UNDERSTANDING DEEP ONE-CLASS CLASSIFICATION: A TRAINING-DYNAMICS VIEWPOINT

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep one-class classification (DOCC) maps normal samples toward a fixed center, yet at test time it must distinguish unseen normal and anomalous inputs. How DOCC training shapes representations of inputs that never appear in the training objective remains poorly understood. We study the training dynamics of DOCC under gradient descent and derive a finite-time first-order description of the representation change of an unseen input. We show that its movement is determined by cumulative normal-sample updates transferred through gradient coupling, while the representations of normal training samples are progressively driven toward the center. This analysis identifies two key factors governing the behavior of unseen inputs: their initial separation from the center and the training-induced displacement. These factors explain representative success and failure cases for both unseen normal and anomalous inputs and yield a sufficient condition for an anomaly to remain detectable after training. We further extend the training-dynamics view to anomaly supervision and explain how labeled anomalies induce additional displacements that can correct anomalies missed by one-class training. Experiments support the predicted representation dynamics, anomaly scores, failure modes, and coupling geometry under Gaussian initialization.

1 INTRODUCTION

Detecting deviations from normality is an integral aspect of many domains, including medical diagnosis, safety, security, and optimization (Yao et al., 2026). Such deviations are generally called *anomalies*. Automating Anomaly Detection (AD) is necessary to deal with the large amounts of complex data.

One of the first methods to provide a neural-network-based solution to AD was Deep One-Class Classification (DOCC), which encloses normal data in a hypersphere in the feature space of a neural network. DOCC is a fundamental element of some of the most significant advancements in the field of AD (Shen et al., 2020; Liznerski et al., 2022), and to this day, it and its extensions, e.g. Deep Semi-Supervised AD (Deep SAD) (Ruff et al., 2020), are important baselines for evaluations in different areas of deep AD (Ye et al., 2025; Tsai et al., 2025; Yao et al., 2026).

A central difficulty in AD is the lack of anomaly labels. Most AD methods, including DOCC, are trained using only normal data, while some extensions incorporate a limited amount of labeled anomalies or auxiliary data from other domains (Ruff et al., 2021). Outlier exposure (OE) Hendrycks et al. (2019), the technique of including out-of-domain data in DOCC through its semi-supervised extension Deep SAD, has improved performance on existing benchmarks by up to 30% AUC-ROC (Liznerski et al., 2021), marking an important advance in the field.

Despite their empirical success, DOCC and deep AD more broadly remain largely unexplored theoretically. Steinwart et al. (2005) established a connection between shallow AD and binary classification of the normal-data distribution vs. a uniform distribution. Building on this connection, Zhou et al. (2026b) proposed a neural-network classifier and showed that its excess risk converges to the optimum identified in Steinwart et al. (2005). Further related work is discussed in Appendix A.

However, for neural networks, the learned detector is shaped by the optimization dynamics: the objective is nonconvex, and the solution reached by gradient descent can depend on the initialization

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.

and training trajectory (Jacot et al., 2018; Du et al., 2019; Arora et al., 2019). These dynamics remain poorly understood for DOCC and Deep SAD.

In this work, we derive a finite-time first-order description of the training dynamics of DOCC and Deep SAD. It shows that the representations of the normal training samples contract toward the center, while the representation of an unseen input is displaced by the accumulated updates of the training samples, transferred through the gradient coupling between that input and the training samples.

Adding labeled anomalies during training, as in Deep SAD, introduces an additional anomaly-induced displacement that can help correct missed anomalies. The analysis yields a sufficient condition for an unseen anomaly to remain detectable after DOCC training and, under the first-order approximation, explains when anomaly supervision can move a missed anomaly outside the normal region. Detection is thus governed by the initial separation from the center and the training-induced displacement, which explain the success and failure cases in Sec. 3.3.

Our contributions are as follows.

- We derive a finite-time first-order description of how DOCC training moves the representations of unseen inputs through gradient coupling, and a sufficient condition for an unseen anomaly to remain detectable (Sec. 3.1–3.3).
- We show that, under Gaussian initialization, the gradient coupling concentrates around a deterministic deep ReLU kernel of the input inner product (Sec. 3.4).
- We extend the analysis to Deep SAD, where labeled anomalies generate repulsive updates transferred through the same coupling, and derive a first-order sufficient condition for these updates to correct missed anomalies (Sec. 4).
- We confirm the predicted representation dynamics, anomaly scores, and failure modes experimentally and examine the coupling geometry under Gaussian initialization (Sec. 5).

2 Deep One-Class Classification

This section introduces the one-class AD problem, the DOCC score and objective, the ReLU network that parameterizes the representation, and the gradient-descent procedure by which it is trained.

One-class AD setting. In classical one-class AD, training data is available only from the normal class. We study the case when additional information is available in Sec. 4. Let P_{N} denote the distribution of normal observations on the input domain $\mathcal{X} \subseteq \mathbb{R}^d$, and let $S = (x_1, \dots, x_n) \sim P_{\mathrm{N}}^n$ be the training sample. Our goal is to learn an anomaly score $s_{\mathbf{W}} : \mathcal{X} \to [0, \infty)$ that assigns small values to normal observations and large values to anomalous ones. Given a threshold $\tau > 0$, an observation x is declared anomalous whenever $s_{\mathbf{W}}(x) > \tau$.

DOCC score and objective. The anomaly score of DOCC (Ruff et al., 2018) is defined as the squared distance of a learned representation $\phi_{\mathbf{W}} : \mathcal{X} \to \mathbb{R}^s$ to a fixed center $\mathbf{c} \in \mathbb{R}^s$: $s_{\mathbf{W}}(x) = \|\phi_{\mathbf{W}}(x) - \mathbf{c}\|_2^2$. With this scoring rule, an observation is accepted as normal when its representation lies within the ball of radius $\sqrt{\tau}$ centered at $\mathbf{c}$. The representation is learned through the DOCC objective

$$\min_{\mathbf{W}} F_n(\mathbf{W}), \qquad F_n(\mathbf{W}) = \frac{1}{2n} \sum_{i=1}^{n} \|\phi_{\mathbf{W}}(x_i) - \mathbf{c}\|_2^2 + \frac{\lambda}{2} \|\mathbf{W}\|_2^2, \tag{1}$$

where $\lambda \geq 0$ is the regularization parameter and $\|\mathbf{W}\|_2$ denotes the Euclidean norm of the vector of all trainable parameters. We fix a nonzero center $\mathbf{c}$ throughout training to exclude the trivial zero solution that would be admissible if the center were zero or jointly optimized alongside $\mathbf{W}$.

Deep ReLU network. We parameterize $\phi_{\mathbf{W}}$ by a fully connected ReLU network with L hidden layers of width m and a fixed linear output map, following Zhou et al. (2026a). The network is bias-free, consistent with the recommendation of Ruff et al. (2018) to avoid trivial constant representations enabled by biases. For an input $x \in \mathcal{X}$, we define the layers of our network $h_{\mathbf{W}}^1(x) = \sqrt{2/m}\,\sigma(W^1 x)$ with $W^1 \in \mathbb{R}^{m \times d}$, and $h_{\mathbf{W}}^\ell(x) = \sqrt{2/m}\,\sigma\left(W^\ell h_{\mathbf{W}}^{\ell-1}(x)\right)$ with $W^\ell \in \mathbb{R}^{m \times m}$ for $\ell = 2, \dots, L$, and set $\phi_{\mathbf{W}}(x) = A h_{\mathbf{W}}^L(x)$. Here $\sigma(u) = \max\{u, 0\}$ is applied coordinate-wise. The entries of $A \in \mathbb{R}^{s \times m}$ are drawn independently from $\mathrm{Unif}\{-1, 1\}$, independently of $\mathbf{W}(0)$,

and A is held fixed throughout training. The trainable parameters are $\mathbf{W} = (W^1, \dots, W^L)$, with $\|\mathbf{W}\|_2^2 = \sum_{\ell=1}^{L} \|W^\ell\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

Gradient descent training. After initializing the weights at $\mathbf{W}(0)$, we minimize $F_n(\mathbf{W})$ by gradient descent (GD): $\mathbf{W}(t+1) = \mathbf{W}(t) - \eta \nabla F_n(\mathbf{W}(t))$ for $t = 0, \dots, T-1$, where $\eta > 0$ is the step size. At nondifferentiable points of the ReLU activation, we use the convention $\sigma'(0) = 0$. After T iterations, the learned anomaly score is $s_T(x) := s_{\mathbf{W}(T)}(x) = \|\phi_{\mathbf{W}(T)}(x) - \mathbf{c}\|_2^2$. Although the training objective is evaluated only on S, the shared parameter updates also change the representations of unseen inputs. We study this effect along the training trajectory.

3 A Training-Dynamics View of Deep One-Class Classification

DOCC is trained only on normal samples, yet must distinguish unseen normal and anomalous inputs at test time. A central question is therefore how normal-data training shapes learned representations, both for the training samples themselves and for inputs that never appear during training.

In this section, we develop a training-dynamics description of these effects and use it to understand when unseen anomalies remain detectable, when detection can fail, and how initialization shapes the transfer of normal-data updates. The formal conditions and finite-width error bounds underlying the analysis are given in Appendix B–E.

3.1 How Normal-Data Training Affects Unseen Inputs

We begin by examining how the update generated by a normal training sample x_i affects an unseen input x. Let $r_i(t) := \phi_{\mathbf{W}(t)}(x_i) - \mathbf{c}$ denote the residual of x_i at iteration t, and let $J_{\mathbf{W}}(x)$ denote the Jacobian of $\phi_{\mathbf{W}}(x)$ with respect to the trainable parameters. From the GD update

$$\mathbf{W}(t+1) - \mathbf{W}(t) = -\frac{\eta}{n} \sum\nolimits_{i=1}^{n} J_{\mathbf{W}(t)}(x_i)^\top r_i(t) - \eta\lambda \mathbf{W}(t),$$

the data-fitting update generated by x_i is $-\frac{\eta}{n} J_{\mathbf{W}(t)}(x_i)^\top r_i(t)$.

Although an unseen input x does not appear in the training objective, the update generated by x_i also changes its representation. Linearizing $\phi_{\mathbf{W}}(x)$ w.r.t. $\mathbf{W}$ around $\mathbf{W}(t)$, the first-order contribution of x_i is $-\frac{\eta}{n} J_{\mathbf{W}(t)}(x) J_{\mathbf{W}(t)}(x_i)^\top r_i(t)$. Thus, the gradient coupling $J_{\mathbf{W}(t)}(x) J_{\mathbf{W}(t)}(x_i)^\top$ controls how this update is transferred to x. We use its value at initialization to obtain a time-independent first-order description:

$$K_0(x, x_i) := J_{\mathbf{W}(0)}(x) J_{\mathbf{W}(0)}(x_i)^\top. \tag{2}$$

We then define $u_{i,t}(x) := -\frac{\eta}{n} K_0(x, x_i) r_i(t)$, which gives the first-order contribution of x_i at iteration t to the representation change of the unseen input x.

We now accumulate these contributions over the full training trajectory. Define

$$u_T(x) := \sum_{t=0}^{T-1} \sum_{i=1}^{n} u_{i,t}(x) = \sum_{i=1}^{n} K_0(x, x_i) \beta_i(T) \quad \text{with} \quad \beta_i(T) := -\frac{\eta}{n} \sum_{t=0}^{T-1} r_i(t). \tag{3}$$

Thus, $u_T(x)$ is the leading-order displacement of the unseen representation induced by training. This suggests the first-order description $\phi_{\mathbf{W}(T)}(x) \approx \phi_{\mathbf{W}(0)}(x) + u_T(x)$. The following theorem makes this approximation precise and quantifies the approximation error. Our derivation builds on standard finite-width NTK-style linearization arguments, adapted to the vector-valued DOCC setting. The formal conditions and the explicit expression of E_T are given in Theorem 5 in Appendix B.

Theorem 1 (How training affects unseen inputs)**.** *For any $x \in \mathcal{X}$, under suitable conditions on the initialization, step size, and regularization parameter,*

$$\phi_{\mathbf{W}(T)}(x) = \phi_{\mathbf{W}(0)}(x) + u_T(x) + \rho_T(x) \quad \textit{with} \quad \|\rho_T(x)\|_2 \le E_T,$$

where E_T is an explicit finite-width error bound.

The remainder $\rho_T(x)$ collects the effects omitted by the first-order description. Correspondingly, E_T consists of three components: the local linearization error, the error from replacing the time-varying

gradient coupling by $K_0(x, x_i)$, and the direct contribution of weight decay, which is not included in $u_T(x)$. For the deep ReLU network under Gaussian initialization studied in Sec. 3.4, we have $E_T \lesssim \mathrm{Poly}(s, L, \eta T)(m^{-1/6} + \lambda m)$. For fixed m, η, and λ, the bound generally increases with T, suggesting that the first-order approximation becomes less accurate later in training, as also observed in Fig. 3. The term λm further motivates choosing λ sufficiently small in our experiments. The precise bound is given in Proposition 2 in the appendix.

Theorem 1 shows that, up to the controlled remainder, the representation of an unseen input after training is governed by two components:

- Initial representation: $\phi_{\mathbf{W}(0)}(x)$ determines where the unseen input x starts before training.
- Training-induced displacement: $u_T(x)$ aggregates the cumulative effects of the normal training samples, with $K_0(x, x_i)$ controlling how strongly each sample influences x.

3.2 How Training Shapes the Representations of Normal Training Samples

In Sec. 3.1, we studied how normal-data training affects unseen inputs. Now, we turn to the normal training samples themselves and study how their representations evolve during optimization. Indeed, we show that training progressively drives the normal training representations toward $\mathbf{c}$.

Let the gradient kernel matrix on the normal training samples be $\mathbf{K}_0 := [K_0(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{ns \times ns}$, and let $\lambda_{\min}(\mathbf{K}_0)$ denote its smallest eigenvalue. If $\mathbf{K}_0$ is non-degenerate, in the sense that $\lambda_{\min}(\mathbf{K}_0) \geq \lambda_0 > 0$, then the linearized dynamics are contractive in every residual direction. The following result formalizes this behavior.

Theorem 2 (Contraction of the Normal Training Representations)**.** *Assume that $\lambda_{\min}(\mathbf{K}_0) \geq \lambda_0 > 0$. Under suitable conditions on the initialization, step size, and regularization parameter,*

$$\frac{1}{2n}\sum_{i=1}^{n} \|\phi_{\mathbf{W}(t)}(x_i) - \mathbf{c}\|_2^2 \lesssim \left(1 - \frac{\eta\lambda_0}{4n}\right)^{2t} \frac{1}{2n}\sum_{i=1}^{n} \|\phi_{\mathbf{W}(0)}(x_i) - \mathbf{c}\|_2^2 + \mathrm{Bias}_\lambda,$$

where Bias_λ is a t-independent additive term induced by weight decay and can be made small by choosing λ sufficiently small. The formal conditions and explicit expression of Bias_λ are given in Appendix C.

Theorem 2 shows that, under a non-degenerate $\mathbf{K}_0$, the average squared distance of the normal training representations to the center decreases geometrically, up to the additive effect of weight decay.

3.3 Why Anomalous Inputs Remain Detectable

At test time, DOCC must distinguish between unseen normal and anomalous inputs, although neither participates in training. We now use the trajectory description in Sec. 3.1 to understand when an anomalous input can remain sufficiently separated from the center after training.

Since DOCC detects an input by thresholding its anomaly score $s_T(x) = \|\phi_{\mathbf{W}(T)}(x) - \mathbf{c}\|_2^2$, an anomalous input remains detectable whenever $\|\phi_{\mathbf{W}(T)}(x) - \mathbf{c}\|_2 > \sqrt{\tau}$, where $\tau > 0$ is the decision threshold. By Theorem 1, we have

$$\|\phi_{\mathbf{W}(T)}(x) - \mathbf{c}\|_2 \geq \|\phi_{\mathbf{W}(0)}(x) - \mathbf{c}\|_2 - \|u_T(x)\|_2 - E_T.$$

The effect of the training-induced displacement depends on both its magnitude and direction. The bound above gives a direction-independent sufficient condition in terms of $\|u_T(x)\|_2$. This yields the following sufficient condition for an anomalous input to remain detectable after training.

Theorem 3 (Why anomalous inputs remain detectable)**.** *Under the conditions of Theorem 1, if*

$$\|\phi_{\mathbf{W}(0)}(x) - \mathbf{c}\|_2 - \|u_T(x)\|_2 - E_T > \sqrt{\tau},$$

then $s_T(x) > \tau$, and hence x is detected as anomalous.

Theorem 3 shows that an anomalous input remains detectable when its initial separation from the center is large enough to dominate the training-induced displacement and the trajectory error. The formal assumptions and proof are given in Appendix D.

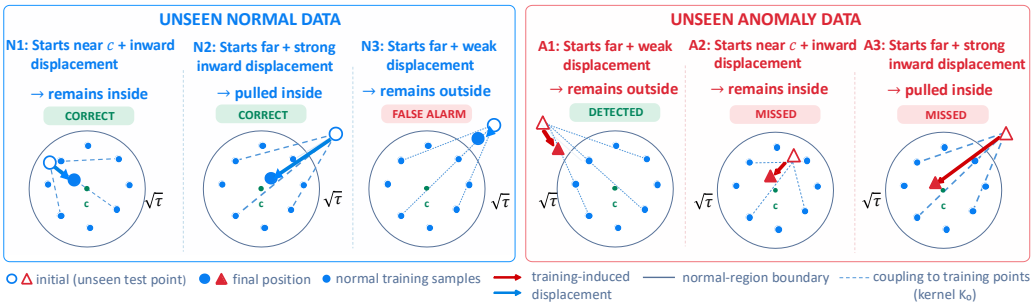


Figure 1: Six representative cases for unseen inputs under DOCC. For unseen normals, acceptance occurs when the input starts close to the center (N1) or is moved sufficiently toward it during training (N2), whereas insufficient movement can lead to a false alarm (N3). For anomalies, detection succeeds when the input remains sufficiently separated from the center (A1), while starting too close (A2) or being moved too far toward the center (A3) can lead to a missed anomaly.

3.3.1 Representative Cases for Unseen Normal and Anomalous Inputs

The preceding analysis identifies two key factors governing the final position of an unseen input: its initial separation from the center and its training-induced displacement $u_T(x)$. Figure 1 illustrates six representative cases based on these two factors.

Unseen normal inputs. For an unseen normal input, successful acceptance occurs when its final representation lies sufficiently close to the center. This can happen either because it starts close to the center or because normal-data training induces a sufficiently large center-directed displacement.

- **Case N1 (correctly accepted normal).** The unseen normal input starts close to the center and remains inside the normal region after training.
- **Case N2 (correctly accepted normal).** The unseen normal input starts farther from the center, but receives a center-directed displacement with sufficiently large magnitude, i.e., large $\|u_T(x)\|_2$, and is brought into the normal region.
- **Case N3 (false alarm).** The unseen normal input starts farther from the center, but the training-induced displacement is too small, i.e., $\|u_T(x)\|_2$ is small, to bring it into the normal region.

Unseen anomalous inputs. For an unseen anomalous input, successful detection occurs when its final representation remains sufficiently separated from the center. Detection can fail either because the anomaly starts too close to the center or because normal-data training induces a sufficiently large center-directed displacement.

- **Case A1 (correctly detected anomaly).** The anomalous input starts sufficiently far from the center and receives only a small training-induced displacement, i.e., $\|u_T(x)\|_2$ is small, so it remains outside the normal region after training.
- **Case A2 (missed anomaly).** The anomalous input starts too close to the center, leaving insufficient separation for detection even when the training-induced displacement is small.
- **Case A3 (missed anomaly).** The anomalous input starts sufficiently far from the center, but receives a center-directed displacement with sufficiently large magnitude, i.e., large $\|u_T(x)\|_2$, and is moved into the normal region.

These cases also highlight the role of initialization in two ways: it determines the initial separation of an unseen input from the center and shapes its coupling with the normal training samples through k_0.

3.4 How Initialization Shapes the Effect of Training

Sec. 3.3 shows that the behavior of an unseen input x depends on the training-induced displacement $u_T(x)$, which is governed by the gradient couplings $K_0(x, x_i)$. To obtain a scalar description of these couplings, we introduce $k_0(x, x') := \langle H_0(x), H_0(x') \rangle_{\mathrm{F}}$ with $H_0(x) = \frac{\partial h^L_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}}$. For the fixed random output matrix A, we have $K_0(x, x') \approx k_0(x, x') I_s$ (see Appendix D for more details). Thus, $k_0(x, x_i)$ provides a scalar measure of how strongly the update generated by x_i is transferred

to x. Since k_0 is defined at initialization, the initialization determines the coupling geometry through which normal-data updates are transferred to unseen inputs. This raises a natural question: *How does initialization shape the coupling $k_0(x, x_i)$ governing this transfer?*

We consider Gaussian initialization. Writing $w_r^\ell(0)^\top$ for the rth row of $W^\ell(0)$, we initialize $w_r^1(0) \sim \mathcal{N}(0, I_d)$ and $w_r^\ell(0) \sim \mathcal{N}(0, I_m)$ for $\ell = 2, \dots, L$, independently across neurons and layers. Under this initialization, $k_0(x, x')$ admits a simple characterization in terms of the input similarity $\langle x, x' \rangle$.

Theorem 4 (Coupling under Gaussian initialization)**.** *Let $\kappa_L : [-1, 1] \to \mathbb{R}$ denote the deterministic depth-L ReLU kernel, which depends only on the depth L and is nondecreasing on $[0, 1]$. Assume $\mathcal{X} \subseteq \mathbb{S}^{d-1}$. Under Gaussian initialization, for sufficiently large width m, with high probability,*

$$\sup_{x,x' \in \mathcal{X}} |k_0(x, x') - \kappa_L(\langle x, x' \rangle)| \lesssim L^{3/2} m^{-1/6} + L(dL \log(m))^{1/2} m^{-1/2} + L^4 m^{-1/3}.$$

The explicit form of κ_L, together with the width and probability conditions, is given in Appendix E.

Theorem 4 shows that, under Gaussian initialization, $k_0(x, x')$ concentrates around the deterministic kernel $\kappa_L(\langle x, x' \rangle)$, which depends only on the network depth L and not on the normal-data distribution or the particular realization of the initialization. Since κ_L is nondecreasing on $[0, 1]$, we have $u_T(x) \approx \sum_{i=1}^n \kappa_L(\langle x, x_i \rangle) \beta_i(T)$, showing that, under Gaussian initialization, raw input similarity governs the transfer strength of normal training updates to unseen inputs.

4 Beyond DOCC Training: The Effect of Anomaly Supervision

This section extends the training-dynamics view of Sec. 3 to Deep Semi-Supervised AD (Deep SAD) (Ruff et al., 2020), which adds a few labeled anomalies to the DOCC objective. Anomaly supervision is known to improve detection (Ruff et al., 2020; Liznerski et al., 2022). We show that labeled anomalies generate repulsive updates, transferred to unseen inputs through the same gradient coupling, whose displacement can help correct missed anomalies.

Deep SAD objective. In addition to the n unlabeled normal samples $\{x_i\}_{i=1}^n$, suppose that $\tilde{n}$ labeled samples $\{(\widetilde{x}_j, \widetilde{y}_j)\}_{j=1}^{\tilde{n}}$ are available, where $\widetilde{y}_j \in \{-1, +1\}$, with $\widetilde{y}_j = +1$ denoting a normal sample and $\widetilde{y}_j = -1$ an anomalous sample. Deep SAD considers

$$\min_{\mathbf{W}} \frac{1}{2(n + \tilde{n})} \Big[\sum_{i=1}^n \|\phi_{\mathbf{W}}(x_i) - \mathbf{c}\|_2^2 + \nu \sum_{j=1}^{\tilde{n}} \big(\|\phi_{\mathbf{W}}(\widetilde{x}_j) - \mathbf{c}\|_2^2 \big)^{\widetilde{y}_j} \Big] + \frac{\lambda}{2} \sum_{\ell=1}^L \|W^\ell\|_{\mathrm{F}}^2,$$

where $\nu > 0$ controls the weight of the labeled term. A labeled normal sample contributes the usual center-seeking loss, whereas a labeled anomaly contributes $\|\phi_{\mathbf{W}}(\widetilde{x}_j) - \mathbf{c}\|_2^{-2}$ and therefore induces a repulsive update away from $\mathbf{c}$.

We now characterize the additional representation displacement induced by anomaly supervision. For a labeled sample $\widetilde{x}_j$, define $\widetilde{r}_j(t) := \phi_{\mathbf{W}(t)}(\widetilde{x}_j) - \mathbf{c}$. Following the same first-order decomposition as in Sec. 3.1, after T iterations,

$$\phi_{\mathbf{W}(T)}(x) \approx \phi_{\mathbf{W}(0)}(x) + u_T^{\mathrm{nor}}(x) + u_T^{\mathrm{ano}}(x). \tag{4}$$

Here, $u_T^{\mathrm{nor}}(x)$ collects the contributions from the unlabeled and labeled normal samples, while the labeled anomalies induce $u_T^{\mathrm{ano}}(x) = \sum_{j:\widetilde{y}_j=-1} K_0(x, \widetilde{x}_j) \widetilde{\beta}_j^{\mathrm{ano}}(T)$, with $\widetilde{\beta}_j^{\mathrm{ano}}(T) = \frac{\eta\nu}{n+\tilde{n}} \sum_{t=0}^{T-1} \widetilde{r}_j(t) \|\widetilde{r}_j(t)\|_2^{-4}$.

Thus, labeled normal samples generate center-seeking updates, whereas labeled anomalies generate updates that push their representations away from $\mathbf{c}$. Indeed, the negative-gradient direction of the anomaly loss satisfies $-\nabla_{\phi_{\mathbf{W}(t)}(\widetilde{x}_j)} \|\phi_{\mathbf{W}(t)}(\widetilde{x}_j) - \mathbf{c}\|_2^{-2} = \frac{2\widetilde{r}_j(t)}{\|\widetilde{r}_j(t)\|_2^4}$, which points away from the center. For an unseen input x, the repulsive update generated by the labeled anomaly $\widetilde{x}_j$ is transferred through the same gradient-coupling mechanism as before. Under the scalarization $K_0(x, x') \approx k_0(x, x') I_s$, the strength of the contribution from a labeled anomaly $\widetilde{x}_j$ is governed by $|k_0(x, \widetilde{x}_j)|$.

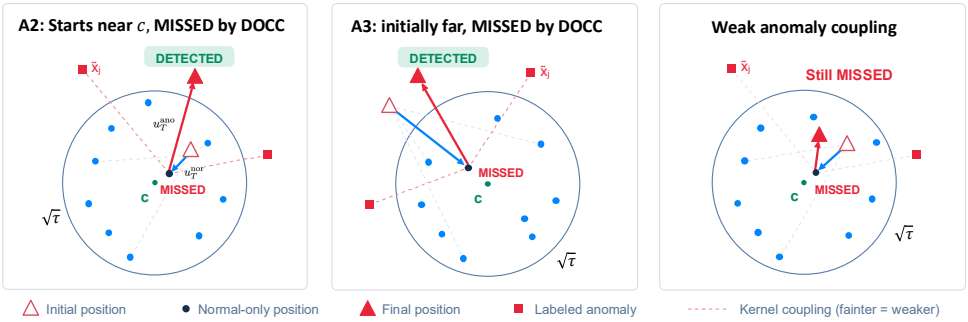


Figure 2: Illustration of how anomaly supervision can correct the A2 and A3 failure modes in Deep SAD. Normal-data training induces an inward displacement that can move an anomaly into the normal region, while anomaly supervision adds an outward displacement u_T^{ano} that can push it back outside. If the coupling to labeled anomalies is too weak, this outward correction may be insufficient.

4.1 Repulsive Displacement Corrects Missed Anomalies

We derive a condition under which the additional displacement induced by anomaly supervision moves an anomaly outside the normal region when the normal contribution alone leaves it inside. Define $q_T^{\text{nor}}(x) := \phi_{\mathbf{W}(0)}(x) - \mathbf{c} + u_T^{\text{nor}}(x)$, which is the first-order residual induced by the normal-sample contribution along the same Deep SAD trajectory, before adding the anomaly-induced contribution.

Lemma 1 (Correction by anomaly supervision)**.** *Under the first-order approximation in Equation 4, suppose that an anomalous input x is missed before adding the anomaly-induced displacement, i.e., $\|q_T^{\text{nor}}(x)\|_2^2 \le \tau$. If*

$$2\langle q_T^{\text{nor}}(x), u_T^{\text{ano}}(x)\rangle + \|u_T^{\text{ano}}(x)\|_2^2 > \tau - \|q_T^{\text{nor}}(x)\|_2^2, \tag{5}$$

then the first-order representation lies outside the normal region, i.e., $\|q_T^{\text{nor}}(x) + u_T^{\text{ano}}(x)\|_2^2 > \tau$.

Lemma 1 gives a sufficient condition under the first-order approximation. Condition 5 admits a simple geometric interpretation. The anomaly-induced displacement changes the squared distance of x from the center by $2\langle q_T^{\text{nor}}(x), u_T^{\text{ano}}(x)\rangle + \|u_T^{\text{ano}}(x)\|_2^2$, while $\tau - \|q_T^{\text{nor}}(x)\|_2^2$ is the remaining squared-distance gap to the detection boundary. Hence, condition 5 states that anomaly supervision corrects the missed anomaly whenever this increase exceeds the remaining gap. This explains how Deep SAD can correct the two missed-anomaly failure modes identified in Figure 1:

- **Case A2: starts near c and remains inside.** The anomaly starts too close to the center and remains inside under the normal-sample contribution. A sufficiently large outward anomaly-induced displacement can then move it outside the normal region and restore detection.
- **Case A3: starts far but is moved inside.** The anomaly starts sufficiently far from the center but is moved inside by the normal-sample contribution. An outward anomaly-induced displacement can counteract this movement and restore detection.

Figure 2 illustrates these mechanisms and shows that weak coupling to labeled anomalies can leave a missed anomaly uncorrected. Our analysis thus explains how anomaly supervision can correct the A2 and A3 failure modes, while a complete analysis of Deep SAD is left for future work.

5 Experiments

We empirically examine the representation dynamics, detection behavior, and coupling structure predicted by our analysis.

Experimental Setup. Following Ruff et al. (2018), we construct ten anomaly-detection setups from MNIST (Deng, 2012), taking each class in turn as normal and the remaining classes as anomalies. Unless stated otherwise, figures show the setup with digit 3 as the normal class. We use 1000 training normals, 1000 held-out normals, and 1000 anomalies, with all inputs normalized to unit Euclidean norm. We use the bias-free ReLU network in Sec. 2 with $L = 3$, $s = 16$, and

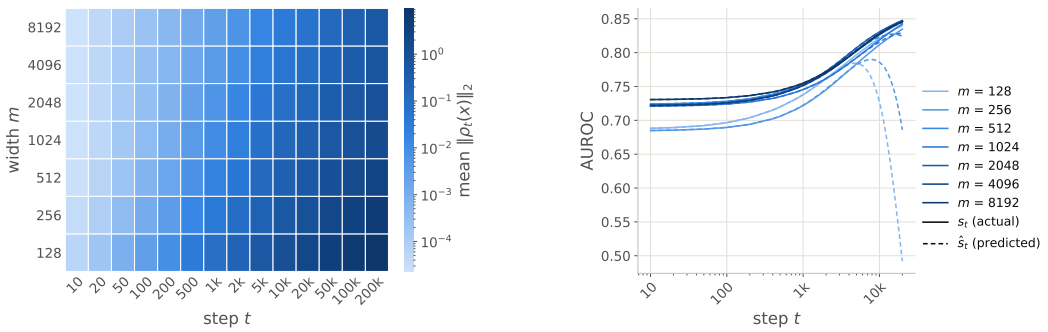


Figure 3: First-order trajectory approximation ($\lambda = 10^{-4}$, $T = 200$k, three seeds). LEFT: mean error $\|\rho_t(x)\|_2$ over held-out normals, averaged over the ten anomaly-detection setups. RIGHT: AUROC of the anomaly score s_t and of its first-order approximation $\widehat{s}_t$ for digit 3, $t \leq 20$k. The error grows with t and shrinks with m, more slowly late in training because the weight-decay part of ρ_t does not shrink with m. The predicted AUROC matches the actual one early and deviates later, earlier for narrower networks.

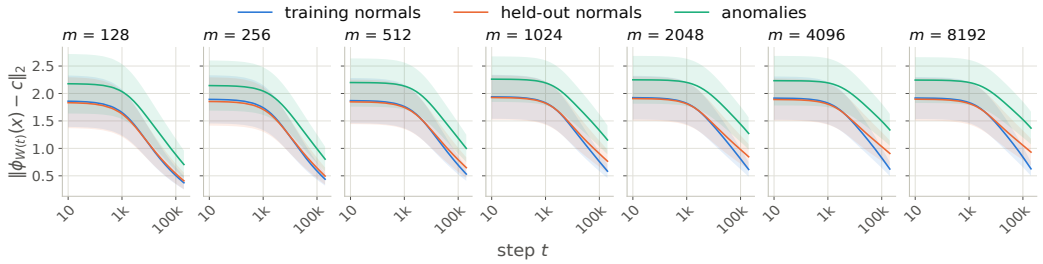


Figure 4: Distance to the center $\mathbf{c}$ during training across widths m (digit 3, $\lambda = 10^{-4}$, per-group means over three seeds, ± 1 std band). As predicted in Sec. 3, training normals contract toward $\mathbf{c}$, and unseen inputs move only through their coupling to them. Held-out normals follow at small m but lag at large m (0.93 vs. 0.63 for training normals at T, $m = 8192$), and anomalies move toward $\mathbf{c}$ but remain separated.

$m \in \{128, 256, \ldots, 8192\}$, unless stated otherwise. The network is Gaussian initialized. Each configuration is run with three random seeds. For anomaly supervision, we train Deep SAD (Sec. 4) with $\tilde{n} = 50$ labeled anomalies from the training split of the other digits and $\nu = 100$, from the same initialization and center as DOCC. Further experimental details are provided in Appendix G. We compare the actual anomaly score $s_t(x) = \|\phi_{\mathbf{W}(t)}(x) - \mathbf{c}\|_2^2$ with its first-order approximation $\widehat{s}_t(x) = \|\phi_{\mathbf{W}(0)}(x) + u_t(x) - \mathbf{c}\|_2^2$. Here, $u_t(x)$ is computed using the realized training residuals, while the coupling is fixed at its initialization value K_0. We report AUROC.

5.1 EMPIRICAL RESULTS

We organize the results around four aspects of our analysis: trajectory approximation, normal-representation contraction, anomaly detection, and initialization-induced coupling.

Accuracy of the trajectory approximation. Under Gaussian initialization, Theorem 1 gives $\|\rho_t(x)\|_2 \leq E_t \lesssim \mathrm{Poly}(s, L, \eta t)(m^{-1/6} + \lambda m)$. The underlying width conditions are conservative and are not expected to be tight at the experimental widths. Figure 3 qualitatively supports the predicted finite-width behavior: the first-order trajectory closely tracks detection performance at early stages, while the anomaly-score approximation becomes less accurate as the remainder grows.

Convergence of normal training representations. Figure 4 supports Theorem 2, which predicts contraction of the normal training representations toward $\mathbf{c}$. The behavior of the held-out samples further illustrates how normal-data training affects unseen inputs differently.

Initial separation, training-induced movement, and anomaly supervision. Figure 5 supports the two factors highlighted in Theorem 3: initial separation and training-induced movement jointly determine whether an anomaly remains detectable. The first-order prediction captures the observed

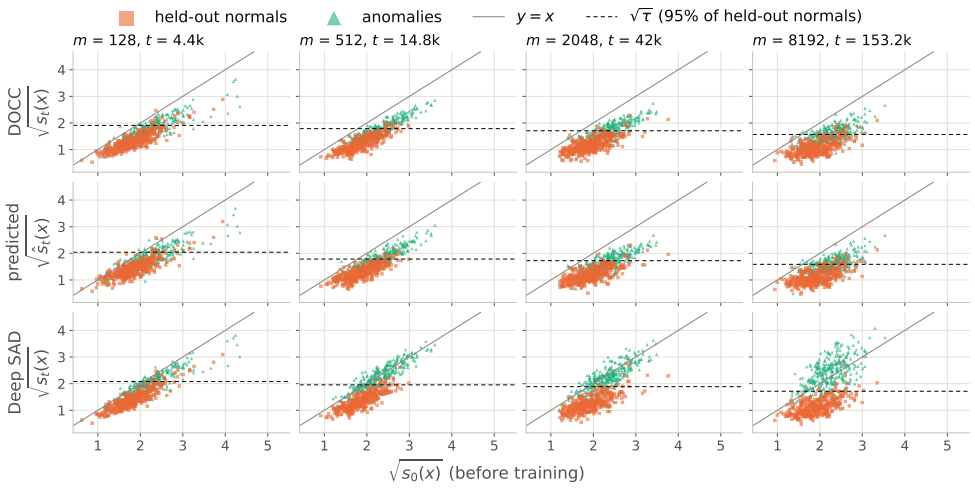


Figure 5: Distance to the center at initialization versus at step t (MNIST, digit 3), where t is the last step at which the first-order AUROC is within 0.01 of the actual AUROC. TOP: DOCC trajectories. MIDDLE: their first-order predictions. BOTTOM: Deep SAD trajectories. Under DOCC, anomalies are missed when they start close to the center (A2) or are moved inside the detection boundary (A3). As Lemma 1 predicts, Deep SAD moves many of them back outside, more so at larger t (wider networks), where the accumulated repulsive displacement is larger.

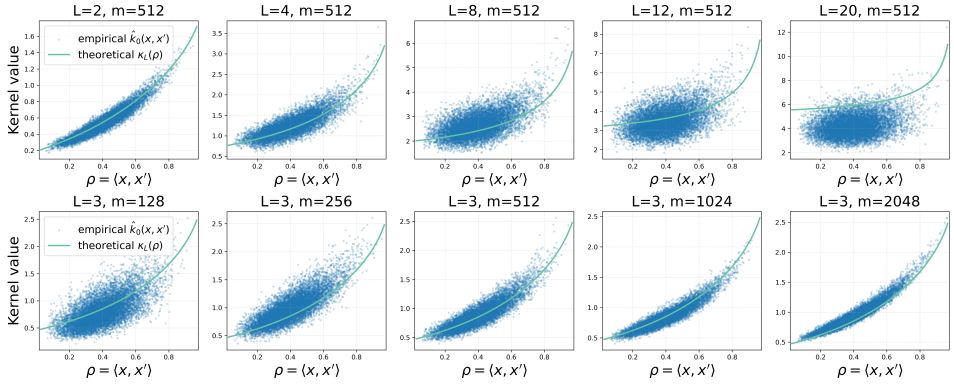


Figure 6: Empirical coupling $\hat{k}_0(x, x')$ against $\rho = \langle x, x' \rangle$ under Gaussian initialization versus the deterministic kernel $\kappa_L(\rho)$ of Theorem 4. As predicted, the coupling concentrates around κ_L as width m grows, and deviates more with depth L.

DOCC dynamics, while Deep SAD corrects many missed anomalies through the additional outward displacement analyzed in Sec. 4.

Coupling under Gaussian initialization. Figure 6 supports Theorem 4, showing that the finite-width coupling converges toward the deterministic deep ReLU kernel and that the approximation becomes less accurate for deeper networks.

6 CONCLUSION AND LIMITATIONS

In this paper, we provide a finite-time training-dynamics view of DOCC, showing how normal-data updates affect unseen representations and, consequently, the final anomaly score. This viewpoint explains representative success and failure cases and extends naturally to Deep SAD.

Our analysis explains how individual unseen representations respond to training through gradient coupling, while the direction of the induced displacement is not explicitly studied. A natural direction for future work is to understand when training induces systematically different responses for unseen normal and anomalous inputs. The extension to anomaly supervision mainly explains how labeled anomalies can improve missed-detection cases such as A2 and A3, while false-alarm cases such as N3 remain open.

AI USE STATEMENT

We used generative AI tools to improve the readability of the manuscript, identify relevant literature, suggest the manuscript structure, and propose titles and keywords. We also used these tools as a research assistant for interpreting results and checking proofs. All AI-assisted material was independently verified, revised, and integrated by the authors. Generative AI was not used to generate experimental data or results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3349–3356, 2020.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net`.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. volume 32, 2019.

Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. In *International Conference on Learning Representations (ICLR)*, 2021.

Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Klaus-Robert MÃžller, and Marius Kloft. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *Transactions on Machine Learning Research*, 2022.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 2018.

Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *International Conference on Learning Representations*, 2020.

Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Klaus-Robert Müller, and Marius Kloft. Rethinking assumptions in deep anomaly detection. In *ICML 2021 Workshop on Uncertainty & Robustness in Deep Learning*, 2021.

Lifeng Shen, Zhuocong Li, and James Kwok. Timeseries anomaly detection using temporal hierarchical one-class network. *Advances in neural information processing systems*, 33:13016–13026, 2020.

Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(2):211, 2005.

Che-Ping Tsai, Ganyu Teng, Phillip Wallis, and Wei Ding. AnoLLM: Large language models for tabular anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7VkHffT5X2.

Régis Vert and Jean-Philippe Vert. Consistency and convergence rates of one-class svms and related algorithms. *The Journal of Machine Learning Research*, 7:817–854, 2006.

Puyu Wang, Yunwen Lei, Di Wang, Yiming Ying, and Ding-Xuan Zhou. Generalization guarantees of gradient descent for shallow neural networks. *Neural Computation*, 37(2):344–402, 2025.

Xu Yao, Siyuan Zhou, Zhenbo Wu, Chaochuan Hou, Shuang Liang, Shiping Wang, Hailiang Huang, Songqiao Han, and Minqi Jiang. Rethinking weak supervision in anomaly detection: A comprehensive benchmark. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 10222–10233, 2026.

Hangting Ye, He Zhao, Wei Fan, Mingyuan Zhou, Dan dan Guo, and Yi Chang. DRL: Decomposed representation learning for tabular anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CJnceDksRd.

Junyu Zhou, Puyu Wang, Yunwen Lei, Yiming Ying, and Ding-Xuan Zhou. Optimal rates for generalization of gradient descent methods with deep neural networks. *arXiv preprint arXiv:2606.06764*, 2026a.

Tian-Yi Zhou, Matthew Lau, Jizhou Chen, Wenke Lee, and Xiaoming Huo. Learning to detect cyber attacks: Neural anomaly detection for cybersecurity with theoretical insights, 2026b. URL https://arxiv.org/abs/2409.08521.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.

Appendix

A Further Related Work

Theoretical studies of anomaly detection. Compared with the large empirical literature on anomaly detection, theoretical analysis remain relatively limited. Steinwart et al. (2005) developed a classification framework for anomaly detection and established consistency guarantees for kernel-based detectors. Closely related to the one-class setting, Vert & Vert (2006) provided a theoretical analysis of one-class SVMs, including consistency and convergence guarantees. More recently, Zhou et al. (2026b) established minimax-optimal excess-risk guarantees for neural anomaly detection with synthetic anomalies. Despite these advances, theoretical understanding of modern deep one-class objectives and their training dynamics remains limited.

Neural tangent kernel and lazy-training analyses. Our analysis builds on the neural tangent kernel (NTK) and lazy-training literature, which studies the evolution of wide neural networks through their linearization around initialization (Jacot et al., 2018; Lee et al., 2019; Chizat et al., 2019; Du et al., 2019; Arora et al., 2019; Wang et al., 2025; Cao & Gu, 2020; Zou et al., 2020; Cao & Gu, 2019). We use this framework as a technical tool to control the finite-time representation trajectory (Arora et al., 2019; Du et al., 2019) and the stability of the gradient coupling for ReLU networks using the estimates of Zhou et al. (2026a). Our focus differs in using these dynamics to study deep one-class classification, where the central question is how updates generated by normal training samples are transferred to unseen normal and anomalous inputs and how this transfer affects detection.

B Formal Guarantees for Sec. 3.1

Define the data-fitting loss

$$\mathcal{L}_n(\mathbf{W}) := \frac{1}{2n}\sum_{i=1}^{n} \|\phi_{\mathbf{W}}(x_i) - \mathbf{c}\|_2^2,$$

so that $F_n(\mathbf{W}) = \mathcal{L}_n(\mathbf{W}) + \frac{\lambda}{2}\|\mathbf{W}\|_2^2$.

Let P denote the total number of trainable parameters. Identifying $\mathbf{W} = (W^1, \dots, W^L)$ with its vectorization in $\mathbb{R}^P$, the Jacobian of the representation $\phi_{\mathbf{W}}(x) = (\phi_{\mathbf{W},1}(x), \dots, \phi_{\mathbf{W},s}(x))^\top$ takes the form

$$J_{\mathbf{W}}(x) := \frac{\partial \phi_{\mathbf{W}}(x)}{\partial \mathbf{W}} = \begin{bmatrix} \nabla_{\mathbf{W}}\phi_{\mathbf{W},1}(x)^\top \\ \vdots \\ \nabla_{\mathbf{W}}\phi_{\mathbf{W},s}(x)^\top \end{bmatrix} \in \mathbb{R}^{s\times P}.$$

In particular, for any $r \in \mathbb{R}^s$, $J_{\mathbf{W}}(x)^\top r = \sum_{j=1}^{s} r_j \nabla_{\mathbf{W}}\phi_{\mathbf{W},j}(x) \in \mathbb{R}^P$.

For $R > 0$, let $\mathcal{B}_R(\mathbf{W}(0)) := \{\mathbf{W} : \|\mathbf{W} - \mathbf{W}(0)\|_{\mathrm{op},\infty} \le R\}$, where $\|\mathbf{W}\|_{\mathrm{op},\infty} := \max_{\ell\in[L]} \|W^\ell\|_{\mathrm{op}}$. For a fixed unseen input $x \in \mathcal{X}$, let $\mathcal{Z}_x := \{x_1, \dots, x_n\} \cup \{x\}$. For brevity, we write $J_0(x) := J_{\mathbf{W}(0)}(x)$ for the Jacobian at initialization.

Assumption 1 (Initialization and local regularity)**.** *Suppose that the initialization satisfies* $\|\mathbf{W}(0)\|_2 \le M_0$ *and* $\max_{z\in\mathcal{Z}_x} \|J_0(z)\|_{\mathrm{op}} \le G_0$. *Moreover, for* $R > 0$*, there exist* $\varepsilon_{\mathrm{grad}}(R), \varepsilon_{\mathrm{lin}}(R) \ge 0$ *such that, for all* $\mathbf{W}, \widetilde{\mathbf{W}} \in \mathcal{B}_R(\mathbf{W}(0))$ *and* $z \in \mathcal{Z}_x$*,*

$$\|J_{\mathbf{W}}(z) - J_0(z)\|_{\mathrm{op}} \le \varepsilon_{\mathrm{grad}}(R),$$
$$\|\phi_{\widetilde{\mathbf{W}}}(z) - \phi_{\mathbf{W}}(z) - J_{\mathbf{W}}(z)(\widetilde{\mathbf{W}} - \mathbf{W})\|_2 \le \varepsilon_{\mathrm{lin}}(R)\|\widetilde{\mathbf{W}} - \mathbf{W}\|_{\mathrm{op},\infty}.$$

The quantities M_0 and G_0 control the size of the initialization and of its Jacobian, while $\varepsilon_{\mathrm{grad}}(R)$ and $\varepsilon_{\mathrm{lin}}(R)$ quantify the Jacobian drift and the departure from linearity near the initialization. We will verify these conditions for the deep ReLU network under Gaussian initialization in Proposition 1 of Appendix E.

Throughout this section, fix an arbitrary input $x \in \mathcal{X}$ and consider Assumption 1 with the corresponding set $\mathcal{Z}_x$.

We first show that the GD iterates remain close to the initialization. Define $R_T := \sqrt{8\eta T\,\mathcal{L}_n(\mathbf{W}(0))}$.

Lemma 2. *Suppose that Assumption 1 holds with $R = R_T$. Assume that $\eta\big(G_0+\varepsilon_{\mathrm{grad}}(R_T)\big)^2 \le 1/8$ and $\varepsilon_{\mathrm{lin}}(R_T)R_T \le \sqrt{\mathcal{L}_n(\mathbf{W}(0))/2}$. Assume further that*

$$\lambda R_T M_0 + \eta\lambda^2(M_0+R_T)^2 \le \frac{1}{2}\mathcal{L}_n(\mathbf{W}(0)).$$

Then, for every $t = 0,\dots,T$,

$$\|\mathbf{W}(t)-\mathbf{W}(0)\|_{\mathrm{op},\infty}^2 \le \|\mathbf{W}(t)-\mathbf{W}(0)\|_2^2 \le 8\eta t\,\mathcal{L}_n(\mathbf{W}(0)). \tag{6}$$

In particular, $\mathbf{W}(t) \in \mathcal{B}_{R_T}(\mathbf{W}(0))$ for all $t \le T$. Moreover,

$$\sum_{t=0}^{T-1}\mathcal{L}_n(\mathbf{W}(t)) \le 8T\,\mathcal{L}_n(\mathbf{W}(0)).$$

Proof. For brevity, write $V_t := \mathbf{W}(t)-\mathbf{W}(0)$ and $\mathcal{L}_t := \mathcal{L}_n(\mathbf{W}(t))$. We prove equation 6 by induction on t. The claim is immediate for $t = 0$. Suppose that $\|V_t\|_2^2 \le 8\eta t\,\mathcal{L}_0$ for some $t \le T-1$. Then $\|V_t\|_{\mathrm{op},\infty} \le \|V_t\|_2 \le R_T$, and hence $\mathbf{W}(t) \in \mathcal{B}_{R_T}(\mathbf{W}(0))$.

Recall that $r_i(t) := \phi_{\mathbf{W}(t)}(x_i) - \mathbf{c}$. Note that $\nabla\mathcal{L}_n(\mathbf{W}(t)) = (1/n)\sum_{i=1}^n J_{\mathbf{W}(t)}(x_i)^\top r_i(t)$ and Assumption 1 implies $\max_{i\in[n]}\|J_{\mathbf{W}(t)}(x_i)\|_{\mathrm{op}} \le G_0+\varepsilon_{\mathrm{grad}}(R_T)$. Therefore, by Jensen's inequality and the Jacobian bound,

$$\|\nabla\mathcal{L}_n(\mathbf{W}(t))\|_2^2 \le \frac{1}{n}\sum_{i=1}^n \|J_{\mathbf{W}(t)}(x_i)^\top r_i(t)\|_2^2 \le 2\big(G_0+\varepsilon_{\mathrm{grad}}(R_T)\big)^2\mathcal{L}_t. \tag{7}$$

We next control the inner product between the gradient and V_t. By the convexity of $u \mapsto \frac{1}{2}\|u-\mathbf{c}\|_2^2$, for each $i \in [n]$,

$$\frac{1}{2}\|\phi_{\mathbf{W}(0)}(x_i)-\mathbf{c}\|_2^2 \ge \frac{1}{2}\|\phi_{\mathbf{W}(t)}(x_i)-\mathbf{c}\|_2^2 + \big\langle r_i(t), \phi_{\mathbf{W}(0)}(x_i)-\phi_{\mathbf{W}(t)}(x_i)\big\rangle.$$

Averaging over the training samples gives

$$\mathcal{L}_0-\mathcal{L}_t \ge \frac{1}{n}\sum_{i=1}^n \big\langle r_i(t), \phi_{\mathbf{W}(0)}(x_i)-\phi_{\mathbf{W}(t)}(x_i)\big\rangle. \tag{8}$$

Since both $\mathbf{W}(t)$ and $\mathbf{W}(0)$ belong to $\mathcal{B}_{R_T}(\mathbf{W}(0))$, Assumption 1 gives

$$\phi_{\mathbf{W}(0)}(x_i)-\phi_{\mathbf{W}(t)}(x_i) = -J_{\mathbf{W}(t)}(x_i)V_t + e_i(t),$$

where $\|e_i(t)\|_2 \le \varepsilon_{\mathrm{lin}}(R_T)R_T$. Substituting this into equation 8 yields

$$\mathcal{L}_0-\mathcal{L}_t \ge -\langle V_t, \nabla\mathcal{L}_n(\mathbf{W}(t))\rangle + \frac{1}{n}\sum_{i=1}^n\langle r_i(t), e_i(t)\rangle.$$

By the Cauchy–Schwarz inequality,

$$-\langle V_t, \nabla\mathcal{L}_n(\mathbf{W}(t))\rangle \le \mathcal{L}_0-\mathcal{L}_t+\varepsilon_{\mathrm{lin}}(R_T)R_T\sqrt{2\mathcal{L}_t}. \tag{9}$$

The GD update yields

$$V_{t+1} = V_t - \eta\nabla\mathcal{L}_n(\mathbf{W}(t)) - \eta\lambda\mathbf{W}(t).$$

Hence,

$$\begin{aligned}\|V_{t+1}\|_2^2 \le \|V_t\|_2^2 &- 2\eta\,\langle V_t, \nabla\mathcal{L}_n(\mathbf{W}(t))\rangle - 2\eta\lambda\langle V_t, \mathbf{W}(t)\rangle \\ &+ 2\eta^2\|\nabla\mathcal{L}_n(\mathbf{W}(t))\|_2^2 + 2\eta^2\lambda^2\|\mathbf{W}(t)\|_2^2.\end{aligned} \tag{10}$$

Since $\mathbf{W}(t) = \mathbf{W}(0)+V_t$, by the Cauchy–Schwarz inequality and the induction hypothesis,

$$-\langle V_t, \mathbf{W}(t)\rangle = -\|V_t\|_2^2 - \langle V_t, \mathbf{W}(0)\rangle \le \|V_t\|_2\|\mathbf{W}(0)\|_2 \le R_T M_0.$$

Moreover, $\|\mathbf{W}(t)\|_2 \le M_0 + R_T$. Substituting equation 7 and equation 9 into equation 10, we obtain

$$\|V_{t+1}\|_2^2 \le \|V_t\|_2^2 + 2\eta\mathcal{L}_0 - 2\eta\mathcal{L}_t + 4\eta^2\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)^2\mathcal{L}_t + 2\eta\varepsilon_{\mathrm{lin}}(R_T)R_T\sqrt{2\mathcal{L}_t} + 2\eta\lambda R_T M_0 + 2\eta^2\lambda^2(M_0 + R_T)^2. \tag{11}$$

By Young's inequality,

$$2\varepsilon_{\mathrm{lin}}(R_T)R_T\sqrt{2\mathcal{L}_t} \le \mathcal{L}_t + 2\varepsilon_{\mathrm{lin}}(R_T)^2R_T^2.$$

The step-size condition gives $4\eta^2\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)^2\mathcal{L}_t \le (\eta/2)\mathcal{L}_t$. Furthermore, $\varepsilon_{\mathrm{lin}}(R_T)R_T \le \sqrt{\mathcal{L}_0/2}$ implies $2\eta\varepsilon_{\mathrm{lin}}(R_T)^2R_T^2 \le \eta\mathcal{L}_0$, while the condition on λ gives

$$2\eta\lambda R_T M_0 + 2\eta^2\lambda^2(M_0 + R_T)^2 \le \eta\mathcal{L}_0.$$

Therefore, equation 11 yields

$$\|V_{t+1}\|_2^2 \le \|V_t\|_2^2 - \frac{\eta}{2}\mathcal{L}_t + 4\eta\mathcal{L}_0. \tag{12}$$

Dropping the nonpositive term and using the induction hypothesis gives

$$\|V_{t+1}\|_2^2 \le 8\eta t\,\mathcal{L}_0 + 4\eta\mathcal{L}_0 \le 8\eta(t+1)\mathcal{L}_0.$$

This closes the induction and proves equation 6.

Finally, summing equation 12 over $t = 0, \dots, T-1$ and using $V_0 = 0$ and $\|V_T\|_2^2 \ge 0$ gives

$$\frac{\eta}{2}\sum_{t=0}^{T-1}\mathcal{L}_t \le 4\eta T\mathcal{L}_0 \quad \text{and hence} \quad \sum_{t=0}^{T-1}\mathcal{L}_t \le 8T\mathcal{L}_0,$$

which completes the proof. □

We next decompose the parameter change into a leading term determined by the Jacobian at initialization and an error caused by Jacobian drift and weight decay. Recall that $\beta_i(T) := -(\eta/n)\sum_{t=0}^{T-1} r_i(t)$.

Lemma 3. *Under the conditions of Lemma 2,*

$$\mathbf{W}(T) - \mathbf{W}(0) = \sum_{i=1}^{n} J_0(x_i)^\top \beta_i(T) + \xi_T,$$

where

$$\|\xi_T\|_2 \le 4\eta T\,\varepsilon_{\mathrm{grad}}(R_T)\sqrt{\mathcal{L}_n(\mathbf{W}(0))} + \eta\lambda T(M_0 + R_T).$$

Proof. By the GD update,

$$\mathbf{W}(t+1) - \mathbf{W}(t) = -\frac{\eta}{n}\sum_{i=1}^{n} J_{\mathbf{W}(t)}(x_i)^\top r_i(t) - \eta\lambda\mathbf{W}(t).$$

Summing over $t = 0, \dots, T-1$ and adding and subtracting $J_0(x_i)$ gives

$$\mathbf{W}(T) - \mathbf{W}(0) = -\frac{\eta}{n}\sum_{t=0}^{T-1}\sum_{i=1}^{n} J_0(x_i)^\top r_i(t) - \frac{\eta}{n}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\big(J_{\mathbf{W}(t)}(x_i) - J_0(x_i)\big)^\top r_i(t) - \eta\lambda\sum_{t=0}^{T-1}\mathbf{W}(t). \tag{13}$$

By the definition of $\beta_i(T)$, the first term on the right-hand side equals $\sum_{i=1}^{n} J_0(x_i)^\top \beta_i(T)$. Hence,

$$\mathbf{W}(T) - \mathbf{W}(0) = \sum_{i=1}^{n} J_0(x_i)^\top \beta_i(T) + \xi_T,$$

where

$$\xi_T = -\frac{\eta}{n}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\big(J_{\mathbf{W}(t)}(x_i) - J_0(x_i)\big)^\top r_i(t) - \eta\lambda\sum_{t=0}^{T-1}\mathbf{W}(t).$$

By Lemma 2, $\mathbf{W}(t) \in \mathcal{B}_{R_T}(\mathbf{W}(0))$ for all $t \le T$. Therefore, Assumption 1 gives $\|J_{\mathbf{W}(t)}(x_i) - J_0(x_i)\|_{\mathrm{op}} \le \varepsilon_{\mathrm{grad}}(R_T)$, while $\|\mathbf{W}(t)\|_2 \le M_0 + R_T$. It follows that

$$\|\xi_T\|_2 \le \frac{\eta\varepsilon_{\mathrm{grad}}(R_T)}{n}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\|r_i(t)\|_2 + \eta\lambda T(M_0 + R_T). \tag{14}$$

For each t, the Cauchy–Schwarz inequality gives

$$\frac{1}{n}\sum_{i=1}^{n}\|r_i(t)\|_2 \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|r_i(t)\|_2^2} = \sqrt{2\mathcal{L}_t}.$$

Hence, using Lemma 2,

$$\frac{\eta}{n}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\|r_i(t)\|_2 \le \eta\sum_{t=0}^{T-1}\sqrt{2\mathcal{L}_t} \le \eta\sqrt{2T\sum_{t=0}^{T-1}\mathcal{L}_t} \le 4\eta T\sqrt{\mathcal{L}_n(\mathbf{W}(0))}.$$

Substituting this bound into equation 14 completes the proof. □

We next translate the parameter trajectory decomposition into a representation trajectory decomposition for the fixed input $x \in \mathcal{X}$. Recall that $K_0(x, x_i) := J_0(x)J_0(x_i)^\top$.

Lemma 4. *Under the conditions of Lemma 2,*

$$\phi_{\mathbf{W}(T)}(x) = \phi_{\mathbf{W}(0)}(x) + \sum_{i=1}^{n}K_0(x, x_i)\beta_i(T) + \rho_T(x),$$

where

$$\|\rho_T(x)\|_2 \le \varepsilon_{\mathrm{lin}}(R_T)R_T + 4\eta T G_0\varepsilon_{\mathrm{grad}}(R_T)\sqrt{\mathcal{L}_n(\mathbf{W}(0))} + \eta\lambda T G_0(M_0 + R_T).$$

Proof. By Lemma 2, both $\mathbf{W}(T)$ and $\mathbf{W}(0)$ belong to $\mathcal{B}_{R_T}(\mathbf{W}(0))$. Hence, Assumption 1 gives

$$\phi_{\mathbf{W}(T)}(x) - \phi_{\mathbf{W}(0)}(x) = J_0(x)\big(\mathbf{W}(T) - \mathbf{W}(0)\big) + e_T(x),$$

where $\|e_T(x)\|_2 \le \varepsilon_{\mathrm{lin}}(R_T)R_T$.

Applying Lemma 3, we obtain

$$\begin{aligned}\phi_{\mathbf{W}(T)}(x) - \phi_{\mathbf{W}(0)}(x) &= J_0(x)\sum_{i=1}^{n}J_0(x_i)^\top\beta_i(T) + J_0(x)\xi_T + e_T(x)\\ &= \sum_{i=1}^{n}K_0(x, x_i)\beta_i(T) + J_0(x)\xi_T + e_T(x).\end{aligned}$$

Thus,

$$\phi_{\mathbf{W}(T)}(x) = \phi_{\mathbf{W}(0)}(x) + \sum_{i=1}^{n}K_0(x, x_i)\beta_i(T) + \rho_T(x),$$

with $\rho_T(x) := J_0(x)\xi_T + e_T(x)$. Since $\|J_0(x)\|_{\mathrm{op}} \le G_0$, we have

$$\|\rho_T(x)\|_2 \le \varepsilon_{\mathrm{lin}}(R_T)R_T + 4\eta T G_0\varepsilon_{\mathrm{grad}}(R_T)\sqrt{\mathcal{L}_n(\mathbf{W}(0))} + \eta\lambda T G_0(M_0 + R_T),$$

which completes the proof. □

Define

$$E_T^{\mathrm{nl}} := \varepsilon_{\mathrm{lin}}(R_T)R_T, \quad E_T^{\mathrm{drift}} := 4\eta T G_0 \varepsilon_{\mathrm{grad}}(R_T)\sqrt{\mathcal{L}_n(\mathbf{W}(0))} \quad E_T^{\mathrm{wd}} := \eta\lambda T G_0(M_0 + R_T).$$

We further write

$$E_T := E_T^{\mathrm{nl}} + E_T^{\mathrm{drift}} + E_T^{\mathrm{wd}}.$$

Theorem 5 (Restatement of Theorem 1)**.** *Fix any $x \in \mathcal{X}$. Under the conditions of Lemma 2,*

$$\phi_{\mathbf{W}(T)}(x) = \phi_{\mathbf{W}(0)}(x) + u_T(x) + \rho_T(x),$$

where $\|\rho_T(x)\|_2 \le E_T$.

Proof. Note $u_T(x) = \sum_{i=1}^n K_0(x, x_i)\beta_i(T)$. By Lemma 4, we have

$$\phi_{\mathbf{W}(T)}(x) = \phi_{\mathbf{W}(0)}(x) + u_T(x) + \rho_T(x),$$

where

$$\|\rho_T(x)\|_2 \le \varepsilon_{\mathrm{lin}}(R_T)R_T + 4\eta T G_0 \varepsilon_{\mathrm{grad}}(R_T)\sqrt{\mathcal{L}_n(\mathbf{W}(0))} + \eta\lambda T G_0(M_0 + R_T) = E_T.$$

The proof is completed. □

C Formal Guarantees for Section 3.2

The convergence argument in this section is architecture-independent, requiring only local Jacobian stability and a positive empirical gradient Gram matrix at initialization. The deep ReLU structure is invoked later only to verify these conditions (see Proposition 1 in Appendix E). The proof follows the standard NTK-style argument of preserving the positive definiteness of the empirical gradient Gram matrix throughout training (Du et al., 2019; Arora et al., 2019).

We now formalize the convergence result in Section 3.2. Recall that $r_i(t) = \phi_{\mathbf{W}(t)}(x_i) - \mathbf{c}$, $\mathbf{K}_0 = [K_0(x_i, x_j)]_{i,j=1}^n$ and

$$\mathcal{L}_n(\mathbf{W}) = \frac{1}{2n}\sum_{i=1}^n \|\phi_{\mathbf{W}}(x_i) - \mathbf{c}\|_2^2.$$

Let

$$\mathbf{r}(t) := (r_1(t)^\top, \dots, r_n(t)^\top)^\top \in \mathbb{R}^{ns}$$

and define

$$\mathbf{K}_t := \left[J_{\mathbf{W}(t)}(x_i)J_{\mathbf{W}(t)}(x_j)^\top\right]_{i,j=1}^n \in \mathbb{R}^{ns\times ns}.$$

Then $\mathbf{K}_t = \mathbf{K}_0$ at $t = 0$ and

$$\mathcal{L}_n(\mathbf{W}(t)) = \frac{1}{2n}\|\mathbf{r}(t)\|_2^2.$$

Further, let

$$\mathbf{J}_t := (J_{\mathbf{W}(t)}(x_1)^\top, \dots, J_{\mathbf{W}(t)}(x_n)^\top)^\top \in \mathbb{R}^{ns\times P},$$

then $\mathbf{K}_t = \mathbf{J}_t\mathbf{J}_t^\top$.

We first show that the empirical gradient Gram matrix remains stable along the training trajectory, so that the positive definiteness at initialization is preserved during optimization.

Lemma 5. *Under the conditions of Lemma 2, for every $t \in \{0, \dots, T\}$, it holds*

$$\|\mathbf{K}_t - \mathbf{K}_0\|_{\mathrm{op}} \le n\varepsilon_{\mathrm{grad}}(R_T)\big(2G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big).$$

Furthermore, if $\lambda_{\min}(\mathbf{K}_0) \ge \lambda_0 > 0$ and $n\varepsilon_{\mathrm{grad}}(R_T)\big(2G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big) \le \frac{\lambda_0}{2}$, then

$$\lambda_{\min}(\mathbf{K}_t) \ge \frac{\lambda_0}{2}, \qquad \textit{for every} \quad t \in \{0, \dots, T\}.$$

Proof. By Lemma 2, it holds $\mathbf{W}(t) \in \mathcal{B}_{R_T}(\mathbf{W}(0))$ for every $t \le T$. Hence, Assumption 1 gives

$$\|\mathbf{J}_t - \mathbf{J}_0\|_{\mathrm{op}} \le \sqrt{n}\,\varepsilon_{\mathrm{grad}}(R_T), \qquad \|\mathbf{J}_0\|_{\mathrm{op}} \le \sqrt{n}\,G_0,$$

and therefore $\|\mathbf{J}_t\|_{\mathrm{op}} \le \sqrt{n}\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)$. It then follows that

$$\begin{aligned}\|\mathbf{K}_t - \mathbf{K}_0\|_{\mathrm{op}} &\le \|\mathbf{J}_t - \mathbf{J}_0\|_{\mathrm{op}}\|\mathbf{J}_t\|_{\mathrm{op}} + \|\mathbf{J}_0\|_{\mathrm{op}}\|\mathbf{J}_t - \mathbf{J}_0\|_{\mathrm{op}} \\ &\le n\varepsilon_{\mathrm{grad}}(R_T)\big(2G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big).\end{aligned}$$

Since $\mathbf{K}_t$ and $\mathbf{K}_0$ are symmetric, we have

$$\lambda_{\min}(\mathbf{K}_t) \ge \lambda_{\min}(\mathbf{K}_0) - \|\mathbf{K}_t - \mathbf{K}_0\|_{\mathrm{op}}.$$

Therefore, if $\lambda_{\min}(\mathbf{K}_0) \ge \lambda_0$ and

$$n\varepsilon_{\mathrm{grad}}(R_T)\big(2G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big) \le \frac{\lambda_0}{2},$$

then $\lambda_{\min}(\mathbf{K}_t) \ge \lambda_0/2$ for every $t \in \{0, \dots, T\}$. This completes the proof. □

We are now ready to establish the geometric convergence of the normal training representations.

Theorem 6 (Restatement of Theorem 2)**.** *Suppose that the conditions of Lemma 2 hold and* $\lambda_{\min}(\mathbf{K}_0) \ge \lambda_0 > 0$*. Assume further that*

$$n\varepsilon_{\mathrm{grad}}(R_T)\big(2G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big) \le \frac{\lambda_0}{2}, \qquad \varepsilon_{\mathrm{lin}}(R_T)\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big) \le \frac{\lambda_0}{4n}.$$

Then, for every $t = 0, \dots, T$*,*

$$\mathcal{L}_n(\mathbf{W}(t)) \le 2\Big(1 - \frac{\eta\lambda_0}{4n}\Big)^{2t}\mathcal{L}_n(\mathbf{W}(0)) + \frac{16\lambda^2 n^2}{\lambda_0^2}\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T) + \varepsilon_{\mathrm{lin}}(R_T)\big)^2(M_0 + R_T)^2$$

with $\mathcal{L}_n(\mathbf{W}) = \frac{1}{2n}\sum_{i=1}^n \|\phi_{\mathbf{W}}(x_i) - \mathbf{c}\|_2^2$ *for any* $\mathbf{W}$*.*

Proof. By Lemma 5, the first additional condition implies that $\lambda_{\min}(\mathbf{K}_t) \ge \lambda_0/2$ for every $t \le T$. Moreover, since $\mathbf{K}_t = \mathbf{J}_t\mathbf{J}_t^\top$ and $\|\mathbf{J}_t\|_{\mathrm{op}} \le \sqrt{n}\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)$,

$$\lambda_{\max}(\mathbf{K}_t) = \|\mathbf{K}_t\|_{\mathrm{op}} \le n\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)^2.$$

Hence, by the step-size condition in Lemma 2,

$$\frac{\eta}{n}\lambda_{\max}(\mathbf{K}_t) \le \eta\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)^2 \le \frac{1}{8}.$$

Since $\mathbf{K}_t$ is symmetric positive semidefinite, all eigenvalues of $I - \frac{\eta}{n}\mathbf{K}_t$ are nonnegative. Therefore,

$$\Big\|I - \frac{\eta}{n}\mathbf{K}_t\Big\|_{\mathrm{op}} = 1 - \frac{\eta}{n}\lambda_{\min}(\mathbf{K}_t) \le 1 - \frac{\eta\lambda_0}{2n}. \tag{15}$$

Let $\Delta_t := \mathbf{W}(t+1) - \mathbf{W}(t)$. Since $\mathbf{W}(t), \mathbf{W}(t+1) \in \mathcal{B}_{R_T}(\mathbf{W}(0))$, Assumption 1 gives, for every $i \in [n]$,

$$r_i(t+1) = r_i(t) + J_{\mathbf{W}(t)}(x_i)\Delta_t + e_{i,t},$$

where $\|e_{i,t}\|_2 \le \varepsilon_{\mathrm{lin}}(R_T)\|\Delta_t\|_{\mathrm{op},\infty}$. Stacking these identities and using the GD update yields

$$\mathbf{r}(t+1) = \Big(I - \frac{\eta}{n}\mathbf{K}_t\Big)\mathbf{r}(t) - \eta\lambda\mathbf{J}_t\mathbf{W}(t) + \mathbf{e}_t, \tag{16}$$

where $\mathbf{e}_t := (e_{1,t}^\top, \dots, e_{n,t}^\top)^\top$ and $\|\mathbf{e}_t\|_2 \le \sqrt{n}\,\varepsilon_{\mathrm{lin}}(R_T)\|\Delta_t\|_2$.

By the GD update rule, equation 7, and Lemma 2, we know

$$\begin{aligned}\|\Delta_t\|_2 &= \eta\big\|\nabla\mathcal{L}_n(\mathbf{W}(t)) + \lambda\mathbf{W}(t)\big\|_2 \le \eta\|\nabla\mathcal{L}_n(\mathbf{W}(t))\|_2 + \eta\lambda\|\mathbf{W}(t)\|_2 \\ &\le \eta\|\nabla\mathcal{L}_n(\mathbf{W}(t))\|_2 + \eta\lambda\|\mathbf{W}(t) - \mathbf{W}(0)\|_2 + \eta\lambda\|\mathbf{W}(0)\|_2 \\ &\le \eta\sqrt{2}\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)\sqrt{\mathcal{L}_n(\mathbf{W}(t))} + \eta\lambda(M_0 + R_T).\end{aligned}$$

Note $\|\mathbf{r}(t)\|_2 = \sqrt{2n\mathcal{L}_n(\mathbf{W}(t))}$, we obtain

$$\|\mathbf{e}_t\|_2 \le \eta\varepsilon_{\mathrm{lin}}(R_T)\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)\|\mathbf{r}(t)\|_2 + \eta\lambda\sqrt{n}\,\varepsilon_{\mathrm{lin}}(R_T)(M_0 + R_T).$$

Combining the above estimate with equation 15, equation 16 and $\|\mathbf{W}(t)\|_2 \le M_0 + R_T$ yields

$$\begin{aligned}
\|\mathbf{r}(t+1)\|_2 &\le \Big\|I - \frac{\eta}{n}\mathbf{K}_t\Big\|_{\mathrm{op}} \|\mathbf{r}(t)\|_2 + \eta\lambda\|\mathbf{J}_t\|_{\mathrm{op}}\|\mathbf{W}(t)\|_2 + \|\mathbf{e}_t\|_2 \\
&\le \Big(1 - \frac{\eta\lambda_0}{2n}\Big)\|\mathbf{r}(t)\|_2 + \eta\lambda\sqrt{n}\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)(M_0 + R_T) \\
&\quad + \eta\varepsilon_{\mathrm{lin}}(R_T)\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)\|\mathbf{r}(t)\|_2 + \eta\lambda\sqrt{n}\,\varepsilon_{\mathrm{lin}}(R_T)(M_0 + R_T) \\
&= \Big(1 - \frac{\eta\lambda_0}{2n} + \eta\varepsilon_{\mathrm{lin}}(R_T)\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big)\Big)\|\mathbf{r}(t)\|_2 \\
&\quad + \eta\lambda\sqrt{n}\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T) + \varepsilon_{\mathrm{lin}}(R_T)\big)(M_0 + R_T).
\end{aligned}$$

Assume $\varepsilon_{\mathrm{lin}}(R_T)\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T)\big) \le \frac{\lambda_0}{4n}$, then

$$\|\mathbf{r}(t+1)\|_2 \le \Big(1 - \frac{\eta\lambda_0}{4n}\Big)\|\mathbf{r}(t)\|_2 + \eta\lambda\sqrt{n}\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T) + \varepsilon_{\mathrm{lin}}(R_T)\big)(M_0 + R_T).$$

Iterating this inequality gives

$$\begin{aligned}
\|\mathbf{r}(t)\|_2 &\le \Big(1 - \frac{\eta\lambda_0}{4n}\Big)^t \|\mathbf{r}(0)\|_2 + \eta\lambda\sqrt{n}\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T) + \varepsilon_{\mathrm{lin}}(R_T)\big)(M_0 + R_T)\sum_{k=0}^{t-1}\Big(1 - \frac{\eta\lambda_0}{4n}\Big)^k \\
&\le \Big(1 - \frac{\eta\lambda_0}{4n}\Big)^t \|\mathbf{r}(0)\|_2 + \frac{4\lambda n^{3/2}}{\lambda_0}\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T) + \varepsilon_{\mathrm{lin}}(R_T)\big)(M_0 + R_T),
\end{aligned}$$

where we have used $\sum_{k=0}^{t-1}\big(1 - \frac{\eta\lambda_0}{4n}\big)^k \le \frac{4n}{\eta\lambda_0}$.

Finally, since $\mathcal{L}_n(\mathbf{W}(t)) = \|\mathbf{r}(t)\|_2^2/2n$, the inequality $(a+b)^2 \le 2a^2 + 2b^2$ gives

$$\mathcal{L}_n(\mathbf{W}(t)) \le 2\Big(1 - \frac{\eta\lambda_0}{4n}\Big)^{2t}\mathcal{L}_n(\mathbf{W}(0)) + \frac{16\lambda^2 n^2}{\lambda_0^2}\big(G_0 + \varepsilon_{\mathrm{grad}}(R_T) + \varepsilon_{\mathrm{lin}}(R_T)\big)^2(M_0 + R_T)^2,$$

which completes the proof. □

D Formal Guarantees for Section 3.3

We next exploit the structure of the fixed output layer to characterize the initialization gradient coupling. Let

$$H_0(x) := \frac{\partial h^L_{\mathbf{W}}(x)}{\partial \mathbf{W}}\Big|_{\mathbf{W}=\mathbf{W}(0)} \in \mathbb{R}^{m\times P}$$

denote the Jacobian of the last hidden representation with respect to the trainable parameters, and define

$$B_0(x, x') := H_0(x)H_0(x')^\top.$$

To isolate the role of the fixed output layer, we first factorize the initialization coupling and identify the scalar quantity around which it concentrates.

Lemma 6. *For any $x, x' \in \mathcal{X}$, it holds that*

$$J_0(x) = AH_0(x), \qquad K_0(x, x') = AB_0(x, x')A^\top.$$

Moreover, conditional on $\mathbf{W}(0)$,

$$\mathbb{E}_A[K_0(x, x')] = k_0(x, x')I_s \qquad \textit{and} \qquad k_0(x, x') := \mathrm{tr}(B_0(x, x')) = \langle H_0(x), H_0(x')\rangle_{\mathrm{F}},$$

where I_s denotes the $s \times s$ identity matrix.

Proof. Since $\phi_{\mathbf{W}}(x) = Ah^L_{\mathbf{W}}(x)$ and A is fixed, we have $J_0(x) = AH_0(x)$, and hence

$$K_0(x, x') = AH_0(x)H_0(x')^\top A^\top = AB_0(x, x')A^\top.$$

Let $a_j^\top$ denote the j-th row of A. Then

$$[K_0(x, x')]_{jk} = a_j^\top B_0(x, x')a_k.$$

For $j \neq k$, independence and zero mean imply $\mathbb{E}_A[a_j^\top B_0(x, x')a_k] = 0$, whereas

$$\mathbb{E}_A[a_j^\top B_0(x, x')a_j] = \operatorname{tr}\big(B_0(x, x')\,\mathbb{E}_A[a_j a_j^\top]\big) = \operatorname{tr}(B_0(x, x')),$$

since $\mathbb{E}_A[a_j a_j^\top] = I_m$. This proves the result. □

To establish concentration of the random output-layer coupling around its mean, we use the following Hanson–Wright inequality.

Lemma 7 (Hanson–Wright inequality)**.** *Let $\xi = (\xi_1, \dots, \xi_N)^\top$ have independent Rademacher entries and let $M \in \mathbb{R}^{N\times N}$ be symmetric. Then, for every $t > 0$,*

$$\mathbb{P}\left(\left|\xi^\top M\xi - \mathbb{E}[\xi^\top M\xi]\right| \geq C\big(\|M\|_{\mathrm{F}}\sqrt{t} + \|M\|_{\mathrm{op}}t\big)\right) \leq 2e^{-ct},$$

where $c, C > 0$ are universal constants.

The preceding factorization identifies $k_0(z, z')I_s$ as the mean gradient coupling induced by the random output layer. We next show that the finite-width kernel concentrates around this scalar form. Define $\Gamma_{\mathrm{F}} := \max_{z,z'\in\mathcal{Z}_x} \|B_0(z, z')\|_{\mathrm{F}}$ and $\Gamma_{\mathrm{op}} := \max_{z,z'\in\mathcal{Z}_x} \|B_0(z, z')\|_{\mathrm{op}}$.

Lemma 8. *Conditional on $\mathbf{W}(0)$, with probability at least $1 - \delta$ over A,*

$$\max_{z,z'\in\mathcal{Z}_x} \|K_0(z, z') - k_0(z, z')I_s\|_{\mathrm{op}} \lesssim \Gamma_{\mathrm{F}}\sqrt{s + \log\big(\tfrac{|\mathcal{Z}_x|}{\delta}\big)} + \Gamma_{\mathrm{op}}\Big(s + \log\big(\tfrac{|\mathcal{Z}_x|}{\delta}\big)\Big) =: \Delta_0.$$

Proof. Fix $z, z' \in \mathcal{Z}_x$ and write $B = B_0(z, z')$. Let $\xi = \mathrm{vec}(A) \in \mathbb{R}^{sm}$, where the rows of A are stacked into a single vector. For any $u, v \in \mathbb{S}^{s-1}$, we have

$$u^\top ABA^\top v = \xi^\top\big((uv^\top) \otimes B\big)\xi.$$

Since a quadratic form depends only on the symmetric part of its coefficient matrix, define $M_{u,v} := \frac{1}{2}\big((uv^\top) \otimes B + (vu^\top) \otimes B^\top\big)$. Then $u^\top ABA^\top v = \xi^\top M_{u,v}\xi$ and

$$\mathbb{E}_A[\xi^\top M_{u,v}\xi] = \operatorname{tr}(M_{u,v}) = \operatorname{tr}(B)u^\top v = k_0(z, z')u^\top v.$$

Moreover, using $\|uv^\top\|_{\mathrm{F}} = \|uv^\top\|_{\mathrm{op}} = 1$ and the corresponding product rules for Kronecker products,

$$\|M_{u,v}\|_{\mathrm{F}} \leq \|B\|_{\mathrm{F}}, \qquad \|M_{u,v}\|_{\mathrm{op}} \leq \|B\|_{\mathrm{op}}.$$

Since the entries of ξ are independent Rademacher random variables, Lemma 7 yields, for every $t > 0$,

$$\mathbb{P}_A\left(\left|u^\top\big(K_0(z, z') - k_0(z, z')I_s\big)v\right| \gtrsim \|B\|_{\mathrm{F}}\sqrt{t} + \|B\|_{\mathrm{op}}t\right) \leq 2e^{-ct}.$$

Let $\mathcal{N}$ be a $1/4$-net of $\mathbb{S}^{s-1}$ with $|\mathcal{N}| \leq 9^s$. For any matrix $C \in \mathbb{R}^{s\times s}$, the standard net argument gives

$$\|C\|_{\mathrm{op}} \leq 2 \max_{u,v\in\mathcal{N}} |u^\top Cv|.$$

Applying the preceding concentration inequality to all $u, v \in \mathcal{N}$ and all $(z, z') \in \mathcal{Z}_x^2$, and then taking a union bound over at most $|\mathcal{N}|^2|\mathcal{Z}_x|^2 \leq 9^{2s}|\mathcal{Z}_x|^2$ events, we find that choosing $t \asymp s + \log(\frac{|\mathcal{Z}_x|}{\delta})$ ensures that, with probability at least $1 - \delta$,

$$\max_{z,z'\in\mathcal{Z}_x} \|K_0(z, z') - k_0(z, z')I_s\|_{\mathrm{op}} \lesssim \Gamma_{\mathrm{F}}\sqrt{s + \log\big(\tfrac{|\mathcal{Z}_x|}{\delta}\big)} + \Gamma_{\mathrm{op}}\Big(s + \log\big(\tfrac{|\mathcal{Z}_x|}{\delta}\big)\Big).$$

This completes the proof. □

Lemma 8 shows that each matrix-valued coupling $K_0(x, x_i)$ is close to $k_0(x, x_i)I_s$. Since the first-order effect $u_T(x)$ is obtained by summing these couplings against the cumulative residual terms $\beta_i(T)$, this immediately converts the matrix-valued description of training into a scalar-weighted one. The following corollary makes this approximation explicit.

Corollary 1. *Fix any $x \in \mathcal{X}$. Conditional on $\mathbf{W}(0)$, with probability at least $1 - \delta$ over A,*

$$\Big\| u_T(x) - \sum_{i=1}^{n} k_0(x, x_i)\beta_i(T) \Big\|_2 \lesssim \Delta_0 \sum_{i=1}^{n} \|\beta_i(T)\|_2.$$

Moreover, if the conditions of Lemma 2 hold, then

$$\Big\| u_T(x) - \sum_{i=1}^{n} k_0(x, x_i)\beta_i(T) \Big\|_2 \lesssim 4\eta T \Delta_0 \sqrt{\mathcal{L}_n(\mathbf{W}(0))}.$$

Proof. By the definition of $u_T(x)$,

$$u_T(x) - \sum_{i=1}^{n} k_0(x, x_i)\beta_i(T) = \sum_{i=1}^{n} \big(K_0(x, x_i) - k_0(x, x_i)I_s\big)\beta_i(T).$$

Therefore, on the event of Lemma 8,

$$\Big\| u_T(x) - \sum_{i=1}^{n} k_0(x, x_i)\beta_i(T) \Big\|_2 \le \sum_{i=1}^{n} \|K_0(x, x_i) - k_0(x, x_i)I_s\|_{\mathrm{op}} \|\beta_i(T)\|_2 \lesssim \Delta_0 \sum_{i=1}^{n} \|\beta_i(T)\|_2.$$

Moreover, by the definition of $\beta_i(T)$,

$$\sum_{i=1}^{n} \|\beta_i(T)\|_2 \le \frac{\eta}{n} \sum_{t=0}^{T-1} \sum_{i=1}^{n} \|r_i(t)\|_2 \le \eta\sqrt{2} \sum_{t=0}^{T-1} \sqrt{\mathcal{L}_n(\mathbf{W}(t))}.$$

Using Cauchy–Schwarz inequality together with $\sum_{t=0}^{T-1} \mathcal{L}_n(\mathbf{W}(t)) \le 8T\mathcal{L}_n(\mathbf{W}(0))$ from Lemma 2, we obtain

$$\sum_{i=1}^{n} \|\beta_i(T)\|_2 \le 4\eta T \sqrt{\mathcal{L}_n(\mathbf{W}(0))}.$$

This completes the proof. □

The above corollary shows that, up to the concentration error induced by the random output layer, the first-order effect of training on an unseen input is determined by the scalar similarities $k_0(x, x_i)$ and the cumulative residual terms $\beta_i(T)$.

We now prove the detection guarantee stated in Section 3.3.

Theorem 7 (Restatement of Theorem 3)**.** *Under the conditions of Theorem 5, let $\tau > 0$ be the decision threshold. If*

$$\|\phi_{\mathbf{W}(0)}(x) - \mathbf{c}\|_2 - \|u_T(x)\|_2 - E_T > \sqrt{\tau},$$

then

$$s_T(x) = \|\phi_{\mathbf{W}(T)}(x) - \mathbf{c}\|_2^2 > \tau,$$

and hence x is detected as anomalous.

Proof. By Theorem 5, we know

$$\phi_{\mathbf{W}(T)}(x) = \phi_{\mathbf{W}(0)}(x) + u_T(x) + \rho_T(x) \quad \text{with} \quad \|\rho_T(x)\|_2 \le E_T.$$

Therefore, the triangle inequality gives

$$\|\phi_{\mathbf{W}(T)}(x) - \mathbf{c}\|_2 \ge \|\phi_{\mathbf{W}(0)}(x) - \mathbf{c}\|_2 - \|u_T(x)\|_2 - \|\rho_T(x)\|_2$$

and hence

$$\|\phi_{\mathbf{W}(T)}(x) - \mathbf{c}\|_2 \ge \|\phi_{\mathbf{W}(0)}(x) - \mathbf{c}\|_2 - \|u_T(x)\|_2 - E_T.$$

Thus, under the stated condition,

$$\|\phi_{\mathbf{W}(T)}(x) - \mathbf{c}\|_2 > \sqrt{\tau},$$

which implies $s_T(x) > \tau$. This completes the proof. □

Finally, note Corollary 1 gives

$$\|u_T(x)\|_2 \le \sum_{i=1}^{n} |k_0(x,x_i)| \|\beta_i(T)\|_2 + \Delta_0 \sum_{i=1}^{n} \|\beta_i(T)\|_2.$$

Therefore, a sufficient condition for detection is

$$\|\phi_{\mathbf{W}(0)}(x) - \mathbf{c}\|_2 - \sum_{i=1}^{n} |k_0(x,x_i)| \|\beta_i(T)\|_2 - \Delta_0 \sum_{i=1}^{n} \|\beta_i(T)\|_2 - E_T > \sqrt{\tau}.$$

E Formal Guarantees for Section 3.4

We specialize the preceding analysis to the deep ReLU representation introduced in Section 2. The main difference from the scalar network analyzed by Zhou et al. (2026a) is that $\phi_{\mathbf{W}}$ has s output coordinates. Following the common normalization used in theoretical analyses of deep ReLU networks, and in particular in Zhou et al. (2026a), we assume throughout this section that $\mathcal{X} \subseteq \mathbb{S}^{d-1}$. We first record a simple lifting argument that converts scalar local estimates into the vector-valued quantities appearing in Assumption 1.

Lemma 9. *Suppose that, for every $j \in [s]$, $z \in \mathcal{Z}_x$, and $\mathbf{W}, \widetilde{\mathbf{W}} \in \mathcal{B}_R(\mathbf{W}(0))$,*

$$\|\nabla_{\mathbf{W}} \phi_{\mathbf{W}(0),j}(z)\|_2 \le g_0, \qquad \|\nabla_{\mathbf{W}} \phi_{\mathbf{W},j}(z) - \nabla_{\mathbf{W}} \phi_{\mathbf{W}(0),j}(z)\|_2 \le \varepsilon^{\mathrm{sc}}_{\mathrm{grad}}(R),$$

and

$$|\phi_{\widetilde{\mathbf{W}},j}(z) - \phi_{\mathbf{W},j}(z) - \langle \nabla_{\mathbf{W}} \phi_{\mathbf{W},j}(z), \widetilde{\mathbf{W}} - \mathbf{W} \rangle| \le \varepsilon^{\mathrm{sc}}_{\mathrm{lin}}(R) \|\widetilde{\mathbf{W}} - \mathbf{W}\|_{\mathrm{op},\infty}.$$

Then the Jacobian and local-linearity bounds in Assumption 1 hold with

$$G_0 = \sqrt{s}\, g_0, \qquad \varepsilon_{\mathrm{grad}}(R) = \sqrt{s}\, \varepsilon^{\mathrm{sc}}_{\mathrm{grad}}(R), \qquad \varepsilon_{\mathrm{lin}}(R) = \sqrt{s}\, \varepsilon^{\mathrm{sc}}_{\mathrm{lin}}(R).$$

Proof. Since the rows of $J_{\mathbf{W}(0)}(z)$ are $\nabla_{\mathbf{W}} \phi_{\mathbf{W}(0),j}(z)^\top$, we have

$$\|J_{\mathbf{W}(0)}(z)\|_{\mathrm{op}} \le \|J_{\mathbf{W}(0)}(z)\|_{\mathrm{F}} = \Big(\sum_{j=1}^{s} \|\nabla_{\mathbf{W}} \phi_{\mathbf{W}(0),j}(z)\|_2^2 \Big)^{1/2} \le \sqrt{s}\, g_0.$$

Similarly, since the rows of $J_{\mathbf{W}}(z) - J_{\mathbf{W}(0)}(z)$ are $\big(\nabla_{\mathbf{W}} \phi_{\mathbf{W},j}(z) - \nabla_{\mathbf{W}} \phi_{\mathbf{W}(0),j}(z)\big)^\top$,

$$\begin{aligned} \|J_{\mathbf{W}}(z) - J_{\mathbf{W}(0)}(z)\|_{\mathrm{op}} &\le \|J_{\mathbf{W}}(z) - J_{\mathbf{W}(0)}(z)\|_{\mathrm{F}} \\ &\le \Big(\sum_{j=1}^{s} \|\nabla_{\mathbf{W}} \phi_{\mathbf{W},j}(z) - \nabla_{\mathbf{W}} \phi_{\mathbf{W}(0),j}(z)\|_2^2 \Big)^{1/2} \le \sqrt{s}\, \varepsilon^{\mathrm{sc}}_{\mathrm{grad}}(R). \end{aligned}$$

Finally, applying the scalar almost-linearity bound coordinatewise gives

$$\begin{aligned} &\|\phi_{\widetilde{\mathbf{W}}}(z) - \phi_{\mathbf{W}}(z) - J_{\mathbf{W}}(z)(\widetilde{\mathbf{W}} - \mathbf{W})\|_2^2 \\ &= \sum_{j=1}^{s} |\phi_{\widetilde{\mathbf{W}},j}(z) - \phi_{\mathbf{W},j}(z) - \langle \nabla_{\mathbf{W}} \phi_{\mathbf{W},j}(z), \widetilde{\mathbf{W}} - \mathbf{W} \rangle|^2 \\ &\le s \big(\varepsilon^{\mathrm{sc}}_{\mathrm{lin}}(R)\big)^2 \|\widetilde{\mathbf{W}} - \mathbf{W}\|_{\mathrm{op},\infty}^2. \end{aligned}$$

Taking square roots yields

$$\|\phi_{\widetilde{\mathbf{W}}}(z) - \phi_{\mathbf{W}}(z) - J_{\mathbf{W}}(z)(\widetilde{\mathbf{W}} - \mathbf{W})\|_2 \le \sqrt{s}\, \varepsilon^{\mathrm{sc}}_{\mathrm{lin}}(R) \|\widetilde{\mathbf{W}} - \mathbf{W}\|_{\mathrm{op},\infty},$$

which completes the proof. □

Recall that the deep ReLU representation introduced in Section 2 is

$$\begin{aligned} h^1_{\mathbf{W}}(x) &= \sqrt{\frac{2}{m}}\, \sigma(W^1 x), \\ h^\ell_{\mathbf{W}}(x) &= \sqrt{\frac{2}{m}}\, \sigma\big(W^\ell h^{\ell-1}_{\mathbf{W}}(x)\big), \qquad \ell = 2, \dots, L, \\ \phi_{\mathbf{W}}(x) &= A h^L_{\mathbf{W}}(x). \end{aligned} \tag{17}$$

The entries of $A \in \{-1,1\}^{s\times m}$ are drawn independently from Unif$\{-1,1\}$, independently of $\mathbf{W}(0)$ and held fixed during training.

We now combine the preceding lifting argument with the scalar deep ReLU estimates of Zhou et al. (2026a) to verify Assumption 1 under Gaussian initialization. The width requirement below is a sufficient condition inherited from these local estimates and is not expected to be tight.

Proposition 1 (Verification of Assumption 1)**.** *Let $\delta \in (0,1)$ and $R \geq 1$. Assume $m \gtrsim \mathrm{poly}(L, d, R, \log(ms/\delta))$. Then, with probability at least $1 - sL\exp\big(\mathcal{O}(dL\log m) - \Omega(m^{1/3})\big) - \delta$ over $(A, \mathbf{W}(0))$, the deep ReLU representation equation 17 satisfies Assumption 1 on $\mathcal{B}_R(\mathbf{W}(0))$ with*

$$M_0 \lesssim \sqrt{L}\, m, \qquad G_0 \lesssim \sqrt{s}\, L^{3/2},$$

and

$$\varepsilon_{\mathrm{grad}}(R) \lesssim \sqrt{s}\, L^{11/6} R^{1/3} m^{-1/6}, \qquad \varepsilon_{\mathrm{lin}}(R) \lesssim \sqrt{s}\, L^{7/3} R^{1/3} m^{-1/6}.$$

Proof. For each $j \in [s]$, the coordinate $\phi_{\mathbf{W},j}(x) = A_j^\top h^L_{\mathbf{W}}(x)$ has the form of the scalar network analyzed by Zhou et al. (2026a), with A_j playing the role of the fixed readout vector.

By part (c) of Lemma 31 in Zhou et al. (2026a), for every $\ell \in [L]$,

$$\sup_{z\in\mathcal{Z}_x} \big\|\nabla_{W^\ell}\phi_{\mathbf{W}(0),j}(z)\big\|_{\mathrm{F}} \lesssim L.$$

The event underlying this bound depends only on $\mathbf{W}(0)$. Indeed, the proof of Lemma 31(c) uses $\|V^\ell_{L,0}(z)A_j\|_2 \leq \|V^\ell_{L,0}(z)\|_{\mathrm{op}}\|A_j\|_2$, while every Rademacher readout satisfies $\|A_j\|_2 = \sqrt{m}$. Hence, the same event gives the bound simultaneously for all $j \in [s]$, without an additional union bound. Since the parameter blocks $W^1, \ldots, W^L$ are disjoint,

$$\|\nabla_{\mathbf{W}}\phi_{\mathbf{W}(0),j}(z)\|_2^2 = \sum_{\ell=1}^{L} \big\|\nabla_{W^\ell}\phi_{\mathbf{W}(0),j}(z)\big\|_{\mathrm{F}}^2,$$

and therefore $g_0 \lesssim L^{3/2}$.

Next, Inequality (27) of Lemma 22 in Zhou et al. (2026a) gives, uniformly for $\mathbf{W} \in \mathcal{B}_R(\mathbf{W}(0))$ and every $\ell \in [L]$,

$$\big\|\nabla_{W^\ell}\phi_{\mathbf{W},j}(z) - \nabla_{W^\ell}\phi_{\mathbf{W}(0),j}(z)\big\|_{\mathrm{F}} \lesssim L^{4/3} R^{1/3} m^{-1/6}.$$

Summing the squared bounds over the L parameter blocks yields

$$\varepsilon^{\mathrm{sc}}_{\mathrm{grad}}(R) \lesssim L^{11/6} R^{1/3} m^{-1/6}.$$

Moreover, Inequality (25) of the same lemma directly gives

$$\varepsilon^{\mathrm{sc}}_{\mathrm{lin}}(R) \lesssim L^{7/3} R^{1/3} m^{-1/6}.$$

Applying Lemma 9 therefore gives

$$G_0 \lesssim \sqrt{s}\, L^{3/2}, \qquad \varepsilon_{\mathrm{grad}}(R) \lesssim \sqrt{s}\, L^{11/6} R^{1/3} m^{-1/6}, \qquad \varepsilon_{\mathrm{lin}}(R) \lesssim \sqrt{s}\, L^{7/3} R^{1/3} m^{-1/6}.$$

To make the estimates from Lemma 22 hold simultaneously for all $j \in [s]$, we apply that lemma to each output coordinate with failure probability of order δ/s and take a union bound. This replaces $\log(m/\delta)$ in its width condition by $\log(ms/\delta)$ and contributes the factor s to the exponentially small failure probability.

It remains to control the initialization norm. Since $\|\mathbf{W}(0)\|_2^2 = \sum_{\ell=1}^{L} \|W^\ell(0)\|_{\mathrm{F}}^2$ is a chi-square random variable with $md + (L-1)m^2 \asymp Lm^2$ degrees of freedom, standard chi-square concentration gives

$$\|\mathbf{W}(0)\|_2 \lesssim \sqrt{md + (L-1)m^2} \lesssim \sqrt{L}\, m$$

except with probability $\exp(-\Omega(Lm^2))$. This failure probability is negligible compared with $sL\exp\big(\mathcal{O}(dL\log m) - \Omega(m^{1/3})\big)$ and can be absorbed into the latter term. Combining the preceding events, and adjusting constant fractions of δ in the individual bounds, gives the stated probability. This completes the proof. □

Proposition 2 (Trajectory error under Gaussian initialization)**.** *Let $\mathcal{L}_0 := \mathcal{L}_n(\mathbf{W}(0))$ and $R_T := \sqrt{8\eta T\mathcal{L}_0}$. Suppose that $R_T \geq 1$, the conditions of Theorem 5 hold, and the width condition in Proposition 1 holds with $R = R_T$. Then, with the same high probability as in Proposition 1,*

$$E_T \lesssim \sqrt{s}L^{7/3}(\eta T)^{2/3}\mathcal{L}_0^{2/3}m^{-1/6} + sL^{10/3}(\eta T)^{7/6}\mathcal{L}_0^{2/3}m^{-1/6} + \eta\lambda T\sqrt{s}L^{3/2}\left(\sqrt{L}\,m + \sqrt{\eta T\mathcal{L}_0}\right).$$

In particular, when $\lambda = 0$ and the remaining parameters are fixed, $E_T = \mathcal{O}(m^{-1/6})$ as $m \to \infty$.

Proof. By Proposition 1, $G_0 \lesssim \sqrt{s}L^{3/2}$, $M_0 \lesssim \sqrt{L}\,m$, $\varepsilon_{\mathrm{grad}}(R_T) \lesssim \sqrt{s}L^{11/6}R_T^{1/3}m^{-1/6}$, and $\varepsilon_{\mathrm{lin}}(R_T) \lesssim \sqrt{s}L^{7/3}R_T^{1/3}m^{-1/6}$. Therefore,

$$E_T^{\mathrm{nl}} = \varepsilon_{\mathrm{lin}}(R_T)R_T \lesssim \sqrt{s}L^{7/3}R_T^{4/3}m^{-1/6},$$

$$E_T^{\mathrm{drift}} = 4\eta T G_0 \varepsilon_{\mathrm{grad}}(R_T)\sqrt{\mathcal{L}_0} \lesssim \eta T\, sL^{10/3}R_T^{1/3}m^{-1/6}\sqrt{\mathcal{L}_0},$$

and

$$E_T^{\mathrm{wd}} = \eta\lambda T G_0(M_0 + R_T) \lesssim \eta\lambda T\sqrt{s}L^{3/2}\big(\sqrt{L}\,m + R_T\big).$$

Substituting $R_T = \sqrt{8\eta T\mathcal{L}_0}$ into these three bounds and using $E_T = E_T^{\mathrm{nl}} + E_T^{\mathrm{drift}} + E_T^{\mathrm{wd}}$ completes the proof. □

For the contraction result in Theorem 6, the additional Gram-stability and local-linearity conditions are satisfied under Gaussian initialization when $m^{1/6} \gtrsim \frac{nsL^{23/6}R_T^{1/3}}{\lambda_0}$, which is the stronger of the two corresponding width requirements.

Now, we characterize the initialization coupling k_0 under Gaussian initialization. For $\rho \in [-1, 1]$, define

$$C(\rho) = \frac{1}{\pi}\left(\sqrt{1-\rho^2} + (\pi - \arccos\rho)\rho\right) \quad \text{and} \quad \dot{C}(\rho) := C'(\rho) = \frac{\pi - \arccos\rho}{\pi}.$$

Set $c_0(\rho) = \rho$, $c_\ell(\rho) = C(c_{\ell-1}(\rho))$, and $q_\ell(\rho) = \dot{C}(c_{\ell-1}(\rho))$ for $\ell \in [L]$. We further define

$$\kappa_L(\rho) = \sum_{\ell=1}^{L} c_{\ell-1}(\rho)\prod_{h=\ell}^{L} q_h(\rho).$$

Theorem 8 (Restatement of Theorem 4)**.** *Suppose that $\mathcal{X} \subseteq \mathbb{S}^{d-1}$ and that the hidden weights are initialized independently according to $w_r^1(0) \sim \mathcal{N}(0, I_d)$ and $w_r^\ell(0) \sim \mathcal{N}(0, I_m)$ for $\ell = 2, \ldots, L$. Then κ_L is a deterministic kernel depending only on the network depth L and is nondecreasing on $[0, 1]$. Moreover, for any $\delta \in (0, 1)$, if $m \gtrsim dL^3\log(m/\delta)$, then, with probability at least $1 - L\exp\big(\mathcal{O}(dL\log m) - \Omega(m^{1/3})\big) - \delta$, we have*

$$\sup_{x,x'\in\mathcal{X}} |k_0(x, x') - \kappa_L(\langle x, x'\rangle)| \lesssim L^{3/2}m^{-1/6} + L\sqrt{\frac{dL\log m}{m}} + L^4 m^{-1/3}.$$

In particular, for fixed d and L, the right-hand side converges to zero as $m \to \infty$.

Proof. The argument follows the proof of Lemma 32 in Zhou et al. (2026a), with two modifications. First, the backward contribution appears as a trace rather than as a random quadratic form in the readout vector, which removes one of their error terms. Second, for unit-norm inputs, we identify the resulting deterministic kernel as a univariate function of $\langle x, x'\rangle$. We indicate these two points below and refer to Zhou et al. (2026a) for the remaining estimates.

For each $\ell \in [L]$, let $H_0^\ell(x)$ denote the derivative of $h_{\mathbf{W}}^L(x)$ with respect to W^ℓ at initialization, and define $k_0^\ell(x, x') = \langle H_0^\ell(x), H_0^\ell(x')\rangle_{\mathrm{F}}$. Since the parameter blocks $W^1, \ldots, W^L$ are disjoint, we have $k_0(x, x') = \sum_{\ell=1}^{L} k_0^\ell(x, x')$.

Following the notation of Zhou et al. (2026a), let $o_0^{\ell-1}(x)$ denote the representation entering layer ℓ at initialization and let $V_{L,0}^{\ell}(x)$ denote the corresponding backward propagation matrix. By the chain rule,

$$k_0^{\ell}(x,x') = \langle o_0^{\ell-1}(x), o_0^{\ell-1}(x')\rangle \operatorname{tr}\big(V_{L,0}^{\ell}(x)^{\top} V_{L,0}^{\ell}(x')\big).$$

In the proof of Lemma 32 in Zhou et al. (2026a), the corresponding scalar-output kernel contains the random quadratic form $(V_{L,0}^{\ell}(x)a)^{\top} V_{L,0}^{\ell}(x')a$. Their error decomposition consists of a forward-covariance term, a term controlling the deviation of this quadratic form from its trace, and a backward-trace term. For k_0^{ℓ}, the second term is absent because the backward contribution is already given by the trace. Since this term is of order $L^2 m^{-1/3}$ and is dominated by the forward-covariance term of order $L^3 m^{-1/3}$, its absence does not improve the resulting rate.

Moreover, Lemma 31 of Zhou et al. (2026a) gives $\sup_x \|V_{L,0}^{\ell}(x)\|_{\mathrm{op}} \lesssim L/\sqrt{m}$. Therefore,

$$\big|\operatorname{tr}\big(V_{L,0}^{\ell}(x)^{\top} V_{L,0}^{\ell}(x')\big)\big| \le m\|V_{L,0}^{\ell}(x)\|_{\mathrm{op}}\|V_{L,0}^{\ell}(x')\|_{\mathrm{op}} \lesssim L^2.$$

Hence, the estimates for the forward-covariance and backward-trace terms in the proof of Lemma 32 apply directly and yield, simultaneously for all $\ell \in [L]$,

$$\sup_{x,x'\in\mathcal{X}} \big|k_0^{\ell}(x,x') - K^{\ell}(x,x')\big| \lesssim \sqrt{L}\, m^{-1/6} + \sqrt{\frac{dL\log m}{m}} + L^3 m^{-1/3},$$

where K^{ℓ} denotes the corresponding deterministic layerwise kernel.

It remains to identify K^{ℓ}. Write $\rho = \langle x, x'\rangle$. By the Gaussian ReLU derivative recursion, we have $q^h(x,x') = q_h(\rho)$ for every $h \in [L]$. For $\ell = 1$, the forward factor is simply $\langle x, x'\rangle = c_0(\rho)$. For $\ell \ge 2$, the Gaussian ReLU covariance recursion gives

$$2\mathbb{E}\left[\sigma(U^{\ell-1}(x))\sigma(U^{\ell-1}(x'))\right] = c_{\ell-1}(\rho).$$

Therefore,

$$K^{\ell}(x,x') = c_{\ell-1}(\rho)\prod_{h=\ell}^{L} q_h(\rho).$$

Summing over ℓ gives $\sum_{\ell=1}^{L} K^{\ell}(x,x') = \kappa_L(\langle x, x'\rangle)$.

Combining this identity with $k_0(x,x') = \sum_{\ell=1}^{L} k_0^{\ell}(x,x')$ and summing the preceding layerwise bound over ℓ yields

$$\sup_{x,x'\in\mathcal{X}} |k_0(x,x') - \kappa_L(\langle x, x'\rangle)| \lesssim L^{3/2} m^{-1/6} + L\sqrt{\frac{dL\log m}{m}} + L^4 m^{-1/3}.$$

For fixed d and L, the right-hand side converges to zero as $m \to \infty$.

It remains to verify the monotonicity of κ_L. On $[0,1]$, both C and $\dot{C}$ are nonnegative and nondecreasing, and C maps $[0,1]$ into $[0,1]$. Since $c_0(\rho) = \rho$, it follows recursively that c_ℓ and q_ℓ are nonnegative and nondecreasing on $[0,1]$. Hence, for every $\ell \in [L]$, the function

$$c_{\ell-1}(\rho)\prod_{h=\ell}^{L} q_h(\rho)$$

is nondecreasing on $[0,1]$. Summing over ℓ shows that κ_L is nondecreasing on $[0,1]$, which completes the proof. □

F Proofs of Section 4

Finally, we give the proof of Lemma 1.

Proof of Lemma 1. It follows directly from

$$\|q_T^{\mathrm{nor}}(x) + u_T^{\mathrm{ano}}(x)\|_2^2 = \|q_T^{\mathrm{nor}}(x)\|_2^2 + 2\langle q_T^{\mathrm{nor}}(x), u_T^{\mathrm{ano}}(x)\rangle + \|u_T^{\mathrm{ano}}(x)\|_2^2.$$

□

G EXPERIMENTS

G.1 SETUP

Dataset. Following Ruff et al. (2018), we construct ten anomaly-detection setups from MNIST (Deng, 2012). In each setup, one digit is treated as the normal class and the remaining nine digits as anomalies. We use the original training and test splits and train only on normal samples from the training split. For each setup, we randomly select $n = 1000$ training normals from the training split, and 1000 held-out normals and 1000 anomalies from the test split. These subsets are drawn once and shared across all runs. Each image is flattened to $x \in \mathbb{R}^{784}$, rescaled to $[0, 1]$, and normalized so that $\|x\|_2 = 1$, as assumed in Sec. E.

Architecture. We use the bias-free fully connected ReLU network in Sec. 2 with depth $L = 3$, output dimension $s = 16$, and width $m \in \{128, 256, \dots, 8192\}$, unless stated otherwise.

Initialization. The hidden weights $\mathbf{W}(0)$ are Gaussian initialized as in Sec. 3.4, and the fixed output matrix A is drawn as in Sec. 2. Following Ruff et al. (2018), the center $\mathbf{c}$ is initialized as the mean network output over the normal training samples. Coordinates satisfying $|c_k| < 0.1$ are set to ± 0.1, preserving their signs. The center is then fixed throughout training. Each configuration is run with three random seeds, which change $\mathbf{W}(0)$ and A. The data subsets are the same for all seeds.

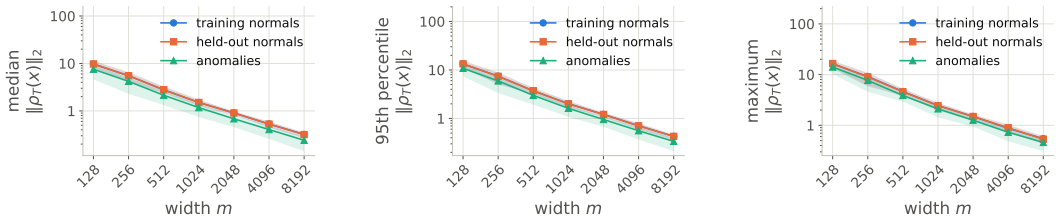
Optimization. We minimize equation 1 by full-batch GD in double precision with step size $\eta = 2 \times 10^{-3}$ for $T = 200$k steps and $\lambda \in \{0, 10^{-4}\}$. Unlike Ruff et al. (2018; 2020), we use no autoencoder pretraining and train with full-batch GD instead of mini-batch Adam, as required by the analysis.

Evaluation. The anomaly score at iteration t is $s_t(x) = \|\phi_{\mathbf{W}(t)}(x) - \mathbf{c}\|_2^2$, and its first-order prediction is $\widehat{s}_t(x) = \|\phi_{\mathbf{W}(0)}(x) + u_t(x) - \mathbf{c}\|_2^2$. The first-order term $u_t(x) = \sum_{i=1}^n K_0(x, x_i)\beta_i(t)$ is computed exactly from one vector–Jacobian and one Jacobian–vector product at $\mathbf{W}(0)$, without forming K_0, at logarithmically spaced steps and every 400 steps. The remainder is $\rho_t(x) = \phi_{\mathbf{W}(t)}(x) - \phi_{\mathbf{W}(0)}(x) - u_t(x)$. For $\lambda > 0$, the weight-decay contribution is part of ρ_t, as in Theorem 1. We report AUROC with anomalies as the positive class. When a detection threshold is required, we set τ to the 95th percentile of the held-out normal scores, corresponding to a 5% false-positive rate. Reported curves are averaged over the three seeds. Scatter plots show seed 0. In the scatter plots, the step shown for each width is the last recorded step before the first-order AUROC of DOCC, averaged over seeds, differs from the actual AUROC by more than 0.01.

Gaussian-coupling experiment. For the experiment corresponding to Theorem 4, we compare the empirical scalar coupling $k_0(x, x')$ with the deterministic kernel $\kappa_L(\langle x, x' \rangle)$. To examine the effect of depth, we fix $m = 512$ and vary $L \in \{2, 4, 8, 12, 20\}$. To examine the effect of width, we fix $L = 3$ and vary $m \in \{128, 256, 512, 1024, 2048\}$. All inputs are normalized to unit Euclidean norm.

Deep SAD experiment. For the Deep SAD comparison in Fig. 5, we use $\tilde{n} = 50$ labeled anomalies sampled from the training split of the remaining MNIST digits and set $\nu = 100$. We chose ν in a pilot over $\nu \in \{1, 10, 100, 1000\}$ (seed 0, $m \in \{128, 1024\}$) as the smallest value with a clear effect at the displayed steps while the normal representations still contract. At $\nu = 1$, the anomaly term is negligible, since the inputs start at a distance of about 2 from $\mathbf{c}$. We use $\lambda = 0$ and train Deep SAD from the same initialization as the corresponding DOCC model. The same held-out samples and training steps are used for the comparison. Each method uses its own detection threshold $\sqrt{\tau}$, defined from the 95th percentile of its held-out normal scores. In Fig. 5, the displayed step for each width is the last step before the first-order DOCC AUROC differs from the actual AUROC by more than 0.01.

Across three seeds, Deep SAD recovers 14%, 44%, 58%, and 76% of the anomalies missed by DOCC at $m = 128$, 512, 2048, and 8192, respectively, while losing at most 4% of the anomalies detected by DOCC. The corresponding AUROC improves from 0.79, 0.84, 0.86, and 0.88 under DOCC to 0.83, 0.92, 0.95, and 0.97 under Deep SAD.



(a) Median. (b) 95th percentile. (c) Maximum.

Figure 7: Scaling of the trajectory remainder with network width ($\lambda = 0$, $T = 200$k). The median, 95th percentile, and maximum of $\|\rho_T(x)\|_2$ are shown for training normals, held-out normals, and anomalies, averaged over the ten anomaly-detection setups. Bands indicate the range across setups. The remainder decreases with width across all three groups, with empirical slopes around -0.8.

Implementation. Experiments use PyTorch 2.6 on NVIDIA A100 GPUs. The width sweep of one setup (7 widths $\times$ 3 seeds) takes about 14 GPU-hours for DOCC and 18 GPU-hours for Deep SAD.

G.2 Further Results

Figure 7 further examines the dependence of the trajectory remainder on network width. Across training normals, held-out normals, and anomalies, the median, 95th percentile, and maximum of $\|\rho_T(x)\|_2$ consistently decrease as m increases, with empirical slopes around -0.8. For comparison, the finite-width term in the Gaussian-initialization bound scales as $m^{-1/6}$.