FINE-GRAINED ANALYSIS OF ADAM'S IMPLICIT BIAS NEAR THE MINIMIZER MANIFOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

The implicit bias of stochastic gradient descent (SGD) has been widely studied. But the implicit bias of Adam, one of the most widely used adaptive gradient methods (AGMs), remains less well understood. A key difficulty is that the model parameters and the adaptive preconditioner evolve on different timescales, depending on how $1-\beta_2$ scales with the learning rate η, where β_2 is the second-moment decay coefficient. Under the scaling $1-\beta_2 = \Theta(\eta^2)$, termed the *2-scheme*, Li et al. (2025) characterized Adam's dynamics near a manifold of loss minimizers and showed that Adam implicitly reduces a form of sharpness different from that associated with SGD. In practice, however, $1-\beta_2$ can be much larger than η^2, motivating the study of the scaling $1-\beta_2 = \Theta(\eta^\alpha)$ with $\alpha < 2$. In this work, we resolve the case where $1 \le \alpha < 2$. We derive the limiting dynamics of AGMs, including Adam, over $O(\eta^{-2})$ iterations and characterize the resulting implicit bias with $1 \le \alpha < 2$, along with the convergence guarantees for approaching and remaining near the minimizer manifold. Under label noise, we recover the sharpness characterization of the 2-scheme for $1 < \alpha < 2$, while an additional drift arising from third order terms leads to a different implicit bias at $\alpha = 1$. To obtain these results, we develop a new framework for analyzing coupled stochastic dynamics with multiple timescales near a manifold. Technically, this framework generalizes the ideas of Katzenberger (1991) and sheds light on the limiting behavior characterization of a broader class of stochastic processes.

1 INTRODUCTION

Neural network training aims to find models that fit the training data and generalize to unseen examples. In overparameterized models, many parameter configurations can fit the training data equally well, yet the resulting models can generalize differently (Zhang et al., 2017; Keskar et al., 2017; Jiang et al., 2020; Saunshi et al., 2024). The optimization algorithm influences which of these solutions is reached. Its preference among solutions is known as its *implicit bias*, and understanding this preference helps explain how optimization affects generalization (Neyshabur et al., 2014; Soudry et al., 2018; Woodworth et al., 2020).

Much of the work on implicit bias has focused on gradient descent (Soudry et al., 2018; Nacson et al., 2019; Lyu & Li, 2020) and stochastic gradient descent (SGD). A line of work (Blanc et al., 2020; Damian et al., 2021; Li et al., 2021b; Gu et al., 2023) shows that the gradient noise in SGD induces a regularization effect to reduce *sharpness*, a metric that has long been observed to be connected with generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Jiang et al., 2020; Foret et al., 2021). Specifically, they consider the case when the training loss is already very small, and the parameter is near the *minimizer manifold*, which can be viewed as a smooth set of local loss minimizers that locally resembles Euclidean space (Garipov et al., 2018; Li et al., 2021b).

To characterize the above behavior of SGD near the minimizer manifold, Li et al. (2021b) build on the framework of Katzenberger (1991) for stochastic processes driven toward a manifold by a large drift. They derive a limiting stochastic differential equation (SDE) for SGD's motion along the minimizer manifold as the learning rate η tends to zero. Gu et al. (2023) refer to this description as a *slow SDE* and establish quantitative weak approximation guarantees over $\mathcal{O}(\eta^{-2})$ iterations. Intuitively, gradient noise perturbs the parameters while the gradient drift pulls them back toward

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the manifold. The slow SDE describes the long term movement of SGD along the manifold driven by the above two effects.

However, the implicit bias of Adam (Kingma & Ba, 2014), one of the most widely used adaptive gradient methods (AGMs), remains less well understood. While SGD updates the parameters simply as $\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k - \eta \nabla \ell_k(\boldsymbol{\theta}_k)$, Adam additionally maintains a momentum variable $\boldsymbol{m}$ and a second moment estimate $\boldsymbol{v}$, which complicates the update rule. Adam's update rule can be stated as

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

Here ℓ_k is the loss for the sample drawn at step k, d is the parameter dimension, $\odot 2$ denotes elementwise squaring, and $\epsilon > 0$ stabilizes the denominator. The main difficulty in analyzing Adam's implicit bias is that, compared with SGD, Adam introduces two additional states $\boldsymbol{m}$ and $\boldsymbol{v}$, which evolve together with the parameters $\boldsymbol{\theta}$. The preconditioner determined by $\boldsymbol{v}$ rescales the parameter update driven by $\boldsymbol{m}$. In turn, changes in $\boldsymbol{\theta}$ affect the gradients used to update $\boldsymbol{m}$ and $\boldsymbol{v}$. Their relative rates of change depend on how $1-\beta_1$ and $1-\beta_2$ scale with η. Therefore, it is difficult to determine how these coupled updates add up over many iterations and influence which solutions Adam reaches near the minimizer manifold.

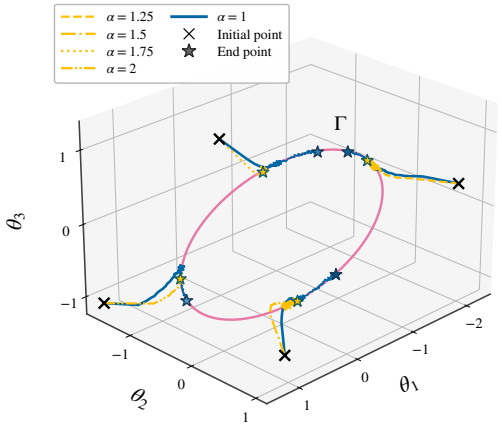


Figure 1: **The 1-scheme has an additional drift compared with the 2-scheme, while schemes $(1,2)$ do not.** In this setting, the loss function $\mathcal{L}$ is designed such that the pink ellipse Γ is the minimizer manifold, and each point on Γ has the same $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H})^{1/2})$. According to Li et al. (2025), the 2-scheme will not induce any drift on manifold after convergence. We observe that schemes $(1,2)$ behaves similarly, but the 1-scheme continues to drift on manifold after convergence. We will later elaborate on this setting in Section 4.1 and Figure 3.

Despite the difficulties in characterizing Adam's implicit bias, several studies have made progress on it. A useful starting point is SGD with momentum, which allows the effect of momentum to be studied separately from Adam's adaptive preconditioner. In this setting, Wang et al. (2023) show that, under suitable scalings, adding momentum does not change the implicit bias of SGD. Further, Li et al. (2025) derive a slow SDE for Adam under the scaling $1-\beta_2 = \Theta(\eta^2)$. Under label noise, they identify a sharpness reduction bias different from that of SGD. In practice, however, $1-\beta_2$ can be much larger than η^2. For example, the default Adam settings in PyTorch are $\beta_2 = 0.999$ and $\eta = 0.001$, so that $1-\beta_2 = \eta$. These values motivate studying faster adaptation of the preconditioner, whose effect on Adam's limiting dynamics and implicit bias remains unclear.

Our contributions. We keep β_1 bounded away from one and study the scaling $1-\beta_2 = c_\beta \eta^\alpha$, where $c_\beta > 0$ is fixed. Our main results are as follows.

- For $1 \le \alpha < 2$, we derive the slow SDE and establish weak approximation guarantees over $\mathcal{O}(\eta^{-2})$ iterations for AGMs including Adam, along with the convergence guarantee for AGMs to get close to the manifold. (Section 3)
- Under label noise, we show that Adam reduces the sharpness measure identified by Li et al. (2025) for $1 < \alpha < 2$. While under the 1-scheme, third-order terms can preventing Adam from minimizing the sharpness measure above. (Section 4)
- Technically, we extend the framework of Katzenberger (1991) to AGMs in which the parameters and the preconditioner change at different timescales. By controlling how re-

peated stochastic parameter updates affect the preconditioner over $\mathcal{O}(\eta^{-2})$ iterations, we derive the limiting dynamics for $1 < \alpha < 2$. These techniques also provide insight into the limiting behavior of a broader class of stochastic processes.

2 PRELIMINARIES

In this section, we make some concepts in Section 1 rigorous, including the assumptions needed on the loss function, the *minimizer manifold*, the class of *Adaptive Gradient Methods (AGMs)*, and the setting of *schemes* whose difference will let the dynamics of AGMs vary. We start from the notations below.

Notations. Let $\mathbb{R}^d_{\geq 0}$ be the set of d-dimensional vectors with non-negative entries, and $\mathbb{S}^d_{++}$ be the set of $d \times d$ symmetric positive definite matrices. Given a vector-valued function $F : \mathbb{R}^d \to \mathbb{R}^d$ and a matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, we define $\partial^2 F(\boldsymbol{\theta})[\boldsymbol{M}]$ as the second-order directional derivative of F at $\boldsymbol{\theta}$ in the direction of $\boldsymbol{M}$ 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 denotes the i-th coordinate of F, and $\boldsymbol{e}_i$ denotes the i-th vector of the standard basis. For any function $f : \mathbb{R}^d \to \mathbb{R}$, we abbreviate $\partial^2(\nabla f)(\boldsymbol{\theta})[\boldsymbol{M}]$ as $\nabla^3 f(\boldsymbol{\theta})[\boldsymbol{M}]$ when the context is clear.

2.1 THE LOSS FUNCTION AND ITS MINIMIZER MANIFOLD

The Loss Function. We now formally define the loss function used in our paper. Let $\mathcal{S}$ be the training dataset. For a data sample ξ, $\xi \sim \mathcal{S}$ means that the sample is uniformly randomly drawn from $\mathcal{S}$. Let $\ell(\boldsymbol{\theta}; \xi)$ be the loss function for a model with parameters $\boldsymbol{\theta}$ and $\mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}_{\xi\sim\mathcal{S}}[\ell(\boldsymbol{\theta}; \xi)]$ be the training loss function. Let $\mathcal{Z}(\boldsymbol{\theta})$ be the distribution of gradient noise $\nabla\ell(\boldsymbol{\theta}; \xi) - \nabla\mathcal{L}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}(\boldsymbol{\theta})}[\boldsymbol{z}\boldsymbol{z}^\top]$ be the noise covariance matrix of gradients at $\boldsymbol{\theta}$. We make the following assumptions on $\mathcal{L}(\boldsymbol{\theta})$ and $\nabla\ell(\boldsymbol{\theta}; \xi)$.

Assumption 2.1. *The loss function $\mathcal{L}$ and the matrix square root of the noise covariance $\boldsymbol{\Sigma}^{1/2}$ are $\mathcal{C}^\infty$-smooth on $\mathbb{R}^d$, i.e. all their partial derivatives up to any finite order exist and are continuous. Moreover, $\mathcal{L}$ is bounded below and ρ-smooth on $\mathbb{R}^d$, i.e. $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) > -\infty$, and for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, $\|\nabla\mathcal{L}(\boldsymbol{\theta}_1) - \nabla\mathcal{L}(\boldsymbol{\theta}_2)\|_2 \leq \rho\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$.*

Assumption 2.2. *The noisy gradients are ℓ_2-bounded, i.e., there exists some constant R s.t. $\forall\boldsymbol{\theta} \in \mathbb{R}^d$, $\|\nabla\ell(\boldsymbol{\theta}; \xi)\|_2 \leq R$ almost surely for data sample $\xi \sim \mathcal{S}$.*

Remark. The $\mathcal{C}^\infty$ in assump 2.1 is a bit tighter than we really need but provides much simplicity. Here, for each fixed $\alpha \in (1, 2)$, the proof of our Theorem 3.3 uses only finitely many derivatives of the loss function $\mathcal{L}$ and matrix square root $\boldsymbol{\Sigma}^{1/2}$. The required order increases to infinity as $\alpha \to 1$. We therefore give a single $\mathcal{C}^\infty$-smoothness assumption to cover the entire interval $(1, 2)$.

The Manifold Assumption. The slow SDE only applies when a manifold of minimizers exists. In fact, the manifold provably exists with mild assumptions: For smooth squared-loss, a full-rank Jacobian of the training residuals generates a local minimizer manifold whose Hessian matrix is non-degenerate in the normal direction (Li et al., 2022). Consequently, we adopt the following manifold assumption saying that a manifold of local minimizers exists, inside whose neighborhood we proceed to study the dynamics of AGMs.

Assumption 2.3. *Let $\mathcal{L}^* \in \mathbb{R}$, and Γ be a $\mathcal{C}^\infty$-smooth, $(d-m)$-dimensional compact submanifold of $\mathbb{R}^d$ such that, for all $\boldsymbol{\zeta} \in \Gamma$, $\mathcal{L}(\boldsymbol{\zeta}) = \mathcal{L}^*$ and rank$(\nabla^2\mathcal{L}(\boldsymbol{\zeta})) = m$. Additionally, there exists an open neighborhood of Γ, denoted as U_Γ, such that $\Gamma = \arg\min_{\boldsymbol{\theta}\in U_\Gamma} \mathcal{L}(\boldsymbol{\theta})$.*

2.2 ADAPTIVE GRADIENT METHODS AND THEIR SCHEMES

A General Class of Adaptive Gradient Methods. Motivated by adaptive optimizers such as AdaGrad (Duchi et al., 2011) and Adam (Kingma & Ba, 2014), we consider a general class of Adaptive Gradient Methods (AGMs) defined 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 $\boldsymbol{\theta}_k, \boldsymbol{m}_k \in \mathbb{R}^d$, $\boldsymbol{v}_k \in \mathbb{R}^D_{\geq 0}$, $\ell_k(\boldsymbol{\theta}_k) := \ell(\boldsymbol{\theta}_k; \xi_k)$ and η is the learning rate. $V : \mathbb{R}^{d\times d} \to \mathbb{R}^D$ is linear and satisfies $V(\boldsymbol{g}\boldsymbol{g}^\top) \in \mathbb{R}^D_{\geq 0}$ for all $\boldsymbol{g} \in \mathbb{R}^d$. $S : \mathbb{R}^D_{\geq 0} \to \mathbb{S}^d_{++}$ is a C^∞-smooth and ρ_s-smooth function such that $S(\boldsymbol{v}) \succeq \frac{1}{R_0}\boldsymbol{I}$ for some $R_0 > 0$ and any $\boldsymbol{v} \in \mathbb{R}^D_{\geq 0}$.

Corollary 2.1. *There exists $R_1 > 0$ such that for any $\boldsymbol{\theta} \in \mathbb{R}^d$, $\mathbf{0} \leq V\left(\nabla \ell(\boldsymbol{\theta};\xi)\nabla \ell(\boldsymbol{\theta};\xi)^\top\right) \leq R_1 \mathbf{1}$.*

Assumption 2.4. *There exists a constant $\bar{\beta}_1 \in (0,1)$ independent of η such that $0 < \beta_1 \leq \bar{\beta}_1$.*

Assumption 2.4 indicates that β_1 is bounded at least $\mathcal{O}(1)$-distance away from 1, a fact that will be used in the convergence theorem to show the concentration of the loss function. In Li et al. (2025) and most practical settings, $\bar{\beta}_1 = 0.9$ suffices.

However, the case is different when considering $1 - \beta_2$. Practically this value is too small to be considered a constant, for instance, PyTorch sets $\beta_2 = 0.999$. In addition, the order α of $1 - \beta_2 = \Theta(\eta^\alpha)$ proves to be a key factor when analyzing the limiting dynamics of AGMs. We therefore give a formal definition of *schemes*, which is referred to as the possible orders of $1 - \beta_2$. More precisely, we study schemes $[1, 2)$ in this paper.

Definition 2.1 (Schemes)**.** *For any $\alpha_0 \in \mathbb{R}$, we refer to the regime where there exists constant $c_\beta > 0$ such that for any $\eta > 0$, $1 - \beta_2 = c_\beta \eta^{\alpha_0}$ as α_0-scheme. For a set $A \subseteq \mathbb{R}$, we refer to the regime where there exists constant $c_\beta > 0$ and $\alpha \in A$ such that for any $\eta > 0$, $1 - \beta_2 = c_\beta \eta^\alpha$ as scheme A.*

3 Main Results

In this section, we introduce the slow SDE of AGMs, including Adam, under schemes $[1, 2)$. We first provide an illustration of the stochastic training process of AGMs near the manifold in Figure 2, covering schemes $[1, 2]$. Then we formally state the slow SDE of AGMs under the 1-scheme (Section 3.3.1) and schemes $(1, 2)$ (Section 3.3.2). These conclusions are significant extensions of the 2-scheme slow SDE solved by Li et al. (2025).

3.1 Convergence Guarantee for AGMs under schemes [0.5, 2]

Our main goal is to characterize the implicit bias of AGMs near the minimizer manifold. To justify analyzing the dynamics in this region, we first establish a convergence guarantee. Specifically, we show that AGMs reach an $\tilde{O}(\sqrt{\eta})$-neighborhood of the minimizer manifold Γ defined in Assumption 2.3 as a prerequisite. The convergence result covers $0.5 \leq \alpha \leq 2$. See Appendix D for the proof.

Theorem 3.1 (Convergence of AGMs, Stated Informally)**.** *Suppose Assumption 2.1–2.4 hold and $1 - \beta_2 = c_\beta \eta^\alpha$, where $c_\beta > 0$ is fixed and $0.5 \leq \alpha \leq 2$. For sufficiently small η, AGMs initialized sufficiently close to Γ with bounded optimizer states satisfy $\mathcal{L}(\boldsymbol{\theta}_{K_\eta}) - \mathcal{L}^* = \widetilde{\mathcal{O}}(\eta), \operatorname{dist}(\boldsymbol{\theta}_{K_\eta}, \Gamma) = \widetilde{\mathcal{O}}(\sqrt{\eta})$ with high probability after $K_\eta = \mathcal{O}(\eta^{-1}\log(1/\eta))$ iterations.*

Theorem 3.1 provides the starting point for analyzing AGM dynamics near Γ. The proofs of the weak approximation results in Theorems 3.2 and 3.3 require the parameters to remain within a vanishing distance of Γ as $\eta \to 0$. The theorem establishes this closeness after the initial convergence period, and Corollary D.1 extends the control to $\mathcal{O}(\eta^{-2})$ iterations. Additional conditions on $\boldsymbol{v}$ are imposed in the corresponding weak approximation results. Although the convergence guarantee covers $0.5 \leq \alpha \leq 2$, our subsequent analysis focuses on $1 \leq \alpha < 2$. See Appendix D for the formal statement and proof.

3.2 An Illustration

Having established that the iterates approach the minimizer manifold, we now examine their subsequent dynamics near it.

As shown in Figure 1, Adam under schemes $(1, 2)$ seems to behave similarly as the 2-scheme, while the 1-scheme does not. In fact, these three regimes all behave differently in dynamics and we offer a heuristic yet intuitive understanding here.

First, recall that the update of $\boldsymbol{v}$ is

$$\boldsymbol{v}_{k+1} := \beta_2 \boldsymbol{v}_k + (1 - \beta_2) V_k = \boldsymbol{v}_k + (1 - \beta_2)(V_k - \boldsymbol{v}_k),$$

where we abbreviate $V\left(\nabla \ell_k(\boldsymbol{\theta}_k)\nabla \ell_k(\boldsymbol{\theta}_k)^\top\right)$ as V_k here. Thus the process of $\boldsymbol{v}$ can be viewed as a stochastic process that moves, in expectation, towards the expectation of V. We denote this expectation as $\bar{U}$ in the sequel, terming it as the *local target* tracked by $\boldsymbol{v}$.

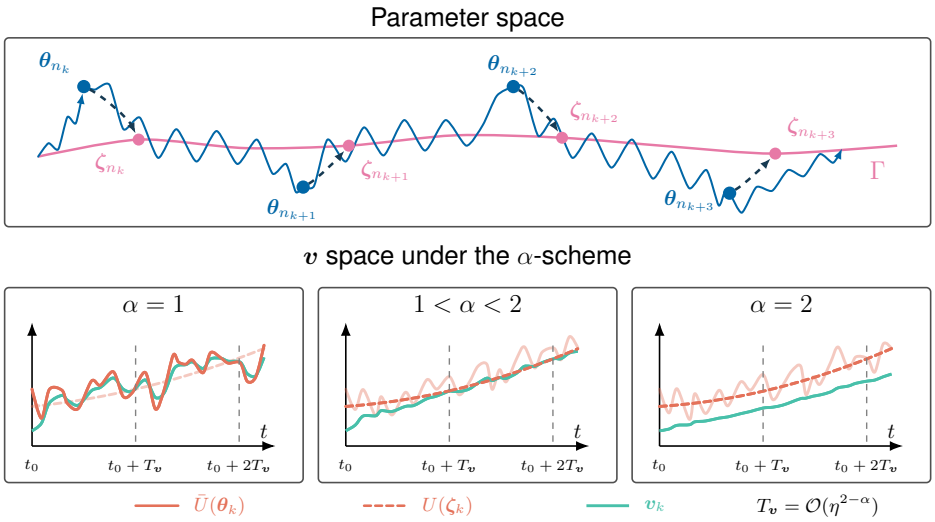


Figure 2: **An illustration for schemes** $[1, 2]$**. Top:** The dynamics of the parameter $\boldsymbol{\theta}$ and its preconditioned flow projection $\boldsymbol{\zeta} := \Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})$ onto the manifold. $\boldsymbol{\zeta}$ nearly stays still locally, only with a $\mathcal{O}(\eta^2)$ moving speed in expectation. **Bottom:** The dynamics of $\boldsymbol{v}$, the local target $\bar{U}(\boldsymbol{\theta})$ and the smoothed local target $U(\boldsymbol{\zeta})$. U is a smoothed version of $\bar{U}$ via short-term EMA. $\bar{U}$ moves at speed $\mathcal{O}(\eta)$, U moves at speed $\mathcal{O}(\eta^2)$ and $\boldsymbol{v}$ tracks $\bar{U}$ at speed $\mathcal{O}(\eta^\alpha)$.

Definition 3.1. *For any* $\boldsymbol{\theta} \in \mathbb{R}^d$*, define the* local target*,* $\bar{U}(\boldsymbol{\theta})$*, as*

$$\bar{U}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}(\boldsymbol{\theta})}\left[V\left(\nabla\ell(\boldsymbol{\theta};\xi)\nabla\ell(\boldsymbol{\theta};\xi)^\top\right)\right] = V\left(\boldsymbol{\Sigma}(\boldsymbol{\theta}) + \nabla\mathcal{L}(\boldsymbol{\theta})\nabla\mathcal{L}(\boldsymbol{\theta})^\top\right).$$

Notice that $\bar{U}(\boldsymbol{\theta})$ moves at a speed of $\mathcal{O}(\eta)$. This is because $\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -\eta S(\boldsymbol{v}_{k+1})\boldsymbol{m}_{k+1} = \mathcal{O}(\eta)$ and $\bar{U}$ is a smooth function. On the other hand, $\boldsymbol{v}$ is tracking its local target $\bar{U}(\boldsymbol{\theta})$ at a speed of $1 - \beta_2 = \mathcal{O}(\eta^\alpha)$. So when $\alpha = 1$, $\boldsymbol{v}$ keeps track of $\bar{U}$ but cannot catch up with it, as illustrated at the bottom left of Figure 2.

When $\alpha > 1$, $\boldsymbol{v}$ can no longer keep track of $\bar{U}(\boldsymbol{\theta})$. However, if we average $\bar{U}(\boldsymbol{\theta})$ along a time period, we will actually find it moving much slower at a speed of $\mathcal{O}(\eta^2)$. We make this point clear below.

Definition 3.2 (Preconditioned Flow Projection)**.** *Fix a point* $\boldsymbol{\theta}_{\text{null}} \notin \Gamma$*. Given a positive definite matrix* $\boldsymbol{S}$*, for* $\boldsymbol{x} \in \mathbb{R}^d$ *consider the preconditioned gradient flow* $\dot{\boldsymbol{x}}(t) = -\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{x}(t))$ *with* $\boldsymbol{x}(0) = \boldsymbol{x}$*. We define* $\Phi_{\boldsymbol{S}}(\boldsymbol{x}) := \lim_{t\to\infty}\boldsymbol{x}(t)$ *if the limit exists and belongs to* Γ*, and set* $\Phi_{\boldsymbol{S}}(\boldsymbol{x}) = \boldsymbol{\theta}_{\text{null}}$ *otherwise. We refer to* $\Phi_{\boldsymbol{S}}(\boldsymbol{x})$ *simply as the* preconditioned flow projection *when the choices of* $\boldsymbol{S}$ *and* $\mathcal{L}$ *is clear from context.*

Consider first the evolution of $\boldsymbol{\theta}$. Each update is based on a stochastic gradient $\nabla\ell(\boldsymbol{\theta};\xi)$, which consists of the full gradient $\nabla\mathcal{L}(\boldsymbol{\theta})$ and the gradient noise $\nabla\ell(\boldsymbol{\theta};\xi) - \nabla\mathcal{L}(\boldsymbol{\theta})$. While the gradient noise keeps driving $\boldsymbol{\theta}$ away from Γ, the full-gradient part points towards the minimizer manifold and brings $\boldsymbol{\theta}$ back close to Γ within $\mathcal{O}(\eta^{-1})$ iterations. Repeating this process many times, the effect of the zero-mean gradient noise nearly vanishes by concentration, and we focus on the effect driven by the full gradient below.

At each step k, the full gradient is preconditioned by $S(\boldsymbol{v}_k)$ which gives $-\eta S(\boldsymbol{v}_k)\nabla\mathcal{L}(\boldsymbol{\theta}_k)$, of which the momentum is used to update the parameter $\boldsymbol{\theta}_k$. Notice that $S(\boldsymbol{v}_k)\nabla\mathcal{L}(\boldsymbol{\theta}_k)$ is exactly the direction given by the preconditioned flow, which keeps $\boldsymbol{\zeta}_k := \Phi_{S(\boldsymbol{v}_k)}(\boldsymbol{\theta}_k)$ unchanged by Definition 3.2. But here the update is discrete instead of a flow, so by a Taylor's expansion of $\boldsymbol{\zeta}_k$ we can derive that $\boldsymbol{\zeta}_k$ moves along Γ with a speed of $\mathcal{O}(\eta^2)$, since the first order term is canceled out.

This suggests we use the smoothed local target $V(\boldsymbol{\Sigma}(\boldsymbol{\zeta}_k))$, which proves to be an approximation of the short-term Exponential Moving Average (EMA) of $\bar{U}(\boldsymbol{\theta})$ within $\mathcal{O}(\eta^{-\alpha})$ iterations, to describe the averaged motion of $\bar{U}(\boldsymbol{\theta})$.

Definition 3.3. *For any* $\boldsymbol{\zeta} \in \Gamma$*, define the smoothed local target* $U(\boldsymbol{\zeta}) := V(\boldsymbol{\Sigma}(\boldsymbol{\zeta}))$.

Lemma 3.1. $\bar{U}(\boldsymbol{\zeta}) = U(\boldsymbol{\zeta})$ *since* $\nabla\mathcal{L}(\boldsymbol{\zeta}) = \mathbf{0}$. *Under schemes* $(1, 2]$, *for fixed* $T > 0$,

$$\sup_{n \le T\eta^{-2}} \left\| \sum_{k=0}^{n-1} \beta_2^{n-1-k}(1-\beta_2)(\bar{U}(\boldsymbol{\theta}_k) - U(\boldsymbol{\zeta}_k)) \right\| = o_{\mathbb{P}}(1).$$

This is a result in the proof of Lemma F.20.

Now we revisit the dynamics of $\boldsymbol{v}$ where $\alpha > 1$. When $1 < \alpha < 2$, $\boldsymbol{v}$ tracks $\bar{U}(\boldsymbol{\theta})$ slower, but still more quickly than $U(\boldsymbol{\zeta})$ moves away, where $\boldsymbol{\zeta} = \Phi_{\boldsymbol{S}(\boldsymbol{v})}(\boldsymbol{\theta})$ is the preconditioned flow projection of $\boldsymbol{\theta}$ onto Γ, and $U(\boldsymbol{\zeta})$ proves to be the short-term Exponential Moving Average (EMA) of $\bar{U}(\boldsymbol{\theta})$. As a result, $\boldsymbol{v}$ nearly keeps up with $U(\boldsymbol{\zeta})$, as depicted at the bottom center of Figure 2. When $\alpha = 2$, the speed of $\boldsymbol{v}$ becomes the same order as $U(\boldsymbol{\zeta})$, making it unable to catch up with $U(\boldsymbol{\zeta})$, as shown at the bottom right of Figure 2.

To summarize, the key intuition is to compare three speeds: $\mathcal{O}(\eta^1), \mathcal{O}(\eta^\alpha), \mathcal{O}(\eta^2)$, corresponding respectively to how quickly $\boldsymbol{\theta}$ returns close to Γ, how quickly $\boldsymbol{v}$ responds to its target, and how quickly the projected point $\boldsymbol{\zeta}$ moves along Γ. If $\boldsymbol{v}$ responds quickly enough, it tracks the local target $\bar{U}(\boldsymbol{\theta})$; if not, it tracks the smoothed local target $U(\boldsymbol{\zeta})$.

3.3 Slow SDE for Adam under different schemes

For each $\eta > 0$, we describe the AGM training process by a stochastic process $\{(\boldsymbol{\theta}_\eta(t), \boldsymbol{v}_\eta(t))\}$, where $t > 0$ is the timestamp. In the sequel, when the context is clear, we omit t and the subscript η. We now present the formal formulation of our results. The slow SDEs under the 1-scheme and schemes $(1, 2)$ are different as discussed and therefore should be introduced separately.

3.3.1 The 1-Scheme

In this section, we introduce our main results under the 1-scheme, i.e. $1 - \beta_2 = c_\beta\eta$ for some constant $c_\beta > 0$. In this regime, $\boldsymbol{\theta}$ and $\boldsymbol{v}$ have the same velocity scale, so we must describe their dynamics together. To establish the slow SDE characterization of AGM process, we first give a rigorous definition of the coupled preconditioned projection of $\boldsymbol{\theta}$ and $\boldsymbol{v}$, as an extension of the preconditioned flow projection in Li et al. (2025).

Definition 3.4 (Coupled Preconditioned Flow Projection)**.** *Consider the following coupled ODE*

$$\dot{\boldsymbol{\theta}} = -S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta}), \qquad \dot{\boldsymbol{v}} = c_\beta\left(\bar{U}(\boldsymbol{\theta}) - \boldsymbol{v}\right). \tag{1}$$

Fix a $\boldsymbol{\zeta}_{\text{null}} \in \Gamma$. *For a state* $\boldsymbol{x} = (\boldsymbol{\theta}, \boldsymbol{v}) \in \mathbb{R}^{d+D}$, *let* $(\boldsymbol{\theta}_{\boldsymbol{x}}(s), \boldsymbol{v}_{\boldsymbol{x}}(s))$ *denote the solution of the above ODE with initial state* $\boldsymbol{x}$. *Define the coupled preconditioned flow projection* $\Pi(\cdot)$ *as*

$$\Pi(\boldsymbol{x}) = (\pi(\boldsymbol{x}), U(\pi(\boldsymbol{x}))),$$

if the coupled solution of the ODE exists for all $s \ge 0$ *and converges to* $(\boldsymbol{\zeta}, U(\boldsymbol{\zeta}))$ *for some* $\boldsymbol{\zeta} = \pi(\boldsymbol{x}) \in \Gamma$. *Otherwise, let* $\pi(\boldsymbol{x}) = \boldsymbol{\zeta}_{\text{null}}$ *and* $\Pi(\boldsymbol{x}) := (\boldsymbol{\zeta}_{\text{null}}, U(\boldsymbol{\zeta}_{\text{null}}))$.

To see why we refer the mapping $\Pi(\cdot)$ as a projection, we define the coupled manifold $\mathcal{M} := \{(\boldsymbol{\zeta}, U(\boldsymbol{\zeta})) : \boldsymbol{\zeta} \in \Gamma\}$. One could notice that, by definition, $\Pi(\cdot)$ is the projection function that projects the coupled state $(\boldsymbol{\theta}, \boldsymbol{v})$ to $\mathcal{M}$ following the ODE (1). The local existence and smoothness of this projection are established in Appendix E.1.

If the ODE (1) start with some point $(\boldsymbol{\zeta}, U(\boldsymbol{\zeta})) \in \mathcal{M}$, both updates of $\boldsymbol{\theta}$ and $\boldsymbol{v}$ vanish since $\nabla\mathcal{L}(\boldsymbol{\zeta}) = \mathbf{0}$ and $U(\boldsymbol{\zeta}) = \bar{U}(\boldsymbol{\zeta})$. However, stochastic updates can still move the state away from $\mathcal{M}$, which leads to the slow SDE we study. Now we give the formulation of the slow SDE of the 1-scheme.

Definition 3.5. *Following the notation of Li et al. (2025, Definition 4.2), we introduce an Itô projection operator* $\boldsymbol{P}_{\boldsymbol{\zeta}}$ *associated with the coupled flow. For any* $\boldsymbol{\zeta} \in \Gamma$ *and any differential form* $\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{a}\mathrm{d}t$ *in Itô calculus, where* $\boldsymbol{A} \in \mathbb{R}^{(d+D)\times(d+D)}$, *and* $\boldsymbol{a} \in \mathbb{R}^{d+D}$, *we use* $P_{\boldsymbol{\zeta}}\left(\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{a}\mathrm{d}t\right)$ *as a shorthand for the differential form* $\partial\pi(\iota(\boldsymbol{\zeta}))\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \left(\partial\pi(\iota(\boldsymbol{\zeta}))\boldsymbol{a} + \frac{1}{2}\partial^2\pi(\iota(\boldsymbol{\zeta}))[\boldsymbol{A}\boldsymbol{A}^\top]\right)\mathrm{d}t$, *where* $\iota(\boldsymbol{\zeta}) := (\boldsymbol{\zeta}, U(\boldsymbol{\zeta}))$.

Definition 3.6 (Slow SDE under the 1-scheme)**.** *Given an initial state* $\boldsymbol{\zeta}_0 \in \Gamma$, *we define the slow SDE under the* 1*-scheme as*

$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}(t)}\left(\boldsymbol{C}^{1/2}_{\boldsymbol{\zeta}(t)}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}_{\boldsymbol{\zeta}(t)}\mathrm{d}t\right). \tag{2}$$

where $\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0$. *The diffusion term* $\boldsymbol{C}_{\boldsymbol{\zeta}}^{1/2}$ *and the drift term* $\boldsymbol{b}_{\boldsymbol{\zeta}}$ *are defined with* $\boldsymbol{S}_*(\boldsymbol{\zeta}) := S(U(\boldsymbol{\zeta}))$ *and* $\boldsymbol{h}(\boldsymbol{\theta}, \xi) := V\left(\nabla\ell(\boldsymbol{\theta};\xi)\nabla\ell(\boldsymbol{\theta};\xi)^\top\right)$ *as below:*

$$\boldsymbol{C}_{\boldsymbol{\zeta}} := \mathbb{E}_{\xi\sim\mathcal{S}}\left[\boldsymbol{q}(\boldsymbol{\zeta},\xi)\boldsymbol{q}(\boldsymbol{\zeta},\xi)^\top\right], \qquad \boldsymbol{q}(\boldsymbol{\zeta},\xi) := \begin{pmatrix} -\boldsymbol{S}_*(\boldsymbol{\zeta})\nabla\ell(\boldsymbol{\zeta};\xi) \\ c_\beta\left(\boldsymbol{h}(\boldsymbol{\zeta},\xi) - U(\boldsymbol{\zeta})\right)\end{pmatrix},$$

$$\boldsymbol{b}_{\boldsymbol{\zeta}} := \begin{pmatrix} -c_\beta\mathbb{E}_{\xi\sim\mathcal{S}}\left[\partial S(U(\boldsymbol{\zeta}))[\boldsymbol{h}(\boldsymbol{\zeta},\xi) - U(\boldsymbol{\zeta})]\nabla\ell(\boldsymbol{\zeta};\xi)\right] \\ \mathbf{0}_D \end{pmatrix} \in \mathbb{R}^{d+D}.$$

The matrix $\boldsymbol{C}_{\boldsymbol{\zeta}}$ is the coupled covariance of $\boldsymbol{\theta}$ and $\boldsymbol{v}$ update generated by the sampled gradient at $\iota(\boldsymbol{\zeta})$. It records the variability of each update and their correlations. Through the curvature of the coupled projection Π, it contribute the drift $\frac{1}{2}\partial^2\pi[\boldsymbol{C}_{\boldsymbol{\zeta}}]$. The vector $\boldsymbol{b}_{\boldsymbol{\zeta}}$ is the mean second order correction caused by updating the preconditioner with the sampled gradient at $\iota(\boldsymbol{\zeta})$ whose contribution to the slow SDE is $\partial\pi\boldsymbol{b}_{\boldsymbol{\zeta}}$. The following theorem makes this description precise by comparing expectations of smooth observables over a fixed interval of slow time.

Theorem 3.2 (Weak approximation for the 1-scheme)**.** *Let* $T > 0$ *and suppose Assumption 2.1–2.4 hold. Additionally, assume that there exists an open neighborhood* $\mathcal{N}$ *of* Γ *such that for every* ξ*, the sample loss* $\ell(\cdot;\xi)$ *is* $\mathcal{C}^3$*-smooth on* $\mathcal{N}$*. There exists* $\epsilon_0, C, M > 0$ *such that the following statements hold:*

Define step $K_0 := \lfloor C\eta^{-1}\log(1/\eta)\rfloor$ *and* $K_\eta := \lfloor T\eta^{-2}\rfloor$*. Assume that the initialization* $(\boldsymbol{\theta}_0, \boldsymbol{m}_0, \boldsymbol{v}_0)$ *satisfies that* $\operatorname{dist}((\boldsymbol{\theta}_0, \boldsymbol{v}_0), \mathcal{M}) \le \epsilon_0$*,* $\boldsymbol{v}_0 \in [0, R_1]^D$ *and* $\|\boldsymbol{m}_0\| \le M$*. Run AGM for* $K_0 + K_\eta$ *iterations and obtain* $\{(\boldsymbol{\theta}_0, \boldsymbol{v}_0)\}_{k=0}^{K_0+K_\eta}$*. Let* $\boldsymbol{\zeta}(t)$ *be the solution to* (2) *on* $t \in [0, T]$ *with initial condition* $\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0 := \pi(\boldsymbol{\theta}_{K_0}, \boldsymbol{v}_{K_0})$*, and let* $\boldsymbol{X}(t) = \iota(\boldsymbol{\zeta}(t)) := (\boldsymbol{\zeta}(t), U(\boldsymbol{\zeta}(t)))$*. For any* $k \in \{0, 1, \cdots, K_\eta\}$*, let* $\bar{\boldsymbol{X}}_k = \Pi(\boldsymbol{\theta}_{K_0+k}, \boldsymbol{v}_{K_0+k})$ *be the AGM state at step* $K_0 + k$ *projected onto* $\mathcal{M}$*. Then for any* C^3 *function* $g : \mathbb{R}^{d+D} \to \mathbb{R}$*,*

$$\max_{0\le k\le K_\eta}\left|\mathbb{E}\left[g(\bar{\boldsymbol{X}}_k)\right] - \mathbb{E}\left[g(\boldsymbol{X}(k\eta^2))\right]\right| \to 0 \quad \text{as } \eta \to 0.$$

Theorem 3.2 shows that with sufficiently small η, the long-horizon behavior within $\mathcal{O}(\eta^2)$ iterations of AGM after approaching the minimizer manifold is described by the slow SDE in Equation (2). The detailed proof of Theorem 3.2 is provided in Appendix E and the proof sketch can be found in Appendix C.

3.3.2 SCHEMES $(1, 2)$

In this subsubsection, we consider schemes $(1, 2)$. In this regime, $\boldsymbol{\theta}$ fluctuates around the manifold Γ so fast that the normal displacement of $\boldsymbol{\theta}$ cancels out when $\boldsymbol{v}$ tracks its target. Therefore, the preconditioner $S(\boldsymbol{v})$ is approximately frozen during the recovery of each normal displacement over slow time. The relevant flow can therefore be viewed as a deterministic flow $\Phi_{\boldsymbol{S}_*(\boldsymbol{\zeta})}$. We now give the formulation of the slow SDE under schemes $(1, 2)$.

Definition 3.7. *Similar to Definition 3.5, we define the Itô preconditioned projection operator* $\boldsymbol{P}_{\boldsymbol{\zeta},\boldsymbol{S}}$*. For any* $\boldsymbol{\zeta} \in \Gamma$*, a fixed positive definite matrix* $\boldsymbol{S} \in \mathbb{S}^d_{++}$ *and any differential form* $\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{a}\mathrm{d}t$ *in Itô calculus, where* $\boldsymbol{A} \in \mathbb{R}^{d\times d}$*, and* $\boldsymbol{a} \in \mathbb{R}^d$*, we use* $\boldsymbol{P}_{\boldsymbol{\zeta},\boldsymbol{S}}(\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{a}\mathrm{d}t)$ *as a shorthand for the differential form* $\partial\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \left(\partial\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})\boldsymbol{a} + \frac{1}{2}\partial^2\Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})[\boldsymbol{A}\boldsymbol{A}^\top]\right)\mathrm{d}t$*.*

Definition 3.8 (Slow SDE under schemes $(1, 2)$)**.** *Given an initial state* $\boldsymbol{\zeta}_0 \in \Gamma$*, we define the slow SDE under schemes* $(1, 2)$ *as:*

$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}(t),\boldsymbol{S}_*(\boldsymbol{\zeta}(t))}\left(\boldsymbol{S}_*(\boldsymbol{\zeta}(t))\sigma(\boldsymbol{\zeta}(t))\mathrm{d}\boldsymbol{W}_t\right). \tag{3}$$

The drift supplied to $\boldsymbol{P}_{\boldsymbol{\zeta},\boldsymbol{S}_*(\boldsymbol{\zeta})}$ vanishes since the contribution of the update of $S(\boldsymbol{v})$ is $\mathcal{O}(\eta(1-\beta_2)) = o(\eta^2)$ per iteration while for a fixed preconditioner $\boldsymbol{S}$, invariance of $\Phi_{\boldsymbol{S}}$ along the preconditioned flow generated by $\boldsymbol{S}\nabla\mathcal{L}$ gives $\partial\Phi_{\boldsymbol{S}}(\boldsymbol{\theta})\boldsymbol{S}\nabla\mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}$.

Theorem 3.3 (Weak approximation under schemes $(1, 2)$)**.** *Let* $T > 0$ *and suppose Assumption 2.1–2.4 hold. There exist constants* $\epsilon_0, C, M > 0$ *such that the following statement holds: Define step* $K_0 := \lfloor C\eta^{-\alpha}\log(1/\eta)\rfloor$ *and* $K_\eta := \lfloor T\eta^{-2}\rfloor$*. Run an AGM for* $K_0 + K_\eta$ *iterations and obtain* $\{(\boldsymbol{\theta}_k, \boldsymbol{v}_k)\}_{k=0}^{K_0+K_\eta}$*, where the initial parameters satisfy* $\operatorname{dist}(\boldsymbol{\theta}_0, \Gamma) \le \epsilon_0$*,* $\boldsymbol{v}_0 \in [0, R_1]^D$ *and*

$\|\boldsymbol{m}_0\| \le M$. *Let* $\boldsymbol{\zeta}(t)$ *be the solution to* (3) *on* $t \in [0,T]$ *with initial condition* $\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0 := \Phi_{\boldsymbol{S}(\boldsymbol{v}_{K_0})}(\boldsymbol{\theta}_{K_0})$. *For any* $k \in \{0,1,\dots,K_\eta\}$, *let* $\bar{\boldsymbol{X}}_k := \left(\Phi_{\boldsymbol{S}(\boldsymbol{v}_{K_0+k})}(\boldsymbol{\theta}_{K_0+k}), \boldsymbol{v}_{K_0+k}\right)$ *be the AGM state at step* K_0+k *projected onto* Γ. *Let* $\boldsymbol{X}(t) := (\boldsymbol{\zeta}(t), U(\boldsymbol{\zeta}(t)))$ *be the state of the SDE* (3) *at time* t. *Then for any* C^3 *function* $g : \mathbb{R}^{d+D} \to \mathbb{R}$,

$$\max_{0\le k\le K_\eta} \left|\mathbb{E}\left[g\left(\bar{\boldsymbol{X}}_k\right)\right] - \mathbb{E}\left[g\left(\boldsymbol{X}(k\eta^2)\right)\right]\right| \to 0 \quad \text{as } \eta \to 0.$$

Theorem 3.3 extends the results of Theorem 3.2 to schemes $(1,2)$, as it describes the long horizon behavior of AGMs within $\mathcal{O}(\eta^2)$ iterations near the minimizer manifold with small η by the slow SDE given in Equation (3). The two theorems, together with the results under the 2-scheme given in Li et al. (2025) provide a fine-grained analysis of the slow SDE of AGMs within $\mathcal{O}(\eta^{-2})$ iterations near the minimizer manifold under schemes $[1,2]$. The detailed proof of Theorem 3.3 is provided in Appendix F and the proof sketch can also be found in Appendix C.

4 Implicit Biases for Different Schemes under Label Noise Condition

In this section, we present Adam's implicit bias under the label noise conditions. The main result is that the 1-scheme has an extra third-order term, while schemes $(1,2)$ have the same implicit bias as the 2-scheme studied by Li et al. (2025). We first recall the definition of the *label noise* condition in Blanc et al. (2020).

Label Noise Condition. *label noise* represents the condition that for all $\boldsymbol{\theta} \in \Gamma$, the covariance matrix $\boldsymbol{\Sigma}$ satisfies $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \tau\nabla^2\mathcal{L}(\boldsymbol{\theta})$ for some constant $\tau > 0$. This condition typically occurs when training an overparameterized model that can fit the training data perfectly, but where each step of training introduces new noise into the target labels. See Li et al. (2025, Section 5.1) for a quick derivation where introducing zero-mean noise to the label in a simple regression task can lead to this condition. This condition greatly simplifies the analysis of the implicit bias of SGD and related optimizers as adopted by Blanc et al. (2020), Damian et al. (2021), Li et al. (2021b), Gu et al. (2023) and Li et al. (2025).

Under the label noise condition, Li et al. (2025) introduce the implicit bias of Adam under the 2-scheme. Before we introduce the implicit bias under schemes $[1,2)$, we first recall some notations. For any point $\boldsymbol{\zeta} \in \Gamma$, the hessian matrix of it is defined as $\boldsymbol{H}_{\boldsymbol{\zeta}} := \nabla^2\mathcal{L}(\boldsymbol{\zeta})$. Recall the Sharpness Regularizer in the 2-scheme (Li et al., 2025), which is denoted as

$$\mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}) := \begin{cases} \operatorname{tr}\Big(\operatorname{Diag}(\boldsymbol{H}_{\boldsymbol{\zeta}})^{1/2} - \frac{\epsilon}{\sqrt{\tau}}\log\Big(\frac{\sqrt{\tau}}{\epsilon}\operatorname{Diag}(\boldsymbol{H}_{\boldsymbol{\zeta}})^{1/2} + \boldsymbol{I}\Big)\Big), & \epsilon > 0, \\ \operatorname{tr}\big(\operatorname{Diag}(\boldsymbol{H}_{\boldsymbol{\zeta}})^{1/2}\big), & \epsilon = 0, \end{cases}$$

where $\epsilon \ge 0$ is the ϵ term in Adam (Kingma & Ba, 2014). We now give the informal version of the implicit bias of 1-scheme and scheme $(1,2)$.

Theorem 4.1 (Implicit Bias of Adam under Schemes $[1,2)$, Stated Informally)**.** *Under the label noise condition, Adam's limiting dynamics have the following properties:*

- *For* $1 < \alpha < 2$, *every fixed point* $\boldsymbol{\zeta}^*$ *satisfies* $\nabla_\Gamma\mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}^*) = \boldsymbol{0}$.
- *For* $\alpha = 1$, *every fixed point* $\boldsymbol{\zeta}^*$ *satisfies* $\nabla_\Gamma\mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}^*) = \boldsymbol{d}_{\epsilon,\tau}(\boldsymbol{\zeta}^*)$, *where* $\boldsymbol{d}_{\epsilon,\tau}(\boldsymbol{\zeta}^*) \not\equiv \boldsymbol{0}$ *is an additional drift induced by third moments of gradients that is not necessarily* $\boldsymbol{0}$. *(See the formal definition of* $\boldsymbol{d}(\boldsymbol{\zeta}^*)$ *in Lemma G.1).*

See Appendix G for the formal statements and proofs.

We now give an intuitive interpretation of the theorem in terms of the loss minimizers selected by Adam. In particular, we explain how the additional drift in the 1-scheme can change where the limiting dynamics stop.

For $1 < \alpha < 2$**, Adam's limiting dynamics reduce sharpness under label noise.** They decrease $\mathcal{R}_{\epsilon,\tau}$ along Γ, favoring loss minimizers with smaller values of this sharpness measure. When $\epsilon = 0$,

it takes the simpler form $\mathrm{tr}(\mathrm{Diag}(\boldsymbol{H}_{\boldsymbol{\zeta}})^{1/2})$. If these dynamics converge to a point $\boldsymbol{\zeta}^*$, it must satisfy $\nabla_\Gamma \mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}^*) = \mathbf{0}$. This is the same fixed point condition as in the 2-scheme.

Under the 1-scheme, third order terms can drive Adam away from sharpness minima. At a local minimum of sharpness along Γ, the sharpness gradient vanishes, but the additional drift $\boldsymbol{d}_{\epsilon,\tau}(\boldsymbol{\zeta}^*)$ under the 1-scheme can still move the parameters. The dynamics may instead settle at another loss minimizer where this drift is canceled by the update toward lower sharpness. To understand why this additional drift appears, recall that Adam uses the second moment variable $\boldsymbol{v}$ to rescale the parameter update. Since the same stochastic gradient enters the parameter update and contributes its squared coordinates to $\boldsymbol{v}$, this rescaling produces terms whose expectations involve third moments. The second moment update changes $\boldsymbol{v}$ by order $1-\beta_2$, which also changes the preconditioner by this order. This change contributes $-\eta\big(S(\boldsymbol{v}_{k+1}) - S(\boldsymbol{v}_k)\big)\boldsymbol{m}_{k+1}$ to the parameter update, giving a correction of order $\eta(1-\beta_2)$. Over $\mathcal{O}(\eta^{-2})$ iterations, their accumulated scale is $\eta^{-2}\eta(1-\beta_2) = c_\beta \eta^{\alpha-1}$. They can therefore contribute to the limiting dynamics when $\alpha = 1$, whereas their contribution vanishes when $\alpha > 1$.

4.1 Illustration on a Constant-Sharpness Manifold

We revisit Figure 1 as a demonstration of the distinction between schemes $(1, 2)$ and the 1-scheme indicated by Theorem 4.1. Our loss construction starts from two constraints, $g_1 = 0$ defining an elliptic cylinder and $g_2 = 0$ defining a plane, whose intersection is the minimizer ellipse Γ. We then multiply both g_1 and g_2 by a smooth, strictly positive weight function $w(\boldsymbol{\theta})$, obtaining the weighted least-squares term $\mathcal{L}_w(\boldsymbol{\theta}) = \frac{1}{2}[(w(\boldsymbol{\theta})g_1(\boldsymbol{\theta}))^2 + (w(\boldsymbol{\theta})g_2(\boldsymbol{\theta}))^2]$. This preserves Γ as the minimizer manifold.

Now consider the Hessian matrix on Γ. Since $g_1 = g_2 = 0$ on Γ, any derivative of w will vanish in Hessian, giving $H_{ii} = w^2[(\partial_i g_1)^2 + (\partial_i g_2)^2]$. Choosing $w = (\sum_i \sqrt{(\partial_i g_1)^2 + (\partial_i g_2)^2})^{-1}$ therefore makes $\mathcal{R}_{0,\tau} = \sum_i \sqrt{H_{ii}} = 1$ along Γ. Thus $\nabla_\Gamma \mathcal{R}_{0,\tau} = \mathbf{0}$, and the limiting dynamics for schemes $(1, 2]$ are stationary on Γ, consistent with the yellow trajectories in Figure 1. In contrast, the additional drift $\boldsymbol{d}_{0,\tau}$ of the 1-scheme need not vanish on Γ. Its limiting dynamics stop only where $\boldsymbol{d}_{0,\tau} = \nabla_\Gamma \mathcal{R}_{0,\tau} = \mathbf{0}$. Figure 3 displays the magnitude of this difference and marks its zeros, toward which the blue trajectories drift.

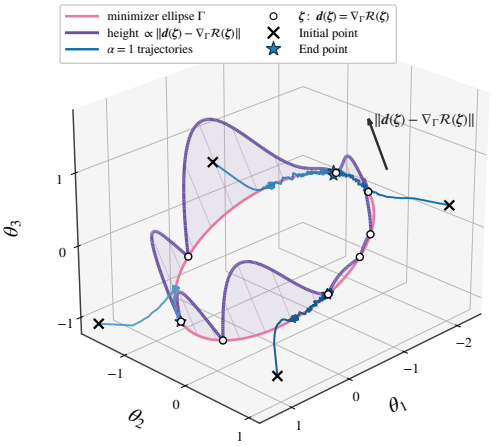


Figure 3: In Figure 1 's setting, we display $\|\boldsymbol{d}(\boldsymbol{\zeta}) - \nabla_\Gamma \mathcal{R}(\boldsymbol{\zeta})\|$ by purple curves. Its zeros (white circles) satisfy the 1-scheme fixed-point condition in Theorem 4.1, and the blue trajectories drift toward these points.

5 Conclusion

In this work, we derive the fine-grained analysis of the AGM's slow SDE and Adam's implicit bias near the minimizer manifold under schemes $[1, 2)$. We show that the 1-scheme has an additional drift that leads to a differnt implicit bias while schemes $(1, 2)$ recover the sharpness characterization of Li et al. (2025). Despite these results, several questions remain for future work. Although our local convergence guarantees also cover schemes $[0.5, 1)$, characterizing the limiting dynamics under these schemes remains an open problem for future work. Another direction is to analyze different hyperparameter scaling. For example, the regime where $1 - \beta_1$ decreases with η, rather than keeping β_1 bounded away from one. Finally, our analysis cannot cover optimizers such as Muon (Jordan et al., 2024), which approximately orthogonalizes the update matrix. Characterizing how this transformation alters the limiting dynamics and whether it induces a different effective sharpness regularizer is an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with checking mathematical claims, theoretical derivations, and proofs, and with analyzing and interpreting experimental results. We did not use generative AI tools to design experiments or implement methods. Additionally, we used these tools for manuscript drafting and editing, figure preparation, literature searches, and LaTeX formatting. We reviewed all AI-assisted work, checking the text for accuracy, comparing experimental analyses with the original results, and examining theoretical derivations and proofs for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 483–513. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/blanc20a.html.

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

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27449–27461. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e6af401c28c1790eaef7d55c92ab6ab6-Paper.pdf.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

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.

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

Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pp. 2315–2357. PMLR, 2021.

Morris W. Hirsch, Charles C. Pugh, and Michael Shub. *Invariant Manifolds*, volume 583 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 1977. doi: 10.1007/BFb0092042.

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.

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.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Gary S. Katzenberger. Solutions of stochastic differential equations forced onto a manifold by a large drift. *The Annals of Probability*, 19(4):1587–1628, 1991. doi: 10.1214/aop/1176990225.

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.

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.

Xinghan Li, Haodong Wen, and Kaifeng Lyu. Adam reduces a unique form of sharpness: Theoretical insights near the minimizer manifold. In *Advances in Neural Information Processing Systems*, 2025.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 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.

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

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.

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.

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 (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 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.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 71437–71464. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/837bc5db12f3d394d220815a7687340c-Paper-Conference.pdf.

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.

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*, pp. 10849–10858. PMLR, 2021.

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.

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*, pp. 3635–3673. PMLR, 2020.

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.

CONTENTS

1 Introduction **1**

2 Preliminaries **3**

2.1 The Loss Function and its Minimizer Manifold 3

2.2 Adaptive Gradient Methods and their Schemes 3

3 Main Results **4**

3.1 Convergence Guarantee for AGMs under schemes [0.5, 2] 4

3.2 An Illustration 4

3.3 Slow SDE for Adam under different schemes 6

3.3.1 The 1-Scheme 6

3.3.2 Schemes $(1, 2)$ 7

4 Implicit Biases for Different Schemes under Label Noise Condition **8**

4.1 Illustration on a Constant-Sharpness Manifold 9

5 Conclusion **9**

A Related Work **15**

B Detailed Background of Katzenberger's Theorem **15**

B.1 Intuition of Limiting Flow Near Manifold From Katzenberger (1991) 15

B.2 Mathematical formulation of Katzenberger (1991) 16

C Our Approach: Multiscale Manifold Reduction **17**

D Proof of the convergence result **18**

D.1 Proof of the convergence theorem 22

E The 1-scheme **23**

E.1 Joint flow and projection 23

E.2 Momentum reduction 25

E.3 Weak convergence 29

F Schemes $(1, 2)$ **35**

F.1 Working zone and adaptive projection 35

F.2 Preliminaries for the proof 38

F.3 Semimartingale decomposition 41

F.4 Collapse to the minimizer manifold 43

F.5 Normal-space averaging 45

F.6 Projection reduction and fast tracking of the preconditioner 53

F.6.1 Fast tracking of the preconditioner 57
F.7 Local characteristics and identification of the limiting SDE 59
F.7.1 Martingale problem and identification of the limiting SDE 61
F.8 Completion of the proof of the main theorem 65

G Implicit bias under label noise **65**
G.1 The 1-scheme 66
G.2 Proof for The 1-scheme 66
G.3 Schemes $(1, 2)$ 68
G.4 Proof for Schemes $(1, 2)$ 69

A Related Work

Implicit Bias of Adam. Previous analyzes on implicit bias of optimizers have been conducted primarily with SGD (Blanc et al., 2020; Woodworth et al., 2020; HaoChen et al., 2021; Damian et al., 2021; Li et al., 2021b; Gu et al., 2023; Wang et al., 2023). Compared with the vast literature of SGD, the implicit bias of Adam still lacks a thorough theoretical understanding, despite the widespread use of Adam in practical deep learning, especially for large language models. However, many efforts have already been made to address this issue. Wang et al. (2021) identified the same implicit regularizer for SGD and Adam under the assumption that all gradient coordinates have magnitude at most ϵ. However, this assumption is typically unfeasible in practice. Using IGR, Cattaneo et al. (2024) showed that full-batch Adam with constant learning rate approximately follows an ODE which anti-regularizes sharpness when $\beta_1 < \beta_2$. In contrast, Li et al. (2025) characterized Adam's long-term dynamics by showing that with gradient noise, Adam can be characterized by a slow SDE that regularizes sharpness in the long term. They further proved that Adam reduces $\mathrm{tr}\left(\mathrm{Diag}(\mathrm{H})^{1/2}\right)$ under the label noise condition.

SDE Approximations and Slow Dynamics. SDE approximations provide continuous models for the discrete updates of SGD and adaptive methods (Li et al., 2018; 2021a; Malladi et al., 2024). For implicit bias near a minimizer manifold, the relevant question is how stochastic updates affect the parameters after the training loss has become small. In the regimes considered here, motion along the manifold accumulates over $\mathcal{O}(\eta^{-2})$ iterations, so an approximation valid only on a shorter horizon does not by itself characterize this behavior. For this longer horizon, Li et al. (2021b) adapted the framework of Katzenberger (1991) to derive a limiting SDE for SGD near a minimizer manifold. By projecting the iterates onto the manifold, their analysis characterizes how gradient noise and the geometry of the loss determine the motion among minimizers. Building on this perspective, Gu et al. (2023) established quantitative weak-approximation guarantees for slow SDEs over $\mathcal{O}(\eta^{-2})$ iterations, and used them to compare the implicit bias of local SGD and SGD. We also adopt the framework of Katzenberger (1991) in our analysis for AGMs under schemes $[1, 2)$ over $\mathcal{O}(\eta^{-2})$ iterations.

Comparison with Li et al. (2025) The analysis of AGMs near a minimizer manifold by Li et al. (2025) is the closest to ours. They study the 2-scheme, and show that the projected parameter $\boldsymbol{\zeta}$ and adaptive variable $\boldsymbol{v}$ converge jointly to a coupled slow SDE over $\mathcal{O}(\eta^{-2})$ iterations. Our work further refines their results to schemes $[1, 2)$, which is a more practical regime. In their work, they used blockwise moment estimates to obtain quantitative weak-approximation bounds. However, we can hardly obtain quantitative weak-approximation bounds in our settings. We instead adapt the framework of Katzenberger (1991) to establish asymptotic weak convergence, using normal-space averaging to separate the process of $\boldsymbol{\theta}$ from the process of second moment variable $\boldsymbol{v}$ under schemes $(1, 2)$.

Comparison with Katzenberger (1991). Katzenberger (1991) studied stochastic processes with a large deterministic drift and a stochastic perturbation, and characterized the the limiting dynamics. For the 1-scheme, we follow this framework using the coupled flow of $(\boldsymbol{\theta}, \boldsymbol{v})$. For $1 < \alpha < 2$, however, the processes of $\boldsymbol{\theta}$ and $\boldsymbol{v}$ occur on distinct timescales while depending on each other. Our analysis therefore requires additional steps. We show that the accumulated effect of the update of $\boldsymbol{\theta}$ on $\boldsymbol{v}$ vanishes and then show that the accumulated effect of the update of $\boldsymbol{v}$ on the projection $\boldsymbol{\zeta} := \Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})$ vanishes. These estimates allows us to replace $\boldsymbol{v}$ with the smoothed local target $U(\boldsymbol{\zeta})$ so that we can adopt the framework of Katzenberger (1991).

B Detailed Background of Katzenberger's Theorem

B.1 Intuition of Limiting Flow Near Manifold From Katzenberger (1991)

Intuitively, Katzenberger (1991) studied a sequence of stochastic processes $\{\boldsymbol{X}_n(t)\}_{n=1}^{\infty}$ whose dynamics are determined by two effects: a deterministic drift and a stochastic perturbation. The deterministic drift drives the state toward an asymptotically stable manifold of equilibria of the flow, denoted by Γ, while the stochastic perturbation keeps pushing the state away from it. As n increases,

the drift becomes increasingly strong. As a result, whenever the noise moves the state away from Γ, the drift quickly pulls it back. In the limit $n \to \infty$, the process is therefore confined to Γ.

However, the repeated effect of the stochastic perturbation and the deterministic drift does not disappear. Instead, it gives rise to a nontrivial stochastic motion along Γ, which is similar to the motion of $\boldsymbol{\theta}$ Figure 2 discussed in Section 3.2. Katzenberger characterized this limiting process with $n \to \infty$ on the manifold and derive the corresponding SDE.

B.2 Mathematical formulation of Katzenberger (1991)

We now formulate the stochastic processes and the proof ideas behind Katzenberger (1991) mathematically. Katzenberger (1991) studies the limiting behavior of a sequence of stochastic processes $\{\boldsymbol{X}_n\}$ satisfying

$$\boldsymbol{X}_n(t) = \boldsymbol{X}_0 + \int_0^t \sigma(\boldsymbol{X}_n(s-))\,\mathrm{d}\boldsymbol{Z}_n(s) + \int_0^t F(\boldsymbol{X}_n(s-))\,\mathrm{d}A_n(s). \tag{4}$$

The equation decomposes the evolution of $\boldsymbol{X}_n(t)$ into two parts. The first integral represents the accumulated stochastic perturbation, while the second describes the motion generated by a deterministic vector field F. Here $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times e}$, and $\boldsymbol{Z}_n$ is an $\mathbb{R}^e$-valued stochastic process. Thus, the term $\sigma(\boldsymbol{X}_n(s-))\,\mathrm{d}\boldsymbol{Z}_n(s)$ describes how the noise changes the state at time s. Here $\boldsymbol{X}_n(s-)$ denotes the state immediately before time s, which is needed because the stochastic process may have jumps. To understand the second term, first ignore the noise and consider the ODE $\frac{\mathrm{d}\boldsymbol{x}(u)}{\mathrm{d}u} = F(\boldsymbol{x}(u))$. For an initial point $\boldsymbol{x} \in \mathbb{R}^d$, let $\boldsymbol{\phi}_F(\boldsymbol{x}, u)$ denote its solution at time u, so that

$$\partial_u \boldsymbol{\phi}_F(\boldsymbol{x}, u) = F(\boldsymbol{\phi}_F(\boldsymbol{x}, u)), \qquad \boldsymbol{\phi}_F(\boldsymbol{x}, 0) = \boldsymbol{x}.$$

The map $u \mapsto \boldsymbol{\phi}_F(\boldsymbol{x}, u)$ describes the trajectory obtained by following the vector field F from $\boldsymbol{x}$.

The manifold Γ consists of the equilibrium points of this flow. Equivalently,

$$\Gamma = \{\boldsymbol{x} \in \mathbb{R}^d : F(\boldsymbol{x}) = 0\} = \left\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{\phi}_F(\boldsymbol{x}, u) = \boldsymbol{x} \text{ for all } u \geq 0\right\}.$$

We assume that Γ is asymptotically stable, so trajectories starting near Γ are driven toward it by the flow of F. The process A_n controls how quickly this deterministic flow acts and for which we will call it as a *clock*. It is nondecreasing with $A_n(0) = 0$, and for every fixed $T > \varepsilon > 0$,

$$\inf_{0\leq s\leq T-\varepsilon} (A_n(s+\varepsilon) - A_n(s)) \longrightarrow \infty.$$

The clock condition means that the time elapsed along the flow of F during any fixed interval of length ε grows without bound as $n \to \infty$. Since this flow drives nearby states toward Γ, the drift pulls the process back toward the manifold on an increasingly short timescale.

Now suppose that the initial state $\boldsymbol{X}_n(0)$ is already close to the manifold Γ. Recall that $\boldsymbol{Z}_n$ controls the accumulation of stochastic perturbations over time. We are interested in the limiting dynamics of $\boldsymbol{X}_n$ when $\boldsymbol{Z}_n$ converges in distribution to a Brownian motion $\boldsymbol{W}$ as $n \to \infty$.

Since the drift vanishes on Γ, one might naively expect that the limiting dynamics are determined only by the noise projected onto the manifold. The key observation of Katzenberger (1991) is that this **is not quite correct**. A stochastic perturbation first moves the state slightly away from Γ, after which the strong drift almost instantaneously pulls it back. The combination of these two effects can produce an additional deterministic motion along the manifold.

To describe this effect, define

$$\Phi_F(\boldsymbol{x}) = \lim_{u\to\infty} \boldsymbol{\phi}_F(\boldsymbol{x}, u),$$

which is the point on Γ reached by following the flow generated by F from $\boldsymbol{x}$. Thus, Φ_F maps a point near Γ to the equilibrium point toward which the deterministic flow drives it.

Consider a state $\boldsymbol{X}_n(t)$ close to Γ. During an infinitesimal time interval $\mathrm{d}t$, the stochastic perturbation changes the state by

$$\delta = \sigma(\boldsymbol{X}_n(t))\,\mathrm{d}\boldsymbol{W}_t.$$

The strong drift then rapidly sends the perturbed state $\boldsymbol{X}_n(t)+\delta$ to $\Phi_F(\boldsymbol{X}_n(t)+\delta)$. Expanding Φ_F around $\boldsymbol{X}_n(t)$ gives

$$\Phi_F(\boldsymbol{X}_n(t)+\delta)-\Phi_F(\boldsymbol{X}_n(t)) \approx \partial\Phi_F(\boldsymbol{X}_n(t))\,\delta+\frac{1}{2}\partial^2\Phi_F(\boldsymbol{X}_n(t))[\delta,\delta].$$

Since δ is of order $\sqrt{\mathrm{d}t}$, the first term is of order $\sqrt{\mathrm{d}t}$ while the second term is of order $\mathrm{d}t$. Using $\delta=\sigma(\boldsymbol{X}_n(t))\,\mathrm{d}\boldsymbol{W}_t$ and the Itô relation $\mathrm{d}\boldsymbol{W}_t\,\mathrm{d}\boldsymbol{W}_t^\top=\boldsymbol{I}\,\mathrm{d}t$, we obtain

$$\begin{aligned}\Phi_F(\boldsymbol{X}_n(t)+\delta)-\Phi_F(\boldsymbol{X}_n(t)) \approx{}& \partial\Phi_F(\boldsymbol{X}_n(t))\sigma(\boldsymbol{X}_n(t))\,\mathrm{d}\boldsymbol{W}_t\\ &+\frac{1}{2}\partial^2\Phi_F(\boldsymbol{X}_n(t))\left[\sigma(\boldsymbol{X}_n(t))\sigma(\boldsymbol{X}_n(t))^\top\right]\mathrm{d}t.\end{aligned}$$

Here, for a matrix $\boldsymbol{C}$,

$$\partial^2\Phi_F(\boldsymbol{x})[\boldsymbol{C}]:=\sum_{i,j}C_{ij}\,\partial_{ij}\Phi_F(\boldsymbol{x})$$

denotes the concentration of the second derivative of Φ_F with $\boldsymbol{C}$.

This expansion shows the two effects that remain in the limit. The first order term gives the stochastic motion along Γ, while the second order term produces an additional deterministic drift. Although each stochastic perturbation has zero mean, its second order effect does not vanish because a Brownian increment has size $\sqrt{\mathrm{d}t}$ and hence its square has size $\mathrm{d}t$. Under the convergence assumptions of Katzenberger (1991), if $\boldsymbol{Z}_n \Rightarrow \boldsymbol{W}$, where $\boldsymbol{W}$ is Brownian motion, the limiting process $\boldsymbol{X}$ remains on Γ and satisfies

$$\mathrm{d}\boldsymbol{X}=\partial\Phi_F(\boldsymbol{X})\sigma(\boldsymbol{X})\,\mathrm{d}\boldsymbol{W}_t+\frac{1}{2}\partial^2\Phi_F(\boldsymbol{X})\left[\sigma(\boldsymbol{X})\sigma(\boldsymbol{X})^\top\right]\mathrm{d}t, \qquad \boldsymbol{X}\in\Gamma,\quad t\le\lambda. \tag{5}$$

Here λ is a localization time that keeps the process inside the neighborhood where the assumptions on the flow and the projection Φ_F hold.

C Our Approach: Multiscale Manifold Reduction

In this section, we explain how we derive the slow SDEs in Section 3, focusing on the case $1\le\alpha<2$. Appendix B provides the detailed argument. The picture in Section 3.2 suggests using the relative timescales of $\boldsymbol{\theta}$ and $\boldsymbol{v}$ to analyze their coupled updates. The main difficulty is to justify this analysis over $\mathcal{O}(\eta^{-2})$ iterations, since small errors in individual updates may accumulate. We address this difficulty through *multiscale manifold reduction*, which controls the accumulated effect of parameter deviations from Γ on the preconditioner. The momentum reduction in Appendix E.2 allows us to analyze $(\boldsymbol{\theta},\boldsymbol{v})$ without keeping the momentum variable $\boldsymbol{m}$ as a separate state.

Under the 1-scheme, $1-\beta_2=c_\beta\eta$, so the parameter and second moment updates have step sizes of the same order. We therefore describe their coupled dynamics through the coupled ODE (1) and its projection Π onto $\mathcal{M}=\{(\boldsymbol{\zeta},U(\boldsymbol{\zeta})):\boldsymbol{\zeta}\in\Gamma\}$. Following the framework of Katzenberger (1991), we expand the projected stochastic updates to second order and give the slow SDE (2). Appendix B recalls the framework, and Appendix E.3 gives the proof.

For $1<\alpha<2$, the two variables no longer change at comparable rates. As discussed in Section 3.2, the gradient drift brings $\boldsymbol{\theta}$ back toward Γ on the η^{-1} iteration scale, while $\boldsymbol{v}$ responds to its target on the $\eta^{-\alpha}$ scale. Over $\mathcal{O}(\eta^{-1})$ iterations, $\boldsymbol{v}$ changes by at most $\mathcal{O}(\eta^{\alpha-1})$, which tends to zero. This motivates first holding $\boldsymbol{v}$ fixed and considering the preconditioned flow generated by $-S(\boldsymbol{v})\nabla\mathcal{L}$. However, this argument only explains why $\boldsymbol{v}$ can be treated as fixed over a short interval. We still need to control the errors over $\mathcal{O}(\eta^{-2})$ iterations, over which $\boldsymbol{v}$ could change non-negligibly as $\mathcal{O}(\eta^{\alpha-2})\overset{\eta\to0}{\to}\infty$.

The key step is to control the accumulated error from replacing the local target $\bar{U}(\boldsymbol{\theta}_k)$ by $U(\boldsymbol{\zeta}_k)$, where we recall that $\boldsymbol{\zeta}_k:=\Phi_{S(\boldsymbol{v}_k)}(\boldsymbol{\theta}_k)$ is the projection of $\boldsymbol{\theta}_k$ onto Γ following the preconditioned flow. Each error $\|\bar{U}(\boldsymbol{\theta}_k)-U(\boldsymbol{\zeta}_k)\|$ enters the update of $\boldsymbol{v}$ with a factor $1-\beta_2$, so estimation of a single error does not establish that its cumulative contribution vanishes. Corollary F.4 (Lemma 3.1) shows that the accumulated error vanishes after controlling the linear contribution of $\boldsymbol{\theta}_k-\boldsymbol{\zeta}_k$ in Lemma F.14, which implies that replacing $\bar{U}(\boldsymbol{\theta}_k)$ with $U(\boldsymbol{\zeta}_k)$ produces a vanishing error.

Next, we use this estimate to show that $\boldsymbol{v}_k$ tracks $U(\boldsymbol{\zeta}_k)$. Recall that $\boldsymbol{v}$ is an exponential moving average of the sampled second moments. The preceding estimate allows us to replace their local mean $\bar{U}(\boldsymbol{\theta}_k)$ by $U(\boldsymbol{\zeta}_k)$ in this average. We also show that the averaged noise vanishes and that $U(\boldsymbol{\zeta})$ changes negligibly over the averaging window (Lemma F.19). Together with the initial closeness established in Lemma F.7, these estimates show that $\boldsymbol{v}$ remains close to $U(\boldsymbol{\zeta})$ throughout the subsequent training horizon (Lemma F.20), and replacing $\boldsymbol{v}$ by $U(\boldsymbol{\zeta}_k)$ provides a vanishing error (Appendix F.6). The two steps can be summarized as follows.

$$(\boldsymbol{\theta}_k, \boldsymbol{v}_k) \xrightarrow{\Phi_{S(\boldsymbol{v}_k)}} (\boldsymbol{\zeta}_k, \boldsymbol{v}_k) \xrightarrow{\boldsymbol{v}_k \approx U(\boldsymbol{\zeta}_k)} (\boldsymbol{\zeta}_k, U(\boldsymbol{\zeta}_k)),$$

where the arrows show that we can reduce the dynamics on the left side to the dynamics on the right, which we also discussed in Section 3.2 with Figure 2.

Finally, we derive the limiting equation for the projected position. The estimations that the accumulated error from replacing $\boldsymbol{v}_k$ by $U(\boldsymbol{\zeta}_k)$ vanishes implies that we can replace $\boldsymbol{v}_k$ by $U(\boldsymbol{\zeta}_k)$ throughout our analysis. We can then replace $S(\boldsymbol{v})$ by $\boldsymbol{S}_*(\boldsymbol{\zeta}) = S(U(\boldsymbol{\zeta}))$ in the limiting coefficients. The remaining stochastic gradient noise and the second order terms in the projection expansion give the diffusion and drift of the slow SDE (3), as derived in Appendix F.7.

D Proof of the convergence result

In this section, we state our convergence results Theorem 3.1 formally and give the proof. The following Theorem D.1 is the formal version of the convergence results.

Theorem D.1 (Convergence to a near–manifold neighborhood)**.** *Suppose Assumption 2.1–2.4 hold. Assume that $1 - \beta_2 = c_\beta \eta^\alpha$ for some constant $c_\beta > 0$ and $0.5 \le \alpha \le 2$. Fix $T, M > 0$. There exist $\epsilon_0, C_0, C > 0$ such that, for every $\delta \in (0,1)$, there exists $\eta_0(\delta) > 0$ with the following property for every $0 < \eta < \eta_0(\delta)$. Initialize the AGM i.i.d. uniformly with*

$$\operatorname{dist}(\boldsymbol{\theta}, \Gamma) \le \epsilon_0, \qquad \|\boldsymbol{m}_0\| \le M, \qquad \boldsymbol{v}_0 \in [0, R_1]^D.$$

Let $K_\eta := \lceil C_0 \eta^{-1} \log(1/\eta) \rceil$. With probability at least $1 - \delta$, $\Phi_{S(\boldsymbol{v}_{K_\eta})}(\boldsymbol{\theta}_{K_\eta}) \in \Gamma$ and the following holds:

$$\mathcal{L}(\boldsymbol{\theta}_{K_\eta}) - \mathcal{L}^* = \mathcal{O}\left(\eta \log \frac{2}{\delta}\right),$$

$$\|\boldsymbol{\theta}_{K_\eta} - \Phi_{S(\boldsymbol{v}_{K_\eta})}(\boldsymbol{\theta}_{K_\eta})\|_2 = \mathcal{O}\left(\sqrt{\eta \log \frac{2}{\delta}}\right).$$

We first give some notations we need for the proof.

Definition D.1. *Define $F(\boldsymbol{\theta}) := \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^*, \lambda_\eta := 1 - \beta_2 = c_\beta \eta^\alpha, \mathcal{F}_k := \sigma(\boldsymbol{\theta}_0, \boldsymbol{m}_0, \boldsymbol{v}_0, \xi_0, \ldots, \xi_{k-1})$. For the AGM iterates, define*

$$\boldsymbol{g}_k := \nabla \ell(\boldsymbol{\theta}_k; \xi_k), \qquad \boldsymbol{S}_k := S(\boldsymbol{v}_k), \qquad c_1 := \frac{\beta_1}{1 - \beta_1}, \qquad \boldsymbol{y}_k := \boldsymbol{\theta}_k - \eta c_1 \boldsymbol{S}_k \boldsymbol{m}_k.$$

Write $\Delta \boldsymbol{y}_k := \boldsymbol{y}_{k+1} - \boldsymbol{y}_k$ and define the centered corrected increment by

$$\boldsymbol{e}_k := \eta^{-1}\big(\Delta \boldsymbol{y}_k - \mathbb{E}[\Delta \boldsymbol{y}_k \mid \mathcal{F}_k]\big).$$

Lemma D.1 (Local quadratic growth and PL inequality)**.** *There exist $r, a_1, a_2, a_3 > 0$ such that the tube $\mathcal{U} := \{\boldsymbol{y} : \operatorname{dist}(\boldsymbol{y}, \Gamma) < r\}$ is bounded and, for every $\boldsymbol{y} \in \mathcal{U}$,*

$$a_1 \operatorname{dist}(\boldsymbol{y}, \Gamma)^2 \le F(\boldsymbol{y}) \le a_2 \operatorname{dist}(\boldsymbol{y}, \Gamma)^2, \qquad \|\nabla F(\boldsymbol{y})\|^2 \ge a_3 F(\boldsymbol{y}).$$

Proof. By Assumption 2.3, F vanishes on Γ and each point of Γ is a local minimizer. Thus $\nabla F(\boldsymbol{\zeta}) = \boldsymbol{0}$ and $\nabla^2 F(\boldsymbol{\zeta})$ is positive semidefinite for $\boldsymbol{\zeta} \in \Gamma$. Differentiating $\nabla F = \boldsymbol{0}$ along Γ shows that the tangent space of Γ lies in the Hessian kernel. The rank assumption 2.3 implies that $T_\zeta \Gamma = \ker \nabla^2 F(\zeta)$ since they both have dimension $d - m$. Compactness of Γ and continuity of the Hessian provide uniform positive lower and finite upper bounds on its normal eigenvalues

$\nabla^2 F(\boldsymbol{\zeta}), \boldsymbol{\zeta} \in \Gamma$. Choose a tubular radius on which nearest-point coordinates $\boldsymbol{y} = \boldsymbol{\zeta} + \boldsymbol{n}$, with $\boldsymbol{n} \in N_{\boldsymbol{\zeta}}\Gamma$, are well defined. Taylor expansions give

$$F(\boldsymbol{\zeta} + \boldsymbol{n}) = \tfrac{1}{2}\boldsymbol{n}^\top \nabla^2 F(\boldsymbol{\zeta})\boldsymbol{n} + \mathcal{O}(\|\boldsymbol{n}\|^3), \qquad \nabla F(\boldsymbol{\zeta} + \boldsymbol{n}) = \nabla^2 F(\boldsymbol{\zeta})\boldsymbol{n} + \mathcal{O}(\|\boldsymbol{n}\|^2).$$

After decreasing the radius, these expansions imply $a_1\|\boldsymbol{n}\|^2 \le F(\boldsymbol{y}) \le a_2\|\boldsymbol{n}\|^2$ and $\|\nabla F(\boldsymbol{y})\| \ge c\|\boldsymbol{n}\|$ for fixed positive constants. Since $\|\boldsymbol{n}\| = \operatorname{dist}(\boldsymbol{y}, \Gamma)$, choosing $a_3 = c^2/a_2$ proves the claims. □

Lemma D.2. *There exist $s_-, s_+, B, C, \eta_1 > 0$ such that, for every $\eta < \eta_1$, the following bounds hold almost surely for every $k \ge 0$:*

$$\|\boldsymbol{m}_k\| \le M + R, \qquad \boldsymbol{v}_k \in [0, R_1]^D, \qquad \|\boldsymbol{v}_{k+1} - \boldsymbol{v}_k\| \le C\lambda_\eta.$$

Moreover,

$$s_-\boldsymbol{I} \preceq \boldsymbol{S}_k \preceq s_+\boldsymbol{I}, \qquad \|\boldsymbol{S}_{k+1} - \boldsymbol{S}_k\| \le C\lambda_\eta,$$

and there exists an $\mathcal{F}_k$-measurable vector $\boldsymbol{r}_k$ such that

$$\mathbb{E}[\Delta\boldsymbol{y}_k \mid \mathcal{F}_k] = -\eta\boldsymbol{S}_k\nabla F(\boldsymbol{y}_k) + \eta\boldsymbol{r}_k, \qquad \|\boldsymbol{r}_k\| \le C(\lambda_\eta + \eta).$$

Proof. For sufficiently small η, $0 < \beta_2 < 1$. The momentum and second-moment recursions are convex averages, so induction gives $\|\boldsymbol{m}_k\| \le M + R, \boldsymbol{v}_k \in [0, R_1]^D$ and $\|\boldsymbol{v}_{k+1} - \boldsymbol{v}_k\| \le C\lambda_\eta$. The smooth map S and its derivative are bounded on the compact box $[0, R_1]^D$. Together with $S(\boldsymbol{v}) \succeq \frac{1}{R_0}\boldsymbol{I}$, this gives the stated spectral bounds and the estimate for $\boldsymbol{S}_{k+1} - \boldsymbol{S}_k$.

We now construct $\boldsymbol{r}_k$. Note that

$$\Delta\boldsymbol{y}_k = -\eta(1 + c_1)\boldsymbol{S}_{k+1}\boldsymbol{m}_{k+1} + \eta c_1\boldsymbol{S}_k\boldsymbol{m}_k = -\eta\boldsymbol{S}_k\boldsymbol{g}_k - \frac{\eta}{1 - \beta_1}(\boldsymbol{S}_{k+1} - \boldsymbol{S}_k)\boldsymbol{m}_{k+1}.$$

Therefore $\|\Delta\boldsymbol{y}_k\| \le C\eta$, $\mathbb{E}[\boldsymbol{e}_k \mid \mathcal{F}_k] = \boldsymbol{0}$, and $\|\boldsymbol{e}_k\| \le B$ by the bound of $\boldsymbol{S}$. Since $\boldsymbol{S}_k$ is $\mathcal{F}_k$-measurable and ξ_k is independent of $\mathcal{F}_k$, $\mathbb{E}[\boldsymbol{g}_k \mid \mathcal{F}_k] = \nabla F(\boldsymbol{\theta}_k)$. Define

$$\boldsymbol{r}_k := \boldsymbol{S}_k\big(\nabla F(\boldsymbol{y}_k) - \nabla F(\boldsymbol{\theta}_k)\big) - \frac{1}{1 - \beta_1}\mathbb{E}[(\boldsymbol{S}_{k+1} - \boldsymbol{S}_k)\boldsymbol{m}_{k+1} \mid \mathcal{F}_k].$$

Global Lipschitz continuity of ∇F bounds the first term by $C\eta$; $\|\boldsymbol{S}_{k+1} - \boldsymbol{S}_k\| \le C\lambda_\eta$ bounds the second by $C\lambda_\eta$. This proves both the conditional identity and its uniform error bound. □

Lemma D.3. *Let r, a_1, a_2, a_3 be as in Lemma D.1 and s_-, s_+ as in Lemma D.2. There exist $r_\Phi \in (0, r)$ and $C_\Phi > 0$ such that, for every fixed matrix $s_-\boldsymbol{I} \preceq \boldsymbol{S} \preceq s_+\boldsymbol{I}$ and every $\boldsymbol{x}$ with $\operatorname{dist}(\boldsymbol{x}, \Gamma) < r_\Phi$, the flow $\dot{\boldsymbol{x}}(t) = -\boldsymbol{S}\nabla F(\boldsymbol{x}(t))$, $\boldsymbol{x}(0) = \boldsymbol{x}$, converges to $\Phi_{\boldsymbol{S}}(\boldsymbol{x}) \in \Gamma$ and satisfies*

$$\operatorname{dist}(\boldsymbol{x}, \Gamma) \le \|\boldsymbol{x} - \Phi_{\boldsymbol{S}}(\boldsymbol{x})\| \le C_\Phi\sqrt{F(\boldsymbol{x})}.$$

Proof. We first prove that the flow stays inside $\mathcal{U}$. The vector field $-\boldsymbol{S}\nabla F$ is globally Lipschitz, so the flow has a unique solution for all $t \ge 0$. While it remains in $\mathcal{U}$ and $F(\boldsymbol{x}(t)) > 0$, Lemma D.1 gives

$$-\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{F(\boldsymbol{x}(t))} = \frac{\nabla F(\boldsymbol{x}(t))^\top \boldsymbol{S}\nabla F(\boldsymbol{x}(t))}{2\sqrt{F(\boldsymbol{x}(t))}} \ge \frac{s_-\sqrt{a_3}}{2}\|\nabla F(\boldsymbol{x}(t))\|.$$

Since $\|\dot{\boldsymbol{x}}(t)\| \le s_+\|\nabla F(\boldsymbol{x}(t))\|$, integration bounds the length of the trajectory by $C_\Phi\sqrt{F(\boldsymbol{x})}$, where $C_\Phi := 2s_+/(s_-\sqrt{a_3})$. If the loss reaches zero, the gradient vanishes and the flow remains constant, so the same length bound continues to hold. Choose $r_\Phi > 0$ such that $(1 + C_\Phi\sqrt{a_2})r_\Phi < r$. Before a possible first exit from $\mathcal{U}$, the distance to Γ is at most $\operatorname{dist}(\boldsymbol{x}, \Gamma) + C_\Phi\sqrt{F(\boldsymbol{x})} < (1 + C_\Phi\sqrt{a_2})r_\Phi < r$. Continuity of the flow therefore implies we must have that for every $t > 0$, $\boldsymbol{x}(t) \in \mathcal{U}$.

The PL inequality now gives $F(\boldsymbol{x}(t)) \le e^{-s_-a_3 t}F(\boldsymbol{x})$ for all $t \ge 0$. The finite total length makes the trajectory Cauchy as $t \to \infty$. Therefore the limit of the flow lies strictly inside $\mathcal{U}$ and has zero loss. By Lemma D.1, the limit belongs to Γ and equals the flow projection. The length bound therefore gives the upper inequality. The lower bound simply comes from the definition of dist. □

Lemma D.4. *There exist $b, c, C > 0$ such that, for every $\eta < \eta_1$, $\boldsymbol{y}_k \in \mathcal{U}$ and $\boldsymbol{h}_k := \nabla F(\boldsymbol{y}_k)$,*

$$F(\boldsymbol{y}_{k+1}) \le F(\boldsymbol{y}_k) - b\eta\|\boldsymbol{h}_k\|^2 + C\eta(\lambda_\eta^2 + \eta) + \eta\langle \boldsymbol{h}_k, \boldsymbol{e}_k\rangle. \tag{6}$$

Consequently,

$$\mathbb{E}[F(\boldsymbol{y}_{k+1}) \mid \mathcal{F}_k] \le (1 - c\eta)F(\boldsymbol{y}_k) + C\eta(\lambda_\eta^2 + \eta).$$

Proof. Smoothness and $\|\Delta \boldsymbol{y}_k\| \le C\eta$ in Lemma D.2 imply

$$F(\boldsymbol{y}_{k+1}) \le F(\boldsymbol{y}_k) + \langle \boldsymbol{h}_k, \Delta \boldsymbol{y}_k\rangle + C\eta^2 = F(\boldsymbol{y}_k) - \eta \boldsymbol{h}_k^\top \boldsymbol{S}_k \boldsymbol{h}_k + \eta\langle \boldsymbol{h}_k, \boldsymbol{r}_k + \boldsymbol{e}_k\rangle + C\eta^2.$$

Young's inequality gives $\langle \boldsymbol{h}_k, \boldsymbol{r}_k\rangle \le (s_-/2)\|\boldsymbol{h}_k\|^2 + C(\lambda_\eta^2 + \eta^2)$. Taking $b = s_-/2$ and using $\eta \le 1$ proves (6). Conditional expectation removes the centered term $\eta\langle \boldsymbol{h}_k, \boldsymbol{e}_k\rangle$. Applying Lemma D.1 then proves the corollary inequality with $c = ba_3$. □

Lemma D.5 (Upcrossing against a negative drift)**.** *Let Z_k be a real process with integrable increments adapted to a filtration $\mathcal{G}_k$, and set $\sigma := \inf\{k \ge 0 : Z_k \ge 2a\}$. Suppose $a, d, J > 0$, $Z_0 \le a/2$, $2J\eta < a$, and $|Z_{k+1} - Z_k| \le J\eta$ on $\{k < \sigma\}$. Assume also that*

$$\mathbb{E}[Z_{k+1} - Z_k \mid \mathcal{G}_k] \le -d\eta \quad \text{on } \{k < \sigma,\ a \le Z_k < 2a\}.$$

Then, for every integer $N \ge 0$,

$$\mathbb{P}(\sigma \le N) \le (N+1)\exp\left(-\frac{da}{2J^2\eta}\right).$$

Proof. The case $N = 0$ is immediate. Assume that $N > 0$. Let $r_u = d/(J^2\eta)$. Conditional Hoeffding's lemma gives that for $k < \sigma,\ a \le Z_k < 2a$,

$$\mathbb{E}[e^{r_u(Z_{k+1} - Z_k)} \mid \mathcal{G}_k] \le \exp\big(-r_u d\eta + \tfrac{1}{2}r_u^2 J^2\eta^2\big) \le 1.$$

Fix $s \in \{1, \dots, N\}$ and consider the $\mathcal{G}_s$-measurable entry event $E_s := \{s < \sigma,\ Z_{s-1} < a \le Z_s\}$. On E_s, bounded increments give $a \le Z_s \le a + J\eta < 2a$. Stop $e^{r_u Z_k}$ at $T_s := \inf\{k \ge s : Z_k < a \text{ or } Z_k \ge 2a\} \wedge N$. The process $\mathbf{1}_{E_s} e^{r_u Z_{k \wedge T_s}}$, $k \ge s$, is a nonnegative supermartingale. Optional stopping at the bounded time T_s gives

$$\mathbb{E}[\mathbf{1}_{E_s} e^{r_u Z_{T_s}} \mid \mathcal{G}_s] \le \mathbf{1}_{E_s} e^{r_u Z_s}.$$

On E_s, the conditional probability of an upper exit before or at N is therefore at most

$$e^{r_u(Z_s - 2a)} \le e^{-r_u(a - J\eta)} \le e^{-r_u a/2}.$$

Any path with $\sigma \le N$ must have an entry time s followed by an upper exit before returning below a, since $Z_0 < a$ and one increment cannot cross from below a to $2a$. A union bound over the at most N candidate entry times proves the stated bound. □

Definition D.2 (Working region and exit time)**.** *Fix $0 < r_0 < r_\Phi/2$ and $0 < a < a_1 r_0^2/2$, with constants from Lemmas D.1 and D.3. Define*

$$\mathcal{U}_0 := \{\boldsymbol{y} : \operatorname{dist}(\boldsymbol{y}, \Gamma) < r_0\}, \qquad \mathcal{D} := \{\boldsymbol{y} \in \mathcal{U}_0 : F(\boldsymbol{y}) < 2a\}, \qquad \tau := \inf\{k \ge 0 : \boldsymbol{y}_k \notin \mathcal{D}\}.$$

Lemma D.6. *There exist constants $\epsilon_0, c_*, J, \eta_2 > 0$ such that the following holds for every $\eta \le \eta_2$. If $\operatorname{dist}(\boldsymbol{\theta}_0, \Gamma) \le \epsilon_0$, then $\boldsymbol{y}_0 \in \mathcal{D}$ and $F(\boldsymbol{y}_0) \le a/2$ almost surely. For $k < \tau$, we have $\boldsymbol{y}_{k+1} \in \mathcal{U}_0$, $\operatorname{dist}(\boldsymbol{\theta}_k, \Gamma) < r_\Phi$, and $|F(\boldsymbol{y}_{k+1}) - F(\boldsymbol{y}_k)| \le J\eta$. On $\{\tau < \infty\}$, $F(\boldsymbol{y}_\tau) \ge 2a$. For every integer $N \ge 0$,*

$$\mathbb{P}(\tau \le N) \le (N+1)e^{-c_*/\eta}.$$

Proof. By Lemma D.1, $\overline{\mathcal{D}}$ is compact and satisfies $\operatorname{dist}(\boldsymbol{y}, \Gamma) \le \sqrt{2a/a_1} < r_0$ for $\boldsymbol{y} \in \overline{\mathcal{D}}$. Thus $\overline{\mathcal{D}}$ has positive distance from the complement of $\mathcal{U}_0$. Choose $\epsilon_0 \le \min\{r_0/4, \sqrt{a/(8a_2)}\}$ and then take $\eta_2 \le \eta_1$ small enough that $\|\boldsymbol{\theta}_0 - \boldsymbol{y}_0\| \le \epsilon_0$ by Lemma D.2. The initial distance of $\boldsymbol{y}_0$ is at most $2\epsilon_0$, and hence $F(\boldsymbol{y}_0) \le 4a_2\epsilon_0^2 \le a/2$. The global bound $\|\Delta \boldsymbol{y}_k\| \le C\eta$ ensures that every step from $\mathcal{D}$ remains in $\mathcal{U}_0$ for $\eta \le \eta_2$. The bound $\|\boldsymbol{\theta}_k - \boldsymbol{y}_k\| \le C\eta$ and $r_0 < r_\Phi/2$ likewise imply the claimed bound on $\boldsymbol{\theta}_k$. An exit from $\mathcal{D}$ therefore occurs through $F \ge 2a$, rather than through the boundary of $\mathcal{U}_0$. The population gradient $\nabla F = \nabla \mathcal{L}$ is bounded by R since it is the expectation of the bounded

sample gradient, so $|F(\boldsymbol{y}_{k+1}) - F(\boldsymbol{y}_k)| \le R\|\Delta \boldsymbol{y}_k\| \le J\eta$. On $\{k < \tau,\ a \le F(\boldsymbol{y}_k) < 2a\}$, Lemma D.4 gives

$$\mathbb{E}[F(\boldsymbol{y}_{k+1}) - F(\boldsymbol{y}_k) \mid \mathcal{F}_k] \le -ca\eta + C\eta(\lambda_\eta^2 + \eta) \le -(ca/2)\eta$$

for sufficiently small η, since $\lambda_\eta^2 + \eta \to 0$. Apply Lemma D.5 to $Z_k := F(\boldsymbol{y}_{k\wedge\tau})$, with $d = ca/2$. The first hit of $[2a, \infty)$ by this process is precisely τ. Taking $c_* = da/(2J^2)$ proves the desired bound of $\mathbb{P}(\tau \le N)$ and completes the proof. □

Definition D.3. *For $s > 0$ and the exit time τ in Definition D.2, define*

$$W_k^{(s)} := e^{sF(\boldsymbol{y}_k)/\eta}\mathbf{1}_{\{\tau > k\}}, \qquad A_k^{(s)} := \mathbb{E}[W_k^{(s)}].$$

Lemma D.7 (Exponential-moment concentration)**.** *Following the hypotheses and initialization of Lemma D.6, there exist constants $s, c_e, C_e, \eta_3 > 0$ such that, for every $\eta \le \eta_3$, we have $q := 1 - c_e\eta \in (0,1)$ and that on $\{k < \tau\}$,*

$$\mathbb{E}[e^{sF(\boldsymbol{y}_{k+1})/\eta} \mid \mathcal{F}_k] \le \exp\left(\frac{sqF(\boldsymbol{y}_k)}{\eta} + C_e\eta\right).$$

For the moments in Definition D.3, it follows that

$$A_{k+1}^{(s)} \le e^{C_e\eta}(A_k^{(s)})^q, \qquad A_k^{(s)} \le \exp\left(q^k\frac{sa}{2\eta} + \frac{C_e}{c_e}\right), \qquad k \ge 0. \tag{7}$$

Proof. Take b from Lemma D.4, and take B from Lemma D.2. Recall that $\boldsymbol{h}_k = \nabla F(\boldsymbol{y}_k)$ and the scalar $\langle \boldsymbol{h}_k, \boldsymbol{e}_k\rangle$ has conditional mean zero and lies in $[-B\|\boldsymbol{h}_k\|, B\|\boldsymbol{h}_k\|]$. Conditional Hoeffding's lemma applied to (6) therefore yields that on $\{k < \tau\}$, for $s > 0$,

$$\mathbb{E}[e^{sF(\boldsymbol{y}_{k+1})/\eta} \mid \mathcal{F}_k] \le \exp\left(\frac{sF(\boldsymbol{y}_k)}{\eta} - (sb - \frac{1}{2}s^2B^2)\|\boldsymbol{h}_k\|^2 + Cs(\lambda_\eta^2 + \eta)\right).$$

For $0.5 \le \alpha \le 2$ and $\eta \le 1$, $\lambda_\eta^2 = c_\beta^2\eta^{2\alpha} \le c_\beta^2\eta$. Take $s = b/B^2$, $c_e = ba_3/2$, and $C_e = Cs(1 + c_\beta^2)$. Then Lemma D.1 gives

$$(sb - \tfrac{1}{2}s^2B^2)\|\boldsymbol{h}_k\|^2 = \tfrac{1}{2}sb\|\boldsymbol{h}_k\|^2 \ge sc_eF(\boldsymbol{y}_k),$$

which proves the conditional exponential bound. Restricting to $\{\tau > k+1\}$ will not increase the positive exponential, and $W_k^{(s)} = 0$ on $\{\tau \le k\}$. Consequently, for $0 < \eta < \min\{1, c_e^{-1}\}$ and $q = 1 - c_e\eta$,

$$\mathbb{E}[W_{k+1}^{(s)} \mid \mathcal{F}_k] \le \mathbf{1}_{\{\tau > k\}}\exp\left(sqF(\boldsymbol{y}_k)/\eta + C_e\eta\right) = (W_k^{(s)})^q\exp(C_e\eta).$$

The moments are finite because $F(\boldsymbol{y}_k) < 2a$ on $\{\tau > k\}$. Jensen's inequality for the concave function x^q proves the first inequality in (7). Since $A_0^{(s)} \le e^{sa/(2\eta)}$, iteration gives

$$A_k^{(s)} \le \exp\left(q^k\frac{sa}{2\eta} + C_e\eta\sum_{j=0}^{k-1}q^j\right) \le \exp\left(q^k\frac{sa}{2\eta} + \frac{C_e}{c_e}\right).$$

This proves the second inequality. □

Lemma D.8 (Tail bound at the burn-in iteration)**.** *Following the hypotheses and initialization of Lemma D.6, define $C_0 = c_e^{-1}$ and $A_* = \exp(sa/2 + C_e/c_e)$. Recall that $K_\eta := \lceil C_0\eta^{-1}\log(1/\eta)\rceil$, we have $A_{K_\eta}^{(s)} \le A_*$ for $\eta \le \eta_3$. For every $u > 0$,*

$$\mathbb{P}\big(\tau \le K_\eta \text{ or } F(\boldsymbol{y}_{K_\eta}) > u\big) \le (K_\eta + 1)e^{-c_*/\eta} + A_*e^{-su/\eta}, \tag{8}$$

Proof. For $\eta < 1$, we have $q^{K_\eta} \le e^{-c_e\eta K_\eta} \le \eta$, so Lemma D.7 gives $A_{K_\eta}^{(s)} \le \exp(sa/2 + C_e/c_e) = A_*$. On $\{\tau > K_\eta,\ F(\boldsymbol{y}_{K_\eta}) > u\}$, we have $W_{K_\eta}^{(s)} > e^{su/\eta}$. Markov's inequality bounds this event by $A_*e^{-su/\eta}$. Adding the exit bound from Lemma D.6 proves (8). □

D.1 Proof of the convergence theorem

Now we are ready to prove Theorem D.1.

Proof of Theorem D.1. Choose ϵ_0 from Lemma D.6 and C_0 from Lemma D.8 for the desired constants. Fix $\delta \in (0,1)$ and let $u := (\eta/s)\log(2A_*/\delta)$. Since $K_\eta = \mathcal{O}(\eta^{-1}\log(1/\eta))$, we have

$$\lim_{\eta \downarrow 0}(K_\eta + 1)e^{-c_*/\eta} = 0.$$

Therefore there exists $\eta_4(\delta) > 0$ such that $(K_\eta+1)e^{-c_*/\eta} \le \delta/2$ for every $0 < \eta < \eta_4(\delta)$. Let $\eta_0(\delta) := \min\{1, \eta_2, \eta_3, \eta_4(\delta)\}$ and assume $0 < \eta < \eta_0(\delta)$. By Lemma D.8, the event

$$\mathcal{A}_\eta := \{\tau > K_\eta,\ F(\boldsymbol{y}_{K_\eta}) \le u\}$$

has probability at least $1-\delta$. On $\mathcal{A}_\eta$, Lemma D.6 places $\boldsymbol{\theta}_{K_\eta}$ inside the projection tube of Lemma D.3. Using Lemmas D.1 and D.2, we obtain

$$0 \le F(\boldsymbol{\theta}_{K_\eta}) \le a_2\big(\operatorname{dist}(\boldsymbol{y}_{K_\eta}, \Gamma) + C\eta\big)^2 \le \frac{2a_2}{a_1}F(\boldsymbol{y}_{K_\eta}) + C\eta^2 \le C\eta\log\frac{2}{\delta}.$$

In the last inequality, we used $\log(2A_*/\delta) \le (1 + \log A_*/\log 2)\log(2/\delta)$ and $\eta \le 1$. By Lemma D.3, the flow projection belongs to Γ, and on the same event,

$$\operatorname{dist}(\boldsymbol{\theta}_{K_\eta}, \Gamma) \le \|\boldsymbol{\theta}_{K_\eta} - \Phi_{S(\boldsymbol{v}_{K_\eta})}(\boldsymbol{\theta}_{K_\eta})\| \le C_\Phi\sqrt{F(\boldsymbol{\theta}_{K_\eta})} \le C\sqrt{\eta\log\frac{2}{\delta}},$$

which completes the proof. □

Corollary D.1 (Common localization and normal-energy estimates)**.** *Assume Assumption 2.1–2.4, and fix $M, c_\beta > 0$, $1 \le \alpha < 2$, and $\varrho \in (0, r_\Phi)$. In Definition D.2, take $r_0 := \varrho/2$ and $a := a_1\varrho^2/64$. There exist constants $\epsilon_0, c_*, c_e, C, C_y, \eta_* > 0$ that satisfy the following properties: Suppose the initial state is independent of subsequent samples with*

$$\operatorname{dist}(\boldsymbol{\theta}_0, \Gamma) \le \epsilon_0, \qquad \|\boldsymbol{m}_0\| \le M, \qquad \boldsymbol{v}_0 \in [0, R_1]^D.$$

1. *For every $0 < \eta < \eta_*$, integer $N \ge 1$, and $\delta \in (0,1)$, let $q := 1 - c_e\eta \in (0,1)$. Then*

$$\mathbb{P}\left(\tau > N,\ F(\boldsymbol{\theta}_k) \le C\left(q^k a + \eta\log\frac{C(N+1)}{\delta}\right) \text{ for all } k \le N\right) \ge 1 - \delta - (N+1)e^{-c_*/\eta}.$$

2. *$\|\boldsymbol{\theta}_k - \boldsymbol{y}_k\| \le C_y\eta$ for every k, and on $\{\tau > N\}$, $\max_{k\le N}\operatorname{dist}(\boldsymbol{y}_k, \Gamma) < \varrho/4$, $\max_{k\le N}\operatorname{dist}(\boldsymbol{\theta}_k, \Gamma) < \varrho/2$.*

3. *For every fixed $C_T > 0$, $C_{\mathrm{b}} \ge c_e^{-1}$, and integer sequence $k_{\mathrm{b}} \le N \le C_T\eta^{-2}$, where $k_{\mathrm{b}} := \lceil C_{\mathrm{b}}\eta^{-1}\log(1/\eta)\rceil$, we have*

$$\max_{k_{\mathrm{b}}\le k\le N} F(\boldsymbol{\theta}_k) = \mathcal{O}_{\mathbb{P}}(\eta\log(1/\eta)), \qquad \max_{k_{\mathrm{b}}\le k\le N}\operatorname{dist}(\boldsymbol{\theta}_k, \Gamma) = \mathcal{O}_{\mathbb{P}}(\sqrt{\eta\log(1/\eta)}). \tag{9}$$

Proof. We first prove assertion (2). For every $\boldsymbol{y} \in \mathcal{D}$, the definition of $\mathcal{D}$ and $r_0 < r_\Phi < r$ imply

$$\operatorname{dist}(\boldsymbol{y}, \Gamma) < r_0 < r, \qquad \boldsymbol{y} \in \mathcal{U}, \qquad F(\boldsymbol{y}) < 2a.$$

Thus Lemma D.1 gives

$$a_1\operatorname{dist}(\boldsymbol{y}, \Gamma)^2 \le F(\boldsymbol{y}) < 2a, \qquad \sup_{\boldsymbol{y}\in\mathcal{D}}\operatorname{dist}(\boldsymbol{y}, \Gamma) \le \sqrt{\frac{2a}{a_1}} = \sqrt{\frac{2}{a_1}\frac{a_1\varrho^2}{64}} = \frac{\varrho}{4\sqrt{2}} < \frac{\varrho}{4}.$$

Next, by the definition of $\boldsymbol{y}_k$ and Lemma D.2, for every $k \ge 0$,

$$\|\boldsymbol{\theta}_k - \boldsymbol{y}_k\| = \eta\frac{\beta_1}{1-\beta_1}\|S(\boldsymbol{v}_k)\boldsymbol{m}_k\| \le \eta\frac{\bar\beta_1}{1-\bar\beta_1}s_+(M+R) =: C_y\eta.$$

Taking the infimum over $\boldsymbol{\zeta} \in \Gamma$ in the triangle inequality yields

$$\operatorname{dist}(\boldsymbol{\theta}_k, \Gamma) = \inf_{\boldsymbol{\zeta}\in\Gamma}\|\boldsymbol{\theta}_k - \boldsymbol{\zeta}\| \le \|\boldsymbol{\theta}_k - \boldsymbol{y}_k\| + \inf_{\boldsymbol{\zeta}\in\Gamma}\|\boldsymbol{y}_k - \boldsymbol{\zeta}\| \le C_y\eta + \operatorname{dist}(\boldsymbol{y}_k, \Gamma).$$

Apply Lemmas D.6 and D.7 with the specified r_0, a, and let $\eta_* := \min\{1, \eta_2, \eta_3, \varrho/(4C_y)\}$. By the definition of τ, on $\{\tau > N\}$ we have $\boldsymbol{y}_k \in \mathcal{D}$ for every $k \le N$, and hence

$$\max_{k\le N} \operatorname{dist}(\boldsymbol{y}_k, \Gamma) \le \frac{\varrho}{4\sqrt{2}} < \frac{\varrho}{4}, \qquad \max_{k\le N} \operatorname{dist}(\boldsymbol{\theta}_k, \Gamma) \le \frac{\varrho}{4\sqrt{2}} + C_y\eta < \frac{\varrho}{4\sqrt{2}} + \frac{\varrho}{4} < \frac{\varrho}{2}.$$

This proves assertion (2). For assertion (1), by (7), for $u > 0$,

$$\mathbb{P}\left(\tau > k,\ F(\boldsymbol{y}_k) > q^k a/2 + u\right) \le e^{C_e/c_e - su/\eta}.$$

Let $u = (\eta/s)\log((N+1)e^{C_e/c_e}/\delta)$. Union bound implies that

$$\mathbb{P}\left(\tau \le N \text{ or } \exists k \le N : F(\boldsymbol{y}_k) > q^k a/2 + u\right) \le (N+1)e^{-c_*/\eta} + \delta,$$

$$F(\boldsymbol{\theta}_k) \le a_2\left(\sqrt{F(\boldsymbol{y}_k)/a_1} + C_y\eta\right)^2 \le (2a_2/a_1)F(\boldsymbol{y}_k) + 2a_2C_y^2\eta^2 \quad (k < \tau).$$

After increasing C to absorb the fixed coefficients and the η^2 term, this proves assertion (1). For assertion (3), take $\delta = \eta$ and restrict to $\eta < \min\{\eta_*, e^{-1}\}$. The definitions of k_{b} and N imply

$$q^{k_{\mathrm{b}}} \le e^{-c_e\eta k_{\mathrm{b}}} \le \eta, \qquad \log\frac{C(N+1)}{\eta} \le \log\big(C(C_T+1)\big) + 3\log(1/\eta) \le C_T'\log(1/\eta).$$

Consequently, assertions (1) and (2), together with Lemma D.1, give an event of probability at least $1 - \eta - (C_T\eta^{-2}+1)e^{-c_*/\eta}$ on which

$$0 \le \max_{k_{\mathrm{b}}\le k\le N} F(\boldsymbol{\theta}_k) \le C_T''\eta\log(1/\eta), \qquad \max_{k_{\mathrm{b}}\le k\le N} \operatorname{dist}(\boldsymbol{\theta}_k, \Gamma) \le \sqrt{C_T''\eta\log(1/\eta)/a_1}.$$

Since $\eta + (C_T\eta^{-2}+1)e^{-c_*/\eta} \to 0$, these bounds prove assertion (3). □

E THE 1-SCHEME

In this section, we will study the 1-scheme and give a detailed proof for Theorem 3.2.

E.1 JOINT FLOW AND PROJECTION

We first describe the coupled dynamics of $(\boldsymbol{\theta}, \boldsymbol{v})$, starting from the coupled flow as defined below.

Definition E.1 (coupled flow)**.** *Let*

$$\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v}) := \begin{pmatrix} -S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta}) \\ c_\beta\big(\bar{U}(\boldsymbol{\theta}) - \boldsymbol{v}\big) \end{pmatrix}.$$

Define the flow generated by $\boldsymbol{f}$ by

$$\frac{\mathrm{d}}{\mathrm{d}s}\varphi(\boldsymbol{x}, s) = \boldsymbol{f}(\varphi(\boldsymbol{x}, s)), \qquad \varphi(\boldsymbol{x}, 0) = \boldsymbol{x}.$$

A point $\boldsymbol{x}$ is an equilibrium of this flow if $\varphi(\boldsymbol{x}, s) = \boldsymbol{x}$ for every s. Equivalently, $\boldsymbol{x}$ is an equilibrium if and only if $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$.

Recall that in Definition 3.4 we define that $\mathcal{M} := \{(\boldsymbol{\zeta}, U(\boldsymbol{\zeta})) : \boldsymbol{\zeta} \in \Gamma\}$. Then $\mathcal{M}$ is a set of equilibrium of φ.

Lemma E.1. *In a sufficiently small neighborhood of $\mathcal{M}$, the equilibrium set of the flow φ is exactly $\mathcal{M}$. For every $\boldsymbol{\zeta} \in \Gamma$,*

$$T_{(\boldsymbol{\zeta}, U(\boldsymbol{\zeta}))}\mathcal{M} = \{(\boldsymbol{u}, \partial U(\boldsymbol{\zeta})\boldsymbol{u}) : \boldsymbol{u} \in T_{\boldsymbol{\zeta}}\Gamma\}. \tag{10}$$

Proof. For any equilibrium $(\boldsymbol{\theta}, \boldsymbol{v})$, we have $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v}) = \boldsymbol{0}$. Therefore $\nabla\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{0}$. Let $\mathcal{C} := (\nabla\mathcal{L})^{-1}(\boldsymbol{0})$ and $\boldsymbol{H}_{\boldsymbol{\zeta}} := \nabla^2\mathcal{L}(\boldsymbol{\zeta})$. Then $T_{\boldsymbol{\zeta}}\Gamma \subseteq \ker(\boldsymbol{H}_{\boldsymbol{\zeta}})$. The rank assumption gives $\dim\ker(\boldsymbol{H}_{\boldsymbol{\zeta}}) = d - m = \dim T_{\boldsymbol{\zeta}}\Gamma$ and hence $\ker(\boldsymbol{H}_{\boldsymbol{\zeta}}) = T_{\boldsymbol{\zeta}}\Gamma$.

Choose a matrix $\boldsymbol{R}_{\boldsymbol{\zeta}} \in \mathbb{R}^{m\times d}$ such that $\operatorname{rank}(\boldsymbol{R}_{\boldsymbol{\zeta}}\boldsymbol{H}_{\boldsymbol{\zeta}}) = m$, and define

$$F_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) := \boldsymbol{R}_{\boldsymbol{\zeta}}\nabla\mathcal{L}(\boldsymbol{\theta}), \qquad \mathcal{N}_{\boldsymbol{\zeta}} := \{\boldsymbol{\theta} \in \mathcal{O}_{\boldsymbol{\zeta}}^0 : F_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) = \boldsymbol{0}\},$$

where $\mathcal{O}_{\boldsymbol{\zeta}}^0$ is a sufficiently small open neighborhood of $\boldsymbol{\zeta}$. The implicit function theorem shows that $\mathcal{N}_{\boldsymbol{\zeta}}$ is a manifold of dimension $d-m$, with

$$T_{\boldsymbol{\zeta}}\mathcal{N}_{\boldsymbol{\zeta}} = \ker(\boldsymbol{R}_{\boldsymbol{\zeta}}\boldsymbol{H}_{\boldsymbol{\zeta}}) = \ker(\boldsymbol{H}_{\boldsymbol{\zeta}}) = T_{\boldsymbol{\zeta}}\Gamma.$$

Since $\nabla\mathcal{L}$ vanishes on Γ, we have $\Gamma \cap \mathcal{O}_{\boldsymbol{\zeta}}^0 \subseteq \mathcal{N}_{\boldsymbol{\zeta}}$. Both sets are embedded manifolds of dimension $d-m$, so $\Gamma \cap \mathcal{O}_{\boldsymbol{\zeta}}^0$ is open in $\mathcal{N}_{\boldsymbol{\zeta}}$. After shrinking to an open neighborhood $\mathcal{O}_{\boldsymbol{\zeta}} \subseteq \mathcal{O}_{\boldsymbol{\zeta}}^0$, we obtain

$$\mathcal{C} \cap \mathcal{O}_{\boldsymbol{\zeta}} = \mathcal{N}_{\boldsymbol{\zeta}} \cap \mathcal{O}_{\boldsymbol{\zeta}} = \Gamma \cap \mathcal{O}_{\boldsymbol{\zeta}}.$$

By compactness, choose $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_J \in \Gamma$ such that

$$\Gamma \subseteq \mathcal{O} := \bigcup_{j=1}^{J} \mathcal{O}_{\boldsymbol{\zeta}_j}, \qquad \mathcal{C} \cap \mathcal{O} = \Gamma. \tag{11}$$

Let $\mathcal{V}$ be a sufficiently small open neighborhood of $U(\Gamma)$ and define

$$\mathcal{U} := \mathcal{O} \times \mathcal{V},$$
$$\mathrm{Eq}_{\mathcal{U}}(\varphi) := \{\boldsymbol{x} \in \mathcal{U} : \varphi(\boldsymbol{x}, s) = \boldsymbol{x} \text{ for every } s\} = \{\boldsymbol{x} \in \mathcal{U} : \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}\}.$$

For $(\boldsymbol{\theta}, \boldsymbol{v}) \in \mathcal{U}$, positive definiteness of $S(\boldsymbol{v})$ and (11) give

$$\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v}) = \boldsymbol{0} \Leftrightarrow \boldsymbol{\theta} \in \mathcal{C} \text{ and } \boldsymbol{v} = \bar{U}(\boldsymbol{\theta}) \Leftrightarrow \boldsymbol{\theta} \in \Gamma \text{ and } \boldsymbol{v} = U(\boldsymbol{\theta}) \Leftrightarrow (\boldsymbol{\theta}, \boldsymbol{v}) \in \mathcal{M}.$$

Thus $\mathrm{Eq}_{\mathcal{U}}(\varphi) = \mathcal{M}$.

Finally, define the graph embedding

$$\iota : \Gamma \to \mathbb{R}^{d+D}, \qquad \iota(\boldsymbol{\zeta}) := (\boldsymbol{\zeta}, U(\boldsymbol{\zeta})).$$

Since $\mathcal{M} = \iota(\Gamma)$,

$$T_{\iota(\boldsymbol{\zeta})}\mathcal{M} = \mathrm{Ran}(\partial\iota(\boldsymbol{\zeta})) = \{(\boldsymbol{u}, \partial U(\boldsymbol{\zeta})\boldsymbol{u}) : \boldsymbol{u} \in T_{\boldsymbol{\zeta}}\Gamma\},$$

which is (10). □

Lemma E.2. *For every $\boldsymbol{\zeta} \in \Gamma$,*

$$\boldsymbol{A}_{\boldsymbol{\zeta}} := \partial\boldsymbol{f}(\iota(\boldsymbol{\zeta})) = \begin{pmatrix} -S(U(\boldsymbol{\zeta}))\nabla^2\mathcal{L}(\boldsymbol{\zeta}) & \boldsymbol{0} \\ c_\beta\partial\bar{U}(\boldsymbol{\zeta}) & -c_\beta\boldsymbol{I}_D \end{pmatrix},$$

with kernel $\ker(\boldsymbol{A}_{\boldsymbol{\zeta}}) = T_{\iota(\boldsymbol{\zeta})}\mathcal{M}$. Let $\lambda_1(\boldsymbol{\zeta}), \dots, \lambda_m(\boldsymbol{\zeta}) > 0$ be the positive eigenvalues of $S(U(\boldsymbol{\zeta}))\nabla^2\mathcal{L}(\boldsymbol{\zeta})$, counted with algebraic multiplicity. The eigenvalues of $\boldsymbol{A}_{\boldsymbol{\zeta}}$ are zero with multiplicity $d-m$, $-\lambda_1(\boldsymbol{\zeta}), \dots, -\lambda_m(\boldsymbol{\zeta})$, and $-c_\beta$ with multiplicity D. The zero eigenvalue is semisimple. Moreover, $\lambda_i(\boldsymbol{\zeta}) \geq \gamma$ and $c_\beta \geq \gamma$ for some $\gamma > 0$ independent of $\boldsymbol{\zeta}$ and i.

Proof. Since the matrix $S(U(\boldsymbol{\zeta}))\nabla^2\mathcal{L}(\boldsymbol{\zeta})$ is similar to the symmetric positive semidefinite matrix

$$S(U(\boldsymbol{\zeta}))^{1/2}\nabla^2\mathcal{L}(\boldsymbol{\zeta})S(U(\boldsymbol{\zeta}))^{1/2}.$$

It has m positive eigenvalues and $d-m$ zero eigenvalues. Their positive lower bound is uniform because the Hessian has constant rank and Γ is compact.

Since $\boldsymbol{A}_{\boldsymbol{\zeta}}$ is block lower triangular, for every $r \in \mathbb{C}$,

$$\det(r\boldsymbol{I}_{d+D} - \boldsymbol{A}_{\boldsymbol{\zeta}}) = \det\left(r\boldsymbol{I}_d + S(U(\boldsymbol{\zeta}))\nabla^2\mathcal{L}(\boldsymbol{\zeta})\right)(r + c_\beta)^D.$$

This proves the stated list of eigenvalues and the uniform bound. It remains to check the eigenvalue zero. Solving $\boldsymbol{A}_{\boldsymbol{\zeta}}(\boldsymbol{u}, \boldsymbol{w}) = \boldsymbol{0}$ gives

$$\nabla^2\mathcal{L}(\boldsymbol{\zeta})\boldsymbol{u} = \boldsymbol{0}, \qquad \boldsymbol{w} = \partial\bar{U}(\boldsymbol{\zeta})\boldsymbol{u} = \partial U(\boldsymbol{\zeta})\boldsymbol{u}.$$

The kernel is therefore $T_{\iota(\boldsymbol{\zeta})}\mathcal{M}$ and has dimension $d-m$, equal to the algebraic multiplicity of zero. Thus zero is semisimple. □

Recall that in Definition 3.4 we defined Π to be the coupled limiting point. The following lemma proves the attraction of the coupled flow, which is needed in Katzenberger (1991).

Lemma E.3. *There exists a neighborhood $\mathcal{U}_{\mathcal{M}}$ of $\mathcal{M}$ and constants $C, a > 0$ such that every $\boldsymbol{x} = (\boldsymbol{\theta}, \boldsymbol{v}) \in \mathcal{U}_{\mathcal{M}}$ has a well-defined limiting point $\Pi(\boldsymbol{x}) \in \mathcal{M}$ which satisfies*

$$\|\varphi(\boldsymbol{x}, s) - \Pi(\boldsymbol{x})\| \le Ce^{-as}\|\boldsymbol{x} - \Pi(\boldsymbol{x})\|, \qquad s \ge 0. \tag{12}$$

After decreasing $\mathcal{U}_{\mathcal{M}}$ if necessary, Π is C^3 and is constant along every trajectory of the flow.

Proof. For $\boldsymbol{\zeta} \in \Gamma$, let $E^s_{\boldsymbol{\zeta}}$ be the stable spectral subspace of $\boldsymbol{A}_{\boldsymbol{\zeta}}$. By Lemmas E.1 and E.2,

$$\mathbb{R}^{d+D} = T_{\iota(\boldsymbol{\zeta})}\mathcal{M} \oplus E^s_{\boldsymbol{\zeta}}, \qquad \partial\varphi(\iota(\boldsymbol{\zeta}), s) = e^{s\boldsymbol{A}_{\boldsymbol{\zeta}}}.$$

The zero eigenvalue is semisimple and the remaining spectrum is bounded away from zero uniformly on Γ. Compactness therefore gives constants $C_0, a_0 > 0$ such that

$$e^{s\boldsymbol{A}_{\boldsymbol{\zeta}}}\boldsymbol{u} = \boldsymbol{u} \quad \text{for } \boldsymbol{u} \in T_{\iota(\boldsymbol{\zeta})}\mathcal{M}, \qquad \left\|e^{s\boldsymbol{A}_{\boldsymbol{\zeta}}}|_{E^s_{\boldsymbol{\zeta}}}\right\| \le C_0e^{-a_0 s} \quad \text{for } s \ge 0.$$

Since the vector field $\boldsymbol{f}$ and the graph $\mathcal{M}$ are C^3 and Since the flow on $\mathcal{M}$ is stationary, the inequality gives the normal hyperbolicity estimates through order three. The stable foliation theorem for normally hyperbolic invariant manifolds (Hirsch et al., 1977, pp. 39–53) therefore gives a positively invariant neighborhood $\mathcal{U}_{\mathcal{M}}$ and a C^3 family of stable leaves satisfying

$$\mathcal{U}_{\mathcal{M}} = \bigsqcup_{\boldsymbol{p}\in\mathcal{M}} W^s_{\mathrm{loc}}(\boldsymbol{p}), \qquad \varphi(W^s_{\mathrm{loc}}(\boldsymbol{p}), s) \subseteq W^s_{\mathrm{loc}}(\boldsymbol{p}).$$

Then $\Pi(\boldsymbol{x}) = \boldsymbol{p}$ when $\boldsymbol{x} \in W^s_{\mathrm{loc}}(\boldsymbol{p})$. After decreasing $\mathcal{U}_{\mathcal{M}}$, the stable leaf estimate is uniform and gives

$$\|\varphi(\boldsymbol{x}, s) - \boldsymbol{p}\| \le Ce^{-as}\|\boldsymbol{x} - \boldsymbol{p}\|, \qquad \boldsymbol{x} \in W^s_{\mathrm{loc}}(\boldsymbol{p}), \quad s \ge 0.$$

Consequently, $\varphi(\boldsymbol{x}, s) \to \Pi(\boldsymbol{x})$ and $\Pi(\varphi(\boldsymbol{x}, s)) = \Pi(\boldsymbol{x})$. The C^3 dependence of the stable leaves on their base points gives $\Pi \in C^3(\mathcal{U}_{\mathcal{M}}, \mathcal{M})$. □

Corollary E.1. *For every $\boldsymbol{x} \in \mathcal{U}_{\mathcal{M}}$ and $s \ge 0$,*

$$\Pi(\varphi(\boldsymbol{x}, s)) = \Pi(\boldsymbol{x}), \qquad \partial\pi(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}.$$

Corollary E.2. *For $\boldsymbol{\zeta} \in \Gamma$, recall that*

$$\boldsymbol{J}_{\boldsymbol{\zeta}} := \lim_{s\to\infty} \exp\bigl(-sS(U(\boldsymbol{\zeta}))\nabla^2\mathcal{L}(\boldsymbol{\zeta})\bigr),$$

which is the spectral projection of $S(U(\boldsymbol{\zeta}))\nabla^2\mathcal{L}(\boldsymbol{\zeta})$ onto its kernel. At $\iota(\boldsymbol{\zeta})$,

$$\partial\pi(\iota(\boldsymbol{\zeta}))[\boldsymbol{u}, \boldsymbol{w}] = \boldsymbol{J}_{\boldsymbol{\zeta}}\boldsymbol{u}, \qquad \partial\Pi(\iota(\boldsymbol{\zeta}))[\boldsymbol{u}, \boldsymbol{w}] = \bigl(\boldsymbol{J}_{\boldsymbol{\zeta}}\boldsymbol{u}, \partial U(\boldsymbol{\zeta})\boldsymbol{J}_{\boldsymbol{\zeta}}\boldsymbol{u}\bigr).$$

In particular,

$$\partial_{\boldsymbol{v}}\pi(\iota(\boldsymbol{\zeta})) = \boldsymbol{0}, \qquad \boldsymbol{J}^2_{\boldsymbol{\zeta}} = \boldsymbol{J}_{\boldsymbol{\zeta}}, \qquad \mathrm{Ran}(\boldsymbol{J}_{\boldsymbol{\zeta}}) = T_{\boldsymbol{\zeta}}\Gamma.$$

E.2 Momentum reduction

Let

$$\mathcal{F}_k := \sigma(\boldsymbol{\theta}_0, \boldsymbol{v}_0, \boldsymbol{m}_0, \xi_0, \ldots, \xi_{k-1}),$$

where $\xi_i \sim \mathcal{S}$. Let $\lambda_\eta := 1 - \beta_2$ and recall that $0 \le \beta_1 \le \bar{\beta}_1 < 1$. The following correction of the first moment variable is also used in schemes $(1, 2)$.

Lemma E.4 (First moment momentum corrector)**.** *Let*

$$S_k := S(\boldsymbol{v}_k), \qquad \boldsymbol{g}_k := \nabla\ell(\boldsymbol{\theta}_k; \xi_k), \qquad c_1 := \frac{\beta_1}{1 - \beta_1},$$

and define

$$\boldsymbol{Q}_{\eta,k} := \eta c_1 S_k \boldsymbol{m}_k, \qquad e_{\eta,k+1} := -\frac{\eta}{1 - \beta_1}(S_{k+1} - S_k)\boldsymbol{m}_{k+1}.$$

Then, for every $k \ge 0$,

$$\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -\eta S_k \boldsymbol{g}_k + (\boldsymbol{Q}_{\eta,k+1} - \boldsymbol{Q}_{\eta,k}) + e_{\eta,k+1}. \tag{13}$$

Equivalently, let $\widehat{\boldsymbol{\theta}}_k := \boldsymbol{\theta}_k - \boldsymbol{Q}_{\eta,k}$*,*

$$\widehat{\boldsymbol{\theta}}_{k+1} - \widehat{\boldsymbol{\theta}}_k = -\eta S_{k+1}\boldsymbol{g}_k - \eta c_1(S_{k+1} - S_k)\boldsymbol{m}_k. \tag{14}$$

In a sufficiently small neighborhood of $\mathcal{M}$*,*

$$\|\boldsymbol{m}_k\| \le \|\boldsymbol{m}_0\| + R, \qquad \sup_{k\le N}\|\boldsymbol{Q}_{\eta,k}\| \le C\eta(1 + \|\boldsymbol{m}_0\|),$$

where R *is the upper bound of* $\|\nabla\ell(\boldsymbol{\theta};\xi)\|_2$ *defined in Assumption 2.2. Moreover, assume that* $\sup_{0\le k\le N}\|\boldsymbol{m}_k\| = \mathcal{O}_{\mathbb{P}}(1)$*, then*

$$\sup_{k\le N}\|\boldsymbol{Q}_{\eta,k}\| = \mathcal{O}_{\mathbb{P}}(\eta), \qquad \sup_{k<N}\|e_{\eta,k+1}\| = \mathcal{O}_{\mathbb{P}}(\eta\lambda_\eta), \qquad \sum_{k=0}^{N-1}\|e_{\eta,k+1}\| = \mathcal{O}_{\mathbb{P}}(N\eta\lambda_\eta). \tag{15}$$

Proof. The momentum recursion gives

$$\boldsymbol{g}_k = \boldsymbol{m}_{k+1} + c_1(\boldsymbol{m}_{k+1} - \boldsymbol{m}_k).$$

Therefore

$$\begin{aligned} -\eta S_{k+1}\boldsymbol{m}_{k+1} &= -\eta S_k\boldsymbol{g}_k + \eta c_1 S_k(\boldsymbol{m}_{k+1} - \boldsymbol{m}_k) - \eta(S_{k+1} - S_k)\boldsymbol{m}_{k+1} \\ &= -\eta S_k\boldsymbol{g}_k + (\boldsymbol{Q}_{\eta,k+1} - \boldsymbol{Q}_{\eta,k}) + e_{\eta,k+1}, \end{aligned}$$

which proves (13), and subtracting $\boldsymbol{Q}_{\eta,k+1} - \boldsymbol{Q}_{\eta,k}$ proves (14).

Iterating the recursion for $\boldsymbol{m}_k$ gives

$$\|\boldsymbol{m}_k\| \le \beta_1^k\|\boldsymbol{m}_0\| + (1-\beta_1)\sum_{j=0}^{k-1}\beta_1^{k-1-j}R \le \|\boldsymbol{m}_0\| + R.$$

The second moment recursion gives $\|\boldsymbol{v}_{k+1} - \boldsymbol{v}_k\| \le C\lambda_\eta$. The Lipschitz continuity of S therefore yields $\|S_{k+1} - S_k\| \le C\lambda_\eta$. Together with $c_1 \le \bar{\beta}_1/(1-\bar{\beta}_1)$, these estimates prove the rest of the lemma. □

Now let $\widehat{\boldsymbol{X}}_k := (\widehat{\boldsymbol{\theta}}_k, \boldsymbol{v}_k)$. For $\boldsymbol{x} = (\boldsymbol{\theta}, \boldsymbol{v})$. Recall that in Definition 3.6 we defined $\boldsymbol{q}(\boldsymbol{\zeta};\xi)$ and $\boldsymbol{b}_{\boldsymbol{\zeta}}$. For $\boldsymbol{x} = (\boldsymbol{\theta}, \boldsymbol{v})$, we further define a more general form:

$$\boldsymbol{q}(\boldsymbol{x},\xi) := \begin{pmatrix} -S(\boldsymbol{v})\left(\nabla\ell(\boldsymbol{\theta},\boldsymbol{v}) - \nabla\mathcal{L}(\boldsymbol{\theta})\right) \\ c_\beta\left(\boldsymbol{h}(\boldsymbol{\theta};\xi) - \bar{U}(\boldsymbol{\theta})\right) \end{pmatrix}$$

and

$$\boldsymbol{b}^0(\boldsymbol{\theta},\boldsymbol{v}) := \begin{pmatrix} -c_\beta\mathbb{E}_{\xi\sim\mathcal{S}}\left[\partial S(\boldsymbol{v})[\boldsymbol{h}(\boldsymbol{\theta},\xi) - \boldsymbol{v}]\nabla\ell(\boldsymbol{\theta};\xi)\right] \\ \boldsymbol{0} \end{pmatrix}.$$

We can now give the recursion of $\widehat{\boldsymbol{X}}_k$.

Lemma E.5. *In a sufficiently small neighborhood of* $\mathcal{M}$*,*

$$\widehat{\boldsymbol{X}}_{k+1} - \widehat{\boldsymbol{X}}_k = \eta\boldsymbol{f}(\widehat{\boldsymbol{X}}_k) + \eta\boldsymbol{q}(\widehat{\boldsymbol{X}}_k,\xi_k) + \eta^2\boldsymbol{b}^0(\widehat{\boldsymbol{X}}_k) + \eta^2\boldsymbol{r}_{\eta,k}^{\mathrm{cen}} + \eta^2\boldsymbol{r}_{\eta,k}^{\mathrm{mom}} + \boldsymbol{R}_{\eta,k}, \tag{16}$$

where $\mathbb{E}[\boldsymbol{q}(\widehat{\boldsymbol{X}}_k,\xi_k) \mid \mathcal{F}_k] = \boldsymbol{0}$*,*

$$\boldsymbol{r}_{\eta,k}^{\mathrm{cen}} := \begin{pmatrix} -c_\beta\big(\partial S(\boldsymbol{v}_k)[\boldsymbol{h}(\widehat{\boldsymbol{\theta}}_k;\xi_k) - \boldsymbol{v}_k]\nabla\ell(\widehat{\boldsymbol{\theta}}_k;\xi_k) - \mathbb{E}[\partial S(\boldsymbol{v}_k)[\boldsymbol{h}(\widehat{\boldsymbol{\theta}}_k;\xi) - \boldsymbol{v}_k]\nabla\ell(\widehat{\boldsymbol{\theta}}_k;\xi)]\big) \\ \boldsymbol{0} \end{pmatrix},$$

and

$$\boldsymbol{r}_{\eta,k}^{\mathrm{mom}} := \begin{pmatrix} -c_1 S(\boldsymbol{v}_k)\partial_{\boldsymbol{\theta}}\nabla\ell(\widehat{\boldsymbol{\theta}}_k;\xi_k)[S(\boldsymbol{v}_k)\boldsymbol{m}_k] - c_\beta c_1\partial S(\boldsymbol{v}_k)[\boldsymbol{h}(\widehat{\boldsymbol{\theta}}_k;\xi_k) - \boldsymbol{v}_k]\boldsymbol{m}_k \\ c_\beta c_1\partial_{\boldsymbol{\theta}}\boldsymbol{h}(\widehat{\boldsymbol{\theta}}_k;\xi_k)[S(\boldsymbol{v}_k)\boldsymbol{m}_k] \end{pmatrix},$$

where

$$\mathbb{E}[\boldsymbol{r}_{\eta,k}^{\mathrm{cen}} \mid \mathcal{F}_k] = \boldsymbol{0}, \qquad \|\boldsymbol{r}_{\eta,k}^{\mathrm{cen}}\| \le C, \qquad \|\boldsymbol{R}_{\eta,k}\| \le C\eta^3(1 + \|\boldsymbol{m}_k\|^2).$$

Proof. Taylor expansion gives

$$\nabla \ell(\boldsymbol{\theta}_k;\xi_k) = \nabla \ell(\widehat{\boldsymbol{\theta}}_k;\xi_k) + \eta \partial_{\boldsymbol{\theta}} \nabla \ell(\widehat{\boldsymbol{\theta}}_k;\xi_k)[c_1 S(\boldsymbol{v}_k)\boldsymbol{m}_k] + \mathcal{O}(\eta^2\|\boldsymbol{m}_k\|^2),$$

$$\boldsymbol{h}(\boldsymbol{\theta}_k;\xi_k) = \boldsymbol{h}(\widehat{\boldsymbol{\theta}}_k;\xi_k) + \eta \partial_{\boldsymbol{\theta}} \boldsymbol{h}(\widehat{\boldsymbol{\theta}}_k;\xi_k)[c_1 S(\boldsymbol{v}_k)\boldsymbol{m}_k] + \mathcal{O}(\eta^2\|\boldsymbol{m}_k\|^2).$$

Therefore

$$\begin{aligned}\boldsymbol{v}_{k+1} - \boldsymbol{v}_k &= c_\beta \eta \left(\boldsymbol{h}(\boldsymbol{\theta}_k;\xi_k) - \boldsymbol{v}_k\right) \\ &= c_\beta \eta \big(\boldsymbol{h}(\widehat{\boldsymbol{\theta}}_k;\xi_k) - \boldsymbol{v}_k\big) + c_\beta c_1 \eta^2 \partial_{\boldsymbol{\theta}} \boldsymbol{h}(\widehat{\boldsymbol{\theta}}_k;\xi_k)[S(\boldsymbol{v}_k)\boldsymbol{m}_k] + \mathcal{O}(\eta^3\|\boldsymbol{m}_k\|^2).\end{aligned}$$

Expanding $S(\boldsymbol{v}_{k+1})$ in (14) gives

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{k+1} - \widehat{\boldsymbol{\theta}}_k = &-\eta S(\boldsymbol{v}_k)\nabla \ell(\widehat{\boldsymbol{\theta}}_k;\xi_k) - \eta^2 c_1 S(\boldsymbol{v}_k)\partial_{\boldsymbol{\theta}} \nabla \ell(\widehat{\boldsymbol{\theta}}_k;\xi_k)[S(\boldsymbol{v}_k)\boldsymbol{m}_k] \\ &- c_\beta \eta^2 \partial S(\boldsymbol{v}_k)[\boldsymbol{h}(\widehat{\boldsymbol{\theta}}_k;\xi_k) - \boldsymbol{v}_k]\nabla \ell(\widehat{\boldsymbol{\theta}}_k;\xi_k) \\ &- c_\beta c_1 \eta^2 \partial S(\boldsymbol{v}_k)[\boldsymbol{h}(\widehat{\boldsymbol{\theta}}_k;\xi_k) - \boldsymbol{v}_k]\boldsymbol{m}_k + \mathcal{O}(\eta^3(1+\|\boldsymbol{m}_k\|^2)).\end{aligned}$$

Substitute all terms and simplify immediately gives the desired recursion. Moreover, the bounded third derivatives make every remainder uniform. □

On $\mathcal{M}$, $\boldsymbol{q}(\boldsymbol{x},\xi) = \boldsymbol{q}(\boldsymbol{\zeta},\xi)$ and $\boldsymbol{b}^0(\boldsymbol{\theta},\boldsymbol{v})$. In the block form, the coupled covariance is

$$\boldsymbol{C}_{\boldsymbol{\zeta}} = \begin{pmatrix} \boldsymbol{S}_*(\boldsymbol{\zeta})\boldsymbol{\Sigma}(\boldsymbol{\zeta})\boldsymbol{S}_*(\boldsymbol{\zeta}) & -c_\beta \boldsymbol{S}_*(\boldsymbol{\zeta})\mathbb{E}[\nabla \ell(\boldsymbol{\zeta};\xi)(\boldsymbol{h}(\boldsymbol{\zeta};\xi) - U(\boldsymbol{\zeta}))^\top] \\ -c_\beta \mathbb{E}[(\boldsymbol{h}(\boldsymbol{\zeta};\xi) - U(\boldsymbol{\zeta}))\nabla \ell(\boldsymbol{\zeta};\xi)^\top]\boldsymbol{S}_*(\boldsymbol{\zeta}) & c_\beta^2 \mathbb{E}[(\boldsymbol{h}(\boldsymbol{\zeta};\xi) - U(\boldsymbol{\zeta}))(\boldsymbol{h}(\boldsymbol{\zeta};\xi) - U(\boldsymbol{\zeta}))^\top] \end{pmatrix},$$

We first record the elementary smoothing estimate used below.

Lemma E.6 (Exponential smoothing)**.** *Let $0 \le \rho < 1$, and let $\{\boldsymbol{a}_k\}_{k\ge 0}$ be a sequence in a finite-dimensional normed space. Define $\mathcal{S}_n := \sum_{k=0}^{n-1} \boldsymbol{a}_k$ and $\mathcal{S}_0 := \boldsymbol{0}$. Then, for every $N \ge 1$,*

$$\sup_{n\le N} \left\| \sum_{k=0}^{n-1} \rho^{n-1-k} \boldsymbol{a}_k \right\| \le 2 \sup_{n\le N} \|\mathcal{S}_n\|.$$

Proof. By Abel summation,

$$\sum_{k=0}^{n-1} \rho^{n-1-k} \boldsymbol{a}_k = \mathcal{S}_n - (1-\rho) \sum_{j=1}^{n-1} \rho^{n-1-j} \mathcal{S}_j.$$

Therefore

$$\left\| \sum_{k=0}^{n-1} \rho^{n-1-k} \boldsymbol{a}_k \right\| \le \sup_{j\le n} \|\mathcal{S}_j\| \left(1 + (1-\rho) \sum_{j=1}^{n-1} \rho^{n-1-j} \right) \le 2 \sup_{j\le n} \|\mathcal{S}_j\|.$$

Taking the supremum over $n \le N$ gives the claim. □

Lemma E.7. *Fix $T < \infty$ and assume that the process stays in a sufficiently small neighborhood of $\mathcal{M}$. If $\sup_\eta \mathbb{E}\|\boldsymbol{m}_0\|^4 < \infty$ and*

$$\sup_{0\le t\le T} \operatorname{dist}(\widehat{\boldsymbol{X}}_{\lfloor t/\eta^2 \rfloor}, \mathcal{M}) = o_{\mathbb{P}}(1),$$

then

$$\sup_{0\le t\le T} \left\| \eta^2 \sum_{k<t/\eta^2} \boldsymbol{r}_{\eta,k}^{\mathrm{mom}} \right\| = o_{\mathbb{P}}(1).$$

The conclusion is uniform over $0 \le \beta_1 \le \bar{\beta}_1 < 1$.

Proof. First localize to $\|\boldsymbol{m}_0\| \le M$. The bounded gradient assumption gives $\sup_{k\le N_\eta}\|\boldsymbol{m}_k\| \le M+R$. All estimates below are therefore uniform up to a constant depending on M. For $\boldsymbol{x}=(\boldsymbol{\theta},\boldsymbol{v})$, define the linear map

$$\boldsymbol{A}(\boldsymbol{x})\boldsymbol{m} := \begin{pmatrix} -c_1 S(\boldsymbol{v})\nabla^2\mathcal{L}(\boldsymbol{\theta})[S(\boldsymbol{v})\boldsymbol{m}] - c_\beta c_1 \partial S(\boldsymbol{v})[\bar{U}(\boldsymbol{\theta})-\boldsymbol{v}]\boldsymbol{m} \\ c_\beta c_1 \partial_{\boldsymbol{\theta}}\bar{U}(\boldsymbol{\theta})[S(\boldsymbol{v})\boldsymbol{m}] \end{pmatrix}.$$

Since $\boldsymbol{m}_k$ and $\widehat{\boldsymbol{X}}_k$ are $\mathcal{F}_k$ measurable, independence of ξ_k gives

$$\mathbb{E}[\boldsymbol{r}^{\mathrm{mom}}_{\eta,k} \mid \mathcal{F}_k] = \boldsymbol{A}(\widehat{\boldsymbol{X}}_k)\boldsymbol{m}_k.$$

Thus

$$\boldsymbol{r}^{\mathrm{sam}}_{\eta,k} := \boldsymbol{r}^{\mathrm{mom}}_{\eta,k} - \boldsymbol{A}(\widehat{\boldsymbol{X}}_k)\boldsymbol{m}_k$$

is a bounded martingale difference after localization. Doob's inequality gives

$$\mathbb{E}\left[\eta^4 \sup_{n\le T\eta^{-2}}\left\|\sum_{k=0}^{n-1}\boldsymbol{r}^{\mathrm{sam}}_{\eta,k}\right\|^2\right] \le C_M T\eta^2.$$

Therefore

$$\eta^2 \sup_{n\le T\eta^{-2}}\left\|\sum_{k=0}^{n-1}\boldsymbol{r}^{\mathrm{sam}}_{\eta,k}\right\| = \mathcal{O}_{\mathbb{P}}(\eta).$$

It remains to control $\sum_k \boldsymbol{A}(\widehat{\boldsymbol{X}}_k)\boldsymbol{m}_k$. Write

$$\bar{\boldsymbol{g}}_k := \nabla\mathcal{L}(\boldsymbol{\theta}_k), \qquad \boldsymbol{z}_k := \nabla\ell(\boldsymbol{\theta}_k;\xi_k) - \bar{\boldsymbol{g}}_k.$$

Iterating the momentum recursion gives

$$\boldsymbol{m}_k = \beta_1^k\boldsymbol{m}_0 + (1-\beta_1)\sum_{j=0}^{k-1}\beta_1^{k-1-j}\bar{\boldsymbol{g}}_j + (1-\beta_1)\sum_{j=0}^{k-1}\beta_1^{k-1-j}\boldsymbol{z}_j.$$

We estimate these three terms separately. The contribution of the initial momentum $\boldsymbol{m}_0$ is bounded by

$$\eta^2 \sup_{n\le T\eta^{-2}}\left\|\sum_{k=0}^{n-1}\beta_1^k\boldsymbol{A}(\widehat{\boldsymbol{X}}_k)\boldsymbol{m}_0\right\| \le \frac{C_M\eta^2}{1-\bar{\beta}_1}.$$

For the second term, $\sup_{0\le t\le T}\operatorname{dist}(\widehat{\boldsymbol{X}}_{\lfloor t/\eta^2\rfloor},\mathcal{M}) = o_{\mathbb{P}}(1)$ imply

$$\sup_{k\le N_\eta}\|\bar{\boldsymbol{g}}_k\| \le C\sup_{k\le N_\eta}\operatorname{dist}(\widehat{\boldsymbol{X}}_k,\mathcal{M}) + \mathcal{O}_{\mathbb{P}}(\eta) = o_{\mathbb{P}}(1).$$

Hence the accumulated mean gradient contribution is bounded by

$$(1-\beta_1)\sum_{j=0}^{k-1}\beta_1^{k-1-j}\cdot\eta^2 N_\eta C_M \sup_{k\le N_\eta}\|\bar{\boldsymbol{g}}_k\| = o_{\mathbb{P}}(1).$$

For the last term, write $\boldsymbol{A}_k := \boldsymbol{A}(\widehat{\boldsymbol{X}}_k)$. We have

$$\|\widehat{\boldsymbol{X}}_{k+1} - \widehat{\boldsymbol{X}}_k\| \le C_M\eta, \qquad \|\boldsymbol{A}_k - \boldsymbol{A}_j\| \le C_M\eta|k-j|.$$

Therefore

$$(1-\beta_1)\sum_{j=0}^{k-1}\beta_1^{k-1-j}\boldsymbol{A}_k\boldsymbol{z}_j = (1-\beta_1)\sum_{j=0}^{k-1}\beta_1^{k-1-j}\boldsymbol{A}_j\boldsymbol{z}_j + \boldsymbol{r}^{\mathrm{chg}}_{\eta,k},$$

where $\|\boldsymbol{r}^{\mathrm{chg}}_{\eta,k}\| \le C_M\eta$ and $\boldsymbol{A}_j\boldsymbol{z}_j$ is a martingale difference. Changing the order of summation gives

$$\sum_{k=0}^{n-1}(1-\beta_1)\sum_{j=0}^{k-1}\beta_1^{k-1-j}\boldsymbol{A}_j\boldsymbol{z}_j = \sum_{j=0}^{n-2}(1-\beta_1^{n-1-j})\boldsymbol{A}_j\boldsymbol{z}_j.$$

The process $\boldsymbol{M}_n := \sum_{j=0}^{n-1} \boldsymbol{A}_j \boldsymbol{z}_j$ is a martingale. Lemma E.6 gives

$$\sup_{n \le T\eta^{-2}} \left\| \sum_{j=0}^{n-2} \beta_1^{n-1-j} \boldsymbol{A}_j \boldsymbol{z}_j \right\| \le 2 \sup_{n \le N_\eta} \|\boldsymbol{M}_n\|.$$

Therefore applying Doob's inequality to $\boldsymbol{M}_n$ gives

$$\mathbb{E}\left[\eta^4 \sup_{n \le T\eta^{-2}} \|\boldsymbol{M}_n\|^2 \right] \le C_M T \eta^2.$$

Therefore

$$\eta^2 \sup_{n \le N_\eta} \left\| \sum_{k=0}^{n-1} (1-\beta_1) \sum_{j=0}^{k-1} \beta_1^{k-1-j} \boldsymbol{A}_j \boldsymbol{z}_j \right\| = \mathcal{O}_{\mathbb{P}}(\eta).$$

Combining all the estimates above proves the desired result. Moreover, $\sup_\eta \mathbb{E}\|\boldsymbol{m}_0\|^4 < \infty$ gives

$$\sup_\eta \mathbb{P}(\|\boldsymbol{m}_0\| > M) \to 0$$

as $M \to \infty$, which removes the localization. □

E.3 Weak convergence

We first specify the neighborhood on which the local estimates and the coupled projection are used, inspired by the working zone of Li et al. (2025).

Lemma E.8 (coupled working zone)**.** *Under the assumptions of Theorem 3.2, there exist bounded open sets, independent of η, such that*

$$\mathcal{M} \subset \mathcal{W}_0 \Subset \mathcal{W}_1 \Subset \mathcal{W}_2 \Subset \mathcal{U}_{\mathcal{M}}.$$

On $\overline{\mathcal{W}}_2$, the attraction estimate (12) *holds uniformly, the derivatives of* Π *through order three are bounded. The local coefficients and their derivatives required in* (16) *are uniformly bounded. There exists a smooth cutoff supported in $\mathcal{U}_{\mathcal{M}}$ and equal to one on a neighborhood of $\overline{\mathcal{W}}_2$. Moreover, the separation distances*

$$d_0 := \operatorname{dist}(\overline{\mathcal{W}}_0, \mathcal{W}_1^c), \qquad d_1 := \operatorname{dist}(\overline{\mathcal{W}}_1, \mathcal{W}_2^c), \qquad d_{\mathcal{M}} := \operatorname{dist}(\mathcal{M}, \mathcal{W}_1^c)$$

are strictly positive.

Proof. Compactness of $\mathcal{M}$ and openness of $\mathcal{U}_{\mathcal{M}}$ give the nested neighborhoods, positive separation distances, and smooth cutoff. The uniform attraction estimate and C^3 regularity of Π follow from Lemma E.3. The derivative bounds follow from compactness of $\overline{\mathcal{W}}_2$ and the coefficient regularity assumptions. □

The parameter localization is supplied by Corollary D.1. The additional estimate below controls the adaptive coordinate in the coupled working zone.

Lemma E.9 (coupled burn-in and localization)**.** *Assume Assumption 2.1–2.4. Fix $T, M > 0$, assume $\|\boldsymbol{m}_0\| \le M$ and $\boldsymbol{v}_0 \in [0, R_1]^D$, and let $1 - \beta_2 = c_\beta \eta$. There exist $\epsilon_0, C > 0$ such that, if $\operatorname{dist}((\boldsymbol{\theta}_0, \boldsymbol{v}_0), \mathcal{M}) \le \epsilon_0$, then with*

$$K_0 := \lfloor C\eta^{-1} \log(1/\eta) \rfloor, \qquad N := K_0 + \lfloor T\eta^{-2} \rfloor, \qquad \widehat{\boldsymbol{X}}_k := (\boldsymbol{y}_k, \boldsymbol{v}_k),$$

we have

$$\mathbb{P}(\widehat{\boldsymbol{X}}_k \in \mathcal{W}_1 \text{ for all } k \le N) \to 1, \qquad \sup_{K_0 \le k \le N} \operatorname{dist}(\widehat{\boldsymbol{X}}_k, \mathcal{M}) = \mathcal{O}_{\mathbb{P}}(\sqrt{\eta \log(1/\eta)}). \tag{17}$$

The same distance bound holds for the original state $(\boldsymbol{\theta}_k, \boldsymbol{v}_k)$, and $K_0 \eta^2 \to 0$.

Proof. Fix a closed tube $\{\boldsymbol{y} : \operatorname{dist}(\boldsymbol{y}, \Gamma) \le r_\Phi\}$ and constants $L_U, G < \infty$ such that, on this tube,

$$\|\partial \bar{U}(\boldsymbol{y})\| \le L_U, \qquad \|\nabla F(\boldsymbol{y})\| \le G \operatorname{dist}(\boldsymbol{y}, \Gamma).$$

Let s_+ be the preconditioner bound and set $A := 1 + L_U + L_U s_+ G/c_\beta$. Choose

$$0 < \varrho < \min\{r_\Phi/2, d_{\mathcal{M}}/(8A)\}, \qquad 0 < \epsilon_0 \le \min\{\epsilon_{\mathrm{loc}}(\varrho, M), r_\Phi/8, d_{\mathcal{M}}/(8(1+L_U))\},$$

where $\epsilon_{\mathrm{loc}}(\varrho, M)$ is the initial radius in Corollary D.1 for the indicated ϱ and M. Use its explicit $r_0, a, \mathcal{D}, \tau$ throughout this proof. The corrected parameter satisfies $\boldsymbol{y}_k = \widehat{\boldsymbol{\theta}}_k$; define $\boldsymbol{u}_k := \boldsymbol{v}_k - \bar{U}(\boldsymbol{y}_k)$ and $\boldsymbol{g}_k := \nabla \ell(\boldsymbol{\theta}_k; \xi_k)$. By Lemma D.2, $\|\Delta \boldsymbol{y}_k\| \le C_\Delta \eta$. For $k < \tau$ and $t \in [0,1]$, $\operatorname{dist}(\boldsymbol{y}_k + t\Delta \boldsymbol{y}_k, \Gamma) \le \varrho/4 + C_\Delta \eta < r_\Phi/2$. For $k < \tau$, let

$$\boldsymbol{r}_{k+1}^U := \int_0^1 (1-t)\partial^2 \bar{U}(\boldsymbol{y}_k + t\Delta \boldsymbol{y}_k)[\Delta \boldsymbol{y}_k, \Delta \boldsymbol{y}_k] \, \mathrm{d}t,$$

$$\boldsymbol{r}_{k+1}^v := c_\beta \eta \big(\bar{U}(\boldsymbol{\theta}_k) - \bar{U}(\boldsymbol{y}_k)\big) - \eta \partial \bar{U}(\boldsymbol{y}_k)\boldsymbol{r}_k - \boldsymbol{r}_{k+1}^U.$$

Taylor expansion and Lemma D.2 give

$$\begin{aligned}
\boldsymbol{u}_{k+1} &= \beta_2 \boldsymbol{u}_k + \eta \partial \bar{U}(\boldsymbol{y}_k)\boldsymbol{S}_k \nabla F(\boldsymbol{y}_k) + \eta \boldsymbol{w}_{k+1} + \boldsymbol{r}_{k+1}^v, \\
\boldsymbol{w}_{k+1} &:= c_\beta \big(V(\boldsymbol{g}_k \boldsymbol{g}_k^\top) - \bar{U}(\boldsymbol{\theta}_k)\big) - \partial \bar{U}(\boldsymbol{y}_k)\boldsymbol{e}_k, \qquad \mathbb{E}[\boldsymbol{w}_{k+1} \mid \mathcal{F}_k] = \boldsymbol{0}, \\
\|\boldsymbol{r}_{k+1}^v\| &\le c_\beta \eta \|\bar{U}(\boldsymbol{\theta}_k) - \bar{U}(\boldsymbol{y}_k)\| + \eta \|\partial \bar{U}(\boldsymbol{y}_k)\boldsymbol{r}_k\| + C\|\Delta \boldsymbol{y}_k\|^2 \le C_M \eta^2.
\end{aligned}$$

Here $\boldsymbol{e}_k$ and $\boldsymbol{r}_k$ are the centered increment and conditional drift error from Definition D.1 and lemma D.2. For the exit index τ of the common parameter region, define

$$\boldsymbol{z}_n^{\mathrm{av}} := \eta \sum_{j<n} \beta_2^{n-1-j} \mathbf{1}_{\{j<\tau\}} \boldsymbol{w}_{j+1}.$$

For $1 \le i \le D$, put $Z_{k+1,i} := \mathbf{1}_{\{k<\tau\}}(\boldsymbol{w}_{k+1})_i$. The bounds on the fixed tube give a constant B_w independent of ϱ such that

$$\mathbb{E}[Z_{k+1,i} \mid \mathcal{F}_k] = 0, \qquad |Z_{k+1,i}| \le B_w, \qquad \mathbb{E}[e^{tZ_{k+1,i}} \mid \mathcal{F}_k] \le e^{t^2 B_w^2/2} \quad (t \in \mathbb{R}).$$

Applying this inequality with the deterministic weights $\eta\beta_2^{n-1-j}$, and then taking a union bound, gives

$$\eta^2 \sum_{j<n} \beta_2^{2(n-1-j)} \le \frac{\eta^2}{1-\beta_2^2} \le C\eta, \qquad \mathbb{P}\left(\max_{n \le N} \|\boldsymbol{z}_n^{\mathrm{av}}\| > x\right) \le C_D(N+1)e^{-cx^2/\eta}.$$

Thus $\max_{n \le N} \|\boldsymbol{z}_n^{\mathrm{av}}\| = \mathcal{O}_{\mathbb{P}}(\sqrt{\eta \log(1/\eta)})$. On $\{\tau > N\}$, variation of constants, $\|\nabla F(\boldsymbol{y})\| \le C \operatorname{dist}(\boldsymbol{y}, \Gamma)$, and $\eta \sum_{j<n} \beta_2^{n-1-j} \le c_\beta^{-1}$ give

$$\max_{n \le N} \|\boldsymbol{u}_n\| \le \|\boldsymbol{u}_0\| + \frac{L_U s_+ G}{c_\beta} \max_{j<N} \operatorname{dist}(\boldsymbol{y}_j, \Gamma) + \max_{n \le N} \|\boldsymbol{z}_n^{\mathrm{av}}\| + C_M \eta.$$

For a nearest point $\boldsymbol{\zeta} \in \Gamma$ to $\boldsymbol{y}$, Lipschitz continuity of $\bar{U}$ gives

$$\operatorname{dist}((\boldsymbol{y}, \boldsymbol{v}), \mathcal{M}) \le \|\boldsymbol{y} - \boldsymbol{\zeta}\| + \|\boldsymbol{v} - U(\boldsymbol{\zeta})\| \le (1 + L_U) \operatorname{dist}(\boldsymbol{y}, \Gamma) + \|\boldsymbol{v} - \bar{U}(\boldsymbol{y})\|.$$

The initial coupled-distance bound also gives

$$\|\boldsymbol{u}_0\| \le (1 + L_U)\epsilon_0 + L_U C_y \eta.$$

On $\{\tau > N\}$, the preceding estimates and the explicit radius choices give

$$\max_{n \le N} \operatorname{dist}(\widehat{\boldsymbol{X}}_n, \mathcal{M}) \le A\varrho + (1 + L_U)\epsilon_0 + \max_{n \le N} \|\boldsymbol{z}_n^{\mathrm{av}}\| + C_M \eta \le d_{\mathcal{M}}/4 + \max_{n \le N} \|\boldsymbol{z}_n^{\mathrm{av}}\| + C_M \eta.$$

For $C_M \eta \le d_{\mathcal{M}}/4$, this implies the explicit exit bound

$$\mathbb{P}\left(\exists n \le N : \widehat{\boldsymbol{X}}_n \notin \mathcal{W}_1\right) \le (N+1)e^{-c_*/\eta} + C_D(N+1)e^{-cd_{\mathcal{M}}^2/(4\eta)} \longrightarrow 0.$$

For the sharper estimate, fix $C_{\mathrm{b}} \geq c_e^{-1}$, set $k_{\mathrm{b}} = \lceil C_{\mathrm{b}} \eta^{-1} \log(1/\eta) \rceil$, and take $C = C_{\mathrm{b}} + 2/c_\beta$ in K_0. For $\eta \leq \min\{e^{-1}, (2c_\beta)^{-1}\}$,

$$\beta_2^{K_0 - k_{\mathrm{b}}} \leq e^{-c_\beta \eta (K_0 - k_{\mathrm{b}})} \leq e^{2c_\beta \eta} \eta^2 \leq \eta.$$

Restarting the same recursion at k_{b} gives, for $K_0 \leq n \leq N$ on $\{\tau > N\}$,

$$\|\boldsymbol{u}_n\| \leq \beta_2^{n-k_{\mathrm{b}}} \|\boldsymbol{u}_{k_{\mathrm{b}}}\| + C \sup_{k_{\mathrm{b}} \leq j \leq N} \operatorname{dist}(\boldsymbol{y}_j, \Gamma) + \|\boldsymbol{z}_n^{\mathrm{av}} - \beta_2^{n-k_{\mathrm{b}}} \boldsymbol{z}_{k_{\mathrm{b}}}^{\mathrm{av}}\| + C_M \eta = \mathcal{O}_{\mathbb{P}}(\sqrt{\eta \log(1/\eta)}),$$

uniformly in n, by (9) and boundedness of $\boldsymbol{u}_{k_{\mathrm{b}}}$ before parameter exit. Finally, $\|\boldsymbol{\theta}_k - \boldsymbol{y}_k\| \leq C_M \eta$ transfers this bound to the original coupled state. □

For each η, define the restarted variables and filtration by

$$(\boldsymbol{\theta}'_k, \boldsymbol{m}'_k, \boldsymbol{v}'_k) := (\boldsymbol{\theta}_{K_0+k}, \boldsymbol{m}_{K_0+k}, \boldsymbol{v}_{K_0+k}), \qquad \mathcal{F}'_k := \mathcal{F}_{K_0+k}, \qquad \xi'_k := \xi_{K_0+k}.$$

Then ξ'_k is independent of $\mathcal{F}'_k$, and the restarted variables satisfy the same AGM recursions, with $\|\boldsymbol{m}'_0\| \leq M + R$. Write $\widehat{\boldsymbol{X}}'_k := (\boldsymbol{\theta}'_k - \eta c_1 S(\boldsymbol{v}'_k)\boldsymbol{m}'_k, \boldsymbol{v}'_k)$ and $d_{\mathcal{M},0} := \operatorname{dist}(\mathcal{M}, \mathcal{W}_0^{\mathrm{c}}) > 0$. Lemma E.9 implies

$$\operatorname{dist}(\widehat{\boldsymbol{X}}'_0, \mathcal{M}) = o_{\mathbb{P}}(1), \qquad \mathbb{P}(\widehat{\boldsymbol{X}}'_0 \notin \mathcal{W}_0) \leq \mathbb{P}(\operatorname{dist}(\widehat{\boldsymbol{X}}'_0, \mathcal{M}) \geq d_{\mathcal{M},0}) = o(1).$$

In the remainder of this subsection, omit the primes from these restarted variables and their filtration.

We call $\mathcal{W}_1$ the *coupled working zone*. With $\widehat{\boldsymbol{X}}_k = (\boldsymbol{\theta}_k - \boldsymbol{Q}_{\eta,k}, \boldsymbol{v}_k)$, define

$$N_\eta(t) := \lfloor t/\eta^2 \rfloor, \qquad \widehat{\boldsymbol{X}}_\eta(t) := \widehat{\boldsymbol{X}}_{N_\eta(t)},$$
$$\tau_\eta := \inf\{t \geq 0 : \widehat{\boldsymbol{X}}_\eta(t) \notin \mathcal{W}_1\}, \qquad A_\eta(t) := \eta N_\eta(t).$$

The burn-in estimate and Lemma E.4 give $\mathbb{P}(\widehat{\boldsymbol{X}}_0 \in \mathcal{W}_0) \to 1$ and a uniform $\mathcal{O}_{\mathbb{P}}(\eta)$ state correction before exit. Together with (16), the separation distance d_1 between $\mathcal{W}_1$ and $\mathcal{W}_2$ keeps the original states before exit and the corrected exit state inside $\mathcal{W}_2$ with probability tending to one. To define the auxiliary dynamics outside this neighborhood, we multiply the local coefficients by a smooth cutoff that equals one near $\overline{\mathcal{W}}_2$. This extends the coefficients without changing the updates used in the localized analysis. We stop the process immediately after the update that first carries the corrected state outside $\mathcal{W}_1$, retaining the resulting exit state and holding it constant thereafter. Thus sums at time t end at $N_\eta(t \wedge \tau_\eta) - 1$, and drift integrals stop at $t \wedge \tau_\eta$. The unstopped clock satisfies $A_\eta(t + \delta) - A_\eta(t) \to \infty$ uniformly for $0 \leq t \leq T - \delta$. Lemma E.12 will remove the localization on each fixed slow-time interval. We are now ready to prove Theorem 3.2 with the framework of Katzenberger (1991).

Lemma E.10. *For the stopped process, write* (16) *as*

$$\widehat{\boldsymbol{X}}_\eta(t \wedge \tau_\eta) = \widehat{\boldsymbol{X}}_\eta(0) + \int_0^{t \wedge \tau_\eta} \boldsymbol{f}(\widehat{\boldsymbol{X}}_\eta(s-)) \, \mathrm{d}A_\eta(s) + \boldsymbol{Z}_\eta(t),$$

where $\boldsymbol{Z}_\eta$ has a decomposition $\boldsymbol{Z}_\eta = \boldsymbol{M}_\eta + \boldsymbol{B}_\eta$ into a local martingale and a predictable finite variation process. For every fixed T,

$$\sup_{t \leq T} \|\Delta \boldsymbol{Z}_\eta(t)\| = o_{\mathbb{P}}(1), \qquad \sup_\eta \mathbb{E} \operatorname{tr}\langle \boldsymbol{M}_\eta \rangle_T < \infty, \qquad \sup_\eta \mathbb{E} \operatorname{TV}(\boldsymbol{B}_\eta)_T < \infty,$$

where

$$\langle \boldsymbol{M} \rangle_T := \sum_{k < N_\eta(T \wedge \tau_\eta)} \mathbb{E}[\Delta \boldsymbol{M}(k+1) \Delta \boldsymbol{M}(k+1)^\top \mid \mathcal{F}_k],$$
$$\operatorname{TV}(\boldsymbol{B})_T = \sum_{k < N_\eta(T \wedge \tau_\eta)} \|\Delta \boldsymbol{B}(k+1)\|.$$

Proof. By (16), it is natural to set

$$\boldsymbol{M}_\eta(t) := \sum_{k < N_\eta(t \wedge \tau_\eta)} \big(\eta \boldsymbol{q}(\widehat{\boldsymbol{X}}_k, \xi_k) + \eta^2 \boldsymbol{r}_{\eta,k}^{\mathrm{cen}}\big)$$
$$+ \sum_{k < N_\eta(t \wedge \tau_\eta)} \big(\eta^2 (\boldsymbol{r}_{\eta,k}^{\mathrm{mom}} - \mathbb{E}[\boldsymbol{r}_{\eta,k}^{\mathrm{mom}} \mid \mathcal{F}_k]) + \boldsymbol{R}_{\eta,k} - \mathbb{E}[\boldsymbol{R}_{\eta,k} \mid \mathcal{F}_k]\big).$$

$$\boldsymbol{B}_\eta(t) := \sum_{k<N_\eta(t\wedge\tau_\eta)} \Big(\eta^2 \boldsymbol{b}^0(\widehat{\boldsymbol{X}}_k) + \eta^2 \mathbb{E}[\boldsymbol{r}^{\mathrm{mom}}_{\eta,k} \mid \mathcal{F}_k] + \mathbb{E}[\boldsymbol{R}_{\eta,k} \mid \mathcal{F}_k]\Big).$$

It remains to prove the estimations. Recall that for $k < N_\eta(\tau_\eta)$,

$$\begin{aligned} \|\boldsymbol{q}(\widehat{\boldsymbol{X}}_k, \xi_k)\| + \|\boldsymbol{b}^0(\widehat{\boldsymbol{X}}_k)\| + \|\boldsymbol{r}^{\mathrm{cen}}_{\eta,k}\| &\le C, \\ \|\boldsymbol{r}^{\mathrm{mom}}_{\eta,k}\| &\le C(1 + \|\boldsymbol{m}_0\|), \\ \|\boldsymbol{R}_{\eta,k}\| &\le C\eta^3(1 + \|\boldsymbol{m}_0\|^2). \end{aligned}$$

Since $\mathbb{E}[\boldsymbol{q}(\widehat{\boldsymbol{X}}_k, \xi_k)|\mathcal{F}_k] = \mathbb{E}[\boldsymbol{r}^{\mathrm{cen}}_{\eta,k}|\mathcal{F}_k] = \boldsymbol{0}$, by the conditional variance inequality,

$$\mathbb{E}\,\mathrm{tr}\langle \boldsymbol{M}_\eta\rangle_T \le C N_\eta(T)\left(\eta^2 + \eta^4\big(1 + \mathbb{E}\|\boldsymbol{m}_0\|^2\big) + \eta^6\big(1 + \mathbb{E}\|\boldsymbol{m}_0\|^4\big)\right) \le C_T.$$

By Jensen's inequality,

$$\mathbb{E}\,\mathrm{TV}(\boldsymbol{B}_\eta)_T \le C N_\eta(T)\left(\eta^2\big(1 + \mathbb{E}\|\boldsymbol{m}_0\|\big) + \eta^3\big(1 + \mathbb{E}\|\boldsymbol{m}_0\|^2\big)\right) \le C_T.$$

The uniform fourth moment assumption on $\boldsymbol{m}_0$ makes the right hand sides uniform in η. Finally,

$$\mathbb{E}\sup_{t\le\tau_\eta} \|\Delta \boldsymbol{Z}_\eta(t)\| \le C\left(\eta + \eta^2\big(1 + \mathbb{E}\|\boldsymbol{m}_0\|\big) + \eta^3\big(1 + \mathbb{E}\|\boldsymbol{m}_0\|^2\big)\right) \to 0,$$

which completes the proof. □

Lemma E.11. *For every fixed T,*

$$\sup_{0\le t\le T} \mathrm{dist}(\widehat{\boldsymbol{X}}_\eta(t\wedge\tau_\eta), \mathcal{M}) = o_{\mathbb{P}}(1). \tag{18}$$

The family of projected stopped paths $\{\pi(\widehat{\boldsymbol{X}}_\eta(\cdot\wedge\tau_\eta))\}_\eta$ is tight in $D([0,T],\Gamma)$, and every limit has continuous paths.

Proof. Recall that the processes in this subsection are restarted at K_0. By (17),

$$\sup_{0\le t\le T} \mathrm{dist}(\widehat{\boldsymbol{X}}_\eta(t), \mathcal{M}) = \max_{0\le k\le \lfloor T\eta^{-2}\rfloor} \mathrm{dist}(\widehat{\boldsymbol{X}}^{\mathrm{orig}}_{K_0+k}, \mathcal{M}) = \mathcal{O}_{\mathbb{P}}(\sqrt{\eta\log(1/\eta)}) = o_{\mathbb{P}}(1),$$

where $\widehat{\boldsymbol{X}}^{\mathrm{orig}}$ denotes the corrected state before restarting. Since $0 \le t\wedge\tau_\eta \le T$,

$$\sup_{0\le t\le T} \mathrm{dist}(\widehat{\boldsymbol{X}}_\eta(t\wedge\tau_\eta), \mathcal{M}) \le \sup_{0\le t\le T} \mathrm{dist}(\widehat{\boldsymbol{X}}_\eta(t), \mathcal{M}) = o_{\mathbb{P}}(1).$$

This proves (18). For tightness, the geometric hypotheses are provided by Appendix E.1, and the perturbation bounds by Lemma E.10. The compact containment and tightness results of Katzenberger (1991, Section 6) therefore apply to the projected stopped paths. The vanishing jump estimate in Lemma E.10 makes every projected limit continuous. □

Lemma E.12. *For every fixed $T < \infty$,*

$$\mathbb{P}(\tau_\eta \le T) \longrightarrow 0. \tag{19}$$

Proof. The coupled working-zone construction in Lemma E.8 gives

$$\delta := d_{\mathcal{M}} = \mathrm{dist}(\mathcal{M}, \mathcal{W}_1^c) > 0.$$

The maximal jump before $T\wedge\tau_\eta$ is $o_{\mathbb{P}}(1)$ by (16) and Lemma E.4. On $\{\tau_\eta \le T\}$, the stopped process therefore has a point at distance at least $\delta/2$ from $\mathcal{M}$ with probability tending to one. Consequently,

$$\mathbb{P}(\tau_\eta \le T) \le \mathbb{P}\left(\sup_{t\le T} \mathrm{dist}(\widehat{\boldsymbol{X}}_\eta(t\wedge\tau_\eta), \mathcal{M}) \ge \delta/2\right) + o(1),$$

and the right hand side converges to zero by (18). □

Lemma E.13. *Define the generator*

$$\mathcal{A}_1 f(\boldsymbol{\zeta}) = \partial(f \circ \pi)(\iota(\boldsymbol{\zeta}))\boldsymbol{b}_{\boldsymbol{\zeta}} + \frac{1}{2}\partial^2(f \circ \pi)(\iota(\boldsymbol{\zeta}))[\boldsymbol{C}_{\boldsymbol{\zeta}}]. \tag{20}$$

If $\boldsymbol{\zeta}$ is a subsequential limit of the projected stopped process, then for every $f \in C^3(\Gamma)$,

$$M_f(t) := f(\boldsymbol{\zeta}(t)) - f(\boldsymbol{\zeta}(0)) - \int_0^t \mathcal{A}_1 f(\boldsymbol{\zeta}(s))\,\mathrm{d}s \tag{21}$$

is a martingale with respect to the natural filtration of $\boldsymbol{\zeta}$. This martingale problem is well posed and its solution is the process in (2)*.*

Proof. By (18) and continuity of $\boldsymbol{b}^0$, we have

$$\eta^2 \sum_{k<t/\eta^2} \boldsymbol{b}^0(\widehat{\boldsymbol{X}}_k) - \eta^2 \sum_{k<t/\eta^2} \boldsymbol{b}_{\pi(\widehat{\boldsymbol{X}}_k)} = o_{\mathbb{P}}(1)$$

uniformly in $t < T$. Recall that in Lemma E.10 we have $\|\boldsymbol{r}^{\mathrm{cen}}_{\eta,k}\| \le C$. Doob's inequality give

$$\mathbb{E}\sup_{t\le T}\left\|\eta^2 \sum_{k<t/\eta^2} \boldsymbol{r}^{\mathrm{cen}}_{\eta,k}\right\|^2 \le C\eta^4 N_\eta(T) \le CT\eta^2 \to 0.$$

For $\boldsymbol{x} = (\boldsymbol{\theta}, \boldsymbol{v})$, let

$$\boldsymbol{C}^0(\boldsymbol{x}) := \mathbb{E}\big[\boldsymbol{q}(\boldsymbol{x},\xi)\boldsymbol{q}(\boldsymbol{x},\xi)^\top\big].$$

Then for $\boldsymbol{\zeta} \in \Gamma$, $\boldsymbol{C}^0(\iota(\boldsymbol{\zeta})) = \boldsymbol{C}_{\boldsymbol{\zeta}}$. Since $\widehat{\boldsymbol{X}}_k$ is $\mathcal{F}_k$ measurable and ξ_k is independent of $\mathcal{F}_k$, the conditional covariance of the leading martingale increment is

$$\mathbb{E}\left[\eta\boldsymbol{q}(\widehat{\boldsymbol{X}}_k,\xi_k)\big(\eta\boldsymbol{q}(\widehat{\boldsymbol{X}}_k,\xi_k)\big)^\top \mid \mathcal{F}_k\right] = \eta^2\boldsymbol{C}^0(\widehat{\boldsymbol{X}}_k).$$

Hence (18) and continuity of $\boldsymbol{C}^0$ imply

$$\sup_{t\le T}\left\|\eta^2 \sum_{k<t/\eta^2}\Big(\boldsymbol{C}^0(\widehat{\boldsymbol{X}}_k) - \boldsymbol{C}_{\pi(\widehat{\boldsymbol{X}}_k)}\Big)\right\| = o_{\mathbb{P}}(1).$$

Along every subsequence for which $\pi(\widehat{\boldsymbol{X}}_\eta) \Rightarrow \boldsymbol{\zeta}$, the remaining sum is a Riemann sum, so

$$\eta^2 \sum_{k<t/\eta^2} \boldsymbol{C}_{\pi(\widehat{\boldsymbol{X}}_k)} \Rightarrow \int_0^t \boldsymbol{C}_{\boldsymbol{\zeta}(s)}\,\mathrm{d}s \quad \text{in } D([0,T],\mathbb{R}^{(d+D)\times(d+D)}).$$

Let $f \in C^3(\Gamma)$ and $F = f \circ \pi$. Then $\partial F\boldsymbol{f} = \boldsymbol{0}$. We first control the momentum term after multiplication by ∂F. Recall the map $\boldsymbol{A}$ in the proof of Lemma E.7, for which

$$\mathbb{E}[\boldsymbol{r}^{\mathrm{mom}}_{\eta,k} \mid \mathcal{F}_k] = \boldsymbol{A}(\widehat{\boldsymbol{X}}_k)\boldsymbol{m}_k.$$

At $\iota(\boldsymbol{\zeta}) = (\boldsymbol{\zeta}, U(\boldsymbol{\zeta}))$, we have $\bar{U}(\boldsymbol{\zeta}) = U(\boldsymbol{\zeta})$ and $\boldsymbol{J}_{\boldsymbol{\zeta}}\boldsymbol{S}_*(\boldsymbol{\zeta})\nabla^2\mathcal{L}(\boldsymbol{\zeta}) = \boldsymbol{O}$. Corollary E.2 therefore gives, for every $\boldsymbol{m}$,

$$\partial\pi(\iota(\boldsymbol{\zeta}))\boldsymbol{A}(\iota(\boldsymbol{\zeta}))\boldsymbol{m} = -c_1\boldsymbol{J}_{\boldsymbol{\zeta}}\boldsymbol{S}_*(\boldsymbol{\zeta})\nabla^2\mathcal{L}(\boldsymbol{\zeta})\boldsymbol{S}_*(\boldsymbol{\zeta})\boldsymbol{m} = \boldsymbol{0}.$$

Since $F = f \circ \pi$, smoothness on the working zone implies

$$\|\partial F(\boldsymbol{x})\boldsymbol{A}(\boldsymbol{x})\| \le C\,\mathrm{dist}(\boldsymbol{x},\mathcal{M}).$$

The uniform bound on $\boldsymbol{m}_k$ and (18) now yield

$$\begin{aligned}\eta^2 \sum_{k<N_\eta(T\wedge\tau_\eta)} \Big|\partial F(\widehat{\boldsymbol{X}}_k)\mathbb{E}[\boldsymbol{r}^{\mathrm{mom}}_{\eta,k} \mid \mathcal{F}_k]\Big| &\le CT\sup_{t\le T}\mathrm{dist}(\widehat{\boldsymbol{X}}_\eta(t\wedge\tau_\eta),\mathcal{M}) \\ &= o_{\mathbb{P}}(1).\end{aligned}$$

Thus the momentum contribution to the projected conditional drift is negligible even after multiplication by ∂F. Since $\boldsymbol{f}$ vanishes on $\mathcal{M}$,

$$\eta^2 \sum_{k<T/\eta^2} \left(\|\boldsymbol{f}(\widehat{\boldsymbol{X}}_k)\| + \|\boldsymbol{f}(\widehat{\boldsymbol{X}}_k)\|^2 \right)$$
$$\leq CT \left(\sup_{t\leq T} \operatorname{dist}(\widehat{\boldsymbol{X}}_\eta(t\wedge\tau_\eta), \mathcal{M}) + \sup_{t\leq T} \operatorname{dist}(\widehat{\boldsymbol{X}}_\eta(t\wedge\tau_\eta), \mathcal{M})^2 \right) = o_{\mathbb{P}}(1).$$

Thus all second order terms containing $\boldsymbol{f}(\widehat{\boldsymbol{X}}_k)$ vanish after summation. Let

$$\Delta \boldsymbol{Z}_{\eta,k+1} := \boldsymbol{Z}_\eta((k+1)\eta^2) - \boldsymbol{Z}_\eta(k\eta^2).$$

The conditional second order expansion in the projection formula gives

$$\mathbb{E}\left[F(\widehat{\boldsymbol{X}}_{k+1}) - F(\widehat{\boldsymbol{X}}_k) \mid \mathcal{F}_k \right]$$
$$= \partial F(\widehat{\boldsymbol{X}}_k)\mathbb{E}[\Delta \boldsymbol{Z}_{\eta,k+1} \mid \mathcal{F}_k] + \frac{1}{2}\partial^2 F(\widehat{\boldsymbol{X}}_k)\left[\mathbb{E}[\Delta \boldsymbol{Z}_{\eta,k+1}\Delta \boldsymbol{Z}_{\eta,k+1}^\top \mid \mathcal{F}_k]\right] + \varepsilon^f_{\eta,k}$$
$$= \eta^2 \partial F(\widehat{\boldsymbol{X}}_k)\boldsymbol{b}^0(\widehat{\boldsymbol{X}}_k) + \frac{\eta^2}{2}\partial^2 F(\widehat{\boldsymbol{X}}_k)[\boldsymbol{C}^0(\widehat{\boldsymbol{X}}_k)] + \widetilde{\varepsilon}^f_{\eta,k},$$

where the two error arrays are cumulatively negligible in the sense that

$$\sup_{t\leq T}\left| \sum_{k<t/\eta^2} \varepsilon^f_{\eta,k} \right| + \sup_{t\leq T}\left| \sum_{k<t/\eta^2} \widetilde{\varepsilon}^f_{\eta,k} \right| = o_{\mathbb{P}}(1).$$

Consequently, the predictable compensator in the projection formula of Katzenberger (1991) is asymptotically equal to

$$A_{\eta,f}(t) := \eta^2 \sum_{k<t/\eta^2} \left(\partial(f\circ\pi)(\iota(\pi(\widehat{\boldsymbol{X}}_k)))\boldsymbol{b}_{\pi(\widehat{\boldsymbol{X}}_k)} + \frac{1}{2}\partial^2(f\circ\pi)(\iota(\pi(\widehat{\boldsymbol{X}}_k)))[\boldsymbol{C}_{\pi(\widehat{\boldsymbol{X}}_k)}] \right).$$

The convergence gives

$$A_{\eta,f}(t) \Rightarrow \int_0^t \left(\partial(f\circ\pi)(\iota(\boldsymbol{\zeta}(s)))\boldsymbol{b}_{\boldsymbol{\zeta}(s)} + \frac{1}{2}\partial^2(f\circ\pi)(\iota(\boldsymbol{\zeta}(s)))[\boldsymbol{C}_{\boldsymbol{\zeta}(s)}] \right) \mathrm{d}s = \int_0^t \mathcal{A}_1 f(\boldsymbol{\zeta}(s))\,\mathrm{d}s.$$

The projection formula makes $f(\pi(\widehat{\boldsymbol{X}}_\eta(t))) - f(\pi(\widehat{\boldsymbol{X}}_\eta(0))) - A_{\eta,f}(t)$ an asymptotic martingale. Passing to the limit proves (21), whose compensator has integrand (20).

By Corollary E.2, the diffusion covariance in the parameter component is

$$\partial\pi(\iota(\boldsymbol{\zeta}))\boldsymbol{C}_{\boldsymbol{\zeta}}\partial\pi(\iota(\boldsymbol{\zeta}))^\top = \boldsymbol{J}_{\boldsymbol{\zeta}}\boldsymbol{S}_*(\boldsymbol{\zeta})\boldsymbol{\Sigma}(\boldsymbol{\zeta})\boldsymbol{S}_*(\boldsymbol{\zeta})\boldsymbol{J}_{\boldsymbol{\zeta}}^\top.$$

The drift is

$$\partial_{(\boldsymbol{\theta},\boldsymbol{v})}\pi(\iota(\boldsymbol{\zeta}))\boldsymbol{b}_{\boldsymbol{\zeta}} + \frac{1}{2}\partial^2_{(\boldsymbol{\theta},\boldsymbol{v})}\pi(\iota(\boldsymbol{\zeta}))[\boldsymbol{C}_{\boldsymbol{\zeta}}],$$

which is exactly the drift in (2). Applying the second order chain rule to $f\circ\pi$ combines this drift with the diffusion covariance and recovers the generator (20).

It remains to prove that every subsequential limit of the projected stopped process solves the martingale problem. Since the coefficients are locally Lipschitz after a smooth extension from the compact manifold Γ to a tubular neighborhood, it remains to verify that a solution starting on Γ stays there.

Let ρ be any local defining function for Γ. Differentiating $\rho(\pi(\boldsymbol{x})) = 0$ once and twice gives

$$\partial\rho(\boldsymbol{\zeta})\boldsymbol{A}^{\mathrm{diff}}_{\boldsymbol{\zeta}} = 0,$$

and

$$\partial\rho(\boldsymbol{\zeta})\boldsymbol{b}^{\mathrm{lim}}_{\boldsymbol{\zeta}} + \frac{1}{2}\partial^2\rho(\boldsymbol{\zeta})[\boldsymbol{A}^{\mathrm{diff}}_{\boldsymbol{\zeta}}(\boldsymbol{A}^{\mathrm{diff}}_{\boldsymbol{\zeta}})^\top] = 0,$$

where $\boldsymbol{A}^{\mathrm{diff}}_{\boldsymbol{\zeta}} := \boldsymbol{J}_{\boldsymbol{\zeta}}\boldsymbol{S}_*(\boldsymbol{\zeta})\sigma(\boldsymbol{\zeta})$ and $\boldsymbol{b}^{\mathrm{lim}}_{\boldsymbol{\zeta}}$ is the limiting drift. Itô's formula gives $\rho(\boldsymbol{\zeta}(t)) = 0$. Existence and uniqueness for the extended equation therefore give uniqueness in law on Γ. Consequently, the entire family of projected stopped paths converges to this solution. □

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. We use the restarted variables introduced after Lemma E.9, so that step zero here corresponds to step K_0 in the theorem. Write $\boldsymbol{\zeta}_0^\eta := \pi(\boldsymbol{\theta}_0, \boldsymbol{v}_0)$ and let $\boldsymbol{X}^\eta = \iota(\boldsymbol{\zeta}^\eta)$ denote the comparison SDE initialized at this random point, with its driving Brownian motion independent of the restarted initial information. The superscript records the possible dependence of its initial law on η; the SDE coefficients do not depend on η.

Fix an arbitrary sequence $\eta_n \to 0$. Since Γ is compact, we may pass to a subsequence such that

$$\mathrm{Law}(\boldsymbol{\zeta}_0^{\eta_n}) \Rightarrow \mu$$

for a probability measure μ on Γ. Let $\boldsymbol{\zeta}^\mu$ solve (2) with initial law μ, and set $\boldsymbol{X}^\mu := \iota(\boldsymbol{\zeta}^\mu)$. The bounded restarted momentum and the momentum correction give

$$\sup_{0\le t\le T} \|\boldsymbol{\theta}_{\lfloor t/\eta^2\rfloor} - \widehat{\boldsymbol{\theta}}_{\lfloor t/\eta^2\rfloor}\| \le C\eta(1+\|\boldsymbol{m}_0\|) = o_{\mathbb{P}}(1).$$

By burn-in, the original and corrected initial states lie in the coupled working zone with probability tending to one. Local smoothness of π therefore shows that their projected initial points differ by $o_{\mathbb{P}}(1)$, so both projected initial laws converge to μ along this subsequence.

The stopped-process tightness and limit identification in Lemma E.13, together with uniqueness of the martingale problem, identify every subsequential limit with $\boldsymbol{X}^\mu$. By Lemma E.12, stopping changes the path on $[0,T]$ only on an event whose probability tends to zero. The vanishing momentum correction and local continuity of Π then give

$$\bar{\boldsymbol{X}}_{\eta_n} \Rightarrow \boldsymbol{X}^\mu \quad \text{in } D([0,T],\mathcal{M}), \qquad \bar{\boldsymbol{X}}_\eta(t) := \Pi(\boldsymbol{\theta}_{\lfloor t/\eta^2\rfloor}, \boldsymbol{v}_{\lfloor t/\eta^2\rfloor}).$$

Here the fixed value assigned to Π outside its attraction domain keeps the projected process in $\mathcal{M}$; only its local continuity is used on the event of no exit.

The SDE coefficients admit locally Lipschitz extensions near the compact invariant manifold Γ. A cutoff yields globally Lipschitz coefficients agreeing there, whose solutions depend continuously in path law on their initial distributions. Uniqueness and invariance of Γ thus imply

$$\boldsymbol{X}^{\eta_n} \Rightarrow \boldsymbol{X}^\mu \quad \text{in } C([0,T],\mathcal{M}).$$

It remains to make the comparison uniform over the time grid. If the claimed maximum did not tend to zero, there would exist $\varepsilon > 0$, a sequence $\eta_n \to 0$, and indices $0 \le k_n \le K_{\eta_n}$ such that

$$\left|\mathbb{E}[g(\bar{\boldsymbol{X}}_{\eta_n}(t_n))] - \mathbb{E}[g(\boldsymbol{X}^{\eta_n}(t_n))]\right| \ge \varepsilon, \qquad t_n := k_n\eta_n^2.$$

Pass to a further subsequence for which the projected initial laws converge and $t_n \to t \in [0,T]$. The two path convergences above and continuity of the limiting paths imply

$$\bar{\boldsymbol{X}}_{\eta_n}(t_n) \Rightarrow \boldsymbol{X}^\mu(t), \qquad \boldsymbol{X}^{\eta_n}(t_n) \Rightarrow \boldsymbol{X}^\mu(t).$$

Since g is bounded and continuous on the compact set $\mathcal{M}$, both expectations converge to $\mathbb{E}[g(\boldsymbol{X}^\mu(t))]$, contradicting the displayed lower bound. Restoring the shift by K_0 proves the unconditional weak approximation in the theorem. □

F SCHEMES $(1,2)$

In this section, we will study schemes $(1,2)$ and give a detailed proof for Theorem 3.3.

F.1 WORKING ZONE AND ADAPTIVE PROJECTION

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. Following the construction of the working zone of Li et al. (2025), we construct the working zones around any local minimizer manifold Γ such that iterations inside the working zones will be captured by the manifold and we obtain certain properties that support the analysis of slow SDE. We first recall two lemmas from Li et al. (2025).

Definition F.1 (Manifold neighborhood)**.** *For a manifold Γ and $\epsilon > 0$, define its ϵ-neighborhood by*

$$\Gamma^\epsilon := \{\boldsymbol{\theta} : \mathrm{dist}(\boldsymbol{\theta}, \Gamma) \le \epsilon\}.$$

Definition F.2 (Preconditioned gradient flow, restatement of definition 3.2)**.** *For any differentiable function $\mathcal{L}$ and any matrix $\boldsymbol{S} \in \mathbb{R}^{d\times d}$, the* $\boldsymbol{S}$-preconditioned gradient flow *of $\mathcal{L}$ is the ordinary differential equation*

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

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

Lemma F.1 (Lemma C.1 of Li et al. (2025))**.** *Let $C_1, C_2 > 0$ with $C_1 \le C_2$, and let $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ be ρ-smooth and satisfy the μ-PL condition. For any symmetric matrix $\boldsymbol{S}$ with $C_1\boldsymbol{I} \preceq \boldsymbol{S} \preceq C_2\boldsymbol{I}$, consider the $\boldsymbol{S}$-preconditioned gradient flow of $\mathcal{L}$ starting at $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$. Then for any $T > 0$,*

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

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

Lemma F.2 (Lemma C.2 of Li et al. (2025))**.** *Assume that $\boldsymbol{v}_0 \in \mathbb{R}^D_{\ge 0}$ is uniformly bounded. There exist constants $R_1, R_2 > 0$ such that, almost surely,*

$$\mathbf{0} \le \boldsymbol{v}_k \le R_1\mathbf{1}, \qquad k \ge 0,$$

and

$$S(\boldsymbol{v}) \preceq R_2\boldsymbol{I}, \qquad \boldsymbol{v} \in [0, R_1]^D.$$

Moreover, S is Lipschitz on $[0, R_1]^D$.

We now introduce the objects used in the local analysis.

Definition F.3. *For every $\epsilon > 0$, define $\mathcal{X}^\epsilon := \Gamma^\epsilon \times [0, R_1]^D$.*

Definition F.4 (Choice of ϵ_4)**.** *Let U be the open neighborhood in Assumption 2.3 on which $\Gamma = \arg\min_{\boldsymbol{\theta}\in U} \mathcal{L}(\boldsymbol{\theta})$ and let $\epsilon_4 > 0$ such that $\Gamma^{\epsilon_4} \subset U$.*

Such ϵ_4 exists since Γ is compact and U is open. For the smoothness assertions below. Since S is C^∞-smooth, it admits a C^∞ extension to an open neighborhood of the compact box $[0, R_1]^D$.

Definition F.5 (Adaptive gradient-flow projection)**.** *For every $(\boldsymbol{\theta}, \boldsymbol{v})$ for which the $S(\boldsymbol{v})$-preconditioned gradient flow from $\boldsymbol{\theta}$ converges to a point of Γ, define*

$$\Psi(\boldsymbol{\theta}, \boldsymbol{v}) := \Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta}).$$

The Working Zone Lemma below will show that the domain of Ψ contains $\mathcal{X}^{\epsilon_2}$ for some $\epsilon_2 > 0$.

Definition F.6 (Slow interpolation and projected coordinates)**.** *For $\eta > 0$ and $t \ge 0$, define*

$$\boldsymbol{\theta}_\eta(t) := \boldsymbol{\theta}_{\lfloor t/\eta^2 \rfloor}, \qquad \boldsymbol{v}_\eta(t) := \boldsymbol{v}_{\lfloor t/\eta^2 \rfloor}.$$

Whenever $\Psi(\boldsymbol{\theta}_\eta(t), \boldsymbol{v}_\eta(t))$ is defined, set

$$\boldsymbol{\zeta}_\eta(t) := \Psi(\boldsymbol{\theta}_\eta(t), \boldsymbol{v}_\eta(t)), \qquad \boldsymbol{r}_\eta(t) := \boldsymbol{\theta}_\eta(t) - \boldsymbol{\zeta}_\eta(t).$$

Definition F.7 (Working-zone exit time)**.** *For $\epsilon > 0$ and a fixed horizon $T > 0$, define*

$$\tau_\eta^\epsilon := \inf\{t \ge 0 : (\boldsymbol{\theta}_\eta(t), \boldsymbol{v}_\eta(t)) \notin \mathcal{X}^\epsilon\} \wedge T.$$

For the inner radius ϵ_1 furnished by Lemmas F.3 and F.4 below, we abbreviate $\tau_\eta := \tau_\eta^{\epsilon_1}$.

We will show that for any $T > 0$, $\mathbb{P}(\tau_\eta < T) \to 0$ as $\eta \to 0$.

Recall that we have for every $\boldsymbol{v} \in [0, R_1]^D$, $R_0^{-1}\boldsymbol{I} \preceq S(\boldsymbol{v}) \preceq R_2\boldsymbol{I}$. We now construct nested working zones $(\Gamma^{\epsilon_1}, \Gamma^{\epsilon_2}, \Gamma^{\epsilon_3})$:

Lemma F.3 (Working Zone Lemma)**.** *Under Assumption 2.1– 2.3, there exist positive constants $\epsilon_1, \epsilon_2, \epsilon_3, \mu$ such that $\epsilon_1 < \epsilon_2 < \epsilon_3 < \epsilon_4$ and the following properties hold:*

1. *The loss is μ-PL on Γ^{ϵ_3}:*

$$2\mu\big(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^*\big) \le \|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2, \qquad \boldsymbol{\theta} \in \Gamma^{\epsilon_3}. \tag{22}$$

2. *For every constant matrix $\boldsymbol{S}$ satisfying $R_0^{-1}\boldsymbol{I} \preceq \boldsymbol{S} \preceq R_2\boldsymbol{I}$, the $\boldsymbol{S}$-preconditioned gradient flow starting from any $\boldsymbol{\theta}_0 \in \Gamma^{\epsilon_2}$ remains in Γ^{ϵ_3} and converges to a point of Γ. Consequently, Ψ is well-defined on $\mathcal{X}^{\epsilon_2}$ and takes values in Γ.*

The inner radius ϵ_1 may be decreased without affecting either property.

Proof. Apply Lemmas D.1 and D.3 with $s_- = R_0^{-1}$ and $s_+ = R_2$. Let

$$\epsilon_3 := \tfrac{1}{2}\min\{r, \epsilon_4\}, \qquad \mu := a_3/2, \qquad \epsilon_2 := \tfrac{1}{2}\min\{r_\Phi, \epsilon_3/(1 + C_\Phi\sqrt{a_2})\}, \qquad \epsilon_1 := \epsilon_2/2.$$

Then $0 < \epsilon_1 < \epsilon_2 < \epsilon_3 < \epsilon_4$ and $(1 + C_\Phi\sqrt{a_2})\epsilon_2 \le \epsilon_3/2$. The local PL bound follows from $\|\nabla\mathcal{L}\|^2 \ge a_3 F$. For the frozen-$\boldsymbol{S}$ flow, the trajectory-length estimate in that proof gives, for all $t \ge 0$,

$$\operatorname{dist}(\boldsymbol{\theta}(t), \Gamma) \le \operatorname{dist}(\boldsymbol{\theta}_0, \Gamma) + C_\Phi\sqrt{F(\boldsymbol{\theta}_0)} \le (1 + C_\Phi\sqrt{a_2})\epsilon_2 < \epsilon_3, \qquad \lim_{t\to\infty}\boldsymbol{\theta}(t) = \Phi_{\boldsymbol{S}}(\boldsymbol{\theta}_0) \in \Gamma.$$

Both conclusions remain valid for every replacement $\widetilde{\epsilon}_1 \in (0, \epsilon_1]$ since neither bound depends on ϵ_1. □

Lemma F.4 (Smooth adaptive projection)**.** *Under the assumptions of Lemma F.3, after decreasing ϵ_1 if necessary, there exists an open neighborhood $\mathcal{U}_\Psi$ of $\mathcal{X}^{\epsilon_1}$ on which Ψ is C^∞. For every fixed integer $q \ge 1$, there is a constant $C_{\Phi,q} < \infty$ such that*

$$\max_{1\le j\le q}\sup_{(\boldsymbol{\theta},\boldsymbol{v})\in\mathcal{X}^{\epsilon_1}} \|\partial^j\Psi(\boldsymbol{\theta}, \boldsymbol{v})\| \le C_{\Phi,q}.$$

Moreover, the adaptive projection satisfies

$$\partial_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta}, \boldsymbol{v})S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}, \qquad (\boldsymbol{\theta}, \boldsymbol{v}) \in \mathcal{X}^{\epsilon_1}, \tag{23}$$

and, for $\boldsymbol{\zeta} \in \Gamma$ and $\boldsymbol{v} \in [0, R_1]^D$,

$$\Psi(\boldsymbol{\zeta}, \boldsymbol{v}) = \boldsymbol{\zeta}, \qquad \partial_{\boldsymbol{v}}\Psi(\boldsymbol{\zeta}, \boldsymbol{v}) = 0.$$

Consequently, there exists $C_\Phi < \infty$ such that

$$\|\partial_{\boldsymbol{v}}\Psi(\boldsymbol{\theta}, \boldsymbol{v})\| \le C_\Phi \operatorname{dist}(\boldsymbol{\theta}, \Gamma), \qquad (\boldsymbol{\theta}, \boldsymbol{v}) \in \mathcal{X}^{\epsilon_1}. \tag{24}$$

Finally, writing $\boldsymbol{r}(\boldsymbol{\theta}, \boldsymbol{v}) := \boldsymbol{\theta} - \Psi(\boldsymbol{\theta}, \boldsymbol{v})$, there exist constants $0 < c_r < C_r < \infty$ such that

$$c_r \operatorname{dist}(\boldsymbol{\theta}, \Gamma) \le \|\boldsymbol{r}(\boldsymbol{\theta}, \boldsymbol{v})\| \le C_r \operatorname{dist}(\boldsymbol{\theta}, \Gamma), \qquad (\boldsymbol{\theta}, \boldsymbol{v}) \in \mathcal{X}^{\epsilon_1}. \tag{25}$$

Proof. Consider the C^∞ vector field

$$\mathcal{F}(\boldsymbol{\theta}, \boldsymbol{v}) := \big(-S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta}), \mathbf{0}\big)$$

on an open neighborhood of $\Gamma \times [0, R_1]^D$. At $(\boldsymbol{\zeta}, \boldsymbol{v}) \in \Gamma \times [0, R_1]^D$, its linearization in the $\boldsymbol{\theta}$-coordinate is $-S(\boldsymbol{v})\boldsymbol{H}_{\boldsymbol{\zeta}}$, where $\boldsymbol{H}_{\boldsymbol{\zeta}} := \nabla^2\mathcal{L}(\boldsymbol{\zeta})$. Since Γ consists of local minimizers and $\operatorname{rank}(\boldsymbol{H}_{\boldsymbol{\zeta}}) = m$, we have $\ker(\boldsymbol{H}_{\boldsymbol{\zeta}}) = T_{\boldsymbol{\zeta}}\Gamma$, while the positive spectrum on $N_{\boldsymbol{\zeta}}\Gamma$ is uniformly bounded away from zero. The compactness of $\Gamma \times [0, R_1]^D$ and the uniform spectral bounds on $S(\boldsymbol{v})$ make this normal attraction uniform in $(\boldsymbol{\zeta}, \boldsymbol{v})$.

The parameter-dependent stable-foliation construction used in Lemma C.3 of Li et al. (2025) therefore gives an open neighborhood of $\Gamma \times [0, R_1]^D$ on which the asymptotic-phase map $(\boldsymbol{\theta}, \boldsymbol{v}) \mapsto \Psi(\boldsymbol{\theta}, \boldsymbol{v})$ is C^∞. Compactness allows us to decrease ϵ_1 so that $\mathcal{X}^{\epsilon_1}$ lies in this neighborhood, and it gives the stated derivative bounds.

Let $\boldsymbol{\theta}(t)$ be the $S(\boldsymbol{v})$-preconditioned flow from $\boldsymbol{\theta}$. Its limiting point is unchanged by a time shift, so

$$\Psi(\boldsymbol{\theta}(t), \boldsymbol{v}) = \Psi(\boldsymbol{\theta}, \boldsymbol{v}).$$

Differentiating at $t = 0$ proves (23). If $\boldsymbol{\zeta} \in \Gamma$, the flow remains at $\boldsymbol{\zeta}$, so $\Psi(\boldsymbol{\zeta}, \boldsymbol{v}) = \boldsymbol{\zeta}$ for every $\boldsymbol{v}$ and $\partial_{\boldsymbol{v}} \Psi(\boldsymbol{\zeta}, \boldsymbol{v}) = 0$. The mean-value theorem and the uniform mixed-derivative bound give

$$\|\partial_{\boldsymbol{v}} \Psi(\boldsymbol{\theta}, \boldsymbol{v})\| \le C_\Phi \operatorname{dist}(\boldsymbol{\theta}, \Gamma).$$

Finally, let $\boldsymbol{\zeta}^* \in \Gamma$ be a nearest point to $\boldsymbol{\theta}$. Since $\Psi(\boldsymbol{\theta}, \boldsymbol{v}) \in \Gamma$,

$$\operatorname{dist}(\boldsymbol{\theta}, \Gamma) \le \|\boldsymbol{\theta} - \Psi(\boldsymbol{\theta}, \boldsymbol{v})\|.$$

The uniform Lipschitz bound for Ψ and $\Psi(\boldsymbol{\zeta}^*, \boldsymbol{v}) = \boldsymbol{\zeta}^*$ also imply

$$\|\boldsymbol{\theta} - \Psi(\boldsymbol{\theta}, \boldsymbol{v})\| \le \|\boldsymbol{\theta} - \boldsymbol{\zeta}^*\| + \|\Psi(\boldsymbol{\theta}, \boldsymbol{v}) - \Psi(\boldsymbol{\zeta}^*, \boldsymbol{v})\| \le C_r \operatorname{dist}(\boldsymbol{\theta}, \Gamma).$$

This proves (25). □

The next lemma collects the projection identities from Lemmas E.2 and E.3 of Li et al. (2025) that are used below.

Lemma F.5 (Action of the projection derivative)**.** *For every $\boldsymbol{\zeta} \in \Gamma$, $\boldsymbol{v} \in [0, R_1]^D$, and $\boldsymbol{H}_{\boldsymbol{\zeta}} := \nabla^2 \mathcal{L}(\boldsymbol{\zeta})$, the following identities hold:*

$$\partial_{\boldsymbol{\theta}} \Psi(\boldsymbol{\zeta}, \boldsymbol{v}) \boldsymbol{u} = \boldsymbol{u}, \qquad \boldsymbol{u} \in T_{\boldsymbol{\zeta}} \Gamma,$$

$$\partial_{\boldsymbol{\theta}} \Psi(\boldsymbol{\zeta}, \boldsymbol{v}) S(\boldsymbol{v}) \boldsymbol{w} = \boldsymbol{0}, \qquad \boldsymbol{w} \in N_{\boldsymbol{\zeta}} \Gamma.$$

Consequently,

$$\partial_{\boldsymbol{\theta}} \Psi(\boldsymbol{\zeta}, \boldsymbol{v}) S(\boldsymbol{v}) \boldsymbol{H}_{\boldsymbol{\zeta}} = \boldsymbol{0}. \tag{26}$$

Lemma F.6 (Second-order Taylor expansion of the adaptive projection)**.** *Let*

$$\boldsymbol{\zeta}_k := \Psi(\boldsymbol{\theta}_k, \boldsymbol{v}_k), \qquad \Delta\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k, \qquad \Delta\boldsymbol{v}_{k+1} := \boldsymbol{v}_{k+1} - \boldsymbol{v}_k,$$

and use the shorthand $\partial^j \Psi_k := \partial^j \Psi(\boldsymbol{\theta}_k, \boldsymbol{v}_k)$. Whenever the line segment joining $(\boldsymbol{\theta}_k, \boldsymbol{v}_k)$ and $(\boldsymbol{\theta}_{k+1}, \boldsymbol{v}_{k+1})$ lies in the smooth extension neighborhood furnished by Lemma F.4,

$$\begin{aligned} \boldsymbol{\zeta}_{k+1} - \boldsymbol{\zeta}_k = {} & \partial_{\boldsymbol{\theta}} \Psi_k[\Delta\boldsymbol{\theta}_{k+1}] + \partial_{\boldsymbol{v}} \Psi_k[\Delta\boldsymbol{v}_{k+1}] + \frac{1}{2} \partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \Psi_k[\Delta\boldsymbol{\theta}_{k+1}, \Delta\boldsymbol{\theta}_{k+1}] \\ & + \partial^2_{\boldsymbol{\theta}\boldsymbol{v}} \Psi_k[\Delta\boldsymbol{\theta}_{k+1}, \Delta\boldsymbol{v}_{k+1}] + \frac{1}{2} \partial^2_{\boldsymbol{v}\boldsymbol{v}} \Psi_k[\Delta\boldsymbol{v}_{k+1}, \Delta\boldsymbol{v}_{k+1}] + \mathcal{R}^{\Psi}_{\eta, k+1}. \end{aligned} \tag{27}$$

Moreover, uniformly on every compact subset of that neighborhood,

$$\|\mathcal{R}^{\Psi}_{\eta, k+1}\| \le C\big(\|\Delta\boldsymbol{\theta}_{k+1}\| + \|\Delta\boldsymbol{v}_{k+1}\|\big)^3. \tag{28}$$

Proof. This is the second-order multivariate Taylor formula for the smooth map $(\boldsymbol{\theta}, \boldsymbol{v}) \mapsto \Psi(\boldsymbol{\theta}, \boldsymbol{v})$, with its third-order integral remainder. The stated uniform bound follows from the uniform third-derivative bound on compact subsets of the extension neighborhood. □

F.2 Preliminaries for the proof

Definition F.8. *Throughout the whole proof, write $\lambda_\eta := 1 - \beta_2 = c_\beta \eta^\alpha$ where $1 < \alpha < 2$ and $c_\beta > 0$ are constants. For $t \in [0, T]$, write*

$$N_\eta(t) := \left\lfloor \frac{t}{\eta^2} \right\rfloor.$$

We work on the slow time scale $t = k\eta^2$ and define the two increasing clocks

$$A_\eta(t) := \eta N_\eta(t), \qquad B_\eta(t) := \lambda_\eta N_\eta(t) = c_\beta \eta^\alpha N_\eta(t).$$

where clock A_η corresponds to the gradient drift in the $\boldsymbol{\theta}$-equation and B_η corresponds to the preconditioner drift in the $\boldsymbol{v}$-equation.

For every fixed $t > 0$,

$$A_\eta(t) \sim \frac{t}{\eta}, \qquad B_\eta(t) \sim c_\beta t\, \eta^{\alpha-2}.$$

Since $1 < \alpha < 2$,

$$A_\eta(t) \to \infty, \qquad B_\eta(t) \to \infty, \qquad \frac{B_\eta(t)}{A_\eta(t)} = \frac{\lambda_\eta}{\eta} = c_\beta \eta^{\alpha-1} \to 0.$$

Thus the $\boldsymbol{\theta}$-normal concentration is faster than the $\boldsymbol{v}$-tracking, which is in turn faster than the projected slow motion on Γ.

For any predictable sequence $\boldsymbol{f}_k$, we use the shorthand

$$\int_0^t \boldsymbol{f}_\eta(s-)\, \mathrm{d}A_\eta(s) := \sum_{k=0}^{N_\eta(t)-1} \eta \boldsymbol{f}_k, \qquad \int_0^t \boldsymbol{f}_\eta(s-)\, \mathrm{d}B_\eta(s) := \sum_{k=0}^{N_\eta(t)-1} \lambda_\eta \boldsymbol{f}_k.$$

The first moment correction is independent of the scaling of $1 - \beta_2$. Recall the definition of S_k, c_1, $\boldsymbol{Q}_{\eta,k}$, $e_{\eta,k+1}$, and $\widehat{\boldsymbol{\theta}}_k$ from Lemma E.4. In the present scheme, $\lambda_\eta = c_\beta \eta^\alpha$ and $N_\eta(T) \le T\eta^{-2}$. Thus (15) gives

$$\begin{aligned} &\sup_{k \le N_\eta(T)} \|\boldsymbol{Q}_{\eta,k}\| = \mathcal{O}_{\mathbb{P}}(\eta), \qquad \sup_{k < N_\eta(T)} \|e_{\eta,k+1}\| = \mathcal{O}_{\mathbb{P}}(\eta^{1+\alpha}), \\ &\sum_{k < N_\eta(T)} \|e_{\eta,k+1}\| = \mathcal{O}_{\mathbb{P}}(\eta^{\alpha-1}) = o_{\mathbb{P}}(1). \end{aligned} \tag{29}$$

Lemma F.7 (Burn-in localization and adaptive matching)**.** *Assume Assumption 2.1–2.4, fix $M, c_\beta > 0$ and $1 < \alpha < 2$. There exist constants $\epsilon_0 > 0$ and $C > 0$ such that the following holds. Suppose the initial state is independent of subsequent samples and satisfies*

$$\operatorname{dist}(\boldsymbol{\theta}_0, \Gamma) \le \epsilon_0, \qquad \boldsymbol{v}_0 \in [0, R_1]^D, \qquad \|\boldsymbol{m}_0\| \le M.$$

Let

$$K_0 := \left\lfloor C\eta^{-\alpha} \log(1/\eta) \right\rfloor, \qquad \boldsymbol{\zeta}_k := \Psi(\boldsymbol{\theta}_k, \boldsymbol{v}_k)$$

whenever the iterate lies in the working zone. Then, as $\eta \to 0$,

$$\mathbb{P}\left((\boldsymbol{\theta}_k, \boldsymbol{v}_k) \in \mathcal{X}^{\epsilon_1} \text{ for every } 0 \le k \le K_0\right) \longrightarrow 1, \tag{30}$$

$$\mathcal{L}(\boldsymbol{\theta}_{K_0}) - \mathcal{L}^* = \mathcal{O}_{\mathbb{P}}\big(\eta \log(1/\eta)\big), \tag{31}$$

$$\operatorname{dist}(\boldsymbol{\theta}_{K_0}, \Gamma) = \mathcal{O}_{\mathbb{P}}\big(\sqrt{\eta \log(1/\eta)}\big), \tag{32}$$

$$\boldsymbol{v}_{K_0} - U(\boldsymbol{\zeta}_{K_0}) = o_{\mathbb{P}}(1). \tag{33}$$

Moreover, $\beta_1^{K_0} \|\boldsymbol{m}_0\| = o(1)$.

Proof. Apply Corollary D.1 with $\varrho := \min\{\epsilon_1/2, r_\Phi/2\}$, and take its initial radius as ϵ_0. For its explicit $\mathcal{D}, \tau$ and all $\eta < \eta_*$, convexity of $[0, R_1]^D$ under the adaptive recursion gives

$$\boldsymbol{v}_k \in [0, R_1]^D \quad (k \ge 0), \qquad \{\tau > K_0\} \subseteq \left\{ \max_{k \le K_0} \operatorname{dist}(\boldsymbol{\theta}_k, \Gamma) < \varrho/2 < \epsilon_1 \right\}.$$

Consequently,

$$\mathbb{P}\left(\exists k \le K_0 : (\boldsymbol{\theta}_k, \boldsymbol{v}_k) \notin \mathcal{X}^{\epsilon_1}\right) \le \mathbb{P}(\tau \le K_0) \le (K_0 + 1) e^{-c_*/\eta} = o(1).$$

Let $k_{\mathrm{b}} = \lceil C_{\mathrm{b}} \eta^{-1} \log(1/\eta) \rceil$ be as in Corollary D.1. Since $1 < \alpha < 2$, $k_{\mathrm{b}} = o(K_0)$ and $K_0 = o(\eta^{-2})$. Equation (9) therefore gives

$$\sup_{k_{\mathrm{b}} \le k \le K_0} F(\boldsymbol{\theta}_k) = \mathcal{O}_{\mathbb{P}}(\eta \log(1/\eta)), \qquad \sup_{k_{\mathrm{b}} \le k \le K_0} \operatorname{dist}(\boldsymbol{\theta}_k, \Gamma) = \mathcal{O}_{\mathbb{P}}(\sqrt{\eta \log(1/\eta)}). \tag{34}$$

These estimates prove (30)–(32) directly from our convergence proof.

We next record the short-time tangential estimate needed for adaptive matching. Recall the Taylor expansion (27) and its remainder bound (28). On the localized event in (30), the adjacent increments

satisfy $\|\Delta\boldsymbol{\theta}_{k+1}\| \le C\eta$ and $\|\Delta\boldsymbol{v}_{k+1}\| \le C\lambda_\eta$. Hence, for all sufficiently small η, the line segments between adjacent states lie in the smooth extension neighborhood required by Lemma F.6.

For $k_{\mathrm{b}} \le k < K_0$, write

$$A_k := \partial_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta}_k, \boldsymbol{v}_k), \qquad B_k := \partial_{\boldsymbol{v}}\Psi(\boldsymbol{\theta}_k, \boldsymbol{v}_k), \qquad \boldsymbol{z}_k := \nabla\ell_k(\boldsymbol{\theta}_k) - \nabla\mathcal{L}(\boldsymbol{\theta}_k).$$

Use $S_k, \boldsymbol{Q}_{\eta,k}$, and $e_{\eta,k+1}$ from Lemma E.4. Substituting the exact momentum identity (13) into (27), and using

$$\partial_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta}, \boldsymbol{v})S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta}) = \mathbf{0},$$

eliminates the deterministic-gradient part of the first-order term and gives

$$\begin{aligned}\boldsymbol{\zeta}_{k+1} - \boldsymbol{\zeta}_k = -\eta A_k[S_k\boldsymbol{z}_k] + A_k[\boldsymbol{Q}_{\eta,k+1} - \boldsymbol{Q}_{\eta,k}] + A_k[e_{\eta,k+1}] \\ + B_k[\Delta\boldsymbol{v}_{k+1}] + \mathcal{T}^{\Psi}_{\eta,k+1},\end{aligned} \tag{35}$$

where $\mathcal{T}^{\Psi}_{\eta,k+1}$ is the sum of the three quadratic terms in (27) and $\mathcal{R}^{\Psi}_{\eta,k+1}$. On the localized event, bounded gradients give a uniform bound on $\boldsymbol{m}_k$, while Lipschitz continuity of S and $\|\Delta\boldsymbol{v}_{k+1}\| \le C\lambda_\eta$ give $\|S_{k+1} - S_k\| \le C\lambda_\eta$. Thus the range estimates in Lemma E.4 give

$$\|\boldsymbol{Q}_{\eta,k}\| \le C\eta, \qquad \|e_{\eta,k+1}\| \le C\eta\lambda_\eta.$$

Together with $\lambda_\eta = o(\eta)$ and the Taylor remainder bound, this implies

$$\|\mathcal{T}^{\Psi}_{\eta,k+1}\| \le C\eta^2.$$

It remains to sum the momentum-corrector term in (35). For $k_{\mathrm{b}} < n \le K_0$, discrete summation by parts gives

$$\begin{aligned}\sum_{k=k_{\mathrm{b}}}^{n-1} A_k[\boldsymbol{Q}_{\eta,k+1} - \boldsymbol{Q}_{\eta,k}] = A_{n-1}[\boldsymbol{Q}_{\eta,n}] - A_{k_{\mathrm{b}}}[\boldsymbol{Q}_{\eta,k_{\mathrm{b}}}] \\ - \sum_{k=k_{\mathrm{b}}+1}^{n-1} (A_k - A_{k-1})[\boldsymbol{Q}_{\eta,k}].\end{aligned}$$

The smoothness of $\partial_{\boldsymbol{\theta}}\Psi$ implies

$$\|A_k - A_{k-1}\| \le C\big(\|\Delta\boldsymbol{\theta}_k\| + \|\Delta\boldsymbol{v}_k\|\big) \le C\eta.$$

Thus the two boundary terms are $\mathcal{O}(\eta)$, while the last sum is $\mathcal{O}(K_0\eta^2)$. The remaining terms in (35) satisfy, uniformly in $k_{\mathrm{b}} \le n \le K_0$,

$$\begin{aligned}\sup_n \|\eta \sum_{k=k_{\mathrm{b}}}^{n-1} A_k[S_k\boldsymbol{z}_k]\| &= \mathcal{O}_{\mathbb{P}}(\eta\sqrt{K_0}), \\ \sum_{k=k_{\mathrm{b}}}^{K_0-1} \|A_k[e_{\eta,k+1}]\| &\le CK_0\eta\lambda_\eta, \\ \sum_{k=k_{\mathrm{b}}}^{K_0-1} \|B_k[\Delta\boldsymbol{v}_{k+1}]\| &= \mathcal{O}_{\mathbb{P}}\Big(K_0\lambda_\eta\sqrt{\eta\log(1/\eta)}\Big), \\ \sum_{k=k_{\mathrm{b}}}^{K_0-1} \|\mathcal{T}^{\Psi}_{\eta,k+1}\| &\le CK_0\eta^2.\end{aligned}$$

Here the first line follows from a maximal martingale inequality (applied to the process stopped at the first working-zone exit). For the second line, Lipschitz continuity of S gives $\|e_{\eta,k+1}\| \le C\eta\lambda_\eta$. For the third line, we used $\|B_k\| \le C\,\mathrm{dist}(\boldsymbol{\theta}_k, \Gamma)$ together with (34). Summing (35) therefore yields

$$\begin{aligned}\sup_{k_{\mathrm{b}} \le n \le K_0} \|\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_{k_{\mathrm{b}}}\| = \mathcal{O}_{\mathbb{P}}\big(\eta\sqrt{K_0} + K_0\eta^2 + \eta + K_0\eta\lambda_\eta \\ + K_0\lambda_\eta\sqrt{\eta\log(1/\eta)}\big) = o_{\mathbb{P}}(1).\end{aligned} \tag{36}$$

The five terms are respectively the projected gradient-noise martingale; the quadratic and cubic Taylor terms together with the interior coefficient variation in the summation-by-parts formula; the boundary momentum corrector; the preconditioner coefficient-lag error; and the first-order $\boldsymbol{v}$-term. The required limits follow from

$$\eta\sqrt{K_0} = \mathcal{O}\big(\eta^{1-\alpha/2}\sqrt{\log(1/\eta)}\big), \quad K_0\eta^2 = \mathcal{O}\big(\eta^{2-\alpha}\log(1/\eta)\big), \quad K_0\eta\lambda_\eta = \mathcal{O}\big(\eta\log(1/\eta)\big),$$

and $K_0\lambda_\eta\sqrt{\eta\log(1/\eta)} = \mathcal{O}(\sqrt{\eta}\,\log(1/\eta)^{3/2})$.

It remains to prove (33). Variation of constants in the $\boldsymbol{v}$-recursion gives the exact identity

$$\boldsymbol{v}_{K_0} - U(\boldsymbol{\zeta}_{K_0}) = \beta_2^{K_0}\big(\boldsymbol{v}_0 - U(\boldsymbol{\zeta}_{K_0})\big) \tag{37}$$

$$+ \lambda_\eta \sum_{j=0}^{K_0-1} \beta_2^{K_0-1-j}\left\{V(\boldsymbol{g}_j\boldsymbol{g}_j^\top) - \bar{U}(\boldsymbol{\theta}_j)\right\} \tag{38}$$

$$+ \lambda_\eta \sum_{j=0}^{K_0-1} \beta_2^{K_0-1-j}\left\{\bar{U}(\boldsymbol{\theta}_j) - U(\boldsymbol{\zeta}_{K_0})\right\}. \tag{39}$$

The first term is $o(1)$ because $\lambda_\eta K_0 \asymp \log(1/\eta)$. The second term is a martingale transform whose second moment is bounded by

$$C\lambda_\eta^2 \sum_{r\geq 0}\beta_2^{2r} \leq C\lambda_\eta,$$

so it is $\mathcal{O}_{\mathbb{P}}(\eta^{\alpha/2})$. For the last term, the indices $j < k_{\mathrm{b}}$ have total weight at most

$$\lambda_\eta k_{\mathrm{b}} = \mathcal{O}\big(\eta^{\alpha-1}\log(1/\eta)\big) = o(1).$$

For $j \geq k_{\mathrm{b}}$, smoothness of $\bar{U}$ and U, together with (34) and (36), gives

$$\sup_{k_{\mathrm{b}}\leq j\leq K_0} \|\bar{U}(\boldsymbol{\theta}_j) - U(\boldsymbol{\zeta}_{K_0})\| = o_{\mathbb{P}}(1).$$

Since $\lambda_\eta\sum_{r\geq 0}\beta_2^r = 1$, the last term in (39) is $o_{\mathbb{P}}(1)$. This proves the matching assertion. The final claim follows immediately from $\beta_1 \leq \bar{\beta}_1 < 1$ and $K_0 \to \infty$. □

Define $(\boldsymbol{\theta}'_k, \boldsymbol{m}'_k, \boldsymbol{v}'_k) := (\boldsymbol{\theta}_{K_0+k}, \boldsymbol{m}_{K_0+k}, \boldsymbol{v}_{K_0+k})$, $\mathcal{F}'_k := \mathcal{F}_{K_0+k}$, and $\xi'_k := \xi_{K_0+k}$. The future samples ξ'_k are independent of $\mathcal{F}'_k$, and these variables satisfy the same AGM recursions. With $\boldsymbol{\zeta}'_0 := \boldsymbol{\zeta}_{K_0}$, the burn-in lemma gives

$$F(\boldsymbol{\theta}'_0) = o_{\mathbb{P}}(1), \qquad \boldsymbol{v}'_0 - U(\boldsymbol{\zeta}'_0) = o_{\mathbb{P}}(1), \qquad \|\boldsymbol{m}'_0\| \leq M + R, \qquad \mathbb{P}((\boldsymbol{\theta}'_0, \boldsymbol{v}'_0) \in \mathcal{X}^{\epsilon_1}) \to 1.$$

For all subsequent estimates, omit the primes and define τ_η by Definition F.7 using these restarted variables.

F.3 Semimartingale decomposition

With the clocks and momentum corrector from We derive the stopped semimartingale decompositions (40) and (41) from the discrete AGM updates. We use the right-continuous step interpolations $\boldsymbol{\theta}_\eta(t) := \boldsymbol{\theta}_{N_\eta(t)}$ and $\boldsymbol{v}_\eta(t) := \boldsymbol{v}_{N_\eta(t)}$. At the update time $(k+1)\eta^2$, their left limits are $\boldsymbol{\theta}_k$ and $\boldsymbol{v}_k$, while the clocks jump by η and λ_η, respectively. Thus the integral-as-sum convention in Section F.2 reproduces the discrete drift sums exactly. We first collect the momentum-corrector increment and the coefficient-lag error into the remainder.

Corollary F.1 (Cumulative momentum-remainder estimates)**.** *With $\boldsymbol{Q}_{\eta,k}$ and $e_{\eta,k+1}$ defined in Lemma E.4, define*

$$\Delta R^\theta_{\eta,k+1} := \boldsymbol{Q}_{\eta,k+1} - \boldsymbol{Q}_{\eta,k} + e_{\eta,k+1}.$$

Then

$$\sup_{n\leq N^\tau_\eta}\left\|\sum_{k=0}^{n-1}\Delta R^\theta_{\eta,k+1}\right\| = \mathcal{O}_{\mathbb{P}}(\eta + \eta^{\alpha-1}) = o_{\mathbb{P}}(1).$$

For every fixed $\gamma > 0$, the exponentially weighted remainder also satisfies

$$\sup_{n\leq N^\tau_\eta}\left\|\sum_{k=0}^{n-1} e^{-\gamma\eta(n-1-k)}\Delta R^\theta_{\eta,k+1}\right\| = o_{\mathbb{P}}(1).$$

Proof. Up to the working-zone exit, the hypotheses of Lemma E.4 hold and $N_\eta^\tau \le N_\eta(T)$. Its range estimate therefore gives the first equation.

For the weighted estimate, apply Lemma E.6 with $\rho = e^{-\gamma\eta}$ and $\boldsymbol{a}_k = \Delta R^\theta_{\eta,k+1}$. The first estimate then gives

$$\sup_{n\le N_\eta^\tau} \left\| \sum_{k=0}^{n-1} e^{-\gamma\eta(n-1-k)} \Delta R^\theta_{\eta,k+1} \right\| = \mathcal{O}_{\mathbb{P}}(\eta + \eta^{\alpha-1}) = o_{\mathbb{P}}(1).$$

□

Consequently, at the level of the cumulative semimartingale decomposition we may write

$$\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -\eta S(\boldsymbol{v}_k)\boldsymbol{g}_k + \Delta R^\theta_{\eta,k+1},$$

where the partial sums of the remainder converge to zero uniformly up to N_η^τ. The corrector increments $\boldsymbol{Q}_{\eta,k+1} - \boldsymbol{Q}_{\eta,k}$ need not have vanishing total variation. Therefore the recursive formula must not be inserted term by term into a nonlinear second-order Taylor expansion. For a fixed first-moment parameter, the linear corrector and the quadratic Taylor terms have to be reduced jointly, as in the finite-memory moment calculation for AGMs.

Define the $\boldsymbol{\theta}$-martingale increment

$$\Delta M^\theta_{\eta,k+1} := -\eta S(\boldsymbol{v}_k)\boldsymbol{z}_k(\boldsymbol{\theta}_k),$$

then $\mathbb{E}_k[\Delta M^\theta_{\eta,k+1}] = \mathbf{0}$. Let

$$M^\theta_\eta(t) := \sum_{k=0}^{N_\eta(t\wedge\tau_\eta)-1} \Delta M^\theta_{\eta,k+1}, \quad R^\theta_\eta(t) := \sum_{k=0}^{N_\eta(t\wedge\tau_\eta)-1} \Delta R^\theta_{\eta,k+1}.$$

Thus M^θ_η is the stopped martingale. The stopped $\boldsymbol{\theta}$-equation admits the decomposition

$$\boldsymbol{\theta}_\eta(t\wedge\tau_\eta) = \boldsymbol{\theta}_\eta(0) - \int_0^{t\wedge\tau_\eta} S(\boldsymbol{v}_\eta(s-))\nabla\mathcal{L}(\boldsymbol{\theta}_\eta(s-))\,\mathrm{d}A_\eta(s) + M^\theta_\eta(t) + R^\theta_\eta(t). \tag{40}$$

Under the step interpolation convention there is no additional interpolation error. By corollary F.1,

$$\sup_{0\le t\le T} \|R^\theta_\eta(t)\| = o_{\mathbb{P}}(1).$$

The predictable quadratic variation of M^θ_η is

$$\langle M^\theta_\eta\rangle_t = \sum_{k=0}^{N_\eta(t\wedge\tau_\eta)-1} \eta^2 S(\boldsymbol{v}_k)\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)S(\boldsymbol{v}_k) = \int_0^{t\wedge\tau_\eta} S(\boldsymbol{v}_\eta(s-))\boldsymbol{\Sigma}(\boldsymbol{\theta}_\eta(s-))S(\boldsymbol{v}_\eta(s-))\,\mathrm{d}s + o_{\mathbb{P}}(1).$$

In particular, $\langle M^\theta_\eta\rangle_T = \mathcal{O}_{\mathbb{P}}(1)$. Thus the $\boldsymbol{\theta}$-martingale survives in the slow limit and produces the Brownian term in the final SDE.

Next consider the $\boldsymbol{v}$-equation. Define

$$\Delta M^v_{\eta,k+1} := \lambda_\eta \left(V(\boldsymbol{g}_k\boldsymbol{g}_k^\top) - \bar{U}(\boldsymbol{\theta}_k) \right).$$

Then $\mathbb{E}_k[\Delta M^v_{\eta,k+1}] = \mathbf{0}$. Let

$$M^v_\eta(t) := \sum_{k=0}^{N_\eta(t\wedge\tau_\eta)-1} \Delta M^v_{\eta,k+1}.$$

Since

$$\boldsymbol{v}_{k+1} - \boldsymbol{v}_k = \lambda_\eta \left(\bar{U}(\boldsymbol{\theta}_k) - \boldsymbol{v}_k \right) + \Delta M^v_{\eta,k+1},$$

the stopped $\boldsymbol{v}$-equation is

$$\boldsymbol{v}_\eta(t\wedge\tau_\eta) = \boldsymbol{v}_\eta(0) + \int_0^{t\wedge\tau_\eta} \Big(\bar{U}(\boldsymbol{\theta}_\eta(s-)) - \boldsymbol{v}_\eta(s-) \Big)\,\mathrm{d}B_\eta(s) + M^v_\eta(t). \tag{41}$$

Using the boundedness of $V(\boldsymbol{g}\boldsymbol{g}^\top)$ in the working zone, the quadratic variation of M_η^v satisfies

$$\mathbb{E}\left[[M_\eta^v]_T\right] \le C\left\lfloor \frac{T}{\eta^2} \right\rfloor \lambda_\eta^2 = \mathcal{O}(\eta^{2\alpha-2}).$$

Since $\alpha > 1$, Doob's inequality gives

$$\sup_{0\le t\le T} \|M_\eta^v(t)\| = o_{\mathbb{P}}(1). \tag{42}$$

Thus $\boldsymbol{v}_\eta$ has no Brownian fluctuation in the limiting slow dynamics. It is a deterministic fast-tracking variable.

Combining (40) and (41), the stopped $1 < \alpha < 2$ scheme has the two-clock form

$$\mathrm{d}\boldsymbol{\theta}_\eta = -S(\boldsymbol{v}_\eta)\nabla\mathcal{L}(\boldsymbol{\theta}_\eta)\,\mathrm{d}A_\eta + \mathrm{d}M_\eta^\theta + \mathrm{d}R_\eta^\theta, \tag{43}$$

$$\mathrm{d}\boldsymbol{v}_\eta = \Big(\bar{U}(\boldsymbol{\theta}_\eta) - \boldsymbol{v}_\eta\Big)\,\mathrm{d}B_\eta + \mathrm{d}M_\eta^v. \tag{44}$$

where all processes in this display are evaluated at $t \wedge \tau_\eta$.

The rest of our proof uses (43) to prove the collapse of $\boldsymbol{\theta}_\eta$ to Γ, and then uses (44) to prove the fast tracking

$$\boldsymbol{v}_\eta(t) \approx U(\boldsymbol{\zeta}_\eta(t)).$$

F.4 Collapse to the minimizer manifold

The first step is to show that the large gradient drift in the $\boldsymbol{\theta}$-equation forces the parameter to stay close to Γ. This is the analogue of the large-drift collapse in Katzenberger's theorem, with the A_η-clock playing the role of the large drift clock.

Let $h(\boldsymbol{\theta}) := \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^*$. On Γ^{ϵ_1}, the local manifold assumptions and the PL condition imply that

$$h(\boldsymbol{\theta}) \asymp \operatorname{dist}(\boldsymbol{\theta},\Gamma)^2, \qquad \|\nabla\mathcal{L}(\boldsymbol{\theta})\| \asymp \operatorname{dist}(\boldsymbol{\theta},\Gamma).$$

More explicitly, there exist constants $0 < c_h < C_h < \infty$ such that

$$c_h \operatorname{dist}(\boldsymbol{\theta},\Gamma)^2 \le h(\boldsymbol{\theta}) \le C_h \operatorname{dist}(\boldsymbol{\theta},\Gamma)^2, \qquad \forall \boldsymbol{\theta} \in \Gamma^{\epsilon_1}.$$

Lemma F.8 (A_η-clock Lyapunov estimate)**.** *For all $(\boldsymbol{\theta},\boldsymbol{v}) \in \mathcal{X}^{\epsilon_1}$,*

$$\nabla h(\boldsymbol{\theta})^\top\big(-S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta})\big) \le -\frac{2\mu}{R_0} h(\boldsymbol{\theta}).$$

Proof. Since $h = \mathcal{L} - \mathcal{L}^*$, we have $\nabla h(\boldsymbol{\theta}) = \nabla\mathcal{L}(\boldsymbol{\theta})$. Using $S(\boldsymbol{v}) \succeq \frac{1}{R_0}\boldsymbol{I}$, we get

$$\nabla h(\boldsymbol{\theta})^\top\big(-S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta})\big) = -\nabla\mathcal{L}(\boldsymbol{\theta})^\top S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta}) \le -\frac{1}{R_0}\|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2.$$

The local PL inequality in Lemma F.3 gives $\|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2 \ge 2\mu h(\boldsymbol{\theta})$. Combining the two inequalities completes the proof. □

Now we introduce the Lyapunov function $h(\boldsymbol{\theta}_\eta(t))$ along the A_η-clock. The next lemma shows that if $h(\boldsymbol{\theta}_\eta(0))$ is small, then $h(\boldsymbol{\theta}_\eta(t))$ remains small for all $t \le \tau_\eta$, and moreover, the integral of the distance to Γ squared with respect to the A_η-clock is bounded in probability.

Lemma F.9 (A_η-clock collapse)**.** *Use the restarted initialization of Section F.2, for which $h(\boldsymbol{\theta}_\eta(0)) = o_{\mathbb{P}}(1)$. Then*

$$\sup_{0\le t\le\tau_\eta} \operatorname{dist}\big(\boldsymbol{\theta}_\eta(t),\Gamma\big) = o_{\mathbb{P}}(1). \tag{45}$$

Moreover,

$$\int_0^{\tau_\eta} \operatorname{dist}\big(\boldsymbol{\theta}_\eta(s-),\Gamma\big)^2\,\mathrm{d}A_\eta(s) = \mathcal{O}_{\mathbb{P}}(1).$$

Equivalently, by the distance equivalence in Lemma F.3,

$$\int_0^{\tau_\eta} \|\boldsymbol{r}_\eta(s-)\|^2\,\mathrm{d}A_\eta(s) = \mathcal{O}_{\mathbb{P}}(1). \tag{46}$$

Proof. We use the restarted initialization from Section F.2. In particular, for a fixed $M_1 := M + R$,

$$\|\boldsymbol{m}_0\| \le M_1, \qquad \boldsymbol{v}_0 \in [0, R_1]^D, \qquad \mathbb{P}(\boldsymbol{\theta}_0 \in \Gamma^{\epsilon_1}) \to 1, \qquad F(\boldsymbol{\theta}_0) = h(\boldsymbol{\theta}_0) = o_{\mathbb{P}}(1).$$

Fix $\varepsilon > 0$ and choose $0 < \varrho < \min\{r_\Phi, \epsilon_1, 2\varepsilon\}$. Let $\epsilon_{\mathrm{loc}} > 0$, $\mathcal{D}$, and τ be supplied by Corollary D.1 for M_1, ϱ, and set $N := \lfloor T\eta^{-2} \rfloor$ and $E_\eta := \{\mathrm{dist}(\boldsymbol{\theta}_0, \Gamma) \le \epsilon_{\mathrm{loc}}\} \in \mathcal{F}_0$. The local quadratic growth bound gives

$$\mathbb{P}(E_\eta^c) \le \mathbb{P}(\boldsymbol{\theta}_0 \notin \Gamma^{\epsilon_1}) + \mathbb{P}(F(\boldsymbol{\theta}_0) > c_h \epsilon_{\mathrm{loc}}^2) = o(1).$$

Conditioning on E_η preserves independence of the future samples. The localization lemma and assertion (2) of the common corollary yield

$$\mathbb{P}(\tau \le N \mid E_\eta) \le (N+1)e^{-c_*/\eta}, \qquad \max_{k \le N} \mathrm{dist}(\boldsymbol{\theta}_k, \Gamma) < \varrho/2 < \varepsilon \quad \text{on } E_\eta \cap \{\tau > N\}.$$

Consequently,

$$\mathbb{P}\left(\sup_{t \le \tau_\eta} \mathrm{dist}(\boldsymbol{\theta}_\eta(t), \Gamma) > \varepsilon \right) \le \mathbb{P}(E_\eta^c) + (N+1)e^{-c_*/\eta} \longrightarrow 0.$$

This proves (45), including the exit state.

For the energy estimate, fix one $\varrho \in (0, \min\{r_\Phi, \epsilon_1\})$ and use the corresponding $E_\eta, \mathcal{D}, \tau$. Write $\mathbb{E}_E[\cdot] := \mathbb{E}[\cdot \mid E_\eta]$ and $\sigma := N \wedge \tau$. By Lemma D.6, $F(\boldsymbol{y}_0) \le a/2$ on E_η and $F(\boldsymbol{y}_\sigma) \ge 0$. The conditional descent inequality in Lemma D.4, with $\lambda_\eta^2 \le c_\beta^2 \eta$ for $\eta \le 1$, implies

$$\mathbb{E}[F(\boldsymbol{y}_{(k+1)\wedge\sigma}) - F(\boldsymbol{y}_{k\wedge\sigma}) \mid \mathcal{F}_k] \le \mathbf{1}_{\{k<\sigma\}}\big(-c\eta F(\boldsymbol{y}_k) + C\eta^2\big).$$

Summing conditional expectations and using $N\eta^2 \le T$ gives

$$c\,\mathbb{E}_E \sum_{k<\sigma} \eta F(\boldsymbol{y}_k) \le \mathbb{E}_E[F(\boldsymbol{y}_0) - F(\boldsymbol{y}_\sigma)] + CN\eta^2 \le a/2 + CT.$$

The geometry and correction lemmas give, for $k < \tau$,

$$\mathrm{dist}(\boldsymbol{\theta}_k, \Gamma)^2 \le 2\,\mathrm{dist}(\boldsymbol{y}_k, \Gamma)^2 + 2C_y^2\eta^2 \le (2/a_1)F(\boldsymbol{y}_k) + 2C_y^2\eta^2.$$

Thus, for a constant C_T independent of η,

$$\mathbb{E}_E \sum_{k<\sigma} \eta\,\mathrm{dist}(\boldsymbol{\theta}_k, \Gamma)^2 \le \frac{2(a/2 + CT)}{a_1 c} + 2C_y^2 N\eta^3 \le C_T.$$

Set $I_\eta := \int_0^{\tau_\eta} \mathrm{dist}(\boldsymbol{\theta}_\eta(s-), \Gamma)^2\,\mathrm{d}A_\eta(s)$. On $E_\eta \cap \{\tau > N\}$, the integral-as-sum convention and $N_\eta(\tau_\eta) \le N = \sigma$ imply

$$I_\eta = \sum_{k < N_\eta(\tau_\eta)} \eta\,\mathrm{dist}(\boldsymbol{\theta}_k, \Gamma)^2 \le \sum_{k<\sigma} \eta\,\mathrm{dist}(\boldsymbol{\theta}_k, \Gamma)^2.$$

For every $L > 0$, Markov's inequality therefore gives

$$\mathbb{P}(I_\eta > L) \le \mathbb{P}(E_\eta^c) + (N+1)e^{-c_*/\eta} + C_T/L, \qquad \lim_{L\to\infty} \limsup_{\eta \downarrow 0} \mathbb{P}(I_\eta > L) = 0.$$

Hence $I_\eta = \mathcal{O}_{\mathbb{P}}(1)$. Finally, (25) holds before the working-zone exit, so

$$\int_0^{\tau_\eta} \|\boldsymbol{r}_\eta(s-)\|^2\,\mathrm{d}A_\eta(s) \le C_r^2 I_\eta = \mathcal{O}_{\mathbb{P}}(1),$$

which proves (46). □

The normal-energy estimate in Lemma F.9 has an immediate consequence on the B_η-clock. Indeed, recall that

$$\mathrm{d}B_\eta = \frac{\lambda_\eta}{\eta}\,\mathrm{d}A_\eta = c_\beta \eta^{\alpha-1}\,\mathrm{d}A_\eta.$$

Since $\alpha > 1$, the A_η-energy bound therefore yields the following stronger smallness under $\mathrm{d}B_\eta$.

Corollary F.2 (B_η-smallness of the normal energy)**.** *Under the assumptions of Lemma F.9,*

$$\int_0^{\tau_\eta} \operatorname{dist}\big(\boldsymbol{\theta}_\eta(s-), \Gamma\big)^2 \, \mathrm{d}B_\eta(s) = o_{\mathbb{P}}(1).$$

Equivalently,

$$\int_0^{\tau_\eta} \|\boldsymbol{r}_\eta(s-)\|^2 \, \mathrm{d}B_\eta(s) = o_{\mathbb{P}}(1). \tag{47}$$

We shall also use the following consequence of Corollary F.2. Since $\|\nabla \mathcal{L}(\boldsymbol{\theta})\| \asymp \operatorname{dist}(\boldsymbol{\theta}, \Gamma)$, we have

$$\int_0^{\tau_\eta} \|\nabla \mathcal{L}(\boldsymbol{\theta}_\eta(s-))\|^2 \, \mathrm{d}B_\eta(s) = o_{\mathbb{P}}(1).$$

The uniform collapse in Lemma F.9 also shows that the working-zone localization is asymptotically inactive.

Corollary F.3 (Vanishing exit probability)**.** *Under the assumptions of Lemmas F.2 and F.9, for every fixed $T > 0$,*

$$\mathbb{P}(\tau_\eta < T) = o(1).$$

Proof. Assume that $\tau_\eta < T$. By Lemma F.2, the adaptive coordinate remains in $[0, R_1]^D$. Hence the exit must occur through the $\boldsymbol{\theta}$-coordinate. By Lemma F.9, we have

$$\sup_{0 \le t \le \tau_\eta} \operatorname{dist}\big(\boldsymbol{\theta}_\eta(t), \Gamma\big) = o_{\mathbb{P}}(1).$$

Therefore, for any fixed $\epsilon_1 > 0$, we have

$$\{\tau_\eta < T\} \subseteq \left\{ \sup_{0 \le t \le \tau_\eta} \operatorname{dist}\big(\boldsymbol{\theta}_\eta(t), \Gamma\big) \ge \epsilon_1 \right\}.$$

Taking probabilities and using the preceding convergence, we get

$$\mathbb{P}(\tau_\eta < T) \le \mathbb{P}\left(\sup_{0 \le t \le \tau_\eta} \operatorname{dist}\big(\boldsymbol{\theta}_\eta(t), \Gamma\big) \ge \epsilon_1 \right) = o(1),$$

which proves the claim. □

Thus, on every fixed slow-time horizon, any convergence statement proved for the stopped processes transfers to the original processes with probability tending to one.

In the next section, we will strengthen these quadratic estimates to first-order normal-space averaging under $\mathrm{d}B_\eta$.

F.5 Normal-space averaging

The key estimate for the $1 < \alpha < 2$ scheme is a normal-space averaging result under the B_η-clock. Since the $\boldsymbol{\theta}$-normal dynamics evolves on the faster A_η-clock, while $B_\eta \ll A_\eta$, the first-order normal displacement averages out when integrated against $\mathrm{d}B_\eta$.

A minor subtlety is that, for AGMs, the stable direction is not necessarily the Euclidean normal space. It is twisted by the preconditioner $S(\boldsymbol{v})$. We therefore formulate the argument using stable fibers of the $S(\boldsymbol{v})$-preconditioned gradient flow.

For fixed $\boldsymbol{\zeta} \in \Gamma$ and $\boldsymbol{v} \in [0, R_1]^D$, define the local stable fiber

$$\mathcal{W}^{\mathrm{s}}_{\boldsymbol{\zeta}, \boldsymbol{v}} := \left\{ \boldsymbol{\theta} \in \Gamma^{\epsilon_1} : \Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta}) = \boldsymbol{\zeta} \right\}.$$

Its tangent space at $\boldsymbol{\zeta}$ is denoted by

$$E^{\mathrm{s}}_{\boldsymbol{\zeta}, \boldsymbol{v}} := T_{\boldsymbol{\zeta}} \mathcal{W}^{\mathrm{s}}_{\boldsymbol{\zeta}, \boldsymbol{v}}.$$

Since $\Phi_{S(\boldsymbol{v})}(\boldsymbol{\zeta}) = \boldsymbol{\zeta}$, we have $E^{\mathrm{s}}_{\boldsymbol{\zeta}, \boldsymbol{v}} = \ker \partial_{\boldsymbol{\theta}} \Phi_{S(\boldsymbol{v})}(\boldsymbol{\zeta})$. Moreover, by the projection identities, we have

$$\partial_{\boldsymbol{\theta}} \Phi_{S(\boldsymbol{v})}(\boldsymbol{\zeta}) \boldsymbol{u} = \boldsymbol{u}, \qquad \boldsymbol{u} \in T_{\boldsymbol{\zeta}} \Gamma,$$

and

$$\partial_{\boldsymbol{\theta}}\Phi_{S(\boldsymbol{v})}(\boldsymbol{\zeta})S(\boldsymbol{v})\boldsymbol{u}=\mathbf{0},\qquad \boldsymbol{u}\in N_{\boldsymbol{\zeta}}\Gamma.$$

Therefore $E^{\mathrm{s}}_{\boldsymbol{\zeta},\boldsymbol{v}}=S(\boldsymbol{v})N_{\boldsymbol{\zeta}}\Gamma$. This is the stable fiber direction of the $S(\boldsymbol{v})$-preconditioned flow.

Let $\boldsymbol{H}_{\boldsymbol{\zeta}}:=\nabla^2\mathcal{L}(\boldsymbol{\zeta})$. Fix $\boldsymbol{v}$ and consider the deterministic preconditioned gradient flow $\dot{\boldsymbol{\theta}}=-S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta})$. Writing $\boldsymbol{\theta}=\boldsymbol{\zeta}+\boldsymbol{y}$ with $\boldsymbol{y}\in E^{\mathrm{s}}_{\boldsymbol{\zeta},\boldsymbol{v}}$, its first-order transverse dynamics at $\boldsymbol{y}=\mathbf{0}$ is

$$\dot{\boldsymbol{y}}=-S(\boldsymbol{v})\boldsymbol{H}_{\boldsymbol{\zeta}}\boldsymbol{y}.$$

Restricted to $E^{\mathrm{s}}_{\boldsymbol{\zeta},\boldsymbol{v}}$, the operator $\boldsymbol{A}_{\boldsymbol{\zeta},\boldsymbol{v}}:=S(\boldsymbol{v})\boldsymbol{H}_{\boldsymbol{\zeta}}$ has spectrum with real parts uniformly bounded away from zero. Equivalently, after possibly shrinking ϵ_1, there exist a constant $C_A>0$ such that

$$\|\boldsymbol{A}^{-1}_{\boldsymbol{\zeta},\boldsymbol{v}}\|_{\mathrm{op}}\le C_A,\qquad \boldsymbol{\zeta}\in\Gamma,\quad \boldsymbol{v}\in[0,R_1]^D,$$

where $\boldsymbol{A}^{-1}_{\boldsymbol{\zeta},\boldsymbol{v}}$ is understood as the inverse on the stable fiber $E^{\mathrm{s}}_{\boldsymbol{\zeta},\boldsymbol{v}}$.

Let $\boldsymbol{\theta}\in\Gamma^{\epsilon_1}$, $\boldsymbol{v}\in[0,R_1]^D$ and $\boldsymbol{\zeta}=\Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta})\in\Gamma$. Since $\boldsymbol{\theta}\in\mathcal{W}^{\mathrm{s}}_{\boldsymbol{\zeta},\boldsymbol{v}}$, there exists a unique stable coordinate

$$\boldsymbol{y}=\boldsymbol{y}(\boldsymbol{\theta},\boldsymbol{v})\in E^{\mathrm{s}}_{\boldsymbol{\zeta},\boldsymbol{v}}$$

such that $\boldsymbol{\theta}=\Theta(\boldsymbol{\zeta},\boldsymbol{v},\boldsymbol{y})$, where Θ is a smooth local diffeomorphism satisfying

$$\Theta(\boldsymbol{\zeta},\boldsymbol{v},\mathbf{0})=\boldsymbol{\zeta},\qquad \partial_{\boldsymbol{y}}\Theta(\boldsymbol{\zeta},\boldsymbol{v},\mathbf{0})=\boldsymbol{I}\quad\text{on } E^{\mathrm{s}}_{\boldsymbol{\zeta},\boldsymbol{v}}.$$

Θ can be chosen such that every fixed finite-order derivative is uniformly bounded on the working zone. Consequently, if

$$\boldsymbol{r}(\boldsymbol{\theta},\boldsymbol{v}):=\boldsymbol{\theta}-\Phi_{S(\boldsymbol{v})}(\boldsymbol{\theta}),$$

then

$$\boldsymbol{r}(\boldsymbol{\theta},\boldsymbol{v})=\boldsymbol{y}(\boldsymbol{\theta},\boldsymbol{v})+\mathcal{O}(\|\boldsymbol{y}(\boldsymbol{\theta},\boldsymbol{v})\|^2),\tag{48}$$

and, on the working zone,

$$\|\boldsymbol{y}(\boldsymbol{\theta},\boldsymbol{v})\|\asymp\|\boldsymbol{r}(\boldsymbol{\theta},\boldsymbol{v})\|\asymp\mathrm{dist}(\boldsymbol{\theta},\Gamma).\tag{49}$$

For a fixed first-moment parameter, we first pass to the corrected coordinate introduced by Lemma E.4. Define the original stable coordinate by $\boldsymbol{y}_k:=\boldsymbol{y}(\boldsymbol{\theta}_k,\boldsymbol{v}_k)$. Recall that $\boldsymbol{Q}_{\eta,k}=\eta c_1 S_k\boldsymbol{m}_k$ is the linear corrector. We introduce

$$\widehat{\boldsymbol{\theta}}_k:=\boldsymbol{\theta}_k-\boldsymbol{Q}_{\eta,k},\qquad \widehat{\boldsymbol{\zeta}}_k:=\Phi_{S(\boldsymbol{v}_k)}(\widehat{\boldsymbol{\theta}}_k),\qquad \widehat{\boldsymbol{r}}_k:=\widehat{\boldsymbol{\theta}}_k-\widehat{\boldsymbol{\zeta}}_k,$$

and let

$$\widehat{\boldsymbol{y}}_k:=\boldsymbol{y}(\widehat{\boldsymbol{\theta}}_k,\boldsymbol{v}_k)\in E^{\mathrm{s}}_{\widehat{\boldsymbol{\zeta}}_k,\boldsymbol{v}_k}.$$

Since $\boldsymbol{Q}_{\eta,k}=\mathcal{O}_{\mathbb{P}}(\eta)$ uniformly and all coordinate maps have bounded derivatives,

$$\sup_{k\le N_\eta(\tau_\eta)}\Big(\|\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k\|+\|\widehat{\boldsymbol{\zeta}}_k-\boldsymbol{\zeta}_k\|+\|\widehat{\boldsymbol{r}}_k-\boldsymbol{r}_k\|+\|\widehat{\boldsymbol{y}}_k-\boldsymbol{y}_k\|\Big)=\mathcal{O}_{\mathbb{P}}(\eta).\tag{50}$$

In particular, for every uniformly bounded test matrix $\boldsymbol{Q}_k$,

$$\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=0}^{n-1}\lambda_\eta\boldsymbol{Q}_k(\widehat{\boldsymbol{y}}_k-\boldsymbol{y}_k)\right\|=\mathcal{O}_{\mathbb{P}}\left(\frac{\lambda_\eta}{\eta}\right)=o_{\mathbb{P}}(1).\tag{51}$$

It therefore suffices to prove normal-space averaging in the corrected coordinates. For readability, throughout the remainder of this section we drop the hats and use $(\boldsymbol{\theta}_k,\boldsymbol{\zeta}_k,\boldsymbol{r}_k,\boldsymbol{y}_k)$ instead of $(\widehat{\boldsymbol{\theta}}_k,\widehat{\boldsymbol{\zeta}}_k,\widehat{\boldsymbol{r}}_k,\widehat{\boldsymbol{y}}_k)$. The corresponding original AGM coordinates are denoted by $(\boldsymbol{\theta}^\circ_k,\boldsymbol{\zeta}^\circ_k,\boldsymbol{r}^\circ_k,\boldsymbol{y}^\circ_k)$ throughout this section. At the end of the argument, (50)-(51) transfer the conclusions back to the original coordinates. The corresponding step interpolation is denoted by

$$\boldsymbol{y}_\eta(t):=\boldsymbol{y}_{\lfloor t/\eta^2\rfloor}.$$

The equivalence (49), the estimates (46) and (47), and the comparison (50) imply

$$\int_0^{\tau_\eta}\|\boldsymbol{y}_\eta(s-)\|^2\,\mathrm{d}B_\eta(s)=o_{\mathbb{P}}(1),\tag{52}$$

and

$$\int_0^{\tau_\eta} \|\boldsymbol{y}_\eta(s-)\|^2 \,\mathrm{d}A_\eta(s) = \mathcal{O}_{\mathbb{P}}(1). \tag{53}$$

Indeed, the additional squared contribution caused by passing from $\boldsymbol{y}^\circ$ is bounded by $C\eta^2 B_\eta(T) = \mathcal{O}(\eta^\alpha)$ and $C\eta^2 A_\eta(T) = \mathcal{O}(\eta)$, respectively, and the cross terms follow from Cauchy-Schwarz. We now record the one-step stable-coordinate expansion.

Lemma F.10 (Stable-coordinate expansion)**.** *For $k < N_\eta(\tau_\eta)$, there exist uniformly bounded predictable linear maps $\boldsymbol{A}_k, \boldsymbol{C}_k, \boldsymbol{K}_k$, with*

$$\boldsymbol{A}_k = S(\boldsymbol{v}_k)\nabla^2\mathcal{L}(\boldsymbol{\zeta}_k) \quad \text{on } E^{\mathrm{s}}_{\boldsymbol{\zeta}_k,\boldsymbol{v}_k},$$

such that

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \eta\boldsymbol{A}_k\boldsymbol{y}_k + \eta\boldsymbol{C}_k\boldsymbol{z}_k + \lambda_\eta\boldsymbol{K}_k\boldsymbol{y}_k + \nu^v_{\eta,k+1} + \chi_{\eta,k+1}.$$

Here $\boldsymbol{A}_k^{-1}$ is uniformly bounded on the stable fiber,

$$\mathbb{E}_k[\boldsymbol{C}_k\boldsymbol{z}_k] = 0, \qquad \mathbb{E}_k[\nu^v_{\eta,k+1}] = 0, \qquad \|\nu^v_{\eta,k+1}\| \le C\lambda_\eta\|\boldsymbol{y}_k\|,$$

and the finite-variation remainder satisfies

$$\frac{\lambda_\eta}{\eta}\sum_{k<N_\eta(\tau_\eta)} \|\chi_{\eta,k+1}\| = o_{\mathbb{P}}(1).$$

Moreover, for every uniformly bounded predictable linear-map process $\mathcal{T}_k$ of compatible dimensions,

$$\frac{\lambda_\eta}{\eta}\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=0}^{n-1}\mathcal{T}_k[\nu^v_{\eta,k+1}]\right\| = o_{\mathbb{P}}(1).$$

Proof. Recall that $\boldsymbol{\theta}^\circ_k := \boldsymbol{\theta}_k + Q_{\eta,k}$. From Lemma E.4,

$$\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -\eta S(\boldsymbol{v}_k)\big(\nabla\mathcal{L}(\boldsymbol{\theta}_k) + \boldsymbol{z}_k\big) + d^m_{\eta,k+1},$$

where $\boldsymbol{z}_k := \boldsymbol{z}_k(\boldsymbol{\theta}^\circ_k)$ is conditionally centered and

$$d^m_{\eta,k+1} := -\eta S(\boldsymbol{v}_k)\big(\nabla\mathcal{L}(\boldsymbol{\theta}^\circ_k) - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\big) + e_{\eta,k+1}.$$

After localizing the tight family $\boldsymbol{m}_0$, the uniform C^2-bounds and Lemma E.4 give

$$\|d^m_{\eta,k+1}\| \le C(\eta^2 + \eta\lambda_\eta). \tag{54}$$

Also

$$\Delta\boldsymbol{v}_{k+1} = \lambda_\eta\big(\bar{U}(\boldsymbol{\theta}^\circ_k) - \boldsymbol{v}_k\big) + \Delta M^v_{\eta,k+1}.$$

Apply second-order Taylor expansion to $(\boldsymbol{\theta},\boldsymbol{v}) \mapsto \boldsymbol{y}(\boldsymbol{\theta},\boldsymbol{v})$. Write $\Delta\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k$ and $\Delta\boldsymbol{v}_{k+1} := \boldsymbol{v}_{k+1} - \boldsymbol{v}_k$, and evaluate all derivatives below at $(\boldsymbol{\theta}_k, \boldsymbol{v}_k)$. Then

$$\begin{aligned}\boldsymbol{y}_{k+1} - \boldsymbol{y}_k = {}& \partial_{\boldsymbol{\theta}}\boldsymbol{y}_k[\Delta\boldsymbol{\theta}_{k+1}] + \partial_{\boldsymbol{v}}\boldsymbol{y}_k[\Delta\boldsymbol{v}_{k+1}] + \frac{1}{2}\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\boldsymbol{y}_k[\Delta\boldsymbol{\theta}_{k+1},\Delta\boldsymbol{\theta}_{k+1}] \\ & + \partial^2_{\boldsymbol{\theta}\boldsymbol{v}}\boldsymbol{y}_k[\Delta\boldsymbol{\theta}_{k+1},\Delta\boldsymbol{v}_{k+1}] + \frac{1}{2}\partial^2_{\boldsymbol{v}\boldsymbol{v}}\boldsymbol{y}_k[\Delta\boldsymbol{v}_{k+1},\Delta\boldsymbol{v}_{k+1}] + \mathcal{R}^y_{\eta,k+1},\end{aligned} \tag{55}$$

where

$$\|\mathcal{R}^y_{\eta,k+1}\| \le C\big(\|\Delta\boldsymbol{\theta}_{k+1}\| + \|\Delta\boldsymbol{v}_{k+1}\|\big)^3 = \mathcal{O}(\eta^3).$$

Substituting the corrected increments, the two first-order terms in (55) are

$$\begin{aligned}& -\eta\partial_{\boldsymbol{\theta}}\boldsymbol{y}_k[S(\boldsymbol{v}_k)\nabla\mathcal{L}(\boldsymbol{\theta}_k)] - \eta\partial_{\boldsymbol{\theta}}\boldsymbol{y}_k[S(\boldsymbol{v}_k)\boldsymbol{z}_k] \\ & \quad + \partial_{\boldsymbol{\theta}}\boldsymbol{y}_k[d^m_{\eta,k+1}] + \lambda_\eta\partial_{\boldsymbol{v}}\boldsymbol{y}_k[\bar{U}(\boldsymbol{\theta}^\circ_k) - \boldsymbol{v}_k] + \partial_{\boldsymbol{v}}\boldsymbol{y}_k[\Delta M^v_{\eta,k+1}].\end{aligned}$$

Moreover, the stable-coordinate parametrization gives

$$\boldsymbol{\theta}_k = \boldsymbol{\zeta}_k + \boldsymbol{y}_k + \mathcal{O}(\|\boldsymbol{y}_k\|^2), \qquad \nabla\mathcal{L}(\boldsymbol{\theta}_k) = \boldsymbol{H}_{\boldsymbol{\zeta}_k}\boldsymbol{y}_k + \mathcal{O}(\|\boldsymbol{y}_k\|^2),$$

and $\partial_{\boldsymbol{\theta}}\boldsymbol{y}(\boldsymbol{\zeta}_k,\boldsymbol{v}_k)$ is the identity on $E^{\mathrm{s}}_{\boldsymbol{\zeta}_k,\boldsymbol{v}_k}$. Hence

$$-\eta\partial_{\boldsymbol{\theta}}\boldsymbol{y}_k[S(\boldsymbol{v}_k)\nabla\mathcal{L}(\boldsymbol{\theta}_k)] = -\eta\boldsymbol{A}_k\boldsymbol{y}_k + \mathcal{O}(\eta\|\boldsymbol{y}_k\|^2).$$

Defining $\boldsymbol{C}_k := -\partial_{\boldsymbol{\theta}} \boldsymbol{y}_k \circ S(\boldsymbol{v}_k)$, the stochastic-gradient term is $\eta \boldsymbol{C}_k \boldsymbol{z}_k$, with predictable $\boldsymbol{C}_k$ and conditional mean zero.

We next linearize the deterministic $\boldsymbol{v}$-drift contribution in the stable coordinate. Since $\partial_{\boldsymbol{v}} \boldsymbol{y}(\boldsymbol{\zeta}, \boldsymbol{v}) = 0$ on Γ, define, for $\boldsymbol{u} \in E^{\mathrm{s}}_{\boldsymbol{\zeta},\boldsymbol{v}}$,

$$\boldsymbol{K}(\boldsymbol{\zeta}, \boldsymbol{v})\boldsymbol{u} := \partial^2_{\boldsymbol{\theta}\boldsymbol{v}} \boldsymbol{y}(\boldsymbol{\zeta}, \boldsymbol{v})[\boldsymbol{u}, \bar{U}(\boldsymbol{\zeta}) - \boldsymbol{v}]. \tag{56}$$

Using $\boldsymbol{\theta}_k = \boldsymbol{\zeta}_k + \boldsymbol{y}_k + \mathcal{O}(\|\boldsymbol{y}_k\|^2)$ and $\boldsymbol{\theta}^\circ_k = \boldsymbol{\theta}_k + Q_{\eta,k}$, Taylor's theorem gives

$$\partial_{\boldsymbol{v}} \boldsymbol{y}(\boldsymbol{\theta}_k, \boldsymbol{v}_k)[\bar{U}(\boldsymbol{\theta}^\circ_k) - \boldsymbol{v}_k] = \boldsymbol{K}(\boldsymbol{\zeta}_k, \boldsymbol{v}_k)\boldsymbol{y}_k + \boldsymbol{r}^K_{\eta,k+1}, \qquad \|\boldsymbol{r}^K_{\eta,k+1}\| \le C\big(\|\boldsymbol{y}_k\|^2 + \eta\|\boldsymbol{y}_k\|\big) \tag{57}$$

after localization. The term $\lambda_\eta \boldsymbol{r}^K_{\eta,k+1}$ is included in $\chi_{\eta,k+1}$; indeed, Young's inequality bounds it by terms already present in (58). We therefore set $\boldsymbol{K}_k := \boldsymbol{K}(\boldsymbol{\zeta}_k, \boldsymbol{v}_k)$. The field $\boldsymbol{K}$ is smooth with uniformly bounded derivatives of every fixed finite order on the compact working zone, and it may be smoothly extended from the stable bundle to the ambient space. Let

$$\nu^v_{\eta,k+1} := \partial_{\boldsymbol{v}} \boldsymbol{y}(\boldsymbol{\theta}_k, \boldsymbol{v}_k)[\Delta M^v_{\eta,k+1}].$$

Its coefficient is predictable, so it is a martingale difference, and $\|\nu^v_{\eta,k+1}\| \le C\lambda_\eta \|\boldsymbol{y}_k\|$.

The three second-order terms and $\mathcal{R}^y_{\eta,k+1}$ in (55), together with $\partial_{\boldsymbol{\theta}} \boldsymbol{y}_k[d^m_{\eta,k+1}]$ and the nonlinear errors displayed above, are collected in $\chi_{\eta,k+1}$. Using (54), $\|\Delta\boldsymbol{\theta}_{k+1}\| \le C\eta$, and $\|\Delta\boldsymbol{v}_{k+1}\| \le C\lambda_\eta$, they satisfy

$$\|\chi_{\eta,k+1}\| \le C\eta\|\boldsymbol{y}_k\|^2 + C\lambda_\eta\|\boldsymbol{y}_k\|^2 + C\eta^2 + C\eta\lambda_\eta + C\lambda^2_\eta. \tag{58}$$

This proves the finite-variation remainder estimate in Lemma F.10. Finally, the predictable quadratic variation of the scaled martingale asserted in Lemma F.10 is bounded by

$$C\left(\frac{\lambda_\eta}{\eta}\right)^2 \sum_{k<N_\eta(\tau_\eta)} \lambda^2_\eta \|\boldsymbol{y}_k\|^2 = C\frac{\lambda^4_\eta}{\eta^3} \int_0^{\tau_\eta} \|\boldsymbol{y}_\eta(s-)\|^2 \,\mathrm{d}A_\eta(s) = o_{\mathbb{P}}(1).$$

Doob's inequality after localization proves the martingale assertion of the last inequality. □

We first isolate the projected-increment expansion used both here and in the subsequent slow-limit calculation. For $k < N_\eta(\tau_\eta)$, write $\Delta\boldsymbol{\zeta}_{k+1} := \boldsymbol{\zeta}_{k+1} - \boldsymbol{\zeta}_k$. For every derivative appearing below, use the shorthand

$$\partial^j \Psi_k := \partial^j \Psi(\boldsymbol{\theta}_k, \boldsymbol{v}_k).$$

Recall the Taylor expansion (27) and the remainder bound (28) from Lemma F.6. On the stopped working zone they imply

$$\sum_{k<N_\eta(\tau_\eta)} \|\mathcal{R}^\Psi_{\eta,k+1}\| \le C\eta^{-2}(\eta + \lambda_\eta)^3 = o(1).$$

Lemma F.11 (Projected-increment decomposition)**.** *Define the leading martingale increment*

$$\Delta M^\zeta_{\eta,k+1} := -\eta\, \partial_{\boldsymbol{\theta}} \Psi_k\big[S(\boldsymbol{v}_k)\boldsymbol{z}_k\big] + \partial_{\boldsymbol{v}} \Psi_k[\Delta M^v_{\eta,k+1}].$$

Then

$$\mathbb{E}_k[\Delta M^\zeta_{\eta,k+1}] = \boldsymbol{0}, \qquad \|\Delta M^\zeta_{\eta,k+1}\| \le C\eta.$$

Moreover,

$$\|\Delta\boldsymbol{\zeta}_{k+1} - \Delta M^\zeta_{\eta,k+1}\| \le C\eta^2 + C\lambda_\eta\|\boldsymbol{y}_k\|.$$

The corresponding square-increment bounds are

$$\sum_{k<N_\eta(\tau_\eta)} \|\Delta\boldsymbol{\zeta}_{k+1}\|^2 = \mathcal{O}_{\mathbb{P}}(1),$$

and

$$\sum_{k<N_\eta(\tau_\eta)} \|\Delta\boldsymbol{v}_{k+1}\|^2 = \mathcal{O}_{\mathbb{P}}(\eta^{2\alpha-2}) = o_{\mathbb{P}}(1).$$

Proof. Recall the increment identities

$$\Delta\boldsymbol{\theta}_{k+1} = -\eta S(\boldsymbol{v}_k)\big(\nabla\mathcal{L}(\boldsymbol{\theta}_k) + \boldsymbol{z}_k\big) + d^m_{\eta,k+1},$$
$$\Delta\boldsymbol{v}_{k+1} = \lambda_\eta\big(\bar{U}(\boldsymbol{\theta}^\circ_k) - \boldsymbol{v}_k\big) + \Delta M^v_{\eta,k+1},$$

and the identity

$$\partial_{\boldsymbol{\theta}}\Psi_k[S(\boldsymbol{v}_k)\nabla\mathcal{L}(\boldsymbol{\theta}_k)] = \mathbf{0}$$

Substituting the two increment identities into the Taylor expansion (27) gives

$$\begin{aligned}\Delta\boldsymbol{\zeta}_{k+1} - \Delta M^\zeta_{\eta,k+1} &= \partial_{\boldsymbol{\theta}}\Psi_k[d^m_{\eta,k+1}] + \lambda_\eta\partial_{\boldsymbol{v}}\Psi_k[\bar{U}(\boldsymbol{\theta}^\circ_k) - \boldsymbol{v}_k] \\ &\quad + \frac{1}{2}\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi_k[\Delta\boldsymbol{\theta}_{k+1}, \Delta\boldsymbol{\theta}_{k+1}] + \partial^2_{\boldsymbol{\theta}\boldsymbol{v}}\Psi_k[\Delta\boldsymbol{\theta}_{k+1}, \Delta\boldsymbol{v}_{k+1}] \\ &\quad + \frac{1}{2}\partial^2_{\boldsymbol{v}\boldsymbol{v}}\Psi_k[\Delta\boldsymbol{v}_{k+1}, \Delta\boldsymbol{v}_{k+1}] + \mathcal{R}^\Psi_{\eta,k+1}.\end{aligned}$$

Since the linear $\boldsymbol{v}$-drift contributes $C\lambda_\eta\|\boldsymbol{y}_k\|$, the quadratic terms contribute $C\eta^2$, and the cubic remainder is absorbed into these terms, this proves the remainder bound above.

It remains to prove the square-increment bounds. Since $\|\Delta M^\zeta_{\eta,k+1}\| \le C\eta$, the leading martingale increments satisfy

$$\sum_{k<N_\eta(\tau_\eta)} \|\Delta M^\zeta_{\eta,k+1}\|^2 = \mathcal{O}_{\mathbb{P}}(1).$$

For the remaining part, the preceding estimate and (53) give

$$\sum_{k<N_\eta(\tau_\eta)} \|\Delta\boldsymbol{\zeta}_{k+1} - \Delta M^\zeta_{\eta,k+1}\|^2 \le C\eta^2 + C\frac{\lambda_\eta^2}{\eta}\int_0^{\tau_\eta} \|\boldsymbol{y}_\eta(s-)\|^2\,\mathrm{d}A_\eta(s) = o_{\mathbb{P}}(1).$$

This proves the bound for $\Delta\boldsymbol{\zeta}$. Finally, $\|\Delta\boldsymbol{v}_{k+1}\| \le C\lambda_\eta$ on the stopped working zone, so

$$\sum_{k<N_\eta(\tau_\eta)} \|\Delta\boldsymbol{v}_{k+1}\|^2 \le C\eta^{-2}\lambda_\eta^2 = \mathcal{O}(\eta^{2\alpha-2}) = o(1).$$

□

We shall also use the following rough projected-increment estimate to prove the normal-space averaging result.

Lemma F.12 (Rough projected increment estimates)**.** *Let $\mathcal{B}_k$ be any uniformly bounded predictable bilinear-map process of compatible dimensions. Then*

$$\frac{\lambda_\eta}{\eta}\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=1}^{n-1}\mathcal{B}_{k-1}[\Delta M^\zeta_{\eta,k}, \boldsymbol{y}_k]\right\| = o_{\mathbb{P}}(1),$$

$$\frac{\lambda_\eta}{\eta}\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=1}^{n-1}\mathcal{B}_{k-1}\big[\Delta\boldsymbol{\zeta}_k - \Delta M^\zeta_{\eta,k}, \boldsymbol{y}_k\big]\right\| = o_{\mathbb{P}}(1).$$

Proof. For the first assertion of the lemma, the second argument $\boldsymbol{y}_k$ is not predictable for the increment $\Delta M^\zeta_{\eta,k}$. By bilinearity,

$$\mathcal{B}_{k-1}[\Delta M^\zeta_{\eta,k}, \boldsymbol{y}_k] = \mathcal{B}_{k-1}[\Delta M^\zeta_{\eta,k}, \boldsymbol{y}_{k-1}] + \mathcal{B}_{k-1}[\Delta M^\zeta_{\eta,k}, \boldsymbol{y}_k - \boldsymbol{y}_{k-1}].$$

The first term is a genuine martingale transform. By (53), its scaled predictable quadratic variation is bounded by

$$C\left(\frac{\lambda_\eta}{\eta}\right)^2\sum_{k<N_\eta(\tau_\eta)}\big(\eta^2\|\boldsymbol{y}_{k-1}\|^2 + \lambda_\eta^2\|\boldsymbol{y}_{k-1}\|^4\big) = o_{\mathbb{P}}(1).$$

where Doob's inequality can be applied to this predictable transform. For the second term, $\|\boldsymbol{y}_k - \boldsymbol{y}_{k-1}\| \le C\eta$, so

$$\frac{\lambda_\eta}{\eta} \sum_{k<N_\eta(\tau_\eta)} \|\Delta M^{\zeta}_{\eta,k}\| \|\boldsymbol{y}_k - \boldsymbol{y}_{k-1}\| \le C\frac{\lambda_\eta}{\eta} = o_{\mathbb{P}}(1).$$

This proves the first assertion.

For the second assertion, use the final bound in Lemma F.11 with $k-1$. Multiplying by $(\lambda_\eta/\eta)\|\boldsymbol{y}_k\|$ and summing, the $C\eta^2$ term is controlled by Cauchy-Schwarz:

$$C\frac{\lambda_\eta}{\eta} \sum_k \eta^2 \|\boldsymbol{y}_k\| \le C\frac{\lambda_\eta}{\eta} \left(\sum_k \eta \|\boldsymbol{y}_k\|^2\right)^{1/2} \left(\sum_k \eta^3\right)^{1/2} = \mathcal{O}_{\mathbb{P}}(\eta^{\alpha-1/2}).$$

The other term gives

$$C\frac{\lambda_\eta}{\eta} \sum_k \lambda_\eta \|\boldsymbol{y}_{k-1}\| \|\boldsymbol{y}_k\| \le C\frac{\lambda_\eta}{\eta} \sum_k \lambda_\eta \big(\|\boldsymbol{y}_{k-1}\|^2 + \|\boldsymbol{y}_k\|^2\big) = o_{\mathbb{P}}(1),$$

and hence the second assertion follows. □

The following elementary lemma isolates the finite iteration used in Lemma F.14.

Lemma F.13 (Finite absorption bootstrap)**.** *Let $\mathscr{S}_\eta$ be nonnegative random functionals on a class of coefficient fields, and let $\mathscr{R}_1, \mathscr{R}_2$ map that class into itself. Fix a coefficient $\boldsymbol{Q}$, and suppose that every coefficient $\boldsymbol{R}$ obtained from $\boldsymbol{Q}$ by finitely many applications of $\mathscr{R}_1, \mathscr{R}_2$ satisfies*

$$\mathscr{S}_\eta(\boldsymbol{R}) \le \rho_\eta(\boldsymbol{R}) + C_{\boldsymbol{R}} \frac{\lambda_\eta}{\eta} \left(\mathscr{S}_\eta(\mathscr{R}_1\boldsymbol{R}) + \mathscr{S}_\eta(\mathscr{R}_2\boldsymbol{R})\right),$$

where $C_{\boldsymbol{R}}$ is independent of η, $\rho_\eta(\boldsymbol{R}) = o_{\mathbb{P}}(1)$ and $\mathscr{S}_\eta(\boldsymbol{R}) = \mathcal{O}_{\mathbb{P}}(\lambda_\eta \eta^{-3/2})$. Then

$$\mathscr{S}_\eta(\boldsymbol{Q}) = o_{\mathbb{P}}(1).$$

Proof. Since $\alpha > 1$, choose a fixed integer $J \ge 1$ such that $J(\alpha-1) + \alpha - \frac{3}{2} > 0$. For a word $w = (w_1, \dots, w_j) \in \{1,2\}^j$, write

$$\boldsymbol{Q}_w := \mathscr{R}_{w_j} \cdots \mathscr{R}_{w_1} \boldsymbol{Q},$$

where $\boldsymbol{Q}_\varnothing := \boldsymbol{Q}$. Only the finitely many fields with $|w| \le J$ occur below, so their constants $C_{\boldsymbol{Q}_w}$ may be bounded by a single deterministic constant C_0. Iterating the recursive inequality in the lemma J times and including the finite branching factor in C_0 gives

$$\mathscr{S}_\eta(\boldsymbol{Q}) \le \sum_{j=0}^{J-1} \left(C_0 \frac{\lambda_\eta}{\eta}\right)^j \max_{|w|=j} \rho_\eta(\boldsymbol{Q}_w) + \left(C_0 \frac{\lambda_\eta}{\eta}\right)^J \max_{|w|=J} \mathscr{S}_\eta(\boldsymbol{Q}_w).$$

The finite sum is $o_{\mathbb{P}}(1)$ and the last term is

$$\mathcal{O}_{\mathbb{P}}\left(\left(\frac{\lambda_\eta}{\eta}\right)^J \lambda_\eta \eta^{-3/2}\right) = \mathcal{O}_{\mathbb{P}}\left(\eta^{J(\alpha-1)+\alpha-3/2}\right) = o_{\mathbb{P}}(1)$$

by the choice of J. This proves the claim. □

We now prove the normal-space averaging estimate. The statement is formulated for test matrices that may depend on both $\boldsymbol{\zeta}$ and $\boldsymbol{v}$, since this is the form needed later when handling $\partial_{\boldsymbol{v}} \Phi_{S(\boldsymbol{v})} \Delta \boldsymbol{v}$-terms.

Lemma F.14 (Normal-Space B_η-averaging)**.** *Let*

$$\boldsymbol{Q} : \Gamma \times [0, R_1]^D \to \mathbb{R}^{p \times d}$$

be smooth with bounded derivatives of every fixed finite order. Then

$$\sup_{0 \le t \le \tau_\eta} \left\| \int_0^t \boldsymbol{Q}(\boldsymbol{\zeta}_\eta(s-), \boldsymbol{v}_\eta(s-)) \boldsymbol{r}_\eta(s-) \, \mathrm{d}B_\eta(s) \right\| = o_{\mathbb{P}}(1).$$

Equivalently, in discrete form,

$$\sup_{n \le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \lambda_\eta \boldsymbol{Q}(\boldsymbol{\zeta}_k, \boldsymbol{v}_k) \boldsymbol{r}_k \right\| = o_{\mathbb{P}}(1).$$

Proof. By (48), $\boldsymbol{r}_k = \boldsymbol{y}_k + \mathcal{O}(\|\boldsymbol{y}_k\|^2)$. Therefore

$$\sup_{n\le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \lambda_\eta \boldsymbol{Q}(\boldsymbol{\zeta}_k, \boldsymbol{v}_k)(\boldsymbol{r}_k - \boldsymbol{y}_k) \right\| \le C \sum_{k<N_\eta(\tau_\eta)} \lambda_\eta \|\boldsymbol{y}_k\|^2 = o_{\mathbb{P}}(1)$$

by (52). Hence, it suffices to prove

$$\sup_{n\le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \lambda_\eta \boldsymbol{Q}(\boldsymbol{\zeta}_k, \boldsymbol{v}_k)\boldsymbol{y}_k \right\| = o_{\mathbb{P}}(1).$$

Let $\boldsymbol{B}(\boldsymbol{\zeta}, \boldsymbol{v})$ be a smooth extension of $\boldsymbol{A}_{\boldsymbol{\zeta},\boldsymbol{v}}^{-1}$ from the stable fiber to $\mathbb{R}^d$, chosen with uniformly bounded derivatives of every fixed finite order, and write $\boldsymbol{B}_k := \boldsymbol{B}(\boldsymbol{\zeta}_k, \boldsymbol{v}_k)$. By construction,

$$\boldsymbol{B}_k \boldsymbol{A}_k \boldsymbol{y}_k = \boldsymbol{y}_k.$$

Recall the recursion in Lemma F.10,

$$\eta \boldsymbol{A}_k \boldsymbol{y}_k = \boldsymbol{y}_k - \boldsymbol{y}_{k+1} + \eta \boldsymbol{C}_k \boldsymbol{z}_k + \lambda_\eta \boldsymbol{K}_k \boldsymbol{y}_k + \nu^v_{\eta,k+1} + \chi_{\eta,k+1}.$$

Hence, writing $\boldsymbol{Q}_k := \boldsymbol{Q}(\boldsymbol{\zeta}_k, \boldsymbol{v}_k)$, we get

$$\begin{aligned}\lambda_\eta \boldsymbol{Q}_k \boldsymbol{y}_k = \frac{\lambda_\eta}{\eta} \boldsymbol{Q}_k \boldsymbol{B}_k (\boldsymbol{y}_k - \boldsymbol{y}_{k+1}) + \lambda_\eta \boldsymbol{Q}_k \boldsymbol{B}_k \boldsymbol{C}_k \boldsymbol{z}_k \\ + \frac{\lambda_\eta^2}{\eta} \boldsymbol{Q}_k \boldsymbol{B}_k \boldsymbol{K}_k \boldsymbol{y}_k + \frac{\lambda_\eta}{\eta} \boldsymbol{Q}_k \boldsymbol{B}_k \nu^v_{\eta,k+1} + \frac{\lambda_\eta}{\eta} \boldsymbol{Q}_k \boldsymbol{B}_k \chi_{\eta,k+1}.\end{aligned}$$

Let $\boldsymbol{L}_k := \boldsymbol{Q}_k \boldsymbol{B}_k$. First, consider the telescoping term. We have

$$\sum_{k=0}^{n-1} \boldsymbol{L}_k (\boldsymbol{y}_k - \boldsymbol{y}_{k+1}) = \boldsymbol{L}_0 \boldsymbol{y}_0 - \boldsymbol{L}_{n-1} \boldsymbol{y}_n + \sum_{k=1}^{n-1} (\boldsymbol{L}_k - \boldsymbol{L}_{k-1}) \boldsymbol{y}_k.$$

The boundary contribution is bounded by

$$C \frac{\lambda_\eta}{\eta} \sup_{k\le N_\eta(\tau_\eta)} \|\boldsymbol{y}_k\| = o_{\mathbb{P}}(1).$$

It remains to bound the last term in the preceding Abel decomposition. Since $\boldsymbol{L}_k$ is smooth in $(\boldsymbol{\zeta}_k, \boldsymbol{v}_k)$, we have

$$\boldsymbol{L}_k - \boldsymbol{L}_{k-1} = \boldsymbol{D}^\zeta_{k-1}[\boldsymbol{\zeta}_k - \boldsymbol{\zeta}_{k-1}] + \boldsymbol{D}^v_{k-1}[\boldsymbol{v}_k - \boldsymbol{v}_{k-1}] + \mathcal{O}\left(\|\boldsymbol{\zeta}_k - \boldsymbol{\zeta}_{k-1}\|^2 + \|\boldsymbol{v}_k - \boldsymbol{v}_{k-1}\|^2\right),$$

where the tensors $\boldsymbol{D}^\zeta_{k-1}$ and $\boldsymbol{D}^v_{k-1}$ are uniformly bounded and predictable. The contribution of the $\boldsymbol{\zeta}$-increment in the last term of the Abel decomposition is controlled by the first two estimates in Lemma F.12, applied with

$$\mathcal{B}_{k-1}[\boldsymbol{u}, \boldsymbol{w}] := \big(\boldsymbol{D}^\zeta_{k-1}[\boldsymbol{u}]\big)\boldsymbol{w},$$

we obtain

$$\frac{\lambda_\eta}{\eta} \sup_{n\le N_\eta(\tau_\eta)} \left\| \sum_{k=1}^{n-1} \big(\boldsymbol{D}^\zeta_{k-1}[\boldsymbol{\zeta}_k - \boldsymbol{\zeta}_{k-1}]\big)\boldsymbol{y}_k \right\| = o_{\mathbb{P}}(1).$$

For the $\boldsymbol{v}$-increment, use the one-step form of the semimartingale decomposition (41):

$$\boldsymbol{v}_k - \boldsymbol{v}_{k-1} = \lambda_\eta\big(\bar{U}(\boldsymbol{\theta}^\circ_{k-1}) - \boldsymbol{v}_{k-1}\big) + \Delta M^v_{\eta,k}.$$

We split

$$\Delta M^v_{\eta,k} \boldsymbol{y}_k = \Delta M^v_{\eta,k} \boldsymbol{y}_{k-1} + \Delta M^v_{\eta,k} (\boldsymbol{y}_k - \boldsymbol{y}_{k-1}).$$

The first term is a martingale transform, whose scaled predictable quadratic variation is

$$\mathcal{O}_{\mathbb{P}}\left(\left(\frac{\lambda_\eta}{\eta}\right)^2 \sum_{k<N_\eta(\tau_\eta)} \lambda_\eta^2 \|\boldsymbol{y}_{k-1}\|^2 \right) = \mathcal{O}_{\mathbb{P}}\left(\frac{\lambda_\eta^4}{\eta^3}\right) = o_{\mathbb{P}}(1).$$

The second term is bounded pathwise after localization by

$$C\frac{\lambda_\eta}{\eta}\sum_{k<N_\eta(\tau_\eta)}\lambda_\eta\eta=\mathcal{O}\left(\frac{\lambda_\eta^2}{\eta^2}\right)=o(1).$$

The leading finite-variation contribution is more explicitly

$$\frac{\lambda_\eta}{\eta}\sum_{k<n}\lambda_\eta\boldsymbol{Q}^{(v)}(\boldsymbol{\zeta}_k,\boldsymbol{v}_k)\boldsymbol{y}_k,\qquad \boldsymbol{Q}^{(v)}(\boldsymbol{\zeta},\boldsymbol{v}):=\partial_{\boldsymbol{v}}(\boldsymbol{QB})(\boldsymbol{\zeta},\boldsymbol{v})[U(\boldsymbol{\zeta})-\boldsymbol{v}].$$

Indeed, replacing $\bar{U}(\boldsymbol{\theta}^\circ_{k-1})$ by $U(\boldsymbol{\zeta}_{k-1})$ produces only terms bounded by a constant times $\frac{\lambda_\eta}{\eta}\sum_k\lambda_\eta\|\boldsymbol{y}_k\|^2$, together with the already controlled momentum-correction error. Replacing the index $k-1$ by k produces terms of the same type as the martingale and quadratic Taylor terms just estimated. All of these errors are $o_{\mathbb{P}}(1)$. Thus the displayed term is the only part that must be reinserted into the averaging argument. Moreover, the quadratic Taylor terms $\mathcal{O}\left(\|\boldsymbol{\zeta}_k-\boldsymbol{\zeta}_{k-1}\|^2+\|\boldsymbol{v}_k-\boldsymbol{v}_{k-1}\|^2\right)\boldsymbol{y}_k$ are negligible by the last two estimates in Lemma F.11. Thus the full telescoping term is the sum of an $o_{\mathbb{P}}(1)$ remainder and $\frac{\lambda_\eta}{\eta}\mathscr{S}_\eta(\boldsymbol{Q}^{(v)})$, where

$$\mathscr{S}_\eta(\boldsymbol{R}):=\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=0}^{n-1}\lambda_\eta\boldsymbol{R}(\boldsymbol{\zeta}_k,\boldsymbol{v}_k)\boldsymbol{y}_k\right\|,$$

for any smooth test field $\boldsymbol{R}$ with bounded derivatives of every fixed finite order.

Second, the martingale term

$$\sum_{k=0}^{n-1}\lambda_\eta\boldsymbol{Q}_k\boldsymbol{B}_k\boldsymbol{C}_k\boldsymbol{z}_k$$

has centered increments. Its quadratic variation is bounded by $CN_\eta(T)\lambda_\eta^2=o(1)$. Therefore, by Doob's inequality,

$$\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=0}^{n-1}\lambda_\eta\boldsymbol{Q}_k\boldsymbol{B}_k\boldsymbol{C}_k\boldsymbol{z}_k\right\|=o_{\mathbb{P}}(1).$$

Third, the perturbative term is

$$\sup_{n\le N_\eta(\tau_\eta)}\left\|\frac{\lambda_\eta^2}{\eta}\sum_{k=0}^{n-1}\boldsymbol{Q}_k\boldsymbol{B}_k\boldsymbol{K}_k\boldsymbol{y}_k\right\|=\frac{\lambda_\eta}{\eta}\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=0}^{n-1}\lambda_\eta\boldsymbol{Q}^{(K)}(\boldsymbol{\zeta}_k,\boldsymbol{v}_k)\boldsymbol{y}_k\right\|,$$

where $\boldsymbol{Q}^{(K)}:=\boldsymbol{QBK}$. This is the second term that must be reinserted.

Fourth, the additional $\boldsymbol{v}$-martingale term is negligible by the martingale estimate in Lemma F.10, applied with $\mathcal{T}_k[\boldsymbol{u}]:=\boldsymbol{Q}_k\boldsymbol{B}_k\boldsymbol{u}$:

$$\frac{\lambda_\eta}{\eta}\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=0}^{n-1}\boldsymbol{Q}_k\boldsymbol{B}_k\nu^v_{\eta,k+1}\right\|=o_{\mathbb{P}}(1).$$

Fifth, by the finite-variation remainder estimate in Lemma F.10,

$$\sup_{n\le N_\eta(\tau_\eta)}\left\|\frac{\lambda_\eta}{\eta}\sum_{k=0}^{n-1}\boldsymbol{Q}_k\boldsymbol{B}_k\chi_{\eta,k+1}\right\|\le C\frac{\lambda_\eta}{\eta}\sum_{k<N_\eta(\tau_\eta)}\|\chi_{\eta,k+1}\|=o_{\mathbb{P}}(1).$$

We now apply Lemma F.13. Define the two coefficient transformations identified above by

$$\mathscr{R}_v\boldsymbol{R}:=\partial_{\boldsymbol{v}}(\boldsymbol{RB})[U-\boldsymbol{v}],\qquad \mathscr{R}_K\boldsymbol{R}:=\boldsymbol{RBK}.$$

Repeating the preceding decomposition with $\boldsymbol{R}$ in place of $\boldsymbol{Q}$ gives a nonnegative remainder $\rho_\eta(\boldsymbol{R})=o_{\mathbb{P}}(1)$ such that

$$\mathscr{S}_\eta(\boldsymbol{R})\le\rho_\eta(\boldsymbol{R})+C_{\boldsymbol{R}}\frac{\lambda_\eta}{\eta}\left(\mathscr{S}_\eta(\mathscr{R}_v\boldsymbol{R})+\mathscr{S}_\eta(\mathscr{R}_K\boldsymbol{R})\right).$$

Here $\rho_\eta(\boldsymbol{R})$ consists exactly of the boundary, martingale, quadratic Taylor, ν^v, and χ terms estimated above. In particular, the two terms multiplied by $\frac{\lambda_\eta}{\eta}$ are precisely the terms to which the finite absorption bootstrap applies. Every field obtained from $\boldsymbol{Q}$ by finitely many applications of $\mathscr{R}_v$ and $\mathscr{R}_K$ is again smooth. For each such fixed field, boundedness on the stopped working zone and Cauchy–Schwarz give the crude estimate

$$\begin{aligned}\mathscr{S}_\eta(\boldsymbol{R}) &\le C \sum_{k<N_\eta(\tau_\eta)} \lambda_\eta \|\boldsymbol{y}_k\| \\ &\le C\frac{\lambda_\eta}{\eta}\left(\sum_{k<N_\eta(\tau_\eta)} \eta\right)^{1/2}\left(\sum_{k<N_\eta(\tau_\eta)} \eta\|\boldsymbol{y}_k\|^2\right)^{1/2} \\ &= \mathcal{O}_{\mathbb{P}}(\lambda_\eta \eta^{-3/2}),\end{aligned}$$

where we used (53). Hence the preceding bootstrap inequality and the crude estimate verify all hypotheses of Lemma F.13, with $\mathscr{R}_1 = \mathscr{R}_v$ and $\mathscr{R}_2 = \mathscr{R}_K$. The lemma therefore gives

$$\sup_{n\le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \lambda_\eta \boldsymbol{Q}(\boldsymbol{\zeta}_k, \boldsymbol{v}_k)\boldsymbol{y}_k \right\| = o_{\mathbb{P}}(1).$$

As explained at the beginning of the proof, replacing $\boldsymbol{y}_k$ by $\boldsymbol{r}_k$ only introduces the negligible error, which proves both assertions of the lemma for the corrected coordinates. □

Recall that the same statements for the original coordinates follow from (51) and the uniform continuity of $\boldsymbol{Q}$. We use $(\boldsymbol{\theta}^\circ_\eta, \boldsymbol{\zeta}^\circ_\eta, \boldsymbol{r}^\circ_\eta)$ for their corresponding step interpolations.

The following corollary evaluates the B_η-drift target in the $\boldsymbol{v}$-equation at the original projected process $\boldsymbol{\zeta}^\circ_\eta$.

Corollary F.4.

$$\sup_{0\le t\le\tau_\eta} \left\| \int_0^t \left(\bar{U}(\boldsymbol{\theta}^\circ_\eta(s-)) - U(\boldsymbol{\zeta}^\circ_\eta(s-))\right) \mathrm{d}B_\eta(s) \right\| = o_{\mathbb{P}}(1).$$

Proof. Since $\bar{U}$ is smooth in the working zone and $\bar{U}(\boldsymbol{\zeta}) = U(\boldsymbol{\zeta})$ for $\boldsymbol{\zeta} \in \Gamma$, Taylor expansion gives

$$\bar{U}(\boldsymbol{\theta}^\circ_\eta) - U(\boldsymbol{\zeta}^\circ_\eta) = \partial_{\boldsymbol{\theta}}\bar{U}(\boldsymbol{\zeta}^\circ_\eta)\boldsymbol{r}^\circ_\eta + \mathcal{O}(\|\boldsymbol{r}^\circ_\eta\|^2).$$

Therefore

$$\begin{aligned}&\sup_{0\le t\le\tau_\eta} \left\| \int_0^t \left(\bar{U}(\boldsymbol{\theta}^\circ_\eta(s-)) - U(\boldsymbol{\zeta}^\circ_\eta(s-))\right) \mathrm{d}B_\eta(s) \right\| \\ &\le \sup_{0\le t\le\tau_\eta} \left\| \int_0^t \partial_{\boldsymbol{\theta}}\bar{U}(\boldsymbol{\zeta}^\circ_\eta(s-))\boldsymbol{r}^\circ_\eta(s-)\,\mathrm{d}B_\eta(s) \right\| + C\int_0^{\tau_\eta} \|\boldsymbol{r}^\circ_\eta(s-)\|^2\,\mathrm{d}B_\eta(s).\end{aligned}$$

The first term converges to zero by the original-coordinate conclusion of Lemma F.14 with $\boldsymbol{Q}(\boldsymbol{\zeta}, \boldsymbol{v}) = \partial_{\boldsymbol{\theta}}\bar{U}(\boldsymbol{\zeta})$, while the second term converges to zero by Corollary F.2. Hence, the claim follows. □

F.6 Projection reduction and fast tracking of the preconditioner

We continue with the momentum-corrected coordinates from Section F.5. Recall that the superscript $\circ$ denotes the original AGM coordinates, while

$$\boldsymbol{\zeta}^\circ_k := \Psi(\boldsymbol{\theta}^\circ_k, \boldsymbol{v}_k), \qquad \boldsymbol{\theta}_k := \boldsymbol{\theta}^\circ_k - \boldsymbol{Q}_{\eta,k}, \qquad \boldsymbol{\zeta}_k := \Psi(\boldsymbol{\theta}_k, \boldsymbol{v}_k),$$
$$\boldsymbol{r}_k := \boldsymbol{\theta}_k - \boldsymbol{\zeta}_k, \qquad \boldsymbol{z}_k := \boldsymbol{z}_k(\boldsymbol{\theta}^\circ_k).$$

where $\mathbb{E}_k[\boldsymbol{z}_k] = 0$ and

$$\Delta\boldsymbol{\theta}_{k+1} = -\eta S(\boldsymbol{v}_k)\big(\nabla\mathcal{L}(\boldsymbol{\theta}_k) + \boldsymbol{z}_k\big) + d^m_{\eta,k+1}, \quad \|d^m_{\eta,k+1}\| \le C(\eta^2 + \eta\lambda_\eta).$$

Recall the second-order Taylor expansion (27) and its remainder bound (28). After substituting the corrected increments, Lemma F.11 cancels the deterministic gradient term and gives the required one-step bound. It remains to show that the momentum mismatch and the terms containing $\Delta\boldsymbol{v}_{k+1}$ vanish in the slow limit.

Lemma F.15 (Vanishing of the projected momentum mismatch)**.** *Let*

$$\delta\boldsymbol{\theta}_{k+1} := -\eta S(\boldsymbol{v}_k)\big(\nabla\mathcal{L}(\boldsymbol{\theta}_k) + \boldsymbol{z}_k\big).$$

Then

$$\sup_{n\le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \partial_{\boldsymbol{\theta}} \Psi_k[d^m_{\eta,k+1}] \right\| = o_{\mathbb{P}}(1), \tag{59}$$

and

$$\sum_{k<N_\eta(\tau_\eta)} \big\| \partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \Psi_k[\Delta\boldsymbol{\theta}_{k+1}, \Delta\boldsymbol{\theta}_{k+1}] - \partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \Psi_k[\delta\boldsymbol{\theta}_{k+1}, \delta\boldsymbol{\theta}_{k+1}] \big\| = o_{\mathbb{P}}(1). \tag{60}$$

Proof. Recall that in Lemma F.10, we defined

$$d^m_{\eta,k+1} = -\eta S(\boldsymbol{v}_k)\big(\nabla\mathcal{L}(\boldsymbol{\theta}_k + \boldsymbol{Q}_{\eta,k}) - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\big) + e_{\eta,k+1}.$$

where the coefficient-lag part is negligible by Lemma E.4:

$$\sum_{k<N_\eta(\tau_\eta)} \|e_{\eta,k+1}\| = \mathcal{O}_{\mathbb{P}}(\eta^{\alpha-1}) = o_{\mathbb{P}}(1).$$

For the remaining part, Taylor expansion around $\boldsymbol{\zeta}_k$ and the identity

$$\partial\Phi_{S(\boldsymbol{v}_k)}(\boldsymbol{\zeta}_k) S(\boldsymbol{v}_k) \nabla^2\mathcal{L}(\boldsymbol{\zeta}_k) = 0$$

give

$$\|\partial_{\boldsymbol{\theta}}\Psi_k S(\boldsymbol{v}_k)[\nabla\mathcal{L}(\boldsymbol{\theta}_k + \boldsymbol{Q}_{\eta,k}) - \nabla\mathcal{L}(\boldsymbol{\theta}_k)]\| \le C\big(\|\boldsymbol{r}_k\|\|\boldsymbol{Q}_{\eta,k}\| + \|\boldsymbol{Q}_{\eta,k}\|^2\big).$$

Since $\sup_k \|\boldsymbol{Q}_{\eta,k}\| = \mathcal{O}_{\mathbb{P}}(\eta)$ and the corrected collapse follows from (50) and Lemma F.9,

$$\sup_{n\le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \partial_{\boldsymbol{\theta}} \Psi_k[d^m_{\eta,k+1}] \right\| \le o_{\mathbb{P}}(1) + C \sum_{k<N_\eta(\tau_\eta)} (\eta^2\|\boldsymbol{r}_k\| + \eta^3) = o_{\mathbb{P}}(1).$$

This proves (59).

Finally, $\Delta\boldsymbol{\theta}_{k+1} = \delta\boldsymbol{\theta}_{k+1} + d^m_{\eta,k+1}$, $\|\delta\boldsymbol{\theta}_{k+1}\| \le C\eta$, and $\|d^m_{\eta,k+1}\| \le C(\eta^2 + \eta\lambda_\eta)$. Hence the left-hand side of (60) is bounded by $C\eta^{-2}\left(\eta^3 + \eta^2\lambda_\eta + \eta^4 + \eta^2\lambda_\eta^2\right) = o_{\mathbb{P}}(1)$. □

The next lemma is the main point of this section. It shows that the first-order $\boldsymbol{v}$-dependence of the adaptive projection does not create an additional drift.

Lemma F.16 (Vanishing of the first-order $\boldsymbol{v}$-projection term)**.**

$$\sup_{n\le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \partial_{\boldsymbol{v}} \Psi_k[\Delta\boldsymbol{v}_{k+1}] \right\| = o_{\mathbb{P}}(1). \tag{61}$$

Proof. Recall the $\boldsymbol{v}$-equation. We decompose

$$\partial_{\boldsymbol{v}}\Psi_k[\Delta\boldsymbol{v}_{k+1}] = \lambda_\eta \partial_{\boldsymbol{v}}\Psi_k[\bar{U}(\boldsymbol{\theta}^\circ_k) - \boldsymbol{v}_k] + \partial_{\boldsymbol{v}}\Psi_k[\Delta M^v_{\eta,k+1}].$$

We first treat the finite-variation part. Write

$$\bar{U}(\boldsymbol{\theta}^\circ_k) - \boldsymbol{v}_k = \big(U(\boldsymbol{\zeta}_k) - \boldsymbol{v}_k\big) + \big(\bar{U}(\boldsymbol{\theta}_k) - U(\boldsymbol{\zeta}_k)\big) + \big(\bar{U}(\boldsymbol{\theta}^\circ_k) - \bar{U}(\boldsymbol{\theta}_k)\big).$$

Since $\partial_{\boldsymbol{v}}\Psi(\boldsymbol{\zeta}, \boldsymbol{v}) = \mathbf{0}$ on Γ, Taylor expansion in the stable direction gives, uniformly in the working zone,

$$\partial_{\boldsymbol{v}}\Psi(\boldsymbol{\theta}_k, \boldsymbol{v}_k)[U(\boldsymbol{\zeta}_k) - \boldsymbol{v}_k] = \boldsymbol{Q}_1(\boldsymbol{\zeta}_k, \boldsymbol{v}_k)\boldsymbol{r}_k + \mathcal{O}(\|\boldsymbol{r}_k\|^2),$$

where $\boldsymbol{Q}_1$ is a bounded smooth matrix-valued map. First, Lemma F.14 gives

$$\sup_{n\le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \lambda_\eta \boldsymbol{Q}_1(\boldsymbol{\zeta}_k, \boldsymbol{v}_k)\boldsymbol{r}_k \right\| = o_{\mathbb{P}}(1).$$

Second, the Taylor remainder and Corollary F.2 imply

$$\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=0}^{n-1}\lambda_\eta\left(\partial_{\boldsymbol{v}}\Psi_k[U(\boldsymbol{\zeta}_k)-\boldsymbol{v}_k]-\boldsymbol{Q}_1(\boldsymbol{\zeta}_k,\boldsymbol{v}_k)\boldsymbol{r}_k\right)\right\|\le C\sum_{k<N_\eta(\tau_\eta)}\lambda_\eta\|\boldsymbol{r}_k\|^2=o_{\mathbb{P}}(1).$$

Combining the two inequalities yields

$$\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=0}^{n-1}\lambda_\eta\partial_{\boldsymbol{v}}\Psi_k[U(\boldsymbol{\zeta}_k)-\boldsymbol{v}_k]\right\|=o_{\mathbb{P}}(1). \tag{62}$$

For the second finite-variation term, by $\|\partial_{\boldsymbol{v}}\Psi_k\|\le C\|\boldsymbol{r}_k\|$ and the Taylor expansion of $\bar{U}$:

$$\|\partial_{\boldsymbol{v}}\Psi_k[\bar{U}(\boldsymbol{\theta}_k)-U(\boldsymbol{\zeta}_k)]\|\le C\|\boldsymbol{r}_k\|\|\bar{U}(\boldsymbol{\theta}_k)-U(\boldsymbol{\zeta}_k)\|\le C\|\boldsymbol{r}_k\|^2.$$

Therefore by Corollary F.2,

$$\sum_{k<N_\eta(\tau_\eta)}\lambda_\eta\|\partial_{\boldsymbol{v}}\Psi_k[\bar{U}(\boldsymbol{\theta}_k)-U(\boldsymbol{\zeta}_k)]\|\le C\int_0^{\tau_\eta}\|\boldsymbol{r}_\eta(s-)\|^2\,\mathrm{d}B_\eta(s)=o_{\mathbb{P}}(1). \tag{63}$$

The last term is also negligible. Indeed,

$$\|\bar{U}(\boldsymbol{\theta}_k^\circ)-\bar{U}(\boldsymbol{\theta}_k)\|\le C\|Q_{\eta,k}\|=\mathcal{O}_{\mathbb{P}}(\eta)$$

uniformly after localization, while $\|\partial_{\boldsymbol{v}}\Psi_k\|\le C\|\boldsymbol{r}_k\|$. Therefore

$$\sum_{k<N_\eta(\tau_\eta)}\lambda_\eta\|\partial_{\boldsymbol{v}}\Psi_k[\bar{U}(\boldsymbol{\theta}_k^\circ)-\bar{U}(\boldsymbol{\theta}_k)]\|\le C\frac{\lambda_\eta}{\eta}\sup_{k<N_\eta(\tau_\eta)}\|\boldsymbol{r}_k\|=o_{\mathbb{P}}(1). \tag{64}$$

It remains to control the martingale part

$$\mathcal{M}_{\eta,n}^{\Psi}:=\sum_{k=0}^{n-1}\partial_{\boldsymbol{v}}\Psi_k[\Delta M_{\eta,k+1}^{v}].$$

To control this martingale, use $\|\partial_{\boldsymbol{v}}\Psi_k\|\le C\|\boldsymbol{r}_k\|$ and the boundedness of $V(\boldsymbol{g}\boldsymbol{g}^\top)$. Its optional quadratic variation is bounded pathwise by

$$[\mathcal{M}_\eta^{\Psi}]_{N_\eta(\tau_\eta)}\le C\sum_{k<N_\eta(\tau_\eta)}\lambda_\eta^2\|\boldsymbol{r}_k\|^2=C\lambda_\eta\sum_{k<N_\eta(\tau_\eta)}\lambda_\eta\|\boldsymbol{r}_k\|^2.$$

The last term converges to zero in probability by (47) and $\lambda_\eta\to 0$. Hence

$$[\mathcal{M}_\eta^{\Psi}]_{N_\eta(\tau_\eta)}=o_{\mathbb{P}}(1).$$

Applying the Lenglart–Rebolledo inequality coordinatewise gives

$$\sup_{n\le N_\eta(\tau_\eta)}\|\mathcal{M}_{\eta,n}^{\Psi}\|=o_{\mathbb{P}}(1). \tag{65}$$

Combining all the estimates above proves (61). □

Lemma F.17 (Vanishing of mixed and second-order $\boldsymbol{v}$-terms)**.**

$$\sum_{k<N_\eta(\tau_\eta)}\left\|\partial^2_{\boldsymbol{\theta}\boldsymbol{v}}\Psi_k[\Delta\boldsymbol{\theta}_{k+1},\Delta\boldsymbol{v}_{k+1}]\right\|=o_{\mathbb{P}}(1),$$

and

$$\sum_{k<N_\eta(\tau_\eta)}\left\|\partial^2_{\boldsymbol{v}\boldsymbol{v}}\Psi_k[\Delta\boldsymbol{v}_{k+1},\Delta\boldsymbol{v}_{k+1}]\right\|=o_{\mathbb{P}}(1).$$

Proof. This is a direct consequence of $\|\Delta\boldsymbol{\theta}_{k+1}\|=\mathcal{O}_{\mathbb{P}}(\eta)$, $\|\Delta\boldsymbol{v}_{k+1}\|=\mathcal{O}_{\mathbb{P}}(\lambda_\eta)$, and the smoothness of Ψ. □

Lemma F.18 (Reduced projected-increment expansion)**.** *For $k < N_\eta(\tau_\eta)$, there exist increments $\Delta\mathcal{E}^{\zeta}_{\eta,k+1}$ such that, in the corrected coordinates,*

$$\Delta\boldsymbol{\zeta}_{k+1} = \underbrace{\partial_{\boldsymbol{\theta}}\Psi_k[-\eta S(\boldsymbol{v}_k)\boldsymbol{z}_k]}_{=:\,\Delta\mathcal{M}^{\zeta}_{\eta,k+1}} + \underbrace{\frac{1}{2}\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi_k\begin{bmatrix}-\eta S(\boldsymbol{v}_k)(\nabla\mathcal{L}(\boldsymbol{\theta}_k)+\boldsymbol{z}_k),\\ -\eta S(\boldsymbol{v}_k)(\nabla\mathcal{L}(\boldsymbol{\theta}_k)+\boldsymbol{z}_k)\end{bmatrix}}_{=:\,\mathcal{Q}^{\zeta}_{\eta,k+1}} + \Delta\mathcal{E}^{\zeta}_{\eta,k+1}. \tag{66}$$

where

$$\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=0}^{n-1}\Delta\mathcal{E}^{\zeta}_{\eta,k+1}\right\| = o_{\mathbb{P}}(1). \tag{67}$$

The same conclusions hold in the original coordinates, after replacing $(\boldsymbol{\theta}_k,\boldsymbol{\zeta}_k)$ by $(\boldsymbol{\theta}^\circ_k,\boldsymbol{\zeta}^\circ_k)$, evaluating every derivative of Ψ at $(\boldsymbol{\theta}^\circ_k,\boldsymbol{v}_k)$, and enlarging the error.

Proof. Start from the second-order Taylor expansion (27) and substitute $\Delta\boldsymbol{\theta}_{k+1} = \delta\boldsymbol{\theta}_{k+1} + d^m_{\eta,k+1}$. The linear $\boldsymbol{\theta}$-term becomes

$$\partial_{\boldsymbol{\theta}}\Psi_k[\Delta\boldsymbol{\theta}_{k+1}] = -\eta\partial_{\boldsymbol{\theta}}\Psi_k[S(\boldsymbol{v}_k)\boldsymbol{z}_k] + \partial_{\boldsymbol{\theta}}\Psi_k[d^m_{\eta,k+1}],$$

because $\partial_{\boldsymbol{\theta}}\Psi_k[S(\boldsymbol{v}_k)\nabla\mathcal{L}(\boldsymbol{\theta}_k)] = \mathbf{0}$. Thus the first term on the right is exactly $\Delta\mathcal{M}^{\zeta}_{\eta,k+1}$. The $\boldsymbol{\theta}\boldsymbol{\theta}$-term is split as

$$\frac{1}{2}\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi_k[\Delta\boldsymbol{\theta}_{k+1},\Delta\boldsymbol{\theta}_{k+1}] = \mathcal{Q}^{\zeta}_{\eta,k+1} + \frac{1}{2}\left(\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi_k[\Delta\boldsymbol{\theta}_{k+1},\Delta\boldsymbol{\theta}_{k+1}] - \partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi_k[\delta\boldsymbol{\theta}_{k+1},\delta\boldsymbol{\theta}_{k+1}]\right).$$

Accordingly, define the discarded increment explicitly by

$$\begin{aligned}\Delta\mathcal{E}^{\zeta}_{\eta,k+1} :=\ & \partial_{\boldsymbol{\theta}}\Psi_k[d^m_{\eta,k+1}] + \frac{1}{2}\left(\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi_k[\Delta\boldsymbol{\theta}_{k+1},\Delta\boldsymbol{\theta}_{k+1}] - \partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi_k[\delta\boldsymbol{\theta}_{k+1},\delta\boldsymbol{\theta}_{k+1}]\right)\\ &+ \partial_{\boldsymbol{v}}\Psi_k[\Delta\boldsymbol{v}_{k+1}] + \partial^2_{\boldsymbol{\theta}\boldsymbol{v}}\Psi_k[\Delta\boldsymbol{\theta}_{k+1},\Delta\boldsymbol{v}_{k+1}] + \frac{1}{2}\partial^2_{\boldsymbol{v}\boldsymbol{v}}\Psi_k[\Delta\boldsymbol{v}_{k+1},\Delta\boldsymbol{v}_{k+1}] + \mathcal{R}^{\Psi}_{\eta,k+1}.\end{aligned}$$

The partial sums of the first term are $o_{\mathbb{P}}(1)$ by (59), and the sum of the norms of the bracketed quadratic difference is $o_{\mathbb{P}}(1)$ by (60). The partial sums of the linear $\boldsymbol{v}$-term are $o_{\mathbb{P}}(1)$ by Lemma F.16; the mixed $\boldsymbol{\theta}\boldsymbol{v}$-term and the quadratic $\boldsymbol{v}\boldsymbol{v}$-term are absolutely summable with total $o_{\mathbb{P}}(1)$ by Lemma F.17. Finally, the Taylor-remainder estimate (28) gives $\sum_k\|\mathcal{R}^{\Psi}_{\eta,k+1}\| = o(1)$. Hence the partial sums of $\Delta\mathcal{E}^{\zeta}_{\eta,k+1}$ satisfy (67), which proves the asserted expansion in the corrected coordinates.

The same decomposition holds in the original coordinates. Set

$$D_{\eta,k} := \boldsymbol{\zeta}^\circ_k - \boldsymbol{\zeta}_k, \qquad \partial^j\Psi^\circ_k := \partial^j\Psi(\boldsymbol{\theta}^\circ_k,\boldsymbol{v}_k).$$

By (50), $\sup_{k<N_\eta(\tau_\eta)}\|D_{\eta,k}\| = \mathcal{O}_{\mathbb{P}}(\eta)$, so the increment $\Delta D_{\eta,k+1}$ has uniformly $\mathcal{O}_{\mathbb{P}}(\eta)$ partial sums. Moreover, the stated smoothness bounds on Ψ and $\sup_k\|Q_{\eta,k}\| = \mathcal{O}_{\mathbb{P}}(\eta)$ give

$$\begin{aligned}&\sup_{n\le N_\eta(\tau_\eta)}\left\|\sum_{k=0}^{n-1}\eta\big(\partial_{\boldsymbol{\theta}}\Psi^\circ_k - \partial_{\boldsymbol{\theta}}\Psi_k\big)[S(\boldsymbol{v}_k)\boldsymbol{z}_k]\right\| = o_{\mathbb{P}}(1),\\ &\sum_{k<N_\eta(\tau_\eta)}\|\mathcal{Q}^{\zeta,\circ}_{\eta,k+1} - \mathcal{Q}^{\zeta}_{\eta,k+1}\| = \mathcal{O}_{\mathbb{P}}(\eta),\end{aligned}$$

where $\mathcal{Q}^{\zeta,\circ}_{\eta,k+1}$ is obtained from $\mathcal{Q}^{\zeta}_{\eta,k+1}$ by replacing $\boldsymbol{\theta}_k$ with $\boldsymbol{\theta}^\circ_k$ and evaluating the derivatives at $(\boldsymbol{\theta}^\circ_k,\boldsymbol{v}_k)$. The first assertion follows from the martingale maximal inequality: its quadratic variation is $\mathcal{O}_{\mathbb{P}}(\eta^2)$. The second follows because the difference per step is $\mathcal{O}_{\mathbb{P}}(\eta^3)$, uniformly on the stopped interval. Consequently, (66)–(67) remain valid, after enlarging the error, with every corrected variable replaced by its original counterpart. □

We now prove the small-window continuity estimate used in the fast-tracking argument. Choose a fixed constant $q_0 > \alpha$, and set

$$R_\eta := \left\lceil q_0 \lambda_\eta^{-1} \log(1/\eta) \right\rceil, \qquad \delta_\eta := R_\eta \eta^2. \tag{68}$$

Since $1 < \alpha < 2$,

$$\delta_\eta = \mathcal{O}\big(\eta^{2-\alpha} \log(1/\eta)\big) = o(1).$$

Lemma F.19 (Small-window continuity of $\boldsymbol{\zeta}_\eta$)**.** *With R_η and δ_η defined in* (68)*,*

$$\omega_\eta(\delta_\eta) := \sup_{\substack{0 \le s \le t \le \tau_\eta \\ t-s \le \delta_\eta}} \|\boldsymbol{\zeta}_\eta(t) - \boldsymbol{\zeta}_\eta(s)\| = o_{\mathbb{P}}(1). \tag{69}$$

The same conclusion holds for the original interpolation $\boldsymbol{\zeta}_\eta^\circ(t)$.

Proof. By Lemma F.18, it suffices to control the leading martingale increments $\Delta\mathcal{M}_{\eta,k+1}^{\boldsymbol{\zeta}}$ and the quadratic increments $\mathcal{Q}_{\eta,k+1}^{\boldsymbol{\zeta}} = \mathcal{O}(\eta^2)$ defined in (66). The jumps of $\mathcal{M}_\eta^{\boldsymbol{\zeta}}$ are bounded by $C\eta$, and its predictable quadratic variation over a slow-time interval of length δ is bounded by $C\delta + o_{\mathbb{P}}(1)$. The number of grid windows is $O(T/\delta_\eta)$. Since $\delta_\eta \to 0$ and $\eta = o(\delta_\eta)$, Freedman's inequality gives

$$\mathbb{P}\left(\sup_{|t-s| \le \delta_\eta} \|\mathcal{M}_\eta^{\boldsymbol{\zeta}}(t) - \mathcal{M}_\eta^{\boldsymbol{\zeta}}(s)\| > \varepsilon \right) \le C\delta_\eta^{-1} \exp\left(-\frac{c\varepsilon^2}{\delta_\eta + \eta\varepsilon} \right) = o(1).$$

The finite-variation part contributes at most $C\delta_\eta + o_{\mathbb{P}}(1)$ on such windows, which also converges to zero. The accumulated error in (67) is already $o_{\mathbb{P}}(1)$ in supremum norm. This proves the claim for the corrected process. Finally,

$$\sup_{t \le \tau_\eta} \|\boldsymbol{\zeta}_\eta^\circ(t) - \boldsymbol{\zeta}_\eta(t)\| = \mathcal{O}_{\mathbb{P}}(\eta)$$

by (50); hence the original modulus is bounded by $\omega_\eta(\delta_\eta) + \mathcal{O}_{\mathbb{P}}(\eta)$, and has the same limit. □

The momentum-corrected coordinates have now served their purpose. In the remainder of this section and thereafter, $(\boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{r})$ again denote the original AGM coordinates, and the superscript $\circ$ is dropped.

F.6.1 Fast tracking of the preconditioner

We next prove that, on the slow time scale, the preconditioner variable is no longer an independent slow state. Under the post-burn-in initialization of Theorem 3.3, it tracks the instantaneous target

$$U(\boldsymbol{\zeta}_\eta(t)).$$

Throughout this subsection we work on the stopped interval $[0, \tau_\eta]$. The stopping will be removed using Corollary F.3.

Recall the memory length R_η, the slow-time window δ_η, and the modulus $\omega_\eta(\delta_\eta)$ defined in (68)-(69). Lemma F.19 gives $\omega_\eta(\delta_\eta) = o_{\mathbb{P}}(1)$. The following lemma states that $\boldsymbol{\zeta}_\eta$ is asymptotically continuous on the memory window of the $\boldsymbol{v}$-equation.

Lemma F.20 (Fast tracking of $\boldsymbol{v}_\eta$)**.** *Assume that*

$$\boldsymbol{v}_\eta(0) - U(\boldsymbol{\zeta}_\eta(0)) = o_{\mathbb{P}}(1), \tag{70}$$

Then

$$\sup_{0 \le t \le \tau_\eta} \|\boldsymbol{v}_\eta(t) - U(\boldsymbol{\zeta}_\eta(t))\| = o_{\mathbb{P}}(1). \tag{71}$$

Proof. For $k < N_\eta(\tau_\eta)$, write

$$\boldsymbol{e}_k := \boldsymbol{v}_k - U(\boldsymbol{\zeta}_k), \qquad \boldsymbol{e}_\eta(t) := \boldsymbol{e}_{\lfloor t/\eta^2 \rfloor}.$$

$$\Delta U_{k+1} := U(\boldsymbol{\zeta}_{k+1}) - U(\boldsymbol{\zeta}_k), \qquad \boldsymbol{D}_k := \bar{U}(\boldsymbol{\theta}_k) - U(\boldsymbol{\zeta}_k).$$

Using

$$\boldsymbol{v}_{k+1} = \boldsymbol{v}_k + \lambda_\eta\big(\bar{U}(\boldsymbol{\theta}_k) - \boldsymbol{v}_k\big) + \Delta M^v_{\eta,k+1},$$

we obtain

$$\boldsymbol{e}_{k+1} = \beta_2 \boldsymbol{e}_k + \lambda_\eta \boldsymbol{D}_k - \Delta U_{k+1} + \Delta M^v_{\eta,k+1}.$$

Iterating this recursion, for every $n \le N_\eta(\tau_\eta)$, we have

$$\boldsymbol{e}_n = \beta_2^n \boldsymbol{e}_0 + \sum_{k=0}^{n-1} \beta_2^{n-1-k} \lambda_\eta \boldsymbol{D}_k - \sum_{k=0}^{n-1} \beta_2^{n-1-k} \Delta U_{k+1} + \sum_{k=0}^{n-1} \beta_2^{n-1-k} \Delta M^v_{\eta,k+1}. \tag{72}$$

Since $\beta_2 < 1$, $\beta_2^n \|\boldsymbol{e}_0\| = o_{\mathbb{P}}(1)$ by the initial condition. We control other three terms on the right-hand side.

Target error. By Corollary F.4,

$$\sup_{n \le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \lambda_\eta \boldsymbol{D}_k \right\| = o_{\mathbb{P}}(1).$$

Applying Lemma E.6 with $\boldsymbol{a}_k = \lambda_\eta \boldsymbol{D}_k$ and $\rho = \beta_2$, we obtain

$$\sup_{n \le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \beta_2^{n-1-k} \lambda_\eta \boldsymbol{D}_k \right\| = o_{\mathbb{P}}(1).$$

Martingale error. Recall that (42) gives $\sup_{0 \le t \le T} \|M^v_\eta(t)\| = o_{\mathbb{P}}(1)$, equivalently,

$$\sup_{n \le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \Delta M^v_{\eta,k+1} \right\| = o_{\mathbb{P}}(1).$$

Another application of Lemma E.6 gives

$$\sup_{n \le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \beta_2^{n-1-k} \Delta M^v_{\eta,k+1} \right\| = o_{\mathbb{P}}(1).$$

Motion of $U(\boldsymbol{\zeta})$. It remains to show that

$$\sup_{n \le N_\eta(\tau_\eta)} \left\| \sum_{k=0}^{n-1} \beta_2^{n-1-k} \Delta U_{k+1} \right\| = o_{\mathbb{P}}(1).$$

Let

$$C_U := \sup_{\boldsymbol{\zeta} \in \Gamma} \|U(\boldsymbol{\zeta})\|, \qquad L_U := \sup_{\boldsymbol{\zeta} \in \Gamma} \|\partial U(\boldsymbol{\zeta})\|.$$

Split the sum into the last R_η terms and the remaining tail. Since $\|\Delta U_{k+1}\| \le 2C_U$, the tail satisfies

$$\sum_{\ell \ge R_\eta} \beta_2^\ell 2C_U \le \frac{2C_U \beta_2^{R_\eta}}{1-\beta_2} \le C \frac{\eta^{q_0}}{\lambda_\eta} = C\eta^{q_0-\alpha} = o(1),$$

where we used $q_0 > \alpha$ in the definition of R_η in (68).

For the last R_η terms, summation by parts gives the deterministic bound

$$\left\| \sum_{k=(n-R_\eta)_+}^{n-1} \beta_2^{n-1-k} \Delta U_{k+1} \right\| \le 2 \sup_{(n-R_\eta)_+ \le j \le n} \|U(\boldsymbol{\zeta}_j) - U(\boldsymbol{\zeta}_n)\| \le 2L_U \omega_\eta(\delta_\eta).$$

Lemma F.19 shows that the last expression is $o_{\mathbb{P}}(1)$, proving the required bound on the motion of $U(\boldsymbol{\zeta})$.

Combining the estimates for the four terms in (72) proves (71). □

Corollary F.3 gives $\mathbb{P}(\tau_\eta < T) \to 0$, so the stopped and unstopped interpolations agree on $[0, T]$ with probability tending to one. Therefore Lemma F.20 yields the corresponding fast-tracking estimate for the original AGM process on every $[\tau, T]$, $0 < \tau < T$. Under the post-burn-in initialization in Theorem 3.3, the matching condition (70) holds, and hence the estimate extends to $[0, T]$, as required by the theorem.

F.7 Local characteristics and identification of the limiting SDE

We identify the local characteristics directly from the reduced-increment expansion in Lemma F.18, specifically (66). Consistent with the convention at the end of Section F.6, all unadorned variables in this section denote the original AGM coordinates. In particular, we use

$$\partial^j \Psi_k := \partial^j \Psi(\boldsymbol{\theta}_k, \boldsymbol{v}_k), \qquad \boldsymbol{z}_k := \nabla \ell_k(\boldsymbol{\theta}_k) - \nabla \mathcal{L}(\boldsymbol{\theta}_k).$$

We first separate the negligible accumulated error in Section F.6 from the part whose conditional characteristics will be used below. Set

$$\mathcal{R}_{\eta,n} := \sum_{k=0}^{n-1} \Delta \mathcal{E}^{\zeta}_{\eta,k+1}, \quad \widetilde{\boldsymbol{\zeta}}_n := \boldsymbol{\zeta}_n - \mathcal{R}_{\eta,n}, \quad \widetilde{\boldsymbol{\zeta}}_\eta(t) := \widetilde{\boldsymbol{\zeta}}_{N_\eta(t \wedge \tau_\eta)}.$$

Recall that (67) gives

$$\sup_{n \le N_\eta(\tau_\eta)} \|\widetilde{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_n\| = \sup_{n \le N_\eta(\tau_\eta)} \|\mathcal{R}_{\eta,n}\| = o_{\mathbb{P}}(1). \tag{73}$$

Thus the two stopped processes have the same tightness and weak-limit behavior.

Let $\mathcal{U}_\Psi$ be an open neighborhood of $\mathcal{X}^{\epsilon_1}$ on which the smooth extension of Ψ furnished by Lemma F.3 is defined. Extend U from Γ to this neighborhood by the same formula

$$U(\boldsymbol{\theta}) := V(\boldsymbol{\Sigma}(\boldsymbol{\theta})).$$

For $(\boldsymbol{\theta}, \boldsymbol{v}) \in \mathcal{U}_\Psi$, write $\boldsymbol{S}_{\boldsymbol{v}} := S(\boldsymbol{v})$ and define

$$\mathfrak{a}(\boldsymbol{\theta}, \boldsymbol{v}) := \partial_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}, \boldsymbol{v}) \boldsymbol{S}_{\boldsymbol{v}} \boldsymbol{\Sigma}(\boldsymbol{\theta}) \boldsymbol{S}_{\boldsymbol{v}} \partial_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}, \boldsymbol{v})^\top, \tag{74}$$

$$\mathfrak{b}(\boldsymbol{\theta}, \boldsymbol{v}) := \frac{1}{2} \partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}, \boldsymbol{v})[\boldsymbol{S}_{\boldsymbol{v}} \boldsymbol{\Sigma}(\boldsymbol{\theta}) \boldsymbol{S}_{\boldsymbol{v}}]. \tag{75}$$

For $\boldsymbol{\zeta} \in \Gamma$, the identity $\Psi(\boldsymbol{\zeta}, \boldsymbol{v}) = \Phi_{\boldsymbol{S}_{\boldsymbol{v}}}(\boldsymbol{\zeta}) = \boldsymbol{\zeta}$ shows that $\mathfrak{a}(\boldsymbol{\zeta}, \boldsymbol{v})$ and $\mathfrak{b}(\boldsymbol{\zeta}, \boldsymbol{v})$ are respectively the covariance and drift of the explicit Itô differential

$$\partial \Phi_{\boldsymbol{S}_{\boldsymbol{v}}}(\boldsymbol{\zeta}) \boldsymbol{S}_{\boldsymbol{v}} \sigma(\boldsymbol{\zeta}) \, \mathrm{d}\boldsymbol{W}_t + \frac{1}{2} \partial^2 \Phi_{\boldsymbol{S}_{\boldsymbol{v}}}(\boldsymbol{\zeta})[\boldsymbol{S}_{\boldsymbol{v}} \boldsymbol{\Sigma}(\boldsymbol{\zeta}) \boldsymbol{S}_{\boldsymbol{v}}] \, \mathrm{d}t$$

. On $\mathcal{U}_\Psi$, define the equilibrium coefficients by

$$\mathfrak{a}_*(\boldsymbol{\theta}) := \mathfrak{a}(\boldsymbol{\theta}, U(\boldsymbol{\theta})), \qquad \mathfrak{b}_*(\boldsymbol{\theta}) := \mathfrak{b}(\boldsymbol{\theta}, U(\boldsymbol{\theta})). \tag{76}$$

These maps are C^3 on $\mathcal{U}_\Psi$ and bounded on the compact working zone.

Lemma F.21 (Coefficient replacement)**.** *Under condition* (70)*,*

$$\int_0^{\tau_\eta} \|\mathfrak{a}(\widetilde{\boldsymbol{\zeta}}_\eta(s), \boldsymbol{v}_\eta(s)) - \mathfrak{a}_*(\widetilde{\boldsymbol{\zeta}}_\eta(s))\| \, \mathrm{d}s = o_{\mathbb{P}}(1), \tag{77}$$

and

$$\int_0^{\tau_\eta} \|\mathfrak{b}(\widetilde{\boldsymbol{\zeta}}_\eta(s), \boldsymbol{v}_\eta(s)) - \mathfrak{b}_*(\widetilde{\boldsymbol{\zeta}}_\eta(s))\| \, \mathrm{d}s = o_{\mathbb{P}}(1). \tag{78}$$

Proof. After localization to a fixed compact neighborhood of the working zone contained in $\mathcal{U}_\Psi$, the maps $\mathfrak{a}$ and $\mathfrak{b}$ are uniformly continuous in $\boldsymbol{v}$. Lemma F.20, the local Lipschitz continuity of U, and (73) imply

$$\sup_{0 \le s \le \tau_\eta} \|\boldsymbol{v}_\eta(s) - U(\widetilde{\boldsymbol{\zeta}}_\eta(s))\| = o_{\mathbb{P}}(1).$$

Uniform continuity then gives convergence of the integrands in probability, uniformly on $[0, \tau_\eta]$. Since the coefficients are bounded and $T < \infty$, dominated convergence in probability yields (77) and (78). □

By the smoothness on $\mathcal{U}_\Psi$ and uniform continuity on a fixed compact neighborhood of $\mathcal{X}^{\epsilon_1}$, replacing $\boldsymbol{\zeta}$ by $\widetilde{\boldsymbol{\zeta}}$ in a slow-time Riemann sum changes it by $o_{\mathbb{P}}(1)$.

Recall from (66) that

$$\Delta\mathcal{M}^{\zeta}_{\eta,k+1} = \partial_{\boldsymbol{\theta}}\Psi_k[-\eta S(\boldsymbol{v}_k)\boldsymbol{z}_k],$$

$$\mathcal{Q}^{\zeta}_{\eta,k+1} = \frac{1}{2}\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi_k[-\eta S(\boldsymbol{v}_k)(\nabla\mathcal{L}(\boldsymbol{\theta}_k)+\boldsymbol{z}_k), -\eta S(\boldsymbol{v}_k)(\nabla\mathcal{L}(\boldsymbol{\theta}_k)+\boldsymbol{z}_k)],$$

then the cleaned increment is exactly

$$\Delta\widetilde{\boldsymbol{\zeta}}_{k+1} = \Delta\mathcal{M}^{\zeta}_{\eta,k+1} + \mathcal{Q}^{\zeta}_{\eta,k+1}.$$

Define

$$b_{\eta,k} := \mathbb{E}_k[\mathcal{Q}^{\zeta}_{\eta,k+1}], \qquad \Delta M_{\eta,k+1} := \Delta\mathcal{M}^{\zeta}_{\eta,k+1} + \mathcal{Q}^{\zeta}_{\eta,k+1} - b_{\eta,k}.$$

The second expression is a martingale difference. On the stopped working zone,

$$\|b_{\eta,k}\| \le C\eta^2, \qquad \|\Delta M_{\eta,k+1}\| \le C\eta.$$

Lemma F.22 (Convergence of the local characteristics)**.** *Uniformly for $n \le N_\eta(\tau_\eta)$,*

$$\sum_{k=0}^{n-1} b_{\eta,k} - \sum_{k=0}^{n-1}\eta^2\mathfrak{b}(\widetilde{\boldsymbol{\zeta}}_k, \boldsymbol{v}_k) = o_{\mathbb{P}}(1), \tag{79}$$

$$\sum_{k=0}^{n-1}\mathbb{E}_k[\Delta M_{\eta,k+1}\Delta M_{\eta,k+1}^\top] - \sum_{k=0}^{n-1}\eta^2\mathfrak{a}(\widetilde{\boldsymbol{\zeta}}_k, \boldsymbol{v}_k) = o_{\mathbb{P}}(1). \tag{80}$$

In fact, the sums of the norms of the corresponding one-step errors converge to zero in probability. Moreover,

$$\sum_{k<N_\eta(\tau_\eta)}\mathbb{E}_k\|\Delta M_{\eta,k+1}\|^3 = \mathcal{O}(\eta).$$

Proof. Conditional centering of $\boldsymbol{z}_k = \boldsymbol{z}_k(\boldsymbol{\theta}_k)$ gives

$$b_{\eta,k} = \frac{\eta^2}{2}\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi_k\big[S(\boldsymbol{v}_k)\{\nabla\mathcal{L}(\boldsymbol{\theta}_k)\nabla\mathcal{L}(\boldsymbol{\theta}_k)^\top + \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)\}S(\boldsymbol{v}_k)\big].$$

The deterministic-gradient part is summably negligible since

$$\sum_{k<N_\eta(\tau_\eta)}\eta^2\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2 \le C\eta\int_0^{\tau_\eta}\|\nabla\mathcal{L}(\boldsymbol{\theta}_\eta(s-))\|^2\,\mathrm{d}A_\eta(s) = C\eta\mathcal{O}_{\mathbb{P}}(1).$$

Recall that $\boldsymbol{\theta}_k = \boldsymbol{\zeta}_k + \boldsymbol{r}_k$, smoothness gives the pointwise bound

$$\|\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi_k - \partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi(\widetilde{\boldsymbol{\zeta}}_k, \boldsymbol{v}_k)\| + \|\boldsymbol{\Sigma}(\boldsymbol{\theta}_k) - \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\zeta}}_k)\| \le C\left(\|\boldsymbol{r}_k\| + \|\widetilde{\boldsymbol{\zeta}}_k - \boldsymbol{\zeta}_k\|\right).$$

Recall that

$$\sup_{k\le N_\eta(\tau_\eta)}\|\boldsymbol{r}_k\| = o_{\mathbb{P}}(1),$$

which follows directly from (45) and (25). Consequently, using $N_\eta(\tau_\eta)\eta^2 \le C$ and (73),

$$\begin{aligned}
&\sum_{k<N_\eta(\tau_\eta)}\|b_{\eta,k} - \frac{\eta^2}{2}\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\Psi(\widetilde{\boldsymbol{\zeta}}_k, \boldsymbol{v}_k)[S(\boldsymbol{v}_k)\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\zeta}}_k)S(\boldsymbol{v}_k)]\| \\
&\quad\le C\sum_{k<N_\eta(\tau_\eta)}\eta^2\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2 + C\sum_{k<N_\eta(\tau_\eta)}\eta^2\left(\|\boldsymbol{r}_k\| + \|\widetilde{\boldsymbol{\zeta}}_k - \boldsymbol{\zeta}_k\|\right) \\
&\quad\le o_{\mathbb{P}}(1) + C\left(\sup_{k\le N_\eta(\tau_\eta)}\|\boldsymbol{r}_k\| + \sup_{k\le N_\eta(\tau_\eta)}\|\widetilde{\boldsymbol{\zeta}}_k - \boldsymbol{\zeta}_k\|\right) = o_{\mathbb{P}}(1).
\end{aligned}$$

This proves (79), including the asserted absolute summability.

Let $\boldsymbol{G}_k := \partial_{\boldsymbol{\theta}} \Psi_k S(\boldsymbol{v}_k)$ and set

$$\boldsymbol{q}_{\eta,k+1} := \mathcal{Q}^{\zeta}_{\eta,k+1} - b_{\eta,k}.$$

By the definition of $b_{\eta,k}$, we have $\mathbb{E}_k[\boldsymbol{q}_{\eta,k+1}] = \mathbf{0}$. Moreover, the boundedness of the noisy gradients and the second derivatives of Ψ on the stopped working zone gives

$$\|\boldsymbol{q}_{\eta,k+1}\| \le C\eta^2.$$

Since $\Delta\mathcal{M}^{\zeta}_{\eta,k+1} = -\eta \boldsymbol{G}_k \boldsymbol{z}_k$, the martingale increment can be written as

$$\Delta M_{\eta,k+1} = -\eta \boldsymbol{G}_k \boldsymbol{z}_k + \boldsymbol{q}_{\eta,k+1}.$$

Expanding its conditional second moment yields

$$\begin{aligned}\mathbb{E}_k[\Delta M_{\eta,k+1}\Delta M_{\eta,k+1}^\top] = \eta^2 \boldsymbol{G}_k \mathbb{E}_k[\boldsymbol{z}_k \boldsymbol{z}_k^\top] \boldsymbol{G}_k^\top - \eta \boldsymbol{G}_k \mathbb{E}_k[\boldsymbol{z}_k \boldsymbol{q}_{\eta,k+1}^\top] \\ - \eta \mathbb{E}_k[\boldsymbol{q}_{\eta,k+1} \boldsymbol{z}_k^\top] \boldsymbol{G}_k^\top + \mathbb{E}_k[\boldsymbol{q}_{\eta,k+1} \boldsymbol{q}_{\eta,k+1}^\top].\end{aligned}$$

Conditional centering of the gradient noise gives $\mathbb{E}_k[\boldsymbol{z}_k \boldsymbol{z}_k^\top] = \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)$. Uniform boundedness, and hence uniform boundedness of the required conditional third moments, gives

$$\|\mathbb{E}_k[\boldsymbol{z}_k \boldsymbol{q}_{\eta,k+1}^\top]\| + \|\mathbb{E}_k[\boldsymbol{q}_{\eta,k+1} \boldsymbol{z}_k^\top]\| \le C\eta^2, \qquad \|\mathbb{E}_k[\boldsymbol{q}_{\eta,k+1} \boldsymbol{q}_{\eta,k+1}^\top]\| \le C\eta^4.$$

Therefore, uniformly for $k < N_\eta(\tau_\eta)$,

$$\mathbb{E}_k[\Delta M_{\eta,k+1}\Delta M_{\eta,k+1}^\top] = \eta^2 \boldsymbol{G}_k \boldsymbol{\Sigma}(\boldsymbol{\theta}_k) \boldsymbol{G}_k^\top + \mathcal{O}(\eta^3).$$

The sum of the $\mathcal{O}(\eta^3)$ terms is $\mathcal{O}(\eta)$. Replacing $\boldsymbol{\theta}_k$ by $\widetilde{\boldsymbol{\zeta}}_k$ in the remaining term has total error bounded by

$$C \sum_{k < N_\eta(\tau_\eta)} \eta^2 \big(\|\boldsymbol{r}_k\| + \|\widetilde{\boldsymbol{\zeta}}_k - \boldsymbol{\zeta}_k\|\big) = o_{\mathbb{P}}(1).$$

This proves (80). Finally, $\mathbb{E}_k\|\Delta M_{\eta,k+1}\|^3 \le C\eta^3$, and there are $\mathcal{O}(\eta^{-2})$ stopped steps, which proves the third-moment assertion. □

F.7.1 Martingale problem and identification of the limiting SDE

We now identify the weak limit of the projected process. In this subsection it is important to distinguish the stopped projected process from the cleaned process. Following Definition F.6, let

$$\boldsymbol{\zeta}^\tau_\eta(t) := \boldsymbol{\zeta}_\eta(t \wedge \tau_\eta) = \Psi\big(\boldsymbol{\theta}_\eta(t \wedge \tau_\eta), \boldsymbol{v}_\eta(t \wedge \tau_\eta)\big), \qquad 0 \le t \le T,$$

be the original-coordinate projected interpolation, stopped and then held constant at τ_η. Let $\widetilde{\boldsymbol{\zeta}}_\eta$ denote the cleaned stopped interpolation defined above. By (73),

$$\sup_{0 \le t \le T} \|\widetilde{\boldsymbol{\zeta}}_\eta(t) - \boldsymbol{\zeta}^\tau_\eta(t)\| = o_{\mathbb{P}}(1). \tag{81}$$

The semimartingale calculation is performed for $\widetilde{\boldsymbol{\zeta}}_\eta$, and all path-space conclusions are then transferred to $\boldsymbol{\zeta}^\tau_\eta$ using (81).

For $f \in C^3(\Gamma)$, choose an arbitrary C^3 extension, still denoted by f, to the fixed tubular working zone. Define

$$\mathcal{A}f(\boldsymbol{\zeta}, \boldsymbol{v}) := \nabla f(\boldsymbol{\zeta})^\top \mathfrak{b}(\boldsymbol{\zeta}, \boldsymbol{v}) + \frac{1}{2} \operatorname{Tr}\left(\mathfrak{a}(\boldsymbol{\zeta}, \boldsymbol{v}) \nabla^2 f(\boldsymbol{\zeta})\right),$$

and the limiting generator

$$\mathcal{A}_* f(\boldsymbol{\zeta}) := \mathcal{A}f(\boldsymbol{\zeta}, U(\boldsymbol{\zeta})). \tag{82}$$

The definition of $\mathcal{A}_* f$ on Γ is independent of the extension of f. Indeed, the coefficients in (74)-(76) are obtained by applying Itô's formula to the gradient-flow projection. Equivalently, $\mathcal{A}_*$ annihilates every C^3 function that vanishes on Γ.

Lemma F.23 (One-step martingale decomposition)**.** *Let $n \le N_\eta(\tau_\eta)$, and let $b_{\eta,k}$, $\Delta M_{\eta,k+1}$ be defined as above. There is a martingale $\mathcal{N}^f_{\eta,n}$ such that*

$$f(\widetilde{\boldsymbol{\zeta}}_n) - f(\widetilde{\boldsymbol{\zeta}}_0) = \mathcal{N}^f_{\eta,n} + \sum_{k=0}^{n-1} \eta^2 \mathcal{A} f(\widetilde{\boldsymbol{\zeta}}_k, \boldsymbol{v}_k) + \rho^f_{\eta,n}, \tag{83}$$

where

$$\sup_{n \le N_\eta(\tau_\eta)} |\rho^f_{\eta,n}| = o_{\mathbb{P}}(1).$$

After localization, we also have

$$\mathbb{E}\left[\sup_{n \le N_\eta(\tau_\eta)} |\rho^f_{\eta,n}|\right] = o(1).$$

Proof. Let

$$C_{\eta,k} := \mathbb{E}_k[\Delta M_{\eta,k+1} \Delta M_{\eta,k+1}^\top].$$

Recall that by definition, $\Delta\widetilde{\boldsymbol{\zeta}}_{k+1} = \Delta M_{\eta,k+1} + b_{\eta,k}$. Taylor expansion of f over one step gives

$$f(\widetilde{\boldsymbol{\zeta}}_{k+1}) - f(\widetilde{\boldsymbol{\zeta}}_k) = \nabla f(\widetilde{\boldsymbol{\zeta}}_k)^\top \Delta M_{\eta,k+1} + \nabla f(\widetilde{\boldsymbol{\zeta}}_k)^\top b_{\eta,k} + \frac{1}{2} \Delta M_{\eta,k+1}^\top \nabla^2 f(\widetilde{\boldsymbol{\zeta}}_k) \Delta M_{\eta,k+1} + r^f_{\eta,k+1}.$$

The terms involving $b_{\eta,k}$ in the quadratic part have been included in $r^f_{\eta,k+1}$, which is bounded by

$$\sum_{k < N_\eta(\tau_\eta)} |r^f_{\eta,k+1}| \le C N_\eta(\tau_\eta) \eta^3 = \mathcal{O}(\eta).$$

Define

$$\Delta\mathcal{N}^f_{\eta,k+1} := \nabla f(\widetilde{\boldsymbol{\zeta}}_k)^\top \Delta M_{\eta,k+1} + \frac{1}{2}\left\{\Delta M_{\eta,k+1}^\top \nabla^2 f(\widetilde{\boldsymbol{\zeta}}_k) \Delta M_{\eta,k+1} - \mathrm{Tr}\big(C_{\eta,k} \nabla^2 f(\widetilde{\boldsymbol{\zeta}}_k)\big)\right\}.$$

This is a martingale difference. Summing the Taylor expansion and using the absolute-summability part of Lemma F.22 proves (83) and the displayed remainder estimate. All errors are bounded on the stopped working zone. Their convergence in probability therefore upgrades to L^1 after the same localization used in the preceding subsection. □

Lemma F.24 (Asymptotic martingale problem)**.** *Under condition* (70)*, for every $f \in C^3(\Gamma)$, let*

$$\mathcal{M}^f_\eta(t) := f(\boldsymbol{\zeta}^\tau_\eta(t)) - f(\boldsymbol{\zeta}^\tau_\eta(0)) - \int_0^{t \wedge \tau_\eta} \mathcal{A}_* f(\boldsymbol{\zeta}^\tau_\eta(s))\, \mathrm{d}s, \qquad t \in [0, T].$$

For every $0 \le s < t \le T$, every $0 \le t_1 < \cdots < t_q \le s$, and every bounded continuous $g : (\mathbb{R}^d)^q \to \mathbb{R}$,

$$\mathbb{E}\left[\big(\mathcal{M}^f_\eta(t) - \mathcal{M}^f_\eta(s)\big) g\big(\boldsymbol{\zeta}^\tau_\eta(t_1), \ldots, \boldsymbol{\zeta}^\tau_\eta(t_q)\big)\right] = o(1). \tag{84}$$

Proof. Apply Lemma F.23 between the grid indices associated with $s \wedge \tau_\eta$ and $t \wedge \tau_\eta$. Since all coefficients are bounded and continuous on the stopped working zone,

$$\sum_k \eta^2 \mathcal{A} f(\widetilde{\boldsymbol{\zeta}}_k, \boldsymbol{v}_k)$$

is the corresponding slow-time Riemann sum up to an error $\mathcal{O}(\eta^2)$. Lemma F.21, together with (81), gives

$$\int_{s \wedge \tau_\eta}^{t \wedge \tau_\eta} \left| \mathcal{A} f(\widetilde{\boldsymbol{\zeta}}_\eta(u), \boldsymbol{v}_\eta(u)) - \mathcal{A}_* f(\boldsymbol{\zeta}^\tau_\eta(u)) \right| \mathrm{d}u = o_{\mathbb{P}}(1).$$

The integrand is bounded, so this convergence also holds in L^1. Likewise, (81) and the bounded derivatives of f allow $\widetilde{\boldsymbol{\zeta}}_\eta$ to be replaced by $\boldsymbol{\zeta}^\tau_\eta$ in the endpoint terms. For simplicity, let

$$n_r := N_\eta(r \wedge \tau_\eta), \qquad G_\eta := g\big(\boldsymbol{\zeta}^\tau_\eta(t_1), \ldots, \boldsymbol{\zeta}^\tau_\eta(t_q)\big).$$

Lemma F.23 implies

$$f(\widetilde{\boldsymbol{\zeta}}_{n_t}) - f(\widetilde{\boldsymbol{\zeta}}_{n_s}) = \mathcal{N}^f_{\eta,n_t} - \mathcal{N}^f_{\eta,n_s} + \sum_{k=n_s}^{n_t-1} \eta^2 \mathcal{A}f(\widetilde{\boldsymbol{\zeta}}_k, \boldsymbol{v}_k) + \rho^f_{\eta,n_t} - \rho^f_{\eta,n_s}.$$

Consequently,

$$\mathcal{M}^f_\eta(t) - \mathcal{M}^f_\eta(s) = \mathcal{N}^f_{\eta,n_t} - \mathcal{N}^f_{\eta,n_s} + \varepsilon^f_{\eta;s,t},$$

where

$$\begin{aligned}\varepsilon^f_{\eta;s,t} := \big[f(\boldsymbol{\zeta}^\tau_\eta(t)) - f(\widetilde{\boldsymbol{\zeta}}_{n_t})\big] - \big[f(\boldsymbol{\zeta}^\tau_\eta(s)) - f(\widetilde{\boldsymbol{\zeta}}_{n_s})\big] + \rho^f_{\eta,n_t} - \rho^f_{\eta,n_s} \\ + \sum_{k=n_s}^{n_t-1} \eta^2 \mathcal{A}f(\widetilde{\boldsymbol{\zeta}}_k, \boldsymbol{v}_k) - \int_{s\wedge\tau_\eta}^{t\wedge\tau_\eta} \mathcal{A}_* f(\boldsymbol{\zeta}^\tau_\eta(u))\,\mathrm{d}u.\end{aligned}$$

The endpoint comparison, the L^1 remainder estimate in Lemma F.23, the coefficient-replacement estimate, and the two boundary intervals of length at most η^2 give

$$\begin{aligned}\mathbb{E}|\varepsilon^f_{\eta;s,t}| \le C\mathbb{E}\left[1 \wedge \sup_{0\le r\le T} \|\widetilde{\boldsymbol{\zeta}}_\eta(r) - \boldsymbol{\zeta}^\tau_\eta(r)\|\right] + 2\mathbb{E}\left[\sup_{n\le N_\eta(\tau_\eta)} |\rho^f_{\eta,n}|\right] \\ + \mathbb{E}\int_{s\wedge\tau_\eta}^{t\wedge\tau_\eta} \Big|\mathcal{A}f(\widetilde{\boldsymbol{\zeta}}_\eta(u), \boldsymbol{v}_\eta(u)) - \mathcal{A}_* f(\boldsymbol{\zeta}^\tau_\eta(u))\Big|\,\mathrm{d}u + C\eta^2 = o(1).\end{aligned}$$

Since $t_q \le s$, G_η is $\mathcal{F}_{n_s}$-measurable. Optional sampling for the discrete martingale $\mathcal{N}^f_{\eta,n}$ therefore yields

$$\mathbb{E}\left[(\mathcal{N}^f_{\eta,n_t} - \mathcal{N}^f_{\eta,n_s})G_\eta\right] = 0.$$

It follows that

$$\left|\mathbb{E}\left[(\mathcal{M}^f_\eta(t) - \mathcal{M}^f_\eta(s))G_\eta\right]\right| \le \|g\|_\infty \mathbb{E}|\varepsilon^f_{\eta;s,t}| = o(1),$$

which proves (84). □

We recall the tightness criterion used below.

Theorem F.1 (Aldous' tightness criterion)**.** *Let E be a Polish space with metric d_E, and let $\{X_n\}$ be a sequence of E-valued adapted càdlàg processes on $[0,T]$. Assume the compact-containment condition: for every $\varepsilon > 0$, there is a compact set $K_\varepsilon \subset E$ such that*

$$\inf_n \mathbb{P}\big(X_n(t) \in K_\varepsilon \text{ for all } 0 \le t \le T\big) \ge 1 - \varepsilon.$$

Assume also that, for every $\varepsilon > 0$,

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \sup_{\sigma_n\le\tau_n\le(\sigma_n+\delta)\wedge T} \mathbb{P}\big(d_E(X_n(\tau_n), X_n(\sigma_n)) > \varepsilon\big) = 0,$$

where the supremum is over stopping times σ_n and τ_n. Then $\{X_n\}$ is tight in $D([0,T],E)$. If, in addition, the maximum jump size converges to zero in probability, then the sequence is C-tight, meaning that every weak limit is supported on $C([0,T],E)$. For grid-based step processes, it is enough to verify the stopping-time condition for grid-valued stopping times.

Lemma F.25 (Tightness of the projected process)**.** *The sequence $\{\boldsymbol{\zeta}^\tau_\eta\}_\eta$ is tight in $D([0,T],\Gamma)$, and every limit point has continuous paths.*

Proof. We first prove C-tightness of the cleaned process in $D([0,T],\mathbb{R}^d)$. Recall that

$$\widetilde{\boldsymbol{\zeta}}_n - \widetilde{\boldsymbol{\zeta}}_0 = \sum_{k=0}^{n-1} \Delta M_{\eta,k+1} + \sum_{k=0}^{n-1} b_{\eta,k}.$$

Let $\kappa_\eta \le \ell_\eta \le N_\eta(\tau_\eta)$ be bounded grid-valued stopping times such that $(\ell_\eta - \kappa_\eta)\eta^2 \le \delta + \eta^2$. Recall that $b_{\eta,k} = \mathcal{O}(\eta^2)$, so the drift term is bounded by

$$\left\|\sum_{k=\kappa_\eta}^{\ell_\eta-1} b_{\eta,k}\right\| \le C(\delta + \eta^2).$$

Moreover, martingale orthogonality and the bound $\mathbb{E}_k[\Delta M_{\eta,k+1}\Delta M_{\eta,k+1}^\top] \preceq C\eta^2 \boldsymbol{I}$ give

$$\mathbb{E}\left[\left\|\sum_{k=\kappa_\eta}^{\ell_\eta-1}\Delta M_{\eta,k+1}\right\|^2\right] \le C(\delta+\eta^2).$$

Fix $\varepsilon > 0$. For all sufficiently small δ and η, the drift bound is at most $\varepsilon/2$, and Chebyshev's inequality gives

$$\begin{aligned}\mathbb{P}\left(\|\widetilde{\boldsymbol{\zeta}}_{\ell_\eta} - \widetilde{\boldsymbol{\zeta}}_{\kappa_\eta}\| > \varepsilon\right) &\le \mathbb{P}\left(\left\|\sum_{k=\kappa_\eta}^{\ell_\eta-1}\Delta M_{\eta,k+1}\right\| > \frac{\varepsilon}{2}\right)\\ &\le \frac{4}{\varepsilon^2}\mathbb{E}\left[\left\|\sum_{k=\kappa_\eta}^{\ell_\eta-1}\Delta M_{\eta,k+1}\right\|^2\right]\\ &\le \frac{C}{\varepsilon^2}(\delta+\eta^2).\end{aligned}$$

Therefore

$$\lim_{\delta\downarrow 0}\limsup_{\eta\to 0}\sup_{\substack{\kappa_\eta\le\ell_\eta\le N_\eta(\tau_\eta)\\(\ell_\eta-\kappa_\eta)\eta^2\le\delta+\eta^2}}\mathbb{P}\left(\|\widetilde{\boldsymbol{\zeta}}_{\ell_\eta} - \widetilde{\boldsymbol{\zeta}}_{\kappa_\eta}\| > \varepsilon\right) = 0,$$

which is the stopping-time modulus condition in Theorem F.1. Also,

$$\sup_{k<N_\eta(\tau_\eta)}\|\Delta\widetilde{\boldsymbol{\zeta}}_{k+1}\| \le C\eta \longrightarrow 0.$$

Compact containment follows from (81) and compactness of Γ. Thus the cleaned stopped processes are C-tight.

Finally, (81) transfers C-tightness to $\boldsymbol{\zeta}_\eta^\tau$. Since $\boldsymbol{\zeta}_\eta^\tau$ takes values in the compact manifold Γ, this is tightness in $D([0,T],\Gamma)$, and all limit points are continuous. Corollary F.3 transfers the same conclusion to the unstopped projected interpolation. □

Proposition F.1 (Identification of the limit)**.** *Consider the restarted process in Section F.2 and suppose*

$$\mathrm{Law}(\boldsymbol{\zeta}_\eta^\tau(0)) \Rightarrow \mu, \qquad \mu(\Gamma) = 1,$$

where $\mathrm{Law}(X)(A) := \mathbb{P}(X \in A)$ *for a random variable* X *and a measurable set* A*. Then every subsequential limit solves the* $\mathcal{A}_*$ *martingale problem with initial distribution* μ*, and hence is a weak solution of* (3)*.*

Proof. By Lemma F.7, we may condition on the restarted initial state lying in the working zone. This event has probability $1 - o(1)$, so removing the conditioning changes bounded expectations by $o(1)$ and preserves the limiting initial distribution. Lemma F.25 gives, along a subsequence,

$$\boldsymbol{\zeta}_\eta^\tau \Rightarrow \boldsymbol{\zeta}, \qquad \boldsymbol{\zeta} \in C([0,T],\Gamma), \qquad \mathrm{Law}(\boldsymbol{\zeta}(0)) = \mu.$$

For $f \in C^3(\Gamma)$, Corollary F.3 gives

$$\mathbb{E}\left|\int_{t\wedge\tau_\eta}^{t}\mathcal{A}_* f(\boldsymbol{\zeta}_\eta^\tau(u))\,\mathrm{d}u\right| \le T\|\mathcal{A}_* f\|_\infty \mathbb{P}(\tau_\eta < T) = o(1).$$

Thus Lemma F.24 and continuity of the limiting paths imply, with

$$\mathcal{M}^f(t) := f(\boldsymbol{\zeta}(t)) - f(\boldsymbol{\zeta}(0)) - \int_0^t \mathcal{A}_* f(\boldsymbol{\zeta}(u))\,\mathrm{d}u,$$

that, for $0 \le t_1 < \cdots < t_q \le s < t \le T$ and bounded continuous g,

$$\mathbb{E}\left[(\mathcal{M}^f(t) - \mathcal{M}^f(s))g(\boldsymbol{\zeta}(t_1),\ldots,\boldsymbol{\zeta}(t_q))\right] = 0.$$

Boundedness of f, $\mathcal{A}_* f$, and g justifies passage to expectations. A monotone class argument proves the martingale property in the natural filtration, and the drift and covariance formulas identify (3). □

F.8 Completion of the proof of the main theorem

Proof of theorem 3.3. Use the restart at K_0 from Section F.2. Lemma F.7 gives

$$\operatorname{dist}(\boldsymbol{\theta}_\eta(0), \Gamma) = o_{\mathbb{P}}(1), \qquad \boldsymbol{v}_\eta(0) - U(\boldsymbol{\zeta}_\eta(0)) = o_{\mathbb{P}}(1).$$

The smooth coefficients of (3) admit globally Lipschitz extensions, and the projection identities preserve Γ as in Lemma E.13. The SDE is therefore unique in law, with continuous dependence on the initial distribution. Choose the extensions to vanish near the existing point $\boldsymbol{\theta}_{\mathrm{null}}$ and use $U = V(\boldsymbol{\Sigma})$ there. This defines the comparison process when the initial projection fails, an event of probability $o(1)$.

Given any sequence $\eta_n \to 0$, compactness allows a subsequence such that

$$\operatorname{Law}(\boldsymbol{\zeta}_{\eta_n}(0)) \Rightarrow \mu, \qquad \mu(\Gamma) = 1.$$

Let $\boldsymbol{\zeta}^\mu$ solve (3) with initial distribution μ and set $\boldsymbol{X}^\mu = (\boldsymbol{\zeta}^\mu, U(\boldsymbol{\zeta}^\mu))$. Lemma F.25, Proposition F.1, and uniqueness identify the projected limit. Corollary F.3 and Lemma F.20 give

$$\mathbb{P}(\tau_{\eta_n} < T) = o(1), \qquad \sup_{0 \le t \le T} \|\boldsymbol{v}_{\eta_n}(t) - U(\boldsymbol{\zeta}_{\eta_n}(t))\| = o_{\mathbb{P}}(1).$$

Hence, using $\bar{\boldsymbol{X}}_k$ from the theorem,

$$\bar{\boldsymbol{X}}_{N_{\eta_n}(\cdot)} \Rightarrow \boldsymbol{X}^\mu \quad \text{in } D([0,T], \mathbb{R}^{d+D}).$$

Take the comparison Brownian motion independent of the information at K_0. The comparison SDE has the same projected initial distribution, so continuous dependence gives, along the same subsequence,

$$\boldsymbol{X}(\cdot) \Rightarrow \boldsymbol{X}^\mu(\cdot) \quad \text{in } C([0,T], \mathbb{R}^{d+D}),$$

where $\boldsymbol{X}$ corresponds to η_n. Both processes lie in $(\Gamma \cup \{\boldsymbol{\theta}_{\mathrm{null}}\}) \times [0, R_1]^D$, so g is bounded on their ranges.

If the claimed maximum does not tend to zero, there exist $\varepsilon > 0$ and indices $0 \le k_n \le K_{\eta_n}$ such that

$$\left|\mathbb{E}[g(\bar{\boldsymbol{X}}_{k_n})] - \mathbb{E}[g(\boldsymbol{X}(k_n \eta_n^2))]\right| \ge \varepsilon.$$

Pass to a subsequence as above with $k_n \eta_n^2 \to t \in [0,T]$. Continuity of the limiting paths and boundedness of g yield

$$\mathbb{E}[g(\bar{\boldsymbol{X}}_{k_n})] \longrightarrow \mathbb{E}[g(\boldsymbol{X}^\mu(t))], \qquad \mathbb{E}[g(\boldsymbol{X}(k_n \eta_n^2))] \longrightarrow \mathbb{E}[g(\boldsymbol{X}^\mu(t))].$$

This contradiction proves the approximation uniformly over the time grid. □

G Implicit bias under label noise

In this section, we present Adam's implicit bias under the corresponding regimes and label noise conditions, following the label noise setting of Li et al. (2025). All points on Γ minimize the training loss, but they can have different curvatures and different noise statistics. Implicit bias describes how the remaining slow motion distinguishes between these minima. We first recall the definition of the *label noise* condition in Blanc et al. (2020) and Li et al. (2025).

Label Noise Condition. For all $\boldsymbol{\theta} \in \Gamma$, the covariance matrix $\boldsymbol{\Sigma}$ satisfies $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \tau \nabla^2 \mathcal{L}(\boldsymbol{\theta})$ for some constant $\tau > 0$.

Let $\boldsymbol{H}_{\boldsymbol{\zeta}} := \nabla^2 \mathcal{L}(\boldsymbol{\zeta})$ and

$$\mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}) := \begin{cases} \operatorname{tr}\Big(\operatorname{Diag}(\boldsymbol{H}_{\boldsymbol{\zeta}})^{1/2} - \frac{\epsilon}{\sqrt{\tau}} \log\Big(\frac{\sqrt{\tau}}{\epsilon}\operatorname{Diag}(\boldsymbol{H}_{\boldsymbol{\zeta}})^{1/2} + \boldsymbol{I}\Big)\Big), & \epsilon > 0, \\ \operatorname{tr}\big(\operatorname{Diag}(\boldsymbol{H}_{\boldsymbol{\zeta}})^{1/2}\big), & \epsilon = 0. \end{cases}$$

The quantity $\mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta})$ measures a form of sharpness relevant to Adam's diagonal preconditioner. For the case $\epsilon = 0$, it sums the square roots of diagonal Hessian entries.

For the Adam specialization used below, we let

$$V(\boldsymbol{M}) = \operatorname{diag}(\boldsymbol{M}), \qquad S(\boldsymbol{v}) = \operatorname{Diag}\left(\frac{1}{\sqrt{\boldsymbol{v}} + \epsilon \mathbf{1}}\right), \qquad \epsilon \ge 0,$$

where $\operatorname{diag}(\boldsymbol{M})$ denotes the vector consisting of its diagonal entries for a square matrix $\boldsymbol{M}$ and $\operatorname{Diag}(\boldsymbol{M}) := \operatorname{Diag}(\operatorname{diag}(\boldsymbol{M}))$ where $\operatorname{Diag}(\boldsymbol{v})$ denotes the diagonal matrix with $\boldsymbol{v}$ on its diagonal for a vector $\boldsymbol{v}$.

G.1 THE 1-SCHEME

Assume that $1-\beta_2 = c_\beta \eta$. The next definition collects the quantities needed to describe the joint effect of $(\boldsymbol{\theta}, \boldsymbol{v})$.

Definition G.1. *For every $\boldsymbol{\zeta} \in \Gamma$, define*

$$\boldsymbol{J}_{\boldsymbol{\zeta}} := \partial_{\boldsymbol{\theta}} \pi(\iota(\boldsymbol{\zeta})) = \lim_{s\to\infty} \exp\left(-s \boldsymbol{S}_*(\boldsymbol{\zeta}) \boldsymbol{H}_{\boldsymbol{\zeta}}\right)$$

and

$$\boldsymbol{M}_{\boldsymbol{\zeta}} := \mathbb{E}_{\xi\sim\mathcal{S}}\left[\nabla \ell(\boldsymbol{\zeta};\xi)\big(\boldsymbol{h}(\boldsymbol{\zeta},\xi) - U(\boldsymbol{\zeta})\big)^\top\right].$$

Define $\boldsymbol{Y}_{\boldsymbol{\zeta}}$ by

$$(\boldsymbol{H}_{\boldsymbol{\zeta}} \boldsymbol{S}_*(\boldsymbol{\zeta}) + c_\beta \boldsymbol{I}) \boldsymbol{Y}_{\boldsymbol{\zeta}} = c_\beta^2 \boldsymbol{M}_{\boldsymbol{\zeta}} + \frac{\tau c_\beta}{2} \boldsymbol{H}_{\boldsymbol{\zeta}} \boldsymbol{S}_*(\boldsymbol{\zeta}) \partial \bar{U}(\boldsymbol{\zeta})^\top, \tag{85}$$

where $\boldsymbol{H}_{\boldsymbol{\zeta}} \boldsymbol{S}_*(\boldsymbol{\zeta}) + c_\beta \boldsymbol{I}$ is invertible because $\boldsymbol{H}_{\boldsymbol{\zeta}} \boldsymbol{S}_*(\boldsymbol{\zeta})$ has nonnegative real eigenvalues.

$\boldsymbol{J}_{\boldsymbol{\zeta}}$ describes which parameter displacements survive the flow toward the manifold. If we perturb the parameter slightly near $\boldsymbol{\zeta}$, the attracting flow removes the stable component of that perturbation. Applying $\boldsymbol{J}_{\boldsymbol{\zeta}}$ gives the remaining displacement tangent to Γ. $\boldsymbol{M}_{\boldsymbol{\zeta}}$ measures the correlation between the noise of $\boldsymbol{\theta}$ and the noise of $\boldsymbol{v}$. For Adam, it involves third moments of the gradient noise, so it contains information beyond the covariance specified by the label-noise condition. $\boldsymbol{Y}_{\boldsymbol{\zeta}}$ collects the effective part of the joint correlation that contributes to the additional drift in the 1-scheme. Now we give the following formulated result of the ODE and implicit bias of Adam under the label noise condition.

Theorem G.1. *Under the label noise condition, the Slow SDE for AGMs in equation* (2) *becomes the following ODE:*

$$\dot{\boldsymbol{\zeta}}(t) = -\frac{\tau}{4} \boldsymbol{J}_{\boldsymbol{\zeta}(t)} \boldsymbol{S}_*(\boldsymbol{\zeta}(t)) \nabla^3 \mathcal{L}(\boldsymbol{\zeta}(t))[\boldsymbol{S}_*(\boldsymbol{\zeta}(t))] - \boldsymbol{J}_{\boldsymbol{\zeta}(t)} \sum_{i=1}^{D} \partial S(U(\boldsymbol{\zeta}(t)))[\boldsymbol{e}_i](\boldsymbol{Y}_{\boldsymbol{\zeta}(t)})_{\cdot i}. \tag{86}$$

Lemma G.1. *Under the label noise condition, let $\boldsymbol{\delta}_{\epsilon,\boldsymbol{\zeta}}$ such that*

$$(\boldsymbol{\delta}_{\epsilon,\boldsymbol{\zeta}})_i := \begin{cases} \frac{(\boldsymbol{Y}_{\boldsymbol{\zeta}})_{ii}}{\sqrt{U_i(\boldsymbol{\zeta})}\left(\sqrt{U_i(\boldsymbol{\zeta})}+\epsilon\right)}, & U_i(\boldsymbol{\zeta}) > 0, \\ 0, & U_i(\boldsymbol{\zeta}) = 0. \end{cases}$$

Then the Adam specialization of (86) *agrees on $\Gamma_+ := \{\boldsymbol{\zeta} \in \Gamma : U_i(\boldsymbol{\zeta}) > 0, i = 1, \ldots, d\}$ with the following ODE on Γ:*

$$\dot{\boldsymbol{\zeta}} = -\frac{\tau}{4} \boldsymbol{J}_{\boldsymbol{\zeta}} \boldsymbol{S}_*(\boldsymbol{\zeta}) \nabla^3 \mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{S}_*(\boldsymbol{\zeta})] + \frac{1}{2} \boldsymbol{J}_{\boldsymbol{\zeta}} \boldsymbol{S}_*(\boldsymbol{\zeta}) \boldsymbol{\delta}_{\epsilon,\boldsymbol{\zeta}}. \tag{87}$$

Every fixed point $\boldsymbol{\zeta}^*$ of Equation (87) *satisfies $\nabla_\Gamma \mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}^*) = \frac{1}{\sqrt{\tau}} \operatorname{proj}_{T_{\boldsymbol{\zeta}^*}\Gamma} \boldsymbol{\delta}_{\epsilon,\boldsymbol{\zeta}^*}$, where ∇_Γ denotes the projection onto Γ of the gradient and $\operatorname{proj}_{T_{\boldsymbol{\zeta}^*}\Gamma}$ denotes the projection onto $T_{\boldsymbol{\zeta}^*}\Gamma$.*

G.2 PROOF FOR THE 1-SCHEME

Proof of Theorem G.1. Fix $\boldsymbol{\zeta} \in \Gamma$. For simplicity, let

$$c := c_\beta, \quad \boldsymbol{H} := \boldsymbol{H}_{\boldsymbol{\zeta}}, \quad \boldsymbol{S} := \boldsymbol{S}_*(\boldsymbol{\zeta}), \quad \boldsymbol{C} := \boldsymbol{C}_{\boldsymbol{\zeta}}, \quad \boldsymbol{J} := \boldsymbol{J}_{\boldsymbol{\zeta}}, \quad \boldsymbol{M} := \boldsymbol{M}_{\boldsymbol{\zeta}}, \quad \boldsymbol{Y} := \boldsymbol{Y}_{\boldsymbol{\zeta}}.$$

Recall that by Corollary E.2, $\partial\pi(\iota(\boldsymbol{\zeta}))[\boldsymbol{u}, \boldsymbol{w}] = \boldsymbol{J}\boldsymbol{u}$. Therefore the Brownian term in (2) is $\boldsymbol{J}\boldsymbol{S}\sigma(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}_t$. If $\boldsymbol{u} \in T_{\boldsymbol{\zeta}}\Gamma = \ker \boldsymbol{H}$, then

$$\mathbb{E}[(\boldsymbol{u}^\top \nabla \ell(\boldsymbol{\zeta};\xi))^2] = \boldsymbol{u}^\top \boldsymbol{\Sigma}(\boldsymbol{\zeta}) \boldsymbol{u} = \tau \boldsymbol{u}^\top \boldsymbol{H} \boldsymbol{u} = 0.$$

Hence $\nabla \ell(\boldsymbol{\zeta};\xi) \in \operatorname{Ran}(\boldsymbol{H})$ almost surely. Since

$$\boldsymbol{J}\boldsymbol{S}\boldsymbol{H} = \lim_{s\to\infty} \exp(-s\boldsymbol{S}\boldsymbol{H})\boldsymbol{S}\boldsymbol{H} = \boldsymbol{O},$$

this gives $\boldsymbol{J}\boldsymbol{S}\nabla\ell(\boldsymbol{\zeta};\xi) = \boldsymbol{0}$ almost surely. Since $\boldsymbol{\Sigma}(\boldsymbol{\zeta}) = \sigma(\boldsymbol{\zeta})\sigma(\boldsymbol{\zeta})^\top$, we have

$$\big(\boldsymbol{J}\boldsymbol{S}\sigma(\boldsymbol{\zeta})\big)\big(\boldsymbol{J}\boldsymbol{S}\sigma(\boldsymbol{\zeta})\big)^\top = \boldsymbol{J}\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\zeta})\boldsymbol{S}\boldsymbol{J}^\top = \boldsymbol{O}.$$

Thus the Brownian term in (2) vanishes.

It remains to identify the deterministic drift. By Lemma E.2, the joint linearization at $\iota(\boldsymbol{\zeta})$ is

$$\boldsymbol{A} = \begin{pmatrix} -\boldsymbol{S}\boldsymbol{H} & \boldsymbol{O} \\ c\partial\bar{U}(\boldsymbol{\zeta}) & -c\boldsymbol{I} \end{pmatrix}.$$

Let $\boldsymbol{Q}$ be the symmetric matrix such that

$$\boldsymbol{A}\boldsymbol{Q} + \boldsymbol{Q}\boldsymbol{A}^\top = -\boldsymbol{C}.$$

In fact, we can let

$$\boldsymbol{Q} := \begin{pmatrix} \boldsymbol{Q}_{\boldsymbol{\theta}\boldsymbol{\theta}} & \boldsymbol{Q}_{\theta v} \\ \boldsymbol{Q}_{\theta v}^\top & \boldsymbol{Q}_{\boldsymbol{v}\boldsymbol{v}} \end{pmatrix}$$

with

$$\boldsymbol{Q}_{\boldsymbol{\theta}\boldsymbol{\zeta}} = \frac{\tau}{2}\boldsymbol{S}, \qquad \boldsymbol{Q}_{\theta v} := c(\boldsymbol{S}\boldsymbol{H} + c\boldsymbol{I})^{-1}\boldsymbol{S}\left(\frac{\tau}{2}\partial\bar{U}(\boldsymbol{\zeta})^\top - \boldsymbol{M}\right),$$

$$\boldsymbol{Q}_{\boldsymbol{v}\boldsymbol{v}} := \frac{1}{2c}(\boldsymbol{C}_{\boldsymbol{\zeta}})_{vv} + \frac{1}{2}\left(\partial\bar{U}(\boldsymbol{\zeta})\boldsymbol{Q}_{\theta v} + \boldsymbol{Q}_{\theta v}^\top\partial\bar{U}(\boldsymbol{\zeta})^\top\right).$$

Recall that $\varphi(\boldsymbol{x}, s)$ is the joint deterministic flow with $\pi(\varphi(\boldsymbol{x}, s)) = \pi(\boldsymbol{x})$. Differentiating this identity twice at $\iota(\boldsymbol{\zeta})$ and contracting with $\boldsymbol{Q}$ yields

$$\partial^2\pi(\iota(\boldsymbol{\zeta}))[\boldsymbol{A}\boldsymbol{Q} + \boldsymbol{Q}\boldsymbol{A}^\top] + \partial\pi(\iota(\boldsymbol{\zeta}))\partial^2\boldsymbol{f}(p_{\boldsymbol{\zeta}})[\boldsymbol{Q}] = \boldsymbol{0},$$

where we recall that $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v}) := \begin{pmatrix} -S(\boldsymbol{v})\nabla\mathcal{L}(\boldsymbol{\theta}) \\ c_\beta\left(\bar{U}(\boldsymbol{\theta}) - \boldsymbol{v}\right) \end{pmatrix}$ is defined in Definition E.1. Since $\partial\pi(\iota(\boldsymbol{\zeta}))[\boldsymbol{u}, \boldsymbol{w}] = \boldsymbol{J}\boldsymbol{u}$,

$$\begin{aligned}
\partial^2\pi(\iota(\boldsymbol{\zeta}))[\boldsymbol{C}] &= \partial\pi(\iota(\boldsymbol{\zeta}))\partial^2\boldsymbol{f}(p_{\boldsymbol{\zeta}})[\boldsymbol{Q}] \\
&= \partial\pi(\iota(\boldsymbol{\zeta}))\left[\sum_{r=1}^{d+D}\boldsymbol{e}_r\sum_{a,b=1}^{d+D}\frac{\partial^2\boldsymbol{f}_r}{\partial\boldsymbol{x}_a\partial\boldsymbol{x}_b}(\iota(\boldsymbol{\zeta}))\boldsymbol{Q}_{ab}\right] \\
&= -\boldsymbol{J}\left(\boldsymbol{S}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{Q}_{\theta\theta}] + 2\sum_{i=1}^{d}\partial S(U(\boldsymbol{\zeta}))[\boldsymbol{e}_i]\boldsymbol{H}(\boldsymbol{Q}_{\theta v})_{\cdot i}\right)
\end{aligned}$$

Then we compute $\partial\pi(\iota(\boldsymbol{\zeta}))\boldsymbol{b}_{\boldsymbol{\zeta}}$.

$$\begin{aligned}
\partial\pi(\iota(\boldsymbol{\zeta}))\boldsymbol{b}_{\boldsymbol{\zeta}} &= \boldsymbol{J}\boldsymbol{b}_{\boldsymbol{\zeta}}^{\boldsymbol{\theta}} \\
&= -c\boldsymbol{J}\sum_{i=1}^{d}\partial S(U(\boldsymbol{\zeta}))[\boldsymbol{e}_i]\mathbb{E}\left[\nabla\ell(\boldsymbol{\zeta};\xi)(\boldsymbol{h}(\boldsymbol{\zeta},\xi) - U(\boldsymbol{\zeta}))^\top\right] \\
&= -c\boldsymbol{J}\sum_{i=1}^{d}\partial S(U(\boldsymbol{\zeta}))[\boldsymbol{e}_i]\boldsymbol{M}_{\cdot i},
\end{aligned}$$

where we recall that $\boldsymbol{M}_{\boldsymbol{\zeta}} := \mathbb{E}\left[\nabla\ell(\boldsymbol{\zeta};\xi)\big(\boldsymbol{h}(\boldsymbol{\zeta},\xi) - U(\boldsymbol{\zeta})\big)^\top\right]$ defined in Definition G.1. Therefore the drift in (2) is

$$\begin{aligned}
&\partial\pi(\iota(\boldsymbol{\zeta}))\boldsymbol{b}_{\boldsymbol{\zeta}} + \frac{1}{2}\partial^2\pi(\iota(\boldsymbol{\zeta}))[\boldsymbol{C}] \\
=& -\frac{1}{2}\boldsymbol{J}\boldsymbol{S}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{Q}_{\theta\theta}] - \boldsymbol{J}\sum_{i=1}^{d}\partial S(U(\boldsymbol{\zeta}))[\boldsymbol{e}_i]\left(c\boldsymbol{M} + \boldsymbol{H}\boldsymbol{Q}_{\theta v}\right)_{\cdot i} \\
=& -\frac{\tau}{4}\boldsymbol{J}\boldsymbol{S}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{S}] - \boldsymbol{J}\sum_{i=1}^{d}\partial S(U(\boldsymbol{\zeta}))[\boldsymbol{e}_i]\left(c\boldsymbol{M} + \boldsymbol{H}\boldsymbol{Q}_{\theta v}\right)_{\cdot i}
\end{aligned}$$

In fact, we claim that $\boldsymbol{Y} := c\boldsymbol{M} + \boldsymbol{H}\boldsymbol{Q}_{\theta v}$. Since $\boldsymbol{H}(\boldsymbol{S}\boldsymbol{H} + c\boldsymbol{I}_d)^{-1} = (\boldsymbol{H}\boldsymbol{S} + c\boldsymbol{I}_d)^{-1}\boldsymbol{H}$, we have

$$\begin{aligned}(\boldsymbol{H}\boldsymbol{S} + c\boldsymbol{I}_d)\boldsymbol{Y} &= c\boldsymbol{H}\boldsymbol{S}\boldsymbol{M} + c^2\boldsymbol{M} + c(\boldsymbol{H}\boldsymbol{S} + c\boldsymbol{I}_d)\boldsymbol{H}(\boldsymbol{S}\boldsymbol{H} + c\boldsymbol{I})^{-1}\boldsymbol{S}\left(\frac{\tau}{2}\partial\bar{U}(\boldsymbol{\zeta})^\top - \boldsymbol{M}\right)\\ &= c\boldsymbol{H}\boldsymbol{S}\boldsymbol{M} + c^2\boldsymbol{M} + c\boldsymbol{H}\boldsymbol{S}\left(\frac{\tau}{2}\partial\bar{U}(\boldsymbol{\zeta})^\top - \boldsymbol{M}\right)\\ &= c^2\boldsymbol{M} + \frac{\tau c}{2}\boldsymbol{H}\boldsymbol{S}\partial\bar{U}(\boldsymbol{\zeta})^\top,\end{aligned}$$

which is precisely (85), and therefore (86) holds. □

Proof of Lemma G.1. We continue to use the abbreviations in the previous proof. For Adam, since

$$V(\boldsymbol{M}) = \mathrm{diag}(\boldsymbol{M}), \qquad S(\boldsymbol{v}) = \mathrm{Diag}\left(\frac{1}{\sqrt{\boldsymbol{v}} + \epsilon\mathbf{1}}\right),$$

together with the label noise condition we have

$$U(\boldsymbol{\zeta}) = \tau\mathrm{diag}(\boldsymbol{H}), \qquad \boldsymbol{S} = \mathrm{Diag}\left(\frac{1}{\sqrt{\tau\mathrm{diag}(\boldsymbol{H})} + \epsilon\mathbf{1}}\right).$$

Direct differentiation gives

$$\partial S(U(\boldsymbol{\zeta}))[\boldsymbol{e}_i] = -\frac{\boldsymbol{e}_i\boldsymbol{e}_i^\top}{2\sqrt{U_i(\boldsymbol{\zeta})}(\sqrt{U_i(\boldsymbol{\zeta})} + \epsilon)^2}$$

and

$$\nabla\mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}) = \frac{\sqrt{\tau}}{2}\sum_{i=1}^{d}\frac{\nabla H_{ii}(\boldsymbol{\zeta})}{\sqrt{\tau H_{ii}(\boldsymbol{\zeta})} + \epsilon} = \frac{\sqrt{\tau}}{2}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{S}].$$

Recall that

$$(\boldsymbol{\delta}_{\epsilon,\boldsymbol{\zeta}})_i := \frac{(\boldsymbol{Y}_{\boldsymbol{\zeta}})_{ii}}{\sqrt{U_i(\boldsymbol{\zeta})}\big(\sqrt{U_i(\boldsymbol{\zeta})} + \epsilon\big)}.$$

Since $\boldsymbol{e}_i^\top\boldsymbol{Y}_{\cdot i} = \boldsymbol{Y}_{ii}$ and $\boldsymbol{S}\boldsymbol{e}_i = (\sqrt{U_i(\boldsymbol{\zeta})} + \epsilon)^{-1}\boldsymbol{e}_i$, we obtain

$$\sum_{i=1}^{d}\partial S(U(\boldsymbol{\zeta}))[\boldsymbol{e}_i]\boldsymbol{Y}_{\cdot i} = -\frac{1}{2}\sum_{i=1}^{d}\frac{\boldsymbol{Y}_{ii}\boldsymbol{e}_i}{\sqrt{U_i(\boldsymbol{\zeta})}(\sqrt{U_i(\boldsymbol{\zeta})} + \epsilon)^2} = -\frac{1}{2}\boldsymbol{S}\boldsymbol{\delta}_{\epsilon,\boldsymbol{\zeta}}.$$

Under label noise, by (86) we have

$$\begin{aligned}\dot{\boldsymbol{\zeta}} &= -\frac{\tau}{4}\boldsymbol{J}\boldsymbol{S}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{S}] - \boldsymbol{J}\sum_{i=1}^{d}\partial S(U(\boldsymbol{\zeta}))[\boldsymbol{e}_i]\boldsymbol{Y}_{\cdot i}\\ &= -\frac{\tau}{4}\boldsymbol{J}\boldsymbol{S}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{S}] + \frac{1}{2}\boldsymbol{J}\boldsymbol{S}\boldsymbol{\delta}_{\epsilon,\boldsymbol{\zeta}}\\ &= -\frac{\sqrt{\tau}}{2}\boldsymbol{J}\boldsymbol{S}\left(\nabla\mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}) - \frac{1}{\sqrt{\tau}}\boldsymbol{\delta}_{\epsilon,\boldsymbol{\zeta}}\right).\end{aligned}$$

Finally, the spectral projection satisfies $\ker\boldsymbol{J} = \mathrm{Ran}(\boldsymbol{S}\boldsymbol{H})$. Since $\boldsymbol{S}$ is invertible,

$$\ker(\boldsymbol{J}\boldsymbol{S}) = \mathrm{Ran}(\boldsymbol{H}) = N_{\boldsymbol{\zeta}}\Gamma.$$

Therefore $\dot{\boldsymbol{\zeta}}$ vanishes exactly when the tangent component of the vector inside parentheses vanishes, as desired. □

G.3 Schemes $(1, 2)$

Assume that $1 - \beta_2 = c_\beta\eta^\alpha$ with constants $c_\beta > 0$ and $1 < \alpha < 2$ for any $\eta > 0$. Note that the impact of $\boldsymbol{v}$ can be replaced by $U(\boldsymbol{\zeta})$. Therefore the joint correlation in the 1-scheme cancels out and absent from the resulting slow ODE. We give the following formulated result of the ODE and implicit bias of Adam under the label noise condition.

Theorem G.2. *Under the label noise condition, the Slow SDE for AGMs in equation* (3) *becomes the following ODE:*

$$\mathrm{d}\boldsymbol{\zeta}(t) = -\frac{\tau}{4}\partial\Phi_{\boldsymbol{S}_*(\boldsymbol{\zeta}(t))}(\boldsymbol{\zeta}(t))\boldsymbol{S}_*(\boldsymbol{\zeta}(t))\nabla^3\mathcal{L}(\boldsymbol{\zeta}(t))[\boldsymbol{S}_*(\boldsymbol{\zeta}(t))]\,\mathrm{d}t. \tag{88}$$

Lemma G.2. *Under the label noise condition and the specialization to Adam, every fixed point $\boldsymbol{\zeta}^*$ of Equation* (88) *for Adam satisfies $\nabla_\Gamma\mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}^*) = \mathbf{0}$.*

G.4 Proof for Schemes $(1, 2)$

Lemma G.3. *Fix $\boldsymbol{\zeta} \in \Gamma$ and a positive definite matrix $\boldsymbol{S}$. Let $\boldsymbol{H} := \nabla^2 \mathcal{L}(\boldsymbol{\zeta})$ and $\boldsymbol{P} := \partial \Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})$. Then*

$$\partial^2 \Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})[\boldsymbol{SHS}] = -\frac{1}{2}\boldsymbol{PS}\nabla^3 \mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{S}]. \tag{89}$$

Proof. Use the linear change of variables $\boldsymbol{\theta} = \boldsymbol{S}^{1/2}\boldsymbol{x}$ and define

$$\widetilde{\mathcal{L}}(\boldsymbol{x}) := \mathcal{L}(\boldsymbol{S}^{1/2}\boldsymbol{x}), \qquad \widetilde{\Phi}(\boldsymbol{x}) := \boldsymbol{S}^{-1/2}\Phi_{\boldsymbol{S}}(\boldsymbol{S}^{1/2}\boldsymbol{x}).$$

Then

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{S}^{-1/2}\dot{\boldsymbol{\theta}}(t) = -\boldsymbol{S}^{1/2}\nabla \mathcal{L}(\boldsymbol{S}^{1/2}\boldsymbol{x}(t)) = -\nabla \widetilde{\mathcal{L}}(\boldsymbol{x}(t)).$$

At $\widetilde{\boldsymbol{\zeta}} := \boldsymbol{S}^{-1/2}\boldsymbol{\zeta}$, let $\widetilde{\boldsymbol{H}} := \nabla^2 \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\zeta}}) = \boldsymbol{S}^{1/2}\boldsymbol{H}\boldsymbol{S}^{1/2}$ and $\widetilde{\boldsymbol{P}} := \partial \widetilde{\Phi}(\widetilde{\boldsymbol{\zeta}})$. The projection is constant along the gradient flow, so $\partial \widetilde{\Phi}(\boldsymbol{x})\nabla \widetilde{\mathcal{L}}(\boldsymbol{x}) = \boldsymbol{0}$ near $\widetilde{\boldsymbol{\zeta}}$. Let $\{\boldsymbol{e}_i\}_{i=1}^d$ be an orthonormal eigenbasis of $\widetilde{\boldsymbol{H}}$, with eigenvalues $\lambda_i \geq 0$. Differentiating $\partial \widetilde{\Phi}(\boldsymbol{x})\nabla \widetilde{\mathcal{L}}(\boldsymbol{x}) = \boldsymbol{0}$ twice at $\widetilde{\boldsymbol{\zeta}}$ gives that

$$0 = \left|\frac{\mathrm{d}^2}{\mathrm{d}s^2}\right|_{s=0} \left\{\partial \widetilde{\Phi}(\widetilde{\boldsymbol{\zeta}} + s\boldsymbol{e}_i)\nabla \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\zeta}} + s\boldsymbol{e}_i)\right\} = 2\lambda_i \partial^2 \widetilde{\Phi}(\widetilde{\boldsymbol{\zeta}})[\boldsymbol{e}_i, \boldsymbol{e}_i] + \widetilde{\boldsymbol{P}}\nabla^3 \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\zeta}})[\boldsymbol{e}_i, \boldsymbol{e}_i].$$

Summing this equation over all i implies

$$\partial^2 \widetilde{\Phi}(\widetilde{\boldsymbol{\zeta}})[\widetilde{\boldsymbol{H}}] = -\frac{1}{2}\widetilde{\boldsymbol{P}}\nabla^3 \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\zeta}})[\boldsymbol{I}].$$

Substitute the original variables back proves (89). □

Proof of Theorem G.2. Fix $\boldsymbol{\zeta} \in \Gamma$ and abbreviate $\boldsymbol{H} := \boldsymbol{H}_{\boldsymbol{\zeta}}$, $\boldsymbol{S} := \boldsymbol{S}_*(\boldsymbol{\zeta})$ and $\boldsymbol{P} := \partial \Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})$. Recall that in Lemma F.5 we have (26), namely

$$\partial_{\boldsymbol{\theta}} \Psi(\boldsymbol{\zeta}, \boldsymbol{v}) S(\boldsymbol{v}) \boldsymbol{H} = \boldsymbol{0}.$$

Since $\boldsymbol{\Sigma}(\boldsymbol{\zeta}) = \tau \boldsymbol{H}$, we have

$$\boldsymbol{PS}\boldsymbol{\Sigma}(\boldsymbol{\zeta})\boldsymbol{SP}^\top = \tau \boldsymbol{PSHSP}^\top = \boldsymbol{0}.$$

Hence the $\mathrm{d}\boldsymbol{W}_t$ term in (3) vanishes. Applying Lemma G.3 to the remaining $\mathrm{d}t$ term gives

$$\frac{1}{2}\partial^2 \Phi_{\boldsymbol{S}}(\boldsymbol{\zeta})[\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\zeta})\boldsymbol{S}] = -\frac{\tau}{4}\boldsymbol{PS}\nabla^3 \mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{S}],$$

which proves (88). □

Proof of Lemma G.2. We continue to use the previous notation. For Adam, recall that $V(\boldsymbol{M}) = \mathrm{Diag}(\boldsymbol{M})$ and $S(\boldsymbol{v}) = \mathrm{Diag}\left(\frac{1}{\sqrt{\boldsymbol{v}} + \epsilon \boldsymbol{1})}\right)$. Together with the label noise condition we have

$$U(\boldsymbol{\zeta}) = \tau \mathrm{diag}(\boldsymbol{H}), \qquad \boldsymbol{S} = \mathrm{Diag}\left(\frac{1}{\sqrt{\tau \mathrm{diag}(\boldsymbol{H})} + \epsilon \boldsymbol{1}}\right).$$

Because $\boldsymbol{S}$ is diagonal, direct differentiation gives

$$\nabla \mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}) = \frac{\sqrt{\tau}}{2}\sum_{i=1}^d \frac{\nabla H_{ii}(\boldsymbol{\zeta})}{\sqrt{\tau H_{ii}(\boldsymbol{\zeta})} + \epsilon} = \frac{\sqrt{\tau}}{2}\nabla^3 \mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{S}].$$

Therefore the Adam slow ODE can be written as

$$\mathrm{d}\boldsymbol{\zeta}(t) = -\frac{\sqrt{\tau}}{2}\boldsymbol{PS}\nabla \mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}(t))\mathrm{d}t.$$

It remains to identify the zero set of this vector field. For any $\boldsymbol{w} \in \mathbb{R}^d$, write $\boldsymbol{w} = \boldsymbol{w}_T + \boldsymbol{w}_N$, where $\boldsymbol{w}_T \in T_{\boldsymbol{\zeta}}\Gamma$ and $\boldsymbol{w}_N \in N_{\boldsymbol{\zeta}}\Gamma$. By Lemma F.5, $\boldsymbol{PS}\boldsymbol{w}_N = \boldsymbol{0}$. Moreover, $\ker \boldsymbol{P} = \boldsymbol{S}N_{\boldsymbol{\zeta}}\Gamma$ since both spaces have dimension m. If $\boldsymbol{w}_{\mathrm{T}} \in T_{\boldsymbol{\zeta}}\Gamma$ and $\boldsymbol{PS}\boldsymbol{w}_{\mathrm{T}} = \boldsymbol{0}$, then $\boldsymbol{S}\boldsymbol{w}_{\mathrm{T}} \in \boldsymbol{S}N_{\boldsymbol{\zeta}}\Gamma$, so $\boldsymbol{w}_{\mathrm{T}} \in T_{\boldsymbol{\zeta}}\Gamma \cap N_{\boldsymbol{\zeta}}\Gamma = \{\boldsymbol{0}\}$. Therefore $\boldsymbol{PS}\boldsymbol{w} = \boldsymbol{0} \Leftrightarrow \boldsymbol{w}_T = \boldsymbol{0}$. At a fixed point $\boldsymbol{\zeta}^*$, by the Adam slow ODE and the equivalence with $\boldsymbol{w} = \nabla \mathcal{R}_{\epsilon,\tau}(\boldsymbol{\zeta}^*)$ completes the proof. □