HiLD 2024: 2nd Workshop on High-dimensional Learning Dynamics 1–58

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

Xinghan Li* XH-LI22@MAILS.TSINGHUA.EDU.CN
Institute for Interdisciplinary Information Sciences, Tsinghua University

Haodong Wen*† HERRYWENH@GMAIL.COM
Institute for Interdisciplinary Information Sciences, Tsinghua University

Kaifeng Lyu‡ KLYU@MAIL.TSINGHUA.EDU.CN
Institute for Interdisciplinary Information Sciences, Tsinghua University

Abstract

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

1. Introduction

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

* Equal contribution; alphabet ordering.
† Most work was done while Haodong was at Xi'an Jiaotong University.
‡ Corresponding author.

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

All these implicit bias characterizations hold, or can be readily extended, to the stochastic variant of gradient descent, *i.e.*, Stochastic Gradient Descent (SGD). Another line of works (Blanc et al., 2020; Damian et al., 2021; Li et al., 2021b) demonstrated that the gradient noise in SGD induces an additional form of implicit bias that reduces the *sharpness* of the solutions, a generalization measure that has been long observed to correlate with generalization (Hochreiter and Schmidhuber, 1997; Keskar et al., 2017; Jiang et al., 2020; Foret et al., 2021). More specifically, these works focus on the dynamics of SGD when the training loss is already small and the iterates are close to a manifold of minimizers. Li et al. (2021b) introduced a general framework to analyze the dynamics of SGD near the minimizer manifold, showing that SGD will not stop at arbitrary global minimizers, but drift and diffuse around the manifold, driving the iterates towards flatter regions of the loss landscape. This behavior is mathematically characterized by a Stochastic Differential Equation (SDE), termed as *slow SDE* (Gu et al., 2023a), which accurately tracks the projected dynamics of SGD near the minimizer manifold over a timescale of $\mathcal{O}(\eta^{-2})$. The resulting dynamics reveal that SGD behaves like a gradient method on the manifold that takes semi-gradients to minimize a specific sharpness measure determined by the Hessian and gradient noise. See Appendix B 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 2, 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 Appendix D, we prove theoretically the generalization benefit of Adam under label noise settings. We show that under label noise setting, the implicit regularizer of Adam will reduce to $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{1/2})$ where $\boldsymbol{H}$ is the Hessian matrix. Compared to the $\mathrm{tr}(\boldsymbol{H})$ of SGD, this new kind of sharpness reduction usually aligns better with sparsity regularization, thus utilizing data more efficiently when the model is required to fit a sparse ground truth. We verify this anticipation experimentally through the diagonal net setting (Woodworth et al., 2020). We also demonstrated the discrepancy of the implicit biases of Adam and SGD through the matrix factorization setting in Appendix E.

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 $\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^*$.

2. 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. We provide a more detailed theoretical background in Appendix B, and readers can refer to it if needed.

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

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

where $S : \mathbb{R}^d \longrightarrow \mathbb{R}^{d\times d}$ is ρ_s-smooth, positive definite and satisfies $S(\boldsymbol{v}) \preceq \epsilon^{-1} I$ for some $\epsilon > 0$ and any $\boldsymbol{v} \in \mathbb{R}^d$, and $V : \mathbb{R}^{d\times d} \longrightarrow \mathbb{R}^d$ is linear. A number of currently used optimization algorithms, such as RMSProp, Adam, Adam-mini, Adafactor[1], 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 Appendix D as a tool to tune the implicit bias of Adam.

2.1. Slow SDE Analysis for AGMs

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

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

Definition 2.1 (Preconditioner Flow Projection) *Fix a point $\theta_{null} \notin \Gamma$. Given a Positive Semi-Definite matrix $\boldsymbol{S}$. For $x \in \mathbb{R}^d$, consider the preconditioner flow $\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\boldsymbol{S}\nabla\mathcal{L}(x(t))$ with $x(0) = x$. We denote the preconditioner flow projection of x as $\Phi_{\boldsymbol{S}}(x)$, i.e. $\Phi_{\boldsymbol{S}}(x) := \lim_{t\to+\infty} x(t)$ if the limit exists and belongs to Γ, and $\Phi_{\boldsymbol{S}}(x) = \theta_{null}$ otherwise.*

Definition 2.2 *For any $\boldsymbol{\zeta} \in \Gamma$ and any differential form $\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}\mathrm{d}t$ in Itô calculus, where $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, and $b \in \mathbb{R}^d$. We use $\boldsymbol{P}_{\boldsymbol{\zeta},\boldsymbol{S}}(\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}\mathrm{d}t)$ as a shorthand for the differential form $\partial\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{S}\left(\partial\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})\boldsymbol{b} + \frac{1}{2}\partial^2\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})[\boldsymbol{A}\boldsymbol{A}^T]\right)\mathrm{d}t$.*

Definition 2.3 (Slow SDE for AGMs) *given learning rate η, $\frac{1-\beta_2}{\eta^2} = c$, $\boldsymbol{v}_0 \in \mathbb{R}^d$, $\boldsymbol{S}_t := S(\boldsymbol{v}(t))$, and $\boldsymbol{\zeta}_0 \in \Gamma$, $\boldsymbol{v}_0 \in \mathbb{R}^d$, we define $\boldsymbol{\zeta}(t)$ as the solution of the following SDE with initial point $(\boldsymbol{\zeta}(0), \boldsymbol{v}(0)) = (\boldsymbol{\zeta}_0, \boldsymbol{v}_0)$:*

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

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

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

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

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

We could prove that $\boldsymbol{\zeta}(t)$ always stays on the manifold Γ. And next, we present our main theorem, and show that the above SDE in Equation (1) track the trajectory of Adam in a weak approximation sense.

Assumption 2.1 *The loss function $\mathcal{L}(\cdot)$ and the matrix square root of the noise covariance $\boldsymbol{\Sigma}^{1/2}(\cdot)$ are $\mathcal{C}^\infty$-smooth. Besides, we assume that $\|\nabla\ell(\boldsymbol{\theta};\xi)\|_2$ is bounded by a constant for all $\boldsymbol{\theta}$ and ξ.*

Assumption 2.2 *Γ is a compact manifold.*

Theorem 2.1 *Let Assumptions B.1, 2.1 and 2.2 hold. Let $T > 0$ be a constant and let $\boldsymbol{X}(t) = (\boldsymbol{\zeta}(t), \boldsymbol{v}(t))$ be the solution to Equation (1) with initial condition:*

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

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

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

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

Theorem 2.1 shows that, in the small η regime, once Adam approaches the minimizer manifold, its long-horizon behavior within $\widetilde{\mathcal{O}}(\frac{1}{\eta^2})$ steps can be well approximated by the SDE defined in Equation (1).

2.2. Interpretation of The Slow SDEs for AGMs

Adaptive Projection Operator. Whereas Equation (2) employs a fixed projection operator $P_{\boldsymbol{\zeta}}$ to constrain the SDE to the manifold, the AGM slow–SDE uses an adaptive projection $P_{\boldsymbol{\zeta},\boldsymbol{S}(t)}$ that depends on the current preconditioner $S(\boldsymbol{v}(t))$. In other words, SGD's projection is 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 $\boldsymbol{\Sigma}$ is filtered through the preconditioner $\boldsymbol{S}(t)$ into $\boldsymbol{S}(t)\boldsymbol{\Sigma}\boldsymbol{S}(t)$ and then contributes to the SDE. Over a long time horizon, this modified noise term alters the deterministic drift direction, further distinguishing AGM's dynamics from those of vanilla SGD.

2.3. Technical Difficulties and Proof Insights

2.3.1. 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 2.2 (Convergence Bound of the AGM Framework, Stated Informally) *Under mild assumptions, for any* $\delta \in (0,1)$, *there exists some* $K = \mathcal{O}\left(\frac{1}{\eta}\log\frac{1}{\eta}\right)$, *such that* $\mathcal{L}(\boldsymbol{\theta}_K) - \mathcal{L}^* = \tilde{\mathcal{O}}(\eta)$ *holds with probability at least* $1-\delta$.

2.3.2. 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 $S(\boldsymbol{v}_k)$, the rotation equivariance is not satisfied here.

To resolve this, we generalize the gradient flow projection in Gu et al. (2023a); Li et al. (2021b) into a varying preconditioner flow projection. Utilizing this definition, after doing a reperemeterization 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.

3. Discussion

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

References

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

David GT Barrett and Benoit Dherin. Implicit gradient regularization. *arXiv preprint arXiv:2009.11162*, 2020.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

David Brandfonbrener and Joan Bruna. Geometric insights into the convergence of nonlinear td learning. *arXiv preprint arXiv:1905.12185*, 2019.

Matias D. Cattaneo and Boris Shigida. How memory in optimization algorithms implicitly modifies the loss, 2025. URL https://arxiv.org/abs/2502.02132.

Matias D. Cattaneo, Jason M. Klusowski, and Boris Shigida. On the implicit bias of adam, 2024. URL https://arxiv.org/abs/2309.00079.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

Steffen Dereich and Arnulf Jentzen. Convergence rates for the adam optimizer. *arXiv preprint arXiv:2407.21078*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Aijun Du and Jinqiao Duan. Invariant manifold reduction for stochastic dynamical systems. *arXiv preprint math/0607366*, 2006.

Johannes Jisse Duistermaat and Johan AC Kolk. *Lie groups*. Springer Science & Business Media, 2012.

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

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

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

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns, 2018. URL https://arxiv.org/abs/1802.10026.

Khashayar Gatmiry, Zhiyuan Li, Tengyu Ma, Sashank Reddi, Stefanie Jegelka, and Ching-Yao Chuang. What is the inductive bias of flatness regularization? a study of deep matrix factorization models. *Advances in Neural Information Processing Systems*, 36:28040–28052, 2023.

Xinran Gu, Kaifeng Lyu, Longbo Huang, and Sanjeev Arora. Why (and when) does local sgd generalize better than sgd? *arXiv preprint arXiv:2303.01215*, 2023a.

Xinran Gu, Kaifeng Lyu, Longbo Huang, and Sanjeev Arora. Why (and when) does local sgd generalize better than sgd?, 2023b. URL https://arxiv.org/abs/2303.01215.

Xinran Gu, Kaifeng Lyu, Sanjeev Arora, Jingzhao Zhang, and Longbo Huang. A quadratic synchronization rule for distributed deep learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=yroyhkhWS6.

Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. Unified convergence analysis for adaptive optimization with moving average estimator, 2025. URL https://arxiv.org/abs/2104.14840.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 01 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.1.1. URL https://doi.org/10.1162/neco.1997.9.1.1.

Yusu Hong and Junhong Lin. High probability convergence of adam under unbounded gradients and affine variance noise, 2023. URL https://arxiv.org/abs/2311.02000.

Hideaki Iiduka. Theoretical analysis of adam using hyperparameters close to one without lipschitz smoothness, 2022. URL https://arxiv.org/abs/2206.13290.

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

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

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

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17176–17186. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c76e4b2fa54f8506719a5c0dc14c2eb9-Paper.pdf.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Rong Ge, and Sanjeev Arora. Explaining landscape connectivity of low-cost solutions for multilayer nets. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/46a4378f835dc8040c8057beb6a2da52-Paper.pdf.

Qianxiao Li, Cheng Tai, et al. Dynamics of stochastic gradient algorithms. *CoRR*, 2015.

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

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations, 2018. URL https://arxiv.org/abs/1811.01558.

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

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12712–12725. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/69f629564298659092 1fa916d61c1f8-Paper.pdf.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021b.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22188–22214. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liu23ao.html.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeLIgBKPS.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35:34689–34708, 2022.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S Du, Jason D Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

Chao Ma, Daniel Kunin, and Lexing Ying. A quasistatic derivation of optimization algorithms' exploration on minima manifolds, 2023. URL https://openreview.net/forum?id=UDbNL0_W-3x.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scaling rules for adaptive gradient algorithms. *Advances in Neural Information Processing Systems*, 35:7697–7711, 2022.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scaling rules for adaptive gradient algorithms, 2024. URL https://arxiv.org/abs/2205.10287.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Mor Shpigel Nacson, Suriya Gunasekar, Jason Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4683–4692. PMLR, 09–15 Jun 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

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

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

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

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

Nikunj Saunshi, Stefani Karp, Shankar Krishnan, Sobhan Miryoosefi, Sashank J. Reddi, and Sanjiv Kumar. On the inductive bias of stacking towards improving reasoning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 71437–71464. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/837bc5db12f3d394d220815a7687340c-Paper-Conference.pdf.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Naichen Shi and Dawei Li. Rmsprop converges with proper hyperparameter. In *International conference on learning representation*, 2021.

Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9058–9067. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/smith20a.html.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pages 10849–10858. PMLR, 2021.

Bohan Wang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, and Wei Chen. On the convergence of adam under non-uniform smoothness: Separability from sgdm and beyond, 2024a. URL https://arxiv.org/abs/2403.15146.

Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 2960–2969, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671718. URL https://doi.org/10.1145/3637528.3671718.

Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate sgd. *arXiv preprint arXiv:2307.15196*, 2023.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective, 2024. URL https://arxiv.org/abs/2410.05192.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞-norm constrained optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54488–54510. PMLR, 21–27 Jul 2024.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/90365351ccc7437a1309dc64e4db32a3-Paper.pdf.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data, 2024a. URL https://arxiv.org/abs/2406.10650.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Qi Zhang, Yi Zhou, and Shaofeng Zou. Convergence guarantees for rmsprop and adam in generalized-smooth non-convex optimization with affine noise variance. *arXiv preprint arXiv:2404.01436*, 2024b.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. *Advances in neural information processing systems*, 35:28386–28399, 2022.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective, 2024c. URL https://arxiv.org/abs/2402.16788.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11127–11135, 2019.

Contents

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

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

Appendix A. 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.

Appendix B. Preliminaries

Notations. Unless otherwise stated, for a square matrix $\boldsymbol{M}$, $\mathrm{diag}(\boldsymbol{M})$ denotes the vector consisting of its diagonal entries. The notation Diag has two usages: For a vector $\boldsymbol{v}$, $\mathrm{Diag}(\boldsymbol{v})$ denotes the diagonal matrix with $\boldsymbol{v}$ on its diagonal; and for a square matrix $\boldsymbol{M}$, $\mathrm{Diag}(\boldsymbol{M})$ denotes the diagonal matrix that only keeps $\boldsymbol{M}$'s diagonal entries and equals 0 elsewhere, i.e. $\mathrm{Diag}(\boldsymbol{M}) \overset{\text{def}}{=} \mathrm{Diag}(\mathrm{diag}(\boldsymbol{M}))$. For two vectors $\boldsymbol{u}$, $\boldsymbol{v}$ with the same dimension d, $\boldsymbol{u}\odot\boldsymbol{v}$ denotes element-wise multiplication $(u_1v_1, \ldots, u_dv_d)$. For any exponent p, $\boldsymbol{v}^{\odot p}$ denotes element-wise exponentiation, i.e. $\boldsymbol{v}^{\odot p} = (v_1^p, \ldots, v_d^p)$, and $\sqrt{\boldsymbol{v}}$ means $\boldsymbol{v}^{\odot 1/2}$.

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

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

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

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

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

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

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

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

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

Assumption B.1 *Γ is $\mathcal{C}^\infty$-smooth, $(d-m)$-dimensional submanifold of $\mathbb{R}^d$, where any $\zeta \in \Gamma$ is a local minimizer of $\mathcal{L}$. For all $\boldsymbol{\zeta} \in \Gamma$, rank$(\nabla^2 \mathcal{L}(\boldsymbol{\zeta})) = m$. Additionally, there exists an open neighborhood of Γ, denoted as U, such that $\Gamma = \arg\min_{\boldsymbol{\theta} \in U} \mathcal{L}(\boldsymbol{\theta})$.*

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

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

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

Definition B.1 (Slow SDE for SGD) *Given $\eta > 0$ and $\boldsymbol{\zeta}_0 \in \Gamma$, define $\boldsymbol{\zeta}(t)$ as the solution of the following SDE with initial condition $\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0$:*

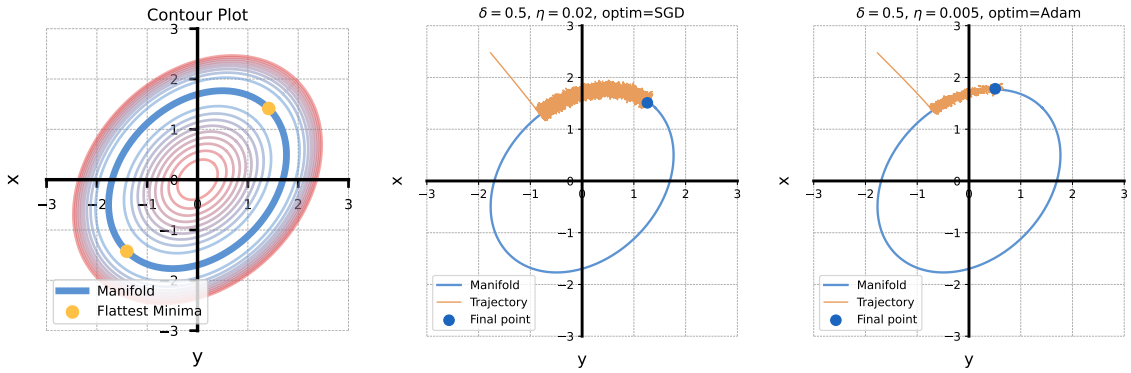
$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}}\Big(\underbrace{\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}_t}_{\text{(a) diffusion}} - \underbrace{\frac{1}{2}\nabla^3 \mathcal{L}(\boldsymbol{\zeta})\left[\widehat{\boldsymbol{\Sigma}}_{\diamondsuit}(\boldsymbol{\zeta})\right]\mathrm{d}t}_{\text{(b)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

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

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

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

Here we use the word "semi-gradient" (Mnih et al., 2015; Brandfonbrener and Bruna, 2019) because it is not exactly the gradient of $\mu(\boldsymbol{\zeta})$ but only the gradient with respect to the first



(a) (b) (c)

Figure 1: **(a)** Contour of the elliptical loss, highlighting the two flattest minima. **(b)** SGD implicitly minimizes $\mathrm{tr}(\boldsymbol{H})$ and converges to a flattest minimum. **(c)** Adam also reduces sharpness but finds a different, sparser minimum.

argument of the inner product. More specifically, define $\mu(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) := \left\langle \nabla^2 \mathcal{L}(\boldsymbol{\zeta}_1), \widehat{\boldsymbol{\Sigma}}_\diamondsuit(\boldsymbol{\zeta}_2) \right\rangle$, then the drift term is essentially $-\frac{1}{2} \nabla_{\boldsymbol{\zeta}_1} \mu(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2)|_{\boldsymbol{\zeta}_1=\boldsymbol{\zeta}, \boldsymbol{\zeta}_2=\boldsymbol{\zeta}}$ after projecting onto the tangent space of Γ at $\boldsymbol{\zeta}$.

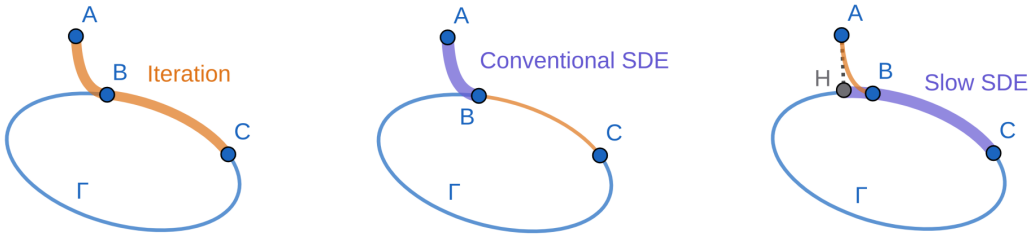
In other words, SGD near manifold takes semi-gradients to minimize the implicit regularizer $\langle \nabla^2 \mathcal{L}(\boldsymbol{\zeta}), \widehat{\boldsymbol{\Sigma}}_\diamondsuit(\boldsymbol{\zeta}) \rangle$ but pretend $\widehat{\boldsymbol{\Sigma}}_\diamondsuit(\boldsymbol{\zeta})$ to be fixed, i.e. ignore the dependency of $\widehat{\boldsymbol{\Sigma}}_\diamondsuit(\boldsymbol{\zeta})$ on $\boldsymbol{\zeta}$.

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

Appendix C. Illustration of the Difference between Conventional SDE and Slow SDE

In this section, we illustrate the difference between conventional SDE and slow SDE. In Fig. 2, let Γ denotes a 1D manifold, then the discrete iteration of the optimization process can be seen as successive steps (orange, Fig. 2(a)) that starts from A, first converge to some point B in Γ and then move along Γ to C.

The main intuition behind slow SDE is that the whole process $A \to B \to C$ can actually be decomposed into two motions: a convergence motion $A \to H$ (dashed, Fig. 2(c)) and an implicit regularization motion $H \to B \to C$. The convergence motion is fast and dominates the dymanics during the convergence phase, but it fades out as soon as convergence phase ends; meanwhile the slow, implicit regularization motion starts to dominate.



(a) Iteration (b) Conventional SDE (c) Slow SDE

Figure 2: Comparison of conventional SDE and slow SDE.

The conventional SDE approximates the convergence phase only, whose unit time corresponds to $\tilde{O}(\eta^{-1})$ steps (Fig. 2(b)). In contrast, slow SDE manages to separate the slow implicit regularization motion from the fast convergence, and approximate the implicit regularization near manifold only (Fig. 2(c)).

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

Appendix D. 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.

D.1. Reduction of Adam's Implicit Regularizer with Label Noise

On an ℓ_2-regression task on dataset $\{\boldsymbol{z}_i, y_i\}_{i=1}^n$, 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 $\boldsymbol{\theta} \in \Gamma$, $\boldsymbol{\Sigma} \equiv \alpha\nabla^2\mathcal{L}$ 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 D.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 $\nabla tr(Diag(\boldsymbol{H})^{1/2}) = 0$.*

Proof Sketch. Under label noise setting, SDE Eq. (1) will be greatly simplified. In fact, the diffusion term of the slow SDE would equal zero, and the drift term could be simplified to

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

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

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 $S(\boldsymbol{v}) = \operatorname{Diag}(1/(\boldsymbol{v}^{\odot\lambda} + \epsilon))$. 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 D.1, and the result is stated below.

Theorem D.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 $\nabla tr(Diag(\boldsymbol{H})^{1-\lambda}) = 0$.*

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

D.2. Example: Sparse Linear Regression with Diagonal Net

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

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

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

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

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

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

This is because only $\widehat{\boldsymbol{w}} = \boldsymbol{u}^{\odot 2}-\boldsymbol{v}^{\odot 2}$ 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 $\operatorname{tr}(\operatorname{Diag}(\boldsymbol{H})^{e_0})$. Thus $u_i=0$ or $v_i=0$ for any i. When this holds, we have $\operatorname{tr}(\operatorname{Diag}(\boldsymbol{H})^{e_0}) = 4^{e_0}\|\boldsymbol{\theta}^{\odot 2e_0}\|_1 = \left(4\left\|\widehat{\boldsymbol{w}}\right\|_{e_0}\right)^{e_0}$. Therefore, implicitly regularizing $\operatorname{tr}(\operatorname{Diag}(\boldsymbol{H})^{e_0})$ 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 advantange 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.

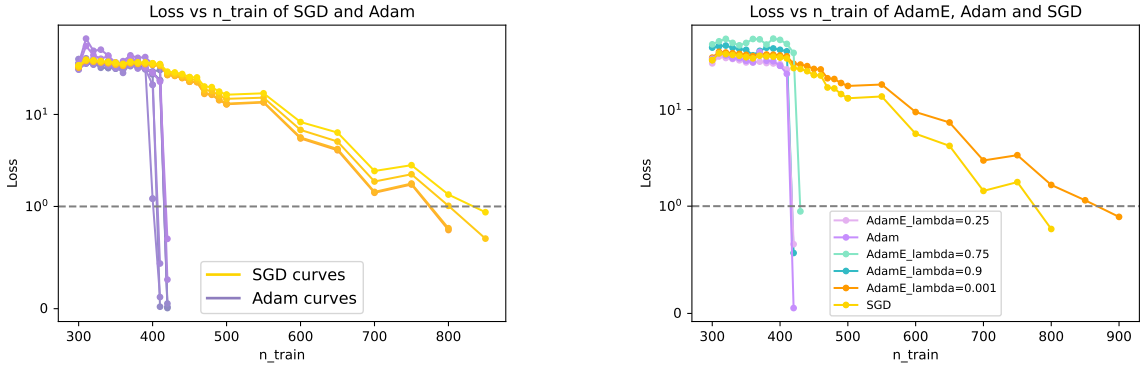
D.2.1. Result: Adam's Implicit Regularizer Facilitates Sparse Ground-truth Recovery

We plot the results of the experiment in Fig. 3. 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. 3(a), 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. 3(b) 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.

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.

Appendix E. Matrix Factorization: Adam Implicitly Regularizes Sharpness Differently

The diagonal-net experiments in Appendix D showed that Adam's implicit bias towards *sparsity* improves generalization relative to SGD. We now turn to supply the potentially negative impact of Adam's implicit bias in another controlled setting: **deep matrix factorization with label noise**, where the relevant implicit regularizers are analytically tractable. In this task, Adam is expected to minimize $\operatorname{tr}(\operatorname{Diag}(\boldsymbol{H})^{1/2})$ rather than $\operatorname{tr}(\boldsymbol{H})$. Leveraging existing theory, we therefore predict that (i) Adam will converge to a solution with $\operatorname{tr}(\boldsymbol{H})$



(a) (b)

Figure 3: 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 varying learning rates and β_2 values. **(b)** Loss comparison between AdamE ($\lambda = 0.001, 0.25, 0.75, 0.9$), Adam, and SGD.

larger—but $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{1/2})$ smaller—than SGD's solution, and (ii) once training reaches the interpolation regime, Adam will *generalize worse* than vanilla SGD in the presence of label noise. Our experiments confirm both predictions (Figure 4).

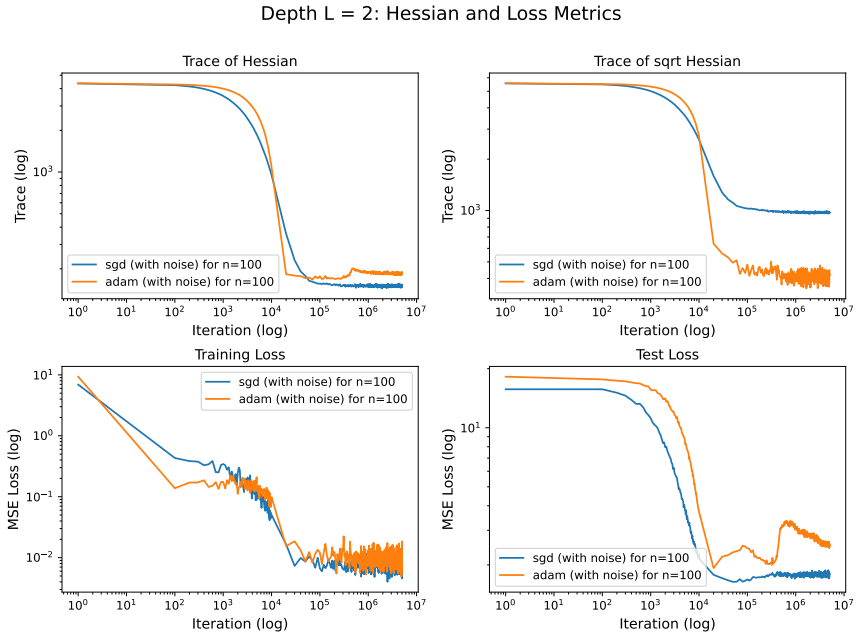


Figure 4: **Deep matrix factorization with label noise.** Adam and SGD are trained on identical data and noise realizations. *Top:* evolution of $\mathrm{tr}(\boldsymbol{H})$ and $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{1/2})$. *Bottom:* training and test MSE. Adam converges to a point with larger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

E.1. Problem setup

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

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

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

Implicit regularization. With small learning rates and additive label noise, SGD asymptotically minimizes $\operatorname{tr}(\boldsymbol{H})$ once it reaches the zero–loss manifold. In matrix factorization, minimizing $\operatorname{tr}(\boldsymbol{H})$ 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 $\boldsymbol{M}^*$ is low rank. Adam, however, implicitly minimizes $\operatorname{tr}(\operatorname{Diag}(\boldsymbol{H})^{1/2})$; it therefore converges to a different point, typically with larger $\operatorname{tr}(\boldsymbol{H})$ and reduced generalization.

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

Takeaway. In deep matrix factorization with label noise, Adam's preference for minimizing $\operatorname{tr}(\operatorname{Diag}(\boldsymbol{H})^{1/2})$ leads to less generalizable—solutions than SGD, reinforcing that Adam's implicit regularization differs qualitatively from SGD's and can hurt performance when $\operatorname{tr}(\boldsymbol{H})$ matters.

Appendix F. Formal Statements of the Main Results

In this section, we give the formal versions of the main results stated in Section 2, where we presented the two principal theorems:

1. The AGM iterates converge to a neighborhood of the manifold (Theorem 2.2);
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 2.1).

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

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

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

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

Assumption F.2 *$\mathcal{L}$ is ρ-smooth on $\mathbb{R}^d$, i.e. $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, $\|\nabla\mathcal{L}(\boldsymbol{\theta}_1) - \nabla\mathcal{L}(\boldsymbol{\theta}_2)\|_2 \le \rho\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ and $\mathcal{L}$ is bounded from below, i.e. $\mathcal{L}^* = inf_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) > -\infty$.*

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

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

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

Assumption F.6 $\beta_1 \le 0.9$.

Remark. The threshold 0.9 in Assumption F.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-word 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 F.1 *Let Assumptions F.2, F.3 and F.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 $\boldsymbol{\theta}_0$ 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 $K_0 = \mathcal{O}(\frac{1}{\eta}\log\frac{1}{\eta})$:*

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

2. *(Formal restatement of Theorem 2.1: Slow SDE tracks AGM's trajectory in a weak approximation sense) Moreover, when Assumptions B.1, 2.2, F.1, F.4 and F.5 hold, we shift the timeline and redefine the final state $(\boldsymbol{\theta}_{K_0}, \boldsymbol{v}_{K_0})$ in conclusion 1 by $(\boldsymbol{\theta}_0, \boldsymbol{v}_0)$. Let*

1. The 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.

$T > 0$ be a constant, $\boldsymbol{X}(t) = (\boldsymbol{\zeta}(t), \boldsymbol{v}(t))$ be the solution to Equation (1) with initial condition:

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

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

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

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

F.1. Convergence Guarantee of AGMs

In the proof, the first part of Theorem F.1 is done by first proving a convergence result with global μ-PL condition, and then arguing that AGM starting near enough to the manifold will stick to the manifold with high probability. As mentioned in Section 2.3.1, the convergence under μ-PL condition can be seen as a separate technical contribution of our paper, which is stated below.

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

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

Theorem F.2 (Formal restatement of Theorem 2.2) *Let Assumptions F.2, F.3 and F.6 be satisfied, and $\mathcal{L}$ satisfy the μ-PL condition. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following holds for some $K = \mathcal{O}(\frac{1}{\eta} \log \frac{1}{\eta})$:*

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

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

Appendix G. Constructing the Working Zones

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

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

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

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

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

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

Then for any $t < T$ we have

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

Integrating both sides gives us

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

The above equations complete the proof. ■

Since the gradients and gradient noises are assumed to be bounded and S and V both satisfy Lipschitz-ness, we deduce that any $\boldsymbol{v}$ produced in the iteration is bounded. Specifically, there exists some constant R_1 such that $\boldsymbol{v}_k \leq R_1$ for any $k \geq 0$. For all algorithms listed

in Table 1, setting $R_1 = R^2$ is sufficient. Similarly, all the outputs of S are also bounded, and we denote R_0 as a constant satisfying $S^{-1}(\boldsymbol{v}) \preceq R_0$ for all $\boldsymbol{v}$ in the iteration. Note that S is Lipschitz on $\{\boldsymbol{v} : 0 \leq \boldsymbol{v} \leq R_1\}$ from this assumption, since it is a compact set, and the derivative of S is bounded.

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

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

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

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

View $\Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})$ as a function defined on support $\mathcal{X}^{\epsilon_2}$. If Assumptions F.1 and F.5 hold, then $\Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})$ is $\mathcal{C}^4$ on $\mathcal{X}^{\epsilon_1}$.

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

Let $C_1 = 1/R_0$ and $C_2 = 1/\epsilon$. Let ϵ_2 be some constant such that $\epsilon_2 + \sqrt{\frac{\rho}{\mu}} \cdot \frac{C_2}{C_1}\epsilon_2 < \epsilon_3$. For any starting point $\boldsymbol{\theta}_0 \in \Gamma^{\epsilon_2}$, and any preconditioning matrix $\boldsymbol{S}$ satisfying $C_1\boldsymbol{I} \preceq \boldsymbol{S} \preceq C_2\boldsymbol{I}$, assume on the contrary that the preconditioned gradient flow starting from $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$ will leave Γ^{ϵ_3} at some finite time. Then let $T = \inf\{t : \boldsymbol{\theta}(t) \notin \Gamma^{\epsilon_3}\} < \infty$. Using Lemma G.1 and combining the μ-PL condition, we conclude that $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}(T)\|_2 \leq \frac{2C_2}{\sqrt{2\mu C_1}}\sqrt{\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*} \leq \frac{2C_2}{\sqrt{2\mu C_1}} \cdot \sqrt{\frac{\mu}{2}}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2 = \sqrt{\frac{\rho}{\mu}} \cdot \frac{C_2}{C_1}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2$ for any $\boldsymbol{\theta}^* \in \Gamma$. Hence $\boldsymbol{\theta}(T) \in \Gamma^{\epsilon_3}$, a contradition.

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

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

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

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

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 neighborbood of $\boldsymbol{\theta}$ stays in the ϵ_2-neighborhoods of two different points on Γ. Taking union of all $\boldsymbol{\theta}_0 \in \Gamma$, we conclude that γ_δ is $\mathcal{C}^4$ on $\mathcal{X}^{\epsilon_1}$. Finally, we use Theorem 6.4 in Falconer (1983) to conclude that $\Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})$ is $\mathcal{C}^4$ on $\mathcal{X}^{\epsilon_1}$. ■

Appendix H. Proof of the Convergence of AGMs

In this section, we aim to prove Theorem F.2, and consequently the first part of Theorem F.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 F.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 $\boldsymbol{m}_0 = \boldsymbol{v}_0 = \boldsymbol{0}$. For every $k \geq 0$, we use **step** $k+1$ to refer to the process of obtaining the noisy gradient $\nabla\ell_k(\boldsymbol{\theta}_k)$ and then $\boldsymbol{m}_{k+1}, \boldsymbol{v}_{k+1}$ and $\boldsymbol{\theta}_{k+1}$. For any $k \geq 0$, to simplify the notation, we denote that

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

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 filteration generated by the history of optimization, where each $\mathcal{F}_k = \sigma\left(\boldsymbol{\theta}_0, \boldsymbol{z}_0, \cdots, \boldsymbol{z}_{k-1}\right)$ can be interpreted as "all the information available up to time k". We use the notation $\mathbb{E}_k$ to denote the expectation conditioned on $\mathcal{F}_k$.

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

From Lemma H.1 throughout Lemma H.5, we will assume that $\mathcal{L}$ satisfies μ-PL condition everywhere, and Theorem F.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 F.1.

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

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

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

1. $|Y_k| \leq C_{1a}\|\nabla\mathcal{L}(\boldsymbol{\theta}_{k-1})\|_2 \cdot \eta^2$ *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 $S(\tilde{\boldsymbol{v}}_k)$ part off the $S(\boldsymbol{v}_k)$ term:

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

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

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

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 $\langle \nabla \mathcal{L}\left(\boldsymbol{\theta}_{k-1}\right), S(\tilde{\boldsymbol{v}}_k)\boldsymbol{g}_{k-1}\rangle$ can also be decomposed into a deterministic part and a random part as:

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

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

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

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

for some constant C_2.

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

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

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

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

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

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

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

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

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

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

Plugging in, we get

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

for some constant C_2. ■

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

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

for some constant C_3.

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

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

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

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

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:

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

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

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

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 formula for $\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^*$:

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

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

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

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

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

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

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

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

for some constant C_4.

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

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

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

Since $\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^* \le f(i,\delta')$ 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:

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

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

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

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

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

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

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

and

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

where $C_4 = 2C_{1b}\sqrt{\rho}$. ■

Lemma H.5 (Convergence Bound of the AGM Framework) *Let η be a small learning rate satisfying $\frac{\beta_1}{\gamma} = \beta_1/(1-\frac{2\eta\mu(1-\beta_1)}{R_0}) \le 0.95$ and $\tilde{\epsilon}_2 + \eta R/\epsilon < \epsilon_2$. Let $\boldsymbol{\theta}_0 \in \Gamma^{\tilde{\epsilon}_2}$, and $K = \mathcal{O}(\mathrm{poly}(1/\eta))$. Under mild restrictions on K, for any $k \le K$, $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

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

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

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

$$(C_{5a} + C_{5b}\eta)\left(\log K + \log\frac{1}{\delta}\right) \ge \mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*,$$

where setting $C_{5a} = \frac{\mathcal{L}(\boldsymbol{\theta}_0)-\mathcal{L}^*}{\log K}$ suffices. Now assume that the statement holds for $0, 1, \cdots, k-1$. From Lemma H.3, we have

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

We can bound the coefficients by

$$\sum_{j=i}^{k}\gamma^{k-j}\beta_1^{j-i} = \gamma^{k-i}\sum_{j=0}^{k-i}\left(\frac{\beta_1}{\gamma}\right)^j$$
$$\le \gamma^{k-i}\cdot\frac{1}{1-\frac{\beta_1}{\gamma}}$$
$$\le 20\gamma^{k-i},$$

where the last inequality is due to our assumption $\frac{\beta_1}{\gamma} \le 0.95$. Let $\tilde{X}_i := \frac{\sum_{j=i}^{k}\gamma^{k-j}\beta_1^{j-i}}{20\gamma^{k-i}}X_i$, then $\left\{\tilde{X}_i\right\}_{i=1}^{k}$ is also a martingale difference sequence and $\left|\tilde{X}_i\right| \le |X_i| \le C_{1b}\|\nabla\mathcal{L}(\boldsymbol{\theta}_i)\|_2$ a.s.

From Lemma H.4, with probability $1-\delta$, $\mathcal{L}(\boldsymbol{\theta}_{i-1}) - \mathcal{L}^* \le f\left(i, \frac{\delta}{2k}\right)$ holds for all $i \in [k]$ and $\left|\sum_{i=1}^k \gamma^{k-i}\tilde{X}_i\right| \le C_4\sqrt{\sum_{i=1}^k \gamma^{2k-2i} f\left(i, \frac{\delta}{2k}\right)\log\frac{4}{\delta}}$ holds. If this happens, we have

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

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

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

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

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

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

$$\begin{cases} C_{5a} & \geq C_{5c} + \mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^* & = 5C_4\sqrt{\frac{6C_{5a}R_0(1-\beta_1)}{\mu}} + \mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^* \\ C_{5b} & \geq C_{5d} + C_3 & = 5C_4\sqrt{\frac{6C_{5a}R_0(1-\beta_1)}{\mu}} + 10C_4\sqrt{\frac{6C_{5b}R_0(1-\beta_1)}{\mu}} + C_3. \end{cases}$$

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

Proof [Proof of Theorem F.2] This is a direct corollary following from Lemma H.5, where letting $\gamma^K = \mathcal{O}(\eta)$ gives $K = \mathcal{O}(\frac{1}{\eta}\log\frac{1}{\eta})$, completing the proof. ■

Proof [Proof of the first part of Theorem F.1] Fix $K = \lfloor (T+1)\eta^{-2} \rfloor$. By Lemma G.2, there exists 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 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 $\mathcal{L}_m = \min\left\{\mathcal{L}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Gamma^{\epsilon_2}, \boldsymbol{\theta} \notin \Gamma^{0.5\epsilon_2}\right\}$. Plugging in $C_{5a} = \frac{\mathcal{L}(\boldsymbol{\theta}_0)-\mathcal{L}^*}{\log K}$, $K \leq (T+1)\eta^{-2}$ and $\log\frac{1}{\delta}$ being upper bounded by $200\log\frac{1}{\eta}$ into Lemma H.5, we conclude that there exists some $\epsilon < 0.5\epsilon_2$ such that if $\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^* \leq \sup\left\{\mathcal{L}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Gamma^{\epsilon}\right\} - \mathcal{L}^*$ and η is sufficiently small, then with probability $1-\delta$, 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 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 H.5 holds, all iterations of AGM stay inside Γ_2^{ϵ}. Then substituting $K_0 = \lceil \log\rceil_\gamma \eta\rceil = \mathcal{O}(\frac{1}{\eta}\log\frac{1}{\eta})$ gives the result.

■

Appendix I. Proof of the SDE Approximation of AGMs

In this section, we present a detailed derivation of our slow SDE approximation of the AGM framework as shown in Theorem 2.1.

Remark I.1 *Without causing confusion, we reword the definition of $\boldsymbol{\theta}_0$ and $\boldsymbol{v}_0$ in this section. In the following calculation in Appendix I, $\boldsymbol{\theta}_0$ and $\boldsymbol{v}_0$ do not represent the parameters that are initialized at the real beginning of training, instead they represent the $\boldsymbol{\theta}_{K_0}$ and $\boldsymbol{v}_{K_0}$ yielded by the first part of Theorem F.1, we define $\boldsymbol{\theta}_0$ as the parameter near the minimizer manifold such that $\|\boldsymbol{\theta}_0 - \boldsymbol{\phi}_0\|_2 = \mathcal{O}(\sqrt{\eta\log\frac{1}{\eta}})$, and $\boldsymbol{v}_0$ is the velocity vector as the corresponding time step as $\boldsymbol{\theta}_0$. Our SDE approximation then describes AGM's dynamics after reaching such a state $(\boldsymbol{\theta}_0, \boldsymbol{v}_0)$.*

I.1. Lemmas for Adaptive Manifold Projection

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

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

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

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

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

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

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

Taking $t = 0$ completes the proof. ■

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

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

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

The above identity completes the proof. ■

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

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

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

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

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

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

A slight modification using the above transformations gives

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

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

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

We completes the proof. ■

I.2. Iteration Stays Near Manifold

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

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

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

for some constant C_6.

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

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

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

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

A key observation is that $\partial\Phi_{S(\boldsymbol{v}_{\bar{k}})}(\boldsymbol{\theta}_{\bar{k}})S(\boldsymbol{v}_{\bar{k}})\nabla\mathcal{L}(\boldsymbol{\theta}_{\bar{k}}) = 0$ from Lemma I.2, which allows us to view $\boldsymbol{\phi}_{k+1} - \boldsymbol{\phi}_k$ as $\sum_{i=\bar{k}}^{k}\tilde{\boldsymbol{z}}_{k,i} + \mathcal{O}(\eta^2\log\frac{1}{\eta})$ where $\tilde{\boldsymbol{z}}_{k,i} = \partial\Phi_{S(\boldsymbol{v}_{\bar{k}})}(\boldsymbol{\theta}_{\bar{k}})(\eta(1-\beta_1)\beta_1^{k-i}S(\boldsymbol{v}_{\bar{k}})\boldsymbol{z}_i)$.

Note that $\tilde{\boldsymbol{z}}_{k,i}$ is $\mathcal{F}_{i+1}$-measurable and its mean is $\mathbf{0}$, since $\tilde{\boldsymbol{z}}_{k,i}$ just applies a linear tensor transformation to $\boldsymbol{z}_i$. If we define a constant $C_{6a} := \sup\left\{\|\partial\Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})\|_2 \mid (\boldsymbol{v},\boldsymbol{\theta}) \in \mathcal{X}^{\epsilon_2}\right\} \cdot (1-\beta_1)\cdot\epsilon^{-1}$ that is independent of k and i, then $\|\tilde{\boldsymbol{z}}_{k,i}\|_2$ is almost surely bounded by $\eta\beta_1^{k-i}C_{6a}\|\boldsymbol{z}_i\|_2$.

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

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

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

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

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

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

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

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

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

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

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

Proof Denote $\delta := \eta^{100}$. From Lemma H.5, with probability $1-\delta/2$, all $k \in [0, K]$ satisfy $\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}^* \le (C_{5a} + C_{5b})\eta \log \frac{2K}{\delta}$. Combining Lemma G.1, this implies $\|\boldsymbol{\theta}_k - \boldsymbol{\phi}_k\|_2 \le \frac{2C_2}{\sqrt{2\mu C_1}} \cdot \sqrt{(C_{5a} + C_{5b})\eta \log \frac{2K}{\delta}}$ for any $k \in [0, K]$. When η is small enough such that $\frac{2C_2}{\sqrt{2\mu C_1}} \cdot \sqrt{(C_{5a} + C_{5b})\eta \log \frac{2K}{\delta}} + \eta R/\epsilon < \epsilon_2$, any $\boldsymbol{\phi}_k \in \Gamma$ with $k \ge 0$ will imply $\boldsymbol{\phi}_{k+1} \in \Gamma$, since $\boldsymbol{\theta}_{k+1}$ cannot escape Γ^{ϵ_2}. Giving $\phi_0 \in \Gamma$ and using induction, we conclude that all $\boldsymbol{\phi}_k \in \Gamma$ for $k \ge 0$.

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

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

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

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

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

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

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

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

I.3. Moment Calculation of AGMs Near Manifold

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

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

We further introduce a list of notations:

$\hat{\boldsymbol{g}}_t^{(s)} := \nabla \ell_t^{(s)}(\hat{\boldsymbol{\theta}}_t^{(s)})$, $\hat{\boldsymbol{S}}_k = S(\hat{\boldsymbol{v}}_k)$, $\hat{\boldsymbol{S}}_t^{(s)} := S(\hat{\boldsymbol{v}}_t^{(s)})$, $\hat{\boldsymbol{S}}^{(s)} := \hat{\boldsymbol{S}}_0^{(s)}$, $\hat{\boldsymbol{\phi}}^{(s)} := \boldsymbol{\Phi}_{\hat{\boldsymbol{S}}^{(s)}}(\hat{\boldsymbol{\theta}}_0^{(s)})$,
$\hat{\boldsymbol{x}}_t^{(s)} := \hat{\boldsymbol{\theta}}_t^{(s)} - \hat{\boldsymbol{\phi}}^{(s)}$, $\Delta\hat{\boldsymbol{\phi}}^{(s)} := \hat{\boldsymbol{\phi}}^{(s)} - \hat{\boldsymbol{\phi}}^{(0)}$, $\boldsymbol{\Sigma}_0 := \boldsymbol{\Sigma}(\hat{\boldsymbol{\phi}}^{(0)})$, $\boldsymbol{P}_{\|} := \partial\boldsymbol{\Phi}_{\hat{\boldsymbol{S}}^{(0)}}(\hat{\boldsymbol{\phi}}^{(0)})$, $\boldsymbol{P}_{\perp} := \boldsymbol{I} - \boldsymbol{P}_{\|}$,
$\hat{\boldsymbol{q}}_t^{(s)} := \mathbb{E}[\hat{\boldsymbol{x}}_t^{(s)}]$, $\hat{\boldsymbol{A}}_t^{(s)} := \mathbb{E}[\hat{\boldsymbol{x}}_t^{(s)}\hat{\boldsymbol{x}}_t^{(s)\top}]$, $\hat{\boldsymbol{B}}_t^{(s)} := \mathbb{E}[\hat{\boldsymbol{x}}_t^{(s)}\Delta\hat{\boldsymbol{\phi}}^{(s)\top}]$.

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

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

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

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

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

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

for some constant C_{10}.

Proof We have

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

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

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

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

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

After that, notice that

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

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

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

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

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

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

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

which concludes the proof. ■

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

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

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

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

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

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

Note that E_1 is neglegible:

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

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

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

Combining the results of E_1 and E_2 gives

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

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

I.3.1. Moment Calculation Within a Giant Step

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

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

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

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

Proof We write the update rule:

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

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

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

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

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

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

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

which concludes the proof. ■

After getting the update rule of $\hat{\boldsymbol{\theta}}_k$, we then derive the moment during the single round with $\boldsymbol{H}$ steps. To this end, we recap our modification of manifold projection from a "Gradient Flow" manner to a "Preconditioned Flow" manner in Definition 2.1.

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

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

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

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

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

Proof First, we consider the scenario when $R_0 < s < R_{\mathrm{grp}}$. By Lemma I.8, the update rule holds. And we consider an auxiliary process $\{\hat{\boldsymbol{\theta}}'_t\}$. Let $L'(\boldsymbol{x}) := L(\boldsymbol{P}\boldsymbol{x})$, then

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

For an one-step GD update, we have that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Notice that

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

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

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

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

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

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

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

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

Proof First we prove the first moment change as

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

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

And for the second moment, we have

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

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

Next, we proceed with the updates of v.

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

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

Proof We have

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

Note that

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

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

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

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

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

Thus we have

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

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

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

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

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

Proof

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

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

The last equation uses

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

and

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

The above equation completes the proof. ■

I.4. Weak Approximation

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

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

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

$$\begin{cases}\mathrm{d}\boldsymbol{\zeta}=\partial\Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\zeta})\boldsymbol{S}(\boldsymbol{v})\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}_t+\frac{1}{2}\boldsymbol{S}(\boldsymbol{v})\partial^2\Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\zeta})\left[\boldsymbol{S}(\boldsymbol{v})\boldsymbol{\Sigma}(\boldsymbol{\zeta})\boldsymbol{S}(\boldsymbol{v})\right]\mathrm{d}t,\\ \mathrm{d}\boldsymbol{v}(t)=c\left(V(\boldsymbol{\Sigma}(\boldsymbol{\zeta}))-\boldsymbol{v}\right)\mathrm{d}t.\end{cases}\tag{5}$$

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

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

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

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

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

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

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

The last equation uses Lemma I.3. Again, applying Lemma I.3 gives

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

The above equation completes the proof. ■

To establish Theorem 2.1, we give an equivalent theorem, which capture the closeness of $\boldsymbol{X}(t)$ and $\bar{\boldsymbol{X}}_t$ in a long horizon. Also, for the proof of Theorem 2.1, it suffices to prove the following lemma, whose proof would be shown at the end of Appendix I.4.

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

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

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

I.4.1. PRELIMINARY AND ADDITIONAL NOTATIONS

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

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

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

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

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

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

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

$$\boldsymbol{b}(\boldsymbol{X}_t, t) = \left(\left(\frac{1}{2}\partial^2 \Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\zeta}) \left[\boldsymbol{\Sigma}(\boldsymbol{\zeta}, \boldsymbol{S}(\boldsymbol{v})) \right] \right)^\top, c\left(V(\boldsymbol{\Sigma}(\boldsymbol{\zeta})) - \boldsymbol{v} \right)^\top \right)^\top,$$

$$\sigma(\boldsymbol{X}_t, t) = \begin{pmatrix} \partial \Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\zeta}) \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\zeta}, \boldsymbol{S}(\boldsymbol{v})), & \mathbf{0} \\ \mathbf{0}, & \mathbf{0} \end{pmatrix}.$$

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

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

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

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

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

Proof (i) By Lemma I.13, the first half solution $\boldsymbol{\zeta}(t)$ in $\boldsymbol{X}(t)$ of Equation (5) stays in the manifold almost surely when $\boldsymbol{\zeta}(0) \in \Gamma$. (ii) We assume that $\mathcal{L} \in \mathcal{C}^5$, so $\boldsymbol{b}, \sigma \in \mathcal{C}^4$. (iii) We know that Γ is compact by Assumption 2.2. Then we can directly apply Lemma B.3 in Malladi et al. (2022) and Lemma 26 in Li et al. (2019). ■

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

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

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

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

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

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

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

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

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

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

Proof According to Appendix I.2, we have that

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

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

$$\left|\mathbb{E}[\Delta_i^{(n)} - \eta_e b_i^{(n)}]\right| \le c_2 \left(\eta_e^{a_1}(\log \frac{1}{\eta_e})^b + \eta_e^{a_2}(\log \frac{1}{\eta_e})^b\right), \tag{6}$$
$$\left|\mathbb{E}[\Delta_i^{(n)}\Delta_j^{(n)} - \eta_e \sum_{l=1}^{d} \sigma_{i,l}^{(n)} \sigma_{l,j}^{(n)}]\right| \le c_2 \left(\eta_e^{a_1}(\log \frac{1}{\eta_e})^b + \eta_e^{a_2}(\log \frac{1}{\eta_e})^b\right) \tag{7}$$
$$\mathbb{E}\left[\left|\prod_{s=1}^{6} \Delta_{i_s}^{(n)}\right|\right] \le c_2^2 \eta_e^{2a_1} (\log \frac{1}{\eta_e})^{2b}. \tag{8}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$I_{1,1} \le dc_1C_{g,4}(\eta_e^{a_1}+\eta_e^{a_2})(\log\frac{1}{\eta_e})^b,\ I_{1,2} \le \frac{d^2}{2}c_1C_{g,4}(\eta_e^{a_1}+\eta_e^{a_2})(\log\frac{1}{\eta_e})^b.$$

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

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

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

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

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

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

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

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

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

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

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

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

Appendix J. Proof of Theorems in Appendix D

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

In this part, we give the proof of Theorem D.1 and Theorem D.2.

Proof [Proof of Theorem D.1] With label noise, the gradient covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\zeta}) = \alpha\nabla^2\mathcal{L}(\boldsymbol{\zeta})$ for any $\boldsymbol{\zeta} \in \Gamma$.

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

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

And $\boldsymbol{P} := \boldsymbol{S}_0^{1/2}$. By applying Lemma I.2, and replacing Σ with $\alpha\nabla^2\mathcal{L}$, we can reduce the SDE formula as

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

Then we can write out the constraint for the fixed point $(\boldsymbol{\zeta}^*, \boldsymbol{v}^*)$ of this ODE as

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

$$V(\boldsymbol{\Sigma}(\boldsymbol{\zeta}^*)) - \boldsymbol{v}^* = 0. \tag{11}$$

Solving Equation (10) and Equation (11) gives

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

Integrating by parts gives us

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

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

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

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

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

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

$$\boldsymbol{S}\nabla\mathrm{tr}\left(\left(\mathrm{Diag}\boldsymbol{H}\right)^{\frac{1}{2}}\right) + O(\boldsymbol{S}\epsilon\mathrm{tr}(\boldsymbol{H})), \tag{14}$$

forming a preconditioned gradient flow that implicitly minimizes $\mathrm{tr}\left(\left(\mathrm{Diag}\boldsymbol{H}\right)^{\frac{1}{2}}\right)$ when $\epsilon \to 0$. Combining Equation (12) and Eq. (14) gives the result in Theorem D.1. ■

Proof [Proof of Theorem D.2] Now we consider the optimizer AdamE-λ, the variant of Adam proposed as a verification case of our main results, whose update rule is

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

Now for AdamE-λ, the precondition matrix $\boldsymbol{S} = (\mathrm{Diag}\boldsymbol{H})^{-\lambda} + O(\epsilon)$, and $\boldsymbol{P} = (\mathrm{Diag}\boldsymbol{H})^{-\lambda/2} + O(\sqrt{\epsilon})$, which gives

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

Also, similar to the case for Adam, we have

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

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

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

Evaluating the above equation when $\epsilon \to 0$ completes the proof. ■

J.2. Proof of Lemma D.1

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

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

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

Hence, we have that

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

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

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

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

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