

THEORETICAL ANALYSIS OF CONTRASTIVE LEARNING UNDER IMBALANCED DATA: FROM TRAINING DYNAMICS TO A PRUNING SOLUTION

Haixu Liao

New Jersey Institute of Technology
hl534@njit.edu

Yating Zhou

Cornell University
yz3554@cornell.edu

Songyang Zhang

University of Louisiana at Lafayette
songyang.zhang@louisiana.edu

Meng Wang

Rensselaer Polytechnic Institute
wangm7@rpi.edu

Shuai Zhang

New Jersey Institute of Technology
sz457@njit.edu

ABSTRACT

Contrastive learning has emerged as a powerful framework for learning generalizable representations, yet its theoretical understanding remains limited, particularly under imbalanced data distributions that are prevalent in real-world applications. Such an imbalance can degrade representation quality and induce biased model behavior, yet a rigorous characterization of these effects is lacking. In this work, we develop a theoretical framework to analyze the training dynamics of contrastive learning with Transformer-based encoders under imbalanced data. Our results reveal that neuron weights evolve through three distinct stages of training, with different dynamics for majority features, minority features, and noise. We further show that minority features reduce representational capacity, increase the need for more complex architectures, and hinder the separation of ground-truth features from noise. Inspired by these neuron-level behaviors, we show that pruning restores performance degraded by imbalance and enhances feature separation, offering both conceptual insights and practical guidance. Major theoretical findings are validated through numerical experiments.

1 INTRODUCTION

Contrastive learning has emerged as a powerful paradigm in representation learning, effectively leveraging unlabeled data without relying on labels. Within this framework, samples with similar semantic meaning are treated as positive pairs, while those with different semantics are considered negative pairs. By pulling positive pairs closer together and pushing negative pairs farther apart in the representation space, contrastive learning enables models to capture rich and discriminative features. Compared with supervised learning, the resulting representations are often more robust and less sensitive to noise (Xue et al., 2022; Ghosh & Lan, 2021; Zhong et al., 2022a; Jiang et al., 2020; Yang & Xu, 2020; Kang et al., 2020). This approach has demonstrated remarkable success across a wide range of applications (Zhong et al., 2022b; Zhang et al., 2022; Jiang et al., 2023; Luo et al., 2023) and has been particularly influential in multi-modal learning (Nakada et al., 2023; Khan et al., 2025), driving major advances in the early development of vision-language models (Radford et al., 2021; Li et al., 2022; 2023).

Despite its strengths, contrastive learning struggles with class imbalance in real-world datasets Jiang et al. (2021), where majority classes dominate pair formation and minority classes are underrepresented. This imbalance hinders the capture of discriminative features for minority classes and degrades representation quality. Conventional approaches to class imbalance in supervised learning

typically rely on re-weighting and re-sampling, and these ideas have inspired analogous methods in contrastive learning. Re-weighting strategies adjust the contribution of pairs or instances to reduce the dominance of majority classes (Cui et al., 2019; Huang et al., 2016), while resampling methods construct more balanced training batches by oversampling minority samples or undersampling majority ones (Drummond & Holte, 2003; He & Garcia, 2009; Peng et al., 2020). Although these approaches have shown effectiveness in certain cases, their application in contrastive settings remains challenging, as they often rely on accurate class labels that are unavailable in self-supervised learning. To address this limitation, an alternative line of research has proposed pruning-based methods, which have been empirically validated to enhance the representation of underrepresented classes (Jiang et al., 2021; Qian et al., 2022).

Despite the progress made by these approaches, most efforts have been largely empirical, relying on heuristic methods to alleviate the imbalance problem. While these techniques often provide performance gains in practice, they do not explain why or how imbalance undermines the quality of learned representations. Recent work has begun to develop theoretical understandings of contrastive learning, primarily addressing questions such as its superiority over traditional generative approaches like GANs (Ji et al., 2023), the necessity of data augmentation for effective representation learning (Wen & Li, 2021), and its ability to produce representations that reduce the sample complexity of downstream tasks (Garg & Liang, 2020). Nonetheless, these studies have not considered the implications of imbalanced data distributions.

In this work, we provide a theoretical analysis of how neurons learn feature representations through contrastive training. We study a simplified but representative setting: a Transformer-MLP framework with a single-head attention mechanism followed by an MLP with bilateral ReLU activations. To make the analysis clear, we use a structured data model where each input includes majority and minority features with different frequencies. This setup highlights the key role of feature frequencies and helps us describe their impact on training dynamics and how neurons learn features. In turn, the model allows us to formalize how contrastive learning enhances majority features and drives neurons to learn purer feature representations. Overall, our paper makes three main **contributions**:

First, we develop a theoretical framework to characterize the training dynamics of contrastive learning under Transformer-based encoders with an imbalanced data distribution. We show that learning proceeds in three stages: first, neuron weights grow in feature directions while non-feature components are suppressed; second, Lucky neurons then specialize in single features, and ordinary neurons learn a mix of features; finally, each neuron converges in a way that guarantees a small training loss, becoming strongly aligned with one or more features, weakly aligned with other features, and remaining small in non-feature directions. See Figure 1 for reference.

Second, we quantitatively characterize how the presence of minority features influences neurons’ learning capacity and, consequently, representation learning. Our analysis reveals that imbalance degrades representation performance in multiple ways: it slows the learning of minority features, decreases the number of neurons that specialize in a single feature, and produces a chain effect that necessitates a more complex model to adequately capture all features.

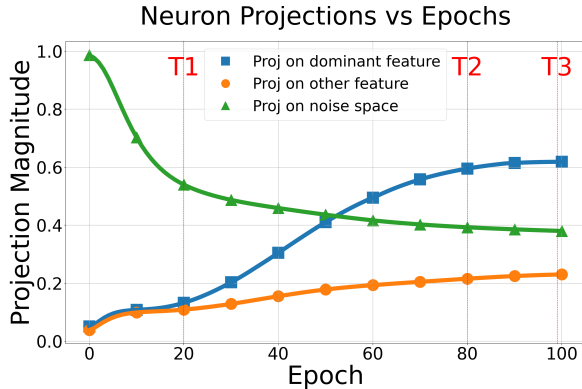


Figure 1: Neuron projection dynamics over training epochs. The blue curve shows the growth of a neuron’s projection onto its dominant feature, the orange curve shows the projection onto a non-dominant feature, and the green curve shows the projection onto the noise space direction (which remains larger than the projections onto other features). In the first stage, the neuron grows mainly along feature directions while suppressing noise. In the second stage, the projection onto the dominant feature grows faster than all other features, creating clear separation. In the third stage, as training approaches T_3 , the neuron converges, and its final representation is dominated by the learned feature, with negligible components in other directions.

Third, magnitude-based pruning can enhance the learning of minority features. Our results reveal that magnitude-based pruning enhances updates along minority feature directions, encouraging more neurons to specialize in pure minority features and thereby yielding more robust and balanced representations. Intuitively, neurons with small magnitudes are more sensitive to samples containing minority features, which implicitly allows pruning to amplify their contribution.

1.1 RELATED WORK

Data Imbalance in Self-Supervised Learning: Data imbalance or long-tail data has been a long-standing challenge since the early development of supervised learning (Chu et al., 2020; Liu et al., 2020; Yang et al., 2022; Chawla et al., 2002). At a high level, tackling data imbalance follows a simple principle: balancing the influence of different groups of data during weight updates, typically through re-sampling (Buda et al., 2018; Choi et al., 2018), which alters the data distribution, or re-weighting (Mahajan et al., 2018), which adjusts loss contributions across classes. These methods all require label information (Cui et al., 2021; Zhu et al., 2022). However, without label information, as in self-supervised learning (SSL), these strategies are far more difficult to apply, and only a few works have addressed the imbalance. Beyond re-weighting and re-sampling (Lin et al., 2017; Shrivastava et al., 2016; Shang et al., 2025; Shen et al., 2016), other alternative approaches have been proposed: optimization-based regularization for rare samples (Liu et al.), mixup for implicit rebalancing (Li & Jia, 2025), and pruning as an implicit means of detecting long-tail data (Jiang et al., 2021; Qian et al., 2022).

Convergence and Generalization Analysis of Contrastive Learning: Despite its empirical success, contrastive learning lacks a mature theoretical understanding, largely due to the complexity of its loss function. Early research investigates why augmentation is essential for the success of contrastive learning, showing that such an alignment between augmented positive pairs facilitates learning useful representations (Saunshi et al., 2022; Tian et al., 2020; Saunshi et al., 2019; Wen & Li, 2021). Tian et al. (2021); Wang et al. (2023) establishes a connection between the gradients of contrastive learning and graph neural networks, highlighting interpretability through a graph-theoretic perspective. HaoChen et al. (2021) also explores the connections between contrastive learning and graph theory, proposing a new loss function linked to graph spectral clustering to help explain its success. Wen & Li (2021) emphasizes the necessity of data augmentation for breaking dependencies on spurious noise. None of these works has explored how imbalanced data influences the training dynamics of contrastive learning.

Feature Learning Paradigm: The mathematical framework in this paper is closely related to the feature learning paradigm. Specifically, we assume the data follow a sparse coding model, which is a mixture of latent features, and study the training dynamics of model weights to examine how they align with these features. Most prior works focus on supervised learning (Allen-Zhu & Li, 2022; Zhang et al., 2023; Li et al., 2025; Cao et al., 2022; Chowdhury et al., 2023; Shandirasegaran et al., 2025), where features are tied to ground-truth labels; however, such settings cannot be directly extended to contrastive learning. Because of the complexity of analyzing fine-grained training dynamics, existing studies are typically limited to simple one-hidden-layer neural networks, with some recent efforts exploring Transformers but still restricted to a single layer (Huang et al., 2024; Oymak et al., 2023; Li et al., 2024), even under supervised settings. The most relevant works are Wen & Li (2021); Sun et al. (2025), which analyze the training dynamics of contrastive learning with one-hidden-layer feedforward networks. In contrast, our paper studies Transformer architectures under a different data model, and further incorporates data imbalance, providing a comprehensive analysis of how it influences the model’s ability to decouple features, rather than being only a direct extension through feature magnitude changes.

2 PROBLEM FORMULATION AND ALGORITHM

Contrastive Learning Framework. Let $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}] \in \mathbb{R}^{d_1 \times L}$ or $\mathbf{Y} \in \mathbb{R}^{d_1 \times L}$ be an input sequence with L tokens. The goal of contrastive learning is to learn a mapping $\mathbf{f}(\cdot) : \mathbb{R}^{d_1 \times L} \rightarrow \mathbb{R}^m$ that outputs a meaningful embedding from the input sequence.

Let $(\mathbf{X}_n, \mathbf{Y}_n)$ denote a *positive pair* (e.g., derived from the same objective or sharing semantic meaning), and let \mathfrak{N} denote a set of corresponding *negative samples* (e.g., random samples). The

InfoNCE loss with temperature parameter $\tau > 0$ is defined as:

$$\ell(\mathbf{f}_\theta, \mathbf{X}_n, \mathbf{Y}_n, \mathfrak{N}) := -\log \left(\frac{e^{\text{sim}_{\mathbf{f}_\theta}(\mathbf{X}_n, \mathbf{Y}_n)/\tau}}{\sum_{\mathbf{X} \in \{\mathbf{Y}_n\} \cup \mathfrak{N}} e^{\text{sim}_{\mathbf{f}_\theta}(\mathbf{X}_n, \mathbf{X})/\tau}} \right), \quad (1)$$

where the similarity function is given by

$$\text{sim}_{\mathbf{f}_\theta}(\mathbf{X}_n, \mathbf{Y}_n) := \langle \mathbf{f}_\theta(\mathbf{X}_n), \text{StopGrad}(\mathbf{f}_\theta(\mathbf{Y}_n)) \rangle, \quad (2)$$

and $\text{StopGrad}(\cdot)$ acts as the identity in forward pass while blocking gradients in backpropagation.

Then, the learning objective is to minimize an empirical risk with l_2 -regularizer, i.e.,

$$\widehat{L}_{\text{aug}}(\mathbf{f}_\theta) = \widehat{L}(\mathbf{f}_\theta) + \frac{\lambda}{2} \|\theta\|_F^2 = \frac{1}{K} \sum_{k=1}^K \ell(\mathbf{f}_\theta, \mathbf{X}_k, \mathbf{Y}_k, \mathfrak{N}_k) + \frac{\lambda}{2} \|\theta\|_F^2, \quad (3)$$

where θ denotes the neural network parameters and $K = \text{poly}(d_1)$.

Model Architecture: Transformer-MLP. We employ a simplified single-head self-attention mechanism on top of an MLP layer. Each input sequence is passed through the attention layer, where every token serves as a query. Then, it is followed by a bilateral ReLU (BReLU) activation in the MLP layer, where $\text{BReLU}_b(s) = \text{ReLU}(s - b) - \text{ReLU}(-s - b)$. Specifically, the embedding function f is expressed as

$$\begin{aligned} \mathbf{f}_\theta(\mathbf{X}_n) &= (h_1(\mathbf{X}_n), \dots, h_m(\mathbf{X}_n))^\top \in \mathbb{R}^m, \\ \text{with } h_i(\mathbf{X}_n) &= \sum_{r=1}^L \text{BReLU}_{b_i^{(t)}} \left(\langle \mathbf{w}_i^{(t)}, \text{Attention}(\mathbf{W}_Q \mathbf{x}_n^{(r)}, \mathbf{W}_K \mathbf{X}_n, \mathbf{W}_V \mathbf{X}_n) \rangle \right). \end{aligned} \quad (4)$$

Pruning Algorithm. To address the issue of data imbalance, we revisit (Jiang et al., 2021; Qian et al., 2022) and propose a pruning algorithm that dynamically removes small-magnitude neuron weights during the forward pass, while retaining all parameters as trainable in the backward pass¹. Specifically, we initialize the MLP layer weights with Gaussian distributions and the attention weights as identity matrices. The binary mask is initially set to all ones, meaning no neurons are pruned at the start. At each epoch, a fraction α of the neurons with the smallest magnitudes are pruned, and the corresponding binary mask is updated. During the forward pass, the masked parameters $\theta_{\text{mk}}^{(t)}$ are used to encode the inputs. In the backward pass, gradients are computed with respect to the pruned model but applied to the full parameter set, namely, the gradient is calculated as

$$g(\theta_t^{(t)}, \mathbf{M}^{(t)}) := \frac{1}{K} \sum_{k=1}^K \left[(\ell'_{p, \theta_{\text{mk}}^{(t)}} - 1) h_i(\mathbf{Y}_k) \nabla_{\theta} h_i(\mathbf{X}_k) + \sum_{\mathbf{X}_{n,s} \in \mathfrak{N}_k} \ell'_{s, \theta_{\text{mk}}^{(t)}} h_i(\mathbf{X}_{n,s}) \nabla_{\theta} h_i(\mathbf{X}_k) \right], \quad (5)$$

where $\ell'_{p, \cdot} := \frac{\exp(\text{Sim}_{\mathbf{f}_\theta}(\mathbf{X}_k, \mathbf{Y}_k)/\tau)}{\sum_{\mathbf{X} \in \{\mathbf{Y}_k\} \cup \mathfrak{N}_k} \exp(\text{Sim}_{\mathbf{f}_\theta}(\mathbf{X}_k, \mathbf{X})/\tau)}$ is the positive logit and $\ell'_{s, \cdot} := \frac{\exp(\text{Sim}_{\mathbf{f}_\theta}(\mathbf{X}_k, \mathbf{X}_{n,s})/\tau)}{\sum_{\mathbf{X} \in \{\mathbf{Y}_k\} \cup \mathfrak{N}_k} \exp(\text{Sim}_{\mathbf{f}_\theta}(\mathbf{X}_k, \mathbf{X})/\tau)}$ is negative logit with respect to the native sample $\mathbf{X}_{n,s}$.

Note that this procedure does not permanently eliminate any neurons for efficiency purposes, even though a reduction in computation cost can be observed. The pruning mask acts as a temporary filter by automatically removing small-magnitude neurons. As shown in Theorem 3.2, these neurons are associated with minority features. Consequently, samples containing such features incur a higher loss, which in turn encourages the model to allocate greater attention to them during training.

3 THEORETICAL ANALYSIS

3.1 KEY INSIGHTS OF THE FINDINGS

We first give a summary of the key insights from our analysis before turning to the data model and the formal theoretical results. Our findings show how neurons gradually learn feature representations across different stages of training. In particular, we have

¹We do not introduce a new algorithm; instead, we adapt established approaches to our theoretical setting.

Algorithm 1 Forward Magnitude Pruning with Backward Unmasked Update**Require:** Training dataset $\{(\mathbf{X}_k, \mathbf{Y}_k, \mathfrak{N}_k)\}_{k=1}^K$ (positive pairs $(\mathbf{X}_k, \mathbf{Y}_k)$ and negative set \mathfrak{N}_k)**Require:** Pruning ratio α **Require:** Training epochs T , weight decay parameter λ , temperature τ

- 1: Initialize network parameters $\mathbf{w}_i^{(0)} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_{d_1})$, $\mathbf{W}_K^{(0)} = \mathbf{W}_Q^{(0)} = \mathbf{I}$.
- 2: Set the initial pruning mask $\mathbf{M}^{(0)} \leftarrow \mathbf{1}$ with the same shape as $\boldsymbol{\theta}^{(0)}$.
- 3: **for** $t = 0$ to $T - 1$ **do**
- 4: **Magnitude based pruning:** At each iteration t , prune α fraction of neurons with the smallest magnitude in $\boldsymbol{\theta}^{(t)}$ by creating the corresponding binary mask $\mathbf{M}^{(t)}$.
- 5: **Forward (masked):** Apply the mask to obtain $\boldsymbol{\theta}_{\text{mk}}^{(t)} \leftarrow \boldsymbol{\theta}^{(t)} \odot \mathbf{M}^{(t)}$, then encode $\mathbf{X}_k, \mathbf{Y}_k$, and negatives \mathfrak{N}_k using $\mathbf{f}_{\boldsymbol{\theta}_{\text{mk}}^{(t)}}$.
- 6: **Compute loss:** $\hat{L}_{\text{aug}}(\mathbf{f}_{\boldsymbol{\theta}_{\text{mk}}^{(t)}}) = \frac{1}{K} \sum_{k=1}^K \ell(\mathbf{f}_{\boldsymbol{\theta}_{\text{mk}}^{(t)}}, \mathbf{X}_k, \mathbf{Y}_k, \mathfrak{N}_k; \tau) + \frac{\lambda}{2} \|\boldsymbol{\theta}_{\text{mk}}^{(t)}\|_F^2$.
- 7: **Backward and update:** Release the mask $\mathbf{M}^{(t)}$ on the masked parameters and update the full parameters by

$$\boldsymbol{\theta}^{(t+1)} \leftarrow (1 - \eta\lambda)\boldsymbol{\theta}^{(t)} - \eta \cdot g(\boldsymbol{\theta}_t^{(t)}, \mathbf{M}^{(t)})$$

8: **end for**9: **return** $\boldsymbol{\theta}^{(T)}$ **(K1). Training dynamics of contrastive learning based on the Transformer-MLP framework.**

The theory divides the learning process into three stages. In Stage 1 (Lemma 3.1), neuron weights grow in feature directions at rates determined by the feature frequencies ϵ_j , while their components in non-feature directions are suppressed. In Stage 2 (Lemma 3.2), lucky neurons in \mathcal{M}_j^* strengthen their alignment with the feature direction \mathbf{M}_j , and ordinary neurons in \mathcal{M}_j remain bounded by these lucky neurons, so that the learned features become purer and non-feature components remain suppressed. In the final stage, each neuron aligns with a specific set of features \mathcal{N}_i , becoming strongly aligned with some features, weakly with others, and remaining small in non-feature directions.

(K2). Feature frequency ratio controls neuron specialization.

At convergence, each neuron is dominated by features in \mathcal{N}_i , with negligible contribution from other directions. First, the neuron magnitude in \mathcal{N}_i , denoted $\alpha_{i,j}$, scales as $\frac{\epsilon_j}{\epsilon_{\max}}$, so rarer features are learned more weakly. Second, the size of \mathcal{N}_i scales as $d^{1-(\epsilon_{\min}/\epsilon_{\max})^2}$: smaller ratios enlarge \mathcal{N}_i and cause feature mixing, while larger ratios shrink it and yield purer alignment. Third, the number of neurons specializing in purified features scales as $d^{-(\epsilon_{\max}/\epsilon_{\min})^2}$, which decreases as the gap between ϵ_{\max} and ϵ_{\min} grows. Since contrastive learning works best when neurons specialize in purified features, imbalance introduces three interrelated obstacles: minority features are learned with smaller magnitude, neurons mix multiple features instead of staying pure, and the overall number of specialized neurons decreases. Together, these effects weaken representation quality and require larger models to learn all features.

(K3). Pruning enhances minority feature learning.

With pruning ratio α , neurons aligned with minority features gain stronger updates of order $\frac{\alpha}{d}$, while those aligned with non-minority features grow only weakly, with updates of order $\frac{\alpha}{d^2}$. At convergence, the coefficient of neurons learning a minority feature can reach the same order as that of majority features, so the performance downgrade from imbalance is alleviated. Intuitively, minority neurons are pruned more often because their magnitudes are smaller, which in turn amplifies the contribution of samples containing the minority feature in gradient updates. As a result, pruning strengthens the minority feature, makes it clearly distinguished from other contributions, and drives more neurons to specialize in it, leading to more robust representation learning.

3.2 ASSUMPTIONS

Data Model. Our data assumption is adopted from the widely used sparse coding model, which constitutes a common foundation for theoretical analyses of deep learning (Allen-Zhu & Li, 2022; Wen & Li, 2021). Moreover, sparse coding provides a conceptual framework for modeling real-world data across diverse domains, including CV (Protter & Elad, 2008; Yang et al., 2009; Mairal et al., 2014; Liao et al., 2025), NLP (Arora et al., 2018), compressed sensing (Candes & Recht,

Table 1: Summary of main notations

η	Learning rate	λ	Regularization parameter
τ	Temperature coefficient	K	Batch size
\mathfrak{N}	Set of negative samples	\mathfrak{B}	The set of \mathbf{Y}_n and negative samples
ϵ_{\min}	frequency of minority feature	ϵ_{\max}	frequency of majority feature
ϵ_j	Feature frequency for feature j	\mathcal{N}_i	Set of dominate features for neuron i
\mathcal{M}_j	Set of ordinary neurons for feature j	\mathcal{M}_j^*	Set of lucky neurons for feature j

2012; Candès & Tao, 2010), and neuroscience (Vinje & Gallant, 2000; Olshausen & Field, 1997; 2004; Foldiak, 2003).

Assumption 3.1 states that each token within a sample can be expressed as a weighted sum of a subset of features from the dictionary matrix \mathbf{M} , corrupted by additive noise $\boldsymbol{\xi}$. Here, \mathbf{M} denotes the dictionary matrix, \mathbf{z} represents the latent signal, and $\boldsymbol{\xi}$ corresponds to spurious noise. Importantly, in the presence of noise, particularly when the noise level is comparable to or even exceeds the signal magnitude, no linear mapping can recover the latent signal directly from the input. This makes the model simple in form yet intrinsically challenging, thereby providing a favorable abstraction for theoretical analyses of nonlinear neural networks.

Assumption 3.1 (Sparse Coding Model). *For a paired data $(\mathbf{X}_n, \mathbf{Y}_n)$, the data structure is:*

$$\begin{aligned}\mathbf{X}_n &= [\mathbf{M}\mathbf{z}_n^{(1)} + \boldsymbol{\xi}_n^{(1)}, \mathbf{M}\mathbf{z}_n^{(2)} + \boldsymbol{\xi}_n^{(2)}, \dots, \mathbf{M}\mathbf{z}_n^{(L)} + \boldsymbol{\xi}_n^{(L)}] \\ \mathbf{Y}_n &= [\mathbf{M}\mathbf{z}_n^{+(1)} + \boldsymbol{\xi}_n^{+(1)}, \mathbf{M}\mathbf{z}_n^{+(2)} + \boldsymbol{\xi}_n^{+(2)}, \dots, \mathbf{M}\mathbf{z}_n^{+(L)} + \boldsymbol{\xi}_n^{+(L)}]\end{aligned}\quad (6)$$

Here, each $\mathbf{z}_n^{(i)} \in \mathbb{R}^d$ represents the latent signal at the i -th token, and $\boldsymbol{\xi}_n^{(i)}$ denotes the additive noise. $\mathbf{M} = [\mathbf{M}_1, \dots, \mathbf{M}_d] \in \mathbb{R}^{d_1 \times d}$ is the dictionary matrix, which is a column-orthonormal matrix and satisfies $\|\mathbf{M}_j\|_\infty \leq \tilde{O}(\frac{1}{\sqrt{d_1}})$, $\forall j \in [d]$. We also assume $d_1 = \text{poly}(d)$.

Assumption 3.2 requires that the latent signal be both bounded and sparse. Sparsity is a standard assumption, introduced primarily to facilitate the theoretical analysis, yet it also agrees with empirical observations that real-world data typically activate only a small subset of latent factors rather than spreading energy across all coordinates. Moreover, the assumption enforces sign consistency across tokens within the same sample, meaning that whenever a particular coordinate is active, its sign remains identical across all tokens. This ensures that different parts of the same sample contribute coherently to the underlying latent feature instead of producing conflicting activations.

Assumption 3.2 (Latent Signal). *We have assumptions on the latent signal $\{\mathbf{z}^{(i)}\}_{i=1}^L$ with $\mathbf{z}^{(i)} = (z_1^{(i)}, \dots, z_j^{(i)}, \dots, z_d^{(i)})^\top$: (i) all $z_j^{(i)}$ are bounded and symmetric around zero over all samples. Moreover, we have $\Pr(|z_{n,j}^{(i)}| \neq 0) = \Theta\left(\frac{\log \log d}{d}\right)$; (ii) $z_j^{(i)}$ share the same sign across all $i \in [L]$.*

Assumption 3.3 states that noise follows Gaussian distributions. This is a mild condition, as no strong restriction is imposed on its variance. In particular, the noise magnitude can exceed that of the sparse signal when $d_1 \gg d$. The assumption is adopted for analytical purposes and demonstrates that contrastive learning can recover meaningful latent representations even in regimes where the signal is dominated by noise.

Assumption 3.3 (Noise). *Here each noise term $\boldsymbol{\xi}_n^{(\ell)}$ and $\boldsymbol{\xi}_n^{+(\ell)}$ for $\ell \in [L]$ is independently drawn from the same distribution $\boldsymbol{\xi}_n^{(\ell)} \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{I}_{d_1})$, with variance $\sigma_\xi^2 = \Theta\left(\frac{\sqrt{\log d}}{d}\right)$.*

Assumption 3.4 states that a pair of positive samples shares the same set of features when aggregated over all tokens within the sample. Intuitively, this means that the two samples encode the same semantic structure, even though their individual token-level representations may differ. In contrast, a negative pair is formed by two random samples whose latent signals are completely independent.

Assumption 3.4 (Positive and Negative Pairs). *A pair of samples \mathbf{X}_n and \mathbf{Y}_n form a positive pair if and only if $\text{supp}(\sum_{\ell=1}^L \mathbf{z}_n^{(\ell)}) = \text{supp}(\sum_{\ell=1}^L \mathbf{z}_n^{+(\ell)})$, $\text{sign}(\sum_{\ell=1}^L \mathbf{z}_n^{(\ell)}) = \text{sign}(\sum_{\ell=1}^L \mathbf{z}_n^{+(\ell)})$. By contrast, negative pairs are defined such that the corresponding latent signals are independent.*

Definition 3.1 states that each feature is controlled by ϵ_j . Intuitively, ϵ_j characterizes how often feature j appears across the data. When ϵ_j is small, feature j is regarded as a minority feature.

Definition 3.1 (Majority and minority features). *For each feature index $j \in [d]$, and for all $i \in [L]$ and all samples, the activation probability of the sparse signal satisfies: $\Pr(|z_j^{(i)}| \neq 0) = \Theta\left(\epsilon_j \frac{\log \log d}{d}\right)$. We define the majority features as those associated with $\epsilon_{\max} = \max_{j \in [d]} \epsilon_j$, and the minority features as those associated with $\epsilon_{\min} = \min_{j \in [d]} \epsilon_j$.*

3.3 FORMAL THEORETICAL RESULTS

Theorem 3.1 analyzes the vanilla contrastive learning algorithm without pruning, showing how data imbalance affects performance. Lemmas 3.1 and 3.2 provide intermediate steps toward its proof and reveal how training dynamics evolve, despite the algorithm appearing to follow a consistent gradient-based procedure. Theorem 3.2 then gives the results with pruning, showing how pruning improves performance under imbalance.

3.3.1 VANILLA CONTRASTIVE LEARNING

Lemma 3.1 shows two main effects of contrastive learning in the first training stage: (a) neuron weights grow in feature directions but are suppressed in non-feature directions, and (b) the growth rate in a feature direction M_j depends on its frequency ϵ_j , with larger ϵ_j leading to faster growth and smaller ϵ_j making the feature harder to capture early in training. We can find the Proof of Lemma 3.1 in Appendix C.4.

Lemma 3.1 (Stage 1). *During the first training stage, the update of neuron weights $w_i^{(t)}$ can be bounded for all $t \in [0, T_1]$ as follows, where C_z denotes positive constants and $T_1 = \Theta\left(\frac{d_1 \log d}{\eta \log \log d}\right)$.*

$$|\langle w_i^{(t+1)}, M_j \rangle| \geq |\langle w_i^{(t)}, M_j \rangle| (1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d}) - \tilde{O}\left(\frac{\eta \|w_i^{(t)}\|_2}{\text{poly}(d_1)}\right), \quad (7)$$

$$\text{and } |\langle w_i^{(t+1)}, M_j^\perp \rangle| \leq (1 - \eta\lambda) |\langle w_i^{(t)}, M_j^\perp \rangle| + \tilde{O}\left(\frac{\eta \|w_i^{(t)}\|_2}{\text{poly}(d_1)}\right). \quad (8)$$

Before presenting the theoretical results in Stage 2, we first categorize neurons into two groups. The *ordinary neurons* \mathcal{M}_j strongly align with a certain direction, while the *lucky neurons* \mathcal{M}_j^* form a special subset that aligns with only one feature direction (see Appendix B for the formal definition). In Stage 2: (a) lucky neurons in \mathcal{M}_j^* grow significantly in alignment with M_j , controlled by ϵ_j , though their number remains small; (b) ordinary neurons in \mathcal{M}_j are bounded by the feature components of lucky neurons up to a constant factor. We can find the Proof of Lemma 3.2 in Appendix D.4.

Lemma 3.2 (Stage 2). *During the second training stage, the update of neuron weights $w_i^{(t)}$ can be bounded for all $t \in [T_1, T_2]$ as follows, where $T_2 = T_1 + \Theta\left(\frac{d\tau \log d}{\epsilon_{\max} \eta \log \log d}\right)$.*

(a) For each $j \in [d]$, if $i \in \mathcal{M}_j^*$, then:

$$|\langle w_i^{(T_2)}, M_j \rangle|^2 \geq 2 \cdot \frac{\epsilon_j}{\epsilon_{\max}} \cdot \|w_i^{(T_1)}\|_2^2, \quad \text{with } |\mathcal{M}_j^*| \geq m \cdot d^{-\left(\frac{\epsilon_{\max}}{\epsilon_{\min}}\right)^2}. \quad (9)$$

(b) For each $j \in [d]$, if $i' \in \mathcal{M}_j$ and $i \in \mathcal{M}_j^*$, then:

$$|\langle w_{i'}^{(T_2)}, M_j \rangle| \leq O(|\langle w_i^{(T_2)}, M_j \rangle|). \quad (10)$$

Theorem 3.1 establishes the convergence of the algorithm. In particular, (11) shows that the algorithm converges with bounded training error. Moreover, (12) characterizes the structure of the learned neuron weights: upon convergence, they become strongly aligned with a subset of features within \mathcal{N}_j , weakly aligned with the remaining features, and remain small in the non-feature directions. The size of \mathcal{N}_j is bounded as in (14), and only a limited number of neurons specialize in learning a single feature. We can find the Proof of Theorem 3.1 in Appendix E.4.

Theorem 3.1 (Stage 3: Convergence). Let $m = d^{C_m}$ be the number of neurons and $\tau = \text{polylog}(d)$, where C_m denotes positive constants and $\Xi_2 = d^{C_m - (\frac{\epsilon_{\min}}{\epsilon_{\max}})^2}$. Suppose we train the neural net f_{θ} via contrastive learning, and consider iterations $T \in [T_3, T_4]$ with $T_3 = \frac{d^{1.01}}{\eta}$ and $T_4 = \frac{d^{1.99}}{\eta}$. Then the following guarantees hold:

$$\frac{1}{T} \sum_{t \in [T]} L_{\text{aug}}(f_{\theta(t)}) \leq o(1) \quad (11)$$

Moreover, for each neuron $i \in [m]$ and $t \in [T_3, T_4]$, the weight will learn the following set of features:

$$\mathbf{w}_i^{(t)} = \sum_{j \in \mathcal{N}_i} \alpha_{i,j} \mathbf{M}_j + \sum_{j \notin \mathcal{N}_i} \alpha'_{i,j} \mathbf{M}_j + \sum_{j \in [d_1] \setminus [d]} \beta_{i,j} \mathbf{M}_j^{\perp}, \quad (12)$$

where

$$\alpha_{i,j} \in \left[\frac{\epsilon_j}{\epsilon_{\max}} \frac{\tau}{\Xi_2}, \frac{\epsilon_j}{\epsilon_{\max}} \tau \right], \alpha'_{i,j} \leq o\left(\frac{\epsilon_j}{\epsilon_{\max}} \frac{1}{\sqrt{d}}\right) \|\mathbf{w}_i^{(t)}\|_2, |\beta_{i,j}| \leq o\left(\frac{1}{\sqrt{d_1}}\right) \|\mathbf{w}_i^{(t)}\|_2. \quad (13)$$

Furthermore, the size of \mathcal{N}_i is bounded as

$$|\mathcal{N}_i| = O\left(d^{1 - (\frac{\epsilon_{\min}}{\epsilon_{\max}})^2}\right). \quad (14)$$

Finally, for each \mathbf{M}_j , there are at least $\Omega(m \cdot d^{-(\frac{\epsilon_{\max}}{\epsilon_{\min}})^2})$ neurons $i \in [m]$ such that $\mathcal{N}_i = \{j\}$.

Remark 1: For a neuron \mathbf{w}_i , its convergent weights are aligned with a subset of features \mathcal{N}_i . In contrast, all other feature directions are smaller by an order of $\frac{1}{\sqrt{d}}$. Hence, we can say that neuron \mathbf{w}_i is dominated by the features in \mathcal{N}_i . Moreover, the neurons associated with learning feature j are influenced by the frequency of that feature, which intuitively explains how imbalance shapes the distribution of neuron weights.

Remark 2: We emphasize that the success of contrastive learning relies on neurons that specialize in a single feature, referred to as lucky neurons, i.e., $\cup_j \mathcal{M}_j^*$. In contrast, neurons that learn mixtures of features are useful only for a limited subset of downstream tasks. The number of lucky neurons for each feature is lower bounded by $m \cdot d^{-(\frac{\epsilon_{\max}}{\epsilon_{\min}})^2}$, as derived from (9). Consequently, beyond the reduced neuron magnitude in minority feature directions, imbalance also decreases the number of neurons that learn purified features. This, in turn, requires a more complex model with a larger number of neurons to capture all features, leading to higher computational cost. Moreover, the upper bound of $|\mathcal{N}_i|$ increases as the ratio $\frac{\epsilon_{\min}}{\epsilon_{\max}}$ decreases, which is undesirable because it indicates that more neurons learn mixtures of features rather than pure ones.

Remark 3: Theorem 3.1 shows that each underlying semantic feature is captured cleanly by a subset of lucky neurons. When upstream contrastive learning produces a representation in which all semantic features are encoded in pure and separable directions, the resulting feature space becomes highly structured: it contains explicit axes corresponding to every true feature. If a downstream task relies on any subset of these features, a linear probe (or any simple classifier) can easily extract them because the corresponding feature directions are directly represented by the lucky neurons. In this sense, stronger neuron specialization leads to better linear separability and, consequently, improved downstream generalization.

3.3.2 CONTRASTIVE LEARNING WITH PRUNING

Theorem 3.2 describes the training dynamics in the pruning setting, serving as the counterpart to the earlier result obtained without pruning. To highlight the effect more clearly, we focus on stage 3. In particular, pruning amplifies the learning of minority features: (a) for lucky neurons aligned with minority directions, the neuron weights increase in that direction at the order of $\frac{\alpha}{d}$, where α is the pruning ratio. (b) In contrast, neurons associated with non-minority features exhibit much smaller growth, with updates in those directions on the order of $\frac{\alpha}{d^2}$ per iteration. (c) Most importantly, when training converges, the coefficients α_{i,j^*} , projecting neuron weights onto the minority feature \mathbf{M}_{j^*} , become dominant and independent of the ratio $\frac{\epsilon_{\min}}{\epsilon_{\max}}$. We can find the Proof of Theorem 3.2 in Appendix F.4.

Theorem 3.2 (Pruning: Reinforcing Minority Feature Learning). *With pruning ratio α , the following statements hold:*

(a) *When $i^* \in \mathcal{M}_{j^*}^*$, we have*

$$\langle \mathbf{w}_{i^*}^{(t+1)}, \mathbf{M}_{j^*} \rangle \geq \left(1 - \eta\lambda + \Omega \left(\eta\epsilon_{j^*}^2 \alpha \frac{C_z \log \log d}{d} \right) \right) \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle. \quad (15)$$

(b) *When $\forall i$ and $j \neq j^*$, we have*

$$\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle \leq \left(1 + O \left(\eta\epsilon_{j^*}^3 \alpha \frac{C_z \log \log d}{d^2} \right) \right) \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle. \quad (16)$$

(c) *For neuron $i \in \mathcal{M}_{j^*}^*$ and $t = T_5$, contrastive learning learns the following decomposition:*

$$\mathbf{w}_i^{(t)} = \alpha_{i,j^*} \mathbf{M}_{j^*} + \sum_{j \notin \mathcal{N}_i} \alpha'_{i,j} \mathbf{M}_j + \sum_{j \in [d_1] \setminus [d]} \beta_{i,j} \mathbf{M}_j^\perp, \quad (17)$$

where

$$\alpha_{i,j^*} \in \left[\frac{\tau}{\Xi_2}, \tau \right], \quad \alpha'_{i,j} \leq o \left(\left(1 + \frac{1}{d} \right) \cdot \frac{1}{\sqrt{d}} \right) \|\mathbf{w}_i^{(t)}\|_2, \quad |\beta_{i,j}| \leq o \left(\frac{1}{\sqrt{d_1}} \right) \|\mathbf{w}_i^{(t)}\|_2. \quad (18)$$

Finally, for feature \mathbf{M}_{j^*} , there are at least $\Omega(m \cdot d^{-1})$ neurons $i \in [m]$ such that $\mathcal{N}_i = \{j^*\}$.

Remark 1: We would like to clarify two implicit assumptions underlying the results. First, the pruning ratio is implicitly upper bounded by $|\mathcal{M}_{j^*}|$, so that under magnitude-based pruning, we can guarantee that all pruned neurons are those aligned with the minority feature \mathbf{M}_{j^*} . In practice, however, the pruning ratio can be extended to include any neurons that have learned minority features, i.e., any i with $j \in \mathcal{N}_i$. Second, we assume that the magnitude of all non-minority features is comparable. Intuitively, in the general case, neurons associated with the minority feature grow until their magnitude reaches the level of the second-smallest feature. At that point, both the original minority feature and the second-smallest feature effectively become the new minority features, and the process continues inductively across features. A detailed analysis of this extension is omitted for simplicity, so that we can prove and present the pruning benefits in a clear manner.

Remark 2: The difference between neurons learning minority features and those learning majority features arises from their sensitivity to pruning. As shown in Theorem 3.1, the magnitude of a neuron is determined by its dominant feature and the frequency of that feature. For neurons in \mathcal{M}_{j^*} that specialize in purified minority features, their magnitudes are significantly smaller than those of other neurons and are therefore more likely to be pruned. This pruning effect results in relatively smaller positive logits and larger negative logits on samples containing the minority feature (see (5)), thereby increasing the influence of these samples on the gradient updates. Since features are assumed to be independent across the data, such samples have a low probability of simultaneously containing other features, resulting in a difference on the order of $1/d$ in the growth dynamics of these neurons.

Remark 3: Unlike in the vanilla learning paradigm, the magnitude of α_{i,j^*} no longer depends on the ratio $\frac{\epsilon_{\min}}{\epsilon_{\max}}$, which suggests that the representation of the minority feature is not suppressed by data imbalance. Although the coefficients $\alpha'_{i,j}$ for other features may grow slightly due to the extended number of iterations required for convergence, their increase remains only on the order of $1/d$. Consequently, $\alpha_{i,j^*} \gg \alpha'_{i,j}$, which suggests that the minority feature is strongly amplified and clearly distinguished from other contributions. This, in turn, drives more neurons to specialize in the purified minority feature, leading to more robust and effective representation learning.

4 NUMERICAL EXPERIMENTS

Experiments on CIFAR10-LT, CIFAR100-LT, and ImageNet-LT. Table 2 reports the results of linear probe evaluation on CIFAR10-LT, CIFAR100-LT, and ImageNet-LT under long-tailed settings, comparing vanilla contrastive learning (w/o pruning) against our proposed approach (w/ pruning). Following the setup in (Jiang et al., 2021; Kang et al., 2020; Chen et al., 2020), models are

first pretrained and then evaluated using a linear probe, where a linear classifier is trained on frozen representations. The imbalance ratio, ρ , is defined as the ratio between the number of samples in the majority and minority classes, with larger values indicating more severe imbalance. Two evaluation metrics are considered: overall classification accuracy (%) and the accuracy gap (Δ_{20}) between the top 20% head classes and the bottom 20% tail classes. The results show that pruning consistently improves accuracy across all datasets, with improvements becoming more substantial as ρ increases. Furthermore, pruning generally reduces Δ_{20} , indicating better balance between head and tail classes. These results indicate that pruning not only enhances overall downstream task performance but also reduces the performance gap between head and tail classes. We also provide additional synthetic data experiments to support our theoretical insights; due to space limitations, these results are deferred to Appendix A.2.

Table 2: Linear probe accuracy (%) on CIFAR10-LT, CIFAR100-LT, and ImageNet-LT. Δ_{20} denotes the accuracy gap between the top 20% head classes and bottom 20% tail classes.

Dataset	ρ	Accuracy		Δ_{20}	
		w/o pruning	w/ pruning	w/o pruning	w/ pruning
CIFAR10-LT	1	90.93	91.52	1.54	1.28
	10	79.25 \pm 1.03	84.92 \pm 0.67	3.42 \pm 1.02	2.99 \pm 0.92
	50	75.58 \pm 0.84	83.60 \pm 1.02	3.92 \pm 1.21	3.35 \pm 0.76
	100	74.24 \pm 0.82	81.31 \pm 0.94	5.69 \pm 1.35	5.62 \pm 0.99
CIFAR100-LT	10	51.21 \pm 1.21	56.33 \pm 1.51	2.45 \pm 0.57	1.37 \pm 0.46
	50	49.32 \pm 0.45	56.12 \pm 0.32	4.95 \pm 1.02	2.57 \pm 0.92
	100	47.12 \pm 0.51	54.93 \pm 0.50	7.11 \pm 0.45	4.38 \pm 0.22
ImageNet-LT	256	63.21	65.12	8.47	7.21

5 LIMITATION

Our work has two main limitations. The first concerns studying the pruning ratio and pruning scheme in magnitude-based pruning. Providing a fully precise characterization of how performance varies across different ratios and schemes is highly nontrivial, and doing so would require making more precise assumptions about the data distribution. This will be part of our future work. Furthermore, existing theoretical results in our feature learning framework focus on a single, simplified architectural setting. Extending the analysis to more complex or realistic models will be another direction for future work, and may require fundamentally different derivations and analytical tools.

6 CONCLUSION

This work provides a theoretical analysis of the training dynamics of a Transformer-MLP model in learning feature representations through contrastive learning under imbalanced data settings. Specifically, we quantitatively characterize how the presence of minority features reduces the number of neurons that capture those features, as well as the number of “lucky neurons” that specialize in a single feature. This reduction, in turn, harms the overall representation learning ability of the model. Motivated by this theoretical characterization, we revisit the magnitude-based pruning approach to address data imbalance. In particular, we theoretically demonstrate that pruning can enhance gradient updates along the minority feature direction. This encourages more neurons to specialize in pure minority features, thereby yielding more robust and balanced representations. Looking ahead, promising directions include exploring alternative strategies beyond pruning that could further promote minority-feature learning.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation (NSF) under Grants #2349879, #2349878, #2425811, and #2430223. Part of Yating’s work was completed while she was a Ph.D.

student at Rensselaer Polytechnic Institute (RPI) and was supported in part by the Army Research Office (ARO) under Grant W911NF-25-1-0020, as well as by the Rensselaer-IBM Future of Computing Research Collaboration (<http://airc.rpi.edu>). We also thank the anonymous reviewers for their constructive and insightful comments.

LLM USAGE DISCLOSURE

We used large-language models (ChatGPT) to aid in polishing the writing of this paper. For numerical experiments, we employed AI-assisted coding tools (GitHub Copilot and ChatGPT) to support code development.

REFERENCES

- Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 977–988. IEEE, 2022.
- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.
- Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.
- Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.
- Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE transactions on information theory*, 56(5):2053–2080, 2010.
- Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.
- Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.
- Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.
- Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6074–6114. PMLR, 2023.
- Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *European conference on computer vision*, pp. 694–710. Springer, 2020.
- Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 715–724, 2021.
- Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- Chris Drummond and Robert C. Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Workshop on Learning from Imbalanced Datasets II*, 2003.

- Peter Foldiak. Sparse coding in the primate cortex. In *The Handbook of Brain Theory and Neural Networks*. MIT Press, 2003.
- Siddhant Garg and Yingyu Liang. Functional regularization for representation learning: A unified theoretical perspective. In *Advances in Neural Information Processing Systems*, 2020.
- Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2703–2708, 2021.
- Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.
- Haibo He and Eduardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009.
- Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Yu Huang, Zixin Wen, Yuejie Chi, and Yingbin Liang. Transformers provably learn feature-position correlations in masked image modeling. *CoRR*, 2024.
- Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis. *Journal of Machine Learning Research*, 2023.
- Yangqin Jiang, Chao Huang, and Lianghao Huang. Adaptive graph contrastive learning for recommendation. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4252–4261, 2023.
- Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Advances in neural information processing systems*, 33:16199–16210, 2020.
- Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *International conference on machine learning*, pp. 4927–4939. PMLR, 2021.
- Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International conference on learning representations*, 2020.
- Asifullah Khan, Laiba Asmatullah, Anza Malik, Shahzaib Khan, and Hamna Asif. A survey on self-supervised contrastive learning for multimodal text-image analysis. *arXiv preprint arXiv:2503.11101*, 2025.
- Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? In *International Conference on Machine Learning*, pp. 28734–28783. PMLR, 2024.
- Hongkang Li, Yihua Zhang, Shuai Zhang, Pin-Yu Chen, Sijia Liu, and Meng Wang. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. In *International Conference on Learning Representations*, 2025.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.
- Zhixin Li and Yuheng Jia. Conmix: Contrastive mixup at representation level for long-tailed deep clustering. In *The Thirteenth International Conference on Learning Representations*, 2025.

- Haixu Liao, Yating Zhou, Songyang Zhang, and Shuai Zhang. On the training dynamics of contrastive learning with imbalanced feature distributions: A theoretical study of feature learning. In *UniReps: 3rd Edition of the Workshop on Unifying Representations in Neural Models*, 2025.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *International Conference on Learning Representations*.
- Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- Dongsheng Luo, Wei Cheng, Yingheng Wang, Dongkuan Xu, Jingchao Ni, Wenchao Yu, Xuchao Zhang, Yanchi Liu, Yuncong Chen, Haifeng Chen, et al. Time series contrastive learning with information-aware augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4534–4542, 2023.
- Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, 2018.
- Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 2014.
- Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4348–4380. PMLR, 2023.
- Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 1997.
- Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 2004.
- Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, pp. 26724–26768. PMLR, 2023.
- Junran Peng, Xingyuan Bu, Ming Sun, Zhaoxiang Zhang, Tieniu Tan, and Junjie Yan. Large-scale object detection in the wild from imbalanced multi-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- Matan Protter and Michael Elad. Image sequence denoising via sparse and redundant representations. *IEEE Transactions on Image Processing*, 2008.
- Yiyue Qian, Chunhui Zhang, Yiming Zhang, Qianlong Wen, Yanfang Ye, and Chuxu Zhang. Co-modality graph contrastive learning for imbalanced node classification. *Advances in Neural Information Processing Systems*, 35:15862–15874, 2022.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 5628–5637. PMLR, 2019.

- Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pp. 19250–19286. PMLR, 2022.
- Mugunthan Shandirasegaran, Yating Zhou, Songyang Zhang, and Shuai Zhang. Theoretical analysis of the selection mechanism in mamba: Training dynamics and generalization. *NeurIPS 2025 Workshop WCTD*, 2025. URL https://transformertheory.github.io/pdf/30_shandirasegaran_et_al.pdf.
- Ziqiao Shang, Bin Liu, Fengmao Lv, Fei Teng, Tianrui Li, and Lan-Zhe Guo. Learning contrastive feature representations for facial action unit detection. *Pattern Recognition*, pp. 112746, 2025.
- Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *Computer Vision – ECCV 2016*, 2016.
- Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- Jiawei Sun, Shuai Zhang, Hongkang Li, and Meng Wang. Contrastive learning with data misalignment: Feature purity, training dynamics and theoretical generalization guarantees. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- Yuangdong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.
- Yuangdong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021.
- William E Vinje and Jack L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 2000.
- Y. Wang, Q. Zhang, T. Du, J. Yang, Z. Lin, and Y. Wang. A message passing perspective on learning dynamics of contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2023.
- Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pp. 11112–11122. PMLR, 2021.
- Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*, pp. 24851–24871. PMLR, 2022.
- Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *2009 IEEE Conference on computer vision and pattern recognition*, pp. 1794–1801. IEEE, 2009.
- Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(7):1837–1872, 2022.
- Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020.
- Shuai Zhang, Meng Wang, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Miao Liu. Joint edge-model sparse learning is provably efficient for graph neural networks. *The Eleventh International Conference on Learning Representations*, 2023.
- Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

- Y Zhong, H Tang, J Chen, J Peng, and Y-X Wang. Is self-supervised learning more robust than supervised learning? In *Proc ICML Workshop on Pre-training*, 2022a.
- Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803, 2022b.
- Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6908–6917, 2022.

A OVERVIEW OF THE APPENDIX AND PROOF SKETCH

The appendices are organized systematically to provide supporting materials for the main text. Appendix B introduces key notations and definitions, along with basic lemmas describing properties at initialization. Appendices C, D, and E present the proofs of the training dynamics of vanilla contrastive learning (without pruning) under the imbalanced data setting. Specifically, Appendix C contains the proof of Stage 1, corresponding to Lemma 3.1 in the main text; Appendix D contains the proof of Stage 2, corresponding to Lemma 3.2; and Appendix E contains the proof of Stage 3, corresponding to Theorem 3.1, which concludes the analysis with the final convergence results. Appendix F then provides the proof of our proposed algorithm (with pruning), corresponding to Theorem 3.2 in the main text. We recommend that readers first consult the proof sketch before examining the detailed lemmas and proofs in the appendices.

In addition, Appendices G-K collect the proofs of the lemmas referenced throughout the earlier appendices. To maintain readability, some of these lemma proofs are included only in the supplementary material. While these details are not essential for following the main arguments, we provide them in full for completeness.

A.1 PROOF SKETCH

In Stage 1, we analyze how neurons learn the features. Each neuron gradually learns the relevant feature directions while hardly learning the non-feature directions. Concretely, the projection of a neuron weights onto the feature subspace, though small at the beginning, grows rapidly during training and becomes significant, reaching the order of $\Omega(\|\mathbf{w}_i^{(T_1)}\|_2^2)$ (see Appendix D, Theorem C.1), while the projection onto the non-feature subspace stays nearly unchanged. The reason why the neuron weights grow toward the feature subspace is that the latent variable $z_{n,j}^{(i)}$ and $z_{n,j}^{+(i)}$ are dependent. This dependence produces an incremental term of order: $\epsilon_j \frac{\eta C_z \log \log d}{d}$, which accumulates during training and drives the neuron weights further into the feature space. In contrast, because the feature are orthogonal to the non-feature directions, and the latent variable $z_{n,j}^{(i)}$ is independent of the noise, the weights in the non-feature subspace remain essentially unchanged. The only variation that appears there is a negligible increment of size about $\frac{1}{\text{poly}(d_1)}$. (see Appendix C, Lemma C.1).

In Stage 2, the lucky neurons with large projection on a feature direction become activated and align clearly with that feature. If a neuron does not belong to \mathcal{M}_j , its projection on feature j remains small, so it cannot be activated and has only weak alignment. The projection on non-feature directions stays very small, so neurons do not learn the non-feature components (Appendix D, Lemma D.1). As a result, if neuron i is lucky for feature j , the projection of $\mathbf{w}_i^{(T_2)}$ onto \mathcal{M}_j is on the order of the $\Omega(1)\|\mathbf{w}_i^{(T_2)}\|_2$, meaning the neuron has already focused on \mathcal{M}_j (see Appendix D, Theorem D.1).

In Stage 3, neurons in \mathcal{M}_j^* continue to strengthen their projection on the corresponding feature j , and this projection remains the dominant part of their weight. Neurons not in \mathcal{M}_j keep only a small projection on feature j , so they cannot be activated. The projections on non-feature directions stay negligible throughout. Overall, the growth of neurons continues along the same directions established earlier, and the network starts to converge around T_3 . At this point, each neuron weight vector \mathbf{w}_i eventually aligns with a set of features \mathcal{N}_i , which corresponds to the features that already had some degree of alignment with \mathbf{w}_i at initialization.

In pruning stage, we rigorously show that pruning the neurons which have learned minority features enhances the learning of those features. After pruning, the gradients in backpropagation for neurons aligned with minority features become significantly stronger, which forces these neurons to further learn the minority features. To some extent, this reinforcement compensates for their lower frequency ϵ_{j^*} compared to majority features. In contrast, for neurons associated with majority features, pruning does not change their gradients, so they continue to update in the same speed and direction as before. As a result, the decomposition of neurons aligned with minority features becomes concentrated on those features, while contributions from other features and from non-feature directions remain suppressed and negligible.

A.2 SYNTHETIC EXPERIMENTAL SETTINGS

In this subsection, we provide the detailed settings of our synthetic experiments. We follow the standard sparse coding model to generate synthetic data, consistent with our main paper. Each generated data sample is passed into a Transformer to obtain a token embeddings, which is then processed by an MLP trained with a contrastive objective. After training, we evaluate the alignment of the learned neurons to the minority feature. Specifically, we report: (i) the number of neurons with alignment above a threshold (Figure 2); (ii) the maximum alignment value (Figure 3); (iii) the mean cosine similarity between positive pairs on the test set (Figure 4); and (iv) the regression test mean squared error (MSE) (Figure 5).

Experiment 1–2 (Alignment with the minority feature). We evaluate how well the learned neurons align with the minority feature. Specifically, for each w_i , we compute its normalized projection onto the minority feature. Figure 2 reports the number of neurons with projection larger than 0.3, while Figure 3 shows the maximum projection value across neurons. We vary ε_{\min} from 0.1 to 1.0, and consider different noise-to-signal ratio (NSR) levels, where $\text{NSR} = \sigma^2 d_1$ with $\sigma^2 \in \{(1/100)^2, (3/100)^2, (5/100)^2\}$ and $d_1 = 500$. Each experiment is independently repeated 100 times, and we report the mean results. The results demonstrate that as ε_{\min} increases, both the number of aligned neurons and the maximum alignment consistently grow, providing direct empirical support for our theoretical results. The detailed hyperparameter settings can be found in the code.

Experiment 3 (Average cosine similarity on the test set). We evaluate performance on the test set using the average cosine similarity between positive pairs. At test time, we keep the feature space identical. For each configuration, we generate 5000 test pairs with a fixed test seed and report the mean cosine similarity. We vary ε_{\min} from 0.05 to 0.5 in increments of 0.05, and use $\sigma^2 \in (5/100)^2, (7.5/100)^2, (10/100)^2$ to compute the corresponding NSR levels. Each configuration is independently repeated 100 times, and the averaged results are reported. The results in Figure 4 show that the average test cosine similarity consistently increases as ε_{\min} grows, indicating a stronger ability to learn the minority feature. Consequently, the quality of the learned features on the test set is enhanced, the model generalizes better, and the test performance becomes stronger, which provides further empirical support for our theoretical results. Detailed hyperparameter settings can be found in the code.

Experiment 4 (Test MSE on the downstream regression task). We evaluate the performance of the downstream regression task on the test set, measured by Test MSE. Both the downstream training stage and the test stage use a unified feature space. A linear regression head is trained on the representations obtained from upstream learning, using 1000 training pairs, and then evaluated on 5000 test pairs with a fixed test seed. In the setup, we vary ε_{\min} from 0.05 to 0.5 with a step size of 0.05, and use $\sigma^2 \in \{(3/100)^2, (5/100)^2, (7.5/100)^2\}$ to compute the corresponding NSR levels. Each configuration is independently repeated 100 times, and the averaged results are reported (Figure 5). The results show that as ε_{\min} increases, the test MSE consistently decreases, indicating a stronger ability to learn the minority feature. Consequently, the model achieves better overall learning and stronger generalization in downstream tasks, which is consistent with our theoretical analysis. Detailed hyperparameter settings can be found in the code.

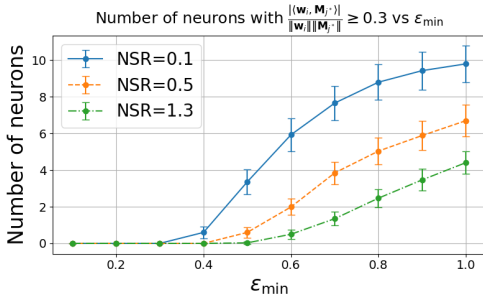


Figure 2: Number of neurons with $\frac{|\langle w_i, M_j \rangle|}{\|w_i\| \|M_j\|} \geq 0.3$ vs ε_{\min} for different NSR values.

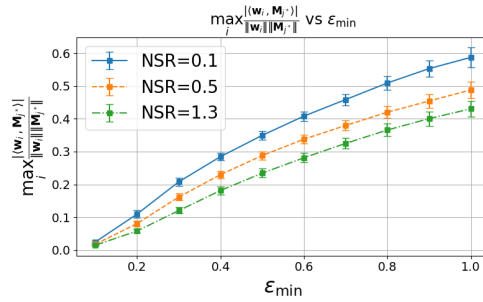


Figure 3: Maximum $\frac{|\langle w_i, M_j \rangle|}{\|w_i\| \|M_j\|}$ vs ε_{\min} for different NSR values.

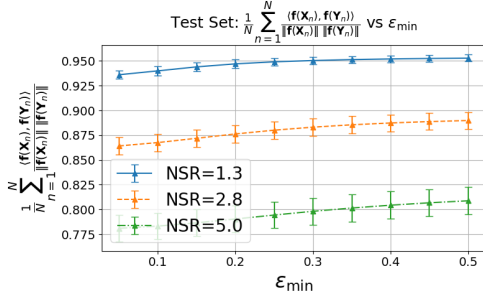


Figure 4: $\frac{1}{N} \sum_{n=1}^N \frac{\langle f(X_n), f(Y_n) \rangle}{\|f(X_n)\| \|f(Y_n)\|}$ vs ϵ_{\min} for different NSR values.

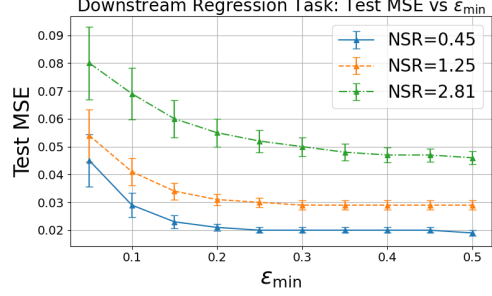


Figure 5: Downstream regression task: Test MSE vs ϵ_{\min} for different NSR values.

Experiment 5 (lucky vs. mixed-feature neurons).

The following heatmap visualizes the alignment values of 24 neurons (indexed 0-23) across 9 features (indexed 0-8), where the first five features are majority features and the last four are minority features.

Example of a lucky neuron: In our experiment, $d_1 = 500$, so the expected alignment from random initialization is approximately 0.002. After training, Neuron 0 exhibits a strong alignment with Feature 4 (around 0.3), while its alignment with the remaining eight features is negligible. Thus, Neuron 0 can be viewed as a lucky neuron for Feature 4.

Example of a mixed-feature neuron: Neuron 10, in contrast, does not exhibit a dominant alignment with any single feature. Instead, it shows moderate alignment with multiple features, specifically, its values on Features 0, 2, and 7 are 0.13, 0.18, and 0.15 respectively, while remaining negligible on all other features. This behavior corresponds to a mixed-feature neuron.

Additional examples: Further instances observed in our experiments include, but are not limited to. Lucky neurons: Neuron 6 and Neuron 17 for Feature 0; Neuron 3 and Neuron 18 for Feature 1; Neuron 4 for Feature 2; Neuron 13 for Feature 3. Mixed-feature neurons: Neuron 1 (Features 1 and 2), Neuron 2 (Features 1 and 3), Neuron 21 (Features 3 and 8).

These empirical patterns closely reflect the specialization and superposition behaviors predicted by our theoretical analysis.

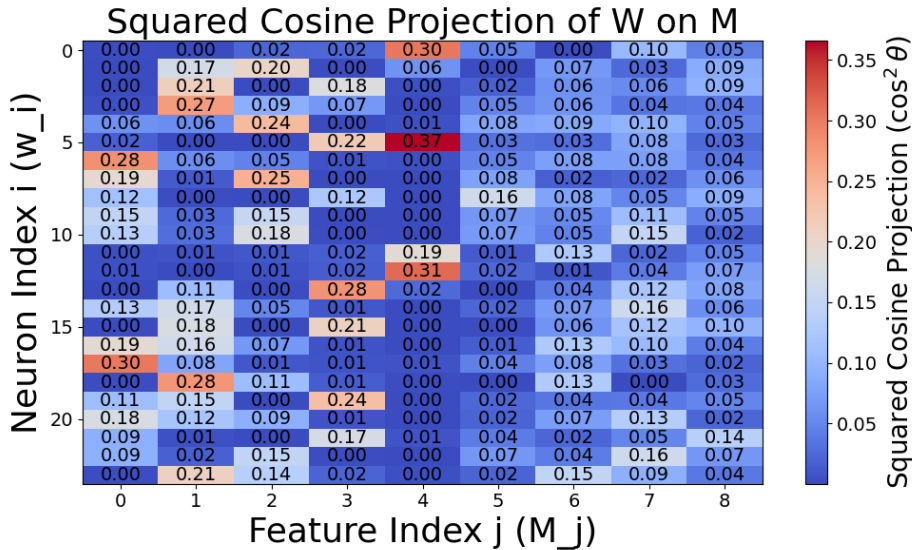


Figure 6: Squared cosine alignment heatmap between 24 neurons and 9 features, illustrating lucky neurons (single strong alignment) and mixed-feature neurons (multiple features alignment).

B NOTATIONS AND LEMMAS

To streamline the presentations, we begin by introducing the key notations and outlining key fundamental derivations that will serve as the basis for the subsequent analysis.

Notations. First, we introduce the notations that will appear in the appendix.

Let $\mathbf{z}_X^{(r)}$ denote the representation of the r -th token of data sample \mathbf{X}_n after passing through the transformer. Similarly, $\mathbf{z}_Y^{(s)}$ denotes the s -th token of data sample \mathbf{Y}_n after the transformer.

Empirical Gradient. To facilitate the calculation of the gradient of the loss function $\ell(\mathbf{f}_\theta, \mathbf{X}_k, \mathfrak{B}_k)$ with respect to the weights $\{\mathbf{w}_i^{(t)}\}_{i \in [m]}$, we introduce the following notation. We denote the positive logit by $\ell'_{p,t}(\mathbf{X}_n, \mathfrak{B})$ and the negative logits by $\ell'_{s,t}(\mathbf{X}_n, \mathfrak{B})$.

$$\ell'_{p,t}(\mathbf{X}_n, \mathfrak{B}) := \frac{\exp(\text{Sim}_{f_t}(\mathbf{X}_n, \mathbf{Y}_n)/\tau)}{\sum_{\mathbf{X} \in \mathfrak{B}} \exp(\text{Sim}_{f_t}(\mathbf{X}_n, \mathbf{X})/\tau)}, \quad (19)$$

$$\ell'_{s,t}(\mathbf{X}_n, \mathfrak{B}) := \frac{\exp(\text{Sim}_{f_t}(\mathbf{X}_n, \mathbf{X}_{n,s})/\tau)}{\sum_{\mathbf{X} \in \mathfrak{B}} \exp(\text{Sim}_{f_t}(\mathbf{X}_n, \mathbf{X})/\tau)}. \quad (20)$$

For convenience, we simplify the *positive logit* $\ell'_{p,\theta^{(t)}}(\mathbf{X}_n, \mathbf{Y}_n, \mathfrak{N}_n)$ as $\ell'_{p,t}$, and the *negative logit* $\ell'_{s,\theta^{(t)}}(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{X}_{n,s}, \mathfrak{N}_n)$ as $\ell'_{s,t}$. For clarity of exposition, we suppress the dependence on $(\mathbf{X}_n, \mathfrak{B})$ when it can be inferred from the context.

Then, the gradient of the empirical risk function $\widehat{L}(f_t)$ with respect to the weight $\mathbf{w}_i^{(t)}$ at iteration t is given by:

$$\begin{aligned} \nabla_{\mathbf{w}_i} \widehat{L}(f_t) = & \frac{1}{K} \sum_{n=1}^K \left[(\ell'_{p,t} - 1) h_i(\mathbf{Y}_n) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_X^{(r)} \rangle| \geq b_i} \mathbf{z}_X^{(r)} \right. \\ & \left. + \sum_{\mathbf{X}_{n,s} \in \mathfrak{N}} \ell'_{s,t} h_i(\mathbf{X}_{n,s}) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_X^{(r)} \rangle| \geq b_i} \mathbf{z}_X^{(r)} \right]. \end{aligned} \quad (21)$$

Population Gradient. Similar to the empirical gradient, the gradient of the population risk function $L(f_t)$ with respect to the weight $\mathbf{w}_i^{(t)}$ at iteration t is given by:

$$\begin{aligned} \nabla_{\mathbf{w}_i} L(f_t) = & \mathbb{E}[(\ell'_{p,t} - 1) h_i(\mathbf{Y}_n) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_X^{(r)} \rangle| \geq b_i} \mathbf{z}_X^{(r)} \\ & + \sum_{\mathbf{X}_{n,s} \in \mathfrak{N}} \ell'_{s,t} h_i(\mathbf{X}_{n,s}) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_X^{(r)} \rangle| \geq b_i} \mathbf{z}_X^{(r)}], \end{aligned} \quad (22)$$

where L is the population risk function as

$$L(f_t) = \mathbb{E}[\ell(\mathbf{f}_\theta, \mathbf{X}_n, \mathbf{Y}_n, \mathfrak{N})]. \quad (23)$$

Stop Gradient. Note that the similarity measure explicitly uses the `StopGrad` operation to block gradient flow through the second input. The similarity is computed as

$$\text{Sim}_{f_t}(\mathbf{X}_1, \mathbf{X}_2) = \langle \mathbf{f}_t(\mathbf{X}_1), \text{StopGrad}(\mathbf{f}_t(\mathbf{X}_2)) \rangle. \quad (24)$$

Concentration Bound. The following lemma shows that, given a sufficiently large number of samples, the approximation error between the empirical gradient and the population gradient remains bounded with high probability. Building on this principle, we will first analyze the training dynamics under the population gradient, and subsequently account for the deviation arising from the empirical gradient. The proof of Lemma B.1 follows standard techniques based on sub-Gaussian tail bounds and is therefore omitted.

Lemma B.1 (Approximation of empirical gradients by population gradients). *Suppose that $\|\mathbf{W}^{(t)}\|_F^2 \leq \text{poly}(d)$. Then there exists some $K = \text{poly}(d_1)$ such that, with high probability, the difference between the empirical gradients and the population gradients is bounded for every iteration t :*

$$\left\| \nabla_{\mathbf{w}_i} \hat{L}_{\text{aug}}(f_t) - \nabla_{\mathbf{w}_i} L_{\text{aug}}(f_t) \right\|_2 \leq \frac{\|\mathbf{w}_i^{(t)}\|_2}{\text{poly}(d_1)}, \quad \forall i \in [m]. \quad (25)$$

This Definition B.1 divides neurons into two categories, ordinary neurons and lucky neurons, based on their initial alignment with feature vectors \mathbf{M}_j . These sets will serve as the foundation for our later analysis.

Definition B.1 (Characterization of Neurons). *We define the following sets of neurons, which will be useful for analyzing the stochastic gradient descent trajectory in later sections:*

(a) For each $j \in [d]$, we define the set of ordinary neurons $\mathcal{M}_j \subseteq [m]$ as:

$$\mathcal{M}_j := \left\{ i \in [m] : \langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle^2 \geq \frac{c_2 \log d}{d} \|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2^2 \right\}, \quad \forall j \in [d] \quad (26)$$

(b) For each $j \in [d]$, we define the set of lucky neurons $\mathcal{M}_j^* \subseteq [m]$ as:

$$\mathcal{M}_j^* := \left\{ i \in [m] : \begin{aligned} &\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle^2 \geq \frac{c_1 \log d}{d} \|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2^2, \\ &\langle \mathbf{w}_i^{(0)}, \mathbf{M}_{j'} \rangle^2 \leq \frac{c_2 \log d}{d} \|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2^2, \quad \forall j' \in [d], j' \neq j \end{aligned} \right\}, \quad (27)$$

where

$$c_1 = \left(\frac{\epsilon_{\max}}{\epsilon_{\min}} \right)^2 \cdot 2(1 + \gamma), \quad c_2 = \left(\frac{\epsilon_{\min}}{\epsilon_{\max}} \right)^2 \cdot 2(1 - \gamma), \quad \gamma \text{ is a small constant.} \quad (28)$$

Properties at initialization: At initialization ($t = 0$), we note key facts about the neurons for later analysis of the SGD trajectory.

Before presenting Lemma B.2, we outline its essential idea: (a) Each $\mathbf{w}_i^{(0)}$ has magnitude in the order of $\sigma_0^2 d_1$; (b) Each $\mathbf{w}_i^{(0)}$ has a projection onto the feature subspace in the order of $\sigma_0^2 d$; (c) For each feature, the numbers of lucky and ordinary neurons are influenced by the frequencies of the majority and minority features; (d) For each neuron, the number of aligned features forms only a limited subset, typically of size smaller than d . We defer the proof of Lemma B.2 to Appendix G for the clarification of presentation.

Lemma B.2. *At initialization ($t = 0$), the following properties hold:*

(a) With high probability, for every $i \in [m]$,

$$\|\mathbf{w}_i^{(0)}\|_2^2 \in \left[\sigma_0^2 d_1 \left(1 - \tilde{O}\left(\frac{1}{\sqrt{d_1}}\right) \right), \sigma_0^2 d_1 \left(1 + \tilde{O}\left(\frac{1}{\sqrt{d_1}}\right) \right) \right]. \quad (29)$$

(b) With high probability, for every $i \in [m]$,

$$\|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2^2 \in \left[\sigma_0^2 d \left(1 - \tilde{O}\left(\frac{1}{\sqrt{d}}\right) \right), \sigma_0^2 d \left(1 + \tilde{O}\left(\frac{1}{\sqrt{d}}\right) \right) \right]. \quad (30)$$

(c) Let $m = d^{C_m}$ be the number of neurons. With probability at least $1 - o\left(\frac{1}{d^4}\right)$, for each $j \in [d]$,

$$|\mathcal{M}_j^*| \geq \Omega(d^{\omega_1}) =: \Xi_1, \quad |\mathcal{M}_j| \leq O(d^{\omega_2}) =: \Xi_2. \quad (31)$$

where

$$\omega_1 = C_m - \left(\frac{\epsilon_{\max}}{\epsilon_{\min}} \right)^2 (1 + \gamma), \quad \omega_2 = C_m - \left(\frac{\epsilon_{\min}}{\epsilon_{\max}} \right)^2 (1 - \gamma). \quad (32)$$

(d) For each $i \in [m]$, there are at most $O\left(d^{1 - \left(\frac{\epsilon_{\min}}{\epsilon_{\max}}\right)^2 \cdot (1 - \gamma)}\right)$ indices $j \in [d]$ such that $i \in \mathcal{M}_j$.

C THEOREM C.1

In this section we analyze the training process at the initial stage. Here we define the stage transition time

$$T_1 = \Theta\left(\frac{d_1 \log d}{\eta \log \log d}\right) \quad (33)$$

to be the iteration when

$$\|MM^\top \mathbf{w}_i^{(t)}\|_2^2 \geq \frac{1}{2} \|\mathbf{w}_i^{(t)}\|_2^2, \quad (34)$$

where the neuron weights are more concentrated in the feature space.

C.1 THEOREM C.1

Before stating Theorem C.1, we give a short description of its parts: (a) For all neurons, most of the weights lie in the feature subspace; (b) Lucky neurons are strongly aligned with their associated feature directions; (c) Neurons not in the set \mathcal{M}_j have only weak alignment with feature j ; (d) Each neuron can have strong alignment with only a limited number of features; and (e) All neuron weights have only small components in non-feature directions.

Theorem C.1 (Initial feature decoupling). *At iteration $t = T_1$, we have the following results:*

(a) For all $i \in [m]$,

$$\|MM^\top \mathbf{w}_i^{(T_1)}\|_2^2 \geq \frac{1}{2} \|\mathbf{w}_i^{(T_1)}\|_2^2. \quad (35)$$

(b) For each $j \in [d]$, and each $i \in \mathcal{M}_j^*$,

$$|\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| \geq \sqrt{1+\gamma} \frac{\sqrt{2 \log d}}{\sqrt{d}} \|\mathbf{w}_i^{(T_1)}\|_2. \quad (36)$$

(c) For each $j \in [d]$, and each $i \notin \mathcal{M}_j$,

$$|\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| \leq \sqrt{1-\gamma} \frac{\sqrt{2 \log d}}{\sqrt{d}} \|\mathbf{w}_i^{(T_1)}\|_2. \quad (37)$$

(d) For each $i \in [m]$,

$$|\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| \geq \frac{\log^{1/4} d}{\sqrt{d}} \|\mathbf{w}_i^{(T_1)}\|_2, \quad \text{for at most } \mathcal{O}\left(\frac{d}{2\sqrt{\log d}}\right) \text{ indices } j \in [d]. \quad (38)$$

(e) For each $i \in [m]$ and $j \in [d_1] \setminus [d]$,

$$|\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j^\perp \rangle| \leq \mathcal{O}\left(\sqrt{\frac{\log d}{d_1}}\right) \|\mathbf{w}_i^{(T_1)}\|_2. \quad (39)$$

C.2 USEFUL LEMMAS

In Lemma C.1, we show that for each neuron $i \in [m]$, the weight vector \mathbf{w}_i largely disregards the non-feature components M^\perp and instead focuses on the relevant features M .

We first describe Lemma C.1: (a) The projection of $\mathbf{w}_i^{(t)}$ onto the feature subspace, though initially small, grows rapidly during training and reaches the order of d_1 relative to its initialization. (b) The component of $\mathbf{w}_i^{(t)}$ in the non-feature subspace remains essentially unchanged, up to negligible variation.

Lemma C.1. *For all $t \leq T_1$, the following properties hold:*

(a)

$$\begin{aligned} \|MM^\top \mathbf{w}_i^{(t)}\|_2^2 &\leq \|MM^\top \mathbf{w}_i^{(0)}\|_2^2 \left(1 + \epsilon_{\max} \frac{\eta C_z \log \log d}{d}\right)^{2t} + \mathcal{O}\left(\frac{1}{d}\right) \|MM^\top \mathbf{w}_i^{(0)}\|_2^2, \\ \text{moreover, } \|MM^\top \mathbf{w}_i^{(t)}\|_2^2 &\leq \mathcal{O}\left(\|\mathbf{w}_i^{(0)}\|_2^2\right). \end{aligned} \quad (40)$$

(b)

$$\left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(t)} \right\|_2^2 \geq \left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)} \right\|_2^2 \left(1 - \eta\lambda + \epsilon_{\min} \frac{\eta C_z \log \log d}{d} \right)^{2t} - O\left(\frac{1}{d}\right) \left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)} \right\|_2^2. \quad (41)$$

(c)

$$\left\| \mathbf{M}^\perp (\mathbf{M}^\perp)^\top \mathbf{w}_i^{(t)} \right\|_2^2 \leq \left(1 + O\left(\frac{1}{\text{poly}(d)}\right) \right) \left\| \mathbf{M}^\perp (\mathbf{M}^\perp)^\top \mathbf{w}_i^{(0)} \right\|_2^2. \quad (42)$$

Lemma C.2. For each $i \in [m]$, there are at most $O(2^{-\sqrt{\log d}} d)$ indices $j \in [d]$ such that

$$|\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle| \geq \Omega(\sigma_0 \log^{1/4} d). \quad (43)$$

C.3 PROOF OF THEOREM C.1

Proof of Theorem C.1(a): The result (a) can be derived from Lemma C.1 (c). We have,

$$\begin{aligned} \left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(T_1)} \right\|_2^2 &= \left\| \mathbf{w}_i^{(T_1)} \right\|_2^2 - \left\| \mathbf{M}^\perp (\mathbf{M}^\perp)^\top \mathbf{w}_i^{(T_1)} \right\|_2^2 \\ &\geq \left\| \mathbf{w}_i^{(T_1)} \right\|_2^2 - \left(1 + \frac{1}{\text{poly}(d)} \right) \left\| \mathbf{M}^\perp (\mathbf{M}^\perp)^\top \mathbf{w}_i^{(0)} \right\|_2^2 \\ &\geq \left\| \mathbf{w}_i^{(T_1)} \right\|_2^2 - \left\| \mathbf{w}_i^{(0)} \right\|_2^2 \\ &\geq \left\| \mathbf{w}_i^{(T_1)} \right\|_2^2 - \frac{\left\| \mathbf{w}_i^{(T_1)} \right\|_2^2}{(1 + \epsilon_{\min} C_z \log d)} \\ &\geq \frac{1}{2} \left\| \mathbf{w}_i^{(T_1)} \right\|_2^2. \end{aligned} \quad (44)$$

□

Proof of Theorem C.1(b): Note that from similar gradient calculations to those in the proof of Lemma C.1 (b), we have, for $j \in [d]$ and $i \in \mathcal{M}_j^*$:

$$\begin{aligned} &|\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| \\ &= |\langle \mathbf{w}_i^{(T_1-1)}, \mathbf{M}_j \rangle - \eta \langle \nabla_{\mathbf{w}_i} L_{\text{aug}}(f_{T_1-1}), \mathbf{M}_j \rangle \pm \frac{\left\| \mathbf{w}_i^{((T_1-1))} \right\|_2}{\text{poly}(d_1)}| \\ &\geq |\langle \mathbf{w}_i^{(T_1-1)}, \mathbf{M}_j \rangle| \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d} \right) - \tilde{O} \left(\frac{\eta \left\| \mathbf{w}_i^{((T_1-1))} \right\|_2}{\text{poly}(d_1)} \right) \\ &\geq |\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle| \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d} \right)^{T_1} - \tilde{O} \left(\frac{\eta T_1 \left\| \mathbf{w}_i^{(T_1)} \right\|_2}{\text{poly}(d_1)} \right). \end{aligned} \quad (45)$$

These gradient descent steps above can be derived from the last few inequalities in the proof of Lemma C.1(b).

$$\begin{aligned} &|\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| \\ &\stackrel{\textcircled{1}}{\geq} \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)} \right\|_2 \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d} \right)^{T_1} - \tilde{O} \left(\frac{\eta T_1 \left\| \mathbf{w}_i^{(T_1)} \right\|_2}{\text{poly}(d_1)} \right) \\ &\stackrel{\textcircled{2}}{\geq} \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)} \right\|_2 \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d} \right)^{T_1} - \tilde{O} \left(\frac{\left\| \mathbf{w}_i^{(T_1)} \right\|_2}{\text{poly}(d)} \right) \\ &\stackrel{\textcircled{3}}{\geq} \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)} \right\|_2 \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d} \right)^{T_1} - \tilde{O} \left(\frac{\left\| \mathbf{w}_i^{(0)} \right\|_2}{\text{poly}(d)} \right) \\ &\stackrel{\textcircled{4}}{\geq} \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)} \right\|_2 \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d} \right)^{T_1} - \tilde{O} \left(\frac{\sqrt{\frac{d_1}{d}} \left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)} \right\|_2}{\text{poly}(d)} \right) \\ &\geq \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)} \right\|_2 \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d} \right)^{T_1} - \frac{\left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)} \right\|_2}{\text{poly}(d)} \\ &\stackrel{\textcircled{5}}{\geq} \sqrt{1 + \gamma} \frac{\sqrt{\log d}}{\sqrt{d}} \left\| \mathbf{w}_i^{(T_1)} \right\|_2. \end{aligned} \quad (46)$$

① is because Definition B.1 (b). ② is because $\frac{\eta T_1}{\text{poly}(d_1)} \leq \frac{1}{\text{poly}(d)}$. ③ is because $\|\mathbf{w}_i^{(t)}\|_2^2 \leq O(1) \|\mathbf{w}_i^{(0)}\|_2^2$ (equation 258). ④ is because Lemma B.2 (a) (b). ⑤ holds because the following equation is valid:

$$\begin{aligned}
& |\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| \\
& \geq \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2 \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d} \right)^{T_1} - \frac{\|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2}{\text{poly}(d)} \\
& = \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2 \left(1 - \eta\lambda + \epsilon_j \frac{d_1}{d} C_z \log d \right) - \frac{\|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2}{\text{poly}(d)} \\
& \geq \left(\frac{\epsilon_j}{\epsilon_{\max}} \right) \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2 \left(1 - \eta\lambda + \epsilon_{\max} \frac{d_1}{d} C_z \log d \right) - \frac{\|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2}{\text{poly}(d)} \\
& \geq \left(\frac{\epsilon_j}{\epsilon_{\max}} \right) \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(T_1)}\|_2 \\
& \stackrel{\textcircled{6}}{\geq} \frac{1}{\sqrt{2}} \left(\frac{\epsilon_j}{\epsilon_{\max}} \right) \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \|\mathbf{w}_i^{(T_1)}\|_2 \\
& \geq \frac{1}{\sqrt{2}} \left(\frac{\epsilon_{\min}}{\epsilon_{\max}} \right) \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \|\mathbf{w}_i^{(T_1)}\|_2 \\
& \geq \sqrt{1 + \gamma} \frac{\sqrt{\log d}}{\sqrt{d}} \|\mathbf{w}_i^{(T_1)}\|_2.
\end{aligned} \tag{47}$$

⑥ holds because of the conclusion of Theorem C.1(a). □

Proof of Theorem C.1(c): Theorem C.1(c) can be verified using Definition B.1 (b), Lemma B.2 (a) (b) together with the proof of Lemma C.1(a).

$$\begin{aligned}
& |\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| \\
& \leq |\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle| \left(1 + \epsilon_j \frac{\eta C_z \log \log d}{d} + \tilde{O}\left(\frac{\eta}{d^2}\right) \right)^{T_1} + \tilde{O}\left(\frac{\eta T_1 \|\mathbf{w}_i^{(T_1)}\|_2}{\text{poly}(d_1)}\right).
\end{aligned} \tag{48}$$

The above equation can be obtained from the first inequality in the proof of Lemma C.1(a).

$$\begin{aligned}
& |\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| \\
& \leq |\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle| \left(1 + \epsilon_j \frac{\eta C_z \log \log d}{d} + \tilde{O}\left(\frac{\eta}{d^2}\right) \right)^{T_1} + \frac{\|\mathbf{M}(\mathbf{M})^\top \mathbf{w}_i^{(0)}\|_2}{\text{poly}(d)} \\
& \leq \sqrt{\frac{c_2 \log d}{d}} \|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)}\|_2 \left(1 + \epsilon_j \frac{\eta C_z \log \log d}{d} + \tilde{O}\left(\frac{\eta}{d^2}\right) \right)^{T_1} + \frac{\|\mathbf{M}(\mathbf{M})^\top \mathbf{w}_i^{(0)}\|_2^2}{\text{poly}(d)} \\
& \leq \frac{\epsilon_j}{\epsilon_{\min}} \sqrt{\frac{c_2 \log d}{d}} \|\mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(T_1)}\|_2 + O\left(\frac{\|\mathbf{M}(\mathbf{M})^\top \mathbf{w}_i^{(0)}\|_2^2}{\text{poly}(d)}\right) \\
& \leq \frac{\epsilon_j}{\epsilon_{\min}} \sqrt{\frac{c_2 \log d}{d}} \|\mathbf{w}_i^{(T_1)}\|_2 + O\left(\frac{\|\mathbf{w}_i^{(T_1)}\|_2}{\text{poly}(d)}\right) \\
& \leq \frac{\epsilon_{\max}}{\epsilon_{\min}} \sqrt{\frac{c_2 \log d}{d}} \|\mathbf{w}_i^{(T_1)}\|_2 + O\left(\frac{\|\mathbf{w}_i^{(T_1)}\|_2}{\text{poly}(d)}\right) \\
& \leq \sqrt{1 + \gamma} \frac{\sqrt{2 \log d}}{\sqrt{d}} \|\mathbf{w}_i^{(T_1)}\|_2.
\end{aligned} \tag{49}$$

□

Proof of Theorem C.1(d): First, by Lemma C.2 we obtain that for each $i \in [m]$, there are at most $O(2^{-\sqrt{\log d}} d)$ indices $j \in [d]$ such that:

$$|\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle| \geq \Omega(\sigma_0 \log^{1/4} d). \quad (50)$$

Next, we proceed to the formal calculation:

$$\begin{aligned} |\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| &\geq |\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle| \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d} \right)^{T_1} \\ &\geq \Omega(\sigma_0 \log^{1/4} d) \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d} \right)^{T_1} \\ &\geq \Omega\left(\frac{\|\mathbf{w}_i^{(0)}\|_2}{\sqrt{d_1}} \log^{1/4} d\right) \Theta\left(\frac{d_1}{d}\right) \\ &\geq \frac{\log^{1/4} d}{\sqrt{d}} \|\mathbf{w}_i^{(T_1)}\|_2. \end{aligned} \quad (51)$$

□

Proof of Theorem C.1(e). At initialization we have

$$\mathbf{w}_i^{(0)} \sim \mathcal{N}(0, \sigma_0^2 I_{d_1}). \quad (52)$$

Hence for any unit vector \mathbf{M}_j^\perp , the projection satisfies

$$\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j^\perp \rangle \sim \mathcal{N}(0, \sigma_0^2). \quad (53)$$

By the standard Gaussian tail bound (sub-Gaussian with parameter σ_0),

$$\mathbb{P}\left(|\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j^\perp \rangle| > \sigma_0 \sqrt{2 \log d}\right) \leq 2 \exp\left(-\frac{t^2}{2\sigma_0^2}\right) = \frac{2}{d}. \quad (54)$$

Therefore, with high probability,

$$|\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j^\perp \rangle| \leq \sigma_0 \cdot O(\sqrt{\log d}). \quad (55)$$

Moreover, since $\|\mathbf{w}_i^{(0)}\|_2 = \Theta(\sigma_0 \sqrt{d_1})$ with high probability, the above bound is equivalently

$$|\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j^\perp \rangle| \leq O\left(\sqrt{\frac{\log d}{d_1}}\right) \cdot \|\mathbf{w}_i^{(0)}\|_2. \quad (56)$$

We have

$$\begin{aligned} &\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j^\perp \rangle \\ &= (1 - \eta\lambda) \langle \mathbf{w}_i^{(T_1-1)}, \mathbf{M}_j^\perp \rangle \pm \tilde{O}\left(\frac{\eta \sum_{i \in [m]} \|\mathbf{w}_i^{(t)}\|_2^2}{\tau d} \cdot \|\mathbf{w}_i^{(t)}\|_2\right) \\ &\leq (1 - \eta\lambda) \langle \mathbf{w}_i^{(T_1-1)}, \mathbf{M}_j^\perp \rangle + \tilde{O}\left(\frac{\eta \|\mathbf{w}_i^{(t)}\|_2}{\text{poly}(d_1)}\right) \\ &\leq |\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j^\perp \rangle| + O(T_1 \eta) \cdot \max_{t \leq T_1} \tilde{O}\left(\frac{\|\mathbf{w}_i^{(t)}\|_2}{\text{poly}(d_1)}\right) \\ &\leq O\left(\sqrt{\frac{\log d}{d_1}}\right) \cdot \|\mathbf{w}_i^{(0)}\|_2 + O(T_1 \eta) \cdot \max_{t \leq T_1} \tilde{O}\left(\frac{\|\mathbf{w}_i^{(t)}\|_2}{\text{poly}(d_1)}\right) \\ &\leq O\left(\sqrt{\frac{\log d}{d_1}}\right) \cdot \|\mathbf{w}_i^{(T_1)}\|_2 + O(T_1 \eta) \cdot \max_{t \leq T_1} \tilde{O}\left(\frac{\|\mathbf{w}_i^{(t)}\|_2}{\text{poly}(d_1)}\right) \\ &\stackrel{\textcircled{v}}{\leq} O\left(\sqrt{\frac{\log d}{d_1}}\right) \|\mathbf{w}_i^{(T_1)}\|_2. \end{aligned} \quad (57)$$

⑦ is because $\frac{T_1 \eta}{\text{poly}(d_1)} \ll \sqrt{\frac{\log d}{d_1}}$ □

Lemma 3.1 can be viewed as an informal version of Theorem C.1. In particular, part (a) of Lemma 3.1 corresponds to the first inequality in the proof of Theorem C.1(b), while part (b) of Lemma 3.1 corresponds to the first inequality in the proof of Theorem C.1(e). Hence, Lemma 3.1 is essentially a simplified restatement of the more general Theorem C.1.

C.4 PROOF OF LEMMA 3.1

Proof of Lemma 3.1. For $j \in [d]$ and $i \in [m]$, the following bounds hold for all $t \in [0, T_1]$:

(a) Lower bound:

$$|\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle| \geq |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| (1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log \log d}{d}) - \tilde{O}\left(\frac{\eta \|\mathbf{w}_i^{(t)}\|_2}{\text{poly}(d_1)}\right). \quad (58)$$

(b) Orthogonal component:

$$|\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j^\perp \rangle| \leq (1 - \eta\lambda) |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j^\perp \rangle| + \tilde{O}\left(\frac{\eta \|\mathbf{w}_i^{(t)}\|_2}{\text{poly}(d_1)}\right). \quad (59)$$

□

D THEOREM D.1

The second stage is defined as the iterations $t \geq T_1$ but $t \leq T_2$, where

$$T_2 = T_1 + \Theta\left(\frac{d\tau \log d}{\epsilon_{\max} \eta \log \log d}\right) \quad (60)$$

is defined as the iteration when one of the neuron $i \in [m]$ satisfies

$$\left\| \mathbf{w}_i^{(T_2)} \right\|_2^2 \geq d \left\| \mathbf{w}_i^{(T_1)} \right\|_2^2. \quad (61)$$

D.1 THEOREM D.1

We first provide an explanation of Theorem D.1: (a) If a neuron i is a lucky neuron for feature j , then the projection of $\mathbf{w}_i^{(T_2)}$ onto \mathbf{M}_j is very large, on the order of the full neuron weight $\|\mathbf{w}_i^{(T_2)}\|_2$. In other words, such neurons have already “focused” on \mathbf{M}_j . (b) The bias term $\mathbf{b}_i^{(T_2)}$ grows proportionally with the neuron weight $\|\mathbf{w}_i^{(T_2)}\|_2$, and at iteration T_2 it reaches at least $\frac{\text{polylog}(d)}{\sqrt{d}} \|\mathbf{w}_i^{(T_2)}\|_2$. In other words, the continuously increasing bias effectively controls the activation of the neuron $\mathbf{w}_i^{(T_2)}$. (c) Among the lucky neurons in \mathcal{M}_j^* , there exists one neuron $\mathbf{w}_i^{(T_2)}$ whose projection onto \mathbf{M}_j is the largest, and this neuron has a larger projection than all the other neurons in \mathcal{M}_j .

Theorem D.1 (Emergence of singletons). *For each neuron $i \in [m]$, the following conditions hold at iteration $t = T_2$:*

(a) For each $j \in [d]$, if $i \in \mathcal{M}_j^*$, then

$$|\langle \mathbf{w}_i^{(T_2)}, \mathbf{M}_j \rangle| \geq \Omega\left(\frac{\epsilon_{\min}}{\epsilon_{\max}}\right) \|\mathbf{w}_i^{(T_2)}\|_2. \quad (62)$$

(b)

$$\mathbf{b}_i^{(T_2)} \geq \frac{\text{polylog}(d)}{\sqrt{d}} \|\mathbf{w}_i^{(T_2)}\|_2. \quad (63)$$

(c) Let

$$\alpha_j^* = \max_{i \in \mathcal{M}_j^*} |\langle \mathbf{w}_i^{(T_2)}, \mathbf{M}_j \rangle|, \quad (64)$$

then there exists a constant $C_j = \Theta(1)$ such that

$$|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| \leq C_j \alpha_j^*, \quad \forall i \in \mathcal{M}_j. \quad (65)$$

D.2 USEFUL LEMMAS

Next, we discuss Lemma D.1. For example, the first item illustrates how each feature \mathbf{M}_j can be captured by certain subsets of neurons, a process influenced by the stochastic nature of initialization. We elaborate on the full content of Lemma D.1 below.

(a) Lucky neurons have large projection on their feature direction, which means they can be activated and are clearly aligned with that feature. (b) If a neuron does not belong to \mathcal{M}_j , then its projection on feature j stays small, which means it cannot be activated and has only weak alignment. (c) A neuron can only be well aligned with a small number of features, not with many at the same time. (d) The projection of a neuron weight on non-feature directions is very small, which means the neuron does not learn the non-feature directions. (e) The size of each neuron weight is controlled by its bias, so the weight does not grow without limit.

Lemma D.1. *For all iterations $t \in (T_1, T_2]$, the neurons $i \in [m]$ satisfy the following properties:*

(a) *For $j \in [d]$, if $i \in \mathcal{M}_j^*$, then*

$$\left| \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \right| \geq \sqrt{1 + \gamma} \mathbf{b}_i^{(t)}. \quad (66)$$

(b) *For $j \in [d]$, if $i \notin \mathcal{M}_j$, then*

$$\left| \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \right| \leq \sqrt{1 - \gamma} \mathbf{b}_i^{(t)}, \quad (67)$$

and furthermore,

$$\left| \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \right| \leq \tilde{\mathcal{O}} \left(\frac{\|\mathbf{w}_i^{(t)}\|_2}{\sqrt{d}} \right). \quad (68)$$

(c) *For each $i \in [m]$, there are at most $\mathcal{O}(2^{-\sqrt{\log d}} d)$ many $j \in [d]$ such that*

$$\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 \geq \frac{(\mathbf{b}_i^{(t)})^2}{\sqrt{\log d}}. \quad (69)$$

(d) *For each $i \in [m]$, and for all $j \in [d_1] \setminus [d]$,*

$$\left| \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j^\perp \rangle \right| \leq \tilde{\mathcal{O}} \left(\frac{\|\mathbf{w}_i^{(t)}\|_2}{\sqrt{d_1}} \right). \quad (70)$$

(e) *For all $i \in [m]$,*

$$\|\mathbf{w}_i^{(t)}\|_2^2 \leq \frac{d(\mathbf{b}_i^{(t)})^2}{\log d}. \quad (71)$$

Lemma D.2. *For each $i \in [m]$, define*

$$\Lambda_i := \left\{ j \in [d] : |\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle| \leq \frac{\sigma_0}{d} \right\} \subseteq [d]. \quad (72)$$

Then

$$|\Lambda_i| = \mathcal{O} \left(\frac{d}{\text{polylog}(d)} \right). \quad (73)$$

D.3 PROOF OF THEOREM D.1

Proof of Theorem D.1: We follow similar analysis as in the proof of Lemma D.1. In order to prove (a)-(c), we have to discuss the two substages of the learning process below.

When all $\|\mathbf{w}_i^{(t)}\|_2 \leq (1 + \frac{\varepsilon_{\min}}{\varepsilon_{\max}}) \|\mathbf{w}_i^{(T_1)}\|_2$: From similar analysis in the proof of Lemma D.1, the iteration complexity for a neuron $i \in [m]$ to reach $\|\mathbf{w}_i^{(t)}\|_2 \geq (1 + \frac{\varepsilon_{\min}}{\varepsilon_{\max}}) \|\mathbf{w}_i^{(T_1)}\|_2$ is no smaller than

$$T'_{i,1} := \max \left\{ T_1 + \Omega \left(\frac{d \log d}{\varepsilon_{\max} \eta \log \log d} \right), T_2 \right\}. \quad (74)$$

When some $\|\mathbf{w}_i^{(t)}\|_2 \geq (1 + \frac{\epsilon_j}{\epsilon_{\max}})\|\mathbf{w}_i^{(T_1)}\|_2$.

We first prove Theorem D.1(a). In the first stage, for $j \notin \mathcal{N}_i$, we have

$$\begin{aligned} \sum_{j \in [d], j \notin \mathcal{N}_i} \langle \mathbf{w}_i^{(T'_{i,1})}, \mathbf{M}_j \rangle^2 &\leq \sum_{j \in [d], j \notin \mathcal{N}_i} \langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle^2 \left(1 + \epsilon_j \frac{O(\eta)}{d \text{polylog}(d)}\right)^{T_2} + \tilde{O}\left(\frac{\|\mathbf{w}_i^{(T_1)}\|_2^2}{d^{3/2}}\right) \\ &\leq (1 + o(1)(\frac{\epsilon_j}{\epsilon_{\max}}) + o(1)(\frac{\epsilon_j}{\epsilon_{\max}})^2) \|\mathbf{M}\mathbf{M}^\top \mathbf{w}_i^{(T_1)}\|_2^2, \end{aligned} \quad (75)$$

where we used the fact that $\|\mathbf{w}_i^{(T_1)}\|_2 \lesssim \|\mathbf{M}\mathbf{M}^\top \mathbf{w}_i^{(T_1)}\|_2$.

For $j \in [d_1] \setminus [d]$ we have

$$\begin{aligned} &\sum_{j \in [d_1] \setminus [d]} \langle \mathbf{w}_i^{(T'_{i,1})}, \mathbf{M}_j^\perp \rangle^2 \\ &\leq \sum_{j \in [d_1] \setminus [d]} \langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j^\perp \rangle^2 + O\left(\frac{\eta(T'_{i,1} - T_1)}{d}\right) e^{-\Omega(\log^{1/4} d)} \max_{t' \in [T_1, T'_{i,1}]} \|\mathbf{w}_i^{(t')}\|_2^2 \\ &\leq \left(1 + o\left(\frac{\epsilon_j}{\epsilon_{\max}}\right)\right) \|\mathbf{M}^\perp (\mathbf{M}^\perp)^\top \mathbf{w}_i^{(T_1)}\|_2^2. \end{aligned} \quad (76)$$

Typically, if $i \in \mathcal{M}_j^*$, there exists $t \leq T_2$ such that $\|\mathbf{w}_i^{(t)}\|_2 \geq (1 + \frac{\epsilon_j}{\epsilon_{\max}})\|\mathbf{w}_i^{(T_2)}\|_2$, as we have argued in the proof of Lemma D.1. Thus, we have

$$\begin{aligned} |\langle \mathbf{w}_i^{(T'_{i,1})}, \mathbf{M}_j \rangle|^2 &\geq \|\mathbf{w}_i^{(T'_{i,1})}\|_2^2 - \sum_{j \in [d], j \notin \mathcal{N}_i} \langle \mathbf{w}_i^{(T'_{i,1})}, \mathbf{M}_j \rangle^2 - \sum_{j \in [d_1] \setminus [d]} \langle \mathbf{w}_i^{(T'_{i,1})}, \mathbf{M}_j^\perp \rangle^2 \\ &\geq (1 + \frac{\epsilon_j}{\epsilon_{\max}})^2 \|\mathbf{w}_i^{(T_1)}\|_2^2 - (1 + o(1)(\frac{\epsilon_j}{\epsilon_{\max}}) + o(1)(\frac{\epsilon_j}{\epsilon_{\max}})^2) \|\mathbf{w}_i^{(T_1)}\|_2^2 \\ &\geq \frac{\epsilon_j}{\epsilon_{\max}} \cdot (2 - o(1)) \|\mathbf{w}_i^{(T_1)}\|_2^2, \end{aligned} \quad (77)$$

which proves the claim.

In the second stage, if $i \in \mathcal{M}_j^*$, then from similar calculations as above, we can prove by induction that starting from $t = T'_{i,1}$, it holds:

$$\begin{aligned} |\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle| &\geq |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| \left(1 + \Omega\left(\epsilon_j \frac{\eta \log \log d}{d}\right)\right) \\ &\geq \|\mathbf{w}_i^{(t)}\|_2 \left(1 + \Omega\left(\epsilon_j \frac{\eta \log \log d}{d}\right)\right) \\ \sum_{j' \in [d], j' \neq j} \langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_{j'} \rangle^2 &\leq \sum_{j' \in [d], j' \neq j} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_{j'} \rangle^2 \left(1 + \epsilon_j \frac{O(\eta)}{d \text{polylog}(d)}\right)^2 \\ \sum_{j \in [d_1] \setminus [d]} \langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j^\perp \rangle^2 &\leq \sum_{j \in [d_1] \setminus [d]} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j^\perp \rangle^2 \left(1 + \frac{O(\eta)}{d \text{polylog}(d)}\right)^2, \end{aligned} \quad (78)$$

which implies

$$|\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle| \geq |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| \cdot \frac{\|\mathbf{w}_i^{(t+1)}\|_2}{\|\mathbf{w}_i^{(t)}\|_2} \geq (1 - o(1)) \|\mathbf{w}_i^{(t+1)}\|_2. \quad (79)$$

Next, we prove Theorem D.1(b). In the first stage, the bias growth is large, i.e.,

$$\begin{aligned} \mathbf{b}_i^{(T'_{i,1})} &\geq \mathbf{b}_i^{(T_1)} \left(1 + \frac{\eta}{d}\right)^{T'_{i,1} - T_1} \geq \mathbf{b}_i^{(T_1)} \cdot \text{polylog}(d) \\ &\geq \frac{\text{polylog}(d)}{\sqrt{d}} \|\mathbf{w}_i^{(T_1)}\|_2 \geq \frac{\text{polylog}(d)}{\sqrt{d}} \|\mathbf{w}_i^{(T'_{i,1})}\|_2. \end{aligned} \quad (80)$$

In the second stage, the bias is large consistently, i.e.,

$$\mathbf{b}_i^{(t+1)} \geq \mathbf{b}_i^{(t)} \cdot \frac{\|\mathbf{w}_i^{(t+1)}\|_2}{\|\mathbf{w}_i^{(t)}\|_2} \geq \frac{\text{polylog}(d)}{\sqrt{d}} \|\mathbf{w}_i^{(t+1)}\|_2 \geq \frac{1}{4} \|\mathbf{w}_i^{(T'_{i,1})}\|_2. \quad (81)$$

Finally, we prove Theorem D.1(c): Assuming $\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle > 0$ (the opposite case is similar), from $t = T_1$, for $i \in \mathcal{M}_j^*$, we have

$$\begin{aligned} \langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle &= \left(\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle - \mathbf{b}_i^{(t)} \right) \left(1 + \epsilon_j \frac{\eta C_z \log \log d}{d} \right) \pm O \left(\frac{\eta |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle|}{d \text{polylog}(d)} \right) \\ &\geq \Omega(1) \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \left(1 + \epsilon_j \frac{\eta C_z \log \log d}{d} \left(1 - \frac{1}{\text{polylog}(d)} \right) \right) \\ &\geq \Omega(1) \langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle \left(1 + \epsilon_j \frac{\eta C_z \log \log d}{d} \left(1 - \frac{1}{\text{polylog}(d)} \right) \right)^{t+1-T_1}. \end{aligned} \quad (82)$$

which implies that after certain iteration $t = T_1 + T'$, where $T' = \Theta\left(\frac{d}{\eta}\right)$, we shall have

$$|\langle \mathbf{w}_i^{(T_1+T')}, \mathbf{M}_j \rangle| \geq \log \log d \cdot |\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| \geq \mathbf{b}_i^{(T_1)} \cdot \log \log d. \quad (83)$$

However, at iteration $t = T_1 + \Theta\left(\frac{d}{\eta}\right)$, we can see from previous analysis that $\|\mathbf{w}_i^{(t)}\|_2 \leq (1 + o(1))\|\mathbf{w}_i^{(T_1)}\|_2$, so the bias growth can be bounded as

$$\begin{aligned} \mathbf{b}_i^{(t)} &\leq \mathbf{b}_i^{(T_1)} \left(1 + \frac{\eta}{d} \right)^{\Theta\left(\frac{d}{\eta}\right)} \cdot \max \left\{ \frac{\|\mathbf{w}_i^{(t)}\|_2}{\|\mathbf{w}_i^{(T_1)}\|_2}, 1 \right\} \\ &\leq \mathbf{b}_i^{(T_1)} \left(1 + \frac{\eta}{d} \cdot \Theta\left(\frac{d}{\eta}\right) \right) \cdot \max \{ (1 + o(1)), 1 \} \\ &\leq O(\mathbf{b}_i^{(T_1)}). \end{aligned} \quad (84)$$

Now from our initialization properties in Lemma D.2, we have that $\langle \mathbf{w}_{i'}^{(0)}, \mathbf{M}_j \rangle^2 \leq O(\sigma_0^2 \log d)$ for all $i \in [m]$. Thus via similar arguments, we also have

$$|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle| \leq |\langle \mathbf{w}_{i'}^{(0)}, \mathbf{M}_j \rangle| \left(1 + \epsilon_j \frac{\eta C_z \log \log d}{d} \left(1 \pm \frac{1}{\text{polylog}(d)} \right) \right)^t. \quad (85)$$

holds for all $i' \in [m]$. Now it is easy to see that for $t \leq T_2 = T_1 + \Theta\left(\frac{d \tau \log d}{\eta \log \log d}\right)$, we have

$$\begin{aligned} \frac{|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle|} &\geq \Omega(1) \cdot \frac{|\langle \mathbf{w}_i^{(T_1)}, \mathbf{M}_j \rangle| \left(1 + \epsilon_j \frac{\eta C_z \log \log d}{d} \left(1 - \frac{\eta}{\text{polylog}(d)} \right) \right)^{t-T_1}}{|\langle \mathbf{w}_{i'}^{(T_1)}, \mathbf{M}_j \rangle| \left(1 + \epsilon_j \frac{\eta C_z \log \log d}{d} + \frac{\eta}{d \text{polylog}(d)} \right)^{t-T_1}} \\ &\geq \left(1 - O \left(\epsilon_j \frac{\eta \log \log d}{d \text{polylog}(d)} \right) \right)^{t-T_1} \geq \Omega(1). \end{aligned} \quad (86)$$

Thus, the last claim is proved. \square

D.4 PROOF OF LEMMA 3.2

Lemma 3.2 can be viewed as an informal version of Theorem D.1. In particular, part (a) of Lemma 3.2 corresponds to (77) and Lemma B.2 (c), while part (b) of Lemma 3.2 corresponds to another formulation of Theorem D.1 (c).

E THEOREM E.1

E.1 THEOREM E.1

At the final stage, we show that sparse activation of neurons naturally leads to convergence toward sparse solutions, thereby guaranteeing sparse representations. For all $t \geq T_2$:

Theorem E.1. *For all iterations t , the neurons $i \in [m]$ satisfy the following properties:*

(a) *For $j \in [d]$, if $i \in \mathcal{M}_j^*$, then*

$$|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| \geq \Omega(1) \|\mathbf{w}_i^{(t)}\|_2. \quad (87)$$

(b) *For $i \in [m]$, we have*

$$\|\mathbf{w}_i^{(t)}\|_2 \leq O(1). \quad (88)$$

(c) *For each $j \in [d]$,*

$$\mathfrak{F}_j^{(t)} := \sum_{i \in \mathcal{M}_j} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 = \Theta\left(\left(\frac{\epsilon_j}{\epsilon_{\max}}\right)^2 \tau \log^3 d\right). \quad (89)$$

(d) *Let $j \in [d]$ and $i \in \mathcal{M}_j^*$, then there exists $C = \Theta(1)$ such that*

$$|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| \geq C \max_{i' \in \mathcal{M}_j} |\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle|. \quad (90)$$

(e) *For $i \notin \mathcal{M}_j$, it holds*

$$|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| \leq O\left(\frac{\epsilon_j}{\epsilon_{\max}} \frac{1}{\sqrt{d} \Xi_2^5}\right) \|\mathbf{w}_i^{(t)}\|_2. \quad (91)$$

(f) *For any $i \in [m]$ and any $j \in [d_1] \setminus [d]$, it holds*

$$|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j^\perp \rangle| \leq O\left(\frac{1}{\sqrt{d_1} \Xi_2^5}\right) \|\mathbf{w}_i^{(t)}\|_2. \quad (92)$$

(g) *For all $i \in [m]$, the bias satisfies*

$$\mathbf{b}_i^{(t)} \geq \frac{\text{polylog}(d)}{\sqrt{d}} \|\mathbf{w}_i^{(t)}\|_2. \quad (93)$$

E.2 USEFUL LEMMAS

When all the conditions in Theorem E.1 hold for some iteration $t \geq T_2$, we have the following fact, which is a simple corollary of Lemma E.9.

Lemma E.1. *For any $i \in [m]$, we denote $\mathcal{N}_i = \{j \in [d] : i \in \mathcal{M}_j\}$. Suppose Theorem E.1 holds at iteration $t \geq T_2$, then with high probability over $x \in \mathcal{D}_x$:*

$$\max_{x \in \{\mathbf{X}_n, \mathbf{Y}_n\}} \mathbf{1}_{h_{i,t}(x) \neq 0} \leq \sum_{j \in \mathcal{N}_i} \mathbf{1}_{|\hat{z}_{p,j}| \neq 0}, \quad (94)$$

which implies that

$$\max_{x \in \{\mathbf{X}_n, \mathbf{Y}_n\} \cup \mathfrak{N}} \Pr(h_{i,t}(x) \neq 0) \leq O\left(\frac{\log \log d}{d}\right). \quad (95)$$

Now for the simplicity of calculations, we define the following notations which are used through out this section

Definition E.1 (Expansion of gradient). For each $i \in [m]$, $j \in [d]$, we expand $\langle \nabla_{\mathbf{w}_i} L(f_t), \mathbf{M}_j \rangle$ as

$$\begin{aligned} & \langle \nabla_{\mathbf{w}_i} L(f_t), \mathbf{M}_j \rangle \\ &= \mathbb{E} \left[\left((\ell'_{p,t} - 1) h_{i,t}(\mathbf{Y}_n) + \sum_{\mathbf{X}_{n,s} \in \mathfrak{N}} \ell'_{s,t} h_{i,t}(\mathbf{X}_{n,s}) \right) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq b_i} \langle \mathbf{z}_{\mathbf{X}}^{(r)}, \mathbf{M}_j \rangle \right], \end{aligned} \quad (96)$$

and

$$\langle \nabla_{\mathbf{w}_i} L(f_t), \mathbf{M}_j \rangle = \Psi_{i,j}^{(t)} + \Phi_{i,j}^{(t)} + \mathcal{E}_{i,j}^{(t)}, \quad (97)$$

where the $\Psi^{(t)}$, $\Phi^{(t)}$, $\mathcal{E}^{(t)}$ are defined as follows. For each

$$\mathbf{z}_{\mathbf{X}} = \frac{1}{L} \left(\sum_j \mathbf{M}_j \tilde{\mathbf{z}}_{n,j} + \tilde{\xi}_n \right) \sim \mathcal{D}_{\mathbf{z}_{\mathbf{X}}}, \quad \mathbf{z}_{\mathbf{Y}} = \frac{1}{L} \left(\sum_j \mathbf{M}_j \tilde{\mathbf{z}}_{n,j}^+ + \tilde{\xi}_n^+ \right) \sim \mathcal{D}_{\mathbf{z}_{\mathbf{Y}}}, \quad (98)$$

we write

$$\begin{aligned} \psi_{i,j}^{(t)}(\mathbf{Y}_n) &= \sum_{s=1}^L \left[\left(\frac{1}{L} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \tilde{\mathbf{z}}_{n,j}^{+(s)} - \mathbf{b}_i^{(t)} \right) \mathbf{1}_{\langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{Y}}^{(s)} \rangle > \mathbf{b}_i^{(t)}} \right. \\ &\quad \left. - \left(\frac{1}{L} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \tilde{\mathbf{z}}_{n,j}^{+(s)} + \mathbf{b}_i^{(t)} \right) \mathbf{1}_{\langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{Y}}^{(s)} \rangle < -\mathbf{b}_i^{(t)}} \right], \end{aligned} \quad (99)$$

$$\phi_{i,j}^{(t)}(\mathbf{Y}_n) = \sum_{s=1}^L \langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{Y}}^{(s)} \rangle \mathbf{1}_{\langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{Y}}^{(s)} \rangle > \mathbf{b}_i^{(t)}} - \langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{Y}}^{(s)} \rangle \mathbf{1}_{\langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{Y}}^{(s)} \rangle < -\mathbf{b}_i^{(t)}}. \quad (100)$$

Now we define

$$\Psi_{i,j}^{(t)} := \mathbb{E} \left[\left((\ell'_{p,t} - 1) \cdot \psi_{i,j}^{(t)}(\mathbf{Y}_n) + \sum_{\mathbf{X}_{n,s} \in \mathfrak{N}} \ell'_{s,t} \cdot \psi_{i,j}^{(t)}(\mathbf{X}_{n,s}) \right) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq b_i} \tilde{\mathbf{z}}_{n,j}^{(r)} \right], \quad (101)$$

$$\Phi_{i,j}^{(t)} := \mathbb{E} \left[\left((\ell'_{p,t} - 1) \cdot \phi_{i,j}^{(t)}(\mathbf{Y}_n) + \sum_{\mathbf{X}_{n,s} \in \mathfrak{N}} \ell'_{s,t} \cdot \phi_{i,j}^{(t)}(\mathbf{X}_{n,s}) \right) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq b_i} \tilde{\mathbf{z}}_{n,j}^{(r)} \right], \quad (102)$$

$$\mathcal{E}_{i,j}^{(t)} := \mathbb{E} \left[\left((\ell'_{p,t} - 1) \cdot h_{i,t}(\mathbf{Y}_n) + \sum_{\mathbf{X}_{n,s} \in \mathfrak{N}} \ell'_{s,t} \cdot h_{i,t}(\mathbf{X}_{n,s}) \right) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq b_i} \langle \mathbf{M}_j, \tilde{\xi}_n^{(r)} \rangle \right]. \quad (103)$$

Moreover, for $j \in [d_1] \setminus [d]$, we can similarly define

$$\Psi_{i,j}^{(t)}, \Phi_{i,j}^{(t)} \equiv 0, \quad (104)$$

$$\mathcal{E}_{i,j}^{(t)} := \mathbb{E} \left[\left((\ell'_{p,t} - 1) \cdot h_{i,t}(\mathbf{Y}_n) + \sum_{\mathbf{X}_{n,s} \in \mathfrak{N}} \ell'_{s,t} \cdot h_{i,t}(\mathbf{X}_{n,s}) \right) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq b_i} \langle \mathbf{M}_j^\perp, \tilde{\xi}_n^{(r)} \rangle \right]. \quad (105)$$

Equipped with the above definition, we are ready to characterize the training process at the final stage.

Lemma E.2 (Lower bound for $\Psi_1^{(t)}$). Suppose Theorem E.1 holds at iteration t . For $j \in [d]$ and $i \in \mathcal{M}_j^*$, there exists $G_1 = \Theta(1)$ such that if

$$\mathfrak{F}_j^{(t)} := \sum_{i' \in \mathcal{M}_j} \langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle^2 \left(\sum_{r=1}^L \tilde{\mathbf{z}}_{n,j}^{(r)} \right)^2 \leq \left(\frac{\epsilon_j}{\epsilon_{\max}} \right)^2 G_1 \tau \log d, \quad (106)$$

then we have

$$\begin{aligned} & \Psi_{i,j}^{(t)} \cdot \text{sign} \left(\sum_{s=1}^L \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \tilde{\mathbf{z}}_{n,j}^{+(s)} \right) \\ & \geq \frac{\mathbb{E} \left[\sum_{r=1}^L |\tilde{\mathbf{z}}_{n,j}^{(r)}| \right]}{\text{polylog}(d)} \left(1 - O \left(\frac{1}{\Xi_2^3} \right) \right) \left(\sum_{s=1}^L |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \tilde{\mathbf{z}}_{n,j}^{+(s)} - \mathbf{b}_i^{(t)}| \right). \end{aligned} \quad (107)$$

Lemma E.3 (Upper bound for $\Psi_{i,j}^{(t)}$). *Let $j \in [d]$ and $i \in \mathcal{M}_j^*$. Suppose Theorem E.1 holds at iteration t , then there exists a constant $G_2 = \Theta(1)$ such that if*

$$\mathfrak{F}_j^{(t)} := \sum_{j: i \in \mathcal{M}_j} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 \left(\sum_{s=1}^L \tilde{\mathbf{z}}_{n,j}^{+(s)} \right)^2 \geq \left(\frac{\epsilon_j}{\epsilon_{\max}} \right)^2 G_2 \tau \log d, \quad (108)$$

we have

$$\Psi_{i,j}^{(t)} \leq \frac{1}{\text{poly}(d)} \sum_{s=1}^L |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \tilde{\mathbf{z}}_{n,j}^{+(s)}|. \quad (109)$$

Similarly, for $i \in \mathcal{M}_j$, we have

$$\Psi_{i,j}^{(t)} \leq \frac{1}{\text{poly}(d)} \sum_{s=1}^L |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \tilde{\mathbf{z}}_{n,j}^{+(s)}| + O \left(\frac{1}{d^2} \right) \mathbf{b}_i^{(t)}. \quad (110)$$

Lemma E.4. *At iteration $t \geq T_2$, let $j \in [d]$ and $i \in [m]$. Suppose Theorem E.1 holds at t . Then for each $j \in [d_1]$, we have*

$$|\mathcal{E}_{i,j}^{(t)}| \leq O \left(\frac{\Xi_2^2 \|\mathbf{w}_i^{(t)}\|_2}{d^2 \tau} \right) \cdot \max_{i' \in [m]} \left(|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle| \right). \quad (111)$$

Lemma E.5 (Reduction of $\Phi^{(t)}$ to the bounds of $\Psi^{(t)}$). *Let $j \in [d]$ and $i \in \mathcal{M}_j$. Suppose Theorem E.1 holds for all iterations before $t \in \left[\frac{d^{1.01}}{\eta}, \frac{d^{1.99}}{\eta} \right]$ and after T_2 . Also suppose that for all $l \in [d]$, we have*

$$\mathfrak{F}_l^{(t')} = \Omega \left(\left(\frac{\epsilon_j}{\epsilon_{\max}} \right)^2 \tau \log d \right) \quad \text{at some } t' = \Theta(T_2). \quad (112)$$

Then the following bounds hold:

For iteration $t \in \left[\frac{d^{1.01}}{\eta}, \frac{d^{1.495}}{\eta} \right]$,

$$\Phi_{i,j}^{(t)} \leq \tilde{O} \left(\frac{\epsilon_j}{\epsilon_{\max}} \cdot \frac{\Xi_2^2}{d^{3/2}} \right) \|\mathbf{w}_i^{(t)}\|_2. \quad (113)$$

For iteration $t \in \left[\frac{d^{1.495}}{\eta}, \frac{d^{1.99}}{\eta} \right]$,

$$\Phi_{i,j}^{(t)} \leq \tilde{O} \left(\frac{1}{d^{1.98}} \right) \|\mathbf{w}_i^{(t)}\|_2. \quad (114)$$

Definition E.2 (Optimal Learner). *We define a learner network that we deem as the optimal feature map for this task. Let $\kappa > 0$, we define $\theta^* := \{\theta_i^*\}_{i \in [m]}$ as follows:*

$$\theta_i^* = \begin{cases} \frac{\sqrt{\tau} \kappa}{|\mathcal{M}_j^*|} \mathbf{M}_j \cdot \text{sign}(\langle \mathbf{w}_i^{(T_2)}, \mathbf{M}_j \rangle), & \text{if } i \in \mathcal{M}_j^*, \\ 0, & \text{if } i \notin \bigcup_{j \in [d]} \mathcal{M}_j^*. \end{cases} \quad (115)$$

Furthermore, we define the optimal feature map f_t^* as follows. For $i \in [m]$, the i -th neuron of $f_{t,\theta}$ given weight $\theta_i \in \mathbb{R}^{d_1}$ is

$$f_{t,\theta,i}(\mathbf{X}_n) = \sum_{r=1}^L \left[(\langle \theta_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle - \mathbf{b}_i) \mathbf{1}_{\langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle \geq \mathbf{b}_i} - (-\langle \theta_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle - \mathbf{b}_i) \mathbf{1}_{-\langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle \geq \mathbf{b}_i} \right]. \quad (116)$$

Finally, we write $f_{t,\theta}$ as the concatenation

$$f_{t,\theta}(\cdot) = (f_{t,\theta,1}(\cdot), \dots, f_{t,\theta,m}(\cdot))^\top. \quad (117)$$

Lemma E.6 (Optimality). *Let $\{\theta_i^*\}_{i \in [m]}$ and $f_{t,\theta}$ be defined as in Definition E.1. When Theorem E.1, define the pseudo loss function*

$$\tilde{L}(f_{t,\theta^*}, f_t) := \mathbb{E} \left[-\tau \log \left(\frac{e^{\langle f_{t,\theta^*}(\mathbf{X}_n), f_t(\mathbf{Y}_n) \rangle / \tau}}{\sum_{\mathbf{X} \in \mathfrak{B}} e^{\langle f_{t,\theta^*}(\mathbf{X}_n), f_t(\mathbf{X}) \rangle / \tau}} \right) \right]. \quad (118)$$

Then by choosing $\kappa = \Theta(\Xi_2)$, and assuming

$$\sum_{i \in \mathcal{M}_j^*} |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| \geq \Omega \left(\frac{\sqrt{\tau}}{\Xi_2} \right), \quad (119)$$

we obtain the following loss guarantee:

$$\tilde{L}(f_{t,\theta^*}, f_t) \leq O \left(\frac{1}{\log d} \right). \quad (120)$$

Lemma E.7 (Pre-activation size I). *Let $\mathbf{z}_{\mathbf{X}}^{(r)} = \frac{1}{L} (\mathbf{M} \tilde{\mathbf{z}}_n^{(r)} + \tilde{\xi}_n^{(r)}) \sim \mathcal{D}_{\mathbf{z}_{\mathbf{X}}}$, $\mathbf{w}_i \in \mathbb{R}^{d_1}$. Define $\mathbf{z}_{\mathbf{X}}^{(r) \setminus j} = \frac{1}{L} (\sum_{j' \neq j, j' \in [d]} \mathbf{M}_{j'} \tilde{\mathbf{z}}_{n,j'}^{(r)} + \tilde{\xi}_n^{(r)})$. Then the following results hold:*

(a) *Naive Chebyshev bound: For any $\lambda > 0$,*

$$\Pr_{\tilde{\mathbf{z}}_n^{(r) \setminus j}, \tilde{\xi}_n^{(r)}} \left(\left(\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r) \setminus j} \rangle + \frac{1}{L} \langle \mathbf{w}_i, \mathbf{M}_j \tilde{\mathbf{z}}_{n,j}^{(r)} \rangle \right)^2 > \frac{\lambda \|\mathbf{w}_i\|_2^2 \sqrt{\log d}}{d} \right) \leq O \left(\frac{1}{\lambda} \right). \quad (121)$$

The same tail bound applies to $\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle$, $\langle \mathbf{w}_i, \frac{\mathbf{z}_{\mathbf{X}}^{(s)} - \mathbf{z}_{\mathbf{X}}^{(r)}}{2} \rangle$, and $\langle \mathbf{w}_i, \tilde{\xi}_n^{(r)} \rangle$.

(b) *High probability bound for sparse signal:*

$$\Pr \left(\langle \mathbf{w}_i, \mathbf{M} \tilde{\mathbf{z}}_n^{(r)} \rangle^2 > \|\mathbf{w}_i\|_2^2 \cdot \max_{j \in [d]} \|\mathbf{M}_j\|_\infty^2 \log^4 d \right) \lesssim e^{-\Omega(\log^2 d)}. \quad (122)$$

(c) *High probability bound for dense signal: Let $Z = \langle \mathbf{w}_i, \tilde{\xi}_n^{(r)} \rangle$. Then*

$$\Pr \left(Z^2 \geq \frac{\|\mathbf{w}_i\|_2^2 \log^4 d}{d} \right) \lesssim e^{-\Omega(\log^2 d)}. \quad (123)$$

Lemma E.8 (Pre-activation size II). *Suppose the following conditions hold:*

$$\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 \geq \Omega((\mathbf{b}_i^{(t)})^2) \quad \text{for at most } O(1) \text{ indices } j \in [d], \quad (124)$$

$$\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 \geq \Omega \left(\frac{(\mathbf{b}_i^{(t)})^2}{\sqrt{\log d}} \right) \quad \text{for at most } O(e^{-\Omega(\sqrt{\log d})} d) \text{ indices } j \in [d], \quad (125)$$

$$\|\mathbf{w}_i^{(t)}\|_2^2 \leq O \left(\frac{d(\mathbf{b}_i^{(t)})^2}{\log d} \right). \quad (126)$$

Then, for any $\lambda \geq 0.0001$,

$$\Pr \left(|\langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq \lambda \mathbf{b}_i^{(t)} \right) \lesssim e^{-\Omega(\log^{1/4} d)}, \quad (127)$$

and

$$\Pr \left(\left| \left\langle \mathbf{w}_i^{(t)}, \frac{\mathbf{z}_{\mathbf{X}}^{(r)} + \mathbf{z}_{\mathbf{X}}^{(s)}}{2} \right\rangle \right| \geq \lambda \mathbf{b}_i^{(t)} \right) \lesssim e^{-\Omega(\log^{1/4} d)}. \quad (128)$$

Lemma E.9 (Pre-activation size III). *Let $i \in [m]$. Suppose there exists a set $\mathcal{N}_i \subseteq [d]$ with $|\mathcal{N}_i| = O(1)$ such that*

$$\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 \leq O\left(\frac{(\mathbf{b}_i^{(t)})^2}{\text{polylog}(d)}\right), \quad \forall j \notin \mathcal{N}_i, \quad (129)$$

and

$$\|\mathbf{w}_i^{(t)}\|_2^2 \leq O\left(\frac{d(\mathbf{b}_i^{(t)})^2}{\text{polylog}(d)}\right). \quad (130)$$

Then, for any $\lambda \in [0.01, 0.99]$,

$$\Pr\left[\left|\sum_{j \notin \mathcal{N}_i} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \tilde{\mathbf{z}}_{n,j}^{(r)} + \langle \mathbf{w}_i, \tilde{\xi}_n^{(r)} \rangle\right| \geq \lambda \mathbf{b}_i^{(t)}\right] \lesssim e^{-\Omega(\log^2 d)}. \quad (131)$$

Lemma E.10 (Gradient for sparse features). *Suppose D.1 holds at iteration $t \geq 0$. For $j \in [d]$, we denote events*

$$\begin{aligned} A_1 &:= \left\{ S_{i,t}^{\setminus j} \geq \mathbf{b}_i^{(t)} - \alpha_{i,j}^{(t)} C_{\tilde{\mathbf{z}}} \right\}, \\ A_2 &:= \left\{ \bar{S}_{i,t}^{\setminus j} \geq \mathbf{b}_i^{(t)} - \bar{\alpha}_{i,j}^{(t)} C_{\tilde{\mathbf{z}}} \right\}, \\ A_3 &:= \left\{ \left| \bar{S}_{i,t}^{\setminus j} + \bar{\alpha}_{i,j}^{(t)} C_{\tilde{\mathbf{z}}} \right| \geq \frac{1}{2} \left(\alpha_{i,j}^{(t)} C_{\tilde{\mathbf{z}}} - \mathbf{b}_i^{(t)} \right) \right\}, \\ A_4 &:= \left\{ S_{i,t}^{\setminus j} \geq \frac{1}{2} \left(\alpha_{i,j}^{(t)} C_{\tilde{\mathbf{z}}} - \mathbf{b}_i^{(t)} \right) \right\}; \end{aligned} \quad (132)$$

and quantities L_1, L_2, L_3, L_4 as

$$\begin{aligned} L_1 &:= \sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t}^{\setminus j}|^2 (\mathbf{1}_{A_1} + \mathbf{1}_{A_2})]}{\mathbb{E}[\langle \mathbf{w}_i^{(t)}, \tilde{\xi} \rangle^2]}}, \quad L_2 := \Pr(A_1), \\ L_3 &:= \sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t}^{\setminus j}|^2 (\mathbf{1}_{A_3} + \mathbf{1}_{A_4})]}{\mathbb{E}[\langle \mathbf{w}_i^{(t)}, \tilde{\xi} \rangle^2]}}, \quad L_4 := \Pr(A_3). \end{aligned} \quad (133)$$

Then we have the following results:

(a) (all features) For all $i \in [m]$, if $\alpha_{i,j}^{(t)} \geq 0$, we have (when $\alpha_{i,j}^{(t)} \leq 0$ the opposite inequality holds)

$$\begin{aligned} &\mathbb{E}\left[h_i(\mathbf{Y}_n) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq \mathbf{b}_i} \tilde{\mathbf{z}}_{n,j}^{(r)}\right] \\ &\leq \frac{1}{L} \alpha_{i,j}^{(t)} \cdot \mathbb{E}\left[\sum_{s=1}^L \tilde{\mathbf{z}}_{n,j}^{+(s)} \sum_{r=1}^L \tilde{\mathbf{z}}_{n,j}^{(r)} \mathbf{1}_{|\langle \mathbf{w}_i^{(t)}, \frac{\mathbf{z}_X + \mathbf{z}_Y}{2} \rangle| \geq \mathbf{b}_i + |\langle \mathbf{w}_i^{(t)}, \mathbf{z}_X - \frac{\mathbf{z}_X + \mathbf{z}_Y}{2} \rangle|}\right] \\ &\pm \left(\alpha_{i,j}^{(t)} + O\left(\sqrt{\mathbb{E}[\bar{\alpha}_{i,j}^{(t)}]^2}\right)\right) \cdot \mathbb{E}\left[\sum_{s=1}^L \sum_{r=1}^L \left|\frac{\tilde{\mathbf{z}}_{n,j}^{(r)} + \tilde{\mathbf{z}}_{n,j}^{+(s)}}{2}\right| |\tilde{\mathbf{z}}_{n,j}^{(r)}|\right] \cdot O(L_1 + L_2). \end{aligned} \quad (134)$$

(b) (lucky features) If $\alpha_{i,j}^{(t)} > \mathbf{b}_i^{(t)}$, we have

$$\begin{aligned} &\mathbb{E}\left[h_i(\mathbf{Y}_n) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq \mathbf{b}_i} \tilde{\mathbf{z}}_{n,j}^{(r)}\right] \\ &\leq \frac{1}{L} \left(\alpha_{i,j}^{(t)} - \mathbf{b}_i^{(t)}\right) \cdot \mathbb{E}\left[\sum_{s=1}^L \tilde{\mathbf{z}}_{n,j}^{+(s)} \sum_{r=1}^L \tilde{\mathbf{z}}_{n,j}^{(r)} \mathbf{1}_{|\langle \mathbf{w}_i^{(t)}, \frac{\mathbf{z}_X + \mathbf{z}_Y}{2} \rangle| \geq \mathbf{b}_i + |\langle \mathbf{w}_i^{(t)}, \mathbf{z}_X - \frac{\mathbf{z}_X + \mathbf{z}_Y}{2} \rangle|}\right] \\ &\pm \left(\alpha_{i,j}^{(t)} + O\left(\sqrt{\mathbb{E}[\bar{\alpha}_{i,j}^{(t)}]^2}\right)\right) \cdot \mathbb{E}\left[\sum_{s=1}^L \sum_{r=1}^L \left|\frac{\tilde{\mathbf{z}}_{n,j}^{(r)} + \tilde{\mathbf{z}}_{n,j}^{+(s)}}{2}\right| |\tilde{\mathbf{z}}_{n,j}^{(r)}|\right] \cdot O(L_3 + L_4). \end{aligned} \quad (135)$$

If $\alpha_{i,j}^{(t)} < -\mathbf{b}_i^{(t)}$, then the opposite inequality holds with $(\alpha_{i,j}^{(t)} - \mathbf{b}_i^{(t)})$ replaced by $(\alpha_{i,j}^{(t)} + \mathbf{b}_i^{(t)})$.

Lemma E.11 (Gradient from dense signals). *Let $i \in [m]$ and $j \in [d]$. Suppose D.1 holds for the current iteration t . Then*

$$\left| \mathbb{E} \left[h_i(\mathbf{Y}_n) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq b_i^{(t)}} \langle \tilde{\xi}_n^{(r)}, \mathbf{M}_j \rangle \right] \right| \leq \tilde{O} \left(\frac{\|\mathbf{w}_i^{(t)}\|_2}{d^2} \right) \cdot \Pr(h_{i,t}(\mathbf{Y}_n) \neq 0). \quad (136)$$

For dense features \mathbf{M}_j^\perp , $j \in [d_1] \setminus [d]$, we have a similar result:

$$\left| \mathbb{E} \left[h_i(\mathbf{Y}_n) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq b_i^{(t)}} \langle \tilde{\xi}_n^{(r)}, \mathbf{M}_j^\perp \rangle \right] \right| \leq \tilde{O} \left(\frac{\|\mathbf{w}_i^{(t)}\|_2}{d\sqrt{d_1}} \right) \cdot \Pr(h_{i,t}(\mathbf{Y}_n) \neq 0). \quad (137)$$

E.3 PROOF OF THEOREM E.1

Proof of Theorem E.1: First we need to prove all the Theorem E.1 hold for $t = T_2$. Indeed, (1), (4), (5), (6), (7) is valid at T_2 from Lemma E.9. and Theorem D.1; (2) and (3) holds at T_2 obviously.

Now suppose it hold for some $t \geq T_2$, we will prove that it still hold for $t + 1$. We first deal with the case where $j \in [d]$ and $i \notin \mathcal{M}_j$, where it holds that

$$\begin{aligned} \langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle &= \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle (1 - \eta\lambda) + \eta \mathbb{E}[h_{i,t}(\mathbf{Y}_n) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq b_i} \langle \mathbf{z}_{\mathbf{X}}^{(r)}, \mathbf{M}_j \rangle] \\ &\quad - \eta \mathbb{E} \left[\sum_{\mathbf{X}_{n,s} \in \mathfrak{N}} \ell'_{s,t} \cdot h_{i,t}(\mathbf{X}_{n,s}) \sum_{r=1}^L \mathbf{1}_{|\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle| \geq b_i} \langle \mathbf{z}_{\mathbf{X}}^{(r)}, \mathbf{M}_j \rangle \right] \pm \frac{\eta}{\text{poly}(d_1)}. \end{aligned} \quad (138)$$

In this case, to calculate the expectation, we need to use Lemma E.10, Lemma E.4. First we compute the probability of events $A_1 - A_4$ by using Lemma E.7, Lemma E.8, Lemma E.9 and our Theorem E.1 to obtain

$$\Pr(A_1), \Pr(A_2) \leq \frac{1}{\text{poly}(d)^{\Omega(\log d)}}, \quad (139)$$

which implies

$$L_1, L_2 \leq \frac{1}{\text{poly}(d)^{\Omega(\log d)}}. \quad (140)$$

Furthermore, from Fact E.1, we also have

$$\mathbb{E} \left[\sum_{s=1}^L \tilde{\mathbf{z}}_{n,j}^{+(s)} \sum_{r=1}^L \tilde{\mathbf{z}}_{n,j}^{(r)} \mathbf{1}_{|\langle \mathbf{w}_i, \frac{\mathbf{z}_{\mathbf{X}}^{(r)} + \mathbf{z}_{\mathbf{Y}}^{(s)}}{2} \rangle| \geq b_i + |\langle \mathbf{w}_i, \mathbf{z}_{\mathbf{X}}^{(r)} - \frac{\mathbf{z}_{\mathbf{X}}^{(r)} + \mathbf{z}_{\mathbf{Y}}^{(s)}}{2} \rangle|} \right] \leq \epsilon_j \frac{1}{\text{poly}(d)^{\Omega(\log d)}}. \quad (141)$$

Now we further take into considerations Lemma E.11, Lemma E.4. We can obtain

$$|\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle| \leq \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle (1 - \eta\lambda) + \tilde{O} \left(\frac{\eta \Xi_2^2 \|\mathbf{w}_i^{(t)}\|_2}{d^2} \right) \pm \frac{\eta}{\text{poly}(d_1)}. \quad (142)$$

Indeed, since we have chosen learning rate $\eta = \frac{1}{\text{poly}(d)}$ and $\lambda \in [\frac{1}{d^{1.01}}, \frac{1}{d^{1.49}}]$, it is easy to prove (5) as follows:

• For $i \notin \mathcal{M}_j$, $|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| \leq O \left(\frac{\epsilon_j \|\mathbf{w}_i^{(t)}\|_2}{\epsilon_{\max} \sqrt{d \Xi_2^5}} \right)$: This is easy since by using Lemma E.10, Lemma E.4, we can prove the following inequality by contradiction

$$\begin{aligned} |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| &\leq |\langle \mathbf{w}_i^{(t-1)}, \mathbf{M}_j \rangle| (1 + \epsilon_j \frac{\eta}{d^2} - \eta\lambda) + \tilde{O} \left(\frac{\eta \Xi_2^2}{d^2} \right) \|\mathbf{w}_i^{(t)}\|_2 \\ &\leq \dots \leq O \left(\frac{\epsilon_j \|\mathbf{w}_i^{(t)}\|_2}{\epsilon_{\max} \sqrt{d \Xi_2^5}} \right). \end{aligned} \quad (143)$$

Now we begin to prove (6). For all $i \in [m]$, we have $\max_{j \in [d_1] \setminus [d]} |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j^\perp \rangle| \leq O\left(\frac{\|\mathbf{w}_i^{(t)}\|_2}{\sqrt{d_1} \Xi_2^5}\right)$ at iteration $t = T_2$. Now, by expanding the gradient updates of $\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j^\perp \rangle$, we can see that

$$\begin{aligned} |\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j^\perp \rangle| &\leq |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j^\perp \rangle| (1 - \eta\lambda) + |\Psi_{i,j}^{(t)}| + |\Phi_{i,j}^{(t)}| + |\mathcal{E}_{i,j}^{(t)}| \\ &\leq |\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j^\perp \rangle| (1 - \eta\lambda) + \tilde{O}\left(\frac{\Xi_2^5}{\tau \sqrt{d_1} d^2}\right) \|\mathbf{w}_i^{(t)}\|_2. \end{aligned} \quad (144)$$

where the last inequality are obtained as follows: From Lemma E.4 we have

$$\begin{aligned} |\mathcal{E}_{i,j}^{(t)}| &\leq O\left(\frac{\|\mathbf{w}_i^{(t)}\|_2 \Xi_2^2}{d^2 \tau}\right) \cdot \max_{i' \in [m]} (|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j^\perp \rangle|) \\ &\leq O\left(\frac{\|\mathbf{w}_i^{(t)}\|_2 \Xi_2^2}{d^2 \tau}\right) \cdot \tilde{O}\left(\frac{1}{\sqrt{d_1}}\right) \quad (\text{since } \max_{i' \in [m]} |\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j^\perp \rangle| \leq \tilde{O}\left(\frac{1}{\sqrt{d_1} \Xi_2^5}\right)) \\ &\leq \tilde{O}\left(\frac{\Xi_2^5}{\tau \sqrt{d_1} d^2}\right) \|\mathbf{w}_i^{(t)}\|_2. \end{aligned} \quad (145)$$

After (5) and (6) are proven, it is easy to observe (1) is true at t . Below we shall prove (2), (3) and (4), after which (7) can be also trivially proven.

Indeed, (2) is a corollary of (3) and (4), since if $\mathfrak{F}_j^{(t)} \leq O(\tau \log^3 d)$ and (4) holds, we simply have

$$\begin{aligned} \|\mathbf{w}_i^{(t)}\|_2^2 &= \sum_{j \in \mathcal{N}_i} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 + \sum_{j \notin \mathcal{N}_i, j \in [d]} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 + \sum_{j \in [d_1] \setminus [d]} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j^\perp \rangle^2 \\ &\leq \sum_{j \in \mathcal{N}_i} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 + O(d) \cdot O\left(\left(\frac{\epsilon_j}{\epsilon_{\max}}\right)^2 \frac{\|\mathbf{w}_i^{(t)}\|_2^2}{d \Xi_2^{10}}\right) + O(d_1) \cdot O\left(\frac{\|\mathbf{w}_i^{(t)}\|_2^2}{d_1 \Xi_2^{10}}\right) \\ &\leq \sum_{j \in \mathcal{N}_i} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 + o\left(\left(\frac{\epsilon_j}{\epsilon_{\max}}\right)^2 \frac{1}{\Xi_2^{10}} \|\mathbf{w}_i^{(t)}\|_2^2\right), \end{aligned} \quad (146)$$

which implies (2).

$$\begin{aligned} \|\mathbf{w}_i^{(t)}\|_2^2 &\leq \sum_{j \in \mathcal{N}_i} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 + o\left(\left(\frac{\epsilon_j}{\epsilon_{\max}}\right)^2 \frac{1}{\Xi_2^{10}} \|\mathbf{w}_i^{(t)}\|_2^2\right) \\ &\leq \sum_{j \in \mathcal{N}_i} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 \\ &\leq O(1) O\left(\frac{\text{poly}(\log(d))}{d^c}\right) \\ &\leq O(1). \end{aligned} \quad (147)$$

Thus we only need to prove (3) and (4). Indeed, for (3), letting $i \in \mathcal{M}_j$, we proceed as follows: we first write the updates of $\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle$ as

$$\begin{aligned} \langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle &= \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle (1 - \eta\lambda) + \Psi_{i,j}^{(t)} + \Phi_{i,j}^{(t)} + \mathcal{E}_{i,j}^{(t)} \\ &= \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle (1 - \eta\lambda) + \Psi_{i,j}^{(t)} + \tilde{O}\left(\frac{\Xi_2^2}{d^2}\right) \|\mathbf{w}_i^{(t)}\|_2. \end{aligned} \quad (148)$$

where the last inequality comes again from Lemma E.4. Now suppose for some t we have $\mathfrak{F}_j^{(t)} \geq \Omega\left(\left(\frac{\epsilon_j}{\epsilon_{\max}}\right)^2 \tau \log^3 d\right)$, by Lemma E.3, we have

$$\begin{aligned} \langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle &= \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \left(1 + \epsilon_j \frac{1}{\text{poly}(d)} - \eta\lambda\right) + \tilde{O}\left(\frac{\Xi_2^2}{d^2}\right) \|\mathbf{w}_i^{(t)}\|_2 \\ &\leq \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \left(1 + \epsilon_j \frac{1}{\text{poly}(d)} - \frac{\eta\lambda}{2}\right). \end{aligned} \quad (149)$$

which means that $\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle \leq \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle$. This in fact gives $\mathfrak{F}_j^{(t+1)} \leq \mathfrak{F}_j^{(t)}$, so that (3) is proven.

Now for (4), we need to induct as follows: for $t \leq T'_j := \frac{d \log d}{\eta \log \log d}$ which is the specific iteration when $\mathfrak{F}_j^{(t)} \geq G_2 \tau \log d$, where G_2 is defined in Lemma E.3. The induction of (4) follows from similar proof in Theorem D.1. After T'_j , we discuss as follows

- When $t \in [T'_j, \frac{d^{1.49}}{\eta}]$, from above calculations, for each $i' \in \mathcal{M}_j$, we have

$$\frac{|\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(t+1)}, \mathbf{M}_j \rangle|} = \frac{|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| (1 - \eta\lambda) + \eta \Psi_{i,j}^{(t)} \pm \mathcal{O}\left(\frac{\sqrt{\Xi_2}}{t\sqrt{d}}\right) \|\mathbf{w}_i^{(t)}\|_2}{|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle| (1 - \eta\lambda) + \eta \Psi_{i',j}^{(t)} \pm \mathcal{O}\left(\frac{\sqrt{\Xi_2}}{t\sqrt{d}}\right) \|\mathbf{w}_{i'}^{(t)}\|_2}. \quad (150)$$

On one hand, for those $i' \in \mathcal{M}_j$ such that $|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle| \leq \mathbf{b}_i^{(t)} \Xi_2^2 \leq \mathcal{O}\left(\frac{\Xi_2^2}{\sqrt{d}} \|\mathbf{w}_i^{(t)}\|_2\right)$, we can safely get $|\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle| \gg |\langle \mathbf{w}_{i'}^{(t+1)}, \mathbf{M}_j \rangle|$. On the other hand, if $|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle| \geq \mathbf{b}_i^{(t)} \Xi_2^2$, then we have

$$\left| \frac{\Psi_{i,j}^{(t)}}{\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle} - \frac{\Psi_{i',j}^{(t)}}{\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle} \right| = \frac{O(\frac{\mathbf{b}_i^{(t)}}{d^2})}{\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle} \leq O\left(\frac{1}{d^2 \Xi_2^2}\right) \leq O\left(\frac{\Xi_2}{t\sqrt{d}\eta} \mathbf{b}_i^{(t)}\right). \quad (151)$$

Thus by letting $\tilde{\Psi}_j := \frac{\Psi_{i,j}^{(t)}}{\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle}$, then

$$\frac{|\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(t+1)}, \mathbf{M}_j \rangle|} = \frac{|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| (1 + \eta \tilde{\Psi}_j^{(t)} - \eta\lambda) \pm \mathcal{O}\left(\frac{\Xi_2}{t\sqrt{d}}\right) \|\mathbf{w}_i^{(t)}\|_2}{|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle| (1 + \eta \tilde{\Psi}_j^{(t)} - \eta\lambda) \pm \mathcal{O}\left(\frac{\Xi_2}{t\sqrt{d}}\right) \|\mathbf{w}_{i'}^{(t)}\|_2}. \quad (152)$$

Since at iteration $t \in [T'_j, \frac{d^{1.49}}{\eta}]$, it is easy to obtain that $|\tilde{\Psi}_j^{(t)} - \lambda| \leq O\left(\frac{\Xi_2}{\eta t}\right)$.

Thus we have

$$\begin{aligned} & \frac{|\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(t+1)}, \mathbf{M}_j \rangle|} \\ & \geq \frac{|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| (1 + \eta(\tilde{\Psi}_j^{(t)} - \lambda)(1 - \frac{\Xi_2^2}{\sqrt{d}}))}{|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle| (1 + \eta(\tilde{\Psi}_j^{(t)} - \lambda)(1 + \frac{\Xi_2^2}{\sqrt{d}}))} \\ & \geq \left(1 + \eta(\tilde{\Psi}_j^{(t)} - \lambda)(1 - \frac{\Xi_2^2}{\sqrt{d}}) - \eta(\tilde{\Psi}_j^{(t)} - \lambda)(1 + \frac{\Xi_2^2}{\sqrt{d}})\right) \cdot \frac{|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle|} \\ & \geq \left(1 - \eta(\tilde{\Psi}_j^{(t)} - \lambda)(\frac{\Xi_2^2}{\sqrt{d}})\right) \cdot \frac{|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle|} \\ & \geq \left(1 - \frac{\Xi_2^2}{t\sqrt{d}}\right) \cdot \frac{|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle|} \\ & \geq \prod_{t'=T'_j}^{t-1} \left(1 - O\left(\frac{\Xi_2^2}{t'\sqrt{d}}\right)\right) \cdot \frac{|\langle \mathbf{w}_i^{(T'_j)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(T'_j)}, \mathbf{M}_j \rangle|} \geq \Omega(1). \end{aligned} \quad (153)$$

where in the last inequality we have used our Theorem E.1 at T'_j

- The proof for iterations $t \in \left[\frac{d^{1.49}}{\eta}, \frac{d^{1.99}}{\eta}\right]$ is largely similar to the above. The only difference here is that we rely on a slightly different comparison here: Indeed, we have

$$\frac{|\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(t+1)}, \mathbf{M}_j \rangle|} = \frac{|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| (1 + \eta \tilde{\Psi}_j^{(t)} - \eta\lambda) \pm O\left(\frac{\Xi_2^2}{d^2}\right) \|\mathbf{w}_i^{(t)}\|_2}{|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle| (1 + \eta \tilde{\Psi}_j^{(t)} - \eta\lambda) \pm O\left(\frac{\Xi_2^2}{d^2}\right) \|\mathbf{w}_{i'}^{(t)}\|_2}. \quad (154)$$

Here we can use similar techniques as above to require $|\tilde{\Psi}_j^{(t)} - \lambda| \leq O\left(\frac{\Xi_2}{t\eta}\right)$. Now we also have

$$\begin{aligned} \frac{|\langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(t+1)}, \mathbf{M}_j \rangle|} &\geq \frac{|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle| (1 + \eta(\tilde{\Psi}_j^{(t)} - \lambda)(1 - \frac{\Xi_2^2}{\sqrt{d}}))}{|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle| (1 + \eta(\tilde{\Psi}_j^{(t)} - \lambda)(1 + \frac{\Xi_2^2}{\sqrt{d}}))} \\ &\geq \left(1 - \frac{\Xi_2^2}{t\sqrt{d}}\right) \cdot \frac{|\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(t)}, \mathbf{M}_j \rangle|} \\ &\geq \prod_{t'=d^{1.49}/\eta}^{t-1} \left(1 - \frac{\Xi_2^2}{t'd^{0.01}}\right) \cdot \frac{|\langle \mathbf{w}_i^{(d^{1.49}/\eta)}, \mathbf{M}_j \rangle|}{|\langle \mathbf{w}_{i'}^{(d^{1.49}/\eta)}, \mathbf{M}_j \rangle|} \geq \Omega(1). \end{aligned} \quad (155)$$

Now (4) are proven. (7) is an immediate result of our update scheme. \square

E.4 PROOF OF THEOREM 3.1

The first part proves the convergence of the loss function. The second part is a further extension of Theorem E.1.

Proof of Theorem 3.1. We start with the proof of convergence ((11) in Theorem 3.1).

Denote $w^{(t)} = (w_1^{(t)}, \dots, w_m^{(t)})$, since our update is

$$w^{(t+1)} = w^{(t)} - \nabla_w L_{\text{aug}}(f_t) + \frac{1}{\text{poly}(d_1)}, \quad (156)$$

we have

$$\begin{aligned} &\eta \langle \nabla_w L_{\text{aug}}(f_t), w^{(t)} - \theta^* \rangle \\ &= \frac{\eta^2}{2} \|\nabla_w L_{\text{aug}}(f_t)\|_F^2 + \frac{1}{2} \|w^{(t)} - \theta^*\|_F^2 - \frac{1}{2} \|w^{(t+1)} - \theta^*\|_F^2 + \frac{\eta^2}{\text{poly}(d_1)} \\ &\leq \eta^2 \text{poly}(d) + \frac{1}{2} \|w^{(t)} - \theta^*\|_F^2 - \frac{1}{2} \|w^{(t+1)} - \theta^*\|_F^2 + \frac{\eta^2}{\text{poly}(d_1)}, \end{aligned} \quad (157)$$

where the inequality comes from

$$\|\nabla_w L_{\text{aug}}(f_t)\|_F^2 = \sum_{i=1}^m \|\nabla_{\mathbf{w}_i} L_{\text{aug}}(f_t)\|^2. \quad (158)$$

Each term is $O(1)$, and since $m = \text{poly}(d)$, the overall complexity is $\text{poly}(d)$.

Now we will use the tools from online learning to obtain a loss guarantee: define a pseudo objective for parameter θ

$$\begin{aligned} \tilde{L}_{\text{aug}_t}(\theta) &:= \tilde{L}(f_t, \theta, f_t) + \frac{\lambda}{2} \sum_{i \in [m]} \|\theta_i\|_2^2 \\ &= \mathbb{E} \left[-\tau \log \left(\frac{e^{\langle f_t, \theta(\mathbf{X}_n), f_t(\mathbf{Y}_n) \rangle / \tau}}{\sum_{\mathbf{X} \in \mathfrak{B}} e^{\langle f_t, \theta(\mathbf{X}_n), f_t(\mathbf{X}) \rangle / \tau}} \right) \right] + \frac{\lambda}{2} \sum_{i \in [m]} \|\theta_i\|_2^2. \end{aligned} \quad (159)$$

Which is a convex function over θ since it is linear in θ (for a fixed f_t , we can consider $\tilde{L}(f_t, \theta, f_t)$ to be convex with respect to θ , because $f_{t,\theta}(x)$ is linear, and softmax + log is a convex composition; the regularization term is convex).

Moreover, we have

$$\tilde{L}_{\text{aug}_t}(w^{(t)}) = L_{\text{aug}}(f_t), \quad (160)$$

and

$$\nabla_{\theta_i} \tilde{L}_{\text{aug}_t}(\mathbf{w}_i^{(t)}) = \nabla_{\mathbf{w}_i} L_{\text{aug}}(f_t). \quad (161)$$

Thus we have

$$\begin{aligned} & \eta \langle \nabla_w L_{\text{aug}}(f_t), w^{(t)} - \theta^* \rangle \\ &= \eta \langle \nabla_{\theta} \tilde{L}_{\text{aug}_t}(w^{(t)}), w^{(t)} - \theta^* \rangle \\ &\stackrel{\textcircled{1}}{\geq} \tilde{L}_{\text{aug}_t}(w^{(t)}) - \tilde{L}_{\text{aug}_t}(\theta^*) \\ &\geq \tilde{L}_{\text{aug}_t}(w^{(t)}) - \mathbb{E} \left[-\tau \log \left(\frac{e^{\langle f_t, \theta^*(\mathbf{X}_n), f_t(\mathbf{Y}_n) \rangle / \tau}}{\sum_{\mathbf{X} \in \mathfrak{B}} e^{\langle f_t, \theta^*(\mathbf{X}_n), f_t(\mathbf{X}) \rangle / \tau}} \right) \right] - \frac{\lambda}{2} \sum_{i \in [m]} \|\theta_i^*\|_2^2 \\ &\stackrel{\textcircled{2}}{\geq} \tilde{L}_{\text{aug}_t}(w^{(t)}) - O\left(\frac{1}{\log d}\right) - \sum_{i \in [m]} O(\lambda \|\theta_i^*\|_2^2) \\ &\geq L_{\text{aug}}(f_t) - O\left(\frac{1}{\log d}\right). \end{aligned} \quad (162)$$

① is because the surrogate objective function \tilde{L}_{aug_t} is a convex function with respect to θ , so we can use a first-order convex lower bound: $f(\theta) - f(\theta') \leq \langle \nabla f(\theta), \theta - \theta' \rangle$. ② is because $\sum_{i \in [m]} \lambda \|\theta_i^*\|_2^2 = \sum_{j \in [d]} \sum_{i \in \mathcal{M}_j^*} \lambda \|\theta_i^*\|_2^2 = \sum_{j \in [d]} \sum_{i \in \mathcal{M}_j^*} \lambda \frac{\tau \kappa^2}{|\mathcal{M}_j^*|^2} = \sum_{j \in [d]} \lambda \frac{\tau \kappa^2}{|\mathcal{M}_j^*|} = \frac{\lambda \tau \kappa^2}{|\mathcal{M}_j^*|}$

Now choosing $\kappa = \Theta(\Xi_2) \leq \frac{1}{\lambda d}$ (so that $\sum_{i \in [m]} \lambda \|\theta_i^*\|_2^2 < \frac{1}{\log d}$), and by a telescoping summation, we have

$$\begin{aligned} \frac{1}{T} \sum_{t=T_3}^{T_3+T-1} \left(L_{\text{aug}}(f_t) - O\left(\frac{1}{\log d}\right) \right) &\leq \frac{1}{T} \sum_{t=T_3}^{T_3+T-1} \eta \langle \nabla_w L_{\text{aug}}(f_t), w^{(t)} - \theta^* \rangle \\ &\leq \frac{O(\|w^{(T_3)} - \theta^*\|_F^2)}{T\eta} \\ &= \frac{O(\|w^{(T_3)}\|_F^2 + \|\theta^*\|_F^2 - 2 \text{Tr}((w^{(T_3)})^\top \theta^*))}{T\eta} \\ &\leq \frac{O(\|w^{(T_3)}\|_F^2 + \|\theta^*\|_F^2)}{T\eta} \\ &\leq \frac{O(m\|\mathbf{w}_i^{(T_3)}\|_2^2)}{T\eta} \\ &\leq O\left(\frac{m\Xi_2}{T\eta}\right). \end{aligned} \quad (163)$$

Since $T\eta \geq m\Xi_2^{10}$, this proves the claim.

For (12) in Theorem 3.1, we have

$$\begin{aligned}
\mathbf{w}_i^{(t)} &= \sum_{j \in \mathcal{N}_i, j \in [d]} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \mathbf{M}_j + \sum_{j \notin \mathcal{N}_i, j \in [d]} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \mathbf{M}_j + \sum_{j \in [d_1] \setminus [d]} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j^\perp \rangle \mathbf{M}_j^\perp \\
&\leq \sum_{j \in \mathcal{N}_i, j \in [d]} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \mathbf{M}_j + \sum_{j \notin \mathcal{N}_i, j \in [d]} O\left(\frac{\epsilon_j}{\epsilon_{\max}} \frac{\|\mathbf{w}_i^{(t)}\|_2}{\sqrt{d} \Xi_2^5}\right) \mathbf{M}_j + \sum_{j \in [d_1] \setminus [d]} O\left(\frac{\|\mathbf{w}_i^{(t)}\|_2}{\sqrt{d_1} \Xi_2^5}\right) \mathbf{M}_j^\perp \\
&= \sum_{j \in \mathcal{N}_i, j \in [d]} \alpha_{i,j} \mathbf{M}_j + \sum_{j \notin \mathcal{N}_i, j \in [d]} \alpha'_{i,j} \mathbf{M}_j + \sum_{j \in [d_1] \setminus [d]} \beta_{i,j} \mathbf{M}_j^\perp.
\end{aligned} \tag{164}$$

From Lemma B.2(c), we know that for each $j \in [d]$, there is at least one neuron that can fully learn the feature \mathbf{M}_j , and at most Ξ_2 neurons can learn the feature \mathbf{M}_j . Combining this with Theorem E.1(c):

$$\sum_{i \in \mathcal{M}_j} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle^2 = \Theta\left(\left(\frac{\epsilon_j}{\epsilon_{\max}}\right)^2 \tau \log^3 d\right), \tag{165}$$

we can conclude that the range of $\langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle$ is $[\frac{\epsilon_j}{\epsilon_{\max}} \frac{\tau}{\Xi_2}, \frac{\epsilon_j}{\epsilon_{\max}} \tau]$, and hence the range of $\alpha_{i,j}$ is $[\frac{\epsilon_j}{\epsilon_{\max}} \frac{\tau}{\Xi_2}, \frac{\epsilon_j}{\epsilon_{\max}} \tau]$. Furthermore, from Theorem E.1 (e) and (f), we can obtain that $\alpha'_{i,j} \leq o(\frac{\epsilon_j}{\epsilon_{\max}} \frac{1}{\sqrt{d}})$ and $\beta_{i,j} \leq o(\frac{1}{\sqrt{d_1}})$ respectively.

Next, we compute the upper bound of $|\mathcal{N}_i|$. As a first step, we calculate the expectation of $|\mathcal{N}_i|$.

$$\begin{aligned}
\mathbb{E}[|\mathcal{N}_i|] &= \frac{1}{m} \sum_{i=1}^m |\mathcal{N}_i| = \frac{1}{m} \sum_{j=1}^d |\mathcal{M}_j| \leq \frac{1}{m} \cdot d \cdot O(d^{\omega_2}) \\
&= \frac{1}{m} \cdot O(d^{1+\omega_2}) = \frac{O(d^{1+\omega_2})}{d^{C_m}} = O(d^{1+\omega_2-C_m}) \\
&= O\left(d^{1-(\frac{\epsilon_{\min}}{\epsilon_{\max}})^2 \cdot (1-\gamma)}\right).
\end{aligned} \tag{166}$$

Fix a neuron i , we have: $\mu_i := \mathbb{E}[|\mathcal{N}_i|]$. By Bernstein's inequality,

$$\Pr[|\mathcal{N}_i| \geq \mu_i + t] \leq \exp\left(-\frac{t^2}{2(\mu_i + t/3)}\right), \quad t \geq 0. \tag{167}$$

We set $t = 3(\sqrt{\mu_i L} + L)$ and plug this into the inequality above. Then we obtain

$$\Pr\left[|\mathcal{N}_i| \geq \mu_i + 3(\sqrt{\mu_i L} + L)\right] \leq e^{-L}. \tag{168}$$

Hence, for any constant $c > 0$, taking $L = c \log d$ yields

$$\Pr\left[|\mathcal{N}_i| \leq \mu_i + 3(\sqrt{\mu_i c \log d} + c \log d)\right] \geq 1 - d^{-c}. \tag{169}$$

Next, we apply the union bound. For the event

$$A_i := \left\{ |\mathcal{N}_i| \leq \mu_i + 3(\sqrt{\mu_i L} + L) \right\}, \tag{170}$$

the union bound gives

$$\Pr\left[\bigcap_{i=1}^m A_i\right] \geq 1 - \sum_{i=1}^m \Pr(A_i^c) \geq 1 - m e^{-L}. \tag{171}$$

Taking $L = c \log(md)$, we obtain

$$\Pr\left[\forall i \in [m], |\mathcal{N}_i| \leq \mu_i + 3(\sqrt{\mu_i c \log(md)} + c \log(md))\right] \geq 1 - (md)^{-c}. \tag{172}$$

We know $\mu_i \gg \log(md)$, so we have

$$\begin{aligned} |\mathcal{N}_i| &= \mu_i \left(1 \pm O\left(\sqrt{\frac{\log(md)}{\mu_i}}\right) \right) \\ &= \mu_i \left(1 \pm o(1) \right) \\ &\leq O\left(d^{1-\left(\frac{\epsilon_{\min}}{\epsilon_{\max}}\right)^2 \cdot (1-\gamma)}\right) \quad \text{with probability at least } 1 - (md)^{-c} \end{aligned} \quad (173)$$

Finally, for each dictionary atom \mathbf{M}_j , there are at least $\Omega(d^{\omega_1})$ neurons $i \in [m]$ such that $\mathcal{N}_i = \{j\}$. From Lemma B.2 (c), we recall that $|\mathcal{M}_j^*| \geq \Omega(d^{\omega_1})$. Moreover, if a neuron belongs to \mathcal{M}_j^* , then it cannot belong to $\mathcal{M}_{j'}$.

For (12) in Theorem 3.1, our proof is complete. \square

F THEOREM F.1

From Lemma B.2(c), we know that for each $j \in [d]$, there is at least one neuron that can fully learn the minority feature \mathbf{M}_{j^*} . When we prune out the lucky neurons that learn these minority features during the forward pass, the network will force the lucky neurons to further strengthen their feature learning ability on the minority features during the backward pass.

After magnitude pruning, neurons encoding a specific minority feature are removed. Pruning these lucky neurons reduces $\text{sim}_{f_\theta}(\mathbf{X}_n, \mathbf{Y}_n)$ during the forward pass. The decrease in similarity reduces the positive logit $\ell'_{p, \theta_{\text{mask}}^{(t)}}$, which in turn increases the gradient of the loss function, thereby encouraging these lucky neurons to further enhance their learning ability on the minority features.

Fix one specific minority feature \mathbf{M}_{j^*} , and let $\mathcal{M}_{j^*}^* \subseteq [m]$ denote the subset of neurons primarily aligned with it, with $|\mathcal{M}_{j^*}^*| = n$. For a pruning rate $\alpha \in [1/m, n/m]$, the number of pruned neurons is $\alpha m \leq n$. Let $\mathcal{P} \subseteq \mathcal{M}_{j^*}^*$ be the pruned set with $|\mathcal{P}| = \alpha m$.

F.1 THEOREM F.1

Theorem F.1 (Feature Dynamics After Pruning). *Starting from the pruning stage T_4 with pruning ratio α , the following statements hold.*

(a) *When $i^* \in \mathcal{M}_{j^*}^*$, we have*

$$\langle \mathbf{w}_{i^*}^{(t+1)}, \mathbf{M}_{j^*} \rangle \geq \left(1 - \eta\lambda + \eta\epsilon_{j^*} \frac{C_z \log \log d}{d} \left(\Theta\left(\frac{1}{\text{polylog}(d)}\right) + \Omega\left(\frac{\alpha m \epsilon_{j^*} \log \log d}{d \Xi_2^2}\right) \right) \right) \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle. \quad (174)$$

(b) *When $i \notin \mathcal{M}_{j^*}^*$ and $j \neq j^*$, we have*

$$\begin{aligned} \langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle &\leq \left(1 - \eta\lambda + \eta\epsilon_j \frac{C_z \log \log d}{d} \left(\Theta\left(\frac{1}{\text{polylog}(d)}\right) \right. \right. \\ &\quad \left. \left. + \epsilon_{j^*} \frac{\log \log d}{d} \left(\Theta\left(\frac{1}{\text{polylog}(d)}\right) + O\left(\frac{\alpha m \epsilon_{j^*} \log \log d}{d \Xi_2^2}\right) \right) \right) \right) \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle. \end{aligned} \quad (175)$$

(c) *For each neuron $i \in \mathcal{P}$ and $t \in [T_4, T_5]$, contrastive learning learns the following decomposition:*

$$\mathbf{w}_i^{(t)} = \alpha_{i,j^*} \mathbf{M}_{j^*} + \sum_{j \notin \mathcal{N}_i} \alpha'_{i,j} \mathbf{M}_j + \sum_{j \in [d_1] \setminus [d]} \beta_{i,j} \mathbf{M}_j^\perp, \quad (176)$$

where

$$\alpha_{i,j^*} \in \left[\frac{\tau}{\Xi_2}, \tau \right], \quad \alpha'_{i,j} \leq o\left(\left(1 + \frac{1}{d}\right) \frac{1}{\sqrt{d}}\right) \|\mathbf{w}_i^{(t)}\|_2, \quad |\beta_{i,j}| \leq o\left(\frac{1}{\sqrt{d_1}}\right) \|\mathbf{w}_i^{(t)}\|_2. \quad (177)$$

F.2 USEFUL LEMMAS

Lemma F.1 (Expected values of neuron activations after T_4). *From T_4 onward, the following results hold:*

(a) *For positive pair,*

$$\mathbb{E} \left[\sum_{i \in \mathcal{P}} h_i(\mathbf{X}_n) h_i(\mathbf{Y}_n) \right] \geq \Omega \left(\alpha m \frac{\tau^2}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d} \right). \quad (178)$$

(b) *For negative pair,*

$$\mathbb{E} \left[\sum_{i \in \mathcal{P}} h_i(\mathbf{X}_n) h_i(\mathbf{X}_{n,s}) \right] = 0. \quad (179)$$

(c) *For negative pair,*

$$\mathbb{E} \left[h_{i,t}(\mathbf{X}_{n,s}) \langle \nabla_{\mathbf{w}_i} h_i(\mathbf{X}_n), \mathbf{M}_{j^*} \rangle \right] = 0. \quad (180)$$

Lemma F.2 (Effect of Pruning on Positive Logit Weight). *At the pruning stage, for the data following distribution D_1 , the post-pruning positive logit $\ell'_{p, \theta_{\text{mask}}^{(t)}}$ satisfies*

$$\mathbb{E} \left[1 - \ell'_{p, \theta_{\text{mask}}^{(t)}} \right] \geq \Theta \left(\frac{1}{\tau} \right) + \Omega \left(\frac{\alpha m}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d} \right). \quad (181)$$

Lemma F.3 (Positive gradient). *Let $h_{i,t}(\cdot)$ denote the i -th neuron at iteration $t \leq T_1$ (so that $\mathbf{b}_i^{(t)} = 0$). Then the following hold:*

(a) *For each $j \in [d]$,*

$$\mathbb{E} [h_{i,t}(\mathbf{Y}_n) \langle \nabla_{\mathbf{w}_i} h_{i,t}(\mathbf{X}_n), \mathbf{M}_j \rangle] = \frac{1}{L^2} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \mathbb{E} [\hat{\mathbf{z}}_{n,j}^+ \hat{\mathbf{z}}_{n,j}]. \quad (182)$$

(b) *For each $j \in [d_1] \setminus [d]$,*

$$\mathbb{E} [h_{i,t}(\mathbf{Y}_n) \langle \nabla_{\mathbf{w}_i} h_{i,t}(\mathbf{X}_n), \mathbf{M}_j^\perp \rangle] = 0. \quad (183)$$

F.3 PROOF OF THEOREM F.1

Overview of the proof: first, the data can be divided into two parts: the samples that contain \mathbf{M}_{j^*} and those that do not. The former follow distribution D_1 , while the latter follow distribution D_2 . Next, let us examine $\ell'_{p, \theta_{\text{mask}}^{(t)}}$. The values of $\ell'_{p, \theta_{\text{mask}}^{(t)}}$ differ depending on the distribution: for samples from D_1 , we have $\ell'_{p, \theta_{\text{mask}}^{(t)}} = 1 - \Theta(\frac{1}{\tau}) - \Omega(\frac{\alpha m}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d})$, whereas for samples from D_2 , $\ell'_{p, \theta_{\text{mask}}^{(t)}} = 1 - \Theta(\frac{1}{\tau})$. Since the latter do not contain \mathbf{M}_{j^*} , pruning does not affect them.

Proof of Theorem F.1. For any neuron $i^* \in \mathcal{P}$ we have

$$\begin{aligned} & \langle \mathbf{w}_{i^*}^{(t+1)}, \mathbf{M}_{j^*} \rangle \\ &= \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle - \eta \langle \nabla_{\mathbf{w}_{i^*}} L_{\text{aug}}(f_t), \mathbf{M}_{j^*} \rangle \pm \frac{\|\mathbf{w}_{i^*}^{(t)}\|_2}{\text{poly}(d)} \\ &= (1 - \eta\lambda) \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle \\ &+ \eta \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n} \left[\left(1 - \ell'_{p, \theta_{\text{mask}}^{(t)}}(\mathbf{X}_n, \mathfrak{B}) \right) \cdot h_{i^*,t}(\mathbf{Y}_n) \langle \nabla_{\mathbf{w}_{i^*}} h_{i^*}(\mathbf{X}_n), \mathbf{M}_{j^*} \rangle \right] \\ &- \eta \sum_{\mathbf{X}_{n,s} \in \mathfrak{N}} \mathbb{E} \left[l'_{s,t}(\mathbf{X}_n, \mathfrak{B}) h_{i^*,t}(\mathbf{X}_{n,s}) \langle \nabla_{\mathbf{w}_{i^*}} h_{i^*}(\mathbf{X}_n), \mathbf{M}_{j^*} \rangle \right] \pm \frac{\|\mathbf{w}_{i^*}^{(t)}\|_2}{\text{poly}(d)} \end{aligned} \quad (184)$$

At stage T_4 , pruning is applied. We regard $\ell'_{p, \theta_{\text{mask}}^{(t)}}$ and $\ell'_{s, \theta_{\text{mask}}^{(t)}}$ as fixed, and by combining Lemma F.1(c) with the law of total probability, we obtain

$$\begin{aligned}
& \langle \mathbf{w}_{i^*}^{(t+1)}, \mathbf{M}_{j^*} \rangle \\
&= (1 - \eta\lambda) \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle \\
&+ \eta \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n} \left[\left(1 - \ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}\right) \right] \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n} \left[h_{i^*, t}(\mathbf{Y}_n) \langle \nabla_{\mathbf{w}_{i^*}} h_{i^*}(\mathbf{X}_n), \mathbf{M}_{j^*} \rangle \right] \\
&= (1 - \eta\lambda) \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle \\
&+ \eta \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_1} \left[\left(1 - \ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}\right) \right] \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_1} \left[h_{i^*, t}(\mathbf{Y}_n) \langle \nabla_{\mathbf{w}_{i^*}} h_{i^*}(\mathbf{X}_n), \mathbf{M}_{j^*} \rangle \right] \cdot \mathbb{P}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_1} \\
&+ \eta \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_2} \left[\left(1 - \ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}\right) \right] \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_2} \left[h_{i^*, t}(\mathbf{Y}_n) \langle \nabla_{\mathbf{w}_{i^*}} h_{i^*}(\mathbf{X}_n), \mathbf{M}_{j^*} \rangle \right] \cdot \mathbb{P}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_2}
\end{aligned} \tag{185}$$

Combining Lemma F.3(a) with (181) in Lemma F.2, we obtain

$$\begin{aligned}
& \langle \mathbf{w}_{i^*}^{(t+1)}, \mathbf{M}_{j^*} \rangle \\
&= (1 - \eta\lambda) \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle \\
&+ \eta \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_1} \left[\left(1 - \ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}\right) \right] \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_1} \left[\frac{1}{L^2} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \mathbb{E}[\hat{\mathbf{z}}_{n,j}^+ \hat{\mathbf{z}}_{n,j^*}] \right] \cdot \mathbb{P}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_1} \\
&+ \eta \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_2} \left[\left(1 - \ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}\right) \right] \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_2} \left[\frac{1}{L^2} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \mathbb{E}[\hat{\mathbf{z}}_{n,j}^+ \hat{\mathbf{z}}_{n,j^*}] \right] \cdot \mathbb{P}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_2} \\
&= (1 - \eta\lambda) \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle \\
&+ \eta \left(\Theta\left(\frac{1}{\tau}\right) + \Omega\left(\frac{\alpha m}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d}\right) \right) \cdot \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \cdot \epsilon_{j^*} \frac{\log \log d}{d} \\
&+ \eta \cdot \Theta\left(\frac{1}{\tau}\right) \cdot 0 \cdot 1 \cdot \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle \\
&= (1 - \eta\lambda) \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle + \eta \left(\Theta\left(\frac{1}{\tau}\right) + \Omega\left(\frac{\alpha m}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d}\right) \right) \cdot \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle \cdot \epsilon_{j^*} \frac{\log \log d}{d}
\end{aligned} \tag{186}$$

Hence, the post-pruning one-step update along \mathbf{M}_{j^*} is

$$\langle \mathbf{w}_{i^*}^{(t+1)}, \mathbf{M}_{j^*} \rangle \geq \left(1 - \eta\lambda + \eta \epsilon_{j^*} \frac{C_z \log \log d}{d} \left(\Theta\left(\frac{1}{\text{polylog}(d)}\right) + \Omega\left(\frac{\alpha m \epsilon_{j^*} \log \log d}{d \Xi_2^2}\right) \right) \right) \langle \mathbf{w}_{i^*}^{(t)}, \mathbf{M}_{j^*} \rangle. \tag{187}$$

Similarly to (186), for any neuron $i \notin \mathcal{P}$, we have:

$$\begin{aligned}
& \langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle \\
&= (1 - \eta\lambda) \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \\
&+ \eta \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_1} \left[\left(1 - \ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}\right) \right] \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_1} \left[\frac{1}{L^2} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \mathbb{E}[\hat{\mathbf{z}}_{n,j}^+ \hat{\mathbf{z}}_{n,j}] \right] \cdot \mathbb{P}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_1} \\
&+ \eta \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_2} \left[\left(1 - \ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}\right) \right] \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_2} \left[\frac{1}{L^2} \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \mathbb{E}[\hat{\mathbf{z}}_{n,j}^+ \hat{\mathbf{z}}_{n,j}] \right] \cdot \mathbb{P}_{\mathbf{X}_n, \mathbf{Y}_n \sim D_2} \\
&= (1 - \eta\lambda) \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \\
&+ \eta \left(\Theta\left(\frac{1}{\tau}\right) + \Omega\left(\frac{\alpha m}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d}\right) \right) \cdot \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle \cdot \epsilon_j \frac{\log \log d}{d} \epsilon_{j^*} \frac{\log \log d}{d} \\
&+ \eta \cdot \Theta\left(\frac{1}{\tau}\right) \cdot \epsilon_j \frac{\log \log d}{d} \cdot 1 \cdot \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle
\end{aligned} \tag{188}$$

Hence, the post-pruning one-step update along M_j is

$$\begin{aligned} \langle \mathbf{w}_i^{(t+1)}, \mathbf{M}_j \rangle &\leq \left(1 - \eta\lambda + \eta\epsilon_j \frac{C_z \log \log d}{d} \left(\Theta\left(\frac{1}{\text{polylog}(d)}\right) \right. \right. \\ &\quad \left. \left. + \epsilon_{j^*} \frac{\log \log d}{d} \left(\Theta\left(\frac{1}{\text{polylog}(d)}\right) + O\left(\frac{\alpha m \epsilon_{j^*} \log \log d}{d \Xi_2^2}\right) \right) \right) \right) \langle \mathbf{w}_i^{(t)}, \mathbf{M}_j \rangle. \end{aligned} \quad (189)$$

The above constitutes the proof of Theorem F.1 regarding pruning. \square

F.4 PROOF OF THEOREM 3.2

Theorem 3.2 (a) and (b) can be derived as simplifications of Theorem F.1 (a) and (b). Theorem 3.2 (c) coincides with Theorem F.1 (c). By taking the elapsed time $T = \frac{((\epsilon_{\max}/\epsilon_{j^*})-1)d}{\eta\alpha\epsilon_{j^*}^2 C_z \log \log d}$ and simplifying (a) and (b), then substituting into the conclusion of Theorem 3.1, the proof follows.

F.5 PROOF OF LEMMA F.1:

Proof of Lemma F.1: The alignment with the target minority feature \mathbf{M}_{j^*} is $\langle \mathbf{w}_i, \mathbf{M}_{j^*} \rangle$, and we have $|\langle \mathbf{w}_i, \mathbf{M}_{j^*} \rangle| \geq \Omega(\frac{\tau}{\Xi_2})$ at T_4 (This is the conclusion of Theorem 3.1, which can be found in the second part of the proof of Theorem 3.1. For the positive pair $(\mathbf{X}_n, \mathbf{Y}_n)$, the latent variables z_{n,j^*} and z_{n,j^*}^+ are correlated through the augmentation process. For a negative sample $\mathbf{X}_{n,s}$, its latent variable z_{n,s,j^*} is independent of those of the positive pair (z_{n,j^*}, z_{n,j^*}^+) , so we have:

$$(z_{n,j^*}, z_{n,j^*}^+) \perp\!\!\!\perp z_{n,s,j^*}. \quad (190)$$

For the anchor \mathbf{X}_n and its positive \mathbf{Y}_n , we have

$$h_i(\mathbf{X}_n) = \sum_{r=1}^L \langle \mathbf{w}_i, z_Y^{(r)} \rangle = \frac{1}{L} \langle \mathbf{w}_i, \mathbf{M} \sum_{r=1}^L \tilde{z}_n^{(r)} + \sum_{r=1}^L \tilde{\xi}_n^{(r)} \rangle, \quad (191)$$

$$h_i(\mathbf{Y}_n) = \sum_{s=1}^L \langle \mathbf{w}_i, z_Y^{(s)} \rangle = \frac{1}{L} \langle \mathbf{w}_i, \mathbf{M} \sum_{s=1}^L \tilde{z}_n^{+(s)} + \sum_{s=1}^L \tilde{\xi}_n^{+(s)} \rangle, \quad (192)$$

$$\hat{z}_n := \sum_{r=1}^L \tilde{z}_n^{(r)}, \quad \hat{z}_n^+ := \sum_{s=1}^L \tilde{z}_n^{+(s)}, \quad \hat{\xi}_n := \sum_{r=1}^L \tilde{\xi}_n^{(r)}, \quad \hat{\xi}_n^+ := \sum_{s=1}^L \tilde{\xi}_n^{+(s)}. \quad (193)$$

We can write the outputs as:

$$h_i(\mathbf{X}_n) = \frac{1}{L} \langle \mathbf{w}_i, \mathbf{M} \hat{z}_n + \hat{\xi}_n \rangle, \quad h_i(\mathbf{Y}_n) = \frac{1}{L} \langle \mathbf{w}_i, \mathbf{M} \hat{z}_n^+ + \hat{\xi}_n^+ \rangle. \quad (194)$$

For a negative sample $\mathbf{X}_{n,s}$: $\hat{z}_{n,s} := \sum_{q=1}^L \tilde{z}_{n,s}^{(q)}$, $\hat{\xi}_{n,s} := \sum_{q=1}^L \tilde{\xi}_{n,s}^{(q)}$, the output is:

$$h_i(\mathbf{X}_{n,s}) = \frac{1}{L} \langle \mathbf{w}_i, \mathbf{M} \hat{z}_{n,s} + \hat{\xi}_{n,s} \rangle. \quad (195)$$

We first establish a lower bound for $\mathbb{E}[h_i(\mathbf{X}_n)h_i(\mathbf{Y}_n)]$.

Expanding and using zero-mean and independence of latent variables and noises, we have

$$\begin{aligned} \mathbb{E}[h_i(\mathbf{X}_n)h_i(\mathbf{Y}_n)] &= \frac{1}{L^2} \mathbb{E}[\langle \mathbf{w}_i, \mathbf{M} \hat{z}_n \rangle \langle \mathbf{w}_i, \mathbf{M} \hat{z}_n^+ \rangle] \\ &= \frac{1}{L^2} \sum_{j=1}^d \langle \mathbf{w}_i, \mathbf{M}_j \rangle^2 \mathbb{E}[\hat{z}_{n,j} \hat{z}_{n,j}^+] \\ &\geq \frac{1}{L^2} \langle \mathbf{w}_i, \mathbf{M}_{j^*} \rangle^2 \mathbb{E}[\hat{z}_{n,j^*} \hat{z}_{n,j^*}^+] \\ &\geq \Omega\left(\frac{\tau^2}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d}\right). \end{aligned} \quad (196)$$

Therefore

$$\mathbb{E}[h_i(\mathbf{X}_n)h_i(\mathbf{Y}_n)] \geq \Omega\left(\frac{\tau^2}{\Xi_2^2}\epsilon_{j^*}\frac{\log \log d}{d}\right). \quad (197)$$

Next, we compute the expectation of $h_i(\mathbf{X}_n), h_i(\mathbf{X}_{n,s}).h_i(\mathbf{X}_n) h_i(\mathbf{X}_{n,s})$,

$$\mathbb{E}[h_i(\mathbf{X}_n)h_i(\mathbf{X}_{n,s})] = \mathbb{E}\left[\left(\frac{1}{L}\langle \mathbf{w}_i, \mathbf{M}\hat{\mathbf{z}}_n + \hat{\xi}_n \rangle\right)\left(\frac{1}{L}\langle \mathbf{w}_i, \mathbf{M}\hat{\mathbf{z}}_{n,s} + \hat{\xi}_{n,s} \rangle\right)\right]. \quad (198)$$

By the assumption, the latent variables of \mathbf{X}_n are independent of those of the negative $\mathbf{X}_{n,s}$, and all noises are mean-zero and independent. Therefore,

$$\mathbb{E}[\langle \mathbf{w}_i, \mathbf{M}\hat{\mathbf{z}}_n \rangle \langle \mathbf{w}_i, \mathbf{M}\hat{\mathbf{z}}_{n,s} \rangle] = 0, \quad \mathbb{E}[\langle \mathbf{w}_i, \hat{\xi}_n \rangle \langle \mathbf{w}_i, \hat{\xi}_{n,s} \rangle] = 0 \quad (199)$$

Therefore, we conclude that

$$\mathbb{E}[h_i(\mathbf{X}_n)h_i(\mathbf{X}_{n,s})] = 0 \quad (200)$$

Let \mathcal{P} be the pruned set with $|\mathcal{P}| = \alpha m$. Summing the per-neuron bounds over $i \in \mathcal{P}$, we obtain

$$\mathbb{E}\left[\sum_{i \in \mathcal{P}} h_i(\mathbf{X}_n)h_i(\mathbf{Y}_n)\right] \geq \Omega(\alpha m \frac{\tau^2}{\Xi_2^2} p_{j^*}), \quad (201)$$

$$\mathbb{E}\left[\sum_{i \in \mathcal{P}} h_i(\mathbf{X}_n)h_i(\mathbf{X}_{n,s})\right] = 0. \quad (202)$$

This completes the proof of Lemma F.1 (a)(b).

Finally, we compute the expectation of $h_{i,t}(\mathbf{X}_{n,s}), \langle \nabla_{\mathbf{w}_i} h_i(\mathbf{X}_n), \mathbf{M}_{j^*} \rangle$, and we have

$$\begin{aligned} & \mathbb{E}\left[h_{i,t}(\mathbf{X}_{n,s}) \langle \nabla_{\mathbf{w}_i} h_i(\mathbf{X}_n), \mathbf{M}_{j^*} \rangle\right] \\ &= \mathbb{E}\left[\left(\frac{1}{L}\langle \mathbf{w}_i, \mathbf{M}\hat{\mathbf{z}}_{n,s} + \hat{\xi}_{n,s} \rangle\right)\left(\frac{1}{L}\langle \mathbf{M}\hat{\mathbf{z}}_n + \hat{\xi}_n, \mathbf{M}_{j^*} \rangle\right)\right] \\ &= \mathbb{E}\left[\left(\frac{1}{L^2}\langle \mathbf{w}_i, \mathbf{M}\hat{\mathbf{z}}_{n,s} + \hat{\xi}_{n,s} \rangle\right)\left(\hat{\mathbf{z}}_{n,j^*} + \langle \hat{\xi}_n, \mathbf{M}_{j^*} \rangle\right)\right] \\ &= \frac{1}{L^2} \mathbb{E}[\langle \mathbf{w}_i, \mathbf{M}\hat{\mathbf{z}}_{n,s} \rangle \cdot \hat{\mathbf{z}}_{n,j^*}] \\ &= 0. \end{aligned} \quad (203)$$

This completes the proof of Lemma F.1(c),

$$\mathbb{E}\left[h_{i,t}(\mathbf{X}_{n,s}) \langle \nabla_{\mathbf{w}_i} h_i(\mathbf{X}_n), \mathbf{M}_{j^*} \rangle\right] = 0. \quad (204)$$

□

F.6 PROOF OF LEMMA F.2:

Proof of Lemma F.2: We link the logit to the pruning ratio and plug it into the gradient growth. Recall the softmax weights and partial derivatives

$$\ell'_p = \frac{e^{u_p/\tau}}{e^{u_p/\tau} + \sum_{s=1}^S e^{u_s/\tau}}, \quad \ell'_s = \frac{e^{u_s/\tau}}{e^{u_p/\tau} + \sum_{s=1}^S e^{u_s/\tau}}, \quad \sum_{s=1}^S \ell'_s = 1 - \ell'_p, \quad (205)$$

$$u_p = \text{Sim}_f(\mathbf{X}_n, \mathbf{Y}_n), \quad u_s = \text{Sim}_f(\mathbf{X}_n, \mathbf{X}_{n,s}), \quad (206)$$

$$\frac{\partial \ell'_p}{\partial u_p} = \frac{1}{\tau} \ell'_p (1 - \ell'_p), \quad \frac{\partial \ell'_p}{\partial u_s} = -\frac{1}{\tau} \ell'_p \ell'_s. \quad (207)$$

Pruning the size αm changes the similarities by

$$\Delta u_p = - \sum_{i \in \mathcal{P}} h_i(\mathbf{X}_n) h_i(\mathbf{Y}_n), \quad \Delta u_s = - \sum_{i \in \mathcal{P}} h_i(\mathbf{X}_n) h_i(\mathbf{X}_{n,s}). \quad (208)$$

Next, calculate the first order change of ℓ'_p , we know:

$$\mathbf{u} = (u_p, u_1, \dots, u_S), \quad \Delta \mathbf{u} = (\Delta u_p, \Delta u_1, \dots, \Delta u_S). \quad (209)$$

Using multivariate Taylor expansion up to second order with remainder:

$$\ell'_p(\mathbf{u} + \Delta \mathbf{u}) - \ell'_p(\mathbf{u}) = \nabla \ell'_p(\mathbf{u})^\top \Delta \mathbf{u} + \frac{1}{2} \Delta \mathbf{u}^\top H_p(\mathbf{u}) \Delta \mathbf{u} + o(\|\Delta \mathbf{u}\|^2), \Delta \mathbf{u} \rightarrow 0. \quad (210)$$

By a first order Taylor expansion, we have

$$\begin{aligned} \Delta \ell'_p &= \frac{\partial \ell'_p}{\partial u_p} \Delta u_p + \sum_{s=1}^S \frac{\partial \ell'_p}{\partial u_s} \Delta u_s + o(\|\Delta \mathbf{u}\|) \\ &= \frac{1}{\tau} \ell'_p (1 - \ell'_p) \Delta u_p - \frac{1}{\tau} \ell'_p \sum_{s=1}^S \ell'_s \Delta u_s. \end{aligned} \quad (211)$$

We note that at T_4 , by the convergence of the loss function, we obtain $\ell'_p = 1 - \Theta(\frac{1}{\tau})$, and both ℓ'_p and ℓ'_s take fixed values. Then, by taking expectations over $\Delta \ell'_p$ and using the relation $\sum_s \ell'_s = 1 - \ell'_p$, we obtain:

$$\mathbb{E}[\Delta \ell'_p] = -\Theta\left(\frac{1}{\tau^2}\right) \left(\mathbb{E}\left[\sum_{i \in \mathcal{P}} h_i(\mathbf{X}_n) h_i(\mathbf{Y}_n)\right] - \mathbb{E}\left[\sum_{i \in \mathcal{P}} h_i(\mathbf{X}_n) h_i(\mathbf{X}_{n,s})\right] \right). \quad (212)$$

Also, by Lemma F.1, given that

$$\mathbb{E}\left[\sum_{i \in \mathcal{P}} h_i(\mathbf{X}_n) h_i(\mathbf{Y}_n)\right] - \mathbb{E}\left[\sum_{i \in \mathcal{P}} h_i(\mathbf{X}_n) h_i(\mathbf{X}_{n,s})\right] \geq \Omega(\alpha m \frac{\tau^2}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d}). \quad (213)$$

Hence,

$$\mathbb{E}[\Delta \ell'_p] = -\Omega\left(\frac{\alpha m}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d}\right) < 0. \quad (214)$$

Hence,

$$\begin{aligned} \mathbb{E}[\ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}] &= \mathbb{E}[\ell'_{p, \boldsymbol{\theta}^{(t)}}] + \mathbb{E}[\Delta \ell'_p] \\ &= \mathbb{E}[\ell'_{p, \boldsymbol{\theta}^{(t)}}] - \Omega\left(\frac{\alpha m}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d}\right) \\ &= 1 - \Theta\left(\frac{1}{\tau}\right) - \Omega\left(\frac{\alpha m}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d}\right). \end{aligned} \quad (215)$$

Now, converting to the form of $1 - \ell'$:

$$\mathbb{E}[1 - \ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}] = 1 - \mathbb{E}[\ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}]. \quad (216)$$

Substituting the previous expression gives

$$\mathbb{E}[1 - \ell'_{p, \boldsymbol{\theta}_{\text{mask}}^{(t)}}] = \Theta\left(\frac{1}{\tau}\right) + \Omega\left(\frac{\alpha m}{\Xi_2^2} \epsilon_{j^*} \frac{\log \log d}{d}\right). \quad (217)$$

□

F.7 PROOF OF LEMMA F.3(A):

Proof of Lemma F.3(a):

$$\begin{aligned}
& \mathbb{E} [h_i(\mathbf{Y}_n) \langle \nabla_{\mathbf{w}_i} h_i(\mathbf{X}_n), \mathbf{M}_j \rangle] \\
&= \mathbb{E} \left[h_i(\mathbf{Y}_n) \left\langle \sum_{r=1}^L \mathbf{1}^{(r)} \left| \langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle \right| \geq 0 \cdot \mathbf{z}_{\mathbf{X}}^{(r)}, \mathbf{M}_j \right\rangle \right] \\
&= \mathbb{E} \left[\sum_{s=1}^L \langle \mathbf{w}_i, \mathbf{z}_{\mathbf{Y}}^{(s)} \rangle \cdot \left(\sum_{r=1}^L \langle \mathbf{z}_{\mathbf{X}}^{(r)}, \mathbf{M}_j \rangle \right) \right] \\
&= \frac{1}{L^2} \mathbb{E} \left[\sum_{s=1}^L \langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^{+(s)} + \tilde{\xi}_n^{+(s)} \rangle \cdot \left(\sum_{r=1}^L \langle \mathbf{M} \hat{\mathbf{z}}_n^{(r)} + \tilde{\xi}_n^{(r)}, \mathbf{M}_j \rangle \right) \right] \\
&= \frac{1}{L^2} \mathbb{E} \left[\sum_{s=1}^L \langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^{+(s)} \rangle \cdot \left(\sum_{r=1}^L \langle \mathbf{M} \hat{\mathbf{z}}_n^{(r)}, \mathbf{M}_j \rangle \right) \right] \\
&= \frac{1}{L^2} \mathbb{E} [\langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^+ \rangle \cdot (\langle \mathbf{M} \hat{\mathbf{z}}_n, \mathbf{M}_j \rangle)] \\
&= \frac{1}{L^2} \mathbb{E} \left[\langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^+ \rangle \cdot \left\langle \sum_{j' \in [d]} \mathbf{M}_{j'} \hat{\mathbf{z}}_{n,j'}, \mathbf{M}_j \right\rangle \right] \\
&= \frac{1}{L^2} \mathbb{E} \left[\langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^+ \rangle \cdot \sum_{j' \in [d]} \langle \mathbf{M}_{j'}, \mathbf{M}_j \rangle \hat{\mathbf{z}}_{n,j'} \right] \\
&= \frac{1}{L^2} \mathbb{E} [\langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^+ \rangle \cdot \hat{\mathbf{z}}_{n,j}] \\
&= \frac{1}{L^2} \mathbb{E} \left[\sum_{j'' \in [d]} \langle \mathbf{w}_i, \mathbf{M}_{j''} \rangle \hat{\mathbf{z}}_{n,j''}^+ \hat{\mathbf{z}}_{n,j} \right] \\
&= \frac{1}{L^2} \sum_{j'' \in [d]} \langle \mathbf{w}_i, \mathbf{M}_{j''} \rangle \mathbb{E} [\hat{\mathbf{z}}_{n,j''}^+ \hat{\mathbf{z}}_{n,j}].
\end{aligned} \tag{218}$$

In the final step, we have

$$\frac{1}{L^2} \sum_{j'' \in [d]} \langle \mathbf{w}_i, \mathbf{M}_{j''} \rangle \mathbb{E} [\hat{\mathbf{z}}_{n,j''}^+ \hat{\mathbf{z}}_{n,j}] = \frac{1}{L^2} \langle \mathbf{w}_i, \mathbf{M}_j \rangle \mathbb{E} [\hat{\mathbf{z}}_{n,j}^+ \hat{\mathbf{z}}_{n,j}]. \tag{219}$$

This completes the proof. \square

F.8 PROOF OF LEMMA F.3(B):

Proof of Lemma F.3(b):

$$\begin{aligned}
& \mathbb{E} [h_i(\mathbf{Y}_n) \langle \nabla_{\mathbf{w}_i} h_i(\mathbf{X}_n), \mathbf{M}_j^\perp \rangle] \\
&= \mathbb{E} \left[h_i(\mathbf{Y}_n) \left\langle \sum_{r=1}^L \mathbf{1}_{\left| \langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle \right| \geq 0} \cdot \mathbf{z}_{\mathbf{X}}^{(r)}, \mathbf{M}_j^\perp \right\rangle \right] \\
&= \mathbb{E} \left[h_i(\mathbf{Y}_n) \cdot \left(\sum_{r=1}^L \langle \mathbf{1}_{\left| \langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle \right| \geq 0} \cdot \mathbf{z}_{\mathbf{X}}^{(r)}, \mathbf{M}_j^\perp \rangle \right) \right] \\
&= \mathbb{E} \left[\sum_{s=1}^L \langle \mathbf{w}_i, \mathbf{z}_{\mathbf{Y}}^{(s)} \rangle \mathbf{1}_{\left| \langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{X}}^{(s)} \rangle \right| \geq 0} \cdot \left(\sum_{r=1}^L \langle \mathbf{1}_{\left| \langle \mathbf{w}_i^{(t)}, \mathbf{z}_{\mathbf{X}}^{(r)} \rangle \right| \geq 0} \cdot \mathbf{z}_{\mathbf{X}}^{(r)}, \mathbf{M}_j^\perp \rangle \right) \right] \\
&= \mathbb{E} \left[\sum_{s=1}^L \langle \mathbf{w}_i, \mathbf{z}_{\mathbf{Y}}^{(s)} \rangle \cdot \left(\sum_{r=1}^L \langle \mathbf{z}_{\mathbf{X}}^{(r)}, \mathbf{M}_j^\perp \rangle \right) \right] \\
&= \frac{1}{L^2} \mathbb{E} \left[\sum_{s=1}^L \langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^{+(s)} + \tilde{\xi}_n^{+(s)} \rangle \cdot \left(\sum_{r=1}^L \langle \mathbf{M} \hat{\mathbf{z}}_n^{(r)} + \tilde{\xi}_n^{(r)}, \mathbf{M}_j^\perp \rangle \right) \right] \\
&= \frac{1}{L^2} \mathbb{E} \left[\sum_{s=1}^L \langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^{+(s)} \rangle \cdot \left(\sum_{r=1}^L \langle \mathbf{M} \hat{\mathbf{z}}_n^{(r)}, \mathbf{M}_j^\perp \rangle \right) \right] \\
&= \frac{1}{L^2} \mathbb{E} \left[\langle \mathbf{w}_i, \mathbf{M} \sum_{s=1}^L \hat{\mathbf{z}}_n^{+(s)} \rangle \cdot \left(\langle \mathbf{M} \sum_{r=1}^L \hat{\mathbf{z}}_n^{(r)}, \mathbf{M}_j^\perp \rangle \right) \right] \\
&= \frac{1}{L^2} \mathbb{E} [\langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^+ \rangle \cdot (\langle \mathbf{M} \hat{\mathbf{z}}_n, \mathbf{M}_j^\perp \rangle)] \\
&= \frac{1}{L^2} \mathbb{E} \left[\langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^+ \rangle \cdot \left\langle \sum_{j' \in [d]} \mathbf{M}_{j'} \hat{\mathbf{z}}_{n,j'}, \mathbf{M}_j^\perp \right\rangle \right] \\
&= \frac{1}{L^2} \mathbb{E} \left[\langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^+ \rangle \cdot \left\langle \sum_{j' \in [d]} \mathbf{M}_{j'}, \mathbf{M}_j^\perp \right\rangle \hat{\mathbf{z}}_{n,j'} \right] \\
&= \frac{1}{L^2} \mathbb{E} \left[\langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^+ \rangle \cdot \sum_{j' \in [d]} \langle \mathbf{M}_{j'}, \mathbf{M}_j^\perp \rangle \hat{\mathbf{z}}_{n,j'} \right] \\
&= \frac{1}{L^2} \mathbb{E} [\langle \mathbf{w}_i, \mathbf{M} \hat{\mathbf{z}}_n^+ \rangle \cdot 0] \\
&= 0.
\end{aligned} \tag{220}$$

□

G PROOF OF LEMMAS IN APPENDIX B

G.1 PROOF OF LEMMA B.2(A):

Proof of Lemma B.2(a): At initialization, the neuron weight $\mathbf{w}_i^{(0)}$ is a high dimensional Gaussian vector :

$$\mathbf{w}_i^{(0)} \sim \mathcal{N}(0, \sigma_0^2 I_{d_1}), \tag{221}$$

with $\mathbf{w}_i^{(0)} \in \mathbb{R}^{d_1}$ and each coordinate $\mathbf{w}_i^{(0)}(k) \sim \mathcal{N}(0, \sigma_0^2)$, i.i.d.

$$\left\| \mathbf{w}_i^{(0)} \right\|_2^2 = \sum_{k=1}^{d_1} \left(\mathbf{w}_i^{(0)}(k) \right)^2. \quad (222)$$

We know that $\mathbf{w}_i^{(0)}(k) \sim \mathcal{N}(0, \sigma_0^2)$, so:

$$\frac{1}{\sigma_0^2} \left\| \mathbf{w}_i^{(0)} \right\|_2^2 \sim \chi^2(d_1). \quad (223)$$

According to the concentration inequality of the chi-square distribution:

If $X \sim \chi^2(d_1)$, then for any $0 < \varepsilon < 1$, we have:

$$\Pr \left[\left| \frac{X}{d_1} - 1 \right| \geq \varepsilon \right] \leq 2 \exp \left(-\frac{d_1 \varepsilon^2}{4} \right). \quad (224)$$

Therefore, we have:

$$\Pr \left[\left| \frac{\left\| \mathbf{w}_i^{(0)} \right\|_2^2}{\sigma_0^2 d_1} - 1 \right| \geq \varepsilon \right] \leq 2 \exp \left(-\frac{d_1 \varepsilon^2}{4} \right). \quad (225)$$

Choose a suitable ε to derive the precision range and we choose: $\varepsilon = \tilde{O} \left(\frac{1}{\sqrt{d_1}} \right)$.

At this time, the probability of deviation is:

$$\Pr \left[\left| \left\| \mathbf{w}_i^{(0)} \right\|_2^2 - \sigma_0^2 d_1 \right| \leq \tilde{O}(\sigma_0^2 \sqrt{d_1}) \right] \geq 1 - \frac{1}{\text{poly}(d)}. \quad (226)$$

That is:

$$\left\| \mathbf{w}_i^{(0)} \right\|_2^2 \in \left[\sigma_0^2 d_1 \left(1 - \tilde{O} \left(\frac{1}{\sqrt{d_1}} \right) \right), \sigma_0^2 d_1 \left(1 + \tilde{O} \left(\frac{1}{\sqrt{d_1}} \right) \right) \right]. \quad (227)$$

This holds with high probability $(1 - \frac{1}{\text{poly}(d)})$. \square

G.2 PROOF OF LEMMA B.2(B):

Proof of Lemma B.2(b): Let:

$$\mathbf{Z}_i := \frac{1}{\sigma_0} \mathbf{w}_i^{(0)} \sim \mathcal{N}(0, \mathbf{I}_{d_1}). \quad (228)$$

Then we have:

$$\left\| \mathbf{M} \mathbf{M}^\top \mathbf{w}_i^{(0)} \right\|_2^2 = \sigma_0^2 \cdot \left\| \mathbf{M} \mathbf{M}^\top \mathbf{Z}_i \right\|_2^2. \quad (229)$$

We regard $\mathbf{M} \mathbf{M}^\top$ as a rank- d projection matrix, projecting $\mathbf{Z}_i \in \mathbb{R}^{d_1}$ onto the column space of \mathbf{M} so we can use the following property:

If $\mathbf{M} \mathbf{M}^\top$ is a fixed rank- d projection matrix, and $\mathbf{Z}_i \sim \mathcal{N}(0, \mathbf{I}_{d_1})$, then:

$$\left\| \mathbf{M} \mathbf{M}^\top \mathbf{Z}_i \right\|_2^2 = \mathbf{Z}_i^\top (\mathbf{M} \mathbf{M}^\top)^\top \mathbf{M} \mathbf{M}^\top \mathbf{Z}_i = \mathbf{Z}_i^\top \mathbf{M} \mathbf{M}^\top \mathbf{Z}_i = \left\| \mathbf{M}^\top \mathbf{Z}_i \right\|_2^2, \quad (230)$$

$$\mathbf{M}^\top \mathbf{Z}_i \sim \mathcal{N}(0, \mathbf{I}_d). \quad (231)$$

Therefore, we can conclude:

$$\left\| \mathbf{M} \mathbf{M}^\top \mathbf{Z}_i \right\|_2^2 \sim \chi^2(d) \implies \mathbb{E} \left[\left\| \mathbf{M} \mathbf{M}^\top \mathbf{Z}_i \right\|_2^2 \right] = d. \quad (232)$$

And it satisfies the following Chi-square concentration inequality:

$$\mathbb{P}\left(\left|\|MM^\top \mathbf{Z}_i\|_2^2 - d\right| \leq \varepsilon d\right) \geq 1 - 2\exp(-c\varepsilon^2 d). \quad (233)$$

Choose $\varepsilon = \tilde{O}(1/\sqrt{d})$, and the result holds with high probability. We substitute back $\mathbf{w}_i^{(0)}$

$$\left\|MM^\top \mathbf{w}_i^{(0)}\right\|_2^2 = \sigma_0^2 \cdot \left\|MM^\top \mathbf{Z}_i\right\|_2^2 \in \left[\sigma_0^2 d \left(1 - \tilde{O}\left(\frac{1}{\sqrt{d}}\right)\right), \sigma_0^2 d \left(1 + \tilde{O}\left(\frac{1}{\sqrt{d}}\right)\right)\right]. \quad (234)$$

□

G.3 PROOF OF LEMMA B.2(C):

Proof of Lemma B.2(c): Recall if g is standard Gaussian, then for every $t > 0$,

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{t}{t^2 + 1} e^{-t^2/2} < \Pr_{g \sim \mathcal{N}(0,1)}[g > t] < \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{t} e^{-t^2/2}. \quad (235)$$

Therefore, for every $i \in [m]$ and $j \in [d]$,

$$\begin{aligned} p_1 &= \Pr\left[\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle^2 \geq \frac{c_1 \log d}{d} \|MM^\top \mathbf{w}_i^{(0)}\|_2^2\right] \\ &= \Pr\left[\frac{\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle}{\sigma_0} \geq \sqrt{c_1 \log d}\right] \\ &\geq \Omega\left(\frac{1}{d^{c_1/2}}\right) \\ &= \Omega\left(\frac{1}{d^{(\frac{\varepsilon_{\max}}{\varepsilon_{\min}})^2 \cdot (1+\gamma)}}\right), \end{aligned} \quad (236)$$

and

$$\begin{aligned} p_2 &= \Pr\left[\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle^2 \geq \frac{c_2 \log d}{d} \|MM^\top \mathbf{w}_i^{(0)}\|_2^2\right] \\ &= \Pr\left[\frac{\langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle}{\sigma_0} \geq \sqrt{c_2 \log d}\right] \\ &\leq O\left(\frac{1}{\sqrt{\log d}}\right) \cdot \frac{1}{d^{c_2/2}} \\ &= O\left(\frac{1}{\sqrt{\log d}}\right) \cdot \frac{1}{d^{(\frac{\varepsilon_{\min}}{\varepsilon_{\max}})^2 \cdot (1-\gamma)}}. \end{aligned} \quad (237)$$

We define the following events in definition B.1:

- A_i : Lucky neuron i satisfies conditions 1 (i.e., the response is large enough and in the correct direction)
- B_i : for all $j' \neq j$, lucky neuron i satisfies condition 2 (i.e., small responses in other directions)

We now compute the probability of the intersection event $A_i \cap B_i$:

$$\begin{aligned} \Pr[A_i] &= \frac{p_1}{2} = \Omega\left(d^{-\left(\frac{\varepsilon_{\max}}{\varepsilon_{\min}}\right)^2 \cdot (1+\gamma)}\right), \\ \Pr[B_i] &= (1 - p_2)^{d-1} = e^{-(d-1)p_2} = e^{-(d-1)d^{-a}} = e^{-d^{1-a}} = 1, \\ \Pr[A_i \cap B_i] &= \frac{p_1}{2} \cdot (1 - p_2)^{d-1} = \Omega\left(\frac{1}{\sqrt{\log d}} \cdot d^{-\left(\frac{\varepsilon_{\max}}{\varepsilon_{\min}}\right)^2 \cdot (1+\gamma)}\right). \end{aligned} \quad (238)$$

(1) We now have $m = d^{C_m}$ neurons. Therefore, the expected number is:

$$\begin{aligned}\mathbb{E}[|\mathcal{M}_j^*|] &= m \cdot \Pr[A_i \cap B_i] = d^{C_m} \cdot \Omega\left(d^{-\left(\frac{\epsilon_{\max}}{\epsilon_{\min}}\right)^2 \cdot (1+\gamma)}\right) \\ &= \Omega\left(d^{C_m - \left(\frac{\epsilon_{\max}}{\epsilon_{\min}}\right)^2 \cdot (1+\gamma)}\right).\end{aligned}\quad (239)$$

Chernoff bound (Lower-tail form): For any $\delta \in (0, 1)$, we have:

$$\Pr\left[\sum X_i < (1 - \delta)\mu\right] \leq e^{-\frac{\delta^2}{2}\mu}.$$
 (240)

Let $\delta = \frac{1}{2}$, we obtain:

$$\begin{aligned}\Pr\left[\sum X_i < \frac{1}{2}\mu\right] &\leq e^{-\mu/8} \\ \Pr[|\mathcal{M}_j^*| < O(d^{\omega_1})] &\leq e^{-\Omega(d^{\omega_1})} \\ \Pr[|\mathcal{M}_j^*| > \Omega(d^{\omega_1})] &\geq 1 - e^{-\Omega(d^{\omega_1})}.\end{aligned}\quad (241)$$

(2) We now have $m = d^{C_m}$ neurons. Therefore, the expected number is:

$$\begin{aligned}\mathbb{E}[|\mathcal{M}_j|] &= m \cdot p_2 = d^{C_m} \cdot \mathcal{O}\left(\frac{1}{\sqrt{\log d}} \cdot d^{-\left(\frac{\epsilon_{\min}}{\epsilon_{\max}}\right)^2 \cdot (1-\gamma)}\right) \\ &= \mathcal{O}\left(\frac{1}{\sqrt{\log d}} \cdot d^{C_m - \left(\frac{\epsilon_{\min}}{\epsilon_{\max}}\right)^2 \cdot (1-\gamma)}\right).\end{aligned}\quad (242)$$

Chernoff bound (upper tail) tells us that for any $0 < \delta < 1$, we have:

$$\begin{aligned}\Pr\left[\sum X_i > (1 + \delta)\mu\right] &\leq e^{-\Omega(\delta^2\mu)} \\ \Pr\left[|\mathcal{M}_j| > \Omega\left(\frac{1}{\sqrt{\log d}} d^{\omega_2}\right)\right] &\leq e^{-\Omega\left(\frac{1}{\sqrt{\log d}} d^{\omega_2}\right)} = o\left(\frac{1}{d^4}\right) \\ \Pr\left[|\mathcal{M}_j| < O\left(\frac{1}{\sqrt{\log d}} d^{\omega_2}\right)\right] &\geq 1 - o\left(\frac{1}{d^4}\right) \\ \Pr[|\mathcal{M}_j| < O(d^{\omega_2})] &\geq 1 - o\left(\frac{1}{d^4}\right).\end{aligned}\quad (243)$$

□

G.4 PROOF OF LEMMA B.2(D):

Proof of Lemma B.2(d): We know: $|\mathcal{M}_j| \leq O(d^{\omega_2})$. There are d indices $j \in [d]$. Therefore, the total number of pairs (i, j) such that $i \in \mathcal{M}_j$ is at most:

$$\sum_{j=1}^d |\mathcal{M}_j| \leq d \cdot O(d^{\omega_2}) = O(d^{1+\omega_2}). \quad (244)$$

On the other hand, the total number of neurons is $m = d^{C_m}$. So for any fixed i , we define:

$$\mathcal{N}_i := \{j \in [d] : i \in \mathcal{M}_j\}. \quad (245)$$

Then,

$$\sum_{i=1}^m |N_i| = \sum_{j=1}^d |\mathcal{M}_j| \leq O(d^{1+\omega_2}). \quad (246)$$

Therefore,

$$\mathbb{E}[|N_i|] = \frac{1}{m} \sum_{i=1}^m |N_i| \leq O(d^{1+\omega_2 - C_m}) = O\left(d^{1 - \left(\frac{\epsilon_{\min}}{\epsilon_{\max}}\right)^2 \cdot (1-\gamma)}\right). \quad (247)$$

Then:

$$\Pr \left[\left| \langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle \right| \geq \Omega(\sigma_0 \log^{1/4} d) \right] \leq 2 \exp \left(-\frac{t^2}{2\sigma_0^2} \right) = 2^{-\Omega(\sqrt{\log d})}. \quad (248)$$

Fix $i \in [m]$, and consider d different j . Each has probability $2^{-\Omega(\sqrt{\log d})}$ to exceed the threshold. Therefore, the expectation is:

$$\mathbb{E} \left[\left| \left\{ j \in [d] \mid \left| \langle \mathbf{w}_i^{(0)}, \mathbf{M}_j \rangle \right| \geq \Omega(\sigma_0 \log^{1/4} d) \right\} \right| \right] = O \left(2^{-\sqrt{\log d}} \cdot d \right). \quad (249)$$

□

H PROOF OF LEMMAS IN APPENDIX C

This section can be found in the Supplementary Material.

I PROOF OF LEMMAS IN APPENDIX D

This section can be found in the Supplementary Material.

J PROOF OF LEMMAS IN APPENDIX E

This section can be found in the Supplementary Material.

K PROOF OF ADDITIONAL LEMMAS

This section can be found in the Supplementary Material.