UNDERSTANDING REPRESENTATION ALIGNMENT IN NEURAL NETWORKS THROUGH NORMALIZED GRADIENT DYNAMICS WITH PERIODIC RESETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how neural networks develop useful representations during training remains an open problem. We study this question by introducing feature-target alignment, a scale-invariant quantity that measures how well a learned feature map captures the target function. We show that normalized feature learning initially induces an alignment-improving direction in the training. Motivated by this observation, we propose normalized gradient descent with periodic resets, which repeatedly returns training to a regime that promotes target-aligned representation learning. We prove that the resulting updates monotonically improve feature-target alignment for a broad class of parameterized feature maps over any given finite horizon, in both population and finite-sample settings. We further study a full-matrix representation model and show that the same learning principle strictly decreases the misspecified risk, providing a concrete example in which representation learning improves prediction. We also provide empirical support for our theoretical results.

1 INTRODUCTION

While deep neural networks have achieved remarkable empirical success, theoretical understanding of how feature representations evolve during training remains limited. One influential approach is the neural tangent kernel (NTK) regime, in which gradient descent on an infinitely wide neural network is described by kernel gradient descent with a limiting NTK (Jacot et al., 2018). This connection enables researchers to apply the mature theory of kernel methods to neural networks, including generalization rates, optimality guarantees, and saturation phenomena for regularized and early-stopped kernel algorithms (Bauer et al., 2007; Yao et al., 2007; Zhang et al., 2024).

These classical guarantees are nevertheless static: the kernel, and hence the feature map, is fixed throughout training. They cannot explain how gradient-based training improves the representation itself. Moreover, neural networks do not always remain close to their initialization. Depending on the initialization scale, training may either behave like a fixed kernel or substantially change the learned features, and these two regimes can generalize differently (Woodworth et al., 2020). For three-layer ReLU networks, Gatmiry et al. (2021) further obtain adaptive generalization bounds that optimize over a family of kernels containing the NTK, in a regime beyond the network's linear approximation. These results suggest that a fixed NTK is not by itself sufficient to explain the benefits of neural-network training.

Kernel learning offers a useful clue: high alignment can certify the existence of effective predictors (Cortes et al., 2012). Studies of evolving neural kernels likewise find that feature learning can improve target alignment (Shan & Bordelon, 2021; Atanasov et al., 2022). However, rigorous monotonicity results for target alignment remain confined to idealized or linear settings. Atanasov et al. (2022) analyze the beginning of training for a small-initialization, whitened two-layer linear network, while Bordelon & Pehlevan (2023) derive the full dynamics of a similar model. For general nonlinear networks, evidence that alignment increases remains largely empirical. Hence, a theoretical framework is needed to capture the alignment-improving dynamics of general networks.

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.

1.1 Related Work: Theoretical Perspectives on Neural Networks

Fixed-kernel theory. A lot of literature has focused on prediction and generalization in fixed kernel settings, where the representation remains unchanged during the learning process. This includes both the NTK regime of wide neural networks (Jacot et al., 2018) and classical theories of regularized and early-stopped kernel methods under source conditions, covering convergence rates, optimality, and saturation effects (Bauer et al., 2007; Yao et al., 2007; Zhang et al., 2024). As a result, these works reveal how well a fixed representation can perform, but do not address the representation learning that occurs in neural networks.

Adaptive kernels and target alignment. Another line of literature allows the kernel to adapt to the learning task. An important measure of the quality of a kernel is its alignment with the target function, which is proposed by Cortes et al. (2012). For general kernel methods, Cortes et al. (2012) develop algorithms that select combinations of base kernels by optimizing their alignment with the labels. In a neural-network setting, Gatmiry et al. (2021) show that projected SGD on a three-layer ReLU network admits a generalization bound that adapts to the best kernel in a family containing the NTK. Shan & Bordelon (2021) provide a normative explanation for target alignment. In a model where features are updated to accelerate subsequent prediction, they show that the induced kernel becomes more aligned with the target. Atanasov et al. (2022), in contrast, study how such alignment emerges from neural-network gradient descent. Under small initialization and whitened data, they prove that for two-layer linear networks, the NTK initially changes its eigendirections toward the target. For nonlinear networks, they report similar behavior empirically. Using a dynamical mean-field description of infinite-width networks, Bordelon & Pehlevan (2023) derive exact hidden-feature dynamics. Furthermore, in a setting similar to that of Atanasov et al. (2022), their limiting dynamics recover the phenomenon reported by Atanasov et al. (2022). More recently, Atanasov et al. (2025) provide a broader empirical analysis and report similar silent-alignment behavior across architectures and feature-learning strengths.

Feature-map views of neural networks. Another line of work considers the feature map as the main object of analysis. In the Bayesian setting, Zavatone-Veth et al. (2021) explicitly decompose a network into a feature map containing all hidden layers and a linear readout, and characterize leading finite-width corrections to the resulting feature kernels for several tractable architectures. In a diagonal sequence-model abstraction, Li & Lin (2024) parameterize the feature map as $\Phi_a(x) = (a_j e_j(x))_{j\geq 1}$ and show that over-parameterized gradient flow adapts the eigenvalues to the target structure. More recently, Nam et al. (2026) define an L^2 kernel operator from the penultimate-layer feature map and compare it with the function space spanned by the learned outputs, thereby separating dynamical richness from representation quality.

1.2 Our contributions

An alignment measure from the perspective of feature maps. In this paper, we study neural networks from the perspective of feature maps. From this perspective, we introduce the scale-invariant feature-target alignment Q, which measures the alignment between a feature map and the target function. Here, alignment means that the feature map assigns more representation capacity to directions that are more relevant to the target function. In the fixed-kernel setting, Q recovers the uncentered population version of kernel–target alignment defined in (Cortes et al., 2012; Atanasov et al., 2022), up to constants. Since Q is defined directly through the parameterized feature map, its dependence on the trainable parameters is explicit, making it a natural measure for studying how feature-target alignment evolves along the training trajectory. We also discuss its relationship with source conditions.

Repeated alignment improvement through periodic reset. We show that normalized feature learning, starting from zero readout, initially exhibits an alignment-improving phase, and propose normalized gradient descent with periodic reset to repeatedly recover this phase. In contrast to earlier exact results that are limited to early training and whitened linear models (Atanasov et al., 2022), our analysis establishes strict stepwise Q-improvement for a broad class of parameterized feature maps. By periodically resetting the linear readout, the method reinstates the alignment-improving regime and steadily improves Q over any fixed finite collection of training blocks. We further provide a finite-sample high-probability guarantee for our method.

Prediction improvement in the full-matrix case. For a full-matrix feature parameterization, we apply the proposed normalized updates with resets directly to the matrix parameter and show that every update strictly decreases the optimally refitted misspecified risk whenever the current feature space captures a nonzero but incomplete component of the target. This provides a concrete setting in which the proposed representation-learning procedure directly improves optimal linear prediction.

2 PRELIMINARIES

Notations. Let $\mathcal{H}_p = \mathbb{R}^p$ when $p < \infty$ and $\mathcal{H}_p = \ell^2$ when $p = \infty$. Inner products and Hilbert norms are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_2$. For an operator T between Hilbert spaces, $T^\top$ denotes its adjoint and $\|T\|_{\mathrm{HS}}$ its Hilbert–Schmidt norm.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the bounded input space. Let ρ be an unknown distribution on $\mathcal{X} \times \mathbb{R}$ with marginal distribution ρ_X, and suppose

$$y = f^*(x) + \epsilon, \qquad \mathbb{E}[\epsilon \mid x] = 0, \qquad \mathbb{E}[\epsilon^2] = \sigma^2 < \infty, \qquad f^* \in L^2(\mathcal{X}, \rho_X). \tag{1}$$

The observed data $\{(x_i, y_i)\}_{i=1}^n$ are sampled independently from ρ.

We consider a neural network with a linear output layer,

$$\widehat{y} = \langle \Phi_\theta(x), \beta \rangle_{\mathcal{H}_p}, \tag{2}$$

where $\Phi_\theta : \mathcal{X} \to \mathcal{H}_p$ is parameterized by $\theta \in \Theta \subseteq \mathbb{R}^D$, and $\beta \in \mathcal{H}_p$ is the linear head. The map Φ_θ represents all hidden layers of the network.

Write $\Phi_\theta = (\phi_{\theta,1}, \phi_{\theta,2}, \ldots)^\top$, where each $\phi_{\theta,j} \in L^2(\mathcal{X}, \rho_X)$. When $p = \infty$, assume in addition that $\sum_{j \geq 1} \|\phi_{\theta,j}\|^2_{L^2(\rho_X)} < \infty$. Fix an orthonormal basis $\Psi = \{\psi_i\}_{i \geq 1}$ of $L^2(\mathcal{X}, \rho_X)$, and denote

$$\Phi_\theta = P_\theta \Psi, \qquad f^* = f^\top \Psi,$$

where $f = (f_1, f_2, \ldots)^\top \in \ell^2$.

The feature covariance operator and the feature–target signal vector are

$$\Sigma_\theta := \mathbb{E}_{x \sim \rho_X}[\Phi_\theta(x)\Phi_\theta(x)^\top] = P_\theta P_\theta^\top, \qquad \mu_f^\theta := \mathbb{E}_{x \sim \rho_X}[\Phi_\theta(x) f^*(x)] = P_\theta f. \tag{3}$$

We assume Σ_θ is nonzero. Since P_θ is Hilbert–Schmidt, $\Sigma_\theta = P_\theta P_\theta^\top$ is trace class and Hilbert–Schmidt.

2.1 FEATURE-TARGET ALIGNMENT

Definition 2.1 (Feature-target alignment)**.** For any nonzero Hilbert–Schmidt operator P, define

$$Q(P, f) := \frac{\|Pf\|_2^2}{\|PP^\top\|_{\mathrm{HS}}}. \tag{4}$$

For a feature map $\Phi_\theta = P_\theta \Psi$, denote

$$Q(\Phi_\theta, f^*) := Q(P_\theta, f) = \frac{\|P_\theta f\|_2^2}{\|P_\theta P_\theta^\top\|_{\mathrm{HS}}}. \tag{5}$$

The definition is basis independent and scale invariant: Q remains unchanged under a different orthonormal basis Ψ, and $Q(c\Phi_\theta, f^*) = Q(\Phi_\theta, f^*)$ for every $c \neq 0$.

Remark 2.2 (Kernel interpretation of Q). Assume $f^* \neq 0$. Let the kernel induced by Φ_θ have Mercer expansion $K_\theta(x, x') = \sum_{j \geq 1} \lambda_{\theta,j} e_{\theta,j}(x) e_{\theta,j}(x')$, and set $f_{\theta,j} = \langle f^*, e_{\theta,j} \rangle_{L^2}$. Then

$$Q(\Phi_\theta, f^*) = \frac{\sum_{j \geq 1} \lambda_{\theta,j} f_{\theta,j}^2}{\left(\sum_{j \geq 1} \lambda_{\theta,j}^2\right)^{1/2}}, \tag{6}$$

which is the uncentered population counterpart of kernel–target alignment used in the kernel-learning and silent-alignment literature (Cortes et al., 2012; Shan & Bordelon, 2021; Atanasov et al., 2022). For certain kernel methods, Cortes et al. (2012) relate larger alignment to tighter generalization bounds. Atanasov et al. (2022) prove that such alignment grows at the beginning of training for a specific whitened two-layer linear model.

2.2 Similarity to the Source Condition

To provide further intuition for Q, let $\Psi_k = (\psi_{k,1}, \psi_{k,2}, \ldots)^\top$, $k \in \{0, 1\}$ be two orthonormal bases of $L^2(\mathcal{X}, \rho_X)$. Consider the feature maps $\Phi_k = \Lambda^{1/2}\Psi_k$ with the same eigenvalues

$$\Lambda := \mathrm{diag}(\lambda_1, \lambda_2, \ldots), \qquad \lambda_i = \frac{i^{-\nu}}{\sqrt{\zeta(2\nu)}}, \qquad \nu > 1.$$

Let $\|f^*\|_{L^2} = 1$ and write $f^* = g_k^\top \Psi_k$, where $g_k = (g_{k,i})_{i\geq 1}$, $g_{k,i}^2 = i^{-\alpha_k}/\zeta(\alpha_k)$, and $\alpha_k > 1$. Here $\zeta(\cdot)$ denotes the Riemann-zeta function. Then

$$Q_k := Q(\Phi_k, f^*) = \frac{\zeta(\nu + \alpha_k)}{\sqrt{\zeta(2\nu)}\,\zeta(\alpha_k)}. \tag{7}$$

Define the source condition

$$s_k := \sup\left\{ s \geq 0 : \sum_{i\geq 1} \frac{g_{k,i}^2}{\lambda_i^s} < \infty \right\} = \frac{\alpha_k - 1}{\nu}. \tag{8}$$

s_k measures how smooth the target is under the corresponding feature kernel $k(x, x') = \langle \Phi_k(x), \Phi_k(x') \rangle_{\mathcal{H}_p}$, which is often used to evaluate the performance of specific kernels (Caponnetto & De Vito, 2007; Fischer & Steinwart, 2020). We obtain the following equivalence.

Corollary 2.3 (Equivalence of source condition and Q)**.** *Under the shared-spectrum power-law setting above,*

$$Q_0 > Q_1 \quad \Longleftrightarrow \quad \alpha_0 > \alpha_1 \quad \Longleftrightarrow \quad s_0 > s_1. \tag{9}$$

Under standard kernel-regression assumptions, the optimal excess risk convergence rate in this setting is $n^{-\nu s/(\nu s+1)}$ (Caponnetto & De Vito, 2007; Fischer & Steinwart, 2020; Li et al., 2026). Hence, within this power-law family, larger Q corresponds to a faster excess-risk convergence rate.

3 Normalized Feature Learning with Periodic Resets

3.1 Learning Algorithm

We study a gradient-descent-based algorithm with two modifications. First, every prediction and gradient evaluation uses a normalized feature map. Second, training is divided into blocks of T joint updates, and the linear output layer β is periodically reset to zero at the beginning of each block.

Starting from the unnormalized feature map $\Phi_\theta = P_\theta \Psi$, define the scalar normalization

$$\bar{P}_\theta = \mathcal{N}(P_\theta) := \frac{P_\theta}{\|P_\theta P_\theta^\top\|_{\mathrm{HS}}^{1/2}}, \qquad \bar{\Phi}_\theta(x) := \bar{P}_\theta \Psi(x) = \frac{\Phi_\theta(x)}{\|\Sigma_\theta\|_{\mathrm{HS}}^{1/2}}. \tag{10}$$

Then $\|\bar{P}_\theta \bar{P}_\theta^\top\|_{\mathrm{HS}} = 1$ and, by scale invariance,

$$Q(\bar{\Phi}_\theta, f^*) = Q(\Phi_\theta, f^*) = \|\bar{P}_\theta f\|_2^2.$$

For brevity, we shall write $Q(\theta, f) := Q(\bar{\Phi}_\theta, f^*)$ without causing any ambiguity.

Why normalize? Notice that feature-map performance should be invariant under global rescaling: multiplying the feature map by a nonzero constant does not change its expressive ability. However, Mukhoti et al. (2020), Atanasov et al. (2022), and Hayou et al. (2026) report growth of feature or induced-kernel norms during training in various settings. Specifically, Atanasov et al. (2022) show that the growth of the kernel norm plays a leading role after several steps. Our normalization removes this irrelevant scale variation and focuses the analysis on changes in feature-target alignment. Furthermore, various normalization techniques have been widely used in practical neural-network training (Ioffe & Szegedy, 2015; Ba et al., 2016; Salimans & Kingma, 2016), further underscoring the practical importance of normalization.

Why reset? Atanasov et al. (2022) identify a silent-alignment regime in which target-aligned kernel structure develops only early in training. Our analysis also shows that, for normalized features, learning β from zero induces an alignment-improving feature update (see Theorem 3.3 for details). Hence, by periodically resetting β, our method restarts the silent-alignment phase and allows the feature map to evolve further. Our periodic reset can be regarded as a form of optimization intervention in neural network training, similar in spirit to dropout (Srivastava et al., 2014).

3.2 Population-Level Analysis

We first analyze gradient descent at the population level, where the loss and covariance scale are known exactly, and the error ϵ is omitted. The normalized population loss is

$$\bar{L}(\theta, \beta) = \mathbb{E}\Big[\big(f^*(x) - \langle \bar{\Phi}_\theta(x), \beta\rangle\big)^2\Big] = \|f - \bar{P}_\theta^\top \beta\|_2^2.$$

We need the following assumption so that the gradient descent is well defined.

Assumption 3.1 (Differentiability of the feature maps)**.** The parameter dimension D is finite. The maps $\theta \mapsto \Phi_\theta$ and $\theta \mapsto \bar{\Phi}_\theta$, viewed as maps to $L^2(\rho_X; \mathcal{H}_p)$, and $\theta \mapsto \bar{P}_\theta = \mathcal{N}(P_\theta)$, viewed in operator norm, are Fréchet differentiable almost everywhere on the parameter domain. Let Θ_{d} denote their common full-measure differentiability set, and assume that every parameter after a gradient update belongs to Θ_{d}. At each $\theta \in \Theta_{\mathrm{d}}$, for $j = 1, \ldots, D$, we write

$$J_j^\theta := \frac{\partial \Phi_\theta}{\partial \theta_j}, \qquad \bar{A}_j(\theta) := \frac{\partial \bar{P}_\theta}{\partial \theta_j}, \qquad g_j^\theta := \frac{\partial \bar{\Phi}_\theta}{\partial \theta_j}. \tag{11}$$

Remark 3.2. Assumption 3.1 allows us to adopt the standard gradient-descent update rule for the feature parameters. For simplicity, we further assume that all derivatives encountered in the remainder of the paper (including population and sample analyses) exist.

At the beginning of each resetting block, set $\beta_0 = 0$. Each inner iteration first updates the head and then the feature parameters:

$$\beta_{t+1} = \beta_t - \eta_1 \nabla_\beta \bar{L}(\theta_t, \beta_t), \qquad \theta_{t+1} = \theta_t - \eta_2 \nabla_\theta \bar{L}(\theta_t, \beta_{t+1}). \tag{12}$$

Algorithm 1 gives the population version of normalized gradient descent with periodic reset.

Algorithm 1 Normalized gradient descent with periodic reset

Require: Step sizes $\eta_1, \eta_2 > 0$, number of blocks $K \geq 1$, block length $T \geq 1$, final head-fit steps $J \geq 0$, initial parameter θ_0

$\theta \leftarrow \theta_0$
for $k = 1, \ldots, K$ **do**
  Reset $\beta \leftarrow 0$
  for $t = 0, \ldots, T-1$ **do**
    $\beta \leftarrow \beta - \eta_1 \nabla_\beta \bar{L}(\theta, \beta)$
    $\theta \leftarrow \theta - \eta_2 \nabla_\theta \bar{L}(\theta, \beta)$
    Use $\bar{P}_\theta = \mathcal{N}(P_\theta)$ at the next evaluation
  end for
end for
for $j = 1, \ldots, J$ **do**
  $\beta \leftarrow \beta - \eta_1 \nabla_\beta \bar{L}(\theta, \beta)$
end for
return (θ, β)

At each differentiable point, define $a_j(\theta) := f^\top \bar{P}_\theta^\top \bar{A}_j(\theta) f$. Denote $a(\theta) := (a_1(\theta), \ldots, a_D(\theta))^\top \in \mathbb{R}^D$. Notice that $\nabla_\theta Q(\theta, f) = 2a(\theta)$, thus $a(\theta) \neq 0$ is the non-criticality condition for Q. We first show that the first feature update with $\beta_0 = 0$ improves Q, which inspires us to periodically reset β.

Theorem 3.3 (Initial one-step alignment improvement)**.** *Suppose Assumption 3.1 holds, θ is a differentiable point, and $a(\theta) \neq 0$. For normalized gradient descent starting with θ and $\beta_0 = 0$, when the step sizes $\eta_1, \eta_2 > 0$ are sufficiently small, the first update satisfies*

$$Q(\theta^+, f) - Q(\theta, f) > 0.$$

Here θ^+ is the feature parameter after one step of the update in (12).

Remark 3.4 (Relation to silent alignment). At the level of feature-target alignment, Theorem 3.3 recovers the silent-alignment phenomenon of Atanasov et al. (2022), which report that neural networks initially enlarge the target alignment in training. Moreover, our result proves this silent-alignment phenomenon on a broader class of parameterized feature maps in our setting (notice that reset only affects the dynamics after the reset block), while the theoretical analysis of Atanasov et al. (2022) is only established for linear networks under small initialization and whitened data.

For further updates in the resetting block, we need a stronger assumption.

Assumption 3.5 (Uniform block regularity)**.** Assume that during the iterations, θ lies in a bounded convex open neighborhood U whose closure is contained in the parameter domain. Assume further that there are constants $L_U, L_g < \infty$ and $\alpha \in (0, 1]$ such that

$$\|\bar{P}_v - \bar{P}_w\|_{\mathrm{op}} \le L_U \|v - w\|_2, \qquad v, w \in U.$$

Writing $U_{\mathrm{d}} := U \cap \Theta_{\mathrm{d}}$ for the common differentiable points from Assumption 3.1, assume also that, for every $v, w \in U_{\mathrm{d}}$,

$$\left(\sum_{j=1}^{D} \|g_j^w - g_j^v\|^2_{L^2(\rho_X;\mathcal{H}_p)} \right)^{1/2} \le L_g \|w - v\|_2^{\alpha}. \tag{13}$$

Remark 3.6. The Lipschitz bound is used to control the propagation of the iterates, while (13) provides uniform first-order control of the normalized feature map. Together, these conditions imply the expansion of $\bar{P}_\theta f$ used in the proof of Theorem 3.7. Notice that (13) can be simply derived from Assumption 3.8, hence for smooth and ReLU networks, α can be chosen as 1 and $\frac{1}{2}$ respectively (we defer a detailed discussion to Remark 3.9).

Theorem 3.7 (Monotone improvement over finite resetting blocks)**.** *Under Assumptions 3.1 and 3.5, fix $T < \infty$ and suppose $a(\theta_0) \neq 0$. Then for sufficiently small step sizes $\eta_1, \eta_2 > 0$, the following holds:*

$$Q(\theta_{t+1}, f) - Q(\theta_t, f) > 0, \qquad t = 0, \dots, T-1.$$

The same conclusion holds over any fixed finite number of repeated-reset blocks if a is nonzero at the beginning of every block and the same neighborhood and differentiability conditions hold.

3.3 Sample-Level Analysis

The population derivation above uses the exact loss and covariance scale, while empirically we only have access to samples. We now construct its sample-level counterpart.

The first obstacle when transitioning to the sample level is the unknown population covariance norm of the feature map. We estimate it by splitting the samples into two parts:

$$\mathcal{S}_{\mathrm{N}} = \{(z_r, z_r')\}_{r=1}^{m}, \qquad \mathcal{S}_{\mathrm{G}} = \{(x_i, y_i)\}_{i=1}^{n}.$$

Here the m pairs are i.i.d. in $\rho_X \otimes \rho_X$, independently of $\mathcal{S}_{\mathrm{G}}$. Denote $h_\theta := \|\Sigma_\theta\|^2_{\mathrm{HS}}$. Since

$$\mathbb{E}_{z,z'} \langle \Phi_\theta(z), \Phi_\theta(z') \rangle^2 = \mathrm{tr}(\Sigma_\theta^2) = h_\theta,$$

the normalized feature map is given by

$$\widehat{h}_\theta = \frac{1}{m} \sum_{r=1}^{m} \langle \Phi_\theta(z_r), \Phi_\theta(z_r') \rangle^2, \qquad \widehat{\bar{\Phi}}_\theta(x) = \frac{\Phi_\theta(x)}{\widehat{h}_\theta^{1/4}}. \tag{14}$$

Whenever $\widehat{h}_\theta > 0$, the estimator is well defined. The empirical normalizer $\widehat{h}_\theta$ participates in the gradient computation, so the θ-gradient differentiates through $\widehat{h}_\theta$.

Hold both $\mathcal{S}_{\mathrm{N}}$ and $\mathcal{S}_{\mathrm{G}}$ fixed throughout a resetting block and define the sample loss

$$\widehat{L}(\theta, \beta; \mathcal{S}_{\mathrm{G}}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left\langle \widehat{\bar{\Phi}}_\theta(x_i), \beta \right\rangle_{\mathcal{H}_p} \right)^2. \tag{15}$$

At the beginning of each resetting block, set $\widehat{\beta}_0 = 0$ and let $\widehat{\theta}_0$ be the current feature parameter. Each inner iteration first updates the head and then the feature parameters:

$$\widehat{\beta}_{t+1} = \widehat{\beta}_t - \eta_1 \nabla_\beta \widehat{L}(\widehat{\theta}_t, \widehat{\beta}_t; \mathcal{S}_{\mathrm{G}}), \qquad \widehat{\theta}_{t+1} = \widehat{\theta}_t - \eta_2 \nabla_\theta \widehat{L}(\widehat{\theta}_t, \widehat{\beta}_{t+1}; \mathcal{S}_{\mathrm{G}}). \tag{16}$$

This gives the sample analogue of Algorithm 1. We now provide assumptions needed for sample-level analysis.

Assumption 3.8. On the bounded convex neighborhood U in Assumption 3.5, there are constants $R, M, L < \infty$, $r > 0$, and $\alpha \in (0, 1]$ such that, for every $v \in U_{\mathrm{d}}$ and almost every x,

$$\|\Phi_v(x)\|_2 \le R, \qquad \left(\sum_{j=1}^{D} \|\partial_{v_j}\Phi_v(x)\|_2^2\right)^{1/2} \le M, \qquad \|\Sigma_v\|_{\mathrm{HS}} \ge r^2.$$

Define the local oscillation

$$\Omega_{v,\tau}(x) := \sup_{\substack{w \in U_{\mathrm{d}} \\ \|w-v\|_2 \le \tau}} \left(\|\Phi_w(x) - \Phi_v(x)\|_2^2 + \sum_{j=1}^{D} \|\partial_{w_j}\Phi_w(x) - \partial_{v_j}\Phi_v(x)\|_2^2\right)^{1/2}. \tag{17}$$

We assume that for every $v \in U_{\mathrm{d}}$ and $\tau \in (0, 1]$,

$$\|\Omega_{v,\tau}\|_{L^2(\rho_X)} \le L\tau^{\alpha}. \tag{18}$$

Remark 3.9 (Smooth and ReLU feature maps). For a smooth feature map with bounded controls, the local oscillation condition in Assumption 3.8 generally holds with $\alpha = 1$. For a finite ReLU network, Proposition A.7 verifies the assumption with $\alpha = 1/2$ under boundedness, covariance nondegeneracy, and a local activation-boundary condition.

Assumption 3.10. The target function satisfies $|f^*(x)| \le R_1$ almost surely.

Assumption 3.11 (Sub-Gaussian noise)**.** The noise variables are independent of the inputs, mean zero, and sub-Gaussian with $\|\epsilon_i\|_{\psi_2} \le K_\epsilon$.

We now state the main sample-level result.

Theorem 3.12 (Finite-sample block improvement)**.** *Suppose Assumptions 3.1, 3.5, 3.8, 3.10, and 3.11 hold. Fix $T < \infty$ and $\delta \in (0, 1)$, and suppose $\widehat{\theta}_0 \in U_{\mathrm{d}}$ and $a_0 := \|a(\widehat{\theta}_0)\|_2 > 0$. Set $\bar{\gamma} := \min\{1, a_0/2\}$. There exist constants $C \ge 1$ and $\eta_* > 0$ such that, whenever*

$$m, n \ge C\bar{\gamma}^{-4}\left[D\log\left(C\bar{\gamma}^{-2/\alpha}\right) + \log\frac{C}{\delta}\right], \tag{19}$$

then, for $0 < \eta_1, \eta_2 \le \eta_*$, with probability at least $1 - \delta$,

$$Q(\widehat{\theta}_{t+1}, f) - Q(\widehat{\theta}_t, f) > 0, \qquad t = 0, \dots, T-1.$$

Here C and η_* do not depend on m, n, δ.

The one-block guarantee can be applied to repeated-reset blocks using fresh samples at each block, with the current feature parameter serving as the new initialization. The sample-size and step-size requirements depend on the corresponding new block-start value of a_0, but not on the total number of blocks.

Remark 3.13 (Main difficulties in the sample-level proof). The proof of Theorem 3.12 is more than regular concentration analysis. First, $\widehat{\theta}_t$ differs when the training sample differs, so a uniform concentration bound is needed. As a result, we use an ε-net argument to derive such uniform concentration. Second, the sample update replaces the population normalization by its empirical counterpart, so the concentrations of the empirical normalizer and its derivative are both needed, which makes the entire proof more complicated.

4 Special Case: Full-Matrix Feature Operators

The preceding results establish stepwise improvement of feature-target alignment for a general parameterization $\theta \mapsto P_\theta$ under normalized gradient descent with periodic reset. Beyond this general guarantee, different choices of parameterization may induce additional structure in the learning dynamics. For example, Li & Lin (2024) analyze the diagonal parameterization $P_\theta = \mathrm{diag}(\lambda_1, \lambda_2, \dots)$ and characterize its resulting dynamics.

We now consider a full-matrix parameterization in which the feature operator itself is learned, and show that the misspecified risk decreases. Specifically, let $p < \infty$ and consider

$$P \in \mathcal{S}_2(\ell^2, \mathbb{R}^p), \qquad \Phi_P = \sum_{i \ge 1}(Pe_i)\psi_i, \qquad \widehat{y}(x) = \langle \Phi_P(x), \beta\rangle_{\mathbb{R}^p},$$

where $\mathcal{S}_2(\ell^2, \mathbb{R}^p)$ is the Hilbert space of Hilbert–Schmidt operators from ℓ^2 to $\mathbb{R}^p$, and the series defining Φ_P converges in $L^2(\rho_X; \mathbb{R}^p)$; here e_i is the ith canonical basis vector of ℓ^2. Equivalently, we write $\Phi_P = P\Psi$ using the L^2 shorthand introduced above.

We first consider the **population case.** For an arbitrary nonzero operator P, define the normalized population loss by

$$\bar{L}(P, \beta) := \|f - \mathcal{N}(P)^\top \beta\|_2^2.$$

Here the full operator P is the feature parameter. At the beginning of each resetting block, set $\beta_0 = 0$. Starting from a normalized P_t, each iteration uses the following normalized update:

$$\beta_{t+1} = \beta_t - \eta_1 \nabla_\beta \bar{L}(P_t, \beta_t) = \beta_t + 2\eta_1 P_t(f - P_t^\top \beta_t). \tag{20}$$

Denote $r_t := f - P_t^\top \beta_{t+1}$. The feature update is

$$\begin{aligned} \widetilde{P}_{t+1} &= P_t - \eta_2 \nabla_P \bar{L}(P_t, \beta_{t+1}) \\ &= P_t + 2\eta_2 \beta_{t+1} r_t^\top - 2\eta_2 \big(\beta_{t+1}^\top P_t r_t\big) P_t P_t^\top P_t, \qquad P_{t+1} = \mathcal{N}(\widetilde{P}_{t+1}). \end{aligned} \tag{21}$$

The calculation of the update is given in the proof of Theorem 4.1.

We evaluate the feature map Φ_P through the best linear predictor it can express. Define the misspecified risk of the feature map by

$$\mathcal{E}_{\mathrm{opt}}(P) := \inf_{\beta \in \mathbb{R}^p} \|f - P^\top \beta\|_2^2 = \|f - \Pi_P f\|_2^2, \qquad \Pi_P := P^\top (PP^\top)^{-1} P. \tag{22}$$

Here $\mathcal{E}_{\mathrm{opt}}(P)$ evaluates the representation after optimal linear refitting. We assume that P has full row rank for brevity, so $PP^\top$ is invertible, Π_P is the orthogonal projector onto $\mathrm{row}(P) \subset \ell^2$, and the optimal β is $\beta_P^* = (PP^\top)^{-1} Pf$.

The following theorem shows that, under this normalized full-matrix update, the misspecified risk strictly decreases after every update, provided that the current feature space captures a nonzero but incomplete component of the target.

Theorem 4.1 (Population full-matrix block improvement)**.** *Let $P_0 \in \mathcal{S}_2(\ell^2, \mathbb{R}^p)$ be normalized and have full row rank, and suppose $0 < \|\Pi_{P_0} f\|_2 < \|f\|_2$. Fix $T < \infty$. Then, for sufficiently small step sizes $\eta_1, \eta_2 > 0$, the iterates (20)–(21) remain full row rank and satisfy*

$$\mathcal{E}_{\mathrm{opt}}(P_{t+1}) < \mathcal{E}_{\mathrm{opt}}(P_t), \qquad t = 0, \ldots, T-1.$$

The same conclusion holds over any fixed finite number of resetting blocks when the head is reset to zero at the beginning of each block.

Sample-level setting. In reality, we can only update a finite number of parameters. Hence, at the sample level, we restrict the feature operator to the first $d_{\mathrm{tr}} < \infty$ basis coordinates while leaving f^* unchanged. As a result, its block updates decrease the misspecified risk with high probability. A detailed discussion is deferred to Appendix A.7.

5 Experiments

We compare the performance of our normalized and periodic resetting gradient descent method with ordinary gradient descent on the following model:

Let $x \sim \mathrm{Unif}[0, 1]$, $f^*(x) = \sin(\pi x)$, and $y = f^*(x) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.1^2)$. For $j \geq 1$, define

$$\begin{aligned} [\Phi_0(x)]_j &= \sqrt{\lambda_{0,j}}\sqrt{2}\sin(j\pi x), & \lambda_{0,j} &= \frac{1}{\pi^2 j^2}, \\ [\Phi_1(x)]_j &= \sqrt{\lambda_{1,j}}\sqrt{2}\sin\left(\frac{(2j-1)\pi x}{2}\right), & \lambda_{1,j} &= \frac{4}{\pi^2(2j-1)^2}, \end{aligned}$$

and let $\Phi_\theta = (1-\theta)\Phi_0 + \theta\Phi_1,\ \theta \in [0, 1]$. The target is proportional to the first eigenfunction of Φ_0, so smaller θ generally corresponds to a more target-adapted representation.

Figure 1 shows that resetting and normalization drive all trajectories toward zero and hence yield a more target-adapted feature map, while ordinary gradient descent remains sensitive to initialization. This indicates that our proposed method can find better feature maps more steadily.

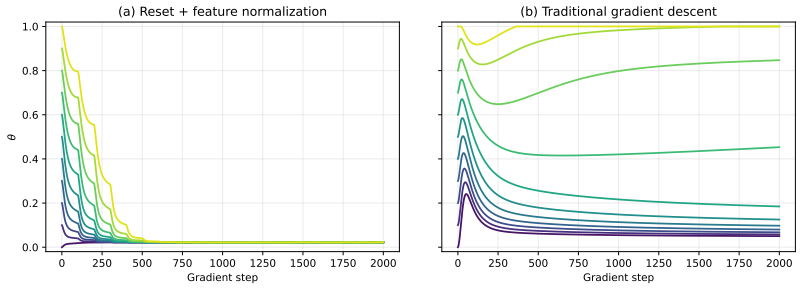

Figure 1: Feature-parameter trajectories from $\theta_0 \in \{0, 0.1, \ldots, 1\}$.

Figure 2 compares our method with ordinary gradient descent, and shows an overall increase in Q under our method, which offers empirical support for our theoretical results. Although not proved directly, it suggests that improved feature–target alignment is indicative of lower prediction error.

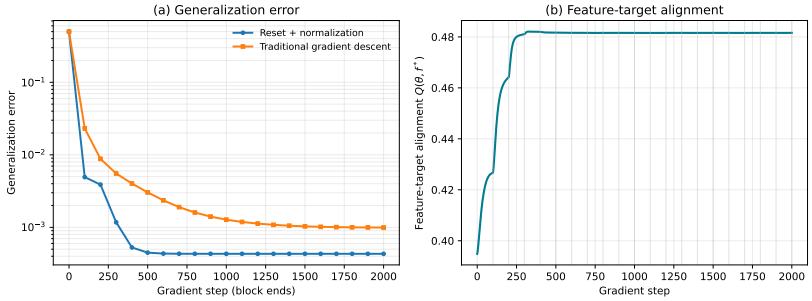

Figure 2: Results for $\theta_0 = 0.8$. Left: prediction error at the end of each block. Right: the feature-target alignment Q under the proposed method; vertical lines indicate resetting times.

We provide further experiments in Appendix B, including: ablation studies on the model provided above, numerical illustration of the full-matrix result in Section 4, experiments on empirical datasets. We provide a summary of further experiments, while deferring detailed results to the appendix.

- **Ablation study.** Appendix B.1 shows that resetting with normalization has better performance and robustness than either resetting or normalization alone.
- **Full-matrix experiment.** Appendix B.2 illustrates that the misspecified risk decreases over the training process, which is consistent with the theoretical results.
- **Experiments on empirical datasets.** Appendix B.3 reports validation performance on six datasets (CIFAR-10, SVHN, Fashion-MNIST, MNIST, UCI HAR, Wine Quality) over 20 seeds. As a result, our method achieves the best average rank among the four methods (traditional, reset only, normalize only, reset + normalize).

6 Conclusion

In this work, we introduce the feature-target alignment Q for feature maps. We propose a normalized gradient descent method with periodic resets, and show that it steadily improves the feature-target alignment Q. Then, we consider a full-matrix parameterization and prove that, in this special case, updates strictly decrease the misspecified risk.

Future work includes characterizing the long-term limit of the learned feature map or the value of Q. It would also be interesting to investigate whether other structured feature parameterizations, beyond the full-matrix example studied here and diagonal matrices studied by Li & Lin (2024), yield additional structural properties. Finally, establishing a general relationship between feature alignment and generalization error remains an important open problem.

REFERENCES

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. *International Conference on Learning Representations*, 2022.

Alexander Atanasov, Alexandru Meterez, James Simon, and Cengiz Pehlevan. The optimization landscape of sgd across the feature learning strength. In *International Conference on Learning Representations*, volume 2025, pp. 5740–5779, 2025.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in neural information processing systems*, 35:9768–9783, 2022.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11): 114009, 2023.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. doi: 10.1007/s10208-006-0196-8.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13:795–828, 2012.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

Khashayar Gatmiry, Stefanie Jegelka, and Jonathan Kelner. Optimization and adaptive generalization of three layer neural networks. In *International Conference on Learning Representations*, 2021.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Plop: Precise lora placement for efficient finetuning of large models. In *International Conference on Learning Representations*, volume 2026, pp. 65166–65191, 2026.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

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

Yicheng Li and Qian Lin. Improving adaptivity via over-parameterization in sequence models. *Advances in Neural Information Processing Systems*, 37:52438–52480, 2024.

Yicheng Li, Weiye Gan, Zuoqiang Shi, and Qian Lin. Generalization error curves for analytic spectral algorithms under power-law decay. *Applied and Computational Harmonic Analysis*, pp. 101920, 2026.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, 33:15288–15299, 2020.

Yoonsoo Nam, Nayara Fonseca, Seok Hyeong Lee, Chris Mingard, Niclas Göring, Ouns El Harzli, Abdurrahman Erturk, Soufiane Hayou, and Ard Louis. Decoupling dynamical richness from representation learning: Towards practical measurement. In *International Conference on Learning Representations*, volume 2026, pp. 63327–63369, 2026.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Haozhe Shan and Blake Bordelon. A theory of neural tangent kernel alignment and its influence on training. *arXiv preprint arXiv:2105.14301*, 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive approximation*, 26(2):289–315, 2007.

Jacob Zavatone-Veth, Abdulkadir Canatar, Ben Ruben, and Cengiz Pehlevan. Asymptotics of representation learning in finite bayesian neural networks. *Advances in neural information processing systems*, 34:24765–24777, 2021.

Haobo Zhang, Yicheng Li, and Qian Lin. On the optimality of misspecified spectral algorithms. *Journal of Machine Learning Research*, 25(188):1–50, 2024.

AI use statement In this work, we used generative AI tools for editing a research paper to improve readability, helping design experiments, assisting in the writing of proofs, summarizing and identifying related literature. We have not used generative AI tools for other tasks. We have reviewed all AI-assisted work, including checking the code, proof, references and all other information provided by LLM. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

A Proofs

A.1 Proof of Corollary 2.3

Proof. Let

$$h(\alpha) = \frac{\zeta(\alpha+\nu)}{\zeta(\alpha)}, \qquad \alpha > 1.$$

Because $h(\alpha) > 0$, it is enough to determine the sign of its logarithmic derivative. For $s > 1$, differentiating the absolutely convergent Euler product gives

$$\frac{\zeta'(s)}{\zeta(s)} = -\sum_{q \text{ prime}} \frac{\log q}{q^s - 1}.$$

Each summand $\log q/(q^s-1)$ is strictly decreasing in s, so $\zeta'(s)/\zeta(s)$ is strictly increasing. Therefore

$$\frac{d}{d\alpha}\log h(\alpha) = \frac{\zeta'(\alpha+\nu)}{\zeta(\alpha+\nu)} - \frac{\zeta'(\alpha)}{\zeta(\alpha)} > 0,$$

and h is strictly increasing.

For the common spectrum in Section 2.2, (7) can be written as

$$Q_k = \frac{1}{\sqrt{\zeta(2\nu)}} h(\alpha_k).$$

The prefactor is positive and independent of k, and hence

$$Q_0 > Q_1 \quad \Longleftrightarrow \quad \alpha_0 > \alpha_1.$$

Equation 8 gives $s_k = (\alpha_k - 1)/(\nu)$. Because $\nu > 0$, this is a strictly increasing affine function of α_k, and hence

$$\alpha_0 > \alpha_1 \quad \Longleftrightarrow \quad s_0 > s_1.$$

Combining the two equivalences proves (9). □

A.2 Proof of Theorem 3.3

Proof. Write $\bar{P} = \bar{P}_\theta$ and $\bar{A}_j = \bar{A}_j(\theta)$. Scalar normalization gives $\|\bar{P}\bar{P}^\top\|_{\mathrm{HS}} = 1$, so

$$Q(\theta, f) = \|\bar{P}f\|_2^2.$$

Since $\theta \mapsto \bar{P}_\theta$ is Fréchet differentiable, for every $j = 1, \dots, D$, we have

$$\partial_{\theta_j} Q(\theta, f) = 2(\bar{P}f)^\top \bar{A}_j f = 2a_j(\theta). \tag{23}$$

By differentiating the population loss, we get

$$\nabla_\beta \bar{L}(\theta, \beta) = -2\bar{P}f + 2\bar{P}\bar{P}^\top\beta, \tag{24}$$

$$\partial_{\theta_j} \bar{L}(\theta, \beta) = -2\beta^\top \bar{A}_j\big(f - \bar{P}^\top\beta\big). \tag{25}$$

Starting from $\beta_0 = 0$, the update of β is

$$\beta^+ = 2\eta_1 \bar{P}f. \tag{26}$$

Define $u = \bar{P}f$. Substituting (26) into (25) gives

$$\begin{aligned}\theta_j^+ - \theta_j &= 2\eta_2(\beta^+)^\top \bar{A}_j\big(f - \bar{P}^\top\beta^+\big) \\ &= 4\eta_1\eta_2 u^\top \bar{A}_j f - 8\eta_1^2\eta_2 u^\top \bar{A}_j \bar{P}^\top u \\ &= 4\eta_1\eta_2 a_j(\theta) + O(\eta_1^2\eta_2).\end{aligned}$$

Consequently,

$$\theta^+ - \theta = 4\eta_1\eta_2 a(\theta) + O(\eta_1^2\eta_2). \tag{27}$$

Hence, by expanding Q at θ, we get

$$\begin{aligned}Q(\theta^+, f) - Q(\theta, f) &= \nabla_\theta Q(\theta, f)^\top(\theta^+ - \theta) + o(\|\theta^+ - \theta\|_2) \\ &= 8\eta_1\eta_2\|a(\theta)\|_2^2 + O(\eta_1^2\eta_2) + o(\eta_1\eta_2),\end{aligned}$$

where we used (23) and (27). If $a(\theta) \neq 0$, the coefficient of $\eta_1\eta_2$ is strictly positive. Taking η_1 small enough controls the $O(\eta_1^2\eta_2)$ term, and then taking the total parameter increment small enough controls the little-o term. □

A.3 Proof of Theorem 3.7

We first prove a lemma that gives a first order expansion of $\bar{P}_\theta f$ with a uniform remainder bound.

Lemma A.1 (First-order consequence of Jacobian regularity)**.** *Under Assumptions 3.1 and 3.5, for every $v, w \in U_{\mathrm{d}}$,*

$$\begin{aligned}&\left\|(\bar{P}_w - \bar{P}_v)f - \sum_{j=1}^{D}(w_j - v_j)\bar{A}_j(v)f\right\|_2 \\ &\leq \frac{L_g\|f\|_2}{1+\alpha}\|w - v\|_2^{1+\alpha}.\end{aligned} \tag{28}$$

Moreover, there is a constant $1 \leq C_a < \infty$ such that

$$\|a(w) - a(v)\|_2 \leq C_a\|w - v\|_2^\alpha, \qquad v, w \in U_{\mathrm{d}}. \tag{29}$$

Proof. Let $u(v) := \bar{P}_v f$. For $v \in U_{\mathrm{d}}$, define the linear map $B(v) : \mathbb{R}^D \to \mathcal{H}_p$ by

$$B(v)[z] := \sum_{j=1}^{D} z_j \bar{A}_j(v)f.$$

The coefficient representation of the feature map gives

$$\bar{A}_j(v)f = \mathbb{E}_x[f^*(x)g_j^v(x)].$$

Consequently, Cauchy–Schwarz and (13) imply

$$\begin{aligned}\|B(w) - B(v)\|_{\mathrm{op}} &\le \left(\sum_{j=1}^{D} \|(\bar{A}_j(w) - \bar{A}_j(v))f\|_2^2\right)^{1/2} \\ &\le \|f\|_2 \left(\sum_{j=1}^{D} \|g_j^w - g_j^v\|_{L^2(\rho_X;\mathcal{H}_p)}^2\right)^{1/2} \\ &\le L_g \|f\|_2 \|w - v\|_2^{\alpha}. \end{aligned} \tag{30}$$

The set U_{d} has full measure and is dense in U, so (30) gives a unique Hölder-continuous extension of B to all of U. We continue to denote this extension by B. Moreover, the Lipschitz part of Assumption 3.5 gives

$$\|u(w) - u(v)\|_2 \le L_U \|f\|_2 \|w - v\|_2, \qquad v, w \in U.$$

Thus u is locally Lipschitz, its weak derivative agrees almost everywhere with its Fréchet derivative, and that derivative is B on U_{d}. Because B is continuous on U and agrees almost everywhere with the weak derivative of u, the standard weak-to-classical regularity result (obtained, for example, by mollification) gives $u \in C^1(U; \mathcal{H}_p)$ with $Du = B$ throughout U.

Now fix $v, w \in U_{\mathrm{d}}$ and denote $\Delta = w - v$. The fundamental theorem of calculus along the line segment from v to w gives

$$u(w) - u(v) = \int_0^1 B(v + s\Delta)[\Delta] \mathrm{d}s.$$

Therefore, by (30),

$$\begin{aligned}&\|u(w) - u(v) - B(v)[\Delta]\|_2 \\ &\le \|\Delta\|_2 \int_0^1 \|B(v + s\Delta) - B(v)\|_{\mathrm{op}} \, \mathrm{d}s \\ &\le L_g \|f\|_2 \|\Delta\|_2^{1+\alpha} \int_0^1 s^{\alpha} \, \mathrm{d}s = \frac{L_g \|f\|_2}{1 + \alpha} \|\Delta\|_2^{1+\alpha},\end{aligned}$$

which proves (28).

It remains to verify (29). For every $j = 1, \ldots, D$, the definition of a_j gives

$$a_j(w) - a_j(v) = \left\langle (\bar{P}_w - \bar{P}_v)f, \bar{A}_j(w)f \right\rangle + \left\langle \bar{P}_v f, (\bar{A}_j(w) - \bar{A}_j(v))f \right\rangle.$$

Consequently, the triangle inequality and Cauchy–Schwarz give

$$\begin{aligned}\|a(w) - a(v)\|_2 \le \|(\bar{P}_w - \bar{P}_v)f\|_2 &\left(\sum_{j=1}^{D} \|\bar{A}_j(w)f\|_2^2\right)^{1/2} \\ &+ \|\bar{P}_v f\|_2 \left(\sum_{j=1}^{D} \|(\bar{A}_j(w) - \bar{A}_j(v))f\|_2^2\right)^{1/2}.\end{aligned}$$

Using Assumption 3.5 and (30) therefore yields

$$\|a(w) - a(v)\|_2 \le C\big(\|w - v\|_2 + \|w - v\|_2^{\alpha}\big) \le C_a \|w - v\|_2^{\alpha},$$

where the last inequality uses $\alpha \in (0, 1]$ and the boundedness of U. □

Now we begin to prove Theorem 3.7.

Proof. By Assumption 3.5, there is a constant $L_U < \infty$ such that

$$\|\bar{P}_v - \bar{P}_w\|_{\mathrm{op}} \le L_U \|v - w\|_2, \qquad v, w \in U.$$

Throughout the proof, all constants may depend on T, U, and the constants in Assumption 3.5.

We first **bound β_t and the drift of θ_t** over the block.

Equation 24 gives the exact recurrence

$$\beta_{t+1} = \beta_t + 2\eta_1 \bar{P}_{\theta_t} f - 2\eta_1 \bar{P}_{\theta_t} \bar{P}_{\theta_t}^\top \beta_t. \tag{31}$$

Scalar normalization implies $\|\bar{P}_v\|_{\mathrm{op}}^2 \le \|\bar{P}_v \bar{P}_v^\top\|_{\mathrm{HS}} = 1$. At every differentiable point θ_t, local Lipschitz continuity gives

$$\left(\sum_{j=1}^{D} \|\bar{A}_j(\theta_t)\|_{\mathrm{op}}^2 \right)^{1/2} \le \sqrt{D}\, L_U =: B_A.$$

Taking norms in (31) yields

$$\|\beta_{t+1}\|_2 \le (1 + 2\eta_1)\|\beta_t\|_2 + 2\eta_1 \|f\|_2.$$

Since $\beta_0 = 0$ and T is fixed, iterating this inequality gives

$$\max_{0 \le t \le T} \|\beta_t\|_2 \le C_T \eta_1. \tag{32}$$

Using (25) and the preceding bounds, we get

$$\begin{aligned} \|\nabla_\theta \bar{L}(\theta_t, \beta_{t+1})\|_2 &\le 2B_A \|\beta_{t+1}\|_2 \big(\|f\|_2 + \|\beta_{t+1}\|_2\big) \\ &\le C_T' \eta_1. \end{aligned}$$

Therefore

$$\|\theta_{t+1} - \theta_t\|_2 \le C_T' \eta_1 \eta_2. \tag{33}$$

Summing over at most T steps gives $\|\theta_t - \theta_0\|_2 \le T C_T' \eta_1 \eta_2$. Since all iterates lie in U, the common Lipschitz bound gives

$$\|\bar{P}_{\theta_t} - \bar{P}_{\theta_0}\|_{\mathrm{op}} \le L_U \|\theta_t - \theta_0\|_2 = O_T(\eta_1 \eta_2). \tag{34}$$

Since U is open and $a(\theta_0) \neq 0$, choose $\rho > 0$ such that $\overline{B}(\theta_0, \rho) \subset U$ and $C_a \rho^\alpha \le \|a(\theta_0)\|_2/2$, and set $\gamma := \|a(\theta_0)\|_2/2$. Equation 29 then gives $\|a(v)\|_2 \ge \gamma$ for every $v \in B(\theta_0, \rho) \cap U_{\mathrm{d}}$. By taking the step sizes sufficiently small that $T C_T' \eta_1 \eta_2 < \rho$, the drift bound (33) ensures

$$\theta_t \in B(\theta_0, \rho), \qquad \|a(\theta_t)\|_2 \ge \gamma, \qquad t = 0, \dots, T-1. \tag{35}$$

Next, we **derive the main term of β_t**.

Using (32) and (34) in (31) gives

$$\beta_{t+1} = \beta_t + 2\eta_1 \bar{P}_{\theta_0} f + O_T(\eta_1^2 + \eta_1^2 \eta_2).$$

Summing from 0 to t and using $\beta_0 = 0$, we get

$$\beta_{t+1} = 2(t+1)\eta_1 \bar{P}_{\theta_0} f + O_T(\eta_1^2 + \eta_1^2 \eta_2). \tag{36}$$

Equation (34) also gives

$$2(t+1)\eta_1 (\bar{P}_{\theta_0} - \bar{P}_{\theta_t}) f = O_T(\eta_1^2 \eta_2).$$

Hence (36) can be rewritten as

$$\beta_{t+1} = 2(t+1)\eta_1 \bar{P}_{\theta_t} f + O_T(\eta_1^2 + \eta_1^2 \eta_2). \tag{37}$$

Finally, we **derive the parameter update and the change in Q**.

Substituting (37) into the exact feature-gradient formula (25), we obtain the main term

$$4(t+1)\eta_1 \eta_2 (\bar{P}_{\theta_t} f)^\top \bar{A}_j(\theta_t) f = 4(t+1)\eta_1 \eta_2 a_j(\theta_t).$$

The error in (37) contributes $O_T(\eta_1^2 \eta_2)$, and the term quadratic in β_{t+1} has the same order. Thus, uniformly for $t = 0, \dots, T-1$,

$$\theta_{t+1,j} - \theta_{t,j} = 4(t+1)\eta_1 \eta_2 a_j(\theta_t) + O_T(\eta_1^2 \eta_2). \tag{38}$$

Write $\Delta_t := \theta_{t+1} - \theta_t$. By (33), $\|\Delta_t\|_2 \le C_T \eta_1 \eta_2$. Lemma A.1 gives

$$(\bar{P}_{\theta_{t+1}} - \bar{P}_{\theta_t}) f = \sum_{j=1}^{D} \Delta_{t,j} \bar{A}_j(\theta_t) f + r_{P,t}, \qquad \|r_{P,t}\|_2 \le C_T (\eta_1 \eta_2)^{1+\alpha}. \tag{39}$$

Since $Q(v, f) = \|\bar{P}_v f\|_2^2$, direct expansion gives

$$\begin{aligned} &Q(\theta_{t+1}, f) - Q(\theta_t, f) \\ &= 2 \left\langle \bar{P}_{\theta_t} f, (\bar{P}_{\theta_{t+1}} - \bar{P}_{\theta_t}) f \right\rangle + \|(\bar{P}_{\theta_{t+1}} - \bar{P}_{\theta_t}) f\|_2^2. \end{aligned} \tag{40}$$

For the first term, substituting (39) and using the definition of $a(\theta_t)$ yields

$$\begin{aligned} &2 \left\langle \bar{P}_{\theta_t} f, (\bar{P}_{\theta_{t+1}} - \bar{P}_{\theta_t}) f \right\rangle \\ &= 2a(\theta_t)^\top \Delta_t + 2\langle \bar{P}_{\theta_t} f, r_{P,t} \rangle. \end{aligned} \tag{41}$$

Scalar normalization gives $\|\bar{P}_{\theta_t} f\|_2 \le \|f\|_2$, so

$$|2\langle \bar{P}_{\theta_t} f, r_{P,t} \rangle| \le C_T (\eta_1 \eta_2)^{1+\alpha}. \tag{42}$$

For the second term, Assumption 3.5 gives

$$\|(\bar{P}_{\theta_{t+1}} - \bar{P}_{\theta_t}) f\|_2^2 \le L_U^2 \|f\|_2^2 \|\Delta_t\|_2^2 \le C_T (\eta_1 \eta_2)^2. \tag{43}$$

Finally, substituting (38) into (41) and combining the preceding bounds, we obtain

$$\begin{aligned} Q(\theta_{t+1}, f) - Q(\theta_t, f) = 8(t+1)\eta_1 \eta_2 \|a(\theta_t)\|_2^2 \\ + O_T(\eta_1^2 \eta_2) + O_T\big((\eta_1 \eta_2)^{1+\alpha}\big) + O_T\big((\eta_1 \eta_2)^2\big). \end{aligned} \tag{44}$$

By (35), the leading term is positive and dominates the remainder terms when $\eta_1, \eta_2 > 0$ are sufficiently small. This proves the claim for one block. For any fixed finite number of repeated-reset blocks at whose initial points a is nonzero, the same argument applies to each block and hence to the resulting finite horizon. □

A.4 Proof of Theorem 3.12

We first give a quantitative uniform concentration lemma. Its Hilbert-space form allows $\mathcal{H}_p = \ell^2$. For $N \ge 1$ and $\delta \in (0, 1)$, define

$$\begin{aligned} R_U &:= \max\{1, \operatorname{diam}(U)\}, & \tau_N &:= \min\{1, N^{-1/(2\alpha)}\}, \\ \Lambda_N(\delta) &:= D \log\Big(1 + 2R_U N^{1/(2\alpha)}\Big) + \log \frac{4}{\delta}, & \mathfrak{r}_N(\delta) &:= \sqrt{\frac{\Lambda_N(\delta)}{N}} + \frac{\Lambda_N(\delta)}{N}. \end{aligned} \tag{45}$$

Lemma A.2 (Quantitative uniform concentration). *Let $V \subseteq U$ be a bounded subset of $\mathbb{R}^D$, let $Z_1, \dots, Z_N$ be independent copies of Z, and let $\{X_v(Z) : v \in V\}$ be measurable random elements of a separable Hilbert space $\mathcal{H}$. Define the measurable local oscillation*

$$\Omega_{v,\tau}^X(Z) := \sup_{\substack{w \in V \\ \|w - v\|_2 \le \tau}} \|X_w(Z) - X_v(Z)\|_{\mathcal{H}}.$$

Suppose that, for some $B_X, L_X < \infty$ and $\alpha \in (0, 1]$,

$$\sup_{v \in V} \big\| \|X_v(Z)\|_{\mathcal{H}} \big\|_{\psi_1} \le B_X, \qquad \sup_{\substack{v \in V \\ 0 < \tau \le 1}} \|\Omega_{v,\tau}^X(Z)\|_{\psi_1} \le B_X,$$

and

$$\sup_{v \in V} \mathbb{E}\Omega_{v,\tau}^X(Z) \le L_X \tau^\alpha, \qquad \tau \in (0, 1].$$

For $\tau \in (0, 1]$ and $\delta \in (0, 1)$, set

$$\Gamma_N(\tau, \delta) := D \log\left(1 + \frac{2R_U}{\tau}\right) + \log \frac{4}{\delta}.$$

Then, with probability at least $1-\delta$*,*

$$\sup_{v\in V}\left\|\frac{1}{N}\sum_{i=1}^{N}X_v(Z_i)-\mathbb{E}X_v(Z)\right\|_{\mathcal{H}}\le 2L_X\tau^{\alpha}+CB_X\left\{\sqrt{\frac{\Gamma_N(\tau,\delta)}{N}}+\frac{\Gamma_N(\tau,\delta)}{N}\right\}. \tag{46}$$

In particular, taking $\tau=\tau_N$ *derives*

$$\sup_{v\in V}\left\|\frac{1}{N}\sum_{i=1}^{N}X_v(Z_i)-\mathbb{E}X_v(Z)\right\|_{\mathcal{H}}\le C(L_X+B_X)\mathfrak{r}_N(\delta) \tag{47}$$

with probability at least $1-\delta$*, where* $\mathfrak{r}_N$ *is defined in (45).*

Proof. Choose a τ-net $\{v_1,\dots,v_{N_\tau}\}\subset V$. We have

$$N_\tau\le\left(1+\frac{2R_U}{\tau}\right)^{D}. \tag{48}$$

The ψ_1 envelope implies, for the centered Hilbert-valued variable $X_{v_k}(Z)-\mathbb{E}X_{v_k}(Z)$ and every integer $q\ge 2$, a Bernstein moment condition

$$\mathbb{E}\|X_{v_k}(Z)-\mathbb{E}X_{v_k}(Z)\|_{\mathcal{H}}^{q}\le\frac{q!}{2}V_XB_0^{q-2},\qquad V_X\le CB_X^2,\quad B_0\le CB_X.$$

The Hilbert-space Bernstein inequality therefore gives, with probability at least $1-\delta/2$,

$$\max_{k\le N_\tau}\left\|\frac{1}{N}\sum_{i=1}^{N}X_{v_k}(Z_i)-\mathbb{E}X_{v_k}(Z)\right\|_{\mathcal{H}}\le CB_X\left\{\sqrt{\frac{\Gamma_N(\tau,\delta)}{N}}+\frac{\Gamma_N(\tau,\delta)}{N}\right\}. \tag{49}$$

The centered scalar variables $\Omega^X_{v_k,\tau}(Z)-\mathbb{E}\Omega^X_{v_k,\tau}(Z)$ have ψ_1 norm at most CB_X. Scalar Bernstein and another union bound give, with probability at least $1-\delta/2$,

$$\max_{k\le N_\tau}\left|\frac{1}{N}\sum_{i=1}^{N}\Omega^X_{v_k,\tau}(Z_i)-\mathbb{E}\Omega^X_{v_k,\tau}(Z)\right|\le CB_X\left\{\sqrt{\frac{\Gamma_N(\tau,\delta)}{N}}+\frac{\Gamma_N(\tau,\delta)}{N}\right\}. \tag{50}$$

For every $v\in V$, choose v_k with $\|v-v_k\|_2\le\tau$. On the intersection of these two events, the triangle inequality gives

$$\begin{aligned}\left\|\frac{1}{N}\sum_{i=1}^{N}X_v(Z_i)-\mathbb{E}X_v(Z)\right\|_{\mathcal{H}}\le{}&\left\|\frac{1}{N}\sum_{i=1}^{N}X_{v_k}(Z_i)-\mathbb{E}X_{v_k}(Z)\right\|_{\mathcal{H}}\\&+\frac{1}{N}\sum_{i=1}^{N}\Omega^X_{v_k,\tau}(Z_i)+\mathbb{E}\Omega^X_{v_k,\tau}(Z).\end{aligned}$$

Equations 49 and 50, together with $\mathbb{E}\Omega^X_{v_k,\tau}\le L_X\tau^{\alpha}$, prove (46). For $\tau=\tau_N$, the oscillation term is $N^{-1/2}$ and is absorbed by the square-root term in (45), proving (47). □

We now give the main lemmas needed for the proof of Theorem 3.12.

Lemma A.3 (Uniform empirical approximations)**.** *Under the assumptions and notation of Theorem 3.12, there are constants* $C, C_T\ge 1$ *with the following property. Set*

$$\varepsilon:=\min\left\{\frac{r^4}{2},1,\frac{1}{C_T}\right\}\bar{\gamma}^2.$$

If the sample-size condition (19) holds with the constant C, then there is an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ such that

$$\sup_{v\in U_{\mathrm{d}}} \left(|\widehat{h}_v - h_v| + \|\nabla_v \widehat{h}_v - \nabla_v h_v\|_2 \right) \leq \varepsilon, \tag{51}$$

$$\sup_{v\in U_{\mathrm{d}}} \operatorname*{ess\,sup}_x \left[\|\widehat{\bar{\Phi}}_v(x) - \bar{\Phi}_v(x)\|_2 + \left(\sum_{j=1}^{D} \|\widehat{g}_j^v(x) - g_j^v(x)\|_2^2 \right)^{1/2} \right] \leq C_T \varepsilon, \tag{52}$$

$$\sup_{v\in U_{\mathrm{d}}} \left\| \frac{1}{n} \sum_{i=1}^{n} y_i \widehat{\bar{\Phi}}_v(x_i) - \bar{P}_v f \right\|_2 \leq C_T \varepsilon, \tag{53}$$

$$\sup_{v\in U_{\mathrm{d}}} \left(\sum_{j=1}^{D} \left\| \frac{1}{n} \sum_{i=1}^{n} y_i \widehat{g}_j^v(x_i) - \bar{A}_j(v) f \right\|_2^2 \right)^{1/2} \leq C_T \varepsilon. \tag{54}$$

On the same event, the empirical normalized feature map and its derivatives are uniformly bounded:

$$\sup_{v\in U_{\mathrm{d}}} \operatorname*{ess\,sup}_x \left\{ \|\widehat{\bar{\Phi}}_v(x)\|_2 + \left(\sum_{j=1}^{D} \|\widehat{g}_j^v(x)\|_2^2 \right)^{1/2} \right\} \leq C. \tag{55}$$

The constants C and C_T depend only on T and the regularity quantities in the assumptions, and are independent of m, n, D, δ, η_1, and η_2.

The following lemmas are all on $\mathcal{E}$. Also, we may ignore the difference of η_* in different lemmas for brevity. As a result, we shall choose the η_* in Lemma A.6 for the final proof of Theorem 3.12.

Lemma A.4 (Finite-sample head expansion)**.** *Suppose the assumptions and sample-size condition of Lemma A.3 hold. There exists $\eta_* > 0$ such that for $0 < \eta_1, \eta_2 \leq \eta_*$ and $t = 0, \ldots, T-1$,*

$$\begin{aligned} &\widehat{\theta}_t, \widehat{\theta}_{t+1} \in U_{\mathrm{d}}, \\ &\|a(\widehat{\theta}_t)\|_2 \geq \bar{\gamma}, \\ &\widehat{\beta}_{t+1} = 2(t+1)\eta_1 \bar{P}_{\widehat{\theta}_t} f + \widetilde{r}_{\beta,t}, \\ &\|\widetilde{r}_{\beta,t}\|_2 \leq C_T \left\{ \eta_1^2 + \eta_1^2 \eta_2 + \eta_1 \varepsilon \right\}. \end{aligned} \tag{56, 57}$$

Lemma A.5 (Finite-sample feature-parameter expansion)**.** *Suppose the assumptions and sample-size condition of Lemma A.3 hold. There exists $\eta_* > 0$ such that for $0 < \eta_1, \eta_2 \leq \eta_*$ and $t = 0, \ldots, T-1$,*

$$\begin{aligned} \widehat{\theta}_{t+1} - \widehat{\theta}_t &= 4(t+1)\eta_1 \eta_2 a(\widehat{\theta}_t) + r_{\theta,t}, \\ \|r_{\theta,t}\|_2 &\leq C_T \eta_1 \eta_2 \left\{ \eta_1 + \varepsilon \right\}, \end{aligned} \tag{58}$$

Lemma A.6 (Finite-sample alignment expansion)**.** *Suppose the assumptions and sample-size condition of Lemma A.3 hold. There exists $\eta_* > 0$ such that for $0 < \eta_1, \eta_2 \leq \eta_*$ and $t = 0, \ldots, T-1$,*

$$\begin{aligned} &\left| Q(\widehat{\theta}_{t+1}, f) - Q(\widehat{\theta}_t, f) - 8(t+1)\eta_1 \eta_2 \|a(\widehat{\theta}_t)\|_2^2 \right| \\ &\leq C_T \eta_1 \eta_2 \left\{ \eta_1 + \varepsilon + (\eta_1 \eta_2)^{\alpha} + \eta_1 \eta_2 \right\}. \end{aligned} \tag{59}$$

Moreover, η_* can be chosen so that

$$C_T \{ \eta_* + \eta_*^{2\alpha} + \eta_*^2 \} \leq \bar{\gamma}^2. \tag{60}$$

Proof of Lemma A.3. Let U_{d} be the set of differentiable points in U. Throughout this proof, C denotes constants depending only on the regularity quantities, and $C_T \geq 1$ is fixed large enough. Neither depends on m, n, D, δ, η_1, or η_2. For $v \in U_{\mathrm{d}}$, write

$$u(v) := \bar{P}_v f = \mathbb{E}[f^*(x) \bar{\Phi}_v(x)], \qquad b_j(v) := \bar{A}_j(v) f = \mathbb{E}[f^*(x) g_j^v(x)].$$

We first control the raw normalizer and its derivative. Specifically, we bound

$$\sup_{v\in U_{\mathrm{d}}} \left\{ |\widehat{h}_v - h_v| + \|\nabla_v \widehat{h}_v - \nabla_v h_v\|_2 \right\},$$

the common normalization error that will also be used for the feature-parameter update.

For $v \in U_{\mathrm{d}}$, denote

$$J_v(x) := (\partial_{v_j}\Phi_v(x))_{j=1}^D \in \mathcal{H}_p^D, \qquad \|J_v(x)\|_{\mathcal{H}_p^D} = \left(\sum_{j=1}^D \|\partial_{v_j}\Phi_v(x)\|_2^2\right)^{1/2},$$

and denote

$$\kappa_v(z,z') := \langle \Phi_v(z), \Phi_v(z')\rangle, \qquad H_v(z,z') := \kappa_v(z,z')^2, \qquad \dot{H}_v(z,z') := \nabla_v H_v(z,z').$$

Notice that $|\kappa_v(z,z')| \le R^2$. If $w \in U_{\mathrm{d}}$ and $\|w-v\|_2 \le \tau$, we have

$$\begin{aligned}|\kappa_w(z,z') - \kappa_v(z,z')| &\le \|\Phi_w(z)-\Phi_v(z)\|_2\|\Phi_w(z')\|_2 \\ &\quad + \|\Phi_v(z)\|_2\|\Phi_w(z')-\Phi_v(z')\|_2 \\ &\le R\{\Omega_{v,\tau}(z)+\Omega_{v,\tau}(z')\}.\end{aligned}$$

For $j = 1, \dots, D$, we have

$$\partial_{v_j}\kappa_v(z,z') = \langle \partial_{v_j}\Phi_v(z), \Phi_v(z')\rangle + \langle \Phi_v(z), \partial_{v_j}\Phi_v(z')\rangle.$$

Consequently,

$$\|\nabla_v \kappa_v(z,z')\|_2 \le R\|J_v(z)\|_{\mathcal{H}_p^D} + R\|J_v(z')\|_{\mathcal{H}_p^D} \le 2MR.$$

For each j, we have

$$\begin{aligned}&\partial_{w_j}\kappa_w(z,z') - \partial_{v_j}\kappa_v(z,z') \\ &= \left\langle \partial_{w_j}\Phi_w(z) - \partial_{v_j}\Phi_v(z), \Phi_w(z')\right\rangle + \left\langle \partial_{v_j}\Phi_v(z), \Phi_w(z')-\Phi_v(z')\right\rangle \\ &+ \left\langle \Phi_w(z)-\Phi_v(z), \partial_{w_j}\Phi_w(z')\right\rangle + \left\langle \Phi_v(z), \partial_{w_j}\Phi_w(z')-\partial_{v_j}\Phi_v(z')\right\rangle.\end{aligned}$$

Taking the Euclidean norm over $j = 1, \dots, D$ and using $\|\Phi_v(x)\|_2 \le R$ and $\|J_v(x)\|_{\mathcal{H}_p^D} \le M$, we get

$$\begin{aligned}&\|\nabla_w\kappa_w(z,z') - \nabla_v\kappa_v(z,z')\|_2 \\ &\le R\|J_w(z) - J_v(z)\|_{\mathcal{H}_p^D} + M\|\Phi_w(z') - \Phi_v(z')\|_2 \\ &+ M\|\Phi_w(z) - \Phi_v(z)\|_2 + R\|J_w(z') - J_v(z')\|_{\mathcal{H}_p^D}\end{aligned}$$

By the definition of $\Omega_{v,\tau}$, both $\|\Phi_w(x)-\Phi_v(x)\|_2$ and $\|J_w(x)-J_v(x)\|_{\mathcal{H}_p^D}$ are bounded by $\Omega_{v,\tau}(x)$ for $x = z, z'$. Hence

$$\|\nabla_w\kappa_w(z,z') - \nabla_v\kappa_v(z,z')\|_2 \le (R+M)\{\Omega_{v,\tau}(z)+\Omega_{v,\tau}(z')\}.$$

Since

$$\dot{H}_v(z,z') = 2\kappa_v(z,z')\nabla_v\kappa_v(z,z'),$$

we have

$$\begin{aligned}&|H_w - H_v| = |\kappa_w - \kappa_v|\,|\kappa_w + \kappa_v|, \\ &\|\dot{H}_w - \dot{H}_v\|_2 \le 2|\kappa_w-\kappa_v|\|\nabla_w\kappa_w\|_2 + 2|\kappa_v|\|\nabla_w\kappa_w - \nabla_v\kappa_v\|_2.\end{aligned}$$

Hence we get

$$\begin{aligned}&|H_w(z,z') - H_v(z,z')| + \|\dot{H}_w(z,z') - \dot{H}_v(z,z')\|_2 \\ &\qquad \le C\{\Omega_{v,\tau}(z)+\Omega_{v,\tau}(z')\}. \end{aligned} \tag{61}$$

Taking the supremum gives

$$\sup_{v\in U_{\mathrm{d}}} \left\| \sup_{\substack{w\in U_{\mathrm{d}} \\ \|w-v\|_2\le\tau}} \left(|H_w - H_v|^2 + \|\dot{H}_w - \dot{H}_v\|_2^2\right)^{1/2} \right\|_{L^2(\rho_X^{\otimes 2})} \le C\tau^\alpha,$$

where Assumption 3.8 was applied separately to z and z'.

We next transfer the oscillation bounds through normalization. Assumption 3.8 controls the unnormalized feature map and its derivatives. We transfer those bounds through the normalization and control the local oscillations $\|\bar{\Phi}_w - \bar{\Phi}_v\|_2$ and $\left(\sum_{j=1}^D \|g_j^w - g_j^v\|_2^2\right)^{1/2}$ of the normalized population feature map and its derivatives. Notice that

$$h_v = \mathbb{E}_{z,z'} H_v(z,z'), \qquad \nabla h_v = \mathbb{E}_{z,z'} \dot{H}_v(z,z'), \qquad \|\nabla h_v\|_2 \le 4MR^3.$$

Integrating (61) shows that, uniformly for $w \in U_{\mathrm{d}}$ with $\|w - v\|_2 \le \tau$,

$$|h_w - h_v| + \|\nabla h_w - \nabla h_v\|_2 \le C\tau^\alpha. \tag{62}$$

Denote

$$c_v := h_v^{-1/4}, \qquad d_v := h_v^{-5/4}, \qquad q_v := \nabla h_v.$$

Because $h_v, h_w \ge r^4$, by simple calculation we get

$$|c_w - c_v| + |d_w - d_v| + \|q_w - q_v\|_2 \le C\tau^\alpha. \tag{63}$$

Writing $q \otimes \Phi := (q_j \Phi)_{j=1}^D \in \mathcal{H}_p^D$, direct differentiation gives

$$\bar{\Phi}_v = c_v \Phi_v, \qquad (g_j^v)_{j=1}^D = c_v J_v - \frac{1}{4} d_v (q_v \otimes \Phi_v).$$

Therefore

$$\begin{aligned}
\|\bar{\Phi}_w(x) - \bar{\Phi}_v(x)\|_2 &\le c_w \|\Phi_w(x) - \Phi_v(x)\|_2 + |c_w - c_v| \|\Phi_v(x)\|_2, \\
\|(g_j^w(x) - g_j^v(x))_{j=1}^D\|_{\mathcal{H}_p^D} &\le c_w \|J_w(x) - J_v(x)\|_{\mathcal{H}_p^D} + |c_w - c_v| \|J_v(x)\|_{\mathcal{H}_p^D} \\
&\quad + \frac{1}{4}|d_w - d_v| \|q_w\|_2 \|\Phi_w(x)\|_2 \\
&\quad + \frac{1}{4} d_v \|q_w - q_v\|_2 \|\Phi_w(x)\|_2 \\
&\quad + \frac{1}{4} d_v \|q_v\|_2 \|\Phi_w(x) - \Phi_v(x)\|_2.
\end{aligned}$$

Hence, we get the following bound:

$$\left\| \sup_{\substack{w \in U_{\mathrm{d}} \\ \|w-v\|_2 \le \tau}} \left(\|\bar{\Phi}_w - \bar{\Phi}_v\|_2^2 + \sum_{j=1}^D \|g_j^w - g_j^v\|_2^2 \right)^{1/2} \right\|_{L^2(\rho_X)} \le C\tau^\alpha. \tag{64}$$

In particular, when $\|w - v\|_2 \le 1$, taking $\tau = \|w - v\|_2$ and retaining only the derivative term gives

$$\left(\sum_{j=1}^D \|g_j^w - g_j^v\|_{L^2(\rho_X;\mathcal{H}_p)}^2 \right)^{1/2} \le C\|w - v\|_2^\alpha, \qquad v, w \in U_{\mathrm{d}}. \tag{65}$$

The pointwise bounds on Φ_v and J_v, together with $h_v \ge r^4$, also give

$$\sup_{v \in U_{\mathrm{d}}} \operatorname*{ess\,sup}_x \|\bar{\Phi}_v(x)\|_2 \le \frac{R}{r}, \qquad \sup_{v \in U_{\mathrm{d}}} \operatorname*{ess\,sup}_x \left(\sum_{j=1}^D \|g_j^v(x)\|_2^2 \right)^{1/2} \le B_g, \tag{66}$$

where one may take

$$B_g := \frac{M}{r} + \frac{MR^4}{r^5}.$$

We then establish uniform concentration of the normalizer. We verify the conditions of Lemma A.2 for $X_v = (H_v, \dot{H}_v)$ in $\mathbb{R} \oplus \mathbb{R}^D$. Since $|H_v(z,z')| \le R^4$ and $\|\dot{H}_v(z,z')\|_2 \le 4MR^3$, both $\|X_v\|_2$ and the local oscillation

$$\sup_{\substack{w \in U_{\mathrm{d}} \\ \|w-v\|_2 \le \tau}} \|X_w - X_v\|_2$$

are deterministically bounded by constants depending only on R and M; their ψ_1 norms are therefore uniformly bounded. Moreover, (61) and Cauchy–Schwarz give

$$\sup_{v\in U_{\mathrm{d}}} \mathbb{E} \sup_{\substack{w\in U_{\mathrm{d}} \\ \|w-v\|_2\le\tau}} \|X_w - X_v\|_2 \le C\tau^{\alpha}.$$

These are the envelope and expected-oscillation conditions in Lemma A.2. Set

$$\varepsilon := \min\left\{\frac{r^4}{2}, 1, \frac{1}{C_T}\right\}\bar{\gamma}^2.$$

With net radius $(\varepsilon/(3C_T))^{1/\alpha}$, Lemma A.2 gives, with probability at least $1-\delta/4$,

$$\begin{aligned}&\sup_{v\in U_{\mathrm{d}}}\left(|\widehat{h}_v - h_v| + \|\nabla_v\widehat{h}_v - \nabla_v h_v\|_2\right)\\ &\le \frac{\varepsilon}{3} + C\left[\sqrt{\frac{D\log(C\bar{\gamma}^{-2/\alpha})+\log(C/\delta)}{m}} + \frac{D\log(C\bar{\gamma}^{-2/\alpha})+\log(C/\delta)}{m}\right] \le \varepsilon, \end{aligned} \tag{67}$$

where the last inequality follows from (19). Denote this event by $\mathcal{E}_{\mathrm{N}}$.

We transfer normalizer concentration to the normalized features. We convert the normalizer estimates into uniform bounds for

$$\sup_{v\in U_{\mathrm{d}}} \operatorname*{ess\,sup}_x \|\widehat{\bar{\Phi}}_v(x) - \bar{\Phi}_v(x)\|_2$$

and

$$\sup_{v\in U_{\mathrm{d}}} \operatorname*{ess\,sup}_x \left(\sum_{j=1}^{D} \|\widehat{g}_j^v(x) - g_j^v(x)\|_2^2\right)^{1/2}.$$

For $v\in U_{\mathrm{d}}$, recall

$$h_v = \|\Sigma_v\|_{\mathrm{HS}}^2 = \mathbb{E}_{z,z'}\langle\Phi_v(z),\Phi_v(z')\rangle^2.$$

Since $\|\Sigma_v\|_{\mathrm{HS}} \ge r^2$, we have $h_v \ge r^4$. On $\mathcal{E}_{\mathrm{N}}$, uniformly over $v\in U_{\mathrm{d}}$,

$$\widehat{h}_v \ge r^4 - \varepsilon.$$

The definition of ε hence implies

$$\inf_{v\in U_{\mathrm{d}}} \widehat{h}_v \ge r^4/2. \tag{68}$$

Direct differentiation gives

$$\partial_{v_j} h_v = 2\mathbb{E}_{z,z'}\left[\kappa_v(z,z')\big(\langle\partial_{v_j}\Phi_v(z),\Phi_v(z')\rangle + \langle\Phi_v(z),\partial_{v_j}\Phi_v(z')\rangle\big)\right].$$

Also,

$$\partial_{v_j}\widehat{h}_v = \frac{2}{m}\sum_{r=1}^{m}\kappa_v(z_r,z_r')\big(\langle\partial_{v_j}\Phi_v(z_r),\Phi_v(z_r')\rangle + \langle\Phi_v(z_r),\partial_{v_j}\Phi_v(z_r')\rangle\big).$$

In particular, $\|\nabla_v h_v\|_2 \le 4MR^3$ and $\|\nabla_v\widehat{h}_v\|_2 \le 4MR^3$.

Since $\bar{\Phi}_v = h_v^{-1/4}\Phi_v$, direct differentiation yields

$$\begin{aligned} g_j^v(x) &= h_v^{-1/4}\partial_{v_j}\Phi_v(x) - \frac{1}{4}h_v^{-5/4}(\partial_{v_j}h_v)\Phi_v(x),\\ \widehat{g}_j^v(x) &:= \partial_{v_j}\widehat{\bar{\Phi}}_v(x)\\ &= \widehat{h}_v^{-1/4}\partial_{v_j}\Phi_v(x) - \frac{1}{4}\widehat{h}_v^{-5/4}(\partial_{v_j}\widehat{h}_v)\Phi_v(x).\end{aligned}$$

On the interval $[r^4/2,\infty)$, the absolute derivatives of $s\mapsto s^{-1/4}$ and $s\mapsto s^{-5/4}$ are bounded by constants that depend only on r. Hence, on $\mathcal{E}_{\mathrm{N}}$, we have

$$\sup_{v\in U_{\mathrm{d}}}\left\{|\widehat{h}_v^{-1/4} - h_v^{-1/4}| + |\widehat{h}_v^{-5/4} - h_v^{-5/4}| + \|\nabla_v\widehat{h}_v - \nabla_v h_v\|_2\right\} \le C_T\varepsilon. \tag{69}$$

The normalized-feature difference is

$$\widehat{\bar{\Phi}}_v(x) - \bar{\Phi}_v(x) = (\widehat{h}_v^{-1/4} - h_v^{-1/4})\Phi_v(x).$$

For the derivative, the difference is

$$\begin{aligned}
&(\widehat{g}_j^v(x) - g_j^v(x))_{j=1}^D \\
&= (\widehat{h}_v^{-1/4} - h_v^{-1/4})J_v(x) \\
&- \frac{1}{4}(\widehat{h}_v^{-5/4} - h_v^{-5/4})(\nabla_v \widehat{h}_v \otimes \Phi_v(x)) \\
&- \frac{1}{4}h_v^{-5/4}((\nabla_v \widehat{h}_v - \nabla_v h_v) \otimes \Phi_v(x)).
\end{aligned}$$

Since $\|\nabla_v h_v\|_2, \|\nabla_v \widehat{h}_v\|_2 \le 4MR^3$, (69) yields

$$\sup_{v \in U_{\mathrm{d}}} \operatorname*{ess\,sup}_x \|\widehat{\bar{\Phi}}_v(x) - \bar{\Phi}_v(x)\|_2 \le C_T\varepsilon,$$

$$\sup_{v \in U_{\mathrm{d}}} \operatorname*{ess\,sup}_x \left(\sum_{j=1}^D \|\widehat{g}_j^v(x) - g_j^v(x)\|_2^2\right)^{1/2} \le C_T\varepsilon. \tag{70}$$

Moreover,

$$\|\widehat{\bar{\Phi}}_v(x)\|_2 \le (r^4/2)^{-1/4}R$$

and

$$\|(\widehat{g}_j^v(x))_{j=1}^D\|_{\mathcal{H}_p^D} \le (r^4/2)^{-1/4}M + \frac{1}{4}(r^4/2)^{-5/4}(4MR^3)R.$$

Thus there is a constant C such that, uniformly over $v \in U_{\mathrm{d}}$,

$$\operatorname*{ess\,sup}_x \left\{\|\widehat{\bar{\Phi}}_v(x)\|_2 + \left(\sum_{j=1}^D \|\widehat{g}_j^v(x)\|_2^2\right)^{1/2}\right\} \le C. \tag{71}$$

We next control the empirical head signal. We first control the population-normalized empirical head-signal error

$$\sup_{v \in U_{\mathrm{d}}} \left\| \frac{1}{n}\sum_{i=1}^n y_i \bar{\Phi}_v(x_i) - \bar{P}_v f \right\|_2.$$

Notice that

$$\mathbb{E}[y\bar{\Phi}_v(x)] = \bar{P}_v f = u(v).$$

Moreover,

$$\mathbb{E}[y^2 \mid x] = (f^*(x))^2 + \mathbb{E}\epsilon^2 \le R_1^2 + CK_\epsilon^2.$$

Equation 64 and conditional Cauchy–Schwarz give

$$\begin{aligned}
&\sup_{v \in U_{\mathrm{d}}} \mathbb{E}\left[|y| \sup_{\substack{w \in U_{\mathrm{d}} \\ \|w-v\|_2 \le \tau}} \|\bar{\Phi}_w(x) - \bar{\Phi}_v(x)\|_2\right] \\
&\le \{R_1^2 + CK_\epsilon^2\}^{1/2} \sup_{v \in U_{\mathrm{d}}} \left\| \sup_{\substack{w \in U_{\mathrm{d}} \\ \|w-v\|_2 \le \tau}} \|\bar{\Phi}_w - \bar{\Phi}_v\|_2 \right\|_{L^2(\rho_X)} \le C\tau^\alpha.
\end{aligned}$$

Since $|f^*| \le R_1$ and $\|\epsilon\|_{\psi_2} \le K_\epsilon$, there is a constant $K_y \le C(R_1 + K_\epsilon)$ such that $\|y\|_{\psi_2} \le K_y$ and hence $\|y\|_{\psi_1} \le CK_y$. For $X_v(x,y) = y\bar{\Phi}_v(x)$, (66) gives

$$\sup_{v \in U_{\mathrm{d}}} \big\| \|X_v(x,y)\|_2 \big\|_{\psi_1} \le \frac{R}{r}\|y\|_{\psi_1} \le C.$$

Its local oscillation satisfies

$$\sup_{\substack{v\in U_{\mathrm{d}}\\0<\tau\le 1}}\left\|\,|y|\sup_{\substack{w\in U_{\mathrm{d}}\\\|w-v\|_2\le\tau}}\|\bar{\Phi}_w(x)-\bar{\Phi}_v(x)\|_2\right\|_{\psi_1}\le\frac{2R}{r}\|y\|_{\psi_1}\le C.$$

Together with the preceding $C\tau^\alpha$ expectation bound, these verify all three conditions of Lemma A.2. Applying the lemma with the same net radius gives, with probability at least $1-\delta/4$,

$$\begin{aligned}&\sup_{v\in U_{\mathrm{d}}}\left\|\frac{1}{n}\sum_{i=1}^n y_i\bar{\Phi}_v(x_i)-u(v)\right\|_2\\&\le\frac{\varepsilon}{3}+C\left[\sqrt{\frac{D\log(C\bar{\gamma}^{-2/\alpha})+\log(C/\delta)}{n}}+\frac{D\log(C\bar{\gamma}^{-2/\alpha})+\log(C/\delta)}{n}\right]\le\varepsilon.\end{aligned}\tag{72}$$

Denote this event by $\mathcal{E}_\mu$.

Moreover, scalar Bernstein applied to the sub-exponential variable $|y|$ shows that the event

$$\mathcal{E}_y:=\left\{\frac{1}{n}\sum_{i=1}^n|y_i|\le\mathbb{E}|y|+CK_y\left[\sqrt{\frac{\log(8/\delta)}{n}}+\frac{\log(8/\delta)}{n}\right]\right\}\tag{73}$$

has probability at least $1-\delta/4$. The sample-size condition makes the bracketed deviation bounded by a fixed constant; hence the empirical average in (73) is also bounded by a fixed constant.

On $\mathcal{E}_\mu$, the population sampling component satisfies

$$\sup_{v\in U_{\mathrm{d}}}\left\|\frac{1}{n}\sum_{i=1}^n y_i\bar{\Phi}_v(x_i)-u(v)\right\|_2\le\varepsilon.$$

On $\mathcal{E}_{\mathrm{N}}\cap\mathcal{E}_y$, the normalization component satisfies, by (70),

$$\begin{aligned}&\sup_{v\in U_{\mathrm{d}}}\left\|\frac{1}{n}\sum_{i=1}^n y_i\{\hat{\bar{\Phi}}_v(x_i)-\bar{\Phi}_v(x_i)\}\right\|_2\\&\le\left(\frac{1}{n}\sum_{i=1}^n|y_i|\right)\sup_{v\in U_{\mathrm{d}}}\operatorname*{ess\,sup}_x\|\hat{\bar{\Phi}}_v(x)-\bar{\Phi}_v(x)\|_2\le C_T\varepsilon.\end{aligned}$$

On $\mathcal{E}_{\mathrm{N}}\cap\mathcal{E}_\mu\cap\mathcal{E}_y$, adding these two components gives

$$\begin{aligned}&\sup_{v\in U_{\mathrm{d}}}\left\|\frac{1}{n}\sum_{i=1}^n y_i\hat{\bar{\Phi}}_v(x_i)-u(v)\right\|_2\\&\le C_T\varepsilon.\end{aligned}\tag{74}$$

We then control the empirical feature-gradient signal. For every $j=1,\dots,D$,

$$\mathbb{E}[yg_j^v(x)]=\bar{A}_j(v)f=b_j(v).$$

The derivative part of (64) gives

$$\begin{aligned}&\sup_{v\in U_{\mathrm{d}}}\mathbb{E}\left[|y|\sup_{\substack{w\in U_{\mathrm{d}}\\\|w-v\|_2\le\tau}}\left(\sum_{j=1}^D\|g_j^w(x)-g_j^v(x)\|_2^2\right)^{1/2}\right]\\&\le\{R_1^2+CK_\epsilon^2\}^{1/2}\sup_{v\in U_{\mathrm{d}}}\left\|\sup_{\substack{w\in U_{\mathrm{d}}\\\|w-v\|_2\le\tau}}\left(\sum_{j=1}^D\|g_j^w-g_j^v\|_2^2\right)^{1/2}\right\|_{L^2(\rho_X)}\le C\tau^\alpha.\end{aligned}$$

By (66), the norm of $(yg_j^v(x))_{j=1}^D$ in $\mathcal{H}_p^D$ is at most $B_g|y|$. Therefore, for $X_v(x,y) = (yg_j^v(x))_{j=1}^D$,

$$\sup_{v\in U_{\mathrm{d}}} \left\| \|X_v(x,y)\|_{\mathcal{H}_p^D} \right\|_{\psi_1} \le B_g \|y\|_{\psi_1} \le C.$$

Its local oscillation satisfies

$$\sup_{\substack{v\in U_{\mathrm{d}}\\ 0<\tau\le 1}} \left\| |y| \sup_{\substack{w\in U_{\mathrm{d}}\\ \|w-v\|_2\le\tau}} \left(\sum_{j=1}^D \|g_j^w(x) - g_j^v(x)\|_2^2 \right)^{1/2} \right\|_{\psi_1} \le 2B_g\|y\|_{\psi_1} \le C.$$

Together with the preceding $C\tau^\alpha$ expectation bound, these verify all three conditions of Lemma A.2. Applying the lemma in the separable Hilbert space $\mathcal{H}_p^D$, with the same net radius, gives, with probability at least $1-\delta/4$,

$$\begin{aligned}
&\sup_{v\in U_{\mathrm{d}}} \left(\sum_{j=1}^D \left\| \frac{1}{n}\sum_{i=1}^n y_i g_j^v(x_i) - b_j(v) \right\|_2^2 \right)^{1/2} \\
&\le \frac{\varepsilon}{3} + C\left[\sqrt{\frac{D\log(C\bar\gamma^{-2/\alpha}) + \log(C/\delta)}{n}} + \frac{D\log(C\bar\gamma^{-2/\alpha}) + \log(C/\delta)}{n} \right] \le \varepsilon. \qquad (75)
\end{aligned}$$

Denote this event by $\mathcal{E}_g$. Define

$$e_{g,j}(v) := \frac{1}{n}\sum_{i=1}^n y_i \widehat{g}_j^v(x_i) - b_j(v).$$

On $\mathcal{E}_{\mathrm{N}} \cap \mathcal{E}_g \cap \mathcal{E}_y$, the direct-sum triangle inequality gives

$$\begin{aligned}
&\sup_{v\in U_{\mathrm{d}}} \left(\sum_{j=1}^D \|e_{g,j}(v)\|_2^2 \right)^{1/2} \\
&\le \varepsilon + \left(\frac{1}{n}\sum_{i=1}^n |y_i| \right) \sup_{v\in U_{\mathrm{d}}} \operatorname*{ess\,sup}_x \left(\sum_{j=1}^D \|\widehat{g}_j^v(x) - g_j^v(x)\|_2^2 \right)^{1/2} \\
&\le C_T\varepsilon. \qquad (76)
\end{aligned}$$

Consequently, at every iterate $\widehat{\theta}_t$, the first factor has the expansion

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n y_i \widehat{g}_{j,t}(x_i) &= \bar{A}_j(\widehat{\theta}_t) f + e_{g,j}(\widehat{\theta}_t), \\
\left(\sum_{j=1}^D \|e_{g,j}(\widehat{\theta}_t)\|_2^2 \right)^{1/2} &\le C_T\varepsilon. \qquad (77)
\end{aligned}$$

Thus the first factor has main part $\bar{A}_j(\widehat{\theta}_t)f$ and remainder $e_{g,j}(\widehat{\theta}_t)$.

We finally assemble the common high-probability event. Define

$$\mathcal{E} := \mathcal{E}_{\mathrm{N}} \cap \mathcal{E}_\mu \cap \mathcal{E}_g \cap \mathcal{E}_y. \qquad (78)$$

The four-event union bound gives $\mathbb{P}(\mathcal{E}) \ge 1-\delta$. Equations (67), (70), (74), (76), and (71) are precisely the bounds stated in Lemma A.3. This completes the proof of Lemma A.3. □

Proof of Lemma A.4. Let U_{d} be the set of differentiable points in U. Throughout this proof, C denotes constants depending only on the regularity quantities, and $C_T \ge 1$ is fixed large enough. Neither depends on m, n, D, δ, η_1, or η_2.

Local alignment margin and notation. Define

$$\rho := \min\left\{1, \frac{\operatorname{dist}(\widehat{\theta}_0, \mathbb{R}^D \setminus U)}{2}, \left(\frac{a_0}{2C_a}\right)^{1/\alpha}\right\},$$

where C_a is the Hölder constant in Lemma A.1. By (29), every $v \in B(\widehat{\theta}_0, \rho) \cap U_{\mathrm{d}}$ satisfies

$$\|a(v)\|_2 \geq a_0 - C_a\|v - \widehat{\theta}_0\|_2^{\alpha} \geq \frac{a_0}{2} \geq \bar{\gamma}. \tag{79}$$

Write

$$\widehat{\phi}_t(x) := \widehat{\bar{\Phi}}_{\widehat{\theta}_t}(x), \qquad \widehat{g}_{j,t}(x) := \widehat{g}_j^{\widehat{\theta}_t}(x).$$

For $v \in U_{\mathrm{d}}$, put

$$u(v) := \bar{P}_v f = \mathbb{E}[f^*(x)\bar{\Phi}_v(x)], \qquad b_j(v) := \bar{A}_j(v)f = \mathbb{E}[f^*(x)g_j^v(x)].$$

Head expansion. Differentiating the empirical loss with respect to the head gives

$$\widehat{\beta}_{t+1} = \widehat{\beta}_t + \frac{2\eta_1}{n}\sum_{i=1}^n y_i\widehat{\phi}_t(x_i) - \frac{2\eta_1}{n}\sum_{i=1}^n \widehat{\phi}_t(x_i)\langle\widehat{\phi}_t(x_i), \widehat{\beta}_t\rangle. \tag{80}$$

The expansion can be written as

$$\widehat{\beta}_{t+1} - \widehat{\beta}_t = 2\eta_1 \bar{P}_{\widehat{\theta}_t} f + r_{\beta,t}^{\mathrm{inc}}, \tag{81}$$

where

$$r_{\beta,t}^{\mathrm{inc}} := 2\eta_1\left\{\frac{1}{n}\sum_{i=1}^n y_i\widehat{\phi}_t(x_i) - \bar{P}_{\widehat{\theta}_t} f\right\} - \frac{2\eta_1}{n}\sum_{i=1}^n \widehat{\phi}_t(x_i)\langle\widehat{\phi}_t(x_i), \widehat{\beta}_t\rangle. \tag{82}$$

On the event $\mathcal{E}$, the first term in (82) is controlled by (53).

For the **second** term, the uniform feature bound in (55) gives

$$\left\|\frac{2\eta_1}{n}\sum_{i=1}^n \widehat{\phi}_t(x_i)\langle\widehat{\phi}_t(x_i), \widehat{\beta}_t\rangle\right\|_2 \leq \frac{2\eta_1}{n}\sum_{i=1}^n \|\widehat{\phi}_t(x_i)\|_2^2\|\widehat{\beta}_t\|_2 \leq C\eta_1\|\widehat{\beta}_t\|_2. \tag{83}$$

Combining this bound with (53) in the exact decomposition (82) gives, on $\mathcal{E}$,

$$\|r_{\beta,t}^{\mathrm{inc}}\|_2 \leq C_T\eta_1\varepsilon + C_T\eta_1\|\widehat{\beta}_t\|_2. \tag{84}$$

Since $\|\bar{P}_{\widehat{\theta}_t} f\|_2 \leq \|f\|_2$, equations (81) and (84) yield

$$\|\widehat{\beta}_{t+1}\|_2 \leq (1 + C\eta_1)\|\widehat{\beta}_t\|_2 + C\eta_1$$

where $\varepsilon \leq 1$ by its definition. Starting from $\widehat{\beta}_0 = 0$ and applying this inequality successively for $t = 0, \ldots, T-1$ gives

$$\max_{0\leq t\leq T}\|\widehat{\beta}_t\|_2 \leq C_T\eta_1. \tag{85}$$

Substitution into (84) gives

$$\|r_{\beta,t}^{\mathrm{inc}}\|_2 \leq C_T\{\eta_1\varepsilon + \eta_1^2\}.$$

Define $r_{\beta,t} := \sum_{s=0}^t r_{\beta,s}^{\mathrm{inc}}$. Summing (81) from $s = 0$ to t therefore gives

$$\widehat{\beta}_{t+1} = 2\eta_1\sum_{s=0}^t \bar{P}_{\widehat{\theta}_s} f + r_{\beta,t}, \qquad \|r_{\beta,t}\|_2 \leq C_T\{\eta_1\varepsilon + \eta_1^2\}. \tag{86}$$

A rough feature-parameter drift. Differentiating the empirical loss gives the following exact increment, separated into two terms:

$$\widehat{\theta}_{t+1} - \widehat{\theta}_t = \underbrace{2\eta_2\left(\left\langle\frac{1}{n}\sum_{i=1}^n y_i\widehat{g}_{j,t}(x_i), \widehat{\beta}_{t+1}\right\rangle\right)_{j=1}^D}_{\text{signal term}}$$
$$\underbrace{-\frac{2\eta_2}{n}\sum_{i=1}^n\left(\langle\widehat{\phi}_t(x_i), \widehat{\beta}_{t+1}\rangle\langle\widehat{g}_{j,t}(x_i), \widehat{\beta}_{t+1}\rangle\right)_{j=1}^D}_{\text{small term}}. \tag{87}$$

The signal term is an inner product. To identify its leading direction, it suffices to control its two factors

$$\frac{1}{n}\sum_{i=1}^{n} y_i \widehat{g}_{j,t}(x_i) \qquad \text{and} \qquad \widehat{\beta}_{t+1}.$$

We establish the required main-term-plus-remainder expansion for each factor in turn.

Refining the head expansion. To express $\widehat{\beta}_{t+1}$ using the current feature operator $\bar{P}_{\widehat{\theta}_t}$, we first need a rough bound on the parameter drift. The uniform gradient and feature bounds give the following estimates for the two terms in (87):

$$\left\| 2\eta_2 \left(\left\langle \frac{1}{n}\sum_{i=1}^{n} y_i \widehat{g}_{j,t}(x_i), \widehat{\beta}_{t+1} \right\rangle \right)_{j=1}^{D} \right\|_2 \le C\eta_2 \|\widehat{\beta}_{t+1}\|_2,$$

$$\left\| \frac{2\eta_2}{n}\sum_{i=1}^{n} \left(\langle \widehat{\phi}_t(x_i), \widehat{\beta}_{t+1}\rangle \langle \widehat{g}_{j,t}(x_i), \widehat{\beta}_{t+1}\rangle \right)_{j=1}^{D} \right\|_2 \le C\eta_2 \|\widehat{\beta}_{t+1}\|_2^2.$$

Combining these inequalities with (85) gives

$$\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|_2 \le C_T \eta_1 \eta_2. \tag{88}$$

Fix $\eta_* > 0$ such that

$$C_T \eta_*^2 \le \frac{\rho}{2}, \qquad C_T\{\eta_* + \eta_*^{2\alpha} + \eta_*^2\} \le \bar{\gamma}^2. \tag{89}$$

We now prove by induction that the iterates remain in $B(\widehat{\theta}_0, \rho)$. This holds at $t = 0$. Suppose it holds through time $t < T$. Because all preceding bounds are uniform over U_{d}, they apply to these iterates and give

$$\begin{aligned} \|\widehat{\theta}_{t+1} - \widehat{\theta}_0\|_2 &\le \sum_{s=0}^{t} \|\widehat{\theta}_{s+1} - \widehat{\theta}_s\|_2 \\ &\le C_T \eta_1 \eta_2 \le C_T \eta_*^2 \le \frac{\rho}{2} < \rho. \end{aligned}$$

Thus $\widehat{\theta}_{t+1}$ also lies in the ball, closing the induction. Since ρ is at most half the distance from $\widehat{\theta}_0$ to $\mathbb{R}^D \setminus U$, this ball is contained in U. Moreover, (79) gives

$$\widehat{\theta}_t \in B(\widehat{\theta}_0, \rho), \qquad \|a(\widehat{\theta}_t)\|_2 \ge \bar{\gamma}, \qquad t = 0, \dots, T-1. \tag{90}$$

Assumption 3.5 and (88) give

$$\begin{aligned} \|\bar{P}_{\widehat{\theta}_s} - \bar{P}_{\widehat{\theta}_t}\|_{\mathrm{op}} &\le L_U \|\widehat{\theta}_s - \widehat{\theta}_t\|_2 \\ &\le L_U \sum_{k=\min\{s,t\}}^{\max\{s,t\}-1} \|\widehat{\theta}_{k+1} - \widehat{\theta}_k\|_2. \end{aligned}$$

Consequently,

$$\|\bar{P}_{\widehat{\theta}_s} - \bar{P}_{\widehat{\theta}_t}\|_{\mathrm{op}} \le C_T |s-t| \eta_1 \eta_2, \qquad 0 \le s, t \le T. \tag{91}$$

The first component needed to replace the past feature operators by the current one is $r_{\beta,t}$, which is controlled in (86). For the second component, (91) gives

$$\begin{aligned} \left\| 2\eta_1 \sum_{s=0}^{t} (\bar{P}_{\widehat{\theta}_s} - \bar{P}_{\widehat{\theta}_t}) f \right\|_2 &\le 2\eta_1 \|f\|_2 \sum_{s=0}^{t} C_T |s-t| \eta_1 \eta_2 \\ &\le C_T \eta_1^2 \eta_2. \end{aligned}$$

Define

$$\widetilde{r}_{\beta,t} := r_{\beta,t} + 2\eta_1 \sum_{s=0}^{t} (\bar{P}_{\widehat{\theta}_s} - \bar{P}_{\widehat{\theta}_t}) f.$$

Combining the two component bounds gives

$$\widehat{\beta}_{t+1} = 2(t+1)\eta_1 \bar{P}_{\widehat{\theta}_t} f + \widetilde{r}_{\beta,t}, \qquad \|\widetilde{r}_{\beta,t}\|_2 \le C_T \left\{\eta_1^2 + \eta_1^2\eta_2 + \eta_1\varepsilon\right\}. \tag{92}$$

This proves that the second factor has main part $2(t+1)\eta_1 \bar{P}_{\widehat{\theta}_t} f$ and remainder $\widetilde{r}_{\beta,t}$. This proves Lemma A.4. □

Proof of Lemma A.5. Work on the event $\mathcal{E}$ from Lemma A.3. Write

$$\widehat{\phi}_t(x) := \widehat{\bar{\Phi}}_{\widehat{\theta}_t}(x), \qquad \widehat{g}_{j,t}(x) := \widehat{g}_j^{\widehat{\theta}_t}(x).$$

The exact feature-parameter update is (87). Its feature-gradient factor is given by (77), and its updated head is given by Lemma A.4.

Combining the two factors in the signal term. Equations (77) and (92) give, for every j,

$$\begin{aligned} &2\eta_2 \left\langle \frac{1}{n}\sum_{i=1}^n y_i \widehat{g}_{j,t}(x_i), \widehat{\beta}_{t+1} \right\rangle \\ &= 4(t+1)\eta_1\eta_2 \left\langle \bar{A}_j(\widehat{\theta}_t) f, \bar{P}_{\widehat{\theta}_t} f \right\rangle + (r_{\mathrm{sig},t})_j. \end{aligned}$$

Collecting the D coordinates, the main term is

$$4(t+1)\eta_1\eta_2 \left(\left\langle \bar{A}_j(\widehat{\theta}_t) f, \bar{P}_{\widehat{\theta}_t} f \right\rangle \right)_{j=1}^D = 4(t+1)\eta_1\eta_2 a(\widehat{\theta}_t).$$

The vector remainder obtained by multiplying out the two factors is

$$\begin{aligned} r_{\mathrm{sig},t} = {} & 4(t+1)\eta_1\eta_2 \left(\left\langle e_{g,j}(\widehat{\theta}_t), \bar{P}_{\widehat{\theta}_t} f \right\rangle \right)_{j=1}^D \\ & + 2\eta_2 \left(\left\langle \bar{A}_j(\widehat{\theta}_t) f + e_{g,j}(\widehat{\theta}_t), \widetilde{r}_{\beta,t} \right\rangle \right)_{j=1}^D. \end{aligned}$$

The first component pairs the remainder of the empirical feature-gradient signal with the main part of the head. By the direct-sum Cauchy–Schwarz inequality and (77),

$$4(t+1)\eta_1\eta_2 \|\bar{P}_{\widehat{\theta}_t} f\|_2 \left(\sum_{j=1}^D \|e_{g,j}(\widehat{\theta}_t)\|_2^2 \right)^{1/2} \le C_T \eta_1\eta_2\varepsilon.$$

The second component contains the head remainder. The same inequality and (92) give

$$\begin{aligned} \left(\sum_{j=1}^D \|\bar{A}_j(\widehat{\theta}_t) f\|_2^2 \right)^{1/2} &\le C, \\ \left(\sum_{j=1}^D \|e_{g,j}(\widehat{\theta}_t)\|_2^2 \right)^{1/2} &\le C_T\varepsilon. \end{aligned}$$

Consequently,

$$\begin{aligned} &2\eta_2 \|\widetilde{r}_{\beta,t}\|_2 \left\{C + C_T\varepsilon\right\} \\ &\le C_T \left\{\eta_1^2\eta_2 + \eta_1\eta_2\varepsilon\right\}. \end{aligned}$$

Here we used $\eta_2 \le 1$ and $\varepsilon \le 1$. Combining the two components yields

$$\|r_{\mathrm{sig},t}\|_2 \le C_T \eta_1\eta_2 \left\{\eta_1 + \varepsilon\right\}.$$

The small term. This term contains two factors of $\widehat{\beta}_{t+1}$. The pointwise bounds in (71) and the head bound (85) give

$$\begin{aligned} &\frac{2\eta_2}{n} \sum_{i=1}^n \|\widehat{\phi}_t(x_i)\|_2 \|\widehat{\beta}_{t+1}\|_2^2 \left(\sum_{j=1}^D \|\widehat{g}_{j,t}(x_i)\|_2^2 \right)^{1/2} \\ &\qquad \le C_T \eta_1^2 \eta_2. \end{aligned}$$

Thus it is strictly smaller than the leading $O_T(\eta_1\eta_2)$ signal term in order.

Now we combine the two terms. Accordingly, define the total parameter remainder by

$$r_{\theta,t} := r_{\mathrm{sig},t} - \frac{2\eta_2}{n}\sum_{i=1}^{n}\big(\langle\widehat{\phi}_t(x_i),\widehat{\beta}_{t+1}\rangle\langle\widehat{g}_{j,t}(x_i),\widehat{\beta}_{t+1}\rangle\big)_{j=1}^{D}.$$

The exact decomposition (87) and the two preceding bounds therefore give

$$\widehat{\theta}_{t+1} - \widehat{\theta}_t = 4(t+1)\eta_1\eta_2 a(\widehat{\theta}_t) + r_{\theta,t}, \qquad \|r_{\theta,t}\|_2 \le C_T\eta_1\eta_2\{\eta_1+\varepsilon\}. \tag{93}$$

This proves Lemma A.5. □

Proof of Lemma A.6. Work on the event $\mathcal{E}$ from Lemma A.3.

Alignment expansion. Since $Q(v,f) = \|\bar{P}_v f\|_2^2$, direct expansion of the difference gives

$$\begin{aligned} &Q(\widehat{\theta}_{t+1},f) - Q(\widehat{\theta}_t,f) \\ &= 2\left\langle \bar{P}_{\widehat{\theta}_t} f, (\bar{P}_{\widehat{\theta}_{t+1}} - \bar{P}_{\widehat{\theta}_t})f\right\rangle + \|(\bar{P}_{\widehat{\theta}_{t+1}} - \bar{P}_{\widehat{\theta}_t})f\|_2^2. \end{aligned} \tag{94}$$

We control the two terms in this exact identity separately.

The first term. By Lemma A.4, both $\widehat{\theta}_t$ and $\widehat{\theta}_{t+1}$ lie in U_{d}. Therefore Lemma A.1, applied directly to these two points, gives

$$(\bar{P}_{\widehat{\theta}_{t+1}} - \bar{P}_{\widehat{\theta}_t})f = \sum_{j=1}^{D}(\widehat{\theta}_{t+1,j} - \widehat{\theta}_{t,j})\bar{A}_j(\widehat{\theta}_t)f + r_{P,t}, \qquad \|r_{P,t}\|_2 \le C_T(\eta_1\eta_2)^{1+\alpha}, \tag{95}$$

where the remainder bound also uses (88). Substitution into the first term of (94) now gives

$$2\left\langle \bar{P}_{\widehat{\theta}_t} f, (\bar{P}_{\widehat{\theta}_{t+1}} - \bar{P}_{\widehat{\theta}_t})f\right\rangle = 2a(\widehat{\theta}_t)^\top(\widehat{\theta}_{t+1} - \widehat{\theta}_t) + R_{Q,t}^{(1)}, \qquad |R_{Q,t}^{(1)}| \le C_T(\eta_1\eta_2)^{1+\alpha}. \tag{96}$$

The second term. Notice that

$$0 \le R_{Q,t}^{(2)} := \|(\bar{P}_{\widehat{\theta}_{t+1}} - \bar{P}_{\widehat{\theta}_t})f\|_2^2 \le L_U^2\|f\|_2^2\|\widehat{\theta}_{t+1} - \widehat{\theta}_t\|_2^2 \le C_T(\eta_1\eta_2)^2. \tag{97}$$

Thus the second term is of strictly smaller order than the leading $O_T(\eta_1\eta_2)$ increment.

Now we combine the two terms. Substituting the parameter expansion (93) into (96), we get

$$\begin{aligned} &Q(\widehat{\theta}_{t+1},f) - Q(\widehat{\theta}_t,f) \\ &= 8(t+1)\eta_1\eta_2\|a(\widehat{\theta}_t)\|_2^2 + 2a(\widehat{\theta}_t)^\top r_{\theta,t} + R_{Q,t}^{(1)} + R_{Q,t}^{(2)}. \end{aligned} \tag{98}$$

The uniform operator bounds give $\sup_{v\in U_{\mathrm{d}}}\|a(v)\|_2 < \infty$. Hence

$$|2a(\widehat{\theta}_t)^\top r_{\theta,t}| \le C_T\eta_1\eta_2\{\eta_1+\varepsilon\}.$$

Combining this bound with (96) and (97) yields

$$\begin{aligned} &\left|Q(\widehat{\theta}_{t+1},f) - Q(\widehat{\theta}_t,f) - 8(t+1)\eta_1\eta_2\|a(\widehat{\theta}_t)\|_2^2\right| \\ &\le C_T\eta_1\eta_2\{\eta_1+\varepsilon+(\eta_1\eta_2)^\alpha+\eta_1\eta_2\}. \end{aligned} \tag{99}$$

Using (90) in (99), we obtain

$$\begin{aligned} &Q(\widehat{\theta}_{t+1},f) - Q(\widehat{\theta}_t,f) \\ &\ge \eta_1\eta_2\left[8(t+1)\bar{\gamma}^2 - C_T\{\eta_1+\varepsilon+(\eta_1\eta_2)^\alpha+\eta_1\eta_2\}\right]. \end{aligned} \tag{100}$$

The remainder estimate (99) proves the alignment expansion. The choice of η_* in (89) also gives (60). This proves Lemma A.6. □

Proof of Theorem 3.12. Apply Lemmas A.4 and A.6. By the definition of ε, $C_T\varepsilon \le \bar{\gamma}^2$. Moreover, (60) gives

$$C_T\{\eta_1 + (\eta_1\eta_2)^\alpha + \eta_1\eta_2\} \le \bar{\gamma}^2.$$

Therefore, (56) and (59) yield

$$Q(\widehat{\theta}_{t+1},f) - Q(\widehat{\theta}_t,f) \ge 4\eta_1\eta_2\bar{\gamma}^2 > 0, \qquad t = 0,\dots,T-1.$$

The event $\mathcal{E}$ has probability at least $1-\delta$, which proves the claim. For repeated-reset blocks, the same one-block argument is applied successively with fresh samples and the corresponding new block-start value of a_0. □

A.5 VERIFICATION FOR FINITE RELU FEATURE MAPS

We now give sufficient conditions under which a finite ReLU network satisfies the sample regularity assumption with $\alpha = 1/2$. Let $d_0 = d$ and consider the depth-L_0 feature network

$$h_0^v(x) = x, \qquad a_\ell^v(x) = W_\ell^v h_{\ell-1}^v(x) + b_\ell^v, \qquad h_\ell^v(x) = \sigma(a_\ell^v(x)), \tag{101}$$

for $\ell = 1, \dots, L_0$, where $\sigma(s) = \max\{s, 0\}$ is applied coordinatewise, $W_\ell^v \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, $b_\ell^v \in \mathbb{R}^{d_\ell}$, and $a_\ell^v(x), h_\ell^v(x) \in \mathbb{R}^{d_\ell}$. We set $\Phi_v = h_{L_0}^v$. The parameter $v \in \mathbb{R}^D$ collects all weight matrices and biases, where

$$D = \sum_{\ell=1}^{L_0} d_\ell(d_{\ell-1} + 1).$$

Write $a_{\ell k}^v$ for the preactivation of unit k in layer ℓ.

Proposition A.7 (ReLU 1/2-Hölder regularity)**.** *Suppose $\|x\|_2 \le R_x < \infty$ almost surely and $U \subset \mathbb{R}^D$ is bounded and convex. Write U_{d} for the differentiable points in U, and, for $v \in U_{\mathrm{d}}$ and $\tau \in (0, 1]$, define the local activation-boundary event*

$$\mathcal{B}_U(v, \tau) := \{x : \ \text{there exist } w \in U,\ \|w - v\|_2 \le \tau, \text{ and } (\ell, k) \text{ such that } a_{\ell k}^w(x) = 0\ \}. \tag{102}$$

Assume that, for some $C_{\mathrm{bd}} < \infty$,

$$\sup_{v \in U_{\mathrm{d}}} \mathbb{P}_x\big(\mathcal{B}_U(v, \tau)\big) \le C_{\mathrm{bd}}\tau, \qquad \tau \in (0, 1]. \tag{103}$$

If, in addition, $\inf_{v \in U} \|\Sigma_v\|_{\mathrm{HS}} \ge r^2 > 0$, then the bounds in Assumption 3.8 hold with $\alpha = 1/2$ at all differentiable points in U.

Remark A.8 (Interpretation of the ReLU conditions). Condition (103) is a standard activation-margin condition. It requires that, under a parameter perturbation of radius τ, the set of inputs for which at least one ReLU gate can cross zero has probability $O(\tau)$. Under such conditions, the input x lies on no activation boundaries almost surely, hence guaranteeing that the samplewise ReLU Jacobian is well defined almost surely. This boundary-mass estimate can be easily verified in related works considering gate changes in ReLU networks, such as Bietti et al. (2022).

Proof. Since U and the input support are bounded, all weight matrices, biases, and hidden states in (101) are uniformly bounded on U. This follows inductively from

$$\|h_\ell^v(x)\|_2 \le \|W_\ell^v\|_{\mathrm{op}} \|h_{\ell-1}^v(x)\|_2 + \|b_\ell^v\|_2.$$

The same induction gives constants $R, L_\Phi < \infty$ such that, writing $\delta = \|v - w\|_2$, uniformly over $v, w \in U$ and almost every x,

$$\|\Phi_v(x)\|_2 \le R, \tag{104}$$

$$\|\Phi_v(x) - \Phi_w(x)\|_2 \le L_\Phi \delta, \tag{105}$$

Let

$$G_\ell^v(x) := \mathrm{diag}\big(\mathbf{1}\{a_{\ell k}^v(x) > 0\} : 1 \le k \le d_\ell\big)$$

be the gate matrix, and the derivative at the origin is chosen as 0. For a parameter direction $\xi = ((\dot W_\ell, \dot b_\ell))_{\ell=1}^{L_0}$, define $J_{0,v}(x)[\xi] = 0$ and recursively define

$$J_{\ell,v}(x)[\xi] = G_\ell^v(x)\left(\dot W_\ell h_{\ell-1}^v(x) + W_\ell^v J_{\ell-1,v}(x)[\xi] + \dot b_\ell\right). \tag{106}$$

We use $J_v(x) := J_{L_0,v}(x)$ as a common version of the feature Jacobian for all (v, x). Since $w = v$ is allowed in (102), (103) and $\tau \downarrow 0$ imply

$$\mathbb{P}_x\big(a_{\ell k}^v(x) = 0 \text{ for some } (\ell, k)\big) = 0, \qquad v \in U_{\mathrm{d}}.$$

Thus $\Phi_v(x)$ is differentiable almost surely at each $v \in U_{\mathrm{d}}$, and we set $J_v(x) = \mathcal{D}_v\Phi_v(x)$ there.

Uniform Jacobian bound. We choose $H, K_W, L_h < \infty$ such that, uniformly over all layers, $v, w \in U$, and almost every x,

$$\|h_\ell^v(x)\|_2 \le H, \qquad \|W_\ell^v\|_{\mathrm{op}} \le K_W, \qquad \|h_\ell^v(x) - h_\ell^w(x)\|_2 \le L_h\|v - w\|_2. \tag{107}$$

For a direction $\xi = ((\dot{W}_\ell, \dot{b}_\ell))_{\ell=1}^{L_0}$ with $\|\xi\|_2 \le 1$, we have

$$\|\dot{W}_\ell\|_{\mathrm{op}} \le \|\dot{W}_\ell\|_{\mathrm{HS}} \le 1, \qquad \|\dot{b}_\ell\|_2 \le 1. \tag{108}$$

Define

$$K_{J,0} := 0, \qquad K_{J,\ell} := H + 1 + K_W K_{J,\ell-1}. \tag{109}$$

Since $\|G_\ell^v(x)\|_{\mathrm{op}} \le 1$, (106), (107), and (108) give

$$\begin{aligned}\|J_{\ell,v}(x)[\xi]\|_2 &\le \|\dot{W}_\ell\|_{\mathrm{op}}\|h_{\ell-1}^v(x)\|_2 + \|W_\ell^v\|_{\mathrm{op}}\|J_{\ell-1,v}(x)[\xi]\|_2 + \|\dot{b}_\ell\|_2 \\ &\le H + K_W\|J_{\ell-1,v}(x)[\xi]\|_2 + 1.\end{aligned}$$

Therefore, we have

$$\sup_{v\in U} \operatorname*{ess\,sup}_x \|J_{\ell,v}(x)\|_{\mathrm{op}} \le K_{J,\ell}. \tag{110}$$

If $e_1, \dots, e_D$ is the standard basis of the parameter space, then

$$\|J_{\ell,v}(x)\|_{\mathrm{HS}}^2 = \sum_{j=1}^{D} \|J_{\ell,v}(x)[e_j]\|_2^2 \le D K_{J,\ell}^2.$$

In particular, with

$$M_J := \sqrt{D} K_{J,L_0}, \qquad \operatorname*{ess\,sup}_x \sup_{v\in U} \|J_v(x)\|_{\mathrm{HS}} \le M_J. \tag{111}$$

Stability away from activation boundaries. Fix $v \in U_{\mathrm{d}}$ and $\tau \in (0, 1]$. By (103),

$$\mathbb{P}_x\big(\mathcal{B}_U(v,\tau)\big) \le C_{\mathrm{bd}}\tau. \tag{112}$$

For $x \notin \mathcal{B}_U(v,\tau)$ and $w \in U$ with $\|w - v\|_2 \le \tau$, we have

$$G_\ell^v(x) = G_\ell^w(x), \qquad \ell = 1, \dots, L_0. \tag{113}$$

On this event, we have

$$\{J_{\ell,w}(x) - J_{\ell,v}(x)\}[\xi] = G_\ell^v(x) \begin{pmatrix} \dot{W}_\ell\{h_{\ell-1}^w(x) - h_{\ell-1}^v(x)\} \\ +(W_\ell^w - W_\ell^v)J_{\ell-1,w}(x)[\xi] \\ +W_\ell^v\{J_{\ell-1,w}(x) - J_{\ell-1,v}(x)\}[\xi] \end{pmatrix}.$$

Taking the supremum over $\|\xi\|_2 \le 1$, we get

$$\begin{aligned}\|J_{\ell,w}(x) - J_{\ell,v}(x)\|_{\mathrm{op}} \le \{L_h + K_{J,\ell-1}\}\|w - v\|_2 \\ + K_W\|J_{\ell-1,w}(x) - J_{\ell-1,v}(x)\|_{\mathrm{op}}.\end{aligned} \tag{114}$$

Define

$$L_{J,0} := 0, \qquad L_{J,\ell} := L_h + K_{J,\ell-1} + K_W L_{J,\ell-1}. \tag{115}$$

Equations (114) and (115) imply inductively that

$$\|J_{\ell,w}(x) - J_{\ell,v}(x)\|_{\mathrm{op}} \le L_{J,\ell}\|w - v\|_2.$$

Setting $\widetilde{L}_J := \sqrt{D} L_{J,L_0}$, we have

$$\sup_{\substack{w\in U_{\mathrm{d}} \\ \|w-v\|_2\le\tau}} \|J_w(x) - J_v(x)\|_{\mathrm{HS}} \le \widetilde{L}_J\tau, \qquad x \notin \mathcal{B}_U(v,\tau). \tag{116}$$

The L^2 oscillation bound. On $\mathcal{B}_U(v,\tau)$, (111) gives

$$\sup_{\substack{w\in U_{\mathrm{d}} \\ \|w-v\|_2\le\tau}} \|J_w(x) - J_v(x)\|_{\mathrm{HS}} \le 2M_J.$$

Splitting the expectation over $\mathcal{B}_U(v,\tau)$ and its complement, we get

$$\begin{aligned}\mathbb{E}_x\left[\sup_{\substack{w\in U_{\mathrm{d}}\\ \|w-v\|_2\le\tau}}\|J_w(x)-J_v(x)\|_{\mathrm{HS}}^2\right]\\ \le \widetilde{L}_J^2\tau^2+4M_J^2\mathbb{P}_x\big(\mathcal{B}_U(v,\tau)\big)\\ \le \widetilde{L}_J^2\tau^2+4M_J^2C_{\mathrm{bd}}\tau. \end{aligned}\tag{117}$$

At differentiable points, we have

$$\|J_w(x)-J_v(x)\|_{\mathrm{HS}}^2=\sum_{j=1}^{D}\|\partial_{w_j}\Phi_w(x)-\partial_{v_j}\Phi_v(x)\|_2^2.\tag{118}$$

Moreover, (105) gives

$$\sup_{\substack{w\in U_{\mathrm{d}}\\ \|w-v\|_2\le\tau}}\|\Phi_w(x)-\Phi_v(x)\|_2^2\le L_\Phi^2\tau^2.\tag{119}$$

Combining (17), (117), (118), and (119) yields

$$\|\Omega_{v,\tau}\|_{L^2(\rho_X)}^2\le(L_\Phi^2+\widetilde{L}_J^2)\tau^2+4M_J^2C_{\mathrm{bd}}\tau\le(L_\Phi^2+\widetilde{L}_J^2+4M_J^2C_{\mathrm{bd}})\tau,$$

where the last inequality uses $\tau\le 1$. Therefore

$$\|\Omega_{v,\tau}\|_{L^2(\rho_X)}\le L_{\mathrm{ReLU}}\tau^{1/2},\qquad L_{\mathrm{ReLU}}:=\left(L_\Phi^2+\widetilde{L}_J^2+4M_J^2C_{\mathrm{bd}}\right)^{1/2}.\tag{120}$$

The constants are independent of v. Thus (104), (111), the assumed covariance lower bound, and (120) verify Assumption 3.8 with $\alpha=1/2$. □

A.6 Proof of Theorem 4.1

Lemma A.9 (Local regularity of the refitted risk)**.** *Let $\bar{P}\in\mathcal{S}_2(\ell^2,\mathbb{R}^p)$ have full row rank. There exist a neighborhood V of $\bar{P}$ and a constant $C<\infty$ such that every $P,Q\in V$ have full row rank. Write*

$$\beta_P^*:=(PP^\top)^{-1}Pf,\qquad s_P:=P^\top\beta_P^*,\qquad u_P:=f-s_P=f-\Pi_Pf.$$

Then

$$\|\beta_P^*-\beta_Q^*\|_2+\|s_P-s_Q\|_2+\|u_P-u_Q\|_2+\|\mathcal{N}(P)-\mathcal{N}(Q)\|_{\mathrm{HS}}\le C\|P-Q\|_{\mathrm{HS}},\tag{121}$$

$$|\mathcal{E}_{\mathrm{opt}}(P+H)-\mathcal{E}_{\mathrm{opt}}(P)-\mathrm{d}\mathcal{E}_{\mathrm{opt}}(P)[H]|\le C\|H\|_{\mathrm{HS}}^2\tag{122}$$

whenever the segment $P+uH$, $u\in[0,1]$, lies in V. Moreover,

$$\mathrm{d}\mathcal{E}_{\mathrm{opt}}(P)[H]=-2(\beta_P^*)^\top Hu_P.\tag{123}$$

Proof. Choose a sufficiently small Hilbert–Schmidt ball V around $\bar{P}$. Since $\bar{P}$ has full row rank and is nonzero, the ball can be chosen so that there are constants $B,\lambda,c_0>0$ satisfying

$$\sup_{P\in V}\|P\|_{\mathrm{HS}}\le B,\qquad\inf_{P\in V}\lambda_{\min}(PP^\top)\ge\lambda,\qquad\inf_{P\in V}\|PP^\top\|_{\mathrm{HS}}^{1/2}\ge c_0.$$

For $P,Q\in V$,

$$\begin{aligned}\|PP^\top-QQ^\top\|_{\mathrm{HS}}\le(\|P\|_{\mathrm{HS}}+\|Q\|_{\mathrm{HS}})\|P-Q\|_{\mathrm{HS}},\\ (PP^\top)^{-1}-(QQ^\top)^{-1}=(PP^\top)^{-1}(QQ^\top-PP^\top)(QQ^\top)^{-1}.\end{aligned}$$

It follows that

$$\|(PP^\top)^{-1}-(QQ^\top)^{-1}\|_{\mathrm{op}}\le C\|P-Q\|_{\mathrm{HS}}.$$

Using

$$\begin{aligned}\beta_P^*-\beta_Q^*=(PP^\top)^{-1}(P-Q)f\\ +\big\{(PP^\top)^{-1}-(QQ^\top)^{-1}\big\}Qf\end{aligned}$$

and the uniform bounds on P, Q and their inverses gives $\|\beta_P^* - \beta_Q^*\|_2 \le C\|P-Q\|_{\mathrm{HS}}$. Consequently,

$$\|s_P - s_Q\|_2 \le \|(P-Q)^\top \beta_P^*\|_2 + \|Q^\top(\beta_P^* - \beta_Q^*)\|_2 \le C\|P-Q\|_{\mathrm{HS}},$$
$$\|u_P - u_Q\|_2 = \|s_P - s_Q\|_2 \le C\|P-Q\|_{\mathrm{HS}}.$$

Write $c(P) := \|PP^\top\|_{\mathrm{HS}}^{1/2}$. The lower bound $c(P), c(Q) \ge c_0$ and the preceding covariance estimate imply

$$|c(P) - c(Q)| \le C\|P-Q\|_{\mathrm{HS}}.$$

Hence

$$\begin{aligned}\|\mathcal{N}(P) - \mathcal{N}(Q)\|_{\mathrm{HS}} &\le \frac{\|P-Q\|_{\mathrm{HS}}}{c(P)} + \|Q\|_{\mathrm{HS}}\left|\frac{1}{c(P)} - \frac{1}{c(Q)}\right| \\ &\le C\|P-Q\|_{\mathrm{HS}},\end{aligned}$$

which proves (121).

For the risk, write

$$\mathcal{E}_{\mathrm{opt}}(P) = \|f\|_2^2 - (Pf)^\top (PP^\top)^{-1} Pf.$$

Let $G_P := PP^\top$ and $z_P := Pf$. For a Hilbert–Schmidt perturbation H,

$$\mathrm{d}G_P[H] = HP^\top + PH^\top, \qquad \mathrm{d}z_P[H] = Hf, \qquad \mathrm{d}(G_P^{-1})[H] = -G_P^{-1}\mathrm{d}G_P[H]G_P^{-1}.$$

Since $G_P^{-1} z_P = \beta_P^*$, direct substitution gives

$$\begin{aligned}\mathrm{d}\mathcal{E}_{\mathrm{opt}}(P)[H] &= -2(\beta_P^*)^\top Hf + (\beta_P^*)^\top (HP^\top + PH^\top)\beta_P^* \\ &= -2(\beta_P^*)^\top H(f - P^\top \beta_P^*) = -2(\beta_P^*)^\top H u_P,\end{aligned}$$

which proves (123). Differentiating the displayed formula once more produces only finite sums of products of P, H, G_P^{-1}, and f. The bounds $\|P\|_{\mathrm{HS}} \le B$ and $\|G_P^{-1}\|_{\mathrm{op}} \le \lambda^{-1}$ therefore bound the second derivative uniformly on V. Taylor's formula with integral remainder yields (122). □

Proof of Theorem 4.1. Write

$$s_0 := \Pi_{P_0} f, \qquad u_0 := f - s_0.$$

The assumption $0 < \|\Pi_{P_0} f\|_2 < \|f\|_2$ implies that both s_0 and u_0 are nonzero. We first directly calculate the feature gradient. Set $c(P) := \|PP^\top\|_{\mathrm{HS}}^{1/2}$, so $\mathcal{N}(P) = P/c(P)$. If P is normalized, then for every Hilbert–Schmidt perturbation H of sufficiently small norm,

$$\begin{aligned}c(P+H)^4 &= \mathrm{tr}\left(\left[(P+H)(P+H)^\top\right]^2\right) \\ &= 1 + 4\langle PP^\top P, H\rangle_{\mathrm{HS}} + O(\|H\|_{\mathrm{HS}}^2),\end{aligned}$$

and hence

$$c(P+H) = 1 + \langle PP^\top P, H\rangle_{\mathrm{HS}} + O(\|H\|_{\mathrm{HS}}^2).$$

For fixed β, write $r = f - P^\top \beta$. The normalized prediction then satisfies

$$\mathcal{N}(P+H)^\top \beta = P^\top \beta + H^\top \beta - \langle PP^\top P, H\rangle_{\mathrm{HS}} P^\top \beta + O(\|H\|_{\mathrm{HS}}^2),$$

where the remainder is measured in ℓ^2. Substituting this expansion into $\bar{L}$ gives

$$\bar{L}(P+H,\beta) - \bar{L}(P,\beta) = \left\langle -2\beta r^\top + 2(\beta^\top Pr)PP^\top P, H\right\rangle_{\mathrm{HS}} + O(\|H\|_{\mathrm{HS}}^2).$$

Therefore,

$$\nabla_P \bar{L}(P,\beta) = -2\beta r^\top + 2(\beta^\top Pr)PP^\top P. \tag{124}$$

We next verify that the finite block remains in one regular neighborhood. Apply Lemma A.9 on a ball V around P_0, and choose $r_V > 0$ such that the closed Hilbert–Schmidt ball of radius $2r_V$ around P_0 is contained in V. Let $L_{\mathcal{N}}$ be a Lipschitz constant for $\mathcal{N}$ on V. We prove by induction that

$$\|P_s - P_0\|_{\mathrm{HS}} < r_V, \qquad s = 0, \dots, T. \tag{125}$$

The claim holds at $s = 0$. Suppose it holds through some $t < T$. Every P_s, $s \le t$, is normalized, and hence

$$\|P_s\|_{\mathrm{op}} \le 1, \qquad \|P_s P_s^\top P_s\|_{\mathrm{HS}}^2 = \mathrm{tr}\big((P_s P_s^\top)^3\big) \le \|P_s P_s^\top\|_{\mathrm{op}} \|P_s P_s^\top\|_{\mathrm{HS}}^2 \le 1.$$

The exact head recurrence is the same normalized-operator recurrence as (31). Starting from $\beta_0 = 0$, its dimension-free bound gives

$$\max_{0 \le s \le t} \|\beta_{s+1}\|_2 \le C_T \eta_1.$$

Thus $\|r_s\|_2 \le C_T$ for $s \le t$. Define the raw feature increment

$$H_s := \widetilde{P}_{s+1} - P_s = 2\eta_2 \left[\beta_{s+1} r_s^\top - \big(\beta_{s+1}^\top P_s r_s\big) P_s P_s^\top P_s\right]. \tag{126}$$

The preceding bounds imply

$$\|H_s\|_{\mathrm{HS}} \le C_T \eta_1 \eta_2, \qquad s = 0, \dots, t. \tag{127}$$

Choose the step sizes sufficiently small that $C_T \eta_1 \eta_2 < r_V$ and $T L_{\mathcal{N}} C_T \eta_1 \eta_2 < r_V$. Then every segment $P_s + uH_s$, $u \in [0,1]$, lies in the ball of radius $2r_V$ and hence in V. Since $\mathcal{N}(P_s) = P_s$, the Lipschitz bound gives

$$\|P_{s+1} - P_s\|_{\mathrm{HS}} = \|\mathcal{N}(P_s + H_s) - \mathcal{N}(P_s)\|_{\mathrm{HS}} \le L_{\mathcal{N}} \|H_s\|_{\mathrm{HS}} \le L_{\mathcal{N}} C_T \eta_1 \eta_2.$$

Consequently,

$$\|P_{t+1} - P_0\|_{\mathrm{HS}} \le \sum_{s=0}^{t} \|P_{s+1} - P_s\|_{\mathrm{HS}} < r_V,$$

which closes the induction. In particular, every iterate and every raw update segment remains in V, and

$$\max_{0 \le t \le T} \|P_t - P_0\|_{\mathrm{HS}} \le C_T \eta_1 \eta_2. \tag{128}$$

Since the Hilbert–Schmidt bound in (128) also bounds the operator norm, the same dimension-free recurrence argument as (36) gives, for $0 < \eta_2 \le 1$,

$$\beta_{t+1} = 2(t+1)\eta_1 P_0 f + O_T(\eta_1^2). \tag{129}$$

Lemma A.9 then implies

$$(\beta_{P_t}^*)^\top \beta_{t+1} = 2(t+1)\eta_1 \|s_0\|_2^2 + O_T(\eta_1^2), \tag{130}$$

$$\|u_{P_t}\|_2^2 = \|u_0\|_2^2 + O_T(\eta_1 \eta_2), \tag{131}$$

where $(\beta_{P_0}^*)^\top P_0 f = \|P_0^\top \beta_{P_0}^*\|_2^2 = \|s_0\|_2^2$.

We now derive the change in the misspecified risk using the raw increment H_t from (126). Since $P_t u_{P_t} = 0$, the second term satisfies

$$(P_t P_t^\top P_t) u_{P_t} = P_t P_t^\top (P_t u_{P_t}) = 0.$$

Moreover,

$$\langle r_t, u_{P_t} \rangle = \langle f - P_t^\top \beta_{t+1}, u_{P_t} \rangle = \langle f, u_{P_t} \rangle = \|u_{P_t}\|_2^2,$$

where the last equality follows from $f = s_{P_t} + u_{P_t}$ and $s_{P_t} \perp u_{P_t}$.

Equation 123 therefore gives the exact identity

$$\mathrm{d}\mathcal{E}_{\mathrm{opt}}(P_t)[H_t] = -4\eta_2 (\beta_{P_t}^*)^\top \beta_{t+1} \|u_{P_t}\|_2^2. \tag{132}$$

Scalar normalization leaves the row space unchanged, so $\mathcal{E}_{\mathrm{opt}}(P_{t+1}) = \mathcal{E}_{\mathrm{opt}}(P_t + H_t)$. Combining (122), (130), (131), and (132) yields

$$\begin{aligned} &\mathcal{E}_{\mathrm{opt}}(P_{t+1}) - \mathcal{E}_{\mathrm{opt}}(P_t) \\ &\quad = -8(t+1)\eta_1 \eta_2 \|s_0\|_2^2 \|u_0\|_2^2 + O_T(\eta_1^2 \eta_2 + \eta_1^2 \eta_2^2). \end{aligned}$$

The leading coefficient is strictly negative, proving the claimed decrease for sufficiently small positive step sizes within one block.

For the repeated-block statement, fix the number of blocks $K < \infty$. By Lemma A.9, we may shrink V so that, for every $P \in V$,

$$\|\Pi_P f\|_2 \geq \frac{1}{2}\|s_0\|_2, \qquad \|f - \Pi_P f\|_2 \geq \frac{1}{2}\|u_0\|_2.$$

Repeat the preceding neighborhood induction over the total finite horizon KT, resetting only β at the beginning of each block. The head bound restarts from zero in every block, so the constants in the raw-increment bound remain uniform. The step sizes can therefore be chosen so that every iterate and raw update segment remains in V. Applying the one-block expansion with the feature operator at the beginning of each block in place of P_0 gives a strictly negative leading coefficient, uniformly over all K blocks. This proves the repeated-block claim.

□

A.7 Sample full-matrix updates

We restrict the sample-level feature operator to the first $d_{\mathrm{tr}} < \infty$ basis coordinates while leaving the population target f^* unchanged. Let

$$\Psi_{\mathrm{tr}}(x) := (\psi_1(x), \dots, \psi_{d_{\mathrm{tr}}}(x))^\top, \qquad 1 \leq p \leq d_{\mathrm{tr}}, \qquad P \in \mathbb{R}^{p \times d_{\mathrm{tr}}}, \qquad \Phi_P(x) = P\Psi_{\mathrm{tr}}(x).$$

Let $\pi_{\mathrm{tr}} : \ell^2 \to \mathbb{R}^{d_{\mathrm{tr}}}$ be the coordinate projection and let $\Pi_{\mathrm{tr}} := \pi_{\mathrm{tr}}^\top \pi_{\mathrm{tr}}$ be the orthogonal projection in ℓ^2 onto the first d_{tr} coordinates. We identify the finite matrix P with its zero-extended operator $P\pi_{\mathrm{tr}} : \ell^2 \to \mathbb{R}^p$ when evaluating the population risk. Write $f_{\mathrm{tr}} := (f_1, \dots, f_{d_{\mathrm{tr}}})^\top \in \mathbb{R}^{d_{\mathrm{tr}}}$ for the first d_{tr} coefficients of the unchanged population target f. From one labeled sample $\{(x_i, y_i)\}_{i=1}^n$, reused throughout the resetting block, define

$$\widehat{f}_{\mathrm{tr}} := \frac{1}{n}\sum_{i=1}^n y_i \Psi_{\mathrm{tr}}(x_i), \qquad \widehat{C}_{\mathrm{tr}} := \frac{1}{n}\sum_{i=1}^n \Psi_{\mathrm{tr}}(x_i)\Psi_{\mathrm{tr}}(x_i)^\top.$$

The normalized sample loss is the ordinary empirical mean-squared error

$$\begin{aligned}\widehat{\widehat{L}}_{n,\mathrm{tr}}(P, \beta) &:= \frac{1}{n}\sum_{i=1}^n \left(y_i - \langle \mathcal{N}(P)\Psi_{\mathrm{tr}}(x_i), \beta\rangle\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n y_i^2 - 2\beta^\top \mathcal{N}(P)\widehat{f}_{\mathrm{tr}} + \beta^\top \mathcal{N}(P)\widehat{C}_{\mathrm{tr}}\mathcal{N}(P)^\top \beta. \end{aligned} \tag{133}$$

At the beginning of the block, set $\widehat{\beta}_0 = 0$ and $\widehat{P}_0 = P_0$. Each iteration first updates the head, then updates the feature matrix using (133), and finally normalizes it:

$$\widehat{\beta}_{t+1} = \widehat{\beta}_t - \eta_1 \nabla_\beta \widehat{\widehat{L}}_{n,\mathrm{tr}}(\widehat{P}_t, \widehat{\beta}_t) = \widehat{\beta}_t + 2\eta_1 \widehat{P}_t\big(\widehat{f}_{\mathrm{tr}} - \widehat{C}_{\mathrm{tr}}\widehat{P}_t^\top \widehat{\beta}_t\big). \tag{134}$$

Define

$$\widehat{r}_t := \widehat{f}_{\mathrm{tr}} - \widehat{C}_{\mathrm{tr}}\widehat{P}_t^\top \widehat{\beta}_{t+1}.$$

Then

$$\begin{aligned}\widehat{\widetilde{P}}_{t+1} &= \widehat{P}_t - \eta_2 \nabla_P \widehat{\widehat{L}}_{n,\mathrm{tr}}(\widehat{P}_t, \widehat{\beta}_{t+1}) \\ &= \widehat{P}_t + 2\eta_2 \widehat{\beta}_{t+1}\widehat{r}_t^\top - 2\eta_2\big(\widehat{\beta}_{t+1}^\top \widehat{P}_t \widehat{r}_t\big)\widehat{P}_t\widehat{P}_t^\top \widehat{P}_t, \qquad \widehat{P}_{t+1} = \mathcal{N}(\widehat{\widetilde{P}}_{t+1}). \end{aligned} \tag{135}$$

All quantities in these updates are finite-dimensional, so the empirical gradient is well defined. Since every row of the zero-extended operator lies in the first d_{tr} coordinates, the misspecified risk decomposes as

$$\mathcal{E}_{\mathrm{opt}}(P) = \|(I - \Pi_{\mathrm{tr}})f\|_2^2 + \|f_{\mathrm{tr}} - \Pi_P f_{\mathrm{tr}}\|_2^2, \qquad \Pi_P := P^\top(PP^\top)^{-1}P \quad \text{on } \mathbb{R}^{d_{\mathrm{tr}}}. \tag{136}$$

The first term is fixed, while the second can change under the sample update. To control its population direction, suppose $|f^*(x)| \leq R_f$, the noise is conditionally mean-zero and conditionally sub-Gaussian with norm at most K_ϵ, and $B_{\mathrm{tr}} := \max_{k \leq d_{\mathrm{tr}}} \|\psi_k\|_\infty < \infty$. Define

$$R_{\Psi,\mathrm{tr}} := \left(\sum_{k=1}^{d_{\mathrm{tr}}} \|\psi_k\|_\infty^2\right)^{1/2} \leq B_{\mathrm{tr}}\sqrt{d_{\mathrm{tr}}}.$$

For $\delta \in (0,1)$, let

$$\varepsilon_{f,n,\mathrm{tr}}(\delta) := CB_{\mathrm{tr}}(R_f + K_\epsilon)\sqrt{\frac{d_{\mathrm{tr}}\log(4d_{\mathrm{tr}}/\delta)}{n}}, \tag{137}$$

$$\varepsilon_{C,n,\mathrm{tr}}(\delta) := C\left\{R_{\Psi,\mathrm{tr}}\sqrt{\frac{\log(4d_{\mathrm{tr}}/\delta)}{n}} + R^2_{\Psi,\mathrm{tr}}\frac{\log(4d_{\mathrm{tr}}/\delta)}{n}\right\}, \tag{138}$$

where C is a universal constant.

Now we are ready to state the sample-level version of Theorem 4.1.

Theorem A.10. *Let P_0 be normalized and have full row rank, write $s_0 := \Pi_{P_0} f_{\mathrm{tr}}$ and $u_0 := f_{\mathrm{tr}} - \Pi_{P_0} f_{\mathrm{tr}}$, and suppose $\|s_0\|_2\|u_0\|_2 > 0$. Fix $T < \infty$ and $\delta \in (0,1)$. Under the preceding boundedness and noise conditions, if*

$$\varepsilon_{f,n,\mathrm{tr}}(\delta/2) \le \frac{1}{2}\min\{\|s_0\|_2, \|u_0\|_2\}, \qquad \varepsilon_{C,n,\mathrm{tr}}(\delta/2) \le 1, \tag{139}$$

then, with probability at least $1-\delta$, for sufficiently small positive step sizes the sample iterates (134)–(135) remain full row rank and satisfy

$$\mathcal{E}_{\mathrm{opt}}(\widehat{P}_{t+1}) < \mathcal{E}_{\mathrm{opt}}(\widehat{P}_t), \qquad t = 0, \dots, T-1.$$

Proof. **Concentration.** For every $k \le d_{\mathrm{tr}}$, orthonormality and the conditional mean-zero noise give $\mathbb{E}[y\psi_k(x)] = f_k$. Since $\widehat{f}_{\mathrm{tr}}$ is sub-Gaussian, we have

$$\|\widehat{f}_{\mathrm{tr}} - f_{\mathrm{tr}}\|_2 \le \varepsilon_{f,n,\mathrm{tr}}(\delta/2)$$

with probability at least $1-\delta/2$. Similarly, $\mathbb{E}[\Psi_{\mathrm{tr}}(x)\Psi_{\mathrm{tr}}(x)^\top] = I_{d_{\mathrm{tr}}}$ and $\|\Psi_{\mathrm{tr}}(x)\|_2 \le R_{\Psi,\mathrm{tr}}$, so matrix Bernstein's inequality gives

$$\|\widehat{C}_{\mathrm{tr}} - I_{d_{\mathrm{tr}}}\|_{\mathrm{op}} \le \varepsilon_{C,n,\mathrm{tr}}(\delta/2)$$

with probability at least $1-\delta/2$. We work on the intersection of these two events, which has probability at least $1-\delta$.

The margin condition then implies

$$\|\widehat{C}_{\mathrm{tr}}\|_{\mathrm{op}} \le 2, \qquad \langle s_0, \widehat{f}_{\mathrm{tr}}\rangle \ge \frac{1}{2}\|s_0\|_2^2, \qquad \langle u_0, \widehat{f}_{\mathrm{tr}}\rangle \ge \frac{1}{2}\|u_0\|_2^2, \tag{140}$$

and $\|\widehat{f}_{\mathrm{tr}}\|_2 \le 3\|f_{\mathrm{tr}}\|_2/2$.

Control over the block. For a full-row-rank P, write

$$\beta^*_P := (PP^\top)^{-1}Pf_{\mathrm{tr}}, \qquad s_P := \Pi_P f_{\mathrm{tr}}, \qquad u_P := f_{\mathrm{tr}} - s_P.$$

Apply Lemma A.9 with target f_{tr} and choose a neighborhood V of P_0 on which all these quantities and $\mathcal{N}$ are Lipschitz. By shrinking V if necessary (so that $s_P - s_0$ and $u_P - u_0$ are sufficiently small), the empirical margin bounds (140) ensure that

$$\langle s_P, \widehat{f}_{\mathrm{tr}}\rangle \ge \frac{1}{4}\|s_0\|_2^2, \qquad \langle u_P, \widehat{f}_{\mathrm{tr}}\rangle \ge \frac{1}{4}\|u_0\|_2^2, \qquad P \in V. \tag{141}$$

In the rest of the proof, take $\eta_1, \eta_2 \le 1$. The notation C_T denotes a finite constant that may depend on the fixed block length T, the neighborhood V, and fixed problem quantities; its value may change from line to line, and the same convention applies to $O(\cdot)$. Since every $\widehat{P}_t$ is normalized,

$$\|\widehat{P}_t\|_{\mathrm{op}} \le 1, \qquad \|\widehat{P}_t\widehat{P}_t^\top\widehat{P}_t\|_{\mathrm{HS}} \le 1.$$

Choose $r_V > 0$ such that the closed Hilbert–Schmidt ball of radius $2r_V$ around P_0 is contained in V. We prove by induction that $\|\widehat{P}_t - P_0\|_{\mathrm{HS}} < r_V$ for every $t = 0, \dots, T$. The claim holds at $t = 0$. Suppose it holds up to some $t < T$. Then $\widehat{P}_s \in V$ for $s = 0, \dots, t$, so the exact head recurrence gives

$$\begin{aligned}\|\widehat{\beta}_{s+1}\|_2 &\le \|\widehat{\beta}_s\|_2 + 2\eta_1\|\widehat{f}_{\mathrm{tr}}\|_2 + 2\eta_1\|\widehat{C}_{\mathrm{tr}}\|_{\mathrm{op}}\|\widehat{\beta}_s\|_2 \\ &\le (1+4\eta_1)\|\widehat{\beta}_s\|_2 + 3\eta_1\|f_{\mathrm{tr}}\|_2, \qquad s = 0, \dots, t.\end{aligned}$$

Starting from $\widehat{\beta}_0 = 0$ and iterating this inequality gives

$$\max_{0\le s\le t} \|\widehat{\beta}_{s+1}\|_2 \le 3\eta_1 \|f_{\mathrm{tr}}\|_2 \sum_{k=0}^{t} (1+4\eta_1)^k \le C_T \eta_1. \tag{142}$$

For $s = 0, \dots, t$, the definition of $\widehat{r}_s$ then gives

$$\|\widehat{r}_s\|_2 \le \|\widehat{f}_{\mathrm{tr}}\|_2 + \|\widehat{C}_{\mathrm{tr}}\|_{\mathrm{op}} \|\widehat{P}_s\|_{\mathrm{op}} \|\widehat{\beta}_{s+1}\|_2 \le C_T.$$

The corresponding raw feature increment is

$$H_s := \widehat{\widetilde{P}}_{s+1} - \widehat{P}_s = 2\eta_2 \left[\widehat{\beta}_{s+1} \widehat{r}_s^\top - \left(\widehat{\beta}_{s+1}^\top \widehat{P}_s \widehat{r}_s\right) \widehat{P}_s \widehat{P}_s^\top \widehat{P}_s \right].$$

Hence we have

$$\|H_s\|_{\mathrm{HS}} \le 4\eta_2 \|\widehat{\beta}_{s+1}\|_2 \|\widehat{r}_s\|_2 \le C_T \eta_1 \eta_2, \qquad s = 0, \dots, t.$$

Take the step sizes small enough that $TC_T\eta_1\eta_2 < r_V$. The induction hypothesis and the bound on H_s ensure that every segment $\widehat{P}_s + uH_s$, $u \in [0,1]$, lies in the ball of radius $2r_V$ and hence in V. Since $\mathcal{N}(\widehat{P}_s) = \widehat{P}_s$ and $\mathcal{N}$ is Lipschitz on V,

$$\|\widehat{P}_{s+1} - \widehat{P}_s\|_{\mathrm{HS}} \le C_T \|H_s\|_{\mathrm{HS}} \le C_T \eta_1 \eta_2.$$

Consequently,

$$\|\widehat{P}_{t+1} - P_0\|_{\mathrm{HS}} \le \sum_{s=0}^{t} \|\widehat{P}_{s+1} - \widehat{P}_s\|_{\mathrm{HS}} \le (t+1) C_T \eta_1 \eta_2 < r_V.$$

Hence, we get

$$\max_{0\le t<T} \|\widehat{\beta}_{t+1}\|_2 \le C_T \eta_1, \tag{143}$$

and

$$\max_{0\le t\le T} \|\widehat{P}_t - P_0\|_{\mathrm{HS}} \le C_T \eta_1 \eta_2. \tag{144}$$

In particular, every iterate and every raw update segment remains in V, so all iterates have full row rank.

Expansion of the misspecified risk update. Notice that scalar normalization preserves the row space, thus

$$\mathcal{E}_{\mathrm{opt}}(\widehat{P}_{t+1}) = \mathcal{E}_{\mathrm{opt}}(\widehat{P}_t + H_t).$$

The first term in (136) is constant in P. Applying the Taylor bound in Lemma A.9 to the second term gives

$$\begin{aligned} &\mathcal{E}_{\mathrm{opt}}(\widehat{P}_{t+1}) - \mathcal{E}_{\mathrm{opt}}(\widehat{P}_t) \\ &= \mathrm{d}\mathcal{E}_{\mathrm{opt}}(\widehat{P}_t)[H_t] + R_t, \qquad |R_t| \le C_T \|H_t\|_{\mathrm{HS}}^2 \le C_T \eta_1^2 \eta_2^2. \end{aligned} \tag{145}$$

Since $u_{\widehat{P}_t} = f_{\mathrm{tr}} - \Pi_{\widehat{P}_t} f_{\mathrm{tr}}$ is orthogonal to the row space of $\widehat{P}_t$, we have $\widehat{P}_t u_{\widehat{P}_t} = 0$. Thus the second term in H_t vanishes when applied to $u_{\widehat{P}_t}$, and hence

$$H_t u_{\widehat{P}_t} = 2\eta_2 \widehat{\beta}_{t+1} \langle \widehat{r}_t, u_{\widehat{P}_t} \rangle.$$

The derivative formula in Lemma A.9 therefore yields

$$\mathcal{E}_{\mathrm{opt}}(\widehat{P}_{t+1}) - \mathcal{E}_{\mathrm{opt}}(\widehat{P}_t) = -4\eta_2 (\beta^*_{\widehat{P}_t})^\top \widehat{\beta}_{t+1} \langle u_{\widehat{P}_t}, \widehat{r}_t \rangle + R_t. \tag{146}$$

It remains to control the two factors in the leading term.

Notice that

$$\begin{aligned} \widehat{\beta}_{t+1} - \widehat{\beta}_t = 2\eta_1 P_0 \widehat{f}_{\mathrm{tr}} + 2\eta_1 (\widehat{P}_t - P_0) \widehat{f}_{\mathrm{tr}} \\ - 2\eta_1 \widehat{P}_t \widehat{C}_{\mathrm{tr}} \widehat{P}_t^\top \widehat{\beta}_t. \end{aligned}$$

By (143) and (144), the second and third terms are, respectively, $O(\eta_1^2\eta_2)$ and $O(\eta_1^2)$. Summing over the fixed block therefore gives

$$\widehat{\beta}_{t+1} = 2(t+1)\eta_1 P_0 \widehat{f}_{\mathrm{tr}} + O(\eta_1^2). \tag{147}$$

The local Lipschitz continuity of $P \mapsto \beta_P^*$ gives $\beta^*_{\widehat{P}_t} = \beta^*_{P_0} + O(\eta_1\eta_2)$. Since $(\beta^*_{P_0})^\top P_0 \widehat{f}_{\mathrm{tr}} = \langle s_0, \widehat{f}_{\mathrm{tr}} \rangle$, we obtain, uniformly for $t = 0, \dots, T-1$,

$$(\beta^*_{\widehat{P}_t})^\top \widehat{\beta}_{t+1} = 2(t+1)\eta_1 \langle s_0, \widehat{f}_{\mathrm{tr}} \rangle + O(\eta_1^2) \geq \frac{1}{2}\eta_1 \|s_0\|_2^2 \tag{148}$$

for sufficiently small η_1.

By (141),

$$\begin{aligned} \langle u_{\widehat{P}_t}, \widehat{r}_t \rangle &= \langle u_{\widehat{P}_t}, \widehat{f}_{\mathrm{tr}} \rangle - \left\langle u_{\widehat{P}_t}, \widehat{C}_{\mathrm{tr}} \widehat{P}_t^\top \widehat{\beta}_{t+1} \right\rangle \\ &\geq \frac{1}{4}\|u_0\|_2^2 - \left| \left\langle u_{\widehat{P}_t}, \widehat{C}_{\mathrm{tr}} \widehat{P}_t^\top \widehat{\beta}_{t+1} \right\rangle \right|. \end{aligned}$$

The second term satisfies

$$\left| \left\langle u_{\widehat{P}_t}, \widehat{C}_{\mathrm{tr}} \widehat{P}_t^\top \widehat{\beta}_{t+1} \right\rangle \right| \leq \|u_{\widehat{P}_t}\|_2 \|\widehat{C}_{\mathrm{tr}}\|_{\mathrm{op}} \|\widehat{P}_t\|_{\mathrm{op}} \|\widehat{\beta}_{t+1}\|_2 \leq C_T \eta_1.$$

Consequently, for sufficiently small η_1,

$$\langle u_{\widehat{P}_t}, \widehat{r}_t \rangle \geq \frac{1}{8}\|u_0\|_2^2. \tag{149}$$

Substituting (148), (149), and the remainder bound in (145) into (146) gives

$$\mathcal{E}_{\mathrm{opt}}(\widehat{P}_{t+1}) - \mathcal{E}_{\mathrm{opt}}(\widehat{P}_t) \leq -\frac{1}{4}\eta_1\eta_2 \|s_0\|_2^2 \|u_0\|_2^2 + O(\eta_1^2\eta_2^2) < 0$$

for sufficiently small positive step sizes. □

B FURTHER EXPERIMENTS

B.1 ABLATION STUDY

We consider two resetting periods $T = 100$ and $T = 400$, in the regression model of Section 5. For $T = 100$, we run each method for 2000 gradient steps; for $T = 400$, we run 8000 gradient steps. We compare the following four methods: traditional gradient descent, resetting only, normalization only and our resetting + normalization method. We choose $\eta_1 = 0.1$ and $\eta_2 = 0.05$.

B.1.1 RESETTING PERIOD $T = 100$

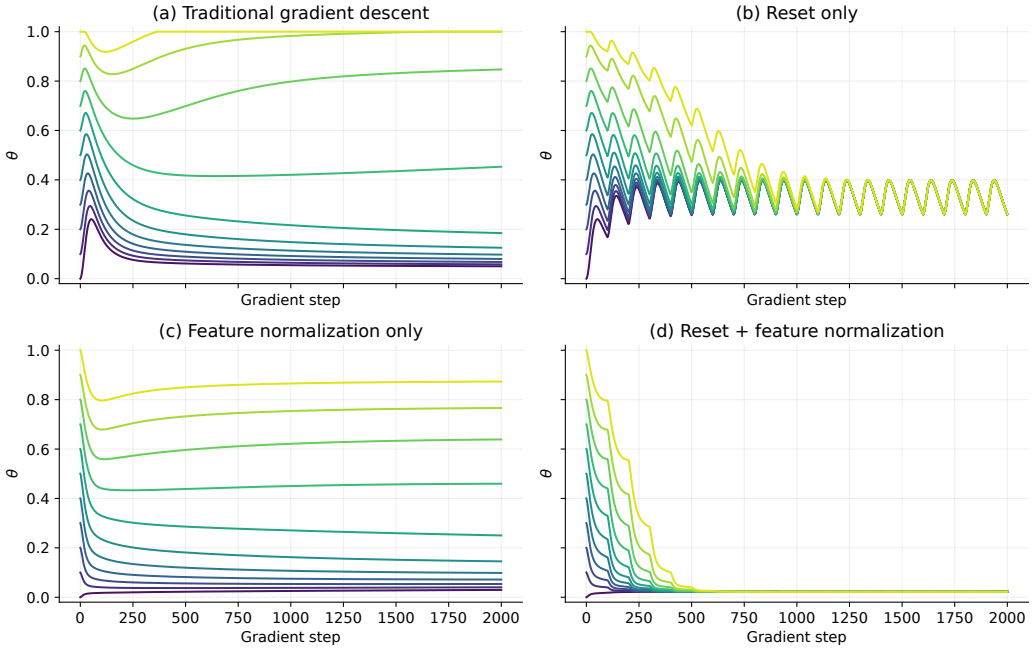

Figure 3: Trajectories over 2000 gradient steps with $T = 100$, from $\theta_0 \in \{0, 0.1, \ldots, 1\}$. All methods share the same data and learning rates.

Initialization dependence and feature dynamics. Figure 3 shows that traditional gradient descent and feature normalization alone remain sensitive to initialization. For the resetting-only method, while it drives all initial values in the same direction, it exhibits persistent oscillation and fails to bring parameters close to zero. In contrast, when resetting is combined with normalization, all trajectories are driven near zero, and the oscillation problem is effectively resolved through normalization.

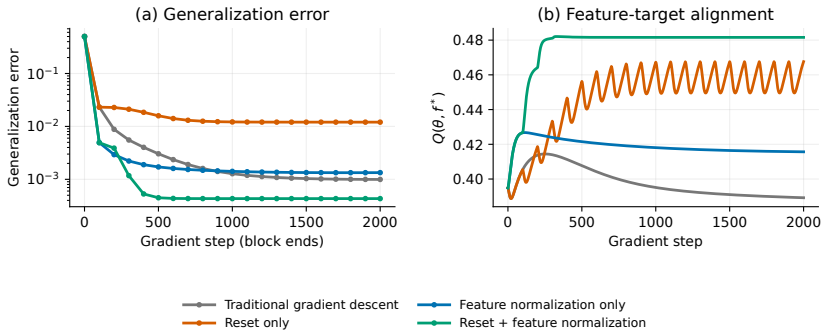

Figure 4: Results with $T = 100$ and $\theta_0 = 0.8$. Left: generalization error evaluated at block ends. Right: feature-target alignment Q.

Generalization error and feature-target alignment. Figure 4 shows that resetting combined with feature normalization attains the lowest generalization error and the highest feature-target alignment. The block-end prediction error also depends on the current optimization state of the readout, while Q directly measures the representation-level alignment studied in the theory. The resetting-only method generally improves feature-target alignment, yet it still suffers from oscillation. One may notice that the generalization error of the resetting-only method is extremely high, which can be attributed to the insufficient training of β during the 100 steps for this method. Hence we also consider a longer block period.

B.1.2 Resetting period $T = 400$

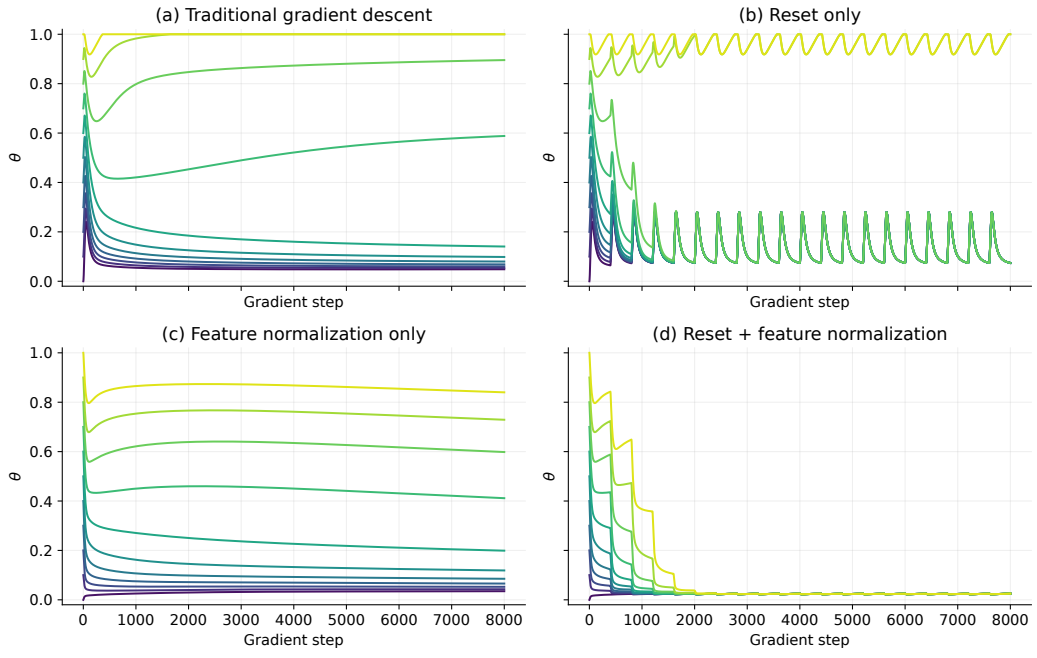

Figure 5: Trajectories over 8000 gradient steps with $T = 400$, from $\theta_0 \in \{0, 0.1, \ldots, 1\}$.

Initialization dependence and feature dynamics. Figure 5 shows that increasing T does not resolve the oscillations of the resetting-only method. Furthermore, it reveals that the resetting-only method is also sensitive to initialization. In contrast, the combination of resetting and normalization remains robust in large resetting periods. For traditional methods and normalization-only methods, the trajectories remain sensitive to initialization.

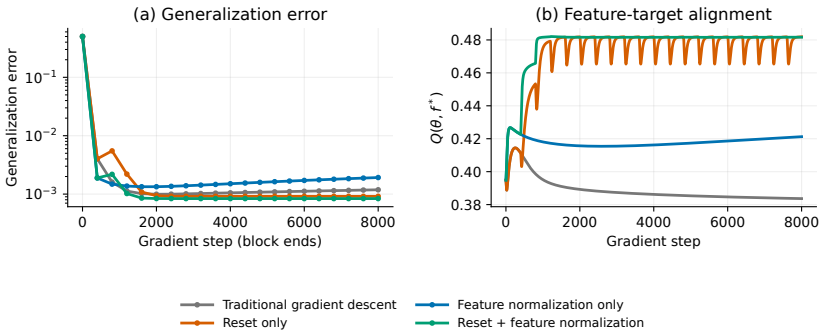

Figure 6: Results over 8000 gradient steps with $T = 400$ and $\theta_0 = 0.8$. Left: generalization error evaluated at block ends. Right: feature-target alignment Q.

Generalization error and feature-target alignment. Figure 6 shows that the resetting-only method performs better when T is larger, while our resetting + normalization method still achieves the best performance in terms of both generalization error and feature-target alignment. Notice that there is a small decrease in Q for the resetting + normalization method. This does not contradict our theoretical findings, as the monotonic increase in feature-target alignment is only guaranteed under sufficiently small step sizes. Consequently, larger T may require smaller step sizes to ensure monotonic increase of feature-target alignment.

B.2 Experiment for Section 4 and Appendix A.7

We provide a numerical illustration of the sample full-matrix result in Appendix A.7. Let $x \sim \mathrm{Unif}[0,1]$. We use the bounded orthonormal Fourier basis on $[0,1]$ ordered as $\psi_1(x) = 1$, $\psi_{2k}(x) = \sqrt{2}\cos(2\pi kx)$ and $\psi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx)$ $(k \geq 1)$. We consider

$$f^*(x) = \sum_{j \geq 1} f_j \psi_j(x), \qquad f_j \propto j^{-3/2}, \qquad \|f\|_2 = 1.$$

The learned feature matrix is restricted to the first $d_{\mathrm{tr}} = 20$ coordinates and has feature dimension $p = 5$. We draw samples from $y_i = f^*(x_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$. The initialization is constructed so that $\|\Pi_{P_0} f_{\mathrm{tr}}\|_2^2 = \frac{1}{2}\|f_{\mathrm{tr}}\|_2^2$. We choose $\eta_1 = \eta_2 = 0.005$, $n = 2000$, and a resetting period of $T = 100$; $\widehat{\beta}$ is reset to zero at the beginning of each block.

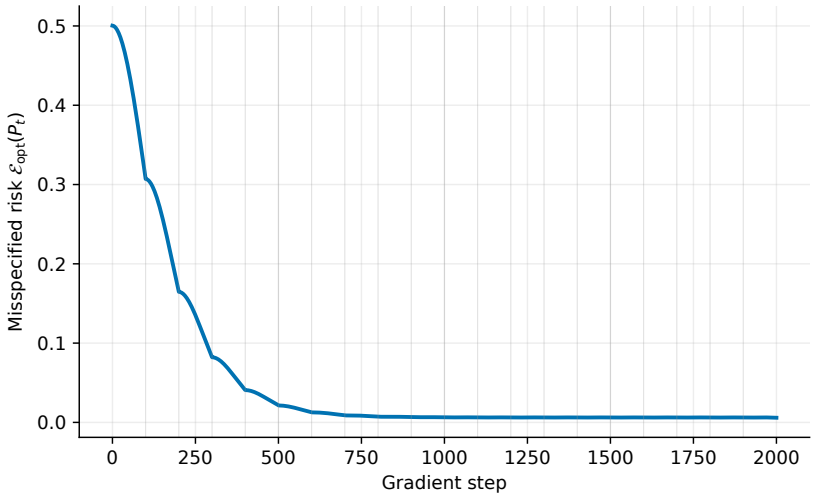

Figure 7: Sample full-matrix dynamics corresponding to Section 4, Appendix A.7. The grey vertical lines mark the resetting blocks, at which the linear head is reset.

Figure 7 reports the resulting misspecified risk over the repeated resetting blocks, which is in accordance with the theoretical results of Theorem A.10. In each resetting block, the decrease is slow at first and then accelerates. This is because β starts small in every block, resulting in only minor updates to the feature matrix. As β grows, the update of the feature matrix becomes larger.

B.3 Experiments on Empirical Datasets

We also evaluate the method on CIFAR-10, SVHN, Fashion-MNIST, MNIST, UCI HAR, and Wine Quality. Although our theory is developed for squared-loss regression, we also empirically evaluate the same training principle on classification tasks.

Experimental setup. MNIST and Fashion-MNIST use a two-convolution CNN with a 128-dimensional feature layer for 100 epochs, with $T = 20$; CIFAR-10 and SVHN use CIFAR ResNet-20 with a 64-dimensional feature layer for 200 epochs, with $T = 50$; UCI HAR uses a 561-256-128-64 MLP for 100 epochs, with $T = 20$; and Wine Quality uses an 11-64-32 MLP for 200 epochs, with $T = 40$. Classification models are trained with cross-entropy, whereas Wine Quality is trained with squared loss. Every minibatch first updates the head with the current features held fixed and then updates the backbone using the new head. At a resetting time, only the output linear head and its optimizer state are reset. For a minibatch feature matrix $Z \in \mathbb{R}^{B\times p}$, define its empirical second-moment matrix as $\widehat{\Sigma}_B = Z^\top Z/B$. We normalize the features according to

$$\bar{Z} = \frac{Z}{\sqrt{\|\widehat{\Sigma}_B\|_{\mathrm{F}} + 10^{-8}}},$$

so that $\|\bar{Z}^\top \bar{Z}/B\|_{\mathrm{F}} \approx 1$. Within each of the 20 seeds, all four methods share the data split, preprocessing, initialization, data order, and remaining hyperparameters. We report the mean and standard deviation of their validation performance.

During training, autograd differentiates through $\sqrt{\|\widehat{\Sigma}_B\|_{\mathrm{F}} + 10^{-8}}$. For validation and test, the scale is recomputed from a fixed calibration set of 1024 examples using the current backbone, rather than from the evaluation minibatch.

Table 1: Performance across six representative datasets. We report the mean across 20 seeds, with the standard deviation shown in parentheses. Classification results report accuracies (%), while Wine Quality is reported using R^2. The last column reports the mean difference between Reset + Normalization and Standard.

Dataset	Standard	Reset	Normalization	Reset + Norm.	R+N − Std.
CIFAR-10	92.659 (0.225)	92.745 (0.249)	92.661 (0.249)	**92.792 (0.269)**	+0.133
SVHN	95.794 (0.103)	95.784 (0.103)	95.824 (0.113)	**95.858 (0.158)**	+0.064
Fashion-MNIST	91.613 (0.269)	**91.963 (0.213)**	91.573 (0.221)	91.607 (0.295)	-0.007
MNIST	99.228 (0.084)	99.248 (0.089)	**99.348 (0.052)**	99.333 (0.047)	+0.105
UCI HAR	98.105 (0.149)	98.254 (0.148)	98.546 (0.203)	**98.757 (0.199)**	+0.652
Wine Quality	0.3947 (0.0206)	0.3801 (0.0136)	0.3900 (0.0215)	**0.3948 (0.0192)**	+0.00007
Average rank ↓	3.17	2.83	2.50	**1.50**	–

As shown in Table 1, Reset + Normalization achieves the highest mean performance on four of the six representative datasets. The reset-only method performs best on Fashion-MNIST, while the normalization-only method performs best on MNIST. Overall, Reset + Normalization achieves the best average rank across all datasets.