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

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Despite the popularity of Adam optimizer in practice, most theoretical analyses
2 study SGD as a proxy and little is known about how the solutions found by Adam
3 differ. In this paper, we show that Adam reduces a specific form of sharpness
4 measure shaped by its adaptive updates, leading to qualitatively different solutions
5 from SGD. When the training loss is small, Adam wanders around the manifold
6 of minimizers and takes semi-gradients to minimize this sharpness measure in
7 an adaptive manner, a behavior we rigorously characterize via a continuous-time
8 approximation using stochastic differential equations. We further illustrate how
9 this behavior differs from that of SGD in a well-studied setting: When training
10 overparameterized models with label noise, SGD has been shown to minimize
11 the trace of the Hessian matrix, $\text{tr}(\mathbf{H})$, whereas we prove that Adam minimizes
12 $\text{tr}(\text{Diag}(\mathbf{H})^{1/2})$ instead. In solving sparse linear regression with diagonal linear
13 networks, Adam provably achieves better sparsity and generalization than SGD
14 due to this difference. Finally, we note that our proof framework applies not only to
15 Adam but also to many other adaptive gradient methods, including but not limited
16 to RMSProp, Adam-mini, and Adalayer. This provides a unified perspective for
17 analyzing how adaptive optimizers reduce sharpness and may offer insights for
18 future optimizer design.

19 1 Introduction

20 Due to the non-convexity of the loss landscape, neural networks trained in different ways can perform
21 very differently on the test set, even if they achieve the same training loss or accuracy (Zhang et al.,
22 2017; Keskar et al., 2017; Liu et al., 2023; Saunshi et al., 2024). To mathematically understand the
23 generalization of neural networks, especially for over-parameterized models that admit many global
24 minimizers, a key step is to understand the *implicit bias* of optimization methods (Neyshabur et al.,
25 2014; Soudry et al., 2018). That is, beyond just minimizing the training loss, *what kinds of solutions*
26 *are different optimizers implicitly biased toward?*

27 Many theoretical works on implicit bias focused on (full-batch) gradient descent or its continuous
28 variant, gradient flow. This includes the works on the implicit bias towards max-margin classi-
29 fiers (Soudry et al., 2018; Nacson et al., 2019; Lyu and Li, 2020; Ji and Telgarsky, 2020), implicit
30 bias towards min-norm solutions (Lyu et al., 2024), and equivalence to kernel methods (Jacot et al.,
31 2018; Chizat et al., 2019).

32 All these implicit bias characterizations hold, or can be readily extended, to the stochastic variant
33 of gradient descent, *i.e.*, Stochastic Gradient Descent (SGD). Another line of works (Blanc et al.,
34 2020; Damian et al., 2021; Li et al., 2021b) demonstrated that the gradient noise in SGD induces an
35 additional form of implicit bias that reduces the *sharpness* of the solutions, a generalization measure
36 that has been long observed to correlate with generalization (Hochreiter and Schmidhuber, 1997;
37 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 $\text{tr}(\text{Diag}(\mathbf{H})^{1/2})$ where \mathbf{H} is the Hessian matrix. Compared to the $\text{tr}(\mathbf{H})$ 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 B.
3. Technically, our analysis holds for a general class of adaptive gradient methods (AGMs), including Adam, RMSProp, Adam-mini, and Adalayer. 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 $\mathcal{L}(\theta_k) - \mathcal{L}^*$.

2 Related Work

Implicit Bias of SGD. Parallel work on *implicit gradient regularization* (IGR) derives higher-order terms for full-batch GD (Barrett and Dherin, 2020) and extends to Adam (Cattaneo et al., 2024; Cattaneo and Shigida, 2025). While Cattaneo et al. (2024) argued that Adam anti-regularizes sharpness when $\beta_1 < \beta_2$, our $\mathcal{O}(\eta^{-2})$ -time SDE analysis shows Adam still regularizes sharpness under these settings, overturning their conclusion.

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.

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.

3 Preliminaries

Notations. Unless otherwise stated, for a square matrix M , $\text{diag}(M)$ denotes the vector consisting of its diagonal entries. The notation Diag has two usages: For a vector v , $\text{Diag}(v)$ denotes the diagonal matrix with v on its diagonal; and for a square matrix M , $\text{Diag}(M)$ denotes the diagonal matrix that only keeps M 's diagonal entries and equals 0 elsewhere, i.e. $\text{Diag}(M) \stackrel{\text{def}}{=} \text{Diag}(\text{diag}(M))$. For two vectors u, v with the same dimension d , $u \odot v$ denotes element-wise multiplication $(u_1 v_1, \dots, u_d v_d)$. For any exponent p , $v^{\odot p}$ denotes element-wise exponentiation, i.e. $v^{\odot p} = (v_1^p, \dots, v_d^p)$, and \sqrt{v} means $v^{\odot 1/2}$.

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

Loss Functions. Define $\ell(\theta; \xi)$ as the loss function for a data sample ξ for a model with parameters θ . Define $\mathcal{L}(\theta) := \mathbb{E}_{\xi \sim \mathcal{S}}[\ell(\theta; \xi)]$ 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 $\mathcal{L}^* := \min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta)$ be the minimum of training loss. Let $\mathcal{Z}(\theta)$ be the distribution of gradient noise $\nabla \ell(\theta; \xi) - \nabla \mathcal{L}(\theta)$, which is a random variable that depends on θ . We define $\Sigma(\theta) := \mathbb{E}_{z \sim \mathcal{Z}(\theta)}[zz^\top]$ as the noise covariance matrix of gradients at θ .

SGD and Adam. SGD is an iterative method that starts from an initial point θ_0 and updates the parameters as $\theta_{k+1} := \theta_k - \eta \nabla \ell_k(\theta_k)$ for all $k \geq 0$, where η is the learning rate, $\ell_k(\theta)$ 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:

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

Note that in practice, it is common to normalize m_{k+1} and 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 $d\theta_t = b(\theta_t)dt + \sigma(\theta_t)dW_t$ where $b : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the drift vector field, $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times m}$ is the diffusion matrix, and $\{W_t\}_{t \geq 0}$ is an m -dimensional Wiener process. A line of works (Li et al., 2015; Jastrzbski 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*:

$$d\theta_t = -\nabla \mathcal{L}(\theta_t)dt + \sqrt{\eta} \Sigma^{1/2}(\theta_t) dW_t,$$

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

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 Φ, P_ζ for points and differential forms respectively. Consider the gradient flow $\frac{d\mathbf{x}(t)}{dt} = -\nabla \mathcal{L}(\mathbf{x}(t))$ with $\mathbf{x}(0) = \mathbf{x}$, and fix some point $\theta_{\text{null}} \notin \Gamma$, we define the gradient flow projection of any \mathbf{x} , $\Phi(\mathbf{x})$, as $\lim_{t \rightarrow +\infty} \mathbf{x}(t)$ if the limit exists and belongs to Γ , and θ_{null} otherwise. It can be shown by simple calculus (Li et al., 2021b) that $\partial\Phi(\zeta)$ equals the projection matrix onto the tangent space of Γ at ζ . We decompose the noise covariance $\Sigma(\zeta)$ for $\zeta \in \Gamma$ into two parts: the noise in the tangent space $\Sigma_{\parallel}(\zeta) := \partial\Phi(\zeta)\Sigma(\zeta)\partial\Phi(\zeta)$ and the noise in the rest $\Sigma_{\diamond}(\zeta) := \Sigma(\zeta) - \Sigma_{\parallel}(\zeta)$.

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

Definition 3.1 (Slow SDE for SGD). Given $\eta > 0$ and $\zeta_0 \in \Gamma$, define $\zeta(t)$ as the solution of the following SDE with initial condition $\zeta(0) = \zeta_0$:

$$d\zeta(t) = \underbrace{P_\zeta\left(\Sigma_{\parallel}^{1/2}(\zeta)d\mathbf{W}_t\right)}_{(a) \text{ diffusion}} - \underbrace{\frac{1}{2}\nabla^3\mathcal{L}(\zeta)[\widehat{\Sigma}_{\diamond}(\zeta)]dt}_{(b) \text{ drift}}. \quad (1)$$

Here $\widehat{\Sigma}_{\diamond}(\zeta)$ is defined as $\sum_{i,j: \lambda_i \neq 0 \vee \lambda_j \neq 0} \frac{1}{\lambda_i + \lambda_j} \langle \Sigma_{\diamond}(\zeta), \mathbf{v}_i \mathbf{v}_j^\top \rangle \mathbf{v}_i \mathbf{v}_j^\top$, where $\{\mathbf{v}_i\}_{i=1}^d$ 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 $P_\zeta(\Sigma_{\parallel}^{1/2}(\zeta)d\mathbf{W}_t)$ injecting noise in the tangent space, and a *drift* term $-\frac{1}{2}P_\zeta(\nabla^3\mathcal{L}(\zeta)[\widehat{\Sigma}_{\diamond}(\zeta)]dt)$ that can be seen as the negative *semi-gradient* of the following sharpness measure:

$$\mu(\zeta) := \left\langle \nabla^2 \mathcal{L}(\zeta), \widehat{\Sigma}_{\diamond}(\zeta) \right\rangle.$$

Here we use the word “semi-gradient” (Mnih et al., 2015; Brandfonbrener and Bruna, 2019) because it is not exactly the gradient of $\mu(\zeta)$ but only the gradient with respect to the first argument of the inner product. More specifically, define $\mu(\zeta_1, \zeta_2) := \left\langle \nabla^2 \mathcal{L}(\zeta_1), \widehat{\Sigma}_{\diamond}(\zeta_2) \right\rangle$, then the drift term is essentially $-\frac{1}{2} \nabla_{\zeta_1} \mu(\zeta_1, \zeta_2)|_{\zeta_1=\zeta, \zeta_2=\zeta}$ after projecting onto the tangent space of Γ at ζ . In other words, SGD near manifold takes semi-gradients to minimize the implicit regularizer $\langle \nabla^2 \mathcal{L}(\zeta), \widehat{\Sigma}_{\diamond}(\zeta) \rangle$ but pretend $\widehat{\Sigma}_{\diamond}(\zeta)$ to be fixed, i.e. ignore the dependency of $\widehat{\Sigma}_{\diamond}(\zeta)$ on ζ .

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 $\{-0.5, 0.5\}$

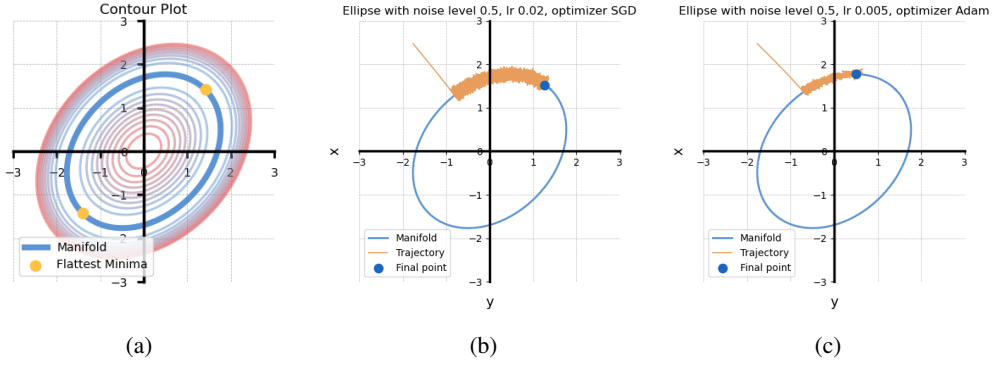


Figure 1: **(a)**: Contour of the elliptical loss, from which we can see the two tips as the flattest minima. **(b)**: SGD implicitly minimizes $\text{tr}(\mathbf{H})$ and converges to the flattest minima. **(c)**: Adam reduces sharpness too but converges to a different and sparser minimum.

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:

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

where $S : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ is ρ_s -smooth, positive definite and satisfies $S(\mathbf{v}) \preceq \epsilon^{-1} I$ for some $\epsilon > 0$ and any $\mathbf{v} \in \mathbb{R}^d$, and $V : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^d$ is linear. A number of currently used optimization algorithms, such as RMSProp, Adam, Adam-mini, Adafactor¹, Adalayer and AdaSGD 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.

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 Φ_S and the SDE projection formula as an extension to the Φ and P_ζ mentioned in Section 3, after which we present the SDE for AGMs we derived.

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

Definition 4.2. For any $\zeta \in \Gamma$ and any differential form $\mathbf{A} d\mathbf{W}_t + \mathbf{b} dt$ in Itô calculus, where $\mathbf{A} \in \mathbb{R}^{d \times d}$, and $\mathbf{b} \in \mathbb{R}^d$. We use $P_{\zeta, S}(\mathbf{A} d\mathbf{W}_t + \mathbf{b} dt)$ as a shorthand for the differential form $\partial \Phi_S(\zeta) \mathbf{A} d\mathbf{W}_t + (\partial \Phi_S(\zeta) \mathbf{b} + \frac{1}{2} \partial^2 \Phi_S(\zeta) [\mathbf{A} \mathbf{A}^T]) dt$.

Definition 4.3 (Slow SDE for AGMs). given learning rate η , $\frac{1-\beta_2}{\eta^2} = c$, $\mathbf{v}_0 \in \mathbb{R}^d$, $S_t := S(\mathbf{v}(t))$, and $\zeta_0 \in \Gamma$, $\mathbf{v}_0 \in \mathbb{R}^d$, we define $\zeta(t)$ as the solution of the following SDE with initial point

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

Table 1: Examples of V, S functions for some optimizers in the AGM Framework.

Optimizer	Function V	Function S	Remarks
Adam	$V(\mathbf{M}) = \text{diag}(\mathbf{M})$	$S(\mathbf{v}) = \text{Diag}(1/(\sqrt{\mathbf{v}} + \epsilon))$	
Adam-mini	$V(\mathbf{M})_i = \frac{1}{ B(i) } \sum_{j \in B(i)} M_{jj}$	$S(\mathbf{v}) = \text{Diag}(1/(\sqrt{\mathbf{v}} + \epsilon))$	Parameters partitioned; i belongs to block $B(i)$. i belongs to layer $L(i)$ in the model.
Adalayer	$V(\mathbf{M})_i = \frac{1}{ L(i) } \sum_{j \in L(i)} M_{jj}$	$S(\mathbf{v}) = \text{Diag}(1/(\sqrt{\mathbf{v}} + \epsilon))$	
AdamE- λ	$V(\mathbf{M}) = \text{diag}(\mathbf{M})$	$S(\mathbf{v}) = \text{Diag}(1/(\mathbf{v}^{\odot \lambda} + \epsilon))$	

210 $(\zeta(0), \mathbf{v}(0)) = (\zeta_0, \mathbf{v}_0)$:

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

211 $\Sigma_{\diamond}(\zeta; S) = S \Sigma(\zeta) S - \Sigma_{\parallel}(\zeta; S)$, $\Sigma_{\parallel}(\zeta; S) = \partial \Phi_S(\zeta) S \Sigma(\zeta) S \partial \Phi_S(\zeta)$.

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

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

214 as if $\Sigma_{\diamond}(\zeta(t); S(t))$ has no dependence on ζ ; also this gradient flow is preconditioned by a positive
215 definite matrix $S(t)$. Recall that the drift term in the slow SDE for SGD can be seen as a semi-
216 gradient descent. In the AGM framework, it takes $\Theta(\eta^{-2})$ time for the preconditioner $S(t)$ to make
217 a significant (i.e. $\Theta(1)$) change, which coincides with the moving speed of the slow SDE of ζ .
218 Therefore, compared to that of SGD, our SDE includes a new formula that tracks the motion of the
219 preconditioner and injects adaptiveness accordingly in the semi-gradient descent process.

220 We could prove that $\zeta(t)$ always stays on the manifold Γ . And next, we present our main theorem,
221 and show that the above SDE in Equation (2) track the trajectory of Adam in a weak approximation
222 sense.

223 **Assumption 4.1.** The loss function $\mathcal{L}(\cdot)$ and the matrix square root of the noise covariance $\Sigma^{1/2}(\cdot)$
224 are C^∞ -smooth. Besides, we assume that $\|\nabla \ell(\theta; \xi)\|_2$ is bounded by a constant for all θ and ξ .

225 **Assumption 4.2.** Γ is a compact manifold.

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

$$\zeta(0) = \Phi(\theta_0) \in \Gamma, \quad \mathbf{v}(0) = \mathbf{v}_0 \in \mathbb{R}^d,$$

228 and we define that the parameters of Adam as $\bar{\mathbf{X}}_t := (\Phi_{S_t}(\theta_t), \mathbf{v}_t)$. For any C^3 -smooth function
229 $g(\theta)$,

$$\max_{0 \leq t \leq \frac{T}{\eta^2}} \left| \mathbb{E}[g(\bar{\mathbf{X}}_t)] - \mathbb{E}[g(\mathbf{X}(t\eta^2))] \right| = \tilde{O}(\eta^{0.25}),$$

230 where $\tilde{O}(\cdot)$ hides logarithmic factors and constants that are independent of η but may depend on
231 $g(\theta)$.

232 Theorem 4.1 shows that, in the small η regime, once Adam approaches the minimizer manifold,
233 its long-horizon behavior within $\tilde{O}(\frac{1}{\eta^2})$ steps can be well approximated by the SDE defined in
234 Equation (2).

235 4.2 Interpretation of The Slow SDEs for AGMs

236 **Adaptive Projection Operator.** Whereas Equation (1) employs a fixed projection operator P_{ζ} to
237 constrain the SDE to the manifold, the AGM slow-SDE uses an adaptive projection $P_{\zeta, S(t)}$ that

depends on the current preconditioner $S(v(t))$. In other words, SGD’s projection is static and 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. It is well known that, near the manifold, SGD’s wandering around is 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). However, the AGM trajectory is influenced by its preconditioner. Concretely, the gradient-noise covariance matrix Σ is filtered through the preconditioner $S(t)$ into $S(t)\Sigma S(t)$ 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.

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. But before we can study this, in order to make our study meaningful, we first need to show that Adam can converge to the neighborhood of the minimizer manifold, which itself is already non-trivial. Unfortunately, Adam can not provably converge to the minimizer manifold without any constraint. In fact, the convergence issue of Adam has been debated from its birth, Reddi et al. (2018) shows that Adam does not converge to the optimal solution even in some simple convex setting. Recent work (Dereich and Jentzen, 2024) gives Adam’s ODE and shows that this ODE does not necessarily converge to the immobile point of the gradient flow. So We present a statement of convergence first.

Theorem 4.2 (Convergence Bound of the AGM Framework, Stated Informally). *Let $K = \mathcal{O}\left(\frac{1}{\eta} \log \frac{1}{\eta}\right)$. Under mild assumptions, for any $k < K$, $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that $\mathcal{L}(\theta_k) - \mathcal{L}^* = \tilde{\mathcal{O}}(\eta)$.*

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 launch 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, but unlike the usual SDE approximation, the SDEs we use here can track the AGMs for a much longer period of time, up to $\tilde{\mathcal{O}}(\frac{1}{\eta^2})$ rather than the $\tilde{\mathcal{O}}(\frac{1}{\eta})$ that was 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 rotation 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 $S(v_k)$, 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. Utilizing this definition, after doing a reparameterization to the original space, we can regain the calculation simplicity in previous works (Gu et al., 2023a; Li et al., 2021b).

The second reason is that when β_2 is too far from 1, the preconditioner moves too fast, making it very hard to characterize the change of moments. In contrast, when β_2 is too close to 1, then the change of precondition is almost negligible, which is also not practical. To this end, we consider the case where $1 - \beta_2 = \mathcal{O}(\eta^2)$. And we term it “2-scheme”. The subtlety here is that this proximity does not make the change in the preconditioner negligible; rather, the change in the preconditioner affects the form of the SDE, and because the change in the preconditioner is slow enough that we can track its change.

5 Adam’s Provable Generalization Benefit with Label Noise

In this section, we will prove that under label noise setting, 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

On an ℓ_2 -regression task on dataset $\{z_i, y_i\}_{i=1}^n$, adding *label noise* means adding a noise sampled i.i.d. from $\{\pm\delta\}$ to any true label y before feeding forward to the network. A crucial property of the label noise setting is that when $\theta \in \Gamma$, $\Sigma \equiv \alpha \nabla^2 \mathcal{L}$ for some constant α (Blanc et al., 2020), which simplifies the setting and has been largely used (Blanc et al., 2020; Damian et al., 2021; Li et al., 2021b; Gu et al., 2023a) to analyze the implicit bias of SGD and other optimizers.

Theorem 5.1 (Adam’s Implicit Bias with Label Noise, Stated Informally). *Adam’s SDE becomes an ODE under the label noise setting, and when ϵ is small, the fixed point of this ODE must satisfy $\nabla \text{tr}(\text{Diag}(\mathbf{H})^{1/2}) = 0$.*

Proof Sketch. Under label noise setting, SDE Eq. (2) will be greatly simplified. In fact, the diffusion term of the slow SDE would equal zero, and the drift term could be simplified to

$$\begin{cases} d\mathbf{v}(t) = c(V(\Sigma(\zeta)) - \mathbf{v}) dt, \\ d\zeta(t) = -\frac{\alpha}{2} S(\mathbf{v}) \nabla^3 \mathcal{L}(\zeta) [S(\mathbf{v})] dt, \end{cases} \quad (3)$$

the proof of which is deferred to Appendix. With our SDE becoming an ODE, we consider the fixed point of this ODE, which should satisfy $\mathbf{v} = V(\Sigma(\zeta))$ and $\nabla^3 \mathcal{L}(\zeta) [S(\mathbf{v})] = 0$ since $S(\mathbf{v})$ is invertible. Denote $\mathbf{H} = \nabla^2 \mathcal{L}(\zeta) = \Sigma(\zeta)/\alpha$. In the case of Adam, $\mathbf{v} = \text{diag}(\Sigma) = \alpha \cdot \text{diag}(\mathbf{H})$, and $S(\mathbf{v}) = \text{Diag}(1/(\sqrt{\mathbf{v}} + \epsilon))$. Then we integrate by parts and obtain $\nabla^3 \mathcal{L}(\zeta) [S(\mathbf{v})] = \nabla[(\mathbf{H}, S(\mathbf{v}))] - \nabla(S(\mathbf{v}))[\mathbf{H}]$. A straightforward simplification gives the result. \square

The proof of this theorem also inspires us of a simple way to directly adjust the implicit bias of Adam. Specifically, for any $\lambda \in [0, 1]$, we define *AdamE- λ* as an optimizer identical with Adam, except that $S(\mathbf{v}) = \text{Diag}(1/(\mathbf{v}^{\odot \lambda} + \epsilon))$. Obviously AdamE- $\frac{1}{2}$ reduces to Adam and all AdamE- λ ’s belong to the AGM framework. To compute the implicit bias of AdamE- λ with label noise, we can apply the same method as in Theorem 5.1, and the result is stated below.

Theorem 5.2 (AdamE- λ ’s Implicit Bias with Label Noise, Stated Informally). *For $\lambda \in [0, 1]$, AdamE- λ ’s SDE becomes an ODE under the label noise setting, and when ϵ is small, the fixed point of this ODE must satisfy $\nabla \text{tr}(\text{Diag}(\mathbf{H})^{1-\lambda}) = 0$.*

Theorem 5.2 indicates that tuning the exponent of the second-order moment in adam will exactly result in tuning the exponent of $\text{diag}(\nabla^2 \mathcal{L}(\zeta))$ 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 will relate the implicit bias with sparsity and compare the performance of Adam, AdamE and SGD in a simple experimental setup.

5.2 Problem Setup: the 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 $\mathbf{w}^* \in \mathbb{R}^d$ be an unknown κ -sparse ground truth vector. Let $\{(z_i, y_i)\}_{i \in [n]}$ be the training dataset where each $z_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\{\pm 1\}^d$, and each y_i is generated by $\langle z_i, \mathbf{w}^* \rangle$. Our parameter is defined as $\theta = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \in \mathbb{R}^{2d}$. For any function g defined on \mathbb{R}^{2d} , we write $g(\theta)$ and $g(\mathbf{u}, \mathbf{v})$ exchangeably. The loss function is defined as:

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

where a label noise is added to the true label y during training. This setting can be viewed as using estimation $\hat{\mathbf{w}} = \mathbf{u}^{\odot 2} - \mathbf{v}^{\odot 2}$ to approximate the ground truth vector \mathbf{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. $\Gamma = \{\theta | \langle z_i, \mathbf{u}^{\odot 2} - \mathbf{v}^{\odot 2} \rangle = y_i, \forall i \in [n]\}$.

331 This setting allows us to relate the implicit bias directly to the sparsity of the output. It's straightforward to verify that $\nabla^2 \mathcal{L}(\theta) = \frac{4}{n} \sum_{i=1}^n \begin{pmatrix} \mathbf{z}_i \odot \mathbf{u} \\ -\mathbf{z}_i \odot \mathbf{v} \end{pmatrix} \begin{pmatrix} \mathbf{z}_i \odot \mathbf{u} \\ -\mathbf{z}_i \odot \mathbf{v} \end{pmatrix}^\top$ when $\theta \in \Gamma$, so $\text{diag}(\nabla^2 \mathcal{L}(\theta)) = 4\theta^{\odot 2}$.
 333 Then note the following property:

334 **Lemma 5.1.** *Let some optimum θ satisfy that $\theta \in \arg \min_{\theta' \in \Gamma} \text{tr}(\text{Diag}(\mathbf{H})^{1/2})$, then we also have $\theta \in \arg \min_{\theta' \in \Gamma} \|\mathbf{u}^{\odot 2} - \mathbf{v}^{\odot 2}\|_{1/2}$. Similarly, for any $e_0 \in (0, 1]$, if $\theta \in \arg \min_{\theta' \in \Gamma} \text{tr}(\text{Diag}(\mathbf{H})^{e_0})$, then $\theta \in \arg \min_{\theta' \in \Gamma} \|\mathbf{u}^{\odot 2} - \mathbf{v}^{\odot 2}\|_{e_0}$.*

337 This is because only $\hat{\mathbf{w}} = \mathbf{u}^{\odot 2} - \mathbf{v}^{\odot 2}$ matters in the evaluation of train loss, so if $u_i \neq 0$ and $v_i \neq 0$ for some i , then we can decrease the absolute value of both u_i and v_i while keeping $u_i^2 - v_i^2$ unchanged, and we will get another optimum with smaller $\text{tr}(\text{Diag}(\mathbf{H})^{e_0})$. Thus $u_i = 0$ or $v_i = 0$ for any i . When this holds, we have $\text{tr}(\text{Diag}(\mathbf{H})^{e_0}) = 4^{e_0} \|\theta^{\odot 2e_0}\|_1 = (4 \|\hat{\mathbf{w}}\|_{e_0})^{e_0}$. Therefore, implicitly regularizing $\text{tr}(\text{Diag}(\mathbf{H})^{e_0})$ can be viewed as regularizing the ℓ_{e_0} norm of the output: Adam regularizes $\ell_{0.5}$, SGD regularizes ℓ_1 , and AdamE- λ regularizes $\ell_{1-\lambda}$ norm of the output. Just as lasso (ℓ_1) regression's advantage over ridge (ℓ_2) regression in sparse ground truth recovery, we argue that Adam and AdamE with large λ 's will recover ground truth more efficiently than SGD, and AdamE with small λ 's on this task.

346 5.3 Result: Adam's Implicit Regularizer Facilitates Sparse Ground-truth Recovery

347 We plot the results of the experiment in Fig. 2. We gradually increase the number of training points, and train Adam, SGD and AdamE with different configurations until convergence. We identify a training configuration as 'recovered the groundtruth' when the test loss ends up below 1. As depicted in Fig. 2a, Adam's test loss plunges towards zero around $n_{\text{train}} = 420$, while SGD's test loss decreases gradually with the increase of training data. As an attempt to interpolate between different implicit biases, we also train AdamE with different λ 's. Fig. 2b shows that AdamE-0.001's performance is similar to that of SGD, and all AdamE with larger λ 's exhibit the same sudden recovery behavior as Adam.

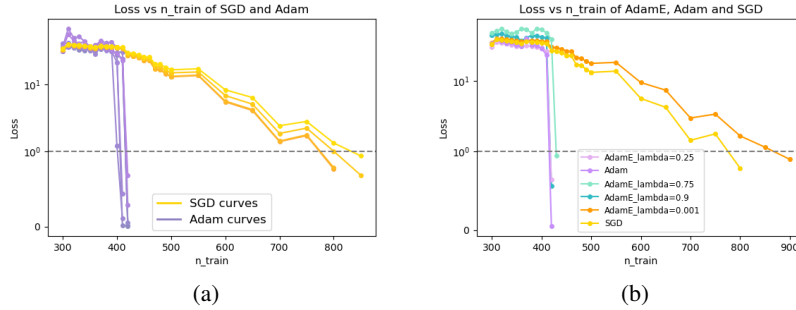


Figure 2: The curve of final test loss vs. scale of training data with $d = 10000, \kappa = 50$. (a): Loss comparison between SGD with different learning rates, and Adam with different learning rates and β_2 's. (b): Loss comparison between AdamE with $\lambda = 0.001, 0.25, 0.75, 0.9$, Adam and SGD.

354

355 **Takeaway.** In the diagonal net setting, 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.

361 6 Discussion

362 We show that Adam implicitly minimizes the sharpness measure $\text{tr}(\text{Diag}(\mathbf{H})^{1/2})$, 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.

References

- Ashish, V. (2017). Attention is all you need. *Advances in neural information processing systems*, 30:1.
- 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., Podoprikin, 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

488 Saunshi, N., Karp, S., Krishnan, S., Miryoosefi, S., Reddi, S. J., and Kumar, S. (2024). On the inductive bias
489 of stacking towards improving reasoning. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U.,
490 Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages
491 71437–71464. Curran Associates, Inc.

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

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

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

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

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

503 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.
504 (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

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

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

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

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

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

518 Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N.
519 (2020). Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages
520 3635–3673. PMLR.

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

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

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

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

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

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

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

537 Zou, F., Shen, L., Jie, Z., Zhang, W., and Liu, W. (2019). A sufficient condition for convergences of adam and
538 rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages
539 11127–11135.

540	Contents	
541	1 Introduction	1
542	2 Related Work	2
543	3 Preliminaries	3
544	4 Theoretical Analysis of Adam	5
545	4.1 Slow SDE Analysis for AGMs	5
546	4.2 Interpretation of The Slow SDEs for AGMs	6
547	4.3 Technical Difficulties and Proof Insights	7
548	4.3.1 Convergence Guarantee of AGMs	7
549	4.3.2 Key Insights in the Derivation of Slow SDEs for AGMs	7
550	5 Adam’s Provable Generalization Benefit with Label Noise	8
551	5.1 Reduction of Adam’s Implicit Regularizer with Label Noise	8
552	5.2 Problem Setup: the Diagonal Net	8
553	5.3 Result: Adam’s Implicit Regularizer Facilitates Sparse Ground-truth Recovery . . .	9
554	6 Discussion	9
555	A Illustration of the Difference between Conventional SDE and Slow SDE	16
556	B Matrix Factorization: Adam Implicitly Regularizes Sharpness Differently	16
557	B.1 Problem setup	16
558	B.2 Results	17
559	C Formal Statements of the Main Results	17
560	C.1 Convergence Guarantee of AGMs	19
561	D Constructing the Working Zones	19
562	E Proof of the Convergence of AGMs	21
563	F Proof of the SDE Approximation of AGMs	26
564	F.1 Lemmas for Adaptive Manifold Projection	27
565	F.2 Iteration Stays Near Manifold	28
566	F.3 Moment Calculation of AGMs Near Manifold	30
567	F.3.1 Moment Calculation Within a Giant Step	32
568	F.4 Weak Approximation	37
569	F.5 Preliminary and Additional Notations	38
570	F.6 Proof of the Approximation for Slow SDE of AGMs	39
571	G Proof of Theorems in Section 5	42

572	G.1 Proof of Adam and AdamE- λ 's Implicit Biases with Label Noise	42
573	G.2 Proof of Lemma 5.1	43

574 A Illustration of the Difference between Conventional SDE and Slow SDE

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

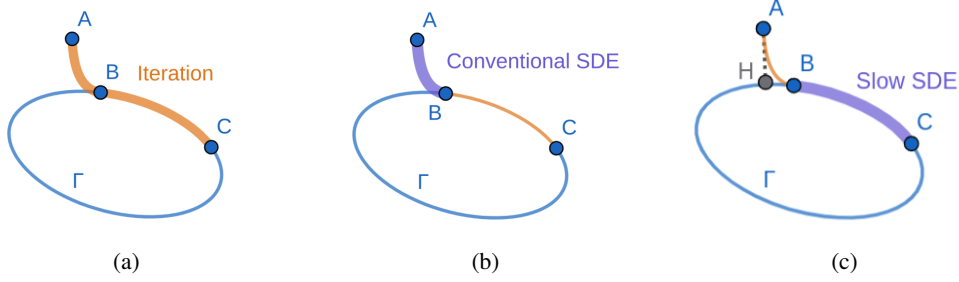


Figure 3: Comparison of conventional SDE and slow SDE.

578

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

584 The conventional SDE approximates the convergence phase only, whose unit time corresponds to
 585 $\tilde{O}(\eta^{-1})$ steps (Fig. 3b). In contrast, slow SDE manages to separate the slow implicit regularization
 586 motion from the fast convergence, and approximate the implicit regularization near manifold only
 587 (Fig. 3c).

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

593 B Matrix Factorization: Adam Implicitly Regularizes Sharpness Differently

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

602 B.1 Problem setup

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

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

607 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) .

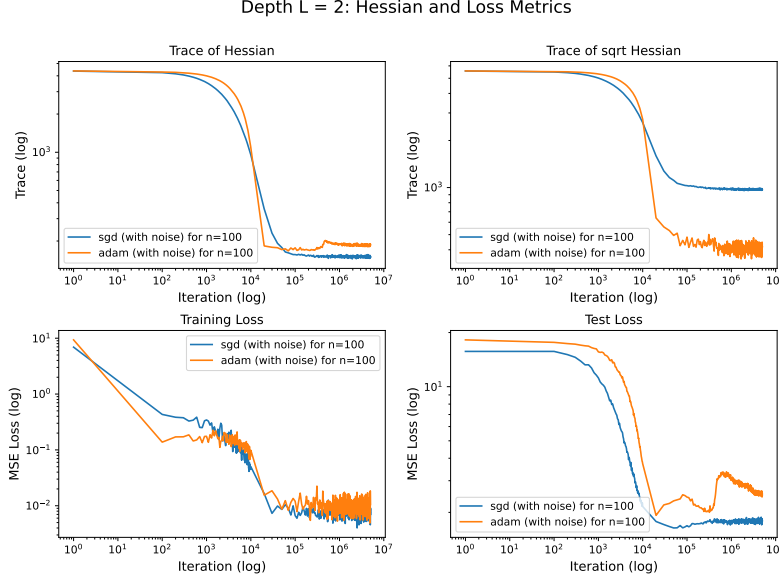


Figure 4: **Deep matrix factorization with label noise.** Adam and SGD are trained on identical data and noise realizations. *Top*: evolution of $\text{tr}(\mathbf{H})$ and $\text{tr}(\text{Diag}(\mathbf{H})^{1/2})$. *Bottom*: training and test MSE. Adam converges to a point with larger overall curvature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

Implicit regularization. With small learning rates and additive label noise, SGD asymptotically minimizes $\text{tr}(\mathbf{H})$ once it reaches the zero-loss manifold. In matrix factorization, minimizing $\text{tr}(\mathbf{H})$ is nearly equivalent to minimizing the nuclear norm of the recovered matrix (Gatmiry et al., 2023), which promotes low rank and hence better generalization when \mathbf{M}^* is low rank. Adam, however, implicitly minimizes $\text{tr}(\text{Diag}(\mathbf{H})^{1/2})$; it therefore converges to a different point, typically with larger $\text{tr}(\mathbf{H})$ and reduced generalization.

B.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 $\text{tr}(\text{Diag}(\mathbf{H})^{1/2})$ sharply downward while $\text{tr}(\mathbf{H})$ 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.

C 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 principal 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 θ_k to θ_{k+1} is defined as:

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

where $S : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ is ρ_s -smooth, positive definite (p.d) and satisfies $S(\mathbf{v}) \preceq \epsilon^{-1} I$ for some $\epsilon > 0$ and any $\mathbf{v} \in \mathbb{R}^d$, and $V : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^d$ is linear. Before formalizing the results, we introduce some technical assumptions first.

Assumption C.1. \mathcal{L} is C^5 -smooth on \mathbb{R}^d , i.e. all partial derivatives of \mathcal{L} up to order 5 exist and are continuous.

Assumption C.2. \mathcal{L} is ρ -smooth on \mathbb{R}^d , i.e. $\forall \theta_1, \theta_2 \in \mathbb{R}^d$, $\|\nabla \mathcal{L}(\theta_1) - \nabla \mathcal{L}(\theta_2)\|_2 \leq \rho \|\theta_1 - \theta_2\|_2$ and \mathcal{L} is bounded from below, i.e. $\mathcal{L}^* = \inf_{\theta} \mathcal{L}(\theta) > -\infty$.

Assumption C.3. The noisy gradients are L2-bounded, i.e., there exists some constant R s.t. $\forall \theta \in \mathbb{R}^d$, $\|\nabla \ell(\theta; \xi)\|_2 \leq R$ almost surely for $\xi \sim \mathcal{S}$.

Assumption C.4. For any $\mathbf{g} \in \mathbb{R}^d$, all entries of $V(\mathbf{g}\mathbf{g}^\top)$ are non-negative.

Assumption C.5. The function S is C^4 -smooth on $\{\mathbf{v} \in \mathbb{R}^d : \mathbf{v} \geq 0\}$, i.e. the subspace where all entries of \mathbf{v} are non-negative.

Assumption C.6. $\beta_1 \leq 0.9$.

Remark. The threshold 0.9 in Assumption C.6 can also be replaced by any constant below 1, and the approximation rate in our result will remain unaffected. Note that β_1 is usually no more than 0.9 in real-world areas such as NLP (Devlin et al., 2019; Radford et al., 2018; Vaswani et al., 2017) or CV (Isola et al., 2017; Dosovitskiy et al., 2020; Ronneberger et al., 2015), so our assumption aligns with common practice.

Theorem C.1. Let Assumptions C.2, C.3 and C.6 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 $\epsilon > 0$, independent of η , such that for any initial point θ_0 whose L2 distance from Γ^ϵ does not exceed ϵ , and any $\delta \in (\eta^{200}, 1)$,¹ with probability at least $1 - \delta$, the following holds for some $K_0 = \mathcal{O}(\frac{1}{\eta} \log \frac{1}{\eta})$:

$$\begin{aligned}\mathcal{L}(\theta_{K_0}) - \mathcal{L}^* &= \mathcal{O}\left(\eta \log \frac{1}{\eta \delta}\right), \\ \|\theta_{K_0} - \Phi_{\mathcal{S}_{K_0}}(\theta_{K_0})\|_2 &= \mathcal{O}\left(\sqrt{\eta \log \frac{1}{\eta \delta}}\right).\end{aligned}$$

2. (Formal restatement of Theorem 4.1: Slow SDE tracks AGM's trajectory in a weak approximation sense) Moreover, when Assumptions 3.1, 4.2, C.1, C.4 and C.5 hold, we shift the timeline and redefine the final state $(\theta_{K_0}, \mathbf{v}_{K_0})$ in conclusion 1 by (θ_0, \mathbf{v}_0) . Let $T > 0$ be a constant, $\mathbf{X}(t) = (\zeta(t), \mathbf{v}(t))$ be the solution to Equation (2) with initial condition:

$$\zeta(0) = \Phi(\theta_0) \in \Gamma, \quad \mathbf{v}(0) = \mathbf{v}_0 \in \mathbb{R}^d,$$

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

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

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

¹The exponent here, along with the exponents related to the δ -goodness in section F.2, can be arbitrary large constant, which does not affect the order of following derivations.

666 C.1 Convergence Guarantee of AGMs

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

Definition C.1 (Polyak-Łojasiewicz Condition). *For some $\mu > 0$, we say some function $\mathcal{L} : \mathbb{R}^d \rightarrow \mathbb{R}$ is μ -Polyak-Łojasiewicz condition (abbreviated as μ -PL), if and only if $\forall \boldsymbol{\theta} \in \mathbb{R}^d$:*

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

671 **Theorem C.2** (Formal restatement of Theorem 4.2). *Let Assumptions C.2, C.3 and C.6 be satisfied,*
 672 *and \mathcal{L} satisfy the μ -PL condition. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following*
 673 *holds for some $K = \mathcal{O}(\frac{1}{\eta} \log \frac{1}{\eta})$:*

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}_K) - \mathcal{L}^* &= \mathcal{O}\left(\eta \log \frac{1}{\eta\delta}\right), \\ \|\boldsymbol{\theta}_K - \Phi_{\mathcal{S}_K}(\boldsymbol{\theta}_K)\|_2 &= \mathcal{O}\left(\sqrt{\eta \log \frac{1}{\eta\delta}}\right). \end{aligned}$$

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

683 D Constructing the Working Zones

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

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

$$\exists \boldsymbol{\zeta} \in \Gamma, \quad \|\boldsymbol{\theta} - \boldsymbol{\zeta}\|_2 \leq \epsilon.$$

692 **Lemma D.1.** *Assume that $C_1 < C_2$ are two positive constants, and \mathcal{L} be a function that satisfies both*
 693 *ρ -smoothness and μ -PL. For any matrix \mathbf{S} satisfying $C_1 \mathbf{I} \preceq \mathbf{S} \preceq C_2 \mathbf{I}$, consider the preconditioned*
 694 *gradient flow $\frac{d\boldsymbol{\theta}(t)}{dt} = -\mathbf{S}\nabla \mathcal{L}(\boldsymbol{\theta}(t))$ starting at $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$. For any $T > 0$, we have $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}(T)\|_2 \leq$*
 695 *$\frac{2C_2}{\sqrt{2\mu C_1}} \sqrt{\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*}$.*

696 *Proof.* Since $C_1 \mathbf{I} \preceq \mathbf{S} \preceq C_2 \mathbf{I}$, we have $\|\mathbf{S}\nabla \mathcal{L}(\boldsymbol{\theta})\|_2 \leq C_2 \|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2$ and $\langle \nabla \mathcal{L}(\boldsymbol{\theta}), \mathbf{S}\nabla \mathcal{L}(\boldsymbol{\theta}) \rangle \geq$
 697 $C_1 \|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2$ for any $\boldsymbol{\theta}$, which implies

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

698 Then for any $t < T$ we have

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

699 Integrating both sides gives us

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

700 The above equations complete the proof. \square

701 Since the gradients and gradient noises are assumed to be bounded and S and V both satisfy Lipschitz-
702 ness, we deduce that any \mathbf{v} produced in the iteration is bounded. Specifically, there exists some
703 constant R_1 such that $\mathbf{v}_k \leq R_1$ for any $k \geq 0$. For all algorithms listed in Table 1, setting $R_1 = R^2$ is
704 sufficient. Similarly, all the outputs of S are also bounded, and we denote R_0 as a constant satisfying
705 $S^{-1}(\mathbf{v}) \preceq R_0$ for all \mathbf{v} in the iteration. Note that S is Lipschitz on $\{\mathbf{v} : 0 \leq \mathbf{v} \leq R_1\}$ from this
706 assumption, since it is a compact set, and the derivative of S is bounded.

707 Denote the minimal distance of Γ and any other local minimizer manifold as ϵ_4 . We construct nested
708 working zones $(\Gamma^{\epsilon_1}, \Gamma^{\epsilon_2}, \Gamma^{\epsilon_3})$ in the following way:

709 **Lemma D.2 (Working Zone Lemma).** *There exist positive constants $\epsilon_1, \epsilon_2, \epsilon_3$ such that $\epsilon_1 < \epsilon_2 <$
710 $\epsilon_3 < \epsilon_4$ and $\Gamma^{\epsilon_1}, \Gamma^{\epsilon_2}, \Gamma^{\epsilon_3}$ satisfy the following properties:*

- 711 1. \mathcal{L} is μ -PL in Γ^{ϵ_3} for some $\mu > 0$.
- 712 2. For any matrix $\mathbf{S} \in \mathbb{R}^{d \times d}$ such that $\frac{1}{R_0} \mathbf{I} \preceq \mathbf{S} \preceq \frac{1}{\epsilon} \mathbf{I}$, any gradient flow preconditioned by
713 \mathbf{S} and starting from Γ^{ϵ_2} will converge to some point in Γ .
- 714 3. For any $\epsilon > 0$, define \mathcal{X}^ϵ as the subset

$$\mathcal{X}^\epsilon := \{(\mathbf{v}, \boldsymbol{\theta}) : 0 \leq \mathbf{v} \leq R_1, \boldsymbol{\theta} \in \Gamma^\epsilon\}.$$

714 View $\Phi_{S(\mathbf{v})}(\boldsymbol{\theta})$ as a function defined on support \mathcal{X}^{ϵ_2} . If Assumptions C.1 and C.5 hold, then
715 $\Phi_{S(\mathbf{v})}(\boldsymbol{\theta})$ is \mathcal{C}^4 on \mathcal{X}^{ϵ_1} .

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

718 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
719 starting point $\boldsymbol{\theta}_0 \in \Gamma^{\epsilon_2}$, and any preconditioning matrix \mathbf{S} satisfying $C_1 \mathbf{I} \preceq \mathbf{S} \preceq C_2 \mathbf{I}$, assume on
720 the contrary that the preconditioned gradient flow starting from $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$ will leave Γ^{ϵ_3} at some
721 finite time. Then let $T = \inf \{t : \boldsymbol{\theta}(t) \notin \Gamma^{\epsilon_3}\} < \infty$. Using Lemma D.1 and combining the μ -PL
722 condition, we conclude that $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}(T)\|_2 \leq \frac{2C_2}{\sqrt{2\mu}C_1} \sqrt{\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*} \leq \frac{2C_2}{\sqrt{2\mu}C_1} \cdot \sqrt{\frac{\mu}{2}} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2 =$
723 $\sqrt{\frac{\rho}{\mu}} \cdot \frac{C_2}{C_1} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2$ for any $\boldsymbol{\theta}^* \in \Gamma$. Hence $\boldsymbol{\theta}(T) \in \Gamma^{\epsilon_3}$, a contradiction.

Next we begin the construction of Γ^{ϵ_1} with Assumptions C.1 and C.5. Define a function $f(\mathbf{v}, \boldsymbol{\theta}) :$
 $\mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$ as

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

724 then f is \mathcal{C}^4 on $\{(\mathbf{v}, \boldsymbol{\theta}) : 0 \leq \mathbf{v} \leq R_1, \boldsymbol{\theta} \in \mathbb{R}^d\}$. Let \tilde{r} be a constant such that $\tilde{r} > \epsilon_2$. Substituting
725 $f_0 = f$, $r = \sqrt{\tilde{r}^2 + d \cdot R_1^2}$, $x_0 = (\mathbf{v}_0, \boldsymbol{\theta}_0)$ such that each entry of \mathbf{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 $\gamma_\delta(\mathbf{v}, \boldsymbol{\theta})$ defined by:

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

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 hypercube and a ball 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 $\|\nabla^2\mathcal{L}\|_2$ can be uniformly bounded. Take $\epsilon_1 = 0.9\epsilon_2$, then for any $\boldsymbol{\theta} \in \Gamma^{\epsilon_1}$, a small open neighborhood 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}^{ϵ_1} . Finally, we use Theorem 6.4 in Falconer (1983) to conclude that $\Phi_{S(\mathbf{v})}(\boldsymbol{\theta})$ is \mathcal{C}^4 on \mathcal{X}^{ϵ_1} . \square

E Proof of the Convergence of AGMs

In this section, we aim to prove Theorem C.2, and consequently the first part of Theorem C.1. Specifically, for some $\gamma = 1 - \Theta(\eta)$, we will 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 C.1; However, this convergence analysis works for any $K = \mathcal{O}(\text{poly}(1/\eta))$, 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 $\mathbf{m}_0 = \mathbf{v}_0 = \mathbf{0}$. 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 \mathbf{m}_{k+1} , \mathbf{v}_{k+1} and $\boldsymbol{\theta}_{k+1}$. For any $k \geq 0$, to simplify the notation, we denote that

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

Time k refers 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 filtration generated by the history of optimization, where each $\mathcal{F}_k = \sigma(\boldsymbol{\theta}_0, \mathbf{z}_0, \dots, \mathbf{z}_{k-1})$ 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 E.1 throughout Lemma E.5, we will assume that \mathcal{L} satisfies μ -PL condition everywhere, and Theorem C.2 follows directly from the result. After that, we argue that even if the loss function only satisfies μ -PL 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 C.1.

Lemma E.1. Define $\tilde{\mathbf{v}}_k := \beta_2\mathbf{v}_{k-1} + (1 - \beta_2)\mathbb{E}_{k-1}[V(\mathbf{g}_{k-1}\mathbf{g}_{k-1}^\top)]$. There exist constants C_{1a} , C_{1b} such that for any $k \geq 1$,

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

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

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

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

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

Define Y_k as $Y_k = -\langle \nabla \mathcal{L}(\boldsymbol{\theta}_{k-1}), (S(\mathbf{v}_k) - S(\tilde{\mathbf{v}}_k)) \mathbf{g}_{k-1} \rangle$, then it holds almost surely that $|Y_k| \leq \|\nabla \mathcal{L}(\boldsymbol{\theta}_{k-1})\|_2 \| (S(\mathbf{v}_k) - S(\tilde{\mathbf{v}}_k)) \mathbf{g}_{k-1} \|_2$. Since S is Lipschitz, V is linear and

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

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

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

Now we only need to let $X_k = \langle \mathbf{z}_{k-1}, S(\tilde{\mathbf{v}}_k)^\top \nabla \mathcal{L}(\boldsymbol{\theta}_{k-1}) \rangle$. It's easy to see that $\mathbb{E}_{k-1}[X_k] = 0$ and $|X_k| \leq C_{1b} \|\nabla \mathcal{L}(\boldsymbol{\theta}_{k-1})\|_2$ a.s. for some constant C_{1b} , which completes the proof. \square

Lemma E.2 (Descent Lemma of the AGM Framework). *For any $k \geq 1$ it holds that*

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

for some constant C_2 .

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

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

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

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

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

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

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

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

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

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

779 Plugging in, we get

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

780 for some constant C_2 . □

781 **Lemma E.3.** Define $\gamma := 1 - \frac{2\eta\mu(1-\beta_1)}{R_0}$. For any $k \geq 0$, we have

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

782 for some constant C_3 .

783 *Proof.* We start from Lemma E.2 and plug in Lemma E.1:

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

Note that $S(\tilde{\mathbf{v}}_i) \succeq \frac{1}{R_0}$, so $\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})^\top S(\tilde{\mathbf{v}}_i) \nabla \mathcal{L}(\boldsymbol{\theta}_{i-1}) \geq \frac{1}{R_0} \|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\|_2^2$ for any i . Combining with the μ -PL property $\|\nabla \mathcal{L}(\boldsymbol{\theta}_{i-1})\|_2^2 \geq 2\mu(\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^*)$ inside the working zone Γ^{ϵ_3} , we have

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k-1}) \leq C_2 \eta^2 - \frac{2\eta\mu(1-\beta_1)}{R_0} \sum_{i=1}^k \beta_1^{k-i} (\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^*) + \eta(1-\beta_1) \sum_{i=1}^k \beta_1^{k-i} (Y_i + X_i).$$

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

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

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

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

786 for some constant \tilde{C}_3 . Note that we can expand the $\mathcal{L}(\boldsymbol{\theta}_{k-1}) - \mathcal{L}^*$ term iteratively to obtain a generic
787 formula for $\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^*$:

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

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

789 **Lemma E.4.** Let $k \leq K = \mathcal{O}(\text{poly}(1/\eta))$ and $f : (\{0, 1, \dots, k-1\} \times (0, 1)) \rightarrow \mathbb{R}^+$ be a
 790 function. Let $\{X_i\}_{i=1}^k$ be any martingale difference sequence such that:

791 1. X_i is \mathcal{F}_i -measurable and $\mathbb{E}_{i-1}[X_i] = 0$;

792 2. $|X_i| \leq C_{1b} \|\nabla \mathcal{L}(\theta_{i-1})\|_2$ 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

$$\mathcal{L}(\theta_{i-1}) - \mathcal{L}^* \leq f(i, \delta),$$

then $\forall \delta \in (0, 1)$, with probability $1 - \delta$, we have $\mathcal{L}(\theta_{i-1}) - \mathcal{L}^* \leq f(i, \frac{\delta}{2k})$ for all $i \in [k]$, and that

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

793 for some constant C_4 .

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

797 *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
 798 and $\exists C_{1b}$ s.t. every $|X_i|$ is bounded by $C_{1b} \|\nabla \mathcal{L}(\theta_{i-1})\|_2$ (Lemma E.1), we have

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

Since $\mathcal{L}(\theta_{i-1}) - \mathcal{L}^* \leq f(i, \delta')$ 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:

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

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

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

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

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

Let $\delta' = \frac{\delta}{2k}$, by union bound, $\mathcal{L}(\theta_{i-1}) - \mathcal{L}^* \leq f(i, \frac{\delta}{2k})$ 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 probability $1 - \delta$, the following two statements hold
 simultaneously for all $i \in [k]$:

$$\mathcal{L}(\theta_{i-1}) - \mathcal{L}^* \leq f\left(i, \frac{\delta}{2k}\right)$$

802 and

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

803 where $C_4 = 2C_{1b}\sqrt{\rho}$. □

Lemma E.5 (Convergence Bound of the AGM Framework). *Let η be a small learning rate satisfying $\frac{\beta_1}{\gamma} = \beta_1/(1 - \frac{2\eta\mu(1-\beta_1)}{R_0}) \leq 0.95$ and $\tilde{\epsilon}_2 + \eta R/\epsilon < \epsilon_2$. Let $\theta_0 \in \Gamma^{\tilde{\epsilon}_2}$, and $K = \mathcal{O}(\text{poly}(1/\eta))$. Under mild restrictions on K , for any $k \leq K$, $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that*

$$\mathcal{L}(\theta_k) - \mathcal{L}^* \leq (C_{5a}\gamma^k + C_{5b}\eta) \log \frac{K}{\delta}$$

804 for some constants C_{5a} and C_{5b} .

Proof. We denote the bound with $1 - \delta$ probability as $f(k, \delta) := (C_{5a}\gamma^k + C_{5b}\eta) \log \frac{K}{\delta}$, where the constants C_{5a}, C_{5b} will be specified by us later. We prove by induction. When $k = 0$, we need

$$(C_{5a} + C_{5b}\eta) \left(\log K + \log \frac{1}{\delta} \right) \geq \mathcal{L}(\theta_0) - \mathcal{L}^*,$$

805 where setting $C_{5a} = \frac{\mathcal{L}(\theta_0) - \mathcal{L}^*}{\log K}$ suffices. Now assume that the statement holds for $0, 1, \dots, k-1$.

806 From Lemma E.3, we have

$$\mathcal{L}(\theta_k) - \mathcal{L}^* \leq \gamma^k (\mathcal{L}(\theta_0) - \mathcal{L}^*) + \eta(1 - \beta_1) \sum_{i=1}^k X_i \sum_{j=i}^k \gamma^{k-j} \beta_1^{j-i} + C_3\eta.$$

807 We can bound the coefficients by

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

808 where the last inequality is due to our assumption $\frac{\beta_1}{\gamma} \leq 0.95$. Let $\tilde{X}_i := \frac{\sum_{j=i}^k \gamma^{k-j} \beta_1^{j-i}}{20\gamma^{k-i}} X_i$,

809 then $\{\tilde{X}_i\}_{i=1}^k$ is also a martingale difference sequence and $|\tilde{X}_i| \leq |X_i| \leq C_{1b} \|\nabla \mathcal{L}(\theta_i)\|_2$ a.s.

810 From Lemma E.4, with probability $1 - \delta$, $\mathcal{L}(\theta_{i-1}) - \mathcal{L}^* \leq f(i, \frac{\delta}{2k})$ holds for all $i \in [k]$ and

811 $\left| \sum_{i=1}^k \gamma^{k-i} \tilde{X}_i \right| \leq C_4 \sqrt{\sum_{i=1}^k \gamma^{2k-2i} f(i, \frac{\delta}{2k}) \log \frac{4}{\delta}}$ holds. If this happens, we have

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

812 As long as $K \geq \max\{2\delta^2, 4\}$ (which is a mild restriction on K), we have $\sqrt{\log \frac{2K^2}{\delta} \log \frac{4}{\delta}} \leq$
813 $\sqrt{3 \log^2 \frac{K}{\delta}}$. Plugging in $\frac{1}{1-\gamma} = \frac{R_0}{2\mu(1-\beta_1)} \cdot \frac{1}{\eta}$, we have

$$\begin{aligned} & \eta(1-\beta_1) \sum_{i=1}^k X_i \sum_{j=i}^k \gamma^{k-j} \beta_1^{j-i} \\ & \leq 20C_4(1-\beta_1) \left(\sqrt{\frac{C_{5a}R_0}{2\mu(1-\beta_1)}} \cdot \sqrt{\eta\gamma^k} + \sqrt{\frac{C_{5b}R_0}{2\mu(1-\beta_1)}} \cdot \eta \right) \cdot \sqrt{3} \log \frac{K}{\delta} \\ & \leq 10C_4 \sqrt{\frac{6C_{5a}R_0(1-\beta_1)}{\mu}} \cdot \sqrt{\eta\gamma^k} \log \frac{K}{\delta} + 10C_4 \sqrt{\frac{6C_{5b}R_0(1-\beta_1)}{\mu}} \cdot \eta \log \frac{K}{\delta} \\ & \leq (C_{5c}\gamma^k + C_{5d}\eta) \log \frac{K}{\delta}, \end{aligned}$$

814 where $C_{5c} = 5C_4\sqrt{6C_{5a}R_0(1-\beta_1)/\mu}$ and $C_{5d} = 5C_4\sqrt{6C_{5a}R_0(1-\beta_1)/\mu} +$
815 $10C_4\sqrt{6C_{5b}R_0(1-\beta_1)/\mu}$. Now as long as $K \geq e\delta$ (so that $\log \frac{K}{\delta} \geq 1$), we have

$$\begin{aligned} \mathcal{L}(\theta_k) - \mathcal{L}^* & \leq \gamma^k (\mathcal{L}(\theta_0) - \mathcal{L}^*) + \eta(1-\beta_1) \sum_{i=1}^k X_i \sum_{j=i}^k \gamma^{k-j} \beta_1^{j-i} + C_3\eta \\ & \leq \gamma^k (\mathcal{L}(\theta_0) - \mathcal{L}^*) + (C_{5c}\gamma^k + C_{5d}\eta) \log \frac{K}{\delta} + C_3\eta \\ & \leq (C_{5c} + \mathcal{L}(\theta_0) - \mathcal{L}^*) \gamma^k \log \frac{K}{\delta} + (C_{5d} + C_3) \eta \log \frac{K}{\delta}. \end{aligned}$$

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

$$\begin{cases} C_{5a} & \geq C_{5c} + \mathcal{L}(\theta_0) - \mathcal{L}^* & = 5C_4\sqrt{\frac{6C_{5a}R_0(1-\beta_1)}{\mu}} + \mathcal{L}(\theta_0) - \mathcal{L}^* \\ C_{5b} & \geq C_{5d} + C_3 & = 5C_4\sqrt{\frac{6C_{5a}R_0(1-\beta_1)}{\mu}} + 10C_4\sqrt{\frac{6C_{5b}R_0(1-\beta_1)}{\mu}} + C_3. \end{cases}$$

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

820 *Proof of Theorem C.2.* This is a direct corollary following from Lemma E.5, where letting $\gamma^K =$
821 $\mathcal{O}(\eta)$ gives $K = \mathcal{O}(\frac{1}{\eta} \log \frac{1}{\eta})$, completing the proof. \square

822 *Proof of the first part of Theorem C.1.* Fix $K = \lfloor (T+1)\eta^{-2} \rfloor$. By Lemma D.2, there exists
823 some constant ϵ_3 such that \mathcal{L} is μ -PL in Γ^{ϵ_3} . We can manually define a function $\tilde{\mathcal{L}}$ such
824 that $\tilde{\mathcal{L}} \equiv \mathcal{L}$ inside Γ^{ϵ_3} , and $\tilde{\mathcal{L}}$ still satisfies μ -PL condition outside Γ^{ϵ_3} . Define $\mathcal{L}_m =$
825 $\min\{\mathcal{L}(\theta) : \theta \in \Gamma^{\epsilon_2}, \theta \notin \Gamma^{0.5\epsilon_2}\}$. Plugging in $C_{5a} = \frac{\mathcal{L}(\theta_0) - \mathcal{L}^*}{\log K}$, $K \leq (T+1)\eta^{-2}$ and $\log \frac{1}{\delta}$
826 being upper bounded by $200 \log \frac{1}{\eta}$ into Lemma E.5, we conclude that there exists some $\epsilon < 0.5\epsilon_2$
827 such that if $\mathcal{L}(\theta_0) - \mathcal{L}^* \leq \sup\{\mathcal{L}(\theta) : \theta \in \Gamma^\epsilon\} - \mathcal{L}^*$ and η is sufficiently small, then with probability
828 $1 - \delta$, the loss values at all steps is strictly smaller than \mathcal{L}_m , and $\eta R/\epsilon < 0.5\epsilon_2$ so any single step of
829 update cannot jump from the interior of $\Gamma^{0.5\epsilon_2}$ to the exterior of Γ^{ϵ_2} . So if the statement in Lemma E.5
830 holds, all iterations of AGM stay inside Γ_2^ϵ . Then substituting $K_0 = \lceil \log \frac{1}{\gamma} \eta \rceil = \mathcal{O}(\frac{1}{\eta} \log \frac{1}{\eta})$ gives
831 the result.

832 \square

833 F Proof of the SDE Approximation of AGMs

834 In this section, we present a detailed derivation of our slow SDE approximation of the AGM
835 framework as shown in Theorem 4.1.

Remark F.1. Without causing confusion, we reword the definition of θ_0 and v_0 in this section. In the following calculation in Appendix F, θ_0 and v_0 do not represent the parameters that are initialized at the real beginning of training, instead they represent the θ_{K_0} and v_{K_0} yielded by the first part of Theorem C.1, we define θ_0 as the parameter near the minimizer manifold such that $\|\theta_0 - \phi_0\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$, and v_0 is the velocity vector as the corresponding time step as θ_0 . Our SDE approximation then describes AGM's dynamics after reaching such a state (θ_0, v_0) .

F.1 Lemmas for Adaptive Manifold Projection

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

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

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

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

$$\frac{d\Phi_S(x(t))}{dt} = -\partial \Phi_S(x) S \nabla \mathcal{L}(x) = 0.$$

Further, we take the second derivative of $\Phi_S(x(t))$ with respect to t

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

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

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

Proof. From Lemma C.1 in Li et al. (2021b), we have for $u \in T_x(\Gamma)$, $\nabla^2 \mathcal{L}(x)u = 0$, and for $u \in T_x^\perp(\Gamma)$, it is direct corollary of Lemma 4.3 in Gu et al. (2023b) that

$$\partial \Phi_S(x) S u = 0.$$

The above identity completes the proof. \square

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

$$\partial^2 \Phi_S(x) [uv^T] = -\partial \Phi_S(x) S \partial^2 (\nabla \mathcal{L})(x) [\nabla^2 \mathcal{L}(x)^\dagger S^{-1} uv^T] - S^{-1} \nabla^2 \mathcal{L}(x)^\dagger \partial^2 (\nabla \mathcal{L})(x) [S \partial \Phi(x) uv^T].$$

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

$$\begin{aligned} \partial \Phi'(x') &= P \partial \Phi_S(Px) P \\ \nabla^2 \mathcal{L}'(x') &= P \nabla^2 \mathcal{L}(Px) P \\ \partial^2 (\nabla \mathcal{L}')(x') [M] &= P \partial^2 (\nabla \mathcal{L})(Px) [PMP] \\ \partial^2 \Phi'(x') [M] &= P \partial^2 \Phi(x) [PMP]. \end{aligned}$$

Notice that in the space of x' , the adaptive projection mapping Φ_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

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

A slight modification using the above transformations gives

$$\begin{aligned} \partial^2 \Phi_S(x) [Pv, Pu] &= -\partial \Phi_S(x) S \partial^2 (\nabla \mathcal{L})(x) [Pv, \nabla^2 \mathcal{L}(x)^\dagger S^{-1} Pu] \\ &\quad - S^{-1} \nabla^2 \mathcal{L}(x)^\dagger \partial^2 (\nabla \mathcal{L})(x) [Pv, S \partial \Phi(x) Pu]. \end{aligned}$$

We now redefine $u = Pu$, $v = Pv$, and we organize the above equation

$$\partial^2 \Phi_S(x) [uv^T] = -\partial \Phi_S(x) S \partial^2 (\nabla \mathcal{L})(x) [\nabla^2 \mathcal{L}(x)^\dagger S^{-1} uv^T] - S^{-1} \nabla^2 \mathcal{L}(x)^\dagger \partial^2 (\nabla \mathcal{L})(x) [S \partial \Phi(x) uv^T].$$

We completes the proof. \square

863 F.2 Iteration Stays Near Manifold

864 Now we begin the final preparations before deriving the slow SDE near the manifold. Note that
 865 in the end of convergence analysis, the total steps equal $K = \lfloor (T+1)\eta^{-2} \rfloor$ and the converging
 866 step $K_0 = \mathcal{O}(\frac{1}{\eta} \log \frac{1}{\eta})$. So after time shifting, the high probability convergence of $\lfloor (T+1)\eta^{-2} \rfloor -$
 867 $\mathcal{O}(\frac{1}{\eta} \log \frac{1}{\eta}) > \lfloor T\eta^{-2} \rfloor$ steps are ensured in Lemma E.5. Now denote $K := \lfloor T\eta^{-2} \rfloor$ be the total
 868 number of steps in our analysis. Let β be some constant in $(0, 0.5)$, whose exact value will be
 869 specified later. First, we bound the movement of projected steps by showing that ϕ shifts no more
 870 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
 871 AGMs after the projection.

Lemma F.4. *If ϕ_k stays inside Γ^{ϵ_2} for any $k \in [0, K]$, then for any $\delta = \mathcal{O}(\text{poly}(\eta))$, with probability $1 - \delta$, for any $k \in [0, K - \Delta K]$, $\Delta k \in [\Delta K]$,*

$$\|\phi_{k+\Delta k} - \phi_k\|_2 \leq C_6 \eta^{0.5-0.5\beta} \sqrt{\log \frac{1}{\eta\delta}}$$

872 for some constant C_6 .

873 *Proof.* Recall that $\Phi_{S(\mathbf{v})}(\boldsymbol{\theta})$ is \mathcal{C}^4 on $\mathcal{X}^{\epsilon_2} := \{(\mathbf{v}, \boldsymbol{\theta}) : 0 \preceq \mathbf{v} \preceq R_1, \boldsymbol{\theta} \in \Gamma^{\epsilon_2}\}$, since \mathcal{X}^{ϵ_2} is compact,
 874 $\Phi_{S(\mathbf{v})}(\boldsymbol{\theta})$ is then bounded and Lipschitz on \mathcal{X}^{ϵ_2} . Similarly, $\partial\Phi_{S(\mathbf{v})}(\boldsymbol{\theta})$ is bounded and Lipschitz on
 875 \mathcal{X}^{ϵ_2} . For any $k \in [0, K]$, let $\bar{k} = k - 2 \log_{\beta_1} \eta$, we have:

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

876 where the second equality comes from the fact that one step of update on \mathbf{v} is of $\mathcal{O}(\eta^2)$ and the
 877 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
 878 follows from the boundedness and Lipschitzness of $\partial\Phi$. We can decompose \mathbf{m}_k as:

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

879 A key observation is that $\partial\Phi_{S(\mathbf{v}_{\bar{k}})}(\boldsymbol{\theta}_{\bar{k}})S(\mathbf{v}_{\bar{k}})\nabla \mathcal{L}(\boldsymbol{\theta}_{\bar{k}}) = 0$ from Lemma F.2, which allows us to view
 880 $\phi_{k+1} - \phi_k$ as $\sum_{i=\bar{k}}^k \tilde{\mathbf{z}}_{k,i} + \mathcal{O}(\eta^2 \log \frac{1}{\eta})$ where $\tilde{\mathbf{z}}_{k,i} = \partial\Phi_{S(\mathbf{v}_{\bar{k}})}(\boldsymbol{\theta}_{\bar{k}})(\eta(1 - \beta_1)\beta_1^{k-i}S(\mathbf{v}_{\bar{k}})\mathbf{z}_i)$. Note
 881 that $\tilde{\mathbf{z}}_{k,i}$ is \mathcal{F}_{i+1} -measurable and its mean is $\mathbf{0}$, since $\tilde{\mathbf{z}}_{k,i}$ just applies a linear tensor transformation
 882 to \mathbf{z}_i . If we define a constant $C_{6a} := \sup \{\|\partial\Phi_{S(\mathbf{v})}(\boldsymbol{\theta})\|_2 \mid (\mathbf{v}, \boldsymbol{\theta}) \in \mathcal{X}^{\epsilon_2}\} \cdot (1 - \beta_1) \cdot \epsilon^{-1}$ that is
 883 independent of k and i , then $\|\tilde{\mathbf{z}}_{k,i}\|_2$ is almost surely bounded by $\eta\beta_1^{k-i}C_{6a}\|\mathbf{z}_i\|_2$.

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

$$\begin{aligned} \phi_{k+\Delta k} - \phi_k &= \sum_{j=k}^{k+\Delta k-1} (\phi_{j+1} - \phi_j) \\ &= \sum_{j=k}^{k+\Delta k-1} \left(\sum_{i=j-2 \log_{\beta_1} \eta}^j \tilde{\mathbf{z}}_{j,i} + \mathcal{O}\left(\eta^2 \log \frac{1}{\eta}\right) \right) \end{aligned}$$

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

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

$$\begin{aligned} \phi_{k+\Delta k} - \phi_k &\leq \sqrt{2\eta^2 \left(\frac{C_{6a}R}{1-\beta_1} \right)^2 \cdot (R_{\text{grp}}H + 2\log_{\beta_1}\eta) \cdot \log \frac{2}{\delta}} \\ &\leq C_{6b} \sqrt{\eta^{1-\beta} \log \frac{2}{\delta}} \end{aligned}$$

888 for some constant C_{6b} . Finally, plugging in $\delta' = \frac{\delta}{K \cdot \Delta K}$ and taking union bound over all $k \in$
 889 $[0, K - \Delta K]$ and $\Delta k \in [\Delta K]$ gives the theorem. \square

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

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

- 894 1. For any $k \in [0, \hat{K}]$, $\phi_k \in \Gamma$ and $\|\theta_k - \phi_k\|_2 \leq C_{8a} \sqrt{\eta \log \frac{1}{\eta\delta}}$.
- 895 2. For any $k \in [0, \hat{K} - \Delta K]$, $\Delta k \in [\Delta K]$, $\|\phi_{k+\Delta k} - \phi_k\|_2 \leq C_{8b} \eta^{0.5-0.5\beta} \sqrt{\log \frac{1}{\eta\delta}}$.

896 Here $C_{8a} = \frac{4C_2}{\sqrt{2\mu}C_1} \cdot \sqrt{(C_{5a} + C_{5b})}$ and $C_{8b} = C_6\sqrt{2}$ are two constants.

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

899 *Proof.* Denote $\delta := \eta^{100}$. From Lemma E.5, with probability $1 - \delta/2$, all $k \in [0, K]$ satisfy
 900 $\mathcal{L}(\theta_k) - \mathcal{L}^* \leq (C_{5a} + C_{5b})\eta \log \frac{2K}{\delta}$. Combining Lemma D.1, this implies $\|\theta_k - \phi_k\|_2 \leq$
 901 $\frac{2C_2}{\sqrt{2\mu}C_1} \cdot \sqrt{(C_{5a} + C_{5b})\eta \log \frac{2K}{\delta}}$ for any $k \in [0, K]$. When η is small enough such that
 902 $\frac{2C_2}{\sqrt{2\mu}C_1} \cdot \sqrt{(C_{5a} + C_{5b})\eta \log \frac{2K}{\delta}} + \eta R/\epsilon < \epsilon_2$, any $\phi_k \in \Gamma$ with $k \geq 0$ will imply $\phi_{k+1} \in \Gamma$,
 903 since θ_{k+1} cannot escape Γ^{ϵ_2} . Giving $\phi_0 \in \Gamma$ and using induction, we conclude that all $\phi_k \in \Gamma$ for
 904 $k \geq 0$.

905 When the above holds, the requirement of Lemma F.4 is met. Then with probability $1 - \delta/2$, for any
 906 $k \in [0, K - \Delta K]$, $\Delta k \in [\Delta K]$, we have $\|\phi_{k+\Delta k} - \phi_k\|_2 \leq C_6 \eta^{0.5-0.5\beta} \sqrt{\log \frac{2}{\eta\delta}}$.

907 Finally, we just take the union of Lemma E.5 and Lemma F.4. With $\log \frac{2K}{\delta} \leq 4 \log \frac{1}{\eta\delta}$ and
 908 $\log \frac{2}{\eta\delta} \leq 2 \log \frac{1}{\eta\delta}$ (which are mild restrictions since η is small), we have the theorem. \square

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

912 **Definition F.2** (Well-behaved Sequence). Denote the event of step k being η^{100} -good as \mathcal{E}_k . Let ϕ_{null}
 913 be a fixed point on Γ . Starting from $\hat{\theta}_0 = \theta_0$ and $\hat{\mathbf{v}}_0 = \mathbf{v}_0$, we define a sequence of $(\hat{\theta}_k, \hat{\mathbf{v}}_k, \hat{\mathbf{m}}_k)$ as
 914 follows:

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

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

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

$$\begin{aligned}\hat{\theta}_{k+1} &:= \hat{\theta}_k - \eta S(\hat{v}_{k+1}) \hat{m}_{k+1} \\ &\quad - \underbrace{\mathbf{1}_{\mathcal{E}_k} (\hat{\theta}_k - \eta S(\hat{v}_{k+1}) \hat{m}_{k+1}) + \mathbf{1}_{\mathcal{E}_k} \phi_{\text{null}}}_{:= e_k}.\end{aligned}$$

917 F.3 Moment Calculation of AGMs Near Manifold

918 **Additional Notations.** To utilize the analysis framework in Gu et al. (2023b), we first introduce
919 some notations needed. Consistent with Gu et al. (2023b), we pretend that AGMs proceed with
920 $H = \frac{1}{\eta}$ local steps, as a single worker (without multiple workers). We denote every H steps as one
921 round. Next, we define a “giant step”, which encompasses $R_{\text{grp}} = \frac{1}{\eta^\beta}$ rounds, corresponding to
922 $R_{\text{grp}} \cdot H$ steps. We consider a total timescope of $\frac{T}{\eta^\alpha}$ steps, which corresponds to $\frac{T}{\eta^{1-\beta}}$ giant steps.

923 For any $0 \leq s < R_{\text{grp}}$ and $0 \leq t \leq H$, we use $\hat{\theta}_t^{(s)}$ and $\hat{\theta}_k$ (where $k = sH + t$) exchangeably to
924 denote the parameter we get on the t -th local step of round s , which is also the k -th global step. Also
925 note that for any $0 \leq s < R_{\text{grp}}$, $\hat{\theta}_H^{(s)}$ and $\hat{\theta}_0^{(s+1)}$ refer to the same thing. We define the notation $\hat{v}_t^{(s)}$,
926 $\hat{m}_t^{(s)}$ and $\mathcal{E}_t^{(s)}$ in the same way as we did for θ .

927 We further introduce a list of notations:

$$\begin{aligned}\hat{g}_t^{(s)} &:= \nabla \ell_t^{(s)}(\hat{\theta}_t^{(s)}), \quad \hat{S}_k = S(\hat{v}_k), \quad \hat{S}_t^{(s)} := S(\hat{v}_t^{(s)}), \quad \hat{S}^{(s)} := \hat{S}_0^{(s)}, \quad \hat{\phi}^{(s)} := \Phi_{\hat{S}^{(s)}}(\hat{\theta}_0^{(s)}), \\ \hat{x}_t^{(s)} &:= \hat{\theta}_t^{(s)} - \hat{\phi}^{(s)}, \quad \Delta \hat{\phi}^{(s)} := \hat{\phi}^{(s)} - \hat{\phi}^{(0)}, \quad \Sigma_0 := \Sigma(\hat{\phi}^{(0)}), \quad P_{\parallel} := \partial \Phi_{\hat{S}^{(0)}}(\hat{\phi}^{(0)}), \quad P_{\perp} := I - P_{\parallel}, \\ \hat{q}_t^{(s)} &:= \mathbb{E}[\hat{x}_t^{(s)}], \quad \hat{A}_t^{(s)} := \mathbb{E}[\hat{x}_t^{(s)} \hat{x}_t^{(s)\top}], \quad \hat{B}_t^{(s)} := \mathbb{E}[\hat{x}_t^{(s)} \Delta \hat{\phi}^{(s)\top}].\end{aligned}$$

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

$$\begin{aligned}\|\hat{x}_t^{(s)}\|_2 &\leq C_{9a} \sqrt{\eta \log \frac{1}{\eta}}, \\ \|\hat{\theta}_t^{(s)} - \hat{\theta}_0^{(s)}\|_2 &\leq C_{9b} \sqrt{\eta \log \frac{1}{\eta}}, \\ \|\hat{\phi}^{(s)} - \hat{\phi}^{(0)}\|_2 &\leq C_{9c} \eta^{0.5-0.5\beta} \sqrt{\log \frac{1}{\eta}}.\end{aligned}$$

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

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

939 **Lemma F.6.** *For any $k \geq 0$, we have*

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

940 *for some constant C_{10} .*

941 *Proof.* We have

$$\begin{aligned}\nabla \mathcal{L}(\hat{\theta}_{k+1}) - \nabla \mathcal{L}(\hat{\theta}_k) &= \nabla^2 \mathcal{L}(\hat{\theta}_k)(\hat{\theta}_{k+1} - \hat{\theta}_k) + \mathcal{O}\left(\|\hat{\theta}_{k+1} - \hat{\theta}_k\|_2^2\right) \\ &= \nabla^2 \mathcal{L}(\hat{\theta}_k)(\hat{\theta}_{k+1} - \hat{\theta}_k) + \mathcal{O}(\eta^2) + \mathcal{O}(\|e_k\|_2),\end{aligned}$$

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

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

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

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

945 After that, notice that

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

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

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

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

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

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

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

950 which concludes the proof. □

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

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

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

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

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

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

955 Note that E_1 is negligible:

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

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

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

957 Combining the results of E_1 and E_2 gives

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

958 for some constant C_{11} , which completes the proof. \square

959 **F.3.1 Moment Calculation Within a Giant Step**

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

962 Now there are only a few preparations left before we get into the direct part of the moment calculation.
963 For all $0 \leq s < R_{\text{grp}}$, $0 \leq t \leq H$. Note that $\|\hat{\mathbf{v}}_{k+1} - \hat{\mathbf{v}}_k\|_2 = (1 - \beta_2) \|V(\hat{\mathbf{g}}_k \hat{\mathbf{g}}_k^\top) - \hat{\mathbf{v}}_k\|_2 = \mathcal{O}(1 - \beta_2) = \mathcal{O}(\eta^2)$, so combining with the Lipschitzness of S gives $\|\hat{\mathbf{S}}_{k_2} - \hat{\mathbf{S}}_{k_1}\|_2 = \mathcal{O}((k_2 - k_1)\eta^2)$
964 for any $k_2 > k_1$ and $k_2 - k_1 = o(\eta^{-2})$. Next, we begin our moment calculation analysis starting
965 from the update in one step.
966

967 **Lemma F.8.** For all $2 \log_{\beta}(\eta) \leq k \leq R_{\text{grp}} H$, we have

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

968 *Proof.* We write the update rule:

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

969 We can prove that $\Delta \hat{\theta}_1$ and $\Delta \hat{\theta}_2$ are small in expectation. If $k = 0$ then $\Delta \hat{\theta}_1 = \mathbf{0}$; and if $k > 0$, we
970 can decompose $\mathbb{E} [\Delta \hat{\theta}_1]$ as:

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

971 Here, the second equality holds because \mathbf{z}_k is conditioned on time k , when \hat{S}_{k-1} has already been
972 determined. For $\Delta \hat{\theta}_2$, let $\bar{k} = k - 2 \log_{\beta_1}(\eta)$, we have

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

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

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

974 which concludes the proof. \square

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

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

982 We decompose the preconditioner matrix in the very begining of the giant step as $\hat{S}_0 = \hat{S}(\hat{v}_0) = PP$,
983 where $P = \hat{S}^{1/2}$. Then we give the first moment calculation of $\hat{\phi}$ in the following lemma.

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

$$\mathbb{E} [\hat{\phi}^{(s+1)} - \hat{\phi}^{(s)}] = \begin{cases} -\frac{H\eta^2}{2} \hat{\mathbf{S}}_0 \partial^2 \nabla \mathcal{L}(\phi_{(0)}) [\mathbf{P} \mathcal{V}_{\nabla^2 \mathcal{L}'(\phi'_{(0)})} (\mathbf{P} \Sigma_0 \mathbf{P}) \mathbf{P}] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), & R_0 < s < R_{\text{grp}} \\ \tilde{\mathcal{O}}(\eta), & s \leq R_0 \end{cases}$$

985 where $R_0 := \max \left\{ \left\lceil \frac{10}{\lambda_{\max} \alpha} \log \frac{1}{\eta} \right\rceil, \left\lceil 2 \log_{1/\beta} \frac{1}{\eta} \right\rceil \right\}$.

986 *Proof.* First we consider the scenario when $R_0 < s < R_{\text{grp}}$. By Lemma F.8, the update rule holds.

987 And we consider an auxiliary process $\{\hat{\theta}'_t\}$. Let $L'(\mathbf{x}) := L(\mathbf{P}\mathbf{x})$, then

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

988 For an one-step GD update, we have that

$$\begin{aligned} \hat{\theta}'_{t+1} &= \hat{\theta}'_t - \eta \nabla \mathcal{L}'(\hat{\theta}'_t) \\ &= \hat{\theta}'_t - \eta \mathbf{P} \nabla \mathcal{L}(\mathbf{P}\hat{\theta}'_t). \end{aligned}$$

989 Similarly, we define $\mathbf{A}_t'^{(s)} := \mathbb{E}[\mathbf{x}_t'^{(s)} \mathbf{x}_t'^{(s)\top}]$, $\mathbf{q}_t'^{(s)} := \mathbb{E}[\mathbf{x}_t'^{(s)}]$, and $\mathbf{B}_t'^{(s)} := \mathbb{E}[\mathbf{x}_t'^{(s)} \Delta \phi'^{(s)\top}]$,
990 where $\phi(\mathbf{x})$ is the gradient flow projection of point \mathbf{x} .

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

$$\mathbf{P}\hat{\theta}'_{t+1} = \mathbf{P}\hat{\theta}'_t - \eta \hat{\mathbf{S}}_0 \nabla \mathcal{L}(\mathbf{P}\hat{\theta}'_t). \quad (4)$$

992 We now define $\hat{\theta}'_t := \mathbf{P}^{-1} \hat{\theta}'_t$, then combining Equation (4) and Lemma F.8 gives

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

993 Now we can apply the results in Lemma I.36 from Gu et al. (2023b) for the update of $\hat{\theta}'$, with loss
994 function $\mathcal{L}'(\hat{\theta})$, number of workers $k = 1$ and manifold projection $\Phi'(\hat{\theta})$, which gives

$$\begin{aligned} \mathbf{P} \mathbb{E} [\phi'^{(s+1)} - \phi'^{(s)}] &= \mathbb{E} [\phi^{(s+1)} - \phi^{(s)}] \\ &= -\frac{H\eta^2}{2} \mathbf{P} \mathbf{P} \partial^2 \nabla \mathcal{L}(\phi_{(0)}) [\mathbf{P} \mathcal{V}_{\nabla^2 \mathcal{L}'(\phi'_{(0)})} (\mathbf{P} \Sigma_0 \mathbf{P}) \mathbf{P}] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), \end{aligned}$$

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

997 The proof when $s \leq R_0$ is a direct conclusion of Lemma I.36 in Gu et al. (2023b) since the
998 $R_0 \propto \log \frac{1}{\eta}$ in our case. \square

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

$$\mathbb{E} [\phi^{(s+1)} - \phi^{(s)}] = \begin{cases} \frac{H\eta^2}{2} \hat{\mathbf{S}}_0 \partial^2 \Phi_{\hat{\mathbf{S}}_0}(\phi^{(0)}) [\hat{\mathbf{S}}_0 \Sigma_0 \hat{\mathbf{S}}_0] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), & R_0 < s < R_{\text{grp}} \\ \tilde{\mathcal{O}}(\eta), & s \leq R_0 \end{cases}$$

1000 *Proof.* Notice that for the preconditioned projection, we also have the corresponding transformation

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

1001 The above two equations and Lemma I.36 in Gu et al. (2023b) complete the proof. \square

1002 **Lemma F.10.** *The second moment of the change of manifold projection every round is*

$$\mathbb{E} [(\phi^{(s+1)} - \phi^{(s)})(\phi^{(s+1)} - \phi^{(s)})^\top] = \begin{cases} H\eta^2 \hat{\mathbf{S}}_0 \mathbf{P}_\parallel \hat{\mathbf{S}}_0 \Sigma_0 \hat{\mathbf{S}}_0 \mathbf{P}_\parallel \hat{\mathbf{S}}_0 + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), & R_0 < s < R_{\text{grp}} \\ \tilde{\mathcal{O}}(\eta), & s \leq R_0 \end{cases}$$

1003 where $R_0 := \max \left\{ \left\lceil \frac{10}{\lambda_{\max} \alpha} \log \frac{1}{\eta} \right\rceil, \left\lceil 2 \log_{1/\beta} \frac{1}{\eta} \right\rceil \right\}$.

1004 *Proof.* According to Lemma I.37 in Gu et al. (2023b), we could write the second moment for $\hat{\theta}'$ as

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

1005 Notice that

$$\begin{aligned} \Sigma'_{0,\parallel} &:= \partial \Phi'(\phi^{(0)}) \Sigma'_0 \partial \Phi'(\phi^{(0)}) \\ &= \mathbf{P} \partial \Phi(\phi^{(0)}) \mathbf{P} \mathbf{P} \Sigma_0 \mathbf{P} \mathbf{P} \partial \Phi(\phi^{(0)}) \mathbf{P}. \end{aligned}$$

1006 When $R_0 \leq s < R_{\text{grp}}$,

$$\begin{aligned} \mathbb{E} \left[(\phi^{(s+1)} - \phi^{(s)}) (\phi^{(s+1)} - \phi^{(s)})^\top \right] &= \mathbb{E} \left[\mathbf{P} (\phi^{(s+1)} - \phi^{(s)}) (\phi^{(s+1)} - \phi^{(s)})^\top \mathbf{P} \right] \\ &= \hat{\Sigma}_0 \mathbf{P}_\parallel \hat{\Sigma}_0 \Sigma_0 \hat{\Sigma}_0 \mathbf{P}_\parallel \hat{\Sigma}_0. \end{aligned}$$

1007 The proof when $s \leq R_0$ is a direct conclusion of Lemma I.37 in Gu et al. (2023b) since the
1008 $R_0 \propto \log \frac{1}{\eta}$ in our case. \square

1009 Then we give the moment change of ϕ within a single giant step.

1010 **Theorem F.1.** Given $\|\hat{\theta}^{(0)} - \phi^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$, for $0 < \beta < 0.5$, the first and second moments
1011 of $\Delta \phi^{(R_{\text{grp}})} := \phi^{(R_{\text{grp}})} - \phi^{(0)}$ are as follows:

$$\begin{aligned} \mathbb{E}[\Delta \phi^{(R_{\text{grp}})}] &= \frac{\eta^{1-\beta}}{2} \hat{\Sigma}_0 \partial^2 \Phi_{\hat{\Sigma}_0}(\phi^{(0)}) [\hat{\Sigma}_0 \Sigma_0 \hat{\Sigma}_0] + \tilde{\mathcal{O}}(\eta^{1.5-2\beta}) + \tilde{\mathcal{O}}(\eta), \\ \mathbb{E}[\Delta \phi^{(R_{\text{grp}})}]^\top &= \eta^{1-\beta} \hat{\Sigma}_0 \Sigma_\parallel(\phi^{(0)}, \hat{\Sigma}^{(0)}) \hat{\Sigma}_0 + \tilde{\mathcal{O}}(\eta^{1.5-1.5\beta}) + \tilde{\mathcal{O}}(\eta), \end{aligned}$$

1012 where $\Sigma_\parallel(\phi^{(0)}, \hat{\Sigma}^{(0)}) := \mathbf{P}_\parallel \hat{\Sigma}_0 \Sigma_0 \hat{\Sigma}_0 \mathbf{P}_\parallel$.

1013 *Proof.* First we prove the first moment change as

$$\begin{aligned} \mathbb{E}[\Delta \phi^{(R_{\text{grp}})}] &= \mathbb{E} \left[\sum_{s=0}^{R_{\text{grp}}-1} \phi^{(s+1)} - \phi^{(s)} \right] \\ &= \sum_{s=0}^{R_0} \mathbb{E}[\phi^{(s+1)} - \phi^{(s)}] + \sum_{s=R_0+1}^{R_{\text{grp}}-1} \mathbb{E}[\phi^{(s+1)} - \phi^{(s)}] \\ &= \frac{\eta^{1-\beta}}{2} \hat{\Sigma}_0 \partial^2 \Phi_{\hat{\Sigma}_0}(\phi^{(0)}) [\hat{\Sigma}_0 \Sigma_0 \hat{\Sigma}_0] + \tilde{\mathcal{O}}(\eta^{1.5-2\beta}) + \tilde{\mathcal{O}}(\eta). \end{aligned}$$

1014 The last equation is a direct conclusion of Corollary F.2.

1015 And for the second moment, we have

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

1016 where the last equation uses $\mathbb{E}[(\phi^{(s+1)} - \phi^{(s)})] \mathbb{E}[(\phi^{(s'+1)} - \phi^{(s')})^\top] = \tilde{\mathcal{O}}(\eta^2)$. \square

1017 Next, we proceed with the updates of v .

1018 **Lemma F.11.** Given $c := \frac{1-\beta_2}{\eta^2}$, and we have

$$\mathbb{E} \left[\hat{v}_0^{(R_{\text{grp}})} - \hat{v}_0^{(0)} \right] = c\eta^{1-\beta} \left(V \left(\Sigma_0^{(0)} \right) - \hat{v}_0^{(0)} \right) + \mathcal{O}(\eta^{1.5-1.5\beta}).$$

1019 *Proof.* We have

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

1020 Note that

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

1021 Combining with the linearity of \mathbf{V} , we conclude that

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

1022 To transfer from $\hat{\mathbf{v}}_0^{(0)}$ to arbitrary $\hat{\mathbf{v}}_0^{(s)}$, we simply expand to get the result:

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

1023 Thus we have

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

1024 where the last equation uses the fact that $1 - \beta_2^{R_{\text{grp}}H} = 1 - (1 - c\eta^{1-\beta}) + \mathcal{O}(\eta^{2-2\beta}) = c\eta + \mathcal{O}(\eta^2)$. \square

1025 Also, for the second moment change of $\hat{\mathbf{v}}$, we get the following lemma

1026 **Lemma F.12.** *The second moment change of $\hat{\mathbf{v}}$ over a giant step is*

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

Proof.

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

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

The last equation uses

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

and

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

1028 The above equation completes the proof. \square

1029 F.4 Weak Approximation

1030 After we get the first and second moment changes within a giant step, we now utilize the moment
1031 calculation to prove the SDE approximation part of Theorem C.1. First, we recall our slow SDE for
1032 AGMs

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

1033 We then open the projection mapping $P_{\zeta, \mathbf{S}(t)}$ as

$$\begin{cases} d\zeta = \partial \Phi_{\mathbf{S}(v)}(\zeta) \mathbf{S}(v) \Sigma^{1/2}(\zeta) d\mathbf{W}_t + \frac{1}{2} \partial^2 \Phi_{\mathbf{S}(v)}(\zeta) [\mathbf{S}(v) \Sigma(\zeta) \mathbf{S}(v)] dt, \\ dv(t) = c(V(\Sigma(\zeta)) - v) dt. \end{cases} \quad (5)$$

1034 Now it suffices to prove the SDE in Equation (5) tracks the trajectory in AGMs within $\mathcal{O}(\frac{1}{\eta^2})$ steps in
1035 a weak approximation sense.

1036 First, we have to show that the solution of Equation (5) is close in the minimizer manifold

1037 **Lemma F.13.** *Let $\mathbf{X}(t) := (\zeta(t)^\top, \mathbf{v}(t)^\top)^\top$ be the solution of Equation (5) with $\zeta(0) \in \Gamma$, and*
1038 *$\mathbf{v}(0) \in \mathbb{R}^d$, then we have that $\zeta(t) \in \Gamma$ for all $t \geq 0$.*

1039 *Proof.* According to Filipović (2000); Du and Duan (2006), for a closed manifold \mathcal{M} to be viable for
1040 the SDE $d\mathbf{X}(t) = \mathbf{A}(\mathbf{X}(t))d\mathbf{W}_t + \mathbf{b}(\mathbf{X}(t))dt$, where $\mathbf{A}(\cdot) : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d \times 2d}$ and $\mathbf{b}(\cdot) : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$
1041 are locally Lipschitz, it suffices to show that the following Nagumo type consistency condition holds:

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

1042 Following the argument in Gu et al. (2023b), here we also only need to show that $\mathbf{P}_\perp(\mathbf{x})\mu(\mathbf{x}) = 0$,
1043 where $\mathbf{P}_\perp(\mathbf{x}) := \mathbf{I}_d - \partial \Phi_{\mathbf{I}}(\mathbf{x})$. $\Phi_{\mathbf{I}}(\mathbf{x})$ is also the gradient flow projection at point \mathbf{x} .

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

1044 The last equation uses Lemma F.3. Again, applying Lemma F.3 gives

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

1045 The above equation completes the proof. \square

1046 To establish Theorem 4.1, we give an equivalent theorem, which capture the closeness of $\mathbf{X}(t)$ and
1047 $\bar{\mathbf{X}}_t$ in a long horizon. Also, for the proof of Theorem 4.1, it suffices to prove the following lemma,
1048 whose proof would be shown at the end of Appendix F.4.

1049 **Theorem F.2.** *If $\|\boldsymbol{\theta}^{(0)} - \phi^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$ and $\boldsymbol{\zeta}(0) = \phi^{(0)}$, $\mathbf{v}(0) = \mathbf{v}^{(0)}$, then for a giant
1050 step $R_{\text{grp}} = \lfloor \frac{1}{\eta^{0.25}} \rfloor$, for every test function $g \in \mathcal{C}^3$,*

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

1051 where C_g is a constant independent of η but depends on $g(\cdot)$ and $b > 0$ is a universal constant
1052 independent of $g(\cdot)$ and η .

1053 F.5 Preliminary and Additional Notations

1054 We first introduce some notations and preliminary background. We consider the following stochastic
1055 gradient algorithms (SGAs)

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

1056 where $\mathbf{x}_n \in \mathbb{R}^{2d}$ is the parameter vector, η_e is the effective learning rate, $\mathbf{h}(\cdot, \cdot) : \mathbb{R}^{2d} \times \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$
1057 depend on the current parameter vector \mathbf{x}_n and the noise vector $\boldsymbol{\xi}_n$ sampled from some distribution
1058 $\Xi(\mathbf{x}_n)$.

1059 We also consider the Stochastic Differential Equation (SDE) of the following form:

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

1060 where $\mathbf{b} : \mathbb{R}^{2d} \times \mathbb{R}^+ \rightarrow \mathbb{R}^{2d}$ is the drift vector function and $\sigma : \mathbb{R}^{2d} \times \mathbb{R}^+ \rightarrow \mathbb{R}^{2d \times 2d}$ is the diffusion
1061 matrix function.

1062 According to the moment calculations in Corollary F.2, Lemma F.10, Lemma F.11, and Lemma F.12,
1063 we set $\eta_e = \eta^{1-\beta}$, and

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

1064 Next, we are going to define the one giant step change of the parameter, both for SGAs and SDE.

$$\begin{aligned} \hat{\hat{\mathbf{X}}}^{(lR_{\text{grp}})} &:= (\Phi_{\hat{\mathbf{S}}^{(lR_{\text{grp}})}}(\hat{\boldsymbol{\theta}}, \mathbf{v}^{lR_{\text{grp}}}), \quad \Delta^{(n)} := \hat{\hat{\mathbf{X}}}^{((n+1)R_{\text{grp}})} - \hat{\hat{\mathbf{X}}}^{(nR_{\text{grp}})}, \\ \tilde{\Delta}^{(n)} &:= \mathbf{X}_{(n+1)\eta_e} - \hat{\hat{\mathbf{X}}}^{(nR_{\text{grp}})}, \quad \mathbf{b}^{(n)} := \mathbf{b}(\hat{\hat{\mathbf{X}}}^{(nR_{\text{grp}})}), \quad \sigma^{(n)} := \sigma(\hat{\hat{\mathbf{X}}}^{(nR_{\text{grp}})}). \end{aligned}$$

1065 We now give a lemma to give the approximation of the first, second, and higher-order moment change
1066 of the SDE.

1067 **Lemma F.14.** *There exists a positive constant c_0 independent of η_e and g such that for all $\boldsymbol{\zeta} \in \Gamma$, it
1068 holds for all $1 \leq i \leq d$ that*

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

1069 *Proof.* (i) By Lemma F.13, the first half solution $\zeta(t)$ in $\mathbf{X}(t)$ of Equation (5) stays in the manifold
 1070 almost surely when $\zeta(0) \in \Gamma$. (ii) We assume that $\mathcal{L} \in \mathcal{C}^5$, so $\mathbf{b}, \sigma \in \mathcal{C}^4$. (iii) We know that Γ is
 1071 compact by Assumption 4.2. Then we can directly apply Lemma B.3 in Malladi et al. (2022) and
 1072 Lemma 26 in Li et al. (2019). \square

1073 **Lemma F.15** (Adaption of Lemma I.41 in Gu et al. (2023b)). *Given drift term and diffusion term*
 1074 *$\mathbf{b}, \sigma \in G^\alpha$ and Lipschitz. Let $s \in [0, T]$ and $g \in G^\alpha$. Then for $t \in [s, T]$, we can define:*

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

1075 *where $\mathcal{P}_X(\mathbf{x}, s, t)$ denotes the distribution of \mathbf{X}_t with the initial condition $\mathbf{X}(s) = \mathbf{x}$. Then*
 1076 *$u(\cdot, s, t) \in G^\alpha$ uniformly in s, t .*

1077 F.6 Proof of the Approximation for Slow SDE of AGMs

1078 For the giant step constant $\beta \in (0, 0.5)$, we define several quantities $a_1 = \frac{1.5-2\beta}{1-\beta} \in (1, 1.5)$,
 1079 $a_2 = \frac{1}{1-\beta} \in (1, 2)$, $a_3 = \frac{1.5-1.5\beta}{1-\beta} = 1.5$, and $a_4 = \frac{2-2\beta}{1-\beta} = 2$. In this part, we will show that only
 1080 a_1 and a_2 would impact the error bound in our approximation theorem.

1081 The following lemma captures the difference between the SDEs' and the AGMs' first and second
 1082 moment changes, as a key step to control the approximation error, utilizing the moment calculation
 1083 results from the last section.

1084 **Lemma F.16.** *If $\|\theta^{(0)} - \phi^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$, then it holds for all $0 \leq n \leq \lfloor T/\eta_e \rfloor$ and $1 \leq i \leq d$*
 1085 *that*

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

1086

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

1087 *where c_1 and b are constants independent of η_e and g .*

1088 *Proof.* According to Appendix F.2, we have that

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

1089 We can further use Corollary F.2, Lemma F.10, Lemma F.11, and Lemma F.12, which gives

$$\left| \mathbb{E}[\Delta_i^{(n)} - \eta_e b_i^{(n)}] \right| \leq c_2 \left(\eta_e^{a_1} \left(\log \frac{1}{\eta_e} \right)^b + \eta_e^{a_2} \left(\log \frac{1}{\eta_e} \right)^b \right), \quad (6)$$

$$\left| \mathbb{E}[\Delta_i^{(n)} \Delta_j^{(n)} - \eta_e \sum_{l=1}^d \sigma_{i,l}^{(n)} \sigma_{l,j}^{(n)}] \right| \leq c_2 \left(\eta_e^{a_1} \left(\log \frac{1}{\eta_e} \right)^b + \eta_e^{a_2} \left(\log \frac{1}{\eta_e} \right)^b \right) \quad (7)$$

$$\mathbb{E} \left[\left| \prod_{s=1}^6 \Delta_{i_s}^{(n)} \right| \right] \leq c_2^2 \eta_e^{2a_1} \left(\log \frac{1}{\eta_e} \right)^{2b}. \quad (8)$$

1090 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
 1091 and Lemma F.14 give the Lemma. \square

1092 **Lemma F.17.** For a test function $g \in \mathcal{C}^3$, and we define $u_{l,n}(\mathbf{x}) := u(\mathbf{x}, l\eta_e, n\eta_e) =$
1093 $\mathbb{E}_{\mathbf{X}_t \sim \mathcal{P}(\mathbf{x}, l\eta_e, n\eta_e)}[g(\mathbf{X}_t)]$. If $\|\boldsymbol{\theta}^{(0)} - \phi^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$, then for all $0 \leq l \leq n-1$, and
1094 $1 \leq n \leq \lfloor T/\eta_e \rfloor$, it holds that

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

1095 where $C_{g,3}$ is some positive constant independent of η_e but can depend on g .

1096 *Proof.* Given $g \in \mathcal{C}^3$, by Lemma F.15, we have $u_{l,n}(\mathbf{x}) \in \mathcal{C}^3$ for all l and n . Which is to say that
1097 there exists a function $Q(\cdot) \in G$, such that the partial derivative of $u_{l,n}(\mathbf{X})$ with respect to l, n, \mathbf{x} up
1098 to the third order is bounded by $Q(\mathbf{x})$. By the law of total expectation and triangle inequality,

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

1099 For I_2 and I_3 , due to the compactness of Γ and $\mathbf{v} \preceq R_1$ from Assumption C.3, $Q(\mathbf{x})$ can be bounded
1100 for some constant $C_{g,4}$ independent of η_e but could depend on test function g . Hence, we have that
1101 $I_2 + I_3 \leq C_{g,4}\eta^{100}$.

1102 Using the triangle inequality, we first decompose I_1 into several terms as

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

1103 where the third order remainders \mathcal{R} and $\tilde{\mathcal{R}}$ are

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

1104 where $\alpha, \tilde{\alpha} \in (0, 1)$. Again, notice that the Γ is compact and $\mathbf{v} \preceq R_1$, thus we can bound the
1105 derivatives of $u_{l,n}(\mathbf{x})$ for any \mathbf{X} as

$$\left| \frac{\partial u_{l+1,n}}{\partial \mathbf{X}_i}(\mathbf{X}) \right| \leq C_{g,4}, \quad \left| \frac{\partial^2 u_{l+1,n}}{\partial \mathbf{X}_i \partial \mathbf{X}_j}(\mathbf{X}) \right| \leq C_{g,4}, \quad \left| \frac{\partial^3 u_{l+1,n}}{\partial \mathbf{X}_i \partial \mathbf{X}_j \partial \mathbf{X}_k}(\mathbf{X}) \right| \leq C_{g,4}. \quad (9)$$

1106 For the term $I_{1,1}$ and $I_{1,2}$, by applying Lemma F.16, we have that

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

1107 Next, we bound the remainders \mathcal{R} and $\tilde{\mathcal{R}}$. By Cauchy-Schwarz inequality,

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

1108 where the last inequality uses Lemma F.16 and Equation (9).

1109 Similarly, we can prove that there exists a positive constant $C_{g,5}$ such that

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

1110 Combining the bounds for I_1 , I_2 , and I_3 gives the lemma. \square

1111 Finally, we are ready to prove Theorem F.2.

1112 *Proof of Theorem F.2.* For $0 \leq l \leq n = \lfloor \frac{T}{\eta^{0.75}} \rfloor$, we denote the random variable by $\hat{\mathbf{x}}_{l,n}$ such that
 1113 follows a distribution $\mathcal{P}_{\mathbf{X}}(\hat{\mathbf{X}}^{(lR_{\text{grp}})}, l\eta_e, n\eta_e)$. When we set $l = n$, $\mathcal{P}(\hat{\mathbf{x}}_{n,n} = \hat{\mathbf{X}}^{(nR_{\text{grp}})})$ and setting
 1114 $l = 0$ gives $\hat{\mathbf{x}}_{0,n} \sim \mathbf{X}(n\eta_e)$. Recall the previous definition that $u(\mathbf{x}, s, t) = \mathbb{E}_{\mathbf{X}_t \sim \mathcal{P}_{\mathbf{X}}(\mathbf{x}, s, t)}[g(\mathbf{X}_t)]$,
 1115 and we define that $\mathcal{T}_{l+1,n} := u_{l+1,n}(\hat{\mathbf{X}}^{(lR_{\text{grp}})} + \Delta^{(l)}, (l+1)\eta_e, n\eta_e) - u_{l+1,n}(\hat{\mathbf{X}}^{(lR_{\text{grp}})} + \tilde{\Delta}^{(l)}, (l+1)\eta_e, n\eta_e)$. Using the definition of $\mathbf{x}_{l,n}$, we can rewrite the distance between AGMs and SDE
 1116 measured by a test function g as

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

1118 The above equation uses the law of total expectation and the definition of δ -good event $\mathcal{E}_0^{(nR_{\text{grp}})}$ in
 1119 Definition F.1. Then the Triangle inequality gives

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

1120 where the second last inequality uses Lemma F.17. Recap that $a_1 = \frac{1.5-2\beta}{1-\beta}$, $a_2 = \frac{1}{1-\beta}$, $\beta \in (0, 0.5)$.

1121 Let $\beta = 0.25$, and we complete the proof. \square

1122 G Proof of Theorems in Section 5

1123 G.1 Proof of Adam and AdamE- λ 's Implicit Biases with Label Noise

1124 In this part, we give the proof of Theorem 5.1 and Theorem 5.2.

1125 *Proof of Theorem 5.1.* With label noise, the gradient covariance matrix $\Sigma(\zeta) = \alpha \nabla^2 \mathcal{L}(\zeta)$ for any
1126 $\zeta \in \Gamma$.

1127 We recall the SDE formula in Equation (5) and Lemma F.9

$$\begin{cases} d\zeta(t) = \partial \Phi_{S(v)}(\zeta) S(v) \Sigma^{1/2}(\zeta) dW_t - \frac{1}{2} S_t \partial^2(\nabla \mathcal{L})(\zeta) [P \mathcal{V}_{\nabla^2 \mathcal{L}'(\phi'_{(0)})} (P \Sigma_0 P) P] dt, \\ dv(t) = c(V(\Sigma(\zeta)) - v) dt. \end{cases}$$

1128 And $P := S_0^{1/2}$. By applying Lemma F.2, and replacing Σ with $\alpha \nabla^2 \mathcal{L}$, we can reduce the SDE
1129 formula as

$$\begin{cases} d\zeta(t) = -\frac{\alpha}{2} S_t \partial^2(\nabla \mathcal{L})(\zeta) [S_t] dt, \\ dv(t) = c(V(\Sigma(\zeta)) - v) dt. \end{cases}$$

1130 Then we can write out the constraint for the fixed point (ζ^*, v^*) of this ODE as

$$-\frac{1}{2} S(v^*) \partial^2(\nabla \mathcal{L})(\zeta^*) [S(v^*)] = 0, \quad (10)$$

$$V(\Sigma(\zeta^*)) - v^* = 0. \quad (11)$$

1131 Solving Equation (10) and Equation (11) gives

$$\partial^2(\nabla \mathcal{L})(\zeta^*) [S(V(\Sigma(\zeta^*)))] = 0. \quad (12)$$

1132 Integrating by parts gives us

$$\partial^2(\nabla \mathcal{L})[S] = \nabla [\langle \nabla^2 \mathcal{L}, S \rangle] - \nabla(S) [\nabla^2 \mathcal{L}]. \quad (13)$$

1133 We use H and $\nabla^2 \mathcal{L}$ interchangeably to denote the Hessian matrix. For the first term, note that

$$\begin{aligned} \langle S, H \rangle &= \sum_{j,k} [S]_{jk} H_{jk} \\ &= \sum_{i,j,k} P_{ji} H_{jk} P_{ki} \\ &= \sum_i [PHP]_{ii} \\ &= \text{tr}(PHP) \\ &= \text{tr}\left((\text{Diag } H)^{\frac{1}{2}}\right) + O(\epsilon \text{tr}(H)), \end{aligned}$$

1134 where the last equality comes from the update rule of Adam: $S = (\text{Diag } H)^{-\frac{1}{2}} + O(\epsilon)$, $P =$
1135 $(\text{Diag } H)^{-\frac{1}{4}} + O(\sqrt{\epsilon})$. For the second term, we also plug in the update rule of Adam, and use h_j to
1136 denote H_{jj} , which turns out to be the gradient of the same thing:

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

1137 Summarizing, our drift term can be represented as a constant multiple of

$$S \nabla \text{tr} \left((\text{Diag} \mathbf{H})^{\frac{1}{2}} \right) + O(S \epsilon \text{tr}(\mathbf{H})), \quad (14)$$

1138 forming a preconditioned gradient flow that implicitly minimizes $\text{tr} \left((\text{Diag} \mathbf{H})^{\frac{1}{2}} \right)$ when $\epsilon \rightarrow 0$.
 1139 Combining Equation (12) and Eq. (14) gives the result in Theorem 5.1. \square

1140 *Proof of Theorem 5.2.* Now we consider the optimizer AdamE- λ , the variant of Adam proposed as a
 1141 verification case of our main results, whose update rule is

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

1142 Now for AdamE- λ , the precondition matrix $\mathbf{S} = (\text{Diag} \mathbf{H})^{-\lambda}$, and $\mathbf{P} = (\text{Diag} \mathbf{H})^{-\lambda/2}$, which gives

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

1143 Also, similar to the case for Adam, we have

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

1144 Utilizaing Equation (13), we get the regularization term for AdamE- λ as

$$\begin{aligned} \partial^2(\nabla \mathcal{L})[\mathbf{S}] &= \nabla[\langle \nabla^2 \mathcal{L}, \mathbf{S} \rangle] - \nabla(\mathbf{S}) [\nabla^2 \mathcal{L}] \\ &= \frac{1}{1-\lambda} \text{tr} \left((\text{Diag} \mathbf{H})^{1-\lambda} \right) + O(\epsilon \text{tr}(\mathbf{H})), \end{aligned}$$

1145 Evaluating the above equation when $\epsilon \rightarrow 0$ completes the proof. \square

1146 G.2 Proof of Lemma 5.1

1147 *Proof.* Recall that the manifold is defined as $\Gamma = \{ \boldsymbol{\theta} | \langle \mathbf{z}_i, \mathbf{u}^{\odot 2} - \mathbf{v}^{\odot 2} \rangle = y_i, \forall i \in [n] \}$. So if any
 1148 $\boldsymbol{\theta} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$ belongs to Γ , and another $\tilde{\boldsymbol{\theta}} = \begin{pmatrix} \tilde{\mathbf{u}} \\ \tilde{\mathbf{v}} \end{pmatrix}$ satisfies that $\tilde{u}_i^{\odot 2} - \tilde{v}_i^{\odot 2} = u_i^{\odot 2} - v_i^{\odot 2}$ for any $i \in [d]$,
 1149 then $\tilde{\boldsymbol{\theta}}$ also belongs to Γ .

1150 Next, we derive the explicit expression of the Hessian matrix when $\boldsymbol{\theta} \in \Gamma$:

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

Hence, we have that

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

and $\|\mathbf{u}^{\odot 2} - \mathbf{v}^{\odot 2}\|_{e_0}^{e_0} = \sum_{i=1}^d |u_i^2 - v_i^2|^{e_0}$. Let $e_0 \in (0, 1]$, and we assume that

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

1151 First we prove by contradiction that $u_i = 0$ or $v_i = 0$ for any $i \in [d]$. If there exists some i
 1152 such that $u_i \neq 0$ and $v_i \neq 0$, then denote $s = \min\{|u_i|, |v_i|\}$, we construct $\tilde{\boldsymbol{\theta}} = \begin{pmatrix} \tilde{\mathbf{u}} \\ \tilde{\mathbf{v}} \end{pmatrix}$ by letting
 1153 $\tilde{u}_j = u_j, \tilde{v}_j = v_j$ for $j \neq i$ and $\tilde{u}_i = |u_i| - s, \tilde{v}_i = |v_i| - s$. Then $\tilde{\mathbf{u}}^{\odot 2} - \tilde{\mathbf{v}}^{\odot 2} = \mathbf{u}^{\odot 2} - \mathbf{v}^{\odot 2}$, so
 1154 $\tilde{\boldsymbol{\theta}} \in \Gamma$, but $|\tilde{u}_i|^{2e_0} + |\tilde{v}_i|^{2e_0} < |u_i|^{2e_0} + |v_i|^{2e_0}$, a contradiction.

1155 Now assume $\boldsymbol{\theta} \notin \arg \min_{\boldsymbol{\theta}' \in \Gamma} \|\mathbf{u}^{\odot 2} - \mathbf{v}^{\odot 2}\|_{e_0}$. There must exist some $\tilde{\boldsymbol{\theta}} \in \Gamma$ such that
 1156 $\|\tilde{\mathbf{u}}^{\odot 2} - \tilde{\mathbf{v}}^{\odot 2}\|_{e_0} < \|\mathbf{u}^{\odot 2} - \mathbf{v}^{\odot 2}\|_{e_0}$. WLOG assume for any $i \in [d]$, either $\tilde{u}_i = 0$ or $\tilde{v}_i = 0$,
 1157 else we can construct another minimizer that preserves $\|\tilde{\mathbf{u}}^{\odot 2} - \tilde{\mathbf{v}}^{\odot 2}\|_{e_0}$ as above. But now we have
 1158 $\sum_{i=1}^d |u_i^2 - v_i^2|^{e_0} = \sum_{i=1}^d |u_i|^{2e_0} + |v_i|^{2e_0}$, 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
 1159 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