On the Training Dynamics of Contrastive Learning with Imbalanced Feature Distributions: A Theoretical Study of Feature Learning

Haixu Liao
Department of Data Science
New Jersey Institute of Technology
Newark, NJ 07102
hl534@njit.edu

Yating Zhou
Department of EECS
Rensselaer Polytechnic Institute
Troy, NY 12180
zhouy26@rpi.edu

Songyang Zhang
Department of Electrical and Computer Engineering
University of Louisiana at Lafayette
Lafayette, LA 70504
songyang.zhang@louisiana.edu

Shuai Zhang
Department of Data Science
New Jersey Institute of Technology
Newark, NJ 07102
shuai.zhang@njit.edu

Abstract

Contrastive learning has served as a powerful framework in the early development of vision–language models (VLMs), demonstrating remarkable effectiveness in learning generalizable representations and establishing itself as the foundation for many state-of-the-art systems. However, despite these advances, its theoretical understanding remains limited, particularly under imbalanced data distributions that are prevalent in real-world settings. Such imbalance can degrade representation quality and induce biased model behavior, yet a rigorous characterization of these effects is still 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 differently across three stages of training, with distinct dynamics for majority features, minority features, and the noise. We further show that minority features diminish neurons' representational capacity, increase the need for more complex architectures, and impair the separation of ground-truth features from noise. These findings offer new theoretical insights into how data imbalance shapes learning in contrastive frameworks and serve as an early step towards principled modifications for developing more robust and unbiased representations.

1 Introduction

Contrastive learning has emerged as a powerful paradigm in representation learning, effectively leveraging unlabeled data without relying on human-annotated 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 [40, 9, 43, 15]. This approach has demonstrated remarkable success across a wide range of applications and has been particularly influential in multi-modal learning [21, 16], driving major advances in the early development of vision-language models [27].

Preprint.

Despite its strengths, contrastive learning faces challenges when applied to real-world datasets with class imbalance. In such scenarios, majority classes dominate the sample space, while minority classes with limited samples are underrepresented in both positive and negative pair formation. This imbalance can hinder the learning process, causing the model to under-capture discriminative features associated with minority classes, ultimately degrading representation quality. Several studies have attempted to address the challenge of contrastive learning under imbalanced data. One line of research focuses on sample re-weighting strategies, which aim to balance the influence of minority and majority class samples [5, 12, 19, 32, 30]. Another line of work explores data resampling methods, such as oversampling minority data or undersampling majority data, to achieve a more balanced training distribution [6, 11, 25, 31]. However, both approaches rely heavily on accurate estimation of re-weighting or re-sampling ratios, which is an aspect that is often difficult to characterize precisely and typically depends on human intuition or heuristic methods.

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 [14], the necessity of data augmentation for effective representation learning [38], and its ability to produce representations that reduce the sample complexity of downstream tasks [8]. Nonetheless, these studies have not considered the implications of imbalanced data distributions.

Most existing studies on contrastive learning focus on empirical performance, while its theoretical foundations, especially for feature learning, remain less understood. 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 two 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. Our results show that neuron weights evolve differently when learning majority features, minority features, and noise across the three stages of training.

Second, we quantitatively characterize how the presence of minority features influences neurons' learning capacity and, consequently, representation learning. Our analysis shows that neurons learn majority features more quickly, while minority features are acquired more slowly. Moreover, in the presence of minority features, capturing effective representations requires a more complex neural network, and the neurons' ability to distinguish ground-truth features from noise becomes degraded.

2 Problem Formulation

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

Let $(X_n, X_{n'})$ 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:

$$\ell(f, X_n, X_{n'}, \mathfrak{N}) := -\log\left(\frac{e^{\mathrm{sim}_f(X_n, X_{n'})/\tau}}{\sum_{x \in \{X_{n'}\} \cup \mathfrak{N}} e^{\mathrm{sim}_f(X_n, x)/\tau}}\right), \tag{1}$$

where the similarity function is given by $\mathrm{sim}_f(X_n, X_{n'}) := \big\langle f(X_n), \mathrm{StopGrad}\big(f(X_{n'})\big)\big\rangle$, and $\mathrm{StopGrad}(\cdot)$ 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 over a batch of size K, i.e.,

$$\widehat{L}_{\text{aug}}(f_t) = \widehat{L}(f_t) + \frac{\lambda}{2}\|w\|_F^2 = \frac{1}{K}\sum_{k=1}^{K} \ell\big(f; X_k, X_{k'}, \mathfrak{N}^k\big) + \frac{\lambda}{2}\|w\|_F^2, \tag{2}$$

where w is the neural network parameters.

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

$$\begin{aligned} f(X_n) &= \big(h_1(X_n), \ldots, h_m(X_n)\big)^\top \in \mathbb{R}^m, \\ \text{with} \quad h_i(X_n) &= \sum_{r=1}^{L} \text{BReLU}_{b_i^{(t)}}\Big(\langle w_i^{(t)}, \text{Attention}(W_Q x_n^{(r)}, W_K X_n, W_V X_n)\rangle\Big). \end{aligned} \tag{3}$$

In this early stage of our analysis, we fix attention layer weights to identity matrices and focus on the MLP layer weights. Note that the analysis of this model still differs substantially from a standard feedforward network because the preceding self-attention aggregates information across tokens.

3 Theoretical Analysis

3.1 Key Insights of the Findings

(K1). Training dynamics of contrastive learning based on the Transformer-MLP framework. The theory shows that the learning process can be divided into three stages. In the first stage, neuron weights in feature directions start to increase, while their components in non-feature directions stay almost unchanged. In the second stage, the alignment with feature directions keeps growing, and the learned features become purer, while non-feature directions remain suppressed. In the final stage, each neuron aligns with a specific set of features $\mathcal{N}_i$, on which it already had some degree of alignment at initialization.

(K2). Theoretical characterizations of how imbalanced data in affecting the neuron's learning ability. In the first stage of training, neurons start to increase along feature directions, and the speed of this growth depends on the feature frequency ϵ_j. Features with larger ϵ_j grow faster, so neurons align with them more quickly. Features with smaller ϵ_j grow more slowly, and neurons may need more time to capture them. In the second stage, this difference becomes more visible, as neurons that follow features with larger ϵ_j keep increasing their alignment, while features with smaller ϵ_j continue to evolve at a slower pace.

(K3). The effect of the ratio $\epsilon_{\min}/\epsilon_{\max}$ on the final learning state. In the final stage of training, the feature frequency ratio $\epsilon_{\min}/\epsilon_{\max}$ controls how neurons distribute their weights across different features. For minority features, $\epsilon_j = \epsilon_{\min}$, so the ratio directly determines the size of the coefficient $\alpha_{i,j}$. When the ratio $\epsilon_{\min}/\epsilon_{\max}$ is small, each $\alpha_{i,j}$ for minority features becomes very small. As a result, the set $\mathcal{N}_i$ becomes larger, meaning that each neuron aligns with more features that had some degree of initialization alignment. However, the contribution from each feature is weaker, so the neuron ends up mixing many features together in a more mixed way. In contrast, when the ratio $\epsilon_{\min}/\epsilon_{\max}$ is larger, the coefficients $\alpha_{i,j}$ become stronger. In this case, the set $\mathcal{N}_i$ becomes smaller, and each neuron aligns with fewer features. This makes the final representation more concentrated, and the features learned by each neuron are purer.

3.2 Formal Theoretical Results

Theorem 3.1 describes two main effects of gradient descent in the first stage of training: (i) In the feature directions, the neuron weights increase rapidly as shown in (4), while in the non-feature directions they are suppressed as shown in (5) during training. (ii) The growth of a neuron's weight in a feature direction $\mathbf{M}_j$ depends on the frequency ϵ_j as shown in (4). Larger ϵ_j leads to faster growth, while smaller ϵ_j results in slower growth, making the feature harder to capture in the early stages of training. In short, the feature frequency ϵ_j directly controls how much the inner product $\langle w_i^{(t)}, \mathbf{M}_j\rangle$ increases under gradient descent.

Table 1: Summary of main notations

η	Learning rate	λ	Regularization parameter
τ	Temperature coefficient	K	Batch size
$w_i^{(t)}$	The Neuron i after t steps of GD	$\mathbf{M}_j$	The feature vector of feature j
$\mathfrak{N}$	Set of negative samples	$\mathfrak{B}$	The set of $X_{n'}$ and negative samples
$\epsilon_{\min}$	frequency of minority feature	$\epsilon_{\max}$	frequency of majority feature
ϵ_j	Feature frequency for feature j	$\mathcal{N}_i$	Set of features for ordinary neuron i
$\mathcal{M}_j$	Set of ordinary neurons for feature j	$\mathcal{M}_j^\star$	Set of lucky neurons for feature j

Theorem 3.1 (Stage 1). *During the first training stage, the update of neuron weights $w_i^{(t)}$ at time T_1 can be bounded as follows.*

$$|\langle w_i^{(T_1)}, \mathbf{M}_j\rangle| \geq |\langle w_i^{(0)}, \mathbf{M}_j\rangle|(1+\epsilon_j C_z \log d) - \tilde{O}\Big(\frac{\|w_i^{(T_1)}\|_2}{\mathrm{poly}(d)}\Big) \tag{4}$$

$$|\langle w_i^{(T_1)}, \mathbf{M}_j^\perp\rangle| \leq |\langle w_i^{(0)}, \mathbf{M}_j^\perp\rangle| + \tilde{O}\Big(\frac{\|w_i^{(T_1)}\|_2}{\mathrm{poly}(d)}\Big) \tag{5}$$

Theorem 3.2 describes the gradient descent dynamics in the second stage of training, focusing on how neurons behave in different directions. (i) For neurons that belong to $\mathcal{M}_j^\star$, their inner product with the feature vector keeps increasing as shown in (6) (ii) In contrast, along the noise direction ($\mathbf{M}_j^\perp$), the growth stays almost unchanged as shown in (7). In particular, the value of ϵ_j determines how quickly neurons in the feature directions evolve during training.

Theorem 3.2 (Stage 2). *During the second training stage, the update of neuron weights $w_i^{(t)}$ at time T_2 can be bounded as follows. For each $j \in [d]$, if $i \in \mathcal{M}_j^\star$, then:*

$$|\langle w_i^{(T_2)}, \mathbf{M}_j\rangle| \geq \Omega(1)\|w_i^{(T_2)}\|_2 \tag{6}$$

If along the orthogonal non-feature direction $\mathbf{M}_j^\perp$:

$$|\langle w_i^{(T_2)}, \mathbf{M}_j^\perp\rangle| \leq |\langle w_i^{(T_1)}, \mathbf{M}_j^\perp\rangle| + \tilde{O}\Big(\frac{\|w_i^{(T_2)}\|_2}{\mathrm{poly}(d)}\Big). \tag{7}$$

Theorem 3.3 describes the feature learning behavior of neurons in the final stage. Specifically, we prove that: (i) Each neuron weight vector w_i eventually aligns with a set of features $\mathcal{N}_i$. This set corresponds to the features that already had some degree of alignment with w_i at initialization. (ii) The size of $\mathcal{N}_i$ depends on the ratio $\epsilon_{\min}/\epsilon_{\max}$. A smaller ratio enlarges $|\mathcal{N}_i|$, leading neurons to encode more mixed features, whereas a ratio closer to one yields smaller $|\mathcal{N}_i|$, so neurons capture purer features that benefit representation learning. (iii) For each feature $\mathbf{M}_j$, the number of neurons that contain some degree of initialization component along $\mathbf{M}_j$ admits an upper bound. Moreover, there are at least $\Omega(d^{\omega_1})$ neurons with $\mathcal{N}_i = \{j\}$, where $\omega_1 = C_m - \left(\frac{\epsilon_{\max}}{\epsilon_{\min}}\right)^2(1+\gamma c_0)$, indicating imbalanced data leads to less number of neurons in learning the purified feature.

Theorem 3.3 (Stage 3: Neuron–Feature Alignment in Contrastive Learning). *Let $m = d^{C_m}$ be the number of neurons and $\tau = \mathrm{polylog}(d)$. Suppose we train the neural net f_t 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:*

$$\frac{1}{T}\sum_{t\in[T]} L_{\mathrm{aug}}(f_t) \leq o(1), \quad \frac{1}{T}\sum_{t\in[T]} \mathcal{L}(f_t) \leq o(1). \tag{8}$$

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

$$w_i^{(t)} = \sum_{j\in\mathcal{N}_i} \alpha_{i,j}\mathbf{M}_j + \sum_{j\notin\mathcal{N}_i} \alpha'_{i,j}\mathbf{M}_j + \sum_{j\in[d_1]\setminus[d]} \beta_{i,j}\mathbf{M}_j^\perp, \tag{9}$$

where $\alpha_{i,j} \in \left[\frac{\epsilon_j}{\epsilon_{\max}}\frac{\tau}{\Xi_2}, \frac{\epsilon_j}{\epsilon_{\max}}\tau\right], \alpha'_{i,j} \leq o\Big(\frac{\epsilon_j}{\epsilon_{\max}}\frac{1}{\sqrt{d}}\Big)\|w_i^{(t)}\|_2, |\beta_{i,j}| \leq o\Big(\frac{1}{\sqrt{d_1}}\Big)\|w_i^{(t)}\|_2$. *Furthermore, the size of* $\mathcal{N}_i$ *is bounded by* $|\mathcal{N}_i| = O\Big(d^{1-\left(\frac{\epsilon_{\min}}{\epsilon_{\max}}\right)^2\cdot(1-\gamma c_0)}\Big)$. *Finally, for each dictionary atom* $\mathbf{M}_j$, *there are at least* $\Omega(d^{\omega_1})$ *neurons* $i \in [m]$ *such that* $\mathcal{N}_i = \{j\}$.

4 Numerical Experiments

Following our learning setup, we validate our theoretical insights on synthetic data with parameters $m = 30$ and $d = 9$ (Details can be found in supplementary materials). In Figure 1, the x-axis represents the feature index, and the y-axis represents the neuron index, where we only plot the first 13 neurons to save space. Each entry corresponds to the projection of a neuron's weight onto the direction of the associated feature; larger values indicate stronger alignment between the neuron and that feature. The first five features (columns 1–5) are majority features, while the last four (columns 6–9) are minority features. As the figure illustrates, neurons exhibit significantly larger projections onto majority features. Nearly every neuron is strongly associated with at least one majority feature. At the same time, each majority feature is represented by at least one neuron, and in such cases, the projection onto that feature is substantially larger than onto all others, meaning the feature dominates the neuron's representation. This demonstrates that majority features are easier to learn and tend to be represented by multiple neurons, in contrast to the minority features.

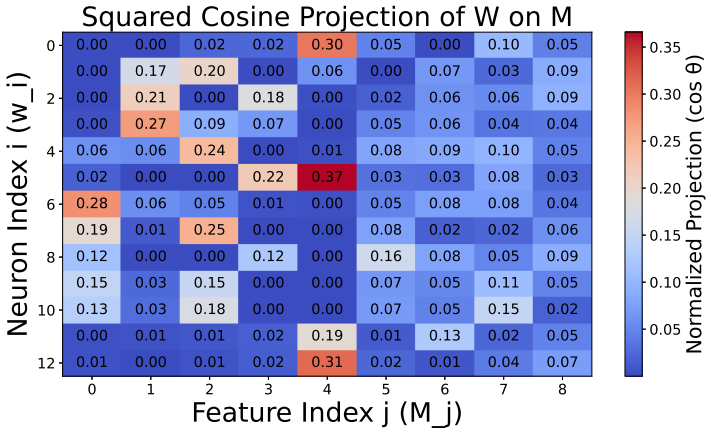

Figure 1: Squared cosine projection of the first 13 neurons (w_i) on 9 dictionary atoms (M_j). The first five atoms are majority features, and the last four are minority features.

5 Conclusion

This work takes a step toward a principled understanding of how imbalanced data shapes the dynamics of contrastive learning in Transformer-based encoders. Our analysis shows that imbalance harms performance: minority features reduce neurons' representational capacity, increase the demand for more complex architectures, and hinder the separation of ground-truth features from noise. Looking ahead, a promising direction is to investigate how these insights can inspire the design of more principled methods or help explain the effectiveness of existing approaches in addressing imbalance in contrastive learning.

Acknowledgments

This work was supported by the National Science Foundation (NSF) #2349879 and #2349878. We also thank all anonymous reviewers for their constructive comments.

References

[1] 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), pages 977–988. IEEE, 2022.

[2] 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.

[3] 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.

[4] 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*, pages 6074–6114. PMLR, 2023.

[5] 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.

[6] 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.

[7] Peter Foldiak. Sparse coding in the primate cortex. In *The Handbook of Brain Theory and Neural Networks*. MIT Press, 2003.

[8] Siddhant Garg and Yingyu Liang. Functional regularization for representation learning: A unified theoretical perspective. In *Advances in Neural Information Processing Systems*, 2020.

[9] 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*, pages 2703–2708, 2021.

[10] 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.

[11] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009.

[12] 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.

[13] Yu Huang, Zixin Wen, Yuejie Chi, and Yingbin Liang. Transformers provably learn feature-position correlations in masked image modeling. *CoRR*, 2024.

[14] 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.

[15] 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.

[16] 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.

[17] Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? *arXiv preprint arXiv:2402.15607*, 2024.

[18] Hongkang Li, Yihua Zhang, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. *arXiv preprint arXiv:2504.10957*, 2025.

[19] 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.

[20] Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 2014.

[21] 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*, pages 4348–4380. PMLR, 2023.

[22] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 1997.

[23] Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 2004.

[24] Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, pages 26724–26768. PMLR, 2023.

[25] 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.

[26] Matan Protter and Michael Elad. Image sequence denoising via sparse and redundant representations. *IEEE Transactions on Image Processing*, 2008.

[27] 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*, pages 8748–8763. PMLR, 2021.

[28] 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*, pages 19250–19286. PMLR, 2022.

[29] 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)*, pages 5628–5637. PMLR, 2019.

[30] Ziqiao Shang, Bin Liu, Fengmao Lv, Fei Teng, and Tianrui Li. Learning contrastive feature representations for facial action unit detection. *arXiv preprint arXiv:2402.06165*, 2024.

[31] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *Computer Vision – ECCV 2016*, 2016.

[32] 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.

[33] Jiawei Sun, Shuai Zhang, Hongkang Li, and Meng Wang. Theoretical guarantees and training dynamics of contrastive learning: How misaligned data influence feature purity. In *High-dimensional Learning Dynamics 2025*, 2025.

[34] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

[35] Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.

[36] William E Vinje and Jack L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 2000.

[37] 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.

[38] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[39] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021.

[40] Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*, pages 24851–24871. PMLR, 2022.

[41] 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*, pages 1794–1801. IEEE, 2009.

[42] S. Zhang, M. Wang, P.-Y. Chen, S. Liu, S. Lu, and M. Liu. Joint edge-model sparse learning is provably efficient for graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

[43] 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*, 2022.

A Related Works

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 [28, 35, 29, 39, 33, 4]. [34, 37] establishes a connection between the gradients of contrastive learning and graph neural networks, highlighting interpretability through a graph-theoretic perspective. [10] 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. [39] 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 [1, 42, 18, 3], where features are tied to ground-truth labels; however, such settings cannot be directly extended to contrastive learning. Due to 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 [13, 24, 17], even in supervised settings. The most relevant work is [39], which analyzes 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.

B Preliminaries

Data Model: Sparse Coding. For the necessity of theoretical proof, we adopt the sparse coding model [22, 23, 7, 36, 23, 26, 41, 20, 2] as a conceptual modeling of real-world data. Specifically, for a paired data $(X_n, X_{n'})$, the data structure is

$$\begin{aligned} X_n &= \left[\mathbf{M}z_n^{(1)} + \xi_n^{(1)},\ \mathbf{M}z_n^{(2)} + \xi_n^{(2)},\ \ldots,\ \mathbf{M}z_n^{(L)} + \xi_n^{(L)}\right] \\ X_{n'} &= \left[\mathbf{M}z_{n'}^{(1)} + \xi_{n'}^{(1)},\ \mathbf{M}z_{n'}^{(2)} + \xi_{n'}^{(2)},\ \ldots,\ \mathbf{M}z_{n'}^{(L)} + \xi_{n'}^{(L)}\right] \end{aligned} \tag{10}$$

Here, each $z_n^{(i)} \in \mathbb{R}^d$ represents the latent signal at the ℓ-th token, $\xi_n^{(i)}$ denotes the additive noise, and $\mathbf{M} = [\mathbf{M}_1, \ldots, \mathbf{M}_d] \in \mathbb{R}^{d_1 \times d}$ is the dictionary matrix. For each index, $z_{n,j}^{(i)} = 0$ indicates that the corresponding feature is absent in the token, while $z_{n,j}^{(i)} \neq 0$ indicates that the feature is present.

For a positive pair, we assume they share the same group of features when counting over all tokens in the sample, whereas negative samples are independent. That is to say, $\sum_{\ell=1}^{L} z_n^{(\ell)}$ and $\sum_{\ell=1}^{L} z_{n'}^{(\ell)}$ share the same support in a positive pair, while $z_n^{(\ell)}$ and $z_{n'}^{(\ell)}$ remain independent in a negative pair.

We first recall a useful concentration property. Whenever the Frobenius norm of the weight matrix satisfies:

$$\|w^{(t)}\|_F^2 = \sum_{i\in[m]} \|w_i^{(t)}\|_2^2 \leq \mathrm{poly}(d), \tag{11}$$

the following estimate can be obtained by applying Bernstein concentration inequalities.

Fact B.1 (Approximation of population gradients by empirical gradients)**.** *Suppose that* $\|w^{(t)}\|_F^2 \leq \mathrm{poly}(d)$. *Then there exists some* $K = \mathrm{poly}(d)$ *such that, with high probability, the difference between*

the empirical gradients and the population gradients is bounded for every iteration t:

$$\left\|\nabla_{w_i}\widehat{\mathrm{Obj}}(f_t) - \nabla_{w_i}\mathrm{Obj}(f_t)\right\|_2 \leq \frac{\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)}, \quad \forall i \in [m]. \tag{12}$$

To facilitate the calculation of the gradient of the loss function $\mathcal{L}(f_t, X_n, \mathcal{B}^\ell)$ with respect to the weights $\{w_i^{(t)}\}_{i\in[m]}$, we introduce the following notation. We denote by $\ell'_{p,t}(X_n, \mathcal{B})$ the positive logit, and by $\ell'_{s,t}(X_n, \mathcal{B})$ the negative logits:

$$\ell'_{p,t}(X_n, \mathcal{B}) := \frac{\exp\left(\mathrm{Sim}_{f_t}(X_n, X_{n'})/\tau\right)}{\sum_{x\in\mathcal{B}}\exp\left(\mathrm{Sim}_{f_t}(X_n, x)/\tau\right)} \tag{13}$$

$$\ell'_{s,t}(X_n, \mathcal{B}) := \frac{\exp\left(\mathrm{Sim}_{f_t}(X_n, X_{n,s})/\tau\right)}{\sum_{x\in\mathcal{B}}\exp\left(\mathrm{Sim}_{f_t}(X_n, x)/\tau\right)} \tag{14}$$

The empirical gradient of $L(f_t)$ with respect to the weight $w_i^{(t)}$ at iteration t is given by (note that the similarity measure Sim_{f_t} makes use of the StopGrad operation):

$$\nabla_{w_i}L(f_t) = \mathbb{E}\left[(\ell'_{p,t}-1)h_i(X_{n'})\sum_{r=1}^{L}\mathbf{1}_{|\langle w_i, z_X^{(r)}\rangle|\geq b_i}z_X^{(r)} + \sum_{X_{n,s}\in\mathfrak{N}}\ell'_{s,t}h_i(X_{n,s})\sum_{r=1}^{L}\mathbf{1}_{|\langle w_i, z_X^{(r)}\rangle|\geq b_i}z_X^{(r)}\right] \tag{15}$$

C Lemmas

Definition C.1 (Characterization of Neurons)**.** *We choose constants*

$$c_1 = \left(\frac{\epsilon_{\max}}{\epsilon_{\min}}\right)^2 \cdot 2(1+\gamma c_0), \quad c_2 = \left(\frac{\epsilon_{\min}}{\epsilon_{\max}}\right)^2 \cdot 2(1-\gamma c_0), \quad \gamma c_0 \in (0, 0.001)$$

We define the following sets of neurons, which will be useful for analyzing the stochastic gradient descent trajectory in later sections:

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

$$\mathcal{M}_j := \left\{ i \in [m] : \langle w_i^{(0)}, \mathbf{M}_j \rangle^2 \geq \frac{c_2 \log d}{d} \left\| \mathbf{M}\mathbf{M}^\top w_i^{(0)} \right\|_2^2 \right\}, \quad \forall j \in [d] \tag{16}$$

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

$$\mathcal{M}_j^\star := \left\{ \begin{array}{c} i \in [m] : \langle w_i^{(0)}, \mathbf{M}_j \rangle^2 \geq \frac{c_1 \log d}{d} \left\| \mathbf{M}\mathbf{M}^\top w_i^{(0)} \right\|_2^2, \\ \langle w_i^{(0)}, \mathbf{M}_{j'} \rangle^2 \leq \frac{c_2 \log d}{d} \left\| \mathbf{M}\mathbf{M}^\top w_i^{(0)} \right\|_2^2, \quad \forall j' \in [d], j' \neq j \end{array} \right\}, \quad \forall j \in [d] \tag{17}$$

Lemma C.1. *At initialization ($t = 0$), the following properties hold:*

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

$$\|w_i^{(0)}\|_2^2 \in \left[\sigma_0^2 d_1 \left(1 - \widetilde{O}\left(\tfrac{1}{\sqrt{d_1}}\right)\right),\ \sigma_0^2 d_1 \left(1 + \widetilde{O}\left(\tfrac{1}{\sqrt{d_1}}\right)\right) \right]. \tag{18}$$

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

$$\|\mathbf{M}\mathbf{M}^\top w_i^{(0)}\|_2^2 \in \left[\sigma_0^2 d \left(1 - \widetilde{O}\left(\tfrac{1}{\sqrt{d}}\right)\right),\ \sigma_0^2 d \left(1 + \widetilde{O}\left(\tfrac{1}{\sqrt{d}}\right)\right) \right]. \tag{19}$$

(c) Let $m = d^{C_m}$ be the number of neurons and we note $\omega_1 = C_m - \left(\frac{\epsilon_{\max}}{\epsilon_{\min}}\right)^2 (1+\gamma c_0)$, $\omega_2 = C_m - \left(\frac{\epsilon_{\min}}{\epsilon_{\max}}\right)^2 (1-\gamma c_0)$. With probability at least $1 - o\left(\frac{1}{d^4}\right)$, for each $j \in [d]$,

$$|\mathcal{M}_j^\star| \geq \Omega(d^{\omega_1}) =: \Xi_1, \qquad |\mathcal{M}_j| \leq O(d^{\omega_2}) =: \Xi_2. \tag{20}$$

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

$$\Lambda_i := \left\{ j \in [d] : |\langle w_i^{(0)}, \mathbf{M}_j \rangle| \leq \tfrac{\sigma_0}{d} \right\} \subseteq [d]. \tag{21}$$

Then

$$|\Lambda_i| = O\left(\tfrac{d}{\mathrm{polylog}(d)}\right). \tag{22}$$

(e) For any $j' \neq j$, we have

$$|\mathcal{M}_{j'} \cap \mathcal{M}_j| \leq O(\log d), \tag{23}$$

with probability at least $1 - o(1/d^4)$.

(f) For each $i \in [m]$, there are at most $O(1)$ indices $j \in [d]$ such that $i \in \mathcal{M}_j$, and at most $O(2^{-\sqrt{\log d}} d)$ indices $j \in [d]$ such that

$$|\langle w_i^{(0)}, \mathbf{M}_j \rangle| \geq \Omega(\sigma_0 \log^{1/4} d). \tag{24}$$

Lemma C.2 (Pre-activation size I)**.** *Let $z_X^{(r)} = \frac{1}{L}\left(\mathbf{M}\tilde{z}_n^{(r)} + \tilde{\xi}_n^{(r)}\right) \sim \mathcal{D}_{z_X}$, $w_i \in \mathbb{R}^{d_1}$. Define $z_X^{(r)\setminus j} = \frac{1}{L}\left(\sum_{j' \neq j,\, j' \in [d]} \mathbf{M}_{j'} \tilde{z}_{n,j'}^{(r)} + \tilde{\xi}_n^{(r)}\right)$. Then the following results hold:*

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

$$\Pr_{\tilde{z}_n^{(r)\setminus j}, \tilde{\xi}_n^{(r)}} \left(\left(\langle w_i, z_X^{(r)\setminus j} \rangle + \tfrac{1}{L} \langle w_i, M_j \rangle \tilde{z}_{n,j}^{(r)} \right)^2 > \tfrac{\lambda \|w_i\|_2^2 \sqrt{\log d}}{d} \right) \leq O\left(\tfrac{1}{\lambda}\right). \tag{25}$$

The same tail bound applies to $\langle w_i, z_X^{(r)}\rangle$, $\langle w_i, \frac{z_Y^{(s)} - z_X^{(r)}}{2}\rangle$, *and* $\langle w_i, \tilde{\xi}_n^{(r)}\rangle$.

(b) High probability bound for sparse signal:

$$\Pr\left(\langle w_i, \mathbf{M}\tilde{z}_n^{(r)}\rangle^2 > \|w_i\|_2^2 \cdot \max_{j\in[d]} \|\mathbf{M}_j\|_\infty^2 \log^4 d\right) \lesssim e^{-\Omega(\log^2 d)}. \tag{26}$$

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

$$\Pr\left(Z^2 \geq \tfrac{\|w_i\|_2^2 \log^4 d}{d}\right) \lesssim e^{-\Omega(\log^2 d)}. \tag{27}$$

Lemma C.3 (Pre-activation size II)**.** *Suppose the following conditions hold:*

$$\langle w_i^{(t)}, \mathbf{M}_j\rangle^2 \geq \Omega\big((b_i^{(t)})^2\big) \quad \textit{for at most } O(1) \textit{ indices } j \in [d], \tag{28}$$

$$\langle w_i^{(t)}, \mathbf{M}_j\rangle^2 \geq \Omega\left(\frac{(b_i^{(t)})^2}{\sqrt{\log d}}\right) \quad \textit{for at most } O\big(e^{-\Omega(\sqrt{\log d})}d\big) \textit{ indices } j \in [d], \tag{29}$$

$$\|w_i^{(t)}\|_2^2 \leq O\left(\frac{d(b_i^{(t)})^2}{\log d}\right). \tag{30}$$

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

$$\Pr\Big(\big|\langle w_i^{(t)}, z_X^{(r)}\rangle\big| \geq \lambda b_i^{(t)}\Big) \lesssim e^{-\Omega(\log^{1/4} d)}, \tag{31}$$

and

$$\Pr\left(\left|\left\langle w_i^{(t)}, \tfrac{z_X^{(r)} + z_X^{(s)}}{2}\right\rangle\right| \geq \lambda b_i^{(t)}\right) \lesssim e^{-\Omega(\log^{1/4} d)}. \tag{32}$$

Lemma C.4 (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*

$$\langle w_i^{(t)}, \mathbf{M}_j\rangle^2 \leq O\left(\frac{(b_i^{(t)})^2}{\mathrm{polylog}(d)}\right), \quad \forall j \notin \mathcal{N}_i, \tag{33}$$

and

$$\|w_i^{(t)}\|_2^2 \leq O\left(\frac{d(b_i^{(t)})^2}{\mathrm{polylog}(d)}\right). \tag{34}$$

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

$$\Pr\left[\left|\sum_{j\notin\mathcal{N}_i} \langle w_i^{(t)}, \mathbf{M}_j\rangle \tilde{z}_{n,j}^{(r)} + \langle w_i, \tilde{\xi}_n^{(r)}\rangle\right| \geq \lambda b_i^{(t)}\right] \lesssim e^{-\Omega(\log^2 d)}. \tag{35}$$

D Theorem 3.1

Lemma D.1 (Positive gradient, stage I)**.** *Let $h_{i,t}(\cdot)$ denote the i-th neuron at iteration $t \le T_1$ (so that $b_i^{(t)} = 0$). Then the following hold:*

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

$$\mathbb{E}[h_{i,t}(X_{n'}) \langle \nabla_{w_i} h_{i,t}(X_n), \mathbf{M}_j \rangle] = \frac{1}{L^2} \langle w_i^{(t)}, \mathbf{M}_j \rangle \, \mathbb{E}[\hat{z}_{n',j''} \hat{z}_{n,j}] \tag{36}$$

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

$$\mathbb{E}\big[h_{i,t}(X_{n'}) \langle \nabla_{w_i} h_{i,t}(X_n), \mathbf{M}_j^{\perp} \rangle\big] = 0 \tag{37}$$

Lemma D.2 (Logits near initialization)**.** *Let $w_i \in \mathbb{R}^{d_1}$ for each $i \in [m]$. Suppose*

$$\sum_{i \in [m]} \left\| w_i^{(t)} \right\|_2^2 \le o\big(\tfrac{\tau}{d}\big) \, . \tag{38}$$

Then, with high probability over the randomness of $X_n, X_{n'}$, and $\mathfrak{N}$, it holds that

$$\left| \ell'_{p,t}(X_n, \mathfrak{B}) - \tfrac{1}{|\mathfrak{B}|} \right| \cdot \left| \ell'_{s,t}(X_n, \mathfrak{B}) - \tfrac{1}{|\mathfrak{B}|} \right| \le \widetilde{\mathcal{O}} \left(\frac{\sum_{i \in [m]} \left\| w_i^{(t)} \right\|_2^2}{\tau |\mathfrak{B}|} \right) \tag{39}$$

Recall that

$$T_1 = \Theta\Big(\frac{d \log d}{\eta \log \log d} \Big) \tag{40}$$

is defined as the iteration when

$$\left\| w_i^{(t)} \right\|_2^2 \ \ge \ (1 + \epsilon_{\min} C_z \log d)^2 \left\| w_i^{(0)} \right\|_2^2, \quad \forall i \in [m], \tag{41}$$

and such a T_1 is indeed of order $\Theta\Big(\frac{d \log d}{\eta \log \log d}\Big)$.

The gradient descent update for the projection of $w_i^{(t)}$ onto $\mathbf{M}_j$ can be written as

$$\begin{aligned} \langle w_i^{(t+1)}, \mathbf{M}_j \rangle &= \langle w_i^{(t)}, \mathbf{M}_j \rangle - \eta \langle \nabla_{w_i} \mathbf{Obj}(f_t), \mathbf{M}_j \rangle \ \ \pm \frac{\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)} \\ &= (1 - \eta\lambda) \langle w_i^{(t)}, \mathbf{M}_j \rangle + \eta \, \mathbb{E}_{X_n, X_{n'}} \big[(1 - \ell'_{p,t}(X_n, \mathfrak{B})) \cdot h_{i,t}(X_{n'}) \langle \nabla_{w_i} h_i(X_n), \mathbf{M}_j \rangle \big] \\ &\quad - \eta \sum_{X_{n,s} \in \mathfrak{N}} \mathbb{E}\big[\ell'_{s,t}(X_n, \mathfrak{B}) \, h_{i,t}(X_{n,s}) \langle \nabla_{w_i} h(X_n), \mathbf{M}_j \rangle \big] \ \ \pm \frac{\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)} \end{aligned} \tag{42}$$

For the positive term: we can use Lemma D.1 and Lemma D.2 to obtain that:

$$\mathbb{E}\big[(1 - \ell'_{p,t}(X_n, \mathfrak{B})) \cdot h_{i,t}(X_{n'}) \langle \nabla_{w_i} h_{i,t}(X_n), \mathbf{M}_j \rangle \big] = \frac{1}{L^2} \langle w_i^{(t)}, \mathbf{M}_j \rangle \, \mathbb{E}[\hat{z}_{n',j} \hat{z}_{n,j}] \tag{43}$$

For the negative term: Here, the bound needs to be verified because Lemma D.2.

$$\begin{aligned} \mathbb{E}\left[\sum_{X \in \mathfrak{N}} \ell'_{s,t} \, h_{i,t}(X) \langle \nabla_{w_i} h(X_n), \mathbf{M}_j \rangle \right] &\overset{(1)}{=} \sum_{X \in \mathfrak{N}} \mathbb{E}\Big[\Big(\ell'_{s,t} - \tfrac{1}{|\mathfrak{B}|} \Big) h_{i,t}(X) \langle \nabla_{w_i} h(X_n), \mathbf{M}_j \rangle \Big] \\ &\overset{(2)}{\le} \sum_{X \in \mathfrak{N}} \mathbb{E}\Big[\Big| \ell'_{s,t} - \tfrac{1}{|\mathfrak{B}|} \Big| \cdot |h_{i,t}(X)| \cdot \big| \langle \nabla_{w_i} h(X_n), \mathbf{M}_j \rangle \big| \Big] \\ &\overset{(3)}{\le} \widetilde{O}\left(\frac{\sum_{i \in [m]} \|w_i^{(t)}\|_2^2}{\tau d} \cdot \|w_i^{(t)}\|_2 \right) \end{aligned} \tag{44}$$

Putting all the above calculations together, we have

$$\begin{aligned}\langle w_i^{(t+1)}, \mathbf{M}_j \rangle = &\left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log\log d}{d}\right) \langle w_i^{(t)}, \mathbf{M}_j \rangle \\ &\pm \widetilde{O}\left(\frac{\eta \sum_{i\in[m]} \|w_i^{(t)}\|_2^2}{\tau d} \cdot \|w_i^{(t)}\|_2\right) \pm \widetilde{O}\left(\frac{\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)}\right)\end{aligned} \tag{45}$$

Prior to the induction step, we establish, by a similar method, the stochastic gradient descent update of w_i along the dense feature direction $\mathbf{M}_j^\perp$. Specifically, we obtain the following update equation:

$$\begin{aligned}\langle w_i^{(t+1)}, \mathbf{M}_j^\perp \rangle &= \langle w_i^{(t)}, \mathbf{M}_j^\perp \rangle - \eta \langle \nabla_{w_i} \mathbf{Obj}(f_t), \mathbf{M}_j^\perp \rangle \\ &= (1-\eta\lambda)\langle w_i^{(t)}, \mathbf{M}_j^\perp \rangle + \eta\, \mathbb{E}\big[(1-\ell'_{p,t})\, h_{i,t}(x_p^{++}) \langle \nabla_{w_i} h_{i,t}(x_p^+), \mathbf{M}_j^\perp \rangle\big] \\ &\quad - \eta \sum_{x_{n,s}\in\mathfrak{N}} \mathbb{E}\big[\ell'_{s,t}\, h_{i,t}(x_{n,s}) \langle \nabla_{w_i} h(x_p^+), \mathbf{M}_j^\perp \rangle\big] + \frac{\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)} \\ &= (1-\eta\lambda)\langle w_i^{(t)}, \mathbf{M}_j^\perp \rangle \pm \widetilde{O}\left(\frac{\eta \sum_{i\in[m]} \|w_i^{(t)}\|_2^2}{\tau d} \cdot \|w_i^{(t)}\|_2\right) \pm \widetilde{O}\left(\frac{\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)}\right)\end{aligned} \tag{46}$$

Proof of Theorem 3.1. For $j \in [d]$ and $i \in [m]$, at iteration T_1 the following bounds hold:

(a) Lower bound:

$$|\langle w_i^{(T_1)}, \mathbf{M}_j \rangle| \geq |\langle w_i^{(0)}, \mathbf{M}_j \rangle| \left(1 - \eta\lambda + \epsilon_j \frac{\eta C_z \log\log d}{d}\right)^{T_1} - \widetilde{O}\left(\frac{\eta T_1 \|w_i^{(T_1)}\|_2}{d_1}\right) \tag{47}$$

(b) Upper bound:

$$|\langle w_i^{(T_1)}, \mathbf{M}_j \rangle| \leq |\langle w_i^{(0)}, \mathbf{M}_j \rangle| \left(1 + \epsilon_j \frac{\eta C_z \log\log d}{d} + \widetilde{O}\Big(\frac{\eta}{d^2}\Big)\right)^{T_1} + \widetilde{O}\left(\frac{\eta T_1 \|w_i^{(T_1)}\|_2}{d_1}\right) \tag{48}$$

(c) Orthogonal component:

$$|\langle w_i^{(T_1)}, \mathbf{M}_j^\perp \rangle| \leq |\langle w_i^{(0)}, \mathbf{M}_j^\perp \rangle| + O(T_1\eta) \cdot \max_{t\leq T_1} O\left(\frac{\|w_i^{(t)}\|_2}{d_1}\right) \tag{49}$$

□

The proof follows by iterating the gradient descent update for w_i along the signal direction $\mathbf{M}_j$ and its orthogonal complement, while controlling the error terms at each step. By substituting T_1 into the recurrence, the bounds in 3.1follow directly.

E Theorem 3.2

In this part, we analyze how each feature $\mathcal{M}_j$ may be captured by certain subsets of neurons, a process that is influenced by the stochastic nature of initialization.

Lemma E.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 $i \in \mathcal{M}_j^\star$, then

$$\left|\langle w_i^{(t)}, \mathbf{M}_j \rangle\right| \geq \sqrt{1+\gamma c_0}\, b_i^{(t)} \tag{50}$$

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

$$\left|\langle w_i^{(t)}, \mathbf{M}_j \rangle\right| \leq \sqrt{1-\gamma c_0}\, b_i^{(t)} \tag{51}$$

and furthermore,

$$\left|\langle w_i^{(t)}, \mathbf{M}_j \rangle\right| \leq \widetilde{\mathcal{O}}\left(\frac{\|w_i^{(t)}\|_2}{\sqrt{d}}\right) \tag{52}$$

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

$$\langle w_i^{(t)}, \mathbf{M}_j \rangle^2 \;\geq\; \frac{(b_i^{(t)})^2}{\sqrt{\log d}} \tag{53}$$

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

$$\left|\langle w_i^{(t)}, \mathbf{M}_j^\perp \rangle\right| \leq \widetilde{\mathcal{O}}\left(\frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}}\right) \tag{54}$$

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

$$\|w_i^{(t)}\|_2^2 \leq \frac{d(b_i^{(t)})^2}{\log d} \tag{55}$$

Definition E.1 (Notations)**.** *For simpler presentation, we define the following notations: given $z_X = \frac{1}{L}(\mathbf{M}\tilde{z}_n + \tilde{\xi}_n) \sim \mathcal{D}_{z_X}$, $z_Y = \frac{1}{L}(\mathbf{M}\tilde{z}_{n'} + \tilde{\xi}_{n'}) \sim \mathcal{D}_{z_Y}$, we let (for each $j \in [d]$):*

$$z_X^{\setminus j} := \frac{1}{L}\left(\sum_{\substack{j' \neq j \\ j' \in [d]}} \mathbf{M}_{j'} \tilde{z}_{n,j'} + \tilde{\xi}_n\right), \quad z_Y^{\setminus j} := \frac{1}{L}\left(\sum_{\substack{j' \neq j \\ j' \in [d]}} \mathbf{M}_{j'} \tilde{z}_{n',j'} + \tilde{\xi}_{n'}\right) \tag{56}$$

$$S_{i,t}^{(r)\setminus j} := \langle w_i^{(t)}, z_X^{(r)\setminus j} \rangle, \quad S_{i,t}^{(s)\setminus j} := \langle w_i^{(t)}, z_Y^{(s)\setminus j} \rangle \tag{57}$$

$$S_{i,t}^{(r,s)\setminus j} := \tfrac{1}{2}\Big(S_{i,t}^{(r)\setminus j} + S_{i,t}^{(s)\setminus j}\Big), \quad \bar{S}_{i,t}^{(r,s)\setminus j} := \tfrac{1}{2}\Big(S_{i,t}^{(s)\setminus j} - S_{i,t}^{(r)\setminus j}\Big) \tag{58}$$

$$\alpha_{i,j}^{(t)} := \langle w_i^{(t)}, \mathbf{M}_j \rangle, \quad \bar{\alpha}_{i,j}^{(r,s)(t)} := \left\langle w_i^{(t)}, \frac{\tilde{z}_{n',j}^{(s)} - \tilde{z}_{n,j}^{(r)}}{\tilde{z}_{n,j}^{(r)} + \tilde{z}_{n',j}^{(s)}} \mathbf{M}_j \right\rangle \tag{59}$$

Whenever the neuron index $i \in [m]$ is clear from the context, we drop the subscript i and the time index t for notational simplicity.

Lemma E.2 (Gradient for sparse features)**.** *Suppose E.1 holds at iteration $t \geq 0$. For $j \in [d]$, we denote events*

$$\begin{aligned} A_1 &:= \Big\{ S_{i,t}^{\setminus j} \geq b_i^{(t)} - \alpha_{i,j}^{(t)} C_{\tilde{z}} \Big\}, \\ A_2 &:= \Big\{ \bar{S}_{i,t}^{\setminus j} \geq b_i^{(t)} - \bar{\alpha}_{i,j}^{(t)} C_{\tilde{z}} \Big\}, \\ A_3 &:= \Big\{ \Big| \bar{S}_{i,t}^{\setminus j} + \bar{\alpha}_{i,j}^{(t)} C_{\tilde{z}} \Big| \geq \tfrac{1}{2} \Big(\alpha_{i,j}^{(t)} C_{\tilde{z}} - b_i^{(t)} \Big) \Big\}, \\ A_4 &:= \Big\{ S_{i,t}^{\setminus j} \geq \tfrac{1}{2} \Big(\alpha_{i,j}^{(t)} C_{\tilde{z}} - b_i^{(t)} \Big) \Big\}; \end{aligned} \tag{60}$$

and quantities L_1, L_2, L_3, L_4 as

$$L_1 := \sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t}^{\setminus j}|^2 (\mathbf{1}_{A_1} + \mathbf{1}_{A_2})]}{\mathbb{E}[\langle w_i^{(t)}, \tilde{\xi} \rangle^2]}}, \quad L_2 := \Pr(A_1), \quad L_3 := \sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t}^{\setminus j}|^2 (\mathbf{1}_{A_3} + \mathbf{1}_{A_4})]}{\mathbb{E}[\langle w_i^{(t)}, \tilde{\xi} \rangle^2]}}, \quad L_4 := \Pr(A_3) \tag{61}$$

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)*

$$\begin{aligned} \mathbb{E}\bigg[h_i(X_{n'}) \sum_{r=1}^{L} \mathbf{1}_{|\langle w_i, z_X^{(r)} \rangle| \geq b_i} \tilde{z}_{n,j}^{(r)} \bigg] \leq{}& \frac{1}{L} \alpha_{i,j}^{(t)} \cdot \mathbb{E}\bigg[\sum_{s=1}^{L} \tilde{z}_{n',j}^{(s)} \sum_{r=1}^{L} \tilde{z}_{n,j}^{(r)} \mathbf{1}_{|\langle w_i^{(t)}, \frac{z_X + z_Y}{2} \rangle| \geq b_i + |\langle w_i^{(t)}, z_X - \frac{z_X + z_Y}{2} \rangle|} \bigg] \\ &\pm \Big(\alpha_{i,j}^{(t)} + O\Big(\sqrt{\mathbb{E}|\bar{\alpha}_{i,j}^{(t)}|^2} \Big) \Big) \cdot \mathbb{E}\bigg[\sum_{s=1}^{L} \sum_{r=1}^{L} \Big| \tfrac{\tilde{z}_{n,j}^{(r)} + \tilde{z}_{n',j}^{(s)}}{2} \Big| |\tilde{z}_{n,j}^{(r)}| \bigg] \cdot O(L_1 + L_2) \end{aligned} \tag{62}$$

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

$$\begin{aligned} \mathbb{E}\bigg[h_i(X_{n'}) \sum_{r=1}^{L} \mathbf{1}_{|\langle w_i, z_X^{(r)} \rangle| \geq b_i} \tilde{z}_{n,j}^{(r)} \bigg] \leq{}& \frac{1}{L} \Big(\alpha_{i,j}^{(t)} - b_i^{(t)} \Big) \cdot \mathbb{E}\bigg[\sum_{s=1}^{L} \tilde{z}_{n',j}^{(s)} \sum_{r=1}^{L} \tilde{z}_{n,j}^{(r)} \mathbf{1}_{|\langle w_i^{(t)}, \frac{z_X + z_Y}{2} \rangle| \geq b_i + |\langle w_i^{(t)}, z_X - \frac{z_X + z_Y}{2} \rangle|} \bigg] \\ &\pm \Big(\alpha_{i,j}^{(t)} + O\Big(\sqrt{\mathbb{E}|\bar{\alpha}_{i,j}^{(t)}|^2} \Big) \Big) \cdot \mathbb{E}\bigg[\sum_{s=1}^{L} \sum_{r=1}^{L} \Big| \tfrac{\tilde{z}_{n,j}^{(r)} + \tilde{z}_{n',j}^{(s)}}{2} \Big| |\tilde{z}_{n,j}^{(r)}| \bigg] \cdot O(L_3 + L_4) \end{aligned} \tag{63}$$

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

Lemma E.3 (Gradient from dense signals)**.** *Let $i \in [m]$ and $j \in [d]$. Suppose E.1 holds for the current iteration t. Then*

$$\bigg| \mathbb{E}\bigg[h_i(X_{n'}) \sum_{r=1}^{L} \mathbf{1}_{|\langle w_i^{(t)}, z_X^{(r)} \rangle| \geq b_i^{(t)}} \langle \tilde{\xi}_n^{(r)}, \mathbf{M}_j \rangle \bigg] \bigg| \leq \widetilde{\mathcal{O}}\bigg(\frac{\|w_i^{(t)}\|_2}{d^2} \bigg) \cdot \Pr\big(h_{i,t}(X_{n'}) \neq 0 \big) \tag{64}$$

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

$$\bigg| \mathbb{E}\bigg[h_i(X_{n'}) \sum_{r=1}^{L} \mathbf{1}_{|\langle w_i^{(t)}, z_X^{(r)} \rangle| \geq b_i^{(t)}} \langle \tilde{\xi}_n^{(r)}, \mathbf{M}_j^{\perp} \rangle \bigg] \bigg| \leq \widetilde{O}\bigg(\frac{\|w_i^{(t)}\|_2}{d\sqrt{d_1}} \bigg) \cdot \Pr\big(h_{i,t}(X_{n'}) \neq 0 \big) \tag{65}$$

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

$$T_2 = \Theta\bigg(\frac{d \log d}{\epsilon_{\max} \eta \log \log d} \bigg) \tag{66}$$

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

$$\Big\| w_i^{(T_2)} \Big\|_2^2 \geq d \Big\| w_i^{(T_1)} \Big\|_2^2 \tag{67}$$

Now we separately discuss three cases:

(a) When $i \in \mathcal{M}_j^\star$, if $\tilde{z}_{n',j}^{(s)}$ and $\tilde{z}_{n,j}^{(r)} \neq 0$, say $\frac{\tilde{z}_{n',j}^{(s)} + \tilde{z}_{n,j}^{(r)}}{2} = C_{\tilde{z}}^{(r,s)}$, we simply have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{L}\tilde{z}_{n',j}^{(s)}\sum_{r=1}^{L}\tilde{z}_{n,j}^{(r)}\mathbf{1}_{\left|\langle w_i, \frac{z_X^{(r)}+z_Y^{(s)}}{2}\rangle\right| \geq b_i + \left|\langle w_i, z_X^{(r)} - \frac{z_X^{(r)}+z_Y^{(s)}}{2}\rangle\right|}\right] \\
=&\mathbb{E}\left[\sum_{s=1}^{L}\tilde{z}_{n',j}^{(s)}\sum_{r=1}^{L}\tilde{z}_{n,j}^{(r)}\right]\cdot \Pr\left(\left|\langle w_i^{(t)}, \tfrac{z_X^{(r)}+z_Y^{(s)}}{2}\rangle\right| \geq b_i + \left|\langle w_i^{(t)}, z_X^{(r)} - \tfrac{z_X^{(r)}+z_Y^{(s)}}{2}\rangle\right|\right) \\
=&\epsilon_j \frac{L^2 C_z \log\log d}{d}\left(1 - \frac{1}{\mathrm{polylog}(d)}\right).
\end{aligned} \tag{68}$$

For $\mathbf{M}_j$ such that $i \in \mathcal{M}_j^\star$, at iteration $t+1$:

$$\begin{aligned}
\langle w_i^{(t+1)}, \mathbf{M}_j\rangle &= \langle w_i^{(t)}, \mathbf{M}_j\rangle - \eta\langle \nabla_{w_i}\mathbf{Obj}(f_t), \mathbf{M}_j\rangle \pm \frac{\eta\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)} \\
&= \langle w_i^{(t)}, \mathbf{M}_j\rangle(1-\eta\lambda) \pm \frac{\eta\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)} \\
&\quad + \eta\mathbb{E}\left[(1-\ell_{p,t}^t)h_{i,t}(X_{n'})\sum_{r=1}^{L}\mathbf{1}_{|\langle w_i, z_X^{(r)}\rangle|\geq b_i}\langle z_X^{(r)}, \mathbf{M}_j\rangle\right] \\
&\quad - \eta\mathbb{E}\left[\sum_{X\in\mathfrak{N}}\ell_{s,t}^t h_{i,t}(X)\sum_{r=1}^{L}\mathbf{1}_{|\langle w_i, z_X^{(r)}\rangle|\geq b_i}\langle z_X^{(r)}, \mathbf{M}_j\rangle\right] \\
&= \langle w_i^{(t)}, \mathbf{M}_j\rangle(1-\eta\lambda) \pm \frac{\eta\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)} \\
&\quad + \eta\frac{1}{L}\mathbb{E}\left[(1-\ell_{p,t}^t)h_{i,t}(X_{n'})\sum_{r=1}^{L}\mathbf{1}_{|\langle w_i, z_X^{(r)}\rangle|\geq b_i}\left(\tilde{z}_{n,j}^{(r)} + \langle\tilde{\xi}_n^{(r)}, \mathbf{M}_j\rangle\right)\right] \\
&\quad - \eta\frac{1}{L}\mathbb{E}\left[\sum_{X\in\mathfrak{N}}\ell_{s,t}^t h_{i,t}(X)\sum_{r=1}^{L}\mathbf{1}_{|\langle w_i, z_X^{(r)}\rangle|\geq b_i}\left(\tilde{z}_{n,j}^{(r)} + \langle\tilde{\xi}_n^{(r)}, \mathbf{M}_j\rangle\right)\right] \\
&\geq \left(\langle w_i^{(t)}, \mathbf{M}_j\rangle - \mathrm{sign}\big(\langle w_i^{(t)}, \mathbf{M}_j\rangle\big)\cdot b_i^{(t)}\right)\cdot\left(1 + \epsilon_j\frac{\eta C_z\log\log d}{d}\left(1 - \frac{\eta}{\mathrm{polylog}(d)}\right)\right)
\end{aligned} \tag{69}$$

Next we compare this growth to the growth of bias $b_i^{(t+1)}$. Since we raise our bias by

$$b_i^{(t+1)} = \max\left\{b_i^{(t)}\left(1+\frac{\eta}{d}\right),\ b_i^{(t)}\frac{\|w_i^{(t+1)}\|_2}{\|w_i^{(t)}\|_2}\right\} \tag{70}$$

We have to prove

$$\frac{|\langle w_i^{(t+1)}, \mathbf{M}_j\rangle|}{|\langle w_i^{(t)}, \mathbf{M}_j\rangle|} \ \geq\ \frac{\|w_i^{(t+1)}\|_2}{\|w_i^{(t)}\|_2}, \qquad i \in \mathcal{M}_j^\star \tag{71}$$

We argue as follows: from previous calculations we have

$$\begin{aligned}
&\sum_{j'\in[d],\, j'\neq j}\langle w_i^{(t+1)}, \mathbf{M}_{j'}\rangle^2 + \sum_{j'\in[d_1]\setminus[d]}\langle w_i^{(t+1)}, \mathbf{M}_{j'}^\perp\rangle^2 \\
&\leq \sum_{j'\in[d],\, j'\neq j}\langle w_i^{(t)}, \mathbf{M}_{j'}\rangle^2\left(1 + \epsilon_{j'}\frac{O(\eta)}{d\,\mathrm{polylog}(d)}\right)^2 \\
&+ \sum_{j'\in[d_1]\setminus[d]}\langle w_i^{(t)}, \mathbf{M}_{j'}^\perp\rangle^2 + \widetilde{\mathcal{O}}\Big(\frac{\eta}{d}\Big)\, e^{-\Omega(\log^{1/4} d)}\|w_i^{(t)}\|_2^2
\end{aligned} \tag{72}$$

Therefore by adding $\langle w_i^{(t+1)}, \mathbf{M}_j\rangle^2$ to the LHS we have

$$\|w_i^{(t+1)}\|_2^2 \;\le\; \|w_i^{(t)}\|_2^2\left(1+\epsilon_{\max}\frac{O(\eta)}{d\cdot\mathrm{polylog}(d)}\right)^2+\left(\frac{|\langle w_i^{(t+1)},\mathbf{M}_j\rangle|}{|\langle w_i^{(t)},\mathbf{M}_j\rangle|}-\frac{O(\eta)}{d\cdot\mathrm{polylog}(d)}\right)|\langle w_i^{(t)},\mathbf{M}_j\rangle|^2 \tag{73}$$

which implies

$$\frac{\|w_i^{(t+1)}\|_2^2}{\|w_i^{(t)}\|_2^2} \;\le\; \left(1+\epsilon_{\max}\frac{O(\eta)}{d\cdot\mathrm{polylog}(d)}\right)^2+\left(\frac{|\langle w_i^{(t+1)},\mathbf{M}_j\rangle|}{|\langle w_i^{(t)},\mathbf{M}_j\rangle|}\right)\frac{|\langle w_i^{(t)},\mathbf{M}_j\rangle|^2}{\|w_i^{(t)}\|_2^2} \tag{74}$$

Therefore,

$$\frac{|\langle w_i^{(t+1)},\mathbf{M}_j\rangle|}{|\langle w_i^{(t)},\mathbf{M}_j\rangle|} \;\ge\; \frac{\|w_i^{(t+1)}\|_2}{\|w_i^{(t)}\|_2} \tag{75}$$

as desired.

(b) When $i\notin\mathcal{M}_j$, we can similarly obtain that

$$\begin{aligned}\mathbb{E}\left[\sum_{s=1}^{L}\tilde{z}_{n',j}^{(s)}\sum_{r=1}^{L}\tilde{z}_{n,j}^{(r)}\mathbf{1}_{\left|\langle w_i,\frac{z_X^{(r)}+z_Y^{(s)}}{2}\rangle\right|\ge b_i+\left|\langle w_i,z_X^{(r)}-\frac{z_X^{(r)}+z_Y^{(s)}}{2}\rangle\right|}\right] &\le \epsilon_j\frac{L^2C_z\log\log d}{d}\left(\frac{1}{\mathrm{polylog}(d)}\right)\\ &= O\left(\epsilon_j\frac{L^2}{d\cdot\mathrm{polylog}(d)}\right)\end{aligned} \tag{76}$$

And similarly we can compute the gradient descent dynamics as follows: For $j\in[d]$ such that $|\langle w_i^{(t)},\mathbf{M}_j\rangle|\ge\frac{\|w_i^{(t)}\|_2 d}{\sqrt{d_1}}$, we have (assume here $\langle w_i^{(t)},\mathbf{M}_j\rangle>0$, the opposite is similar)

$$\begin{aligned}\langle w_i^{(t+1)},\mathbf{M}_j\rangle &= \langle w_i^{(t)},\mathbf{M}_j\rangle-\eta\langle\nabla_{w_i}\mathbf{Obj}(f_t),\mathbf{M}_j\rangle+\frac{\eta\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)}\\ &\le \langle w_i^{(t)},\mathbf{M}_j\rangle\left(1-\eta\lambda+\epsilon_j\frac{O(\eta)}{d\cdot\mathrm{polylog}(d)}\right)\\ &\quad\pm O\left(\frac{\eta\sum_{i'\in[m]}\|w_{i'}^{(t)}\|_2^2\|w_i^{(t)}\|_2}{d\tau}\right)\pm\widetilde{O}\left(\frac{\eta\|w_i^{(t)}\|_2}{d^2}\right)\\ &\le \langle w_i^{(t)},\mathbf{M}_j\rangle\left(1+\epsilon_j\frac{O(\eta)}{d\cdot\mathrm{polylog}(d)}\right)+\widetilde{O}\left(\frac{\eta\|w_i^{(t)}\|_2}{d^2}\right)\end{aligned} \tag{77}$$

It is also worth noting that similar calculations also lead to a lower bound:

$$|\langle w_i^{(t+1)},\mathbf{M}_j\rangle|\ge|\langle w_i^{(t)},\mathbf{M}_j\rangle|(1-\eta\lambda)-\widetilde{\mathcal{O}}\left(\eta\frac{\|w_i^{(t)}\|_2}{d^2}\right) \tag{78}$$

(c) Next we consider the learning dynamics for the dense features.

We can use E.3 to calculate its dynamics by

$$\begin{aligned}
\langle w_i^{(t+1)}, \mathbf{M}_j^\perp \rangle &= \langle w_i^{(t)}, \mathbf{M}_j^\perp \rangle (1-\eta\lambda) \pm \frac{\eta \|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)} \\
&\quad + \eta \mathbb{E}\left[(1-\ell'_{p,t}) h_{i,t}(X_{n'}) \sum_{r=1}^{L} \mathbf{1}_{|\langle w_i, z_X^{(r)} \rangle| \geq b_i} \langle z_X^{(r)}, \mathbf{M}_j^\perp \rangle \right] \\
&\quad - \eta \sum_{X \in \mathfrak{N}} \mathbb{E}\left[\ell'_{s,t} h_{i,t}(X) \sum_{r=1}^{L} \mathbf{1}_{|\langle w_i, z_X^{(r)} \rangle| \geq b_i} \langle z_X^{(r)}, \mathbf{M}_j^\perp \rangle \right] \\
&= \langle w_i^{(t)}, \mathbf{M}_j^\perp \rangle (1-\eta\lambda) \pm \frac{\eta \|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)} \\
&\quad + \eta \mathbb{E}\left[(1-\ell'_{p,t}) h_{i,t}(X_{n'}) \sum_{r=1}^{L} \mathbf{1}_{|\langle w_i, z_X^{(r)} \rangle| \geq b_i} \langle \tilde{\xi}_n^{(r)}, \mathbf{M}_j^\perp \rangle \right] \\
&\quad - \eta \sum_{X \in \mathfrak{N}} \mathbb{E}\left[\ell'_{s,t} h_{i,t}(X) \sum_{r=1}^{L} \mathbf{1}_{|\langle w_i, z_X^{(r)} \rangle| \geq b_i} \langle \tilde{\xi}_n^{(r)}, \mathbf{M}_j^\perp \rangle \right] \\
&= \langle w_i^{(t)}, \mathbf{M}_j^\perp \rangle (1-\eta\lambda) + \widetilde{\mathcal{O}}\left(\frac{\eta \|w_i^{(t)}\|_2}{d\sqrt{d_1}} \right) \cdot \Pr(h_{i,t}(X_{n'}) \neq 0) \\
&\leq \langle w_i^{(t)}, \mathbf{M}_j^\perp \rangle + O\left(\frac{\eta \|w_i^{(t)}\|_2}{d\sqrt{d_1}} e^{-\Omega(\log^{1/4} d)} \right)
\end{aligned} \tag{79}$$

In the proof above, we have depended on the crucial assumption that $T_2 := \min\left\{ t \in \mathbb{N} : \exists i \in [m] \text{ s.t. } \|w_i^{(t)}\|_2^2 \geq d \|w_i^{(T_1)}\|_2^2 \right\}$ is of order $\Theta\left(\frac{d \log d}{\epsilon_{\max} \eta \log\log d} \right)$. Now we verify it as follows. If $i \in \mathcal{M}_j^*$ for some $j \in [d]$ (which also means $j' \notin \mathcal{N}_i$ for $j' \neq j$), we have

$$\begin{aligned}
|\langle w_i^{(t)}, \mathbf{M}_j \rangle| \;&\geq\; |\langle w_i^{(T_1)}, \mathbf{M}_j \rangle| \left(1 + \Omega\left(\epsilon_j \frac{\eta \log\log d}{d} \right) \right)^{t - T_1} \\
&\geq\; d\sqrt{\frac{2\log d}{d}} \, \|w_i^{(T_1)}\|_2 \quad \text{for some } \; t = O\left(\frac{d \log d}{\epsilon_j \eta \log\log d} \right)
\end{aligned} \tag{80}$$

Thus for some $t = O\left(\frac{d \log d}{\epsilon_j \eta \log\log d} \right)$, we have $|\langle w_i^{(t)}, \mathbf{M}_j \rangle|^2 \geq d \|w_i^{(T_1)}\|_2^2$, which proves that $T_2 \leq O\left(\frac{d \log d}{\epsilon_{\max} \eta \log\log d} \right)$.

Conversely, we also have for all $t \leq O\left(\frac{d \log d}{\epsilon_j \eta \log\log d} \right)$

$$\begin{aligned}
&\sum_{j' \in [d]: j' \neq j} \langle w_i^{(t)}, \mathbf{M}_{j'} \rangle^2 + \sum_{j' \in [d_1] \setminus [d]} \langle w_i^{(t)}, \mathbf{M}_{j'}^\perp \rangle^2 \\
&\leq\; \|w_i^{(T_1)}\|_2^2 \left(1 + \epsilon_{j'} \frac{O(\eta)}{d \, \mathrm{polylog}(d)} \right)^{t - T_1} + \max_{t' \leq t} O\left(\frac{\eta(t - T_1)}{d} \right) e^{-\Omega(\log^{1/4} d)} \|w_i^{(t')}\|_2^2 \\
&\leq\; o\Big(d \, \|w_i^{(T_1)}\|_2^2 \Big)
\end{aligned} \tag{81}$$

Except for the principal direction $\mathbf{M}_j$ (i.e., the alignment direction of neuron i), the total growth of squared weights along all other directions remains far below the target scale $d \cdot \left\| w_i^{(T_1)} \right\|_2^2$. And also

$$\begin{aligned}
|\langle w_i^{(t)}, \mathbf{M}_j \rangle| \;&\leq\; |\langle w_i^{(T_1)}, \mathbf{M}_j \rangle| \left(1 + \epsilon_j \frac{C_z \eta \log\log d}{d} \left(1 - \frac{1}{\mathrm{polylog}(d)} \right) \right)^{t - T_1} \\
&\leq\; O\left(\sqrt{\tfrac{\log d}{d}} \, \|w_i^{(T_1)}\|_2 \right) \left(1 + \epsilon_j \frac{C_z \eta \log\log d}{d} \left(1 - \frac{1}{\mathrm{polylog}(d)} \right) \right)^{t - T_1}
\end{aligned} \tag{82}$$

Therefore we at least need $\frac{d\log\left(\Omega\left(\sqrt{d}\sqrt{\frac{d}{\log d}}\right)\right)}{\epsilon_{\max}\eta C_z\log\log d}(1-o(1))$ iteration to let any neuron $i\in[m]$ reach $\|w_i^{(t)}\|_2^2\geq d\|w_i^{(T_1)}\|_2$, which proves that $T_2=\Theta\left(\frac{d\log d}{\epsilon_{\max}\eta\log\log d}\right)$

Proof of Theorem 3.2. **When all $\|w_i^{(t)}\|_2\leq 2\|w_i^{(T_1)}\|_2$:** The iteration complexity for a neuron $i\in[m]$ to reach $\|w_i^{(t)}\|_2\geq 2\|w_i^{(T_1)}\|_2$ is no smaller than

$$T'_{i,1}:=\max\left\{\Omega\left(\frac{d\log d}{\eta\log\log d}\right),T_2\right\}. \tag{83}$$

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

$$\begin{aligned}\sum_{j\in[d_1]\setminus[d]}\langle w_i^{(T'_{i,1})},\mathbf{M}_j^\perp\rangle^2 &\leq \sum_{j\in[d_1]\setminus[d]}\langle w_i^{(T_1)},\mathbf{M}_j^\perp\rangle^2+O\bigg(\frac{\eta(T'_{i,1}-T_1)}{d}\bigg)e^{-\Omega(\log^{1/4}d)}\max_{t'\in[T_1,T'_{i,1}]}\|w_i^{(t')}\|_2^2\\ &\leq (1+o(1))\left\|\mathbf{M}^\perp(\mathbf{M}^\perp)^\top w_i^{(T_1)}\right\|_2^2\end{aligned} \tag{84}$$

If $i\in\mathcal{M}_j^*$, there exist $t\leq T_2$ such that $\|w_i^{(t)}\|_2\geq 2\|w_i^{(T_2)}\|_2$, we have

$$\begin{aligned}|\langle w_i^{(T'_{i,1})},\mathbf{M}_j\rangle|^2 &\geq \|w_i^{(T'_{i,1})}\|_2^2-\sum_{j\in[d],\,j\notin\mathcal{N}_i}\langle w_i^{(T'_{i,1})},\mathbf{M}_j\rangle^2-\sum_{j\in[d_1]\setminus[d]}\langle w_i^{(T'_{i,1})},\mathbf{M}_j^\perp\rangle^2\\ &\geq 2\|w_i^{(T_1)}\|_2^2-(1+o(1))\|w_i^{(T_1)}\|_2^2\\ &\geq (1-o(1))\|w_i^{(T_1)}\|_2^2\end{aligned} \tag{85}$$

which proves the claim.

At this substage, we have: 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:

$$\begin{aligned}|\langle w_i^{(t+1)},\mathbf{M}_j\rangle| &\geq |\langle w_i^{(t)},\mathbf{M}_j\rangle|\left(1+\Omega\left(\epsilon_j\frac{\eta\log\log d}{d}\right)\right)\\ &\geq \|w_i^{(t)}\|_2\left(1+\Omega\left(\epsilon_j\frac{\eta\log\log d}{d}\right)\right)\end{aligned} \tag{86}$$

$$\sum_{j'\in[d],\,j'\neq j}\langle w_i^{(t+1)},\mathbf{M}_{j'}\rangle^2\leq\sum_{j'\in[d],\,j'\neq j}\langle w_i^{(t)},\mathbf{M}_{j'}\rangle^2\left(1+\epsilon_j\frac{O(\eta)}{d\,\mathrm{polylog}(d)}\right)^2 \tag{87}$$

$$\sum_{j\in[d_1]\setminus[d]}\langle w_i^{(t+1)},\mathbf{M}_j^\perp\rangle^2\leq\sum_{j\in[d_1]\setminus[d]}\langle w_i^{(t)},\mathbf{M}_j^\perp\rangle^2\left(1+\frac{O(\eta)}{d\,\mathrm{polylog}(d)}\right)^2 \tag{88}$$

which implies

$$\begin{aligned}|\langle w_i^{(t+1)},\mathbf{M}_j\rangle| &\geq |\langle w_i^{(t)},\mathbf{M}_j\rangle|\cdot\frac{\|w_i^{(t+1)}\|_2}{\|w_i^{(t)}\|_2}\\ &\geq (1-o(1))\,\|w_i^{(t+1)}\|_2\end{aligned} \tag{89}$$

Theorem 3.2 (6) is proved. As for Theorem 3.2 (7), we can revisit case (c) from the three situations discussed earlier and then proceed by iteration. □

F Theorem 3.3

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

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

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

$$\left|\langle w_i^{(t)}, \mathbf{M}_j \rangle\right| \;\geq\; \Omega(1)\, \|w_i^{(t)}\|_2 \tag{90}$$

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

$$\|w_i^{(t)}\|_2 \;\leq\; O(1) \tag{91}$$

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

$$\widehat{\mathfrak{F}}_j^{(t)} := \sum_{i \in \mathcal{M}_j} \langle w_i^{(t)}, \mathbf{M}_j \rangle^2 \;\leq\; O((\frac{\epsilon_j}{\epsilon_{\max}})^2 \tau \log^3 d) \tag{92}$$

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

$$\left|\langle w_i^{(t)}, \mathbf{M}_j \rangle\right| \;\geq\; C \max_{i' \in \mathcal{M}_j} \left|\langle w_{i'}^{(t)}, \mathbf{M}_j \rangle\right| \tag{93}$$

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

$$\left|\langle w_i^{(t)}, \mathbf{M}_j \rangle\right| \;\leq\; O\Big(\frac{\epsilon_j}{\epsilon_{\max}} \frac{1}{\sqrt{d}\,\Xi_2^5}\Big)\, \|w_i^{(t)}\|_2 \tag{94}$$

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

$$\left|\langle w_i^{(t)}, \mathbf{M}_j^\perp \rangle\right| \;\leq\; O\Big(\frac{1}{\sqrt{d_1}\,\Xi_2^5}\Big)\, \|w_i^{(t)}\|_2 \tag{95}$$

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

$$b_i^{(t)} \;\geq\; \frac{\mathrm{polylog}(d)}{\sqrt{d}}\, \|w_i^{(t)}\|_2 \tag{96}$$

Definition F.1 (Optimal Learner)**.** *We define a learner network that we deem as the "optimal" feature map for this task. Let $\kappa > 0$, we define $\theta^\star := \{\theta_i^\star\}_{i \in [m]}$ as follows:*

$$\theta_i^\star = \begin{cases} \dfrac{\sqrt{\tau}\,\kappa}{|\mathcal{M}_j^\star|}\, \mathbf{M}_j \cdot \mathrm{sign}\big(\langle w_i^{(T_2)}, \mathbf{M}_j \rangle\big), & \text{if } i \in \mathcal{M}_j^\star, \\ 0, & \text{if } i \notin \bigcup_{j \in [d]} \mathcal{M}_j^\star \end{cases} \tag{97}$$

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

$$f_{t,\theta,i}(X_n) \;=\; \sum_{r=1}^{L} \Big[\big(\langle \theta_i, z_X^{(r)} \rangle - b_i\big) \mathbf{1}_{\langle w_i^{(t)}, z_X^{(r)} \rangle \geq b_i} - \big(-\langle \theta_i, z_X^{(r)} \rangle - b_i\big) \mathbf{1}_{-\langle w_i^{(t)}, z_X^{(r)} \rangle \geq b_i} \Big]. \tag{98}$$

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

$$f_{t,\theta}(\cdot) \;=\; \big(f_{t,\theta,1}(\cdot), \ldots, f_{t,\theta,m}(\cdot)\big)^\top \tag{99}$$

Lemma F.2 (Optimality)**.** *Let $\{\theta_i^\star\}_{i \in [m]}$ and $f_{t,\theta}$ be defined as in Definition F.1. When Lemma F.1, define the pseudo loss function*

$$\widetilde{\mathcal{L}}(f_{t,\theta^\star}, f_t) := \mathbb{E}\left[-\tau \log \left(\frac{e^{\langle f_{t,\theta^\star}(X_n), f_t(X_{n'}) \rangle / \tau}}{\sum_{x \in \mathfrak{B}} e^{\langle f_{t,\theta^\star}(X_n), f_t(x) \rangle / \tau}} \right) \right] \tag{100}$$

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

$$\sum_{i \in \mathcal{M}_j^\star} |\langle w_i^{(t)}, \mathbf{M}_j \rangle| \;\geq\; \Omega\left(\frac{\sqrt{\tau}}{\Xi_2}\right), \tag{101}$$

we obtain the following loss guarantee:

$$\widetilde{\mathcal{L}}(f_{t,\theta^\star}, f_t) \;\leq\; O\left(\tfrac{1}{\log d}\right) \tag{102}$$

Proof of Theorem 3.3. We start with the proof of convergence Theorem 3.3 (8).

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

$$w^{(t+1)} = w^{(t)} - \nabla_w \mathrm{Obj}(f_t) + \tfrac{1}{\mathrm{poly}(d_1)}, \tag{103}$$

we have

$$\begin{aligned}\eta \langle \nabla_w \mathrm{Obj}(f_t), w^{(t)} - \theta^\star \rangle &= \tfrac{\eta^2}{2} \|\nabla_w \mathrm{Obj}(f_t)\|_F^2 + \tfrac{1}{2}\|w^{(t)} - \theta^\star\|_F^2 - \tfrac{1}{2}\|w^{(t+1)} - \theta^\star\|_F^2 + \tfrac{\eta^2}{\mathrm{poly}(d_1)} \\ &\leq \eta^2 \,\mathrm{poly}(d) + \tfrac{1}{2}\|w^{(t)} - \theta^\star\|_F^2 - \tfrac{1}{2}\|w^{(t+1)} - \theta^\star\|_F^2 + \tfrac{\eta^2}{\mathrm{poly}(d_1)}\end{aligned} \tag{104}$$

The proof of the above equation is as follows:

$$\langle x, y \rangle = \tfrac{1}{2}\left(\|x\|^2 + \|y\|^2 - \|x - y\|^2\right) \tag{105}$$

Let $x = a - c,\ y = a - b$, and substitute into the above equation.

$$\langle a - c, a - b \rangle = \tfrac{1}{2}\left(\|a - c\|^2 + \|a - b\|^2 - \|b - c\|^2\right) \tag{106}$$

Here we substitute the following three quantities into the three point identity:

$$a = w^{(t)}, \qquad b = \theta^\star, \qquad c = w^{(t+1)} = w^{(t)} - \eta \nabla_w \mathrm{Obj}(f_t) \pm \tfrac{\eta}{\mathrm{poly}(d_1)} \tag{107}$$

Thus, the original equation is proved. As for the inequality,

$$\|\nabla_w \mathrm{Obj}(f_t)\|_F^2 = \sum_{i=1}^m \|\nabla_{w_i} \mathrm{Obj}(f_t)\|^2 \tag{108}$$

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

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

$$\begin{aligned}\widetilde{\mathrm{Obj}}_t(\theta) &:= \widetilde{\mathcal{L}}(f_{t,\theta}, f_t) + \tfrac{\lambda}{2} \sum_{i \in [m]} \|\theta_i\|_2^2 \\ &= \mathbb{E}\left[-\tau \log\left(\frac{e^{\langle f_{t,\theta}(X_n), f_t(X_{n'}) \rangle / \tau}}{\sum_{x \in \mathfrak{B}} e^{\langle f_{t,\theta}(X_n), f_t(x) \rangle / \tau}}\right)\right] + \tfrac{\lambda}{2} \sum_{i \in [m]} \|\theta_i\|_2^2\end{aligned} \tag{109}$$

Which is a convex function over θ since it is linear in θ (for a fixed f_t, we can consider $\widetilde{\mathcal{L}}(f_{t,\theta}, f_t)$ 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

$$\widetilde{\mathrm{Obj}}_t(w^{(t)}) = \mathrm{Obj}(f_t), \tag{110}$$

and

$$\nabla_{\theta_i} \widetilde{\mathrm{Obj}}_t(w_i^{(t)}) = \nabla_{w_i} \mathrm{Obj}(f_t) \tag{111}$$

Thus we have

$$\begin{aligned}
\eta\langle\nabla_w \mathrm{Obj}(f_t), w^{(t)} - \theta^\star\rangle &= \eta\langle\nabla_\theta \widetilde{\mathrm{Obj}}_t(w^{(t)}), w^{(t)} - \theta^\star\rangle \\
&\overset{(1)}{\geq} \widetilde{\mathrm{Obj}}_t(w^{(t)}) - \widetilde{\mathrm{Obj}}_t(\theta^\star) \\
&\geq \widetilde{\mathrm{Obj}}_t(w^{(t)}) - \mathbb{E}\left[-\tau \log\left(\frac{e^{\langle f_{t,\theta^\star}(X_n), f_t(X_{n'})\rangle/\tau}}{\sum_{x\in\mathfrak{B}} e^{\langle f_{t,\theta^\star}(X_n), f_t(x)\rangle/\tau}}\right)\right] - \frac{\lambda}{2}\sum_{i\in[m]}\|\theta_i^\star\|_2^2 \\
&\overset{(2)}{\geq} \widetilde{\mathrm{Obj}}_t(w^{(t)}) - O\Big(\tfrac{1}{\log d}\Big) - \sum_{i\in[m]} O(\lambda\|\theta_i^\star\|_2^2) \\
&\geq \mathrm{Obj}(f_t) - O\Big(\tfrac{1}{\log d}\Big) \qquad (112)
\end{aligned}$$

(1) is because the surrogate objective function $\widetilde{\mathrm{Obj}}_t$ 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 $\sum_{i\in[m]}\lambda\|\theta_i^\star\|_2^2 = \sum_{j\in[d]}\sum_{i\in\mathcal{M}_j^\star}\lambda\|\theta_i^\star\|_2^2 = \sum_{j\in[d]}\sum_{i\in\mathcal{M}_j^\star}\lambda\frac{\tau\kappa^2}{|\mathcal{M}_j^\star|^2} = \sum_{j\in[d]}\lambda\frac{\tau\kappa^2}{|\mathcal{M}_j^\star|} = \frac{\lambda\tau\kappa^2}{|\mathcal{M}_j^\star|}$

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

$$\begin{aligned}
\frac{1}{T}\sum_{t=T_3}^{T_3+T-1}\Big(\mathrm{Obj}(f_t) - O\Big(\tfrac{1}{\log d}\Big)\Big) &\leq \frac{1}{T}\sum_{t=T_3}^{T_3+T-1}\eta\langle\nabla_w \mathrm{Obj}(f_t), w^{(t)} - \theta^\star\rangle \\
&\leq \frac{O(\|w^{(T_3)} - \theta^\star\|_F^2)}{T\eta} \\
&= \frac{O\big(\|w^{(T_3)}\|_F^2 + \|\theta^\star\|_F^2 - 2\,\mathrm{Tr}((w^{(T_3)})^\top\theta^\star)\big)}{T\eta} \\
&\leq \frac{O\big(\|w^{(T_3)}\|_F^2 + \|\theta^\star\|_F^2\big)}{T\eta} \\
&\leq \frac{O\Big(m\|w_i^{(T_3)}\|_2^2\Big)}{T\eta} \\
&\leq O\Big(\tfrac{m\Xi_2}{T\eta}\Big) \qquad (113)
\end{aligned}$$

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

For Theorem 3.3 (9), we have

$$\begin{aligned}
w_i^{(t)} &= \sum_{j\in\mathcal{N}_i,\, j\in[d]}\langle w_i^{(t)}, \mathbf{M}_j\rangle\mathbf{M}_j + \sum_{j\notin\mathcal{N}_i,\, j\in[d]}\langle w_i^{(t)}, \mathbf{M}_j\rangle\mathbf{M}_j + \sum_{j\in[d_1]\setminus[d]}\langle w_i^{(t)}, \mathbf{M}_j^\perp\rangle\mathbf{M}_j^\perp \\
&\leq \sum_{j\in\mathcal{N}_i,\, j\in[d]}\langle w_i^{(t)}, \mathbf{M}_j\rangle\mathbf{M}_j + \sum_{j\notin\mathcal{N}_i,\, j\in[d]} O\left(\frac{\epsilon_j}{\epsilon_{\max}}\frac{\|w_i^{(t)}\|_2}{\sqrt{d}\,\Xi_2^5}\right)\mathbf{M}_j + \sum_{j\in[d_1]\setminus[d]} O\left(\frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}\,\Xi_2^5}\right)\mathbf{M}_j^\perp \\
&= \sum_{j\in\mathcal{N}_i,\, j\in[d]}\alpha_{i,j}\mathbf{M}_j + \sum_{j\notin\mathcal{N}_i,\, j\in[d]}\alpha'_{i,j}\mathbf{M}_j + \sum_{j\in[d_1]\setminus[d]}\beta_{i,j}\mathbf{M}_j^\perp
\end{aligned} \qquad (114)$$

Under the condition of $|\mathcal{M}_j| \neq 0$ (if $|\mathcal{M}_j| = 0$, then it is not a target within $\mathcal{N}_i$, and thus it becomes meaningless), and due to Lemma F.1: The proof is complete. For each feature $\mathbf{M}_j$, there are at most $o(m/d)$ many $i \in [m]$ such that $j \in \mathcal{N}_i$: It follows from the proof of Lemma C.1 that $\mathbb{P}[i \in \mathcal{M}_j] = \frac{1}{d^{\Omega(1)}}$, and at least $\Omega(d^{\omega_1})$ many $i \in [m]$ such that $\mathcal{N}_i = \{j\}$: From Lemma C.1, we recall that $|\mathcal{M}_j^\star| \geq \Omega(d^{\omega_1})$ If a neuron belongs to $\mathcal{M}_j^\star$, then it must not belong to $\mathcal{M}_{j'}$

□