsymbol{v})\boldsymbol{\Sigma}(\boldsymbol{\zeta})\boldsymbol{S}(\boldsymbol{v})\right]\mathrm{d}t, \\ \mathrm{d}\boldsymbol{v}(t) = c\left(V(\boldsymbol{\Sigma}(\boldsymbol{\zeta})) - \boldsymbol{v}\right)\mathrm{d}t. \end{cases} \tag{8}$$

Now it suffices to prove the SDE in Equation (8) tracks the trajectory in AGMs within $\mathcal{O}(\frac{1}{\eta^2})$ steps in a weak approximation sense.

First, we have to show that the solution of Equation (8) in close in the minimizer manifold

Lemma G.13. *Let $\boldsymbol{X}(t) := (\boldsymbol{\zeta}(t)^\top, \boldsymbol{v}(t)^\top)^\top$ be the solution of Equation* (8) *with $\boldsymbol{\zeta}(0) \in \Gamma$, and $\boldsymbol{v}(0) \in \mathbb{R}^d$, then we have that $\boldsymbol{\zeta}(t) \in \Gamma$ for all $t \geq 0$.*

Proof. According to Filipović (2000); Du and Duan (2006), for a closed manifold $\mathcal{M}$ to be viable for the SDE $\mathrm{d}\boldsymbol{X}(t) = \boldsymbol{A}(\boldsymbol{X}(t))\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}(\boldsymbol{X}(t))\mathrm{d}t$, where $\boldsymbol{A}(\cdot) : \mathbb{R}^{2d} \to \mathbb{R}^{2d\times 2d}$ and $\boldsymbol{b}(\cdot) : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ are locally Lipchitz, it suffices to show that the following Nagumo type consistency condition holds:

$$\mu(\boldsymbol{x}) := \boldsymbol{b}(\boldsymbol{x}) - \frac{1}{2}\sum_j D[A_j(\boldsymbol{x})]A_j(\boldsymbol{x}) \in T_{\boldsymbol{x}}(\mathcal{M}), \quad A_j(\boldsymbol{x}) \in T_{\boldsymbol{x}}(\mathcal{M}),$$

where $D[\cdot]$ is the Jacobian operator and $A_j(\boldsymbol{x})$ denotes the j-th column of $A(\boldsymbol{x})$.

Following the argument in Gu et al. (2023b), here we also only need to show that $\boldsymbol{P}_{\perp,\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{x})\mu(\boldsymbol{x}) = 0$, where $\boldsymbol{P}_{\perp,\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{x}) := \boldsymbol{I}_d - \partial\Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{x})$.

$$\begin{aligned} \boldsymbol{P}_{\perp,\boldsymbol{S}}(\boldsymbol{x})\sum_j D[A_j(\boldsymbol{x})]A_j(\boldsymbol{x}) &= \boldsymbol{P}_{\perp,\boldsymbol{S}}(\boldsymbol{x})\sum_j D\left[\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\boldsymbol{\Sigma}_j^{1/2}\right]\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\boldsymbol{\Sigma}_j^{1/2} \\ &= \boldsymbol{P}_{\perp,\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\sum_j \partial^2\Phi_{\boldsymbol{S}}(\boldsymbol{x})[\boldsymbol{S}\boldsymbol{\Sigma}_j^{1/2}, \boldsymbol{S}\partial\Phi_{\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\boldsymbol{\Sigma}_j^{1/2}] \\ &= -\boldsymbol{P}_{\perp,\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\boldsymbol{S}^{-1}\nabla^2\mathcal{L}(\boldsymbol{x})^\dagger\partial^2(\nabla\mathcal{L})(\boldsymbol{x})\left[\boldsymbol{S}\boldsymbol{\Sigma}_{\parallel}(\boldsymbol{x},\boldsymbol{S})\right]. \end{aligned}$$

Notice that, since it is clear from the context, here we write $\boldsymbol{S} = \boldsymbol{S}$ for short. The last equation uses Lemma G.3. Agian, applying Lemma G.3 gives

$$\boldsymbol{P}_{\perp,\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{b}(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{P}_{\perp,\boldsymbol{S}}(\boldsymbol{x})\boldsymbol{S}\boldsymbol{S}^{-1}\nabla^2\mathcal{L}(\boldsymbol{x})^\dagger\partial^2(\nabla\mathcal{L})(\boldsymbol{x})\left[\boldsymbol{S}\boldsymbol{\Sigma}_{\parallel}(\boldsymbol{x},\boldsymbol{S})\right].$$

The above equation completes the proof. □

To establish Theorem 4.1, we give an equivalent theorem, which captures the closeness of $\boldsymbol{X}(t)$ and $\bar{\boldsymbol{X}}_t$ in a long horizon. Also, for the proof of Theorem 4.1, it suffices to prove the following lemma, whose proof will be shown in Appendix G.5.

Theorem G.2. *If $\|\boldsymbol{\theta}^{(0)} - \phi^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta\log\frac{1}{\eta}})$ and $\boldsymbol{\zeta}(0) = \phi^{(0)}$, $\boldsymbol{v}(0) = \boldsymbol{v}^{(0)}$, then for a giant step $R_{\mathrm{grp}} = \lfloor\frac{1}{\eta^{0.25}}\rfloor$, for every test function $g \in \mathcal{C}^3$,*

$$\max_{0\leq n\leq\lfloor\frac{T}{\eta^{0.75}}\rfloor}\left|\mathbb{E}\big[g\left(\bar{\boldsymbol{X}}^{(nR_{\mathrm{grp}})}\right)\big] - \mathbb{E}\big[g\left(\boldsymbol{X}(n\eta^{0.75})\right)\big]\right| = C_g\eta^{0.25}(\log\frac{1}{\eta})^b,$$

where C_g is a constant independent of η but depends on $g(\cdot)$ and $b > 0$ is a universal constant independent of $g(\cdot)$ and η.

G.4.1 Preliminary and Additional Notations

We first introduce some notations and preliminary background. We consider the following stochastic gradient algorithms (SGAs)

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \eta_e \boldsymbol{h}(\boldsymbol{x}_n, \boldsymbol{\xi}_n),$$

where $\boldsymbol{x}_n \in \mathbb{R}^{2d}$ is the parameter vector, η_e is the effective learning rate, $\boldsymbol{h}(\cdot,\cdot) : \mathbb{R}^{2d} \times \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ depend on the current parameter vector $\boldsymbol{x}_n$ and the noise vector $\boldsymbol{\xi}_n$ sampled from some distribution $\Xi(\boldsymbol{x}_n)$.

We also consider the Stochastic Differential Equation (SDE) of the following form:

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t)\mathrm{d}t + \sigma(\boldsymbol{X}_t, t)\mathrm{d}\boldsymbol{W}_t,$$

where $\boldsymbol{b} : \mathbb{R}^{2d} \times \mathbb{R}^{+} \to \mathbb{R}^{2d}$ is the drift vector function and $\sigma : \mathbb{R}^{2d} \times \mathbb{R}^{+} \to \mathbb{R}^{2d\times 2d}$ is the diffusion matrix function.

According to the moment calculations in Corollary G.2,Lemma G.10, Lemma G.11, and Lemma G.12, we set $\eta_e = \eta^{1-\beta}$, and

$$\begin{aligned}
\boldsymbol{b}(\boldsymbol{X}_t, t) &= \left(\left(\frac{1}{2} \partial^2 \Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\zeta}) \left[\boldsymbol{\Sigma}(\boldsymbol{\zeta}, \boldsymbol{S}(\boldsymbol{v})) \right] \right)^{\top}, c\left(V(\boldsymbol{\Sigma}(\boldsymbol{\zeta})) - \boldsymbol{v} \right)^{\top} \right)^{\top}, \\
\sigma(\boldsymbol{X}_t, t) &= \begin{pmatrix} \partial \Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\zeta}) \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\zeta}, \boldsymbol{S}(\boldsymbol{v})), & \mathbf{0} \\ \mathbf{0}, & \mathbf{0} \end{pmatrix}.
\end{aligned}$$

Next, we are going to define the one giant step change of the parameter, both for SGAs and SDE.

$$\hat{\hat{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})} := \left(\Phi_{\hat{\boldsymbol{S}}^{(lR_{\mathrm{grp}})}} \left(\hat{\boldsymbol{\theta}} \right)^{\top}, {\boldsymbol{v}^{l\hat{R}_{\mathrm{grp}}}}^{\top} \right)^{\top} \in \mathbb{R}^{2d}, \quad \Delta^{(n)} := \hat{\hat{\boldsymbol{X}}}^{((n+1)R_{\mathrm{grp}})} - \hat{\hat{\boldsymbol{X}}}^{(nR_{\mathrm{grp}})},$$
$$\tilde{\Delta}^{(n)} := \boldsymbol{X}_{(n+1)\eta_e} - \hat{\hat{\boldsymbol{X}}}^{(nR_{\mathrm{grp}})}, \quad \boldsymbol{b}^{(n)} := \boldsymbol{b}(\hat{\hat{\boldsymbol{X}}}^{(nR_{\mathrm{grp}})}), \quad \sigma^{(n)} := \sigma(\hat{\hat{\boldsymbol{X}}}^{(nR_{\mathrm{grp}})}).$$

We now give a lemma to give the approximation of the first, second, and higher-order moment change of the SDE.

Lemma G.14. *There exists a positive constant c_0 independent of η_e and g such that for all $\boldsymbol{\zeta} \in \Gamma$, it holds for all $1 \le i \le d$ that*

$$\begin{aligned}
\left| \mathbb{E}[\tilde{\Delta}_i(\boldsymbol{\zeta}, n)] - \eta_e b_i(\boldsymbol{\zeta}) \right| &\le c_0 \eta_e^2, \\
\left| \mathbb{E}[\tilde{\Delta}_i(\boldsymbol{\zeta}, n)\tilde{\Delta}_j(\boldsymbol{\zeta}, n)] - \eta_e \sum_{l=1}^{d} \sigma_{i,l}(\boldsymbol{\zeta})\sigma_{l,j}(\boldsymbol{\zeta}) \right| &\le c_0 \eta_e^2, \\
\mathbb{E}\left[\left| \prod_{s=1}^{6} \tilde{\Delta}_{i_s}(\boldsymbol{\zeta}, n) \right| \right] &\le c_0 \eta_e^3.
\end{aligned}$$

Proof. (i) By Lemma G.13, the first half solution $\boldsymbol{\zeta}(t)$ in $\boldsymbol{X}(t)$ of Equation (8) stays in the manifold almost surely when $\boldsymbol{\zeta}(0) \in \Gamma$. (ii) We assume that $\mathcal{L} \in \mathcal{C}^5$, so $\boldsymbol{b}, \sigma \in \mathcal{C}^4$. (iii) We know that Γ is compact by Assumption 3.4. Then we can directly apply Lemma B.3 in Malladi et al. (2022) and Lemma 26 in Li et al. (2019). $\square$

Lemma G.15 (Adaption of Lemma I.41 in Gu et al. (2023b))**.** *Given drift term and diffusion term $\boldsymbol{b}, \sigma \in G^{\alpha}$ and Lipschitz. Let $s \in [0, T]$ and $g \in G^{\alpha}$. Then for $t \in [s, T]$, we can define:*

$$u(\boldsymbol{x}, s, t) := \mathbb{E}_{\boldsymbol{X}_t \sim \mathcal{P}_X(\boldsymbol{x}, s, t)}[g(\boldsymbol{X}_t)].$$

where $\mathcal{P}_X(\boldsymbol{x}, s, t)$ denotes the distribution of $\boldsymbol{X}_t$ with the initial condition $\boldsymbol{X}(s) = \boldsymbol{x}$. Then $u(\cdot, s, t) \in G^{\alpha}$ uniformly in s, t.

G.4.2 Proof of the Approximation for Slow SDE of AGMs

For the giant step constant $\beta \in (0, 0.5)$, we define several quantities $a_1 = \frac{1.5-2\beta}{1-\beta} \in (1, 1.5)$, $a_2 = \frac{1}{1-\beta} \in (1, 2)$, $a_3 = \frac{1.5-1.5\beta}{1-\beta} = 1.5$, and $a_4 = \frac{2-2\beta}{1-\beta} = 2$. In this part, we will show that only a_1 and a_2 would impact the error bound in our approximation theorem.

The following lemma captures the difference between the SDEs' and the AGMs' first and second moment changes, as a key step to control the approximation error, utilizing the moment calculation results from the last section.

Lemma G.16. *If $\|\boldsymbol{\theta}^{(0)} - \phi^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$, then it holds for all $0 \le n \le \lfloor T/\eta_e \rfloor$ and $1 \le i \le d$ that*

$$\left|\mathbb{E}[\Delta_i^{(n)} - \tilde{\Delta}_i^{(n)} \mid \mathcal{E}_0^{(nR_{\mathrm{grp}})}]\right| \le c_1 \left(\eta_e^{a_1} (\log \frac{1}{\eta_e})^b + \eta_e^{a_2} (\log \frac{1}{\eta_e})^b\right),$$

$$\left|\mathbb{E}[\Delta_i^{(n)}\Delta_j^{(n)} - \tilde{\Delta}_i^{(n)}\tilde{\Delta}_j^{(n)} \mid \mathcal{E}_0^{(nR_{\mathrm{grp}})}]\right| \le c_1 \left(\eta_e^{a_1} (\log \frac{1}{\eta_e})^b + \eta_e^{a_2} (\log \frac{1}{\eta_e})^b\right),$$

$$\mathbb{E}\left[\left|\prod_{s=1}^{6} \Delta_{i_s}^{(n)} \mid \mathcal{E}^{(nR_{\mathrm{grp}})}\right|\right] \le c_1^2 \eta_e^{2a_1} (\log \frac{1}{\eta_e})^{2b},$$

$$\mathbb{E}\left[\left|\prod_{s=1}^{6} \tilde{\Delta}_{i_s}^{(n)} \mid \mathcal{E}^{(nR_{\mathrm{grp}})}\right|\right] \le c_1^2 \eta_e^{2a_1} (\log \frac{1}{\eta_e})^{2b},$$

where c_1 and b are constants independent of η_e and g.

Proof. According to Appendix G.2, we have that

$$\mathbb{E}\left[\left|\prod_{s=1}^{6} \Delta_{i_s}^{(n)} \mid \mathcal{E}^{(nR_{\mathrm{grp}})}\right|\right] = \mathcal{O}(\eta^{3-3\beta}).$$

We can further use Corollary G.2, Lemma G.10, Lemma G.11, and Lemma G.12, which gives

$$\left|\mathbb{E}[\Delta_i^{(n)} - \eta_e b_i^{(n)}]\right| \le c_2 \left(\eta_e^{a_1} (\log \frac{1}{\eta_e})^b + \eta_e^{a_2} (\log \frac{1}{\eta_e})^b\right), \tag{9}$$

$$\left|\mathbb{E}[\Delta_i^{(n)}\Delta_j^{(n)} - \eta_e \sum_{l=1}^{d} \sigma_{i,l}^{(n)} \sigma_{l,j}^{(n)}]\right| \le c_2 \left(\eta_e^{a_1} (\log \frac{1}{\eta_e})^b + \eta_e^{a_2} (\log \frac{1}{\eta_e})^b\right) \tag{10}$$

$$\mathbb{E}\left[\left|\prod_{s=1}^{6} \Delta_{i_s}^{(n)}\right|\right] \le c_2^2 \eta_e^{2a_1} (\log \frac{1}{\eta_e})^{2b}. \tag{11}$$

Notice that the above equations uses $a_1 < a_3$ and $a_2 < a_4$ for all $\beta \in (0, 0.5)$. These three equations and Lemma G.14 give the Lemma. $\square$

Lemma G.17. *For a test function $g \in \mathcal{C}^3$, and we define $u_{l,n}(\boldsymbol{x}) := u(\boldsymbol{x}, l\eta_e, n\eta_e) = \mathbb{E}_{\boldsymbol{X}_t \sim \mathcal{P}(\boldsymbol{x}, l\eta_e, n\eta_e)}[g(\boldsymbol{X}_t)]$. If $\|\boldsymbol{\theta}^{(0)} - \phi^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$, then for all $0 \le l \le n-1$, and $1 \le n \le \lfloor T/\eta_e \rfloor$, it holds that*

$$\left|\mathbb{E}[u_{l+1,n}(\bar{\boldsymbol{X}}^{(lR_{\mathrm{grp}})} + \Delta^{(l)}) - u_{l+1,n}(\bar{\boldsymbol{X}}^{(lR_{\mathrm{grp}})} + \tilde{\Delta}^{(l)}) \mid \bar{\boldsymbol{X}}^{(lR_{\mathrm{grp}})}]\right| \le C_{g,3}(\eta_e^{a_1} + \eta_e^{a_2}) \log(\frac{1}{\eta_e})^b,$$

where $C_{g,3}$ is some positive constant independent of η_e but can depend on g.

Proof. Given $g \in \mathcal{C}^3$, by Lemma G.15, we have $u_{l,n}(\boldsymbol{x}) \in \mathcal{C}^3$ for all l and n. Which is to say that there exists a function $Q(\cdot) \in G$, such that the partial derivative of $u_{l,n}(\boldsymbol{X})$ with respect to $l, n, \boldsymbol{x}$ up to the third order is bounded by $Q(\boldsymbol{x})$. By the law of total expectation and triangle inequality,

$$
\begin{aligned}
&\left|\mathbb{E}[u_{l+1,n}(\hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})} + \Delta^{(l)}) - u_{l+1,n}(\hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})} + \tilde{\Delta}^{(l)}) \mid \hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})}]\right| \\
\leq &\underbrace{\left|\mathbb{E}[u_{l+1,n}(\hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})} + \Delta^{(l)}) - u_{l+1,n}(\hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})} + \tilde{\Delta}^{(l)}) \mid \hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}]\right|}_{I_1} \\
&+ \eta^{100} \underbrace{\mathbb{E}[\left|u_{l+1,n}(\hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})} + \Delta^{(l)})\right| \mid \hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}]}_{I_2} \\
&+ \eta^{100} \underbrace{\mathbb{E}[\left|u_{l+1,n}(\hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})} + \tilde{\Delta}^{(l)})\right| \mid \hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}]}_{I_3}.
\end{aligned}
$$

For I_2 and I_3, due to the compactness of Γ and $\boldsymbol{v} \preceq R_1$ from Assumption 3.3, $Q(\boldsymbol{x})$ can be bounded for some constant $C_{g,4}$ independent of η_e but could depend on test function g. Hence, we have that $I_2 + I_3 \leq C_{g,4}\eta^{100}$.

Using the triangle inequality, we first decompose I_1 into several terms as

$$
\begin{aligned}
I_1 \leq &\underbrace{\sum_{i=1}^{d} \left|\mathbb{E}\left[\frac{\partial u_{l,n}}{\partial X_i}(\hat{\tilde{\boldsymbol{X}}}^{(R_{\mathrm{grp}})}) \left(\Delta_i^{(l)} - \tilde{\Delta}_i^{(l)}\right) \mid \hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}\right]\right|}_{I_{1,1}} \\
&+ \underbrace{\frac{1}{2}\sum_{1\leq i,j\leq d} \left|\mathbb{E}\left[\frac{\partial^2 u_{l,n}}{\partial X_i \partial X_j}(\hat{\tilde{\boldsymbol{X}}}^{(R_{\mathrm{grp}})}) \left(\Delta_j^{(l)}\Delta_i^{(l)} - \tilde{\Delta}_i^{(l)}\tilde{\Delta}_j^{(l)}\right) \mid \hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}\right]\right|}_{I_{1,2}} \\
&+ |\mathcal{R}| + |\tilde{\mathcal{R}}|,
\end{aligned}
$$

where the third order remainders $\mathcal{R}$ and $\tilde{\mathcal{R}}$ are

$$
\begin{aligned}
\mathcal{R} &= \frac{1}{6}\sum_{1\leq i,j,k\leq d} \left|\mathbb{E}\left[\frac{\partial^3 u_{l,n}}{\partial X_i \partial X_j \partial X_k}(\hat{\tilde{\boldsymbol{X}}}^{(R_{\mathrm{grp}})} + \alpha\Delta^{(l)}) \left(\Delta_j^{(l)}\Delta_i^{(l)}\Delta_k^{(l)}\right) \mid \hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}\right]\right| \\
\tilde{\mathcal{R}} &= \frac{1}{6}\sum_{1\leq i,j,k\leq d} \left|\mathbb{E}\left[\frac{\partial^3 u_{l,n}}{\partial X_i \partial X_j \partial X_k}(\hat{\tilde{\boldsymbol{X}}}^{(R_{\mathrm{grp}})} + \tilde{\alpha}\tilde{\Delta}^{(l)}) \left(\tilde{\Delta}_j^{(l)}\tilde{\Delta}_i^{(l)}\tilde{\Delta}_k^{(l)}\right) \mid \hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}\right]\right|,
\end{aligned}
$$

where $\alpha, \tilde{\alpha} \in (0,1)$. Again, notice that the Γ is compact and $vv \preceq R_1$, thus we can bound the derivatives of $u_{l,n}(\boldsymbol{x})$ for any $\boldsymbol{X}$ as

$$
\left|\frac{\partial u_{l+1,n}}{\partial \boldsymbol{X}_i}(\boldsymbol{X})\right| \leq C_{g,4}, \ \left|\frac{\partial^2 u_{l+1,n}}{\partial \boldsymbol{X}_i \partial \boldsymbol{X}_j}(\boldsymbol{X})\right| \leq C_{g,4}, \ \left|\frac{\partial^3 u_{l+1,n}}{\partial \boldsymbol{X}_i \partial \boldsymbol{X}_j \partial \boldsymbol{X}_k}(\boldsymbol{X})\right| \leq C_{g,4}. \tag{12}
$$

For the term $I_{1,1}$ and $I_{1,2}$, by applying Lemma G.16, we have that

$$
I_{1,1} \leq dc_1 C_{g,4}(\eta_e^{a_1} + \eta_e^{a_2})(\log\frac{1}{\eta_e})^b, \ I_{1,2} \leq \frac{d^2}{2}c_1 C_{g,4}(\eta_e^{a_1} + \eta_e^{a_2})(\log\frac{1}{\eta_e})^b.
$$

Next, we bound the remainders $\mathcal{R}$ and $\tilde{\mathcal{R}}$. By Cauchy-Schwarz inequality,

$$
\begin{aligned}
|\mathcal{R}| \leq &\frac{1}{6}\sum_{1\leq i,j,k\leq d} \sqrt{\mathbb{E}\left[\left(\frac{\partial^3 u_{l,n}}{\partial X_i \partial X_j \partial X_k}(\hat{\tilde{\boldsymbol{X}}}^{(R_{\mathrm{grp}})} + \alpha\Delta^{(l)})\right)^2 \mid \hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}\right]} \times \\
&\sqrt{\mathbb{E}\left[\left(\Delta_j^{(l)}\Delta_i^{(l)}\Delta_k^{(l)}\right)^2 \mid \hat{\tilde{\boldsymbol{X}}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}\right]} \\
\leq &\frac{d^3}{6}C_{g,4}c_1\eta_e^{a_1}\log(\frac{1}{\eta_e})^b,
\end{aligned}
$$

where the last inequality uses Lemma G.16 and Equation (12).

Similarly, we can prove that there exists a positive constant $C_{g,5}$ such that

$$|\tilde{\mathcal{R}}| \leq \frac{d^3}{6} C_{g,5} c_1 \eta_e^{a_1} \log(\frac{1}{\eta_e})^b.$$

Combining the bounds for I_1, I_2, and I_3 gives the lemma. □

G.5 Proof of Theorem G.2

Finally, we are ready to prove Theorem G.2.

Proof of Theorem G.2. For $0 \leq l \leq n = \lfloor \frac{T}{\eta^{0.75}} \rfloor$, we denote the random variable by $\hat{\boldsymbol{x}}_{l,n}$ such that follows a distribution $\mathcal{P}_{\boldsymbol{X}}(\hat{\bar{\boldsymbol{X}}}^{(lR_{\rm grp})}, l\eta_e, n\eta_e)$. When we set $l = n$, $\mathcal{P}(\hat{\boldsymbol{x}}_{n,n} = \hat{\bar{\boldsymbol{X}}}^{(nR_{\rm grp})})$ and setting $l = 0$ gives $\hat{\boldsymbol{x}}_{0,n} \sim \boldsymbol{X}(n\eta_e)$. Recall the previous definition that $u(\boldsymbol{x}, s, t) = \mathbb{E}_{\boldsymbol{X}_t \sim \mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x},s,t)}[g(\boldsymbol{X}_t)]$, and we define that $\mathcal{T}_{l+1,n} := u_{l+1,n}(\hat{\bar{\boldsymbol{X}}}^{(lR_{\rm grp})} + \Delta^{(l)}, (l+1)\eta_e, n\eta_e) - u_{l+1,n}(\hat{\bar{\boldsymbol{X}}}^{(lR_{\rm grp})} + \tilde{\Delta}^{(l)}, (l+1)\eta_e, n\eta_e)$. Using the definition of $\boldsymbol{x}_{l,n}$, we can rewrite the distance between AGMs and SDE measured by a test function g as

$$\begin{aligned}
&\left|\mathbb{E}\left[g(\bar{\boldsymbol{X}}^{(nR_{\rm grp})}) - g(\boldsymbol{X}(n\eta_e))\right]\right| \\
\leq &\left|\mathbb{E}\left[g(\boldsymbol{x}_{n,n}) - g(\boldsymbol{x}_{0,n}) \mid \mathcal{E}_0^{(nR_{\rm grp})}\right]\right| + \mathcal{O}(\eta^{100}).
\end{aligned}$$

The above equation uses the law of total expectation and the definition of δ-good event $\mathcal{E}_0^{(nR_{\rm grp})}$ in Definition G.1. Then the Triangle inequality gives

$$\begin{aligned}
\left|\mathbb{E}\left[g(\boldsymbol{x}_{n,n}) - g(\boldsymbol{x}_{0,n}) \mid \mathcal{E}_0^{(nR_{\rm grp})}\right]\right| &\leq \sum_{l=0}^{n-1} \left|\mathbb{E}\left[g(\hat{\boldsymbol{x}}_{l+1,n}) - g(\hat{\boldsymbol{x}}_{l,n}) \mid \mathcal{E}_0^{(nR_{\rm grp})}\right]\right| + \mathcal{O}(\eta^{100}) \\
&= \sum_{l=0}^{n-1} \left|\mathbb{E}\left[\mathcal{T}_{l+1,n} \mid \mathcal{E}_0^{(nR_{\rm grp})}\right]\right| + \mathcal{O}(\eta^{100}) \\
&= \sum_{l=0}^{n-1} \left|\mathbb{E}\left[\mathbb{E}\left[\mathcal{T}_{l+1,n} \mid \hat{\bar{\boldsymbol{X}}}^{(lR_{\rm grp})}, \mathcal{E}_0^{(nR_{\rm grp})}\right] \mid \mathcal{E}_0^{(nR_{\rm grp})}\right]\right| + \mathcal{O}(\eta^{100}) \\
&\leq \sum_{l=0}^{n-1} \mathbb{E}\left[\left|\mathbb{E}\left[\mathcal{T}_{l+1,n} \mid \hat{\bar{\boldsymbol{X}}}^{(lR_{\rm grp})}, \mathcal{E}_0^{(nR_{\rm grp})}\right]\right| \mid \mathcal{E}_0^{(nR_{\rm grp})}\right] + \mathcal{O}(\eta^{100}) \\
&\leq n C_{g,3}(\eta_e^{a_1} + \eta_e^{a_2}) \log(\frac{1}{\eta_e})^b \\
&\leq T C_{g,3}(\eta_e^{a_1-1} + \eta_e^{a_2-1}) \log(\frac{1}{\eta_e})^b.
\end{aligned}$$

where the second last inequality uses Lemma G.17. Recall that $a_1 = \frac{1.5-2\beta}{1-\beta}$, $a_2 = \frac{1}{1-\beta}$, $\beta \in (0, 0.5)$. Let $\beta = 0.25$, and we complete the proof. □

H Proof of Theorems in Section 5

H.1 Proof of Adam and AdamE's Implicit Biases with Label Noise

In this part, we give the proof of Theorem 5.1, Lemma 5.1 and Lemma 5.2.

Proof of Theorem 5.1. Recall the SDE formula in Equation (8) and Lemma G.9:

$$\begin{cases} \mathrm{d}\boldsymbol{\zeta}(t) = \partial\Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\zeta})\boldsymbol{S}(\boldsymbol{v})\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}_t - \frac{1}{2}\boldsymbol{S}_t\partial\Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\zeta})\boldsymbol{S}_t\partial^2(\nabla\mathcal{L})(\boldsymbol{\zeta})[\boldsymbol{P}\mathcal{V}_{\nabla^2\mathcal{L}'(\phi'_{(0)})}(\boldsymbol{P}\boldsymbol{\Sigma}_0\boldsymbol{P})\boldsymbol{P}]\mathrm{d}t, \\ \mathrm{d}\boldsymbol{v}(t) = c\left(V(\boldsymbol{\Sigma}(\boldsymbol{\zeta})) - \boldsymbol{v}\right)\mathrm{d}t. \end{cases}$$

Plugging in the following:

- The definition that $\boldsymbol{P} := \boldsymbol{S}_0^{1/2}$,
- Lemma G.2: For any $\boldsymbol{\zeta} \in \Gamma$ and $p.d$ matrix $\boldsymbol{S}$, $\partial\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})\boldsymbol{S}\nabla^2\mathcal{L}(\boldsymbol{\zeta}) = \mathbf{0}$,
- The label noise condition: $\boldsymbol{\Sigma}(\boldsymbol{\zeta}) := \alpha\nabla^2\mathcal{L}(\boldsymbol{\zeta})$ for any $\boldsymbol{\zeta} \in \Gamma$

yields the final result:

$$\begin{cases} \mathrm{d}\boldsymbol{\zeta}(t) = -\frac{\alpha}{2}\boldsymbol{S}_t\partial\Phi_{\boldsymbol{S}_t}(\boldsymbol{\zeta})\boldsymbol{S}_t\partial^2(\nabla\mathcal{L})(\boldsymbol{\zeta})[\boldsymbol{S}_t]\mathrm{d}t, \\ \mathrm{d}\boldsymbol{v}(t) = c\,(V(\boldsymbol{\Sigma}(\boldsymbol{\zeta})) - \boldsymbol{v})\,\mathrm{d}t. \end{cases} \tag{13}$$

The above equation completes the proof. $\square$

For Lemma 5.1 and Lemma 5.2, we first give the more general lemma, for which the above two lemmas are direct corollaries. We first recall the update rule of AdamE-λ, the variant of Adam proposed as a verification case of our main results

$$\begin{aligned} \boldsymbol{m}_{k+1} &:= \beta_1\boldsymbol{m}_k + (1-\beta_1)\nabla\ell_k(\boldsymbol{\theta}_k) \\ \boldsymbol{v}_{k+1} &:= \beta_2\boldsymbol{v}_k + (1-\beta_2)\nabla\ell_k(\boldsymbol{\theta}_k)^{\odot 2} \\ \theta_{k+1,i} &:= \theta_{k,i} - \eta\frac{m_{k+1,i}}{(v_{k+1,i})^{\lambda} + \epsilon} \quad \text{for all } i \in [d], \lambda \in (0,1). \end{aligned}$$

Notice that taking $\lambda = 1/2$ reduces AdamE-λ to Adam. Then we give the following Lemma.

Lemma H.1 (Adam and AdamE's Implicit Biases with Label Noise)**.** *With the label noise condition, $\epsilon \geq 0$, the fixed points of* (3) *in the AdamE-λ case satisfiy $\nabla tr(Diag(\boldsymbol{H})^{1-\lambda}) = O(\epsilon)$, where $\lambda \in [0,1)$.*

Proof of Lemma H.1. Consider the a fixed point $(\boldsymbol{\zeta}^*, \boldsymbol{v}^*)$ of the ODE (13). It must satisfy

$$S(\boldsymbol{v}^*)\partial\Phi_{S(\boldsymbol{v}^*)}(\boldsymbol{\zeta}^*)S(\boldsymbol{v}^*)\partial^2(\nabla\mathcal{L})(\boldsymbol{\zeta}^*)[S(\boldsymbol{v}^*)] = 0 \tag{14}$$

and

$$\boldsymbol{v}^* = V(\boldsymbol{\Sigma}(\boldsymbol{\zeta}^*)). \tag{15}$$

We first consider Equation (14). To simplify the notation, we denote that

$$\boldsymbol{S}^* := S(\boldsymbol{v}^*), \boldsymbol{P}_{\parallel}^* := \partial\Phi_{S(\boldsymbol{v}^*)}\boldsymbol{S}^*, \boldsymbol{H}^* = \nabla^2\mathcal{L}(\boldsymbol{\zeta}^*).$$

Integrating by parts gives

$$\partial^2\left(\nabla\mathcal{L}\right)[\boldsymbol{S}] = \nabla\left[\langle\nabla^2\mathcal{L}, \boldsymbol{S}\rangle\right] - \nabla\left(\boldsymbol{S}\right)\left[\nabla^2\mathcal{L}\right]. \tag{16}$$

We use $\boldsymbol{H}$ and $\nabla^2\mathcal{L}$ interchangeably to denote the Hessian matrix. For the first term, note that

$$\begin{aligned} \langle\boldsymbol{S}, \boldsymbol{H}\rangle &= \sum_{j,k}[\boldsymbol{S}]_{jk}\boldsymbol{H}_{jk} = \sum_{i,j,k}\boldsymbol{P}_{ji}\boldsymbol{H}_{jk}\boldsymbol{P}_{ki} \\ &= \sum_i[\boldsymbol{PHP}]_{ii} = \mathrm{tr}\left(\boldsymbol{PHP}\right) \\ &= \mathrm{tr}\left((\mathrm{Diag}\boldsymbol{H})^{1-\lambda}\right) + O(\epsilon\mathrm{tr}(\boldsymbol{H})), \end{aligned}$$

where the last equality comes from the update rule of AdamE-λ:

$$\boldsymbol{S} = (\mathrm{Diag}\boldsymbol{H})^{-\lambda} + O(\epsilon), \boldsymbol{P} = (\mathrm{Diag}\boldsymbol{H})^{-\lambda/2} + O(\sqrt{\epsilon}).$$

For the second term, we also plug in the update rule of Adam, and use h_j to denote $\boldsymbol{H}_{jj}$, which turns out to be the gradient of the same thing:

$$\begin{aligned}
\nabla(\boldsymbol{S})\left[\nabla^2\mathcal{L}\right] &= \sum_{j,k}\nabla\left([\boldsymbol{S}]_{jk}\right)\nabla^2_{jk}\mathcal{L} \\
&= \sum_{j}\nabla\left([\boldsymbol{S}]_{jj}\right)\nabla^2_{jj}\mathcal{L} \\
&= \sum_{j}\nabla\left(h_j^{-\lambda}\right)\cdot h_j + \mathcal{O}\left(\epsilon\sum_j h_j\right) \\
&= \sum_{j}\nabla\left(h_j\right)\cdot -\lambda h_j^{-\lambda} \\
&= -\frac{\lambda}{1-\lambda}\sum_j\nabla\left(-h_j^{1-\lambda}\right) + O(\epsilon\mathrm{tr}(\boldsymbol{H})) \\
&= -\frac{\lambda}{1-\lambda}\nabla\mathrm{tr}\left((\mathrm{Diag}\boldsymbol{H})^{1-\lambda}\right) + O(\epsilon\mathrm{tr}(\boldsymbol{H})).
\end{aligned}$$

Summarizing, our drift term can be represented as a constant multiple of

$$\begin{aligned}
\partial^2(\nabla\mathcal{L})(\boldsymbol{\zeta}^*)[S(\boldsymbol{v}^*)] &= \nabla\left[\langle\boldsymbol{H}^*, \boldsymbol{S}^*\rangle\right] - \nabla\left(\boldsymbol{S}^*\right)\left[\boldsymbol{H}^*\right] && (17)\\
&= \frac{1}{1-\lambda}\nabla\mathrm{tr}\left((\mathrm{Diag}\boldsymbol{H}^*)^{1-\lambda}\right) + O(\epsilon\nabla\mathrm{tr}(\boldsymbol{H}^*)), && (18)
\end{aligned}$$

from which we can rewrite the constant for the fixed points in Equation (14) as

$$\boldsymbol{S}^*\boldsymbol{P}^*_{\|}\nabla\mathrm{tr}\left((\mathrm{Diag}\boldsymbol{H})^{\frac{1}{2}}\right) = O(\epsilon).$$

W.L.O.G, we can decompose $\boldsymbol{P}^*_{\|}$ and $\boldsymbol{H}^*$ into block matrices as

$$\boldsymbol{P}^*_{\|} = \begin{pmatrix}\mathbf{0}, & \mathbf{0}\\ \mathbf{0}, & \boldsymbol{P}^*_{d-m}\end{pmatrix}, \boldsymbol{H}^* = \begin{pmatrix}\mathbf{0}, & \mathbf{0}\\ \mathbf{0}, & \boldsymbol{H}^*_{d-m}\end{pmatrix},$$

where $\boldsymbol{P}^*_{d-m}, \boldsymbol{H}^*_{d-m} \in \mathbb{R}^{(d-m)\times(d-m)}$ are full-rank matrices. Under this decomposition, the first m diagonal elements in $(\mathrm{Diag}(\boldsymbol{H}))^{1/2}$ is 0, and the first m diagonal elements in $\boldsymbol{S}^*$ is $\epsilon \geq 0$. Specifically

$$\boldsymbol{S}^* = \begin{pmatrix}\epsilon\boldsymbol{I}_m, & \mathbf{0}\\ \mathbf{0}, & \boldsymbol{S}^*_{d-m}\end{pmatrix}, \mathrm{Diag}(\boldsymbol{H}^*) = \begin{pmatrix}\mathbf{0}, & \mathbf{0}\\ \mathbf{0}, & \mathrm{Diag}(\boldsymbol{H}^*_{d-m})\end{pmatrix}.$$

Then the constraint in Equation (14) can be reduced into

$$-\nabla_\Gamma\mathrm{tr}\left(\mathrm{Diag}(\boldsymbol{H}^*)^{1-\lambda}\right) = O(\epsilon),$$

where $\nabla_\Gamma f$ stands for the gradient of a function f projected to the tangent space of the manifold Γ.

Then we complete the proof. Also notice that, taking $\epsilon = 0$, gives Lemma 5.2, and futher taking $\lambda = 1/2$ gives the results in Lemma 5.1. $\square$

H.2 Proof of Lemma 5.3

Proof. We only prove the second argument in Lemma 5.3 with any $e_0 \in (0, 1]$, since taking $e_0 = 0.5$ yields the first argument. First, we recall that the minimizer manifold Γ is defined as

$$\Gamma := \left\{\boldsymbol{\theta}|\langle\boldsymbol{z}_i, \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}\rangle = y_i, \forall i \in [n]\right\}.$$

So if any $\boldsymbol{\theta} = \binom{\boldsymbol{u}}{\boldsymbol{v}}$ belongs to Γ, and another $\tilde{\boldsymbol{\theta}} = \binom{\tilde{\boldsymbol{u}}}{\tilde{\boldsymbol{v}}}$ satisfies that $\tilde{u}_i^{\odot 2} - \tilde{v}_i^{\odot 2} = u_i^{\odot 2} - v_i^{\odot 2}$ for any $i \in [d]$, then $\tilde{\boldsymbol{\theta}}$ also belongs to Γ.

Next, we derive the explicit expression of the Hessian matrix when $\boldsymbol{\theta} \in \Gamma$:

$$\nabla^2 \mathcal{L}(\boldsymbol{\theta}) = \frac{2}{n}\sum_{i=1}^{n} 2\binom{\boldsymbol{z}_i\odot\boldsymbol{u}}{-\boldsymbol{z}_i\odot\boldsymbol{v}}\binom{\boldsymbol{z}_i\odot\boldsymbol{u}}{-\boldsymbol{z}_i\odot\boldsymbol{v}}^{\top} + \left(\langle \boldsymbol{z}_i, \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}\rangle - y_i\right)\begin{pmatrix} \mathrm{diag}(\boldsymbol{z}) & 0 \\ 0 & -\mathrm{diag}(\boldsymbol{z}) \end{pmatrix}$$
$$= \frac{4}{n}\sum_{i=1}^{n} \binom{\boldsymbol{z}_i\odot\boldsymbol{u}}{-\boldsymbol{z}_i\odot\boldsymbol{v}}\binom{\boldsymbol{z}_i\odot\boldsymbol{u}}{-\boldsymbol{z}_i\odot\boldsymbol{v}}^{\top}.$$

Hence, we have that

$$\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{e_0}) \propto \sum_{i=1}^{d}(|u_i|^{2e_0} + |v_i|^{2e_0}),$$

and $\left\|\boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}\right\|_{e_0}^{e_0} = \sum_{i=1}^{d}|u_i^2 - v_i^2|^{e_0}$. Let $e_0 \in (0, 1]$, and we recall that our goal is to prove that given the following condition

$$\boldsymbol{\theta} \in \arg\min_{\boldsymbol{\theta}'\in\Gamma} \mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{e_0}) = \arg\min_{\boldsymbol{\theta}'\in\Gamma}\sum_{i=1}^{d}(|u_i|^{2e_0} + |v_i|^{2e_0}), \tag{19}$$

it holds that $\boldsymbol{\theta} \in \arg\min_{\boldsymbol{\theta}'\in\Gamma}\|\widehat{\boldsymbol{w}}\|_{e_0}$.

First, we prove that $u_i = 0 \vee v_i = 0$ holds for any $i \in [d]$. Assume for the contrary that there exists some i such that $u_i \neq 0$ and $v_i \neq 0$, then we construct another reference point $\tilde{\boldsymbol{\theta}} = \binom{\tilde{\boldsymbol{u}}}{\tilde{\boldsymbol{v}}}$ by letting $\tilde{\boldsymbol{u}}$ and $\boldsymbol{u}$ agree on all indices other than i, and that

$$\tilde{u}_i = \sqrt{u_i^2 - v_i^2}, \tilde{v}_i = 0, \quad \text{if } |u_i| \geq |v_i|, \tag{20}$$

$$\tilde{u}_i = 0, \tilde{v}_i = \sqrt{v_i^2 - u_i^2}, \quad \text{otherwise.} \tag{21}$$

With this construction, $\tilde{\boldsymbol{u}}^{\odot 2} - \tilde{\boldsymbol{v}}^{\odot 2} = \boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}$, so $\tilde{\boldsymbol{\theta}} \in \Gamma$. One can observe that

$$|\tilde{u}_i|^{2e_0} + |\tilde{v}_i|^{2e_0} < |u_i|^{2e_0} + |v_i|^{2e_0},$$

which contradicts the condition in Equation (19).

Now we are ready to prove $\boldsymbol{\theta} \in \arg\min_{\boldsymbol{\theta}'\in\Gamma}\|\widehat{\boldsymbol{w}}\|_{e_0}$. Also, we prove this by contradiction. Now assume $\boldsymbol{\theta} \notin \arg\min_{\boldsymbol{\theta}'\in\Gamma}\left\|\boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}\right\|_{e_0}$. There must exist some $\tilde{\boldsymbol{\theta}} \in \Gamma$ such that

$$\left\|\tilde{\boldsymbol{u}}^{\odot 2} - \tilde{\boldsymbol{v}}^{\odot 2}\right\|_{e_0} < \left\|\boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2}\right\|_{e_0}.$$

W.L.O.G., one can assume that for any $i \in [d]$, either $\tilde{u}_i = 0$ or $\tilde{v}_i = 0$, else we can construct another minimizer that preserves $\left\|\tilde{\boldsymbol{u}}^{\odot 2} - \tilde{\boldsymbol{v}}^{\odot 2}\right\|_{e_0}$ as Equation (20) and Equation (21). However, given the condition $u_i = 0 \vee v_i = 0$, we have that

$$\sum_{i=1}^{d}|u_i^2 - v_i^2|^{e_0} = \sum_{i=1}^{d}|u_i|^{2e_0} + |v_i|^{2e_0},$$

and $\sum_{i=1}^{d}|\tilde{u}_i^2 - \tilde{v}_i^2|^{e_0} = \sum_{i=1}^{d}|\tilde{u}_i|^{2e_0} + |\tilde{v}_i|^{2e_0}$, which indicates that $\sum_{i=1}^{d}|\tilde{u}_i|^{2e_0} + |\tilde{v}_i|^{2e_0} < \sum_{i=1}^{d}|u_i|^{2e_0} + |v_i|^{2e_0}$, a contradiction. $\square$

I Regularizers under label noise for AGMs in Table 1

In this section, we provide additional discussions on the regularizer for the AGM optimizers in Table 1 besides Adam, AdamE (refer to Appendix H), and Shampoo (refer to Appendix J).

SGD. Under label noise, the implicit bias of SGD has been extensively studied by previous works; As discussed in Appendix B, approaches such as fixed point analysis (Blanc et al., 2020), slow SDE (Li et al., 2021b) and implicit gradient regularization (Barrett and Dherin, 2020) all agree on the result that SGD implicitly regularizes $\mathrm{tr}(\boldsymbol{H})$ on the minimizer manifold. Our work provides a new insight on the implicit bias of SGD by comparing with that of Adam. Specifically, SGD treats each direction equally which results in a rotation invariant $\mathrm{tr}(\boldsymbol{H})$ as the implicit regularizer, while Adam has the second-order momentum as a denominator, so Adam regularizes the entries with small gradients relatively faster, as is indicated in its implicit bias $\mathrm{tr}((\mathrm{diag}\boldsymbol{H})^{1/2})$.

RMSProp. The RMSProp optimizer Hinton et al. (2012) can be seen as a special case of Adam, where $\beta_1 = 0$. One can observe that β_1 does not appear in the slow SDE system, which implies that as long as $1 - \beta_1$ is of constant order, the choice of β_1 has nothing to do with the dynamics of Adam on the minimizer manifold. The intuition is that, after the iteration approaches the manifold, the gradient $\nabla \mathcal{L}(\boldsymbol{\theta}_k)$ moves very slowly as k proceeds. Since the momentum only captures $\mathcal{O}(\log 1/\eta)$ past steps, the different between momentum and the gradient at that step becomes negligible. Therefore, RMSProp possesses an implicit bias identical to Adam: $\operatorname{tr}((\operatorname{diag}\boldsymbol{H})^{1/2})$.

Adam-mini and Adalayer. Adam-mini (Zhang et al., 2025) and Adalayer (Zhao et al., 2025) belong to the same kind of variant of Adam that partitions the parameters. In Adam-mini the partitions are blocks, and in Adalayer the partitions are layers. In the sequel, we provide a brief derivation of the implicit bias of "Partitioned Adam", which is applicable to any kind of optimizer whose functions V and S can be expressed in the form of

$$V(\boldsymbol{M})_i := \frac{1}{|B_{\pi(i)}|} \sum_{j \in B_{\pi(i)}} M_{jj}$$
$$S(\boldsymbol{v}) := \operatorname{Diag}\bigl(1/(\sqrt{\boldsymbol{v}} + \epsilon)\bigr)$$

where $\mathcal{B} = \{B_1, B_2, \cdots, B_N\}$ is a partition of $[d]$, and for each $i \in [d]$, $\pi(i)$ denotes the index of the set containing i, i.e. $i \in B_{\pi(i)}$. We derive the case for $\epsilon = 0$.

Recall from the proof of Lemma H.1 that the gradient of the implicit regularizer being minimized on manifold can be expressed as

$$\partial^2 (\nabla \mathcal{L}) [\boldsymbol{S}] = \nabla [\langle \boldsymbol{S}, \boldsymbol{H} \rangle] - \nabla (\boldsymbol{S}) [\boldsymbol{H}]. \tag{22}$$

In our case $\boldsymbol{S}$ is diagonal, so we can calculate the contribution of each set in the partition, and add them up. Next we focus on a single set, and re-index it as $\{1, 2, \cdots, G\}$ without loss of generality. In this set $\operatorname{tr}\boldsymbol{H}/G$ is used as a shared second-order momentum, so we have

$$\boldsymbol{S} = \left[\frac{\operatorname{tr}\boldsymbol{H}}{G} \boldsymbol{I}_G\right]^{1/2}.$$

Combining with $\boldsymbol{P} = \boldsymbol{S}^{1/2}$ gives us

$$\langle \boldsymbol{S}, \boldsymbol{H} \rangle = \operatorname{tr}(\boldsymbol{PHP}) = \sqrt{G \cdot \operatorname{tr}\boldsymbol{H}}.$$

For the second term, we again denote $h_j := \boldsymbol{H}_{jj}$, and we further denote $t := \operatorname{tr}\boldsymbol{H}/G = \frac{1}{G}\sum_{j=1}^{G} h_j$.

$$\begin{aligned}
\nabla (\boldsymbol{S}) [\boldsymbol{H}] &= \sum_j \nabla \left(t^{-1/2}\right) \cdot h_j \\
&= \sum_j \nabla (t) \cdot h_j \cdot -\frac{1}{2} t^{-3/2} \\
&= G \cdot \nabla (t) \cdot -\frac{1}{2} t^{-1/2} \\
&= -G \nabla \left(t^{1/2}\right) = -\nabla \sqrt{G \cdot \operatorname{tr}\boldsymbol{H}}.
\end{aligned}$$

Plugging into (22) gives the implicit bias contributed by this set as $\sqrt{G \cdot \operatorname{tr}\boldsymbol{H}}$. Finally, summing up all the sets, we conclude the overall implicit bias as

$$\sum_{i \in [N]} \sqrt{|B_i| \cdot \operatorname{tr}\boldsymbol{H}_{B_i}}.$$

Here, $\boldsymbol{H}_{B_i}$ means the submatrix of $\boldsymbol{H}$ if we restrict the rows and columns to B_i.

J Shampoo Optimizer as an AGM

Inspired by the matrix form of the parameter instead of the vector form, the Shampoo optimizer was proposed for faster convergence in optimization (Gupta et al., 2018).

J.1 Shampoo under the AGM Framework

Specifically, given a matrix gradient $\boldsymbol{G} \in \mathbb{R}^{d_1 \times d_2}$, whose vectorization is $\boldsymbol{g} = \mathrm{vec}(\boldsymbol{G}) \in \mathbb{R}^{d \times 1}$, where $d = d_1 \cdot d_2$. The original Shampoo considers a Kronecker-factored approximation $(\boldsymbol{S}_1)^{2p} \otimes (\boldsymbol{S}_2)^{2p}$ for the second moment for the flattened gradient $\mathbb{E}\left[\boldsymbol{g}\boldsymbol{g}^\top\right]$, where $\boldsymbol{S}_1 := \mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$, $\boldsymbol{S}_2 := \mathbb{E}[\boldsymbol{G}^\top\boldsymbol{G}]$, the operator $\otimes$ represents the Kronecker product, and p is some positive constant. In practice, we often approximate the second moment $\mathbb{E}[\boldsymbol{g}\boldsymbol{g}^\top]$ with an exponentially moving average (EMA) on the outer product (Morwani et al., 2024; Lin et al., 2025). And for the exponent p, the original Shampoo uses $p = 1/4$ and later works (Anil et al., 2020; Shi et al., 2023; Morwani et al., 2024) suggest using $p = 1/2$. Here, we consider the most practically used version of Shampoo, whose vector version follows the update rule with EMA, and $p = 1/2$:

$$\boldsymbol{S}_1 \leftarrow (1-\beta_2)\boldsymbol{S}_1 + \beta_2 \boldsymbol{G}\boldsymbol{G}^\top, \quad \boldsymbol{S}_2 \leftarrow (1-\beta_2)\boldsymbol{S}_2 + \beta_2 \boldsymbol{G}^\top\boldsymbol{G},$$
$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \boldsymbol{S}^{-1/2}\boldsymbol{g},$$

where $\boldsymbol{S} := \boldsymbol{S}_1^{2p} \otimes \boldsymbol{S}_2^{2p}$. And we recall our AGM framework with respect to the update rule for the second moment and the training parameter that

$$\boldsymbol{v}_{k+1} := \beta_2 \boldsymbol{v}_k + (1-\beta_2) V\left(\nabla \ell_k(\boldsymbol{\theta}_k) \nabla \ell_k(\boldsymbol{\theta}_k)^\top\right)$$
$$\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k - \eta S(\boldsymbol{v}_{k+1}) \boldsymbol{m}_{k+1}.$$

Now we specify the function V, S, and the dimension d' for Shampoo. First, we defne two functions $V_1(\boldsymbol{g}\boldsymbol{g}^\top) := \mathrm{vec}(\boldsymbol{G}\boldsymbol{G}^\top)$, and $V_2(\boldsymbol{g}\boldsymbol{g}^\top) := \mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{G})$. Thus, we can write out $V(\boldsymbol{g}\boldsymbol{g}^\top)$ as

$$V(\boldsymbol{g}\boldsymbol{g}^\top) = \left(V_1(\boldsymbol{g}\boldsymbol{g}^\top)^\top, V_2(\boldsymbol{g}\boldsymbol{g}^\top)^\top\right)^\top \in \mathbb{R}^{d_1^2 + d_2^2} \tag{23}$$

One may notice that here $\boldsymbol{v} = (\mathrm{vec}(\boldsymbol{S}_1)\top, \mathrm{vec}(\boldsymbol{S}_2)\top)\top$. Finally, we define the matrix reshape function $P : \mathbb{R}^{m^2 \times n^2} \to \mathbb{R}^{mn \times mn}$ as

$$P(\boldsymbol{M})[mj + i, mj' + i'] = \boldsymbol{M}[mi + i', nj + j'],$$

where $i, i' \in [0, 1, 2, \ldots, m-1]$ and $j, j' \in [0, 1, 2, \ldots, n-1]$, and since P is a bijective by its definition, its inverse mapping P^{-1} is thus well-defined. Also, one important property (Morwani et al., 2024) for the reshape function P and Kronecker product is that

$$\boldsymbol{H} = \boldsymbol{A} \otimes \boldsymbol{B} \iff P^{-1}(\boldsymbol{H}) = \mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{B})^\top$$

Thus for $S(\boldsymbol{v})$, we have

$$S(\boldsymbol{v}) = (\boldsymbol{S}_1 \otimes \boldsymbol{S}_2)^{-1/2} = P\left(\mathrm{vec}(\boldsymbol{S}_1)\mathrm{vec}(\boldsymbol{S}_2)^\top\right)^{-1/2}. \tag{24}$$

Then we give the expressions for V and S via Equation (23) and Equation (24).

J.2 Discussion on Shampoo's Implicit Bias under Label Noise

Recall that for all AGMs, under label noise, Equation (14) and Equation (15) hold as

$$\boldsymbol{S}^* \partial \boldsymbol{P}_\parallel^* \partial^2 (\nabla \mathcal{L})(\boldsymbol{\zeta}^*)[\boldsymbol{S}^*] = 0$$
$$\boldsymbol{v}^* = V(\boldsymbol{\Sigma}(\boldsymbol{\zeta}^*)).$$

To see if shampoo has and explicit regularizer, we have to calculate the following term as the first step

$$\boldsymbol{A}(\boldsymbol{\zeta}^*) := \partial^2 \left(\nabla \mathcal{L}^*\right)[\boldsymbol{S}^*] = \nabla\left[\langle \boldsymbol{H}^*, \boldsymbol{S}^* \rangle\right] - \nabla\left(\boldsymbol{S}^*\right)[\boldsymbol{H}^*], \tag{25}$$

where we denote $\nabla \mathcal{L}^* = \nabla \mathcal{L}(\boldsymbol{\zeta}^*)$. And the question: Does Shampoo have an explicit regularizer under label noise? is equivalent to the question: Does the vector field $\partial^2 \left(\nabla \mathcal{L}\right)[\boldsymbol{S}]$ has the corresponding potential function?

The above equivalence is obvious since if we can find the corresponding potential function $\psi(\boldsymbol{\zeta}^*)$ for the vector field $\partial^2 \left(\nabla \mathcal{L}^*\right)[\boldsymbol{S}^*] \in \mathbb{R}^d$, then we can write our the constraint in Equation (14) as

$$\nabla_\Gamma \psi(\boldsymbol{\zeta}^*) = 0,$$

and the regularizer is exactly $\psi(\boldsymbol{\zeta}^*)$, and vice versa.

Unfortunately, even in a simple case as a sanity check, where $\boldsymbol{H}^*$ is assumed to be diagonal (for example, the scenario of diagonal net), the potential function ψ does not exist. We assume that that the Hessian $\boldsymbol{H}^*$ has the following form.

$$\boldsymbol{H}^* = \mathrm{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_d),$$

where $\lambda_i \geq 0$ for any $1 \leq i \leq d$. And we have $\Sigma^* = \alpha \boldsymbol{H}^*$. In addition, a direct calculation gives

$$\begin{aligned}
\Sigma_R &:= \mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top] = \mathrm{Diag}\left(\sum_{i=0}^{d_2-1}(\lambda_1 + id_1), \sum_{i=0}^{d_2-1}(\lambda_2 + id_1), \ldots, \sum_{i=0}^{d_2-1}(\lambda_{d_1} + id_1)\right) \\
&=: \mathrm{Diag}(r_1, r_2, \ldots, r_{d_1}) \in \mathbb{R}^{d_1 \times d_1}, \\
\Sigma_L &:= \mathbb{E}[\boldsymbol{G}^\top\boldsymbol{G}] = \mathrm{Diag}\left(\sum_{i=0}^{d_1-1}(\lambda_1 + id_1), \sum_{i=0}^{d_1-1}(\lambda_2 + id_1), \ldots, \sum_{i=0}^{d_1-1}(\lambda_{d_2} + id_1)\right) \\
&=: \mathrm{Diag}(l_1, l_2, \ldots, l_{d_2}) \in \mathbb{R}^{d_2 \times d_2}.
\end{aligned}$$

Therefore, the preconditioner matrix $\boldsymbol{S}^*$ in Equation (24) can be written as

$$\boldsymbol{S}^* = \mathrm{Diag}\left(l_1^{-1/2} \cdot \Sigma_R^{-1/2}, l_2^{-1/2} \cdot \Sigma_R^{-1/2}, \ldots, l_{d_2}^{-1/2} \cdot \Sigma_R^{-1/2}\right) \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}.$$

One can easily verify that the curl $\nabla \times \boldsymbol{A}(\boldsymbol{\zeta}^*) \neq 0$, then there does not exist a potential function ψ such that $\boldsymbol{A}(\boldsymbol{\zeta}^*) = \nabla\psi(\boldsymbol{\zeta}^*)$, which is a direct corollary by the Stokes-Cartan theorem. (Theorem 16.11 in Lee (2012)). Thus, in general the regularization effect of Shampoo under label noise cannot be reduced to an explicit regularizer for shampoo, for which the diagonal case is a counterexample.