IMPLICIT BIAS OF SGD IN RELU AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

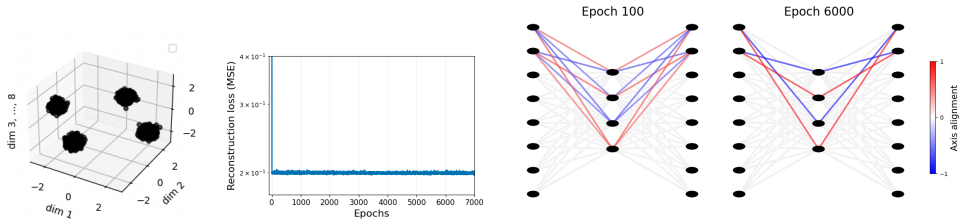
How can useful features emerge without label supervision? We study this question theoretically using shallow ReLU and linear autoencoders trained on data from a 4-component Gaussian mixture model. Under stochastic gradient descent (SGD), the ReLU autoencoder learns neurons aligned with the cluster-mean directions, whereas the linear autoencoder recovers only the signal subspace, without selecting a preferred orientation. Consequently, the ReLU model can learn representations better suited to downstream classification, even though both architectures attain the same optimal reconstruction loss. We show that the implicit directional bias emerges during late-stage SGD but is absent under population gradient flow on the same ReLU model. We explain this implicit bias of SGD by characterizing the global-minimizer manifold of the ReLU autoencoder as a *normally attracting stationary manifold* (NASM) for population gradient flow. Building on this geometry, we develop an analysis of SGD with a small step size η over $O(\eta^{-2})$ iterations and prove that stochastic gradient fluctuations induce an effective tangential drift along the global-minimizer manifold. In our setting, the high-dimensional drift reduces to a scalar angular ordinary differential equation whose stable equilibria correspond to the cluster-mean directions. This analysis reveals a mode of unsupervised feature learning during late-stage SGD, in which the representation continues to "get better" even after the unsupervised training loss has plateaued.

1 INTRODUCTION

There is significant interest in understanding feature learning in neural networks and the associated implicit biases of gradient descent (GD) and stochastic gradient descent (SGD). A long line of work has precisely characterized feature learning in shallow neural networks in the context of learning Boolean functions, such as k-parity (Abbe & Sandon, 2020; Daniely & Malach, 2020; Barak et al., 2022; Glasgow, 2024), or Gaussian mixture classification (Refinetti et al., 2021; Shi et al., 2023; Frei et al., 2023; Xu et al., 2024). While early work on feature learning showed the benefits of neural networks over kernel or random feature models (Daniely & Malach, 2020; Refinetti et al., 2021), other studies have shown which features are learned in over-parameterized networks (Frei et al., 2023; Glasgow, 2024), demonstrated the sample efficiency of SGD over GD (Abbe & Sandon, 2020), identified directional biases of SGD (Wu et al., 2021), and explained grokking as a consequence of late training dynamics (Xu et al., 2024), among others. However, the current literature on feature learning is predominantly focused on supervised learning, where the network is trained to predict target labels, imposing an explicit bias to learn features that are better suited for label prediction.

By contrast, despite the empirical success of the unsupervised representation learning (Ericsson et al., 2021), considerably less is understood about how optimization implicitly shapes features learned without supervision. We investigate the implicit bias of SGD in a canonical unsupervised model: a shallow, tied-weight autoencoder (AE) with a bottleneck layer, comparing linear and ReLU activations. Linear autoencoders, in particular, have long admitted rigorous characterizations of both their loss landscapes and training dynamics (Baldi & Hornik, 1989; Fukumizu, 1998). This classical theory, however, largely characterizes feature learning through the lens of principal component analysis (PCA). Linear AEs recover a principal subspace while identifying their hidden representations only up to an invertible transformation (Baldi & Hornik, 1989) and up to an orthogonal transformation even with regularization (Kunin et al., 2019). Recent analyses of shallow nonlinear autoencoders, trained with SGD, provide a similar PCA-type characterization of the subspace learned at the hidden layer using mean-field or high-dimensional order-parameter dynamics (Nguyen, 2021;

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.



(a) 4-GMM data (b) Training loss (c) Evolution of learned ReLU weights

Figure 1: Feature learning by an (unsupervised) tied-weight ReLU autoencoder trained by SGD to minimize reconstruction loss. We draw 2000 samples from a 4-GMM in $\mathbb{R}^8$, where centers $\{\pm\mu_1, \pm\mu_2\}$ are along the first and second axes. Edge colors indicate the cosine between each weight vector and the corresponding coordinate axis: red denotes positive alignment, blue negative alignment, and white near-zero alignment. Early in training (epoch 100), the weights concentrate on the first two input coordinates, indicating recovery of $\mathrm{Span}(\mu_1, \mu_2)$. By epoch 6000, each hidden neuron aligns almost exclusively with one signed coordinate and hence with a cluster-mean direction. Remarkably, the reconstruction loss remains nearly constant throughout this realignment, and the autoencoder never observes the labels—yet SGD recovers the same mean-aligned feature organization observed under direct supervision (Frei et al., 2023), where the label-dependent objective explicitly reward such alignment (see Figure 4 in Appendix E.2).

Refinetti & Goldt, 2022). Still, this leaves unresolved whether nonlinearity and SGD select a particular organization of features within the recovered subspace. This leads us to ask: ***Do shallow autoencoders with nonlinear activation in the bottleneck layer learn beyond PCA or spectral projections?*** Although the widespread use of autoencoders in practice shows that they are indeed useful for representation learning (Vincent et al., 2010), the following theoretical questions remain open.

> Which projections (or features) are learned by nonlinear architectures in the context of unsupervised representation learning? What are the causes and implicit biases that lead to the learned features?

We answer these questions precisely in the context of a shallow ReLU autoencoder trained to reconstruct samples from a 4-component Gaussian mixture model (4-GMM) with means $\{\pm\mu_1, \pm\mu_2\}$ and isotropic covariance $\sigma^2 I_d$, where $\mu_1^\top \mu_2 = 0$ and $\|\mu_1\| = \|\mu_2\|$. For the specific 4-GMM, one can verify that a linear autoencoder learns the principal subspace spanned by $\{\mu_1, \mu_2\}$, and the convergence of SGD to this solution is closely related to the SGD convergence for PCA (Shamir, 2016). Notably, rotating the hidden features within the recovered subspace leaves the optimal reconstruction error unchanged. As noted earlier, recent works on shallow autoencoders—despite incorporating nonlinearity and SGD—only characterize the learned subspace $\mathrm{Span}(\mu_1, \mu_2)$. On the other hand, the 4-GMM model has become a theoretical benchmark for studying feature learning in supervised neural networks under the so-called Gaussian XOR model Refinetti et al. (2021); Frei et al. (2023); Xu et al. (2024). Here, the nonlinear separation of the classes forces the hidden-layer to exactly learn the cluster centers $\{\pm\mu_1, \pm\mu_2\}$, instead of arbitrary orientations of the subspace. Figure 1 reveals a striking unsupervised analogue. A ReLU autoencoder trained solely through reconstruction nevertheless learns its hidden weights aligned with the cluster-mean directions, matching the organization observed under direct supervision even though the reconstruction objective is indifferent to orientation. A closer inspection also reveals that the early training dynamics of SGD learn an orthogonal basis for $\mathrm{Span}(\mu_1, \mu_2)$, whereas the late training dynamics *rotate* the features within this subspace to align with the cluster centers, with the training error remaining nearly constant. The remainder of this paper focuses on rigorously characterizing this phenomenon.

Main technical contributions. Considering the problem of minimizing the reconstruction error of samples from a 4-GMM using shallow autoencoders, we make the following technical contributions.

(A) In Section 2, we first present a precise neuron-level characterization of the global minimizers of undercomplete, tied-weight ReLU autoencoders (AE), which serves as the foundation for the subsequent characterization of the training dynamics. A few simple but notable observations follow:

(i) while a linear AE requires only a 2-neuron bottleneck to recover the signal subspace, a ReLU AE requires a bottleneck width of 4; (ii) at their respective minimal widths, both linear and ReLU AEs attain the same optimal reconstruction error, and the global minima are invariant to orthogonal transformations; (iii) the manifold of global minima in the ReLU AE contains representations in which the hidden weights form antipodal pairs along mutually orthogonal directions.

(B) In Section 3, we formally state the implicit bias of SGD in ReLU AE towards specific representations among the global minima. In particular, we rigorously show that if a ReLU AE is trained with SGD using a small step size η, then after $O(\eta^{-2})$ iterations, the learned features align with the cluster centers with high probability. The key geometric result is that the global-minimizer manifolds of the ReLU and linear autoencoders are *normally attracting stationary manifolds* (NASM), which is a special case of a normally hyperbolic invariant manifold (NHIM) (Wiggins, 2013).

(C) Near NASM, the dynamics exhibit a natural separation of time scales, where population gradient flow (pGF) contracts rapidly in the normal directions, while stochastic effects accumulate slowly along the tangential direction. In Section 4, we exploit this structure, and develop an analysis of SGD applicable for $O(\eta^{-2})$ steps near NASM, matching the time scale of the global analysis of Li et al. (2021) while requiring weaker regularity assumptions that accommodate the ReLU population loss. Applied to the ReLU autoencoder on the 4-GMM, our reduction eliminates the $4d-1$ fast normal degrees of freedom and reduces the slow motion along the manifold to a scalar angular ordinary differential equation (ODE). The corresponding tangential drift vanishes for the linear autoencoder.

(D) Section 5 discusses the implications of our results for downstream prediction and their relevance beyond the current data setting. We first bound the XOR classification error achievable by a linear probe on the learned representations, connecting the derived implicit bias to downstream utility. Second, experiments on Fashion-MNIST provide evidence of similar feature-selection behavior beyond Gaussian mixtures. Finally, experiments on symmetric GMMs with more components reveal a surprising departure from the 4-component case. More concretely, SGD continues to select distinguished features, but these need not align with the cluster means.

2 Manifold of Global Minima for Autoencoders

As a first step toward understanding which solutions SGD selects, we characterize the global-minimizer manifolds of ReLU and Linear AEs under the settings in Section 2.1. This characterization extends to a more general setting, as discussed in Remark 2.3 and detailed in Appendix B.1.

2.1 Autoencoders and Training setup

Following Xu et al. (2024), we study the data consists of a symmetric four-components Gaussian-mixture embedded in $\mathbb{R}^d$, denoted as $x \sim \frac{1}{4}\sum_{i=1}^{2}\left[\mathcal{N}(\mu_i, \sigma^2 I_d) + \mathcal{N}(-\mu_i, \sigma^2 I_d)\right]$, where we assume $\|\mu_1\| = \|\mu_2\| = s$, $\mu_1^\top \mu_2 = 0$. For $Q := [\mu_1/s, \mu_2/s] \in \mathbb{R}^{d\times 2}$ and $Q_\perp \in \mathbb{R}^{d\times(d-2)}$ spans the subspace orthogonal to $\mathrm{Span}(Q)$, we denote $\Sigma_x := \mathbb{E}[xx^\top] = \lambda_s QQ^\top + \lambda_n Q_\perp Q_\perp^\top$, where $\lambda_s := \sigma^2 + \frac{s^2}{2}$ and $\lambda_n := \sigma^2$.[1] Our ReLU and linear AEs have widths of 4 and 2, respectively. Let $W = [w_1, \ldots, w_4] \in \mathbb{R}^{d\times 4}$, and $W_{\mathrm{lin}} = [\bar{w}_1, \bar{w}_2] \in \mathbb{R}^{d\times 2}$ denote their respective weight matrices.[2] The corresponding reconstruction maps are $f(x; W) = \sum_{i=1}^{4} w_i \phi(w_i^\top x)$ for ReLU, and $f_{\mathrm{lin}}(x; W_{\mathrm{lin}}) = \sum_{i=1}^{2} \bar{w}_i \bar{w}_i^\top x$ for linear, where $\phi(t) = \max\{t, 0\}$ acts entrywise. For $\lambda \in (\lambda_n, \lambda_s)$, the corresponding population losses are $L_\triangle(W_\triangle) = \frac{1}{2}\mathbb{E}\,\|x - f_\triangle(x; W_\triangle)\|_2^2 + \frac{\lambda}{2}\|W_\triangle\|_F^2$, for $\triangle \in \{\mathrm{ReLU}, \mathrm{lin}\}$. We use the row-major vectorization $h : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$, where $h(A) = [A_{11}, \ldots, A_{mn}]^\top$. Throughout, we identify each weight matrix $W \in \mathbb{R}^{d\times m}$ with its vectorization in $\mathbb{R}^{dm}$. Since vectorization is a linear isometry between the Frobenius and Euclidean inner-product spaces, we use the same notation $\mathcal{M}$ for a matrix manifold and its vectorized image. By default, we omit the subscript ReLU across all notations when clear from context. We denote $\varphi_t(z)$ as the population gradient-flow (pGF) solution initialized at z, namely, $\frac{d}{dt}\varphi_t(z) = -\nabla L(\varphi_t(z)), \varphi_0(z) = z$.

[1] Our result is coordinate-free and holds for any $\lambda_n \in (0, \lambda_s)$. For ease of presentation, we set $\lambda_n = \sigma^2$.

[2] This choice of width corresponds to the minimum required width to reconstruct the 2-dimensional signal subspace, where both autoencoders achieve the same reconstruction error at their global minimum.

2.2 Manifolds of Global Minimizers and Neuron Pairing

We first recall a standard result on global minimizers of linear autoencoders (Kunin et al., 2019). Here, $\mathrm{O}(2)$ denotes the group of 2×2 orthogonal matrices, and $\rho^2 := 1 - \lambda/\lambda_s$.

Fact 2.1 *Let $\mathcal{M}_{\mathrm{lin}} := \{\rho QR \mid R \in \mathrm{O}(2)\}$. Under the settings of Section 2.1, the following holds: a weight matrix W_{lin} is a global minimizer of $L_{\mathrm{lin}}(W_{\mathrm{lin}})$ if and only if $W_{\mathrm{lin}} \in \mathcal{M}_{\mathrm{lin}}$.*

It is well known that a linear neuron can be realized by a pair of ReLU neurons through the identity $z = \phi(z) - \phi(-z)$. The following result shows that, at minimal width, this paired structure is not merely sufficient but in fact necessary. Every global minimizer has this form up to permutation.

Proposition 2.2 (Manifold of Global Minimizers for ReLU Autoencoders). *Let $\mathcal{M}_{\mathrm{ReLU}} := \{\rho[QR, -QR] \mid R \in \mathrm{O}(2)\}$. Under the settings of Section 2.1, up to permutation, a weight matrix $W \in \mathbb{R}^{d\times 4}$ is a global minimizer of $L(W)$ if and only if $W \in \mathcal{M}_{\mathrm{ReLU}}$.*

(Proof). It follows from Proposition B.6, where we set $r = 2$ and $D = \rho I_2$. □

Remark 2.3 (Generality of Neuron Pairing). *Previous analyses of ReLU autoencoders describe PCA-type reconstruction behavior in their respective settings (Nguyen, 2021; Refinetti & Goldt, 2022). Motivated by this, in Appendix B.1.1, we characterize the weight structure of untied ReLU autoencoders that realize the PCA reconstruction at minimal width. This analysis shows that neuron pairing extends beyond Gaussian models: our distributional assumption only requires the law of x to be mutually absolutely continuous with respect to Lebesgue measure on the signal subspace (Assumption B.1). This condition is strictly more general than the additive Gaussian-noise setting (Refinetti & Goldt, 2022) and the Gaussian setting (Nguyen, 2021). However, this characterization is conditional on realizing the PCA reconstruction and does not by itself establish global optimality for the population loss. Under the additional assumption of central symmetry stated in Appendix B.1.2, we show that the paired-orthogonal configurations form the global-minimizer manifold.*

At global optimality, both models learn the same reconstruction map: for $W = \rho[QR, -QR] \in \mathcal{M}$,

$$f(x; W) = \rho^2 \left((QR)^\top \phi(QRx) + (-QR)^\top \phi(-QRx)\right) = \rho^2 (QR)^\top (QR)x = f_{\mathrm{lin}}(x; QR).$$

Note that this equivalence concerns only the reconstructed output, not the hidden representation. The ReLU autoencoders learn the same reconstruction as its linear counterpart, but through a more complex organization of hidden features.[3] There is, however, yet another equivalence among the solutions within each architecture: different orientations yield the same reconstruction map. Are these equivalent-risk solutions also equally favored by SGD?

3 Main Result: Implicit Bias of SGD in ReLU Autoencoder

We study SGD initialized in a sufficiently small neighborhood of the global-minimizer manifold with a radius independent of the step size η (in contrast to the shrinking initialization neighborhoods used in the local analysis of Blanc et al. (2020)).Theorem 3.1 shows that SGD induces an effective tangential drift that drives the ReLU autoencoder toward mean-aligned orientations with high probability. Figure 2 illustrates this geometry.

SGD setting The following notation applies separately to either autoencoder. Let $\vartheta \in \mathbb{R}^p$ denote the vectorized parameter matrix and let $\mathcal{L}$ denote the corresponding population loss. Write $\mathcal{L}(\vartheta) = \mathbb{E}_x[\ell(\vartheta; x)]$, where $\ell(\vartheta; x)$ is the single-sample loss. Define the sample gradient by $g(\vartheta; x) = \nabla_\vartheta \ell(\vartheta; x)$. Assume that, for the parameters under consideration, $\mathbb{E}[g(\vartheta; x)] = \nabla\mathcal{L}(\vartheta)$, $\mathbb{E}\|g(\vartheta; x)\|^2 < \infty$ (both hold in our setting). Given i.i.d. samples $(x_k)_{k\geq 0}$ independent of the initialization ϑ_0^η, define the filtration $\mathcal{F}_k^\eta := \sigma\left(\vartheta_0^\eta, x_0, \ldots, x_{k-1}\right)$. Define the gradient noise and its covariance by $\xi(\vartheta; x) := g(\vartheta; x) - \nabla\mathcal{L}(\vartheta)$ and $\Sigma(\vartheta) := \mathbb{E}[\xi(\vartheta; x)\xi(\vartheta; x)^\top]$. Writing $\xi_k^\eta := \xi(\vartheta_k^\eta; x_k)$, the noise increments satisfy $\mathbb{E}[\xi_k^\eta \mid \mathcal{F}_k^\eta] = 0$, $\mathbb{E}[\xi_k^\eta(\xi_k^\eta)^\top \mid \mathcal{F}_k^\eta] = \Sigma(\vartheta_k^\eta)$. Thus, the SGD update with step size $\eta > 0$ can be written as $\vartheta_{k+1}^\eta = \vartheta_k^\eta - \eta g(\vartheta_k^\eta; x_k) = \vartheta_k^\eta - \eta\nabla\mathcal{L}(\vartheta_k^\eta) - \eta\xi_k^\eta$.

[3] The implication of this for downstream XOR classification is discussed in Section 5.

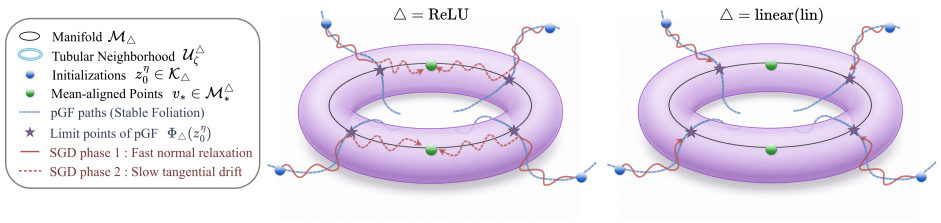


Figure 2: Illustration of our main results. Starting anywhere in the stable foliation (definition in Theorem C.2), SGD initially tracks the trajectory of population gradient flow (pGF), rapidly enters the relevant tubular neighborhood $\mathcal{U}_\zeta^\triangle$ and contracts further until the normal displacement is $O(\sqrt{\eta})$ (Phase 1). For the ReLU autoencoder, SGD subsequently induces a deterministic tangential drift, driving a collective rotation of the features toward alignment with the cluster-mean directions (Phase 2). In contrast, pGF converges to an initialization-dependent point on the manifold. For linear autoencoder, no corresponding limiting tangential drift occurs even with SGD (no Phase 2).

Because the model and the SGD updates are equivariant under permutations of the hidden units, we fix a column ordering by choosing the convention for R_θ so that the hidden-weight direction in the first quadrant is proportional to $(\mu_1/s)\sin\theta + (\mu_2/s)\cos\theta$. Thus, $\theta \in [0, \pi/2]$ is its angle with μ_2. Define the global minimizers, and the mean-aligned target set by

$$\mathcal{M}^\theta_{\mathrm{ReLU}} := \{v(\theta) := \rho[QR_\theta, -QR_\theta] \,|\, \theta \in [0, \pi/2]\}, \qquad \mathcal{M}_* := \{v(0),\ v(\pi/2)\}.$$

We denote $\mathrm{dist}(z, \mathcal{Y}) := \inf_{y\in\mathcal{Y}} \|z - y\|$. Below is our main result (see D.3 for the full proof).

Theorem 3.1 (Implicit bias of the ReLU AE under SGD). *Consider the settings of Section 2.1 and the SGD recursions defined above. All initializations below are kept fixed as the step size η varies.*

(ReLU autoencoder). *There exists $\sigma_0 > 0$ such that, for every $\sigma \in (0, \sigma_0]$, there exists a compact neighborhood $\mathcal{K}$ of $\mathcal{M}^\theta$, independent of η, on which the gradient-flow projection $\Phi : \mathcal{K} \to \mathcal{M}^\theta$ is well-defined. Initialize $z_0^\eta = z_0$, where $z_0 \in \mathcal{K}$ almost surely and $\mathbb{P}(\Phi(z_0) = v(\pi/4)) = 0$.[4] Then, for every $\tau > 0$ and $\delta \in (0, 1)$, there exist $T < \infty$ and $\eta_0 > 0$ such that, for every $\eta \in (0, \eta_0)$,*

$$\mathbb{P}\left(\mathrm{dist}(z^\eta_{\lfloor T/\eta^2 \rfloor}, \mathcal{M}_*) < \tau\right) \geq 1 - \delta.$$

(Linear autoencoder). *Similarly, there exists an η-independent compact neighborhood $\mathcal{K}_{\mathrm{lin}}$ of $\mathcal{M}_{\mathrm{lin}}$ on which the gradient-flow projection $\Phi_{\mathrm{lin}} : \mathcal{K}_{\mathrm{lin}} \to \mathcal{M}_{\mathrm{lin}}$ is well defined. Initialize $z^\eta_{\mathrm{lin},0} = z_{\mathrm{lin},0}$, where $z_{\mathrm{lin},0} \in \mathcal{K}_{\mathrm{lin}}$ almost surely. Then, for every $0 < t_0 < T' < \infty$,*

$$\sup_{t\in[t_0,T']} \|z^\eta_{\mathrm{lin},\lfloor t/\eta^2 \rfloor} - \Phi_{\mathrm{lin}}(z_{\mathrm{lin},0})\| \to_{\mathbb{P}} 0, \qquad \text{as } \eta \downarrow 0.$$

Thus, on the η^{-2} time scale, SGD induces no additional drift for the linear autoencoder beyond the phase selected by the population gradient flow (pGF). For the ReLU autoencoder, by contrast, SGD systematically selects the mean-aligned target set. What explains these separations—between the ReLU and linear architectures, and between SGD and pGF? We clarify the underlying mechanisms through the proof sketch in Section 4. The following geometric result provides the key starting point.

Lemma 3.2 *For each $\triangle \in \{\mathrm{ReLU}, \mathrm{lin}\}$, $\mathcal{M}_\triangle$ is a compact C^∞ normally attracting stationary manifold for the population gradient flow $\dot{z} = -\nabla L_\triangle(z)$.*

(Proof). Since the map $R \mapsto \rho[Q_2 R, -Q_2 R]$ smoothly embeds the compact Lie group $\mathrm{O}(2)$ into parameter space, $\mathcal{M}_{\mathrm{ReLU}}$ is C^∞. The rest of the proof is provided in Appendix D.4. □

Lemma 3.2 reveals the fast-slow geometry underlying our analysis. In NHIM theory (Wiggins, 2013), the dynamics (pGF in our setting) decompose into rapid normal motion toward or away from the invariant manifold and slower tangential motion along it. Crucially, the domination of the normal rates over the tangential rates allows the long-time behavior to be described through *reduced*

[4] This holds for any initialization supported in $\mathcal{K}$ with an ambient Lebesgue density.

dynamics on the manifold. From this viewpoint, we therefore expect SGD with a sufficiently small step size to initially shadow pGF during the fast contraction phase. Stationarity, however, implies that pGF induces no deterministic motion once the manifold is reached. How, then, can nontrivial tangential motion arise near this manifold (NASM)? In the next section, we outline the stochastic mechanism responsible for this motion and explain how we characterize it on the global η^{-2} time scale despite the absence of the C^3 regularity on the loss function assumed by Li et al. (2021).

4 Proof Outline: NASM Reduction of SGD to Angular Dynamics

We briefly sketch the analysis of SGD under the NASM framework applied to autoencoders (AE). The full NASM-based analysis, applicable to more general models, is provided in Appendix C. We first introduce few notations. When an analysis holds for both ReLU and linear AE, we suppress the subscript, for instance, $\mathcal{M} \in \{\mathcal{M}_{\mathrm{ReLU}}, \mathcal{M}_{\mathrm{lin}}\}$. Let $\mathcal{U}_{\mathrm{st}}$ be a neighborhood of $\mathcal{M}$ such that population gradient flow (pGF) initialized at any $z \in \mathcal{U}_{\mathrm{st}}$ converges to $\Phi(z) \in \mathcal{M}$, where $\Phi : \mathcal{U}_{\mathrm{st}} \to \mathcal{M}$ is the asymptotic pGF limit map.[5] For $v \in \mathcal{M}$, let $P_T(v)$ and $P_N(v)$ denote the orthogonal projections onto the tangent space $T_v\mathcal{M}$ and the normal space $N_v\mathcal{M}$, respectively. For the corresponding population loss L, which is C^2 near $\mathcal{M}$, define the normal Hessian $H_N(v) := \nabla^2 L(v)|_{N_v\mathcal{M}}$ and the normal covariance $\Sigma_N(v) := \mathbb{E}[g(v;x)g(v;x)^\top]|_{N_v\mathcal{M}}$, where $A|_{N_v\mathcal{M}}$ denotes the restriction of A to the normal space. With this notation, we state two properties of the loss and sample gradients of both autoencoders that, together with the key geometric lemma 3.2, form the basis of the reduction.

1. Tangential regularity: The normal Hessian $H_N(v)$ varies C^1-smoothly along $\mathcal{M}$. For $n \in N_v\mathcal{M}$, write $L(v+n) = L(v) + \frac{1}{2}\langle n, H_N(v)n\rangle + R(v,n)$ (since $\nabla L(v) = 0$). The remainder vary negligibly along the manifold: $\left|\frac{d}{ds}R(v(s),n(s))\right| = o(\|n(s)\|^2)$, as $\|n(s)\| \to 0$.

2. Absence of leading tangential noise: At every $v \in \mathcal{M}$, the sample gradient is normal to the manifold, meaning that $P_T(v)g(v;x) = 0$ almost surely.

The first property is critical to overcome C^3 regularity limitation assumed in the previous work (Li et al., 2021) (see Remark 4.4). The second property arises naturally when individual sample losses are invariant along the minimizer manifold. Specifically, if $\ell(v;x)$ is constant along $\mathcal{M}$, then, for every tangent direction $\tau \in T_v\mathcal{M}$ and $v(s) \in \mathcal{M}$ satisfying $v(0) = v$ and $\dot{v}(0) = \tau$, we have $\langle g(v;x), \tau\rangle = \frac{d}{ds}\big|_{s=0}\ell(v(s);x) = 0$. Hence $P_T(v)g(v;x) = 0$. For the autoencoders, orthogonal changes of the hidden basis leave both the reconstruction map and the weight-decay penalty unchanged. This makes the sample loss constant along each minimizer manifold. The same condition also holds in supervised learning with fresh label noise on an interpolation manifold (Li et al., 2021; Blanc et al., 2020; HaoChen et al., 2021). It turns out that these properties, together with the regularity conditions stated in Appendix C, yield a general reduction theorem for SGD near a NASM (Theorem C.6) that makes no explicit architectural assumptions. Both autoencoders fall within this framework. Before stating the resulting corollary 4.1, we introduce two quantities.

1. Nearest-point projection π: By the tubular-neighborhood theorem (Lee, 2003) and compactness of $\mathcal{M}$, there exists a uniform radius $\zeta > 0$ such that every $z \in \mathcal{U}_\zeta := \{z \mid \mathrm{dist}(z,\mathcal{M}) < \zeta\}$ has a unique nearest point $\pi(z) \in \mathcal{M}$. The map $\pi : \mathcal{U}_\zeta \to \mathcal{M}$ is C^∞. The pair $(v,n) := (\pi(z), z - \pi(z))$ defines the *tubular coordinates* of z, so that $z = v + n$, $n \in N_v\mathcal{M}$, and $\|n\| = \mathrm{dist}(z,\mathcal{M})$.

2. Effective tangential drift b : For $v \in \mathcal{M}$, let Π_v be the unique solution of the Lyapunov equation $H_N(v)\Pi_v + \Pi_v H_N(v) = \Sigma_N(v)$. Define the tangent vector field $b(v) \in T_v\mathcal{M}$ by $\langle b(v), \tau\rangle = \mathrm{Tr}((\nabla^\perp_\tau H_N)(v)\Pi_v)$, where $\tau \in T_v\mathcal{M}$ and $\nabla^\perp_\tau H_N$ denotes the derivative of the normal Hessian along the manifold, with normal vectors defined using parallel transport in the normal bundle (this is well-defined due to the tangential regularity).

Applying Theorem C.6 to the two autoencoders yields the following corollary.

Corollary 4.1 (Effective tangential dynamics of SGD). *Consider the corresponding SGD recursion under the setting of Section 2.1. Initialize $z_0^\eta = z_0$ for every $\eta > 0$, where its law is supported on a compact subset of $\mathcal{U}_{\mathrm{st}}$. Then, for every $0 < t_0 < T < \infty$, $\sup_{t\in[t_0,T]} \|\pi(z^\eta_{\lfloor t/\eta^2\rfloor}) - v(t)\| \to_{\mathbb{P}} 0$ as $\eta \downarrow 0$, where $v(t) \in \mathcal{M}$ is the unique solution of $\dot{v}(t) = -\frac{1}{2}b(v(t))$ with $v(0) = \Phi(z_0)$.*

[5] Such a neighborhood and a continuous limit map Φ exist by the stable-foliation theorem for compact NASMs (Hirsch et al., 1970; Wiggins, 2013).

(Proof sketch). To explain this corollary, we outline the main ideas behind Theorem C.6 using the notation introduced above (the full proof is provided in Appendix C). This proceeds in two steps, exploiting the fast-slow nature near NASM to derive the tangential drift on the $O(\eta^{-2})$ time scale. The proof combines classical stochastic approximation and Poisson-equation averaging (Pavliotis & Stuart, 2008), with tubular coordinates around a NASM.

PHASE 1: BURN-IN INTO A LOCAL TUBE. Uniform normal attraction of NASM, together with our regularity assumptions on sample gradient, ensures that SGD initialized in a compact subset of $\mathcal{U}_{\mathrm{st}}$ enters any sufficiently small fixed tube $\mathcal{U}_\zeta$ within $O(\eta^{-1})$ iterations, with probability tending to one. During this entrance phase, SGD tracks population gradient flow and preserves its asymptotic phase up to a vanishing error (see Lemma C.8). This justifies using tubular coordinates for the subsequent local analysis, without an additional order-one phase shift induced by SGD. Moreover, within $\tilde{O}(\eta^{-1})$ iterations, SGD reaches the neighborhood with the normal variables scale $O_{\mathbb{P}}(\sqrt{\eta})$.

PHASE 2: STOCHASTIC REDUCTION. After phase 1, SGD enters the regime in which its normal displacement is $O_{\mathbb{P}}(\sqrt{\eta})$. Write $z_k = v_k + n_k$, where $v_k = \pi(z_k) \in \mathcal{M}$, and define $q_k := \eta^{-1/2} n_k$, so that the rescaled normal fluctuations remain of order one. For $\tau \in T_v\mathcal{M}$, define the symmetric bilinear map $\mathcal{B}_v : N_v\mathcal{M} \times N_v\mathcal{M} \to T_v\mathcal{M}$ by $\langle \mathcal{B}_v[n_1, n_2], \tau \rangle := \langle n_1, (\nabla^\perp_\tau H_N)(v) n_2 \rangle$.[6] Decomposing the sample gradient as $g_{k+1} = \nabla L(z_k) + \xi_{k+1}$, and Taylor expanding π gives $v_{k+1} - v_k = -\eta D\pi(z_k)\nabla L(z_k) - \eta D\pi(z_k)\xi_{k+1} +$ (remainder terms). Lemma C.14 identifies the leading population-gradient contribution $-\eta D\pi(z_k)\nabla L(z_k) = -\frac{\eta^2}{2}\mathcal{B}_{v_k}[q_k, q_k] + o(\eta^2)$ uniformly for bounded q_k. The absence of leading tangential sample noise, together with the moment and martingale estimates, shows that the remaining terms vanish after accumulation on the $O(\eta^{-2})$ steps. Consequently, $v_N - v_0 = -(\eta^2/2)\sum_{k<N} \mathcal{B}_{v_k}[q_k, q_k] + o_{\mathbb{P}}(1)$ uniformly for $N \le T/\eta^2$.

To identify the fast normal dynamics, we temporarily hold the base point $v \in \mathcal{M}$ fixed. On the $s = k\eta$ time scale, the rescaled normal variable is approximated by the Ornstein-Uhlenbeck (OU) process $\mathrm{d}q_s = -H_N(v)q_s\,\mathrm{d}s + \Sigma_N(v)^{1/2}\mathrm{d}B_s$, whose invariant covariance Π_v solves $H_N(v)\Pi_v + \Pi_v H_N(v) = \Sigma_N(v)$. *Because the base point changes negligibly on this normal-relaxation time scale, this separation of time scales suggests replacing* $\sum_{k<N}\mathcal{B}_{v_k}[q_k, q_k]$ *by its local stationary average* $b(v) = \mathbb{E}_{q\sim\mathcal{N}(0,\Pi_v)}[\mathcal{B}_v[q, q]]$. We justify this replacement using a Poisson-equation argument, which accounts for the difficulty of dealing with the changing base point v_k: For each fixed v, let χ_v be the quadratic corrector solving $\mathscr{L}_v\chi_v(q) = -(\mathcal{B}_v[q, q] - b(v))$, where $\mathscr{L}_v$ is the generator of the normal OU process. Along the actual iterates, its increment decomposes as $\chi_{v_{k+1}}(q_{k+1}) - \chi_{v_k}(q_k) = (\chi_{v_k}(q_{k+1}) - \chi_{v_k}(q_k)) + (\chi_{v_{k+1}}(q_{k+1}) - \chi_{v_k}(q_{k+1}))$. The first term is controlled by the OU generator with v_k held fixed. The second term accounts for the motion of the base point. The Lipschitz dependence of the corrector on v gives $|\chi_{v_{k+1}}(q_{k+1}) - \chi_{v_k}(q_{k+1})| \le C\|v_{k+1} - v_k\|\|q_{k+1}\|^2$. Because v_k moves slowly and the moments of q_k remain uniformly bounded, this moving-base contribution vanishes after accumulation. Summing the remaining terms produces a telescoping corrector and a martingale, both of which are negligible on the slow time scale.[7] Consequently, $\sup_{N\le T/\eta^2}\|\eta^2\sum_{k<N}(\mathcal{B}_{v_k}[q_k, q_k] - b(v_k))\| \to_{\mathbb{P}} 0$. It follows that $v_N - v_0 = -(\eta^2/2)\sum_{k<N} b(v_k) + o_{\mathbb{P}}(1)$, which is an Euler approximation of the effective dynamics $\dot{v}(t) = -\frac{1}{2}b(v(t))$. □

Computing limiting autoencoder drift in Corollary 4.1. The limiting motion $\dot{v}(t)$ is governed by averaging the quadratic tangential response over the stationary normal fluctuations, which is $4d-1$ dimensional in ReLU AE (the tangent space is 1-dimensional). We now apply Corollary 4.1 to each autoencoder. This starts from analyzing each normal space. For $v \in \rho[U, -U] \in \mathcal{M}_{\mathrm{ReLU}}$, $U = QR$, the high-dimensional normal space admits the splitting $N_v\mathcal{M}_{\mathrm{ReLU}} = \bigoplus_{\alpha\in\{S,K,B,C\}} N_\alpha$, where

$$N_S := \{[US, -US] \,|\, S = S^\top, S \in \mathbb{R}^{2\times 2}\},\ N_K := \{[Q_\perp K, -Q_\perp K] \,|\, K \in \mathbb{R}^{(d-2)\times 2}\},$$
$$N_B := \{[UB, UB] \,|\, B \in \mathbb{R}^{2\times 2}\},\ N_C := \{[Q_\perp C, Q_\perp C] \,|\, C \in \mathbb{R}^{(d-2)\times 2}\}.$$

Notice that the S, K-blocks preserve the antipodal pairing, whereas the B, C-blocks break it. Let P_α be the orthogonal projection onto N_α, and define the block operators on the full normal space

[6] Thus, $\mathcal{B}_v$ measures how the normal curvature of the loss changes along the minimizer manifold.

[7] The relevance of this equation to averaging comes from the definition of the generator. Over one normal update, $\mathbb{E}[\chi_{v_k}(q_{k+1}) - \chi_{v_k}(q_k) \mid \mathcal{F}_k] = -\eta(\mathcal{B}_{v_k}[q_k, q_k] - b(v_k)) + o(\eta)$.

by $H_\alpha := P_\alpha H_N P_\alpha$, $\Sigma_\alpha := P_\alpha \Sigma_N P_\alpha$. Our model-specific calculations show that H_N and Σ_N are block diagonal, and hence $H_N = \sum_\alpha H_\alpha$, $\Sigma_N = \sum_\alpha \Sigma_\alpha$. The unique solution of the normal Lyapunov equation is therefore also block diagonal. For $\Pi_\alpha := P_\alpha \Pi_N P_\alpha$, we have $\Pi_N = \sum_\alpha \Pi_\alpha$, with $H_\alpha \Pi_\alpha + \Pi_\alpha H_\alpha = \Sigma_\alpha$ on N_α. Note that, for $\tau \in T_v\mathcal{M}$, $\langle b_{\mathrm{ReLU}}(v), \tau \rangle = \mathrm{Tr}[(\nabla^\perp_\tau H_N)(v)\Pi_N(v)]$. Define $\langle b_\alpha(v), \tau \rangle := \mathrm{Tr}[(\nabla^\perp_\tau H_\alpha)(v)\Pi_\alpha(v))]$. Linearity gives $\nabla^\perp_\tau H_N = \sum_\alpha \nabla^\perp_\tau H_\alpha$. The cross terms vanish: for $\alpha \neq \beta$, $\mathrm{Tr}((\nabla^\perp_\tau H_\alpha)\Pi_\beta) = 0$ (see Appendix D.2). Consequently, $\langle b_{\mathrm{ReLU}}(v), \tau \rangle = \sum_\alpha \mathrm{Tr}[(\nabla^\perp_\tau H_\alpha)(v)\Pi_\alpha(v)]$, and therefore $b_{\mathrm{ReLU}} = b_S + b_K + b_B + b_C$.

Lemma 4.2 (Block reduction of the late-phase dynamics). *The S, K-block normal Hessians are covariantly constant along the minimizer manifold: $\nabla^\perp_\tau H_S = \nabla^\perp_\tau H_K = 0$. Consequently, $b_S = b_K = 0$, and $b_{\mathrm{ReLU}} = b_B + b_C$. For the linear AE, $N_v\mathcal{M}_{\mathrm{lin}} = N_S^{\mathrm{lin}} \oplus N_K^{\mathrm{lin}}$, and the corresponding two normal Hessian blocks are likewise covariantly constant. Hence $b_{\mathrm{lin}} = b_S^{\mathrm{lin}} + b_K^{\mathrm{lin}} = 0$.*

Thus *the pair-breaking B, C-directions are the only possible contributors to the implicit bias*, which explains the absence of the bias for the linear AEs on the $O(\eta^{-2})$ time scale. The next lemma shows how this pair-breaking blocks induce the angular ODE. We denote by $\widehat{\Pi}_B$ and $\widehat{\Pi}_C$ the covariances of the reduced coordinates B and C induced by Π_B and Π_C (details in Appendix D.4).

Lemma 4.3 (Limiting flow of ReLU AE). *Let $v(\theta) \in \mathcal{M}^\theta_{\mathrm{ReLU}}$, $z := Q^\top x$, $y(\theta) := R(\theta)^\top z = (y_1(\theta), y_2(\theta))^\top$, and $u_y(\theta) := \mathbb{E}|y_1(\theta)y_2(\theta)|$. Define $\kappa_B = \frac{1}{2}\mathbb{E}_{B\sim\mathcal{N}(0,\widehat{\Pi}_B)}[\sum_{i=1}^2 B_{i1}B_{i2}]$, and $\kappa_C = \frac{1}{4}\mathbb{E}_{C\sim\mathcal{N}(0,\widehat{\Pi}_C)}[\sum_{i=1}^{d-2} C_{i1}C_{i2}]$. Then $b_B(v(\theta)) + b_C(v(\theta)) = 2(\kappa_B + \kappa_C)u_y'(\theta)\,\partial_\theta v(\theta)$. Consequently, $\mathrm{d}v_t = \partial_\theta v(\theta_t)\,\mathrm{d}\theta_t$, where $\mathrm{d}\theta_t = -(\kappa_B + \kappa_C)u_y'(\theta_t)\,\mathrm{d}t$. Moreover, there exists $\sigma_* > 0$ such that, for every $\sigma \in (0, \sigma_*]$, $\kappa_B + \kappa_C > 0$, and if $\theta_0 \in (0, \pi/4)$, then $\theta_t \to 0$, whereas if $\theta_0 \in (\pi/4, \pi/2)$, then $\theta_t \to \frac{\pi}{2}$. The unique interior equilibrium is $\theta = \frac{\pi}{4}$ and the vector field vanishes at the two endpoints: $\lim_{\theta\downarrow 0} u_y'(\theta) = 0$, $\lim_{\theta\uparrow\pi/2} u_y'(\theta) = 0$.*

Thus SGD asymptotically slows as it approaches either endpoint of $\mathcal{M}_*$. Combining Proposition 2.2, Lemma 3.2, Corollary 4.1, Lemma 4.2 and 4.3 proves Theorem 3.1.

Remark 4.4 (Lack of smoothness of ReLU loss & comparison to Li et al. (2021)). *A central technical ingredient for the proof is a sufficiently smooth projection of the SGD iterates onto the manifold, which allows us to derive effective dynamics along the manifold in the small step-size regime. The main limitation of Li et al. (2021) comes from assuming a C^3 loss and using the resulting C^2 gradient-flow projection Φ. Our ReLU loss does not satisfy this assumption (see Lemma D.3). Our first observation is that **the reduction only requires differentiability of the normal Hessian $\nabla^2 L$ along the manifold**, rather than in all ambient directions, which motivates the first property (tangential regularity). Another key step in overcoming the regularity obstacle is to analyze SGD using **tubular coordinates**. Crucially, for a compact embedded C^k manifold $\mathcal{M} \subset \mathbb{R}^D$ without boundary, **the regularity of π inherits that of the manifold**: $\mathcal{M}$ being C^k guarantees that π is C^{k-1} (Foote, 1984). Thus, taylor expanding smooth π instead of Φ we avoid differentiating Φ completely.*

5 Discussion

In this work, we have characterized which features SGD and ReLU nonlinearity jointly select. In particular, we have highlighted the mechanism that drives this selection through a structured, low-dimensional deterministic drift, *even without explicit guidance from label supervision*. But how can these learned representations be useful, and what are the broader implications?

Effect of ReLU features on XOR downstream classification Empirical studies have shown that representations learned without label supervision can match or even surpass supervised baselines when evaluated using linear probes (Ericsson et al., 2021). We investigate a theoretical counterpart to these observations by assigning XOR labels to our 4-GMM (the two antipodal clusters along one signal direction receive label $+1$, while those along the other receive label -1, this yields a linearly inseparable classification task). In this setting, prior work has proved a separation between shallow ReLU and linear networks trained with label supervision (Frei et al., 2023). Proposition 5.1 recovers an analogous separation through linear probing of features learned solely by reconstruction, without access to the downstream labels. A quantitative statement and its proof are in Appendix E.1.

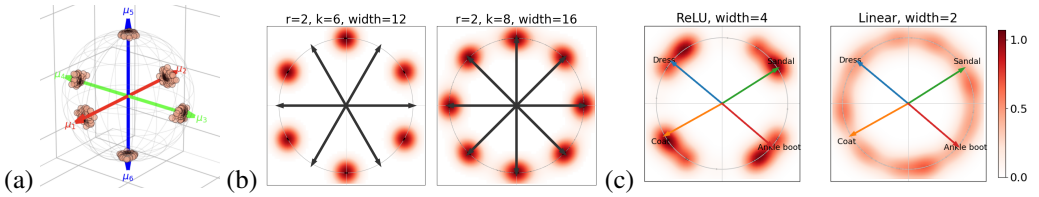


Figure 3: Additional empirical results for autoencoders. Solid lines indicate cluster-mean directions for k-GMMs, and density plots summarize the final feature orientations over 200 runs. **(a)** Our analysis extends readily to $2r$-GMMs with antipodal pairs of means along r orthogonal signal directions (see Appendix D). The panel illustrates the case $r = 3$, where the feature density concentrates around the cluster-mean directions. **(b)** For ReLU AEs trained beyond the minimal-width setting on symmetric GMMs, we consistently observe k attraction basins, where k is the number of clusters. The two panels illustrate contrasting behavior: for $k = 4, 8, \ldots$, features align with the cluster means, whereas for $k = 6, 10, \ldots$, they align with the angular bisectors between neighboring cluster directions. **(c)** The two panels show the behavior on real data (Fashion-MNIST). Features from ReLU AE exhibit directional bias, similar to what our theory predict for GMMs, whereas features learned by linear AE show no such directional bias. More details are provided in Appendix E.2.

Proposition 5.1 (Downstream XOR Classification, Informal). *$\mathcal{M}_{\mathrm{ReLU}}$ contains representations that enable a linear probe to achieve arbitrarily small downstream XOR classification error at sufficiently high signal-to-noise ratio, despite having identical reconstruction risk to the linear counterpart. Specifically, this holds for every fixed orientation $\theta \neq \pi/4$, with the error bound improving as θ approaches the mean-aligned orientations 0 or $\pi/2$.*

This result first establishes the architectural advantage of antipodally paired ReLU features over their linear counterparts in terms of downstream utility, with the classification-error bound minimized at the mean-aligned orientations $\theta = 0$ and $\pi/2$, which SGD selects (Theorem 3.1). Thus, the ReLU autoencoder with SGD selects representations with the strongest downstream guarantee within a family that reconstruction loss alone cannot distinguish.

Feature Identification Whereas population gradient flow on the ReLU AE and SGD on the linear AE retain initialization-dependent orientations, SGD on the ReLU model consistently selects the same feature directions. A natural question is whether this systematic selection persists beyond the minimal widths considered in our theory. Extending the analysis is challenging. The global-minimizer manifold is then parameterized by a Stiefel manifold rather than $\mathrm{O}(r)$, and additional degeneracies in certain normal directions invalidate the uniform normal attraction required by our NASM-based analysis. Nevertheless, the experiments in Figure 3 indicate that SGD continues to select distinguished ReLU directions during late-stage training, motivating the following conjecture.

Conjecture 5.2 (Feature identification by ReLU AE with SGD). *Consider a symmetric GMM with $2m$ sufficiently well-separated components, where $m \geq 2$ is an integer, whose means are equally spaced on a circle in a two-dimensional signal subspace. Under suitable random initialization, a shallow ReLU autoencoder of sufficiently large width, trained by SGD with small step size η, exhibits the following behavior. Every hidden weight that contributes nontrivially to reconstruction is attracted toward a preferred direction: an angular bisector between neighboring cluster directions when m is odd, or a cluster-mean direction when m is even. For any fixed $\varepsilon > 0$, all preferred directions are represented to within angular error ε with high probability within $O(\eta^{-2})$ steps.*

More broadly, our results raise a question in unsupervised representation learning. When an objective admits equally optimal representations, how does training select among them, and when does that selection matter downstream? Related ambiguities also arise in augmentation-based self-supervised learning, where, identifiability guarantees establish recovery of invariant content up to an invertible transformation (Von Kügelgen et al., 2021), or of latent factors up to an orthogonal transformation in contrastive learning (Zimmermann et al., 2021). Such guarantees do not determine which representative a particular training procedure selects. Complementing this perspective, our work provides a concrete example in which architecture and stochastic optimization select distinguished features among equivalent-loss solutions, with consequences for downstream prediction.

AI USE STATEMENT

In this work, we used generative AI tools to generate code for the experiments and to refine our writing, including grammar checking, sentence rephrasing, polishing hypotheses and proofs. We did not use generative AI tools for other core tasks such as conducting automated literature reviews or generating results. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the full statements and their complete proofs in the appendix. The full details of the experiments are provided in Appendix E.2.

REFERENCES

Emmanuel Abbe and Colin Sandon. On the universality of deep learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Boaz Barak, Benjamin L. Edelman, Surbhi Goel, Sham M. Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

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

Amit Daniely and Eran Malach. Learning parities with neural networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Steffen Dereich and Sebastian Kassing. On minimal representations of shallow relu networks. *Neural Networks*, 148:121–128, 2022.

Jaap Eldering et al. *Normally hyperbolic invariant manifolds: the noncompact case*, volume 2. Springer, 2013.

Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5410–5419. IEEE, 2021.

Robert L Foote. Regularity of the distance function. *Proceedings of the American Mathematical Society*, 92(1):153–155, 1984.

Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *Journal of Machine Learning Research*, 24(303): 1–49, 2023.

Kenji Fukumizu. Dynamics of batch learning in multilayer neural networks. In *International Conference on Artificial Neural Networks*, pp. 189–194. Springer, 1998.

Margalit Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. In *International Conference on Learning Representations*, volume 2024, pp. 52419–52430, 2024.

Jonghyun Ham, Maximilian Fleissner, and Debarghya Ghoshdastidar. Impact of bottleneck layers and skip connections on the generalization of linear denoising autoencoders. *arXiv preprint arXiv:2505.24668*, 2025.

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 Chapman Pugh, and Michael Shub. Invariant manifolds. *Bulletin of the American Mathematical Society*, 76(5):1015–1019, 1970.

Gary Shon Katzenberger. *Solutions of a stochastic differential equation forced onto a manifold by a large drift*. The University of Wisconsin-Madison, 1990.

Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In *International conference on machine learning*, pp. 3560–3569. PMLR, 2019.

John M Lee. Smooth manifolds. In *Introduction to smooth manifolds*, pp. 1–29. Springer, 2003.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer Cham, 2018.

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

Phan-Minh Nguyen. Analysis of feature learning in weight-tied autoencoders via the mean field lens. *arXiv preprint arXiv:2102.08373*, 2021.

Thanh V Nguyen, Raymond KW Wong, and Chinmay Hegde. On the dynamics of gradient descent for autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2858–2867. PMLR, 2019.

Grigorios A Pavliotis and Andrew M Stuart. *Multiscale methods: averaging and homogenization*, volume 53. Springer New York, 2008.

Henning Petzka, Martin Trimmel, and Cristian Sminchisescu. Notes on the symmetries of 2-layer relu-networks. In *Proceedings of the northern lights deep learning workshop*, volume 1, pp. 6–6, 2020.

Arnu Pretorius, Steve Kroon, and Herman Kamper. Learning dynamics of linear denoising autoencoders. In *International Conference on Machine Learning*, pp. 4141–4150. PMLR, 2018.

Maria Refinetti and Sebastian Goldt. The dynamics of representation learning in shallow, non-linear autoencoders. In *International Conference on Machine Learning*, pp. 18499–18519. PMLR, 2022.

Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8936–8947. PMLR, 2021.

Ohad Shamir. Convergence of stochastic gradient descent for PCA. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 257–265. JMLR.org, 2016.

Zhenmei Shi, Junyi Wei, and Yingyu Liang. Provable guarantees for neural networks via gradient feature learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Stephen Wiggins. *Normally hyperbolic invariant manifolds in dynamical systems*, volume 105. Springer Science & Business Media, 2013.

David Williams. *Probability with martingales*. Cambridge university press, 1991.

Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Zhiwei Xu, Yutong Wang, Spencer Frei, Gal Vardi, and Wei Hu. Benign overfitting and grokking in relu networks for xor cluster data. In *International Conference on Learning Representations*, volume 2024, pp. 43356–43400, 2024.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International conference on machine learning*, pp. 12979–12990. PMLR, 2021.

Appendix Overview

The appendices are organized as follows. Section B.1 studies neuron pairing in ReLU autoencoders and characterizes their global-minimizer manifold. Its normally attracting stationary manifold (NASM) structure motivates the analysis of SGD near such manifolds in Section C. Although motivated by the ReLU setting, this reduction framework is not ReLU specific and requires only the NASM structure and additional regularity assumptions. Section D then combines these results to prove the main theorems. Each section begins with its motivation and an outline of the argument, providing a guide to the supporting lemmas.

A Additional Related Work **13**

B Results for Neuron Pairing in ReLU Autoencoder Learning **14**

B.1 Pairing Phenomenon 14

B.1.1 ReLU AE achieving PCA-level Reconstruction 14

B.1.2 Global Minimizer defined by Centrally Symmetric Distribution 15

B.2 Proofs 15

B.2.1 Proof of Lemma B.3 15

B.2.2 Proof of Proposition B.5 18

B.2.3 Proof of Proposition B.6 19

C Implicit Bias of SGD near NASM **21**

C.1 Notations 21

C.2 Theory of NHIM 21

C.3 General setting and assumptions 22

C.4 Main Theorem 23

C.5 SGD Phase 1: Fast relaxation with $\tilde{O}(\eta^{-1})$ steps and entrance into the $O(\sqrt{\eta})$ tube 24

C.5.1 Auxiliary Lemmas of Phase 1 26

C.5.2 Proof of Lemma C.3 31
C.6 SGD Phase 2: Stochastic reduction 31
C.6.1 Proof of Proposition C.10 32
C.6.2 Auxiliary Lemmas and Proofs for Phase 2 39

D Application to $\mathcal{M}_{\mathrm{ReLU}}$ **50**
D.1 Notations and Setting 51
D.2 Proof of Lemma 4.2 51
D.3 Proof of Theorem 3.1 52
D.4 $\mathcal{M}_{\mathrm{ReLU}}$ is a Normally Attracting Stationary Manifold 52
D.5 Auxiliary Lemmas and Proofs 53
D.5.1 Verification of the Regularity Assumptions 53
D.5.2 Auxiliary Lemmas for Section D.4 61
D.5.3 Proof of Lemma 4.3 65

E Miscellaneuous **74**
E.1 Proposition 5.1 74
E.2 Additional Experiments and Settings 74

A Additional Related Work

Shallow Autoencoders The feature-learning properties of weight-tied shallow ReLU autoencoders have also been studied in Nguyen (2021); Nguyen et al. (2019); Refinetti & Goldt (2022), although in regimes with different implications from ours. Specifically, Nguyen (2021) analyzes overcomplete autoencoders trained on centered Gaussian data using a mean-field approach. Beyond the mismatch in data settings, their dynamics recover the principal signal subspace—but preserve rotational symmetry within repeated eigenspaces, leaving the SGD-induced selection of mean-aligned representations unaddressed. Refinetti & Goldt (2022) derive high-dimensional order-parameter dynamics for finite-width, undercomplete nonlinear autoencoders using heuristic statistical-physics arguments. Their analysis captures the sequential acquisition of leading PCA subspaces and the corresponding reconstruction dynamics, but under vanilla SGD, the learned representation is characterized only up to rotations of the principal components. Their dependence on second-order statistics makes the analysis technique inapplicable to our setting. They also report that a tied, unbiased ReLU autoencoder performs poorly relative to an untied model with trained biases. This conclusion is specific to their parametrization and width, whereas our work concludes otherwise. Nguyen et al. (2019) instead considers an undercomplete autoencoder for a Gaussian-mixture model whose unit-norm component means are approximately orthogonal. This assumption excludes our setting, which contains antipodal pairs of means with absolute inner product one. Moreover, their convergence guarantee requires that, up to permutation, each hidden weight be initialized close to its corresponding mixture mean. Their result therefore establishes local refinement from an initialization that already identifies the individual components.

Feature learning in shallow linear denoising autoencoders (DAEs) has also been studied. Pretorius et al. (2018) analyze the learning dynamics of linear DAEs under an objective averaged over corruption noise, which acts as a regularizer on the reconstruction map. Ham et al. (2025) instead characterize the learned linear maps and their generalization when fitting finite corrupted datasets in a high-dimensional regime where the data dimension and the number of training samples are comparable, examining the effects of bottleneck layers and skip connections.

Implicit bias of SGD in late training. A related line of work studies SGD near manifolds of minimizers, particularly in overparameterized supervised models that interpolate the training data Blanc et al. (2020); HaoChen et al. (2021); Li et al. (2021). In interpolating models with squared loss, ordinary minibatch noise vanishes at exact interpolation, motivating persistent label perturbations to induce motion. Blanc et al. (2020) pioneer this direction and identify noise-induced drift along the minimizer manifold, while HaoChen et al. (2021) show in a quadratically parameterized regression model, that the parameter-dependent covariance of label-noise SGD can favor sparse solutions where gradient descent or isotropic parameter noise does not. Building on the theory of Katzenberger (1990), Li et al. (2021) establish an $O(\eta^{-2})$-scale limit for smooth minimizer manifolds and general noise covariance, extending the shorter local time horizon of Blanc et al. (2020) and allowing tangential noise. Their C^3 loss assumption does not cover ReLU population loss in general. Their framework however accommodates more general noise structures, including noise with a nonzero tangential component.

B Results for Neuron Pairing in ReLU Autoencoder Learning

Representing a scalar linear function using two sign-paired ReLU neurons is a familiar construction in the study of ReLU network identifiability (Petzka et al., 2020). Dereich & Kassing (2022) characterize minimal-width representations of fixed scalar-valued functions by shallow ReLU networks and describe the smooth geometry of their parameter sets. Here, we investigate when optimal autoencoder reconstruction at minimal width forces neuron pairing and what this structure implies for feature learning. We state the main results in Section B.1 and provide their proofs in Section B.2.

B.1 Pairing Phenomenon

As discussed in Remark 2.3, Refinetti & Goldt (2022) argues that shallow nonlinear autoencoders cannot improve upon the PCA reconstruction error in the setting considered. This motivates a closer examination of the structure of ReLU autoencoders that achieve PCA-level reconstruction. Section B.1.1 connects this reconstruction property to neuron pairing. However, this characterization alone does not establish the global optimality of the resulting solution manifold. In Section B.1.2, we additionally assume central symmetry of the data distribution and tied weights, and show that the paired-orthogonal manifold indeed characterizes global minimizers.

B.1.1 ReLU AE achieving PCA-level Reconstruction

Data. For the centered data x, let $\Sigma_x := \mathbb{E}[xx^\top]$ denote the population covariance matrix, and let $\Sigma_x = U\Gamma U^\top$ be its eigendecomposition, where Γ is diagonal with eigenvalues ordered in descending order. We denote by U_k the matrix formed by the top k eigenvectors of Σ_x. Define the rank-k projector $P_k := U_k U_k^\top$. We assume $k \le r$, where r is the rank of Σ_x.

Assumption B.1 *Let μ_x denote the law of x, and let μ_z denote the law of $z := U_k^\top x$. We assume that μ_z is mutually absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^k$ and that* $\mathrm{span}(U_k) \subseteq \mathrm{supp}(\mu_x)$.

Model. We consider a two-layer ReLU network of the form $f(x) = V\phi(Wx)$, with bottleneck dimension $2k$, [8] where $W \in \mathbb{R}^{2k\times d}$, $V \in \mathbb{R}^{d\times 2k}$, and $\phi(\cdot)$ denotes the ReLU nonlinearity. We assume that this model achieves the k-PCA reconstruction. Recall that the k-PCA reconstruction of the input x is given by $U_k U_k^\top x = P_k x$.

Assumption B.2 (PCA Optimal Solution at Rank k). *The network achieves PCA-level reconstruction in the sense that $f(x) = V\phi(Wx) = P_k x$ almost surely. In other words, $\mathbb{E}\|P_k x - f(x)\|_2^2 = 0$.*

The set of such solutions $\{f\}$ is non-empty, since the PCA reconstruction can always be implemented by a ReLU network, with two ReLU neurons simulating $u_i^\top x$, for each $i \in [k]$.

Lemma B.3 (Pairing under minimal-width PCA reconstruction). *Let $f(x) = V\phi(Wx)$ be a two-layer ReLU autoencoder with hidden width $m = 2k$ satisfying Assumptions B.1 and B.2. Write*

[8] This is the minimal requirement for the bottleneck dimension to acquire the k-PCA reconstruction.

$A := U_k^\top V = [a_1, \dots, a_{2k}]$ and $B := WU_k = [b_1, \dots, b_{2k}]^\top$. Then, up to permutation (or relabeling), the projected encoder and decoder weights satisfy

$$b_{i+k} = -\lambda_i b_i, \qquad a_{i+k} = -\lambda_i^{-1} a_i, \qquad i = 1, \dots, k,$$

for some positive scalars $\lambda_1, \dots, \lambda_k$. Moreover, the paired representatives satisfy the biorthogonality relations $\langle a_i, b_j \rangle = \delta_{ij}$ for $i, j = 1, \dots, k$.

Remark B.4 (Bi-orthogonality). *The condition $\langle a_i, b_j \rangle = \delta_{ij}$ is known as bi-orthogonality. Under this condition, no orthogonality is imposed among $(a_i)_{i=1}^k$ themselves. Instead, a_i and b_j are orthogonal unless $i = j$. This requirement is relatively weak and admits degenerate solutions, such as correlated encoder weight vectors. Consequently, little can be said about the individual structures of W and V. To obtain stronger characterizations, additional regularity assumptions are required.*

Imposing the minimum-norm principle recovers the familiar paired-orthogonal structure.

Proposition B.5 (Paired representation under the minimum-norm principle). *Let $m = 2k$ and suppose B.1 and B.2. Among solutions satisfying $f(x) = P_k x$ almost surely, consider the minimum-norm selection*

$$(W^*, V^*) \in \arg\min_{W,V} \left(\|W\|_F^2 + \|V\|_F^2 \right) \quad \text{s.t.} \quad V\phi(Wx) = P_k x \quad \text{a.s.}$$

Then there exists an orthogonal matrix $Q \in \mathbb{R}^{k \times k}$ such that $W^* = \begin{bmatrix} Q^\top U_k^\top \\ -Q^\top U_k^\top \end{bmatrix}$, and $V^* = (W^*)^\top$.

B.1.2 Global Minimizer defined by Centrally Symmetric Distribution

The above proposition characterizes ReLU features in the minimum-width setting but does not guarantee the optimality of this solution. In this section, we revisit the tied-weight AE (with $2r$-width) and regularized reconstruction loss from the main text under a relaxed data setting, specifically focusing on centrally symmetric data.

Assumptions Assume the centered data with its covariance Σ_x, and that the data distribution is centrally symmetric: $x =_d -x$. Let $\Sigma_x = Q_r \Lambda_s Q_r^\top + Q_\perp \Lambda_\perp Q_\perp^\top$, where $Q_r \in \mathbb{R}^{d \times r}$ and $Q_\perp \in \mathbb{R}^{d \times r}$ are orthogonal matrices that span the signal subspace and its orthogonal complement, respectively. For $\Lambda_s = \operatorname{diag}(\nu_1, \dots, \nu_r)$ and $\Lambda_\perp = \operatorname{diag}(\tau_1, \dots, \tau_r)$, assume the spectral gap between the signal and noise: $\tau_{\max} := \max_a \tau_a < \lambda < \nu_{\min} := \min_b \nu_b$, where λ denotes the regularization parameter). Define $D := \operatorname{diag}\left(\sqrt{1 - \frac{\lambda}{\nu_1}}, \dots, \sqrt{1 - \frac{\lambda}{\nu_r}} \right)$. Write $x = Q_r z + Q_\perp \varepsilon$, where z is the signal component and the ε is orthogonal noise component. Additionally, we invoke Assumption B.1, which ensures that the signal marginal has enough support. Note that we do not impose any further constraints on the covariance matrix (i.e., eigenvalues can be distinct or zero).

Proposition B.6 (Global minimizers for general signal covariance). *Under the assumptions above, every global minimizer of $L(W)$ is, up to permutation of the $2r$ hidden units, of the form*

$$W_* = [Q_r D R, -Q_r D R], \quad R \in \mathrm{O}(r).$$

The proof for this proposition is provided in Section B.2.3.

Remark B.7 (Isotropic case restricted to the signal subspace). *If $\Lambda_s = \lambda_s I_r$, then $D = \rho I_r$, where $\rho^2 = 1 - \frac{\lambda}{\lambda_s}$. Then, $W_* = [\rho Q_r R, -\rho Q_r R]$, for $R \in \mathrm{O}(r)$. This corresponds to the case of the Gaussian-mixture data considered in our work.*

B.2 Proofs

B.2.1 Proof of Lemma B.3

1. Reduction to the top-k subspace Let $P_k := U_k U_k^\top$. Every $x \in \mathbb{R}^d$ can be decomposed as

$$x = U_k z + e, \qquad z := U_k^\top x \in \mathbb{R}^k, \quad e := (\mathbb{I} - P_k)x \in \operatorname{Span}(U_k)^\perp.$$

Since $P_k e = 0$, the PCA reconstruction depends only on z. Thus $P_k x = P_k(U_k z + e) = U_k z$.

Using the assumption $f(x) = P_k x$ a.s, then for almost every pair (z, e) induced by x, we have

$$f(U_k z + e) = P_k(U_k z + e) = U_k z = f(U_k z),$$

so f is almost surely invariant along the orthogonal directions $\{U_k z + e \mid e \in \mathrm{Span}(U_k)^\perp\}$. In other words, any optimal solution must depend only on the projection $z = U_k^\top x$. The function class under consideration is rich enough to realize this invariance and achieve zero reconstruction error. For instance, for any orthonormal basis $\{u_1, \ldots, u_k\}$ of $\mathrm{Span}(U_k)$, $P_k x = \sum_{i=1}^k u_i\big(\phi(u_i^\top x) - \phi(-u_i^\top x)\big)$, which is representable by a two-layer ReLU network. Hence, the feasible set is nonempty. Restricting to inputs of the form $x = U_k z$ yields the necessary condition

$$f(U_k z) = U_k z \quad \text{for almost every } z \in \mathbb{R}^k.$$

Define $A := U_k^\top V$, $B := W U_k$. This condition is then equivalent to $g(z) := A\phi(Bz) = z$, which we analyze from now on.

2. Minimal width and nondegeneracy. By Lemma B.8, exact k-PCA reconstruction requires width $m \geq 2k$. In the minimal-width case $m = 2k$, every hidden unit must contribute to reconstruction. In particular, $a_i \neq 0$ and $b_i \neq 0$ for all $i \in [2k]$, and $a_i \neq a_j$, for all distinct pairs of $i, j \in [2k]$(this also applies to b_i), where $A = [a_1, \ldots, a_{2k}]$ and $B = [b_1, \ldots, b_{2k}]^\top$.

3. Overlapping hyperplane. [9] Assume for contradiction that for each $i \in [2k]$, the hyperplane $\mathcal{H}_i := \{z \in \mathbb{R}^k : b_i^\top z = 0\}$ does not overlap with any other hyperplane $\mathcal{H}_j$, $j \neq i$. Then, since $\mathcal{H}_i \cap \mathcal{H}_j$ is a proper affine subspace of $\mathcal{H}_i$ for each $j \neq i$, and there are finitely many such j, we can choose

$$z_0 \in \mathcal{H}_i \setminus \bigcup_{j \neq i} \mathcal{H}_j,$$

and thus $b_i^\top z_0 = 0$ and $b_j^\top z_0 \neq 0$ for all $j \neq i$. For any activation pattern $\mathcal{I} \in \{0,1\}^m$, define the corresponding activation region

$$\mathcal{R}_\mathcal{I} := \big\{z \mid (\mathbf{1}_{\{b_i^\top z > 0\}})_{i=1}^m = \mathcal{I}\big\}. \tag{1}$$

The hyperplane $\mathcal{H}_i$ separates two adjacent full-dimensional activation regions, which we denote by $\mathcal{R}_\mathcal{I}$ and $\mathcal{R}_{\mathcal{I}'}$, with corresponding activation patterns $\mathcal{I}$ and $\mathcal{I}'$ (since z is absolutely continuous, each of these regions has positive Lebesgue measure). On each region, g is affine, and since $g(z) = z$ holds almost everywhere, it must hold everywhere on $\mathcal{R}_\mathcal{I}$ and $\mathcal{R}_{\mathcal{I}'}$. Consequently, the Jacobian on both regions equals the identity (more detailed explanation of this is in Lemma B.8).

$$\mathbb{J}_\mathcal{I} = A D_\mathcal{I} B = I_k, \qquad \mathbb{J}_{\mathcal{I}'} = A D_{\mathcal{I}'} B = I_k,$$

where $D_\mathcal{I} = \mathrm{diag}(\mathcal{I})$. Now let v be any direction such that $b_i^\top v \neq 0$. For any $\delta_* < \min_{j \neq i} \frac{|b_j^\top z_0|}{|b_j^\top v|}$, the points $z_0 \pm \delta_* v$ lie in $\mathcal{R}_\mathcal{I}$ and $\mathcal{R}_{\mathcal{I}'}$, respectively, differing at exactly one index, which is i. Since $\mathcal{I}$ and $\mathcal{I}'$ differ only at coordinate i, we have $D_\mathcal{I} - D_{\mathcal{I}'} = \pm e_i e_i^\top$ (e_i is the standard basis vector), and therefore

$$\mathbb{J}_\mathcal{I} - \mathbb{J}_{\mathcal{I}'} = A(D_\mathcal{I} - D_{\mathcal{I}'})B = \pm a_i b_i^\top.$$

But $\mathbb{J}_\mathcal{I} = \mathbb{J}_{\mathcal{I}'} = I_k$, hence $a_i b_i^\top = 0$, contradicting the fact that both a_i and b_i are nonzero. This contradiction shows that $\mathcal{H}_i$ must overlap with at least one other hyperplane.

4. Neuron Pairing. By Lemma B.8, overlapping hyperplanes cannot have the same orientation, nor can there be more than one overlapping hyperplane, as either case would reduce the number of effective rank-one components below $2k$. Therefore, for each $i \in [k]$ there exists a unique paired index $i + k$ and scalar $\lambda_i > 0$ such that

$$b_{i+k} = -\lambda_i b_i, \qquad a_{i+k} = -\lambda_i^{-1} a_i.$$

Hence, $\{b_1, \ldots, b_{2k}\} = \{b_1, -\lambda_1 b_1, \ldots, b_k, -\lambda_k b_k\}$, and the encoder and decoder weights are paired. Finally, writing the Jacobian on any full-dimensional activation region as $\mathbb{J} = \sum_{l=1}^k a_l b_l^\top$, Lemma B.9 implies biorthogonality, $\langle a_i, b_j \rangle = \delta_{ij}$, completing the proof. □

[9] A similar argument can also be found in Petzka et al. (2020, Lemma 2).

Lemma B.8 (Width lower bound for k-PCA reconstruction). *Let $f(x) = V\phi(Wx)$ be a two-layer ReLU autoencoder without biases, with width m. Consider the assumptions B.1 and B.2 in Lemma B.3. Then, $m \geq 2k$ is necessary.*

Moreover, in the minimal-width case $m = 2k$, for any z_0 lying in the interior of a full-dimensional activation region, the active index sets

$$\mathcal{I}_+(z_0) := \{i \mid b_i^\top z_0 > 0\}, \qquad \mathcal{I}_-(z_0) := \{i \mid b_i^\top z_0 < 0\},$$

each have cardinality k, and the corresponding rank-one matrices $\{a_i b_i^\top\}_{i\in\mathcal{I}_+(z_0)}$ and $\{a_i b_i^\top\}_{i\in\mathcal{I}_-(z_0)}$ each form a rank-k decomposition of the identity. In particular, each $a_i b_i^\top$ is nonzero and has rank one, and the vectors $\{b_i\}_{i\in\mathcal{I}_+(z_0)}$ (and similarly $\{b_i\}_{i\in\mathcal{I}_-(z_0)}$) define k distinct hyperplanes in $\mathbb{R}^k$.

(Proof of Lemma B.8). Restrict to inputs of the form $x = U_k z$ and define $A := U_k^\top V \in \mathbb{R}^{k\times m}$, $B := WU_k \in \mathbb{R}^{m\times k}$, $g(z) := A\phi(Bz)$. From $f(x) = P_k x$ a.s, we have $g(z) = z$ for almost every $z \in \mathbb{R}^k$.

Let $\mathcal{H} := \bigcup_{i=1}^m \{z : b_i^\top z = 0\}$ denote the union of the ReLU hyperplanes, where $B = [b_1, \ldots, b_m]^\top$. As a finite union of hyperplanes, $\mathcal{H}$ has Lebesgue measure zero. Since the distribution of z is absolutely continuous with respect to Lebesgue measure, $\mathbb{P}(z \in \mathcal{H}) = 0$. Hence with probability one, z lies in the interior of some full-dimensional activation region $\mathcal{R}_\mathcal{I}$ (see Eq. (1) for the definition). On each $\mathcal{R}_\mathcal{I}$, the map g is linear and equals $g(z) = AD_\mathcal{I}Bz$. Moreover, $g(z) = z$ on a subset of $\mathcal{R}_\mathcal{I}$ of positive measure. Since both sides are linear on $\mathcal{R}_\mathcal{I}$, it follows that

$$AD_\mathcal{I}B = I_k \qquad \text{for every full-dimensional cell } \mathcal{R}_\mathcal{I} \text{ with positive measure.} \tag{2}$$

Now choose any z_0 in the interior of such $\mathcal{R}_\mathcal{I}$ and also such that $b_i^\top z_0 \neq 0$ for all i (this holds for all interior points of $\mathcal{R}_\mathcal{I}$). Let

$$\mathcal{I}_+(z_0) := \{i \mid b_i^\top z_0 > 0\}, \qquad \mathcal{I}_-(z_0) := \{i \mid b_i^\top z_0 < 0\}.$$

On the cell containing z_0, the active set is $\mathcal{I}_+(z_0)$, and Eq. (2) yields

$$\sum_{i\in\mathcal{I}_+(z_0)} a_i b_i^\top = I_k, \tag{3}$$

where $A = [a_1, \ldots, a_m]$. Similarly, replacing z_0 by $-z_0$ flips all signs, so the active set at $-z_0$ is exactly $\mathcal{I}_-(z_0)$. Applying the same argument on the region containing $-z_0$ gives

$$\sum_{i\in\mathcal{I}_-(z_0)} a_i b_i^\top = I_k. \tag{4}$$

Each matrix $a_i b_i^\top$ has rank at most 1. Therefore, Eq. (3) implies $|\mathcal{I}_+(z_0)| \geq k$, and Eq. (4) implies $|\mathcal{I}_-(z_0)| \geq k$. Since $\mathcal{I}_+(z_0)$ and $\mathcal{I}_-(z_0)$ are disjoint and partition $[m]$, we conclude

$$m = |\mathcal{I}_+(z_0)| + |\mathcal{I}_-(z_0)| \geq k + k = 2k.$$

□

Lemma B.9 (ReLU Features form bi-orthogonality). *Under the setting of the lemma B.3, suppose that $(a_i)_{i=1}^k$ and $(b_j)_{j=1}^k$ are weight vectors for the decoder and the encoder, respectively. Then, it satisfies biorthogonality, i.e., $\langle b_j, a_i\rangle = \delta_{ji}$.*

(Proof of Lemma B.9). $\mathbb{J} = I_k = \sum_{i\in[k]} a_i b_i^\top$ For all $z \in \mathbb{R}^k$, it satisfies that $z = I_k z = \sum_{i\in[k]} a_i \langle b_i, z\rangle$. Left multiplying $b_j^\top$ gives us that $\langle b_j, z\rangle = \sum_{i\in[k]} \langle b_j, a_i\rangle\langle b_i, z\rangle$. Note that this satisfies for all z, which gives us that $\langle b_j, a_i\rangle = \delta_{ji}$. □

B.2.2 Proof of Proposition B.5

Let $P_k = U_k U_k^\top$. Decompose $W = WP_k + W(I - P_k)$. Since $(I - P_k)U_k = 0$, the term $W(I - P_k)$ does not affect WU_k and hence does not affect the reduced map on z. Moreover,

$$\|W\|_F^2 = \|WP_k\|_F^2 + \|W(I - P_k)\|_F^2 \geq \|WP_k\|_F^2,$$

so any minimum-norm solution must satisfy $W^*(I - P_k) = 0$, i.e, $W^* = W^* P_k$. Equivalently, each row of W^* lies in $\mathrm{Span}(U_k)$, so there exists $\widetilde{B} \in \mathbb{R}^{2k \times k}$ such that $W^* = \widetilde{B} U_k^\top$. Thus $W^* U_k = \widetilde{B}$. Similarly, writing $\widetilde{A} := U_k^\top V^* \in \mathbb{R}^{k \times 2k}$, the reduced map is

$$g(z) = \widetilde{A}\,\phi(\widetilde{B} z), \qquad \text{and} \qquad g(z) = z \text{ a.e. in } \mathbb{R}^k.$$

By Lemma B.3, after permuting hidden units there exist vectors $a_1, \ldots, a_k \in \mathbb{R}^k$, $b_1, \ldots, b_k \in \mathbb{R}^k$ and scalars $\lambda_1, \ldots, \lambda_k > 0$ such that

$$\widetilde{B} = \begin{bmatrix} b_1^\top \\ \vdots \\ b_k^\top \\ -\lambda_1 b_1^\top \\ \vdots \\ -\lambda_k b_k^\top \end{bmatrix}, \qquad \widetilde{A} = \begin{bmatrix} a_1 & \cdots & a_k & -\lambda_1^{-1} a_1 & \cdots & -\lambda_k^{-1} a_k \end{bmatrix}. \tag{5}$$

Moreover, on any full-dimensional activation region the Jacobian equals I_k, which yields the rank-k identity decomposition $\sum_{i=1}^k a_i b_i^\top = I_k$, and hence the biorthogonality relations $\langle a_i, b_j \rangle = \delta_{ij}$ for $i, j \in [k]$.

Fix such a paired representation. Its contribution to the squared Frobenius norms satisfies

$$\|\widetilde{B}\|_F^2 + \|\widetilde{A}\|_F^2 = \sum_{i=1}^k \Big((1 + \lambda_i^2)\|b_i\|_2^2 + (1 + \lambda_i^{-2})\|a_i\|_2^2 \Big).$$

For each i, by the arithmetic-geometric mean(AM-GM) we have

$$(1 + \lambda_i^2)\|b_i\|_2^2 + (1 + \lambda_i^{-2})\|a_i\|_2^2 \geq 2\sqrt{(1 + \lambda_i^2)(1 + \lambda_i^{-2})}\,\|a_i\|_2 \|b_i\|_2.$$

Since $\sqrt{(1 + \lambda^2)(1 + \lambda^{-2})} \geq 2$ for all $\lambda > 0$ with equality iff $\lambda = 1$, it follows that

$$(1 + \lambda_i^2)\|b_i\|_2^2 + (1 + \lambda_i^{-2})\|a_i\|_2^2 \geq 4\|a_i\|_2 \|b_i\|_2,$$

with equality iff $\lambda_i = 1$ and $\|a_i\|_2 = \|b_i\|_2$. Therefore, at a minimum-norm solution we must have

$$\lambda_i = 1 \quad \text{and} \quad \|a_i\|_2 = \|b_i\|_2 \qquad \text{for all } i \in [k]. \tag{6}$$

Next, using $\sum_{i=1}^k a_i b_i^\top = I_k$, we have

$$\mathrm{Tr}(I_k) = k = \mathrm{Tr}\Big(\sum_{i=1}^k a_i b_i^\top \Big) = \sum_{i=1}^k \langle a_i, b_i \rangle \leq \sum_{i=1}^k \|a_i\|_2 \|b_i\|_2 \leq \frac{1}{2} \sum_{i=1}^k (\|a_i\|_2^2 + \|b_i\|_2^2),$$

where the first inequality is Cauchy-Schwarz and the second is AM-GM. At the minimum-norm solution, both inequalities must be tight. Thus $a_i = c b_i$ for $c > 0$, combined with $\langle a_i, b_i \rangle = 1$ and Eq. (6) yields $a_i = b_i$ and $\|a_i\|_2 = \|b_i\|_2 = 1$. Thus the vectors $(a_i)_{i=1}^k$ form an orthonormal basis of $\mathbb{R}^k$. In other words, there exists an orthogonal matrix $Q \in \mathbb{R}^{k \times k}$ such that $a_i = Q e_i$, $b_i = Q e_i$, for $i = 1, \ldots, k$. Substituting $\lambda_i = 1$ and $a_i = b_i = Q e_i$ into Eq. (5) and recalling $W^* = \widetilde{B} U_k^\top$ and $V^* = U_k \widetilde{A}$, we obtain

$$W^* = \begin{bmatrix} Q^\top U_k^\top \\ -Q^\top U_k^\top \end{bmatrix}, \qquad V^* = \begin{bmatrix} U_k Q & -U_k Q \end{bmatrix},$$

as claimed. □

B.2.3 Proof of Proposition B.6

Expanding the loss function, we have $L(W) = \frac{1}{2}\left(\mathbb{E}\|x\|^2 - 2\mathbb{E}\langle x, f_W(x)\rangle + \mathbb{E}\|f_W(x)\|^2\right) + \frac{\lambda}{2}\|W\|_F^2$. We first compute $\mathbb{E}\langle x, f_W(x)\rangle$. Note that $\mathbb{E}\langle x, f_W(x)\rangle = \mathbb{E}\left\langle x, \sum_{i=1}^{2r} w_i \phi(w_i^\top x)\right\rangle = \sum_{i=1}^{2r} \mathrm{Tr}(w_i \mathbb{E}[\phi(w_i^\top x)x^\top])$. Using the identity $\phi(w_i^\top x) = \frac{1}{2}(w_i^\top x + |w_i^\top x|)$, we obtain

$$\begin{aligned}\mathbb{E}\langle x, f_W(x)\rangle &= \frac{1}{2}\sum_{i=1}^{2r} \mathrm{Tr}(w_i \mathbb{E}[w_i^\top x x^\top + |w_i^\top x| x^\top]) \\ &= \frac{1}{2}\sum_{i=1}^{2r} \mathrm{Tr}(w_i w_i^\top \Sigma_x + \mathbb{E}[|w_i^\top x| x^\top]).\end{aligned}$$

Since $x =_d -x$, we have $\mathbb{E}[|w_i^\top x| x^\top] = \mathbb{E}[|w_i^\top x|(-x)^\top]$, and therefore $\mathbb{E}[|w_i^\top x| x^\top] = 0$. Hence, denoting $G_W := WW^\top$, we have $\mathbb{E}\langle x, f_W(x)\rangle = \frac{1}{2}\sum_{i=1}^{2r} \mathrm{Tr}(w_i w_i^\top \Sigma_x) = \frac{1}{2}\mathrm{Tr}(G_W \Sigma_x)$.

For the term $\mathbb{E}\|f_W(x)\|^2$, we use the parallelogram inequality: for $a, b \in \mathbb{R}^d$, $\frac{1}{2}(\|a\|^2 + \|b\|^2) \geq \frac{1}{4}\|a-b\|^2$, with equality if and only if $\|a+b\| = 0$. This gives the lower bound

$$\begin{aligned}\mathbb{E}\|f_W(x)\|^2 &= \frac{1}{2}\mathbb{E}[\|f_W(x)\|^2 + \|f_W(-x)\|^2] \\ &\geq \frac{1}{4}\mathbb{E}[\|f_W(x) - f_W(-x)\|^2] = \frac{1}{4}\mathbb{E}\|G_W x\|^2.\end{aligned}$$

The first equality follows from $x =_d -x$, and the last equality follows from

$$f_W(x) - f_W(-x) = W(\phi(W^\top x) - \phi(-W^\top x)) = WW^\top x.$$

Combining the above results, we obtain the following lower bound on the loss:

$$\begin{aligned}L(W) &= \frac{1}{2}\left(\mathbb{E}\|x\|^2 - 2\mathbb{E}\langle x, f_W(x)\rangle + \mathbb{E}\|f_W(x)\|^2\right) + \frac{\lambda}{2}\|W\|_F^2 \\ &\geq \frac{1}{2}\mathrm{Tr}(\Sigma_x) - \frac{1}{2}\mathrm{Tr}(\Sigma_x G_W) + \frac{1}{8}\mathrm{Tr}(\Sigma_x G_W^2) + \frac{\lambda}{2}\mathrm{Tr}(G_W). \qquad (7)\end{aligned}$$

We denote the right-hand side by $F(G_W)$. Note that inequality becomes an equality if and only if $f_W(x) + f_W(-x) = 0$ a.s. We proceed as follows. First, we minimize the lower bound $F(G_W)$, which gives the condition on $G_W^* := \arg\min_{G_W} F(G_W)$. Thus, any W satisfying $WW^\top = G_W^*$ attains the lowest possible value of the lower bound. This bound is achievable if and only if such a W also satisfies the equality condition $f_W(x) + f_W(-x) = 0$ a.s., which then yields a global minimizer.

We therefore first minimize $F(G_W)$. Without loss of generality, we work on the eigenbasis of Σ_x. We denote

$$\xi_i = \begin{cases}\nu_i, & i \leq r, \\ \tau_{i-r}, & i > r.\end{cases}$$

Note that the off-diagonal terms of G_W only appears in the the third term of Eq. 7. Using the symmetry of G_W, we have that $\mathrm{Tr}(\Sigma_x G_W^2) = \sum_{i,j} \xi_i (G_W)_{ij}^2$. Therefore, the off-diagonal entries appear only quadratically with positive coefficient $\frac{1}{8}(\xi_i + \xi_j)(G_W)_{ij}^2$. In order to make this term vanished, there are two cases: in the case of $\xi_i + \xi_j > 0$, $(G_W)_{ij} = 0$ and the other case is where $\xi_i + \xi_j = 0$, meaning $\xi_i = \xi_j = 0$. Therefore, remaining part is the diagonal entries. A straighforward calculation shows that for each diagonal entry $g \geq 0$ of G_W, which corresponds to covariance eigenvalue ξ, the scalar objective is $\kappa_\xi(s) = -\frac{1}{2}\xi g + \frac{1}{8}\xi g^2 + \frac{\lambda}{2} g$. Its minimizer over $g \geq 0$ is $g_\xi^* = 2\left(1 - \frac{\lambda}{\xi}\right)_+$, where $a_+ := \max(a, 0)$ for $a \in \mathbb{R}$. Since $\tau_{\max} < \lambda < \nu_{\min}$, we obtain

$$g_{\nu_i}^* = 2\left(1 - \frac{\lambda}{\nu_i}\right), \quad i = 1, \ldots, r,$$

and $g_{\nu_j}^* = 0$ for $j = 1, \ldots, d-r$. Thus the minimizer of $F(G_W)$ is $G_W^* = 2Q_r D^2 Q_r^\top$. With this, the conditions for W to be a global minimizer of $L(W)$ are

$$W_* = \arg\min_W L(W) \quad \Leftrightarrow \quad WW^\top = 2Q_r D^2 Q_r^\top, \text{ s.t. } f(x; W) + f(-x; W) = 0 \text{ a.s.}$$

It remains to analyze the condition $f(x;W)+f(-x;W)=0$ a.s. Note that

$$f(x;W)+f(-x;W)=\sum_{i=1}^{2r} w_i\big(\phi(w_i^\top x)+\phi(-w_i^\top x)\big)=\sum_{i=1}^{2r} w_i|w_i^\top x|.$$

Since $WW^\top = 2Q_rD^2Q_r^\top$, each weight vector satisfies $w_j \in \mathrm{Span}(Q_r)$. Hence, there exists $a_j \in \mathbb{R}^r$ such that $w_j = Q_r a_j$. We write the distinct directions of the weight vectors as $\{u_1,\dots,u_l\}$, where each $u_i \in \mathbb{R}^r$ has unit norm $\|u_i\|=1$. For each $j\in[l]$, define

$$\mathcal{I}_j := \{k\in[2r] \,|\, a_k = c_k u_j,\ c_k \neq 0\}.$$

Thus, each $\mathcal{I}_j$ collects the indices of the weight vectors with the same direction as u_j. We also define the partition

$$\mathcal{I}_j^+ := \{k\in\mathcal{I}_j \mid c_k>0\}, \qquad \mathcal{I}_j^- := \{k\in I_j \mid c_k<0\}.$$

Note that $\mathcal{I}_j = \mathcal{I}_j^+ \uplus \mathcal{I}_j^-$. With this notation, for $x = Q_r z + Q_\perp \varepsilon$, we have

$$\begin{aligned}\sum_{i=1}^{2r} w_i|w_i^\top x| &= Q_r \sum_{i=1}^{l}\sum_{j\in I_i} c_j|c_j|(u_i|u_i^\top z|)\\ &= Q_r \sum_{i=1}^{l}\left(\sum_{j\in I_i^+} c_j^2 - \sum_{j\in I_i^-} c_j^2\right) u_i|u_i^\top z|.\end{aligned}$$

We denote the function above by $F(z)$. We claim that $b_i := \sum_{j\in I_i, c_j>0} c_j^2 - \sum_{j\in I_i, c_j<0} c_j^2 = 0$, for every $i\in[l]$.

Once we show this, we have the following consequence: Since $\mathrm{rank}(WW^\top)=r$, we have $l\geq r$. Moreover, since $c_j\neq 0$ for all $j\in[2r]$, in order to make $b_i=0$, it satisfies that $|I_i|\geq 2$, with $|\mathcal{I}_i^+|\geq 1$ and $|\mathcal{I}_i^-|\geq 1$. This means that for each direction u_j, there are at least two weight vectors that share the same direction up to scalar multiplication. Since there are $2r$ weight vectors in total, it follows that $l=r$ and $|\mathcal{I}_i^+| = |\mathcal{I}_i^-| = 1$ for each $i\in[l]$. This directly implies that there are r pairs of weight vectors, where each pair consists of two weight vectors with the same norm and opposite directions. Thus, up to permutation, we may write

$$W = Q_r\,[a_1 \quad \dots \quad a_r \quad -a_1 \quad \dots \quad -a_r] \in \mathbb{R}^{d\times 2r},$$

for $a_i = c_i u_i$. To determine the coefficients, we use the fact that $WW^\top = 2Q_rD^2Q_r^\top$. Since $WW^\top = 2Q_rAA^\top Q_r^\top$, for $A=[a_1,\dots,a_r]\in\mathbb{R}^{r\times r}$, we have $AA^\top = D^2$. Since D is positive definite, it satisfies that $A = DR$, for some $R\in\mathrm{O}(r)$. Therefore, up to permutation, $W = [Q_rDR, -Q_rDR]$ as claimed.

It remains to show that each $b_i=0$. The proof idea is almost the same as the argument in the proof of Lemma B.3, however we repeat this for completeness. Fix $i\in[l]$. Define, for $j\in[l]$, $\mathcal{N}_j = \{p\in\mathbb{R}^r \mid u_i^\top p = 0,\ p\neq 0\}$. Choose $p_0\in\mathcal{H}_i\setminus\bigcup_{j\neq i}\mathcal{H}_j$. This means that $u_i^\top p_0 = 0$, while $u_j^\top p_0\neq 0$ for every $j\neq i$. Now define $\delta_j := |u_j^\top p_0|>0$ for $j\neq i$, and let $\delta := \frac{1}{2}\min_{j\neq i}\delta_j > 0$. For any $p\in\mathbb{R}^r$ satisfying $\|p-p_0\|<\delta$, we have

$$|u_j^\top(p-p_0)| \leq \|p-p_0\| < \delta \leq \frac{1}{2}|u_j^\top p_0|.$$

This implies $\mathrm{sgn}(u_j^\top p) = \mathrm{sgn}(u_j^\top p_0)$, because a change of sign would require $|u_j^\top(p-p_0)| > |u_j^\top p_0|$, which is a contradiction. Hence, $\mathrm{sgn}(u_j^\top p) = \mathrm{sgn}(u_j^\top p_0)$ for all $p\in B(p_0,\delta)$ and all $j\neq i$. Therefore, inside the ball $B(p_0,\delta)$, the only possible sign change is that of $u_i^\top p$.

Now define the two open sets $\mathcal{O}_i^+ := B(p_0,\delta)\cap\{p\in\mathbb{R}^r \mid u_i^\top p>0\}$ and $\mathcal{O}_i^- := B(p_0,\delta)\cap\{p\in\mathbb{R}^r \mid u_i^\top p<0\}$. Then, on $p\in\mathcal{O}_i^+$, we have $\mathrm{sgn}(u_i^\top p)=1$, while on $p\in\mathcal{O}_i^-$, we have $\mathrm{sgn}(u_i^\top p)=-1$. Thus, the sign change occurs only for u_i.

Recall that $F(z) = Q_r\sum_{i=1}^l b_i u_i|u_i^\top z|$. On $z\in\mathcal{O}_i^+$, we have $F_+(z) = b_iQ_ru_iu_i^\top z + Q_r\sum_{j\neq i} b_ju_j|u_j^\top z|$. On the other hand, on $z\in\mathcal{O}_i^-$, we have $F_-(z) = -b_iQ_ru_iu_i^\top z + Q_r\sum_{j\neq i} b_ju_j|u_j^\top z|$. Since $F_+(z)=0$ on the open set $\mathcal{O}_i^+$, F_+ is the zero map. The same argument applies to F_-. Therefore, subtracting the two maps gives $F_+(z)-F_-(z) = 2b_iQ_ru_iu_i^\top z = 0$, for all $z\in\mathbb{R}^r$. Hence, $b_iQ_ru_iu_i^\top = 0$, which implies $b_i=0$. Repeating the same argument for every $k\in[l]$, we conclude that $b_k=0$ for all $k\in[l]$. This completes the proof. □

C IMPLICIT BIAS OF SGD NEAR NASM

In this section, we develop general results for stochastic gradient descent near a compact NASM $\mathcal{M} \subset \mathbb{R}^D$ for the pGF defined by a population loss L, without specifying a specific architecture or data distribution other than the assumptions stated in Section C.3. Subsequent sections verify the required assumptions for the ReLU and linear autoencoders studied in this paper.

Section C.2 first introduces normally hyperbolic invariant manifolds (NHIMs), the stable-foliation theorem, and a few consequences from the theorem. We present the main result in Section C.4. The analysis of the main theorem consists of two stages. First, we show that SGD initialized in a compact subset of a stable neighborhood reaches an $O_{\mathbb{P}}(\sqrt{\eta})$ normal layer around $\mathcal{M}$ while preserving the asymptotic phase selected by the pGF (Section C.5). Second, we average the fast normal fluctuations to derive the effective motion of the projected iterates along $\mathcal{M}$ on the $O(\eta^{-2})$ iteration scale (Section C.6).

C.1 NOTATIONS

We denote $\operatorname{dist}(z, \mathcal{Y}) := \inf_{v \in \mathcal{M}} \|z - v\|$. For a smooth embedded submanifold $\mathcal{M} \subset \mathbb{R}^D$, and for $\zeta > 0$, we denote the tubular neighborhood of $\mathcal{M}$ by $\mathcal{U}_\zeta := \{z \in \mathbb{R}^D \mid \operatorname{dist}(z, \mathcal{M}) < \zeta\}$. We denote the sequence of random variables $(x_n)_{n\in\mathbb{N}}$ is $O_{\mathbb{P}}(a)$ if, for any $\epsilon > 0$, there exists a finite constant $M > 0$ and $N \in \mathbb{N}$ such that for all $n > N$, $\mathbb{P}\left(\left|\frac{x_n}{a}\right| > M\right) < \epsilon$. We write $R_\eta = o_{\mathbb{P}}(1)$ if $R_\eta \to_{\mathbb{P}} 0$ as $\eta \downarrow 0$. For a vector $v \in \mathbb{R}^D$, $\|v\| := \|v\|_2$ and for a matrix $M \in \mathbb{R}^{D_1 \times D_2}$, we denote $\|M\| := \|M\|_2$. For vectors $v_1, v_2 \in \mathbb{R}^D$, we write $v_1 \otimes v_2 := v_1 v_2^\top$ and $v_1^{\otimes 2} := v_1 v_1^\top$.

C.2 THEORY OF NHIM

Definition C.1 (Normally Attracting Invariant Manifold, see (Eldering et al., 2013)). *Let $\dot{z} = f(z)$ be a C^1 dynamical system with flow φ_t. A smooth submanifold $\mathcal{M}$ embedded in the ambient space $\mathcal{P}$ is invariant, if $z \in \mathcal{M} \Rightarrow \varphi_t(z) \in \mathcal{M}$, for all $t \geq 0$. Assume that, along $\mathcal{M}$, the ambient tangent space admits a continuous splitting $T_z P = T_z\mathcal{M} \oplus N_z\mathcal{M}$. The invariant manifold $\mathcal{M}$ is normally attracting if there exist constants $C > 0$, $\lambda > 0$, and $\alpha \geq 0$, with $\alpha < \lambda$, such that for all $t \geq 0$, $v \in T_z\mathcal{M}$, and $n \in N_z\mathcal{M}$, it satisfies that $\|D\varphi_t(z)n\| \leq Ce^{-\lambda t}\|n\|$ and $\|D\varphi_t(z)v\| \leq Ce^{\alpha t}\|v\|$, uniformly over $z \in \mathcal{M}$.*

Importantly, perturbations transverse to $\mathcal{M}$ dominates perturbations tangent to $\mathcal{M}$.

Theorem C.2 (Stable Foliation for NAIM, see Hirsch et al. (1970); Wiggins (2013)). *Let $\mathcal{M}$ be a compact normally attracting invariant manifold (without boundary) of a C^r flow φ_t on an ambient manifold $\mathcal{P}$, with associated invariant splitting $T_{\mathcal{M}}\mathcal{P} = T\mathcal{M} \oplus E^s$, where $T\mathcal{M}, E^s$ are the tangent- and stable bundle of $\mathcal{M}$, respectively. Then there exists a neighborhood $\mathcal{U}$ of $\mathcal{M}$ and, for every $v \in \mathcal{M}$, a local stable fiber $W^s_{\mathrm{loc}}(v) \subset \mathcal{U}$ such that:*

1. *Each $W^s_{\mathrm{loc}}(v)$ is a C^k (where $1 \leq k \leq r$) embedded submanifold satisfying $T_v W^s_{\mathrm{loc}}(v) = E^s_v$ for each $v \in \mathcal{M}$.*
2. *The stable fibers form a local foliation of $\mathcal{U}$: $\mathcal{U} = \bigcup_{v\in\mathcal{M}} W^s_{\mathrm{loc}}(v)$, and every point $w \in \mathcal{U}$ belongs to a unique stable fiber.*
3. *The foliation is equivariant under the flow: $\varphi_t\left(W^s_{\mathrm{loc}}(v)\right) \subseteq W^s_{\mathrm{loc}}\left(\varphi_t(v)\right)$ whenever $\varphi_t(v)$ and the flowed fiber remain inside $\mathcal{U}$.*
4. *There exist constants $C > 0$ and $\lambda > 0$ such that, for every $w \in W^s_{\mathrm{loc}}(v)$ and $v \in \mathcal{M}$, $\operatorname{dist}\left(\varphi_t(w), \varphi_t(v)\right) \leq Ce^{-\lambda t} \operatorname{dist}(w, v)$ for all $t \geq 0$.*

Consequently, the union of the s-dimensional stable fibers over the m-dimensional manifold $\mathcal{M}$ has dimension $m + s = \dim \mathcal{P}$, so in the normally attracting case the local stable manifold of $\mathcal{U}$ forms a full-dimensional neighborhood of $\mathcal{M}$. Nearby trajectories therefore exponentially shadow unique trajectories on $\mathcal{M}$. Due to this fact, the crucial aspect of NAIM is that one can analyze the reduced invariant dynamics on the manifold. However, in our case, the manifold is *stationary*, meaning that $\varphi_t(v) = v$ for $v \in \mathcal{M}$.

Geometric consequences used in the burn-in (Phase 1) analysis. Let $\mathcal{M} \subset \mathbb{R}^D$ be a compact C^4 normally attracting stationary manifold of the population gradient flow $\dot{z} = -\nabla L(z)$, and let $\varphi_t(z)$ denote the corresponding flow. Assume that the stable-foliation theorem applies and yields a stable-foliation neighborhood $\mathcal{U}_{\mathrm{st}}$ of $\mathcal{M}$. Let $\Phi : \mathcal{U}_{\mathrm{st}} \to \mathcal{M}$ denote the associated asymptotic phase map, defined by $\Phi(w) = v \iff w \in W^s_{\mathrm{loc}}(v)$. We have the following consequences from the stable foliation and NHIM structure.

(C1). (Exponential contraction to the asymptotic phase). For $v \in \mathcal{M}$ and $w \in W^s_{\mathrm{loc}}(v)$, following from the stationarity of $\mathcal{M}$, $\varphi_t(v) = v$. Due to the above theorem, by replacing $v = \Phi(w)$, it follows that

$$\|\varphi_t(w) - \Phi(w)\| \le C_{\mathrm{st}} e^{-\lambda_{\mathrm{st}} t} \|w - \Phi(w)\|. \tag{8}$$

Moreover, $\Phi(\varphi_t(w)) = \Phi(w)$ due to the phase invariance of stable-foliation.

(C2). (The phase map annihilates the population gradient). Phase invariance gives $\Phi(\varphi_t(z)) = \Phi(z)$. Differentiating at $t = 0$ yields

$$0 = D\Phi(z) \left.\frac{d}{dt}\right|_{t=0} \varphi_t(z) = -D\Phi(z)\nabla L(z).$$

Therefore, $D\Phi(z)\nabla L(z) = 0$, for $z \in \mathcal{U}_{\mathrm{st}}$.

(C3). (Uniform normal coercivity). Normal attraction of the stationary manifold implies that there exists $c_* > 0$ such that

$$\langle n, \nabla^2 L(v) n \rangle = \langle n, H_N(v) n \rangle \ge c_* \|n\|^2 \tag{9}$$

for every $v \in \mathcal{M}$ and $n \in N_v\mathcal{M}$. The constant can be chosen uniformly because $\mathcal{M}$ is compact.

C.3 General setting and assumptions

SGD setting We briefly recall the SGD setting and notations (see Section 3). Let $L : \mathbb{R}^D \to \mathbb{R}$ be a population loss and consider the population gradient flow $\dot{z} = -\nabla L(z)$, $\varphi_0(z) = z$. The corresponding SGD recursion is

$$z^\eta_{k+1} = z^\eta_k - \eta g(z^\eta_k; x_{k+1}), \tag{10}$$

where $(x_k)_{k\ge1}$ are i.i.d. samples. We denote $g_{k+1} := g(z^\eta_k; x_{k+1})$ and $\xi_{k+1} := g_{k+1} - \nabla L(z^\eta_k)$.

Standing geometric setting Let $\mathcal{M} \subset \mathbb{R}^D$ be a compact C^4 NASM of the population gradient flow. Let $\mathcal{U}_{\mathrm{st}}$ be a stable-foliation neighborhood of $\mathcal{M}$. After decreasing $\zeta > 0$ if necessary, let $\mathcal{U}_\zeta := \{z \in \mathbb{R}^D : \mathrm{dist}(z, \mathcal{M}) < \zeta\}$ be a tubular neighborhood satisfying $\overline{\mathcal{U}_\zeta} \subset \mathcal{U}_{\mathrm{st}}$, on which the nearest-point projection $\pi : \mathcal{U}_\zeta \to \mathcal{M}$ is uniquely defined and C^3. For $v \in \mathcal{M}$, let $P_T(v)$ and $P_N(v)$ denote the orthogonal projections onto $T_v\mathcal{M}$ and $N_v\mathcal{M}$, respectively.

Tubular Neighborhood By Lemma C.3, one can construct a tubular neighborhood with the desired properties stated in the lemma, including the above well-defined nearest-point projection discussed above (this mainly follows from the Tubular Neighborhood Theorem (Lee, 2003)). Throughout the remainder of the paper, any set $\mathcal{U}_\zeta$ is understood to be chosen in this way.

Lemma C.3 (Construction of tubular neighborhood $\mathcal{U}_\zeta$). *Let $\mathcal{M}$ be a compact NASM. Then, there exist constants $\zeta > 0$ and $c > 0$ such that*

$$\mathcal{U}_\zeta := \left\{z \in \mathbb{R}^D \mid \mathrm{dist}(v, \mathcal{M}) < \zeta\right\}$$

is a tubular neighborhood of $\mathcal{M}$ with the following properties: Every $z \in \mathcal{U}_\zeta$ satisfies a unique decomposition $z = v + n$, $v \in \mathcal{M}$, $n \in N_\tau\mathcal{M}$, with $\mathrm{dist}(z, \mathcal{M}) = \|n\|_2$. Moreover, $\langle \nabla L(z), n\rangle \ge c\|n\|_2^2$, for every $z = v + n \in \mathcal{U}_\zeta$.

(Proof). The proof can be found in Section C.5.2. □

We next state the regularity assumptions used in the subsequent stochastic reduction. In contrast to Li et al. (2021), we do not require the population loss to be C^3 in an ambient neighborhood of the minimizer manifold. Instead, we control the variation of the quadratic normal curvature as the base point moves along $\mathcal{M}$.

Assumption C.4 (Population loss regularity). *The following conditions hold.*

(P1). *The gradient ∇L is locally Lipschitz on $\mathcal{U}_{\mathrm{st}}$, and $L \in C^2$ on an open neighborhood of $\overline{\mathcal{U}_\zeta}$.*

(P2). (Tangential regularity of the normal Hessian.). *Define*

$$H_N(v) := P_N(v)\nabla^2 L(v) P_N(v)\big|_{N_v\mathcal{M}}.$$

The operators $H_N(v) : N_v\mathcal{M} \to N_v\mathcal{M}$ form a C^1 family along $\mathcal{M}$.[10]

(P3). (Negligible tangential derivative of the remainder.). *For $v \in \mathcal{M}$ and $n \in N_v\mathcal{M}$, define*

$$R(v,n) := L(v+n) - L(v) - \frac{1}{2}\langle n, H_N(v)n\rangle.$$

There exists a function $\omega_R : [0,\zeta] \to [0,\infty)$ satisfying $\lim_{r\downarrow 0}\omega_R(r) = 0$ such that the following holds uniformly. Let $v \in \mathcal{M}$, $\tau \in T_v\mathcal{M}$, and $n \in N_v\mathcal{M}$ with $\|n\| \le \zeta$. If $v : (-\varepsilon,\varepsilon) \to \mathcal{M}$ is a C^1 curve satisfying $v(0) = v$, $\dot v(0) = \tau$, and $n(s) \in N_{v(s)}\mathcal{M}$ is the parallel normal transport of n (meaning that $n(0) = n$, $P_N(v(s))\dot n(s) = 0$), then

$$\left|\frac{d}{ds}\right|_{s=0} R\big(v(s), n(s)\big)\right| \le \omega_R(\|n\|)\,\|\tau\|\,\|n\|^2. \tag{11}$$

Condition (P3) is a uniform $o(\|\tau\|\|n\|^2)$ condition as $\|n\| \to 0$. Thus, instead of controlling every ambient third-order derivative of L, we control only the change in the quadratic normal approximation when the base point moves along $\mathcal{M}$.

Assumption C.5 (Sample gradient regularity). *The following conditions hold.*

(S1). (Moment bounds). *For some integer $p > 4$ and every compact set $\mathcal{C} \Subset \mathcal{U}_{\mathrm{st}}$,*

$$\sup_{z\in\mathcal{C}} \mathbb{E}\|g(z;x)\|^{2p} < \infty. \tag{12}$$

(S2). (Local L^4 continuity near the manifold).

$$\sup_{\substack{z\in\mathcal{U}_\zeta \\ \mathrm{dist}(z,\mathcal{M})\le r}} \mathbb{E}\,\|g(z;x) - g(\pi(z);x)\|^4 \longrightarrow 0 \qquad \text{as } r \downarrow 0. \tag{13}$$

(S3). (Absence of leading tangential noise). *For every $v \in \mathcal{M}$, $P_T(v)g(v;x) = 0$ almost surely.*

(S4). (Regularity of the normal noise covariance and curvature tensor). *Define the normal noise covariance by $\Sigma_N(v) := \mathbb{E}\,[g(v;x)\otimes g(v;x)]$. For $n_1, n_2 \in N_v\mathcal{M}$, let $\mathcal{B}_v[n_1,n_2] \in T_v\mathcal{M}$ be the tangent vector uniquely determined by*

$$\langle \mathcal{B}_v[n_1,n_2], \tau\rangle = \left\langle n_1, (\nabla^\perp_\tau H_N)(v)n_2\right\rangle, \qquad \tau \in T_v\mathcal{M}. \tag{14}$$

We assume that the maps $v \mapsto \Sigma_N(v)$ and $v \mapsto \mathcal{B}_v$ are locally Lipschitz along $\mathcal{M}$.

C.4 Main Theorem

Theorem C.6 (Stochastic reduction from a stable-foliation neighborhood). *Let $\mathcal{M} \subset \mathbb{R}^D$ satisfy the standing geometric conditions, and let $\mathcal{U}_{\mathrm{st}}$ be its stable-foliation neighborhood with asymptotic phase map $\Phi : \mathcal{U}_{\mathrm{st}} \to \mathcal{M}$. Suppose that Assumptions* P1-P3 *and* S1-S4 *hold. Consider the SGD recursion $z^\eta_{k+1} = z^\eta_k - \eta g(z^\eta_k; x_{k+1})$, where $(x_k)_{k\ge1}$ are i.i.d. Let $\mathcal{K} \Subset \mathcal{U}_{\mathrm{st}}$ be compact, and let z_0 be independent of $(x_k)_{k\ge1}$ and satisfy $\mathbb{P}(z_0 \in \mathcal{K}) = 1$. For every $\eta > 0$, we initialize $z^\eta_0 = z_0$.*

For each $v \in \mathcal{M}$, let $\Pi_v : N_v\mathcal{M} \to N_v\mathcal{M}$ be the unique solution of the Lyapunov equation $H_N(v)\Pi_v + \Pi_v H_N(v) = \Sigma_N(v)$, and define $b(v) := \sum_{a=1}^{\dim N_v\mathcal{M}} \mathcal{B}_v[e_a, \Pi_v e_a]$, where $\{e_a\}_{a=1}^{\dim N_v\mathcal{M}}$ is any orthonormal basis of $N_v\mathcal{M}$.

[10] More precisely, let $v(s) \subset \mathcal{M}$ be a smooth curve and let $\mathsf{P}_s : N_{v(0)}\mathcal{M} \to N_{v(s)}\mathcal{M}$ denote parallel transport in the normal bundle. The assumption means that $\mathsf{P}_s^{-1} H_N(v(s))\mathsf{P}_s$ varies C^1-smoothly with s.

Then there exists a deterministic burn-in time $\mathtt{t}_\eta = O\left(\eta^{-1}\log(\eta^{-1})\right)$ *such that* $\eta^2\mathtt{t}_\eta \longrightarrow 0$. *Moreover, for every fixed* $T > 0$, *with probability tending to one,* $z^\eta_{\mathtt{t}_\eta+k} \in \mathcal{U}_\zeta$ *for every* $0 \le k \le \left\lfloor \frac{T}{\eta^2} \right\rfloor$, *where* $\mathcal{U}_\zeta$ *is the tubular neighborhood used in the late-phase reduction. The shifted projected process*

$$v^\eta(t) := \pi\left(z^\eta_{\mathtt{t}_\eta+\lfloor t/\eta^2\rfloor}\right), \qquad 0 \le t \le T,$$

then satisfies $\sup_{0\le t\le T}\|v^\eta(t)-v(t)\| \xrightarrow{\mathbb{P}} 0$, *where* $v(t)$ *is the unique solution of* $\dot v(t) = -\frac{1}{2}b(v(t))$, *for* $v(0) = \Phi(z_0)$. *All projections above are well defined with probability tending to one.*

(Proof of Theorem C.6). Lemma C.14 provides $\zeta > 0$ such that the projected-gradient expansion

$$D\pi(v+n)\nabla L(v+n) = \frac{1}{2}\mathcal{B}_v[n,n] + r_T(v,n) \tag{15}$$

holds throughout $\mathcal{U}_\zeta$, where $\|r_T(v,n)\| \le C\big(\omega_R(\|n\|) + \|n\|\big)\|n\|^2 = o(\|n\|^2)$ uniformly over $v \in \mathcal{M}$. Fix this tubular neighborhood for the remainder of the proof. We first apply the burn-in result, Lemma C.7, with $\mathcal{U}_\zeta$ as the target tube. It provides a deterministic time, where $\mathtt{t}_\eta = O\left(\eta^{-1}\log(\eta^{-1})\right)$, $\eta^2\ell_\eta \longrightarrow 0$, and events $\mathcal{A}_\eta$ satisfying $\mathbb{P}(\mathcal{A}_\eta) \to 1$ such that, on $\mathcal{A}_\eta$, $z^\eta_{\ell_\eta} \in \mathcal{U}_\zeta$. Define $v^\eta_{\mathrm{in}} := \pi(z^\eta_{\ell_\eta})$, and the rescaled normal variable $q^\eta_{\mathrm{in}} := \eta^{-1/2}\left(z^\eta_{\ell_\eta} - v^\eta_{\mathrm{in}}\right)$. The phase-preservation and high-moment conclusions of the burn-in lemma give $v^\eta_{\mathrm{in}} \xrightarrow{\mathbb{P}} \Phi(z_0)$ and $\mathbb{E}\left[\|q^\eta_{\mathrm{in}}\|^{2p}\mathbf{1}_{\mathcal{A}_\eta}\right] \le C$ (see also Lemma C.9).

As above, define an auxiliary initialization equal to $z^\eta_{\ell_\eta}$ on $\mathcal{A}_\eta$ and equal to a fixed point of $\mathcal{M}$ on $\mathcal{A}_\eta^c$. Its associated base and normal variables satisfy $\widetilde v^\eta_{\mathrm{in}} \xrightarrow{\mathbb{P}} \Phi(z_0)$, $\sup_{\eta>0}\mathbb{E}\|\widetilde q^\eta_{\mathrm{in}}\|^{2p} < \infty$. Thus burn-in supplies precisely the initialization conditions required by Lemma C.10.

At this point, the expansion Eq. (15), together with Assumptions S1-S4, verifies the remaining hypotheses of the late-phase stochastic reduction. Applying Proposition C.10 to the recursion restarted at time $\mathtt{t}_\eta$ yields the claimed convergence to $\dot v(t) = -\frac{1}{2}b(v(t))$, $v(0) = \Phi(z_0)$. Since the auxiliary and original recursions agree on $\mathcal{A}_\eta$, and $\mathbb{P}(\mathcal{A}_\eta^c) \to 0$, the conclusion transfers to the original SGD recursion. □

C.5 SGD Phase 1: Fast relaxation with $\tilde O(\eta^{-1})$ steps and entrance into the $O(\sqrt{\eta})$ tube

Following Lemma C.7 shows that with probability tending to 1, SGD reaches the tubular neighborhood with its radius $O(\sqrt{n})$ around the manifold. While approaching to this neighborhood, it creates $o_{\mathbb{P}}(1)$ tangential movement and it occupies asymptotically zero time on the $O(\eta^{-2})$ time scale.

Lemma C.7 (Burn-in to the stochastic $O(\sqrt{\eta})$ layer). *Suppose the assumptions* P1 *and* S1 *hold. Let* $\mathcal{M} \subset \mathbb{R}^D$ *be the compact* C^4 *NASM of the population gradient flow* $\dot z = -\nabla L(z)$, $\varphi_0(z) = z$. *Let* $\mathcal{U}_{\mathrm{st}}$ *and* $\Phi : \mathcal{U}_{\mathrm{st}} \to \mathcal{M}$ *be the stable-foliation neighborhood and asymptotic phase map, respectively. Let* $\mathcal{K}$ *be any compact set satisfying* $\mathcal{K} \Subset \mathcal{U}_{\mathrm{st}}$ *and assume* $\mathbb{P}(z^\eta_0 \in \mathcal{K}) \longrightarrow 1$.

Then there exists $\delta_* > 0$ *such that the nearest-point projection is uniquely defined and* C^3 *on an open neighborhood of* $\overline{\mathcal{U}_{\delta_*}}$, *where* $\mathcal{U}_{\delta_*} := \{z : \mathrm{dist}(z,\mathcal{M}) < \delta_*\}$, *and* $\overline{\mathcal{U}_{\delta_*}} \subset \mathcal{U}_{\mathrm{st}}$. *Denote this nearest-point projection by* $\pi : \mathcal{U}_{\delta_*} \to \mathcal{M}$.

For every $\delta \in (0,\delta_*]$, *there exist* $T_0 < \infty$ *and* $A_0 > 0$ *such that, with* $n_\eta := \left\lfloor \frac{T_0}{\eta} \right\rfloor$ *and, for any* $A > A_0$, $K_\eta := \left\lceil A\eta^{-1}\log(\eta^{-1})\right\rceil$, *the following conclusions hold:*

$$\sup_{0\le k\le n_\eta}\|z^\eta_k - \varphi_{k\eta}(z^\eta_0)\| \xrightarrow{\mathbb{P}} 0, \quad \mathbb{P}\left(z^\eta_{n_\eta} \in \mathcal{U}_{\delta/3}\right) \longrightarrow 1. \tag{16}$$

Moreover, $\mathbb{P}(z^\eta_{n_\eta+j} \in \mathcal{U}_{2\delta/3} \text{ for every } 0 \le j \le K_\eta) \xrightarrow{\mathbb{P}} 1$, *and*

$$\mathrm{dist}\left(z^\eta_{n_\eta+K_\eta}, \mathcal{M}\right) = O_{\mathbb{P}}(\sqrt{\eta}). \tag{17}$$

The corresponding nearest-point projection satisfies

$$\left\|\pi\left(z^{\eta}_{n_\eta+K_\eta}\right)-\Phi(z^{\eta}_0)\right\| \xrightarrow{\mathbb{P}} 0. \tag{18}$$

Finally, $\eta^2(n_\eta+K_\eta)=O\left(\eta+\eta\log(\eta^{-1})\right)\longrightarrow 0$.

(Proof of Lemma C.7). We first construct the common geometric neighborhood, then apply Lemmas C.8 and C.9, and finally combine their conclusions with the asymptotic phase preservation.

1. Construction of the working tube & Fixed-time entrance. Because $\mathcal{M}$ is a compact C^4 embedded manifold, we apply Lemma C.3, where we construct a tubular neighborhood with its radius $r_{\mathrm{tub}}>0$ with desirable properties stated in the lemma. We call this tubular neighborhood $\mathcal{U}_{r_{\mathrm{tub}}}:=\{z:\mathrm{dist}(z,\mathcal{M})<r_{\mathrm{tub}}\}$. Notably, every point in $\mathcal{U}_{r_{\mathrm{tub}}}$ has a unique nearest-point projection onto $\mathcal{M}$. The resulting map $\pi:\mathcal{U}_{r_{\mathrm{tub}}}\to\mathcal{M}$ is C^3 (Foote, 1984).

Since $\mathcal{M}$ is compact and the stable-foliation $\mathcal{U}_{\mathrm{st}}$ is an open neighborhood of $\mathcal{M}$, there exists $\delta_*>0$ such that the closure of the tubular neighborhood satisfies $\overline{\mathcal{U}_{\delta_*}}\subset\mathcal{U}_{\mathrm{st}}\cap\mathcal{U}_{r_{\mathrm{tub}}}$. Fix any $\delta\in(0,\delta_*]$ for the remainder of the proof.

With the above construction, we invoke Lemma C.8. Therefore, there exists $T_0<\infty$ such that, with $n_\eta:=\left\lfloor\frac{T_0}{\eta}\right\rfloor$, proving Eq. (16). Moreover, with probability tending to one, the iterates remain in $\mathcal{U}_{\mathrm{st}}$ for all $0\le k\le n_\eta$. Define the entrance event $\mathcal{A}_\eta:=\{z^{\eta}_{n_\eta}\in\mathcal{U}_{\delta/3}\}$. Then it follows from the lemma that $\mathbb{P}(\mathcal{A}_\eta)\longrightarrow 1$.

2. Contraction to the stochastic normal layer. Starting at time n_η, define the shifted exit time $\tau_\eta:=\inf\{j\ge 0: z^{\eta}_{n_\eta+j}\notin\mathcal{U}_{2\delta/3}\}$. Conditional on $\mathcal{F}_{n_\eta}$ and on $\mathcal{A}_\eta$, the shifted process $(z^{\eta}_{n_\eta+j})_{j\ge0}$ is an SGD recursion initialized in $\mathcal{U}_{\delta/3}$. Since the future samples are independent of $\mathcal{F}_{n_\eta}$, and all estimates in Lemma C.9 are uniform over initializations in $\mathcal{U}_{\delta/3}$, the lemma applies.

Let $c_p>0$ be its high-moment contraction constant and choose $A>p/c_p$ (see Lemma C.9. Set $K_\eta:=\left\lceil A\eta^{-1}\log(\eta^{-1})\right\rceil$. The no-exit conclusion ($\mathbb{P}(\tau_\eta\le K_\eta)\to 0$) gives $\mathbb{P}\left(\tau_\eta\le K_\eta,\,\mathcal{A}_\eta\right)\longrightarrow 0$, since, using $\mathbb{P}(\tau_\eta\le K_\eta)\le\varepsilon_\eta$ for $\varepsilon_\eta\to 0$ (which holds uniformly in $\mathcal{U}_{\delta/3}$),

$$\begin{aligned}\mathbb{P}(\tau_\eta\le K_\eta,\mathcal{A}_\eta)&=\mathbb{E}[\mathbf{1}_{\mathcal{A}_\eta}\mathbb{P}(\tau_\eta\le K_\eta\mid\mathcal{F}_{n_\eta})]\\&\le\varepsilon_\eta\mathbb{E}[\mathbf{1}_{\mathcal{A}_\eta}]\le\varepsilon_\eta\to 0.\end{aligned}$$

Consequently, for $\mathcal{E}_\eta:=\mathcal{A}_\eta\cap\{\tau_\eta>K_\eta\}$, we have $\mathbb{P}(\mathcal{E}_\eta)\longrightarrow 1$. The high-moment conclusion of the same lemma gives $\mathbb{E}\left[\mathrm{dist}\left(z^{\eta}_{n_\eta+K_\eta},\mathcal{M}\right)^{2p}\mathbf{1}_{\mathcal{E}_\eta}\right]\le C\eta^p$. Therefore, for every $R>0$,

$$\begin{aligned}\mathbb{P}\left(\mathrm{dist}\left(z^{\eta}_{n_\eta+K_\eta},\mathcal{M}\right)>R\sqrt{\eta}\right)&\le\mathbb{P}(\mathcal{E}_\eta^c)+\frac{1}{R^{2p}\eta^p}\mathbb{E}\left[\mathrm{dist}\left(z^{\eta}_{n_\eta+K_\eta},\mathcal{M}\right)^{2p}\mathbf{1}_{\mathcal{E}_\eta}\right]\\&\le\mathbb{P}(\mathcal{E}_\eta^c)+\frac{C}{R^{2p}}.\end{aligned}$$

Taking first $\eta\downarrow 0$ and then $R\uparrow\infty$ proves $\mathrm{dist}\left(z^{\eta}_{n_\eta+K_\eta},\mathcal{M}\right)=O_{\mathbb{P}}(\sqrt{\eta})$.

3. Phase preservation. The preceding estimates show that SGD reaches an $O(\sqrt{\eta})$ neighborhood of $\mathcal{M}$, but they do not identify its location along the manifold. This location determines the initial condition for the late-phase dynamics. We now show that the burn-in phase produces no order-one tangential displacement: the nearest projection of the terminal burn-in iterate converges to the asymptotic gradient-flow phase $\Phi(z^{\eta}_0)$. Thus, any nontrivial order-one feature selection occurs only on the subsequent $O(\eta^{-2})$ time scale.

Let $m_\eta:=n_\eta+K_\eta$. Lemma C.8 gives $\|\Phi(z^{\eta}_{n_\eta})-\Phi(z^{\eta}_0)\|\xrightarrow{\mathbb{P}}0$. For the second stage, the phase-preservation bound in Lemma C.9 is uniform over initializations in $\mathcal{U}_{\delta/3}$. Thus, for every $\varepsilon>0$, there is a deterministic $r_\eta(\varepsilon)\to 0$ such that

$$\begin{aligned}\mathbb{P}\left(\mathcal{A}_\eta\cap\{\|\Phi(z^{\eta}_{m_\eta})-\Phi(z^{\eta}_{n_\eta})\|>\varepsilon\}\right)&=\mathbb{E}\left[\mathbf{1}_{\mathcal{A}_\eta}\mathbb{P}\left(\|\Phi(z^{\eta}_{m_\eta})-\Phi(z^{\eta}_{n_\eta})\|>\varepsilon\middle|\mathcal{F}_{n_\eta}\right)\right]\\&\le r_\eta(\varepsilon)\longrightarrow 0.\end{aligned}$$

Here and below, projection expressions can be extended arbitrarily outside their domains, since these events have probability tending to zero. Because $\mathbb{P}(\mathcal{A}_\eta^c) \to 0$, combining the two stages yields

$$\|\Phi(z_{m_\eta}^\eta) - \Phi(z_0^\eta)\| \le \|\Phi(z_{m_\eta}^\eta) - \Phi(z_{n_\eta}^\eta)\| + \|\Phi(z_{n_\eta}^\eta) - \Phi(z_0^\eta)\| \xrightarrow{\mathbb{P}} 0.$$

Finally, continuity of Φ on the compact set $\overline{\mathcal{U}_\delta}$ gives a modulus of continuity ω_Φ with $\omega_\Phi(r) \to 0$ as $r \downarrow 0$. On $\mathcal{E}_\eta$, $\pi(z_{m_\eta}^\eta) \in \mathcal{M}$ and $\Phi(\pi(z_{m_\eta}^\eta)) = \pi(z_{m_\eta}^\eta)$. Hence

$$\begin{aligned}\|\pi(z_{m_\eta}^\eta) - \Phi(z_0^\eta)\| &\le \|\pi(z_{m_\eta}^\eta) - \Phi(z_{m_\eta}^\eta)\| + \|\Phi(z_{m_\eta}^\eta) - \Phi(z_0^\eta)\| \\ &\le \omega_\Phi\Big(\operatorname{dist}(z_{m_\eta}^\eta, \mathcal{M})\Big) + o_{\mathbb{P}}(1) \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

Since $\mathbb{P}(\mathcal{E}_\eta) \to 1$, this proves Eq. (18).

The total duration satisfies $\eta^2(n_\eta + K_\eta) \le T_0\eta + A\eta\log(\eta^{-1}) + \eta^2 \longrightarrow 0$. Thus the complete burn-in phase reaches the stochastic $O_{\mathbb{P}}(\sqrt{\eta})$ layer while preserving the asymptotic gradient-flow phase selected by the initialization, and occupies asymptotically zero time on the $O(\eta^{-2})$ late-phase scale. □

C.5.1 Auxiliary Lemmas of Phase 1

Lemma C.8 (Fixed-time SGD entrance into a local tube). *Let $\mathcal{M}$ be a compact NASM, and let $\mathcal{U}_{\mathrm{st}}$ be its forward-invariant stable-foliation neighborhood, with continuous asymptotic phase map $\Phi : \mathcal{U}_{\mathrm{st}} \to \mathcal{M}$. Let $\mathcal{K} \Subset \mathcal{U}_{\mathrm{st}}$ be compact and suppose $\mathbb{P}(z_0^\eta \in \mathcal{K}) \longrightarrow 1$. Assume* P1 *and* S1*, with the initialization independent of the future samples. For $\delta > 0$, define $\mathcal{U}_\delta := \{z : \operatorname{dist}(z, \mathcal{M}) < \delta\}$. Let $\delta_0 > 0$ satisfy $\overline{\mathcal{U}_{\delta_0}} \subset \mathcal{U}_{\mathrm{st}}$. Then, for every $\delta \in (0, \delta_0]$, there exists $T_0 < \infty$ such that, with $n_\eta := \lfloor T_0/\eta \rfloor$, $\sup_{0\le k\le n_\eta} \|z_k^\eta - \varphi_{k\eta}(z_0^\eta)\| \xrightarrow{\mathbb{P}} 0$. Consequently, $\mathbb{P}(z_{n_\eta}^\eta \in \mathcal{U}_{\delta/3}) \longrightarrow 1$. Moreover, $\mathbb{P}\left(z_k^\eta \in \mathcal{U}_{\mathrm{st}} \text{ for every } 0 \le k \le n_\eta\right) \longrightarrow 1$, and the fixed-time phase is asymptotically preserved: $\sup_{0\le k\le n_\eta} \|\Phi(z_k^\eta) - \Phi(z_0^\eta)\| \xrightarrow{\mathbb{P}} 0$.*

(Proof of Lemma C.8). Since $\mathcal{M}$ is a NASM, by uniform normal attraction on the compact set $\mathcal{K}$, there are constants $C_{\mathcal{K}}, c_{\mathcal{K}} > 0$ such that

$$\operatorname{dist}\big(\varphi_t(z), \mathcal{M}\big) \le C_{\mathcal{K}} e^{-c_{\mathcal{K}} t} \operatorname{dist}(z, \mathcal{M}), \qquad z \in \mathcal{K}, \quad t \ge 0.$$

Since $\mathcal{K}$ is compact, and thus $D_{\mathcal{K}} := \sup_{z\in\mathcal{K}} \operatorname{dist}(z, \mathcal{M}) < \infty$, we can choose $T_0 < \infty$ sufficiently large so that $C_{\mathcal{K}} e^{-c_{\mathcal{K}} t} D_{\mathcal{K}} \le \delta/5$. It follows that $\varphi_{T_0}(z) \in \mathcal{U}_{\delta/5}$, uniformly over $z \in \mathcal{K}$. Consider the compact set swept out by these deterministic trajectories: $\mathcal{C}_0 := \{\varphi_t(z) \mid z \in \mathcal{K},\ 0 \le t \le T_0\}$. Since $\mathcal{C}_0$ is compact and $\mathcal{C}_0 \subset \mathcal{U}_{\mathrm{st}}$, while $\mathcal{U}_{\mathrm{st}}$ is open, define $d_0 := \operatorname{dist}\left(\mathcal{C}_0, \mathbb{R}^D \setminus \mathcal{U}_{\mathrm{st}}\right) > 0$. Choose $r \in (0, d_0)$ and define $\mathcal{C}_r := \{w : \operatorname{dist}(w, \mathcal{C}_0) \le r\}$. Then $\mathcal{C}_r \subset \mathcal{U}_{\mathrm{st}}$. Moreover, $\mathcal{C}_r$ is closed and bounded, and hence compact.

On $\mathcal{C}_r$, let L_0 be a Lipschitz constant for ∇L, and set $B_0 := \sup_{z\in\mathcal{C}_r} \|\nabla L(z)\| < \infty$. By S1, there is also $M_0 < \infty$ such that $\sup_{z\in\mathcal{C}_r} \mathbb{E}\,\|g(z;x) - \nabla L(z)\|^2 \le M_0$. On the event $\{z_0^\eta \in \mathcal{K}\}$, define $y_k^\eta := \varphi_{k\eta}(z_0^\eta)$, $e_k^\eta := z_k^\eta - y_k^\eta$, and introduce the localization time

$$\sigma_\eta := \inf\left\{0 \le k \le n_\eta : \|e_k^\eta\| \ge r\right\}.$$

Note that for $k < \sigma_\eta$, we have $z_k^\eta \in \mathcal{C}_r$. The population flow satisfies the one-step expansion

$$y_{k+1}^\eta = y_k^\eta - \eta\nabla L(y_k^\eta) + q_{k+1}^\eta, \qquad \|q_{k+1}^\eta\| \le C_0\eta^2, \tag{19}$$

uniformly for $k < n_\eta$. Indeed, $q_{k+1}^\eta = -\int_{k\eta}^{(k+1)\eta} \left[\nabla L\big(\varphi_s(z_0^\eta)\big) - \nabla L\big(\varphi_{k\eta}(z_0^\eta)\big)\right] \mathrm{d}s$. Since $\|\dot\varphi_s\| \le B_0$ on $\mathcal{C}_0$, $\|\varphi_s(z_0^\eta) - \varphi_{k\eta}(z_0^\eta)\| \le B_0(s - k\eta)$, and therefore

$$\|q_{k+1}^\eta\| \le L_0 B_0 \int_{k\eta}^{(k+1)\eta} (s - k\eta)\, ds = \frac{L_0 B_0}{2}\eta^2.$$

Define the centered sample-gradient noise $\xi_{k+1}^\eta := g(z_k^\eta; x_{k+1}) - \nabla L(z_k^\eta)$ (Note that $\mathbb{E}[\xi_{k+1}^\eta \mid \mathcal{F}_k] = 0$). Subtracting Eq. (19) from the sgd recursion $z_{k+1}^\eta = z_k^\eta - \eta g(z_k^\eta; x_{k+1})$ gives, for $k < \sigma_\eta$,

$$e_{k+1}^\eta = e_k^\eta - \eta\left(\nabla L(z_k^\eta) - \nabla L(y_k^\eta)\right) - \eta\xi_{k+1}^\eta - q_{k+1}^\eta. \tag{20}$$

Let $S_m^\eta := \eta \sum_{k<m\wedge\sigma_\eta} \mathbf{1}_{\{z_0^\eta\in\mathcal{K}\}}\xi_{k+1}^\eta$. Because $\mathbf{1}_{\{z_0^\eta\in\mathcal{K}\}}\mathbf{1}_{\{k<\sigma_\eta\}}$ is $\mathcal{F}_k$-measurable, (S_m^η) is a martingale. Its increments are orthogonal in L^2, and hence

$$\begin{aligned}\mathbb{E}\|S_{n_\eta}^\eta\|^2 &= \eta^2 \sum_{k<n_\eta} \mathbb{E}\left[\mathbf{1}_{\{Z_0^\eta\in\mathcal{K}\}}\mathbf{1}_{\{k<\sigma_\eta\}}\|\xi_{k+1}^\eta\|^2\right]\\ &\le M_0\eta^2 n_\eta \le M_0 T_0 \eta.\end{aligned}$$

Doob's maximal inequality therefore yields $\mathbb{E}\left[\max_{m\le n_\eta}\|S_m^\eta\|^2\right] \le 4M_0T_0\eta$. In particular, $\max_{m\le n_\eta}\|S_m^\eta\| \xrightarrow{\mathbb{P}} 0$. Summing Eq. (20) up to $m\wedge\sigma_\eta$ and using the Lipschitz property of ∇L gives $\max_{\ell\le m}\|e_{\ell\wedge\sigma_\eta}^\eta\| \le L_0\eta\sum_{k<m}\max_{\ell\le k}\|e_{\ell\wedge\sigma_\eta}^\eta\| + \max_{\ell\le m}\|S_\ell^\eta\| + C_0T_0\eta$. Discrete Grönwall's inequality implies

$$\max_{m\le n_\eta}\|e_{m\wedge\sigma_\eta}^\eta\| \le e^{L_0T_0}\left(\max_{m\le n_\eta}\|S_m^\eta\| + C_0T_0\eta\right). \tag{21}$$

The right-hand side converges to zero in probability. Consequently,

$$\mathbb{P}(\sigma_\eta \le n_\eta) \le \mathbb{P}\left(\max_{m\le n_\eta}\|e_{m\wedge\sigma_\eta}^\eta\| \ge r\right) \longrightarrow 0.$$

Removing the localization and using $\mathbb{P}(z_0^\eta\in\mathcal{K})\to 1$ proves

$$\sup_{i\le k\le n_\eta}\|z_k^\eta - \varphi_{k_\eta}(z_0^\eta)\| \xrightarrow{\mathbb{P}} 0. \tag{22}$$

It remains to verify entrance into the smaller tube. Since $0 \le T_0 - n_\eta\eta < \eta$ and the population flow has uniformly bounded speed on $\mathcal{C}_0$, $\left\|\varphi_{n_\eta\eta}(z) - \varphi_{T_0}(z)\right\| \le B_0\eta$, $z\in\mathcal{K}$. Combining this with $\varphi_{T_0}(z)\in\mathcal{U}_{\delta/5}$, for all sufficiently small η, $\operatorname{dist}\left(\varphi_{n_\eta\eta}(z),\mathcal{M}\right) \le \frac{\delta}{5} + B_0\eta < \frac{\delta}{4}$. Therefore, on the event $\left\|z_{n_\eta}^\eta - \varphi_{n_\eta\eta}(z_0^\eta)\right\| < \frac{\delta}{12}$, we have $\operatorname{dist}(z_{n_\eta}^\eta,\mathcal{M}) < \frac{\delta}{4} + \frac{\delta}{12} = \frac{\delta}{3}$. Eq. (22) now gives $\mathbb{P}(z_{n_\eta}^\eta\in\mathcal{U}_{\delta/3})\to 1$.

Phase preservation. Since Φ is continuous and $\mathcal{C}_r$ is compact, its modulus of continuity on this set, $\omega_{\Phi,r}(h) := \sup_{z,w\in\mathcal{C}_r,\|z-w\|\le h}\|\Phi(z)-\Phi(w)\|$, satisfies $\omega_{\Phi,r}(h)\to 0$ as $h\downarrow 0$. On $\{\sigma_\eta > n_\eta\}$, both z_k^η and $\varphi_{k\eta}(z_0^\eta)$ lie in $\mathcal{C}_r$. Phase invariance (follows from the stable foliation) gives $\Phi(\varphi_{k\eta}(z_0^\eta)) = \Phi(z_0^\eta)$. Consequently,

$$\begin{aligned}\sup_{0\le k\le n_\eta}\|\Phi(z_k^\eta) - \Phi(z_0^\eta)\| &= \sup_{0\le k\le n_\eta}\|\Phi(z_k^\eta) - \Phi(\varphi_{k\eta}(z_0^\eta))\|\\ &\le \omega_{\Phi,r}\sup_{0\le k\le n_\eta}\|z_k^\eta - \varphi_{k\eta}(z_0^\eta)\| \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

□

Lemma C.9 (Contraction from a fixed tube to the stochastic normal layer). *Let $\mathcal{M}\subset\mathbb{R}^D$ be a compact embedded manifold, and choose $\delta > 0$ such that the nearest-point projection π is well defined and C^3 on an open neighborhood of $\overline{\mathcal{U}_\delta}$. For $z\in\mathcal{U}_\delta$, define $\mathsf{n}(z) := z - \pi(z)$, and a lyapunov function $V(z) := \frac{1}{2}\|\mathsf{n}(z)\|^2$. Suppose that, uniformly over $z\in\mathcal{U}_\delta$,*

$$\langle n(z), \nabla L(z)\rangle \ge \mu\|n(z)\|^2, \quad \text{for some } \mu > 0. \tag{23}$$

Assume that $z_0^\eta\in\mathcal{U}_{\delta/3}$ a.s, and define $\tau_{2\delta/3}^\eta := \inf\left\{k\ge 0 \mid z_k^\eta\notin\mathcal{U}_{2\delta/3}\right\}$. There is a constant $c > 0$, as in Eq. (30) below, such that for every fixed $A > p/c$ and $K_\eta := \left\lceil A\eta^{-1}\log(\eta^{-1})\right\rceil$,

$$\mathbb{P}\left(\tau_{2\delta/3}^\eta \le K_\eta\right) \longrightarrow 0, \quad \mathbb{E}\left[\operatorname{dist}(z_{K_\eta}^\eta,\mathcal{M})^{2p}\mathbf{1}_{\{K_\eta<\tau_{2\delta/3}^\eta\}}\right] \le C\eta^p \tag{24}$$

Consequently, $\operatorname{dist}(z_{K_\eta}^\eta,\mathcal{M}) = O_{\mathbb{P}}(\sqrt{\eta})$. Moreover, the asymptotic phase is preserved:

$$\left\|\Phi(z_{K_\eta}^\eta) - \Phi(z_0^\eta)\right\| \xrightarrow{\mathbb{P}} 0.$$

(Proof of Lemma C.9) We divide the proof into three parts.

1. Proof of $\mathbb{P}\left(\tau^{\eta}_{2\delta/3} \le K_\eta\right) \to 0$. Choose $a \in (0, \delta/3)$, and introduce the event

$$\mathcal{J}_\eta := \left\{\eta\|g_{k+1}\| \le a \text{ for every } 0 \le k < K_\eta \wedge \tau^{\eta}_{2\delta/3}\right\}.$$

If $k < \tau^{\eta}_{2\delta/3}$ and $\eta\|g_{k+1}\| \le a$, the full update segment $z^{\eta}_k - t\eta g_{k+1}$ $(0 \le t \le 1)$ lies in $\mathcal{U}_\delta$. Indeed, distance to $\mathcal{M}$ is 1-Lipschitz, so $\operatorname{dist}(z^{\eta}_k - t\eta g_{k+1}, \mathcal{M}) \le 2\delta/3 + a < \delta$.

The moment assumption (S1) and Markov's inequality give

$$\begin{aligned}\mathbb{P}(\mathcal{J}^c_\eta) &\le \sum_{k<K_\eta} \mathbb{E}\left[\mathbf{1}_{\{k<\tau^{\eta}_{2\delta/3}\}}\mathbb{P}\left(\|g_{k+1}\| > \frac{a}{\eta}\middle|\mathcal{F}_k\right)\right] \\ &\le K_\eta\left(\frac{\eta}{a}\right)^{2p} K_{2p} \le C\eta^{2p-1}\log(\eta^{-1}) \longrightarrow 0. \end{aligned} \tag{25}$$

Because V has bounded second derivative on the closed tube $\overline{\mathcal{U}_\delta}$, Taylor's theorem gives, on $\mathcal{J}_\eta$ and for $k < \tau^{\eta}_{2\delta/3}$,

$$\begin{aligned}V(Z^{\eta}_{k+1}) &\le V(z^{\eta}_k) - \eta\langle n_k, g_{k+1}\rangle + C\eta^2\|g_{k+1}\|^2 \\ &= V(z^{\eta}_k) - \eta\langle n_k, \nabla L(z^{\eta}_k)\rangle - \eta\langle n_k, \xi_{k+1}\rangle + C\eta^2\|g_{k+1}\|^2 \\ &\le V(z^{\eta}_k) - \mu\eta\|n_k\|^2 - \eta\langle n_k, \xi_{k+1}\rangle + C\eta^2\|g_{k+1}\|^2, \end{aligned} \tag{26}$$

where $n_k := n(z^{\eta}_k)$. Define the stopped martingale $M_m := -\eta\sum_{k<m}\mathbf{1}_{\{k<\tau^{\eta}_{2\delta/3}\}}\langle n_k, \xi_{k+1}\rangle$ and the nonnegative remainder $R_m := C\eta^2\sum_{k<m}\mathbf{1}_{\{k<\tau^{\eta}_{2\delta/3}\}}\|g_{k+1}\|^2$. Summing Eq. (26) and dropping the second term gives, on $\mathcal{J}_\eta$,

$$V(z^{\eta}_{m\wedge\tau^{\eta}_{2\delta/3}}) \le V(z^{\eta}_0) + M_{m\wedge\tau^{\eta}_{2\delta/3}} + R_{m\wedge\tau^{\eta}_{2\delta/3}}, \qquad m \le K_\eta. \tag{27}$$

Before $\tau^{\eta}_{2\delta/3}$, we have $\|n_k\| < 2\delta/3$. Moreover, $\mathbb{E}\big[\|\xi_{k+1}\|^2 \mid \mathcal{F}_k\big] \le C$ uniformly in k (due to (S1)). Therefore,

$$\begin{aligned}\mathbb{E}|M_{K_\eta}|^2 &= \eta^2\sum_{k<K_\eta}\mathbb{E}\left[\mathbf{1}_{\{k<\tau^{\eta}_{2\delta/3}\}}\mathbb{E}\left[\langle n_k, \xi_{k+1}\rangle^2\middle|\mathcal{F}_k\right]\right] \\ &\le C\eta^2 K_\eta \le C\eta\log(\eta^{-1}). \end{aligned} \tag{28}$$

Doob's maximal inequality consequently gives $\mathbb{E}\left[\sup_{m\le K_\eta}|M_m|^2\right] \le C\eta\log(\eta^{-1}) \to 0$. Similarly,

$$\begin{aligned}\mathbb{E}R_{K_\eta} &\le C\eta^2\sum_{k<K_\eta}\mathbb{E}\left[\mathbf{1}_{\{k<\tau^{\eta}_{2\delta/3}\}}\|g_{k+1}\|^2\right] \\ &\le C\eta^2 K_\eta \le C\eta\log(\eta^{-1}) \to 0. \end{aligned} \tag{29}$$

Since $z^{\eta}_0 \in \mathcal{U}_{\delta/3}$, $V(z^{\eta}_0) \le \frac{\delta^2}{18}$. If $\tau^{\eta}_{2\delta/3} \le K_\eta$ and $\mathcal{J}_\eta$ occurs, then $z^{\eta}_{\tau^{\eta}_{2\delta/3}} \in \mathcal{U}_\delta$ since, for $\tau := \tau^{\eta}_{2\delta/3}$, $\operatorname{dist}(z_\tau, \mathcal{M}) \le \operatorname{dist}(z_{\tau-1}, \mathcal{M}) + \|z_\tau - z_{\tau-1}\| < 2\delta/3 + \delta/3 = \delta$ and

$$V\left(z^{\eta}_{\tau^{\eta}_{2\delta/3}}\right) = \frac{1}{2}\operatorname{dist}\left(z^{\eta}_{\tau^{\eta}_{2\delta/3}}, \mathcal{M}\right)^2 \ge \frac{1}{2}\left(\frac{2\delta}{3}\right)^2 = \frac{2\delta^2}{9}.$$

It follows from Eq. (27) that $\sup_{m\le K_\eta}|M_m| + R_{K_\eta} \ge \frac{2\delta^2}{9} - \frac{\delta^2}{18} = \frac{\delta^2}{6}$. Hence

$$\mathbb{P}(\tau^{\eta}_{2\delta/3} \le K_\eta) \le \mathbb{P}(\mathcal{J}^c_\eta) + \mathbb{P}\left(\sup_{m\le K_\eta}|M_m| \ge \frac{\delta^2}{12}\right) + \mathbb{P}\left(R_{K_\eta} \ge \frac{\delta^2}{12}\right).$$

The above results combined with Markov's inequalities prove $\mathbb{P}(\tau^{\eta}_{2\delta/3} \le K_\eta) \to 0$.

2. Contraction to the $O(\sqrt{\eta})$ scale. We first show that there exist constants $c, C > 0$ such that

$$\mathbb{E}\left[V(z_{k+1}^\eta)^p \mathbf{1}_{\{k+1<\tau_{2\delta/3}^\eta\}} \mid \mathcal{F}_k\right] \leq (1-c\eta)V(z_k^\eta)^p \mathbf{1}_{\{k<\tau_{2\delta/3}^\eta\}} + C\eta^{p+1}\mathbf{1}_{\{k<\tau_{2\delta/3}^\eta\}}. \tag{30}$$

To derive this, let $\tau := \tau_{2\delta/3}^\eta$, $d_k := \operatorname{dist}(z_k^\eta, \mathcal{M})$. Note that $d_{k+1} \leq \|n_k - \eta g_{k+1}\|$. We apply the standard inequality

$$\|u+h\|^{2p} \leq \|u\|^{2p} + 2p\|u\|^{2p-2}\langle u, h\rangle + C_p(\|u\|^{2p-2}\|h\|^2 + \|h\|^{2p}), \tag{31}$$

for all $u, h \in \mathbb{R}^D$. To see this, for $f(u) = \|u\|^{2p}$, the second-order integral Taylor formula gives $f(u+h) = f(u) + Df(u)[h] + \int_0^1 (1-t)D^2 f(u+th)[h,h]\,\mathrm{d}t$. Since $Df(u)[h] = 2p\|u\|^{2p-2}\langle u, h\rangle$ and $\|D^2 f(w)\| \leq C_p\|w\|^{2p-2}$, the estimate $\|u+th\|^{2p-2} \leq C_p\left(\|u\|^{2p-2} + \|h\|^{2p-2}\right)$ proves Eq. (31).

Applying Eq. (31) with $u = n_k$ and $h = -\eta g_{k+1}$ gives

$$d_{k+1}^{2p} \leq d_k^{2p} - 2p\eta d_k^{2p-2}\langle n_k, g_{k+1}\rangle + C_p\eta^2 d_k^{2p-2}\|g_{k+1}\|^2 + C_p\eta^{2p}\|g_{k+1}\|^{2p}.$$

Note that $\{k+1 < \tau\} \subseteq \{k < \tau\}$, yielding $d_{k+1}^{2p}\mathbf{1}_{\{k+1<\tau\}} \leq d_{k+1}^{2p}\mathbf{1}_{\{k<\tau\}}$. The indicator on the right is $\mathcal{F}_k$-measurable. Therefore, conditional unbiasedness gives

$$\begin{aligned}\mathbb{E}\left[d_{k+1}^{2p}\mathbf{1}_{\{k+1<\tau\}} \mid \mathcal{F}_k\right] \leq \mathbf{1}_{\{k<\tau\}}\Big[&d_k^{2p} - 2p\eta d_k^{2p-2}\,\langle n_k, \nabla L(z_k^\eta)\rangle + \\ &C_p\eta^2 d_k^{2p-2}\mathbb{E}\left[\|g_{k+1}\|^2 \mid \mathcal{F}_k\right] + C_p\eta^{2p}\mathbb{E}\left[\|g_{k+1}\|^{2p} \mid \mathcal{F}_k\right]\Big].\end{aligned} \tag{32}$$

On $\{k < \tau\}$, the iterate lies in $\mathcal{U}_{2\delta/3}$. Using Eq. (23), there exist some $\mu > 0$ such that the negative term in Eq. (32) is bounded by $-2p\mu\eta d_k^{2p}$.(S1) also gives constants $K_2, K_{2p} < \infty$ such that, on $\{k < \tau\}$,

$$\mathbb{E}[\|g_{k+1}\|^2 \mid \mathcal{F}_k] \leq K_2, \qquad \mathbb{E}[\|g_{k+1}\|^{2p} \mid \mathcal{F}_k] \leq K_{2p}.$$

For every $\varepsilon > 0$, Young's inequality gives $\eta^2 d_k^{2p-2} \leq \varepsilon\eta d_k^{2p} + C_{\varepsilon,p}\eta^{p+1}$. Moreover, for $0 < \eta \leq 1$, $\eta^{2p} \leq \eta^{p+1}$. Choosing ε sufficiently small allows the term $\varepsilon\eta d_k^{2p}$ to be absorbed into the negative $-2p\mu\eta d_k^{2p}$ contribution. Hence, for some constants $c_p, C_p > 0$,

$$\mathbb{E}\left[d_{k+1}^{2p}\mathbf{1}_{\{k+1<\tau\}} \mid \mathcal{F}_k\right] \leq (1-c_p\eta)d_k^{2p}\mathbf{1}_{\{k<\tau\}} + C_p\eta^{p+1}\mathbf{1}_{\{k<\tau\}}. \tag{33}$$

Finally, since $d_k^{2p} = 2^p V(z_k^\eta)^p$, dividing Eq. (33) by 2^p and absorbing the resulting constant into C_p proves Eq. (30).

With this, let $a_k := \mathbb{E}\left[V(z_k^\eta)^p \mathbf{1}_{\{k<\tau_{2\delta/3}^\eta\}}\right]$. Taking expectation in Eq. (30) yields $a_{k+1} \leq (1-c\eta)a_k + C\eta^{p+1}$. By telescoping the sum,

$$a_k \leq (1-c\eta)^k a_0 + C\eta^{p+1}\sum_{j=0}^{k-1}(1-c\eta)^j \leq e^{-c\eta k}a_0 + \frac{C}{c}\eta^p. \tag{34}$$

Using $z_0^\eta \in \mathcal{U}_{\delta/3}$, $a_0 \leq \left(\frac{\delta^2}{18}\right)^p$. With $K_\eta = \left\lceil A\eta^{-1}\log(\eta^{-1})\right\rceil$ and $A > p/c$, we have

$$e^{-c\eta K_\eta} \leq e^{-cA\log(\eta^{-1})} = \eta^{cA} \leq \eta^p$$

for all sufficiently small η. Substituting this into Eq. (34) gives $a_{K_\eta} \leq C\eta^p$. Since $\operatorname{dist}(z, \mathcal{M})^{2p} = \|n(z)\|^{2p} = 2^p V(z)^p$, we obtain $\mathbb{E}\left[\operatorname{dist}(z_{K_\eta}^\eta, \mathcal{M})^{2p}\mathbf{1}_{\{K_\eta<\tau_{2\delta/3}^\eta\}}\right] \leq C\eta^p$.

For every $R > 0$, Markov's inequality and the no-exit estimate from the previous step give

$$\mathbb{P}\left(\operatorname{dist}(z_{K_\eta}^\eta, \mathcal{M}) > R\sqrt{\eta}\right) \leq \mathbb{P}(\tau_{2\delta/3}^\eta \leq K_\eta) + \frac{C}{R^{2p}}. \tag{35}$$

Taking first $\eta \downarrow 0$ and then $R \uparrow \infty$ proves $\operatorname{dist}(z_{K_\eta}^\eta, \mathcal{M}) = O_{\mathbb{P}}(\sqrt{\eta})$.

3. Preservation of the asymptotic phase. We retain the original SGD recursion. The argument first compares it with population gradient flow over a fixed time interval and then bounds the remaining displacement up to K_η.

Stationarity on $\mathcal{M}$ and local Lipschitz continuity of ∇L imply that, for some $B < \infty$,

$$\|\nabla L(z)\| = \|\nabla L(z) - \nabla L(\pi(z))\| \le B \operatorname{dist}(z, \mathcal{M}), \qquad z \in \mathcal{U}_\delta.$$

The constant is uniform because $\overline{\mathcal{U}_\delta}$ is a compact subset of $\mathcal{U}_{\mathrm{st}}$. Fix an auxiliary time $S > 0$ and set $J_\eta := \lfloor S/\eta \rfloor$. For sufficiently small η, $J_\eta < K_\eta$. Since $\overline{\mathcal{U}_{\delta/3}}$ is a compact subset of $\mathcal{U}_{\mathrm{st}}$, the conclusion of Lemma C.8 gives

$$\|z^\eta_{J_\eta} - \varphi_S(z^\eta_0)\| \xrightarrow{\mathbb{P}} 0. \tag{36}$$

Define the vector-valued stopped martingale $S^\eta_l := \eta \sum_{k<l} \mathbf{1}_{\{k<\tau\}} \xi_{k+1}$. Its increments have zero conditional mean. The uniform second-moment bound (S1) and Doob's inequality give

$$\mathbb{E} \max_{l \le K_\eta} \|S^\eta_l\|^2 \le C\eta^2 K_\eta \le C\eta \log(\eta^{-1}) \longrightarrow 0. \tag{37}$$

Summing the original SGD updates (rather than expanding Φ) gives the exact stopped identity

$$z^\eta_{K_\eta \wedge \tau} - z^\eta_{J_\eta \wedge \tau} = -\eta \sum_{k=J_\eta}^{K_\eta - 1} \mathbf{1}_{\{k<\tau\}} \nabla L(z^\eta_k) - \big(S^\eta_{K_\eta} - S^\eta_{J_\eta}\big). \tag{38}$$

We now bound the population-gradient contribution using the high moments already established in Part 2. Since $d_k^{2p} = 2^p V(z^\eta_k)^p$, Hölder's inequality and Eq. (34) yield

$$\begin{aligned}\mathbb{E}[d_k \mathbf{1}_{\{k<\tau\}}] &\le \Big(\mathbb{E}[d_k^{2p} \mathbf{1}_{\{k<\tau\}}]\Big)^{1/(2p)} \\ &\le Ce^{-c_p \eta k/(2p)} + C\sqrt{\eta}.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{E}\left[\eta \sum_{k=J_\eta}^{K_\eta - 1} \mathbf{1}_{\{k<\tau\}} \|\nabla L(z^\eta_k)\|\right] &\le B\eta \sum_{k=J_\eta}^{K_\eta - 1} \mathbb{E}[d_k \mathbf{1}_{\{k<\tau\}}] \\ &\le C\eta \sum_{k=J_\eta}^{K_\eta - 1} \Big(e^{-c_p \eta k/(2p)} + \sqrt{\eta}\Big) \\ &\le Ce^{-c_p S/(2p)} + C\sqrt{\eta}\log(\eta^{-1}).\end{aligned} \tag{39}$$

The last step uses the geometric-series bound $\eta \sum_{k \ge J_\eta} e^{-c_p \eta k/(2p)} \le Ce^{-c_p S/(2p)}$. The constants in this bound do not depend on S.

Combining Eq. (38), Eq. (37), Eq. (39), and the no-exit conclusion, Markov's inequality gives, for every $\varepsilon > 0$,

$$\limsup_{\eta \downarrow 0} \mathbb{P}\left(\|z^\eta_{K_\eta} - z^\eta_{J_\eta}\| > \varepsilon\right) \le \frac{C}{\varepsilon} e^{-c_p S/(2p)}. \tag{40}$$

Locally uniform attraction from the stable foliation (C1) gives $\epsilon_\delta(S) := \sup_{z \in \overline{\mathcal{U}_{\delta/3}}} \|\varphi_S(z) - \Phi(z)\| \longrightarrow 0$, as $S \to \infty$. Therefore $\|z^\eta_{K_\eta} - \Phi(z^\eta_0)\| \le \|z^\eta_{K_\eta} - z^\eta_{J_\eta}\| + \|z^\eta_{J_\eta} - \varphi_S(z^\eta_0)\| + \epsilon_\delta(S)$. For each error threshold, first fix S and let $\eta \downarrow 0$, using Eq. (36) and equation 40; then let $S \to \infty$. This proves the stronger endpoint conclusion $\|z^\eta_{K_\eta} - \Phi(z^\eta_0)\| \xrightarrow{\mathbb{P}} 0$. Since Φ is uniformly continuous on $\overline{\mathcal{U}_\delta}$, let ω_Φ be its modulus of continuity there. On $\{\tau > K_\eta\}$, both $z^\eta_{K_\eta}$ and $\Phi(z^\eta_0) \in \mathcal{M}$ belong to this compact set. Hence

$$\begin{aligned}\|\Phi(z^\eta_{K_\eta}) - \Phi(z^\eta_0)\| &= \|\Phi(z^\eta_{K_\eta}) - \Phi(\Phi(z^\eta_0))\| \\ &\le \omega_\Phi\left(\|z^\eta_{K_\eta} - \Phi(z^\eta_0)\|\right) \xrightarrow{\mathbb{P}} 0,\end{aligned}$$

where we used $\Phi(v) = v$ for $v \in \mathcal{M}$. The no-exit estimate removes this event restriction. The nearest-point projection satisfies, on the same event,

$$\|\pi(z^\eta_{K_\eta}) - \Phi(z^\eta_0)\| \le \operatorname{dist}(z^\eta_{K_\eta}, \mathcal{M}) + \|z^\eta_{K_\eta} - \Phi(z^\eta_0)\| \xrightarrow{\mathbb{P}} 0.$$

Note that all estimates are uniform over the initial point in $\mathcal{U}_{\delta/3}$, so they also hold for the stated random initializations. Finally, $\eta^2 K_\eta \le A\eta \log(\eta^{-1}) + \eta^2 \longrightarrow 0$, so this contraction stage occupies asymptotically zero time on the $O(\eta^{-2})$ late-phase scale. □

C.5.2 Proof of Lemma C.3

We denote by $N\mathcal{M} := \{(v, n) \mid v \in \mathcal{M},\ n \in N_v\mathcal{M}\}$ as the normal bundle of $\mathcal{M}$. For a positive continuous function $\delta : \mathcal{M} \to (0, \infty)$, define the open subset $\mathcal{V} := \{(v, n) \in N\mathcal{M} \mid \|n\| < \delta(v)\}$. Since $\mathcal{M}$ is compact, the function δ attains a positive minimum $\delta_1 := \min_{v\in\mathcal{M}} \delta(v) > 0$. Hence,

$$\mathcal{V}_{\delta_1} := \{(v, n) \in N\mathcal{M} \mid \|n\| < \delta_1\} \subset \mathcal{V}.$$

Define the map $E : N\mathcal{M} \to \mathbb{R}^{2dr}$, $E(v, n) = v + n$. By the Tubular Neighborhood Theorem (Lee, 2003), the restriction $E|_{\mathcal{V}_{\delta_1}} : \mathcal{V}_{\delta_1} \to \mathcal{U}_{\delta_1}$ is a diffeomorphism onto its image, where $\mathcal{U}_{\delta_1} := E(\mathcal{V}_{\delta_1})$ is a tubular neighborhood of $\mathcal{M}$. Consequently, every $Z \in \mathcal{U}_{\delta_1}$ admits a unique decomposition

$$z = v + n, \qquad v \in \mathcal{M},\ n \in N_v\mathcal{M}.$$

Let $\pi : N\mathcal{M} \to \mathcal{M}$, $\pi(v, n) = v$, denote the bundle projection. The induced projection onto the manifold is $\Pi := \pi \circ E^{-1} : \mathcal{U}_{\delta_1} \to \mathcal{M}$.

By Exercise 6-5 in (Lee, 2003), after possibly shrinking the tubular neighborhood, there exists $\delta_* \in (0, \delta_1]$ such that, for every $Z = v + n \in \mathcal{U}_{\delta_*}$, the point $v = \Pi(Z)$ is the unique closest point on $\mathcal{M}$, thus $\operatorname{dist}(Z, \mathcal{M}) = \|n\|$.

Since $\mathcal{M}$ has uniformly positive normal curvature, there exists $c > 0$ such that $H_N(v)[n, n] \ge 2c\|n\|^2$ for every $v \in \mathcal{M}$ and every $n \in N_v\mathcal{M}$. Expanding the gradient with respect to v gives

$$\nabla L(v + n) = \nabla L(v) + D(\nabla L(v))[n] + O(\|n\|^2).$$

Since $\nabla L(v) = 0$ for every $v \in \mathcal{M}$, $\langle \nabla L(v+n), n\rangle = H_N(v)[n, n] + O(\|n\|^3) \ge c\|n\|^2 - C\|n\|^3$, for some constant $C > 0$, where the constant can be chosen uniformly over $\mathcal{M}$ by compactness. Consequently, there exists a uniform constant $c_* > 0$ such that, whenever $\|n\| < c_*$, $\langle \nabla L(v + n), n\rangle \ge c\|n\|^2$. Finally, we define $\delta := \min\{\delta_*, c_*\}$, and denote the resulting tubular neighborhood by $\mathcal{U}_\delta$. □

C.6 SGD Phase 2: Stochastic reduction

Proposition C.10 (Late-phase stochastic reduction). *Let $\mathcal{M} \subset \mathbb{R}^D$ satisfy the standing geometric conditions, and suppose that Assumptions* P1-P3 *and* S1-S4 *hold. Assume that $z^\eta_0 = v^\eta_0 + \sqrt{\eta}\, q^\eta_0$, for $v^\eta_0 \in \mathcal{M}$ and $q^\eta_0 \in N_{v^\eta_0}\mathcal{M}$, with $v^\eta_0 \xrightarrow{\mathbb{P}} v_0$, $\sup_{\eta>0} \mathbb{E}\|q^\eta_0\|^{2p} < \infty$.*

For each $v \in \mathcal{M}$, let $\Pi_v : N_v\mathcal{M} \to N_v\mathcal{M}$ be the unique solution of the Lyapunov equation $H_N(v)\Pi_v + \Pi_v H_N(v) = \Sigma_N(v)$. For any orthonormal basis $\{\mathsf{e}_a\}_{a=1}^{\dim N_v\mathcal{M}}$ of $N_v\mathcal{M}$, define $b(v) := \sum_{a=1}^{\dim N_v\mathcal{M}} \mathcal{B}_v[\mathsf{e}_a, \Pi_v\mathsf{e}_a]$. Then there exists an tubular neighborhood $\mathcal{U}_\zeta$ such that, for every fixed $T > 0$, the iterates remain in the tubular neighborhood with probability tending to one: $\mathbb{P}\left(z^\eta_k \in \mathcal{U}_\zeta \text{ for every } 0 \le k \le \lfloor T/\eta^2 \rfloor\right) \longrightarrow 1$. Consequently, the projected process

$$v^\eta(t) := \pi\left(z^\eta_{\lfloor t/\eta^2 \rfloor}\right), \qquad 0 \le t \le T,$$

is well defined and satisfies $\sup_{0\le t\le T} \|v^\eta(t) - v(t)\| \xrightarrow{\mathbb{P}} 0$, where $v(t)$ is the unique solution of $\dot v(t) = -\frac{1}{2} b(v(t))$, for $v(0) = v_0$.

C.6.1 Proof of Proposition C.10

As outlined in Section 4, the leading tangential contribution of the population gradient to one SGD update is $-\frac{\eta^2}{2}\mathcal{B}_{v_k}[q_k, q_k]$. After controlling the accumulated stochastic and Taylor remainder terms, we obtain

$$v_N - v_0 = -\frac{\eta^2}{2}\sum_{k<N}\mathcal{B}_{v_k}[q_k, q_k] + o_{\mathbb{P}}(1),$$

uniformly for $N \le \lfloor T/\eta^2 \rfloor$. The main task is therefore to replace this accumulated quadratic contribution by its stationary average $b(v) := \mathbb{E}_{q\sim\mathcal{N}(0,\Pi_v)}[\mathcal{B}_v[q,q]]$, which exploits the separation of time scales between tangent and normal variables (the normal fluctuations relax rapidly, while the base point moves slowly along the manifold).

We first introduce some notations and setup.

Common localization setup. Denote $d(z) := \mathrm{dist}(z, \mathcal{M})$ and $N_\eta(T) := \lfloor T/\eta^2 \rfloor$. Choose fixed radius $0 < \zeta_1 < \zeta_2 < \zeta$ (sufficiently small that Lemma C.3 applies). Set $\delta := \zeta_2 - \zeta_1$ and define

$$g^\eta(z; x) := g(z; x)\mathbf{1}_{\{\eta\|g(z;x)\|\le\delta\}}. \tag{41}$$

Couple the truncated recursion to the original SGD recursion through the same initialization and samples:

$$\widetilde{z}^\eta_{k+1} = \widetilde{z}^\eta_k - \eta g^\eta(\widetilde{z}^\eta_k; x_{k+1}), \qquad \widetilde{z}^\eta_0 = z^\eta_0. \tag{42}$$

Their respective exit times from the inner tube are

$$\tau_\eta := \inf\{k \ge 0 \mid d(z^\eta_k) \ge \zeta_1\}, \qquad \widetilde{\tau}_\eta := \inf\{k \ge 0 \mid d(\widetilde{z}^\eta_k) \ge \zeta_1\}.$$

Define the killed normal displacement and its rescaling by

$$\overline{n}_k := \begin{cases} z^\eta_k - \pi(z^\eta_k), & k < \tau_\eta, \\ 0, & k \ge \tau_\eta, \end{cases} \qquad \overline{q}_k := \eta^{-1/2}\overline{n}_k. \tag{43}$$

Define $\widetilde{\overline{n}}_k$ and $\widetilde{\overline{q}}_k$ analogously for the truncated chain, using $\widetilde{\tau}_\eta$. We also define the following to bound the biases introduced from the truncated chain.

$$b_\eta(z) := \mathbb{E}[g^\eta(z; x)] - \nabla L(z), \qquad M_2(z) := \mathbb{E}[g(z; x)^{\otimes 2}], \qquad M_2^\eta(z) := \mathbb{E}[g^\eta(z; x)^{\otimes 2}].$$

1. Localization and truncation. As the first technical step towards the goal above, we first establish the estimates needed to analyze SGD updates near $\mathcal{M}$. These estimates control the normal fluctuations and ensure that the iterates remain, with probability tending to one, in a fixed tubular neighborhood satisfying the properties of Lemma C.3.

To conduct Taylor expansions of the nearest-point projection, we ensure that each update segment remains inside a tubular neighborhood. Since the sample gradients need not be bounded, we introduce a truncated recursion. For fixed $0 < \zeta_1 < \zeta_2 < \zeta$, set $\delta := \zeta_2 - \zeta_1$ and replace $g(z; x)$ by $g^\eta(z; x) := g(z; x)\mathbf{1}_{\{\eta\|g(z;x)\|\le\delta\}}$. Every truncated update starting in $\mathcal{U}_{\zeta_1}$ then has its entire segment inside $\mathcal{U}_{\zeta_2}$. Lemma C.11 below shows that truncation introduces negligible first- and second-moment biases, preserves the normal-moment bounds, and leaves the trajectory unchanged with probability tending to one throughout $0 \le k \le N_\eta(T) := \lfloor T/\eta^2 \rfloor$.

Lemma C.11 (Localization, truncation, and normal-moment bounds). *Under the setup above, the following statements hold for every fixed $T > 0$.*

1. (Normal-moment bounds). *There exists $C_{p,T} < \infty$, such that*

$$\sup_{0\le k\le N_\eta(T)} \mathbb{E}\left[\|\overline{n}_k\|^{2p} + \|\widetilde{\overline{n}}_k\|^{2p}\right] \le C_{p,T}\eta^p. \tag{44}$$

Or equivalently,

$$\sup_{0\le k\le N_\eta(T)} \mathbb{E}\left[\|\overline{q}_k\|^{2p} + \|\widetilde{\overline{q}}_k\|^{2p}\right] \le C_{p,T}. \tag{45}$$

2. (Tube confinement). *Both chains remain in the inner tube with probability tending to one, i.e.,*

$$\mathbb{P}\left(\tau_\eta \wedge \widetilde{\tau}_\eta \le N_\eta(T)\right) \longrightarrow 0. \tag{46}$$

3. (Agreement between two chains). *The original and truncated chains agree throughout the time horizon with probability tending to one, meaning that*

$$\mathbb{P}\left(z_k^\eta = \widetilde{z}_k^\eta \text{ for every } 0 \le k \le N_\eta(T)\right) \longrightarrow 1. \tag{47}$$

4. (Truncation biases). *There exists $C < \infty$, such that*

$$\sup_{z \in \mathcal{U}_{\zeta_1}} \|b_\eta(z)\| \le C\eta^{2p-1}, \qquad \sup_{z \in \mathcal{U}_{\zeta_1}} \|M_2^\eta(z) - M_2(z)\| \le C\eta^{2p-2}. \tag{48}$$

Note that all limits are taken as $\eta \downarrow 0$, and constants are independent of (sufficiently small) η.

The proof of the above lemma is in Section C.6.2. For the remainder of the proof, we analyze the truncated chain(the above lemma will allow us to transfer back to the original SGD process). To simplify notation, we suppress the tildes. Denote z_k for the truncated chain and $\tau_\eta := \inf\{k \ge 0 : z_k \notin \mathcal{U}_{\zeta_1}\}$. The notation g^η and M_2^η always denotes truncated quantities, while g and M_2 continue to denote the original sample gradient and its raw second moment. For $k < \tau_\eta$, set

$$v_k := \pi(z_k), \qquad n_k := z_k - v_k, \qquad q_k := \eta^{-1/2} n_k, \qquad g_{k+1}^\eta := g^\eta(z_k; x_{k+1}).$$

As mentioned above, since distance to $\mathcal{M}$ is 1-Lipschitz and $\eta\|g_{k+1}^\eta\| \le \delta$, the entire update segment lies in $\mathcal{U}_{\zeta_2}$. Thus Taylor expansions of π and $n(z) := z - \pi(z)$ are valid, including at the endpoint of an update that exits the inner tube.

2. Fast normal dynamics and slow base-point motion. The preceding localization allows us to expand the normal and projected coordinates along each truncated update. The following lemma identifies the leading conditional drift and second moment of the normal increment and quantifies the slower motion of the base point.

Lemma C.12 (Normal-increment estimates and slow motion of base point). *Under the preceding setup, consider the truncated chain and write the normal increment $\delta q_k := q_{k+1} - q_k$. For $k < \tau_\eta$, define $v_{k+1} := \pi(z_{k+1})$, $q_{k+1} := \eta^{-1/2}(z_{k+1} - v_{k+1})$. Then there exist bounded, nondecreasing functions $\omega_H, \omega_g : [0, \zeta_2] \to [0, \infty)$, independent of η and k, satisfying $\omega_H(r) \to 0$, $\omega_g(r) \to 0$ as $r \downarrow 0$, and a constant $C < \infty$ such that the following estimates hold almost surely on $\{k < \tau_\eta\}$, uniformly in k for sufficiently small η.*

1. (Conditional first moment).

$$\mathbb{E}[\delta q_k \mid \mathcal{F}_k] = -\eta H_N(v_k) q_k + r_k^{(1)}, \tag{49}$$

where

$$\|r_k^{(1)}\| \le C\eta\, \omega_H(\sqrt{\eta}\|q_k\|)\|q_k\| + C\eta^{3/2}(1 + \|q_k\|^2). \tag{50}$$

2. (Conditional second moment).

$$\mathbb{E}[\delta q_k^{\otimes 2} \mid \mathcal{F}_k] = \eta \Sigma_N(v_k) + r_k^{(2)}, \tag{51}$$

where

$$\|r_k^{(2)}\| \le C\eta^{3/2}(1 + \|q_k\|) + C\eta\, \omega_g(\sqrt{\eta}\|q_k\|). \tag{52}$$

3. (Higher conditional moments).

$$\mathbb{E}[\|\delta q_k\|^3 \mid \mathcal{F}_k] \le C\eta^{3/2}, \tag{53}$$

$$\mathbb{E}[\|\delta q_k\|^4 \mid \mathcal{F}_k] \le C\eta^2. \tag{54}$$

4. (Slow motion of the base point). *The weighted increment satisfies*

$$\mathbb{E}\left[\|v_{k+1}-v_k\|(1+\|q_{k+1}\|^2)\mid \mathcal{F}_k\right]\le C\eta(1+\|q_k\|^2). \tag{55}$$

Moreover, for every fixed $R\ge 1$, there exists $C_R<\infty$ such that, on $\{k<\tau_\eta,\ \|q_k\|\le R\}$,

$$\frac{1}{\eta}\mathbb{E}\left[\|v_{k+1}-v_k\|(1+\|q_{k+1}\|^2)\mid \mathcal{F}_k\right]\le \vartheta_R(\eta), \tag{56}$$

where $\vartheta_R(\eta):=C_R[\sqrt{\eta}+\omega_g(\sqrt{\eta}R)]\longrightarrow 0$.

Proof of the above lemma can be found in Section C.6.2.

The conditional-moment estimates identify the leading normal dynamics, while Eq. (56) shows that, for bounded q_k, the expected base-point increment is $o(\eta)$. Together with the normal-moment bounds, this expresses the separation between the fast normal fluctuations and the slow motion along $\mathcal{M}$. We next identify the OU generator obtained by holding the base point fixed. The subsequent Poisson-equation argument will justify using these fixed-base dynamics along the evolving process.

FROZEN DISCRETE GENERATOR. Using the above estimates, we next identify the fast normal dynamics that will be averaged in the following step. For $v\in\mathcal{M}$, $q\in N_v\mathcal{M}$, and sufficiently small η, define the one-step frozen normal increment by (recall that $n(z):=z-\pi(z)$)

$$\Delta^\eta(v,q;x):=\eta^{-1/2}n(v+\sqrt{\eta}q-\eta g^\eta(v+\sqrt{\eta}q;x))-q.$$

For a test function $\varphi\in C_b^3(\mathbb{R}^D)$, define the corresponding discrete generator by (divide by η since one normal update corresponds to a fast-time increment of length η)

$$\mathscr{L}_{v,\eta}\varphi(q):=\frac{1}{\eta}\mathbb{E}\left[\varphi\big(q+\Delta^\eta(v,q;x)\big)-\varphi(q)\right].$$

We now show that this discrete generator converges to the generator of a frozen OU process. For every fixed $R<\infty$, the conditional-moment estimates Eq. (49)-(53), applied at the deterministic state $v+\sqrt{\eta}q$, give

$$\mathbb{E}\left[\Delta^\eta(v,q;x)\right]=-\eta H_N(v)q+r_1^\eta(v,q),\quad \mathbb{E}\big[\Delta^\eta(v,q;x)^{\otimes 2}\big]=\eta\Sigma_N(v)+r_2^\eta(v,q), \tag{57}$$

where, uniformly over $v\in\mathcal{M}$ and $q\in N_v\mathcal{M}$ with $\|q\|\le R$,

$$\|r_1^\eta(v,q)\|\le C_R\eta^{3/2}, \tag{58}$$

$$\|r_2^\eta(v,q)\|\le C_R\eta^{3/2}+C_R\eta\,\omega_g(\sqrt{\eta}R), \tag{59}$$

$$\mathbb{E}\left[\|\Delta^\eta(v,q;x)\|^3\right]\le C_R\eta^{3/2}. \tag{60}$$

The constants can be chosen uniformly because $\{(v,q)\mid v\in\mathcal{M},\ q\in N_v\mathcal{M},\ \|q\|\le R\}$ is compact. Using Taylor's expansion,

$$\varphi(q+\Delta)-\varphi(q)=\langle\nabla_q\varphi(q),\Delta\rangle+\frac{1}{2}\Delta^\top\nabla_q^2\varphi(q)\Delta+R_\varphi(q,\Delta), \tag{61}$$

where $|R_\varphi(q,\Delta)|\le\frac{\|D^3\varphi\|_\infty}{6}\|\Delta\|^3$. Taking expectations in Eq. (61), dividing by η, and $\mathbb{E}\left[\Delta^\top A\Delta\right]=\mathrm{Tr}\big(A\,\mathbb{E}[\Delta^{\otimes 2}]\big)$ gives

$$\mathscr{L}_{v,\eta}\varphi(q)=\left\langle\nabla_q\varphi(q),\frac{1}{\eta}\mathbb{E}[\Delta^\eta]\right\rangle+\frac{1}{2}\,\mathrm{Tr}\left(\nabla_q^2\varphi(q)\frac{1}{\eta}\mathbb{E}[(\Delta^\eta)^{\otimes 2}]\right)+\frac{1}{\eta}\mathbb{E}[R_\varphi(q,\Delta^\eta)]. \tag{62}$$

Substituting Eq. (57) into Eq. (62) yields

$$\mathscr{L}_{v,\eta}\varphi(q)=-\left\langle H_N(v)q,\nabla_q\varphi(q)\right\rangle+\frac{1}{2}\,\mathrm{Tr}\left(\Sigma_N(v)\nabla_q^2\varphi(q)\right)+\mathcal{E}_\varphi^\eta(v,q),$$

where

$$|\mathcal{E}_\varphi^\eta(v,q)|\le\|\nabla\varphi\|_\infty\frac{\|r_1^\eta(v,q)\|}{\eta}+\frac{1}{2\eta}\left|\mathrm{Tr}\left(\nabla_q^2\varphi(q)r_2^\eta(v,q)\right)\right|+\frac{\|D^3\varphi\|_\infty}{6\eta}\mathbb{E}\|\Delta^\eta(v,q;x)\|^3. \tag{63}$$

Consequently, using $|\operatorname{Tr}(AB)| \le \|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}} \le D\|A\|_2\|B\|_2$,

$$\sup_{\substack{v\in\mathcal{M},\ q\in N_v\mathcal{M}\\ \|q\|\le R}} |\mathcal{E}^\eta_\varphi(v,q)| \le C_{\varphi,R}\,[\sqrt{\eta}+\omega_g(\sqrt{\eta}R)] \longrightarrow 0.$$

We have therefore proved $\sup_{v\in\mathcal{M},\ q\in N_v\mathcal{M},\|q\|\le R} |\mathscr{L}_{v,\eta}\varphi(q) - \mathscr{L}_v\varphi(q)| \longrightarrow 0$, where

$$\mathscr{L}_v\varphi(q) = -\langle H_N(v)q, \nabla_q\varphi(q)\rangle + \frac{1}{2}\operatorname{Tr}\left(\Sigma_N(v)\nabla_q^2\varphi(q)\right). \tag{64}$$

Here $H_N(v)$ and $\Sigma_N(v)$ are regarded as operators on the fixed normal space $N_v\mathcal{M}$. The operator $\mathscr{L}_v$ is the generator of the frozen OU process $dq_s = -H_N(v)q_s\,ds + \Sigma_N(v)^{1/2}dB_s$, where $q_s \in N_v\mathcal{M}$. Because there exists $c_* > 0$ such that $H_N(v) \succeq c_*I$ on $N_v\mathcal{M}$, this process has a unique centered Gaussian distribution with covariance

$$\Pi_v = \int_0^\infty e^{-H_N(v)s}\Sigma_N(v)e^{-H_N(v)s}\,ds.$$

To see this, note that the integral converges since $\|e^{-H_N(v)s}\| \le e^{-c_*s}$. Moreover,

$$\begin{aligned} H_N(v)\Pi_v + \Pi_v H_N(v) &= -\int_0^\infty \frac{d}{ds}\left[e^{-H_N(v)s}\Sigma_N(v)e^{-H_N(v)s}\right]\mathrm{d}s \\ &= \Sigma_N(v), \end{aligned}$$

which is the associated Lyapunov equation.

3. Quadratic normal averaging. The stationary covariance Π_v identifies the average quadratic contribution

$$b(v) := \mathbb{E}_{q\sim\mathcal{N}(0,\Pi_v)}[\mathcal{B}_v[q,q]].$$

However, the actual normal fluctuations evolve together with the moving base point v_k. The following lemma justifies replacing their accumulated quadratic contribution by this local stationary average.

Lemma C.13 (Quadratic normal averaging). *Under the standing assumptions and the preceding localization setup, for every fixed $T > 0$,*

$$\sup_{0\le t\le T}\left\|\eta^2\sum_{k<N_\eta(t)\wedge\tau_\eta}(\mathcal{B}_{v_k}[q_k,q_k] - b(v_k))\right\| \xrightarrow{\mathbb{P}} 0 \qquad \text{as } \eta\downarrow 0. \tag{65}$$

As briefly introduced in Section 4, the proof (deferred to Section C.6.2) uses a quadratic solution of the OU Poisson equation at each base point. Evaluating this solution along the iterates converts the centered quadratic contribution into a telescoping term, a martingale, and a remainder caused by the moving base point. The established increment and moment estimates show that these terms vanish after the required scaling.

4. One-step expansion of the projected process. We now connect the quadratic normal fluctuation analyzed in the previous step to the actual motion of the projected iterate. The goal is to prove that, for every fixed $T < \infty$,

$$\sup_{0\le t\le T}\left\|v_{N_\eta(t)\wedge\tau_\eta} - v_0 + \frac{\eta^2}{2}\sum_{k<N_\eta(t)\wedge\tau_\eta}\mathcal{B}_{v_k}[q_k,q_k]\right\| \xrightarrow{\mathbb{P}} 0, \qquad N_\eta(t) := \left\lfloor\frac{t}{\eta^2}\right\rfloor. \tag{66}$$

Define the conditional mean, truncation bias, and centered truncated noise by $\overline{g}^\eta(z) := \mathbb{E}[g^\eta(z;x)] = \nabla L(z) + b_\eta(z)$, $\xi^\eta(z;X) := g^\eta(z;x) - \overline{g}^\eta(z)$. Then

$$\mathbb{E}[\xi^\eta(z_k;x_{k+1}) \mid \mathcal{F}_k] = 0, \qquad \sup_{z\in\mathcal{U}_{\varsigma_1}}\|b_\eta(z)\| \le C\eta^{2p-1} \tag{67}$$

from the step 1.

TAYLOR EXPANSION OF THE PROJECTION. Following the approach in the proof of Lemma C.12, we employ a Taylor expansion. In this case, however, the expansion is performed with respect to the projection, in contrast to the lemma where the normal increments were expanded. Taylor's theorem gives

$$\begin{aligned} v_{k+1} - v_k = &- \eta D\pi(z_k)\nabla L(z_k) - \eta D\pi(z_k) b_\eta(z_k) - \eta D\pi(z_k)\xi^\eta(z_k; x_{k+1}) \\ &+ \frac{\eta^2}{2} D^2\pi(z_k[g^\eta_{k+1}, g^\eta_{k+1}] + R^\pi_{k+1}, \end{aligned} \tag{68}$$

where $\|R^\pi_{k+1}\| \le C\eta^3 \|g^\eta_{k+1}\|^3$. Because truncation does not increase moments, due to the assumption on the moments of g and $\eta^2 N_\eta(T) \le T$,

$$\mathbb{E}\left[\sum_{k<N_\eta(T)\wedge\tau_\eta} \|R^\pi_{k+1}\|\right] \le C\eta^3 N_\eta(T) \sup_{z\in\mathcal{U}_{\zeta_1}} \mathbb{E}\|g(z;X)\|^3 \le C_T\eta \longrightarrow 0. \tag{69}$$

We now identify the leading term of Eq. (68) and control the other terms.

POPULATION DRIFT AND TRUNCATION BIAS. The projected-gradient expansion from Lemma C.14 gives, for $z_k = v_k + n_k$, the leading term $D\pi(z_k)\nabla L(z_k) = \frac{1}{2}\mathcal{B}_{v_k}[n_k, n_k] + r_T(v_k, n_k)$, where $\|r_T(v_k, n_k)\| \le C\|n_k\|^3$. Since $n_k = \sqrt{\eta}\, q_k$,

$$-\eta D\pi(z_k)\nabla L(z_k) = -\frac{\eta^2}{2}\mathcal{B}_{v_k}[q_k, q_k] + r^{\mathrm{pop}}_{k+1},$$

where

$$\|r^{\mathrm{pop}}_{k+1}\| \le C\eta^{5/2}\|Q_k\|^3. \tag{70}$$

The uniform $2p$-moment bound (Eq. (45)) for the killed normal process, with $2p > 3$, implies

$$\mathbb{E}\left[\sum_{k<N_\eta(T)\wedge\tau_\eta} \|r^{\mathrm{pop}}_{k+1}\|\right] \le C\eta^{5/2} N_\eta(T) \sup_{k\le N_\eta(T)} \mathbb{E}\|q_k\|^3 \le C_T\eta^{1/2} \longrightarrow 0. \tag{71}$$

Moreover, boundedness of $D\pi$ and Eq. (67) give

$$\sum_{k<N_\eta(T)\wedge\tau_\eta} \eta\|D\pi(z_k) b_\eta(z_k)\| \le C\eta N_\eta(T)\eta^{2p-1} \le C_T\eta^{2p-2} \longrightarrow 0. \tag{72}$$

PROJECTED LINEAR-NOISE TERM. We first quantify the fact that the sample-gradient noise has no tangential component on the stationary manifold. For the original sample-gradient noise $\xi(z;x) := g(z;x) - \nabla L(z)$, define $\alpha_1(r) := \sup_{z\in\mathcal{U}_{\zeta_2}, \mathrm{dist}(z,\mathcal{M})\le r} \mathbb{E}\|D\pi(z)\xi(z;x)\|^2$. Then $\alpha_1(r) \longrightarrow 0$, as $r \downarrow 0$. To see this, put $v = \pi(z)$. Since $\nabla L(v) = 0$ and the sample gradient is normal to $\mathcal{M}$ at v, $D\pi(v)\xi(v;x) = P_T(v)\xi(v;x) = 0$ a.s. Consequently,

$$D\pi(z)\xi(z;x) = \big(D\pi(z) - P_T(v)\big)\xi(z;x) + P_T(v)\big(\xi(z;x) - \xi(v;x)\big).$$

The first term converges to zero in L^2 by smoothness of $D\pi$ and the uniform moment of ξ, following from the assumed uniform moment on g. For the second term,

$$\begin{aligned} &\mathbb{E}\left\|P_T(v)\big(\xi(z;x) - \xi(v;x)\big)\right\|^2 \\ &\le \mathbb{E}\|\xi(z;x) - \xi(v;x)\|^2 \le 2\left(\mathbb{E}\|g(z;x) - g(v;x)\|^4\right)^{1/2} + 2\|\nabla L(z) - \nabla L(v)\|^2. \end{aligned}$$

The first quantity tends to zero by the assumed L^4-continuity of the sample gradient g, and the second by continuity of the population gradient ∇L. Compactness of $\mathcal{M}$ makes the convergence uniform.

For the truncated noise, define $\alpha_{1,\eta}(r) := \sup_{z\in\mathcal{U}_{\zeta_1}, \mathrm{dist}(z,\mathcal{M})\le r} \mathbb{E}\|D\pi(z)\xi^\eta(z;x)\|^2$. Since $\xi^\eta(z;x) - \xi(z;x) = \big(g^\eta(z;x) - g(z;x)\big) - \mathbb{E}[g^\eta(z;x) - g(z;x)]$, the tail estimate Eq. (87) gives

$$\begin{aligned} \sup_{z\in\mathcal{U}_{\zeta_1}} \mathbb{E}\|\xi^\eta(z;x) - \xi(z;x)\|^2 &\le C \sup_{z\in\mathcal{U}_{\zeta_1}} \mathbb{E}\left[\|g(z;x)\|^2 \mathbf{1}_{\{\eta\|g(z;x)\|>\delta\}}\right] \\ &\le C\eta^{2p-2}. \end{aligned}$$

Thus, using boundedness of $D\pi$, $\alpha_{1,\eta}(r) \le 2\alpha_1(r) + C\eta^{2p-2}$.

The family $\alpha_{1,\eta}$ is uniformly bounded on $[0, \zeta_1]$ by definition. For every fixed $a > 0$, the killed normal moment bound Eq. (45) and Markov's inequality give

$$\begin{aligned}\sup_{k\le N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}}\alpha_{1,\eta}(\|n_k\|)\right] &\le \sup_{0\le r\le a} \alpha_{1,\eta}(r) + C \sup_{k\le N_\eta(T)} \mathbb{P}(\|n_k\| > a) \\ &\le 2 \sup_{0\le r\le a} \alpha_1(r) + C\eta^{2p-2} + C_{p,T}a^{-2p}\eta^p.\end{aligned}$$

Taking first $\eta \downarrow 0$ and then $a \downarrow 0$ yields

$$\sup_{k\le N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}}\alpha_{1,\eta}(\|n_k\|)\right] \longrightarrow 0. \tag{73}$$

Define the stopped martingale increment $\widehat{M}^{(1)}_{k+1} := -\eta \mathbf{1}_{\{k<\tau_\eta\}} D\pi(z_k)\xi^\eta(z_k; x_{k+1})$. Since $\{k < \tau_\eta\} \in \mathcal{F}_k$, this is a martingale-difference sequence. Doob's L^2 inequality Williams (1991) and Eq. (73) give

$$\begin{aligned}\mathbb{E}\left[\sup_{n\le N_\eta(T)} \left\|\sum_{k<n} \widehat{M}^{(1)}_{k+1}\right\|^2\right] &\le 4\eta^2 \sum_{k<N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}}\alpha_{1,\eta}(\|n_k\|)\right] \\ &\le 4T \sup_{k\le N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}}\alpha_{1,\eta}(\|n_k\|)\right] \longrightarrow 0.\end{aligned} \tag{74}$$

QUADRATIC PROJECTION TERM. Define $A^\eta(z, x) := D^2\pi(z)[g^\eta(z; x), g^\eta(z; x)]$. If $v \in \mathcal{M}$, then $g^\eta(v; x)$ is normal to $\mathcal{M}$ almost surely, since truncation only multiplies $g(v; x)$ by a scalar indicator. Moreover, the tubular-neighborhood representation implies that $\pi(v + n) = v$ for every sufficiently small $n \in N_v\mathcal{M}$. Thus, the restriction of π to the local affine normal fiber $v + N_v\mathcal{M}$ is constant. Hence $D^2\pi(v)[n_1, n_2] = 0$, where $n_1, n_2 \in N_v\mathcal{M}$. It follows that $A^\eta(v, x) = 0$ a.s.

To quantify continuity away from the manifold, define the untruncated quantity $A(z, x) := D^2\pi(z)[g(z; x), g(z; x)]$ and the modulus $\alpha_2(r) := \sup_{z\in\mathcal{U}_{\zeta_2}, \mathrm{dist}(z,\mathcal{M})\le r} \mathbb{E}\|A(z, x)\|$. Then we claim that $\alpha_2(r) \longrightarrow 0$ as $r \downarrow 0$. To show this, with $v = \pi(z)$ and $A(v, x) = 0$, the difference $A(z, x) - A(v, x)$ is bounded by

$$\|D^2\pi(z) - D^2\pi(v)\|\|g(z; x)\|^2 + C\|g(z; x) - g(v; x)\|\big(\|g(z; x)\| + \|g(v; x)\|\big).$$

Smoothness of $D^2\pi$, the uniform fourth moment of g, and its L^4-continuity imply the claimed convergence uniformly in v. Furthermore, since $A^\eta(z, x) = A(z, x)\mathbf{1}_{\{\eta\|g(z;x)\|\le\delta\}}$, we have

$$\mathbb{E}\|A^\eta(z, x)\| \le \alpha_2(\mathrm{dist}(z, \mathcal{M})). \tag{75}$$

Exactly as in Eq. (73), the killed normal moment bound and $\alpha_2(r) \to 0$ imply

$$\sup_{k\le N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}}\alpha_2(\|n_k\|)\right] \longrightarrow 0. \tag{76}$$

Decompose the stopped quadratic term into its predictable and centered parts:

$$\frac{\eta^2}{2}\mathbf{1}_{\{k<\tau_\eta\}}A^\eta(z_k, x_{k+1}) = \frac{\eta^2}{2}\mathbf{1}_{\{k<\tau_\eta\}}\mathbb{E}[A^\eta(z_k, x_{k+1}) \mid \mathcal{F}_k] + \widehat{M}^{(2)}_{k+1},$$

where $\widehat{M}^{(2)}_{k+1} := \frac{\eta^2}{2}\mathbf{1}_{\{k<\tau_\eta\}}\Big(A^\eta(z_k, x_{k+1}) - \mathbb{E}[A^\eta(z_k, x_{k+1}) \mid \mathcal{F}_k]\Big)$. The predictable part vanishes in L^1, since, using Eq. (75) and Eq. (76),

$$\begin{aligned}&\mathbb{E}\left[\sup_{n\le N_\eta(T)} \left\|\frac{\eta^2}{2}\sum_{k<n} \mathbf{1}_{\{k<\tau_\eta\}}\mathbb{E}[A^\eta(z_k, x_{k+1}) \mid \mathcal{F}_k]\right\|\right] \\ &\quad\le \frac{\eta^2 N_\eta(T)}{2} \sup_{k\le N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}}\alpha_2(\|n_k\|)\right] \longrightarrow 0.\end{aligned} \tag{77}$$

Also, boundedness of $D^2\pi$ and the uniform fourth moment of g give $\mathbb{E}\left[\|\widehat{M}_{k+1}^{(2)}\|^2 \mid \mathcal{F}_k\right] \le C\eta^4 \mathbf{1}_{\{k<\tau_\eta\}}$. Therefore, Doob's L^2 inequality yields

$$\mathbb{E}\left[\sup_{n\le N_\eta(T)}\left\|\sum_{k<n}\widehat{M}_{k+1}^{(2)}\right\|^2\right] \le 4C\eta^4 N_\eta(T) \le C_T\eta^2 \longrightarrow 0. \tag{78}$$

To conclude, for every deterministic $t \le N_\eta(T)$, sum Eq. (68) over $k < t \wedge \tau_\eta$. The left-hand side telescopes to $v_{t\wedge\tau_\eta} - v_0$. The bounds Eq. (69), Eq. (71), and Eq. (72) control the finite-variation remainders uniformly in t, while Eq. (74), Eq. (77), and Eq. (78) control the stochastic and quadratic-projection terms. This proves Eq. (66) for the stopped truncated chain.

5. Averaging and passage to the ODE. Combining Eq. (65) and Eq. (66) gives

$$v_{t\wedge\tau_\eta} = v_0^\eta - \frac{\eta^2}{2}\sum_{k<t\wedge\tau_\eta} b(v_k) + R_t^\eta, \tag{79}$$

where $\max_{0\le t\le N_\eta(T)}\|R_t^\eta\| \xrightarrow{\mathbb{P}} 0$. Note that b is Lipschitz on the compact manifold $\mathcal{M}$. To see this, for every ambient coordinate j, $\langle e_j, b(v)\rangle = \mathrm{Tr}\left(B_v^{(j)}\Pi_v\right)$. Since $B_v^{(j)}$ and Π_v are Lipschitz and uniformly bounded on $\mathcal{M}$ (see the proof of Lemma C.13), each coordinate of b is Lipschitz. Because the ambient dimension is finite, b is Lipschitz on $\mathcal{M}$. Compactness of $\mathcal{M}$ then allows the Lipschitz constant to be chosen uniformly.

To compare Eq. (79) with an ordinary Euclidean Euler scheme, extend b to a Lipschitz vector field on an ambient neighborhood of $\mathcal{M}$. Note that $b : \mathcal{M} \to \mathbb{R}^D$, $b(v) \in T_v\mathcal{M}$. However, the following ambient Euler scheme

$$\widehat{v}_{k+1}^\eta = \widehat{v}_k^\eta - \frac{\eta^2}{2}b(\widehat{v}_k^\eta), \qquad \widehat{v}_0^\eta = v_0^\eta. \tag{80}$$

shows that $\widehat{v}_{k+1}^\eta$ can leave $\mathcal{M}$, even though $b(\widehat{v}_k^\eta)$ is tangent. This is because a tangent step to a curved manifold need not lie on the manifold. Then $b(\widehat{v}_{k+1}^\eta)$ would be undefined. To avoid this, we therefore define b slightly off the manifold. In the tubular neighborhood, a natural extension is $b(z) := b(\pi(z))$, where $\pi(z)$ is the nearest point on $\mathcal{M}$. Since b is Lipschitz on $\mathcal{M}$ and π is smooth, this extension is Lipschitz on a sufficiently small tubular neighborhood. We continue to denote this extension by b.

For every $t \le N_\eta(T) \wedge \tau_\eta$, subtracting Eq. (80) from Eq. (79) and using the Lipschitz continuity of b gives $\|v_t - \widehat{v}_t^\eta\| \le \max_{j\le t}\|R_j^\eta\| + C\eta^2\sum_{k<t}\|v_k - \widehat{v}_k^\eta\|$. Applying discrete Grönwall for any $v \le N_\eta(T)\wedge\tau_\eta$ and taking $\max$ for both sides therefore yields

$$\max_{t\le N_\eta(T)\wedge\tau_\eta}\|v_t - \widehat{v}_t^\eta\| \le e^{CT}\max_{t\le N_\eta(T)}\|R_t^\eta\| \xrightarrow{\mathbb{P}} 0. \tag{81}$$

Let $v(t)$ be the unique solution of

$$\dot{v}(t) = -\frac{1}{2}b(v(t)), \qquad v(0) = v_0. \tag{82}$$

For every $v \in \mathcal{M}$, $b(v) = \sum_{\alpha=1}^{\dim N_v\mathcal{M}} \mathcal{B}_v\left[\mathsf{e}_\alpha, \Pi_v\mathsf{e}_\alpha\right] \in T_v\mathcal{M}$, where $\{\mathsf{e}_\alpha\}$ is any orthonormal basis of $N_v\mathcal{M}$. Thus, the vector field in Eq. (82) is tangent to $\mathcal{M}$, and the solution initialized at $v_0 \in \mathcal{M}$ remains on $\mathcal{M}$. The standard Euler estimate for a Lipschitz vector field gives

$$\sup_{0\le t\le T}\left\|\widehat{v}_{\lfloor t/\eta^2\rfloor}^\eta - v(t)\right\| \le C_T\left(\eta^2 + \|v_0^\eta - v_0\|\right) \longrightarrow 0. \tag{83}$$

To derive it, let $t_k := k\eta^2$. The ODE satisfies $v(t_{k+1}) = v(t_k) - \frac{1}{2}\int_{t_k}^{t_{k+1}} b(v(s))\,\mathrm{d}s$. By adding and subtracting $\frac{\eta^2}{2}b(v(t_k))$, $v(t_{k+1}) = v(t_k) - \frac{\eta^2}{2}b(v(t_k)) + r_k^\eta$, where $r_k^\eta = -\frac{1}{2}\int_{t_k}^{t_{k+1}}\left(b(v(s)) - b(v(t_k))\right)\mathrm{d}s$. Because b is Lipschitz and bounded on the relevant neighborhood, it satisfies that $\|b(v(s)) - b(v(t_k))\| \le C(s - t_k)$, and therefore

$$\|r_k^\eta\| \le C\int_{t_k}^{t_{k+1}}(s - t_k)\,\mathrm{d}s \le C\eta^4.$$

The Euler recursion is $\widehat{v}_{k+1}^{\eta} = \widehat{v}_k^{\eta} - \frac{\eta^2}{2} b(\widehat{v}_k^{\eta})$. Subtracting it from the exact grid-point recursion gives

$$\|\widehat{v}_{k+1}^{\eta} - v(t_{k+1})\| \le \left(1 + \frac{\eta^2 L_b}{2}\right) \|\widehat{v}_k^{\eta} - v(t_k)\| + C\eta^4.$$

Discrete Grönwall then gives $\max_{k \le N_\eta(T)} \|\widehat{v}_k^{\eta} - v(t_k)\| \le C_T \left(\|v_0^{\eta} - v_0\| + \eta^2\right)$. Combining Eq. (81) and Eq. (83) yields

$$\sup_{0 \le t \le T \wedge \eta^2 \tau_\eta} \left\| v_{\lfloor t/\eta^2 \rfloor} - v(t) \right\| \xrightarrow{\mathbb{P}} 0. \tag{84}$$

6. Removal of stopping and truncation. The convergence in Eq. (84) has so far been proved for the stopped truncated chain. The tube-confinement estimate Eq. (46) gives $\mathbb{P}\big(\tau_\eta \le N_\eta(T)\big) \longrightarrow 0$. Moreover, the full-process equivalence estimate Eq. (47) gives

$$\mathbb{P}\left(z_k^{\eta} \ne \widetilde{z}_k^{\eta} \text{ for some } k \le N_\eta(T)\right) \longrightarrow 0, \tag{85}$$

where z_k^{η} and $\widetilde{z}_k^{\eta}$ denote the original and truncated recursions, respectively.

If neither event in Eq. (46)-Eq. (85) occurs, then the original, truncated, and stopped projected processes coincide throughout $[0, T]$. Consequently, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{0 \le t \le T} \|v^{\eta}(t) - v(t)\| > \varepsilon\right)$$

$$\le \mathbb{P}\left(\sup_{0 \le t \le T \wedge \eta^2 \tau_\eta} \left\| v_{\lfloor t/\eta^2 \rfloor} - v(t) \right\| > \varepsilon\right) + \mathbb{P}\big(\tau_\eta \le N_\eta(T)\big) + \mathbb{P}\left(z_k^{\eta} \ne \widetilde{z}_k^{\eta} \text{ for some } k \le N_\eta(T)\right).$$

Each term on the right-hand side converges to zero. Therefore, $\sup_{0 \le t \le T} \|v^{\eta}(t) - v(t)\| \xrightarrow{\mathbb{P}} 0$, where

$$\dot{v}(t) = -\frac{1}{2} b(v(t)) = -\frac{1}{2} \sum_{\alpha=1}^{\dim N_{v(t)}\mathcal{M}} \mathcal{B}_{v(t)}\big[\mathsf{e}_\alpha, \Pi_{v(t)} \mathsf{e}_\alpha\big], \qquad v(0) = v_0.$$

This completes the proof. □

C.6.2 Auxiliary Lemmas and Proofs for Phase 2

Proof of Lemma C.11 We first establish the tube confinement of original recursion, and then we work on the truncated recursion.

1. Original recursion. Note that the tube confinement and the normal moment bounds of the original recursion follows the derivation of Eq. (33) with τ replaced by τ_n (since it uses only normal coercivity and the uniform sample-gradient moment bound before exit, not the size of the initial displacement). Thus, setting $a_k := \mathbb{E}[d(z_k^{\eta})^{2p} \mathbf{1}_{\{k < \tau_\eta\}}]$, we obtain, by telescoping the sum (see Eq. (34)),

$$a_{k+1} \le (1 - c_p \eta) a_k + C_p \eta^{p+1}, \qquad a_k \le e^{-c_p \eta k} a_0 + \frac{C_p}{c_p} \eta^p.$$

The late-phase initialization gives $a_0 \le \mathbb{E} d(z_0^{\eta})^{2p} \le \eta^p \mathbb{E} \|q_0^{\eta}\|^{2p} \le C\eta^p$. Consequently, $a_k \le C'\eta^p$ uniformly in k (using $e^{-c_p \eta k} \le 1$), which proves the normal-moment bounds of the original recursion.

We next show that the trajectory does not leave the fixed tube. First, Markov's inequality gives $\mathbb{P}(\tau_\eta = 0) \le \zeta^{-2p} \mathbb{E} d(z_0^{\eta})^{2p} \le C\eta^p$. For $k \ge 0$,

$$\begin{aligned} \mathbb{P}(\tau_\eta = k+1) &= \mathbb{E}\left[\mathbf{1}_{\{\tau_\eta > k\}} \mathbf{1}_{\{d(z_{k+1}^{\eta}) \ge \zeta\}}\right] \\ &\le \zeta^{-2p} \mathbb{E}\left[\mathbf{1}_{\{\tau_\eta > k\}} d(z_{k+1}^{\eta})^{2p}\right] \le C_T \eta^p, \end{aligned}$$

where the last inequality follows from the above one step estimate and Eq. (44). Summing over $k < N_\eta(T)$ yields

$$\mathbb{P}\left(\tau_\eta \le N_\eta(T)\right) \le C_T \eta^{p-2} \longrightarrow 0.$$

Thus, the normal displacement has the $O(\sqrt{\eta})$ moment scale at each iteration, while the complete trajectory remains inside the fixed tubular neighborhood with probability tending to one.

2. TRUNCATED CHAIN We first verify the asymptotic agreement of the two chains. Define

$$\sigma_\eta := \inf\left\{m \geq 1 \mid \eta \left\|g(\widetilde{z}^\eta_{m-1}; x_m)\right\| > \delta\right\}.$$

Note that $\tilde{\tau}_\eta$ records when the truncated chain leaves the inner tube $\mathcal{U}_{\zeta_1}$, whereas σ_η records the first time truncation is activated. Then $\{\sigma_\eta > k\} \in \mathcal{F}_k$, and the original and truncated chains agree through time k on this event. Note that since,

$$\{\sigma_\eta \leq N_\eta(T) \wedge \widetilde{\tau}_\eta\} \subseteq \bigcup_{k=0}^{N_\eta(T)-1} \left(\{\sigma_n > k\} \cap \{\widetilde{\tau}_\eta > k\} \cap \{\eta \|g(\widetilde{z}^\eta_k; x_{k+1})\| > \delta\}\right),$$

for $k < N_\eta(T)$, conditional Markov inequality gives

$$\begin{aligned}\mathbb{P}\left(\sigma_\eta \leq N_\eta(T) \wedge \widetilde{\tau}_\eta\right) &\leq \sum_{k<N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{\sigma_\eta>k,\, \widetilde{\tau}_\eta>k\}} \mathbb{P}\left(\eta \|g(\widetilde{z}^\eta_k; x_{k+1})\| > \delta | \mathcal{F}_k\right)\right] \\ &\leq N_\eta(T) \left(\frac{\eta}{\delta}\right)^{2p} \sup_{z\in\mathcal{U}_{\zeta_1}} \mathbb{E}\|g(z;x)\|^{2p}.\end{aligned}$$

Hence

$$\mathbb{P}\left(\sigma_\eta \leq N_\eta(T) \wedge \widetilde{\tau}_\eta\right) \leq C_{p,T}\eta^{2p-2} \longrightarrow 0. \tag{86}$$

For bounding the truncation bounded, let $M_2(z) := \mathbb{E}[g(z;x)^{\otimes 2}]$ be the original raw second moment.[11] We first establish the elementary bound. Let $a := \frac{\delta}{\eta}$. On the event $\{\|g\| > a\}$, write $\|g\|^r = \|g\|^{2p}\|g\|^{-(2p-r)}$. Due to $r \leq 2p$, $2p - r \geq 0$ is nonnegative. Moreover, on $\{\|g\| > a\}$, $\|g\|^{-(2p-r)} \leq a^{-(2p-r)}$. Therefore,

$$\begin{aligned}\|g\|^r \mathbf{1}_{\{\|g\|>a\}} &= \|g\|^{2p}\|g\|^{-(2p-r)}\mathbf{1}_{\{\|g\|>a\}} \\ &\leq a^{-(2p-r)}\|g\|^{2p}\mathbf{1}_{\{\|g\|>a\}} \leq a^{-(2p-r)}\|g\|^{2p}.\end{aligned}$$

Taking expectations gives $\mathbb{E}\left[\|g\|^r \mathbf{1}_{\{\|g\|>a\}}\right] \leq a^{-(2p-r)}\mathbb{E}\|g\|^{2p}$. Substituting $a = \delta/\eta$, $a^{-(2p-r)} = \left(\frac{\eta}{\delta}\right)^{2p-r}$, thus

$$\mathbb{E}\left[\|g\|^r \mathbf{1}_{\{\|g\|>\delta/\eta\}}\right] \leq \left(\frac{\eta}{\delta}\right)^{2p-r} \mathbb{E}\|g\|^{2p}, \qquad r = 1, 2, \tag{87}$$

Since $g^\eta - g = -g\mathbf{1}_{\{\|g\|>\delta/\eta\}}$, the bound gives, uniformly for $z \in \mathcal{U}_{\zeta_1}$, $\|b_\eta(z)\| \leq C\eta^{2p-1}$, $\|M_2^\eta(z) - M_2(z)\| \leq C\eta^{2p-2}$, proving Eq. (48).

We next verify that the confinement and normal-moment bounds also hold for the truncated chain. Here, the same argument used above in the original recursion and in Eq. (33) (now applied to g^η) gives

$$\mathbb{E}[\widetilde{d}^{2p}_{k+1} \mid \mathcal{F}_k] \leq \widetilde{d}^{2p}_k - 2p\eta\widetilde{d}^{2p-2}_k \langle \widetilde{n}_k, \nabla L(\widetilde{z}^\eta_k) + b_\eta(\widetilde{z}^\eta_k)\rangle + C_p\eta^2\widetilde{d}^{2p-2}_k + C_p\eta^{2p}, \tag{88}$$

for $k < \widetilde{\tau}_\eta$. Here the moment terms have the same bounds as before, because $\|g^\eta(z;x)\| \leq \|g(z;x)\|$. The only additional contribution is the truncation bias. Using $\|b_\eta(z)\| \leq C\eta^{2p-1}$ and $\widetilde{d}_k < \zeta_1$ before exit, we obtain

$$2p\eta\widetilde{d}^{2p-2}_k |\langle \widetilde{n}_k, b_\eta(\widetilde{z}^\eta_k)\rangle| \leq C\eta^{2p}\widetilde{d}^{2p-1}_k \leq C\eta^{2p}.$$

Since $\eta^{2p} \leq \eta^{p+1}$ for $p > 1$ and $0 < \eta \leq 1$, this contribution is absorbed into the existing $O(\eta^{p+1})$ term. Normal coercivity and the same application of Young's inequality therefore yield

$$\mathbb{E}[\widetilde{d}^{2p}_{k+1} \mid \mathcal{F}_k] \leq (1 - c_p\eta)\widetilde{d}^{2p}_k + C_p\eta^{p+1}, \qquad k < \widetilde{\tau}_\eta.$$

Thus same argument gives $\sup_{k\leq N_\eta(T)} \mathbb{E}\left[\widetilde{d}^{2p}_k \mathbf{1}_{\{k<\widetilde{\tau}_\eta\}}\right] \leq C_{p,T}\eta^p$, $\mathbb{P}(\widetilde{\tau}_\eta \leq N_\eta(T)) \longrightarrow 0$ This, combined with the estimates of the original recursion, proves the Normal-moment bounds and the tube confinement. □

[11] The first conditional moment of the normal increment determines the restoring drift of the fast process, whereas M_2 determines its diffusion covariance and, ultimately its invariant covariance Π_v.

Proof of Lemma C.12

1. TAYLOR EXPANSION INSIDE THE LARGER TUBE. We first recall the following fact. For $k < \tau_\eta$, it satisfies that $\mathrm{dist}(z_k, \mathcal{M}) < \zeta_1$ and $\eta\|g^\eta(z_k; x_{k+1})\| \le \delta$. Therefore, for every $t \in [0, 1]$,

$$\mathrm{dist}\left(z_k - t\eta g^\eta(z_k; x_{k+1}), \mathcal{M}\right) < \zeta_1 + \delta = \zeta_2.$$

Hence the whole Taylor segment lies in $\mathcal{U}_{\zeta_2}$. Since $\mathcal{M}$ is compact and $\zeta_2 < \zeta$, the closure $\overline{\mathcal{U}_{\zeta_2}}$ is compactly contained in the tubular neighborhood $\mathcal{U}_\zeta$. Moreover, the C^4 regularity of $\mathcal{M}$ implies that the nearest-point projection π and $n(z) := z - \pi(z)$ are C^3 on $\mathcal{U}_\zeta$. Their derivatives through order three are therefore uniformly bounded on $\overline{\mathcal{U}_{\zeta_2}}$.

We next compare the stochastic gradient at z with that at its projection $\pi(z) \in \mathcal{M}$. Define the L^4 modulus

$$\omega_g(r) := \sup_{\substack{z \in \mathcal{U}_{\zeta_2} \\ \mathrm{dist}(z,\mathcal{M}) \le r}} \left(\mathbb{E}\|g(z; x) - g(\pi(z); x)\|^4\right)^{1/4}. \tag{89}$$

The local L^4-continuity assumption (S2) on the sample gradient implies $\omega_g(r) \longrightarrow 0$ as $r \downarrow 0$. Thus, when z approaches $\mathcal{M}$, the stochastic gradient at z approaches the stochastic gradient at its projection uniformly in L^4.

Writing $\Delta g := g(z; x) - g(\pi(z); x)$, we have the identity

$$g(z; x)^{\otimes 2} - g(\pi(z); x)^{\otimes 2} = \Delta g \otimes g(z; x) + g(\pi(z); x) \otimes \Delta g.$$

Since $\|u \otimes v\| = \|u\|\|v\|$, taking expectations and applying the triangle inequality and Cauchy-Schwarz gives

$$\|M_2(z) - M_2(\pi(z))\| \le \left(\mathbb{E}\|\Delta g\|^2\right)^{1/2} \left[\left(\mathbb{E}\|g(z; x)\|^2\right)^{1/2} + \left(\mathbb{E}\|g(\pi(z); x)\|^2\right)^{1/2}\right].$$

The uniform second-moment bound on g bounds the bracketed term by a constant. Moreover, $\left(\mathbb{E}\|\Delta g\|^2\right)^{1/2} \le \left(\mathbb{E}\|\Delta g\|^4\right)^{1/4} \le \omega_g(\mathrm{dist}(z, \mathcal{M}))$. Therefore,

$$\|M_2(z) - M_2(\pi(z))\| \le C\, \omega_g(\mathrm{dist}(z, \mathcal{M})). \tag{90}$$

For $k < \tau_\eta$, set $v_k := \pi(z_k)$, $n_k := n(z_k) = z_k - v_k$, $q_k := \eta^{-1/2} n_k$, $g^\eta_{k+1} := g^\eta(z_k; x_{k+1})$, and let $\delta q_k := q_{k+1} - q_k$. Taylor expansion of $n(z) := z - \pi(z)$ at z_k in the direction $-\eta g^\eta_{k+1}$ gives

$$\delta q_k = -\sqrt{\eta}\, Dn(z_k) g^\eta_{k+1} + \frac{\eta^{3/2}}{2} D^2 n(z_k)[g^\eta_{k+1}, g^\eta_{k+1}] + R^q_{k+1}, \tag{91}$$

where

$$\|R^q_{k+1}\| \le C\eta^{5/2} \|g^\eta_{k+1}\|^3. \tag{92}$$

Since the nearest-point projection is C^3 on the tubular neighborhood, the normal-coordinate map $n(z) := z - \pi(z)$ has uniformly bounded second derivatives on a smaller closed tube. Moreover, $Dn(v) = I - D\pi(v) = P_N(v)$, for $v \in \mathcal{M}$. Consequently, uniformly over $v \in \mathcal{M}$ and sufficiently small $n \in N_v\mathcal{M}$,

$$\left\|Dn(v + n) - P_N(v)\right\| \le C\|n\|. \tag{93}$$

Since $L \in C^2$ near $\mathcal{M}$, compactness gives a modulus $\omega_H(r) \to 0$ as $r \downarrow 0$ such that $\left\|\nabla^2 L(v + tn) - \nabla^2 L(v)\right\| \le \omega_H(\|n\|)$, for $0 \le t \le 1$. Using $\nabla L(v) = 0$ and $\nabla^2 L(v) n = H_N(v) n$ for $n \in N_v\mathcal{M}$, the integral form of Taylor's theorem gives

$$\nabla L(v + n) = H_N(v) n + R_L(v, n), \qquad \|R_L(v, n)\| \le \omega_H(\|n\|)\|n\| = o(\|n\|), \tag{94}$$

uniformly over $v \in \mathcal{M}$.

For the vector-valued map $\mathsf{n} = (\mathsf{n}_1, \ldots, \mathsf{n}_D)$, and a matrix $M \in \mathbb{R}^{D \times D}$, define the vector-valued trace

$$\mathrm{Tr}\big(D^2 \mathsf{n}(z) M\big) := \begin{bmatrix} \mathrm{Tr}\big(D^2 \mathsf{n}_1(z) M\big) \\ \vdots \\ \mathrm{Tr}\big(D^2 \mathsf{n}_D(z) M\big) \end{bmatrix}. \tag{95}$$

Equivalently, for any random vector U with $\mathbb{E}[U^{\otimes 2}] = M$, $\mathrm{Tr}\big(D^2 \mathsf{n}(z) M\big) = \mathbb{E}\big[D^2 \mathsf{n}(z)[U, U]\big]$.

2. CONDITIONAL FIRST MOMENT Taking conditional expectation in Eq. (91) gives

$$\mathbb{E}[\delta q_k \mid \mathcal{F}_k] = -\sqrt{\eta}\, Dn(z_k)\big(\nabla L(z_k) + b_\eta(z_k)\big) + \frac{\eta^{3/2}}{2}\,\mathrm{Tr}\big(D^2 n(z_k) M_2^\eta(z_k)\big) + \mathbb{E}[R_{k+1}^q \mid \mathcal{F}_k]. \tag{96}$$

Here the trace in the second line is understood componentwise according to Eq. (95). Since $n_k = \sqrt{\eta} q_k$, Eq. (93) and Eq. (94) give

$$\begin{aligned} -\sqrt{\eta}\, Dn(z_k)\nabla L(z_k) &= -\sqrt{\eta}\,(P_N(v_k) + O(\sqrt{\eta}\|q_k\|)) \times \big(\sqrt{\eta} H_N(v_k) q_k + O(\eta\|q_k\|^2)\big) \\ &= -\eta H_N(v_k) q_k + O\left(\eta^{3/2}\|q_k\|^2\right). \end{aligned} \tag{97}$$

The remaining terms in Eq. (96) satisfy

$$\begin{aligned} \sqrt{\eta}\,\|Dn(z_k) b_\eta(z_k)\| &\le C\eta^{2p-1/2}, \\ \frac{\eta^{3/2}}{2}\left\|\mathrm{Tr}\left(D^2 n(z_k) M_2^\eta(z_k)\right)\right\| &\le C\eta^{3/2}, \\ \left\|\mathbb{E}[R_{k+1}^q \mid \mathcal{F}_k]\right\| &\le C\eta^{5/2}. \end{aligned}$$

The second inequality follows from boundedness of $D^2 n$ and the uniform second moment of g^η, while the third follows from Eq. (92) and the uniform third moment. Consequently, $\mathbb{E}[\delta q_k \mid \mathcal{F}_k] = -\eta H_N(v_k) q_k + r_k^{(1)}$, where $\|r_k^{(1)}\| \le C\eta^{3/2}(1 + \|q_k\|^2)$.

3. CONDITIONAL SECOND MOMENT The leading term in Eq. (91) is $-\sqrt{\eta}\, Dn(z_k) g_{k+1}^\eta$. Its conditional second moment is $\eta Dn(z_k) M_2^\eta(z_k) Dn(z_k)^\top$. The cross product between this leading term and the quadratic Taylor term has conditional expectation $O(\eta^2)$, by the uniform third moment of g^η. The square of the quadratic term and all terms involving R_{k+1}^q are of smaller order. Hence

$$\mathbb{E}[\delta q_k^{\otimes 2} \mid \mathcal{F}_k] = \eta Dn(z_k) M_2^\eta(z_k) Dn(z_k)^\top + O(\eta^2). \tag{98}$$

We now compare each factor with its value at $v_k = \pi(z_k)$. The truncation estimate gives $\|M_2^\eta(z_k) - M_2(z_k)\| \le C\eta^{2p-2}$ (see the paragraph Truncation Biases), while Eq. (90) gives $\|M_2(z_k) - M_2(v_k)\| \le C\omega_g(\sqrt{\eta}\|q_k\|)$. Moreover, $\|Dn(z_k) - P_N(v_k)\| \le C\sqrt{\eta}\|q_k\|$. By definition, $M_2(v) = \Sigma_N(v)$. Since $\nabla L(v) = 0$, this raw second moment is also the covariance of the sample gradient. Assumption S3 implies $g(v; x) = P_N(v) g(v; x)$ a.s, thus $\Sigma_N(v) = P_N(v)\Sigma_N(v) P_N(v)$.

Expanding the three factors in Eq. (98) around their values at v_k therefore yields $\mathbb{E}[\delta q_k^{\otimes 2} \mid \mathcal{F}_k] = \eta\Sigma_N(v_k) + r_k^{(2)}$, where $\|r_k^{(2)}\| \le C\eta^{3/2}(1 + \|q_k\|) + C\eta\,\omega_g(\sqrt{\eta}\|q_k\|)$. Here, terms of order $\eta^2\|q_k\|^2$ are absorbed into $\eta^{3/2}\|q_k\|$, because $\sqrt{\eta}\|q_k\| = \|n_k\| \le \zeta_1$ before exit. The truncation contribution is $\eta\|M_2^\eta(z_k) - M_2(z_k)\| = O(\eta^{2p-1})$, which is absorbed into $O(\eta^2)$ due to $p > 4$.

4. HIGHER CONDITIONAL MOMENTS For $k < \tau_\eta$, define $\Gamma_k(t) := z_k - t\eta g_{k+1}^\eta$, for $0 \le t \le 1$. The truncation guarantees that $\Gamma_k(t) \in \mathcal{U}_{\zeta_2}$ for every $t \in [0, 1]$. Consequently, since n is C^3 on $\mathcal{U}_\zeta$, $C_n := \sup_{z \in \overline{\mathcal{U}_{\zeta_2}}} \|Dn(z)\| < \infty$. By the fundamental theorem of calculus,

$$\delta q_k = \eta^{-1/2}\,(n(z_{k+1}) - n(z_k)) = -\sqrt{\eta}\int_0^1 Dn(\Gamma_k(t)) g_{k+1}^\eta\, dt.$$

It follows that

$$\|\delta q_k\| \le C_n\sqrt{\eta}\,\|g_{k+1}^\eta\|. \tag{99}$$

For every $r \le 2p$, truncation and the uniform $2p$-moment assumption give

$$\begin{aligned} \mathbb{E}[\|g_{k+1}^\eta\|^r \mid \mathcal{F}_k] &\le \mathbb{E}[\|g(z_k; x_{k+1})\|^r \mid \mathcal{F}_k] \\ &\le \big(\mathbb{E}[\|g(z_k; x_{k+1})\|^{2p} \mid \mathcal{F}_k]\big)^{r/(2p)} \le C_r. \end{aligned}$$

Raising Eq. (99) to the r-th power and taking conditional expectations therefore yields $\mathbb{E}[\|\delta q_k\|^r \mid \mathcal{F}_k] \le C_r\eta^{r/2}$. In particular,

$$\mathbb{E}[\|\delta q_k\|^3 \mid \mathcal{F}_k] \le C\eta^{3/2}, \tag{100}$$

$$\mathbb{E}[\|\delta q_k\|^4 \mid \mathcal{F}_k] \le C\eta^2. \tag{101}$$

5. SLOW MOTION OF THE BASE POINT Taylor expansion of π gives

$$v_{k+1} - v_k = -\eta D\pi(z_k) g^\eta_{k+1} + R^v_{k+1}, \qquad \|R^v_{k+1}\| \le C\eta^2 \|g^\eta_{k+1}\|^2. \tag{102}$$

For the original gradient, define

$$\kappa(r) := \sup_{\substack{z \in \mathcal{U}_{\zeta_2} \\ \operatorname{dist}(z,\mathcal{M}) \le r}} \left(\mathbb{E}\|D\pi(z) g(z;x)\|^2\right)^{1/2}.$$

If $v = \pi(z)$, then $D\pi(z)g(z;x) = (D\pi(z) - P_T(v))\, g(z;x) + P_T(v)\,(g(z;x) - g(v;x))$, since $P_T(v)g(v;x) = 0$ a.s. Smoothness of π, the uniform second moment, and Eq. (89) give

$$\kappa(r) \le Cr + C\omega_G(r) \longrightarrow 0.$$

Since $D\pi(z)g^\eta(z;x) = D\pi(z)g(z;x)\mathbf{1}_{\{\eta\|g(z;x)\| \le \delta\}}$, the same bound holds for the truncated gradient. For every fixed $R < \infty$, Hölder's inequality, Eq. (102), and Eq. (54) give

$$\sup_{\substack{k < \tau_\eta \\ \|q_k\| \le R}} \frac{1}{\eta} \mathbb{E}\left[\|v_{k+1} - v_k\|(1 + \|q_{k+1}\|^2) \mid \mathcal{F}_k\right] \le C_R\left[\kappa(\sqrt{\eta}R) + \eta\right] \longrightarrow 0.$$

□

Proof of Lemma C.13 Lemma C.12 identifies the leading conditional moments of the rescaled normal variable q_k with those of an OU process on the fast time scale $s = k\eta$, with the base point held fixed. By contrast, substituting $n_k = \sqrt{\eta} q_k$ into Lemma C.14 shows that the leading population-gradient contribution to the projected update is $-\frac{\eta^2}{2}\mathcal{B}_{v_k}[q_k, q_k]$. This motivates observing the projected process on the slower time scale $t = k\eta^2$.

Over $O(\eta^{-1})$ iterations, corresponding to an order-one interval of fast time, the slow-motion and normal-moment estimates ensure that the base point changes by $o_{\mathbb{P}}(1)$, while the normal dynamics evolve on their natural time scale. On the longer $O(\eta^{-2})$ iteration scale, this separation suggests replacing the accumulated quadratic contribution by $b(v) := \mathbb{E}_{q \sim \mathcal{N}(0, \Pi_v)}[\mathcal{B}_v[q,q]]$. The purpose of this lemma is to justify that replacement uniformly on the slow time interval, while accounting for the fact that the base point continues to move.

For $v \in \mathcal{M}$, let $(\mathsf{e}_\alpha)_{\alpha=1}^{\dim N_v\mathcal{M}}$ be any orthonormal basis of $N_v\mathcal{M}$, and define $\overline{\mathcal{B}}(v) := \sum_{\alpha=1}^{\dim N_v\mathcal{M}} \mathcal{B}_v[\mathsf{e}_\alpha, \Pi_v \mathsf{e}_\alpha]$ (this is the normal-space trace of the bilinear map $\mathcal{B}_v$ against the covariance Π_v). Our goal is to prove that, for every $T < \infty$,

$$\sup_{0 \le t \le T} \left\| \eta^2 \sum_{k < N_\eta(t) \wedge \tau_\eta} \left(\mathcal{B}_{v_k}[q_k, q_k] - \overline{\mathcal{B}}(v_k)\right) \right\| \xrightarrow{\mathbb{P}} 0, \qquad N_\eta(t) := \left\lfloor \frac{t}{\eta^2} \right\rfloor. \tag{103}$$

The summation over k arises because the slow tangential motion is the cumulative effect of $O(\eta^{-2})$ projected SGD updates, each of size $O(\eta^2)$. Although the instantaneous centered fluctuation $\mathcal{B}_{v_k}[q_k, q_k] - \overline{\mathcal{B}}(v_k)$ does not vanish, its accumulated contribution is expected to cancel as the fast normal process equilibrates. We take the supremum over $t \in [0, T]$ because the limiting result concerns the entire projected trajectory.

For each frozen v, we represent the centered quadratic observable $\mathcal{B}_v[q,q] - \overline{\mathcal{B}}(v)$ as the negative OU generator applied to a quadratic corrector. To explain this, let q^v_s denote the frozen OU process. Its generator is the differential operator $\mathscr{L}_v\varphi(q) = -\langle H_N(v)q, \nabla_q\varphi(q)\rangle + \frac{1}{2}\operatorname{Tr}\left(\Sigma_N(v)\nabla^2_q\varphi(q)\right)$, which gives the infinitesimal expected change of a smooth test function $\mathscr{L}_v\varphi(q) = \lim_{h \downarrow 0} \frac{\mathbb{E}[\varphi(q^v_h) \mid q^v_0 = q] - \varphi(q)}{h}$. Because $\mathcal{B}_v[q,q]$ is vector-valued and since the basis of the tangent and normal space change as v evolves, we apply the argument to each fixed ambient coordinate. For coordinate j, we construct a quadratic function $\chi_j(v,q) = \langle q, C^{(j)}_v q\rangle$ that solves the Poisson equation (Pavliotis & Stuart, 2008) $-\mathscr{L}_v\chi_j(v,q) = \left\langle e_j, \mathcal{B}_v[q,q] - \overline{\mathcal{B}}(v)\right\rangle$. For one discrete SGD step, the corresponding identity is approximately

$$\mathbb{E}\left[\chi_j(v_k, q_{k+1}) - \chi_j(v_k, q_k) \mid \mathcal{F}_k\right] = -\eta\left\langle e_j, \mathcal{B}_{v_k}[q_k, q_k] - \overline{\mathcal{B}}(v_k)\right\rangle.$$

Summing these increments produces a telescoping boundary term, together with martingale and approximation remainders. The rest of this step shows that all three contributions vanish on the $O(\eta^{-2})$ time horizon.

1. COORDINATEWISE POISSON EQUATIONS Let $(e_1, \dots, e_D)$ be the fixed ambient Euclidean basis. For each $j \in \{1, \dots, D\}$, define the symmetric operator $B_v^{(j)} : N_v\mathcal{M} \to N_v\mathcal{M}$ by $\langle e_j, \mathcal{B}_v[q_1, q_2]\rangle = \left\langle q_1, B_v^{(j)} q_2 \right\rangle$. Because $\mathcal{B}_v$ depend smoothly (Assumption S4) on v, $v \mapsto B_v^{(j)}$ is also locally Lipschitz.

The domain and codomain of $B_v^{(j)}$ depend on v, because the normal space $N_v\mathcal{M}$ changes along the manifold. To compare these operators at different base points, we embed them into the fixed ambient parameter space $\mathbb{R}^D$. Every $\varsigma \in \mathbb{R}^D$ has the unique orthogonal decomposition $\varsigma = P_T(v)\varsigma + P_N(v)\varsigma$. Define the ambient extension of $B_v^{(j)}$ by $\widetilde{B}_v^{(j)}\varsigma := B_v^{(j)} P_N(v)\varsigma$. Thus $\widetilde{B}_v^{(j)}$ agrees with $B_v^{(j)}$ on $N_v\mathcal{M}$ and maps every tangent vector to zero. We extend Π_v to $\mathbb{R}^D$ in the same way. To simplify notation, we continue to write $B_v^{(j)}$ and Π_v for these ambient extensions.

The trace identity below is then an ordinary matrix identity on $\mathbb{R}^D$. If (e_α) is an orthonormal basis of $N_v\mathcal{M}$, then

$$\begin{aligned}\left\langle e_j, \overline{\mathcal{B}}(v)\right\rangle &= \sum_\alpha \langle e_j, \mathcal{B}_v[\mathsf{e}_\alpha, \Pi_v \mathsf{e}_\alpha]\rangle \\ &= \sum_\alpha \left\langle \mathsf{e}_\alpha, B_v^{(j)} \Pi_v \mathsf{e}_\alpha \right\rangle = \operatorname{Tr}\left(B_v^{(j)} \Pi_v\right). \end{aligned} \tag{104}$$

The tangent directions contribute zero to the ambient trace, so this is also the trace of $B_v^{(j)}\Pi_v$ restricted to $N_v\mathcal{M}$.

By Assumption S4, the ambient extensions of $B_v^{(j)}$ and $H_N(v)$ are locally Lipschitz in v. The tangent projection $P_T(v)$ is locally Lipschitz as well because $\mathcal{M}$ is C^4. We now establish the local lipschitzness of $C_v^{(j)}$, which is the unique symmetric normal-space solution of

$$H_N(v) C_v^{(j)} + C_v^{(j)} H_N(v) = B_v^{(j)}. \tag{105}$$

To compare these operators at different points without choosing normal coordinates, define the ambient extension $\widehat{H}(v) := H_N(v) + P_T(v)$, where $H_N(v)$ is extended by zero on $T_v\mathcal{M}$. Since $H_N(v) \succeq c_* I$ on $N_v\mathcal{M}$, $\widehat{H}(v) \succeq \widehat{c} I$, where $\widehat{c} := \min\{c_*, 1\} > 0$. Because $B_v^{(j)}$ vanishes on $T_v\mathcal{M}$, the solution of Eq. (105), relative to the orthogonal decomposition $\mathbb{R}^D = T_v\mathcal{M} \oplus N_v\mathcal{M}$, the relevant operators therefore have the block forms

$$\widehat{H}(v) = \begin{bmatrix} I & 0 \\ 0 & H_N(v) \end{bmatrix}, \qquad B_v^{(j)} = \begin{bmatrix} 0 & 0 \\ 0 & B_v^{(j)}|_{N_v\mathcal{M}} \end{bmatrix}.$$

It follows that the normal-space solution, extended by zero on the tangent space, has the ambient representation

$$C_v^{(j)} = \int_0^\infty e^{-\widehat{H}(v)s} B_v^{(j)} e^{-\widehat{H}(v)s}\, ds. \tag{106}$$

To see this, the block structure shows that the integrand, and hence $C_v^{(j)}$, vanishes on the tangent space and maps the normal space into itself. Moreover,

$$\frac{d}{ds}\left(e^{-\widehat{H}(v)s} B_v^{(j)} e^{-\widehat{H}(v)s}\right) = -\widehat{H}(v) e^{-\widehat{H}(v)s} B_v^{(j)} e^{-\widehat{H}(v)s} - e^{-\widehat{H}(v)s} B_v^{(j)} e^{-\widehat{H}(v)s} \widehat{H}(v).$$

Integrating this identity over $[0, \infty)$ and using exponential decay gives

$$\widehat{H}(v) C_v^{(j)} + C_v^{(j)} \widehat{H}(v) = B_v^{(j)}.$$

Since $P_T(v) C_v^{(j)} = C_v^{(j)} P_T(v) = 0$, this ambient equation reduces exactly to Eq. (105) on the normal space.

Now we work on obtaining the local Lipschitz property of this ambient C_v^j. The uniform spectral bound gives $\|e^{-\widehat{H}(v)s}\| \le e^{-\widehat{c}s}$, and hence from Eq. (106),

$$\sup_{v \in \mathcal{M}} \|C_v^{(j)}\|_{\mathrm{op}} \le \frac{1}{2\widehat{c}} \sup_{v \in \mathcal{M}} \|B_v^{(j)}\|_{\mathrm{op}} < \infty. \tag{107}$$

Moreover, Duhamel's formula states that $e^{-As} - e^{-\widetilde{A}s} = -\int_0^s e^{-A(s-u)}(A-\widetilde{A})e^{-\widetilde{A}u}\mathrm{d}u$. Applying this, $e^{-\widehat{H}(v)s} - e^{-\widehat{H}(\widetilde{v})s} = -\int_0^s e^{-\widehat{H}(v)(s-u)}\big(\widehat{H}(v) - \widehat{H}(\widetilde{v})\big)e^{-\widehat{H}(\widetilde{v})u}\,du$. Using this identity in Eq. (106) and the uniform exponential decay yields

$$\|C_v^{(j)} - C_{\widetilde{v}}^{(j)}\| \le C\big(\|B_v^{(j)} - B_{\widetilde{v}}^{(j)}\| + \|\widehat{H}(v) - \widehat{H}(\widetilde{v})\|\big). \tag{108}$$

Using the assumed local Lipschitz regularity and compactness of $\mathcal{M}$,

$$\|C_v^{(j)} - C_{\widetilde{v}}^{(j)}\| \le C\|v - \widetilde{v}\|. \tag{109}$$

Similarly, since the solution of the normal-space Lyapunov equation has the ambient representation $\Pi_v = \int_0^\infty e^{-\widehat{H}(v)s}\Sigma_N(v)e^{-\widehat{H}(v)s}\,ds$, following the same steps as above give us that Π_v is also locally lipschitz.

2. DISCRETE POISSON IDENTITY For each $j \in [D]$, define the quadratic corrector $\chi_j(v,q) := \langle q, C_v^{(j)}q\rangle$, for $q \in N_v\mathcal{M}$. For brevity, write $B_k := B_{v_k}^{(j)}$, $C_k := C_{v_k}^{(j)}$, $\Pi_k := \Pi_{v_k}$, $\delta q_k := q_{k+1} - q_k$. Our goal is to obtain $\chi_j(v,q)$ such that, for the frozen OU generator from Eq. (64),

$$\mathscr{L}_v\chi_j(v,q) = -\left(\langle q, B_v^{(j)}q\rangle - \mathrm{Tr}\left(B_v^{(j)}\Pi_v\right)\right). \tag{110}$$

Indeed, with the quadratic ansatz, $\nabla_q\chi_j(v,q) = 2C_v^{(j)}q$, $\nabla_q^2\chi_j(v,q) = 2C_v^{(j)}$, so $\mathscr{L}_v\chi_j(v,q) = -\langle q, \big(H_N(v)C_v^{(j)} + C_v^{(j)}H_N(v)\big)q\rangle + \mathrm{Tr}\left(C_v^{(j)}\Sigma_N(v)\right)$. Therefore, comparing with Eq. (110), we have the Lyapunov equations $H_N(v)C_v^{(j)} + C_v^{(j)}H_N(v) = B_v^{(j)}$. This, together with the other equation $H_N(v)\Pi_v + \Pi_v H_N(v) = \Sigma_N(v)$ and using the cyclicity of the trace implies the identity

$$\mathrm{Tr}\left(C_v\Sigma_N(v)\right) = \mathrm{Tr}(B_v^{(j)}\Pi_v). \tag{111}$$

Thus the centered quadratic observable in Eq. (103) is the negative frozen generator of χ_j. We next turn this continuous identity into a discrete telescoping identity.

Adding and subtracting $\langle q_{k+1}, C_k q_{k+1}\rangle$, we obtain

$$\chi_j(v_{k+1}, q_{k+1}) - \chi_j(v_k, q_k) = 2\langle C_k q_k, \delta q_k\rangle + \langle \delta q_k, C_k\delta q_k\rangle + \Theta_{k+1}^{(j)}, \tag{112}$$

where $\Theta_{k+1}^{(j)} := \left\langle q_{k+1}, \left(C_{v_{k+1}}^{(j)} - C_{v_k}^{(j)}\right)q_{k+1}\right\rangle$. The first two terms describe the change in the quadratic corrector with v_k frozen ($\chi_j(v_k, q_{k+1}) - \chi_j(v_k, q_k)$), whereas $\Theta_{k+1}^{(j)}$ records the change of its coefficient as the base point moves.

Since $q_k, v_k, C_k \in \mathcal{F}_k$, and using Eq. (49), Eq. (51), Eq. (105), and Eq. (111), we obtain

$$\begin{aligned}\mathbb{E}\left[\chi_j(v_{k+1}, q_{k+1}) - \chi_j(v_k, q_k) \mid \mathcal{F}_k\right] &= 2\langle C_k q_k, \mathbb{E}[\delta q_k \mid \mathcal{F}_k]\rangle + \mathbb{E}\left[\langle \delta q_k, C_k\delta q_k\rangle \mid \mathcal{F}_k\right] + \mathbb{E}[\Theta_{k+1}^{(j)} \mid \mathcal{F}_k] \\ &= -\eta\left(\langle q_k, B_k q_k\rangle - \mathrm{Tr}(B_k\Pi_k)\right) + R_k^{(j)},\end{aligned} \tag{113}$$

where

$$R_k^{(j)} = 2\langle C_k q_k, r_k^{(1)}\rangle + \mathrm{Tr}(C_k r_k^{(2)}) + \mathbb{E}[\Theta_{k+1}^{(j)} \mid \mathcal{F}_k]. \tag{114}$$

3. SUMMABILITY OF THE APPROXIMATION REMAINDER The use of the deterministic function using χ_j is to approximate the centered quadratic observable. For this, we now show that the cumulative contribution of $R_k^{(j)}$ vanishes after multiplication by η. From Eq. (50), Eq. (107), and Eq. (45), since $|\langle C_k q_k, r_k^{(1)}\rangle| \le \|C_k\|\|q_k\|\|r_k^{(1)}\|$, there exist $C, C_T > 0$ such that

$$\begin{aligned}&\eta\sum_{k<N_\eta(T)}\mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}}\left|\langle C_k q_k, r_k^{(1)}\rangle\right|\right] \\ &\quad\le C\eta\sum_{k<N_\eta(T)}\eta^{3/2}\mathbb{E}\left[\|q_k\| + \|q_k\|^3\right] \le C_T\eta^{1/2} \longrightarrow 0.\end{aligned} \tag{115}$$

Similarly with Eq. (52), there exist $C_T, C > 0$ such that

$$\eta \sum_{k<N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}} \left|\mathrm{Tr}(C_k r_k^{(2)})\right|\right] \le C_T \eta^{1/2} + C\eta^2 \sum_{k<N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}} \omega_g(\sqrt{\eta}\,\|q_k\|)\right]. \tag{116}$$

The last term converges to zero. To see this, first note that the uniform fourth-moment bound on g makes ω_g bounded on $[0, \zeta_1]$; denote this bound by $\overline{\omega}$. For every $a > 0$,

$$\mathbb{E}\left[\omega_g(\sqrt{\eta}\,\|q_k\|)\right] \le \sup_{0\le r\le a} \omega_g(r) + \overline{\omega}\,\mathbb{P}(\sqrt{\eta}\,\|q_k\| > a). \tag{117}$$

By Markov's inequality and Eq. (45), $\sup_{k\le N_\eta(T)} \mathbb{P}(\sqrt{\eta}\,\|q_k\| > a) \le C_{p,T} a^{-2p}\eta^p \longrightarrow 0$. Taking first $\eta \downarrow 0$ and then $a \downarrow 0$ in Eq. (117) gives $\sup_{k\le N_\eta(T)} \mathbb{E}\left[\omega_g(\sqrt{\eta}\,\|q_k\|)\right] \longrightarrow 0$. Since $\eta^2 N_\eta(T) \le T$, the last term in Eq. (116) vanishes.

For the coefficient-variation term, the local lipschitzness of $C_v^{(j)}$ Eq. (109) gives

$$|\Theta_{k+1}^{(j)}| \le C\|v_{k+1} - v_k\|\|q_{k+1}\|^2. \tag{118}$$

Fix $R \ge 1$. On $\{\|q_k\| \le R\}$, Eq. (56) implies

$$\mathbb{E}\left[|\Theta_{k+1}^{(j)}|\mathbf{1}_{\{\|q_k\|\le R\}} \mid \mathcal{F}_k\right] \le C\eta\vartheta_R(\eta). \tag{119}$$

To control $\{\|q_k\| > R\}$, we first record a conditional bound valid for every $k < \tau_\eta$. From Eq. (102) and the uniform fourth moment of g_{k+1}^η, $\mathbb{E}\left[\|v_{k+1} - v_k\|^2 \mid \mathcal{F}_k\right] \le C\eta^2$. Moreover, $q_{k+1} = q_k + \delta q_k$ and Eq. (54) imply $\mathbb{E}\left[\|q_{k+1}\|^4 \mid \mathcal{F}_k\right] \le C(1 + \|q_k\|^4)$. Conditional Cauchy-Schwarz therefore yields

$$\begin{aligned}\mathbb{E}\left[\|v_{k+1} - v_k\|\,\|q_{k+1}\|^2 \mid \mathcal{F}_k\right] &\le \mathbb{E}\left[\|v_{k+1} - v_k\|^2 \mid \mathcal{F}_k\right]^{1/2} \mathbb{E}\left[\|q_{k+1}\|^4 \mid \mathcal{F}_k\right]^{1/2} \\ &\le C\eta(1 + \|q_k\|^2).\end{aligned}$$

Because $\{\|q_k\| > R\} \in \mathcal{F}_k$, it follows from equation 118 that $\mathbb{E}\left[|\Theta_{k+1}^{(j)}|\mathbf{1}_{\{\|q_k\|>R\}}\right] \le C\eta\,\mathbb{E}\left[(1 + \|q_k\|^2)\mathbf{1}_{\{\|q_k\|>R\}}\right]$. For $R \ge 1$, $(1 + \|q_k\|^2)\mathbf{1}_{\{\|q_k\|>R\}} \le 2R^{-(2p-2)}\|q_k\|^{2p}$. Hence Eq. (45) gives

$$\mathbb{E}\left[|\Theta_{k+1}^{(j)}|\mathbf{1}_{\{\|q_k\|>R\}}\right] \le C_T\eta R^{-(2p-2)}. \tag{120}$$

Combining Eq. (119) and Eq. (120), $\limsup_{\eta\downarrow 0} \eta \sum_{k<N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}}|\Theta_{k+1}^{(j)}|\right] \le CTR^{-(2p-2)}$, because $\vartheta_R(\eta) \to 0$ for every fixed R. Letting $R \to \infty$ gives

$$\eta \sum_{k<N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}}|\Theta_{k+1}^{(j)}|\right] \longrightarrow 0. \tag{121}$$

Consequently, combining Eq. (115), Eq. (116), and Eq. (121), we obtain

$$\eta \sum_{k<N_\eta(T)} \mathbb{E}\left[\mathbf{1}_{\{k<\tau_\eta\}}|\mathcal{R}_k^{(j)}|\right] \longrightarrow 0. \tag{122}$$

In particular, Markov's inequality gives

$$\sup_{n\le N_\eta(T)} \left|\eta \sum_{k<n} \mathbf{1}_{\{k<\tau_\eta\}}\mathcal{R}_k^{(j)}\right| \xrightarrow{\mathbb{P}} 0, \tag{123}$$

because the supremum is bounded by $\eta \sum_{k<N_\eta(T)} \mathbf{1}_{\{k<\tau_\eta\}}|\mathcal{R}_k^{(j)}|$.

4. MARTINGALE TERM For $k < \tau_\eta$, define the martingale difference

$$M_{k+1}^{(j)} := \chi_j(v_{k+1}, q_{k+1}) - \chi_j(v_k, q_k) - \mathbb{E}\left[\chi_j(v_{k+1}, q_{k+1}) - \chi_j(v_k, q_k) \mid \mathcal{F}_k\right]. \tag{124}$$

For all $k \geq 0$, set $\widehat{M}_{k+1}^{(j)} := \mathbf{1}_{\{k<\tau_\eta\}} M_{k+1}^{(j)}$. Here $M_{k+1}^{(j)}$ may be assigned any value on $\{k \geq \tau_\eta\}$, since the indicator sets the stopped increment to zero. Because $\{k < \tau_\eta\} \in \mathcal{F}_k$, the sequence $(\widehat{M}_{k+1}^{(j)})_{k\geq 0}$ is again a martingale-difference sequence. This explicit stopping is needed when Doob's inequality is applied below. We claim that

$$\mathbb{E}\left[|M_{k+1}^{(j)}|^2 \mid \mathcal{F}_k\right] \leq C\eta(1 + \|q_k\|^4). \tag{125}$$

Indeed, conditional variance is bounded by the conditional second moment of the exact increment in Eq. (112). For its linear term, Eq. (51) and the crude bound $\mathbb{E}[\|\delta q_k\|^2 \mid \mathcal{F}_k] \leq C\eta$ give

$$\mathbb{E}\left[|2\langle C_k q_k, \delta q_k\rangle|^2 \mid \mathcal{F}_k\right] \leq C\eta\|q_k\|^2. \tag{126}$$

For the quadratic term, Eq. (54) gives

$$\mathbb{E}\left[|\langle \delta q_k, C_k \delta q_k\rangle|^2 \mid \mathcal{F}_k\right] \leq C\eta^2. \tag{127}$$

It remains to control $\Theta_{k+1}^{(j)}$. From Eq. (118) and conditional Cauchy-Schwarz,

$$\mathbb{E}\left[|\Theta_{k+1}^{(j)}|^2 \mid \mathcal{F}_k\right] \leq C\mathbb{E}\left[\|v_{k+1} - v_k\|^4 \mid \mathcal{F}_k\right]^{1/2} \mathbb{E}\left[\|q_{k+1}\|^8 \mid \mathcal{F}_k\right]^{1/2}. \tag{128}$$

The update expansion Eq. (102), the bound $\|g_{k+1}^\eta\| \leq \|g_{k+1}\|$, and the uniform eighth moment of the sample gradient give $\mathbb{E}\left[\|v_{k+1} - v_k\|^4 \mid \mathcal{F}_k\right] \leq C\eta^4$. Similarly, $q_{k+1} = q_k + \delta q_k$ and the conditional eighth-moment bound $\mathbb{E}[\|\delta q_k\|^8 \mid \mathcal{F}_k] \leq C\eta^4$, which follows from the estimate $\|\delta q_k\| \leq C\sqrt{\eta}\|G_{k+1}^\eta\|$ and the uniform $2p$-moment bound, imply $\mathbb{E}\left[\|q_{k+1}\|^8 \mid \mathcal{F}_k\right] \leq C(1 + \|q_k\|^8)$. Substitution into Eq. (128) yields

$$\mathbb{E}\left[|\Theta_{k+1}^{(j)}|^2 \mid \mathcal{F}_k\right] \leq C\eta^2(1 + \|q_k\|^4). \tag{129}$$

This is where the assumption $p > 4$ is used. Combining Eq. (126), Eq. (127), and Eq. (129) proves Eq. (125).

5.TELESCOPING AND CONCLUSION By Eq. (113) and the definition of $M_{k+1}^{(j)}$, for every $k < \tau_\eta$,

$$\begin{aligned}&\chi_j(v_{k+1}, q_{k+1}) - \chi_j(v_k, q_k)\\ &\quad = -\eta\left(\langle q_k, B_k q_k\rangle - \mathrm{Tr}(B_k\Pi_k)\right) + R_k^{(j)} + M_{k+1}^{(j)}.\end{aligned}$$

For each deterministic $t \leq N_\eta(T)$, summing this identity up to $t \wedge \tau_\eta$ and multiplying by η gives

$$\begin{aligned}&\eta^2 \sum_{k<t\wedge\tau_\eta} \left(\langle q_k, B_k q_k\rangle - \mathrm{Tr}(B_k\Pi_k)\right)\\ &= \eta\left(\chi_j(v_0, q_0) - \chi_j(Y_{t\wedge\tau_\eta}, q_{t\wedge\tau_\eta})\right) + \eta\sum_{k<t} \widehat{M}_{k+1}^{(j)} + \eta\sum_{k<t} \mathbf{1}_{\{k<\tau_\eta\}} R_k^{(j)}.\end{aligned}$$

The boundary term vanishes uniformly in t. It is enough to verify this on the event $\{\tau_\eta > N_\eta(T)\}$, whose complement has probability tending to zero by the confinement estimate (step 1). On this event, $t \wedge \tau_\eta = t$ for every $t \leq N_\eta(T)$, and $|\chi_j(Y, q)| \leq C\|q\|^2$ gives

$$\sup_{t\leq N_\eta(T)} \eta\,|\chi_j(v_0, q_0) - \chi_j(v_n, q_n)| \leq C\eta \max_{k\leq N_\eta(T)} \|q_k\|^2.$$

Consequently, for every $\varepsilon > 0$, a union bound and the killed moment estimate Eq. (45) give

$$\begin{aligned}\mathbb{P}\left(\{\tau_\eta > N_\eta(T)\} \cap \left\{C\eta \max_{k\leq N_\eta(T)} \|q_k\|^2 > \varepsilon\right\}\right) &\leq \sum_{k\leq N_\eta(T)} \mathbb{P}\left(\tau_\eta > k, \|q_k\|^{2p} > \left(\frac{\varepsilon}{C\eta}\right)^p\right)\\ &\leq C_{\varepsilon,T}\eta^{p-2} \longrightarrow 0.\end{aligned}$$

For the martingale term, Doob's inequality, Eq. (125), and Eq. (45) yield

$$\mathbb{E}\left[\sup_{n\le N_\eta(T)}\left|\eta\sum_{k<n}\widehat{M}_{k+1}^{(j)}\right|^2\right]\le 4\eta^2\sum_{k<N_\eta(T)}\mathbb{E}|\widehat{M}_{k+1}^{(j)}|^2\le C_T\eta\longrightarrow 0.$$

The uniform remainder term vanishes by Eq. (123). Therefore,

$$\sup_{0\le t\le T}\left|\eta^2\sum_{k<N_\eta(t)\wedge\tau_\eta}\left(\langle q_k,B_{v_k}^{(j)}q_k\rangle-\mathrm{Tr}\left(B_{v_k}^{(j)}\Pi_{v_k}\right)\right)\right|\xrightarrow{\mathbb{P}}0. \tag{130}$$

We now reassemble the coordinate wise estimates. For the stopped truncated chain, define $S_\eta^{\tau,\mathrm{tr}}(t):=\eta^2\sum_{k<N_\eta(t)\wedge\tau_\eta}\left(\mathcal{B}_{v_k}[q_k,q_k]-\overline{\mathcal{B}}(v_k)\right)$. By the definition of $B_v^{(j)}$ and Eq. (104), its jth ambient coordinate is

$$\left\langle e_j,S_\eta^{\tau,\mathrm{tr}}(t)\right\rangle=\eta^2\sum_{k<N_\eta(t)\wedge\tau_\eta}\left(\langle q_k,B_{v_k}^{(j)}q_k\rangle-\mathrm{Tr}\left(B_{v_k}^{(j)}\Pi_{v_k}\right)\right).$$

Eq. (130) therefore implies, for each $j=1,\dots,D$, that $\sup_{0\le t\le T}\left|\left\langle e_j,S_\eta^{\tau,\mathrm{tr}}(t)\right\rangle\right|\xrightarrow{\mathbb{P}}0$. Since D is fixed and finite, for every $\varepsilon>0$ the union bound gives

$$\mathbb{P}\left(\sup_{0\le t\le T}\left\|S_\eta^{\tau,\mathrm{tr}}(t)\right\|>\varepsilon\right)\le\sum_{j=1}^{D}\mathbb{P}\left(\sup_{0\le t\le T}\left|\left\langle e_j,S_\eta^{\tau,\mathrm{tr}}(t)\right\rangle\right|>\frac{\varepsilon}{\sqrt{D}}\right)\longrightarrow 0.$$

This proves Eq. (103) for the stopped truncated chain. □

Notations Let $\nabla^\perp$ denote the normal connection, and use the same notation for its natural extension to operator fields on the normal bundle. In particular, $(\nabla_\tau^\perp H_N)n:=\nabla_\tau^\perp(H_Nn)-H_N(\nabla_\tau^\perp n)$.

Lemma C.14 (Projected-gradient expansion). *Let $\mathcal{U}_{\zeta_0}$ denote the fixed tubular neighborhood of $\mathcal{M}$ on which the nearest-point projection π is uniquely defined and C^3. Suppose that the tangential regularity assumptions* P2, P3 *hold on $\mathcal{U}_{\zeta_0}$. For $v\in\mathcal{M}$ and $\tau\in T_v\mathcal{M}$, let $\nabla_\tau^\perp H_N$ denote the covariant derivative of H_N in the direction τ. Define $\mathcal{B}_v:N_v\mathcal{M}\times N_v\mathcal{M}\to T_v\mathcal{M}$ by*

$$\langle\mathcal{B}_v[n_1,n_2],\tau\rangle:=\left\langle n_1,(\nabla_\tau^\perp H_N)(v)n_2\right\rangle \tag{131}$$

for every $n_1,n_2\in N_v\mathcal{M}$ and $\tau\in T_v\mathcal{M}$.

Then $\mathcal{B}_v$ is a well-defined symmetric bilinear map that depends continuously on v. Moreover, there exist $\zeta\in(0,\zeta_0]$ and $C>0$ such that, for every $v\in\mathcal{M}$ and every $n\in N_v\mathcal{M}$ with $\|n\|\le\zeta$,

$$\left\|D\pi(v+n)\nabla L(v+n)-\frac{1}{2}\mathcal{B}_v[n,n]\right\|\le C\big(\omega_R(\|n\|)+\|n\|\big)\|n\|^2. \tag{132}$$

Equivalently, for every $\tau\in T_v\mathcal{M}$,

$$\left|\langle D\pi(v+n)\nabla L(v+n),\tau\rangle-\frac{1}{2}\left\langle n,(\nabla_\tau^\perp H_N)(v)n\right\rangle\right|\le C\big(\omega_R(\|n\|)+\|n\|\big)\|\tau\|\|n\|^2. \tag{133}$$

In particular, $D\pi(v+n)\nabla L(v+n)=\frac{1}{2}\mathcal{B}_v[n,n]+o(\|n\|^2)$ uniformly over $v\in\mathcal{M}$ and $n\in N_v\mathcal{M}$ as $\|n\|\to 0$.

(Proof of Lemma C.14). We divide the proof into four steps.

1. Well-defined $\mathcal{B}_v$. Because the normal space varies along $\mathcal{M}$, we compare normal vectors by differentiating them in the ambient parameter space and projecting the result back onto the normal space. For $\tau\in T_v\mathcal{M}$ and a smooth normal vector field n, define $\nabla_\tau^\perp n:=P_N(v)D_\tau n$. A normal field is parallel along a curve when this derivative vanishes.

Due to (P2), the normal Hessian is a smoothly varying family of self-adjoint operators $H_N(v):N_v\mathcal{M}\to N_v\mathcal{M}$. Its derivative along τ is defined by

$$(\nabla_\tau^\perp H_N)n:=\nabla_\tau^\perp(H_Nn)-H_N(\nabla_\tau^\perp n). \tag{134}$$

This derivative is also self-adjoint. To see this, let $v(s) \subset \mathcal{M}$ satisfy $v(0) = v$ and $\dot{v}(0) = \tau$, and parallel-transport $n_1, n_2 \in N_v\mathcal{M}$ along $v(s)$. Since $H_N(v(s))$ is self-adjoint,

$$\langle n_1(s), H_N(v(s))n_2(s)\rangle = \langle H_N(v(s))n_1(s), n_2(s)\rangle.$$

Differentiating at $s = 0$ and using parallelity gives $\langle n_1, (\nabla^\perp_\tau H_N)n_2\rangle = \langle (\nabla^\perp_\tau H_N)n_1, n_2\rangle$.

Due to the assumption on the normal Hessian $H_N(v) := P_N(v)D^2L(v)P_N(v)\big|_{N_v\mathcal{M}}$ varying C^1-smoothly along $\mathcal{M}$, its normal covariant derivative $\nabla^\perp_\tau H_N$ is well defined, depends continuously on (v, τ), and is linear in the tangent direction τ. Consequently, for fixed $n_1, n_2 \in N_v\mathcal{M}$, the map $\tau \mapsto \langle n_1, (\nabla^\perp_\tau H_N)n_2\rangle$ is a continuous linear functional on $T_v\mathcal{M}$. The Riesz representation theorem therefore defines a unique vector $\mathcal{B}_v[n_1, n_2] \in T_v\mathcal{M}$ such that

$$\langle \mathcal{B}_v[n_1, n_2], \tau\rangle = \langle n_1, (\nabla^\perp_\tau H_N)n_2\rangle .$$

The preceding identity shows that $\mathcal{B}_v$ is symmetric in (n_1, n_2), while bilinearity follows directly from its definition. Finally, continuity of $\nabla^\perp H_N$ and compactness of $\mathcal{M}$ imply $C_\mathcal{B} := \sup_{\substack{v\in\mathcal{M} \\ \tau\in T_v\mathcal{M},\ \|\tau\|=1}} \|\nabla^\perp_\tau H_N\| < \infty$. Taking the supremum over unit tangent vectors in the defining identity gives

$$\|\mathcal{B}_v[n_1, n_2]\| \le C_\mathcal{B}\|n_1\|\|n_2\|. \tag{135}$$

2. Differential of the nearest-point projection. Consider the normal-coordinate map

$$\Psi : N^\zeta\mathcal{M} \to \mathcal{U}_\zeta, \qquad \Psi(v, n) = v + n,$$

where $N^\zeta\mathcal{M} := \{(v, n) : v \in \mathcal{M},\ n \in N_v\mathcal{M},\ \|n\| < \zeta\}$. Ψ is a diffeomorphism onto $\mathcal{U}_\zeta$, and $\pi(\Psi(v, n)) = v$. We first describe how the normal space changes as its base point moves along $\mathcal{M}$. Note that the second fundamental form is the normal-valued bilinear map

$$\mathrm{II}_v(\tau_1, \tau_2) := P_N(v)D_{\tau_1}\widetilde{\tau}_2, \qquad \tau_1, \tau_2 \in T_v\mathcal{M},$$

where $\widetilde{\tau}_2$ is any smooth tangent-vector field extending τ_2 ($\tilde{\tau}_2(v) = \tau_2$). For $n \in N_v\mathcal{M}$, define the corresponding shape operator $A_n : T_v\mathcal{M} \to T_v\mathcal{M}$ by $\langle A_n\tau_1, \tau_2\rangle = \langle \mathrm{II}_v(\tau_1, \tau_2), n\rangle$. Let $v(s) \subset \mathcal{M}$ satisfy $v(0) = v$ and $\dot{v}(0) = \tau$, and let $n(s) \in n_{v(s)}\mathcal{M}$ be parallel in the normal bundle, so that $n(0) = n$ and $P_N(v(s))\dot{n}(s) = 0$. The Weingarten equation gives $\dot{n}(0) = -A_n\tau$. To see this, fix an arbitrary $\tau \in T_v\mathcal{M}$, and let $\tilde{\tau}_2$ be a smooth tangent vector field (see above) note that $\langle n(s), \tilde{\tau}(v(s))\rangle = 0$ for every s. Differentiating at $s = 0$ gives $\langle \dot{n}(0), \tau_2\rangle + \langle n, D_\tau\tilde{\tau}_2\rangle = 0$. Because n is normal, only the normal component of $D_\tau\tilde{\tau}_2$ contributes

$$\langle n, D_\tau\tilde{\tau}_2\rangle = \langle n, P_N(v)D_\tau\tilde{\tau}_2\rangle = \langle n, \mathrm{II}(\tau, \tau_2)\rangle = \langle A_n\tau, \tau_2\rangle.$$

Therefore, $\langle \dot{n}(0) + A_n\tau, \tau_2\rangle = 0$. This applies for every $\tau_2 \in T_v\mathcal{M}$, and parallel normal transport gives $P_N(v)\dot{n}(0) = 0$, meaning that $\dot{n}(0)$ is tangent. Since $A_n\tau$ is also tangent, $\dot{n}(0) + A_n\tau \in T_v\mathcal{M}$ and therefore $\dot{n}(0) = -A_n\tau$. Consequently,

$$\left.\frac{d}{ds}\right|_{s=0} \Psi(v(s), n(s)) = \tau + \dot{n}(0) = (I - A_n)\tau. \tag{136}$$

A vertical perturbation changes the normal displacement while keeping the base point fixed. Thus, for $m \in N_v\mathcal{M}$, $D\Psi_{(v,n)}[0, m] := \frac{d}{ds}\big|_{s=0}\Psi(v, n + sm) = m$. Using the horizontal-vertical decomposition of the tangent space of the normal bundle, we therefore obtain

$$D\Psi_{(v,n)}[\tau, m] = (I - A_n)\tau + m.$$

Since $\pi \circ \Psi(v, n) = v$, differentiating this identity gives $D\pi(v + n)\big[(I - A_n)\tau + m\big] = \tau$. Every ambient vector z can be written uniquely as $z = (I - A_n)\tau + m$, where $\tau \in T_n\mathcal{M}$, $m \in N_v\mathcal{M}$. Taking its tangential projection yields $P_T(v)z = (I - A_n)\tau$. It follows that

$$D\pi(v + n)z = (I - A_n)^{-1}P_T(v)z. \tag{137}$$

Finally, because $\mathcal{M}$ is a compact smooth submanifold, its second fundamental form is uniformly bounded. Since A_n depends linearly on n, there exists $C_A < \infty$ such that $\|A_n\| \le C_A\|n\|$. After decreasing ζ, we may assume that $\|A_n\| \le \frac{1}{2}$, whenever $\|n\| < \zeta$. Therefore, $I - A_n$ is invertible and

$$\|(I - A_n)^{-1}\| \le 2. \tag{138}$$

3. Tangential derivative in normal coordinates. Define the pullback of the loss to the normal bundle by $\widetilde{L}(v,n) := L(v+n)$. For $\tau \in T_v\mathcal{M}$, define its horizontal derivative by moving v in direction τ and transporting n parallelly: $D_v^{\perp}\widetilde{L}(v,n)[\tau] := \frac{d}{ds}\big|_{s=0} L(v(s)+n(s))$. Using Eq. (136),

$$\begin{aligned} D_v^{\perp}\widetilde{L}(v,n)[\tau] &= \langle \nabla L(v+n), (I-A_n)\tau \rangle \\ &= \langle (I-A_n)P_T(v)\nabla L(v+n), \tau \rangle , \end{aligned} \tag{139}$$

where we have used the self-adjointness of the shape operator A_n. Let $G_T(v,n) \in T_v\mathcal{M}$ be the Riesz representative of this horizontal derivative $\langle G_T(v,n), \tau\rangle = D_v^{\perp}\widetilde{L}(v,n)[\tau]$. Eq. (139) gives the exact identity

$$G_T(v,n) = (I-A_n)P_T(v)\nabla L(v+n). \tag{140}$$

By (P3), we have $\widetilde{L}(v,n) = L(v) + \frac{1}{2}\langle n, H_N(v)n\rangle + R(v,n)$. Every point of $\mathcal{M}$ is a global minimizer, so $L(v)$ is constant along $\mathcal{M}$. Therefore,

$$D_v^{\perp}\widetilde{L}(v,n)[\tau] = \frac{1}{2}\left.\frac{d}{ds}\right|_{s=0} \langle n(s), H_N(v(s))n(s)\rangle + \left.\frac{d}{ds}\right|_{s=0} R(v(s),n(s)). \tag{141}$$

Because $n(s)$ is parallel in the normal bundle ($\nabla_s^{\perp} n(s) = 0$), applying the product rule to the normal-space inner product gives

$$\begin{aligned} \frac{d}{ds}\langle n(s), H_N(v(s))n(s)\rangle &= \left\langle \nabla_s^{\perp} n(s), H_N(v(s))n(s)\right\rangle + \left\langle n(s), (\nabla^{\perp}_{\dot v(s)} H_N)n(s)\right\rangle \\ &\quad + \left\langle n(s), H_N(v(s))\nabla_s^{\perp} n(s)\right\rangle \\ &= \left\langle n(s), (\nabla^{\perp}_{\dot v(s)} H_N)n(s)\right\rangle . \end{aligned}$$

Evaluating at $s=0$, where $\dot v(0) = \tau$, therefore yields

$$\left.\frac{d}{ds}\right|_{s=0} \langle n(s), H_N(v(s))n(s)\rangle = \left\langle n, (\nabla_{\tau}^{\perp} H_N)n\right\rangle = \langle \mathcal{B}_v[n,n], \tau\rangle . \tag{142}$$

The tangential remainder assumption yields $\left|\frac{d}{ds}\big|_{s=0} R(v(s),n(s))\right| \le \omega_R(\|n\|)\|\tau\|\|n\|^2$. It follows from Eq. (141),Eq. (142), and Riesz representation that

$$G_T(v,n) = \frac{1}{2}\mathcal{B}_v[n,n] + r_T(v,n), \qquad \|r_T(v,n)\| \le \omega_R(\|n\|)\|\tau\|\|n\|^2. \tag{143}$$

4. Conversion to the projected ambient gradient. Combining Eq. (137) and Eq. (140), we obtain

$$D\pi(v+n)\nabla L(v+n) = (I-A_n)^{-1}P_T(v)\nabla L(v+n) = (I-A_n)^{-2}G_T(v,n). \tag{144}$$

From the bound of the shape operator (see above) and by Eq. (138), and since $(I-A_n)^{-2} - I = (I-A_n)^{-2}(2A_n - A_n^2)$,

$$\left\|(I-A_n)^{-2} - I\right\| \le C\|n\|. \tag{145}$$

Substituting Eq. (143) into Eq. (144) gives

$$D\pi(v+n)\nabla L(v+n) - \frac{1}{2}\mathcal{B}_v[n,n] = (I-A_n)^{-2}r_T(v,n) + \frac{1}{2}\left((I-A_n)^{-2} - I\right)\mathcal{B}_v[n,n].$$

Using Eq. (135), Eq. (138), Eq. (145), and Eq. (143), we conclude that

$$\left\| D\pi(v+n)\nabla L(v+n) - \frac{1}{2}\mathcal{B}_v[n,n]\right\| \le C\omega_R(\|n\|)\|n\|^2.$$

□

D Application to $\mathcal{M}_{\mathrm{ReLU}}$

In this section, we verify the assumptions of Section C for the ReLU autoencoder under the $2r$-GMM, including the NASM property of its global-minimizer manifold in Section D.4. For general $r \ge 2$, Section D.2 proves Lemma 4.2, thus proving that the effective drift vanishes for the linear autoencoder, whereas only the pairing-breaking blocks contribute to the ReLU drift. We then specialize to $r = 2$ and derive the angular ODE in Lemma 4.3 (proved in Section D.5.3). Together with Lemma 4.2, this yields Theorem 3.1, whose proof is completed in Section D.3.

D.1 Notations and Setting

For a differentiable map f, we denote its derivative at W in the direction N by $D_W f(W)[N] := \frac{d}{d\epsilon} f(W+\epsilon N)\big|_{\epsilon=0}$. Similarly, the second derivative is the bilinear map $D^2_W f(W)[N,M] := \frac{\partial}{\partial\delta\partial\epsilon} f(W+\epsilon N+\delta M)\big|_{\epsilon=\delta=0}$. For $z = Q^\top x$, we denote $y := R^\top z$. $D_y := \mathrm{diag}(\mathrm{sgn}(y))$.

We denote $J[N] := D_W f(x;W)[N]$. For the normal blocks $\{S,K,B,C\}$ which correspond to the normal splitting $N_W\mathcal{M} = \bigoplus_{\alpha\in\{S,A,B,C\}} N_\alpha$, $W \in \mathcal{M}_\triangle$, we define $J_\alpha[\alpha] := D_W f(x;W)[\delta N_\alpha]$, where $\delta N_\alpha \in N_\alpha$. We denote $r(x;W) = x - f(x;W)$, thus the population loss $L(W) = \frac{1}{2}\mathbb{E}[\|r(x;W)\|^2] + \frac{1}{2}\|W\|_F^2$. For $\delta N, \delta M \in N_W\mathcal{M}$, the quadratic form of the Hessian $H_W := \nabla^2 L(W)$ is

$$H_W[\delta N,\delta M] = \mathbb{E}\langle J[N], J[M]\rangle - \mathbb{E}\langle r(x;W), D^2_W f(x;W)[\delta N,\delta M] + \lambda\langle \delta N,\delta M\rangle_F. \quad (146)$$

Vectorization. Let $h : \mathbb{R}^{d\times 2r} \to \mathbb{R}^{2rd}$ denote row-major vectorization. Equivalently, $h(A) = \mathrm{vec}(A^\top)$, where vec stacks matrix columns. For $W = [w_1,\dots,w_{2r}]$ and $v = h(W)$, define the sample gradient of neuron i by $g_i := \nabla_{w_i}\ell(v;x) \in \mathbb{R}^d$. Then $g(v;x) := \nabla_v \ell(v;x) = h([g_1,\dots,g_{2r}])$.

Setting We study the data consists of an r-dimensional symmetric Gaussian mixture embedded in $\mathbb{R}^d$, together with isotropic noise in the orthogonal complement. Namely, for gaussian mixture components $Q_r z$, where $Q_r \in \mathbb{R}^{d\times r}$ and $Q_r^\top Q_r = I_r$, we have the noisy data defined as $x = Q_r z + Q_\perp \varepsilon$ and $Q_\perp$ spans the orthogonal complement of $\mathrm{span}(Q_r)$. The signal variable $z \in \mathbb{R}^r$ is distributed as

$$z \sim \frac{1}{2r}\sum_{i=1}^{r}\left[\mathcal{N}(se_i,\sigma^2 I_r) + \mathcal{N}(-se_i,\sigma^2 I_r)\right],$$

where we assume that $e_1,\dots,e_r$ are the coordinate axes in the signal subspace without loss of generality. The orthogonal noise satisfies $\varepsilon \sim \mathcal{N}(0,\lambda_n I_{d-r})$, for $\varepsilon \perp z$.[12] We denote $\Sigma_x := \mathbb{E}[xx^\top]$, where $\mathbb{E}[xx^\top] = (\sigma^2 + \frac{s^2}{r})Q_r Q_r^\top + \lambda_n Q_\perp Q_\perp^2$. We denote $\lambda_s := \sigma^2 + \frac{s^2}{r}$ and $\lambda_n = \lambda_n$.

For the tied-weight AEs, we consider the width of $2r$ for the ReLU case and r for the linear AE. From Proposition B.6, the corresponding global-minimizer manifolds are

$$\mathcal{M}_{\mathrm{ReLU}} = \{\rho[Q_r R, -Q_r R] \mid R \in \mathrm{O}(r)\}, \quad \mathcal{M}_{\mathrm{lin}} = \{\rho Q_r R \mid R \in \mathrm{O}(r)\}.$$

D.2 Proof of Lemma 4.2

We first verify the conditions of Theorem C.6. Lemma 3.2 and its proof in Section D.4 establish that $\mathcal{M}_{\mathrm{ReLU}}$ and $\mathcal{M}_{\mathrm{lin}}$ are compact NASMs. Lemmas D.4 and D.5 verify Assumptions C.4 and C.5 for the ReLU AE.

By Lemma D.6 and D.12, the normal Hessian and sample-gradient covariance are block diagonal. Lemma D.11 then gives the block decomposition of Π_N. Throughout, $H_\alpha := P_\alpha H_N P_\alpha$ and $\Pi_\alpha := P_\alpha \Pi_N P_\alpha$ are operators on the full normal space (extended by zero outside their respective blocks).

For $\tau \in T_v\mathcal{M}$, the effective drift satisfies $\langle b_{\mathrm{ReLU}}(v),\tau\rangle = \mathrm{Tr}\left[(\nabla^\perp_\tau H_N)(v)\Pi_N(v)\right]$. Due to linearity $\nabla^\perp_\tau H_N = \sum_\alpha \nabla^\perp_\tau H_\alpha$. If $\alpha \neq \beta$, then $H_\alpha P_\beta = P_\beta H_\alpha = 0$. Differentiating $P_\beta H_\alpha P_\beta = 0$ by the product rule therefore gives $P_\beta(\nabla^\perp_\tau H_\alpha)P_\beta = 0$. Since $\Pi_\beta = P_\beta\Pi_\beta P_\beta$, the cross-block trace terms vanish. Hence $\langle b_{\mathrm{ReLU}}(v),\tau\rangle = \sum_{\alpha,\beta}\mathrm{Tr}\left[(\nabla^\perp_\tau H_\alpha)(v)\Pi_\beta(v)\right] = \sum_\alpha \mathrm{Tr}\left[(\nabla^\perp_\tau H_\alpha(v)\Pi_\alpha(v)\right]$. Defining $\langle b_\alpha(v),\tau\rangle := \mathrm{Tr}\left[(\nabla^\perp_\tau H_\alpha(v)\Pi_\alpha(v)\right]$ yields $b_{\mathrm{ReLU}} = b_S + b_K + b_B + b_C$.

It remains to show that $b_S = b_K = 0$. Lemma D.6 shows that, on their respective blocks, $H_S = h_S I$ and $H_K = h_K I$, where $h_S = \rho^2\lambda_s$, $h_K = \frac{2-\rho^2}{2}(\lambda_s - \lambda_n)$ (Note that $\|\delta N_\alpha\|^2 = 2\|\alpha\|^2$). These scalars are independent of v. As full normal-space operators, $H_\alpha = h_\alpha P_\alpha$ for $\alpha \in \{S,K\}$. Differentiating $P_\alpha^2 = P_\alpha$ gives $(\nabla^\perp_\tau P_\alpha)P_\alpha + P_\alpha(\nabla^\perp_\tau P_\alpha) = \nabla^\perp_\tau P_\alpha$. Multiplying by P_α for both sides yield $P_\alpha(\nabla^\perp_\tau P_\alpha)P_\alpha = 0$. Therefore, $P_\alpha(\nabla^\perp_\tau H_\alpha)P_\alpha = 0$, and, because Π_α is supported on

[12] Setting $r = 2$, $\sigma_\perp = \sigma$ recovers the data setting in the main text.

that block, $\langle b_\alpha(v), \tau\rangle = \mathrm{Tr}\left[(\nabla^\perp_\tau H_\alpha)(v)\Pi_\alpha(v)\right] = 0$, for $\alpha \in \{S, K\}$. Thus $b_S = b_K = 0$. For the linear autoencoder, only the S- and K-blocks are present, and their Hessian restrictions are likewise constant scalar multiples of the identity. The same argument therefore gives $b_{\mathrm{lin}} = 0$. □

Remark D.1 *For the linear autoencoder, the corresponding regularity conditions follow directly from the polynomial dependence of the sample loss on the parameters and the finite moments of all orders of the Gaussian mixture. We therefore omit the straightforward verification.*

D.3 Proof of Theorem 3.1

Now we set $r = 2$. Fix $\sigma \in (0, \sigma_0]$. For the ReLU autoencoder, write $\mathcal{M} = \mathcal{M}_{\mathrm{ReLU}}$. By Theorem C.6, for every fixed $T > 0$, $\pi(z^\eta_{\lfloor T/\eta^2\rfloor}) \xrightarrow{\mathbb{P}} v_T$, where v_t follows the effective dynamics with $v_0 = \Phi(z_0)$ and $\mathrm{dist}(z^\eta_{\lfloor T/\eta^2\rfloor}, \mathcal{M}) \xrightarrow{\mathbb{P}} 0$. Consequently, on the event where the projection is defined,

$$\|z^\eta_{\lfloor T/\eta^2\rfloor} - v_T\| \le \mathrm{dist}(z^\eta_{\lfloor T/\eta^2\rfloor}, \mathcal{M}) + \|\pi(z^\eta_{\lfloor T/\eta^2\rfloor}) - v_T\|.$$

Since this event has probability tending to one, $z^\eta_{\lfloor T/\eta^2\rfloor} \xrightarrow{\mathbb{P}} v_T$.

Let $\theta_0 \in [0, \pi/2]$ denote the orientation of $v_0 = \Phi(z_0)$. The initialization assumption gives $\mathbb{P}(\theta_0 = \pi/4) = 0$. Lemmas 4.2 and 4.3 identify the effective angular dynamics: $\theta_t \to 0$ when $\theta_0 \in (0, \pi/4)$ and $\theta_t \to \pi/2$ when $\theta_0 \in (\pi/4, \pi/2)$, while the endpoints 0 and $\pi/2$ are stationary. Therefore, $\mathrm{dist}(v_T, \mathcal{M}_*) \to 0$ almost surely as $T \to \infty$.

For any $\tau > 0$ and $\delta \in (0, 1)$, choose a finite $T > 0$ such that $\mathbb{P}(\mathrm{dist}(v_T, \mathcal{M}_*) < \tau/2) \ge 1 - \delta/2$. For this fixed T, convergence in probability gives $\eta_0 > 0$ such that, for every $\eta \in (0, \eta_0)$, $\mathbb{P}(\|z^\eta_{\lfloor T/\eta^2\rfloor} - v_T\| < \tau/2) \ge 1 - \delta/2$. Using $\mathrm{dist}(z^\eta_{\lfloor T/\eta^2\rfloor}, \mathcal{M}_*) \le \|z^\eta_{\lfloor T/\eta^2\rfloor} - v_T\| + \mathrm{dist}(v_T, \mathcal{M}_*)$, union bound yields $\mathbb{P}(\mathrm{dist}(z^\eta_{\lfloor T/\eta^2\rfloor}, \mathcal{M}_*) < \tau) \ge 1 - \delta$.

For the linear autoencoder, Lemma 4.2 gives $b_{\mathrm{lin}} \equiv 0$, so its effective trajectory is $v_{\mathrm{lin}}(t) = \Phi_{\mathrm{lin}}(z_{\mathrm{lin},0})$. Theorem C.6 and following the above argument gives, for every $0 < t_0 < T' < \infty$, $\sup_{t\in[t_0,T']} \left\| z^\eta_{\mathrm{lin},\lfloor t/\eta^2\rfloor} - \Phi_{\mathrm{lin}}(z_{\mathrm{lin},0})\right\| \xrightarrow{\mathbb{P}} 0$. □

D.4 $\mathcal{M}_{\mathrm{ReLU}}$ is a Normally Attracting Stationary Manifold

We prove that $\mathcal{M}_{\mathrm{ReLU}}$ and $\mathcal{M}_{\mathrm{lin}}$ are compact NASM (the case $r = 2$ gives Lemma 3.2). Since these manifolds consist of global minimizers, every point is stationary under the population gradient flow. Moreover, the C^2 regularity of the population loss on an open neighborhood of each manifold ensures that the gradient vector field is C^1. By Definition C.1, it remains to establish uniform contraction in the normal directions, dominating the stationary tangent dynamics. We prove this for $\mathcal{M}_{\mathrm{ReLU}}$ below. The linear case follows by the same argument applied to its S and K blocks. In the following, we denote $\mathcal{M} := \mathcal{M}_{\mathrm{ReLU}}$.

Lemma D.2 (Uniform attraction of $\mathcal{M}_{\mathrm{ReLU}}$). *Consider the data setting D.1 and C^2 loss L. Then there exists a uniform constant $c > 0$, independent of $W \in \mathcal{M}$, such that for every $W \in \mathcal{M}$ and every $\delta N \in N_W\mathcal{M}$,*

$$\left\langle \delta N, \nabla^2 L(W)\delta N\right\rangle \ge c\|\delta N\|^2.$$

Consequently, $\mathcal{M}_{\mathrm{ReLU}}$ is uniformly normally attracting.

(Proof of Lemma D.2). Our goal is to find a constant $c > 0$, independent of $W \in \mathcal{M}$, such that $H_N[\delta N, \delta N] \ge c\|\delta N\|_F^2$. By the orthogonal decomposition in Lemma D.7, $\delta N = \sum_{\alpha\in\{S,K,B,C\}} \delta N_\alpha$ and $\langle \delta N_\alpha, \delta N_\beta\rangle_F = 0$ for $\alpha \ne \beta$. Hence $\|\delta N\|_F^2 = \sum_{\alpha\in\{S,K,B,C\}} \|\delta N_\alpha\|_F^2$. The block representations $\delta N_S = [Q_r RS, -Q_r RS]$, $\delta N_K = [Q_\perp K, -Q_\perp K]$, $\delta N_B = [Q_r RB, Q_r RB]$, and $\delta N_C = [Q_\perp C, Q_\perp C]$, together with the orthonormality of $Q_r R$ and $Q_\perp$, give

$$\|\delta N\|_F^2 = 2\big(\|S\|_F^2 + \|K\|_F^2 + \|B\|_F^2 + \|C\|_F^2\big).$$

It therefore suffices to find a uniform constant $c_1 > 0$ such that $H_N[\delta N, \delta N] \ge c_1(\|S\|_F^2 + \|K\|_F^2 + \|B\|_F^2 + \|C\|_F^2)$, in which case we may take $c = c_1/2$. We first establish positive definiteness at each $W \in \mathcal{M}$, then obtain a uniform lower bound using continuity and compactness.

Pointwise positive definiteness. Fix $W \in \mathcal{M}$. By Lemma D.6 and the assumption $\lambda \in (\lambda_n, \lambda_s)$, there exists $c_2 > 0$ such that $H_\alpha[\delta N_\alpha, \delta N_\alpha] \geq c_2 \|\delta N_\alpha\|_F^2$ for $\alpha \in \{S, K, C\}$.

For the B-block, we prove positive definiteness by showing that $\mathbb{E}\|J_B[B]\|_2^2 = 0$ implies $B = 0$. Suppose $\mathbb{E}\|J_B[B]\|_2^2 = 0$. Then $B|y| + D_y B^\top y = 0$ almost surely, or equivalently, $(BD_y + D_y B^\top)y = 0$, since $|y| = D_y y$. For each sign vector $s \in \{-1, 1\}^r$, define $O_s := \{y : \operatorname{sgn}(y) = s\}$ and $D_s := \operatorname{diag}(s)$. Since $y = R^\top z$ has an everywhere strictly positive density, the continuous linear map $y \mapsto (BD_s + D_s B^\top)y$ must vanish throughout O_s. Since a linear map that vanishes on a nonempty open set is zero, $BD_s + D_s B^\top = 0$ for every s.

Key idea is that every sign pattern occurs with non-zero probability. Taking $D_s = I$ gives $B^\top = -B$ (thus B is skew-symmetric here). Substituting this identity back into the preceding equation yields $BD_s = D_s B$ for every sign matrix D_s. For $i \neq j$, choosing $s_i = 1$ and $s_j = -1$ gives $B_{ij} = 0$, while skew-symmetry gives $B_{ii} = 0$. Thus $B = 0$, and consequently $\delta N_B = 0$. Therefore, $H_B[\delta N_B, \delta N_B] = \mathbb{E}\|J_B[B]\|_2^2 > 0$ whenever $\delta N_B \neq 0$. Together with positivity of the S-, K-, and C-blocks and the block-diagonal decomposition of the Hessian, this shows that $H_N[\delta N, \delta N] > 0$ for every nonzero $\delta N \in N_W \mathcal{M}$.

Uniformity. Since $\mathcal{M} = \{\rho[Q_r R, -Q_r R] \mid R \in O(r)\}$ is the continuous image of the compact set $\mathrm{O}(r)$, it is compact. Consider its unit normal bundle

$$N^1\mathcal{M} := \{(W, \delta N) \mid W \in \mathcal{M},\ \delta N \in N_W \mathcal{M},\ \|\delta N\|_F = 1\}.$$

The normal spaces of the smooth manifold $\mathcal{M}$ vary continuously with W (Lee, 2018, Proposition 2.16). Hence $N^1\mathcal{M}$ is a closed subset of the product of $\mathcal{M}$ and the ambient unit sphere, and is therefore compact. Since $L \in C^2$, the map $(W, \delta N) \mapsto \langle \delta N, \nabla^2 L(W)\delta N\rangle_F$ is continuous on $N^1\mathcal{M}$. By the pointwise positive definiteness established above, it is strictly positive there. Consequently, it attains a strictly positive minimum $c > 0$. By homogeneity,

$$\langle \delta N, \nabla^2 L(W)\delta N\rangle_F \geq c\|\delta N\|_F^2$$

for every $W \in \mathcal{M}$ and every $\delta N \in N_W \mathcal{M}$. Thus the lower bound is uniform over the entire manifold and its corresponding normal spaces.

To connect this bound to normal attraction, note that ∇L vanishes on $\mathcal{M}$. Differentiating along a tangent direction therefore gives $\nabla^2 L(W)\tau = 0$ for every $\tau \in T_W \mathcal{M}$. By self-adjointness,

$$\langle \nabla^2 L(W)\delta N, \tau\rangle_F = \langle \delta N, \nabla^2 L(W)\tau\rangle_F = 0.$$

Hence the Hessian preserves $N_W \mathcal{M}$. Consequently, perturbations that start normal remain normal under the linearized gradient flow, so the coercivity bound applies throughout their evolution.

Let φ_t denote the population gradient flow. Since $W \in \mathcal{M}$ is stationary, $D\varphi_t(W) = \exp(-t\nabla^2 L(W))$. The preceding properties give, for every $t \geq 0$,

$$D\varphi_t(W)\tau = \tau, \qquad \|D\varphi_t(W)\delta N\|_F \leq e^{-ct}\|\delta N\|_F,$$

for $\tau \in T_W \mathcal{M}$ and $\delta N \in N_W \mathcal{M}$. Thus the invariant normal bundle contracts uniformly exponentially, while tangent perturbations remain unchanged. This proves that $\mathcal{M}$ is a NASM. □

D.5 Auxiliary Lemmas and Proofs

D.5.1 Verification of the Regularity Assumptions

Lemmas D.4 and D.5 provide the necessary assumptions to apply Theorem C.6. Near $\mathcal{M}_{\mathrm{ReLU}}$, the obstruction to C^3 regularity of the population loss arises solely from antipodal weight pairs (see intro of the proof in Lemma D.4). The kernel estimates in Lemma D.3 are central to apply the reduction theorem despite the failure of ambient C^3 regularity.

Lemma D.3 (Antipodal kernel estimate). *Let x have a finite Gaussian-mixture distribution with positive-definite component covariances, and define $K(a, b) := \mathbb{E}[\phi(a^\top x)\phi(b^\top x)]$. Fix constants $0 < a_- \leq a_+ < \infty$ and $0 \leq P < \infty$. There exist constants $t_0, \varepsilon_0, C > 0$, with $\varepsilon_0 < a_-/2$, such that the following statements hold.*

Let $a, p, q : I \to \mathbb{R}^d$ be C^1 curves on an open interval I, satisfying $a_- \le \|a(s)\| \le a_+$ and $\|p(s)\| + \|q(s)\| \le P$ for every $s \in I$. Then, for every $s \in I$ and $|t| \le t_0$,

$$K(a(s) + tp(s), -a(s) + tq(s)) \le C|t|^3, \tag{147}$$

$$\left| \frac{d}{ds} K(a(s) + tp(s), -a(s) + tq(s)) \right| \le C|t|^3 \big(\|\dot{a}(s)\| + \|\dot{p}(s)\| + \|\dot{q}(s)\| \big). \tag{148}$$

Here, differentiation with respect to s holds t fixed. Moreover, for every a satisfying $a_- \le \|a\| \le a_+$, the map $(\delta a, \delta b) \mapsto K(a + \delta a, -a + \delta b)$ is C^2 on $\{(\delta a, \delta b) : \|\delta a\|^2 + \|\delta b\|^2 < \varepsilon_0^2\}$. Writing $r := (\|\delta a\|^2 + \|\delta b\|^2)^{1/2}$, its derivatives satisfy

$$\left\| \nabla_{(\delta a, \delta b)} K(a + \delta a, -a + \delta b) \right\| \le C r^2, \tag{149}$$

$$\left\| \nabla^2_{(\delta a, \delta b)} K(a + \delta a, -a + \delta b) \right\| \le C r. \tag{150}$$

All constants are uniform over the stated parameter ranges. In particular, the kernel, its gradient, and its Hessian vanish at every antipodal pair $(a, -a)$ in this range.

Moreover, K is not C^3 at any antipodal pair $(a, -a)$ with $a \neq 0$.

(Proof of Lemma D.3). We first take $t > 0$. Set $e := \frac{a}{\|a\|}$ and decompose

$$p = p_\parallel e + p_\perp, \qquad q = q_\parallel e + q_\perp, \qquad p_\perp, q_\perp \perp e.$$

Every $x \in \mathbb{R}^d$ can be written uniquely as $x = ue + z$ with $z \in e^\perp$. Define $A_t := \|a\| + tp_\parallel$ and $B_t := \|a\| - tq_\parallel$. Since $a_- \le \|a(s)\| \le a_+$ and $|p_\parallel(s)|, |q_\parallel(s)| \le P$, choosing $t_0 > 0$ such that $t_0 P \le \frac{a_-}{2}$ gives $\frac{a_-}{2} \le A_t(s), B_t(s) \le a_+ + \frac{a_-}{2}$ uniformly over every $s \in I$ and $0 < t \le t_0$.

We have $(a + tp)^\top x = A_t u + tp_\perp^\top z = A_t(u - l_t(z))$, where $l_t(z) := -\frac{tp_\perp^\top z}{A_t}$, and $(-a + tq)^\top x = -B_t u + tq_\perp^\top z = B_t(r_t(z) - u)$, where $r_t(z) := \frac{tq_\perp^\top z}{B_t}$. Both ReLU factors are positive precisely when $l_t(z) < u < r_t(z)$. Let $d_t(z) := \frac{q_\perp^\top z}{B_t} + \frac{p_\perp^\top z}{A_t}$, so that $r_t(z) - l_t(z) = t d_t(z)$. Then,

$$K(a + tp, -a + tq) = A_t B_t \int_{e^\perp} \mathbf{1}_{\{d_t(z) > 0\}} \int_{l_t(z)}^{r_t(z)} (u - \ell_t(z))(r_t(z) - u) p_x(ue + z) \, \mathrm{d}u \mathrm{d}z.$$

By changing the variable, $u = \ell_t(z) + t d_t(z) y$ for $y \in [0, 1]$, we obtain

$$K(a + tp, -a + tq) = t^3 A_t B_t \int_{e^\perp} [d_t(z)_+]^3 \int_0^1 y(1 - y) p_x \left(t c_t(y, z) e + z \right) \mathrm{d}y \mathrm{d}z, \tag{151}$$

where $c_t(y, z) := -\frac{p_\perp^\top z}{A_t} + d_t(z) y$.

Note that $|d_t(z)| \le \left(\frac{\|q_\perp\|}{B_t} + \frac{\|p_\perp\|}{A_t} \right) \|z\| \le C_1 \|z\|$, for some $C_1 > 0$, and due to orthogonality, $\|t c_t(y, z) e + z)\| \ge \|z\|^2$. Since p_x is a finite mixture of nondegenerate Gaussian densities, there exist $C_0, c_0 > 0$ such that $p_x(w) \le C_0 e^{-c_0 \|w\|^2}$ for every $w \in \mathbb{R}^d$. Therefore, $p_x(t c_t(y, z) e + z) \le C_0 e^{-c_0 \|z\|^2}$. Hence, Eq. (151) is bounded by

$$t^3 C_2 \int_{e^\perp} \|z\|^3 e^{-c_0 \|z\|^2} \mathrm{d}z \int_0^1 y(1 - y) \mathrm{d}y.$$

This integral is uniformly integrable. To see this (and for the subsequent differentiation with respect to s), we use the following orthonormal coordinates. Locally in s, choose C^1 vector fields $v_2(s), \ldots, v_d(s)$ such that $\{e(s), v_2(s), \ldots, v_d(s)\}$ is an orthonormal basis of $\mathbb{R}^d$, where $e(s) := a(s)/\|a(s)\|$. Define $E_s : \mathbb{R}^{d-1} \to e(s)^\perp$ by $E_s \xi := \sum_{j=1}^{d-1} \xi_j v_{j+1}(s)$.[13] Since

[13] For the derivative estimates, choose the basis vectors to satisfy $\dot{v}_j(s) = -e(s) \langle \dot{e}(s), v_j(s) \rangle$, which gives $\|\dot{E}_s\|_2 \le \|\dot{e}(s)\|$.

$E_s^\top E_s = I_{d-1}$, the substitution $z = E_s\xi$ preserves lengths and $(d-1)$-dimensional volume ($\|z\| = \|\xi\|$ and $dz = d\xi$). Consequently,

$$\int_{e(s)^\perp} \|z\|^3 e^{-c_0\|z\|^2}\,\mathrm{d}z = \int_{\mathbb{R}^{d-1}} \|\xi\|^3 e^{-c_0\|\xi\|^2}\,\mathrm{d}\xi$$
$$= |\mathbb{S}^{d-2}| \int_0^\infty r^{d+1} e^{-c_0 r^2}\,\mathrm{d}r < \infty,$$

where $|\mathbb{S}^{d-2}|$ denotes the surface area of the unit sphere in $\mathbb{R}^{d-1}$. This integral is independent of s and the direction $e(s)$. Since $\int_0^1 y(1-y)\,\mathrm{d}y = 1/6$, the preceding bound yields $K(a+tp, -a+tq) \le C_* t^3$, uniformly over the parameters satisfying the lemma's assumptions and $t \in (0, t_0]$. This proves the first claim.

We now differentiate with respect to s, keeping $t > 0$ fixed. Using the same substitution $z = E_s\xi$, we rewrite Eq. (151) over the fixed domain $\mathbb{R}^{d-1} \times [0,1]$. To differentiate $K(a(s) + tp(s), -a(s) + tq(s))$ under the integral sign, we will bound the s-derivative of the transformed integrand by an integrable function, uniformly on compact subintervals of I. Write $w_t := tc_t(y, E_s\xi)e(s) + E_s\xi$.

Set $\Lambda(s) := \|\dot a(s)\| + \|\dot p(s)\| + \|\dot q(s)\|$. By the choice of the orthonormal frame, $\|\dot E_s\|_{\mathrm{op}} \le \|\dot e(s)\| \le a_-^{-1}\|\dot a(s)\|$. The definitions of A_t, B_t therefore give $|\partial_s A_t| + |\partial_s B_t| \le C\Lambda(s)$. The uniform parameter bounds, the positive lower bounds on A_t, B_t, and the bounds on $\dot E_s$, $\partial_s A_t$, and $\partial_s B_t$ yield, by the product and quotient rules,

$$|d_t(E_s\xi)| + |c_t(y, E_s\xi)| \le C\|\xi\|,$$
$$|\partial_s[d_t(E_s\xi)]| + |\partial_s[c_t(y, E_s\xi)]| \le C\Lambda(s)\|\xi\|.$$

We consequently have $\|\partial_s w_t\| \le C\Lambda(s)\|\xi\|$ and $\|w_t\| \ge \|\xi\|$.

The Gaussian-mixture density satisfies $p_x(w) + \|\nabla p_x(w)\| \le Ce^{-c\|w\|^2}$ for suitable constants $C, c > 0$. Also, $u \mapsto (u_+)^3$ is continuously differentiable. Thus, applying the product and chain rules to the transformed integrand gives

$$\left|\partial_s\left(A_t B_t [d_t(E_s\xi)_+]^3 y(1-y) p_x(w_t)\right)\right| \le C\Lambda(s) y(1-y)(\|\xi\|^3 + \|\xi\|^4) e^{-c\|\xi\|^2}.$$

The right-hand side is integrable in (ξ, y). On every compact subinterval of I, Λ is bounded because the curves are C^1, so dominated convergence justifies differentiation under the integral. The prefactor t^3 remains unchanged, yielding

$$\left|\frac{d}{ds} K(a(s) + tp(s), -a(s) + tq(s))\right| \le Ct^3\big(\|\dot a(s)\| + \|\dot p(s)\| + \|\dot q(s)\|\big).$$

This proves Eq. (148) for $t > 0$.

For $t < 0$, write $t = -h$ with $h > 0$ and apply the preceding argument with perturbation directions $(-p, -q)$. This gives the bound $C|t|^3$ on both sides of zero.

At $t = 0$, the pair is exactly antipodal and $K(a, -a) = \mathbb{E}\left[\phi(a^\top X)\phi(-a^\top X)\right] = 0$. Thus, both the kernel and its s-derivative vanish at $t = 0$, so the cubic estimate extends to that point.

Finally, for a nonzero perturbation $\delta := (\delta a, \delta b)$, write $r := \|\delta\|$ and $\delta = r\widehat\delta$, where $\widehat\delta = (\widehat p, \widehat q)$ belongs to the unit sphere of $\mathbb{R}^d \times \mathbb{R}^d$. The scaled integral representation gives $K(a+\delta a, -a+\delta b) = r^3\Gamma(a, \widehat\delta, r)$. We claim that, for sufficiently small $r_0 > 0$, $|\Gamma| + |\partial_r\Gamma| + \|D_{\widehat\delta}\Gamma\| \le C$ uniformly over $a_- \le \|a\| \le a_+$, $\|\widehat\delta\| = 1$, and $0 \le r \le r_0$. Here, $D_{\widehat\delta}\Gamma$ denotes differentiation along the unit sphere.

The directional derivative bound follows from the preceding differentiation argument, with a fixed and $(p, q) = (\widehat p, \widehat q)$ varying along the unit sphere. For the radial derivative, differentiating A_r, B_r, d_r, c_r with respect to r gives the same bounds. The denominators remain uniformly positive, while d_r, c_r and their radial derivatives are bounded by $C\|\xi\|$. The product and chain rules therefore yield an integrable bound of the form $C(1 + \|\xi\|^4)e^{-c\|\xi\|^2}$ for the differentiated integrand. Differentiation under the integral sign thus justifies the claimed bounds.

Since $\|D_\delta r\| = 1$ and $\|D_\delta \widehat{\delta}\| \le r^{-1}$, the chain rule gives

$$\begin{aligned}\|\nabla_\delta K(a+\delta a, -a+\delta b)\| &\le 3r^2|\Gamma| + r^3|\partial_r \Gamma| + r^2\|D_{\widehat{\delta}}\Gamma\| \\ &\le Cr^2 = C\big(\|\delta a\|^2 + \|\delta b\|^2\big).\end{aligned}$$

At $\delta = 0$, the cubic bound $K(a+\delta a, -a+\delta b) = O(\|\delta\|^3)$ implies differentiability with derivative zero. The preceding gradient bound also shows continuity of the gradient at zero. This proves Eq. (149).

The same integral representation also establishes local C^2 regularity of the kernel near antipodal pairs. Indeed, $u \mapsto (u_+)^3$ is C^2, with derivatives $3(u_+)^2$ and $6u_+$. The remaining factors in the integral defining Γ are smooth in r and the normalized perturbation directions $(\widehat{p}, \widehat{q})$, Moreover, the density $p_x(w)$, its gradient $\nabla_w p_x(w)$, and its Hessian $\nabla_w^2 p_x(w)$ are bounded by Gaussian envelopes. Thus, differentiating the integrand up to twice produces bounds of the form $C(1 + \|\xi\|^m)e^{-c\|\xi\|^2}$ for some finite m. Dominated convergence consequently shows that Γ is C^2 in these variables, with uniformly bounded derivatives through order two.

For $\delta \neq 0$, the representation $K(a+\delta a, -a+\delta b) = r^3\Gamma(a, \widehat{\delta}, r)$ shows that the kernel is C^2. The preceding chain-rule estimates give $\|D_\delta^2 K(a+\delta a, -a+\delta b)\| \le C\|\delta\|$. At $\delta = 0$, the previously established value bound implies $\nabla_\delta K(a, -a) = 0$. The gradient bound then gives

$$\frac{\|\nabla_\delta K(a+\delta a, -a+\delta b) - \nabla_\delta K(a, -a)\|}{\|\delta\|} \le C\|\delta\| \longrightarrow 0.$$

Thus the gradient is differentiable at zero, with $D_\delta^2 K(a, -a) = 0$. The Hessian estimate shows that these second derivatives converge continuously to zero as $\delta \to 0$. Hence the kernel is C^2 throughout the perturbation neighborhood, and Eq. (150) holds there.

Failure of C^3 regularity. Fix $a \neq 0$, set $e := a/\|a\|$, and choose a unit vector $u \in e^\perp$. Consider $F(t) := K(a, -a + tu)$. Applying Eq. (151) with $p = 0$ and $q = u$, and using the same integrable Gaussian envelope, gives

$$\lim_{t \downarrow 0} \frac{F(t)}{t^3} = \frac{1}{6\|a\|} \int_{e^\perp} (u^\top z)_+^3 p_x(z)\, \mathrm{d}z =: c(a, u) > 0.$$

The integral is finite and strictly positive because p_x is everywhere positive and $u^\top z > 0$ on a nonempty open subset of $e^\perp$.

The preceding estimates give $F(0) = F'(0) = F''(0) = 0$. If F were C^3 at zero, Taylor's theorem would imply $F(t) = F'''(0)t^3/6 + o(|t|^3)$. The right limit would then give $F'''(0)/6 = c(a, u) > 0$, forcing $F(t) < 0$ for sufficiently small negative t. This contradicts $F(t) = K(a, -a + tu) \ge 0$. Thus F, and hence K, is not C^3 at the antipodal pair.

□

Lemma D.4 *For $Y \in \mathcal{M}$, and the normal Hessian $H_N(Y) := P_N(Y)\nabla^2 L(Y) P_N(Y)|_{N_Y\mathcal{M}}$, There exist constants $\zeta, C > 0$ such that, for every $Y \in \mathcal{M}$ and every $N \in N_Y\mathcal{M}$ satisfying $\|N\|_F \le \zeta$,*

$$L(Y + N) = L(Y) + \frac{1}{2}\langle N, H_N(N)N\rangle_F + R(Y, N), \tag{152}$$

where $|R(Y, N)| \le C\|N\|_F^3$, and, for every $M \in N_v\mathcal{M}$, $|D_N R(Y, N)[M]| \le C\|N\|_F^2\|M\|_F$.

Moreover, let $Y(s) \subset \mathcal{M}$ be a C^1 curve and let $N(s) \in N_{Y(s)}\mathcal{M}$ be parallel in the normal bundle (meaning $\nabla_s^\perp N(s) := P_N(Y(s))\dot{N}(s) = 0$). Then

$$\left| \frac{d}{ds}\Big|_{s=0} R(Y(s), N(s)) \right| \le C\|\dot{Y}(0)\|_F \|N(0)\|_F^3. \tag{153}$$

In particular, the quadratic form $(Y, N) \mapsto \langle N, H_N(Y)N\rangle_F$ is continuously differentiable in the base point v along $\mathcal{M}$, although the ambient population loss need not be C^3 in a neighborhood of $\mathcal{M}$.

(Proof of Lemma D.4).

Write $W = [w_1, \dots, w_{2r}]$ and define the population ReLU kernel $K(a,b) := \mathbb{E}\left[\phi(a^\top x)\phi(b^\top x)\right]$. Expanding the squared reconstruction error gives

$$L(W) = \frac{1}{2}\mathbb{E}\|x\|^2 - \sum_{i=1}^{2r}\mathbb{E}\left[x^\top w_i \phi(w_i^\top x)\right] + \frac{1}{2}\sum_{i,j=1}^{2r}\langle w_i, w_j\rangle K(w_i, w_j) + \frac{\lambda}{2}\sum_{i=1}^{2r}\|w_i\|^2.$$

The data distribution is centrally symmetric, i.e., $x \overset{d}{=} -x$. Consequently, $\mathbb{E}\left[x^\top w\phi(w^\top x)\right] = \mathbb{E}\left[(w^\top x)^2 \mathbf{1}_{\{w^\top x>0\}}\right] = \frac{1}{2}\mathbb{E}[(w^\top x)^2] = \frac{1}{2}w^\top \Sigma_x w$. It follows that

$$L(W) = \frac{1}{2}\mathbb{E}\|x\|^2 + \frac{1}{2}\sum_{i=1}^{4} w_i^\top(\lambda I - \Sigma_x)w_i + \frac{1}{2}\sum_{i,j=1}^{4}\langle w_i, w_j\rangle K(w_i, w_j), \tag{154}$$

Since the first two terms are smooth, possible nonsmoothness arises only from the pairwise kernels $K(a,b)$. Our goal is to bound this kernel in order to establish Eq. (152). A given pair must fall into one of three mutually exclusive categories: $i = j$, an antipodal pair ($i \neq j$), or a non-antipodal pair ($i \neq j$). As we later show, potential non-smoothness arises exclusively from antipodal pairs, thus our proof strategy partitions the analysis into a smooth component (comprising the $i = j$ and non-antipodal cases) and a potentially non-smooth component (the antipodal case). In this proof, we denote $\sigma_j(A)$ as j-th singular value of a matrix A, and $\lambda_j(C)$ as j-th eigenvalue of a positive semi-definite matrix C.

1. Smoothness away from antipodal pairs. Consider first one Gaussian-mixture component $x \sim \mathcal{N}(\mu, \Sigma_x)$. For $A = [a, b] \in \mathbb{R}^{d\times 2}$, the projected vector $z = A^\top x = \begin{bmatrix} a^\top x & b^\top x \end{bmatrix}^\top$ is Gaussian with mean and covariance $m = A^\top \mu$ and $C = A^\top \Sigma_x A$. If a, b are linearly independent, then for every nonzero $v \in \mathbb{R}^2$, $v^\top C v = (Av)^\top \Sigma_x (Av) > 0$, since $Av \neq 0$ and $\Sigma_x \succ 0$. Hence $C \succ 0$.

At every $Y \in \mathcal{M}$, each distinct non-antipodal pair of columns is orthogonal, with both norms equal to ρ. Thus, for such a pair $A_0 = [a_0, b_0]$, we have $\sigma_{\min}(A_0) = \sigma_{\max}(A_0) = \rho$. For $A = A_0 + \Delta A$, Weyl's inequality gives $\sigma_{\min}(A) \geq \rho - \|\Delta A\|_2$ and $\sigma_{\max}(A) \leq \rho + \|\Delta A\|_2$. Since $\|\Delta A\| \leq \|\Delta A\|_F \leq \|N\|_F < \zeta$, choosing $\zeta \leq \rho/2$ ensures $\sigma_{\min}(A) \geq \rho/2$ and $\sigma_{\max}(A) \leq 3\rho/2$, uniformly over all such reference pairs and perturbations. Consequently, for $C = A^\top \Sigma_x A$,

$$c_- I_2 \preceq C \preceq c_+ I_2, \quad c_- := \frac{\rho^2}{4}\lambda_{\min}(\Sigma_x) > 0, \quad c_+ := \frac{9\rho^2}{4}\lambda_{\max}(\Sigma_x).$$

Moreover, the projected mean satisfies $\|m(A)\| = \|A^\top \mu\| \leq \|A\|\|\mu\| \leq \frac{3\rho}{2}\|\mu\| =: M$.

For this Gaussian component, the kernel therefore admits the fixed-domain representation

$$K_\mu(a,b) = \int_0^\infty \int_0^\infty z_1 z_2\, \gamma(z; m, C)\, \mathrm{d}z_1 \mathrm{d}z_2,$$

where $\gamma(z; m, C) = (2\pi\sqrt{\det C})^{-1}\exp(-\frac{1}{2}(z-m)^\top C^{-1}(z-m))$. The above uniform bounds give $(z-m)^\top C^{-1}(z-m) \geq \frac{\|z-m\|^2}{c_+} \geq \frac{\|z\|^2}{2c_+} - \frac{M^2}{c_+}$. Together with $\det C \geq c_-^2$, this yields $\gamma_A(z) := \gamma(z; m(A), C(A)) \leq C_0 e^{-c\|z\|^2}$ for constants $C_0, c > 0$ independent of the reference pair and perturbation.

For any matrix direction E, the chain rule gives $D_A\gamma_A(z)[E] = \gamma_A(z) D_A(\log\gamma_A(z))[E]$ (for $\gamma_A = \exp(\log\gamma_A)$). Since $\log\gamma_A(z)$ is quadratic in z, repeated application of the product rule shows that each component of $D_A^k\gamma_A(z)$ is the density multiplied by a polynomial in z of degree at most $2k$. Its coefficients are uniformly bounded by the bounds on A and C^{-1}. Hence, for every integer $k \geq 0$,

$$\|D_A^k \gamma_A(z)\|_2 \leq C_k(1 + \|z\|^{2k})e^{-c\|z\|^2}.$$

This bound remains integrable after multiplication by $z_1 z_2 \mathbf{1}_{(0,\infty)}$. Dominated convergence therefore justifies differentiation under the integral at every order and establishes continuity of the resulting derivatives. Thus, K_μ is C^∞ on $\{A \mid \|A - A_0\|_F < \zeta\}$ for every orthogonal reference pair A_0 on $\mathcal{M}$. The derivative bounds are uniform over all such A_0 and all A satisfying $\|A - A_0\|_F < \zeta$. This same logic applies to the other mixture components.

In the special case of a repeated pair where $a = b$, central symmetry gives the exact formula

$$K(a,a) = \mathbb{E}[(a^\top x)^2 \mathbf{1}_{\{a^\top x > 0\}}] = \frac{1}{2} a^\top \Sigma_x a,$$

which is polynomial and thus smooth.

2. Smooth and antipodal parts of the loss. Fix the column labeling of $Y \in \mathcal{M}$, the columns of the second block are the negatives of the corresponding columns of the first block. In a sufficiently small uniform neighborhood of $\mathcal{M}$, write $L = L_{\mathrm{sm}} + L_{\mathrm{anti}}$, where L_{anti} contains exactly the ordered cross-kernel terms whose column pairs are antipodal on $\mathcal{M}$, including their inner-product prefactors. The remaining terms in Eq. (154) form L_{sm}, which is C^∞ on this neighborhood by the preceding arguments.

Lemma D.3, together with the product rule and uniform bounds on the inner-product prefactors and their derivatives, gives

$$|L_{\mathrm{anti}}(Y+\Delta)| \le C\|\Delta\|_F^3, \qquad \|\nabla L_{\mathrm{anti}}(Y+\Delta)\|_F \le C\|\Delta\|_F^2, \tag{155}$$

uniformly over $Y \in \mathcal{M}$ and sufficiently small ambient perturbations Δ. In particular, for $N, M \in N_Y\mathcal{M}$,

$$|L_{\mathrm{anti}}(Y+N)| \le C\|N\|_F^3, \qquad |D_N L_{\mathrm{anti}}(Y+N)[M]| \le C\|N\|_F^2\|M\|_F.$$

Now let $Y(s) \subset \mathcal{M}$ be a C^1 curve and let $N(s)$ be parallel in the normal bundle. Since $P_N(Y(s))\dot{N}(s) = 0$, the vector $\dot{N}(s)$ is tangent, whereas $N(s)$ is normal. Consequently, $\frac{d}{ds}\|N(s)\|_F^2 = 2\langle N(s), \dot{N}(s)\rangle_F = 0$. Differentiating $N(s) = P_N(Y(s))N(s)$ and using the parallel transport condition gives $\dot{N}(s) = DP_N(Y(s))[\dot{Y}(s)]N(s)$. Smoothness of P_N and compactness of $\mathcal{M}$ therefore yield

$$\|\dot{N}(s)\|_F \le C\|\dot{Y}(s)\|_F\|N(s)\|_F. \tag{156}$$

For $r := \|N(s)\|_F > 0$, write $N(s) = r\widehat{N}(s)$. Then r is constant and $\|\dot{\widehat{N}}(s)\|_F \le C\|\dot{Y}(s)\|_F$.

Applying the tangential estimate in Lemma D.3 to each antipodal pair therefore bounds its kernel derivative by $C\|\dot{Y}(s)\|_F r^3$. The inner-product prefactors are uniformly bounded, and their s-derivatives are bounded by $C\|\dot{Y}(s)\|_F$. The product rule consequently gives

$$\left|\frac{d}{ds} L_{\mathrm{anti}}(Y(s)+N(s))\right| \le C\|\dot{Y}(s)\|_F\|N(s)\|_F^3. \tag{157}$$

The same conclusion holds when $r = 0$, since L_{anti} vanishes identically on $\mathcal{M}$. The cubic estimate in Eq. (155) implies $L_{\mathrm{anti}}(Y) = 0$ and $DL_{\mathrm{anti}}(Y) = 0$. Moreover, the gradient estimate gives

$$\frac{\|\nabla L_{\mathrm{anti}}(Y+\Delta) - \nabla L_{\mathrm{anti}}(Y)\|_F}{\|\Delta\|_F} \le C\|\Delta\|_F \to 0.$$

Thus ∇L_{anti} is differentiable at Y with derivative zero, so $D^2 L_{\mathrm{anti}}(Y) = 0$. Since every $Y \in \mathcal{M}$ is stationary for L, it follows that

$$DL_{\mathrm{sm}}(Y) = DL(Y) = 0, \qquad D^2 L_{\mathrm{sm}}(Y) = D^2 L(Y). \tag{158}$$

This also establishes the required tangential regularity of the normal Hessian. Indeed, along $\mathcal{M}$,

$$H_N(Y) = P_N(Y) D^2 L_{\mathrm{sm}}(Y) P_N(Y)\big|_{N_Y\mathcal{M}}.$$

Both $D^2 L_{\mathrm{sm}}(Y)$ and $P_N(Y)$ vary smoothly along each labeled component of $\mathcal{M}$. Hence H_N is a smoothly varying family of operators on the normal spaces.

3. Taylor expansion of the smooth part. Since L_{sm} is C^∞, Taylor's theorem along $t \mapsto Y + tN$ and Eq. (158) give

$$L_{\mathrm{sm}}(Y+N) = L_{\mathrm{sm}}(Y) + \frac{1}{2}\langle N, H_N(Y)N\rangle_F + R_{\mathrm{sm}}(Y,N). \tag{159}$$

, where the integral remainder is

$$R_{\mathrm{sm}}(Y,N)=\frac{1}{2}\int_0^1(1-t)^2D^3L_{\mathrm{sm}}(Y+tN)[N,N,N]\,\mathrm{d}t. \tag{160}$$

Choose $\zeta>0$ so that the closed tubular neighborhood $\overline{\mathcal{U}_\zeta}:=\{Y+N:Y\in\mathcal{M},\ N\in N_Y\mathcal{M},\ \|N\|_F\le\zeta\}$ lies inside the region on which L_{sm} is smooth (since $\mathcal{M}$ is compact, $\overline{\mathcal{U}_\zeta}$ is compact). Consequently, $M_3:=\sup_{Z\in\overline{\mathcal{U}_\zeta}}\|D^3L_{\mathrm{sm}}(Z)\|<\infty$. It follows from Eq. (160) that

$$|R_{\mathrm{sm}}(Y,N)|\le\frac{M_3}{2}\int_0^1(1-t)^2\,\mathrm{d}t\,\|N\|_F^3\le C\|N\|_F^3. \tag{161}$$

We next control differentiation in the normal variable. For $M\in N_Y\mathcal{M}$, differentiating the definition of the remainder gives $D_NR_{\mathrm{sm}}(Y,N)[M]=DL_{\mathrm{sm}}(Y+N)[M]-D^2L_{\mathrm{sm}}(Y)[N,M]$. Taylor's theorem applied to the gradient yields

$$DL_{\mathrm{sm}}(Y+N)[M]=D^2L_{\mathrm{sm}}(Y)[N,M]+\int_0^1(1-t)D^3L_{\mathrm{sm}}(Y+tN)[N,N,M]\,\mathrm{d}t.$$

Hence, $D_NR_{\mathrm{sm}}(Y,N)[M]=\int_0^1(1-t)D^3L_{\mathrm{sm}}(Y+tN)[N,N,M]\,\mathrm{d}t$, and therefore

$$|D_NR_{\mathrm{sm}}(Y,N)[M]|\le C\|N\|_F^2\|M\|_F. \tag{162}$$

It remains to control differentiation in the base point. Fix $Y\in\mathcal{M}$, $\tau\in T_Y\mathcal{M}$, and $N\in N_Y\mathcal{M}$ with $\|N\|_F\le\zeta$. Choose a C^1 curve $Y(s)\subset\mathcal{M}$ satisfying $Y(0)=Y$, $\dot Y(0)=\tau$, and let $N(s)$ be the parallel normal transport of N. As established above, $\|N(s)\|_F$ is constant and Eq. (156) holds. We now differentiate the integral remainder. Define $Z_t(s):=Y(s)+tN(s)$, $\dot Z_t(s)=\dot Y(s)+t\dot N(s)$. Using Eq. (160) and the chain rule gives

$$\begin{aligned}\frac{d}{ds}R_{\mathrm{sm}}(Y(s),N(s))=\frac{1}{2}\int_0^1(1-t)^2\Big(&D^4L_{\mathrm{sm}}(Z_t(s))[\dot Y(s)+t\dot N(s),N(s),N(s),N(s)]\\&+3D^3L_{\mathrm{sm}}(Z_t(s))[\dot N(s),N(s),N(s)]\Big)\,dt.\end{aligned} \tag{163}$$

The first term arises from differentiating the evaluation point $Z_t(s)$, while the second arises from differentiating the three copies of $N(s)$. Since L_{sm} is smooth on the compact neighborhood $\overline{\mathcal{U}_\zeta}$, $M_4:=\sup_{Z\in\overline{\mathcal{U}_\zeta}}\|D^4L_{\mathrm{sm}}(Z)\|<\infty$. Using the uniform bounds on D^3L_{sm} and D^4L_{sm} in above equation, we obtain

$$\left|\frac{d}{ds}R_{\mathrm{sm}}(Y(s),N(s))\right|\le C\Big(\|\dot Y(s)\|_F\|N(s)\|_F^3+\|\dot N(s)\|_F\|N(s)\|_F^3+\|\dot N(s)\|_F\|N(s)\|_F^2\Big).$$

Substituting Eq. (156) gives $\left|\frac{d}{ds}R_{\mathrm{sm}}(Y(s),N(s))\right|\le C\|\dot Y(s)\|_F\left(\|N(s)\|_F^3+\|N(s)\|_F^4\right)$. Finally, since $\|N(s)\|_F\le\zeta$, the fourth-order term can be absorbed into the cubic term. Thus,

$$\left|\frac{d}{ds}R_{\mathrm{sm}}(Y(s),N(s))\right|\le C\|\dot Y(s)\|_F\|N(s)\|_F^3. \tag{164}$$

To complete the proof, define $R(Y,N):=R_{\mathrm{sm}}(Y,N)+L_{\mathrm{anti}}(Y+N)$. Since $L_{\mathrm{anti}}(Y)=0$, Eq. (159) gives $L(Y+N)=L(Y)+\frac{1}{2}\langle N,H_N(Y)N\rangle_F+R(Y,N)$, which proves Eq. (152).

Combining Eq. (D.5.1) with Eq. (161) and Eq. (162), respectively, yields

$$|R(Y,N)|\le C\|N\|_F^3,\qquad |D_NR(Y,N)[M]|\le C\|N\|_F^2\|M\|_F.$$

Finally, along any C^1 curve $Y(s)\subset\mathcal{M}$ with $N(s)$ parallel in the normal bundle, Eq. (157) and Eq. (164) imply $\left|\frac{d}{ds}R(Y(s),N(s))\right|\le C\|\dot Y(s)\|_F\|N(s)\|_F^3$. Evaluating at $s=0$ proves Eq. (153). All constants are uniform over $Y\in\mathcal{M}$ and $\|N\|_F\le\zeta$. □

Lemma D.5 (Verification of the remaining regularity assumptions). *Consider the ReLU autoencoder and Gaussian-mixture distribution in Sections D.1, with fixed $\sigma > 0$. Let $f(x;W) := W\phi(W^\top x)$, $\ell(x;W) := \frac{1}{2}\|x - f(x;W)\|^2 + \frac{\lambda}{2}\|W\|_F^2$, $L(W) := \mathbb{E}[\ell(x;W)]$, and let $g(x;W) := \nabla_W \ell(x;W)$ denote the sample gradient. We identify matrix parameters and gradients with their vectorizations through the isometry h. Then there exists an open neighborhood $\mathcal{V}$ of $\mathcal{M}$ (chosen sufficiently small that every column of every $W \in \mathcal{V}$ is nonzero), such that the following statements hold.*

1. *$L \in C^2(\mathcal{V})$, ∇L is locally Lipschitz on $\mathcal{V}$, and $\mathbb{E}[g(x;W)] = \nabla L(W)$ for every $W \in \mathcal{V}$.*
2. *For every compact $\mathcal{C} \subset \mathcal{V}$ and every finite $p \geq 1$, $\sup_{W\in\mathcal{C}} \mathbb{E}\|g(x;W)\|_F^{2p} < \infty$.*
3. *For every compact $\mathcal{C} \subset \mathcal{V}$, there exists $C_{\mathcal{C}} < \infty$ such that, for $W, V \in \mathcal{C}$ with $\|W - V\|_F \leq 1$, $\mathbb{E}\|g(x;W) - g(x;V)\|_F^4 \leq C_{\mathcal{C}}\|W - V\|_F$. Consequently, for every sufficiently small tubular neighborhood $\mathcal{U}_\zeta$ with $\overline{\mathcal{U}_\zeta} \subset \mathcal{V}$,*
$$\sup_{\substack{W\in\mathcal{U}_\zeta \\ \operatorname{dist}(W,\mathcal{M})\leq r}} \mathbb{E}\|g(x;W) - g(x;\pi(W))\|_F^4 \longrightarrow 0 \qquad \text{as } r \downarrow 0.$$
4. *For every $Y \in \mathcal{M}$, $P_T(Y)g(x;Y) = 0$ almost surely.*
5. *The normal noise covariance $\Sigma_N(Y) := P_N(Y)\mathbb{E}[g(x;Y) \otimes g(x;Y)]P_N(Y)|_{N_Y\mathcal{M}}$ and the bilinear map $\mathcal{B}_Y$ defined in Eq. (131) are locally Lipschitz along $\mathcal{M}$.*

In particular, choosing the local stable neighborhood $\mathcal{U}_{\mathrm{st}} \subset \mathcal{V}$ and a sufficiently small tube satisfying $\overline{\mathcal{U}_\zeta} \subset \mathcal{U}_{\mathrm{st}}$ verifies Assumptions P1 *and* S1-S4.

(Proof of Lemma D.5). Note that here, all constants below may depend on the fixed mixture parameters, the dimension, and the compact parameter set under consideration, but not on the SGD step size.

1. Population regularity. By Eq. (154), it suffices to verify C^2 regularity of the kernel terms appearing in the loss near $\mathcal{M}$. For diagonal terms, central symmetry gives $K(w_i, w_i) = \frac{1}{2}w_i^\top \Sigma_X w_i$, which is polynomial. Distinct columns of $Y \in \mathcal{M}$ are either orthogonal or antipodal. For pairs that are orthogonal on $\mathcal{M}$, their projected Gaussian covariances remain positive definite in a sufficiently small uniform neighborhood. The Gaussian integral argument in the proof of Lemma D.4 therefore shows that these kernel terms are C^∞. For pairs that are antipodal on $\mathcal{M}$, Lemma D.3 gives C^2 regularity on a uniform neighborhood.

Since these neighborhoods are uniform over $Y \in \mathcal{M}$ and there are finitely many kernel terms, choose $\varepsilon > 0$ sufficiently small that all the preceding regularity conclusions hold whenever $\|W - Y\|_F < \varepsilon$ for some $Y \in \mathcal{M}$. Define $\mathcal{V} := \{W \in \mathbb{R}^{d\times 2r} \mid \operatorname{dist}(W, \mathcal{M}) < \varepsilon\}$. Note that this does not have the properties such as stable foliation or nearest point projection. The inner-product prefactors and the remaining terms in Eq. (154) are polynomial, so $L \in C^2(\mathcal{V})$.

Choose the local stable neighborhood sufficiently small that $\mathcal{U}_{\mathrm{st}} \subset \mathcal{V}$. Then ∇L is locally Lipschitz on $\mathcal{U}_{\mathrm{st}}$. By compactness of $\mathcal{M}$, choose $\zeta > 0$, below the tubular radius, such that $\overline{\mathcal{U}_{2\zeta}} \subset \mathcal{U}_{\mathrm{st}}$. Thus L is C^2 on the open neighborhood $\mathcal{U}_{2\zeta}$ of $\overline{\mathcal{U}_\zeta}$, verifying Assumption P1.

2. Sample-gradient formula and moment bounds. Set $r_W(x) := x - f(x;W)$. For every fixed $W \in \mathcal{V}$, the gradient with respect to column w_i is
$$g_i(x;W) = -\phi(w_i^\top x)r_W(x) - \mathbf{1}_{\{w_i^\top x>0\}}\langle w_i, r_W(x)\rangle x + \lambda w_i.$$
Since $\mathcal{V}$ is chosen sufficiently small that all columns are nonzero, $\mathbb{P}(w_i^\top x = 0) = 0$ for every i.

On any bounded parameter set, $\|f(x;W)\| \leq \|W\|_F^2\|x\| \leq C\|x\|$. The sample-gradient formula therefore gives
$$\|g(x;W)\|_F \leq C(1 + \|x\|^2). \tag{165}$$
Since a finite Gaussian mixture has moments of every order, for every compact $\mathcal{C} \subset \mathcal{V}$ and every finite $p \geq 1$, $\sup_{W\in\mathcal{C}} \mathbb{E}\|g(x;W)\|_F^{2p} \leq C_{\mathcal{C},p}\big(1 + \mathbb{E}\|x\|^{4p}\big) < \infty$.

To justify exchanging differentiation and expectation, note that in each fixed parameter direction, local difference quotients of the sample loss are bounded by $C(1+\|x\|^2)$, uniformly for sufficiently small steps. Since this is integrable and the sample loss is differentiable at each fixed $W \in \mathcal{V}$ almost surely, dominated convergence gives $\mathbb{E}[g(x;W)] = \nabla L(W)$. This proves S1.

3. Activation-region estimates. For nonzero a, b, define $A(a,b) := \{x : \mathbf{1}_{\{a^\top x>0\}} \neq \mathbf{1}_{\{b^\top x>0\}}\}$. For every fixed $m \geq 0$ and $\kappa > 0$, there exists $C_{m,\kappa} < \infty$ such that

$$\mathbb{E}[(1+\|x\|^m)\mathbf{1}_{A(a,b)}(x)] \leq C_{m,\kappa}\|a-b\|, \qquad \min\{\|a\|, \|b\|\} \geq \kappa. \tag{166}$$

Indeed, $p_x(x) \leq C_0 e^{-c_0\|x\|^2}$. Let $\vartheta(a,b) \in [0,\pi]$ denote the angle between a and b. The fraction of directions on $\mathbb{S}^{d-1}$ for which the two signs disagree is $\vartheta(a,b)/\pi$. Thus, polar coordinates give

$$\begin{aligned}\mathbb{E}[(1+\|x\|^m)\mathbf{1}_{A(a,b)}(x)] &\leq C_0 \frac{|\mathbb{S}^{d-1}|}{\pi}\,\vartheta(a,b)\int_0^\infty (1+r^m) r^{d-1} e^{-c_0 r^2}\,\mathrm{d}r \\ &\leq C_m \vartheta(a,b).\end{aligned}$$

The radial integral is finite, and $\vartheta(a,b) \leq \frac{\pi}{2}\left\|\frac{a}{\|a\|} - \frac{b}{\|b\|}\right\| \leq \frac{\pi}{\kappa}\|a-b\|$. This proves Eq. (166).

Fix a compact $\mathcal{C} \subset \mathcal{V}$. All column norms on $\mathcal{C}$ are bounded away from zero. For $W, V \in \mathcal{C}$, set $\delta := \|W - V\|_F$. The sample-gradient formula and the Lipschitz property of ReLU give

$$\|g(x;W) - g(x;V)\|_F \leq C(1+\|x\|^2)\left(\delta + \sum_{i=1}^{4} \mathbf{1}_{A(w_i,v_i)}(x)\right). \tag{167}$$

Since each indicator takes values in $\{0,1\}$, the fourth power of the parenthesized expression is bounded by $C(\delta^4 + \sum_{i=1}^{2r} \mathbf{1}_{A(w_i,v_i)}(x))$. Taking expectations and applying Eq. (166) with $m = 8$ therefore yields $\mathbb{E}\|g(x;W) - g(x;V)\|_F^4 \leq C(\delta^4 + \delta) \leq C\delta$, for $\delta \leq 1$. Taking $\mathcal{C} = \overline{\mathcal{U}_\zeta} \subset \mathcal{V}$ and $V = \pi(W)$ gives a uniform bound Cr whenever $\operatorname{dist}(W, \mathcal{M}) \leq r \leq 1$. This proves the required uniform L^4-continuity (S2).

4. Normality of the sample gradient. For $Y = \rho[Q_r R, -Q_r R] \in \mathcal{M}$, the identity $\phi(u) - \phi(-u) = u$ gives $f(x;Y) = \rho^2 Q_r Q_r^\top x$, Thus $\ell(x;Y)$ is independent of R for each fixed x. At every sample where the ambient gradient exists, differentiating along any $\tau \in T_Y\mathcal{M}$ gives $\langle g(x;Y), \tau\rangle_F = 0$ (note that the exceptional samples lie in finitely many hyperplanes and have probability zero). Therefore, $P_T(Y)g(x;Y) = 0$ almost surely, proving S3.

5. Regularity of the covariance and $\mathcal{B}$. Define the ambient second-moment operator $M_2(W) := \mathbb{E}[g(x;W)g(x;W)^\top]$. Using $\|uu^\top - vv^\top\|_2 \leq (\|u\| + \|v\|)\|u - v\|$, together with Eq. (165), Eq. (167), and Eq. (166) with $m = 4$, gives $\|M_2(W) - M_2(V)\|_2 \leq C_{\mathcal{C}}\|W - V\|_F$, for $W, V \in \mathcal{C}$, for every compact $\mathcal{C} \subset \mathcal{V}$. Thus M_2 is locally Lipschitz on $\mathcal{V}$. Since $P_N(Y)$ is smooth along $\mathcal{M}$, the zero extension $P_N(Y)M_2(Y)P_N(Y)$ of $\Sigma_N(Y)$ is locally Lipschitz as well. Moreover, stationarity and unbiasedness give $\mathbb{E}[g(x;Y)] = \nabla L(Y) = 0$. Together with the sample-gradient normality established above, this identifies $M_2(Y)$ as the ambient representation of the normal noise covariance.

Finally, the proof of Lemma D.4 shows that H_N is a smooth family of normal-space operators along $\mathcal{M}$. Its normal covariant derivative is therefore smooth. The defining identity

$$\langle \mathcal{B}_Y[n_1, n_2], \tau\rangle_F = \langle n_1, (\nabla^\perp_\tau H_N)(Y) n_2\rangle_F$$

then shows that $\mathcal{B}_Y$ is smooth as well. Since $P_N(Y)$ is smooth, the ambient extension $(u,v) \mapsto \mathcal{B}_Y[P_N(Y)u, P_N(Y)v]$ depends smoothly (and hence locally lipschitz continuously) on Y.

Together with the preceding steps, these estimates verify the asserted assumptions on the chosen stable neighborhood and tubular neighborhood. □

D.5.2 Auxiliary Lemmas for Section D.4

Lemma D.6 *For any $\delta N \in N_W\mathcal{M}_{\mathrm{ReLU}}$, $W \in \mathcal{M}$,*

$$H_N[\delta N, \delta N] = \sum_{\alpha \in \{S,K,B,C\}} H_\alpha[\delta N_\alpha, \delta N_\alpha],$$

where

$$H_S[\delta N_S, \delta N_S] = 2\rho^2\lambda_S\|S\|_F^2$$
$$H_B[\delta N_B, \delta N_B] = \rho^2\|(BD_y + D_yB^\top)y\|_F^2,$$
$$H_K[\delta N_K, \delta N_K] = (2-\rho^2)(\lambda_s - \lambda_n)\|K\|_F^2,$$
$$H_C[\delta N_C, \delta N_C] = \rho^2(\mathbb{E}\|C|y|\|_F^2 + \lambda_n\|C\|_F^2) + 2(\lambda - \lambda_n)\|C\|_F^2.$$

All the cross terms vanish, thus $H_N = \bigoplus_{\alpha\in\{S,K,B,C\}} H_\alpha \mid_{N_\alpha}$. *Furthermore,*

$$J_S[S] = 2\rho Q_r RSy,\ J_K[K] = \rho(Q_\perp Ky + Q_r RK^\top\varepsilon),$$
$$J_B[B] = \rho Q_r R(B|y| + D_yB^\top y),\ J_C[C] = \rho(Q_\perp C|y| + Q_r RD_yC^\top\varepsilon).$$

For any $\delta\bar{N} \in \mathcal{M}_{\text{lin}}$, $H_N^{\text{lin}}[\delta\bar{N}, \delta\bar{N}] = H_S^{\text{lin}}[\delta\bar{N}_S, \delta\bar{N}_S] + H_K^{\text{lin}}[\delta\bar{N}_K, \delta\bar{N}_K]$, *where* $H_S^{\text{lin}}[\delta\bar{N}_S, \delta\bar{N}_S] = 4\rho^2\lambda_S\|S\|_F^2$, $H_K^{\text{lin}}[\delta\bar{N}_K, \delta\bar{N}_K] = (2-\rho^2)(\lambda_s - \lambda_n)\|K\|_F^2$.

(Proof of Lemma D.6). Suppose that the expressions for $J_\alpha[\alpha]$ are true. Using Eq. (146), we begin by showing that all cross-terms in the first Gauss–Newton block vanish. First, $\mathbb{E}\langle J_S[S], J_K[K]\rangle = \mathbb{E}\langle J_S[S], J_C[C]\rangle = 0$, using the fact that $Q_r \perp Q_\perp$, and that ε is independent from y and is centered. For $\langle J_S[S], J_B[B]\rangle$, note that, due to the symmetry of y, $J_B[B](y)$ is an even function and $J_S[S](y)$ is an odd function. Hence, the expectation vanishes. For $\langle J_K[K], J_B[B]\rangle$, due to $Q_\perp \perp Q_r$, and the centered ε, the terms vanish. For $\langle J_K[K], J_C[C]\rangle$, $\langle J_K[K], J_C[C]\rangle = \rho^2 y^\top K^\top C|y| + \rho^2\varepsilon^\top KD_yC^\top\varepsilon$. This is an odd function in terms of y, thus this term also vanishes. Lastly, B and C blocks vanishes, because of the centered ε which is independent of y.

Now we work on the second term of Eq. (146). First, we denote the normal perturbation $\delta N = [H+P, -H+P]$, where $H = Q_rRS + Q_\perp K$ and $P = Q_rB + Q_\perp C$. For $h_i := Q_rRSe_i + Q_\perp Ke_i$ and $p_i := Q_rBe_i + Q_\perp Ce_i$, we may write each i-th perturbed weight pair $w_i(t), w_{i+r}(t)$ as $w_i(t) = u_i + t(h_i + p_i)$, $w_{i+r}(t) = -u_i + t(-h_i + p_i)$. We denote $\delta w_i = h_i + p_i$ and $\delta w_{i+r} = -h_i + p_i$. Then, we have that

$$\begin{aligned}\nabla_W^2 f(x;W)[\delta N, \delta N] &= \frac{d^2}{dt^2} f(x; W + t\delta N) \mid_{t=0} \\ &= \sum_{i=1}^{2r} \frac{d^2}{dt^2}(w_i + t\delta w_i)\phi((w_i + t\delta w_i)^\top x) \mid_{t=0} \\ &= 2\sum_{i=1}^{2r} \delta w_i \delta w_i^\top x 1_{w_i^\top x > 0}.\end{aligned}$$

For the pair $(i, i+r)$, the contribution is

$$\begin{aligned}&2(h_i + p_i)(h_i + p_i)^\top x 1_{w_i^\top x>0} + 2(-h_i + p_i)(-h_i + p_i)^\top x 1_{w_i^\top x>0} \\ &= 2\left(h_ih_i^\top + p_ip_i^\top + \text{sgn}(w_i^\top x)(h_ip_i^\top + p_ih_i^\top)\right)x.\end{aligned}$$

Thus, we may write

$$\nabla_W^2 f(x;W)[\delta N, \delta N] = 2\sum_{i=1}^{r}\left(h_ih_i^\top + p_ip_i^\top + \text{sgn}(w_i^\top x)(h_ip_i^\top + p_ih_i^\top)\right)x. \tag{168}$$

Since $r(x;W) = (1-\rho^2)Q_rz + Q_\perp\varepsilon$, we split into two parts. First, we consider the inner product with $Q_\perp\varepsilon$.

$$\begin{aligned}\mathbb{E}\langle Q_\perp\varepsilon, \nabla_W^2 f(x;W)[\delta N, \delta N]\rangle &= 2\sum_{i=1}^{2r}\mathbb{E}[\varepsilon^\top(Ke_ie_i^\top K^\top + Ce_ie_i^\top C^\top)\varepsilon \\ &\quad + \text{sgn}(w_i^\top x)\varepsilon^\top(Ke_ie_i^\top A^\top + Ce_ie_i^\top C^\top)\varepsilon] \\ &= 2\lambda_n(\|K\|_F^2 + \|C\|_F^2).\end{aligned}$$

Second,

$$\begin{aligned}
&\mathbb{E}\langle Q_r z, \nabla_W^2 f(x;W)[\delta N, \delta N]\rangle \\
&= 2\sum_{i=1}^{r} \mathbb{E}[y^\top S_{\cdot,i} S_{\cdot,i}^\top y + y^\top S_{\cdot,i} K_{\cdot,i}^\top \varepsilon + y^\top R^\top B_{\cdot,i} C_{\cdot,i}^\top \varepsilon + y^\top R^\top B_{\cdot,i} B_{\cdot,i}^\top R y + \\
&\quad \operatorname{sgn}(y_i)(y^\top R^\top B_{\cdot,i} S_{\cdot,i}^\top y + y^\top R^\top B_{\cdot,i} K_{\cdot,i}^\top \varepsilon + y^\top S_{\cdot,i} C_{\cdot,i}^\top \varepsilon + y^\top S_{\cdot,i} B_{\cdot,i}^\top R y] \\
&= 2E[yy^\top](\|S\|_F^2 + \|B\|_F^2) + 2\sum_{i=1}^{r} \mathbb{E}[\operatorname{sgn}(y_i)(y^\top R^\top B_{\cdot,i} S_{\cdot,i}^\top y + y^\top S_{\cdot,i} B_{\cdot,i}^\top R y)] \\
&= 2E[yy^\top](\|S\|_F^2 + \|B\|_F^2) + 2\mathbb{E}[y^\top R^\top B D_y S^\top y + y^\top S D_y B^\top R y] \\
&= 2E[yy^\top](\|S\|_F^2 + \|B\|_F^2).
\end{aligned}$$

The last equality follows from the oddness of the functions and the symmetry of y. Note that there is no cross-term in this second term. For the final regularization term in Eq. (146), the cross-terms vanish because the S, K, B, C blocks form a direct sum decomposition of the normal space. Therefore, all cross-terms vanish. The remaining terms in the Gauss-Newton block are the squared terms: $\mathbb{E}\|J_S[S]\|^2 = 4\lambda_s \rho^2 \|S\|_F^2$, $\mathbb{E}\|J_K[K]\|^2 = \rho^2(\lambda_s + \lambda_n)\|K\|_F^2$, $\mathbb{E}\|J_B[B]\|^2 = \rho^2 \mathbb{E}\|(BD_y + D_y B^\top)y\|^2$, $\mathbb{E}\|J_C[C]\|^2 = \rho^2(\mathbb{E}\|C|y|\|^2 + \lambda_n \|C\|_F^2$. For the regularization block, $\lambda\|\delta N_\alpha\|^2 = 2\lambda\|\alpha\|_F^2$, for $\alpha \in \{S, K, B, C\}$.

It remains to prove the expressions for $J_\alpha[\alpha]$. Denote $J_{w_i} := D_{w_i} f(x; W)$. Then,

$$J_{w_i}(x;W) = \sigma(w_i^\top x) I_r + \mathbf{1}\{w_i^\top x > 0\} w_i x^\top. \tag{169}$$

Evaluating at $W \in \mathcal{M}$, where $W = \rho[U, -U]$, $U = Q_r R$, $J_{w_i}(x;W) = \rho(\phi(u_i^\top x) I_r + \mathbf{1}\{u_i^\top x > 0\} u_i x^\top)$, and $J_{w_{i+r}}(x; W_*) = \rho(\phi(-u_i^\top x) I_r - \mathbf{1}\{-u_i^\top x > 0\} u_i x^\top)$.

For a pair w_i and w_{i+r}, and $e_i \in \{0,1\}^r$ where i-th component is 1 and the other components are 0, the normal perturbation can be written as $\delta w_i = Q_r R B e_i + Q_\perp C e_i + Q_\perp K + Q_r R S e_i$ and $\delta w_{i+r} = Q_r R B e_i + Q_\perp C e_i - Q_\perp K - Q_r R S e_i$. Let us first consider perturbing only the pairing perturbation B.Denote $b_i = Q_r R B e_i$. Then,

$$J_B[B] = \frac{d}{dt} f(x; W + t\delta N_B)\Big|_{t=0} = \sum_{i=1}^{r} (J_{w_i} \delta w_i + J_{w_{i+r}} \delta w_{i+r}) = \sum_{i=1}^{r} (J_{w_i} + J_{w_{i+r}}) b_i.$$

Note that

$$\begin{aligned}
J_{w_i} b_i / \rho &= \phi(u_i^\top x) b_i + u_i \mathbf{1}\{u_i^\top x > 0\}(x^\top b_i), \\
J_{w_{i+r}} b_i / \rho &= \phi(-u_i^\top x) b_i + (-u_i) \mathbf{1}\{-u_i^\top x > 0\}(x^\top b_i).
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
(J_{w_i} + J_{w_{i+r}}) b_i / \rho &= (\phi(u_i^\top x) + \phi(-u_i^\top x)) b_i + u_i(\mathbf{1}\{u_i^\top x > 0\} + \mathbf{1}\{u_i^\top x < 0\})(x^\top b_i) \\
&= |u_i^\top x| b_i + u_i \operatorname{sgn}(u_i^\top x)(x^\top U B e_i) \\
&= |u_i^\top (Q_r z + Q_\perp \varepsilon)| U B b_i + u_i \operatorname{sgn}(u_i^\top (Q_r z + Q_\perp \varepsilon))((Q_r z + Q_\perp \varepsilon)^\top U B b_i) \\
&= |y_i| U B e_i + u_i \operatorname{sgn}(y_i)(y^\top B b_i).
\end{aligned}$$

Repeating this for $i = 1, \ldots, r$ and summing over i, we have $J_B[B] = \rho U(B|y| + D_y B^\top y)$. Applying the same steps to the other blocks proves the claim. We omit the proof for $\mathcal{M}_{\text{lin}}$ to avoid repetition. □

Lemma D.7 (Tangent and normal space decomposition). *At each $W = \rho[U, -U] \in \mathcal{M}_{\text{ReLU}}$ and $\overline{W} = \rho U \in \mathcal{M}_{\text{lin}}$, we denote the tangent spaces by $T_W \mathcal{M}_{\text{ReLU}}, T_{\overline{W}} \mathcal{M}_{\text{lin}}$ and the normal spaces by $N_W \mathcal{M}_{\text{ReLU}}, N_{\overline{W}} \mathcal{M}_{\text{lin}}$, respectively. Letting $\mathfrak{o}(r)$ be the space of skew-symmetric matrices, i.e., $\mathfrak{o}(r) = \{E \in \mathbb{R}^{r \times r} \mid E^\top = -E\}$, these spaces can be decomposed as follows:*

$$\begin{aligned}
T_W \mathcal{M}_{\text{ReLU}} &= \{\rho[UH, -UH] \mid H \in \mathfrak{o}(r)\}, T_{\overline{W}} \mathcal{M}_{\text{lin}} = \{\rho U H \mid H \in \mathfrak{o}(r)\}, \\
N_W \mathcal{M}_{\text{ReLU}} &= \{[US + Q_\perp K + P, -US - Q_\perp K + P] \mid P \in \mathbb{R}^{d \times r}, K \in \mathbb{R}^{(d-r) \times r}, S = S^\top\} \\
N_{\overline{W}} \mathcal{M}_{\text{lin}} &= \{[US + Q_\perp K] \mid K \in \mathbb{R}^{(d-r) \times r}, S = S^\top\}.
\end{aligned}$$

We denote P as a pairing-breaking block and further decompose it as $P = UB + Q_\perp C$.

(Proof of Lemma D.7). We first construct tangent space, and construct the normal space induced from the Frobenius inner product.

1. Tangent space construction. Consider $\mathcal{M} = \{\rho[Q_r R, -Q_r R] \mid R \in O(r)\} \subset \mathbb{R}^{d\times 2r}$. The map $\Psi : \mathrm{O}(r) \to \mathbb{R}^{d\times 2r}$, $\Psi(R) = \rho[Q_r R, -Q_r R]$, is a smooth embedding. Indeed, it is the restriction of an injective linear map, whose inverse on its image is $[A, -A] \mapsto Q_r^\top A$. Thus $\mathcal{M}$ is a smooth embedded submanifold diffeomorphic to $\mathrm{O}(r)$.

Fix $R_0 \in \mathrm{O}(r)$ and set $U_0 := Q_r R_0$, $W_0 := \rho[U_0, -U_0]$. Let $\mathfrak{o}(r) = \left\{K \in \mathbb{R}^{r\times r} \mid K^\top = -K\right\}$, and choose a sufficiently small neighborhood $\mathcal{V}_\varepsilon = \{K \in \mathfrak{o}(r) : \|K\|_F < \varepsilon\}$. Define

$$\varphi_{R_0} : \mathcal{V}_\varepsilon \to \mathcal{M}, \qquad \varphi_{R_0}(K) = \rho[U_0 e^K, -U_0 e^K].$$

For every $K \in \mathfrak{o}(r)$, $(e^K)^\top e^K = e^{-K} e^K = I_r$, so $R_0 e^K \in \mathrm{O}(r)$ and $\varphi_{R_0}(K) \in \mathcal{M}$. The Lie-group exponential map is a local diffeomorphism at zero. Its differential satisfies $D\exp(0)[H] = \frac{d}{dt}\big|_{t=0} e^{tH} = H$, $H \in \mathfrak{o}(r)$. Consequently, for sufficiently small ε, φ_{R_0} is a smooth local parameterization of $\mathcal{M}$ around W_0. After identifying $\mathfrak{o}(r)$ with $\mathbb{R}^{r(r-1)/2}$, its inverse defines a coordinate chart. For $H \in \mathfrak{o}(r)$,

$$D\varphi_{R_0}(0)[H] = \frac{d}{dt}\Big|_{t=0} \rho[U_0 e^{tH}, -U_0 e^{tH}] = \rho[U_0 H, -U_0 H].$$

Therefore $T_{W_0}\mathcal{M} = D\varphi_{R_0}(0)[\mathfrak{o}(r)] = \left\{\rho[U_0 H, -U_0 H] : H^\top = -H\right\}$. Since $U_0^\top U_0 = I_r$, this differential is injective. Hence $\dim \mathcal{M} = \dim \mathfrak{o}(r) = \frac{r(r-1)}{2}$. Its codimension in the full ambient parameter space is $\mathrm{codim}_{\mathbb{R}^{d\times 2r}} \mathcal{M} = 2dr - \frac{r(r-1)}{2}$.

2. Normal space construction. Let $[Q_r, Q_\perp] \in \mathrm{O}(d)$, $U = Q_r R$, $R \in \mathrm{O}(r)$, and fix $W = \rho[U, -U] \in \mathcal{M}$. The normal space is the Frobenius-orthogonal complement of the tangent space $N_W\mathcal{M} := (T_W\mathcal{M})^\perp$. By the tangent-space characterization, $T_W\mathcal{M} = \left\{\rho[U\Omega, -U\Omega] \mid \Omega^\top = -\Omega\right\}$. Since $\rho \neq 0$, an ambient perturbation $[A_1, A_2] \in \mathbb{R}^{d\times 2r}$ is normal if and only if $0 = \langle [U\Omega, -U\Omega], [A_1, A_2]\rangle_F = \mathrm{Tr}\left(\Omega^\top U^\top (A_1 - A_2)\right)$ for every skew-symmetric Ω. The Frobenius-orthogonal complement of the skew-symmetric matrices is the space of symmetric matrices. Therefore

$$[A_1, A_2] \in N_W\mathcal{M} \quad \Longleftrightarrow \quad U^\top (A_1 - A_2) \text{ is symmetric.} \tag{170}$$

To obtain an explicit parameterization, define $P := \frac{A_1+A_2}{2}$, $H := \frac{A_1-A_2}{2}$. Then $[A_1, A_2] = [P, P] + [H, -H]$. The pair-breaking component $[P, P]$ is orthogonal to every tangent vector, since $\langle [P, P], [U\Omega, -U\Omega]\rangle_F = 0$. If $[A_1, A_2]$ is normal, then $S := U^\top H$ is symmetric by Eq. (170). Set also $K := Q_\perp^\top H$. Using $UU^\top + Q_\perp Q_\perp^\top = I_d$, we obtain $H = UU^\top H + Q_\perp Q_\perp^\top H = US + Q_\perp K$. Consequently, $[A_1, A_2] = [US + Q_\perp K + P, -US - Q_\perp K + P]$. Conversely, any perturbation of this form satisfies $U^\top (A_1 - A_2) = 2S$, which is symmetric, and hence is normal by Eq. (170). Therefore

$$N_W\mathcal{M} = \left\{[US + Q_\perp K + P, -US - Q_\perp K + P] \mid S \in \mathrm{Sym}(r), K \in \mathbb{R}^{(d-r)\times r}, P \in \mathbb{R}^{d\times r}\right\}. \tag{171}$$

The coordinates are uniquely determined by $P = \frac{A_1+A_2}{2}$, $S = \frac{1}{2}U^\top (A_1 - A_2)$, $K = \frac{1}{2}Q_\perp^\top (A_1 - A_2)$. In particular,

$$\begin{aligned}\dim N_W\mathcal{M} &= \frac{r(r+1)}{2} + (d-r)r + dr \\ &= 2dr - \frac{r(r-1)}{2} = \mathrm{codim}_{\mathbb{R}^{d\times 2r}} \mathcal{M}.\end{aligned}$$

Finally, the arbitrary pair-breaking matrix P decomposes uniquely as $P = UB + Q_\perp C$, $B = U^\top P$, $C = Q_\perp^\top P$. Thus $N_W\mathcal{M} = N_S(W) \oplus N_K(W) \oplus N_B(W) \oplus N_C(W)$, where

$$\begin{aligned}N_S(W) &= \{[US, -US] \mid S \in \mathrm{Sym}(r)\}, \\ N_K(W) &= \{[Q_\perp K, -Q_\perp K] \mid K \in \mathbb{R}^{(d-r)\times r}\}, \\ N_B(W) &= \{[UB, UB] \mid B \in \mathbb{R}^{r\times r}\}, \\ N_C(W) &= \{[Q_\perp C, Q_\perp C] \mid C \in \mathbb{R}^{(d-r)\times r}\}.\end{aligned}$$

These four subspaces are mutually orthogonal: pair-preserving and pair-breaking perturbations are orthogonal by construction, while the signal-valued and orthogonal-valued components are orthogonal because $U^\top Q_\perp = 0$. Consequently,

$$T_W\mathbb{R}^{d\times 2r} = T_W\mathcal{M} \oplus N_S(W) \oplus N_K(W) \oplus N_B(W) \oplus N_C(W).$$

The similar but simpler steps can be applied to $\mathcal{M}_{\text{lin}}$, and we omit the proof to avoid repetition. □

D.5.3 Proof of Lemma 4.3

Let $v(s) \subset \mathcal{M}$ be a smooth curve satisfying $v(0) = v$ and $\dot{v}(0) = \tau$, and let $\mathsf{P}_s : N_v\mathcal{M} \to N_{v(s)}\mathcal{M}$ denote parallel transport in the normal bundle. By the proof in Section D.2 and the identity

$$(\nabla^\perp_\tau H_\alpha)(v) = \left.\frac{d}{ds}\right|_{s=0} \mathsf{P}_s^{-1} H_\alpha(v(s))\mathsf{P}_s,$$

we obtain

$$\begin{aligned}
\langle b_\alpha(v), \tau\rangle &= \operatorname{Tr}\left[(\nabla^\perp_\tau H_\alpha)(v)\Pi_\alpha(v)\right] \\
&= \mathbb{E}_{q_\alpha\sim\mathcal{N}(0,\Pi_\alpha(v))}\left\langle q_\alpha, (\nabla^\perp_\tau H_\alpha)(v) q_\alpha\right\rangle \\
&= \left.\frac{d}{ds}\right|_{s=0} \mathbb{E}_{q_\alpha\sim\mathcal{N}(0,\Pi_\alpha(v))}\left\langle \mathsf{P}_s q_\alpha, H_\alpha(v(s))\mathsf{P}_s q_\alpha\right\rangle \\
&= \left.\frac{d}{ds}\right|_{s=0} \mathbb{E}_{q_\alpha\sim\mathcal{N}(0,\Pi_\alpha(v))}\left\langle q_\alpha(s), H_\alpha(v(s)) q_\alpha(s)\right\rangle, \qquad (172)
\end{aligned}$$

where $q_\alpha(s) := \mathsf{P}_s q_\alpha$. Throughout this calculation, the Gaussian law $\mathcal{N}(0, \Pi_\alpha(v))$ is held fixed at the initial point v.

The exchange of expectation and differentiation is justified by the bounded convergence using Assumption P2 and the finite second moment of q_α. Indeed, the C^1 regularity of H_α along $\mathcal{M}$ implies that, on a sufficiently small compact interval,

$$\left|\frac{d}{ds}\left\langle \mathsf{P}_s q_\alpha, H_\alpha(v(s))\mathsf{P}_s q_\alpha\right\rangle\right| \le C\|q_\alpha\|^2.$$

Moreover, $\mathbb{E}\|q_\alpha\|^2 = \operatorname{Tr}(\Pi_\alpha(v)) \le \operatorname{Tr}(\Pi_N(v))$. Using the integral representation $\Pi_N(v) = \int_0^\infty e^{-H_N(v)t}\Sigma_N(v)e^{-H_N(v)t}\,\mathrm{d}t$, the uniform normal coercivity $H_N(v) \succeq c_* I$, and the moment assumption S1, we obtain

$$\operatorname{Tr}(\Pi_N(v)) \le \int_0^\infty e^{-2c_* t}\operatorname{Tr}(\Sigma_N(v))\,\mathrm{d}t = \frac{\operatorname{Tr}(\Sigma_N(v))}{2c_*} < \infty.$$

Before working on *Eq.* (172), note that for $\alpha \in \{K, B, C\}$, parallel transport leaves q_α constant in ambient coordinates, meaning that $q_\alpha(s) = q_\alpha(0) = q_\alpha$ for all $s \ge 0$.This follows because the K, B, and C-blocks are fixed ambient subspaces. For example, choosing a fixed orthonormal basis $Q_\perp$ (we can always rotate to make it fixed), we have $N_K(v(s)) = \left\{[Q_\perp K(s), -Q_\perp K(s)] \mid K \in \mathbb{R}^{(d-r)\times r}\right\}$. For $q_K(s) = [Q_\perp K(s), -Q_\perp K(s)] \in N_K(v(s))$,

$$\dot{q}_K(s) = [Q_\perp \dot{K}(s), -Q_\perp \dot{K}(s)] \in N_K(v(s)) \subset N_{v(s)}\mathcal{M}.$$

Thus its ambient derivative is already normal, so the parallel transport condition becomes $\nabla^\perp_s q_K(s) = \dot{q}_K(s) = 0$. Consequently, constant fields in this block are parallel, and uniqueness of parallel transport implies that $q_K(s)$ is constant. The same argument applies to the B- and C-blocks. Therefore, for $\alpha \in \{K, B, C\}$, Eq. (172) reduces to

$$\langle b_\alpha(v), \tau\rangle = \left.\frac{d}{ds}\right|_{s=0} \mathbb{E}_{q_\alpha\sim\mathcal{N}(0,\Pi_\alpha(v))}\left\langle q_\alpha, H_\alpha(v(s)) q_\alpha\right\rangle,$$

where the Gaussian law is held fixed at $v = v(0)$.

We next examine the quadratic Hessian forms for the B- and C-blocks (see Eq. (146) for the form of the Hessian)), beginning with the B-block and the regularization term. For $q_B = h([UB, UB])$, the regularization contribution satisfies

$$\lambda\langle q_B, q_B\rangle = \lambda\,\langle [UB, UB], [UB, UB]\rangle_F = 2\lambda\|B\|_F^2.$$

Similarly, the calculation in Lemma D.6 gives the residual contribution

$$\mathbb{E}\left\langle r(x; W), \nabla^2 f(x; W)[h^{-1}(q_B), h^{-1}(q_B)]\right\rangle = c_B\|B\|_F^2,$$

where c_B is independent of s. Both contributions are therefore constant along the curve, so their derivatives vanish. The Gauss-Newton term is consequently the only remaining contribution to the B-block drift. By Eq. (146), this term is $\mathbb{E}\|J_B[B]\|^2 = \rho^2\mathbb{E}\left\|\left(BD_y + D_yB^\top\right)y\right\|^2$, Hence $\langle b_B(v), \tau\rangle = \rho^2\,\frac{d}{ds}\big|_{s=0}\,\mathbb{E}_{q_B\sim\mathcal{N}(0,\Pi_B)}\mathbb{E}\left\|\left(BD_y + D_yB^\top\right)y\right\|^2$.

For the last C-block, the regularizer and the residual terms vanish similarly, and we have again the Gauss-Newton term with non-zero gradient

$$\begin{aligned}\mathbb{E}\|J_C[C]\|^2 &= \rho^2\mathbb{E}[|y|^\top C^\top C|y| + \varepsilon^\top CC^\top\varepsilon] \\ &= \rho^2\,\mathrm{Tr}(C^\top C\mathbb{E}[|y||y|^\top]) + \rho^2\lambda_n\|C\|_F^2.\end{aligned}$$

since the second term is independent of s, we have that

$$\langle b_C(v), \tau\rangle = \rho^2\,\frac{d}{ds}\bigg|_{s=0}\,\mathbb{E}_{q_C\sim\mathcal{N}(0,\Pi_C)}\,\mathrm{Tr}(C^\top C\mathbb{E}[|y||y|^\top].$$

We now apply this to the $r = 2$ setting. Here, we parameterize y by θ, i.e., $y(\theta) = R^\top(\theta)z$. The rest of the proof proceeds in two steps. First, we derive the expression for the effective drift $b_{\mathrm{ReLU}}(v(\theta)) = b_B(v(\theta)) + b_C(v(\theta))$, which governs an angular ODE. Then, we explain how the normal fluctuations interact within this expression and find the stable equilibria.

1.1 : B-block. With $B = \begin{bmatrix}\alpha & \beta\\ \gamma & \delta\end{bmatrix}$, we have that

$$(BD_y + D_yB^\top)y = \begin{bmatrix}2\alpha|y_1| + \beta|y_2| + \gamma s_1y_2\\ 2\delta|y_2| + \gamma|y_1| + \beta s_2y_1\end{bmatrix},$$

where we denoted $s_i = \mathrm{sgn}(y_i)$ for $i \in \{1, 2\}$. Expanding the square term and taking the expectation, we have that

$$\begin{aligned}\mathbb{E}\|(BD_y + D_yB^\top)y\|^2 &= \mathbb{E}[4\alpha^2y_1^2 + 4\gamma^2y_2^2 + (\beta^2+\gamma^2)(y_1^2+y_2^2) + 4|y_1y_2|(\alpha\beta+\gamma\delta)+ \\ &\quad 2s_1s_2\beta\gamma(y_1^2+y_2^2) + 4y_1y_2(\alpha\gamma+\beta\gamma)] \\ &= \lambda_s(4\alpha^2 + 4\delta^2 + 2\beta^2 + 2\gamma^2) + 4(\alpha\beta+\gamma\delta)\mathbb{E}[|y_1||y_2|]+ \\ &\quad 2\beta\gamma\mathbb{E}[s_1s_2(y_1^2+y_2^2)] + 4(\alpha\gamma+\beta\gamma)\mathbb{E}[y_1y_2] \\ &= \lambda_s(4\alpha^2 + 4\delta^2 + 2\beta^2 + 2\gamma^2) + 4(\alpha\beta+\gamma\delta)\mathbb{E}[|y_1||y_2|].\end{aligned}$$

The last equality satisfies, since, for $f(y_1, y_2) := s_1s_2(y_1^2+y_2^2)$, $f(y_1, y_2) =_d f(-y_1, y_2)$. This means that $\mathbb{E}[s_1s_2(y_1^2+y_2^2)] = \mathbb{E}[-s_1s_2(y_1^2+y_2^2)]$, thus it follows that the term vanishes. The other term $\mathbb{E}[y_1y_2]$ vanishes with the same reason that this is symmetric ($y_1y_2 =_d -y_1y_2$).

Consequently, writing $u_y(\theta) = \mathbb{E}|y_1(\theta)y_2(\theta)|$,

$$\begin{aligned}\langle b_B(v(\theta_0)), \tau_{\theta_0}\rangle &= \rho^2\,\frac{d}{d\theta}\bigg|_{\theta=\theta_0}\,\mathbb{E}_{q_B,x}\left\|\left(B(\theta)D_{y(\theta)} + D_{y(\theta)}B(\theta)^\top\right)y(\theta)\right\|^2 \\ &= 4\rho^2\mathbb{E}_{q_B}[\alpha\beta+\gamma\delta]\,u_y'(\theta_0),\end{aligned}$$

where $\tau_\theta := \partial_\theta v(\theta)$. Define u on $\mathcal{M}$ by $u(v(\theta)) := u_y(\theta) = \mathbb{E}|y_1(\theta)y_2(\theta)|$. By the definition of the Riemannian gradient, $u_y'(\theta) = Du(v(\theta))[\tau_\theta] = \langle\mathrm{grad}_{\mathcal{M}}\,u(v(\theta)), \tau_\theta\rangle$. Since $T_{v(\theta)}\mathcal{M}$ is spanned by τ_θ,

$$b_B(v(\theta)) = 4\rho^2\mathbb{E}_{q_B\sim\mathcal{N}(0,\Pi_B)}[\alpha\beta+\gamma\delta]\;\mathrm{grad}_{\mathcal{M}}\,u(v(\theta)).$$

Since the tangent space is one-dimensional, $\operatorname{grad}_{\mathcal{M}} u(v(\theta)) = c(\theta)\partial_\theta v$ for some scalar $c(\theta)$. Thus,

$$u_y'(\theta) = c(\theta)\|\partial_\theta v\|^2, \qquad \operatorname{grad}_{\mathcal{M}} u(v(\theta)) = \frac{u_y'(\theta)}{\|\partial_\theta v\|^2}\partial_\theta v. \tag{173}$$

For $v(\theta) = \rho h([Q_r R(\theta), -Q_r R(\theta)])$, $\partial_\theta v = h([\rho Q_r R(\theta) J, -\rho Q_r R(\theta) J])$, where $J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Since Q_r has orthonormal columns, $R(\theta)$ is orthogonal, and $\|J\|_F^2 = 2$, we obtain $\|\partial_\theta v\|^2 = 2\rho^2\|Q_r R(\theta) J\|_F^2 = 4\rho^2$. Consequently, $\operatorname{grad}_{\mathcal{M}} u(v(\theta)) = \frac{u_y'(\theta)}{4\rho^2}\partial_\theta v$.

We next compute $u_y'(\theta)$. For $y(\theta) = R(\theta)^\top z$, we have $y_1(\theta) = \cos\theta\, z_1 + \sin\theta\, z_2$, $y_2(\theta) = -\sin\theta\, z_1 + \cos\theta\, z_2$. Hence $y_1' = y_2$, $y_2' = -y_1$, and $\frac{d}{d\theta}(y_1 y_2) = y_2^2 - y_1^2$. For each fixed θ, the nondegenerate Gaussian-mixture distribution gives $\mathbb{P}(y_1 y_2 = 0) = 0$. Moreover, since $\left|\frac{d}{d\theta}(y_1 y_2)\right| \le y_1^2 + y_2^2 = \|z\|^2$ and $\mathbb{E}\|z\|^2 < \infty$, dominated convergence therefore yields $u_y'(\theta) = \mathbb{E}\left[\operatorname{sgn}(y_1 y_2)(y_2^2 - y_1^2)\right]$. Therefore,

$$b_B(v(\theta)) = \mathbb{E}_{q_B \sim \mathcal{N}(0,\Pi_B)}[\alpha\beta + \gamma\delta] u_y'(\theta)\partial_\theta v.$$

1.2: C-block. For $r = 2$, using the same parameterization as the B-block, $\mathbb{E}|yy|^\top = \begin{bmatrix} \lambda_s & u_y(\theta) \\ u_y(\theta) & \lambda_s \end{bmatrix}$, since $\mathbb{E}[y_1^2] = \mathbb{E}[y_2^2] = \lambda_s$, and $u_y(\theta) := \mathbb{E}|y_1 y_2|$. Now for $C = [c_1\, c_2] \in \mathbb{R}^{(d-2)\times 2}$, $C^\top C = \begin{bmatrix} \|c_1\|^2 & c_1^\top c_2 \\ c_1^\top c_2 & \|c_2\|^2 \end{bmatrix}$. Thus, $\operatorname{Tr}(C^\top C \mathbb{E}|y||y|^\top) = \lambda_s\|C\|_F^2 + 2u_y(\theta)c_1^\top c_2$. Since the first term has no dependence on θ, the derivative is determined by the second term $2\rho^2 u_y(\theta)c_1^\top c_2$. Consequently,

$$\begin{aligned}\langle b_C(v(\theta_0)), \tau_{\theta_0}\rangle &= \rho^2 \left.\frac{d}{d\theta}\right|_{\theta=\theta_0} \mathbb{E}_{q_C} \operatorname{Tr}(C^\top C \mathbb{E}|yy^\top|) \\ &= 2\rho^2 \mathbb{E}_{q_C}[c_1^\top c_2] u_y(\theta_0)\end{aligned}$$

Using the identity derived previously in the B-block for Eq. (173), we have that

$$b_C(v(\theta_0)) = 2\rho^2 \mathbb{E}_{q_C}[c_1^T c_2] \operatorname{grad}_{\mathcal{M}} u(v(\theta)).$$

Combining the two terms, and notating $\kappa_B := \frac{1}{2}\mathbb{E}_{q_B}[b_1^\top b_2]$ ($b_1 := [\alpha, \gamma]^\top, b_2 := [\beta, \delta]^\top$) and $\kappa_C := \frac{1}{4}\mathbb{E}_{q_C}[c_1^\top c_2]$, we have that $\dot{v}(t) = -(\kappa_B + \kappa_C)u_y'(\theta_t)\partial_\theta v(\theta_t)$. On the other hand, by chain rule, we have that $\dot{v}(t) = \partial_\theta v(\theta_t)\dot{\theta}_t$. Comparing the two terms, the effective dynamics reduce to the scalar angular ODE

$$\mathrm{d}\theta_t = -(\kappa_B + \kappa_C)u_y'(\theta_t)\mathrm{d}t.$$

Before proceeding to Step 2, fix $v \in \mathcal{M}$ and define the linear maps

$$\begin{aligned} E_B : \mathbb{R}^{r^2} &\to \mathbb{R}^{2dr}, \qquad h(B) \mapsto h([UB, UB]), \\ E_C : \mathbb{R}^{r(d-r)} &\to \mathbb{R}^{2dr}, \qquad h(C) \mapsto h([Q_\perp C, Q_\perp C]), \end{aligned}$$

where h denotes row-major vectorization for the relevant matrix size. Let E_B^* and E_C^* denote their adjoints. For $A = [A_1, A_2] \in \mathbb{R}^{d\times 2r}$,

$$\begin{aligned}\langle E_B h(B), h(A)\rangle &= \langle [UB, UB], [A_1, A_2]\rangle_F \\ &= \langle B, U^\top(A_1 + A_2)\rangle_F.\end{aligned}$$

Thus $E_B^* h(A) = h(U^\top(A_1 + A_2))$, and $E_B^* E_B h(B) = 2h(B)$ because $U^\top U = I_r$. Similarly, $E_C^* h(A) = h(Q_\perp^\top(A_1 + A_2))$ and $E_C^* E_C = 2I$. These identities allow us to express the block Lyapunov equations in their reduced matrix coordinates.

Lemma D.8 (Coordinate Lyapunov equations). *For $\alpha \in \{B, C\}$, define*

$$\hat{H}_\alpha := \frac{1}{2}E_\alpha^* H_\alpha E_\alpha, \qquad \hat{\Sigma}_\alpha := \frac{1}{4}E_\alpha^* \Sigma_\alpha E_\alpha, \qquad \hat{\Pi}_\alpha := \frac{1}{4}E_\alpha^* \Pi_\alpha E_\alpha.$$

Then $\hat{\Pi}_\alpha$ satisfies the coordinate Lyapunov equation $\hat{H}_\alpha\hat{\Pi}_\alpha + \hat{\Pi}_\alpha\hat{H}_\alpha = \hat{\Sigma}_\alpha$.

(Proof of Lemma D.8). Throughout the proof, the base point v is held fixed. The ambient block Lyapunov equation characterizes the stationary covariance Π_α of the OU process

$$\mathrm{d}q_\alpha(t) = -H_\alpha q_\alpha(t)\,\mathrm{d}t + \sigma_\alpha\,\mathrm{d}B_t, \qquad \sigma_\alpha\sigma_\alpha^\top = \Sigma_\alpha. \tag{174}$$

Take this process in stationarity on $N_\alpha(v)$ and write $q_\alpha(t) = E_\alpha z_\alpha(t)$, where $z_B(t) \in \mathbb{R}^{r^2}$ and $z_C(t) \in \mathbb{R}^{r(d-r)}$. Since $E_\alpha^* E_\alpha = 2I$, $z_\alpha(t) = \frac{1}{2}E_\alpha^* q_\alpha(t)$. Applying $\frac{1}{2}E_\alpha^*$ to Eq. (174) gives

$$\mathrm{d}z_\alpha(t) = -\hat{H}_\alpha z_\alpha(t)\,\mathrm{d}t + \frac{1}{2}E_\alpha^*\sigma_\alpha\,\mathrm{d}B_t.$$

Its noise covariance is $\frac{1}{4}E_\alpha^*\Sigma_\alpha E_\alpha = \hat{\Sigma}_\alpha$, and its stationary covariance is

$$\mathbb{E}[z_\alpha(t)z_\alpha(t)^\top] = \frac{1}{4}E_\alpha^*\mathbb{E}[q_\alpha(t)q_\alpha(t)^\top]E_\alpha = \frac{1}{4}E_\alpha^*\Pi_\alpha E_\alpha = \hat{\Pi}_\alpha.$$

Consequently, $\hat{H}_\alpha\hat{\Pi}_\alpha + \hat{\Pi}_\alpha\hat{H}_\alpha = \hat{\Sigma}_\alpha$, as claimed. □

2: Effect of averaged normal fluctuations and stable equilibria The above analysis show that, when $\kappa_B + \kappa_C > 0$, the effective angular dynamics decrease $u_y(\theta)$ along nonstationary trajectories. Lemma D.9 provides the existence of σ_* which establishes this positivity for every $\sigma \in (0, \sigma_*]$. Then, θ_t moves towards the direction where it minimizes the energy $u_y(\theta)$. Lemma D.10 then identifies $\theta = 0$ and $\theta = \pi/2$ as the stable equilibria. Whenever $\theta_0 \neq \pi/4$, the trajectory converges to one of these minimizing orientations, and its angular drift tends to zero. □

Lemma D.9 (Sign of κ_B and κ_C). *Let $\sigma > 0$ denote the noise scale for the gaussian mixture components. Then, there exists $\sigma_* > 0$, such that for all $\sigma \in (0, \sigma_*]$, $\mathrm{sgn}(\kappa_B) > 0$. For κ_C, we have the following simplication $\kappa_C = \frac{1}{4}(d-r)(\hat{\Pi}_C)_{12}$. Under the standing condition $\lambda \in (\lambda_n, \lambda_s)$, it satisfies that $(\hat{\Pi}_C)_{12} > 0$, thus $\mathrm{sgn}(\kappa_C) > 0$.*

(Proof of Lemma D.9). Our goal is to find a condition to establish $\kappa_B + \kappa_C > 0$, in order to claim that the limiting flow moves in the direction where it minimizes $u_y(\theta)$. We have the following Lyapunov equation on the coordinates B, C: $\hat{H}_\alpha\hat{\Pi}_\alpha + \hat{\Pi}_\alpha\hat{H}_\alpha = \hat{\Sigma}_\alpha, \quad \alpha \in \{B, C\}$.

1. Sign of κ_B. Since $\hat{H}_B = \frac{1}{2}E_B^* H_B E_B$ and $q_B = E_B h(B)$, we have that

$$q_B^\top H_B q_B = h(B)^\top E_B^* H_B E_B h(B) = 2h(B)^\top \hat{H}_B h(B).$$

For $r = 2$,

$$H_B[q_B, q_B] = \rho^2(\lambda_s(4\alpha^2 + 2\beta^2 + 2\gamma^2 + 4\delta^2) + 4u_y(\theta)(\alpha\beta + \gamma\delta)) + 2(\lambda - (1-\rho^2)\lambda_s)\|B\|_F^2.$$

Since $\lambda = (1-\rho^2)\lambda_s$, the last term vanishes. With $h(B) = [\alpha\,\beta\,\gamma\,\delta]^\top$, we have that

$$\hat{H}_B = \frac{\rho^2}{2}\begin{bmatrix} 4\lambda_s & 2u_y & 0 & 0 \\ 2u_y & 2\lambda_s & 0 & 0 \\ 0 & 0 & 2\lambda_s & 2u_y \\ 0 & 0 & 2u_y & 4\lambda_s \end{bmatrix}. \tag{175}$$

Now we obtain $\hat{\Sigma}_B = \frac{1}{4}E_B^*\Sigma_B E_B$. Note that $\Sigma_B = \mathbb{E}[P_B g(P_B g)^\top]$, and from the calculation in Lemma D.12, $P_B g = [\rho(1-\rho^2)Uy|y|^\top, \rho(1-\rho^2)Uy|y|^\top]$. Thus, we have that $E_B^*(P_B g) = 2\rho(1-\rho^2)y|y|^\top$. Therefore, $\hat{\Sigma}_B = \rho^2(1-\rho^2)^2\mathbb{E}[y|y|^\top \otimes y|y|^\top]$. For $r = 2$, this can be written as

$$\hat{\Sigma}_B = \rho^2(1-\rho^2)^2\mathbb{E}\begin{bmatrix} y_1|y_1| \\ y_1|y_2| \\ y_2|y_1| \\ y_2|y_2| \end{bmatrix}\begin{bmatrix} y_1|y_1| & y_1|y_2| & y_2|y_1| & y_2|y_2| \end{bmatrix}^\top$$

$$= \rho^2(1-\rho^2)^2\mathbb{E}\begin{bmatrix} y_1^4 & |y_1|^3|y_2| & 0 & 0 \\ |y_1|^3|y_2| & y_1^2y_2^2 & 0 & 0 \\ 0 & 0 & y_1^2y_2^2 & |y_2|^3|y_1| \\ 0 & 0 & |y_2|^3|y_1| & y_2^4 \end{bmatrix}.$$

Note that terms such as $\mathbb{E}[y_1^3 y_2]$, $\mathbb{E}[y_1^3 y_2]$, and $\mathbb{E}[y_1^2 y_2^2 \operatorname{sgn}(y_1 y_2)]$ vanish, due to the symmetry of y_1 and y_2. Notice that $\hat{H}_B$ and $\hat{\Sigma}_B$ have block diagonal structure. With the following notation

$$\hat{H}_B = \begin{bmatrix} \hat{H}^B_{\alpha\beta} & 0 \\ 0 & \hat{H}^B_{\gamma\delta} \end{bmatrix}, \quad \hat{\Sigma}_B = \begin{bmatrix} \hat{\Sigma}^B_{\alpha\beta} & 0 \\ 0 & \hat{\Sigma}^B_{\gamma\delta} \end{bmatrix}, \quad \hat{\Pi}_B = \begin{bmatrix} \hat{\Pi}^B_{\alpha\beta} & 0 \\ 0 & \hat{\Pi}^B_{\gamma\delta} \end{bmatrix},$$

we may split into two Lyapunov equations with the reduced 2×2 matrices

$$\hat{\Pi}^B_{\alpha\beta}\hat{H}^B_{\alpha\beta} + \hat{H}^B_{\alpha\beta}\hat{\Pi}^B_{\alpha\beta} = \hat{\Sigma}^B_{\alpha\beta}, \quad \hat{\Pi}^B_{\gamma\delta}\hat{H}^B_{\gamma\delta} + \hat{H}^B_{\gamma\delta}\hat{\Pi}^B_{\gamma\delta} = \hat{\Sigma}^B_{\gamma\delta}. \tag{176}$$

Recall that our goal is to calculate $\kappa_B = \frac{1}{2}\mathbb{E}[\alpha\beta + \gamma\delta]$, and notice that since $\hat{\Pi}_B = \mathbb{E}[h(B)h(B)^\top] = \mathbb{E}\begin{bmatrix} \alpha^2 & \alpha\beta & 0 & 0 \\ \alpha\beta & \beta^2 & 0 & 0 \\ 0 & 0 & \gamma^2 & \gamma\delta \\ 0 & 0 & \gamma\delta & \delta^2 \end{bmatrix}$, it is equivalent to compute $(\hat{\Pi}^B_{\alpha\beta})_{12}$ and $(\hat{\Pi}^B_{\gamma\delta})_{12}$. To calculate this, observe a general 2×2 Lyapunov equation, where $H = \begin{bmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{bmatrix}$, $\Sigma = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix}$, and $\Pi = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}$. Solving $H\Pi + \Pi H = \Sigma$ with respect to p_{12} gives that

$$p_{12} = \frac{s_{12} - h_{12}\left(\frac{s_{11}}{2h_{11}} + \frac{s_{22}}{2h_{22}}\right)}{h_{11} + h_{22} - h_{12}^2(h_{11}^{-1} + h_{22}^{-1})}. \tag{177}$$

Using the above Eq. (177), we solve the reduced Lyapunov equations (Eq. (176)). This gives

$$(\hat{\Pi}^B_{\alpha\beta})_{12} = \frac{(1-\rho^2)^2\lambda_s}{3(2\lambda_s^2 - u_y^2)} N_{\alpha\beta}, \quad N_{\alpha\beta} = \mathbb{E}|y_1|^3|y_2| - u_y\left(\frac{\mathbb{E}[y_1^2 y_2^2]}{2\lambda_s} + \frac{\mathbb{E}[y_1^4]}{4\lambda_s}\right), \tag{178}$$

$$(\hat{\Pi}^B_{\gamma\delta})_{12} = \frac{(1-\rho^2)^2\lambda_s}{3(2\lambda_s^2 - u_y^2)} N_{\gamma\delta}, \quad N_{\gamma\delta} = \mathbb{E}|y_1||y_2|^3 - u_y\left(\frac{\mathbb{E}[y_1^2 y_2^2]}{2\lambda_s} + \frac{\mathbb{E}[y_2^4]}{4\lambda_s}\right). \tag{179}$$

Due to the symmetry of y_1 and y_2, we have that $(\hat{\Pi}^B_{\gamma\delta})_{12} = (\hat{\Pi}^B_{\alpha\beta})_{12}$. Note that since $\lambda_s > u_y$, and $(1-\rho^2) > 0$(for $\lambda > 0$), $\operatorname{sgn}(\kappa_B) = \operatorname{sgn}((\hat{\Pi}^B_{\alpha\beta})_{12}) = \operatorname{sgn}(N_{\alpha\beta})$. We invoke Lemma D.13 to conclude that there exists $\sigma_* > 0$, such that for all $\sigma \in (0, \sigma_*]$, $\operatorname{sgn}(\kappa_B) > 0$.

2. Sign of κ_C We now work on solving that $\hat{H}_C\hat{\Pi}_C + \hat{\Pi}_C\hat{H}_C = \hat{\Sigma}_C$. For the C-block coordinate $C = \begin{bmatrix} c_1 \\ \vdots \\ c_{d-r} \end{bmatrix}$, $c_i \in \mathbb{R}^{1\times 2}$ denotes a row vector. With this notation, $\kappa_C = \frac{1}{4}\sum_{i=1}^{d-r}\mathbb{E}[(c_i)_1(c_i)_2]$. The complication arises from the fact that $d - r$ can be large, meaning the corresponding Lyapunov equation is potentially massive. However, due to the fact that $\hat{\Pi}_C = I_{d-r} \otimes \tilde{\Pi}_C$, where $\tilde{\Pi}_C \in \mathbb{R}^{r\times r}$, the above $(d-r)\times(d-r)$ system can be reduced to solving a $r \times r$ Lyapunov equation. Since

$$\begin{aligned} \hat{\Pi}_C &= \mathbb{E}[h(C)h(C)^\top] \\ &= \mathbb{E}\begin{bmatrix} c_1^\top c_1 & \dots & c_1^\top c_{d-r} \\ \vdots & \ddots & \vdots \\ c_{d-r}^\top c_1 & \dots & c_{d-r}^\top c_{d-r} \end{bmatrix} \in \mathbb{R}^{r(d-r)\times r(d-r)}, \end{aligned}$$

It satisfies that $c_i^\top c_j = \delta_{ij}\tilde{\Pi}_C$. Therefore,

$$\kappa_C = \frac{1}{4}\sum_{i=1}^{d-r}\mathbb{E}[(c_i)_1(c_i)_2] = \frac{1}{4}\sum_{i=1}^{d-r}(\tilde{\Pi}_C)_{12} = \frac{d-r}{4}(\tilde{\Pi}_C)_{12}.$$

Thus, it is enough to find the expression of the reduced $\tilde{\Pi}_C$. In order to do this, we first work on the Hessian $\hat{H}_C$. The Gauss-Newton block of the Hessian is $\mathbb{E}\|J_C[C]\|^2 = \rho^2(\mathbb{E}\|C|y|\|^2 +$

$\mathbb{E}\|D_y C^\top \varepsilon\|^2$. For the first term, $\mathbb{E}\|C|y|\|^2 = \mathrm{Tr}(CMC^\top)$, where $M := \mathbb{E}[|y||y|^\top]$. Notice that $\mathrm{Tr}(CMC^\top) = h(C)(I_{d-r} \otimes M)h(C)^\top$. For the second term, since $\mathbb{E}[\varepsilon\varepsilon^\top] = \lambda_n I_{d-r}$, $\mathbb{E}\|D_y C^\top \varepsilon\|^2 = \lambda_n \|C\|_F^2$. For the rest of the quadratic Hessian terms, the residual block gives $-2\lambda_n\|C\|_F^2$, and the regularizer $2\lambda\|C\|_F^2$. Define $h_0 := 2\lambda + (\rho^2 - 2)\lambda_n$. Then, we have that

$$h_0\|C\|_F^2 = h_0 \,\mathrm{Tr}(CC^\top) = h(C)^\top (I_{d-r} \otimes h_0 I_r) h(C).$$

Thus, $H_C[q_C, nqC] = h(C)^\top (I_{d-r} \otimes (M + h_0 I_r))h(C)$. Since $q_C^\top H_C q_C = 2h(C)^\top \hat{H}_C h(C)$,

$$\hat{H}_C = \frac{1}{2} I_{d-r} \otimes (\rho^2 M + h_0 I_r). \tag{180}$$

Second, we work on the covariance $\hat{\Sigma}_C$. Since $\Sigma_C = \mathbb{E}[P_C g (P_C g)^\top]$ and $P_C g = [\frac{\rho(2-\rho^2)}{2} Q_\perp \varepsilon |y|^\top, \frac{\rho(2-\rho^2)}{2} Q_\perp \varepsilon |y|^\top]$, $E_C^* P_C g = \rho(2-\rho^2)\varepsilon|y|^\top$. Therefore, for $h_1 = \frac{1}{4}\rho^2(2-\rho^2)^2$,

$$\begin{aligned}
\hat{\Sigma}_C &= \frac{1}{4} E_C^* \Sigma_C E_C \\
&= \frac{1}{4}\mathbb{E}[\rho(2-\rho^2)\varepsilon|y|^\top \otimes \rho(2-\rho^2)\varepsilon|y|^\top] \\
&= h_1 \mathbb{E}[\varepsilon|y|^\top \otimes \varepsilon|y|^\top].
\end{aligned}$$

Note that, $\hat{\Sigma}_C = h_1 \mathbb{E}[h(\varepsilon|y|^\top) h(\varepsilon|y|^\top)^\top]$, and $h(\varepsilon|y|^\top) = \begin{bmatrix} \varepsilon_1|y| \\ \vdots \\ \varepsilon_{d-r}|y| \end{bmatrix} = \varepsilon \otimes |y|$. Hence,

$$\begin{aligned}
\mathbb{E}[h(\varepsilon|y|^\top) h(\varepsilon|y|^\top)^\top] &= \mathbb{E}[(\varepsilon \otimes |y|)(\varepsilon \otimes |y|)^\top] \\
&= \mathbb{E}[(\varepsilon \otimes |y|)(\varepsilon^\top \otimes |y|^\top)] \\
&= \mathbb{E}[\varepsilon\varepsilon^\top] \otimes \mathbb{E}|y||y|^\top \\
&= \lambda_n I_{d-r} \otimes M.
\end{aligned}$$

We now go back to solving the Lyapunov equation. First, we show that $\hat{\Pi}_C$ can be also written as $I_{d-r} \otimes \tilde{\Pi}_C$, for some $\tilde{\Pi}_C \in \mathbb{R}^{r\times r}$. With $\hat{H}_C = I_{d-r} \otimes \frac{1}{2}(\rho^2 M + h_o I_r)$ and $\Sigma_C = I_{d-r} \otimes \frac{1}{4}\lambda_n M$, we denote $\Pi_C = \begin{bmatrix} \Pi_{1,1} & \dots & \Pi_{1,d-r} \\ \vdots & \ddots & \vdots \\ \Pi_{d-r,1} & \dots & \Pi_{d-r,d-r} \end{bmatrix}$, for $\Pi_{ij} \in \mathbb{R}^{r\times r}$. From $\hat{\Pi}_C \hat{H}_C + \hat{H}_C \hat{\Pi}_C = \hat{\Sigma}_C$, we have that $(I_{d-r} \otimes \frac{1}{2}(\rho^2 M + h_o I_r))\hat{\Pi}_C + \hat{\Pi}_C (I_{d-r} \otimes \frac{1}{2}(\rho^2 M + h_o I_r)) = I_{d-r} \otimes \frac{1}{4}\lambda_n M$. We denote $\tilde{H}_C = \frac{1}{2}(\rho^2 M + h_o I_r)$ and $\tilde{\Sigma}_C = \frac{1}{4}\lambda_n M$. From this, for $i, j \in [d-r]$, we have that $\tilde{H}_C \Pi_{i,j} + \Pi_{i,j}\tilde{H}_C = \delta_{ij}\tilde{\Sigma}_C$. Since $\tilde{H}_C$ is invertible, we have that $\Pi_{i,j} = 0$, for $i \neq j$ and for $i = j$, we have that $\Pi_{1,1} = \cdots = \Pi_{d-r,d-r} =: \tilde{\Pi}_C$. Thus, we are left to solve the reduced Lyapunov equation $\tilde{H}_C \tilde{\Pi}_C + \tilde{\Pi}_C \tilde{H}_C = \tilde{\Sigma}_C$.

In the case of $r = 2$, using the Eq. (177), we have that

$$(\tilde{\Pi}_C)_{12} = \frac{1}{4} \frac{\lambda_n u_y h_0}{(h_0 + \rho^2(\lambda_s + u_y))(h_0 + \rho^2(\lambda_s - u_y))}.$$

Note that $\lambda_s > u_y$ and $u_y > 0$ almost surely. The first one is a simple application of Cauchy-Schwarz inequality, where $u_y = \mathbb{E}|y_1 y_2| \leq \sqrt{\mathbb{E}y_1^2 \mathbb{E}y_2^2} = \lambda_s$. The equality condition holds whenever $|y_1| = c'|y_2|$ for $c' > 0$, but this is a measure zero event for the Gaussian Mixture with $\sigma > 0$. $u_y = 0$ iff $y_1 = y_2 = 0$, which is again a measure zero event. For the standing assumption $\lambda_s > \lambda > \lambda_n$, the condition $\frac{2(1-\rho^2)\lambda_s}{(2-\rho^2)\lambda_n} > 1$ always satisfies, since $1 > \frac{\lambda_n}{\lambda}$ implies $\lambda_s > \frac{2\lambda_n}{2\lambda}\lambda_s > \frac{\lambda_s+\lambda}{2\lambda}\lambda_n = \frac{2-\rho^2}{2(1-\rho^2)}\lambda_n$. This makes $h_0 > 0$ and we conclude that $(\tilde{\Pi}_C)_{12} > 0$ and therefore, $\kappa_C = \frac{1}{4}(d-r)(\tilde{\Pi}_C)_{12}$ has the positive sign. □

Lemma D.10 (Minimization of $\mathbb{E}|y_1 y_2|$). *For $u_y(\theta) := \mathbb{E}|y_1 y_2|$, it satisfies that $u_y(\theta)$ increases monotonically in $\theta \in (0, \pi/4)$, and decreases monotonically in $\theta \in (\pi/4, \pi/2)$. Consequently, $\theta_0 \in (0, \pi/4)$, $\theta_t \to 0$ and $\theta_0 \in (\pi/4, \pi/2)$, $\theta_t \to \pi/2$. Moreover, $\lim_{\theta\downarrow 0} u_y'(\theta) = \lim_{\theta\uparrow\frac{\pi}{2}} u_y'(\theta) = 0$.*

(Proof of Lemma D.10). Let $z = z_0 + \sigma G$, where $z_0 \sim \mathrm{Unif}\{\pm se_1, \pm se_2\}$ and $G \sim \mathcal{N}(0, I_2)$ are independent, and write $y(\theta) = R(\theta)^\top z$. Define $h(x) := \mathbb{E}_{g\sim\mathcal{N}(0,1)}|x + \sigma g|$. Conditionally on z_0, the coordinates of $y(\theta)$ are independent. Since h is even, each mixture component gives the same product, yielding $u_y(\theta) = h(s\sin\theta)h(s\cos\theta)$. This function is $\pi/2$-periodic and satisfies $u_y(\pi/2 - \theta) = u_y(\theta)$. We show that it is strictly increasing on $(0, \pi/4)$ and strictly decreasing on $(\pi/4, \pi/2)$.

Set $a := s\sin\theta$ and $b := s\cos\theta$. Differentiating gives

$$u_y'(\theta) = bh'(a)h(b) - ah(a)h'(b) = abh(a)h(b)\big(r(a) - r(b)\big), \qquad r(x) := \frac{h'(x)}{xh(x)}.$$

For $\theta \in (0, \pi/2)$, the prefactor $abh(a)h(b)$ is strictly positive. It therefore suffices to prove that r is strictly decreasing on $(0, \infty)$.

For $\sigma > 0$, the function h is smooth and positive, with $h'(0) = 0$ and $h''(x) = \sqrt{2/\pi}\,\sigma^{-1}e^{-x^2/(2\sigma^2)}$. Since h'' is strictly decreasing on $(0, \infty)$, $h'(x) = \int_0^x h''(t)\,\mathrm{d}t > xh''(x) > 0$ for every $x > 0$. Consequently,

$$r'(x) = \frac{h(x)\big(xh''(x) - h'(x)\big) - x(h'(x))^2}{x^2h(x)^2} < 0.$$

For $\theta \in (0, \pi/4)$, we have $0 < a < b$, so $r(a) > r(b)$ and $u_y'(\theta) > 0$. For $\theta \in (\pi/4, \pi/2)$, we have $0 < b < a$, so $r(a) < r(b)$ and $u_y'(\theta) < 0$. Thus, on $[0, \pi/2]$, the minima occur exactly at $\theta = 0$ and $\theta = \pi/2$, corresponding to alignment with the cluster-mean directions, while $\theta = \pi/4$ is the unique maximum.

□

Lemma D.11 *Let* $\Pi_\alpha := P_\alpha \Pi_N P_\alpha$ *for* $\alpha \in \{S, K, B, C\}$. *Then* $\Pi_N = \Pi_S + \Pi_K + \Pi_B + \Pi_C$.

(Proof of Lemma D.11). Since $\sum_\alpha P_\alpha = I$ on the normal space, write $\Pi_N = \sum_{\alpha,\beta} P_\alpha \Pi_N P_\beta$ and denote $\Pi_{\alpha\beta} := P_\alpha \Pi_N P_\beta$. By Lemmas D.12 and D.6, both $H_N = \sum_\alpha H_\alpha$ and $\Sigma_N = \sum_\alpha \Sigma_\alpha$ are block diagonal. Projecting the Lyapunov equation $H_N\Pi_N + \Pi_N H_N = \Sigma_N$ onto blocks α and β gives

$$H_\alpha \Pi_{\alpha\beta} + \Pi_{\alpha\beta} H_\beta = P_\alpha \Sigma_N P_\beta.$$

For $\alpha \neq \beta$, the right-hand side is zero. Diagonalize the positive definite H_α and H_β on their respective blocks. In the resulting eigenbases, each entry of $\Pi_{\alpha\beta}$ satisfies $(\lambda_i + \mu_j)(\Pi_{\alpha\beta})_{ij} = 0$, where $\lambda_i, \mu_j > 0$. Hence $\Pi_{\alpha\beta} = 0$ whenever $\alpha \neq \beta$, leaving only the diagonal blocks $\Pi_\alpha = P_\alpha \Pi_N P_\alpha$. □

Lemma D.12 (Block diagonality of the normal sample-gradient covariance Σ_N). *For* $v \in \mathcal{M}$, *let* $g_N(v; x) := P_N(v)g(v; x)$ *and* $g_\alpha(v; x) := P_\alpha(v)g_N(v; x)$, *where* $\alpha \in \{S, K, B, C\}$. *Under the standing data and model assumptions, for* $\alpha \neq \beta$, $\mathbb{E}\left[g_\alpha(v; x)g_\beta(v; x)^\top\right] = 0$. *Consequently, the normal sample-gradient covariance is block diagonal* $\Sigma_N(v) := \mathbb{E}\left[g_N(v; x)g_N(v; x)^\top\right] = \sum_{\alpha\in\{S,K,B,C\}} \Sigma_\alpha(v)$, *where* $\Sigma_\alpha(v) := \mathbb{E}\big[g_\alpha(v; x)g_\alpha(v; x)^\top\big]$. *Similary for the linear case,* $\Sigma_N^{\mathrm{lin}} = \Sigma_S^{\mathrm{lin}} + \Sigma_K^{\mathrm{lin}}$.

(Proof of Lemma D.12). We denote $g_\alpha := g_\alpha(v; x)$. Note that $g_N = P_N g = (P_S + P_K + P_A + P_B)g$. Let the ambient gradient $h^{-1}(g) = [G_1, G_2] \in \mathbb{R}^{d\times r} \times \mathbb{R}^{d\times r}$. For $P_S g$, since this is the projection onto N_S, there exists some symmetric matrix $\hat{S} \in \mathbb{R}^{r\times r}$, such that $h^{-1}(P_S g) = [Q_r R\hat{S}, -Q_r R\hat{S}]$. In order to find $\hat{S}$, we use the fact that $h^{-1}(g) - h^{-1}(P_S g) \perp h^{-1}(N_S)$. For every symmetric $T = T^\top$, it satisfies that $\langle [G_1 - Q_r R\hat{S}, G_2 + Q_r R\hat{S}], [Q_r RT, -Q_r RT]\rangle_F = 0$. Using the fact that $\langle A, B\rangle_F = \mathrm{Tr}(A^\top B)$, and some straightforward calculation, the above term can be simplified as

$$\mathrm{Tr}\left(\left((G_1^\top - G_2^\top)Q_r R - 2\hat{S}\right)T\right) = 0.$$

This means that the matrix $(G_1^\top - G_2^\top)Q_r R - 2\hat{S}$ is a skew-symmetric matrix, thus solving $((G_1^\top - G_2^\top)Q_r R - 2\hat{S})^\top = -((G_1^\top - G_2^\top)Q_r R - 2\hat{S})$ gives us that $\hat{S} =$

$\frac{1}{4}\left((Q_rR)^\top(G_1-G_2)+(G_1-G_2)^\top Q_rR\right)$. For a point $W=\rho[Q_rR,-Q_rR]\in\mathcal{M}$, $(G_1-G_2)_i=2\lambda w_i+J_{w_i}r(x;W)-J_{w_{i+r}}r(x;W)$. Using the Eq. (169), with $y=R^\top z$, we have that

$$(Q_rR)^\top(G_1-G_2)=2\rho((1-\rho^2)yy^\top+\lambda I),$$

thus $\hat{S}=\rho(1-\rho^2)yy^\top+\rho\lambda I$. Finally, we have that

$$h^{-1}(P_Sg)=\rho[Q_rR((1-\rho^2)yy^\top+\lambda I),-Q_rR((1-\rho^2)yy^\top+\lambda I)].$$

Next, we calculate P_Kg in a similar way. For some $\hat{K}\in\mathbb{R}^{(d-r)\times r}$, let $h^{-1}(P_Kg)=[Q_\perp\hat{K},-Q_\perp\hat{K}]$. Since $h^{-1}(g)-h^{-1}(P_Kg)\perp N_K$, for any $T\in\mathbb{R}^{d-r\times r}$, it satisfies that $\langle[G_1-Q_\perp\hat{K},G_2+Q_\perp\hat{K}],[Q_\perp T,-Q_\perp T]\rangle=0$. Solving this gives us that $\hat{K}=\frac{1}{2}Q_\perp^\top(G_1-G_2)=\frac{1}{2}\rho(2-\rho^2)\varepsilon y^\top$. Therefore,

$$h^{-1}(P_Kg)=\left[\frac{1}{2}\rho(2-\rho^2)Q_\perp\varepsilon y^\top,-\frac{1}{2}\rho(2-\rho^2)Q_\perp\varepsilon y^\top\right].$$

For the B-block, we write $h^{-1}(P_Bg)=[Q_r\hat{B},Q_r\hat{B}]$. The orthogonality condition gives us that, for all $T\in\mathbb{R}^{r\times r}$, $\langle[G_1-Q_r\hat{B},G_2-Q_r\hat{B}],[Q_rT,Q_rT]\rangle_F=0$. Solving this gives us that $\hat{B}=\rho(1-\rho^2)Ry|y|^\top$. Thus,

$$h^{-1}(P_Bg)=[\rho(1-\rho^2)Q_rRy|y|^\top,\rho(1-\rho^2)Q_rRy|y|^\top].$$

Following the same steps, the last C-block is

$$h^{-1}(P_Cg)=\left[\frac{\rho(2-\rho^2)}{2}Q_\perp\varepsilon|y|^\top,\frac{\rho(2-\rho^2)}{2}Q_\perp\varepsilon|y|^\top\right].$$

We now prove that $\mathbb{E}[P_\alpha g(P_\beta g)^\top]=0$, for $\alpha\neq\beta$. For $\mathbb{E}[P_Sg(P_Kg)^\top]$ and $\mathbb{E}[P_Sg(P_Cg)^\top]$, ε is independent from y and centered, so these terms vanish. For $\mathbb{E}[P_Sg(P_Bg)^\top]$, Note that $\hat{S}(y)$ is an even function and $\hat{B}$ is an odd function. Thus, they yield an odd function, and due to the symmetry of the distribution of y, this term vanishes. For $\mathbb{E}[P_Kg(P_Bg)^\top]$ and $\mathbb{E}[P_Bg(P_Cg)^\top]$, these vanish, since ε is independent from y and centered. For $\mathbb{E}[P_Kg(P_Cg)^\top]$, note that $\hat{K}(y)$ is an odd function while $\hat{C}(y)$ is an even function. Thus, this term also vanishes. Therefore, $\mathbb{E}[P_\alpha g(P_\beta g)^\top]=0$, for $\alpha\neq\beta$ and it follows that $\Sigma_N=\Sigma_S+\Sigma_K+\Sigma_B+\Sigma_C$.

An analogous calculation for the linear case gives $\mathbb{E}[g_Sg_K^\top]=0$. Consequently, $\Sigma_N^{\text{lin}}=\Sigma_S^{\text{lin}}+\Sigma_K^{\text{lin}}$ (Note that due to Lemma D.7, the normal space has only the S,K blocks). □

Lemma D.13 (Sign of κ_B). *Let $y(\theta)=R^\top(\theta)z$, where $z=z_0+\sigma\varepsilon$, $\varepsilon\sim\mathcal{N}(0,I_2)$, and $z_0\sim\text{Unif}\{\pm se_1,\pm se_2\}$. Then, there exists $\sigma_*>0$, such that $\text{sgn}(\kappa_B)>0$ for all $\sigma\in(0,\sigma_*]$.*

(Proof of Lemma D.13). We write $N_{\alpha\beta}^{(0)}$ for $N_{\alpha\beta}$ evaluated at $\sigma=0$, and similarly for $N_{\gamma\delta}^{(0)}$. Note that, conditioned on z_0, the coordinates y_1 and y_2 are independent.

There are few useful facts coming out of the symmetry of the setting. First, by symmetry of the Gaussian noise, $se_i+\sigma\varepsilon=_d-se_i+\sigma\varepsilon$ for $i\in\{1,2\}$. Second, observe that $\text{sgn}(\kappa_B)=\text{sgn}(N_{\alpha\beta}+N_{\gamma\delta})=\text{sgn}(N_{\alpha\beta})$. Indeed, by the symmetry obtained from exchanging y_1 and y_2, we have $N_{\gamma\delta}=N_{\alpha\beta}$. Hence it suffices to prove that $N_{\alpha\beta}>0$.

The proof is divided into endpoint and interior regimes. By Lemma D.14, $N_{\alpha\beta}$ is continuous in (θ,σ). Moreover, when $\sigma=0$, we have $N_{\alpha\beta}^{(0)}>0$ away from the endpoints $\theta=0$ and $\theta=\pi/2$. Therefore, on any compact subinterval of $(0,\pi/2)$, positivity of $N_{\alpha\beta}$ for sufficiently small σ follows directly from continuity. The only delicate regimes are neighborhoods of the endpoints, where $N_{\alpha\beta}^{(0)}$ vanishes. In these regions, we use Taylor expansions to show that there exist $\delta>0$ and $\sigma_*>0$ such that $N_{\alpha\beta}(\theta,\sigma)>0$ for all $\theta\in[0,\delta]\cup[\pi/2-\delta,\pi/2]$ and all $\sigma\in(0,\sigma_*]$.

Case 1: θ near the endpoints. First consider where θ is near 0. We denote $a = s\cos\theta$ and $b = s\sin\theta$. Note that since θ is small, it satisfies that $a > b$. Let $X_a = a + \sigma G_a$ and $X_b = b + \sigma G_b$, where $G = (G_a, G_b)^\top \sim \mathcal{N}(0, I_2)$, and define $A_k := \mathbb{E}|X_a|^k$ and $B_k := \mathbb{E}|X_b|^k$. Using the conditional independence of y_1 and y_2 given z_0, and symmetry of the noise, we obtain $u_y = \mathbb{E}|y_1 y_2| = A_1 B_1$, $M_{31} := \mathbb{E}|y_1|^3|y_2| = \frac{1}{2}A_3 B_1 + \frac{1}{2}B_3 A_1$, $M_{40} := \mathbb{E}[y_1^4] = \frac{1}{2}(A_4 + B_4)$, and $M_{22} := \mathbb{E}[y_1^2 y_2^2] = A_2 B_2$. Therefore, $N_{\alpha\beta} = M_{31} - u_y\left(\frac{M_{40}}{4\lambda_s} + \frac{M_{22}}{2\lambda_s}\right)$. Let $t := \sin\theta + \sigma$. For θ sufficiently small, we may write $A_i = s^i + O(t^2)$ for $i \in \{1, \dots, 4\}$ (use the expansion $a = s\cos\theta = s - \frac{s}{2}\theta^2 + O(\theta^4)$). For X_b, we have $B_1 = \mathbb{E}|b + \sigma G_b| \geq c(b + \sigma)$ for some $c > 0$, since $\mathbb{E}|b + \sigma G_b| \geq b$ and $\mathbb{E}|b + \sigma G_b| \geq c_0\sigma$ for some $c_0 > 0$. Also, $B_i = O(t^i)$ for $i \in \{2, 3, 4\}$ (also, expand B_i and use the fact that the Gaussian moments are finite).

Using these estimates, we turn to evaluating $N_{\alpha\beta}$. First, $A_3 = s^3 + O(t^2)$, $A_1 = s + O(t^2)$, and $B_3 = O(t^3)$ imply $M_{31} = \frac{1}{2}s^3 B_1 + O(t^2 B_1) + O(t^3)$. Since $B_1 \geq c(s\sin\theta + \sigma)$, the term $O(t^3)$ is absorbed into $O(t^2 B_1)$, and hence $M_{31} = \frac{1}{2}s^3 B_1 + O(t^2 B_1)$. Moreover, $M_{40} = \frac{1}{2}(A_4 + B_4) = \frac{1}{2}s^4 + O(t^2)$ and $\lambda_s = \frac{s^2}{2} + \sigma^2 = \frac{s^2}{2} + O(t^2)$, so $\frac{M_{40}}{4\lambda_s} = \frac{s^2}{4} + O(t^2)$. Since $M_{22} = A_2 B_2 = O(t^2)$, we also have $\frac{M_{22}}{2\lambda_s} = O(t^2)$. Finally, $u_y = A_1 B_1 = sB_1 + O(t^2 B_1)$. Combining these estimates gives $N_{\alpha\beta} = \frac{1}{2}s^3 B_1 + O(t^2 B_1) - \left(sB_1 + O(t^2 B_1)\right)\left(\frac{s^2}{4} + O(t^2)\right) = \frac{1}{4}s^3 B_1 + O(t^2 B_1)$. Thus there exists $C > 0$ such that $N_{\alpha\beta} \geq \left(\frac{1}{4}s^3 - C(s\sin\theta + \sigma)^2\right) B_1$. Choose $\delta_1 > 0$ and $\sigma_1 > 0$ such that $\frac{1}{4}s^3 - C(s\sin\theta + \sigma)^2 \geq \frac{1}{8}s^3$ for all $\theta \in [0, \delta_1]$ and $\sigma \in (0, \sigma_1]$. Then $N_{\alpha\beta} \geq \frac{1}{8}s^3 B_1(\theta, \sigma) > 0$ on this region. By the same argument near the other endpoint, there exist $\delta_2 > 0$ and $\sigma_2 > 0$ such that $N_{\alpha\beta} > 0$ for all $\theta \in [\frac{\pi}{2} - \delta_2, \frac{\pi}{2}]$ and $\sigma \in (0, \sigma_2]$. We set $\delta := \min(\delta_1, \delta_2)$.

Case 2: θ away from the endpoints. Now let $\theta \in [\delta, \frac{\pi}{2} - \delta]$. By Lemma D.14, the map $N_{\alpha\beta}(\theta, \sigma)$ is continuous in (θ, σ). We first show that $N_{\alpha\beta}^{(0)} > 0$. In the case of $\sigma = 0$, we have that $y = R^\top(\theta)z = R^\top(\theta)z_0$, and a direct computation gives $\mathbb{E}[y_1^4] = \mathbb{E}[y_2^4] = \frac{s^4}{2}(\cos^4\theta + \sin^4\theta)$, $\mathbb{E}[y_1^2 y_2^2] = s^4\cos^2\theta\sin^2\theta$, $\lambda_s = \frac{s^2}{2}$, $u_y(\theta) = s^2\cos\theta\sin\theta$, and $\mathbb{E}[|y_1|^3|y_2|] = \mathbb{E}[|y_2|^3|y_1|] = \frac{s^4}{2}\cos\theta\sin\theta$. Therefore,

$$
\begin{aligned}
N_{\alpha\beta}^{(0)} &= \frac{s^4}{2}\cos\theta\sin\theta - s^2\cos\theta\sin\theta\left(\frac{\frac{s^4}{2}(\cos^4\theta + \sin^4\theta)}{2s^2} + \frac{s^4\cos^2\theta\sin^2\theta}{s^2}\right) \\
&= s^4\cos\theta\sin\theta\left(\frac{1}{4} - \frac{1}{2}\cos^2\theta\sin^2\theta\right).
\end{aligned}
$$

Since $\cos^2\theta\sin^2\theta \in [0, \frac{1}{4}]$, we have $\frac{1}{4} - \frac{1}{2}\cos^2\theta\sin^2\theta \geq \frac{1}{8} > 0$. Thus the only possible zeros of $N_{\alpha\beta}^{(0)}$ come from $\cos\theta\sin\theta$, which is impossible on $[\delta, \frac{\pi}{2} - \delta]$. Hence $N_{\alpha\beta}^{(0)}(\theta) > 0$ on this interval. Define $m_\delta := \min_{\theta\in[\delta,\pi/2-\delta]} N_{\alpha\beta}^{(0)}(\theta) > 0$. Since the interval is compact and $N_{\alpha\beta}^{(0)}$ is continuous, this minimum is attained and strictly positive. By continuity in (θ, σ), there exists $\sigma_3 > 0$ such that $|N_{\alpha\beta}^{(\sigma)}(\theta) - N_{\alpha\beta}^{(0)}(\theta)| \leq \frac{1}{2}m_\delta$ for all $\theta \in [\delta, \frac{\pi}{2} - \delta]$ and $\sigma \in [0, \sigma_3]$. Therefore, $N_{\alpha\beta}^{(\sigma)}(\theta) \geq \frac{1}{2}m_\delta > 0$ throughout the interior region.

Finally, choose $\sigma_* := \min(\sigma_1, \sigma_2, \sigma_3)$. Then $N_{\alpha\beta}(\theta, \sigma) > 0$ for all $\theta \in [0, \frac{\pi}{2}]$ and all $\sigma \in (0, \sigma_*]$. Since $\operatorname{sgn}(\kappa_B) = \operatorname{sgn}(N_{\alpha\beta})$, the claim follows.

□

Lemma D.14 *$N_{\alpha\beta}$ is continuous on $[0, \pi/2] \times [0, \sigma_0]$.*

(Proof of Lemma D.14). Let $z_\sigma = z_0 + \sigma g$, where $z_0 \sim \mathrm{Unif}\{\pm se_1, \pm se_2\}$ and $g \sim \mathcal{N}(0, I_2)$. For each $\theta \in [0, \pi/2]$, define $y_{\theta,\sigma} = R(\theta)^\top z_\sigma$ and let $y_{\theta,\sigma} = [y_1(\theta, \sigma), y_2(\theta, \sigma)]^\top$. Then the map $(\theta, \sigma) \mapsto N_{\alpha\beta}(\theta, \sigma)$ is continuous on $[0, \pi/2] \times [0, \sigma_0]$. This is a straightforward application of the dominated convergence theorem.

Since $N_{\alpha\beta} = M_{31} - u_y(\frac{M_{40}}{4\lambda_s} + \frac{M_{22}}{2\lambda_s})$, it is enough to show that each term is continuous. First, for $M_{31} = \mathbb{E}|y_1|^3|y_2|$, we show that, for a convergent sequence $(\theta_n, \sigma_n) \to (\theta, \sigma)$, it applies that

$\mathbb{E}|y_1(\theta_n,\sigma_n)|^3|y_2(\theta_n,\sigma_n)| \to \mathbb{E}|y_1(\theta,\sigma)|^3|y_2(\theta,\sigma)|$. First of all, it is straightforward to see that $y_i(\theta_n,\sigma_n) \to y_i(\theta,\sigma)$ a.s ($\|g\| < \infty$ a.s.) and the map $\phi(y) = |y_1|^3|y_2|$ is continuous, thus $|y_1(\theta_n,\sigma_n)|^3|y_2(\theta_n,\sigma_n)| \xrightarrow{a.s} |y_1(\theta,\sigma)|^3|y_2(\theta,\sigma)|$.

Now, since $\sigma_n \le \sigma_0 + 1$ for a sufficiently large n, $\|y_{\theta_n,\sigma_n}\| = \|z_0 + \sigma_n g\| \le \|z_0\| + (\sigma_0+1)\|g\|$. Since $|y_1|^3|y_2| \le \|y\|^4$, $|y_1(\theta_n,\sigma_n)|^3|y_2(\theta_n,\sigma_n)| \le (\|z_0\| + (\sigma_0+1)\|g\|)^4$. Since $s < \infty$ and $\mathbb{E}\|g\|^4 < \infty$, $\mathbb{E}(\|z_0\| + (\sigma_0+1)\|g\|)^4 < \infty$. Applying the dominated convergence theorem shows the desired result. This simple argument applies for all the other terms such as $\mathbb{E}[y_1^2 y_2^2]$, $\mathbb{E}y_1^4$, since $y_1^2 y_2^2 \le \|y\|^4$ and $y_1^4 \le \|y\|^4$. Thus $N_{\alpha\beta}$ is continuous on $[0,\pi/2] \times [0,\sigma_0]$. □

E MISCELLANEUOUS

Section E.1 provides the formal statement and proof of Proposition 5.1. Section E.2 details the settings for the figures in the main text and presents additional experimental results.

E.1 PROPOSITION 5.1

Proposition E.1 (Downstream XOR Classification with ReLU Features). *Consider the 4-GMM with means $\{\pm se_1, \pm se_2\}$ and component covariance $\sigma^2 I$, where e_1, e_2 are orthonormal. Assign label $y = +1$ to the components centered at $\pm se_1$ and $y = -1$ to those centered at $\pm se_2$. Let $\gamma := s/\sigma$ and $W_\theta = \rho[u_1, u_2, -u_1, -u_2] \in \mathcal{M}_{\mathrm{ReLU}}$, where $u_1 = \cos\theta\, e_1 - \sin\theta\, e_2$ and $u_2 = \sin\theta\, e_1 + \cos\theta\, e_2$. For any $c > 0$, define $u_\theta := \cos\theta - \sin\theta$ and $a_\theta := c\,\mathrm{sgn}(u_\theta)[1,-1,1,-1]^\top$. Then the linear-probe classifier $h_\theta(x) := \mathrm{sgn}\langle a_\theta, \phi(W_\theta^\top x)\rangle$ satisfies $\mathbb{P}(h_\theta(x) \ne y) \le 3\exp\left(-\frac{\gamma^2 u_\theta^2}{8}\right)$.*

(Proof of Proposition E.1). First suppose $\theta \in [0,\pi/4]$ and condition on $x \sim \mathcal{N}(se_1, \sigma^2 I)$. Writing $z_i := u_i^\top x$, the identity $\phi(z) + \phi(-z) = |z|$ gives $h_\theta(x) = \mathrm{sgn}(|z_1| - |z_2|)$. Here $z_1 \sim \mathcal{N}(s\cos\theta, \sigma^2)$ and $z_2 \sim \mathcal{N}(s\sin\theta, \sigma^2)$.

Since $\mathbb{P}(h_\theta(x) \ne y) = \mathbb{P}(|z_2| > |z_1|)$, a union bound gives $\mathbb{P}(h_\theta(x) \ne y) \le \mathbb{P}(|z_2| > t) + \mathbb{P}(|z_1| < t)$ for any $t \ge 0$. Choosing $t := s(\cos\theta + \sin\theta)/2$ yields

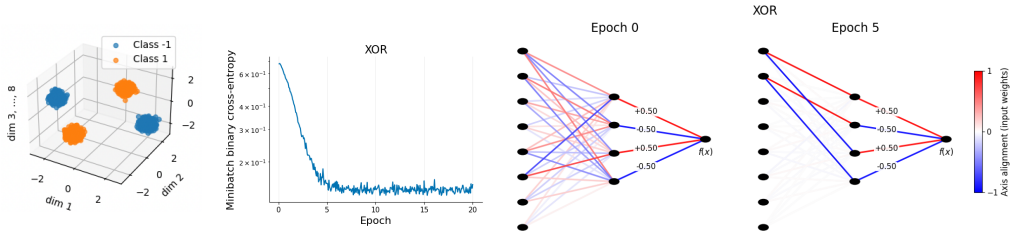
$$\begin{aligned}\mathbb{P}(h_\theta(x) \ne y) &\le \mathbb{P}(z_2 > t) + \mathbb{P}(z_2 < -t) + \mathbb{P}(z_1 < t)\\ &= 2\mathbb{P}\left(G < -\frac{\gamma u_\theta}{2}\right) + \mathbb{P}\left(G < -\frac{\gamma(\cos\theta + 3\sin\theta)}{2}\right)\\ &\le 3\exp\left(-\frac{\gamma^2 u_\theta^2}{8}\right),\end{aligned}$$

where $G \sim \mathcal{N}(0,1)$. The other components satisfy the same bound by sign symmetry and interchange of the two coordinates. For $\theta > \pi/4$, reversing the probe interchanges $|z_1|$ and $|z_2|$, and the same argument applies with θ replaced by $\pi/2 - \theta$. Finally, averaging over the 4 components proves the claim. □

E.2 ADDITIONAL EXPERIMENTS AND SETTINGS

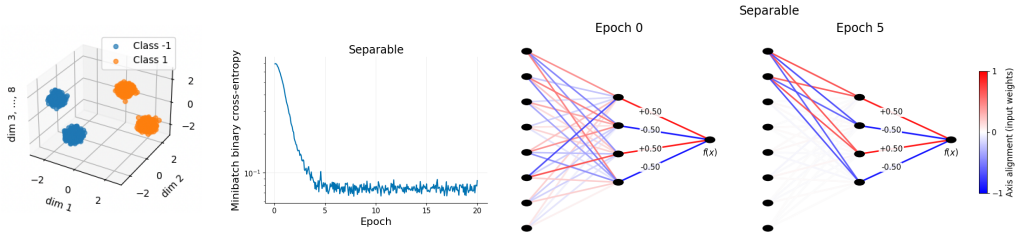
Additional experimental details. For Figure 1, we used shallow tied-weight autoencoders at minimal width, with Gaussian weight initialization, a learning rate of 0.25, and a batch size of 128. For the Gaussian-mixture experiment in Figure 3, we used same Gaussian initialization and the ambient dimension was set to $d = 32$.

For the supervised experiments in Figures 4 and 5, we use ℓ_2 regularization to control weight growth during training. We use the same GMM data as in the autoencoder experiments and a bias-free, two-layer ReLU classifier with four hidden units. The second-layer weights are held fixed at the values displayed in the figures. The first-layer weights are initialized in antipodal pairs. Specifically, we independently sample $v_1, v_2 \sim \mathcal{N}(0, \sigma_{\mathrm{init}}^2 I_d)$ with small σ_{init}, and initialize $(w_1, w_2, w_3, w_4) = (v_1, v_2, -v_1, -v_2)$(this symmetry is imposed only at initialization). These settings were chosen to stabilize optimization and obtain the feature configurations in the supervised experiments in Frei et al. (2023). Note that, we also observe the cluster-mean alignment phenomenon in autoencoders



(a) XOR (b) Training loss (c) Evolution of learned ReLU weights

Figure 4: Supervised feature learning by a two-layer ReLU network trained with SGD and ℓ_2 regularization on the 4-component Gaussian mixture with XOR labels. Mean-aligned features emerge by epoch 5, by which time the classification loss has also approximately plateaued. Both the feature directions and the loss remain nearly unchanged thereafter. This contrasts with the unsupervised autoencoder in Figure 1, whose features continue to realign toward the means while the reconstruction loss hardly changes. The comparison with Figure 5 highlights the role of labeling in selecting the learned feature directions.

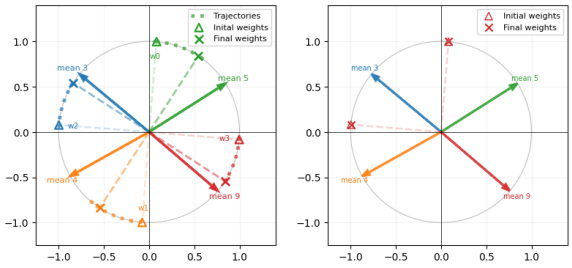


(a) Linearly separable (b) Training loss (c) Evolution of learned ReLU weights

Figure 5: Supervised feature learning using the same setting as Figure 4 with linearly separable labels. The learned features concentrate near the class-mean directions, rather than the individual mixture-component means.

with more unconstrained setting, i.e., even with untied encoder and decoder weights and independent Gaussian initialization (without imposing antipodal pairs).

For the Fashion-MNIST experiments in Figures 3 and 6, we selected four classes whose empirical mean directions are approximately symmetrically arranged in a two-dimensional subspace. The data are not centrally symmetric, and the class means have unequal norms (the displayed mean vectors are normalized to unit norm only for visualization). We randomly initialized the encoder weights near this subspace and trained both models by SGD on their reconstruction objectives. To increase the signal-to-noise ratio and improve class separation, we scaled the selected class means by factors between 3 and 4. No synthetic noise was added. Examples from the remaining classes were retained as additional unlabeled training data, serving as background noise relative to the selected classes.



(a) ReLU (b) Linear

Figure 6: Feature trajectories by ReLU and linear autoencoders on Fashion-MNIST. In the runs shown, the features from the ReLU autoencoder rotates toward the class-mean directions, whereas the linear autoencoder shows no comparable directional selection.