KNOWING LESS, TEACHING MORE: WEAK-TO-STRONG GENERALIZATION FROM PARTIAL KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Weak-to-strong generalization (W2S) occurs when a strong student trained on a weak teacher's labels outperforms the teacher on unseen samples. We study how a strong student can exploit supervision from a weak teacher that reliably recognizes an easy feature but has only partial knowledge of a hard feature, producing stochastic label noise on hard-only samples. For a two-layer ReLU network trained by gradient descent, we show that hard-feature learning approaches a saturation level determined by the feature-strength-weighted balance between correctly and incorrectly labeled hard-only samples. Overlap samples containing both features initially promote hard-feature learning, but their contribution vanishes as the easy feature is learned. Nevertheless, the student retains a positive fraction of the saturation level while fitting incorrect labels through sample-specific noise, which has little effect on unseen samples. Consequently, its test error converges to zero during a prescribed late-training interval, even after fitting every weak label. Experiments with ReLU networks and language models show that, at fixed label accuracy, assigning correct labels to samples with stronger hard features improves feature learning and student accuracy. A teacher-selection study further shows that hard-feature knowledge can be a better criterion for assessing weak-supervision quality than validation accuracy.

1 INTRODUCTION

As models become more capable, obtaining reliable supervision for their continued improvement becomes increasingly challenging. This raises a fundamental question: can weaker models provide useful supervision for stronger ones? Whereas knowledge distillation typically transfers knowledge from a larger teacher to a smaller student (Hinton et al., 2015), weak-to-strong generalization (W2S) occurs when a stronger student trained on a weaker teacher's labels outperforms that teacher (Burns et al., 2024). Understanding this phenomenon requires explaining *how the student learns beyond its weak supervisor rather than merely reproducing its errors*.

Several theoretical studies explain W2S through feature learning. In particular, Shin et al. (2025), Oh et al. (2025) distinguish between easy features that the weak teacher recognizes and hard features to which it is blind. Overlap samples containing both features bridge this gap: the teacher provides correct labels using the easy feature, allowing the student to learn the co-occurring hard feature. Building on this mechanism, Shin et al. (2025) propose increasing the proportion of overlap samples, while Oh et al. (2025) study a linear teacher unable to recognize nonlinear hard-feature patterns. These analyses explain how students learn features their teachers cannot recognize, but leave open how students exploit features their teachers understand only partially.

A weak teacher, however, need not be completely blind to a hard feature. It may recognize the feature imperfectly, providing more reliable labels when the feature is strongly expressed and making more errors when it is subtle. Such partial knowledge can provide noisy yet informative supervision even without overlap samples, suggesting another route to W2S. We therefore ask *how a strong student extracts and preserves hard-feature knowledge from imperfect supervision*. Answering this question requires understanding both how the student aggregates useful signals across noisy labels and whether the acquired knowledge survives as training continues and the student fits the teacher's errors.

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.

We study a two-layer ReLU network trained by gradient descent on fixed weak labels. Each sample contains an easy feature, a hard feature, or both, along with independent sample-specific noise. The teacher reliably labels samples containing the easy feature but has only partial knowledge of the hard feature, modeled through a limited feature response perturbed by stochastic output noise. On hard-only samples, the probability of a correct weak label is nondecreasing in hard-feature strength. This setting allows us to trace how the teacher's partial knowledge on the hard feature shapes the student's learning throughout training.

Our main result explains how hard-feature knowledge acquired early in training survives later noise memorization. The student's hard-feature response initially approaches a saturation level determined by the feature-strength-weighted balance between correctly and incorrectly labeled hard-only samples. Overlap samples also support hard-feature learning, but their contribution diminishes as the easy feature is learned. During later training, the student memorizes noise faster on label-flipped samples, weakening their opposing hard-feature updates and retaining a positive fraction of the saturation level. Since memorized noise has little effect on unseen samples, the student can, under our assumptions, fit all weak training labels while generalizing beyond the teacher, even without overlap samples.

We examine these findings in synthetic experiments with ReLU networks and sentiment classification with large language models. At fixed weak-label accuracy, assigning correct labels to hard-only samples with stronger features increases the learned hard-feature response and improves student accuracy on average. Finally, we compare teacher selection by validation accuracy with selection by the correct-to-flipped hard-feature-strength ratio, using the same hard-only validation subset. This case study identifies settings in which a less accurate teacher produces a more accurate student, demonstrating the practical value of assessing which samples a teacher labels correctly, beyond how often it is correct.

Main contributions.

- **Hard feature learning from partial teacher knowledge.** We quantify how label correctness and feature strength jointly determine the early hard-feature saturation level under a teacher's partial hard-feature knowledge, and characterize the diminishing contribution of overlap samples.
- **Feature preservation and generalization.** We establish conditions under which the student retains a positive fraction of this saturation level while fitting all weak training labels, enabling generalization beyond the teacher even without overlap samples.
- **Empirical validation and teacher selection.** Experiments with ReLU networks and large language models show that assigning correct labels to samples with stronger hard features improves feature learning and student accuracy at fixed weak-label accuracy. A teacher-selection study demonstrates that hard-feature information can identify more effective teachers than validation accuracy alone.

Related work. Theoretical analyses of W2S study when a student can correct or avoid inheriting its teacher's errors. Lang et al. (2024) establish conditions for correcting teacher errors and generalizing to samples left unlabeled by the teacher. Under fixed representations, the geometry of input features determines which errors transfer to the student (Xue et al., 2025). For networks with randomly sampled hidden features, increased width together with early stopping can enable W2S (Medvedev et al., 2025). Other analyses characterize prediction-error gains under regression and classification objectives (Charikar et al., 2024, Mulgund & Pabbaraju, 2025, Ildiz et al., 2025).

Our work is closest to feature-learning analyses that use overlap samples to transfer supervision from easy to hard features (Shin et al., 2025, Oh et al., 2025). We instead study teachers with partial hard-feature knowledge, allowing hard-only samples to provide informative supervision. Related work studies feature acquisition and retention with pretraining (Moniri & Hassani, 2025, Awano & Suzuki, 2026), benign overfitting in neural networks (Kou et al., 2023), and W2S through linear interpolation (Wu & Sahai, 2025). Our contribution connects a feature-strength-dependent saturation point to a quantitative preservation bound during noise memorization, while accounting for the fading contribution of overlap samples.

Practical approaches improve W2S through confidence regularization (Burns et al., 2024), sample selection or weighting (Guo & Yang, 2024, Uzunoğlu et al., 2026), and label refinement (Somerstep et al., 2025). Our teacher-selection study complements these methods by assessing how correct and incorrect labels are distributed across hard-feature strengths, beyond overall label accuracy.

2 PROBLEM SETUP

We call the weak model providing labels the *teacher* and the strong model trained on them the *student*. We ask whether the student predicts true labels on new samples more accurately than the teacher.

Notation. Bold letters denote vectors and matrices, and $[n] = \{1, \dots, n\}$. We use $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ and $\|\boldsymbol{v}\|$ for the Euclidean inner product and norm, respectively. All asymptotics are as the patch dimension $d \to \infty$; $\mathrm{poly}(d)$ and $\mathrm{polylog}(d)$ denote a polynomial in d and a fixed positive power of $\log d$, resp.

Data and features. We study binary classification with true label $y \in \{\pm 1\}$ using a patch model inspired by Oh et al. (2025). For each class $k \in \{\pm 1\}$, let $\boldsymbol{\mu}_k, \boldsymbol{\nu}_k \in \mathbb{R}^d$ denote its easy and hard features, respectively; all four feature vectors are orthonormal. Each input $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}) \in \mathbb{R}^{d \times 3}$ consists of two signal patches containing class features and one noise patch.

Definition 1 (Data distribution)**.** *We generate a sample $(\boldsymbol{X}, y)$ as follows:*

- ***Label and Type.*** *Draw a label $y \in \{\pm 1\}$ uniformly and a sample type $\tau = \mathrm{E}$ (easy-only), H (hard-only), or O (overlap) with probabilities $p_E, p_H > 0$, $p_O \geq 0$, and $p_E + p_H + p_O = 1$.*
- ***Signal patches.*** *Draw the hard-feature strength $\alpha \sim P_\alpha$ on $[\alpha_{\min}, \alpha_{\max}] \subset (0, 1)$ and set*

$$(\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}) = \begin{cases} (\boldsymbol{\mu}_y, \boldsymbol{\mu}_y), & \tau = \mathrm{E}; \\ (\boldsymbol{\mu}_y, \alpha \boldsymbol{\nu}_y), & \tau = \mathrm{O}; \\ (\alpha \boldsymbol{\nu}_y, \alpha \boldsymbol{\nu}_y), & \tau = \mathrm{H}. \end{cases} \tag{1}$$

- ***Input noise.*** *Draw Gaussian input noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_p^2 \boldsymbol{P}_\perp)$ independently of (y, τ, α), where $\sigma_p > 0$ is the input-noise scale and $\boldsymbol{P}_\perp = \boldsymbol{I}_d - \sum_k (\boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top + \boldsymbol{\nu}_k \boldsymbol{\nu}_k^\top)$.*
- ***Patch order.*** *Form $\boldsymbol{X}$ by randomly permuting $(\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \boldsymbol{\xi})$.*

Weak supervision. We denote the output of the trained teacher by $G(\boldsymbol{X})$ and its weak label by $\tilde{y} = \mathrm{sign}(G(\boldsymbol{X}))$. In the following definition, the nonnegative responses b_e and b_h quantify the teacher's easy and hard feature knowledge, respectively.

Definition 2 (Weak teacher response)**.** *Let ϵ be output noise independent of the data, symmetric about zero, and supported on $[-B, B]$, with continuous cumulative distribution function F_ϵ. The teacher's true-label aligned response M_w and signed weak label $\tilde{y}$ are*

$$M_w = yG(\boldsymbol{X}) = \begin{cases} 2b_e + \epsilon, & \tau = \mathrm{E}; \\ b_e + \alpha b_h + \epsilon, & \tau = \mathrm{O}; \\ 2\alpha b_h + \epsilon, & \tau = \mathrm{H}, \end{cases} \qquad \tilde{y} = y\,\mathrm{sign}(M_w). \tag{2}$$

We assume that ($b_e > B \geq |\epsilon|$), so the teacher correctly labels any sample containing the easy feature. On hard-only samples, the label flips whenever the noise ϵ overwhelms the hard-feature response $2\alpha b_h$. Thus, the teacher's weakness is characterized by b_h relative to the noise ϵ.

Proposition 1 (Weak teacher accuracy)**.** *Suppose that $b_e > B$. Then,*

$$\Pr(\tilde{y} = y \mid \tau \in \{\mathrm{E}, \mathrm{O}\}) = 1; \qquad \Pr(\tilde{y} = y \mid \tau = \mathrm{H}, \alpha) = F_\epsilon(2\alpha b_h).$$

When $b_h = 0$, the teacher's accuracy is $F_\epsilon(0) = 1/2$ on hard-only samples ($\tau = \mathrm{H}$). When $b_h > 0$, samples with stronger hard-feature strength α are more likely to receive correct weak labels.

Student model. The student model $F_{\boldsymbol{W}}$ is a two-layer neural network using ReLU activation. For each class $k \in \{\pm 1\}$, it has m trainable weight vectors $\boldsymbol{w}_{k,j} \in \mathbb{R}^d$, indexed by $j \in [m]$. Let $\boldsymbol{W}$ denote the collection of all weight vectors. The student's output is

$$F_{\boldsymbol{W}}(\boldsymbol{X}) = \frac{1}{m} \sum_{j=1}^{m} \sum_{p=1}^{3} \left[\phi(\langle \boldsymbol{w}_{+1,j}, \boldsymbol{x}^{(p)} \rangle) - \phi(\langle \boldsymbol{w}_{-1,j}, \boldsymbol{x}^{(p)} \rangle) \right], \quad \phi(u) = \max(0, u). \tag{3}$$

The student is strong in that it can represent all four feature directions and generalize to the data distribution. However, it also has sufficient capacity to fit incorrect labels by memorizing sample-specific input noise. Our central question is how the student can exploit weak supervision that is noisy along the hard-feature direction to learn the hard feature and generalize beyond the weak teacher.

Training and evaluation. We draw n independent samples $\{(\boldsymbol{X}_i, y_i)\}_{i=1}^n$ from the data distribution and assign the weak label $\tilde{y}_i = \mathrm{sign}(G(\boldsymbol{X}_i))$. For sample i, let y_i, τ_i, α_i, and $\boldsymbol{\xi}_i$ denote its true label, sample type, hard-feature strength, and input noise, respectively. For each true class k, let $\mathcal{E}_k, \mathcal{O}_k, \mathcal{H}_k$ denote the index sets of easy-only, overlap, and hard-only training samples, respectively.

We initialize each weight vector independently as $\boldsymbol{w}_{k,j}^{(0)} \sim \mathcal{N}(0, \sigma_0^2 \boldsymbol{I}_d)$. We train the student by full-batch gradient descent with step size $\eta > 0$ on the empirical logistic loss:

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-\tilde{y}_i F_{\boldsymbol{W}}(\boldsymbol{X}_i)}); \qquad \boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta \nabla \mathcal{L}(\boldsymbol{W}^{(t)}). \tag{4}$$

We write $\boldsymbol{W}^{(t)}$ for the parameters after t updates and train for $T = \mathrm{poly}(d)/\eta$ steps. Training uses only inputs and weak labels; true labels, sample types, and feature strengths are used only in the analysis. We evaluate the trained student on a fresh test sample $(\boldsymbol{X}, y)$ drawn from the same data distribution, measuring test error by $\Pr_{(\boldsymbol{X},y)}(yF_{\boldsymbol{W}}(\boldsymbol{X}) < 0)$.

3 Main Results

In this section, we analyze the student's training dynamics and show how it exploits the teacher's limited knowledge of the hard feature to achieve more accurate predictions. We consider the asymptotic regime ($d \to \infty$) and allow all parameters to depend on d unless stated otherwise. Fig. 1 illustrates the feature-learning and noise-memorization dynamics analyzed below.

Parameter conditions. The main parameter conditions are given below; see Appx. C.2 for details:

(1) **Sample size.** $n = o\big(\sigma_0\sigma_p d/(\log d)^2\big)$, $n = \Omega(\log d)$.
(2) **Width and step size.** $m = \omega(\log d)$, $0 < \eta \le m/2$.
(3) **Initialization.** $\sigma_0(1 + \sigma_p\sqrt{d})\log d = O(1/\mathrm{polylog}(d))$.
(4) **High effective signal-to-noise ratio (SNR).** $n/(\sigma_p^2 d) = \Omega(\mathrm{polylog}(d))$.

Cond. (1) provides enough samples for feature learning while limiting noise interference. Cond. (2) ensures that sufficiently many neurons are active at initialization. Cond. (3) makes the effect of initialization negligible. Cond. (4) ensures that features are learned before noise is memorized.

Feature updates. For each class k, we track the student's easy- and hard-feature knowledge through weight changes along the corresponding feature $\boldsymbol{\mu}_k$ and $\boldsymbol{\nu}_k$:

$$a_k^{(t)} := \frac{1}{m}\sum_{j=1}^{m}\Big(\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\mu}_k\rangle - \langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\mu}_k\rangle\Big), \quad b_k^{(t)} := \frac{1}{m}\sum_{j=1}^{m}\Big(\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\nu}_k\rangle - \langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\nu}_k\rangle\Big).$$

By definition, $a_k^{(0)} = b_k^{(0)} = 0$. For the logistic loss $\ell(u) = \log(1 + e^{-u})$, define the gradient weight $g_i^{(t)} := -\ell'(\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)) = \frac{1}{1+\exp\left(\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)}$. This weight is large when the student's output $F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)$ disagrees with the weak label $\tilde{y}_i$ and small when it agrees confidently.

Lemma 1 (Feature and noise gradient decomposition)**.** *Let $\boldsymbol{v}_i^{(1)}$ and $\boldsymbol{v}_i^{(2)}$ denote the signal patches of sample i, and define $\phi'(u) = \mathbb{1}\{u > 0\}$. For every class $k \in \{\pm 1\}$, neuron j, and iteration t,*

$$-\nabla_{\boldsymbol{w}_{k,j}} \mathcal{L}(\boldsymbol{W}^{(t)}) = \frac{k}{mn}\sum_{i=1}^{n} \tilde{y}_i g_i^{(t)} \Bigg[\underbrace{\sum_{p=1}^{2} \phi'\Big(\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{v}_i^{(p)}\rangle\Big)\boldsymbol{v}_i^{(p)}}_{\textit{feature directions}} + \underbrace{\phi'\Big(\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\xi}_i\rangle\Big)\boldsymbol{\xi}_i}_{\textit{noise direction}}\Bigg].$$

By orthogonality between the feature and noise directions, the updates of $a_k^{(t)}$ and $b_k^{(t)}$ are

$$\begin{aligned} a_k^{(t+1)} - a_k^{(t)} &= \frac{\eta \pi_{e,k}^{(t)}}{mn}\left(2\sum_{i\in\mathcal{E}_k} g_i^{(t)} + \sum_{i\in\mathcal{O}_k} g_i^{(t)}\right), \\ b_k^{(t+1)} - b_k^{(t)} &= \frac{\eta \pi_{h,k}^{(t)}}{mn}\left(\sum_{i\in\mathcal{O}_k} \alpha_i g_i^{(t)} + 2\sum_{i\in\mathcal{H}_k} \alpha_i z_i g_i^{(t)}\right), \end{aligned} \tag{5}$$

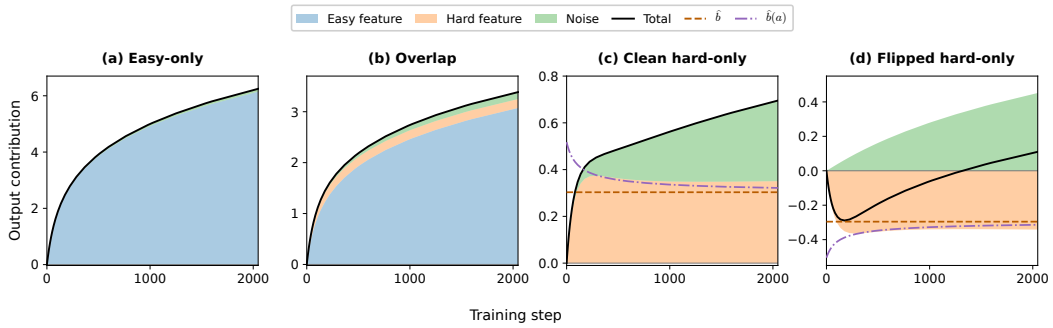


Figure 1: **Feature learning and noise memorization.** The colored regions show the contributions of the easy feature (blue), hard feature (orange), and input noise (green) to the student's output over training for easy-only, overlap, correctly labeled hard-only, and label-flipped hard-only samples. The black curves show their sum. The dashed and dash-dotted reference curves represent the hard-feature saturation points $\widehat{b}$ and $\widehat{b}(a)$, obtained as roots of the functions in (7). See Appx. B.1.1 for details.

where $z_i := y_i\tilde{y}_i$ equals 1 for a correct weak label and -1 for a flipped label. Here, $\pi_{e,k}^{(t)} = \frac{1}{m}\sum_j \phi'(\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\mu}_k \rangle)$ and $\pi_{h,k}^{(t)} = \frac{1}{m}\sum_j \phi'(\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\nu}_k \rangle)$ are the fractions of neurons positively activated by the easy and hard features, respectively. Because easy-only and overlap samples are correctly labeled, both contribute positively to $a_k^{(t)}$. For the hard feature, overlap and correctly labeled hard-only samples increase $b_k^{(t)}$, whereas label-flipped hard-only samples ($z_i = -1$) decrease it.

3.1 Hard-Feature Saturation Under Competing Supervision

Under the high-SNR condition, we define the *noise-free regime* as $t \in [0, T_{\text{noise}}]$, during which input noise contributes at most $O(1/\text{polylog}(d))$ to the student's output on each training sample; see Appx. C.6. We therefore focus on the easy- and hard-feature contributions.

Fix a class k and write $a_t := a_k^{(t)}$ and $b_t := b_k^{(t)}$, omitting the class subscript throughout, including from the sample sets. Because the effect of initialization is also negligible in this regime, a_t and b_t approximate the class-aligned feature responses. The true-label-aligned response $y_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)$ and gradient weight $g_i^{(t)}$ then satisfy

$$y_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) \approx \begin{cases} 2a_t, & i \in \mathcal{E}, \\ a_t + \alpha_i b_t, & i \in \mathcal{O}, \\ 2\alpha_i b_t, & i \in \mathcal{H}, \end{cases} ; \qquad g_i^{(t)} \approx \begin{cases} (1+e^{2a_t})^{-1}, & i \in \mathcal{E}, \\ (1+e^{a_t+\alpha_i b_t})^{-1}, & i \in \mathcal{O}, \\ (1+e^{2\alpha_i z_i b_t})^{-1}, & i \in \mathcal{H}. \end{cases} \tag{6}$$

We first characterize the saturation level of b_t. Using gradient approximations above, define the hard-only and full hard-feature update functions, the latter also accounting for the overlap samples, as

$$\begin{aligned} f_H(b) &:= \sum_{i\in\mathcal{H}:z_i=1} \frac{\alpha_i}{1+e^{2\alpha_i b}} - \sum_{i\in\mathcal{H}:z_i=-1} \frac{\alpha_i}{1+e^{-2\alpha_i b}}, \\ f(a,b) &:= f_H(b) + \frac{1}{2}\sum_{i\in\mathcal{O}} \frac{\alpha_i}{1+e^{a+\alpha_i b}}. \end{aligned} \tag{7}$$

In the noise-free regime, $f(a,b)$ approximates the hard-feature update in (5) up to a constant factor. When both correct and label-flipped hard-only samples are present, we define the *hard-only saturation point* $\widehat{b}$ as the unique root of f_H. Define the correct-to-flipped hard-feature-strength ratio as

$$R := \frac{\sum_{i\in\mathcal{H}:z_i=1}\alpha_i}{\sum_{i\in\mathcal{H}:z_i=-1}\alpha_i}. \tag{8}$$

Under our teacher model, any fixed $b_h > 0$ makes samples with larger hard-feature strength α more likely to be labeled correctly. Thus, with sufficiently many samples, correctly labeled hard-only samples have greater total feature strength than label-flipped samples, yielding $R > 1$ with high probability. When $b_h = 0$, the two groups have equal expected total feature strength, so $R \approx 1$.

Lemma 2 (Hard-only saturation point)**.** *If $R > 1$, the hard-only saturation point $\widehat{b}$ satisfies*

$$\frac{\log R}{2\alpha_{\max}} \le \widehat{b} \le \frac{\log R}{2\alpha_{\min}}. \tag{9}$$

Correctly labeled hard-only samples increase hard feature knowledge, while label-flipped samples decrease it. The ratio R compares the total feature strengths of these two groups and bounds the saturation point at which their updates balance.

How overlap samples shift the balance. An overlap sample contains both easy and hard features and has approximate margin $a_t + \alpha_i b_t$, as shown in (6). Its correct label initially supports learning both features, but its contribution to hard-feature learning diminishes as a_t grows. For fixed a, define the *full saturation point* $\widehat{b}(a)$ as the root of $f(a, \cdot)$. The following lemma characterizes the diminishing influence of overlap samples, consistent with the experiments in Fig. 5 of Appx. B.3.

Lemma 3 (Full saturation point)**.** *For each $a \ge 0$, the full saturation point $\widehat{b}(a)$ satisfies*

$$\widehat{b}(a) \ge \widehat{b}, \qquad \widehat{b}(a) \downarrow \widehat{b} \quad (a \to \infty).$$

Equality holds when $\mathcal{O} = \varnothing$.

Remark 1 (Relation to overlap-based W2S)**.** Prior work explains W2S through correctly labeled overlap samples, which enables the student to learn hard features unknown to the teacher (Shin et al., 2025, Oh et al., 2025). This setting corresponds to $b_h = 0$ in our model. If correctly labeled and label-flipped hard-only samples have equal total feature strength, then $R = 1$ and the hard-only saturation point is $\widehat{b} = 0$. Hard-feature learning therefore relies entirely on overlap samples. However, Lemma 3 shows that their contribution is transient: as the easy-feature response a grows, the full saturation point $\widehat{b}(a) \downarrow 0$. This explains why early stopping is essential in prior works, consistent with the initial hard-feature gains and subsequent degradation observed by Oh et al. (2025). In contrast, our model allows the teacher to have partial hard-feature knowledge, with $b_h > 0$. Correct labels then concentrate on hard-only samples with stronger hard features, yielding $R > 1$ and a positive hard-only saturation $\widehat{b} > 0$. Thus, the student can learn the hard feature without overlap samples and retain this knowledge after overlap-driven updates vanish, as shown in subsequent sections.

3.2 Feature Learning Dynamics in the Noise-Free Regime

The following result characterizes feature learning before noise memorization begins.

Theorem 1 (Feature learning in the noise-free regime)**.** *With probability at least $1 - 1/\mathrm{poly}(d)$, for both classes and all $t \in [0, T_{\mathrm{noise}}]$,*

$$a_t \asymp \log(1 + \eta t/m),$$
$$|b_t - \widehat{b}| \le \underbrace{C b_h e^{-c\eta t/m}}_{\textit{incomplete learning}} + \underbrace{\frac{C p_O}{1 + c p_O \eta t/m}}_{\textit{overlap contribution}} + \underbrace{O\left(\frac{1}{\mathrm{polylog}(d)}\right)}_{\textit{initialization and noise}}. \tag{10}$$

Here $c, C > 0$ are fixed constants, and $\widehat{b}$ is the hard-only saturation point in (9)*.*

Because easy-only and overlap samples have correct weak labels, they provide positive easy-feature updates, and hence a_t increases monotonically. Hard-feature updates instead compete: correctly labeled samples increase b_t, whereas label-flipped samples decrease it. As the easy feature fits the overlap samples, their contribution to hard-feature learning diminishes. Consequently, the gap between b_t and $\widehat{b}$ is governed by incomplete learning and the remaining overlap contribution, together with negligible initialization and noise errors. If the noise-free regime lasts long enough that $\eta t/m \to \infty$, then b_t approaches $\widehat{b}$, where the updates from correctly labeled and label-flipped hard-only samples balance.

Fig. 1 illustrates how these dynamics shape the student's output across the four sample groups when $R > 1$, so that $\widehat{b} > 0$. For easy-only and overlap samples, the growing easy-feature response eventually dominates the output. For correctly labeled hard-only samples, the saturated hard-feature response supports the given label. For label-flipped hard-only samples, however, it opposes the given label, maintaining a larger gradient signal that drives subsequent noise memorization. In the next section, we analyze this later phase and quantify how much hard-feature knowledge survives.

3.3 Noise Memorization and Feature Preservation

We next quantify how much hard-feature knowledge the student retains while memorizing noise. For each training sample i, let $c_i^{(t)}$ track the extent to which its noise patch $\boldsymbol{\xi}_i$ is memorized in the weights of its weak-label class $\tilde{y}_i$. With $c_i^{(0)} = 0$, its dynamics satisfy

$$c_i^{(t)} \approx \frac{1}{m}\sum_{j=1}^{m}\left(\langle \boldsymbol{w}_{\tilde{y}_i,j}^{(t)}, \boldsymbol{\xi}_i\rangle - \langle \boldsymbol{w}_{\tilde{y}_i,j}^{(0)}, \boldsymbol{\xi}_i\rangle\right), \quad c_i^{(t+1)} - c_i^{(t)} = \frac{\eta\|\boldsymbol{\xi}_i\|^2\pi_i^{(t)}}{mn} g_i^{(t)}.$$

Here, $\pi_i^{(t)} := \frac{1}{m}\sum_j \phi'(\langle \boldsymbol{w}_{\tilde{y}_i,j}^{(t)}, \boldsymbol{\xi}_i\rangle)$ is the fraction of neurons in the weak-label class activated by $\boldsymbol{\xi}_i$. The approximation neglects small interference from other samples' noise patches.

For easy-only and overlap samples, the logarithmic growth of a_t in (10) during the noise-free regime yields $g_i^{(t)} = O((1+\eta t/m)^{-1})$. In contrast, hard-only samples retain nonvanishing gradient weights while b_t remains near $\widehat{b}$ and their noise has not yet been memorized. Among hard-only samples with comparable noise power $\|\boldsymbol{\xi}_i\|^2$, the learned hard feature reduces the gradient weight of correctly labeled samples but increases that of label-flipped samples, causing the latter to memorize noise faster. This faster memorization subsequently suppresses their opposing hard-feature updates, preventing the student from losing all its acquired hard-feature knowledge, as illustrated in Fig. 1.(c)–(d). Theorem 2 formalizes this effect by lower-bounding b_t through the end of training.

Theorem 2 (Hard feature preservation)**.** *With probability at least* $1 - 1/\mathrm{poly}(d)$*, for both classes and all* $t \in [T_{\mathrm{noise}}, T]$*,*

$$b_t \geq c\widehat{b}\big(1 - e^{-c'\eta t/m}\big). \tag{11}$$

Here $c, c' > 0$ *are fixed constants.*

Theorem 2 shows that the student retains a positive fraction of $\widehat{b}$ during noise memorization. The scale of this retained knowledge is determined by the correct-to-flipped hard-feature-strength ratio R.

Why weak-to-strong generalization works. By Def. 2, the teacher's output noise ϵ can flip hard-only labels, but its hard-feature knowledge $b_h > 0$ creates a feature-strength-weighted bias toward the true class. With enough samples, this bias yields $R > 1$, so correctly labeled samples initially provide the larger aggregate update and drive b_t toward $\widehat{b}$. Near $\widehat{b}$, label-flipped samples have larger gradient weights and thus memorize noise faster, attenuating their opposing hard-feature updates. This preserves a positive fraction of the learned hard feature, as shown in Theorem 2. On fresh samples, the learned features generalize, whereas memorized noise has little effect due to high-dimensional orthogonality. The following theorem shows that the student can therefore outperform its teacher even after fitting all weak training labels.

Theorem 3 (Weak-to-strong generalization)**.** *With probability at least* $1 - 1/\mathrm{poly}(d)$*, the final student satisfies*

$$\tilde{y}_i F_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}_i) > 0, \qquad i \in [n],$$
$$\Pr_{(\boldsymbol{X},y)}\left(yF_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}) < 0\right) \leq \frac{1}{\mathrm{poly}(d)}.$$

Theorem 3 shows that the student achieves nearly perfect generalization despite the teacher's nonzero error on hard-only samples. When $b_h > 0$, the student preserves hard-feature knowledge while fitting incorrect weak labels through noise memorization. The following corollary considers $b_h = 0$, when the teacher has no hard-feature knowledge.

Corollary 1 (Zero hard-feature knowledge)**.** *Suppose* $b_h = 0$ *and the conditions in Appendix C.13 hold, including* $R = 1$ *for both classes. Then, for both classes and* $t \in [T_{\mathrm{noise}}, T]$*,*

$$|b_t| = O\left(\frac{1}{\log d}\right).$$

When $b_h = 0$ and $R = 1$, the hard-only saturation point is $\widehat{b} = 0$. Overlap samples initially support hard-feature learning, but their contribution decreases as the easy feature is learned. The competing updates from correct and flipped hard-only samples keep b_t near zero throughout noise memorization.

4 EXPERIMENTS

We examine how the correct-to-flipped hard-feature-strength ratio R in (8) shapes hard-feature learning and generalization in experiments on synthetic and real datasets. We vary R by changing which hard-only samples receive correct labels while holding the mean weak-label accuracy fixed. We also show that hard-feature knowledge, as captured by R, can be a better criterion for teacher selection than the weak model's label accuracy. App. B provides further details and additional results.

4.1 EFFECT OF R IN WEAK SUPERVISION FOR A RELU STUDENT

We first test whether correct labels are more useful when assigned to samples with stronger hard features. We use the ReLU student and three sample types from Sec. 2, with correct labels for all easy-only and overlap samples. For each hard-only label accuracy of 40%, 50%, and 60%, we construct three weak-label assignments by selecting hard-only samples with the largest α, a random subset, or the smallest α to receive correct labels. These assignments produce different values of R while keeping the inputs, number of correct labels, and student initialization fixed.

We evaluate the student's hard-feature knowledge, quantified by its response to a unit-strength hard feature, in Fig. 2.(a), as well as its hard-only and overall test accuracies in Fig. 2.(b)–(c), respectively. Assigning correct labels to samples with larger α yields a larger hard-feature response and higher test accuracy. For example, at 50% label accuracy, changing which half of the hard-only samples receives correct labels can even reverse the sign of the learned hard-feature response, substantially altering both hard-only and overall accuracy.

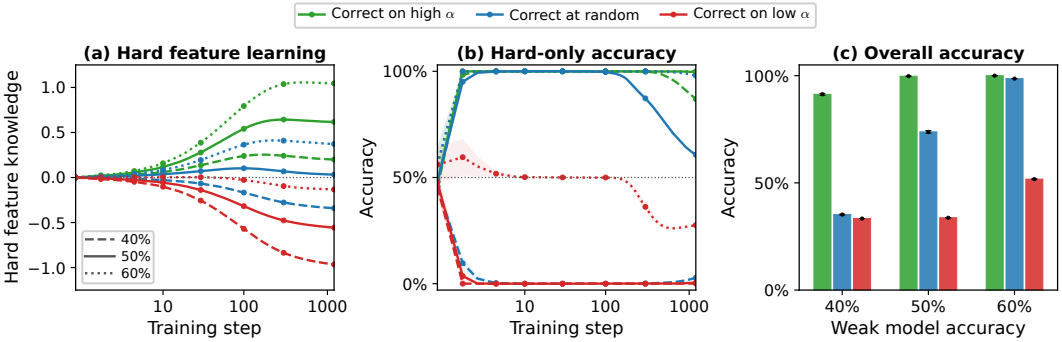


Figure 2: **Same label accuracy, different hard-feature learning.** (a) Student response to a unit hard feature and (b) hard-only test accuracy over training; (c) final overall test accuracy. Colors indicate correct label assignments to high-α, random, or low-α hard-only samples; line styles indicate hard-only weak-label accuracy. Shaded bands and error bars show one standard error over six seeds.

4.2 EFFECT OF R IN WEAK SUPERVISION FOR FINE-TUNING LARGE LANGUAGE MODELS

We test whether sentiment classification on IMDb, SST-2, Rotten Tomatoes (RT), and Yelp exhibits a similar dependence on R. We use four small Qwen teachers and three larger students from the Gemma, Llama, and Phi families, yielding 12 teacher–student combinations. To estimate feature directions, we fine-tune weak and strong reference models on separate sets of 1K true-labeled examples. From the strong model's hidden representations, we estimate the easy-feature direction using samples on which the weak model is highly confident. After removing this direction, we identify a residual hard-feature direction aligned with the true labels and use both directions to categorize samples; see Appx. B.1.2 for details. From unused samples, we construct a 5K-example training set with 4K non-hard-only samples and 1K hard-only samples, split equally between correct and incorrect weak labels. Holding hard-only label accuracy at 50%, we assign correct labels to either high- or low-α samples, where α is estimated from the strong reference model's hard-feature output.

Using the estimated feature strengths and weak labels, we compute the predicted saturation point $\widehat{b}$ and compare it with the student's learned hard-feature knowledge b_T. Fig. 3.(a) shows a mean Spearman correlation of 0.91 between $\widehat{b}$ and b_T, demonstrating that the predicted saturation point closely tracks the hard-feature knowledge acquired through gradient-based training. As shown in

Fig. 3.(b)–(c), larger R also yields higher hard-only and overall test accuracy on average. Assigning correct labels to high-α rather than low-α samples improves these accuracies by 9.99 and 2.27 percentage points, respectively.

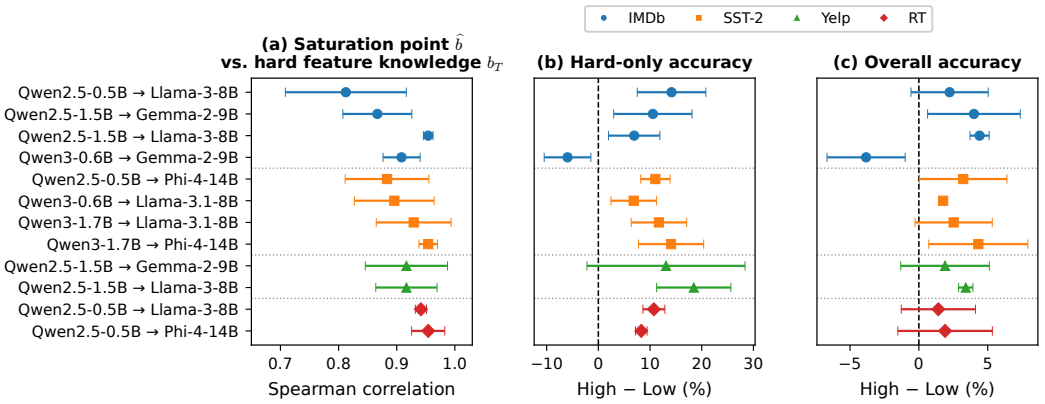


Figure 3: **Hard feature knowledge and generalization in LLMs.** Each row represents a dataset–teacher–student combination. (a) Spearman correlation between the predicted saturation point $\widehat{b}$ and the student's hard feature knowledge b_T. (b,c) High–Low differences in hard-only and overall test accuracy, resp., in percentage points. High and Low assign correct labels to hard-only samples with the largest and smallest α, respectively. Points and error bars show the mean and one std. over 4 runs.

4.3 Teacher Selection Using R

We next evaluate whether R can identify the weak teacher that yields the largest student generalization gain. We consider four Qwen teachers (Q2.5-0.5B/1.5B and Q3-0.6B/1.7B) and two Qwen students (Q2.5-7B and Q3-8B) across four sentiment-classification datasets. We compare selection by R with selection by teacher validation accuracy. To compute R, we estimate hard-feature strengths using reference models, following Sec. 4.2. Both criteria use the same validation set.

Across $4 \times 2 \times 8 = 64$ dataset–student–seed combinations, the two criteria select the same teacher in 56 cases. In the eight cases where they differ, selection by R yields higher student test accuracy in seven and equal accuracy in one. Table 4 (Appendix D.1) summarizes these cases, with full results in Table 1. These findings support R as a practical teacher-selection criterion that captures differences in downstream supervision quality missed by validation accuracy alone.

Table 1: **Comparing teacher-selection criteria.** Overall student test accuracy when selecting the teacher with the highest validation accuracy vs. the largest hard-feature-strength ratio R.

Dataset	Different teachers	Student test accuracy (%)		Gain (pp)
		Validation accuracy	R	
IMDb	1/16	91.96	92.08	+0.115
SST-2	0/16	92.29	92.29	+0.000
RT	2/16	91.03	91.29	+0.255
Yelp	5/16	95.74	95.89	+0.153
Overall	**8/64**	**92.75**	**92.88**	**+0.131**

5 Conclusion

We analyzed W2S under partial teacher knowledge of hard features. The feature-strength ratio between correctly and incorrectly labeled hard-only samples determines the student's early hard-feature saturation level, a positive fraction of which survives subsequent noise memorization. Under our assumptions, the student can therefore fit all weak labels while outperforming the teacher, even without overlap samples. Experiments with ReLU networks and LLMs show that hard-feature information shapes student learning at fixed label accuracy and provides a practical criterion for teacher selection beyond validation accuracy. See Appx. A for limitations and further discussion.

AI USE STATEMENT

Generative AI assisted with literature searches, manuscript editing, proof development and checking, experimental code implementation, synthetic data generation, figure preparation, and result interpretation. The authors reviewed the AI-assisted work and take full responsibility for the manuscript, code, and results.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. URL https://arxiv.org/abs/2412.08905.

AI@Meta. Llama 3 model card, 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Ryoya Awano and Taiji Suzuki. The mechanism of weak-to-strong generalization: Feature elicitation from latent knowledge. *arXiv preprint arXiv:2605.12908*, 2026. URL https://arxiv.org/abs/2605.12908.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ghNRg2mEgN.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126474–126499, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Yue Guo and Yi Yang. Improving weak-to-strong generalization with reliability-aware alignment. *arXiv preprint arXiv:2406.19032*, 2024. URL https://arxiv.org/abs/2406.19032.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

M. Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/43c18853329c7504996b255252b6cb1f-Abstract-Conference.html.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer ReLU convolutional neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17615–17659. PMLR, 2023. URL https://proceedings.mlr.press/v202/kou23a.html.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=HOSh0SKklE.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150. Association for Computational Linguistics, 2011. URL https://aclanthology.org/P11-1015/.

Marko Medvedev, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. Weak-to-strong generalization even in random feature networks, provably. In *Forty-second International Conference on Machine Learning*, pp. 43519–43556, 2025. URL https://openreview.net/forum?id=OUzDIhgiqr.

Behrad Moniri and Hamed Hassani. On the mechanisms of weak-to-strong generalization: A theoretical perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=qR5B1Es9Ty.

Abhijeet Mulgund and Chirag Pabbaraju. Relating misfit to gain in weak-to-strong generalization beyond the squared loss. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=wITGhlMPXz.

Junsoo Oh, Jerry Song, and Chulhee Yun. From linear to nonlinear: Provable weak-to-strong generalization through feature learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/f7977daf15f87e19b56e4d2d363b9002-Abstract-Conference.html.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 115–124. Association for Computational Linguistics, 2005. doi: 10.3115/1219840.1219855. URL https://aclanthology.org/P05-1015/.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Changho Shin, John Cooper, and Frederic Sala. Weak-to-strong generalization through the data-centric lens. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uogG8BfLs2.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, 2013. URL https://aclanthology.org/D13-1170/.

Seamus Somerstep, Felipe Maia Polo, Moulinath Banerjee, Ya'acov Ritov, Mikhail Yurochkin, and Yuekai Sun. A transfer learning framework for weak to strong generalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fa47ffc7b2033fd32d8b8734686f6ff4-Abstract-Conference.html.

Arda Uzunoğlu, Alvin Zhang, and Daniel Khashabi. Trust functions: Near lossless weak-to-strong generalization by learning to trust the weak teacher. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.01000.

David X. Wu and Anant Sahai. Provable weak-to-strong generalization via benign overfitting. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4vzGQcVUG8.

Yihao Xue, Jiping Li, and Baharan Mirzasoleiman. Representations shape weak-to-strong generalization: Theoretical insights and empirical predictions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ypEW077kle.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

A LIMITATIONS

We use a two-layer ReLU network for our theory and controlled experiments so that we can analyze feature learning and noise memorization explicitly. The model has one easy and one hard feature per class, with mutually orthogonal feature directions and input noise orthogonal to all features. Real models can learn many correlated features that interact across layers and change during training. These interactions can affect which features are learned and preserved, and our analysis does not cover this richer setting. Extending the theory to more complex feature structures and deeper architectures remains an open direction.

Our theory assumes the teacher's response without modeling how it arises from training. We complement this analysis with W2S experiments using trained language models. In these experiments, we estimate the hard feature direction from the representations of reference models. Appendix B.4 examines the stability of the resulting feature estimates.

B EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

B.1 EXPERIMENTAL DETAILS

B.1.1 DETAILS OF RELU EXPERIMENTS

Feature learning and noise fitting. For Fig. 1, we train the two-layer ReLU network defined in (3) with $d = 128$, $m = 32$, $\eta = 8$, $\sigma_0 = 10^{-4}$, $\sigma_p = 0.2$, and $n = 1{,}536$ for 2,048 full-batch gradient descent steps. We set $(p_E, p_O, p_H) = (1/6, 1/6, 2/3)$. The hard feature strength α is independently sampled from a uniform distribution on $[0.25, 0.75]$. We generate weak labels once using $b_e = 4$, $b_h = 0.2$, and independent output noise $\epsilon \sim \mathcal{N}(0, 1)$ in (2).

Fig. 1(a)–(d) show easy-only, overlap, correctly labeled hard-only, and flipped hard-only samples, respectively. We decompose $\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)$ into contributions from the easy feature, hard feature, and input noise. We average within each class and sample type, then equally over classes and seeds. Shaded regions show these contributions, and black curves show their sum. In Fig. 1(c) and (d), we numerically compute $\widehat{b}$ and $\widehat{b}(a)$ from (7) in Sec. 3.1.

Same accuracy, different feature learning. For Fig. 2, we train the two-layer ReLU network defined in (3) with $d = 64$, $m = 32$, $\eta = 10$, $\sigma_0 = 0.02$, $\sigma_p = 0.08$, and $n = 768$ for 1,200 full-batch gradient descent steps. We set $(p_E, p_O, p_H) = (1/6, 1/6, 2/3)$. The hard feature strength α is independently sampled from a uniform distribution on $[0.10, 0.90]$ or $[0.25, 0.75]$. For the high-α and low-α allocations, we assign correct labels to hard-only samples with the largest and smallest α values, respectively.

For evaluation, we use 512 easy-only, 512 overlap, and 2,048 hard-only samples per class. We measure a_t, b_t by the increase from initialization in the mean projections of class-k weights onto their easy and hard feature directions.

B.1.2 DETAILS OF LANGUAGE MODEL EXPERIMENTS

Datasets and models. We use IMDb (Maas et al., 2011), SST-2 (Socher et al., 2013), Rotten Tomatoes (RT) (Pang & Lee, 2005), and Yelp Polarity (Zhang et al., 2015) for sentiment classification. The allocation experiments use Qwen2.5-0.5B/1.5B (Qwen Team, 2024) and Qwen3-0.6B/1.7B (Qwen Team, 2025) teachers, with Gemma-2-9B (Gemma Team et al., 2024), Llama-3-8B (AI@Meta, 2024), Llama-3.1-8B (Grattafiori et al., 2024), and Phi-4-14B (Abdin et al., 2024) students.

Estimating feature knowledge. We estimate feature directions from the final hidden layer at the last non-padding token of the prompt. For each sample x, let $\phi_{\text{weak}}(x)$ and $\phi_{\text{strong}}(x)$ denote this normalized representation in the weak and strong reference models, respectively. Using the weak reference model's training examples, we define $\widetilde{\boldsymbol{\mu}}_+$ and $\widetilde{\boldsymbol{\mu}}_-$ as the representative directions of the positive and negative classes, respectively:

$$\widetilde{\boldsymbol{\mu}}_+ = \frac{\sum_{i:y_i=+1} \phi_{\text{weak}}(x_i)}{\left\| \sum_{i:y_i=+1} \phi_{\text{weak}}(x_i) \right\|}, \qquad \widetilde{\boldsymbol{\mu}}_- = \frac{\sum_{i:y_i=-1} \phi_{\text{weak}}(x_i)}{\left\| \sum_{i:y_i=-1} \phi_{\text{weak}}(x_i) \right\|}.$$

For easy feature knowledge, we compare a sample's alignment with the two class representatives:

$$s_{\text{easy}}(x, y) = \frac{2y}{\|\widetilde{\boldsymbol{\mu}}_+ - \widetilde{\boldsymbol{\mu}}_-\|} \left(\langle \phi_{\text{weak}}(x), \widetilde{\boldsymbol{\mu}}_+ \rangle - \langle \phi_{\text{weak}}(x), \widetilde{\boldsymbol{\mu}}_- \rangle \right).$$

A large $s_{\text{easy}}(x, y)$ indicates stronger alignment with the true class mean relative to the other class mean, so we treat such samples as containing a stronger easy feature. We set the easy threshold τ_e to the 25th percentile of s_{easy} over the weak reference model's training examples, pooling both classes.

For hard feature knowledge, we first estimate the easy feature directions in the strong model's hidden representations. We select the weak-reference training samples with $s_{\text{easy}} \geq \tau_e$ and compute their normalized class means in the strong reference representation. Applying QR decomposition to the difference of these means gives an orthonormal basis $\boldsymbol{Q} \in \mathbb{R}^{d_s \times 1}$ for the easy feature direction, where d_s is the strong model's hidden dimension. We remove the easy subspace and normalize the remaining representation:

$$\widetilde{\phi}_{\text{strong}}(x) = \frac{(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\top)\phi_{\text{strong}}(x)}{\|(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\top)\phi_{\text{strong}}(x)\|}.$$

To estimate the hard feature directions, we select strong-reference training samples with $s_{\text{easy}}(x, y) < \tau_e$ to exclude easy-only samples. Let $\mathcal{I}_+$ and $\mathcal{I}_-$ be the indices of the selected samples with true labels $+1$ and -1, respectively. We then estimate the hard feature directions $\widetilde{\boldsymbol{\nu}}_+$ and $\widetilde{\boldsymbol{\nu}}_-$ as

$$\widetilde{\boldsymbol{\nu}}_+ = \frac{\sum_{i \in \mathcal{I}_+} \widetilde{\phi}_{\text{strong}}(x_i)}{\left\| \sum_{i \in \mathcal{I}_+} \widetilde{\phi}_{\text{strong}}(x_i) \right\|}, \qquad \widetilde{\boldsymbol{\nu}}_- = \frac{\sum_{i \in \mathcal{I}_-} \widetilde{\phi}_{\text{strong}}(x_i)}{\left\| \sum_{i \in \mathcal{I}_-} \widetilde{\phi}_{\text{strong}}(x_i) \right\|}.$$

We measure hard feature knowledge by

$$s_{\text{hard}}(x, y) = \frac{2y}{\|\widetilde{\boldsymbol{\nu}}_+ - \widetilde{\boldsymbol{\nu}}_-\|} \left(\langle \widetilde{\phi}_{\text{strong}}(x), \widetilde{\boldsymbol{\nu}}_+ \rangle - \langle \widetilde{\phi}_{\text{strong}}(x), \widetilde{\boldsymbol{\nu}}_- \rangle \right).$$

We set the hard threshold τ_h to the 25th percentile of s_{hard} over all strong-reference training examples, pooling both classes. A sample has the estimated easy feature if $s_{\text{easy}} \geq \tau_e$ and the estimated hard feature if $s_{\text{hard}} \geq \tau_h$; these two tests define easy-only, overlap, hard-only, and neither-feature samples. In particular, hard-only samples satisfy $s_{\text{easy}} < \tau_e$ and $s_{\text{hard}} \geq \tau_h$. We define their hard feature strength by

$$\alpha_i = \max\{s_{\text{hard}}(x_i, y_i) - \tau_h, 0\}.$$

Thus α_i is nonnegative and is positive above the hard threshold; a sample exactly at the threshold has zero strength and contributes zero to the strength-weighted sums. The reference models, feature directions, and thresholds remain fixed during student training.

Assigning correct labels. For the High–Low comparison in Fig. 3, we vary which hard-only training samples receive correct labels while keeping hard-only label accuracy at 50%. We form the 5,000-example training set from 4,000 non-hard-only samples and 1,000 hard-only samples, including 500 originally labeled correctly and 500 originally labeled incorrectly by the weak reference model, which also serves as the teacher. Among the 1,000 hard-only training samples, High assigns correct labels to the 500 samples with the largest α_i, while Low assigns correct labels to the 500 samples with the smallest α_i. This assignment pools both true classes; the non-hard-only samples keep their original weak labels. We separately draw a 256-example hard-only evaluation set with 128 originally correct and 128 originally incorrect teacher labels, and draw the overall evaluation set from the remaining feature pool. This set is used only to measure hard-only accuracy and b_T using true labels. Within each seed, all allocation levels share these splits, and we report the student after five epochs. We repeat the experiment with different training and evaluation samples drawn for each seed.

Hard-only saturation point and estimated hard feature knowledge. Using the strengths α_i defined above, we pool the hard-only training samples into $\mathcal{H}$ and write $z_i = y_i \tilde{y}_i$. The hard-only saturation point $\widehat{b}$ is the root of the hard feature update function f_H in (7) and satisfies

$$\sum_{i \in \mathcal{H}: z_i = 1} \frac{\alpha_i}{1 + e^{2\alpha_i \widehat{b}}} = \sum_{i \in \mathcal{H}: z_i = -1} \frac{\alpha_i}{1 + e^{-2\alpha_i \widehat{b}}}.$$

For a text sample x, let $p_T(+ \mid x)$ and $p_T(- \mid x)$ be the final student's next-token probabilities, normalized over the first tokens of `Positive` and `Negative` at the end of the prompt. We define its output margin by

$$F_T(x) = \log \max\{p_T(+ \mid x), 10^{-12}\} - \log \max\{p_T(- \mid x), 10^{-12}\}.$$

On the hard-only evaluation set $\mathcal{P}$, we estimate hard feature knowledge by fitting the true-label margin to $2\alpha_i$ with a separate intercept for each class:

$$(a_{T,k}, b_{T,k}) = \arg\min_{a,b} \sum_{i \in \mathcal{P}: y_i = k} \big(y_i F_T(x_i) - a - 2\alpha_i b\big)^2, \qquad k \in \{\pm 1\},$$

$$b_T = \frac{b_{T,+1} + b_{T,-1}}{2}.$$

Here b_T is an empirical proxy for hard feature knowledge.

B.1.3 Teacher Selection

Candidates and reference features. We compare Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen3-0.6B, and Qwen3-1.7B as teachers for Qwen2.5-7B and Qwen3-8B students on IMDb, SST-2, Rotten Tomatoes, and Yelp. We report eight seeds for each dataset–student combination, giving 64 teacher-selection comparisons and 256 candidate-trained students. For the Qwen2.5-7B student, we use Qwen2.5-3B/Qwen2.5-7B as the weak/strong reference pair; for Qwen3-8B, we use Qwen3-4B/Qwen3-8B. Using these reference models, we identify hard-only samples and estimate their hard feature strengths to compare all four candidate teachers on the same validation set. Candidate teachers are fixed adapters trained on 1,000 true-labeled examples and are reused across students and seeds.

Validation scores and selection. We use α_i from Appendix B.1.2. For each dataset–student–seed combination, we select a hard-only validation set $\mathcal{V}_H$ of 500 samples, with 250 from each true class. After excluding these validation samples, we draw 5,000 student-training examples and a disjoint evaluation set of 5,000 examples (1,030 for RT). Each candidate labels the same student-training inputs, and we train a fresh student for three epochs. For a teacher T, we define its validation accuracy $A(T)$ and correct-to-flipped hard-feature-strength ratio $R(T)$ on this set as

$$A(T) = \frac{1}{|\mathcal{V}_H|} \sum_{i \in \mathcal{V}_H} \mathbf{1}\{T(x_i) = y_i\}, \qquad R(T) = \frac{\sum_{i \in \mathcal{V}_H : T(x_i) = y_i} \alpha_i}{\sum_{i \in \mathcal{V}_H : T(x_i) \neq y_i} \alpha_i}.$$

Maximizing $A(T)$ selects the teacher with the highest validation accuracy, while maximizing $R(T)$ selects the one whose correctly labeled samples contain the largest total hard feature strength.

B.2 Hard Feature Saturation under Different Label Rules

We test the hard-only saturation equation under the five label rules in Table 2. These rules allow the probability of a correct weak label to increase, decrease, or vary nonmonotonically with the hard feature strength α.

For Fig. 4, we train the two-layer ReLU network in (3) with $d = 128$, $m = 64$, $\eta = 10$, $\sigma_0 = 0.02$, and $\sigma_p = 0.08$ for 1,800 full-batch updates. We use 64 or 256 hard-only samples per class, sample α uniformly from $[0.1, 0.9]$, and generate weak labels once using each rule. We repeat each setting with ten seeds. For each class y, we compute $\widehat{b}$ from $f_H(\widehat{b}) = 0$ in (7) and obtain the hard feature knowledge as

$$b_T = F(\boldsymbol{\nu}_y, y) - F(\boldsymbol{\nu}_y, -y),$$

Fig. 4 shows close agreement between $\widehat{b}$ and the hard feature knowledge b_T across the five rules, with a median absolute error of 0.0082. The network approaches the predicted saturation point even when stronger hard features are not more likely to receive correct labels, suggesting that the update equation describes feature learning beyond the teacher model assumed in our analysis.

Table 2: **Hard-only label rules.** Each rule applies to both classes, with $\alpha \in [0.1, 0.9]$.

Rule	$\Pr(\tilde{y} = y \mid \alpha, \tau = \mathrm{H})$
logistic	$(1 + e^{-1.1\alpha})^{-1}$
decreasing	$0.78 - 0.42\alpha$
U-shaped	$0.45 + 1.25(\alpha - 0.55)^2$
inverted U	$0.72 - 0.72(\alpha - 0.55)^2$
step	$\begin{cases} 0.76, & \alpha < 0.35; \\ 0.43, & 0.35 \le \alpha < 0.72; \\ 0.69, & \alpha \ge 0.72. \end{cases}$

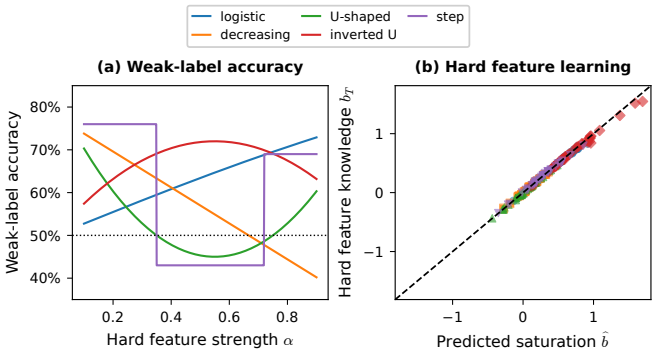


Figure 4: **Hard feature saturation under different label rules.** (a) Correct-label probability under the rules in Table 2. (b) Predicted saturation $\widehat{b}$ versus hard feature knowledge b_T. Each point represents one class; the dashed line marks equality.

B.3 Effect of Overlap Samples on Feature Learning

We fix 32 easy-only and 128 hard-only samples per class and increase the number of overlap samples per class from 0 to 32 and 128. For Fig. 5, we train the two-layer ReLU network in (3) with $d = 8{,}192$, $m = 32$, $\eta = 8$, $\sigma_0 = 10^{-4}$, and $\sigma_p = 0$ for 16,384 full-batch updates. We sample α uniformly from $[0.25, 0.75]$ and generate weak labels using $b_e = 4$, $b_h = 0.2$, and $\epsilon \sim \mathcal{N}(0, 1)$ in (2). We use 17 seeds, keeping the easy-only and hard-only samples, their weak labels, and the initialization fixed across overlap counts within each seed. We measure a_t and b_t as in Appendix B.1.1.

Adding overlap samples increases easy feature learning and initially boosts hard feature learning. These results illustrate that overlap samples can accelerate early hard feature learning, while their influence diminishes as the easy feature is learned.

B.4 Stability of Feature Estimation

We examine the variability of the results when using different samples for feature estimation. For Fig. 6, we use fixed reference-model checkpoints for the eight IMDb and SST-2 model pairs. Following Appendix B.1.2, we estimate features using 500, 1,000, or 2,000 samples per model. We repeat each setting ten times, drawing new samples within each true class.

We evaluate all estimated easy and hard features on the same validation set of 10,000 samples. In Fig. 6(a), we report the Spearman rank correlations of s_{easy} and s_{hard}, each compared with a fixed

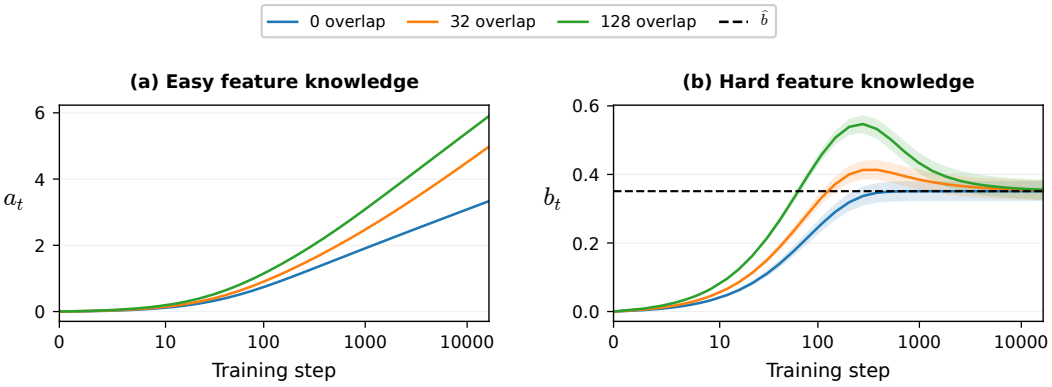


Figure 5: **Feature learning with increasing overlap.** (a) and (b) show easy and hard feature knowledge with 32 easy-only and 128 hard-only samples per class and varying overlap counts. Curves show means across 17 seeds, and shading shows one standard error. The dashed line marks the mean hard-only saturation point $\widehat{b}$.

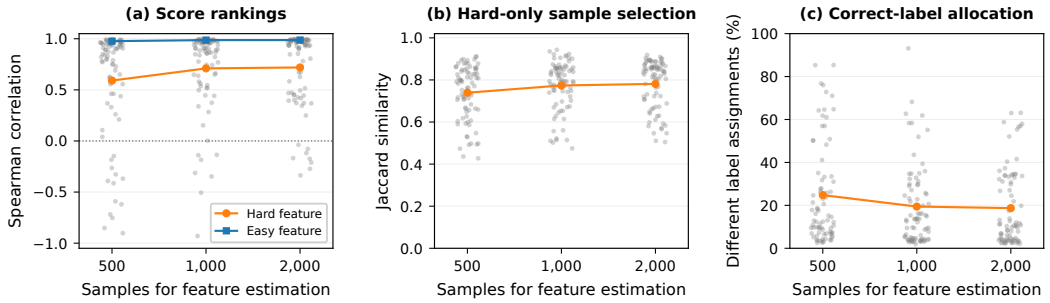


Figure 6: **Stability of feature estimation.** (a) shows rank correlations for the estimated easy and hard features, (b) shows the similarity of hard-only sample sets, and (c) shows the fraction of samples with different correct-label assignments. Points show individual subsamples, and lines show means across eight model pairs.

reference estimate. In Fig. 6(b), we compare the overlap between the hard-only sample sets using Jaccard similarity, the size of their intersection divided by the size of their union. In Fig. 6(c), we compare the correct-label assignments obtained by selecting the stronger half within each class and report the fraction of samples whose assignments differ. We average over the ten repetitions within each model pair and then over the eight model pairs equally.

Increasing the estimation sample size from 500 to 2,000 raises the mean hard feature rank correlation from 0.591 to 0.719, while the mean easy feature rank correlation exceeds 0.976 at every sample size. The mean hard-only set similarity increases from 0.739 to 0.781, and the fraction of samples receiving different label assignments decreases from 24.79% to 18.71%. Using more samples to estimate hard features improves the consistency of feature rankings and label assignments on average.

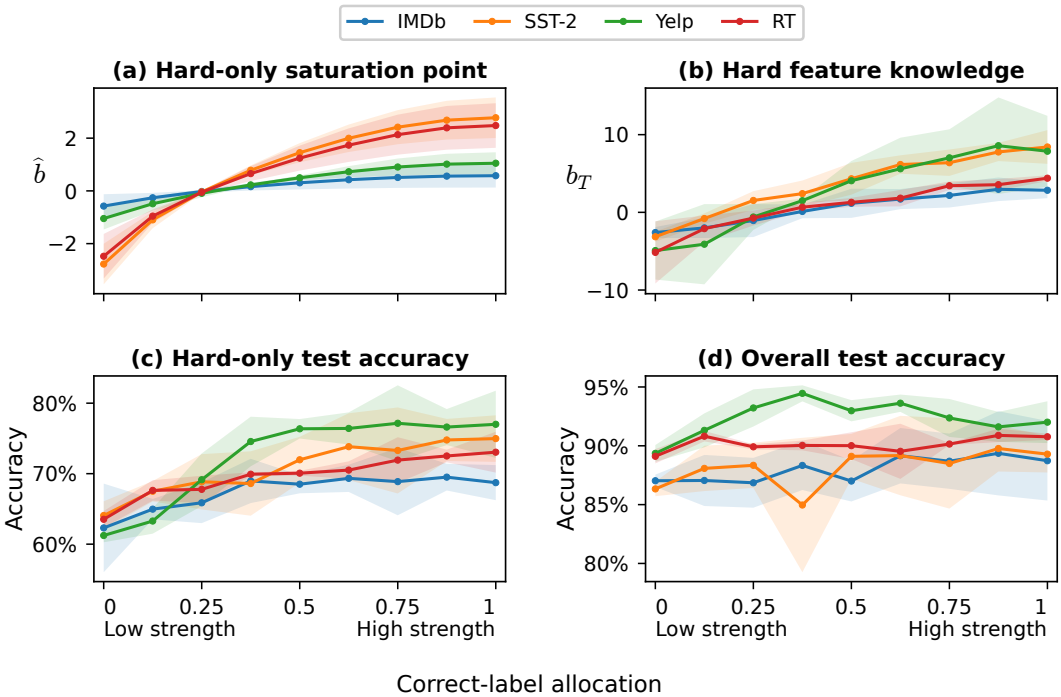


Figure 7: **Correct-label allocation across datasets.** (a) and (b) show the hard-only saturation point $\hat{b}$ and estimated hard feature knowledge b_T. (c) and (d) show hard-only and overall test accuracy, respectively. Hard-only label accuracy stays at 50% as correct labels move from low-α to high-α samples. Curves show dataset means, and shading shows one standard deviation across model-pair means.

B.5 CORRECT-LABEL ALLOCATION ACROSS DATASETS

In Fig. 3(b)–(c), we compare two cases: assigning correct labels to high-α or low-α hard-only samples. In Fig. 7, we extend this comparison to nine allocation levels, gradually moving correct labels from low-α to high-α samples while keeping hard-only label accuracy at 50%.

In Fig. 7(a)–(b), the mean hard-only saturation point $\widehat{b}$ and estimated hard feature knowledge b_T increase from Low to High on all four datasets. In Fig. 7(c)–(d), this shift also improves mean hard-only and overall accuracies, with larger gains on hard-only samples. Assigning correct labels to samples with stronger hard features therefore improves feature learning and generalization on average at fixed weak-label accuracy.

C FULL PROOF

The table below summarizes the notation used in the main text and the proofs.

Notation	Meaning
Data and weak labels	
n, d, m	Sample size, patch dimension, and neurons per class.
$k, k' \in \{\pm 1\}$, $j \in [m]$	Class indices and neuron index.
$\boldsymbol{X}_i, y_i, \tilde{y}_i$	Input, true label, and weak label.
$\boldsymbol{\mu}_k, \boldsymbol{\nu}_k$	Class-k easy and hard features.
$\alpha_i \sim P_\alpha$	Hard feature strength in $[\alpha_{\min}, \alpha_{\max}] \subset (0, 1)$.
τ_i, p_E, p_H, p_O	Sample type and probabilities: E (easy-only), H (hard-only), O (overlap).
$\mathcal{E}_k, \mathcal{O}_k, \mathcal{H}_k$	Class-k easy-only, overlap, and hard-only sets.
$z_i = y_i \tilde{y}_i$	Weak label correctness.
$\mathcal{C}_k, \mathcal{F}_k$; $A_{C,k}, A_{F,k}$	Correct and flipped hard-only sets; sums of their hard feature strengths.
b_e, b_h	Teacher's easy and hard feature knowledge.
ϵ_i, F_ϵ	Teacher output noise for sample i and its CDF.
$\boldsymbol{\xi}_i$, σ_p, $\boldsymbol{P}_\perp$	Input noise patch, its scale, and normalized noise covariance.
Training dynamics	
$\boldsymbol{w}_{k,j}^{(t)}, \sigma_0$	Student weight vector and its initialization scale.
$a_k^{(t)}, b_k^{(t)}$	Student's easy and hard feature knowledge.
$\rho_{k,j,i}^{(t)}$	Noise memorization from sample i in $\boldsymbol{w}_{k,j}$.
$c_i^{(t)}$	Mean of $\rho_{\tilde{y}_i,j,i}^{(t)}$ over j; noise learned in the weak-label class.
$\pi_{e,k}^{(t)}, \pi_{h,k}^{(t)}$	Active neuron fractions on class-k easy and hard features.
$\pi_i^{(t)}$	Active neuron fraction on $\boldsymbol{\xi}_i$ in the weak-label class, $\tilde{y}_i$.
$g_i^{(t)}$	Gradient weight $[1 + \exp\{\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\}]^{-1}$.
Saturation Analysis	
$f_{H,k}(b)$, $f_k(a, b)$	Hard-only and full hard feature update functions.
$\widehat{b}_k$, $\widehat{b}_k(a)$	Hard-only and full saturation points.
$R_k = A_{C,k}/A_{F,k}$	Correct-to-flipped hard-feature-strength ratio.
$\lambda_i = \eta\|\boldsymbol{\xi}_i\|^2/(mn)$	Sample i's noise-update scale.
Training regimes and error bounds	
η, T	Gradient descent step size and total number of training steps.
T_{noise}	End of the noise-free regime.

C.1 PROBLEM SETUP (RESTATED FOR CONVENIENCE)

We call the weak model that supplies training labels the *teacher* and the strong model trained on these labels the *student*. We study whether the student can predict the true labels of new samples more accurately than the teacher.

Notation. Bold letters denote vectors and matrices, and $[n] = \{1, \ldots, n\}$. Intervals of iteration indices contain only integers. We use $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ and $\|\boldsymbol{v}\|$ for the Euclidean inner product and norm, respectively. All asymptotics are as the patch dimension $d \to \infty$; $\mathrm{poly}(d)$ and $\mathrm{polylog}(d)$ denote a polynomial in d and a fixed positive power of $\log d$, resp.

Data and features. We study binary classification with true label $y \in \{\pm 1\}$ using a patch model inspired by Oh et al. (2025). For each class $k \in \{\pm 1\}$, let $\boldsymbol{\mu}_k, \boldsymbol{\nu}_k \in \mathbb{R}^d$ denote its easy and hard features, respectively; all four feature vectors are orthonormal. Each input $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}) \in \mathbb{R}^{d \times 3}$ consists of two signal patches containing class features and one noise patch.

Definition 1 (Data distribution). *We generate a sample* $(\boldsymbol{X}, y)$ *as follows:*

- ***Label and Type.*** *Draw a label $y \in \{\pm 1\}$ uniformly and a sample type $\tau = \mathrm{E}$ (easy-only), H (hard-only), or O (overlap) with probabilities $p_E, p_H > 0$, $p_O \geq 0$, and $p_E + p_H + p_O = 1$.*
- ***Signal patches.*** *Draw the hard-feature strength $\alpha \sim P_\alpha$ on $[\alpha_{\min}, \alpha_{\max}] \subset (0, 1)$ and set*

$$(\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}) = \begin{cases} (\boldsymbol{\mu}_y, \boldsymbol{\mu}_y), & \tau = \mathrm{E}; \\ (\boldsymbol{\mu}_y, \alpha\boldsymbol{\nu}_y), & \tau = \mathrm{O}; \\ (\alpha\boldsymbol{\nu}_y, \alpha\boldsymbol{\nu}_y), & \tau = \mathrm{H}. \end{cases}$$

- ***Input noise.*** *Draw Gaussian input noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_p^2 \boldsymbol{P}_\perp)$ independently of (y, τ, α), where $\sigma_p > 0$ is the input-noise scale and $\boldsymbol{P}_\perp = \boldsymbol{I}_d - \sum_k (\boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top + \boldsymbol{\nu}_k \boldsymbol{\nu}_k^\top)$.*
- ***Patch order.*** *Form $\boldsymbol{X}$ by randomly permuting $(\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \boldsymbol{\xi})$.*

Weak supervision. We denote the output of the trained teacher by $G(\boldsymbol{X})$ and its weak label by $\tilde{y} = \mathrm{sign}(G(\boldsymbol{X}))$. In the following definition, the nonnegative responses b_e and b_h quantify the teacher's easy and hard feature knowledge, respectively.

Definition 2 (Weak teacher response). *Let ϵ be output noise independent of the data, symmetric about zero, and supported on $[-B, B]$, with continuous cumulative distribution function F_ϵ. The teacher's true-label aligned response M_w and signed weak label $\tilde{y}$ are*

$$M_w = yG(\boldsymbol{X}) = \begin{cases} 2b_e + \epsilon, & \tau = \mathrm{E}; \\ b_e + \alpha b_h + \epsilon, & \tau = \mathrm{O}; \\ 2\alpha b_h + \epsilon, & \tau = \mathrm{H}, \end{cases} \qquad \tilde{y} = y\,\mathrm{sign}(M_w).$$

We assume that ($b_e > B \geq |\epsilon|$), so the teacher correctly labels any sample containing the easy feature. On hard-only samples, the label flips whenever the noise ϵ overwhelms the hard-feature response $2\alpha b_h$. Thus, the teacher's weakness is characterized by b_h relative to the noise ϵ.

Proposition 1 (Weak teacher accuracy). *Suppose that $b_e > B$. Then,*

$$\Pr(\tilde{y} = y \mid \tau \in \{\mathrm{E}, \mathrm{O}\}) = 1; \qquad \Pr(\tilde{y} = y \mid \tau = \mathrm{H}, \alpha) = F_\epsilon(2\alpha b_h).$$

When $b_h = 0$, the teacher's accuracy is $F_\epsilon(0) = 1/2$ on hard-only samples ($\tau = \mathrm{H}$). When $b_h > 0$, samples with stronger hard-feature strength α are more likely to receive correct weak labels.

Student model. The student model $F_{\boldsymbol{W}}$ is a two-layer neural network using ReLU activation. For each class $k \in \{\pm 1\}$, it has m trainable weight vectors $\boldsymbol{w}_{k,j} \in \mathbb{R}^d$, indexed by $j \in [m]$. Let $\boldsymbol{W}$ denote the collection of all weight vectors. The student's output is

$$F_{\boldsymbol{W}}(\boldsymbol{X}) = \frac{1}{m} \sum_{j=1}^{m} \sum_{p=1}^{3} \left[\phi(\langle \boldsymbol{w}_{+1,j}, \boldsymbol{x}^{(p)} \rangle) - \phi(\langle \boldsymbol{w}_{-1,j}, \boldsymbol{x}^{(p)} \rangle) \right], \quad \phi(u) = \max(0, u).$$

The student is strong in that it can represent all four feature directions and generalize to the data distribution. However, it also has sufficient capacity to fit incorrect labels by memorizing sample-specific input noise. Our central question is how the student can exploit weak supervision that is noisy along the hard-feature direction to learn the hard feature and generalize beyond the weak teacher.

Training and evaluation. We draw n independent samples $\{(\boldsymbol{X}_i, y_i)\}_{i=1}^n$ from the data distribution and assign the weak label $\tilde{y}_i = \mathrm{sign}(G(\boldsymbol{X}_i))$. For sample i, let y_i, τ_i, α_i, and $\boldsymbol{\xi}_i$ denote its true label, sample type, hard-feature strength, and input noise, respectively. For each true class k, let $\mathcal{E}_k, \mathcal{O}_k, \mathcal{H}_k$ denote the index sets of easy-only, overlap, and hard-only training samples, respectively.

We initialize each weight vector independently as $\boldsymbol{w}_{k,j}^{(0)} \sim \mathcal{N}(0, \sigma_0^2 \boldsymbol{I}_d)$. We train the student by full-batch gradient descent with step size $\eta > 0$ on the empirical logistic loss:

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-\tilde{y}_i F_{\boldsymbol{W}}(\boldsymbol{X}_i)}); \qquad \boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta \nabla \mathcal{L}(\boldsymbol{W}^{(t)}).$$

We write $\boldsymbol{W}^{(t)}$ for the parameters after t updates and train for $T = \mathrm{poly}(d)/\eta$ steps. Training uses only inputs and weak labels; true labels, sample types, and feature strengths are used only in the analysis. We evaluate the trained student on a fresh test sample $(\boldsymbol{X}, y)$ drawn from the same data distribution, measuring test error by $\Pr_{(\boldsymbol{X},y)}(yF_{\boldsymbol{W}}(\boldsymbol{X}) < 0)$.

C.2 Assumptions

All asymptotic notation is with respect to the patch dimension d.

Assumption 1 (Parameters)**.** *We assume the following conditions throughout the proof.*

(A1) ***Data.*** *The type probabilities $p_E, p_H > 0$, $p_O \geq 0$ and the feature-strength bounds $0 < \alpha_{\min} \leq \alpha_{\max} < 1$ are constants independent of d.*

(A2) ***Weak model.*** *The easy feature knowledge $b_e > B$ and the hard feature knowledge $0 < b_h \leq \bar{b}_h$ are fixed constants, and the output-noise density q_ϵ satisfies*

$$0 < q_- \leq q_\epsilon(u) \leq q_+ \qquad \text{for } |u| \leq 2\alpha_{\max}\bar{b}_h. \tag{12}$$

Here $q_-, q_+, \bar{b}_h$ are fixed constants, and for $c_{\mathrm{flip}} \in (0, 1/2)$, we assume $F_\epsilon(2\alpha b_h) \leq 1 - c_{\mathrm{flip}}$ for all $\alpha \in [\alpha_{\min}, \alpha_{\max}]$.

(A3) ***Sample size.***

$$n = o\left(\frac{\sigma_0\sigma_p d}{(\log d)^2}\right), \qquad n = \Omega(\log d). \tag{13}$$

(A4) ***Width and step size.*** *$m = \omega(\log d)$ and $0 < \eta \leq m/2$.*

(A5) ***Initialization.*** *$\sigma_0(1 + \sigma_p\sqrt{d})\log d = O(1/\mathrm{polylog}(d))$.*

(A6) ***High SNR.*** *$n/(\sigma_p^2 d) = \Omega(\mathrm{polylog}(d))$.*

(A7) ***Sufficient training time.*** *$T = \mathrm{poly}(d)/\eta$ and $T \geq mn/(\eta\sigma_p^2)$.*

Bounds implied by the parameters. (A3) and (A5) imply $n = o(\sqrt{d}/\log d)$, since $\sigma_0\sigma_p\sqrt{d} = o(1)$. We also have

$$\frac{n\log d}{\sqrt{d}} = o(1), \qquad \frac{\eta\sigma_p^2 d}{mn} \leq \frac{\sigma_p^2 d}{2n} = o(1), \qquad \log(2 + \eta T/m) = O(\log d). \tag{14}$$

For every fixed $M \geq 1$,

$$\sigma_0(1 + \sigma_p\sqrt{d})\log d + \frac{Mn\log d\,\log(2 + \eta T/m)}{\sqrt{d}} = o(1), \tag{15}$$

$$\frac{Mn\log d\,\log(2 + \eta T/m)}{\sigma_0\sigma_p d} + \sqrt{\frac{\log d}{m}} + \sqrt{\frac{\log d}{d}} = o(1). \tag{16}$$

C.3 A Common Good Event

We now define a good event $\mathcal{G}$ on which all required sample, initialization, and independent test-noise bounds hold. We write $z_i = y_i\tilde{y}_i$, so $z_i = 1$ means that sample i's weak label is correct and $z_i = -1$ means that it is flipped.

Let $\mathcal{V} = \{\boldsymbol{\mu}_{+1}, \boldsymbol{\mu}_{-1}, \boldsymbol{\nu}_{+1}, \boldsymbol{\nu}_{-1}\}$. The hard-only groups and their total hard-feature strengths are

$$\mathcal{C}_k = \{i \in \mathcal{H}_k : z_i = 1\}, \qquad \mathcal{F}_k = \{i \in \mathcal{H}_k : z_i = -1\},$$
$$A_{C,k} = \sum_{i\in\mathcal{C}_k} \alpha_i, \qquad A_{F,k} = \sum_{i\in\mathcal{F}_k} \alpha_i.$$

Lemma C4 (Good events)**.** *Suppose $n, m \leq \mathrm{poly}(d)$. Let $\mathcal{G}$ be the event that all conditions below hold, for fixed small $c_0 > 0$ and large $C_0 > 0$. Each condition holds for all $k \in \{\pm 1\}$, $j \in [m]$, and $i, \ell \in [n]$.*

(B1) ***Sample counts.***

$$\begin{aligned} c_0 n p_E \le |\mathcal{E}_k| \le C_0 n p_E, \\ c_0 n p_H \le |\mathcal{H}_k| \le C_0 n p_H, \\ c_0 n p_O \le |\mathcal{O}_k| \le C_0 n p_O. \end{aligned} \tag{17}$$

(B2) ***Hard feature strengths.***

$$c_0 n \le A_{C,k}, A_{F,k} \le C_0 n, \qquad c_0 n b_h \le A_{C,k} - A_{F,k} \le C_0 n b_h. \tag{18}$$

(B3) ***Weight initialization.*** $\|\boldsymbol{w}_{k,j}^{(0)}\| \le C_0 \sigma_0 \sqrt{d}$.

(B4) ***Initial feature knowledge.*** *For every* $\boldsymbol{v} \in \mathcal{V}$,

$$|\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{v} \rangle| \le C_0 \sigma_0 \log d. \tag{19}$$

(B5) ***Initial noise memorization.*** $|\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_i \rangle| \le C_0 \sigma_0 \sigma_p \sqrt{d} \log d$.

(B6) ***Input noise strength.***

$$\left| \frac{\|\boldsymbol{\xi}_i\|^2}{\sigma_p^2 d} - 1 \right| \le C_0 \sqrt{\frac{\log d}{d}}. \tag{20}$$

(B7) ***Noise correlation.*** *For* $i \ne \ell$, $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_\ell \rangle| \le C_0 \sigma_p^2 \sqrt{d} \log d$.

(B8) ***Initial active neurons.*** *For every* $\boldsymbol{v} \in \mathcal{V}$,

$$\frac{1}{3} \le \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\{\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{v} \rangle > 0\} \le \frac{2}{3}.$$

(B9) ***Noise projection distribution.*** *Let* Φ *be the standard normal CDF. For every threshold* $u \in \mathbb{R}$,

$$\left| \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\left\{ \frac{\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_i \rangle}{\sigma_0 \|\boldsymbol{\xi}_i\|} \le u \right\} - \Phi(u) \right| \le C_0 \sqrt{\frac{\log d}{m}}. \tag{21}$$

(B10) ***Test noise correlations.*** *For independent test noise* $\boldsymbol{\xi}_* \sim \mathcal{N}(0, \sigma_p^2 \boldsymbol{P}_\perp)$,

$$\begin{aligned} \max_{k,j} |\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_* \rangle| \le C_0 \sigma_0 \sigma_p \sqrt{d} \log d, \\ \max_{i} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_* \rangle| \le C_0 \sigma_p^2 \sqrt{d} \log d. \end{aligned}$$

Then $\mathcal{G}$ *holds with high probability, i.e.,* $\Pr(\mathcal{G}) \ge 1 - 1/\mathrm{poly}(d)$.

Proof. **Proof for (B1).** For each class k, the counts $|\mathcal{E}_k|, |\mathcal{H}_k|, |\mathcal{O}_k|$ are binomial with means $np_E/2, np_H/2, np_O/2$, respectively. Chernoff's inequality gives

$$\begin{aligned} \Pr\Big(\Big| |\mathcal{E}_k| - \frac{np_E}{2} \Big| > \frac{np_E}{4} \Big) \le 2e^{-np_E/24}, \\ \Pr\Big(\Big| |\mathcal{H}_k| - \frac{np_H}{2} \Big| > \frac{np_H}{4} \Big) \le 2e^{-np_H/24}, \\ \Pr\Big(\Big| |\mathcal{O}_k| - \frac{np_O}{2} \Big| > \frac{np_O}{4} \Big) \le 2e^{-np_O/24}. \end{aligned}$$

Proof for (B2). For $i \in \mathcal{H}_k$, (A2) gives

$$\begin{aligned} \Pr(z_i = 1) = F_\epsilon(2\alpha_i b_h), \\ \mathbb{E}[z_i] = 2 \int_0^{2\alpha_i b_h} q_\epsilon(u)\, du \in [4q_- \alpha_i b_h, 4q_+ \alpha_i b_h]. \end{aligned} \tag{22}$$

From (B1), we have $|\mathcal{H}_k| \asymp n$. Then $S_k = A_{C,k} - A_{F,k}$ and $\mu_k = \mathbb{E}[S_k]$ satisfy

$$S_k = \sum_{i \in \mathcal{H}_k} \alpha_i z_i,$$
$$4q_- \alpha_{\min}^2 |\mathcal{H}_k| b_h \le \mu_k \le 4q_+ \alpha_{\max}^2 |\mathcal{H}_k| b_h, \qquad \mu_k \asymp n b_h.$$

Hoeffding's inequality yields

$$\Pr\Big(|S_k - \mu_k| > \frac{\mu_k}{2}\Big) \le 2\exp\left(-\frac{\mu_k^2}{8\sum_{i\in\mathcal{H}_k}\alpha_i^2}\right) \le 2e^{-cnb_h^2}.$$

Moreover, $\Pr(z_i = 1) \ge 1/2$ and $\Pr(z_i = -1) \ge c_{\text{flip}}$, so

$$\min\{\mathbb{E}[A_{C,k}], \mathbb{E}[A_{F,k}]\} \ge \alpha_{\min} c_{\text{flip}} |\mathcal{H}_k|,$$
$$\Pr\left(\min\{A_{C,k}, A_{F,k}\} < \frac{\alpha_{\min} c_{\text{flip}} |\mathcal{H}_k|}{2}\right) \le 2\exp\left(-\frac{\alpha_{\min}^2 c_{\text{flip}}^2 |\mathcal{H}_k|}{2\alpha_{\max}^2}\right) \le Ce^{-cn}.$$

We also have $A_{C,k}, A_{F,k} \le \alpha_{\max}|\mathcal{H}_k| \le \alpha_{\max} n$.

Proof for (B3). For $Z \sim \chi_d^2$, $\mathbb{E}[e^{Z/4}] = 2^{d/2}$. Thus

$$\frac{\|\boldsymbol{w}_{k,j}^{(0)}\|^2}{\sigma_0^2} \sim \chi_d^2,$$
$$\Pr\left(\max_{k,j} \|\boldsymbol{w}_{k,j}^{(0)}\| > C_0\sigma_0\sqrt{d}\right) \le 2m\exp\left[-\left(\frac{C_0^2}{4} - \frac{\log 2}{2}\right)d\right] \le Cd^{-c}.$$

Proof for (B4). Gaussian initialization and a union bound give

$$\frac{\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{v}\rangle}{\sigma_0} \sim \mathcal{N}(0,1), \qquad \boldsymbol{v} \in \mathcal{V},$$
$$\Pr\left(\max_{k,j,\boldsymbol{v}\in\mathcal{V}} |\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{v}\rangle| > C_0\sigma_0 \log d\right) \le 16me^{-C_0^2(\log d)^2/2} \le Cd^{-c}.$$

Proof for (B5). From (B3), we have $\sigma_p^2\|\boldsymbol{P}_\perp \boldsymbol{w}_{k,j}^{(0)}\|^2 \le C_0^2\sigma_0^2\sigma_p^2 d$ with probability at least $1 - 1/\mathrm{poly}(d)$. Hence, for each k, j, and i, we have

$$\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_i\rangle \sim \mathcal{N}\Big(0, \sigma_p^2\|\boldsymbol{P}_\perp \boldsymbol{w}_{k,j}^{(0)}\|^2\Big).$$

Thus, the Gaussian tail bound gives

$$\Pr\Big(|\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_i\rangle| > C_0\sigma_0\sigma_p\sqrt{d}\log d\Big) \le 2e^{-(\log d)^2/2}.$$

Finally, a union bound yields

$$\Pr\left(\max_{k,j,i} |\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_i\rangle| > C_0\sigma_0\sigma_p\sqrt{d}\log d\right) \le 4mne^{-(\log d)^2/2} \le C'd^{-c'}.$$

Proof for (B6). Let $Z_i := \|\boldsymbol{\xi}_i\|^2/\sigma_p^2 \sim \chi_{d-4}^2$. For sufficiently large d, the chi-square tail bound and a union bound give

$$\Pr\left(\max_i \left|\frac{Z_i}{d} - 1\right| > C_0\sqrt{\frac{\log d}{d}}\right) \le \sum_{i=1}^n \Pr\left(|Z_i - (d-4)| > \frac{C_0}{2}\sqrt{d\log d}\right)$$
$$\le 2nd^{-cC_0^2} \le 2d^{-2}.$$

The last inequality follows from $n \le \mathrm{poly}(d)$ for sufficiently large C_0.

Proof for (B7). From (B6), we have $\|\boldsymbol{\xi}_i\|^2 \le 2\sigma_p^2 d$ for all i with probability at least $1 - 2d^{-2}$. For fixed $\boldsymbol{\xi}_i$ satisfying this bound and $\ell \neq i$, the noise patches are sampled independently, which yields

$$\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_\ell \rangle \sim \mathcal{N}(0, \sigma_p^2 \|\boldsymbol{\xi}_i\|^2).$$

Thus, the Gaussian tail bound gives

$$\Pr\Big(|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_\ell \rangle| > C_0 \sigma_p^2 \sqrt{d} \log d \Big) \le 2e^{-C_0^2 (\log d)^2/4}.$$

Finally, a union bound yields

$$\Pr\left(\max_{i \neq \ell} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_\ell \rangle| > C_0 \sigma_p^2 \sqrt{d} \log d \right) \le 2d^{-2} + 2n^2 e^{-C_0^2 (\log d)^2/4} \le C d^{-c}.$$

Proof for (B8). For each k and $\boldsymbol{v} \in \mathcal{V}$,

$$\sum_{j=1}^{m} \mathbf{1}\{\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{v} \rangle > 0\} \sim \mathrm{Bin}(m, 1/2).$$

Hoeffding's inequality and $m = \omega(\log d)$ give

$$\begin{aligned} &\Pr\left(\max_{k, \boldsymbol{v} \in \mathcal{V}} \left| \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\{\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{v} \rangle > 0\} - \frac{1}{2} \right| > \frac{1}{6} \right) \\ &\le 16 e^{-m/18} \le C d^{-c}. \end{aligned}$$

Proof for (B9). For fixed k, i, conditional on the noise patches, the variables

$$Z_{k,j,i} = \frac{\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_i \rangle}{\sigma_0 \|\boldsymbol{\xi}_i\|}, \qquad j \in [m],$$

are independent standard Gaussians. Set $\widehat{\Phi}_{k,i}(u) = m^{-1} \sum_j \mathbf{1}\{Z_{k,j,i} \le u\}$. Choose thresholds $u_r = \Phi^{-1}(r/m)$ for $r = 0, \dots, m$, with $u_0 = -\infty$ and $u_m = \infty$. Hoeffding's inequality and a union bound give

$$\Pr\left(\max_{k,i,r} \left| \widehat{\Phi}_{k,i}(u_r) - \frac{r}{m} \right| > \varepsilon \right) \le 4n(m+1) e^{-2m\varepsilon^2}.$$

Monotonicity between consecutive thresholds gives

$$\sup_{u \in \mathbb{R}} |\widehat{\Phi}_{k,i}(u) - \Phi(u)| \le \max_{0 \le r \le m} \left| \widehat{\Phi}_{k,i}(u_r) - \frac{r}{m} \right| + \frac{1}{m}.$$

Take $\varepsilon = (C_0/2)\sqrt{\log d / m}$, so $1/m \le \varepsilon$ for large d. Averaging over the noise patches and using $n, m \le \mathrm{poly}(d)$, we obtain for sufficiently large C_0

$$\Pr\left(\max_{k,i} \sup_{u \in \mathbb{R}} |\widehat{\Phi}_{k,i}(u) - \Phi(u)| > C_0 \sqrt{\frac{\log d}{m}} \right) \le 4n(m+1) d^{-C_0^2/2} \le C d^{-c}.$$

Proof for (B10). From (B3) and (B6), we have $\sigma_p^2 \|\boldsymbol{P}_\perp \boldsymbol{w}_{k,j}^{(0)}\|^2 \le C_0^2 \sigma_0^2 \sigma_p^2 d$ and $\|\boldsymbol{\xi}_i\|^2 \le 2\sigma_p^2 d$. Hence, for each k, j, and i, we have

$$\begin{aligned} \langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_* \rangle &\sim \mathcal{N}\Big(0, \sigma_p^2 \|\boldsymbol{P}_\perp \boldsymbol{w}_{k,j}^{(0)}\|^2\Big), \\ \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_* \rangle &\sim \mathcal{N}(0, \sigma_p^2 \|\boldsymbol{\xi}_i\|^2). \end{aligned}$$

Thus, the Gaussian tail bound gives

$$\begin{aligned} \Pr_{\boldsymbol{\xi}_*}\Big(|\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_* \rangle| > C_0 \sigma_0 \sigma_p \sqrt{d} \log d \Big) &\le 2e^{-(\log d)^2/2}, \\ \Pr_{\boldsymbol{\xi}_*}\Big(|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_* \rangle| > C_0 \sigma_p^2 \sqrt{d} \log d \Big) &\le 2e^{-C_0^2 (\log d)^2/4}. \end{aligned}$$

Finally, a union bound yields

$$\begin{aligned}\Pr_{\boldsymbol{\xi}_*}\Big(\max_{k,j}|\langle \boldsymbol{w}_{k,j}^{(0)},\boldsymbol{\xi}_*\rangle| > C_0\sigma_0\sigma_p\sqrt{d}\log d\Big)\\ + \Pr_{\boldsymbol{\xi}_*}\Big(\max_{i}|\langle \boldsymbol{\xi}_i,\boldsymbol{\xi}_*\rangle| > C_0\sigma_p^2\sqrt{d}\log d\Big)\\ \le 4me^{-(\log d)^2/2} + 2ne^{-C_0^2(\log d)^2/4} \le d^{-c_0}.\end{aligned}$$

Union bounds over all conditions give

$$\Pr(\mathcal{G}^c) \le Cd^{-c} + Ce^{-cn} + Ce^{-cnb_h^2} \le C'd^{-c'} = \frac{1}{\mathrm{poly}(d)}.$$

□

C.4 Bounds Used Throughout the Proof

The training bounds use only (B1)–(B9) in $\mathcal{G}$; (B10) is used for independent test samples. The bounds below control initialization errors, noise interference, and changes in neuron activations.

Initialization. The contribution of initialization is small:

$$\begin{aligned}\max_{\boldsymbol{v}\in\mathcal{V}}|\langle \boldsymbol{w}_{k,j}^{(0)},\boldsymbol{v}\rangle| \le C\sigma_0\log d,\\ |\langle \boldsymbol{w}_{k,j}^{(0)},\boldsymbol{\xi}_i\rangle| + |F_{\boldsymbol{W}^{(0)}}(\boldsymbol{\xi}_i)| \le C\sigma_0\sigma_p\sqrt{d}\log d.\end{aligned} \tag{23}$$

The second line uses $|\phi(u)| \le |u|$.

Noise strengths and interferences. The noise patches have similar strengths and small correlations.

$$\begin{aligned}\frac{\sigma_p^2 d}{2} \le \|\boldsymbol{\xi}_i\|^2 \le 2\sigma_p^2 d, \qquad \left|\frac{\|\boldsymbol{\xi}_i\|^2}{\sigma_p^2 d} - 1\right| \le C\sqrt{\frac{\log d}{d}},\\ \max_i \sum_{\ell\neq i}\frac{|\langle \boldsymbol{\xi}_\ell,\boldsymbol{\xi}_i\rangle|}{\|\boldsymbol{\xi}_\ell\|^2} \le \frac{Cn\log d}{\sqrt{d}} = o(1).\end{aligned} \tag{24}$$

Activation changes. From (B9), for every k, i and $u \ge 0$,

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{|\langle \boldsymbol{w}_{k,j}^{(0)},\boldsymbol{\xi}_i\rangle| \le u\} &\le \Phi\left(\frac{u}{\sigma_0\|\boldsymbol{\xi}_i\|}\right) - \Phi\left(-\frac{u}{\sigma_0\|\boldsymbol{\xi}_i\|}\right) + C\sqrt{\frac{\log d}{m}}\\ &\le C\left(\frac{u}{\sigma_0\sigma_p\sqrt{d}} + \sqrt{\frac{\log d}{m}}\right).\end{aligned} \tag{25}$$

Thus, if each inner product changes by at most u, at most this fraction of neurons can switch between active and inactive.

C.5 Signal-Noise Decomposition

The gradient descent update separates into easy-feature, hard-feature, and noise terms:

$$\begin{aligned}\boldsymbol{w}_{k,j}^{(t+1)} - \boldsymbol{w}_{k,j}^{(t)} &= -\eta\nabla_{\boldsymbol{w}_{k,j}}\mathcal{L}(\boldsymbol{W}^{(t)})\\ &= \frac{\eta}{mn}\sum_{k'\in\{\pm1\}} kk'\phi'\Big(\langle \boldsymbol{w}_{k,j}^{(t)},\boldsymbol{\mu}_{k'}\rangle\Big)\left(2\sum_{i\in\mathcal{E}_{k'}} g_i^{(t)} + \sum_{i\in\mathcal{O}_{k'}} g_i^{(t)}\right)\boldsymbol{\mu}_{k'}\\ &\quad + \frac{\eta}{mn}\sum_{k'\in\{\pm1\}} kk'\phi'\Big(\langle \boldsymbol{w}_{k,j}^{(t)},\boldsymbol{\nu}_{k'}\rangle\Big)\left(2\sum_{i\in\mathcal{H}_{k'}} \alpha_i z_i g_i^{(t)} + \sum_{i\in\mathcal{O}_{k'}} \alpha_i g_i^{(t)}\right)\boldsymbol{\nu}_{k'}\\ &\quad + \frac{\eta}{mn}\sum_{i=1}^n k\tilde{y}_i g_i^{(t)}\phi'\Big(\langle \boldsymbol{w}_{k,j}^{(t)},\boldsymbol{\xi}_i\rangle\Big)\boldsymbol{\xi}_i.\end{aligned} \tag{26}$$

Here $\phi'(u) = \mathbb{1}\{u > 0\}$, and the factors of two come from the repeated signal patches in easy-only and hard-only samples. Summing the noise updates from initialization gives the nonnegative coefficient

$$\rho_{k,j,i}^{(t)} = \lambda_i \sum_{v\in[0,t-1]} g_i^{(v)} \mathbb{1}\{\langle \boldsymbol{w}_{k,j}^{(v)}, \boldsymbol{\xi}_i\rangle > 0\}, \qquad \lambda_i = \frac{\eta\|\boldsymbol{\xi}_i\|^2}{mn}.$$

Write $\Delta\boldsymbol{w}_{k,j}^{(t)} = \boldsymbol{w}_{k,j}^{(t)} - \boldsymbol{w}_{k,j}^{(0)}$ for the accumulated weight change. Summing (26) and using orthogonality between the feature and noise directions gives

$$\begin{aligned}\Delta\boldsymbol{w}_{k,j}^{(t)} = &\underbrace{\sum_{k'\in\{\pm1\}} \langle \Delta\boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\mu}_{k'}\rangle \boldsymbol{\mu}_{k'}}_{\text{easy features}} + \underbrace{\sum_{k'\in\{\pm1\}} \langle \Delta\boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\nu}_{k'}\rangle \boldsymbol{\nu}_{k'}}_{\text{hard features}} \\ &+ \underbrace{\sum_{i=1}^{n} k\tilde{y}_i \rho_{k,j,i}^{(t)} \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|^2}}_{\text{sample-specific noise}}.\end{aligned} \tag{27}$$

The student's easy and hard feature knowledge are the average coefficients along its own class's feature directions:

$$a_k^{(t)} := \frac{1}{m}\sum_{j=1}^{m} \langle \Delta\boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\mu}_k\rangle, \qquad b_k^{(t)} := \frac{1}{m}\sum_{j=1}^{m} \langle \Delta\boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\nu}_k\rangle. \tag{28}$$

Similarly, the learned noise $c_i^{(t)}$ averages sample i's noise coefficients in the weak-label class $\tilde{y}_i$; the corresponding active-neuron fraction is $\pi_i^{(t)}$:

$$c_i^{(t)} := \frac{1}{m}\sum_{j=1}^{m} \rho_{\tilde{y}_i,j,i}^{(t)}, \qquad \pi_i^{(t)} := \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\langle \boldsymbol{w}_{\tilde{y}_i,j}^{(t)}, \boldsymbol{\xi}_i\rangle > 0\}. \tag{29}$$

Thus $a_k^{(0)} = b_k^{(0)} = c_i^{(0)} = 0$. The next lemma gives the exact updates of the feature and noise coefficients at all training times.

Lemma C5 (Feature and noise updates)**.** *For $t \in [0, T]$ and $i \in [n]$, the following identities hold.*

(C1) ***Feature updates.***

$$\begin{aligned} a_k^{(t+1)} - a_k^{(t)} &= \frac{\eta\pi_{e,k}^{(t)}}{mn}\left(2\sum_{i\in\mathcal{E}_k} g_i^{(t)} + \sum_{i\in\mathcal{O}_k} g_i^{(t)}\right), \\ b_k^{(t+1)} - b_k^{(t)} &= \frac{\eta\pi_{h,k}^{(t)}}{mn}\left(2\sum_{i\in\mathcal{H}_k} \alpha_i z_i g_i^{(t)} + \sum_{i\in\mathcal{O}_k} \alpha_i g_i^{(t)}\right), \end{aligned} \tag{30}$$

where the active-neuron fractions are

$$\pi_{e,k}^{(t)} := \frac{1}{m}\sum_{j=1}^{m} \phi'\Big(\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\mu}_k\rangle\Big), \qquad \pi_{h,k}^{(t)} := \frac{1}{m}\sum_{j=1}^{m} \phi'\Big(\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\nu}_k\rangle\Big).$$

(C2) ***Noise learned from each sample.*** *The learned noise satisfies*

$$\rho_{k,j,i}^{(0)} = 0, \qquad \rho_{k,j,i}^{(t+1)} - \rho_{k,j,i}^{(t)} = \lambda_i g_i^{(t)} \mathbb{1}\{\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\xi}_i\rangle > 0\}, \qquad \lambda_i = \frac{\eta\|\boldsymbol{\xi}_i\|^2}{mn}. \tag{31}$$

Then $0 \le \rho_{k,j,i}^{(t)} \le \lambda_i t \le C\frac{\eta t\sigma_p^2 d}{mn}$, and

$$\begin{aligned} \langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\xi}_i\rangle &= \langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_i\rangle + k\tilde{y}_i\rho_{k,j,i}^{(t)} + X_{k,j,i}(t), \\ X_{k,j,i}(t) &= \sum_{\ell\neq i} k\tilde{y}_\ell \rho_{k,j,\ell}^{(t)} \frac{\langle \boldsymbol{\xi}_\ell, \boldsymbol{\xi}_i\rangle}{\|\boldsymbol{\xi}_\ell\|^2}. \end{aligned} \tag{32}$$

(C3) ***Average learned noise.*** *The average learned noise satisfies*

$$c_i^{(t+1)} - c_i^{(t)} = \lambda_i \pi_i^{(t)} g_i^{(t)}, \qquad 0 \le c_i^{(t)} \le C\frac{\eta t \sigma_p^2 d}{mn}. \tag{33}$$

Here $\pi_i^{(t)}$ is the fraction of neurons in class $\tilde{y}_i$ active on $\boldsymbol{\xi}_i$:

$$\pi_i^{(t)} := \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\langle \boldsymbol{w}_{\tilde{y}_i,j}^{(t)}, \boldsymbol{\xi}_i \rangle > 0\}.$$

If $a_k^{(s)}, b_k^{(s)} \ge 0$ for all $s \in [0,t]$, the class-k active set remains unchanged and the opposite-class outputs are bounded by their initial values. Thus, for $\boldsymbol{v} \in \{\boldsymbol{\mu}_k, \boldsymbol{\nu}_k\}$,

$$\begin{aligned} kF_{\boldsymbol{W}^{(t)}}(\boldsymbol{v}) &= \frac{1}{m}\sum_{j=1}^{m}\Big[\langle \Delta\boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{v}\rangle + \phi\Big(\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{v}\rangle\Big) - \phi\Big(\langle \boldsymbol{w}_{-k,j}^{(t)}, \boldsymbol{v}\rangle\Big)\Big] \\ &= \frac{1}{m}\sum_{j=1}^{m}\langle \Delta\boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{v}\rangle + O(\sigma_0 \log d). \end{aligned}$$

By the definitions of $a_k^{(t)}$ and $b_k^{(t)}$,

$$kF_{\boldsymbol{W}^{(t)}}(\boldsymbol{\mu}_k) = a_k^{(t)} + O(\sigma_0 \log d), \qquad kF_{\boldsymbol{W}^{(t)}}(\boldsymbol{\nu}_k) = b_k^{(t)} + O(\sigma_0 \log d). \tag{34}$$

For the noise of sample i, applying the 1-Lipschitz property of ReLU to (32) and $\rho_{k,j,i}^{(t)} \ge 0$ gives

$$\begin{aligned} \Big|\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i) - c_i^{(t)}\Big| &\le \frac{1}{m}\sum_{k,j}\Big(\Big|\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_i\rangle\Big| + |X_{k,j,i}(t)|\Big) \\ &\le C\sigma_0\sigma_p\sqrt{d}\log d + 2\max_{k,j}|X_{k,j,i}(t)|. \end{aligned}$$

Therefore,

$$\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i) = c_i^{(t)} + O\left(\sigma_0\sigma_p\sqrt{d}\log d + \max_{k,j}|X_{k,j,i}(t)|\right). \tag{35}$$

C.6 Noise-Free Epoch

We call $t \in [0, T_{\mathrm{noise}}]$ the *noise-free epoch*, during which the input-noise contribution to every training output is $O(1/\mathrm{polylog}(d))$. The next lemma specifies T_{noise} and proves this bound.

Lemma C6 (Noise-free epoch)**.** *Choose T_{noise} satisfying*

$$1 \le \frac{\eta T_{\mathrm{noise}}}{m}, \qquad \frac{\eta T_{\mathrm{noise}}\sigma_p^2 d}{mn} = O\left(\frac{1}{\mathrm{polylog}(d)}\right). \tag{36}$$

For $t \in [0, T_{\mathrm{noise}}]$ and $i \in [n]$,

$$|F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i)| = O\left(\frac{1}{\mathrm{polylog}(d)}\right). \tag{37}$$

Proof. From (A7) and (36), $T_{\mathrm{noise}}/T = O(1/(d\,\mathrm{polylog}(d)))$, so $2T_{\mathrm{noise}} \le T$ for sufficiently large d. From (35), for $t \in [0, T_{\mathrm{noise}}]$,

$$|F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i)| \le c_i^{(t)} + C\left(\sigma_0\sigma_p\sqrt{d}\log d + \max_{k,j}|X_{k,j,i}(t)|\right).$$

From Lemma C5 and (24), we obtain

$$|F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i)| \le C\left[\sigma_0\sigma_p\sqrt{d}\log d + \frac{\eta t\sigma_p^2 d}{mn}\left(1 + \frac{n\log d}{\sqrt{d}}\right)\right] = O\left(\frac{1}{\mathrm{polylog}(d)}\right).$$

The last equality follows from (A5), (36), and $n\log d/\sqrt{d} = o(1)$. □

Lemma C7 (Feature updates in the noise-free epoch)**.** *For $t \in [0, T_{\text{noise}}]$, the class-$k$ active set on its own features remains unchanged, $a_k^{(t)}$ is nondecreasing, and for a constant $B_0 > 0$,*

$$a_k^{(t)} \geq 0, \qquad 0 \leq b_k^{(t)} \leq B_0. \tag{38}$$

With $g(u) = (1 + e^u)^{-1}$, write

$$f_k(a, b) := \sum_{i \in \mathcal{H}_k} \alpha_i z_i g(2\alpha_i z_i b) + \frac{1}{2} \sum_{i \in \mathcal{O}_k} \alpha_i g(a + \alpha_i b),$$
$$\delta_{T_{\text{noise}}} := \sigma_0(1 + \sigma_p\sqrt{d}) \log d + \frac{\eta T_{\text{noise}} \sigma_p^2 d}{mn} \left(1 + \frac{n \log d}{\sqrt{d}}\right) = O\left(\frac{1}{\text{polylog}(d)}\right). \tag{39}$$

For $t \in [0, T_{\text{noise}} - 1]$, the feature updates satisfy

$$\begin{aligned} a_k^{(t+1)} - a_k^{(t)} &= \frac{\eta \pi_{e,k}}{mn} \left(2|\mathcal{E}_k| g(2a_k^{(t)}) + \sum_{i \in \mathcal{O}_k} g(a_k^{(t)} + \alpha_i b_k^{(t)})\right) + O(\frac{\eta}{m} \delta_{T_{\text{noise}}}), \\ b_k^{(t+1)} - b_k^{(t)} &= \frac{2\eta \pi_{h,k}}{mn} f_k(a_k^{(t)}, b_k^{(t)}) + O(\frac{\eta}{m} \delta_{T_{\text{noise}}}). \end{aligned} \tag{40}$$

Here $\pi_{e,k} = \pi_{e,k}^{(0)}$ and $\pi_{h,k} = \pi_{h,k}^{(0)}$, with $\pi_{e,k}, \pi_{h,k} \in [1/3, 2/3]$.

Proof. From Lemma C5, $a_k^{(t)}$ is nondecreasing and the easy-feature active set stays fixed because easy-only and overlap labels are correct. For the hard feature, suppose $b_k^{(s)} \in [0, B_0]$ and its class-k active set remains unchanged for $s \in [0, t]$. From (34) and Lemma C6,

$$\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) = \begin{cases} 2a_k^{(t)}, & i \in \mathcal{E}_k; \\ a_k^{(t)} + \alpha_i b_k^{(t)}, & i \in \mathcal{O}_k; \\ 2z_i \alpha_i b_k^{(t)}, & i \in \mathcal{H}_k, \end{cases} \quad + e_i(t), \qquad |e_i(t)| \leq C\delta_{T_{\text{noise}}}. \tag{41}$$

Substituting these margins into (30) and using $|g'| \leq 1/4$ gives (40) at time t.

From (B2), we have

$$f_k(a, 0) = \frac{A_{C,k} - A_{F,k}}{2} + \frac{1}{2} \sum_{i \in \mathcal{O}_k} \alpha_i g(a) \geq \frac{c_0 n b_h}{2} > 0, \qquad a \geq 0.$$

In the negative terms of $f_k(a, B_0)$, we use (B2) and $g(-2\alpha_i B_0) \geq 1/2$ for $B_0 > 0$ to obtain

$$-\sum_{i \in \mathcal{F}_k} \alpha_i g(-2\alpha_i B_0) \leq -\frac{A_{F,k}}{2} \leq -\frac{c_0 n}{2}.$$

In the positive terms of $f_k(a, B_0)$, we use $ug(u) \leq 1$ for $u \geq 0$ and the monotonicity of g to obtain

$$\sum_{i \in \mathcal{C}_k} \alpha_i g(2\alpha_i B_0) + \frac{1}{2} \sum_{i \in \mathcal{O}_k} \alpha_i g(a + \alpha_i B_0) \leq \frac{|\mathcal{C}_k| + |\mathcal{O}_k|}{2B_0} \leq \frac{n}{2B_0}.$$

Thus, choosing $B_0 = 2/c_0$ gives, for both classes and every $a \geq 0$,

$$f_k(a, B_0) \leq \frac{n}{2B_0} - \frac{c_0 n}{2} = -\frac{c_0 n}{4} < 0.$$

From $-g'(u) = e^u/(1 + e^u)^2 \leq 1/4$ and $z_i^2 = 1$, we obtain

$$\begin{aligned} 0 \leq -\partial_b f_k(a, b) &= -2 \sum_{i \in \mathcal{H}_k} \alpha_i^2 g'(2\alpha_i z_i b) - \frac{1}{2} \sum_{i \in \mathcal{O}_k} \alpha_i^2 g'(a + \alpha_i b) \\ &\leq \frac{\alpha_{\max}^2}{2} |\mathcal{H}_k| + \frac{\alpha_{\max}^2}{8} |\mathcal{O}_k| \leq \frac{n \alpha_{\max}^2}{2}. \end{aligned} \tag{42}$$

Now we consider the map $b \mapsto b + \frac{2\eta\pi_{h,k}}{mn} f_k(a,b)$ for fixed $a \geq 0$. From (42), $\eta/m \leq 1/2$, and $\pi_{h,k}\alpha_{\max}^2 \leq 1$,

$$\frac{\partial}{\partial b}\left[b + \frac{2\eta\pi_{h,k}}{mn} f_k(a,b)\right] = 1 + \frac{2\eta\pi_{h,k}}{mn}\partial_b f_k(a,b) \geq 1 - \frac{\eta\pi_{h,k}\alpha_{\max}^2}{m} \geq \frac{1}{2} > 0.$$

Thus the map is increasing. Evaluating it at 0 gives the lower bound

$$b_k^{(t+1)} \geq \frac{2\eta\pi_{h,k}}{mn} f_k(a_k^{(t)}, 0) - \frac{C\eta}{m}\delta_{T_{\text{noise}}} \geq 0,$$

and evaluating it at B_0 gives the upper bound

$$b_k^{(t+1)} \leq B_0 + \frac{2\eta\pi_{h,k}}{mn} f_k(a_k^{(t)}, B_0) + \frac{C\eta}{m}\delta_{T_{\text{noise}}} \leq B_0.$$

Thus $b_k^{(t+1)} \in [0, B_0]$, preserving the activation and completing the induction.

□

C.7 Easy Feature Learning

For the fixed class k, write $a_t = a_k^{(t)}$, $b_t = b_k^{(t)}$, $\pi_e = \pi_{e,k}^{(0)}$, and $\pi_h = \pi_{h,k}^{(0)}$. Every sample containing the easy feature has a correct weak label, so every easy feature update is nonnegative. We bound the growth of a_t to show how quickly the easy feature fits overlap samples and reduces their contribution to the hard feature update.

Lemma C8 (Easy feature growth)**.** *For $t \in [0, T_{\text{noise}}]$,*

$$a_t \asymp \log(1 + \eta t/m). \tag{43}$$

Proof. From (41), easy-only samples have margins $2a_t + O(\delta_{T_{\text{noise}}})$, so $g_i^{(t)} \geq ce^{-2a_t}$. Since $|\mathcal{E}_k| \asymp n$ and $\pi_e \geq 1/3$, their updates in (30) give

$$a_{t+1} - a_t \geq \frac{2\eta\pi_e}{mn} \sum_{i \in \mathcal{E}_k} g_i^{(t)} \geq \frac{c\eta}{m} e^{-2a_t},$$

$$e^{2a_{t+1}} - e^{2a_t} \geq 2e^{2a_t}(a_{t+1} - a_t) \geq \frac{c\eta}{m}.$$

Summing from $a_0 = 0$ gives $e^{2a_t} \geq 1 + c\eta t/m$.

Since $b_t \leq B_0$, overlap samples have margins $a_t + O(1)$ and hence $g_i^{(t)} \geq ce^{-a_t}$. Using $|\mathcal{O}_k| \geq cp_O n$ in (30), we obtain

$$a_{t+1} - a_t \geq \frac{\eta\pi_e}{mn} \sum_{i \in \mathcal{O}_k} g_i^{(t)} \geq \frac{c\eta}{m} p_O e^{-a_t},$$

$$e^{a_{t+1}} - e^{a_t} \geq e^{a_t}(a_{t+1} - a_t) \geq \frac{c\eta}{m} p_O.$$

Summing also gives $e^{a_t} \geq 1 + cp_O\eta t/m$. Both sample types contribute nonnegative updates, so these lower bounds hold simultaneously:

$$a_t \geq \max\left\{\frac{1}{2}\log(1 + c\eta t/m),\, \log(1 + cp_O\eta t/m)\right\}.$$

For the upper bound, both easy-only and overlap samples have margins at least $a_t - C\delta_{T_{\text{noise}}}$ because $a_t, b_t \geq 0$. Thus $g_i^{(t)} \leq Ce^{-a_t}$ for these samples. From (30) and $\eta/m \leq 1/2$, we have $a_{t+1} - a_t \leq 1$ and

$$0 \leq a_{t+1} - a_t \leq C\eta e^{-a_t}/m,$$

$$e^{a_{t+1}} - e^{a_t} = \int_{a_t}^{a_{t+1}} e^s\, ds \leq e^{a_{t+1}}(a_{t+1} - a_t)$$

$$\leq e\, e^{a_t}(a_{t+1} - a_t) \leq \frac{C\eta}{m}.$$

Summing gives $e^{a_t} \leq 1 + C\eta t/m$. Therefore, there exist constants $c_0, C_0 > 0$ such that

$$c_0 \log(1 + \eta t/m) \leq a_t \leq C_0 \log(1 + \eta t/m),$$

which proves (43). □

As a_t grows, the easy feature increases each overlap sample's weak-label margin. Since $b_t \geq 0$, each such sample satisfies $g_i^{(t)} \leq Ce^{-a_t} \leq C/(1 + cp_O\eta t/m)$. Its remaining contribution to the hard feature update therefore decreases.

C.8 Hard Feature Learning

For the fixed class k, omit the class subscript and define

$$A_C := \sum_{i\in\mathcal{H}:z_i=1} \alpha_i, \qquad A_F := \sum_{i\in\mathcal{H}:z_i=-1} \alpha_i, \qquad A_O := \sum_{i\in\mathcal{O}} \alpha_i. \tag{44}$$

From (40), write the hard-only and full update functions as

$$\begin{aligned} f_H(b) &= \sum_{i\in\mathcal{H}} \alpha_i z_i g(2\alpha_i z_i b) \\ &= \sum_{i\in\mathcal{H}:z_i=1} \alpha_i g(2\alpha_i b) - \sum_{i\in\mathcal{H}:z_i=-1} \alpha_i g(-2\alpha_i b), \\ f(a,b) &= f_H(b) + \frac{1}{2}\sum_{i\in\mathcal{O}} \alpha_i g(a + \alpha_i b). \end{aligned}$$

The following lemma describes the saturation points of $f_H(b)$.

Lemma C9 (Hard-only saturation point)**.** *Let $R = A_C/A_F > 1$. The update function f_H is strictly decreasing and has a unique root $\widehat{b}$, satisfying*

$$\frac{\log R}{2\alpha_{\max}} \leq \widehat{b} \leq \frac{\log R}{2\alpha_{\min}}. \tag{45}$$

For fixed $K > 0$, the following bounds hold for $|b| \leq K$ and $a \geq 0$, with a constant $c_K > 0$:

$$c_K|\mathcal{H}| \leq -f_H'(b) \leq \frac{\alpha_{\max}^2|\mathcal{H}|}{2}, \qquad c_K|\mathcal{H}| \leq -\partial_b f(a,b) \leq \alpha_{\max}^2\left(\frac{|\mathcal{H}|}{2} + \frac{|\mathcal{O}|}{8}\right). \tag{46}$$

Proof. From the definition of f_H, we have

$$\begin{aligned} f_H(b) &= A_C - \sum_{i\in\mathcal{H}} \frac{\alpha_i e^{2\alpha_i b}}{1 + e^{2\alpha_i b}}, \\ -f_H'(b) &= 2\sum_{i\in\mathcal{H}} \frac{\alpha_i^2 e^{2\alpha_i b}}{(1 + e^{2\alpha_i b})^2} > 0. \end{aligned}$$

Thus f_H strictly decreases from $A_C > 0$ to $-A_F < 0$ as b runs from $-\infty$ to $+\infty$, so it has a unique root.

For $R > 1$, at $b = \log R/(2\alpha_{\max})$, every exponential $e^{2\alpha_i b}$ is at most R; at $b = \log R/(2\alpha_{\min})$, it is at least R. Therefore,

$$\begin{aligned} f_H\left(\frac{\log R}{2\alpha_{\max}}\right) &\geq \frac{A_C - RA_F}{1 + R} = 0, \\ f_H\left(\frac{\log R}{2\alpha_{\min}}\right) &\leq \frac{A_C - RA_F}{1 + R} = 0. \end{aligned}$$

Since f_H is decreasing, its root lies between these two points, proving (45). Also, $-g'(u) = e^u/(1 + e^u)^2$ decreases with $|u|$. Thus, for $|b| \leq K$, the bound $|2\alpha_i b| \leq 2\alpha_{\max}K$ gives

$$\frac{2\alpha_{\min}^2 e^{2\alpha_{\max}K}}{(1 + e^{2\alpha_{\max}K})^2}|\mathcal{H}| \leq -f_H'(b) \leq \frac{\alpha_{\max}^2|\mathcal{H}|}{2}.$$

The overlap contribution adds a nonnegative term bounded by

$$0 \leq -\partial_b f(a,b) + f_H'(b) = \frac{1}{2}\sum_{i\in\mathcal{O}} \frac{\alpha_i^2 e^{a+\alpha_i b}}{(1 + e^{a+\alpha_i b})^2} \leq \frac{\alpha_{\max}^2|\mathcal{O}|}{8}.$$

Combining these bounds proves (46). □

How overlap samples shift the balance. An overlap sample $i \in \mathcal{O}$ contains both easy and hard features and has approximate margin $a_t + \alpha_i b_t$, as shown in (41). Its correct label initially supports learning both features. However, as the easy-feature contribution a_t grows and fits the label, the sample contributes progressively less to hard-feature learning. For fixed a, define the *full saturation point* $\widehat{b}(a)$ as the root of $f(a, \cdot)$.

Lemma C10 (Full saturation point)**.** *For each $a \geq 0$, the full saturation point $\widehat{b}(a)$ satisfies*

$$\widehat{b}(a) \geq \widehat{b}, \qquad \widehat{b}(a) \downarrow \widehat{b} \quad (a \to \infty).$$

Equality holds when $\mathcal{O} = \varnothing$.

Proof. For fixed $a \geq 0$, $f_H(b)$ is strictly decreasing in b, and the overlap contribution is nonincreasing in b. Thus $f(a, b)$ strictly decreases from $A_C + A_O/2 > 0$ to $-A_F < 0$ as b runs from $-\infty$ to $+\infty$, so it has a unique root.

Since $f(a, \widehat{b}) \geq 0$ and the overlap contribution is nonincreasing in a, the roots satisfy

$$\widehat{b} \leq \widehat{b}(a_2) \leq \widehat{b}(a_1), \qquad a_2 \geq a_1 \geq 0.$$

Thus $\widehat{b}(a)$ has a finite limit. It also satisfies

$$\begin{aligned} 0 \leq -f_H(\widehat{b}(a)) &= \frac{1}{2} \sum_{i \in \mathcal{O}} \frac{\alpha_i}{1 + e^{a + \alpha_i \widehat{b}(a)}} \\ &\leq \frac{e^{-a}}{2} \sum_{i \in \mathcal{O}} \alpha_i e^{-\alpha_i \widehat{b}} \longrightarrow 0. \end{aligned}$$

By continuity, the limit is a root of f_H, whose unique root is $\widehat{b}$. If $\mathcal{O} = \varnothing$, then $f(a, b) = f_H(b)$ and $\widehat{b}(a) = \widehat{b}$ for every a. □

C.9 Hard Feature Dynamics

Since correctly labeled hard-only samples have greater total feature strength than flipped samples, b_t initially increases. We first bound this growth and then show how b_t approaches the hard-only saturation point $\widehat{b}$.

Lemma C11 (Initial hard feature growth)**.** *For $t \in [0, T_{\text{noise}}]$,*

$$b_t \geq c b_h (1 - e^{-c' \eta t / m}). \tag{47}$$

Here $c, c' > 0$ are fixed constants.

Proof. From (B2), (42), and $b_t \geq 0$, we have

$$\begin{aligned} f(a_t, b_t) &= f(a_t, 0) + \int_0^{b_t} \partial_b f(a_t, u)\, du \\ &\geq \frac{A_C - A_F}{2} + \frac{A_O}{2} g(a_t) - \frac{n \alpha_{\max}^2}{2} b_t \\ &\geq \frac{n}{2} (c_0 b_h - \alpha_{\max}^2 b_t). \end{aligned}$$

Substituting into (40), with $\pi_h \in [1/3, 2/3]$, $b_t \geq 0$, and $\delta_{T_{\text{noise}}} = o(b_h)$, yields

$$\begin{aligned} b_{t+1} &\geq b_t + \frac{2\eta \pi_h}{mn} f(a_t, b_t) - \frac{C\eta}{m} \delta_{T_{\text{noise}}} \\ &\geq \left(1 - \frac{\eta \pi_h \alpha_{\max}^2}{m}\right) b_t + \frac{\eta}{m} \left(\pi_h c_0 b_h - C \delta_{T_{\text{noise}}}\right) \\ &\geq \left(1 - \frac{\eta \alpha_{\max}^2}{m}\right) b_t + \frac{\eta}{m} \left(\frac{c_0}{3} b_h - C \delta_{T_{\text{noise}}}\right) \\ &\geq \left(1 - \frac{\eta \alpha_{\max}^2}{m}\right) b_t + \frac{c\eta}{m} b_h. \end{aligned}$$

The multiplier lies in $[1/2, 1)$, so iterating from $b_0 = 0$ gives

$$\begin{aligned} b_t &\geq \frac{c\eta b_h}{m} \sum_{s=0}^{t-1} \left(1 - \frac{\eta\alpha_{\max}^2}{m}\right)^s \\ &= \frac{cb_h}{\alpha_{\max}^2} \left[1 - \left(1 - \frac{\eta\alpha_{\max}^2}{m}\right)^t\right] \\ &\geq \frac{cb_h}{\alpha_{\max}^2} \left(1 - e^{-\alpha_{\max}^2 \eta t/m}\right). \end{aligned}$$

□

Approaching the hard-only saturation point. The hard-only update moves b_t toward $\widehat{b}$. As a_t grows, the overlap updates weaken, while initialization and input noise contribute a small error. The following lemma bounds the distance between b_t and $\widehat{b}$ during the noise-free regime.

Lemma C12 (Tracking the hard-only saturation point)**.** *For $t \in [0, T_{\text{noise}}]$,*

$$|b_t - \widehat{b}| \leq Cb_h e^{-c\eta t/m} + \frac{Cp_O}{1 + cp_O\eta t/m} + C\delta_{T_{\text{noise}}}. \tag{48}$$

Proof. Lemma C7 and the monotonicity of f_H give

$$b_t \in [0, B_0], \qquad f_H(0) > 0 > f_H(B_0), \qquad \widehat{b} \in [0, B_0].$$

Recall the derivative bounds in (46):

$$|\mathcal{H}| \asymp n, \qquad cn \leq -f_H'(b) \leq \frac{n\alpha_{\max}^2}{2}, \qquad b \in [0, B_0].$$

Since $f_H(\widehat{b}) = 0$, the mean value theorem gives, for some u_t between b_t and $\widehat{b}$,

$$b_t + \frac{2\eta\pi_h}{mn} f_H(b_t) - \widehat{b} = \left(1 + \frac{2\eta\pi_h}{mn} f_H'(u_t)\right)(b_t - \widehat{b}).$$

The derivative bounds, $\pi_h \in [1/3, 2/3]$, and $\eta/m \leq 1/2$ imply

$$\frac{1}{2} \leq 1 + \frac{2\eta\pi_h}{mn} f_H'(u_t) \leq 1 - \frac{\lambda\eta}{m} =: r < 1$$

for a fixed $\lambda > 0$. Thus each hard-only update reduces the distance to $\widehat{b}$ by a factor of at most r.

For the overlap contribution, $b_t \geq 0$ and $|\mathcal{O}| \leq Cnp_O$ give

$$\begin{aligned} 0 \leq \frac{f(a_t, b_t) - f_H(b_t)}{n} &= \frac{1}{2n} \sum_{i \in \mathcal{O}} \frac{\alpha_i}{1 + e^{a_t + \alpha_i b_t}} \\ &\leq Cp_O e^{-a_t} \leq \frac{Cp_O}{1 + cp_O\eta t/m}. \end{aligned}$$

The last inequality uses $e^{a_t} \geq 1 + cp_O\eta t/m$, proved in Lemma C8. Adding this contribution and the error in (40) yields

$$\begin{aligned} |b_{t+1} - \widehat{b}| &\leq r|b_t - \widehat{b}| + \frac{2\eta\pi_h}{mn}\big(f(a_t, b_t) - f_H(b_t)\big) + \frac{C\eta}{m}\delta_{T_{\text{noise}}} \\ &\leq r|b_t - \widehat{b}| + \frac{C\eta}{m}\left(\frac{p_O}{1 + cp_O\eta t/m} + \delta_{T_{\text{noise}}}\right). \end{aligned} \tag{49}$$

By (B2) and Lemma C9, $\widehat{b} \leq \log R/(2\alpha_{\min}) \leq Cb_h$. Iterating from $b_0 = 0$ and using $(\eta/m)\sum_{j \geq 0} r^j = 1/\lambda$ gives

$$|b_t - \widehat{b}| \leq Cb_h r^t + C\delta_{T_{\text{noise}}} + \frac{C\eta}{m} \sum_{u \in [0, t-1]} r^{t-1-u} \frac{p_O}{1 + cp_O\eta u/m}.$$

To bound the remaining sum, split the overlap updates at $u = t/2$. For $u \in [0, \lceil t/2 \rceil - 1]$, the factor r^{t-1-u} contains at least $\lfloor t/2 \rfloor$ powers of r, so these earlier updates have already been reduced by repeated hard-only updates. For $u \in [\lceil t/2 \rceil, t-1]$, the overlap contribution itself is small, since $1 + cp_O\eta u/m \geq (1 + cp_O\eta t/m)/2$. Applying the same geometric-sum bound to each part gives

$$\begin{aligned}\frac{\eta}{m}\sum_{u\in[0,t-1]} r^{t-1-u}\frac{p_O}{1+cp_O\eta u/m} &\leq \frac{p_O}{\lambda}r^{\lfloor t/2\rfloor} + \frac{2p_O}{\lambda(1+cp_O\eta t/m)} \\ &\leq Cp_O e^{-\lambda\eta t/(2m)} + \frac{Cp_O}{1+cp_O\eta t/m} \\ &\leq \frac{Cp_O}{1+cp_O\eta t/m}.\end{aligned}$$

The last inequality uses $p_O \leq 1$ and the exponential decay of the first term. Combining these bounds with $r^t \leq e^{-\lambda\eta t/m}$ proves (48). □

Conclusion for early training. For both classes and $t \in [0, T_{\mathrm{noise}}]$:

- **The easy feature knowledge grows logarithmically.** $a_t \asymp \log(1 + \eta t/m)$.
- **The hard feature knowledge saturates.**

$$|b_t - \widehat{b}| \leq Cb_h e^{-c\eta t/m} + \frac{Cp_O}{1+cp_O\eta t/m} + O\bigg(\frac{1}{\mathrm{polylog}(d)}\bigg).$$

- **Noise outputs remain small.** $\max_i |F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i)| = O(1/\mathrm{polylog}(d))$.

The first two bounds prove Theorem 1, since $\Pr(\mathcal{G}) \geq 1 - 1/\mathrm{poly}(d)$.

C.10 Noise Memorization and Feature Preservation

Positive hard feature knowledge increases the weak-label margins of correctly labeled hard-only samples and decreases those of flipped samples. We first show that samples learn input noise at nearly equal rates, then use the margin difference to compare the noise learned by the two groups. This comparison keeps the hard feature knowledge above a positive fraction of $\widehat{b}$ from the end of early training through T.

Noise outputs and learning rates. Recall that the learned noise coefficients are

$$\rho_{s,j,i}^{(t)} = \lambda_i \sum_{v\in[0,t-1]} g_i^{(v)}\mathbb{1}\{\langle \boldsymbol{w}_{s,j}^{(v)}, \boldsymbol{\xi}_i\rangle > 0\}, \qquad \lambda_i = \frac{\eta\|\boldsymbol{\xi}_i\|^2}{mn}.$$

Recall that $c_i^{(t)}$ averages the learned noise coefficients $\rho_{\tilde{y}_i,j,i}^{(t)}$ over j.

$$\begin{aligned}\pi_i^{(t)} &:= \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\{\langle \boldsymbol{w}_{\tilde{y}_i,j}^{(t)}, \boldsymbol{\xi}_i\rangle > 0\}, \\ c_i^{(t)} &:= \frac{1}{m}\sum_{j=1}^{m}\rho_{\tilde{y}_i,j,i}^{(t)} = \lambda_i \sum_{v\in[0,t-1]}\pi_i^{(v)}g_i^{(v)}.\end{aligned}$$

By Lemma C5, we already have

$$c_i^{(0)} = 0, \qquad c_i^{(t+1)} - c_i^{(t)} = \lambda_i\pi_i^{(t)}g_i^{(t)}.$$

The coefficient $\lambda_i\pi_i^{(t)}$ is the rate at which sample i learns its noise. The next lemma shows that these rates are nearly equal across samples.

Lemma C13 (Nearly equal noise-learning rates)**.** *Fix a constant $M \geq 1$ and an integer $u \in [1, T+1]$, and suppose*

$$\max_{s,j,i}\rho_{s,j,i}^{(t)} \leq M\log(2 + \eta T/m), \qquad t \in [0, u-1].$$

For $t \in [0, u-1]$ and $i \in [n]$, the noise outputs and active fractions satisfy

$$\begin{aligned} |\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i) - c_i^{(t)}| &\le C\left[\sigma_0(1+\sigma_p\sqrt{d})\log d + \frac{Mn\log d\,\log(2+\eta T/m)}{\sqrt{d}}\right] = o(1), \\ \left|\pi_i^{(t)} - \frac{1}{2}\right| &\le C\left[\frac{Mn\log d\,\log(2+\eta T/m)}{\sigma_0\sigma_p d} + \sqrt{\frac{\log d}{m}}\right] = o(1). \end{aligned} \tag{50}$$

Consequently, all samples have nearly equal noise-learning rates:

$$\lambda_i \pi_i^{(t)} = \frac{\eta\sigma_p^2 d}{2mn}\big(1 + o(1)\big). \tag{51}$$

Proof. From Lemma C5, the inner product with $\boldsymbol{\xi}_i$ splits into its initial value, sample i's own update, and interference from other samples:

$$\begin{aligned} \langle \boldsymbol{w}_{s,j}^{(t)}, \boldsymbol{\xi}_i \rangle &= \langle \boldsymbol{w}_{s,j}^{(0)}, \boldsymbol{\xi}_i \rangle + s\tilde{y}_i \rho_{s,j,i}^{(t)} + X_{s,j,i}(t), \\ X_{s,j,i}(t) &:= \sum_{\ell \ne i} s\tilde{y}_\ell \rho_{s,j,\ell}^{(t)} \frac{\langle \boldsymbol{\xi}_\ell, \boldsymbol{\xi}_i \rangle}{\|\boldsymbol{\xi}_\ell\|^2}. \end{aligned}$$

By (24) and the assumed bound on learned noise,

$$\begin{aligned} |X_{s,j,i}(t)| &\le \max_\ell \rho_{s,j,\ell}^{(t)} \sum_{\ell \ne i} \frac{|\langle \boldsymbol{\xi}_\ell, \boldsymbol{\xi}_i \rangle|}{\|\boldsymbol{\xi}_\ell\|^2} \\ &\le \frac{CMn\log d\,\log(2+\eta T/m)}{\sqrt{d}}, \qquad t \in [0, u-1]. \end{aligned}$$

For the activation fraction to change, (1) an initially active neuron must become inactive or (2) an initially inactive neuron must become active. In either case, the initial projection must be no larger in magnitude than the interference bound:

$$\left|\langle \boldsymbol{w}_{\tilde{y}_i,j}^{(0)}, \boldsymbol{\xi}_i \rangle\right| \le \frac{CMn\log d\,\log(2+\eta T/m)}{\sqrt{d}}.$$

Applying (25) gives

$$|\pi_i^{(t)} - \pi_i^{(0)}| \le C\left(\frac{Mn\log d\,\log(2+\eta T/m)}{\sigma_0\sigma_p d} + \sqrt{\frac{\log d}{m}}\right).$$

From (B9), the initial active fraction also satisfies

$$\left|\pi_i^{(0)} - \frac{1}{2}\right| \le C\sqrt{\frac{\log d}{m}}.$$

Combining the two bounds yields

$$\begin{aligned} \left|\pi_i^{(t)} - \frac{1}{2}\right| &\le |\pi_i^{(t)} - \pi_i^{(0)}| + \left|\pi_i^{(0)} - \frac{1}{2}\right| \\ &\le C\left[\frac{Mn\log d\,\log(2+\eta T/m)}{\sigma_0\sigma_p d} + \sqrt{\frac{\log d}{m}}\right] = o(1). \end{aligned}$$

The noise-norm bound in (24) gives

$$\lambda_i = \frac{\eta\|\boldsymbol{\xi}_i\|^2}{mn} = \frac{\eta\sigma_p^2 d}{mn}\left(1 + O\left(\sqrt{\frac{\log d}{d}}\right)\right).$$

Therefore,

$$\lambda_i \pi_i^{(t)} = \frac{\eta\sigma_p^2 d}{mn}(1+o(1))\left(\frac{1}{2} + o(1)\right) = \frac{\eta\sigma_p^2 d}{2mn}(1+o(1)).$$

Since the coefficients are nonnegative and ReLU is 1-Lipschitz, we have

$$\begin{aligned}
|\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i) - c_i^{(t)}| &\le \frac{1}{m}\sum_{j=1}^{m}\left|\phi(\langle \boldsymbol{w}_{\tilde{y}_i,j}^{(t)}, \boldsymbol{\xi}_i\rangle) - \rho_{\tilde{y}_i,j,i}^{(t)}\right| + \frac{1}{m}\sum_{j=1}^{m}\phi(\langle \boldsymbol{w}_{-\tilde{y}_i,j}^{(t)}, \boldsymbol{\xi}_i\rangle) \\
&\le \frac{1}{m}\sum_{s,j}\Big(|\langle \boldsymbol{w}_{s,j}^{(0)}, \boldsymbol{\xi}_i\rangle| + |X_{s,j,i}(t)|\Big) \\
&\le C\sigma_0\sigma_p\sqrt{d}\log d + 2\max_{s,j}|X_{s,j,i}(t)| \\
&\le C\left[\sigma_0(1+\sigma_p\sqrt{d})\log d + \frac{Mn\log d\,\log(2+\eta T/m)}{\sqrt{d}}\right] = o(1).
\end{aligned}$$

□

Why flipped samples learn more noise. At the same amount of learned noise, a flipped hard-only sample has a smaller weak-label margin than a correctly labeled hard-only sample. Its gradient weight is therefore larger, so its noise update is larger when the learning rates agree. The following lemma shows that this comparison remains valid with small errors in the margins and rates.

Lemma C14 (Noise ordering)**.** *Suppose that, for $t \in [0, v-1]$ and $i \in \mathcal{H}$,*

$$g_i^{(t)} = \frac{1}{1+\exp\{2z_i\alpha_i b_t + c_i^{(t)} + e_i(t)\}}, \qquad c_i^{(t+1)} = c_i^{(t)} + \theta_i(t) g_i^{(t)},$$

with $c_i^{(0)} = 0$, $b_t \ge 0$, $|e_i(t)| \le \delta$, and $\theta_L \le \theta_i(t) \le \theta_U$. Here v is a nonnegative integer, and $0 \le \delta \le 1/4$ and $0 < \theta_L \le \theta_U$ are fixed constants. Then

$$\min_{j\in\mathcal{F}} c_j^{(v)} \ge \max_{i\in\mathcal{C}} c_i^{(v)} - \omega, \qquad \omega = 2\delta + 2\log(\theta_U/\theta_L) + 2\theta_U. \tag{52}$$

Proof. Fix a correct sample $i \in \mathcal{C}$ and a flipped sample $j \in \mathcal{F}$. Write their noise increments as

$$\Delta c_\ell^{(t)} := c_\ell^{(t+1)} - c_\ell^{(t)} = \theta_\ell(t) g_\ell^{(t)} \ge 0, \qquad \ell \in \{i, j\}.$$

Since $z_i = 1$, $z_j = -1$, and the hard feature knowledge is nonnegative, the margin and rate bounds give

$$\begin{aligned}
\Delta c_i^{(t)} &= \frac{\theta_i(t)}{1 + e^{2\alpha_i b_t + c_i^{(t)} + e_i(t)}} \le \frac{\theta_U}{1 + e^{c_i^{(t)} - \delta}}, \\
\Delta c_j^{(t)} &= \frac{\theta_j(t)}{1 + e^{-2\alpha_j b_t + c_j^{(t)} + e_j(t)}} \ge \frac{\theta_L}{1 + e^{c_j^{(t)} + \delta}}.
\end{aligned}$$

Suppose, for a contradiction, that $c_i^{(t+1)} > c_j^{(t+1)} + \omega$ for the first time at $t+1 \in [1, v]$. Since $\Delta c_i^{(t)} \le \theta_U$ and $\Delta c_j^{(t)} \ge 0$, writing $r = \theta_U/\theta_L \ge 1$ gives

$$\begin{aligned}
c_i^{(t)} - c_j^{(t)} &= c_i^{(t+1)} - c_j^{(t+1)} - \Delta c_i^{(t)} + \Delta c_j^{(t)} \\
&> \omega - \Delta c_i^{(t)} + \Delta c_j^{(t)} \\
&\ge \omega - \theta_U \ge 2\delta + 2\log r.
\end{aligned}$$

With $u = c_j^{(t)} + \delta \ge 0$, this implies $c_i^{(t)} - \delta \ge u + 2\log r$, so

$$\begin{aligned}
\frac{\Delta c_i^{(t)}}{\Delta c_j^{(t)}} &\le \frac{\theta_U}{\theta_L}\frac{1 + e^{c_j^{(t)} + \delta}}{1 + e^{c_i^{(t)} - \delta}} \\
&= \frac{r(1+e^u)}{1 + e^{c_i^{(t)} - \delta}} \le \frac{r(1+e^u)}{1 + r^2 e^u} \le 1.
\end{aligned}$$

The last inequality uses $r \ge 1$ and $e^u \ge 1$. Since $c_i^{(0)} = c_j^{(0)} = 0$, the first-crossing assumption now gives the contradiction

$$\omega < c_i^{(t+1)} - c_j^{(t+1)} \le c_i^{(t)} - c_j^{(t)} \le \omega.$$

Since the bound holds for every correct–flipped pair, taking the largest correct-sample coefficient and the smallest flipped-sample coefficient proves (52). □

A positive lower bound for the hard feature knowledge. The previous noise comparison limits how much more noise correct samples can learn than flipped samples. Together with $R > 1$, this makes the hard update nonnegative near zero. We also show that a larger response cannot fall below a positive threshold in one update.

In the application below, δ combines the noise-output error from Lemma C13 with the $O(\sigma_0 \log d)$ feature-output error; both vanish under the logarithmic noise bound, so $\delta = o(1)$. The same lemma gives $\theta_U/\theta_L = 1 + o(1)$ and $\theta_U = o(1)$, so (52) yields

$$\omega = 2\delta + 2\log(\theta_U/\theta_L) + 2\theta_U = o(1).$$

The following lemma shows that the learned hard feature knowledge is preserved.

Lemma C15 (Keeping the hard feature knowledge positive)**.** *For $t \in [0, T]$,*

$$\beta = \frac{\log R}{4\alpha_{\max}} - o(1) > 0, \qquad b_{t+1} \geq \min\{b_t, \beta\}. \tag{53}$$

Proof. We use the following definitions and bounds:

$$c_C = \max_{i \in \mathcal{C}} c_i, \qquad c_F = \min_{i \in \mathcal{F}} c_i, \qquad R = \frac{A_C}{A_F} > 1,$$

$$c_C \leq c_F + \omega, \qquad |e_i| \leq \delta, \qquad \beta := \frac{\log R - \omega - 2\delta}{4\alpha_{\max}} > 0.$$

Write $b_{t+1} = Q(b_t)$, where $Q(x) = x + \frac{2\eta\pi}{mn}\psi(x)$ and

$$\psi(x) = \sum_{i \in \mathcal{H}} \frac{\alpha_i z_i}{1 + e^{2z_i\alpha_i x + c_i + e_i}} + \frac{1}{2}\sum_{i \in \mathcal{O}} \frac{\alpha_i}{1 + e^{a + \alpha_i x + d_i}}.$$

First consider $0 \leq x \leq \beta$. Using $\alpha_i \leq \alpha_{\max}$, the bounds on c_i, e_i, and the nonnegative overlap contribution gives

$$\psi(x) \geq \frac{A_C}{1 + e^{2\alpha_{\max}x + c_C + \delta}} - \frac{A_F}{1 + e^{-2\alpha_{\max}x + c_F - \delta}}.$$

Since $c_C \leq c_F + \omega$ and $x, \omega, \delta \geq 0$, the two denominators satisfy

$$\begin{aligned} 1 + e^{2\alpha_{\max}x + c_C + \delta} &\leq 1 + e^{2\alpha_{\max}x + c_F + \omega + \delta} \\ &\leq e^{4\alpha_{\max}x + \omega + 2\delta}\left(1 + e^{-2\alpha_{\max}x + c_F - \delta}\right). \end{aligned}$$

Therefore,

$$\psi(x) \geq \frac{A_C - A_F e^{4\alpha_{\max}x + \omega + 2\delta}}{1 + e^{2\alpha_{\max}x + c_C + \delta}} \geq 0,$$

where the last inequality follows from the definition of β:

$$4\alpha_{\max}x + \omega + 2\delta \leq 4\alpha_{\max}\beta + \omega + 2\delta = \log R = \log\frac{A_C}{A_F}.$$

Thus the correct-label contribution is at least as large as the flipped-label contribution, so $Q(x) \geq x$ on $[0, \beta]$.

To handle $x \geq \beta$, we show that Q is increasing. Differentiating ψ and using $e^u/(1+e^u)^2 \leq 1/4$ gives

$$0 \leq -\psi'(x) \leq \frac{1}{2}\sum_{i \in \mathcal{H}} \alpha_i^2 + \frac{1}{8}\sum_{i \in \mathcal{O}} \alpha_i^2 \leq \frac{n\alpha_{\max}^2}{2}.$$

Since $\eta \leq m/2$, $\pi \leq 1$, and $\alpha_{\max} \leq 1$, we obtain

$$Q'(x) = 1 + \frac{2\eta\pi}{mn}\psi'(x) \geq 1 - \frac{\eta\pi\alpha_{\max}^2}{m} \geq \frac{1}{2}.$$

Hence, for $x \geq \beta$, monotonicity and the first part give

$$Q(x) \geq Q(\beta) \geq \beta.$$

Combining this with $Q(x) \geq x$ on $[0, \beta]$ and taking $x = b_t$ proves the claim. □

Bounds through the full training period. Using the feature preservation in Lemma C15, we bound b_t and the learned noise through T. We first obtain these bounds before a possible first violation of the noise bound, then show that such a violation cannot occur.

Lemma C16 (Noise memorization)**.** *For $t \in [0, T]$, $s \in \{\pm 1\}$, $i \in [n]$, and $j \in [m]$,*

$$0 \le \rho_{s,j,i}^{(t)} \le \lambda_i \sum_{v \in [0,t-1]} g_i^{(v)} \le C \log(2 + \eta T/m). \tag{54}$$

For $t \in [0, T]$,

$$b_t \le \frac{1}{\alpha_{\min}} \log(1 + C\eta t/m). \tag{55}$$

Proof. For a constant $M \ge 1$, let τ be the first violation of the noise bound:

$$\tau = \min \left\{ t \in [0, T] : \max_{s,j,i} \rho_{s,j,i}^{(t)} > M \log(2 + \eta T/m) \right\}.$$

Set $\tau = T + 1$ if no violation occurs. From Lemma C13, for every sample and $t \in [0, \tau - 1]$,

$$|\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i) - c_i^{(t)}| + C\sigma_0 \log d \le \delta = o(1),$$
$$0 < \theta_L \le \lambda_i \pi_i^{(t)} \le \theta_U = o(1), \qquad \theta_U/\theta_L = 1 + o(1).$$

Hence, noise memorization gap between flipped and correct samples is bounded by

$$\omega = 2\delta + 2\log(\theta_U/\theta_L) + 2\theta_U = o(1).$$

From Lemmas C14 and C15, induction from $b_0 = 0$ gives $b_{t+1} \ge \min\{b_t, \beta\} \ge 0$ for $t \in [0, \tau - 1]$. For all $v \in [0, \tau - 1]$, the weak-label margins satisfy

$$\tilde{y}_i F_{\boldsymbol{W}^{(v)}}(\boldsymbol{X}_i) = \begin{cases} 2a_v, & \tau_i = \mathrm{E}; \\ a_v + \alpha_i b_v, & \tau_i = \mathrm{O}; \\ 2z_i \alpha_i b_v, & \tau_i = \mathrm{H}, \end{cases} + c_i^{(v)} + e_i(v), \qquad |e_i(v)| \le \delta. \tag{56}$$

For $t \in [0, \tau - 1]$, dropping the negative terms in the hard feature update and using $g(u) \le e^{-u}$ gives

$$b_{t+1} - b_t \le \frac{C\eta}{mn} \left(2 \sum_{i \in \mathcal{C}} \alpha_i e^{-2\alpha_i b_t} + \sum_{i \in \mathcal{O}} \alpha_i e^{-\alpha_i b_t} \right) \le \frac{C\eta}{m} e^{-\alpha_{\min} b_t}.$$

For positive increments, $e^x - 1 \le x e^x$ and $\eta/m \le 1/2$ imply

$$e^{\alpha_{\min} b_{t+1}} - e^{\alpha_{\min} b_t} = e^{\alpha_{\min} b_t} \left(e^{\alpha_{\min}(b_{t+1} - b_t)} - 1 \right) \le \frac{C\eta}{m}.$$

The same bound holds for nonpositive increments. Summing from $b_0 = 0$ gives

$$b_t \le \frac{1}{\alpha_{\min}} \log(1 + C\eta t/m), \qquad t \in [0, \tau].$$

We use this upper bound to control noise learning. Only flipped hard-only samples have a negative feature contribution in (56). Define

$$D_T = \frac{2\alpha_{\max}}{\alpha_{\min}} \log(1 + C\eta T/m) + 1 \le C \log(2 + \eta T/m).$$

For every sample and $t \in [0, \tau - 1]$, the weak-label margins satisfy

$$\tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) \ge c_i^{(t)} - D_T, \qquad g_i^{(t)} \le \min\{1, e^{D_T - c_i^{(t)}}\}.$$

From Lemma C13, for sufficiently large d,

$$\lambda_0 := \frac{\eta \sigma_p^2 d}{mn} = o(1), \qquad \lambda_i \pi_i^{(t)} = \left(\frac{1}{2} + o(1) \right) \lambda_0 \le \lambda_0.$$

Hence, the noise increment satisfies

$$0 \le c_i^{(t+1)} - c_i^{(t)} = \lambda_i \pi_i^{(t)} g_i^{(t)} \le \lambda_0 \min\{1, e^{D_T - c_i^{(t)}}\}.$$

Each increment is at most $\lambda_0 = o(1)$, so we can apply $e^u - 1 \le Cu$ for $0 \le u \le \lambda_0$:

$$\begin{aligned} e^{c_i^{(t+1)}} - e^{c_i^{(t)}} &= e^{c_i^{(t)}} \left(e^{c_i^{(t+1)} - c_i^{(t)}} - 1 \right) \\ &\le C e^{c_i^{(t)}} \left(c_i^{(t+1)} - c_i^{(t)} \right) \\ &\le C e^{c_i^{(t)}} \lambda_0 e^{D_T - c_i^{(t)}} = C \lambda_0 e^{D_T}. \end{aligned}$$

Summing from $c_i^{(0)} = 0$ and taking logarithms gives, for $t \in [0, \tau]$,

$$\begin{aligned} e^{c_i^{(t)}} &\le 1 + C\lambda_0 t e^{D_T}, \\ c_i^{(t)} &\le \log(1 + C\lambda_0 t e^{D_T}) \le D_T + \log(1 + C\lambda_0 t). \end{aligned}$$

Since $t \le \tau \le T + 1 \le 2T$ and $\sigma_p^2 d/n \le 1$,

$$\lambda_0 t = \frac{\eta t}{m} \frac{\sigma_p^2 d}{n} \le \frac{2\eta T}{m}.$$

Together with $D_T \le C \log(2 + \eta T/m)$, this gives the logarithmic bound

$$c_i^{(t)} \le D_T + \log(1 + 2C\eta T/m) \le C_c \log(2 + \eta T/m).$$

Here C_c is independent of M, since we used only $|e_i(t)| \le 1$ and $\lambda_i \pi_i^{(t)} \le \lambda_0$.

The average learned noise also controls each neuron's noise coefficient because $\pi_i^{(v)} \ge 1/3$ for $v < \tau$. Thus, for all s, j, i and $t \in [0, \tau]$,

$$\begin{aligned} \lambda_i \sum_{v \in [0, t-1]} g_i^{(v)} &\le 3\lambda_i \sum_{v \in [0, t-1]} \pi_i^{(v)} g_i^{(v)} = 3c_i^{(t)}, \\ \rho_{s,j,i}^{(t)} &\le \lambda_i \sum_{v \in [0, t-1]} g_i^{(v)} \le 3C_c \log(2 + \eta T/m). \end{aligned}$$

Choose $M > 3C_c$. If a first violation occurred at $\tau \le T$, then

$$M \log(2 + \eta T/m) < \max_{s,j,i} \rho_{s,j,i}^{(\tau)} \le 3C_c \log(2 + \eta T/m) < M \log(2 + \eta T/m).$$

This contradiction shows that $\tau = T + 1$, so both upper bounds hold through T. □

Conclusion for noise memorization. For both classes:

- **Noise memorization grows at most logarithmically.**

$$\max_{s,j,i} \rho_{s,j,i}^{(t)} \le C \log(2 + \eta T/m), \qquad t \in [0, T].$$

- **Flipped samples learn more noise.**

$$\min_{i \in \mathcal{F}} c_i^{(t)} \ge \max_{i \in \mathcal{C}} c_i^{(t)} - o(1), \qquad t \in [0, T].$$

- **Hard features are retained.**

$$b_t \ge c\widehat{b}, \qquad t \in [T_{\text{noise}}, T].$$

The last bound proves Theorem 2, since $\Pr(\mathcal{G}) \ge 1 - 1/\text{poly}(d)$.

C.11 Fitting All Weak Labels

We show that every weak training label is fitted under (A7). We first use the noise bound to control the average training loss, then show that the loss cannot increase after T_{noise}.

Lemma C17 (Late training loss)**.** *Let* $\mathcal{I}_T = [\lceil T/2 \rceil, T-1]$. *Then*

$$\min_{t\in\mathcal{I}_T} \mathcal{L}(\boldsymbol{W}^{(t)}) \le \frac{1}{|\mathcal{I}_T|} \sum_{t\in\mathcal{I}_T} \mathcal{L}(\boldsymbol{W}^{(t)}) \le \frac{Cmn\log^2(2+\eta T/m)}{\eta\sigma_p^2 dT}. \tag{57}$$

Proof. For each sample, write its weak-label margin as $u_i^{(t)} = \tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)$. The logistic loss and its gradient weight satisfy

$$\mathcal{L}(\boldsymbol{W}^{(t)}) = \frac{1}{n}\sum_{i=1}^n \ell(u_i^{(t)}), \qquad \ell(u) = \log(1+e^{-u}),$$
$$g_i^{(t)} = g(u_i^{(t)}), \qquad g(u) = -\ell'(u) = \frac{1}{1+e^u}.$$

The weak-label margin has the following lower bound:

$$u_i^{(t)} \ge -2\alpha_{\max} b_{y_i}^{(t)} - 1 \ge -D_T,$$

where

$$D_T = \frac{2\alpha_{\max}}{\alpha_{\min}} \log(1 + C\eta T/m) + 1.$$

For $u \ge 0$, we have

$$\ell(u) = \log(1+e^{-u}) \le e^{-u} = (1+e^{-u})g(u) \le 2g(u).$$

For $-D_T \le u < 0$, we have

$$\ell(u) = -u + \log(1+e^u) \le D_T + \log 2 \le 2(1+D_T)g(u),$$

since $g(u) \ge 1/2$ for $u < 0$. Hence

$$\ell(u_i^{(t)}) \le 2(1+D_T)g_i^{(t)} \le C\log(2+\eta T/m) g_i^{(t)}.$$

From (B6), we have

$$\|\boldsymbol{\xi}_i\|^2 \ge c\sigma_p^2 d, \qquad \lambda_i = \frac{\eta\|\boldsymbol{\xi}_i\|^2}{mn} \ge \frac{c\eta\sigma_p^2 d}{mn}.$$

Combining this with (54) gives

$$\sum_{t\in[0,T-1]} g_i^{(t)} \le \frac{C\log(2+\eta T/m)}{\lambda_i} \le \frac{Cmn\log(2+\eta T/m)}{\eta\sigma_p^2 d}, \qquad i \in [n]. \tag{58}$$

We finally obtain

$$\begin{aligned}
\min_{t\in\mathcal{I}_T} \mathcal{L}(\boldsymbol{W}^{(t)}) &\le \frac{1}{|\mathcal{I}_T|}\sum_{t\in\mathcal{I}_T} \mathcal{L}(\boldsymbol{W}^{(t)}) \\
&\le \frac{C\log(2+\eta T/m)}{n|\mathcal{I}_T|} \sum_{i=1}^n \sum_{t\in[0,T-1]} g_i^{(t)} \\
&\le \frac{Cmn\log^2(2+\eta T/m)}{\eta\sigma_p^2 dT}.
\end{aligned}$$

□

Keeping every label fitted. A small total loss $n\mathcal{L}$ makes every individual loss small. We next show that continued training cannot increase this loss.

Lemma C18 (Nonincreasing training loss)**.** *For $t \in [T_{\text{noise}}, T-1]$ and sufficiently large d,*

$$\mathcal{L}(\boldsymbol{W}^{(t+1)}) \le \mathcal{L}(\boldsymbol{W}^{(t)}).$$

Proof. Fix $t \in [T_{\text{noise}}, T-1]$ and write

$$\begin{aligned} u_i &= \tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i), \qquad g_i = \frac{1}{1+e^{u_i}}, \\ \Delta u_i &= \tilde{y}_i\big(F_{\boldsymbol{W}^{(t+1)}}(\boldsymbol{X}_i) - F_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\big), \\ \Delta \boldsymbol{u} &= (\Delta u_i)_{i=1}^n, \qquad \boldsymbol{g} = (g_i)_{i=1}^n. \end{aligned}$$

Since $\ell'(u_i) = -g_i$ and $0 \le \ell'' \le 1/4$, Taylor's inequality gives

$$\begin{aligned} n\big(\mathcal{L}(\boldsymbol{W}^{(t+1)}) - \mathcal{L}(\boldsymbol{W}^{(t)})\big) &= \sum_{i=1}^n \big(\ell(u_i+\Delta u_i) - \ell(u_i)\big) \\ &\le -\boldsymbol{g}^\top \Delta \boldsymbol{u} + \frac{1}{8}\|\Delta \boldsymbol{u}\|^2. \end{aligned}$$

Let $\boldsymbol{V} \in \mathbb{R}^{n\times 4}$ have columns ordered as $(\boldsymbol{\mu}_{+1}, \boldsymbol{\nu}_{+1}, \boldsymbol{\mu}_{-1}, \boldsymbol{\nu}_{-1})$. In row i, the true-class block is $(2,0)$, $(1,\alpha_i)$, or $(0, 2z_i\alpha_i)$ for an easy-only, overlap, or hard-only sample, respectively; the other block is zero. Define

$$\begin{aligned} \boldsymbol{D} &= \operatorname{diag}\Big(\pi_{e,+1}^{(0)}, \pi_{h,+1}^{(0)}, \pi_{e,-1}^{(0)}, \pi_{h,-1}^{(0)}\Big), \\ \boldsymbol{q}_s &= \big(a_{+1}^{(s)}, b_{+1}^{(s)}, a_{-1}^{(s)}, b_{-1}^{(s)}\big)^\top, \qquad s \in \{t, t+1\}, \\ \boldsymbol{A} &= \frac{\eta}{mn}\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top. \end{aligned}$$

The early feature-growth bounds, feature preservation, and (26) give the exact updates

$$\begin{aligned} \Delta \boldsymbol{q} := \boldsymbol{q}_{t+1} - \boldsymbol{q}_t &= \frac{\eta}{mn}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{g}, \\ \Delta \boldsymbol{u}_{\text{feat}} = \boldsymbol{V}\Delta \boldsymbol{q} &= \frac{\eta}{mn}\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{g} = \boldsymbol{A}\boldsymbol{g}. \end{aligned}$$

Since every row of $\boldsymbol{V}$ has squared norm at most 4 and $0 \preceq \boldsymbol{D} \preceq \boldsymbol{I}_4$,

$$\boldsymbol{A} \succeq 0, \qquad \|\boldsymbol{A}\|_{\text{op}} \le \operatorname{tr}(\boldsymbol{A}) \le \frac{4\eta}{m} \le 2.$$

For the noise updates, write

$$\begin{aligned} v_i &= \tilde{y}_i\big(F_{\boldsymbol{W}^{(t+1)}}(\boldsymbol{\xi}_i) - F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i)\big), \qquad \boldsymbol{v} = (v_i)_{i=1}^n, \\ \Delta \boldsymbol{u} &= \boldsymbol{A}\boldsymbol{g} + \boldsymbol{v}. \end{aligned}$$

Define the interference matrix by

$$E_{ii} = 0, \qquad E_{i\ell} = \frac{\eta|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_\ell\rangle|}{mn} \quad (i \ne \ell).$$

From (26) and the 1-Lipschitz property of ReLU,

$$v_i \ge \lambda_i \pi_i^{(t)} g_i - 2(\boldsymbol{E}\boldsymbol{g})_i, \qquad |v_i| \le 2\lambda_i g_i + 2(\boldsymbol{E}\boldsymbol{g})_i.$$

Define

$$\lambda_0 = \frac{\eta\sigma_p^2 d}{mn} = o(1).$$

From (24) and Lemmas C13 and C16,

$$\begin{aligned} \|\boldsymbol{E}\|_{\text{op}} \le \max_i \sum_{\ell\ne i} E_{i\ell} \le C\lambda_0 \frac{n\log d}{\sqrt{d}} \le \frac{\lambda_0}{24}, \\ \frac{\lambda_0}{2} \le \lambda_i \le 2\lambda_0, \qquad \pi_i^{(t)} \ge \frac{1}{3}. \end{aligned}$$

The two bounds for v_i therefore give

$$\boldsymbol{g}^\top \boldsymbol{v} \geq \sum_i \lambda_i \pi_i^{(t)} g_i^2 - 2\boldsymbol{g}^\top \boldsymbol{E}\boldsymbol{g} \geq \left(\frac{\lambda_0}{6} - \frac{\lambda_0}{12}\right)\|\boldsymbol{g}\|^2 = \frac{\lambda_0}{12}\|\boldsymbol{g}\|^2,$$
$$\|\boldsymbol{v}\| \leq 2\max_i \lambda_i \|\boldsymbol{g}\| + 2\|\boldsymbol{E}\|_{\mathrm{op}}\|\boldsymbol{g}\| \leq 5\lambda_0\|\boldsymbol{g}\|.$$

Since $\boldsymbol{A} \succeq 0$,

$$\|\boldsymbol{A}\boldsymbol{g} + \boldsymbol{v}\|^2 \leq 2\|\boldsymbol{A}\boldsymbol{g}\|^2 + 2\|\boldsymbol{v}\|^2 \leq 2\|\boldsymbol{A}\|_{\mathrm{op}}\boldsymbol{g}^\top \boldsymbol{A}\boldsymbol{g} + 2\|\boldsymbol{v}\|^2.$$

Substituting into the loss bound, we finally obtain

$$\begin{aligned} n\big(\mathcal{L}(\boldsymbol{W}^{(t+1)}) - \mathcal{L}(\boldsymbol{W}^{(t)})\big) &\leq -\left(1 - \frac{\|\boldsymbol{A}\|_{\mathrm{op}}}{4}\right)\boldsymbol{g}^\top \boldsymbol{A}\boldsymbol{g} - \boldsymbol{g}^\top \boldsymbol{v} + \frac{1}{4}\|\boldsymbol{v}\|^2 \\ &\leq -\frac{1}{2}\boldsymbol{g}^\top \boldsymbol{A}\boldsymbol{g} - \left(\frac{\lambda_0}{12} - \frac{25\lambda_0^2}{4}\right)\|\boldsymbol{g}\|^2 \\ &\leq -\frac{1}{2}\boldsymbol{g}^\top \boldsymbol{A}\boldsymbol{g} - \frac{\lambda_0}{24}\|\boldsymbol{g}\|^2 \leq 0. \end{aligned}$$

The last inequality holds for $\lambda_0 \leq 1/150$, since $\lambda_0 = o(1)$. □

C.12 Weak-to-Strong Generalization

The loss bounds show that the student fits every weak training label. The preserved features give correct predictions on independent test samples with high probability, since memorized training noise contributes only a vanishing error.

Theorem 3. *Suppose $\mathcal{G}$ holds. With probability at least $1 - 1/\mathrm{poly}(d)$, the final student satisfies*

$$\tilde{y}_i F_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}_i) > 0, \qquad i \in [n],$$
$$\Pr_{(\boldsymbol{X},y)}\left(yF_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}) < 0\right) \leq \frac{1}{\mathrm{poly}(d)}.$$

Proof. **Fitting weak labels.** Each sample's loss is at most the total training loss:

$$0 \leq \ell(\tilde{y}_i F_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}_i)) \leq n\mathcal{L}(\boldsymbol{W}^{(T)}).$$

Since $\ell(0) = \log 2$ and ℓ is decreasing, it suffices to show that the total loss is below $\log 2$. Since $\mathcal{I}_T \subset [T_{\mathrm{noise}}, T-1]$, Lemmas C17 and C18 give

$$\begin{aligned} n\mathcal{L}(\boldsymbol{W}^{(T)}) &\leq \frac{n}{|\mathcal{I}_T|}\sum_{t\in\mathcal{I}_T}\mathcal{L}(\boldsymbol{W}^{(t)}) \\ &\leq \frac{Cmn^2\log^2(2+\eta T/m)}{\eta\sigma_p^2 dT} \leq \frac{Cn(\log d)^2}{d} = o(1). \end{aligned} \tag{59}$$

The last step follows from (A7), (14), and $n = o(\sqrt{d}/\log d)$. Thus, for sufficiently large d, every sample's loss is below $\log 2$, so

$$\tilde{y}_i F_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}_i) > 0, \qquad i \in [n]. \tag{60}$$

Generalization. By (B10), the independent test noise $\boldsymbol{\xi}_*$ satisfies

$$|\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_*\rangle| \leq C_0\sigma_0\sigma_p\sqrt{d}\log d, \qquad |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_*\rangle| \leq C_0\sigma_p^2\sqrt{d}\log d. \tag{61}$$

The weight decomposition gives

$$|\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\xi}_*\rangle| \leq |\langle \boldsymbol{w}_{k,j}^{(0)}, \boldsymbol{\xi}_*\rangle| + \sum_{i=1}^n \rho_{k,j,i}^{(t)} \frac{|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_*\rangle|}{\|\boldsymbol{\xi}_i\|^2}.$$

Using these bounds, $\|\boldsymbol{\xi}_i\|^2 \asymp \sigma_p^2 d$, and Lemma C16 gives

$$\begin{aligned}\sup_{t\in[0,T]} |F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_*)| &\le 2\max_{k,j} \sup_{t\in[0,T]} |\langle \boldsymbol{w}_{k,j}^{(t)}, \boldsymbol{\xi}_*\rangle| \\ &\le C\left[\sigma_0\sigma_p\sqrt{d}\log d + \frac{n\log d\,\log(2+\eta T/m)}{\sqrt{d}}\right] = o(1).\end{aligned} \tag{62}$$

The last inequality follows from (16).

Since $a_k^{(t)}$ never decreases and hard feature knowledge is preserved, the feature knowledges satisfy

$$a_k^{(T)} \ge a_k^{(T_{\text{noise}})} \ge c_a > 0, \qquad b_k^{(T)} \ge c_b b_h > 0.$$

For a new sample with $y = k$, the feature and noise bounds give

$$yF_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}) = \begin{cases} 2a_k^{(T)}, & \tau = \mathrm{E}; \\ a_k^{(T)} + \alpha b_k^{(T)}, & \tau = \mathrm{O}; \\ 2\alpha b_k^{(T)}, & \tau = \mathrm{H}, \end{cases} \quad + o(1) > 0. \tag{63}$$

The error tends to zero, so the margin stays positive and

$$\Pr_{(\boldsymbol{X},y)}(yF_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}) < 0) \le \frac{1}{\mathrm{poly}(d)}.$$

□

C.13 Zero Hard Feature Knowledge

We consider $b_h = 0$ and retain (A1) and (A3)–(A7), with correct labels on easy-only and overlap samples. Assume $\mathcal{G}$ with (B2) replaced by $A_{C,k} = A_{F,k} > 0$ for both classes, and take $p_O > 0$. Fix a class k and omit its subscript. For a quantitative bound, strengthen the parameter conditions to

$$\begin{aligned}\frac{\sigma_p^2 d}{n} \le \frac{1}{(\log d)^5}, \qquad \sigma_0(1+\sigma_p\sqrt{d}) \le \frac{1}{(\log d)^4}, \\ n \le \frac{\sigma_0\sigma_p d}{(\log d)^4}, \qquad m \ge (\log d)^5.\end{aligned} \tag{64}$$

Choose

$$T_{\text{noise}} := \left\lceil \frac{m}{\eta}(\log d)^2 \right\rceil, \qquad \varepsilon_d := \frac{1}{(\log d)^2}. \tag{65}$$

Then $\delta_{T_{\text{noise}}} = O((\log d)^{-3}) = o(\varepsilon_d)$ in (39) and $T_{\text{noise}} < T$.

Strength balance makes the hard-only saturation point zero, and the early-learning argument shows that b_t approaches it.

Lemma C19 (Approaching zero)**.** *Suppose $A_C = A_F$. Then, for $t \in [0, T_{\text{noise}}]$,*

$$0 \le b_t \le \frac{C}{1+\eta t/m} + \frac{C}{(\log d)^3}. \tag{66}$$

In particular, $b_{T_{\text{noise}}} \le C\varepsilon_d$.

Proof. Since $f_H(0) = (A_C - A_F)/2 = 0$ and f_H is strictly decreasing, $\widehat{b} = 0$. We first show that b_t stays in $[0, B_0]$ for a sufficiently large constant B_0. On an induction prefix in this interval, the easy-feature estimates in the proof of Lemma C8 give

$$\frac{f(a_t, 0)}{n} = \frac{A_O}{2n(1+e^{a_t})} \ge \frac{cp_O}{1+\eta t/m} \ge c\varepsilon_d \gg \delta_{T_{\text{noise}}}.$$

Also, $A_F = A_C \asymp n$ gives $f(a_t, B_0) \le -cn$ as in the proof of Lemma C7. The same increasing update map preserves $[0, B_0]$, since these boundary signs dominate the error.

Next, the hard-only update moves b_t toward zero, while the overlap contribution decreases. From (46) and the easy-feature estimates,

$$f_H(b_t) = \int_0^{b_t} f'_H(u)\,du \le -cnb_t,$$
$$0 \le f(a_t, b_t) - f_H(b_t) \le Cnp_O e^{-a_t} \le \frac{Cn}{1+\eta t/m}.$$

Substituting into (40), we obtain

$$0 \le b_{t+1} \le \left(1 - \frac{c\eta}{m}\right) b_t + \frac{C\eta}{m}\left[\frac{1}{1+\eta t/m} + \delta_{T_{\text{noise}}}\right].$$

Iterating from $b_0 = 0$ and summing as in the proof of Lemma C12 gives

$$\begin{aligned} b_t &\le \frac{C\eta}{m}\sum_{s=0}^{t-1} \frac{(1-c\eta/m)^{t-1-s}}{1+\eta s/m} + C\delta_{T_{\text{noise}}} \\ &\le \frac{C}{1+\eta t/m} + \frac{C}{(\log d)^3}. \end{aligned}$$

□

For $t \ge T_{\text{noise}}$, we compare the hard-feature updates from correctly labeled and flipped hard-only samples. Write the hard-feature output and the weak-label-signed noise output as

$$B_t := kF_{\boldsymbol{W}^{(t)}}(\boldsymbol{\nu}_k), \qquad C_i(t) := \tilde{y}_i F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\xi}_i).$$

Holding the current noise outputs $C_i(t)$ fixed, define the strength-weighted gradient contributions as functions of the hard-feature output b:

$$\begin{aligned} G_C(t,b) &:= \sum_{i\in\mathcal{C}} \frac{\alpha_i}{1+\exp(C_i(t)+2\alpha_i b)}, \\ G_F(t,b) &:= \sum_{i\in\mathcal{F}} \frac{\alpha_i}{1+\exp(C_i(t)-2\alpha_i b)}. \end{aligned} \tag{67}$$

Here $G_C(t, B_t)$ contributes positively and $G_F(t, B_t)$ negatively to the hard-feature update. From (30), the exact update is

$$b_{t+1} - b_t = \frac{2\eta\pi_h^{(t)}}{mn}\left[G_C(t,B_t) - G_F(t,B_t) + \frac{1}{2}\sum_{i\in\mathcal{O}} \alpha_i g_i^{(t)}\right]. \tag{68}$$

With the noise outputs fixed, increasing b decreases $G_C(t,b)$ and increases $G_F(t,b)$:

$$\partial_b\big(G_C(t,b) - G_F(t,b)\big) = -2\sum_{i\in\mathcal{H}} \frac{\alpha_i^2 \exp(C_i(t)+2\alpha_i z_i b)}{\big(1+\exp(C_i(t)+2\alpha_i z_i b)\big)^2} < 0.$$

Strength balance lets us compare the two total contributions by averaging correct–flipped pairs:

$$G_C(t,b) - G_F(t,b) = \frac{1}{A_C}\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{F}} \alpha_i\alpha_j\left[g(C_i(t)+2\alpha_i b) - g(C_j(t)-2\alpha_j b)\right], \tag{69}$$

where $g(v) = (1+e^v)^{-1}$. We show that the flipped contribution dominates at an upper boundary for B_t, including the remaining overlap contribution. The lower boundary is handled in the same way.

Lemma C20 (Small hard-feature knowledge)**.** *For $t \in [T_{\text{noise}}, T]$,*

$$|b_t| = O\left(\frac{1}{\log d}\right). \tag{70}$$

Proof. Choose large constants M, K_2, K_1 in this order, and write

$$\gamma := \frac{\eta\sigma_p^2 d}{2mn}, \qquad \beta_t := K_1\varepsilon_d\,[1+\log(1+\gamma(t-T_{\mathrm{noise}}))] = O(1/\log d).$$

We prove the following bounds for both classes by induction:

$$\begin{gathered}|B_t| \le \beta_t + K_2\varepsilon_d, \qquad |c_j^{(t)} - c_i^{(t)}| \le 2(\alpha_i+\alpha_j)\beta_t \quad (i\in\mathcal{C},\ j\in\mathcal{F}),\\ \max_{s,r,i}\rho_{s,r,i}^{(t)} \le M\log(2+\eta T/m).\end{gathered} \tag{71}$$

From Lemma C19 and the feature-output identity (34), we have

$$B_{T_{\mathrm{noise}}} = O(\varepsilon_d).$$

Also,

$$0 \le c_i^{(t)}, \rho_{s,r,i}^{(t)} \le \lambda_i t = O((\log d)^{-3}), \qquad t\in[0, T_{\mathrm{noise}}].$$

Hence, the induction hypothesis holds at T_{noise}. Suppose the bounds hold through $\tau-1$, where $\tau \in [T_{\mathrm{noise}}+1, T]$. We show that they also hold at τ. From Lemma C13 and (64), before τ,

$$C_i(t) = c_i^{(t)} + e_i(t), \qquad |e_i(t)| = O((\log d)^{-3}) = o(\varepsilon_d), \qquad \lambda_i\pi_i^{(t)} = \gamma(1+O(\varepsilon_d)). \tag{72}$$

For hard-only samples, the margin is $c_i^{(t)} + 2\alpha_i z_i B_t + e_i(t)$, so

$$c_i^{(t+1)} - c_i^{(t)} = \lambda_i\pi_i^{(t)} g_i^{(t)} = \frac{\lambda_i\pi_i^{(t)}}{1+\exp\Big(c_i^{(t)} + 2\alpha_i z_i B_t + e_i(t)\Big)}. \tag{73}$$

Since $|B_t| = o(1)$ and $c_i^{(t)} \ge 0$, the noise recursion gives

$$c\gamma \le e^{c_i^{(t+1)}} - e^{c_i^{(t)}} \le C\gamma, \qquad g_i^{(t)} \asymp e^{-c_i^{(t)}} \asymp \frac{1}{1+\gamma(t-T_{\mathrm{noise}})}.$$

We first show that $c_j^{(\tau)} - c_i^{(\tau)} \le 2(\alpha_i+\alpha_j)\beta_\tau$ for all correct samples $i\in\mathcal{C}$ and flipped samples $j\in\mathcal{F}$. From (73),

$$\frac{\partial c_j^{(\tau)}}{\partial c_j^{(\tau-1)}} = 1 - \lambda_j\pi_j^{(\tau-1)} g_j^{(\tau-1)}\big(1-g_j^{(\tau-1)}\big) \ge 1 - \frac{\lambda_j\pi_j^{(\tau-1)}}{4} > 0.$$

The same bound holds for sample i. Thus, $c_j^{(\tau)} - c_i^{(\tau)}$ is largest when $c_j^{(\tau-1)} - c_i^{(\tau-1)} = 2(\alpha_i + \alpha_j)\beta_{\tau-1}$. At this boundary, the flipped margin minus the correct margin satisfies

$$\begin{aligned}&\big(c_j^{(\tau-1)} - 2\alpha_j B_{\tau-1} + e_j(\tau-1)\big) - \big(c_i^{(\tau-1)} + 2\alpha_i B_{\tau-1} + e_i(\tau-1)\big)\\ &\quad= \big(c_j^{(\tau-1)} - c_i^{(\tau-1)}\big) - 2(\alpha_i+\alpha_j)B_{\tau-1} + e_j(\tau-1) - e_i(\tau-1)\\ &\quad= 2(\alpha_i+\alpha_j)(\beta_{\tau-1} - B_{\tau-1}) + e_j(\tau-1) - e_i(\tau-1)\\ &\quad\ge -2(\alpha_i+\alpha_j)K_2\varepsilon_d - |e_j(\tau-1)| - |e_i(\tau-1)| \ge -C\varepsilon_d.\end{aligned}$$

Hence, the gradient weights at this boundary satisfy

$$g_j^{(\tau-1)} \le \frac{1}{1+\exp\Big(c_i^{(\tau-1)} + 2\alpha_i B_{\tau-1} + e_i(\tau-1) - C\varepsilon_d\Big)} \le e^{C\varepsilon_d} g_i^{(\tau-1)}.$$

Using (72) and $e^{C\varepsilon_d} - 1 = O(\varepsilon_d)$, we obtain

$$\begin{aligned}&\big(c_j^{(\tau)} - c_j^{(\tau-1)}\big) - \big(c_i^{(\tau)} - c_i^{(\tau-1)}\big)\\ &\quad= \lambda_j\pi_j^{(\tau-1)} g_j^{(\tau-1)} - \lambda_i\pi_i^{(\tau-1)} g_i^{(\tau-1)}\\ &\quad= \lambda_j\pi_j^{(\tau-1)}\big(g_j^{(\tau-1)} - g_i^{(\tau-1)}\big) + \big(\lambda_j\pi_j^{(\tau-1)} - \lambda_i\pi_i^{(\tau-1)}\big) g_i^{(\tau-1)}\\ &\quad\le \Big[\lambda_j\pi_j^{(\tau-1)}\big(e^{C\varepsilon_d}-1\big) + \big|\lambda_j\pi_j^{(\tau-1)} - \lambda_i\pi_i^{(\tau-1)}\big|\Big]\, g_i^{(\tau-1)}\\ &\quad\le C\varepsilon_d\gamma g_i^{(\tau-1)} \le \frac{C\varepsilon_d\gamma}{1+\gamma(\tau-1-T_{\mathrm{noise}})}.\end{aligned}$$

Applying the same comparison at the lower boundary gives

$$|c_j^{(\tau)} - c_i^{(\tau)}| \le 2(\alpha_i + \alpha_j)\beta_{\tau-1} + \frac{C\varepsilon_d\gamma}{1+\gamma(\tau-1-T_{\text{noise}})} \le 2(\alpha_i+\alpha_j)\beta_\tau.$$

The last inequality follows from $\beta_\tau - \beta_{\tau-1} \ge K_1\varepsilon_d\gamma/[2(1+\gamma(\tau-1-T_{\text{noise}}))]$ and the choice of K_1. Thus, the noise-difference bound holds at τ.

Next, we show that the overlap contribution diminishes. Let us define J_t, which measures the gradient contribution from overlap samples, as

$$J_t := \frac{1}{n}\sum_{i\in\mathcal{O}} e^{-kF_{\boldsymbol{W}^{(t)}}(\boldsymbol{\mu}_k) - c_i^{(t)}} \asymp \frac{1}{n}\sum_{i\in\mathcal{O}} g_i^{(t)}.$$

The comparison holds for $T_{\text{noise}} \le t < \tau$, since $|B_t| = o(1)$ and the noise-output errors are small. All overlap labels are correct, so their updates increase the easy-feature output:

$$k\big(F_{\boldsymbol{W}^{(t+1)}}(\boldsymbol{\mu}_k) - F_{\boldsymbol{W}^{(t)}}(\boldsymbol{\mu}_k)\big) \ge \frac{c\eta}{m}J_t.$$

Since $c_i^{(t+1)} \ge c_i^{(t)}$, every term in the sum defining J_t shrinks by a factor of at most $e^{-(c\eta/m)J_t}$. Thus,

$$J_{t+1} \le e^{-(c\eta/m)J_t}J_t.$$

Taking reciprocals and using $e^x \ge 1+x$ gives

$$\frac{1}{J_{t+1}} \ge \frac{e^{(c\eta/m)J_t}}{J_t} \ge \frac{1+(c\eta/m)J_t}{J_t} = \frac{1}{J_t} + \frac{c\eta}{m}.$$

Thus $1/J_t$ increases by at least $c\eta/m$ at each step. Starting from $J_{T_{\text{noise}}} \le C\varepsilon_d$, we obtain

$$J_t \le \frac{1}{J_{T_{\text{noise}}}^{-1} + (c\eta/m)(t-T_{\text{noise}})} \le \frac{C}{1+\eta t/m}.$$

Consequently, at $\tau-1$, the overlap contribution to the hard-feature update satisfies

$$\frac{1}{2}\sum_{i\in\mathcal{O}}\alpha_i g_i^{(\tau-1)} \le CnJ_{\tau-1} \le \frac{Cn}{1+\eta(\tau-1)/m} \le \frac{Cn\varepsilon_d}{1+\gamma(\tau-1-T_{\text{noise}})}. \tag{74}$$

We next bound B_τ. Its update satisfies

$$B_\tau = B_{\tau-1} + \kappa_{\tau-1}\left[G_C(\tau-1,B_{\tau-1}) - G_F(\tau-1,B_{\tau-1}) + \frac{1}{2}\sum_{i\in\mathcal{O}}\alpha_i g_i^{(\tau-1)}\right],$$

where $0 \le \kappa_{\tau-1} \le 4\eta/(mn)$. For a clean sample $i\in\mathcal{C}$, a flipped sample $j\in\mathcal{F}$, and $b = \beta_{\tau-1} + K_2\varepsilon_d$, we have

$$\big(C_i(\tau-1)+2\alpha_i b\big) - \big(C_j(\tau-1) - 2\alpha_j b\big) \ge 2(\alpha_i+\alpha_j)K_2\varepsilon_d - o(\varepsilon_d) \ge cK_2\varepsilon_d.$$

Using $e^x - 1 \ge x$ and the bounds on the gradient weights, we obtain

$$\begin{aligned} &g\big(C_i(\tau-1)+2\alpha_i b\big) - g\big(C_j(\tau-1)-2\alpha_j b\big) \\ &\quad= -g\big(C_i(\tau-1)+2\alpha_i b\big)\big[1 - g\big(C_j(\tau-1)-2\alpha_j b\big)\big] \\ &\qquad\times\Big(e^{C_i(\tau-1)-C_j(\tau-1)+2(\alpha_i+\alpha_j)b} - 1\Big) \\ &\quad\le -\frac{cK_2\varepsilon_d}{1+\gamma(\tau-1-T_{\text{noise}})}. \end{aligned}$$

From (69) and $A_C = A_F \ge cn$,

$$\begin{aligned} G_C(\tau-1,b) - G_F(\tau-1,b) &\le -\frac{cK_2\varepsilon_d}{A_C[1+\gamma(\tau-1-T_{\text{noise}})]}\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{F}}\alpha_i\alpha_j \\ &= -\frac{cK_2A_F\varepsilon_d}{1+\gamma(\tau-1-T_{\text{noise}})} \le -\frac{cK_2n\varepsilon_d}{1+\gamma(\tau-1-T_{\text{noise}})}. \end{aligned}$$

For large K_2, the total update is nonpositive at the upper boundary by (74) and nonnegative at the lower boundary. Hence, we have

$$|B_\tau| \le \beta_{\tau-1} + K_2\varepsilon_d \le \beta_\tau + K_2\varepsilon_d.$$

With $|B_t| = o(1)$ and Lemma C16 with $\pi_i^{(t)} \ge 1/3$, we have

$$e^{c_i^{(t+1)}} - e^{c_i^{(t)}} \le C\gamma, \qquad \max_{s,r,i} \rho_{s,r,i}^{(\tau)} \le 3 \max_i c_i^{(\tau)} \le C_0 \log(2 + \eta T/m).$$

The constant C_0 is independent of M, so choosing $M > C_0$ gives the noise bound at τ. All three bounds in (71) therefore hold at τ, completing the induction through T.

Finally, by (68) and the preceding bounds, any own-class hard-feature projection crossing zero after T_{noise} stops within $C\varepsilon_d$ of zero. The opposite updates of the two classes leave at least one average activation at the initialization scale, so

$$|b_t| \le |B_t| + C\varepsilon_d + C\sigma_0 \log d = O\left(\frac{1}{\log d}\right).$$

□

D Additional Experimental Results

D.1 Teacher Selection

Table 4 shows the eight comparisons in which validation accuracy and R select different teachers; the other 56 comparisons select the same teacher. Each row corresponds to one dataset–student–seed combination.

Table 4: **Different teacher choices under the two selection rules.** Validation accuracy and correct-to-flipped hard-feature-strength ratio R of the selected teachers, with the resulting students' overall test accuracies. Validation accuracy is measured on hard-only samples. Q denotes Qwen, and bold indicates the highest test accuracy in each row.

Dataset	Student	Seed	Selected teacher		Val. acc. (%)		R		Test acc. (%)	
			By acc.	By R	By acc.	By R	By acc.	By R	By acc.	By R
IMDb	Q2.5-7B	1	Q2.5-0.5B	Q2.5-1.5B	69.4	69.0	1.9779	2.8280	89.96	**91.80**
RT	Q2.5-7B	6	Q3-1.7B	Q2.5-1.5B	67.4	67.4	2.3735	2.4699	92.62	**93.11**
RT	Q3-8B	3	Q2.5-1.5B	Q3-1.7B	72.0	70.4	2.5392	2.6294	88.16	**91.75**
Yelp	Q2.5-7B	1	Q2.5-0.5B	Q3-1.7B	85.8	84.0	8.3025	8.3085	95.26	**96.14**
Yelp	Q2.5-7B	4	Q3-0.6B	Q3-1.7B	85.8	84.4	7.5323	7.7226	95.36	**96.18**
Yelp	Q2.5-7B	5	Q2.5-0.5B	Q3-1.7B	85.8	85.2	7.4620	9.2940	95.94	**96.42**
Yelp	Q2.5-7B	6	Q2.5-0.5B	Q3-1.7B	82.2	82.0	6.3311	7.3775	96.14	**96.40**
Yelp	Q3-8B	8	Q3-1.7B	Q2.5-0.5B	85.2	84.2	6.3891	6.7178	**96.08**	**96.08**