Early Alignment without Neural Collapse in Two-Layer ReLU Networks on Gaussian XOR

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

A central question in understanding deep networks is how their learned representations acquire strong discriminative structure, a phenomenon illustrated by Neural Collapse (NC). Motivated by recent analyses of two-layer training dynamics, we study Gaussian XOR data and two-layer ReLU networks. We show that, under an extremal-alignment assumption motivated by early-alignment theory, the hidden features do not exhibit neural collapse: the population NC1 metric remains bounded below by a positive constant. We also give a three-layer ReLU construction that can realize low NC1 on the same data, and use experiments to illustrate the two-layer obstruction and the effect of training choices in three-layer networks.

1. Introduction

Neural collapse describes a late-training geometry in which last-layer features concentrate within each class, class means form organized configurations, and classifiers align with these means [9]. Early theory often used unconstrained feature models, where features are optimized directly rather than produced by hidden layers [5, 8, 11, 12]. More recent work moves closer to actual network dynamics; for example, Min et al. [7] prove neural-collapse behavior for shallow ReLU networks under orthogonally separable data. These results suggest that NC can arise from training, but they also raise the question of how collapse depends on data geometry.

Gaussian XOR is a simple nonlinear model in which each class consists of two separated components. It is more complex than linearly or orthogonally separable settings, but still simple enough for population-level calculations. This makes it a useful testbed for asking whether two-layer ReLU networks still exhibit neural collapse when the class geometry is multimodal rather than single-cluster.

Early-alignment results provide dynamical motivation for the feature structure we study. In different settings, Maennel et al. [6], Boursier et al. [2], and Boursier and Flammarion [1] support the idea that small-initialized ReLU networks may first align hidden weights with a small number of data-dependent directions. For XOR-type data, Glasgow [4] and Tsoy and Konstantinov [10] give related evidence and analysis. These works motivate the four XOR signal directions used below, but do not prove an unconditional population-gradient-flow theorem for Gaussian XOR. Motivated by this picture, we analyze the NC1 geometry of two-layer ReLU features under alignment to the natural XOR signal directions.

We also include experiments to illustrate the theory and the role of depth. In two-layer networks, the learned weights align with the XOR signal directions while NC1 stays away from zero. In three-layer networks, we observe that lower-NC1 representations can appear, and that feature normalization and weight decay are useful training ingredients. Overall, the experiments support the view that learning XOR-relevant directions and producing neural-collapse geometry are related but distinct.

Contributions. First, under an extremal-alignment assumption, we prove a population NC1 lower bound for two-layer ReLU features on Gaussian XOR data. Second, we give a three-layer ReLU construction and experiments showing how additional depth can promote more collapse-like geometry.

2. Model and NC1

Gaussian XOR. Let $e_1, e_2 \in \mathbb{R}^d$ be the first two coordinate vectors and set

$$\mu_1 = e_1 + e_2, \qquad \mu_2 = e_1 - e_2.$$

The latent signal and observed sample are

$$z \sim \mathrm{Unif}\{\pm\mu_1, \pm\mu_2\}, \qquad x = z + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I_d/d).$$

We write $\tau = \sigma/\sqrt{d}$. The clean label and observed label are

$$\tilde{y} = \mathrm{sign}(z_1 z_2), \qquad y = \mathrm{sign}(x_1 x_2).$$

Finally, the normalized signal directions are

$$u_1 = \frac{\mu_1}{\|\mu_1\|}, \qquad u_2 = \frac{\mu_2}{\|\mu_2\|}.$$

Two-layer ReLU. For the ReLU activation $\varphi(t) = \max\{t, 0\}$, consider

$$f_\theta(x) = \sum_{j=1}^{p} a_j \varphi(w_j^\top x), \qquad \phi(x) = \big(\varphi(w_1^\top x), \ldots, \varphi(w_p^\top x)\big).$$

The scalar output $f_\theta(x)$ is used for classification, while NC1 is measured on the hidden feature map $\phi(x)$.

Gradient-flow training. For the theory, we use the population logistic risk

$$L(\theta) = \mathbb{E}\left[\log\big(1 + \exp(-y f_\theta(x))\big)\right].$$

Gradient flow is the continuous-time limit

$$\dot{\theta}(t) \in -\partial^\circ L(\theta(t)).$$

In coordinates, this gives $\dot{a}_j \in -\partial^\circ_{a_j} L$ and $\dot{w}_j \in -\partial^\circ_{w_j} L$. Since ReLU is nonsmooth at the origin, the differential inclusion is interpreted using the Clarke subdifferential [3]. The numerical experiments use full-batch gradient descent as a discrete approximation.

NC1. Given a feature map $h(x)$ and labels $y \in \mathcal{C}$, let

$$\mu_c = \mathbb{E}[h(x) \mid y = c], \qquad \mu_G = \mathbb{E}[h(x)].$$

The within-class covariance and between-class covariance are

$$\Sigma_W = \mathbb{E}[(h(x) - \mu_y)(h(x) - \mu_y)^\top], \qquad \Sigma_B = \mathbb{E}[(\mu_y - \mu_G)(\mu_y - \mu_G)^\top],$$

and define

$$\mathrm{NC1} = \frac{\mathrm{Tr}(\Sigma_W)}{\mathrm{Tr}(\Sigma_B)}.$$

Small NC1 means that features of the same class concentrate relative to the separation between class means.

3. A Two-Layer NC1 Lower Bound

Early-alignment results for finite samples and empirical risks motivate the expectation that small-initialized two-layer networks first move toward a small set of extremal directions [1, 2]. For XOR-type problems, the discussion in Boursier and Flammarion [1] and related simplicity-bias results [4, 10] point to the four signed signal directions $\{\pm u_1, \pm u_2\}$. Motivated by these results, the main theorem analyzes the NC1 geometry conditional on such an extremal-alignment structure. A conditional population-level justification is given in Appendix B.

Assumption 1 (Extremal alignment) For each hidden weight w_j, there are $r_j = \|w_j\| > 0$ and $v_j \in \{\pm u_1, \pm u_2\}$ such that

$$\left\| \frac{w_j}{\|w_j\|} - v_j \right\| \le \varepsilon, \qquad 0 < \varepsilon < \sqrt{2}.$$

Remark 2 Appendix B discusses why this assumption is natural in the clean-label signal-plane regime.

Theorem 3 (Two-layer NC1 lower bound) Under the Gaussian XOR model, suppose Assumption 1 holds. Then there is an absolute constant $C > 0$ such that, for all sufficiently small $\tau = \sigma/\sqrt{d}$ and ε,

$$\mathrm{NC1}_{\mathrm{pop}}(y) \ge 2 + \frac{4}{\sqrt{\pi}}\tau - C(\tau^2 + \varepsilon).$$

Consequently, the aligned two-layer ReLU representation does not exhibit neural collapse in the NC1 sense; as $\tau \to 0$, the lower bound tends to 2.

The proof is deferred to Appendix A. The point is geometric. In the aligned two-layer representation, the four signed XOR components are detected, but the two components with the same label are not merged. Thus the features can be discriminative while the within-class scatter remains comparable to the between-class separation.

4. A Three-Layer Construction

The previous result shows that two-layer extremal alignment can learn useful XOR directions while still preserving the four-component geometry, so NC1 remains bounded away from zero. This motivates adding one more hidden ReLU layer. The first layer detects the four signed XOR components, and the second layer combines the two components with the same label, producing a more class-level feature representation. We give an explicit construction based on such picture.

Proposition 4 (Three-layer feature map with small NC1) Consider the Gaussian XOR model in $\mathbb{R}^d$. For any hidden widths $d_1 = 4p$ and $d_2 = 2q$, where $p, q \geq 1$, there exists an explicit three-layer ReLU feature map

$$h(x) = \varphi(W_2\varphi(W_1 x)) \in \mathbb{R}^{d_2}$$

such that, for sufficiently small $\tau = \sigma/\sqrt{d}$,

$$\mathrm{NC1}(h) \leq C\tau^2.$$

Equivalently, $\mathrm{NC1}(h) \leq C\sigma^2/d$. The construction and proof are given in Appendix C.

Remark 5 For example, taking $d_1 = 64$, $d_2 = 36$, $d = 64$, and $\sigma = 0.2$, the same construction gives $\mathrm{NC1}(h) < 10^{-3}$.

The construction first uses p repeated neurons in each of the four directions $u_1, -u_1, u_2, -u_2$, separating the four XOR components into four first-layer blocks. The second ReLU layer then uses q repeated neurons in each of two pooling directions: one pools the two positive-label first-layer blocks, and the other pools the two negative-label blocks. The replication parameters p and q, together with the layer scales, only replicate and rescale the features; they cancel in the NC1 ratio.

5. Numerical Experiments

We conduct experiments to illustrate the two main messages of the paper. First, two-layer ReLU networks learn XOR-relevant directions, but their hidden features do not collapse. Second, three-layer networks can produce smaller NC1, with a visible dependence on normalization and regularization. All experiments use full-batch gradient descent; implementation details and hyperparameters are deferred to Appendix D.

Two-layer alignment without collapse. Figure 1 shows the two-layer behavior. The hidden weights concentrate near the four XOR signal directions $\{\pm u_1, \pm u_2\}$, consistent with the extremal-alignment picture. At the same time, the features remain component-level and the NC1 curve stays bounded away from zero, as predicted by Theorem 3.

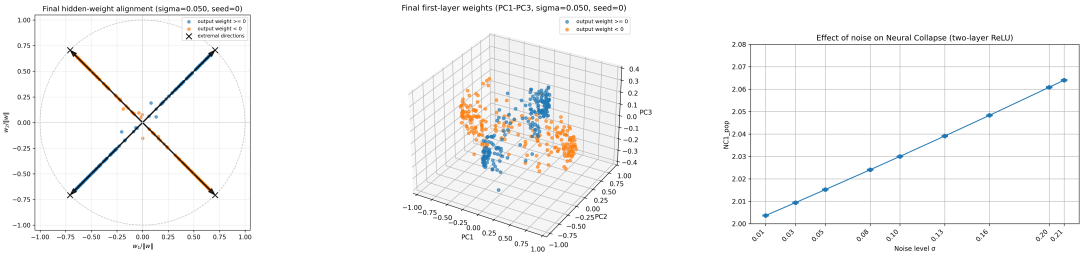

Figure 1: Two-layer ReLU behavior. Left/middle: final hidden-weight directions cluster near the XOR signal directions. Right: population NC1 estimate over noise levels.

Three-layer collapse behavior. We use a two-stage training procedure. In the first stage, a two-layer ReLU network is trained on Gaussian XOR. In the second stage, the trained first-layer weights are copied into a three-layer ReLU network as the first layer. We compare three protocols in this second stage: standard training, training with feature normalization, and training with both feature normalization and weight decay. Figure 2 reports the corresponding NC1 trajectories.

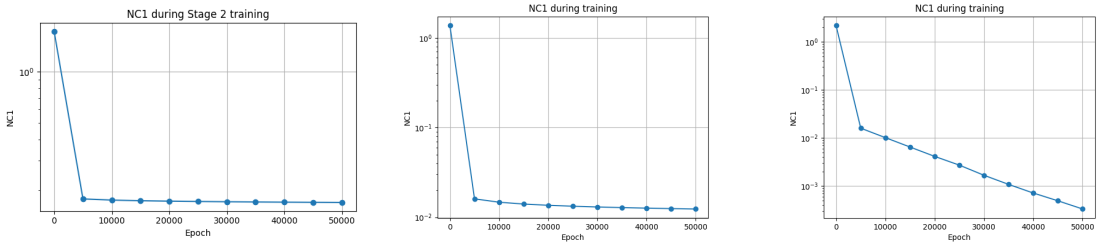

Figure 2: Three-layer NC1 trajectories under three protocols: standard training, feature normalization, and feature normalization with weight decay.

Overall, the experiments suggest that three-layer ReLU networks can reach low NC1 on Gaussian XOR data in the tested training regimes. The results further indicate that feature normalization and weight decay may help stabilize and promote low-NC1 feature geometry.

Appendix A. Proof of Theorem 3

We write $\tau = \sigma/\sqrt{d}$. The proof has three parts: the ideal clean-label computation, stability under ε-cone perturbations, and transfer from clean labels to observed labels.

Lemma 6 (Ideal clean-label computation) Assume ideal extremal alignment,

$$w_j = r_j s_j u_{\ell(j)}, \qquad r_j > 0, \quad s_j \in \{\pm 1\}, \quad \ell(j) \in \{1, 2\}.$$

Let $\widetilde{\mathrm{NC1}}_{\mathrm{pop}}$ be the population NC1 computed with respect to the clean label $\tilde{y}$. Then

$$\widetilde{\mathrm{NC1}}_{\mathrm{pop}} = \frac{2(V_A + V_B)}{(A - B)^2},$$

where, with $a = \sqrt{2}$, $g \sim \mathcal{N}(0, \tau^2)$, and $s \in \{\pm 1\}$ uniform,

$$A = \mathbb{E}[(sa + g)_+], \qquad B = \mathbb{E}[(g)_+],$$

$$V_A = \mathrm{Var}((sa + g)_+), \qquad V_B = \mathrm{Var}((g)_+).$$

Moreover,

$$\widetilde{\mathrm{NC1}}_{\mathrm{pop}} = 2 + \frac{4}{\sqrt{\pi}}\tau + O(\tau^2)$$

as $\tau \to 0$.

Proof Fix a coordinate j. If the clean label matches the direction $u_{\ell(j)}$, then

$$u_{\ell(j)}^\top z = sa, \qquad s \in \{\pm 1\} \text{ uniform},$$

and $\phi_j(x) = r_j(sa + g)_+$. If the clean label does not match this direction, then $u_{\ell(j)}^\top z = 0$, and $\phi_j(x) = r_j(g)_+$. Thus the matched and unmatched class means are $r_j A$ and $r_j B$, and the corresponding variances are $r_j^2 V_A$ and $r_j^2 V_B$. Summing over j gives

$$\sum_k \|\tilde{\phi}_k - \tilde{\bar{\phi}}\|^2 = \frac{1}{2}(A - B)^2 \sum_j r_j^2,$$

and

$$\sum_k \mathbb{E}[\|\phi(x) - \tilde{\phi}_k\|^2 \mid \tilde{y} = k] = (V_A + V_B) \sum_j r_j^2.$$

This proves the displayed NC1 formula.

For the asymptotics, let $c = a/\tau$. The Gaussian-ReLU moment identities give

$$A = \tau\varphi_{\mathrm{g}}(c) + a\Big(\Phi(c) - \frac{1}{2}\Big), \qquad B = \frac{\tau}{\sqrt{2\pi}},$$

$$V_A = \frac{a^2 + \tau^2}{2} - A^2, \qquad V_B = \frac{\tau^2}{2}\left(1 - \frac{1}{\pi}\right),$$

where φ_g and Φ are the standard Gaussian density and distribution function. Since $c \to \infty$, $A = a/2 + O(e^{-a^2/(2\tau^2)})$. Substitution with $a = \sqrt{2}$ yields

$$\widetilde{\mathrm{NC1}}_{\mathrm{pop}} = 2 + \frac{4}{\sqrt{\pi}}\tau + O(\tau^2).$$

■

Lemma 7 (Shift stability for ReLU moments) Let η be symmetric with $\mathbb{E}[\eta^2] < \infty$, and define

$$m(t) = \mathbb{E}[(\eta + t)_+], \qquad q(t) = \mathbb{E}[(\eta + t)_+^2], \qquad v(t) = q(t) - m(t)^2.$$

Then

$$|m(t) - m(t')| \le |t - t'|,$$

and

$$|v(t) - v(t')| \le C|t - t'|\big(|t| + |t'| + \mathbb{E}|\eta|\big)$$

for an absolute constant $C > 0$.

Proof The first bound follows from the 1-Lipschitz property of ReLU. For the variance, use

$$|(x_+)^2 - (y_+)^2| \le |x - y|(|x| + |y|)$$

to control $q(t) - q(t')$, and combine it with $v = q - m^2$ and the first bound. ■

Proof [Proof of Theorem 3] First consider the ideal clean-label model. By Lemma 6,

$$\widetilde{\mathrm{NC1}}^{\mathrm{ideal}}_{\mathrm{pop}} \ge 2 + \frac{4}{\sqrt{\pi}}\tau - C_0\tau^2$$

for sufficiently small τ.

Now suppose Assumption 1 holds. For each j, choose $v_j \in \{\pm u_1, \pm u_2\}$ with

$$\|w_j/\|w_j\| - v_j\| \le \varepsilon, \qquad r_j = \|w_j\|.$$

The matched clean signal shift differs from $\pm\sqrt{2}r_j$ by at most $C\varepsilon r_j$, while the unmatched clean signal shift differs from 0 by at most $C\varepsilon r_j$. Applying Lemma 7 coordinatewise and averaging over the two clean components in a class gives

$$|\tilde{\phi}_{k,j} - \tilde{\phi}^0_{k,j}| \le C\varepsilon r_j,$$

and

$$\big|\mathrm{Var}(\phi_j(x) \mid \tilde{y} = k) - \mathrm{Var}^0(\phi_j(x) \mid \tilde{y} = k)\big| \le C\varepsilon r_j^2,$$

where the superscript 0 denotes the ideal comparison model. Setting $R = \sum_j r_j^2$, the numerator and denominator of clean-label NC1 therefore satisfy

$$\tilde{N} \ge N_0 - C\varepsilon R, \qquad \tilde{D} \le D_0 + C\varepsilon R.$$

Since N_0 and D_0 are both proportional to R, and $D_0 \geq cR$ for small τ, we get

$$\widetilde{\mathrm{NC1}}_{\mathrm{pop}} \geq \widetilde{\mathrm{NC1}}_{\mathrm{pop}}^{\mathrm{ideal}} - C\varepsilon \geq 2 + \frac{4}{\sqrt{\pi}}\tau - C(\tau^2 + \varepsilon).$$

The preceding argument proves the lower bound for the clean-label NC1, i.e., for the numerator and denominator conditioned on $\tilde{y}$. It remains to compare this quantity with the NC1 conditioned on the observed label $y = \mathrm{sign}(x_1 x_2)$. Let $E = \{y \neq \tilde{y}\}$. A label flip can occur only if at least one of the first two Gaussian-noise coordinates crosses the corresponding coordinate hyperplane. Thus, with $\xi_i \sim N(0, \tau^2)$,

$$\mathbb{P}(E) \leq 4\,\mathbb{P}(\xi_1 < -1) \leq Ce^{-c/\tau^2}.$$

The feature contribution on this rare event is also exponentially small. Indeed, $\phi_j(x)^2 \leq (w_j^\top x)^2$, and Gaussian fourth-moment bounds give

$$\mathbb{E}\|\phi(x)\|^4 \leq C\left(\sum_j \|w_j\|^2\right)^2 = CR^2.$$

By Cauchy–Schwarz,

$$\mathbb{E}[\|\phi(x)\|^2 \mathbf{1}_E] \leq \left(\mathbb{E}\|\phi(x)\|^4\right)^{1/2} \mathbb{P}(E)^{1/2} \leq Ce^{-c/\tau^2} R.$$

The NC1 numerator and denominator depend only on conditional first and second moments. Since conditioning on y and on $\tilde{y}$ differs only through E, the preceding bounds imply

$$|N_y - \tilde{N}| \leq Ce^{-c/\tau^2} R, \qquad |D_y - \tilde{D}| \leq Ce^{-c/\tau^2} R.$$

Since the clean-label denominator is bounded below by cR, it follows that

$$\left|\mathrm{NC1}_{\mathrm{pop}}(y) - \widetilde{\mathrm{NC1}}_{\mathrm{pop}}\right| \leq Ce^{-c/\tau^2}.$$

Hence

$$\mathrm{NC1}_{\mathrm{pop}}(y) \geq \widetilde{\mathrm{NC1}}_{\mathrm{pop}} - Ce^{-c/\tau^2}.$$

Absorbing the exponentially small term into $C\tau^2$ gives

$$\mathrm{NC1}_{\mathrm{pop}}(y) \geq 2 + \frac{4}{\sqrt{\pi}}\tau - C(\tau^2 + \varepsilon).$$

■

Appendix B. Conditional Population Alignment Justification

We record the population alignment statement used to motivate Assumption 1. This is a conditional clean-label statement in the signal plane, not a full long-time theorem for empirical Gaussian-XOR training.

Let $\mathcal{P} = \mathrm{span}\{u_1, u_2\}$. For $w_j(t) \neq 0$, write

$$r_j(t) = \|w_j(t)\|, \qquad q_j(t) = \frac{w_j(t)}{\|w_j(t)\|} \in S^1 \subset \mathcal{P},$$

and set $s_j = \mathrm{sign}(a_j(0))$. Thus $q_j(t)$, rather than $w_j(t)$, is the angular variable.

Proposition 8 (Population angular early alignment) Fix $\eta \in (0,1)$ and $\varepsilon \in (0,2)$. Assume balanced signal-plane initialization,

$$a_j(0)^2 = \|w_j(0)\|^2 \neq 0, \qquad w_j(0) \in \mathcal{P}.$$

Suppose that the initial direction of neuron j is separated from the relevant separatrix:

$$a_j(0) > 0 \Rightarrow |q_{j1}(0)| \geq \eta, \qquad a_j(0) < 0 \Rightarrow |q_{j2}(0)| \geq \eta.$$

Under the reduced early-time population angular dynamics, and as long as the usual small-output perturbation bound holds up to the alignment time, $q_j(t)$ enters the ε-neighborhood of

$$\operatorname{sign}(q_{j1}(0))u_1 \quad \text{if } a_j(0) > 0, \qquad \operatorname{sign}(q_{j2}(0))u_2 \quad \text{if } a_j(0) < 0.$$

The statement supports the extremal-alignment assumption in a clean-label, signal-plane regime, and is consistent with finite-sample early-alignment theory, but it should not be read as proving that arbitrary two-layer ReLU training always aligns on Gaussian XOR. Proof [Proof idea] In the clean-label signal plane, the early-time population angular field is governed by the clean logistic correlation potential

$$G^{\mathrm{cl}}_{\log}(w) := \frac{1}{2}\mathbb{E}[Y(w^\top X)_+] = r\, g_{\log}(q_1, q_2), \qquad w = r(q_1 u_1 + q_2 u_2).$$

For Gaussian XOR,

$$g_{\log}(q_1, q_2) = \frac{1}{2}(\psi(q_1) - \psi(q_2)), \qquad \psi'(q) = \frac{\sqrt{2}}{2}\left[\Phi\left(\frac{\sqrt{2}q}{\tau}\right) - \frac{1}{2}\right].$$

Thus $\psi'(q) > 0$ for $q > 0$ and $\psi'(q) < 0$ for $q < 0$. Away from the separatrices $q_1 = 0$ or $q_2 = 0$, this gives a strictly positive angular drift toward the corresponding signed axis. For $a_j(0) > 0$, the relevant drift is toward $\operatorname{sign}(q_{j1}(0))u_1$; for $a_j(0) < 0$, it is toward $\operatorname{sign}(q_{j2}(0))u_2$.

The remaining term in the true population gradient flow comes from the nonzero network output. Under the small-output perturbation bound stated in Proposition 8, this term is dominated by the clean angular drift until the direction enters the prescribed ε-neighborhood. This proves the conditional population-level alignment statement in the clean-label, signal-plane setting. ■

Appendix C. Three-Layer Construction

We prove Proposition 4. In this construction appendix, y denotes the clean XOR label. The data are

$$x = z + \xi, \qquad z \sim \mathrm{Unif}\{\pm\mu_1, \pm\mu_2\}, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I_d/d),$$

where

$$\mu_1 = e_1 + e_2, \qquad \mu_2 = e_1 - e_2, \qquad \tau = \frac{\sigma}{\sqrt{d}}.$$

The clean labels are

$$y = +1 \text{ on } \{\mu_1, -\mu_1\}, \qquad y = -1 \text{ on } \{\mu_2, -\mu_2\}.$$

For a feature map h, we use

$$\mathrm{NC1}(h) = \frac{\sum_{k=\pm 1} \mathbb{E}[\|h(x) - h_k\|^2 \mid y = k]}{\sum_{k=\pm 1} \|h_k - \bar{h}\|^2}, \qquad h_k = \mathbb{E}[h(x) \mid y = k], \qquad \bar{h} = \frac{1}{2}(h_{+1} + h_{-1}).$$

Let

$$u_1 = \frac{\mu_1}{\|\mu_1\|}, \qquad u_2 = \frac{\mu_2}{\|\mu_2\|}.$$

Fix widths $d_1 = 4p$, $d_2 = 2q$, with $p, q \geq 1$. Choose any $\alpha, \beta > 0$. The first layer puts p neurons in each direction $u_1, -u_1, u_2, -u_2$, all with norm α. Writing

$$U = u_1^\top x, \qquad V = u_2^\top x,$$

the first hidden feature is

$$h^{(1)}(x) = \varphi(W_1 x) = \alpha\big(\varphi(U)\mathbf{1}_p, \varphi(-U)\mathbf{1}_p, \varphi(V)\mathbf{1}_p, \varphi(-V)\mathbf{1}_p\big).$$

For the second layer, define the normalized pooling directions

$$q_+ = \frac{1}{\sqrt{2p}}(\mathbf{1}_p, \mathbf{1}_p, 0, 0), \qquad q_- = \frac{1}{\sqrt{2p}}(0, 0, \mathbf{1}_p, \mathbf{1}_p).$$

We put q neurons in direction q_+ and q neurons in direction q_-, all with norm β. Since $\varphi(t) + \varphi(-t) = |t|$,

$$q_+^\top h^{(1)}(x) = \alpha\sqrt{\frac{p}{2}}|U|, \qquad q_-^\top h^{(1)}(x) = \alpha\sqrt{\frac{p}{2}}|V|.$$

Thus the final hidden feature has the form

$$h(x) = K(|U|\mathbf{1}_q, |V|\mathbf{1}_q), \qquad K = \alpha\beta\sqrt{\frac{p}{2}}.$$

Because u_1, u_2 are orthonormal and the noise is isotropic,

$$u_1^\top \xi,\ u_2^\top \xi \sim \mathcal{N}(0, \tau^2)$$

and the two variables are independent. If $y = +1$, then

$$|U| \overset{d}{=} |m + \eta|, \qquad |V| \overset{d}{=} |\zeta|,$$

where $m = \sqrt{2}$ and $\eta, \zeta \sim \mathcal{N}(0, \tau^2)$. If $y = -1$, the two roles are exchanged. Set

$$A = |m + \eta|, \qquad B = |\zeta|, \qquad a = \mathbb{E}A, \qquad b = \mathbb{E}B.$$

Then

$$h(x) \mid y = +1 \overset{d}{=} K(A\mathbf{1}_q, B\mathbf{1}_q), \qquad h(x) \mid y = -1 \overset{d}{=} K(B\mathbf{1}_q, A\mathbf{1}_q),$$

and hence

$$h_+ = K(a\mathbf{1}_q, b\mathbf{1}_q), \qquad h_- = K(b\mathbf{1}_q, a\mathbf{1}_q).$$

The NC1 numerator is

$$2K^2 q\big(\mathrm{Var}(A) + \mathrm{Var}(B)\big),$$

whereas the denominator is

$$K^2 q(a-b)^2.$$

Therefore

$$\mathrm{NC1}(h) = \frac{2(\mathrm{Var}(A) + \mathrm{Var}(B))}{(a-b)^2}.$$

The factors K, p, q, α, β have all canceled, which is why the same construction works for all flexible widths $d_1 = 4p$ and $d_2 = 2q$.

It remains to evaluate the one-dimensional moments. For

$$B = |\zeta|, \qquad \zeta \sim \mathcal{N}(0, \tau^2),$$

we have

$$b = \tau\sqrt{\frac{2}{\pi}}, \qquad \mathrm{Var}(B) = \tau^2\left(1 - \frac{2}{\pi}\right).$$

For $A = |m + \eta|$, the folded-normal moment formula gives

$$a = \tau\sqrt{\frac{2}{\pi}}\exp\left(-\frac{m^2}{2\tau^2}\right) + m\ \mathrm{erf}\left(\frac{m}{\sqrt{2}\tau}\right), \qquad \mathrm{Var}(A) = m^2 + \tau^2 - a^2.$$

Thus, as $\tau \to 0$,

$$a = m + O(e^{-c/\tau^2}), \qquad \mathrm{Var}(A) = \tau^2 + O(e^{-c/\tau^2}),$$

for some constant $c > 0$. Also

$$\mathrm{Var}(A) + \mathrm{Var}(B) = \tau^2\left(2 - \frac{2}{\pi}\right) + O(e^{-c/\tau^2}),$$

and

$$(a-b)^2 = \left(m - \tau\sqrt{\frac{2}{\pi}}\right)^2 + O(e^{-c/\tau^2}).$$

Since $m^2 = 2$, this yields

$$\mathrm{NC1}(h) = 2\left(1 - \frac{1}{\pi}\right)\tau^2 + O(\tau^3).$$

Substituting $\tau = \sigma/\sqrt{d}$ gives

$$\mathrm{NC1}(h) = 2\left(1 - \frac{1}{\pi}\right)\frac{\sigma^2}{d} + O\left(\frac{\sigma^3}{d^{3/2}}\right),$$

which proves Proposition 4.

For the concrete example in the main text, $d_1 = 64$ and $d_2 = 36$ correspond to $p = 16$ and $q = 18$. Taking $d = 64$ and $\sigma = 0.2$, the leading term

$$2\left(1 - \frac{1}{\pi}\right)\frac{\sigma^2}{d}$$

is about 8.5×10^{-4}, and the higher-order term is negligible in the small-noise regime. The choices of $\alpha, \beta > 0$ only set concrete weights and do not affect the NC1 value.

Appendix D. Experimental Details

The two-layer runs use Gaussian XOR with $d = 100$, 100 samples from each of the four components, width 400, initialization scale 0.02, learning rate 0.1, and 5000 full-batch gradient-descent steps. The noise level is varied for the NC1 curve. Population NC1 is estimated by fresh Monte Carlo samples from the same Gaussian XOR distribution.

The three-layer runs use $d = 200$, $\sigma = 0.2$, and 100 samples from each component. Stage 1 trains a two-layer ReLU network of width 300 for 10,000 steps with learning rate 0.1 and initialization scale 0.02. Stage 2 copies the first layer into a three-layer network with second hidden width 100, trains for 50,000 steps with learning rate 0.9, and compares three protocols: standard training, feature normalization, and feature normalization with weight decay 10^{-4}.

Protocol	Empirical NC1	Population NC1
Feature normalization + weight decay	2.1953×10^{-4}	2.1971×10^{-4}
Feature normalization only	1.1366×10^{-2}	1.1371×10^{-2}
Standard three-layer training	1.6990×10^{-1}	1.6990×10^{-1}

The slow-first-layer control experiment and additional curves are omitted from the main text. They can be included in an extended appendix if needed.

References

[1] Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. Journal of Machine Learning Research, 26(108):1–72, 2025.

[2] Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. arXiv preprint arXiv:2206.00939, 2022.

[3] Frank H. Clarke. Optimization and Nonsmooth Analysis, volume 5 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[4] Margalit R. Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. In International Conference on Learning Representations, 2024.

[5] Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J. Su. An unconstrained layer-peeled perspective on neural collapse. In International Conference on Learning Representations, 2022.

[6] Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. arXiv preprint arXiv:1803.08367, 2018.

[7] Hancheng Min, Zhihui Zhu, and Rene Vidal. Neural collapse under gradient flow on shallow relu networks for orthogonally separable data. In Advances in Neural Information Processing Systems (NeurIPS), 2025.

[8] Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. Sampling Theory, Signal Processing, and Data Analysis, 20(2):11, 2022.

[9] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences, 117(40):24652–24663, 2020.

[10] Nikita Tsoy and Nikolaos Konstantinov. Simplicity bias of two-layer networks beyond linearly separable data. In Proceedings of the 41st International Conference on Machine Learning, 2024.

[11] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In Proceedings of the 39th International Conference on Machine Learning, 2022.

[12] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In Advances in Neural Information Processing Systems (NeurIPS), 2021.