UNDERSTANDING FEATURE LEARNING IN GRAPH CONTRASTIVE LEARNING: NEIGHBORHOOD CORRELATION AND THEORY-GUIDED AUGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph contrastive learning (GCL) learns representations by contrasting augmented graph views, where each node representation aggregates information from its local neighborhood. Despite the empirical success of GCL, how neighborhood structure shapes the feature-learning process remains theoretically underexplored. This paper provides a theoretical characterization of feature learning in GCL under a sparse coding model that separates task-relevant features with stronger neighborhood correlation from task-irrelevant features with weaker neighborhood correlation. Our analysis reveals a selective feature-learning mechanism: neighborhood aggregation preferentially amplifies features with stronger neighborhood correlation while suppressing less correlated features. We quantitatively characterize how feature learning and downstream performance depend on feature overlap, the fraction of correlated neighbors, and neighborhood size. Guided by these theoretical insights, we propose a plug-in augmentation strategy, Theory-Guided Neighborhood Augmentation (TGNA), that constructs and selects graph views using an empirical measure of neighborhood correlation. Experiments on synthetic and real-world graphs support our theoretical predictions and show that TGNA generally improves downstream performance across various GCL frameworks while reducing contrastive training cost.

1 INTRODUCTION

Graph neural networks (GNNs) have become a powerful framework for learning representations from graph-structured data (Kipf & Welling, 2016; Xu et al., 2018), with applications across scientific and relational domains (Senior et al., 2020; Shlomi et al., 2021). However, learning effective GNN representations often relies on task-specific labels, which can be expensive or scarce. Graph contrastive learning (GCL) has emerged as a prominent self-supervised paradigm that learns graph representations without task-specific supervision (Zhu et al., 2020b; You et al., 2020; Zhu et al., 2021). A typical GCL method constructs augmented views of an input graph and learns representations by encouraging agreement between representations of the same node across different views. Unlike conventional contrastive learning, each node representation in GCL incorporates information aggregated from its local neighborhood. Consequently, features consistently shared among neighboring nodes may be reinforced through aggregation, whereas less consistent features may be diluted.

Despite the empirical success of GCL, how graph structure shapes feature learning through contrastive training remains theoretically underexplored. In fact, even for standard contrastive learning, theoretical understanding of the feature-learning process and generalization remains limited. Saunshi et al. (2019) establish theoretical guarantees connecting contrastive learning to downstream classification performance. Wang & Isola (2020) characterize contrastive learning through two representation-level properties, alignment and uniformity. Going beyond these representation-level analyses, Wen & Li (2021) study the neural-network feature-learning dynamics of contrastive learning and show how data augmentation can decouple different latent features and affect which features are learned during training. Their analysis, however, considers contrastive learning without graph neighborhood aggregation and does not capture how message passing on graphs changes the learning dynamics of different feature directions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A few recent studies have begun to investigate graph-specific aspects of GCL theoretically. Liu et al. (2023) connect augmentation strength and positive-pair alignment with contrastive loss and downstream generalization. Zhuo et al. (2024) interpret graph augmentations through the message-passing mechanism of GNNs as local feature modifications within node neighborhoods. However, these studies do not characterize how neighborhood aggregation shapes the learning dynamics of different latent features during contrastive training. This leaves a fundamental question:

How does neighborhood aggregation shape the training dynamics of GCL, and how do these dynamics determine which latent features are preferentially learned?

1.1 Major Contributions

In this paper, we develop a theoretical framework to answer this question. Following recent theoretical studies of GNN training dynamics that adopt shallow GNN models for tractable analysis (Zhang et al., 2020; Li et al., 2022; Awasthi et al., 2021), we consider a one-hidden-layer GNN with symmetrized ReLU neurons and analyze its stochastic-gradient-based training dynamics under a contrastive objective. We adopt a sparse coding model for individual node features and introduce graph-specific neighborhood correlation to capture heterogeneous latent features. Specifically, we consider task-relevant features M^1 with stronger neighborhood correlation and task-irrelevant features M^2 with weaker neighborhood correlation. For each feature group $r \in \{1, 2\}$, neighborhood correlation is characterized by the feature overlap α_r and the fraction of correlated neighbors ρ_r, with K denoting the neighborhood size. While our theory focuses on this tractable shallow model, we further validate the resulting theoretical insights on practical GCL frameworks with deep GNN encoders. Our main contributions are summarized as follows.

Neighborhood-induced selective feature learning. We prove that neighborhood aggregation induces a selective feature-learning mechanism in GCL: features in M^1 are preferentially learned, while the less correlated features in M^2 are suppressed. In contrast, standard contrastive learning learns features from both M^1 and M^2 without exploiting their different neighborhood correlations. This selective recovery of M^1 allows GCL to learn the task-relevant features with fewer training iterations and a smaller model size than standard contrastive learning.

Quantitative dependence on neighborhood correlation. We establish explicit guarantees characterizing that stronger neighborhood correlation leads to faster feature learning and improved downstream performance. In particular, the training-iteration for recovering the task-relevant features decreases as $1/\sqrt{\alpha_1\rho_1^2(K+1)}$, and the downstream performance error exhibits an exponential dependence on $\sqrt{\alpha_1\rho_1^2(K+1)}$. Experiments on synthetic graphs support the predicted theoretical scaling.

Theory-guided neighborhood augmentation. Our theoretical results motivate neighborhood correlation as a principle for constructing graph views. Based on this insight, we propose Theory-Guided Neighborhood Augmentation (TGNA), which uses node-feature similarity as an observable proxy for latent neighborhood correlation and constructs graph views to maximize the empirical correlation criterion $\hat{\alpha}\hat{\rho}^2(\hat{K}+1)$, motivated directly by our theoretical analysis. Experiments real-world graphs show that TGNA improves downstream performance across multiple GCL frameworks.

1.2 Other related work

Graph Contrastive Learning and Graph Augmentation. A typical GCL method constructs contrastive views through random graph perturbations, such as node dropping, edge perturbation, and feature masking. Since such random perturbations may discard semantically or structurally important information, recent studies have developed more structured graph-dependent augmentation strategies. One direction explores structure-aware augmentation, which constructs various views based on saliency, community structure, motifs, node similarity, or joint attribute–structure information (Wei et al., 2023; Tan et al., 2024; Shu et al., 2023; Chi & Ma, 2024; Chen et al., 2025). An alternative direction exploits spectral properties to preserve or perturb informative graph components (Mo et al., 2024; Yang et al., 2025; Zheng et al., 2026; Zhang et al., 2023b; Liu et al., 2025; He et al., 2024). More recently, adaptive approaches have been explored through attention, reinforcement learning, subgraph-specific perturbations, learnable edge weights, progressive view generation and multi-view augmentation strategies.(Shen et al., 2023; Liu et al., 2024; Peng et al.,

2025; Xu et al., 2026; Zhao et al., 2025; Li et al., 2024b). Our work complements these approaches by deriving a quantitative view-selection principle directly from the feature-learning analysis.

Training Dynamics and Feature Learning of Neural Networks. A growing body of theoretical work studies how neural networks learn structured features through gradient-based training. Due to the nonconvexity of neural-network optimization, such analyses commonly focus on shallow models, including one-hidden-layer fully connected and convolutional networks, to characterize how individual neurons evolve toward discriminative feature directions (Li et al., 2018; Brutzkus & Globerson, 2021; Karp et al., 2021; Shi et al., 2022; Damian et al., 2022; Allen-Zhu & Li, 2022). This line of analysis has also been extended to graph-structured data, where theoretical studies characterize the learning and generalization of shallow GNNs under gradient-based optimization and quantify how graph aggregation or sampling affects the learning of class-relevant features (Zhang et al., 2020; Li et al., 2022; Zhang et al., 2023a). These works, however, primarily consider supervised learning objectives, where node labels directly guide the learning dynamics. In contrast, our work studies feature-learning dynamics of GNNs under a self-supervised contrastive objective and characterizes how neighborhood aggregation itself differentiates latent features according to their neighborhood correlations.

2 Problem Formulation: Graph Contrastive Learning

We first introduce the graph contrastive learning (GCL) framework considered in this work.

Graph. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, X\}$ denote an undirected graph, where $\mathcal{V} = \{v_k\}_{k=1}^N$ is the set of N nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, and $X = [x_1, x_2, \ldots, x_N]^\top \in \mathbb{R}^{N \times d_1}$ is the node feature matrix, with $x_k \in \mathbb{R}^{d_1}$ being the feature vector associated with node v_k. The one-hop neighborhood of v_k is denoted by $\mathcal{N}(k) = \{v_j \in \mathcal{V} : (v_k, v_j) \in \mathcal{E}\}$.

GNN encoder. Graph Neural Networks (GNNs) learn node representations by recursively aggregating neighborhood information. Let $h_k^{(0)} = x_k$ denote the initial representation of node v_k. At the ℓ-th layer, the embedding associated with node v_k is first aggregated as

$$\bar{h}_k^{(\ell)} = \mathrm{AGGREGATE}^{(\ell)}\left(\left\{h_j^{(\ell-1)} : v_j \in \mathcal{N}(k) \cup \{v_k\}\right\}\right), \qquad \ell = 1, \ldots, L,$$

where $\mathrm{AGGREGATE}^{(\ell)}(\cdot)$ denotes a permutation-invariant neighborhood aggregation operator with with various aggregation rules (Kortvelesy et al., 2023; Cordonnier et al., 2024). Then the node representation at layer ℓ is obtained by $h_k^{(\ell)} = f^{(\ell)}\left(\bar{h}_k^{(\ell)}\right)$, $\ell = 1, \ldots, L$. We denote the final output representation of node v_k by $f_{\mathcal{G}}(v_k) := h_k^{(L)}$.

Our theoretical analysis focuses on a one-layer GNN with average neighborhood aggregation. Specifically, $L = 1$, and

$$\bar{h}_k^{(1)} = (|\mathcal{N}(k)| + 1)^{-1}\big(x_k + \sum_{v_j \in \mathcal{N}(k)} x_j\big), \tag{1}$$

averages the feature of the center node and those of its one-hop neighbors. The one-layer GNN has m symmetrized ReLU neurons, and the output representation is

$$f_{\mathcal{G}}(v_k) = h_k^{(1)} = \left(h_{k,1}^{(1)}, h_{k,2}^{(1)}, \ldots, h_{k,m}^{(1)}\right)^\top \in \mathbb{R}^m, \tag{2}$$

where

$$h_{k,i}^{(1)} = \mathrm{ReLU}\left(\left\langle w_i, \bar{h}_k^{(1)}\right\rangle - b_i\right) - \mathrm{ReLU}\left(-\left\langle w_i, \bar{h}_k^{(1)}\right\rangle - b_i\right), \tag{3}$$

and $w_i \in \mathbb{R}^{d_1}$ and $b_i \in \mathbb{R}$ are the weight and bias of neuron i, respectively.

Shallow nonlinear networks are widely used in theoretical analyses of training dynamics and feature learning, as they enable precise characterization of learned-feature evolution while retaining the essential nonlinear learning mechanisms (Wen & Li, 2021; Zhang et al., 2020; Li et al., 2022; Ba et al., 2022). This setting allows us to isolate the effect of graph neighborhood aggregation on feature learning under contrastive training.

Graph contrastive learning constructs multiple augmented views of the same underlying graph and learns node representations that are consistent across these views. Given an input graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, X\}$, two augmented graph views $\mathcal{G}^A = \{\mathcal{V}^{(1)}, \mathcal{E}^{(1)}, X^{(1)}\}$, and $\mathcal{G}^B = \{\mathcal{V}^{(2)}, \mathcal{E}^{(2)}, X^{(2)}\}$ may be obtained by common augmentation strategies include node or edge dropping, subgraph sampling, and feature masking or perturbation.

For a center node v_k, applying the same GNN encoder to the two augmented views produce the representations $f_{\mathcal{G}^A}(v_k)$ and $f_{\mathcal{G}^B}(v_k)$, which form a positive pair. Representations associated with other nodes are used as negative samples. Let $\mathfrak{N}_k$ denote the set of negative nodes associated with center node v_k. The similarity between the representations of nodes v_k and v_j from the two views as is defined as

$$\mathrm{Sim}_{f_{\mathcal{G}}}\left(\mathcal{G}^A, \mathcal{G}^B; v_k, v_j\right) := \langle f_{\mathcal{G}^A}(v_k), \mathrm{StopGrad}\left(f_{\mathcal{G}^B}(v_j)\right)\rangle\,. \tag{4}$$

Here, $\mathrm{StopGrad}(\cdot)$ means not computing its gradient during optimization. It is commonly used in contrastive learning formulations to prevent gradient propagation (Wen & Li, 2021; Chen & He, 2021; Grill et al., 2020). The contrastive loss at node v_k is defined as

$$\mathcal{L}\left(f_{\mathcal{G}}, \mathcal{G}^A, \mathcal{G}^B, v_k, \mathcal{B}_k^-\right) = -\tau \log \frac{\exp\left(\mathrm{Sim}_{f_{\mathcal{G}}}(\mathcal{G}^A, \mathcal{G}^B; v_k, v_k)/\tau\right)}{\sum_{v_j \in \mathcal{B}_k^- \cup \{v_k\}} \exp\left(\mathrm{Sim}_{f_{\mathcal{G}}}(\mathcal{G}^A, \mathcal{G}^B; v_k, v_j)/\tau\right)}, \tag{5}$$

where $\tau > 0$ is the temperature parameter.

The training objective of a one-layer GNN encoder with ℓ_2 regularization is to minimize

$$\mathrm{Obj}(f_{\mathcal{G}}) = \mathbb{E}_{k, \mathcal{G}^A, \mathcal{G}^B, \mathcal{B}_k^-}\left[\mathcal{L}\left(f_{\mathcal{G}}, \mathcal{G}^A, \mathcal{G}^B, v_k, \mathcal{B}_k^-\right)\right] + \frac{\lambda}{2}\sum_{i=1}^{m}\|w_i\|_2^2, \tag{6}$$

where $\lambda > 0$ denotes the weight-decay parameter. We consider the typical stochastic gradient descent with learning rate η and batch size B_a.

3 Theoretical Training Dynamics and Generalization Analysis

3.1 Main Theoretical Insights

We first summarize our main theoretical insights before formally presenting the setup and results. We consider a sparse coding model that abstracts heterogeneous semantic features in real graph data into two groups: a task-relevant feature dictionary M^1 with stronger neighborhood correlation and a task-irrelevant feature dictionary M^2 with weaker neighborhood correlation. We characterize neighborhood correlation by the feature overlap ratio α_r and the fraction of correlated neighbors ρ_r, with $\alpha_1\rho_1^2 > \alpha_2\rho_2^2$, and let K denote the neighborhood size.

(I1) Training dynamics and feature learning. We characterize the stochastic-gradient-based training dynamics of GCL and show that GCL preferentially learns features in M^1 while suppressing those in M^2 due to neighborhood aggregation. The number of iterations required to learn the task-relevant features scales inversely with $\sqrt{\alpha_1\rho_1^2(K+1)}$, demonstrating that stronger neighborhood correlation accelerates feature learning (Theorem 3.1).

(I2) Downstream performance. For downstream tasks whose labels depend on M^1, the objective decreases exponentially with $\sqrt{\alpha_1\rho_1^2(K+1)}$, showing that selective learning improves the downstream performance (Theorem 3.2).

(I3) Advantage over standard contrastive learning. Compared with standard contrastive learning, which cannot distinguish the importance of two feature groups and need to learn both M^1 and M^2, GCL reduces the number of training iterations by a factor of $\sqrt{\alpha_1\rho_1^2(K+1)}$ and requires a smaller model size (Remark **??**).

3.2 Data Model

Sparse coding. We adopt the standard sparse coding model commonly used in theoretical studies of neural network feature learning (Allen-Zhu & Li, 2022; Wen & Li, 2021), while partitioning the

latent dictionary into task-relevant and task-irrelevant feature blocks. Specifically, let $M = [M^1 \mid M^2] \in \mathbb{R}^{d_1 \times (d+d')}$, be the column orthonormal matrix with $M^1 \in \mathbb{R}^{d_1 \times d}$, $M^2 \in \mathbb{R}^{d_1 \times d'}$ represent the task-relevant and task-irrelevant features, respectively. For each node v_k, its feature $x_k \in \mathbb{R}^{d_1}$ can be written as

$$x_k = Mz_k + \xi_k, \qquad z_k = [(z_k^1)^\intercal, (z_k^2)^\intercal]^\intercal \in \mathbb{R}^{d+d'}, \quad z_k^1 \in \mathbb{R}^d, \quad z_k^2 \in \mathbb{R}^{d'}, \tag{7}$$

and ξ_k is dense noise in $\mathbb{R}^{d_1}$. We set $d' = \Theta(d) > d$ and $d_1 = \mathrm{poly}(d)$. Following the standard sparse coding model, each coordinate of z_k takes values in $\{-1, 0, 1\}$, and satisfies $\Pr(|z_{k,j}| = 1) = O(\log\log d/d)$. Each dictionary atom $M_j \in \mathbb{R}^{d_1}$ satisfies $\|M_j\|_\infty \le O(1/\sqrt{d_1})$, and the dense noise is independently generated as $\xi_k \sim \mathcal{N}(0, \sigma^2 I_{d_1})$ with $\sigma^2 = O\left(\frac{\sqrt{\log d}}{d}\right)$.

Neighborhood correlation. We assume that each node has K neighbors in each view, i.e., $\mathcal{G}^A$ or $\mathcal{G}^B$, and its neighborhoods in the two views are disjoint. Moreover, in each view, at least a ρ_1 fraction of its neighbors share an α_1 fraction of its active features in M^1. The same condition holds for M^2 with parameters ρ_2 and α_2. Furthermore, $\alpha_1\rho_1^2 > \alpha_2\rho_2^2$, so that the task-relevant features in M^1 have a stronger effective neighborhood correlation than the task-irrelevant features in M^2. Specifically, let $\mathcal{N}^q(k)$ denote the neighborhood of v_k in view $\mathcal{G}^q$, for $q \in \{A, B\}$, and let

$$S_k^1 = \mathrm{supp}(z_k^1) \subseteq [d], \qquad S_k^2 = \mathrm{supp}(z_k^2) \subseteq [d'], \tag{8}$$

where supp denotes the support set. We assume

Assumption 3.1. *For every v_k, the two views satisfy*

$$|\mathcal{N}^A(k)| = |\mathcal{N}^B(k)| = K, \qquad \mathcal{N}^A(k) \cap \mathcal{N}^B(k) = \emptyset. \tag{9}$$

Moreover, for every $q \in \{A, B\}$ and $r \in \{1, 2\}$, there exists $\mathcal{N}^{q,r}(k) \subseteq \mathcal{N}^q(k)$ satisfying

$$|\mathcal{N}^{q,r}(k)| \ge \rho_r K, \qquad |S_k^r \cap S_j^r| \ge \alpha_r |S_k^r|, \quad \textit{for every} \quad v_j \in \mathcal{N}^{q,r}(k), \tag{10}$$

where $\alpha_1\rho_1^2 > \alpha_2\rho_2^2$.

The main motivation of our Assumption 3.1 is the homophily and the local homophily concepts widely consider in graph representation learning which suggest that connected nodes in a graph typically share similar labels or semantic information (Loveland et al., 2024; Zhu et al., 2020a; Ma et al., 2021)

3.3 Formal Theoretical Results

We next characterize how neighborhood correlation affects feature learning under GCL. In particular, we show that the stronger neighborhood correlation of M^1 in Assumption 3.1 leads GCL to preferentially learn features in M^1 while suppressing those in M^2.

Theorem 3.1 (Selective Feature Learning under GCL)**.** *Let $m = d^{1.01}$ be the number of neurons, $\tau = \alpha_1\rho_1^2\,\mathrm{polylog}(d)$, $B_a = \mathrm{poly}(d_1)$, and $|\mathcal{B}_k^-| = \mathrm{poly}(d)$ for all v_k. Let the learning rate $\eta = \eta'\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}$ with $\eta' = \frac{1}{\mathrm{poly}(d)}$. Then, under Assumption 3.1, the GCL objective satisfies,*

$$\frac{1}{T}\sum_{t=T_1}^{T_1+T-1} \mathit{Obj}(f_{\mathcal{G}}^t) \le o(1) \tag{11}$$

for any $T_1 + T - 1 \le T_2$, and $T \ge \Omega\big(\frac{d^{1.21}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}\big)$, with $T_1 = \frac{d^{1.01}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}$, and $T_2 = \frac{d^{1.99}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}$.

Moreover, with high probability, for every neuron $i \in [m]$ and $t \in [T_1, T_2]$,

$$w_i^{(t)} = \sum_{j\in\mathcal{I}_i} A_{ij}M_j^1 + \sum_{j\notin\mathcal{I}_i} B_{ij}M_j^1 + \sum_{j\in[d']} B'_{ij}M_j^2 + \sum_{j\in[d_1-(d+d')]} C_{ij}M_j^\perp, \tag{12}$$

where $\mathcal{I}_i \subseteq [d]$, $|\mathcal{I}_i| = O(1)$, and $A_{ij} \in \left[\frac{\mathrm{polylog}(d)}{d^\epsilon}, O(\alpha_1\rho_1^2\,\mathrm{polylog}(d))\right]$, $B_{ij} \le o(\frac{1}{\sqrt{d(K+1)}})\|w_i^{(t)}\|_2$, $B'_{ij} \le o(\frac{\alpha_2\rho_2^2}{\alpha_1\rho_1^2}\frac{1}{\sqrt{d(K+1)}})\|w_i^{(t)}\|_2$, $C_{ij} \le o(\frac{1}{\sqrt{d_1(K+1)}})\|w_i^{(t)}\|_2$. Furthermore, for every M_j^1, at most $o(m/d)$ neurons satisfy $j \in \mathcal{I}_i$, and at least $\Omega(1)$ neurons satisfy $\mathcal{I}_i = \{j\}$.

Theorem 3.1 establishes a feature-selection effect induced by graph structure: neighborhood aggregation preferentially amplifies features that are consistently shared across nearby nodes. Consequently, the learned neurons preferentially align with the more strongly correlated task-relevant features in M^1, while suppressing the task-irrelevant features in M^2.

Remark 1 (Comparison with standard contrastive learning). Our data model can be viewed as the sparse coding model considered in Wen & Li (2021) by treating $M = [M^1 \mid M^2]$ as a single dictionary with $D = d + d'$ latent features. Directly applying Theorem 3.4 of Wen & Li (2021), standard contrastive learning learns the entire dictionary M with network width $m = D^{1.01} = (d + d')^{1.01}$, without distinguishing between the task-relevant features in M^1 and the task-irrelevant features in M^2. In contrast, Theorem 3.1 shows that GCL exploits neighborhood information to selectively recover M^1 with width $m = d^{1.01}$, requiring a smaller network size. Moreover, GCL also requires fewer training iterations. Specifically, relative to the training-iteration range in Theorem 3.4 of Wen & Li (2021), the corresponding range in Theorem 3.1 is reduced by a factor of $\sqrt{\alpha_1 \rho_1^2 (K+1)}$.

Remark 2 (Proof techniques and novelty). Our analysis requires graph-specific techniques beyond existing feature-learning analyses of standard contrastive learning. (i) Neighborhood aggregation in GCL introduces dependencies among node representations through overlapping neighborhoods so that the typical i.i.d. assumption for the training data in contrastive learnign no long holds. We develop concentration arguments that explicitly account for these graph-induced dependencies. (ii) we develop a graph-aware decomposition of the gradient that separates the contributions from M_1, M_2 and spurious feature directions after neighborhood aggregation. (iii) The learning dynamics in GCL depend explicitly on the neighborhood correlation of each feature group. In particular, the effective signal for group r scales with $\alpha_r \rho_r^2 (K+1)$, leading to different growth rates for the M^1 and M^2 features. The detailed proof is in Appendix.

We next evaluate the learned model $f_{\mathcal{G}}^T$ on downstream supervised node classification tasks in which the label depends only on the task-relevant features M_1 but not M_2. Specifically, for each node v_k, its label y_k is determined by $y_k = \operatorname{sign}\left(\langle u^\star, z_k^1 \rangle\right)$, where $u^\star \in \mathbb{R}^d$ is the ground-truth downstream-task vector with $|u_j^\star| = \Theta(1)$for all $j \in [d]$. We train a linear classifier $U \in \mathbb{R}^m$ to predict the label by $\widehat{y}_k = \langle U, f_{\mathcal{G}}^T(x_k) \rangle$. The classifier is trained using the regularized logistic objective

$$\mathrm{Obj}_T^{\mathrm{dwn}}(U) = \mathbb{E}_{v_k}\left[\log\left(1 + \exp\left(-y_k \langle U, f_{\mathcal{G}}^T(x_k)\rangle\right)\right)\right] + \frac{\lambda_d}{2}\|U\|_2^2. \tag{13}$$

We optimize $\mathrm{Obj}_T^{\mathrm{dwn}}(U)$ using SGD with learning rate η_d for S iterations, and denote the classifier at iteration s by $U^{(s)}$. Since Theorem 3.1 guarantees that the learned $f_{\mathcal{G}}^T$ preserves the features in M^1, we next show that a linear classifier trained on this representation achieves a vanishing downstream loss.

Theorem 3.2 (Downstream training convergence). *Let $\lambda_d = O(1/d)$ and $\eta_d = O(1/\operatorname{poly}(d))$. For $S\eta_d = O(1)$, the downstream classifier satisfies*

$$\frac{1}{S}\sum_{s=0}^{S-1} \mathrm{Obj}_T^{\mathrm{dwn}}(U^{(s)}) \le \left(\frac{1}{\log d}\right)^{O\left(\sqrt{\frac{\alpha_1 \rho_1^2 (K+1)}{\log\log d}}\right)}. \tag{14}$$

This result shows that suppressing M^2 does not prevent successful downstream learning. More importantly, stronger neighborhood correlation of the task-relevant features, characterized by $\alpha_1 \rho_1^2$, improves the performance of the downstream task.

4 Theory-Guided Neighborhood Augmentation

Theorem 3.1 suggests that effective graph augmentation should jointly account for feature similarity of task relevant features, the fraction of correlated neighbors, and neighborhood size, as characterized by $\alpha_1 \rho_1^2 (K+1)$. In practice, the latent decomposition into M^1 and M^2 is unobserved. We therefore use node-feature similarity as an observable proxy for neighborhood correlation and propose *Theory-Guided Neighborhood Augmentation (TGNA)*, which selects the augmentation that maximizes the corresponding empirical criterion.

Specifically, for each node v_k, we compute its cosine similarity with each neighbor v_j in $\mathcal{N}(k)$, $s_{kj} = \frac{\langle x_k, x_j \rangle}{\|x_k\|_2 \|x_j\|_2}$. Given a similarity threshold γ, we partition $\mathcal{N}_k$ into similar and dissimilar sets $\mathcal{N}_k^+ = \{v_j \in \mathcal{N}(k) : s_{kj} \geq \gamma\}$, and $\mathcal{N}_k^- = \{v_j \in \mathcal{N}(k) : s_{kj} < \gamma\}$.

Let $2K$ denote the average degree of the original graph $\mathcal{G}$. For a candidate ratio ρ, we set the anchor set as

$$\mathcal{V}_{\text{anc}}(\rho) = \left\{ v_k \in V : |\mathcal{N}_k^+| \geq 2\lfloor \rho K \rfloor,\ |\mathcal{N}_k^-| \geq 2\lfloor (1-\rho)K \rfloor \right\} \tag{15}$$

For each $v_k \in \mathcal{V}_{\text{anc}}(\rho)$, we randomly draw two disjoint subsets $\widetilde{\mathcal{N}}_k^{A+}$ and $\widetilde{\mathcal{N}}_k^{B+}$ of size $\lfloor \rho K \rfloor$ from $\mathcal{N}_k^+$, and similarly two disjoint subsets $\widetilde{\mathcal{N}}_k^{A-}$ and $\widetilde{\mathcal{N}}_k^{B-}$ of size $\lfloor (1-\rho)K \rfloor$ from $\mathcal{N}_k^-$. The neighborhood of v_k in two views are $\widetilde{\mathcal{N}}_k^A = \widetilde{\mathcal{N}}_k^{A+} \cup \widetilde{\mathcal{N}}_k^{A-}$ and $\widetilde{\mathcal{N}}_k^B = \widetilde{\mathcal{N}}_k^{B+} \cup \widetilde{\mathcal{N}}_k^{B-}$. Aggregating these neighborhoods over the anchors produces two directed graph views $\mathcal{G}^A$ and $\mathcal{G}^B$.

For each sampled pair of views, we estimate $\widehat{\alpha}$ by the average cosine similarity between each eligible anchor and its selected similar neighbors, averaged over the eligible anchors and both views. The realized $\widehat{\rho}$ is the average fraction of selected neighbors that belong to the similar sets, while $\widehat{K}$ is the average size of the selected neighborhoods. We evaluate each sampled pair of views using score $\widehat{\mathcal{S}}_{\text{TGNA}} = \widehat{\alpha}\,\widehat{\rho}^2(\widehat{K}+1)$ and select the best pair over candidate values of ρ and multiple random trials. We denote the resulting views and anchor set by $\mathcal{G}^{A\star}$, $\mathcal{G}^{B\star}$, and $\mathcal{V}_{\text{anc}}^\star$. Only nodes in $\mathcal{V}_{\text{anc}}^\star$ participate in the contrastive objective, with negative set $\mathcal{B}_k^- = \mathcal{V}_{\text{anc}}^\star \setminus \{v_k\}$ for each anchor $v_k \in \mathcal{V}_{\text{anc}}^\star$. The complete procedure is summarized in Algorithm 1 in Appendix J.

Plug-in Augmentation. TGNA can be used as a plug-in replacement for the graph augmentation module in existing augmentation-based GCL methods, without modifying their GNN architecture or contrastive objective. $\mathcal{G}^{A\star}$, $\mathcal{G}^{B\star}$, and $\mathcal{V}_{\text{anc}}^\star$ are fixed and reused throughout training, avoiding repeated view generation across training iterations. Moreover, restricting contrastive loss to $\mathcal{V}_{\text{anc}}^\star$ reduces the number of anchors and negative pairs involved in the contrastive objective.

5 Experiments

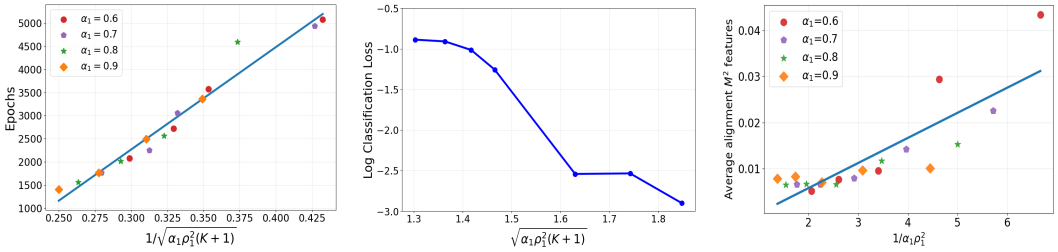
5.1 Evaluation of theoretical results on Synthetic Data

The data are generated according to the sparse coding model in Section 3.2, with $d = 10$, $d' = 20$, and $d_1 = 200$. The encoder contains $m = 128$ neurons. The additive noise is sampled from $\mathcal{N}(0, 0.01 I_{d_1})$. We use $|\mathcal{B}_k^-| = 200$ negative samples for all anchor node v_k and trained using SGD with batch size 100 and learing rate 0.05, The graph has 6000 nodes where 5000 nodes used for training and remaining for testing. We evaluate the impact of neighborhood correlations on the following quantities.

Training convergence. In Figure 1a, We vary $K \in \{10, 15, 20\}$ and $\rho_1, \alpha_1 \in [0.1, 0.9]$, while fixing $\alpha_2 = 0.2$, and $\rho_2 = 0.25$. We report the number of epochs required for the training loss to fall below 0.005. As shown in Figure 1a, the required number of training iterations is linear in as $1/\sqrt{\alpha_1 \rho_1^2 (K+1)}$, consistent with requirement of T in Theorem 3.1.

Downstream performance. After GCL, we freeze the learned encoder and train the linear classifier as in (13). In Figure 1b, we vary both α_1 and ρ_1, while fixing $K = 15$, $\alpha_2 = 0.2$, and $\rho_2 = 0.25$. The logarithm of the downstream loss decreases approximately linearly with $\sqrt{\alpha_1 \rho_1^2 (K+1)}$, verifying Theorem 3.2.

Suppression of task-irrelevant features. In Figure 1c, we evaluate the alignment of the learned neurons with the M^2 features. For each neuron w_i and feature M_j^2, we measure the normalized alignment $|\langle w_i, M_j^2 \rangle| / \|w_i\|_2$ and then average over all neurons and features in M^2. We vary $\alpha_1 \in [0.6 - 0.9]$ $\rho_1 \in [0.5 - 0.9]$, while fixing $K = 15$, $\alpha_2 = 0.2$, and $\rho_2 = 0.25$. The M^2 alignment decreases approximately with $1/(\alpha_1 \rho_1^2)$, consistent with the scaling of B'_{ij} in Theorem 3.1. This shows that stronger neighborhood correlation in the task-relevant features suppresses the learning of task-irrelevant features.



(a) Training epochs (b) Downstream loss. (c) M_2-feature alignment.

Figure 1: Impact of neighborhood correlations: (a) Number of training epochs required (b) Downstream classification loss (c) Neuron alignment with M_2 features

5.2 Experiments on Real Data using TGNA

To evaluate its applicability across different GCL methods, we use TGNA as a plug-in replacement for the original graph augmentation in five representative frameworks: **GRACE** (Zhu et al., 2020b), **GCA** (Zhu et al., 2021), **HomoGCL** (Li et al., 2023), **Str-GCL** (He et al., 2025), and **SpikeGCL** (Li et al., 2024a). For each framework, we retain its original encoder, contrastive objective, and downstream evaluation protocol, and only replace its graph augmentation with the graph views generated by TGNA. We evaluate on five benchmark node-classification datasets: **Cora**, **PubMed** (Sen et al., 2008), **Amazon Photo**, **Amazon Computers**, and **Coauthor-CS** (Shchur et al., 2018). After contrastive training, the encoder is frozen, and a linear classifier is trained for the downstream tasks.

Improved downstream performance. Table 1 reports the node-classification accuracy. TGNA improves the accuracy in 20 out of 25 model–dataset combinations, demonstrating that it can benefit a diverse set of GCL frameworks and datasets without modifying their encoders or contrastive objectives. We further compare with three recent GCL methods, ASGCL (Liu et al., 2025), NDSCL (Li et al., 2024b), and NCLA (Shen et al., 2023), as additional baselines beyond the augmentation-based comparisons. The results show that TGNA-enhanced methods remain competitive with these recent approaches across the evaluated datasets.

Table 1: Mean node classification accuracy (%) on real-world graph datasets. Boldface indicates the better result within each original/TGNA pair. Results for ASGCL, NDSCL, and NCLA are taken from their respective papers.

	Algorithm	Cora	AmazonPhoto	AmazonComputers	PubMed	CS
TGNA plug-in	GRACE	83.54±0.64	91.29±0.19	82.23±0.25	85.90±0.04	**92.65±0.22**
	GRACE+TGNA	**84.04±0.17**	**92.51±0.23**	**85.69±0.29**	**85.97±0.26**	92.36±0.05
	GCA	82.01±0.34	**92.66±0.16**	86.87±0.05	84.81±0.09	**93.04±0.07**
	GCA+TGNA	**83.11±0.17**	92.28±0.28	**87.66±0.21**	**86.56±0.14**	92.20±0.20
	HomoGCL	83.90±0.46	92.16±0.56	87.79±0.14	80.10±0.82	91.73±0.04
	HomoGCL+TGNA	**84.20±0.07**	**92.79±0.41**	**88.05±0.19**	**81.33±0.84**	**92.11±0.16**
	Str-GCL	**84.38±0.48**	92.24±0.09	88.91±0.26	86.72±0.05	92.69±0.12
	Str-GCL+TGNA	84.15±0.13	**93.05±0.33**	**89.44±0.08**	**86.98±0.07**	**93.68±0.11**
	SpikeGCL	86.61±0.52	92.61±0.11	89.01±0.14	87.85±0.29	**92.38±0.19**
	SpikeGCL+TGNA	**86.83±0.48**	**92.62±0.22**	**89.05±0.19**	**88.52±0.17**	91.43±0.33
Baselines	ASGCL (Liu et al., 2025)	85.20±0.30	93.50±0.20	89.90±0.30	84.90±0.30	93.50±0.30
	NDSCL (Li et al., 2024b)	84.34±0.50	90.51±1.00	80.71±2.10	80.51±0.10	92.71±0.20
	NCLA(Shen et al., 2023)	82.20±1.60	90.20±1.30	83.7± 1.10	82.00±1.40	91.50±0.70

Reduced contrastive training cost. Table 2 evaluates the GCL training efficiency of TGNA after the graph views have been selected for a fixed γ. Across the five datasets, TGNA retains only

Table 2: Anchor retention ratio and GCL training time after view selection on real-world graph datasets on real-world graph datasets.

	Method	Cora	AmazonPhoto	AmazonComputers	PubMed	CS
	Anchor ratio (%)	46.86	66.54	66.82	30.93	68.96
Training time (s)	GRACE	10.8	119.4	315.8	777.7	506.5
	GRACE+TGNA	**8.8**	**47.5**	**123.9**	**73.1**	**321.7**
	GCA	37.3	96.5	220.0	332.5	**331.7**
	GCA+TGNA	**20.8**	**68.0**	**118.2**	**175.9**	459.6
	HomoGCL	23.1	19.7	67.9	79.6	**85.8**
	HomoGCL+TGNA	**21.6**	**18.1**	**44.8**	**62.4**	86.6
	Str-GCL	73.4	643.8	445.6	1501.2	773.9
	Str-GCL+TGNA	**17.8**	**151.3**	**403.3**	**739.7**	**425.8**
	SpikeGCL	**10.5**	39.3	22.6	**53.7**	**31.2**
	SpikeGCL+TGNA	10.8	**35.1**	**22.3**	55.2	33.0

30.93%–68.96% of the nodes as anchors, reducing the number of anchors and negative pairs involved in contrastive training. Consequently, TGNA reduces the training time while generally maintaining or improving downstream performance (Table 1). For example, on Amazon Computers, TGNA improves the accuracy of GRACE from 82.23% to 85.69% while reducing its training time from 315.8 s to 123.9 s.

Validation of the theoretical error scaling. We vary the target neighbor ratio ρ to generate a collection of graph views pairs. For each pair of views, we compute $\widehat{\alpha}$, $\widehat{\rho}$, and $\widehat{K}$, train the corresponding GCL model, and evaluate its downstream classification error. As shown in Figure 2, across different GCL frameworks, the logarithm of the error on Cora Dataset decreases approximately linearly with $\sqrt{\widehat{\alpha}\widehat{\rho}^2(\widehat{K}+1)}$, consistent with the exponential dependence predicted by (193) in Theorem 3.2. We have similar results for other datasets as shown in AppendixK.

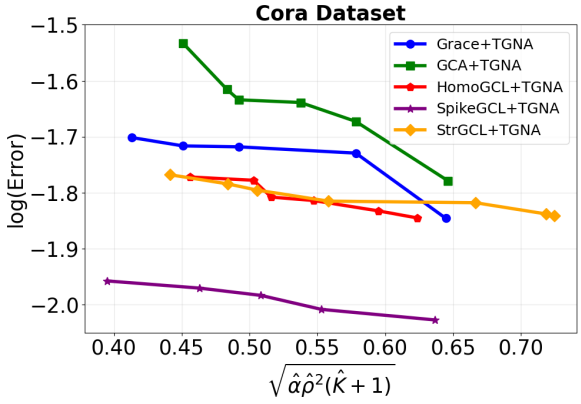


Figure 2: Downstream classification error versus $\sqrt{\hat{\alpha}\hat{\rho}^2(\hat{K}+1)}$ on Cora.

6 Conclusion

This work provides the first theoretical analysis of the training dynamics of graph contrastive learning (GCL).Our analysis shows that neighborhood aggregation amplifies features with stronger neighborhood correlation, enabling GCL to preferentially learn semantically consistent features and improving feature learning over contrastive learning without graph aggregation.

Guided by this insight, we propose Theory-Guided Neighborhood Augmentation (TGNA), which constructs and selects graph views based on a theory-motivated measure of neighborhood correlation. TGNA serves as a plug-in augmentation for existing GCL frameworks without modifying their encoders or contrastive objectives. Experiments on synthetic and real-world graphs validate our theoretical findings and show that TGNA generally improves downstream performance while reducing the number of nodes and pairs used in contrastive training and, consequently, the training time.

LLM USAGE DISCLOSURE

We used large language models to aid in polishing the writing, LaTeX error debugging, and error debugging for the code used to implement the experiments. All research ideas, theoretical analysis, and experiments were completely conducted by the authors.

REPRODUCIBILITY STATEMENT

We provide comprehensive information to enable the reproduction of both theoretical and empirical results. Explicit assumptions and settings are detailed in the main paper, with complete proofs and lemmas included in the appendix. The supplementary material includes source code for the proposed procedures and reproducibility instructions.

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.

Pranjal Awasthi, Abhimanyu Das, and Sreenivas Gollapudi. A convergence analysis of gradient descent on graph neural networks. *Advances in Neural Information Processing Systems*, 34: 20385–20397, 2021.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Alon Brutzkus and Amir Globerson. An optimization and generalization analysis for max-pooling networks. In *Uncertainty in Artificial Intelligence*, pp. 1650–1660. PMLR, 2021.

Jialu Chen, Rui Chen, and Gang Kou. Unifying attribute and structure preservation for enhanced graph contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6): 10479–10490, 2025.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15745–15753. IEEE, 2021.

Hongliang Chi and Yao Ma. Enhancing contrastive learning on graphs with node similarity. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 456–465, 2024.

Matthieu Cordonnier, Nicolas Keriven, Nicolas Tremblay, and Samuel Vaiter. Convergence of message-passing graph neural networks with generic aggregation on large random graphs. *Journal of Machine Learning Research*, 25(406):1–49, 2024.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on learning theory*, pp. 5413–5452. PMLR, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

Dongxiao He, Jitao Zhao, Cuiying Huo, Yongqi Huang, Yuxiao Huang, and Zhiyong Feng. A new mechanism for eliminating implicit conflict in graph contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12340–12348, 2024.

Dongxiao He, Yongqi Huang, Jitao Zhao, Xiaobao Wang, and Zhen Wang. Str-gcl: Structural commonsense driven graph contrastive learning. In *Proceedings of the ACM Web Conference 2025*, pp. 1129–1141, 2025. doi: 10.1145/3696410.3714900.

Stefani Karp, Yuanzhi Li, Sreeram Pramanik, Aarti Singh, and Ju Zhang. Local signal adaptivity: Provable feature learning in neural networks beyond kernels. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24883–24897, 2021.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Ryan Kortvelesy, Steven Morad, and Amanda Prorok. Generalised f-mean aggregation for graph neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=JMrIeKjTAe.

Hongkang Li, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. Generalization guarantee of training graph convolutional networks with graph topology sampling. In *International Conference on Machine Learning*, pp. 13014–13051. PMLR, 2022.

Jintang Li, Huizhe Zhang, Ruofan Wu, Zulun Zhu, Baokun Wang, Changhua Meng, Zibin Zheng, and Liang Chen. A graph is worth 1-bit spikes: When graph contrastive learning meets spiking neural networks. In *International Conference on Learning Representations*, 2024a.

Wen-Zhi Li, Chang-Dong Wang, Hui Xiong, and Jian-Huang Lai. Homogcl: Rethinking homophily in graph contrastive learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1341–1352, 2023. doi: 10.1145/3580305.3599380.

Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter W. Battaglia. Learning deep generative models of graphs. *ArXiv*, abs/1803.03324, 2018. URL https://api.semanticscholar.org/CorpusID:3783953.

Zhe Li, Lei Zhou, Yandong Hou, Min Ji, Zhuanzheng Hang, and Bolun Chen. Node classification in complex networks based on multi-view debiased contrastive learning. *Complex & Intelligent Systems*, 10(4):5683–5701, 2024b.

Jingyu Liu, Huayi Tang, and Yong Liu. Perfect alignment may be poisonous to graph contrastive learning. *arXiv preprint arXiv:2310.03977*, 2023.

Ruyue Liu, Rong Yin, Yong Liu, Xiaoshuai Hao, Haichao Shi, Can Ma, and Weiping Wang. Asgcl: Asymmetric spectral augmentation on graph contrastive learning. *IEEE Transactions on Multimedia*, 2025.

Ziyang Liu, Chaokun Wang, and Cheng Wu. Graph contrastive learning with reinforcement augmentation. In *IJCAI*, pp. 2225–2233, 2024.

Donald Loveland, Jiong Zhu, Mark Heimann, Benjamin Fish, Michael T Schaub, and Danai Koutra. On performance discrepancies across local homophily levels in graph neural networks. In *Learning on Graphs Conference*, pp. 6–1. PMLR, 2024.

Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. Is homophily a necessity for graph neural networks? *arXiv preprint arXiv:2106.06134*, 2021.

Yanhu Mo, Xiao Wang, Shaohua Fan, and Chuan Shi. Graph contrastive invariant learning from the causal perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8904–8912, 2024.

Tianhao Peng, Xuhong Li, Haitao Yuan, Yuchen Li, and Haoyi Xiong. Sola-gcl: Subgraph-oriented learnable augmentation method for graph contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 19875–19883, 2025.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning*, pp. 5628–5637. PMLR, 2019.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–106, 2008.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Xiao Shen, Dewang Sun, Shirui Pan, Xi Zhou, and Laurence T. Yang. Neighbor contrastive learning on learnable graph augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9782–9791, 2023.

Zhenmei Shi, Junyi Wei, and Yingyu Liang. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. *arXiv preprint arXiv:2206.01717*, 2022.

Jonathan Shlomi, Peter Battaglia, and Jean-Roch Vlimant. Graph neural networks in particle physics. *Machine Learning: Science and Technology*, 2(2):021001, 2021.

Lin Shu, Chuan Chen, and Zibin Zheng. Capturing fine-grained semantics in contrastive graph representation learning. *arXiv preprint arXiv:2304.11658*, 2023.

Shiyin Tan, Dongyuan Li, Renhe Jiang, Ying Zhang, and Manabu Okumura. Community-invariant graph contrastive learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47579–47606. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/tan24b.html.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9929–9939. PMLR, 2020.

Chunyu Wei, Yu Wang, Bing Bai, Kai Ni, David Brady, and Lu Fang. Boosting graph contrastive learning via graph contrastive saliency. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36839–36855. PMLR, 2023.

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*, pp. 11112–11122, 2021.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Yu Xu, Guixiang Zhu, Huijia Li, Jia Wu, Jie Cao, and Darko Vukovic. Enhancing contrastive learning on learnable graph augmentation. *IEEE Transactions on Big Data*, 2026.

Liang Yang, Zhenna Li, Jiaming Zhuo, Jing Liu, Ziyi Ma, Chuan Wang, Zhen Wang, and Xiaochun Cao. Graph contrastive learning with joint spectral augmentation of attribute and topology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21983–21991, 2025.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5812–5823, 2020.

Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. Fast learning of graph neural networks with guaranteed generalizability: one-hidden-layer case. In *International Conference on Machine Learning*, pp. 11268–11277. PMLR, 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. *arXiv preprint arXiv:2302.02922*, 2023a.

Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, and Irwin King. Spectral feature augmentation for graph contrastive learning and beyond. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11289–11297, 2023b.

Yuhai Zhao, Yejiang Wang, Zhengkui Wang, Wen Shan, Miaomiao Huang, and Xingwei Wang. Graph contrastive learning with progressive augmentations. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2079–2088, 2025.

Xiangping Zheng, Xiuxin Hao, Bo Wu, Wei Li, Bin Ren, Bin Tang, Yuhui Guo, Xun Liang, and Zhiwen Yu. From semantics to spectrum: A new lens on graph augmentation strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16486–16494, 2026.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33:7793–7804, 2020a.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. In *ICML Workshop on Graph Representation Learning and Beyond*, 2020b.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*, pp. 2069–2080, 2021.

Jiaming Zhuo, Yintong Lu, Hui Ning, Kun Fu, Bingxin Niu, Dongxiao He, Chuan Wang, Yuanfang Guo, Zhen Wang, Xiaochun Cao, et al. Unified graph augmentations for generalized contrastive learning on graphs. *Advances in Neural Information Processing Systems*, 37:37473–37503, 2024.

OVERVIEW OF THE APPENDIX

The appendix contains eleven sections. Section A provides the notations, training objective, gradient formulation, and the generalization bound. Section B provides the proof sketch of our theoretical analysis. In Section C we derive the neighborhood depend generalization bound. Section D introduces definition of lucky neurons, properties and other lemmas used at the initial stage. Section E provides the gradient dynamics computation of M^1, M^2,$M^\perp$ features and other supporting lemmas. Section F analyzes Stage 1, establishing the initial growth of useful features and the learning dynamics. Section G studies Stage 2, showing how subsets of neurons become associated with M^1 features and maintain this specialization during training. Section H analyzes Stage 3, where sparsely activated neurons converge to sparse feature representations and the contrastive objective approaches convergence. Section I provides how well the learned model perform in the downstream task with the similarity score of the graph neighbors. Section J provides our TGNA pseudo algorithm. Finally, Section K we included additional experiments we conducted for both synthetic and real datasets.

A PRELIMINARIES AND NOTATIONS

Table 3: Summary of main notations

Symbol	Description
η	Learning rate
λ	Regularization parameter
τ	Temperature coefficient
K	Number of one hop neighbors
$\mathfrak{B}_k^-$	Set of negative samples of center node k
$\mathcal{B}$	Set of negative and positive
M^1, M^2	Set of majority and minority features
α_1, α_2	Fraction of shared M^1 and M^2feature dimensions
ρ_1, ρ_2	Fraction of neighbours' M^1 and M^2 features similar to the center node

We denote the hidden layer weights by at every iteration $t \geq 0$ as, $w^{(t)} = [w_1^{(t)}, w_1^{(t)}, \ldots, w_m^{(t)}] \in \mathbb{R}^{d_1 \times m}$ and each node feature vector $\{x_k\}_{k=1}^N \in \mathbb{R}^{d_1}$. We denote the feature vector of an each aggregated node with their neighbours by one layer GNN as

$$\bar{h}_k^{(1)} = (|\mathcal{N}(k)| + 1)^{-1}(x_k + \sum_{v_j \in \mathcal{N}(k)} x_j) := \bar{x}_k \tag{16}$$

The output of the one layer GNN for $\bar{x}_k$ is

$$f_{\mathcal{G}}^t(\bar{x}_k) = (h_{1,t}(\bar{x}_k), h_{2,t}(\bar{x}_k), \ldots, h_{m,t}(\bar{x}_k)^\top \in \mathbb{R}^m$$
$$h_{i,t}(\bar{x}_k) = \phi(\langle w_i^{(t)}, \bar{x}_k \rangle - b_i^{(t)}) - \phi(-\langle w_i^{(t)}, \bar{x}_k \rangle - b_i^{(t)})$$

where $\phi(.)$ is the activation function. We use ReLU function $\phi(x) = \max\{0, x\}$ as our activation in the entire contrastive learning framework. For the given $\bar{x}_k$, we denote the $\bar{x}_k^{(1)}, \bar{x}_k^{(2)}$ as the aggregated feature vector of $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ respectively. We use $h_{i,t}(\bar{x}_k)$ as the first layer node representation of the node k represent as $h_{k,i}^{(1)}$ in the paper to avoid complexity.

We define the similarity between two graph views of the GNN output of node k by

$$\mathrm{Sim}_{f_{\mathcal{G}}^t}(\bar{x}_k^{(1)}, \bar{x}_k^{(2)}) = \left\langle f_{\mathcal{G}}^t(\bar{x}_k^{(1)}, \bar{x}_k^{(2)}), \mathrm{StopGrad}\left(f_{\mathcal{G}}^t(\bar{x}_k^{(1)}, \bar{x}_k^{(2)})\right) \right\rangle.$$

Given the positive pair $(\bar{x}_k^{(1)}, \bar{x}_k^{(2)})$, let $\mathfrak{B}_k^- = \{x_{n,s}\}_{s \in [|\mathfrak{B}_k^-|]}$ be a batch of negative nodes, where each $x_{n,s}$ is sampled independently from the graph data distribution. Define $\mathfrak{B} = \{\bar{x}_k^{(2)}\} \cup \mathfrak{B}_k^-$. Then the population contrastive loss is

$$L(f_{\mathcal{G}}^t) = \mathbb{E}_{x_k, \mathfrak{B}}\left[-\mathrm{Sim}_{f_{\mathcal{G}}^t}(\bar{x}_k^{(1)}, \bar{x}_k^{(2)}) + \tau \log\left(\sum_{x_k \in \mathcal{B}} \exp\left(\mathrm{Sim}_{f_{\mathcal{G}}^t}(\bar{x}_k^{(1)}, x_k)/\tau \right) \right) \right]. \tag{17}$$

To compute the gradient of our loss function $\mathcal{L}(f_{\mathcal{G}}^t, \bar{x}_k^{(1)}, x_k, \mathcal{B}^\ell)$ with respect to the weights $\{w_i^{(t)}\}_{i\in[m]}$, we define the following notations: positive logit $\ell'_{p,t}(\bar{x}_k^{(1)}, x_k, \mathcal{B})$ and negative logits $\ell'_{s,t}(\bar{x}_k^{(1)}, x_k, \mathcal{B})$:

$$\begin{aligned}\ell'_{p,t}(\bar{x}_k^{(1)}, x_k, \mathcal{B}) &:= \frac{e^{\mathrm{Sim}_{f_{\mathcal{G}}^t}(\bar{x}_k^{(1)}, x_k, \bar{x}_k^{(2)}, x_k)/\tau}}{\sum_{x\in\mathcal{B}} e^{\mathrm{Sim}_{f_{\mathcal{G}}^t}(\bar{x}_k^{(1)}, x_k, x_k)/\tau}},\\ \ell'_{s,t}(\bar{x}_k^{(1)}, x_k, \mathcal{B}) &:= \frac{e^{\mathrm{Sim}_{f_{\mathcal{G}}^t}(\bar{x}_k^{(1)}, x_k, x_{n,s})/\tau}}{\sum_{x\in\mathcal{B}} e^{\mathrm{Sim}_{f_{\mathcal{G}}^t}(\bar{x}_k^{(1)}, x_k, x_k)/\tau}}.\end{aligned} \tag{18}$$

Then the empirical gradient of $L(f_{\mathcal{G}}^t)$ with respect to weight $w_i^{(t)}$ at iteration t can be expressed as The gradient of the loss is

$$\begin{aligned}\nabla_{w_i}\mathcal{L}(f_{\mathcal{G}}^t) = \mathbb{E}\Bigg[&-(1-\ell'_{p,t})\cdot h_{i,t}(\bar{x}_k^{(2)})\cdot \mathbf{1}_{\left|\langle w_i^{(t)}, \bar{x}_k^{(1)}\rangle\right| \geq b_i^{(t)}}\cdot \bar{x}_k^{(1)}\\ &+ \sum_{\bar{x}_{n,s}\in\mathfrak{B}_k^-} \ell'_{s,t}\cdot h_{i,t}(\bar{x}_{n,s})\cdot \mathbf{1}_{\left|\langle w_i^{(t)}, \bar{x}_k^{(1)}\rangle\right| \geq b_i^{(t)}}\cdot \bar{x}_k^{(1)}\Bigg]\end{aligned} \tag{19}$$

B Proof Sketch

Stage 1: We analyse the how the features learned by the GNN encoder weights. Initially, the neuron weights remain close enough to their random initialization that the contrastive logits are approximately uniform and the gradient can be characterized explicitly. Then, each neuron gradually learns the task-relevant feature directions while scarcely learning the task-irrelevent directions. During this stage, the projection of each neuron onto the desired feature block M^1, although small at initialization, grows multiplicatively at a rate determined by the semantic consistency shared across the two graph views. In particular, the dominant growth term is of order $\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\,\frac{\log\log d}{d}$ which accumulates over training.The competing feature block M^2 also grows, but only at the weaker rate of order $\eta'\alpha_2\rho_2^2\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\,\frac{O(\log\log d)}{d}$(Theorem D.1, Lemma C.1).

$$\|M^1 M^{1\top} w_i^{(t+1)}\|_2^2 \approx \left(1+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\right)^{2t}\|M^1(M^1)^\top w_i^{(0)}\|_2^2 \tag{20}$$

$$\|M^2(M^2)^\top w_i^{(t+1)}\|_2^2 \approx \left[\left(1+\eta'\alpha_2\rho_2^2\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\frac{O(\log\log d)}{d}\right)^{2t}\right. \tag{21}$$

Since $\alpha_1\rho_1^2 > \alpha_2\rho_2^2$, the M^1-components are therefore amplified faster than the M^2-components. In contrast, the projection onto the spurious subspace $M^\perp$ remains essentially unchanged throughout this stage,

$$\|M^\perp(M^\perp)^\top w_i^{(t)}\|_2^2 \approx \left(1+\frac{1}{\mathrm{poly}(d)}\right)\|M^\perp(M^\perp)^\top w_i^{(0)}\|_2^2 \tag{22}$$

After $T_1 = O\left(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\,\log\log d}\right)$ number of iterations, the squared norm of every neuron weight $i\in[m]$ reaches $O(1)|w_i^{(0)}|_2^2$. Consequently, by T_1, the neuron weights become increasingly concentrated in the meaningful feature subspace, with a stronger alignment toward the desired block M^1.

Stage 2: After the initial feature separation established in Stage I, we increase the bias to make neuron activations progressively more selective and drive the neurons toward singleton feature

representations. At the beginning of this stage, bias starts to grow above zero. We set the $b_i^{(t)} = \sqrt{\frac{2\log d}{d}}\,\|w_i^{(t)}\|_2$ and subsequently update it according to

$$b_i^{(t+1)} = b_i^{(t)}(1+\eta_{b,t}), \qquad \eta_{b,t} = \max\left\{\frac{\eta}{d}, \frac{\|w_i^{(t+1)}\|_2}{\|w_i^{(t)}\|_2} - 1\right\} \tag{23}$$

The increasing bias introduces an activation threshold that separates the strongly amplified M^1 features from the weaker feature coordinates. In particular, only those neurons in $\mathcal{M}_j^{1*}$ are sufficiently large to exceed the bias remain active and continue to receive significant gradient component. In contrast, the M^2 features, together with the weaker M^1 coordinates, remain below the activation threshold and therefore receive only negligible updates during this stage(Lemma D.2).Consequently, not all important features are consistently amplified in Stage II. Rather, the projection onto M_j^1 stays small for neurons $i \notin \mathcal{M}_j^{1*}$, however it remains dominating for each lucky neuron $i \in \mathcal{M}_j^{1*}$. As a result, the neuron representations get more and more sparse. By the end of step II, lucky neurons successfully specialize to the corresponding desired M^1-features, creating the singleton structure (Theorem D.6)needed for the last step of convergence.

Stage 3: After $T_3 = O\left(\frac{d^{1.01}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}\right)$ iterations of training,as we start our final stage, the contribution from the negative samples increases in significance while the positive-pair gradient continues reinforcing these learned features. Specifically, the negative contrastive term starts to counterbalance the positive gradient as the dominant M^1-coefficients increase, preventing further unlimited growth. The dominant feature coefficients gradually approach a stable phase as a result of the two contributions balancing (Lemma F.7). Since the singleton in Stage 2 is preserved throughout the final stage, each lucky neuron $i \in \mathcal{M}_j^{1*}$, the projection onto its selected feature M_j^1 remains dominant. In particular, throughout the convergence interval $[T_3, T_4]$ the lucky feature coefficients remain bounded away from zero, while the non-selected and spurious components remain small (Theorem F.1). By comparing the learned representation with an optimal sparse learner and applying a descent argument starting at T_3, it is demonstrated that the average contrastive loss converges to a minimal value.

C Generalization Bounds

At iteration t, the population objective is

$$\mathrm{Obj}(f_{\mathcal{G}}^t) = L(f_{\mathcal{G}}^t) + \lambda\|w^{(t)}\|_F^2. \tag{24}$$

The empirical objective using M independent batches $\{(\bar{x}_{1k,\ell}, \bar{x}_{2k,\ell}, N_\ell)\}_{\ell=1}^M$ is

$$\widehat{\mathrm{Obj}}(f_{\mathcal{G}}^t) = \frac{1}{M}\sum_{\ell=1}^{M} L(f_{\mathcal{G}}^t, \bar{x}_{1k,\ell}, B_\ell) + \lambda\|w^{(t)}\|_F^2. \tag{25}$$

The empirical gradient of neuron i is

$$\nabla_{w_i}\widehat{\mathrm{Obj}}(f_{\mathcal{G}}^t) = \frac{1}{M}\sum_{\ell=1}^{M} g_{i,t}^{(\ell)} + \lambda w_i^{(t)}, \tag{26}$$

where

$$g_{i,t}^{(\ell)} := \nabla_{w_i} L(f_{\mathcal{G}}^t, \bar{x}_{1k,\ell}, B_\ell).$$

The population gradient satisfies

$$\nabla_{w_i}\mathrm{Obj}(f_{\mathcal{G}}^t) = \mathbb{E}[g_{i,t}] + \lambda w_i^{(t)}.$$

Hence

$$\nabla_{w_i}\widehat{\mathrm{Obj}}(f_{\mathcal{G}}^t) - \nabla_{w_i}\mathrm{Obj}(f_{\mathcal{G}}^t) = \frac{1}{M}\sum_{\ell=1}^{M}\big(g_{i,t}^{(\ell)} - \mathbb{E}[g_{i,t}]\big). \tag{27}$$

The gradient of the loss function is

$$\nabla_{w_i}\mathcal{L}(f_{\mathcal{G}}^t) = \mathbb{E}\Bigg[-(1-\ell'_{p,t})\cdot h_{i,t}(\bar{x}_k^{(2)})\cdot \mathbf{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\cdot \bar{x}_k^{(1)} + \sum_{\bar{x}_{n,s}\in\mathfrak{B}_k^-} \ell'_{s,t}\cdot h_{i,t}(\bar{x}_{n,s})\cdot \mathbf{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\cdot \bar{x}_k^{(1)}\Bigg] \tag{28}$$

The gradient contribution of one sampled batch is

$$g_{i,t} = \alpha_{i,t}\cdot \mathbf{1}_{\{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}\}}\,\bar{x}_k^{(1)},$$

where

$$\alpha_{i,t} = (1-\ell'_{p,t})\,h_{i,t}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathfrak{B}_k^-}\ell'_{s,t}\,h_{i,t}(x_{n,s}),$$

and $0\leq \ell'_{p,t},\ell'_{s,t}\leq 1$.

The activation is

$$h_{i,t}(\bar{x}_k) = \mathrm{ReLU}(\langle w_i^{(t)},\bar{x}_k\rangle - b_i^{(t)}) - \mathrm{ReLU}(-\langle w_i^{(t)},\bar{x}_k\rangle - b_i^{(t)}).$$

From the Lemma C.2 we have

$$|h_{i,t}(\bar{x}_k)| \leq \tilde{O}(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}})\text{w.h.p1} - e^{-\Omega(\log^2 d)} \tag{29}$$

$$|\alpha_{i,t}| \leq |h_{i,t}(\bar{x}_k^{(2)})| + \sum_{x_{n,s}\in\mathfrak{B}_k^-}|h_{i,t}(x_{n,s})| \leq \tilde{O}(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}}|\mathfrak{B}_k^-|) \tag{30}$$

Therefore,

$$\|g_{i,t}\|_2 \leq \tilde{O}\Big(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}}|\mathfrak{B}_k^-|\Big)\|\bar{x}_k^{(1)}\|_2. \tag{31}$$

Based on the data distribution and the sparse coding model we have,$\|\bar{\xi}_k\|_2 = \mathrm{poly}(d) >> \|M\bar{z}_k\|_2 \approx \tilde{O}(1)$w.h.p. Hence we can find the bound of $\|\bar{x}_k\|_2$ as

$$\|\bar{\xi}_k\|_2 - \|M\bar{z}_k\|_2 \leq \|\bar{x}_k\|_2 \leq \|M\bar{z}_k\|_2 + \|\bar{\xi}_k\|_2 \tag{32}$$

$$\mathrm{poly}(d) - \tilde{O}(1) \leq \|\bar{x}_k\|_2 \leq \tilde{O}(1) + \mathrm{poly}(d)$$

Hence,

$$\|g_{i,t}\|_2 \leq \tilde{O}\Big(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}}|\mathfrak{B}_k^-|\Big)\mathrm{poly}(d)$$

Since we know the $|\mathfrak{B}_k^-| = \mathrm{poly}(d)$

$$\|g_{i,t}\|_2 \leq \tilde{O}\Big(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}}\Big)(\mathrm{poly}(d))^2 \tag{33}$$

Let

$$X_\ell := g_{i,t}^{(\ell)} - \mathbb{E}[g_{i,t}].$$

The X_ℓ are mean-zero vectors satisfying

$$\|X_\ell\|_2 \leq 2R_t \quad \text{and} \quad \mathbb{E}\|X_\ell\|_2^2 \leq V_t.$$

We consider the partial dependency among nodes coming from noe-hop node aggregation as $1+K^2$ Applying coordinate-wise Hoeffdings's inequality and union bounding over coordinates yields that for any $\varepsilon > 0$,

$$\begin{aligned}\Pr\left(\left\|\frac{1}{M}\sum_{\ell=1}^{M} X_\ell\right\|_2 \geq \varepsilon\right) &\leq \exp\left(-\frac{M\epsilon^2}{2V_t^2(1+K^2)}\right). \\ &\leq \exp\left(-\frac{M\epsilon^2 d(K+1)}{2(1+K^2)\|w_i^{(t)}\|_2^2(\mathrm{poly}(d))^2}\right).\end{aligned} \tag{34}$$

For $M = (1+K^2)\mathrm{poly}(d_1)$ sufficiently large

$$\varepsilon = \frac{\|w_i^{(t)}\|_2}{\sqrt{M(K+1)}}\log d$$

The right-hand side becomes $\exp(-\Omega(\log^2 d))$.Hence, with high probability,

$$\left\|\nabla_{w_i}\widehat{\mathrm{Obj}}(f_{\mathcal{G}}^t) - \nabla_{w_i}\mathrm{Obj}(f_{\mathcal{G}}^t)\right\|_2 \leq \frac{\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)\sqrt{(K+1)}}. \tag{35}$$

D Neuron Classification at Initialization

To facilitate the analysis of the stochastic gradient descent trajectory, we introduce the following partitions of the neuron set $[m]$ based on their alignment with the signal subspace.We choose constants $c_0 \in (0,1)$,$\gamma \in (0, 0.01)$, and $c_1 = 2[1+(1-\gamma)c_0]$. We select $c_2 = c_1 - 2\gamma c_0$ and $c_3' = (\frac{\alpha_1\rho_1^2}{\alpha_2\rho_2^2})^2(c_1 - \gamma c_0)$ to be a samll gap between the initial strong and weak neron alignments.

Definition D.1. We define several sets of neurones that will be used in the nerone alignment of the different features in later sections.

- For each signal coordinate $j \in [d]$, define the set of neurons that strongly activate on the j-th feature:

$$\mathcal{M}_j := \left\{i \in [m] \ : \ \langle w_i^{(0)}, M_j\rangle^2 \geq \frac{c_2 \log d}{d}\|M^1 M^{1\top} w_i^{(0)}\|_2^2\right\}. \tag{36}$$

- For each $j \in [d]$, define the set of neurons that selectively activate only on the j-th feature: For $M_j, M_{j'} \in M^1$For $M_j, M_{j'} \in M^1$

$$\mathcal{M}_j^{1\star} := \left\{ i \in [m] \ : \ \begin{array}{ll} \langle w_i^{(0)}, M_j\rangle^2 \geq \frac{c_1 \log d}{d}\|M^1 M^{1\top} w_i^{(0)}\|_2^2, & \\ \langle w_i^{(0)}, M_{j'}\rangle^2 \leq \frac{c_2 \log d}{d}\|M^1 M^{1\top} w_i^{(0)}\|_2^2, & \forall j' \in [d], j' \neq j \end{array} \right\} \tag{37}$$

- For all $j \in [d']$,$M_j \in M^2$ define the all of the neurons activate:

$$\mathcal{M}_j^2 := \left\{i \in [m] \ : \ \langle w_i^{(0)}, M_j\rangle^2 \leq \frac{c_3' \log d}{d}\|M^1 M^{1\top} w_i^{(0)}\|_2^2\right\} \tag{38}$$

Initialization Properties

The following lemma establishes key properties of the neuron weights at initialization ($t = 0$), which will be instrumental in characterizing the subsequent SGD dynamics.

Lemma D.2. *At initialization time $t = 0$, the following statements hold with high probability*

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

$$\|w_i^{(0)}\|_2^2 \in \left[\sigma_0^2 d_1 \left(1 - \tilde{O}\left(\frac{1}{\sqrt{d_1}}\right)\right),\ \sigma_0^2 d_1 \left(1 + \tilde{O}\left(\frac{1}{\sqrt{d_1}}\right)\right)\right] \tag{39}$$

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

$$\|MM^\top w_i^{(0)}\|_2^2 \in \left[\sigma_0^2 \hat{d} \left(1 - \tilde{O}\left(\frac{1}{\sqrt{\hat{d}}}\right)\right),\ \sigma_0^2 \hat{d} \left(1 + \tilde{O}\left(\frac{1}{\sqrt{\hat{d}}}\right)\right)\right]. \tag{40}$$

Further,

$$\|M^1 M^{1\top} w_i^{(0)}\|_2^2 = \sigma_0^2 d \left(1 \pm \tilde{O}(\frac{1}{\sqrt{d}})\right) \qquad \|M^2 M^{2\top} w_i^{(0)}\|_2^2 = \sigma_0^2 d' \left(1 \pm \tilde{O}(\frac{1}{\sqrt{d'}})\right) \tag{41}$$

Where $\hat{d} = d + d'$

(c) With probability at least $1 - o(d^{-4})$, for each $j \in [d]$,

$$\begin{aligned} &|\mathcal{M}_j^\star| \geq \Omega(d^{\gamma c_0/4}) =: \Xi_1, \qquad |\mathcal{M}_j| \leq O(d^{2\gamma c_0}) =: \Xi_2, \\ &\textit{and} \qquad |\mathcal{M}_j^2| \geq m\big(1 - \tilde{\Theta}(\frac{1}{\sqrt{\log d}}) d^{(\frac{\alpha_1 \rho_1^2}{\alpha_2 \rho_2^2})^2 (c_1 - \gamma c_0)/2}\big). \end{aligned} \tag{42}$$

(d) For each $i \in [m]$, define the set of coordinates with negligible magnitude

$$\Lambda_i^1 := \left\{j \in [d] \ :\ |\langle w_i^{(0)}, M_j^1\rangle| \leq \sigma_0/d\right\}. \textit{ Then } |\Lambda_i^1| = O\left(\frac{d}{\mathrm{polylog}(d)}\right) \tag{43}$$

$$\Lambda_i^2 := \left\{k \in [d'] \ :\ |\langle w_i^{(0)}, M_k^2\rangle| \leq \sigma_0/d'\right\}. \textit{ Then } |\Lambda_i^2| = O\left(\frac{d'}{\mathrm{polylog}(d)}\right) \tag{44}$$

(e) For any distinct $j, j' \in [d]$, the intersection size satisfies

$$|\mathcal{M}_j \cap \mathcal{M}_{j'}| \leq O(\log d), \textit{ with probability at least } 1 - o(d^{-4})$$

(f) For each neuron $i \in [m]$:

- *The number of features containing $i \in \mathcal{M}_j$ is bounded by $O(1)$;*
- *The number of coordinates $j \in [d]$ for which $|\langle w_i^{(0)}, M_j^1\rangle| \geq \Omega(\sigma_0 \log^{1/4} d)$ is at most $O(2^{-\sqrt{\log d}} d)$.*
- *The number of coordinates $k \in [d']$ for which $|\langle w_i^{(0)}, M_k^2\rangle| \geq \Omega(\sigma_0 \log^{1/4} d)$ is at most $O(2^{-\sqrt{\log d}} d')$.*

(g) For each $i \in [m]$ and $j \in [d_1] \setminus [d]$, we have $|\langle w_i^{(0)}, M_j^\perp\rangle| \leq O\left(\sqrt{\frac{\log d}{d_1}}\right) \|w_i^{(0)}\|_2$ w.h.p

Proof of Lemma D.2
At initialization, each neuron weight satisfies $w_i^{(0)} \sim \mathcal{N}\left(0, \sigma_0^2 I_{d_1}\right)$, independently across $i \in [m]$. Since M is orthonormal on its columns, $MM^\top$ is the orthogonal projector onto that d-dimensional subspace.

(a) from the standard chi-square concentration inequality gives, for any $t > 0$,

$$\Pr\left[\left|\|w_i^{(0)}\|_2^2 - \sigma_0^2 d_1\right| \geq \sigma_0^2 \left(2\sqrt{\frac{t}{d_1}} + \frac{2t}{d_1}\right)\right] \leq 2e^{-t}. \tag{45}$$

For $t = \Omega(\log^2 d)$. Then for a fixed i,

$$\|w_i^{(0)}\|_2^2 = \sigma_0^2 d_1 \left(1 \pm \tilde{O}(\frac{1}{\sqrt{d_1}})\right) \quad \text{with probability at least } 1 - e^{\Omega(\log^2 d)}$$

(b) can similarly prove as (a), Applying the same concentration inequality, we get $\mathbb{E}[\|MM^\top w_i^{(0)}\|_2^2] = d\sigma_0^2$

$$\|MM^\top w_i^{(0)}\|_2^2 = \sigma_0^2 d\left(1 \pm \tilde{O}(\frac{1}{\sqrt{d}})\right) \tag{46}$$

(c) This is same as the proof of Lemma B.2 in Allen-Zhu & Li (2022) by using the a standard Gaussian concentration, Using the definition of the neurones and the (b) above we can write for fixed i, j,

For $i \in \mathcal{M}_j$

$$p_1 = \Pr[\langle w_i^{(0)}, M_j\rangle^2 \geq c_2\sigma_0 \log d] = \tilde{\Theta}(\frac{1}{\sqrt{\log d}})d^{-c_2/2}, \tag{47}$$

For $i \in \mathcal{M}_j^{1*}$

$$p_2 = \Pr[\langle w_i^{(0)}, M_j\rangle^2 \geq c_1\sigma_0 \log d] = \tilde{\Theta}(\frac{1}{\sqrt{\log d}})d^{-c_1/2} \tag{48}$$

For $i \notin \mathcal{M}_j^2$

$$p_3 = \Pr[\langle w_i^{(0)}, M_j\rangle^2 \geq c_3'\sigma_0 \log d] = \tilde{\Theta}(\frac{1}{\sqrt{\log d}})d^{-c_3'/2} \tag{49}$$

$$\Pr[i \in \mathcal{M}_j^{1*}] = p_1(1-p_2)^{|d_a|-1} = \tilde{\Theta}(\frac{1}{\sqrt{\log d}})d^{-c_1/2} \tag{50}$$

We can find how many number of lucky neurones available at the initialization aligned with $M_j \in \{M^1\}$ from the sum of m Bernoulli random variables as

$$\mathbb{E}[|\mathcal{M}_j^{1*}|] = m*\Pr[i \in \mathcal{M}_j^{1*}] = \tilde{\Theta}(\frac{d^{\gamma c_0}}{\sqrt{\log d}}) \qquad \mathbb{E}[|\mathcal{M}_j|] = m*\Pr[i \in \mathcal{M}_j] = \tilde{\Theta}(\frac{d^{2\gamma c_0}}{\sqrt{\log d}}) \tag{51}$$

We can show asymptotically for large d,we have $\mathbb{E}[|\mathcal{M}_j^{1*}|] \geq \Omega(d^{\frac{\gamma c_0}{2}}), \mathbb{E}[|\mathcal{M}_j|] \leq \Omega(d^{2\gamma c_0})$,and $\mathbb{E}[|\mathcal{M}_j^2|] \geq m - \Omega(\frac{1}{\sqrt{\log d}}d^{(\frac{\alpha_1\rho_1^2}{\alpha_2\rho_2^2})^2(c_1-\gamma c_0)/2})$. By applying Chernoff concentration we can derived ,$|\mathcal{M}_j^{1*}| \geq \Omega(d^{\frac{\gamma c_0}{4}})$ with $1 - o(\frac{1}{d^3})$ probability.

(d) we can divide the neurones into two sets in such a way that with strongly align and weakly aligned neurones as $\Lambda_i := \left\{j \in [d] \ : \ |\langle w_i^{(0)}, M_j\rangle| \leq \sigma_0/d\right\}$. Here we are interested about the size of this set and it follows a binomial distribution. For a fixed $i \in [m]$, it is the sum of d independent gaussian random variables which satisfy the condition of the Λ_i. Among the d samples we find the probability of nerones satisfy $|\langle w_i^{(0)}, M_j\rangle| \leq \sigma_0/d$.From the Gaussian concentration

$$\Pr\left[|\langle w_i^{(0)}, M_j\rangle| \leq \sigma_0/d\right] = O\left(\frac{1}{d}\right) \tag{52}$$

By applying Chernoff for the $Bino(d, 1/d)$ gives the desired bound on $|\Lambda_i|$.

$$\Pr(|\Lambda_i| \geq t) \leq (\frac{e\mu}{t})^t \tag{53}$$

$$\Pr(|\Lambda_i| \geq \frac{d}{\text{polylog}d}) \leq (\frac{e \cdot \text{polylog}d}{d})^{\frac{d}{\text{polylog}d}} \tag{54}$$

Hence, we can argue that with high probability $|\Lambda_i| \leq O\left(\frac{d}{\text{polylog}d}\right)$ We prove that for $j \neq j'$,

$$|M_j \cap M_{j'}| \leq O(\log d) \tag{55}$$

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

(e) We prove this part by using the part (b) and (c) above. Fix $i \in [m]$. The event $i \in M_j \cap M_{j'}$ means both

$$\langle w_i^{(0)}, M_j\rangle^2 \geq \frac{c_2 \log d}{d}\|M^1 M^{1\top} w_i^{(0)}\|_2^2 \tag{56}$$

and

$$\langle w_i^{(0)}, M_{j'}\rangle^2 \geq \frac{c_2 \log d}{d} \|M^1 M^{1\top} w_i^{(0)}\|_2^2. \tag{57}$$

Using part (b), this implies both Gaussian coordinates exceed a threshold of order $\sqrt{c_2 \log d}$. Since the coordinates $\langle w_i^{(0)}, M_j\rangle$, $\langle w_i^{(0)}, M_{j'}\rangle$ are independent Gaussians, the probability of this double event is roughly the square of the single-event probability:

$$\Pr[i \in M_j \cap M_{j'}] \leq \tilde{O}(d^{-c_2}). \tag{58}$$

Therefore

$$|M_j \cap M_{j'}| = \sum_{i=1}^{m} \mathbf{1}\{i \in M_j \cap M_{j'}\} \tag{59}$$

is a sum of independent Bernoulli variables with mean

$$\mu_{j,j'} \leq m\tilde{O}(d^{-c_2}).$$

For the chosen constants, this mean is $O(1)$. Then a Chernoff bound gives

$$\Pr[|M_j \cap M_{j'}| > C \log d] \leq d^{-6} \tag{60}$$

for large enough constant C. From the union bound over the d^2 pairs (j, j') yields

$$|M_j \cap M_{j'}| \leq O(\log d) \quad \forall j \neq j'$$

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

(f) From (b) and the neuron initialization we have for $i \in M_j$,

$$\langle w_i(0), M_j\rangle^2 \geq c^2 \sigma_0^2 \log d.$$

This is a large Gaussian-tail event and the probability of this happening for one fixed j is very small:

$$\Pr(i \in M_j) \leq O\Big(d^{-1-\Omega(1)}\Big). \tag{61}$$

Now there are only d possible choices of j. Therefore the number of such j exceed the bound behaves like a binomial random variable with very small success probability Binomial$\big(d,\, O(d^{-1-\Omega(1)})\big)$ with expectation of $d \cdot O\big(d^{-1-\Omega(1)}\big) = O\big(d^{-\Omega(1)}\big)$. By Chernoff concentration, the actual number is also at most $O(1)$ with high probability.

For fixed neuron $i \in [m]$,and $j \in [d], \langle w_i^{(0)}, M_j\rangle$ are i.i.d Gaussian random variables with the distribution of $\mathcal{N}(0, \sigma_0^2)$.Here we interest at the event $|\langle w_i^{(0)}, M_j\rangle| \geq \Omega(\sigma_0 \log^{1/4} d)$.

From the gaussian concentration bound we have

$$\Pr\Big(|\langle w_i^{(0)}, M_j^1\rangle| \geq \Omega(\sigma_0 \log^{1/4} d)\Big) \leq 2e^{-\log^{1/2} d} = 2^{-\Omega(\sqrt{\log d})} \tag{62}$$

Since we have $[d]$ cordinates in the $i \in \mathcal{M}_j^{1*}$. we can find the number of $j \in [d]$ s satisfy the above condition by Bernoulli principle by $\leq O(2^{-\sqrt{\log d}} d)$. We can follow the similar approach for $M_k^2, k \in [d']$ to derive the bound.

(g)Similar to the proof of (f), $\langle w_i^{(0)}, M_j^{\perp}\rangle$ is also a gaussian distributed i.i.d random variable.Therefore, we can apply the gaussian concentration bound to upper bound the $\langle w_i^{(0)}, M_j^{\perp}\rangle$ with high probability $1 - e^{-\Omega(\log d)}$.

$$\Pr\Big(|\langle w_i^{(0)}, M_j^{\perp}\rangle| \geq C\sigma_0 \sqrt{\log d}\Big) \leq 2e^{-\Omega(\log d)}. \tag{63}$$

Hence, with high probability,

$$|\langle w_i^{(0)}, M_j^{\perp}\rangle| \leq O\Big(\sqrt{\frac{\log d}{d_1}}\Big) \|w_i^{(0)}\|_2 \tag{64}$$

E GRADIENT COMPUTATIONS

We compute the gradient of the positive samples for each $t \leq T_1$ as the following. In order to minimize the computational complexity, we assume the bias $b_i^{(t)}$ is zero in the first stage.

Lemma E.1. *Let i^{th} neuron at iteration $t \leq T_1$ is taken as $h_{i,t}(.)$, We can derived the general expression considering the the randomness and the independence of the $\bar{\xi}_k^{(1)}$ and $\bar{\xi}_k^{(2)}$ as follows. Notice that $\bar{z}_{k,j}^{(1)}$ and $\bar{z}_{k,j'}^{(2)}$ are independant to each other if $j \neq j'$*

$$\begin{aligned}
&\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq 0}\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq 0}\langle \bar{x}_k^{(1)}, M_j\rangle\Big] \\
&= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[\langle w_i^{(t)}, M\bar{z}_k^{(2)} + \bar{\xi}_k^{(2)}\rangle\langle M\bar{z}_k^{(1)} + \bar{\xi}_k^{(1)}, M_j\rangle\Big] \\
&= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\sum_{j'=1}^{d}\langle w_i^{(t)}, M_{j'}\rangle \bar{z}_{2k,j'}\bar{z}_{k,j}^{(1)}\right] \\
&= \langle w_i^{(t)}, M_j\rangle \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[\bar{z}_{k,j}^{(2)}\bar{z}_{k,j}^{(1)}\Big]
\end{aligned} \tag{65}$$

(a) For each $j \in [d/2], M_j \in M^1$ we have

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq 0}\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq 0}\langle \bar{x}_k^{(1)}, M_j\rangle\Big] \approx \big(\alpha_1\rho_1^2 + \frac{1-\alpha_1\rho_1}{K+1}\big)\frac{O(\log\log d)}{d}\langle w_i^{(t)}, M_j\rangle \tag{66}$$

(b) For each $j \in [d] \setminus [d/2], M_j \in M^2$ we have

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq 0}\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq 0}\langle \bar{x}_k^{(1)}, M_j\rangle\Big] \approx \big(\alpha_2\rho_2^2 + \frac{1-\alpha_2\rho_2}{K+1}\big)\frac{O(\log\log d)}{d}\langle w_i^{(t)}, M_j\rangle \tag{67}$$

(c) For each $j \in [d_1] \setminus [d], M_j \in M^{\perp}$ we have

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq 0}\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq 0}\langle \bar{x}_k^{(1)}, M_j^{\perp}\rangle\Big] = 0 \tag{68}$$

Proof of Lemma E.1

(a) For each $j \in [d/2]$ and at iteration $t \leq T_1$, for all neurones $i \in [m]$ we can simplify

$$\begin{aligned}
&\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq 0}\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq 0}\langle \bar{x}_k^{(1)}, M_j\rangle\Big] \\
&= \langle w_i^{(t)}, M_j\rangle\Big[\big(\frac{\rho_1 K+1}{K+1}\big)^2\Pr(|\bar{z}_{1i,j}| = \frac{\rho_1 K+1}{K+1}) \\
&\quad + \big(\frac{1}{K+1}\big)^2\Pr(|z_{1i,j}| = \frac{1}{K+1})\Big] \\
&= \langle w_i^{(t)}, M_j\rangle\Big[\big(\frac{\rho_1 K+1}{K+1}\big)^2\cdot\frac{\alpha_1 O(\log\log d)}{d} \\
&\quad + \big(\frac{1}{K+1}\big)^2\frac{(K(1-\alpha_1\rho_1)O(\log\log d)}{d}\Big] \\
&\approx \big(\alpha_1\rho_1^2 + \frac{1-\alpha_1\rho_1}{K+1}\big)\frac{O(\log\log d)}{d}\langle w_i^{(t)}, M_j\rangle
\end{aligned} \tag{69}$$

(b) For each $j \in [d] \setminus [d/2]$ and $M_j \in M^2$ we assume $\rho_2 \in (0,1)$ fraction among (K+1) non-zeros of the $\bar{z}_{k,j}^{(1)}$ and $\bar{z}_{k,j}^{(2)}$ overlapping.

$$\begin{aligned}
&\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq 0}\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq 0}\langle \bar{x}_k^{(1)}, M_j\rangle\Big] \\
&\approx \big(\alpha_2\rho_2^2 + \frac{1-\alpha_2\rho_2}{K+1}\big)\frac{O(\log\log d)}{d}\langle w_i^{(t)}, M_j\rangle
\end{aligned} \tag{70}$$

(c) For each $j \in [d_1] \setminus [d]$ and $M_j \in M^{\perp}$ due to the independance and the orthogonality of the $M_j$and $M_j^{\perp}$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq 0}\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq 0}\langle \bar{x}_k^{(1)}, M_j^{\perp}\rangle\right] = 0 \tag{71}$$

Lemma E.2. *Each aggregated node $\bar{x}_k$ can be written using sparse coding model notation $\bar{x}_k = M\bar{z}_k + \bar{\xi}_k$,$w_i \in \mathbb{R}^{d_1}$,we can derive the following results*

a) High probability bound for the sparse signal is

$$\Pr\Big(\langle w_i,\, M\bar{z}_{2k}\rangle \geq \|w_i^{(t)}\|_2 \max_{j\in[d]} \|M_j\|_\infty \bar{z}_{k,j}^{(2)} \log d\Big) \leq e^{-\Omega(\log^2 d)} \tag{72}$$

b) High probability bound for the dense noise is

$$\Pr\Big(\langle w_i, \bar{\xi}_{2k}\rangle \geq \frac{\|w_i^{(t)}\|_2 \log d}{\sqrt{d(K+1)}}\Big) \leq e^{-\Omega(\log^2 d)} \tag{73}$$

c) For any $j \in [d_1] \setminus [d]$, $\big|\langle M\bar{z}_{1k} + \bar{\xi}_{1k},\, M_j^{\perp}\rangle\big| \leq \tilde{O}(1/\sqrt{(K+1)d})$ w.h.p $1 - e^{-\Omega(\log^2 d)}$

d) For any $j \in [d_1] \setminus [d]$, with probability at least $1 - e^{-\Omega(\log^2 d)}$ over the randomness of $\bar{z}_{1k}, \bar{z}_{2k}, \bar{\xi}_{1k}, \bar{\xi}_{2k}$, we have

$$\Big|\big\langle w_i,\, M\bar{z}_{2k} + \bar{\xi}_{2k}\big\rangle \big\langle M\bar{z}_{1k} + \bar{\xi}_{1k},\, M_j^{\perp}\big\rangle\Big| \;\leq\; \widetilde{O}\Big(\frac{\|w_i\|_2}{d^2(K+1)}\Big), \tag{74}$$

Proof of Lemma E.2
a) High probability bound for the sparse signal is

$$\Pr\Big(\langle w_i,\, M\bar{z}_{2k}\rangle \geq \|w_i^{(t)}\|_2 \max_{j\in[d]} \|M_j\|_\infty \bar{z}_{k,j}^{(2)} \log d\Big) \leq e^{-\Omega(\log^2 d)} \tag{75}$$

$\langle w_i,\, M\bar{z}_{2k}\rangle$ can be written as the sum of sub-Gaussian random variables, each with the variance $O(w_{i,s}\mathrm{Supp}(\bar{z}_{k,j}^{(2)})/\sqrt{d_1})$ where $w_{i,s}$ is the s^{th} element where $s \in [d_1]$ of the i^{th} neurone and $\mathrm{Supp}(.)$ is the suprimum of the j^{th} element of z_{2k}. Then we can get the above bound by applying the concentration bound for the subgaussian random variables as below.

$$\Pr\Big(\langle w_i,\, M\bar{z}_{2k}\rangle \geq t\Big) \leq 2\exp\Big(-c\frac{t^2}{\|w_i^{(t)}\|_2^2 \max_{j\in[d]} \|M_j\|_\infty^2 (\bar{z}_{k,j}^{(2)})^2}\Big) \tag{76}$$

We can achieve the above by choosing $t = \tilde{O}(\frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}}\rho)$, we can ensure that $\langle w_i,\, M\bar{z}_{2k}\rangle \leq \tilde{O}(\frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}}\rho)$ w.h.p $1 - e^{-\Omega(\log^2 d)}$

b) High probability bound for the dense noise is

$$\Pr\Big(\langle w_i, \bar{\xi}_{2k}\rangle \geq \frac{\|w_i^{(t)}\|_2 \log d}{\sqrt{d(K+1)}}\Big) \leq e^{-\Omega(\log^2 d)} \tag{77}$$

We can obtain these high probability bounds using the concentration of Gaussian variables since our noise variance $\sigma_{\bar{\xi}_{2k}}^2 = \frac{\sqrt{\log d}}{d(K+1)}$. By combining the results of both (a) and (b), we can obtain

$$\langle w_i,\, M\bar{z}_{2k} + \bar{\xi}_{2k}\rangle \leq \frac{\|w_i^{(t)}\|_2 \log d}{\sqrt{d(K+1)}} \;\text{ w.h.p }\; 1 - e^{-\Omega(\log^2 d)} \tag{78}$$

c) First start with the second term. Since the column-orthonormality of the dictionary matrix M, we have $\langle M\bar{z}_{1k}, M_j^{\perp}\rangle = 0$. Hence, we can simplify the second term as

$$\big\langle M\bar{z}_{1k} + \bar{\xi}_{1k},\, M_j^{\perp}\big\rangle = \big\langle \bar{\xi}_{1k}, M_j^{\perp}\big\rangle \tag{79}$$

By applying the gaussian concentration to bound the above term for all $t \geq 0$

$$\Pr(|\langle M\bar{z}_{1k} + \bar{\xi}_{1k},\, M_j^{\perp}\rangle| > t) \leq 2\exp\left(-c\,\frac{t^2}{\sigma_\xi^2}\right) \tag{80}$$

We can bound the above $|\langle M\bar{z}_{1k} + \bar{\xi}_{1k},\, M_j^{\perp}\rangle| \leq \tilde{O}(1/\sqrt{(K+1)d})$ w.h.p $1 - e^{-\Omega(\log^2 d)}$

For each $j \in [d]$ we have w.h.p

$$\langle w_i,\, M\bar{z}_{2k} + \bar{\xi}_{2k}\rangle \leq \frac{\|w_i^{(t)}\|_2 \log d}{\sqrt{d(K+1)}} \tag{81}$$

(d) We have from the proofs in above, by the union bound we can conclude the proof of Lemma E.2

$$\left|\langle w_i,\, M\bar{z}_{2k} + \bar{\xi}_{2k}\rangle\, \langle M\bar{z}_{1k} + \bar{\xi}_{1k},\, M_j^{\perp}\rangle\right| \;\leq\; \widetilde{O}\Big(\frac{\|w_i\|_2}{d^2(K+1)}\Big),\;\; \text{w.h.p}\;\; 1 - e^{-\Omega(\log^2 d)} \tag{82}$$

Since we know $0 < \alpha, \rho < 1$ and $K, d >> 1$, we can simplify the expressions above..Further we have $\mathbb{E}\left[\sum_{j" \neq j'} \langle w_i, M_{j"}\rangle \bar{z}_{2i,j'} \cdot \bar{z}_{1i,j}\right] = 0$ due to independance.

Lemma E.3. *Let the each view $v \in \{1,2\}$ of the node $l, k \in \mathcal{V}$ with $|\hat{\mathcal{N}}_{v(\nu_k)}| = |\hat{\mathcal{N}}_{v(\nu_l)}| = K$, for all $i \in [m]$ when $\alpha_1\rho_1^2(K+1)$ is comparably larger than $1-\alpha\rho$ and $\alpha_2\rho_2^2(K+1)$, we have following*

1 . For each $M_j \in M^1$ with high probability $1 - e^{-\Omega(\log^2 d)}$

$$|h_{i,t}(\bar{x}_{vl})|\,|\langle \bar{x}_{vk}, M_j\rangle| \leq \tilde{O}\Big(\frac{\|w_i^{(t)}\|_2}{d}\sqrt{\frac{\alpha_1\rho_1^2 \log\log d}{K+1}}\Big) \tag{83}$$

2 . For each $M_j \in M^2$ with high probability $1 - e^{-\Omega(\log^2 d)}$

$$|h_{i,t}(\bar{x}_{vl})|\,|\langle \bar{x}_{vk}, M_j\rangle| \leq \tilde{O}\Big(\frac{\|w_i^{(t)}\|_2}{d}\sqrt{\frac{\alpha_2\rho_2^2 \log\log d}{K+1}}\Big) \tag{84}$$

3 . For each $M_j \in M^{\perp}$ with high probability $1 - e^{-\Omega(\log^2 d)}$

$$|h_{i,t}(\bar{x}_{vl})|\,|\langle \bar{x}_{vk}, M_j^{\perp}\rangle| \leq \tilde{O}\Big(\frac{\|w_i^{(t)}\|_2}{d(K+1)}\Big) \tag{85}$$

4 . For each $M_j \in M^1$ with high probability $1 - e^{-\Omega(\log^2 d)}$

$$\Big|\mathbb{E}\sum_{l\neq k}\big[\ell'_{l,t}(\bar{x}_{vk}, B)h_{i,t}(\bar{x}_{vl})\langle \nabla_{w_i} h(\bar{x}_{vk}), M_j\rangle\Big| \leq \tilde{O}\Big(\frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d}\|w_i^{(t)}\|_2\sqrt{\frac{\alpha_1\rho_1^2\log\log d}{K+1}}\Big) \tag{86}$$

5 . For each $M_j \in M^2$ with high probability $1 - e^{-\Omega(\log^2 d)}$

$$\Big|\mathbb{E}\sum_{l\neq k}\big[\ell'_{l,t}(\bar{x}_{vk}, B)h_{i,t}(\bar{x}_{vl})\langle \nabla_{w_i} h(\bar{x}_{vk}), M_j\rangle\Big| \leq \tilde{O}\Big(\frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d}\|w_i^{(t)}\|_2\sqrt{\frac{\alpha_2\rho_2^2\log\log d}{K+1}}\Big) \tag{87}$$

6 . For each $M_j \in M^{\perp}$ with high probability $1 - e^{-\Omega(\log^2 d)}$

$$\Big|\mathbb{E}\sum_{l\neq k}\big[\ell'_{l,t}(\bar{x}_{vk}, B)h_{i,t}(\bar{x}_{vl})\langle \nabla_{w_i} h(\bar{x}_{vk}), M_j^{\perp}\rangle\Big| \leq \tilde{O}\Big(\frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d(K+1)}\|w_i^{(t)}\|_2\Big) \tag{88}$$

Proof of Lemma E.3
From Lemma C3 Allen-Zhu & Li (2022) Assume $\sum_{r=1}^{m} \|w_r^{(t)}\|_2^2 \leq o(\tau/d)$ at initiallization, Then, with high probability over the draw of the anchor and negatives in B,

$$\Big|\ell'_{a,t}(\bar{x}_{lv}, B) - \frac{1}{|B|}\Big| \;\leq\; \widetilde{\mathcal{O}}\Big(\frac{\sum_{r=1}^{m} \|w_r^{(t)}\|_2^2}{\tau|B|}\Big), \qquad \forall a \in B,\;\; v \in \{1,2\}. \tag{89}$$

Apply Lemma C3 to get $\big|\ell'_{\ell,t}(x_k^{(v)}, B)\big| \leq |B|^{-1}\big(1 + \widetilde{\mathcal{O}}(\sum_r \|w_r^{(t)}\|_2^2/\tau)\big)$. From the Lemma E.2, We can find the bound of the below

$$\Big|\mathbb{E}\sum_{l\neq k}\big[\ell'_{l,t}(\bar{x}_{vk}, B) h_{i,t}(\bar{x}_{vl})\langle\nabla_{w_i} h(\bar{x}_{vk}), M_j\rangle\Big| \leq \sum_{\ell\neq k}\mathbb{E}\Big[\big|\ell'_{\ell,t}(\bar{x}_{vk}, B)\big|\,|h_{i,t}(\bar{x}_{vl})|\,|\langle\bar{x}_{vk}, M_j\rangle|\Big]. \tag{90}$$

We analyze bound of each term with high probability then we can derive the bound for the products w.h.p

$$|h_{i,t}(\bar{x}_{vl})|\,|\langle\bar{x}_{vk}, M_j\rangle| = |\langle w_i^{(t)}, \bar{x}_{vl}\rangle||\langle M\bar{z}_{vk}, M_j\rangle| \tag{91}$$

From the Lemma E.2 (b) and (c), we can bound the $|\langle w_i^{(t)}, \bar{x}_{vl}\rangle| \;\leq\; \tilde{O}(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}})$ w.h.p $1 - e^{-\Omega(\log^2 d)}$
The second term is a random variable with $\mathbb{E}[\langle M\bar{z}_{vk}, M_j\rangle] = 0$. We can find the variance for each $j \in [d]$and $j \in [d']$for given $\|M_j\|_\infty \leq \tilde{O}\Big(\frac{1}{\sqrt{d_1}}\Big)$
For each $j \in [d]$ and $M_j \in M^1$ we have

$$|\langle M\bar{z}_{vk}, M_j\rangle| \leq O\Big(\sqrt{\frac{\big(\alpha_1\rho_1^2(K+1) + 1 - \alpha_1\rho_1\big)\log\log d}{d(K+1)}}\Big) \approx O\Big(\sqrt{\frac{\alpha_1\rho_1^2\log\log d}{d}}\Big) \tag{92}$$

For each $j \in [d] \setminus [d/2]$ and $M_j \in M^2$,we have the following for $\alpha_2\rho_2 \in (0,1)$

$$|\langle M\bar{z}_{vk}, M_j\rangle| \leq O\Big(\sqrt{\frac{\alpha_2\rho_2^2\log\log d}{d}}\Big) \tag{93}$$

From combining the results of the bounds derive in earlier steps we can derive (4),(5) and (6) as in the mentioned in the Lemma.

F Initial Stage Learning

This section we analyse the first-stage training process.The main outcomes we are focusing in here is the growth of the weight $w_i^{(t)}$ for all features $\{M\}_{j\in[d]} \cup \{M^\perp\}_{j\in[d_1]\setminus[d]}$. Here we define the time $T_1 = O\Big(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\log\log d}\Big)$ to complete the training of the first stage as the iteration to satisfy for all the neurones $i \in [m]$, $\|w_i^{(t)}\|_2^2$ reach to $O(\|w_i^{(0)}\|_2^2)$.

Theorem F.1. *At iteration $t = T_1$, we have the following results:*

- *(a) $\|MM^\top w_i^{(T_1)}\|_2^2 \geq \|w_i^{(T_1)}\|_2^2/2$ for all $i \in [m]$, and we write the neuron dependent corresponding projection onto the M^1 subspace as $\|M^1M^{1\top}w_i^{(T_1)}\|_2^2 = L_i\|w_i^{(T_1)}\|_2^2$*
- *(b) For each $j \in [d]$ and each $i \in \mathcal{M}_j^{1*}$, we have $|\langle w_i^{(T_1)}, M_j\rangle| \geq (1 + c_0\gamma)\sqrt{2\frac{\log d}{d}}\|w_i^{(T_1)}\|_2$*
- *(c) For each $j \in [d]$ and each $i \notin \mathcal{M}_j$, we have $|\langle w_i^{(T_1)}, M_j\rangle| \leq (1 - 2c_0\gamma)\sqrt{2\frac{\log d}{d}}\|w_i^{(T_1)}\|_2$*
- *(d) For each $j \in [d']$ and each $i \in \mathcal{M}_j^2$, we have $|\langle w_i^{(T_1)}, M_j\rangle| \leq (1 - c_0\gamma)\sqrt{2\frac{\log d}{d}}\|w_i^{(T_1)}\|_2$*

(e) For each $i \in [m]$, $|\langle w_i^{(T_1)}, M_j\rangle| \geq \frac{\log^{1/4} d}{\sqrt{d}} \|w_i^{(T_1)}\|_2$, for at most $O\left(\frac{d}{2^{\sqrt{\log d}}}\right)$ many $j \in [d/2]$;

(f) For each $i \in [m]$ and $j \in [d_1 - (d' + d)]$, we have $|\langle w_i^{(T_1)}, M_j^{\perp}\rangle| \leq O\left(\sqrt{\frac{\log d}{d_1}}\right) \|w_i^{(T_1)}\|_2$

Suppose we have for $i \in [m]$ we have $\sum_{r=1}^m \|w_r^{(t)}\|_2^2 \leq O(\tau/d)$ For a fixed neurone $i \in [m]$ and $j \in [d], M_j \in M^1$ we can apply the SGD for the computed gradient with each feature M^1, M^2 and $M^{\perp}$

$$
\begin{aligned}
\langle w_i^{(t+1)}, M_j\rangle &= \langle w_i^{(t)}, M_j\rangle - \eta\langle \nabla_{w_i} \mathsf{Obj}(f_{\mathcal{G}}^t), M_j\rangle \\
&= (1-\eta\lambda)\langle w_i^{(t)}, M_j\rangle + \eta \mathbb{E}_{x_p^+, x_p^{++}}\left[(1-\ell'_{p,t}(x_p^+, B)) \cdot h_{i,t}(x_p^{++}) \cdot \langle \nabla_{w_i} h_{i,t}(x_p^+), M_j\rangle\right] \\
&\quad - \eta \sum_{x_{n,s} \in \mathcal{B}_k^-} \mathbb{E}\left[\ell'_{s,t}(x_p, B) \cdot h_{i,t}(x_{n,s}) \cdot \langle \nabla_{w_i} h(x_p^+), M_j\rangle\right] + \frac{\eta \|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)\sqrt{(K+1)}} \\
&= (1 - \eta\lambda + \eta(\alpha_1\rho_1^2 + \frac{1-\alpha_1\rho_1}{K+1})\frac{\log\log d}{d})\langle w_i^{(t)}, M_j\rangle \\
&\quad \pm \tilde{O}\Big(\eta \frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d}\|w_i^{(t)}\|_2 \sqrt{\frac{\alpha_1\rho_1^2 \log\log d}{K+1}}\Big) + \frac{\eta\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)\sqrt{(K+1)}} \\
\langle w_i^{(t+1)}, M_j\rangle &\leq \Big(1 - \eta\lambda + \eta\alpha_1\rho_1^2 \frac{\log\log d}{d}\Big)\langle w_i^{(t)}, M_j\rangle + \tilde{O}\Big(\eta \frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d}\|w_i^{(t)}\|_2 \sqrt{\frac{\alpha_1\rho_1^2 \log\log d}{K+1}}\Big) \\
\langle w_i^{(t+1)}, M_j\rangle &\leq \Big(1 - \eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}} + \eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)\langle w_i^{(t)}, M_j\rangle \\
&\quad + \tilde{O}\Big(\eta' \sqrt{\frac{K+1}{\alpha_1\rho_1^2}} \frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d}\|w_i^{(t)}\|_2 \sqrt{\frac{\alpha_1\rho_1^2 \log\log d}{K+1}}\Big) \\
\langle w_i^{(t+1)}, M_j\rangle &\leq \Big(1 - \eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}} + \eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)\langle w_i^{(t)}, M_j\rangle \\
&\quad + \tilde{O}\Big(\eta' \frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d}\|w_i^{(t)}\|_2 \sqrt{\log\log d}\Big)
\end{aligned} \tag{94}
$$

For $M_j \in M^2$ we used similar approach by applying stochastic gradient step of $w_i^{(t)}$ toward the direction of sparse feature $\{M^2\}_{j \in [d']}$

$$
\begin{aligned}
\langle w_i^{(t+1)}, M_j\rangle &\leq \Big(1 - \eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}} + \eta'\alpha_2\rho_2^2\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\frac{\log\log d}{d}\Big)\langle w_i^{(t)}, M_j\rangle \\
&\quad + \tilde{O}\Big(\eta' \frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d}\|w_i^{(t)}\|_2 \sqrt{\frac{\alpha_2\rho_2^2}{\alpha_1\rho_1^2}\log\log d}\Big)
\end{aligned} \tag{95}
$$

For $M_j^\perp \in M^\perp$ we used similar approach by applying stochastic gradient step of $w_i^{(t)}$ toward the direction of sparse feature $\{M^\perp\}_{j\in[d_1]\setminus[d]}$

$$\begin{aligned}
\langle w_i^{(t+1)}, M_j^\perp\rangle &= \langle w_i^{(t)}, M_j^\perp\rangle - \langle \nabla_{w_i}\text{Obj}(f_{\mathcal{G}}^t), M_j^\perp\rangle \\
&= (1-\eta\lambda)\langle w_i^{(t)}, M_j^\perp\rangle + \eta\mathbb{E}_{x_p^+, x_p^{++}}\left[(1-\ell'_{p,t}(x_p^+, B))\cdot h_{i,t}(x_p^{++})\cdot\langle\nabla_{w_i}h_{i,t}(x_p^+), M_j^\perp\rangle\right] \\
&\quad - \eta\sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}\left[\ell'_{s,t}(x_p, B)\cdot h_{i,t}(x_{n,s})\cdot\langle\nabla_{w_i}h(x_p^+), M_j^\perp\rangle\right] + \frac{\eta\|w_i^{(t)}\|_2}{\text{poly}(d_1)\sqrt{(K+1)}} \\
&= (1-\eta\lambda)\langle w_i^{(t)}, M_j^\perp\rangle \pm \tilde{O}\Big(\eta\frac{\sum_{r=1}^m\|w_r^{(t)}\|_2^2}{\tau d(K+1)}\|w_i^{(t)}\|_2\Big) + \frac{\eta\|w_i^{(t)}\|_2}{\text{poly}(d_1)\sqrt{(K+1)}} \\
\langle w_i^{(t+1)}, M_j^\perp\rangle &\le (1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}})\langle w_i^{(t)}, M_j\rangle + \tilde{O}\Big(\eta'\frac{\sum_{r=1}^m\|w_r^{(t)}\|_2^2}{\tau d\sqrt{\alpha_1\rho_1^2(K+1)}}\|w_i^{(t)}\|_2\Big)
\end{aligned} \tag{96}$$

EARLY STAGE LEARNING DYNAMICS

Through this section, we will prove that for each neuron $i \in [m]$ primarily learns the meaningful features in $M = [M^1|M^2]$ and largely suppresses the spurious directions in $M^\perp$.Here we set $T_1 = O\Big(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\log\log d}\Big)$ to be the required number of iterations to $\|w_i^{(t)}\|_2^2 \ge O(1)\|w_i^{(0)}\|_2^2$ for all $i \in [m]$. From initialization we have $\sum_{r=1}^m\|w_r^{(t)}\|_2^2 \le \frac{1}{\text{poly}(d)}$

Induction Hypothesis F.2. following holds for $t \le T_1$

a) $$\|M^\perp(M^\perp)^\top w_i^{(t+1)}\|_2^2 \le \Big(1+\frac{1}{\text{poly}(d)}\Big)\|M^\perp(M^\perp)^\top w_i^{(0)}\|_2^2 \tag{97}$$

b) $$\|M^1M^{1\top}w_i^{(t+1)}\|_2^2 \le \Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)^{2t}\|M^1(M^1)^\top w_i^{(0)}\|_2^2 + O\Big(\frac{1}{\alpha_1\rho_1^2\tau^2 d(K+1)}\Big)\|M^1M^{1\top}w_i^{(0)}\|_2^2 \tag{98}$$

c) $$\|M^1M^{1\top}w_i^{(t+1)}\|_2^2 \ge \Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)^{2t}\|M^1(M^1)^\top w_i^{(0)}\|_2^2 - O\Big(\frac{1}{\alpha_1\rho_1^2\tau^2 d(K+1)}\Big)\|M^1M^{1\top}w_i^{(0)}\|_2^2 \tag{99}$$

d) $$\|M^2(M^2)^\top w_i^{(t+1)}\|_2^2 \le \Big[\Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\alpha_2\rho_2^2\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\frac{O(\log\log d)}{d}\Big)^{2t} + \tilde{O}\big(\frac{\alpha_2\rho_2^2}{\alpha_1^2\rho_1^4\tau^2 d(K+1)}\big)\Big]\|M^2(M^2)^\top w_i^{(0)}\|_2^2 \tag{100}$$

Proof of Induction Hypothesis F.2:
We first derive the upper and lower bounds of $\|M^1M^{1\top}w_i^{(t+1)}\|_2^2$ at $t \le T_1$. For each $j \in [d/2]$ and

$M_j \in M^1$

$$
\begin{aligned}
\|M^1 M^{1\top} w_i^{(t+1)}\|_2^2 &= \sum_{j\in[d/2]} \Big[\Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)\langle w_i^{(t)}, M_j\rangle \\
&\quad \pm \tilde{O}\Big(\eta'\frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d}\|w_i^{(t)}\|_2\sqrt{\log\log d}\Big)\Big]^2 \\
&\le \sum_{j\in d/2}\langle w_i^{(0)}, M_j\rangle^2\cdot\Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)^{2t} \\
&\quad + \sum_{\substack{j\in[d/2]\\ j\in\mathcal{E}^{(0)}}} O\left(\frac{(t+1)^2\eta'^2\log\log d}{\tau^2 d^2(\mathrm{poly}(d))^2}\max_{t\le t_1}\|w_i\|_2^2\right) \\
&\le \sum_{j\in d/2}\langle w_i^{(0)}, M_j\rangle^2\cdot\Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2}\frac{\log\log d}{d}\Big)^{2t} \\
&\quad + \sum_{\substack{j\in[d/2]\\ j\in\mathcal{E}^{(0)}}} O\left(\frac{(t+1)^2\eta'^2\log\log d}{\tau^2 d^2(\mathrm{poly}(d))^2}\max_{t\le t_1}\|w_i\|_2^2\right) \\
&\le \sum_{j\in d/2}\langle w_i^{(0)}, M_j\rangle^2\cdot\Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)^{2t} \\
&\quad + \sum_{\substack{j\in[d/2]\\ j\in\mathcal{E}^{(0)}}} O\left(\frac{(t+1)^2\eta'^2\log\log d}{\tau^2 d^2(\mathrm{poly}(d))^2}\max_{t\le t_1}\|w_i\|_2^2\right) \\
\|M^1 M^{1\top} w_i^{(t+1)}\|_2^2 &\le \Big[\Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)^{2t} \\
&\quad + O\big(\frac{1}{\alpha_1\rho_1^2\tau^2 d(K+1)}\big)\Big]\|MM^\top w_i^{(0)}\|_2^2
\end{aligned}
\tag{101}
$$

We have derived in above for all $t \in t_1 \le T_1 = O\Big(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)\log\log d}}\Big)$ for the initial stage of the training, above bound can derive by the following calculations.

$$
\sum_{\substack{j\in[d/2]\\ j\in\mathcal{E}^{(0)}}} O\left(\frac{(t+1)^2\eta'^2\log\log d}{\tau^2 d^2(\mathrm{poly}(d))^2}\max_{t\le t_1}\|w_i\|_2^2\right) \le \widetilde{O}\left(\frac{d/2\cdot d^2\max_{t\le t_1}\|w_i^{(t)}\|_2^2}{\alpha_1\rho_1^2\tau^2 d^2 d_1 \mathrm{poly}(d)(K+1)}\right) \tag{102}
$$

$$
\le \widetilde{O}\left(\frac{d/2\max_{t\le t_1}\|w_i^{(t)}\|_2^2}{\alpha_1\rho_1^2\tau^2 d^2 d_1(K+1)}\right) \tag{103}
$$

since $d_1 = poly(d), |\mathcal{E}^{(0)}| \subset [d]$,We define $\mathcal{E}^{(t)}$ as the set of features similar to the Λ_i defined in Lemma D.2 such that $\mathcal{E}^{(t)} := \{j \in [d] : \langle w_i^{(t)}, M_j\rangle^2 < O(\|w_i^{(t)}\|_2^2/d\sqrt{d_1})\}$. We have $\mathcal{E}^{(t)} \subseteq \Lambda_i$

$$
\le \frac{d/2\max_{t\le t_1}\|w_i^{(t)}\|_2^2}{\alpha_1\rho_1^2\tau^2 d^2 d_1(K+1)} \le \frac{d/2\|w_i^{(t)}\|_2^2}{\alpha_1\rho_1^2\tau^2 d d_1(K+1)} \le O\left(\frac{1}{\alpha_1\rho_1^2\tau^2 d(K+1)}\cdot\|M^1M^{1\top}w_i^{(0)}\|_2^2\right)
$$

Here we use the learning rate $\eta = \eta'\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}$ where $\eta' = \frac{1}{\mathrm{poly}(d)}$.

$$
\|w_i^{(t+1)}\|_2^2 = \|M^1(M^1)^\top w_i^{(t+1)}\|_2^2 + \|M^2(M^2)^\top w_i^{(t+1)}\|_2^2 + \|M^\perp(M^\perp)^\top w_i^{(t+1)}\|_2^2 \tag{104}
$$

$$
\|M^1(M^1)^\top w_i^{(t+1)}\|_2^2 \le \|M^1(M^1)^\top w_i^{(0)}\|_2^2\cdot\Big(1-\eta\lambda+\eta\alpha_1\rho_1^2\frac{\log\log d}{d}\Big)^{2t} + error
$$

$$d\|w_i^{(0)}\|_2^2 \le \|w_i^{(0)}\|_2^2 \cdot \Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{O(\log\log d)}{d}\Big)^{2t} + error$$
$$t \ge O\Big(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\log\log d}\Big)$$

$$\begin{aligned}
\|M^1 M^{1\top} w_i^{(t+1)}\|_2^2 &= \sum_{j\in[d/2]}\Big[\Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)\langle w_i^{(t)}, M_j\rangle \\
&\quad \pm \tilde{O}\Big(\eta'\frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d}\|w_i^{(t)}\|_2\sqrt{\log\log d}\Big)\Big]^2 \\
&\le \sum_{j\in d/2}\langle w_i^{(0)}, M_j\rangle^2 \cdot \Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)^{2t} \\
&\quad - \sum_{\substack{j\in[d/2]\\ j\in\mathcal{E}^{(0)}}} O\left(\frac{(t+1)^2\eta'^2\log\log d}{\tau^2 d^2(\mathrm{poly}(d))^2}\max_{t\le t_1}\|w_i\|_2^2\right) \\
\|M^1 M^{1\top} w_i^{(t+1)}\|_2^2 &\ge \Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)^{2t}\|M^1 M^{1\top} w_i^{(0)}\|_2^2 \\
&\quad - O\big(\frac{1}{\alpha_1\rho_1^2\tau^2 d(K+1)}\big)\|M^1 M^{1\top} w_i^{(0)}\|_2^2
\end{aligned} \tag{105}$$

For $M_j \in M^2$ following the same calculations above

$$\begin{aligned}
\|M^2(M^2)^\top w_i^{(t+1)}\|_2^2 &= \sum_{j\in[d]\setminus[d/2]}\Big[\Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\alpha_2\rho_2^2\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\frac{O(\log\log d)}{d}\Big)\langle w_i^{(t)}, M_j\rangle \\
&\quad \pm \tilde{O}\Big(\eta'\frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d}\|w_i^{(t)}\|_2\sqrt{\frac{\alpha_2\rho_2^2}{\alpha_1\rho_1^2}\log\log d}\Big)\Big]^2 \\
\|M^2(M^2)^\top w_i^{(t+1)}\|_2^2 &\le \Big[\Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\alpha_2\rho_2^2\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\frac{O(\log\log d)}{d}\Big)^{2t} \\
&\quad + \tilde{O}\big(\frac{\alpha_2\rho_2^2}{\alpha_1^2\rho_1^4\tau^2 d(K+1)}\big)\Big]\|M^2(M^2)^\top w_i^{(0)}\|_2^2
\end{aligned} \tag{106}$$

Finally the upper bound of $\|M^\perp(M^\perp)^\top w_i^{(t+1)}\|_2^2$

$$\begin{aligned}
\|M^\perp(M^\perp)^\top w_i^{(t+1)}\|_2^2 &= \sum_{j\in[d_1]\setminus[d]}\left[(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}})\langle w_i^{(t)}, M_j\rangle + \tilde{O}\Big(\eta'\frac{\sum_{r=1}^m\|w_r^{(t)}\|_2^2}{\tau d\sqrt{\alpha_1\rho_1^2(K+1)}}\|w_i^{(t)}\|_2\Big)\right]^2 \\
&\le \|M^\perp(M^\perp)^\top w_i^{(0)}\|_2^2 + d_1\max_{j\in[d_1]\setminus[d]}\langle w_i^{(0)}, M_j^\perp\rangle\frac{O(\eta'\max_{t\le t_1}\|w_i^{(t)}\|_2)}{\tau d d_1\sqrt{\alpha_1\rho_1^2(K+1)}} \\
&\quad + O\big(\frac{d^2\|w_i^{(0)}\|_2^2}{\tau^2(\alpha_1\rho_1^2(K+1))^2 d_1}\big) \\
&\le \Big(1+\tilde{O}\big(\frac{\eta'}{\tau\sqrt{\alpha_1\rho_1^2 d_1(K+1)}}\big)\Big)\|M^\perp(M^\perp)^\top w_i^{(0)}\|_2^2 + \tilde{O}\big(\frac{d^2\|w_i^{(0)}\|_2^2}{\tau^2(\alpha_1\rho_1^2(K+1))^2 d_1}\big) \\
\|M^\perp(M^\perp)^\top w_i^{(t+1)}\|_2^2 &\le \Big(1+\frac{1}{\mathrm{poly}(d)}\Big)\|M^\perp(M^\perp)^\top w_i^{(0)}\|_2^2
\end{aligned} \tag{107}$$

We have used the facts at the initialization (1) $|\langle w_i^{(0)}, M_j^{\perp}\rangle| \leq \tilde{O}(\frac{\|w_i^{(0)}\|_2}{\sqrt{d_1}})$ and (2) $\max_{t\leq t_1}\|w_i^{(t)}\|_2^2 \leq O(d)\|w_i^{(0)}\|_2^2$, (3) we have $\|w_i^{(0)}\|_2 \leq O(\|M^{\perp}(M^{\perp})^{\top}w_i^{(t+1)}\|_2)$

Note that,for each neuron $i \in [m]$,From Lemma and the Induction hypothesis, we can prove that at $T_1 = O\Big(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)\log\log d}}\Big)$ all the weights $\|w_i^{(t)}\|_2^2$ reach to $O(1)\|w_i^{(0)}\|_2^2$.

$$\|w_i^{(T_1)}\|_2^2 = a_1^{2T_1}\|M^1(M^1)^{\top}w_i^{(0)}\|_2^2 + a_2^{2T_1}\|M^2(M^2)^{\top}w_i^{(0)}\|_2^2 + \|M^{\perp}(M^{\perp})\top w_i^{(T_1)}\|_2^2. \quad (108)$$

From the Induction Hypothesis D.2 (a),(b),and (d), and the initialization, we can derive that

$$\begin{aligned}\|w_i^{(T_1)}\|_2^2 &\approx a_1^{2T_1}\|M^1(M^1)^{\top}w_i^{(0)}\|_2^2 + a_2^{2T_1}\|M^2(M^2)^{\top}w_i^{(0)}\|_2^2 + \|M^{\perp}(M^{\perp})\top w_i^{(0)}\|_2^2.\\ &\approx (a_1^{2T_1}-1)\|M^1(M^1)^{\top}w_i^{(0)}\|_2^2 + (a_2^{2T_1}-1)\|M^2(M^2)^{\top}w_i^{(0)}\|_2^2 + \|w_i^{(0)}\|_2^2\\ &\approx (a_1^{2T_1}-1)\frac{d}{2d_1}\|w_i^{(0)}\|_2^2 + (a_2^{2T_1}-1)\frac{d}{2d_1}\|w_i^{(0)}\|_2^2 + \|w_i^{(0)}\|_2^2\end{aligned}$$

$$(A-1)\|w_i^{(0)}\|_2^2 \approx (a_1^{2T_1}-1)\frac{d}{2d_1}\|w_i^{(0)}\|_2^2 + (a_2^{2T_1}-1)\frac{d}{2d_1}\|w_i^{(0)}\|_2^2 \text{ Where A=O(1)}$$

Since we know the M^1 grows much faster than M^2, neurone weights learns M^1 rapidly

$$\begin{aligned}O(1)\|w_i^{(0)}\|_2^2 &\approx a_1^{2T_1}\frac{d}{2d_1}\|w_i^{(0)}\|_2^2\\ O(1)\log(d_1/d) &\approx 2T_1\log a_1 \text{ We know } d_1 = \mathrm{poly}(d)\\ T_1 &= O\Big(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\log\log d}\Big)\end{aligned} \quad (109)$$

Proof of Theorem F.1

It is easy to prove (a) using Induction Hypothesis results at iteration $t = T_1$ for the upper and lower bounds of $\|M^1M^{1\top}w_i^{(t+1)}\|_2^2$, $\|M^2M^{2\top}w_i^{(t+1)}\|_2^2$, and $\|M^{\perp}(M^{\perp})^{\top}w_i^{(t+1)}\|_2^2$. Since we know the growth of $\langle w_i^{(t)}, M_j\rangle, M_j \in M$ is incomparable with the growth of $\langle w_i^{(t)}, M_j^{\perp}\rangle$ we have

$$\|M^1M^{1\top}w_i^{(t+1)}\|_2^2 + \|M^2M^{2\top}w_i^{(t+1)}\|_2^2 \geq \|M^{\perp}(M^{\perp})^{\top}w_i^{(t+1)}\|_2^2 \quad (110)$$

$$\begin{aligned}\|M^1M^{1\top}w_i^{(t+1)}\|_2^2 + \|M^2M^{2\top}w_i^{(t+1)}\|_2^2 \geq \frac{1}{2}\Big(&\|M^1M^{1\top}w_i^{(t+1)}\|_2^2 + \|M^2M^{2\top}w_i^{(t+1)}\|_2^2\\ &+ \|M^{\perp}(M^{\perp})^{\top}w_i^{(t+1)}\|_2^2\Big)\end{aligned} \quad (111)$$

$$\|MM^{\top}w_i^{(T_1)}\|_2^2 \geq \frac{1}{2}\|w_i^{(T_1)}\|_2^2 \quad (112)$$

Similarly this implies that

$$2\|M^{\perp}(M^{\perp})^{\top}w_i^{(t+1)}\|_2^2 \leq \|M^1M^{1\top}w_i^{(t+1)}\|_2^2 + \|M^2M^{2\top}w_i^{(t+1)}\|_2^2 + \|M^{\perp}(M^{\perp})^{\top}w_i^{(t+1)}\|_2^2 \quad (113)$$

$$\|M^{\perp}(M^{\perp})^{\top}w_i^{(t+1)}\|_2^2 \leq \frac{1}{2}\|w_i^{(T_1)}\|_2^2 \quad (114)$$

(b),(c) Here we prove the growth of $|\langle w_i^{(t+1)}, M_j\rangle|$ with the trained weight norm for $i \in \mathcal{M}_j^{1*}$ From the upper and lower bounds of the $\|M^1(M^1)^{\top}w_i^{(0)}\|_2$ we have for $i \in \mathcal{M}_j^{1*}$. We choose constants $c_0 \in (0,1)$,$\gamma \in (0, 0.01)$, and $c_1 = 2(1 + c_0(1-\gamma))$. We select $c_1 - c_2 = \gamma c_0$ to be a samll gap between the initial strong and weak neron alignments.

We can write the $\|MM^{\top}w_i^{(T_1)}\|_2^2$ as the summation of the growing rates $(a_1^{T_1}, a_2^{T_1})$ and the error terms $(\mathcal{E}_1, \mathcal{E}_2)$ of $\|M^1(M^1)^{\top}w_i^{(0)}\|_2^2$ and $\|M^2(M^2)^{\top}w_i^{(0)}\|_2^2$ as follows.

$$\|MM^{\top}w_i^{(T_1)}\|_2^2 = (a_1^{T_1} \pm \mathcal{E}_1)^2\|M^1(M^1)^{\top}w_i^{(0)}\|_2^2 + (a_2^{T_1} \pm \mathcal{E}_2)^2\|M^2(M^2)^{\top}w_i^{(0)}\|_2^2. \quad (115)$$

Since $1 < a_2 < a_1$, the M_2-component is growing slowly while the M_1-component is growing rapidly. Recall for $T = T_1$

$$\left(\frac{a_2}{a_1}\right)^T = \left(\frac{1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\alpha_2\rho_2^2\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\frac{\log\log d}{d}}{1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}}\right)^T \tag{116}$$

$$= \left(1-\frac{\eta'\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\left(\alpha_1\rho_1^2-\alpha_2\rho_2^2\right)\frac{\log\log d}{d}}{1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}}\right)^T \tag{117}$$

Also we can write $D := \frac{\eta'\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\left(\alpha_1\rho_1^2-\alpha_2\rho_2^2\right)\frac{\log\log d}{d}}{1+\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}}$ Using $(1-D)^T \approx e^{-DT}$, we get

$$\left(\frac{a_2}{a_1}\right)^T = d^{-O(1-\frac{\alpha_2\rho_2^2}{\alpha_1\rho_1^2})} \leq \frac{\alpha_2\rho_2^2}{\alpha_1\rho_1^2}$$

We use Theorem (a) to get a more simplified, meaningful bounds. We have at $t = T_1$,

$$\|M^1(M^1)^\top w_i^{(T_1)}\|_2^2 = L\|w_i^{(T_1)}\|_2^2$$

(b) First start with $i \in \mathcal{M}_j^{1*}$ for $M_j \in M^1$

$$|\langle w_i^{(t+1)}, M_j\rangle| \geq \left(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\right)^{T_1}|\langle w_i^{(0)}, M_j\rangle| \tag{118}$$

$$-\tilde{O}\left(\frac{\eta' T_1}{\tau d\cdot\text{poly}(d)}\max_{t\leq t_1}\|w_i\|_2\right) \tag{119}$$

$$\geq \sqrt{\frac{c_1\log d}{d}}\|M^1M^{1\top}w_i^{(0)}\|_2\left(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\right)^{T_1}$$

$$-\frac{\|M^1M^{1\top}w_i^{(0)}\|_2}{\tau\sqrt{\alpha_1\rho_1^2(K+1)d_1}}$$

$$\geq \sqrt{\frac{c_1\log d}{d}}\|M^1M^{1\top}w_i^{(T_1)}\|_2 \tag{120}$$

$$\geq \sqrt{\frac{c_1\Theta_{i\in\mathcal{M}_j^{1*}}\log d}{d}}\|w_i^{(T_1)}\|_2 \tag{121}$$

$$\geq (1+\gamma c_0)\sqrt{\frac{2\log d}{d}}\|w_i^{(T_1)}\|_2 \tag{122}$$

(c) For $i \notin \mathcal{M}_j^{1*}$ and $M_j \in M^1$

$$|\langle w_i^{(t+1)}, M_j\rangle| \leq \Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)^{T_1}|\langle w_i^{(0)}, M_j\rangle| \tag{123}$$

$$+\tilde{O}\Big(\frac{\eta' T_1}{\tau d\cdot \mathrm{poly}(d)}\max_{t\leq t_1}\|w_i\|_2\Big)$$

$$\leq \sqrt{\frac{c_2\log d}{d}}\|M^1M^{1\top}w_i^{(0)}\|_2\Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)^{T_1}$$

$$+\frac{\|M^1M^{1\top}w_i^{(0)}\|_2}{\tau\sqrt{\alpha_1\rho_1^2(K+1)\Gamma d_1}}$$

$$\leq \sqrt{\frac{c_2\log d}{d}}\|M^1M^{1\top}w_i^{(T_1)}\|_2 \tag{124}$$

$$\leq \sqrt{\frac{c_2\Theta_{i\notin\mathcal{M}_j^{1*}}\log d}{d}}\|w_i^{(T_1)}\|_2 \tag{125}$$

$$\leq (1-2\gamma c_0)\sqrt{\frac{2\log d}{d}}\|w_i^{(T_1)}\|_2 \tag{126}$$

(d) For $i \in \mathcal{M}_j^2$ $M_j \in M^2$.Following the similar calculation for the upper bound of the neuron initialization in $\mathcal{M}_j^2$, we can derive that for $j' \in [d]\backslash[d/2]$

$$|\langle w_i^{(t+1)}, M_{j'}\rangle| \leq \Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+\eta'\alpha_2\rho_2^2\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\frac{\log\log d}{d}\Big)^{T_1}|\langle w_i^{(0)}, M_{j'}\rangle| \tag{127}$$

$$-\tilde{O}\left(\frac{\eta' T_1\sqrt{\alpha_2\rho_2^2}\max_{t\leq t_1}\|w_i\|_2}{\tau d\cdot\mathrm{poly}(d)\sqrt{\alpha_1\rho_1^2}}\right)$$

$$\leq \frac{\alpha_2\rho_2^2}{\alpha_1\rho_1^2}\sqrt{\frac{2c_3'\log d}{d}}\|w_i^{(T_1)}\|_2 \tag{128}$$

$$\leq (1-\gamma c_0)\sqrt{\frac{2\log d}{d}}\|w_i^{(T_1)}\|_2 \tag{129}$$

$$\frac{|\langle w_i^{(t+1)}, M_j\rangle|}{|\langle w_i^{(t+1)}, M_{j'}\rangle|} \geq \frac{\sqrt{\frac{c_1\log d}{d}}\|w_i^{(T_1)}\|_2}{\sqrt{\frac{c_3'\log d}{d}}\cdot\frac{1}{d^{O(C_\beta)}}\|w_i^{(T_1)}\|_2} \geq \frac{1}{\Big(1-\frac{\alpha_2\rho_2^2}{\alpha_1\rho_1^2}\Big)\log d}\sqrt{\frac{c_1}{c_3'}} \tag{130}$$

This ensures that the growth of the neurones towards the M^1 is exteamly high with respect to the M^2 at $t = T_1$ iterations.

(f)For $i \in [m]$ at initialization $|\langle w_i^{(0)}, M_j^\perp\rangle| \leq O\Big(\sqrt{\frac{\log d}{d_1}}\Big)\|w_i^{(0)}||_2$ w.h.p. From the initial layer training duynamics we have for $(1-\eta\lambda) \leq 1$,

$$\langle w_i^{(t+1)}, M_j^\perp\rangle \leq \Big(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\Big)\langle w_i^{(t)}, M_j^\perp\rangle + O\big(\frac{\eta'\|w_i^{(t)}\|_2^2}{\tau d d_1\sqrt{\alpha_1\rho_1^2(K+1)}}\big) \tag{131}$$

$$\leq \langle w_i^{(0)}, M_j^\perp\rangle + \sum_{t=0}^{t} O\big(\frac{\eta' C_\eta\|w_i^{(t)}\|_2^2}{\tau d d_1\sqrt{\alpha_1\rho_1^2(K+1)}}\big) \tag{132}$$

$$\langle w_i^{(T_1)}, M_j^\perp\rangle \leq \langle w_i^{(0)}, M_j^\perp\rangle + \eta' T_1\max_{t\leq T_1} O\big(\frac{C_\eta\|w_i^{(t)}\|_2^2}{\tau d d_1\sqrt{\alpha_1\rho_1^2(K+1)}}\big) \tag{133}$$

Since $\eta T_1/d_1 \leq 1/poly(d)$ and $\|w_i^{(t)}\|_2^2 \leq O(1)\|w_i^{(0)}\|_2^2$ we can conclude that $\langle w_i^{(T_1)}, M_j^\perp\rangle \leq O\Big(\sqrt{\frac{\log d}{d_1}}\|w_i^{(t)}\|_2\Big)$

G SECOND STAGE TRAINING

Induction Hypothesis G.1. For all neurones $i \in [m]$ satisfies for the iterations $t \in [T_1, T_2]$

(a) For $j \in [d]$ and $i \in \mathcal{M}_j^{1*}$ then $|\langle w_i^{(t)}, M_j\rangle| \geq (1 + c_0)\frac{b_i^{(t)}}{\rho_1}$

(b) For $j \in [d]$, if $i \notin \mathcal{M}_j$, then $|\langle w_i^{(t)}, M_j\rangle| \leq (1 - c_0)\frac{b_i^{(t)}}{\rho_1}$.

(c) For each $i \in [m]$, there are at most $O(2^{-\sqrt{\log d}}d)$ many $j \in [d]$ such that $\langle w_i^{(t)}, M_j\rangle^2 \geq \frac{(b_i^{(t)}/\rho_1)^2}{\sqrt{\log d}}$;

(d) For $j \in [d]$,if $i \notin \mathcal{M}_j$ $|\langle w_i^{(t)}, M_j\rangle| \leq \tilde{O}\left(\frac{\|w_i^{(t)}\|_2}{\sqrt{d}}\right)$

(e) For each $i \in [m]$, we have $|\langle w_i^{(t)}, M_j^{\perp}\rangle| \leq \tilde{O}\left(\frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}}\right)$ for all $j \in [d_1] \setminus [d]$,;

(f) $\|w_i^{(t)}\|_2^2 \leq \frac{d(b_i^{(t)}/\rho_1)^2}{\alpha_1 \log d}$ for all $i \in [m]$.

First we define the following parameters before compute the bounds of the gradient.From the definition of the sparse coding model we deifne the parameter $\bar{x}_k$ as the average of the two views in terms of the dictionary matrix M,the corresponfding sparse vectors in each view $\bar{z}_k^{(1)}, \bar{z}_k^{(2)}$, and the dense noise $\bar{\xi}_k^{(1)}, \bar{\xi}_k^{(2)}$

$$\bar{x}_k = \frac{1}{2}\sum_{v\in\{1,2\}} \bar{x}_{vk} = \frac{1}{2}\sum_{v\in\{1,2\}} \Big(\sum_{\substack{j'\neq j\\ j'\in[d]}} M_{j'}\bar{z}_{vk,j'} + \bar{\xi}_{vk} + M_j\bar{z}_{vk,j}\Big)$$

Define the contribution excluding the j-th dictionary atom as

$$\bar{x}_{vk}^{\setminus j} = \sum_{j'\neq j, j'\in[d]} M_{j'}\bar{z}_{vk,j'} + \bar{\xi}_{vk}, \quad v \in \{1,2\} \qquad \bar{x}_k = \frac{1}{2}\Big[\sum_{v\in\{1,2\}} \bar{x}_{vk}^{\setminus j} + M_j\bar{z}_{vk,j}\Big]$$

For a neuron i at iteration t, define

$$S_{i,t,vk}^{\setminus j} = \langle w_i^{(t)}, \bar{x}_{vk}^{\setminus j}\rangle, \;\; \alpha_{i,j} = \langle w_i^{(t)}, M_j\rangle, \;\; S_{i,t,k}^{\setminus j} = \langle w_i^{(t)}, \bar{x}_k^{\setminus j}\rangle = \frac{1}{2}\sum_{v\in\{1,2\}} S_{i,t,vk}^{\setminus j}$$

Next define the δ as the alignment of the difference between two views with neuron i as follows. Then we can rewrite $\langle w_i^{(t)}, \bar{x}_k^{(1)}\rangle$ and $\langle w_i^{(t)}, \bar{x}_k^{(2)}\rangle$ with $\delta = \langle w_i^{(t)}, \frac{\bar{x}_k^{(2)} - \bar{x}_k^{(1)}}{2}\rangle$,

$$\langle w_i^{(t)}, \bar{x}_k^{(2)}\rangle = \langle w_i^{(t)}, \bar{x}_k\rangle + \delta, \;\; \langle w_i^{(t)}, \bar{x}_k^{(1)}\rangle = \langle w_i^{(t)}, \bar{x}_k\rangle - \delta$$

Finally, define the half-differences of the residual signals and sparse coefficients as

$$\bar{S}_{i,t,vk}^{\setminus j} = \frac{S_{i,t,2k}^{\setminus j} - S_{i,t,1k}^{\setminus j}}{2}, \;\; \bar{z'}_{k,j} = \frac{\bar{z}_{k,j}^{(2)} - \bar{z}_{k,j}^{(1)}}{2}, \;\; \bar{z}_{k,j} = \frac{\bar{z}_{k,j}^{(1)} + \bar{z}_{k,j}^{(2)}}{2}$$

Most of the elements in $|\bar{z'}_{k,j'}|$ are zero due to the similiarity in the first $d/2$ elements. However, in the second $d/2$ elements have a dissimilarity and this creates a non zero expectation.

Let $j \in [d]$ and suppose $M_j \in M^1$. Consider the sparse codes $\bar{z}_k^{(1)}, \bar{z}_k^{(2)} \in \mathbb{R}^d$ constructed as in the data model,where each coordinate is activated with probability $O(\log\log d)/d$ and conditional on activation takes value $\pm\frac{1}{K+1}$ or $\pm\frac{\rho K+1}{K+1}$ depends on the way of the define $\bar{z'}_{k,j}$ and

$\bar{z}_{k,j}$. Then for $K < d$, the following expectations hold.

$$\begin{aligned}
\mathbb{E}[|\bar{z'}_{k,j}|] &= 2\frac{1}{K+1}.\Pr(|\bar{z}^{(1)}_{k,j}| = \frac{1}{K+1}|j \in M^1).\Pr(\bar{z}^{(2)}_{k,j} = 0|j \in M^1)Pr(j \in M^1)^2 \\
&= \frac{K+1-\alpha_1-\alpha_1\rho_1 K}{K+1} \cdot \frac{O(\log\log d)}{d}\Big(1-(K(1-\alpha_1\rho_1)+1)\frac{O(\log\log d)}{d}\Big).\frac{1}{2} \\
&= (1-\alpha_1\rho_1)\cdot\frac{O(\log\log d)}{d} \qquad (134) \\
\mathbb{E}[|\bar{z'}_{k,j}||\bar{z}^{(1)}_{k,j}|] &= \big[\frac{1}{K+1}\big]^2 \cdot \Pr(|\bar{z}^{(1)}_{k,j}| = \frac{1}{K+1}|j \in M^1)\big(1-\Pr(|\bar{z}^{(2)}_{k,j}| = \frac{1}{K+1}|j \in M^1)\big) \\
&= \frac{1-\alpha_1\rho_1}{K+1}\cdot\frac{O(\log\log d)}{d} + \frac{1}{K+1}O\Big(\frac{\log\log d}{d}\Big)^2 \qquad (135) \\
\mathbb{E}[|\bar{z}_{k,j}||\bar{z}^{(1)}_{k,j}|] &= \big[\frac{1}{K+1}\big]^2[\Pr(|\bar{z}^{(1)}_{k,j}| = \frac{1}{K+1}|j \in M^1) + \Pr(|\bar{z}^{(1)}_{k,j}| = \frac{1}{K+1}, |\bar{z}^{(2)}_{k,j}| = \frac{1}{K+1}|j \in M^1)] \\
&\quad + \big[\frac{\rho_1 K+1}{K+1}\big]^2 \cdot \Pr(|\bar{z}^{(1)}_{k,j}| = \frac{1}{K+1}|j \in M^1)\big(1+\Pr(|\bar{z}^{(2)}_{k,j}| = \frac{1}{K+1}|j \in M^1)\big) \\
&= \Big(\alpha_1\rho_1^2 + \frac{1-\alpha_1\rho_1}{K+1}\Big)\frac{O(\log\log d)}{d} \qquad (136)
\end{aligned}$$

Let $j \in [d]\backslash[d/2]$ and suppose $M_j \in M^2$.

$$\begin{aligned}
\mathbb{E}[|\bar{z'}_{k,j}|] &= (|\bar{z'}_{k,j}| = \frac{1}{K+1}).Pr(|\bar{z}^{(1)}_{k,j}| = \frac{1}{K+1}|j \in M^2).Pr(\bar{z}^{(2)}_{k,j} = 0|j \in M^2).Pr(j \in M^2)^2 \\
&= \frac{K+1-\alpha_2-\alpha_2\rho_2 K}{K+1} \cdot \frac{O(\log\log d)}{d}\Big(1-(K(1-\alpha_2\rho_2)+1)\frac{O(\log\log d)}{d}\Big).\frac{1}{2} \\
&\approx (1-\alpha_2\rho_2)\frac{O(\log\log d)}{d} \qquad (137) \\
\mathbb{E}[|\bar{z'}_{k,j}||\bar{z}^{(1)}_{k,j}|] &= (|\bar{z'}_{k,j}| = \frac{1}{2(K+1)}) \cdot (|\bar{z}^{(1)}_{k,j}| = \frac{1}{K+1}) \cdot Pr(|\bar{z}^{(1)}_{k,j}| = \frac{1}{K+1}) \\
&= \frac{1}{(K+1)^2}(K+1-\alpha_2-\alpha_2\rho_2 K)\frac{O(\log\log d)}{d} \\
&= \frac{1-\alpha_2\rho_2}{(K+1)}\frac{O(\log\log d)}{d} \qquad (138) \\
\mathbb{E}[|\bar{z}_{k,j}||\bar{z}^{(1)}_{k,j}|] &= \mathbb{E}\Big[|\frac{\bar{z}^{(1)}_{k,j}+\bar{z}^{(2)}_{k,j}}{2}||\bar{z}^{(1)}_{k,j}|\Big] \\
&= \Big(\alpha_2\rho_2^2 + \frac{1-\alpha_2\rho_2}{K+1}\Big)\frac{O(\log\log d)}{d} \qquad (139)
\end{aligned}$$

Lemma G.2. *Gradient for sparse features at iteration $t \geq 0$ for $j \in [d]$ denote the events and the quantities L_1, L_2, L_3, L_4, L_5*
$E_{1k} = \{S^{\backslash j}_{i,t,1k} \geq b_i - \alpha_{i,j}\bar{z}^{(1)}_{k,j}\}$, $E_{2k} = \{S^{\backslash j}_{i,t,2k} \geq b_i - \alpha_{i,j}\bar{z}^{(2)}_{k,j}\}$
$E^1_k = \{S^{\backslash j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}\}$, $E^2_k = \{\bar{S}^{\backslash j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z'}_{k,j}\}$,
$E_{3k} = \{|\bar{S}^{\backslash j}_{i,t,k} + \alpha_{i,j}\bar{z'}_{k,j}| \geq (\alpha_{i,j}\bar{z}_{k,j} - b^{(t)}_i)/2\}$, $E_{4k} = \{S^{\backslash j}_{i,t,k} \geq (\alpha_{i,j}\bar{z}_{k,j} - b^{(t)}_i)/2\}$

$$L_1 = \sqrt{\frac{\mathbb{E}[|\bar{S}^{\backslash j}_{i,t,k}|^2(\mathbf{1}_{E^1_k}+\mathbf{1}_{E^2_k})]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}},\ L_2 = \Pr(E^1_k),\ L_3 = \sqrt{\frac{\mathbb{E}[|\bar{S}^{\backslash j}_{i,t,k}|^2(\mathbf{1}_{E_{3k}}+\mathbf{1}_{E_{4k}})]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}},\ L_4 = \Pr(E^2_k),$$
$L_5 = \Pr(E_{3k})$

a) For all neurons $i \in [m]$, if $\alpha_{i,j} \geq 0$ we have

For $j \in [d]$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\bar{z}_{k,j}^{(1)}\right] \leq \alpha_{i,j}\Big(\alpha_1\rho_1^2+\frac{1-\alpha_1\rho_1}{K+1}\Big)\frac{O(\log\log d)}{d}\cdot \Big[\Pr(|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|)\pm O(L_1+L_2)\Big] \pm \frac{\alpha_{i,j}(1-\alpha_1\rho_1)}{K+1}\frac{O(\log\log d)}{d}O\big(L_2+L_4+L_1\big) \tag{140}$$

For $j \in [d']$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\bar{z}_{k,j}^{(1)}\right] \leq \alpha_{i,j}\Big(\alpha_2\rho_2^2+\frac{1-\alpha_1\rho_1}{K+1}\Big)\frac{O(\log\log d)}{d}\cdot \Big[\Pr(|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|)\pm O(L_1+L_2)\Big] \pm \frac{\alpha_{i,j}(1-\alpha_2\rho_2)}{K+1}\frac{O(\log\log d)}{d}O\big(L_2+L_4+L_1\big) \tag{141}$$

b) For lucky neurones when $\alpha_{i,j} > b_i^{(t)}$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[h_i(\bar{x}_k^{(2)})\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\right] = \alpha_{i,j}\Big(\alpha_1\rho_1^2+\frac{1-\alpha_1\rho_1}{K+1}\Big)\frac{O(\log\log d)}{d}\Big[\Pr(|\langle w_i,\bar{x}_k\rangle|\geq b_i + |\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|)\pm O\Big(L_5+L_3\Big)\Big] \pm \alpha_{i,j}\frac{(1-\alpha_1\rho_1)}{(K+1)}\frac{O(\log\log d)}{d}O\Big(L_5+L_3\Big) \tag{142}$$

First we start from the part (a) of the Lemma D1 for all neurones $i \in [m]$ in the case of $0 < \alpha_{i,j} < b_i^{(t)}$

Proof of Lemma G.2 (a)

$$\begin{aligned}
&\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[h_i(\bar{x}_k^{(2)})\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\right]\\
&=\mathbb{E}\left[(\langle w_i^{(t)},\bar{x}_k\rangle+\delta)\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq b_i}\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\right] &(143)\\
&\quad-\mathbb{E}\left[b_i\left(\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}-\mathbb{1}_{-\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}\right)\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\right]\\
&=\mathbb{E}\left[\alpha_{i,j}\cdot\bar{z}_{k,j}\bar{z}_{k,j}^{(1)}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\geq b_i}\right] &(144)\\
&\quad+\mathbb{E}\left[\alpha_{i,j}\cdot\bar{z}_{k,j}\bar{z}_{k,j}^{(1)}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\leq -b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\leq -b_i}\right]\\
&\quad+\mathbb{E}\left[(S_{i,t,k}^{\setminus j}-b_i)\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\right] &(145)\\
&\quad+\mathbb{E}\left[(S_{i,t,k}^{\setminus j}+b_i)\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\leq -b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\leq -b_i}\cdot\bar{z}_{k,j}^{(1)}\right]\\
&\quad+\mathbb{E}\left[(\alpha_j\bar{z}_{k,j}^{(1)}+S_{i,t,k}^{\setminus j}-b_i)\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\leq -b_i}\cdot\bar{z}_{k,j}^{(1)}\right]\\
&\quad+\mathbb{E}\left[(\alpha_j\bar{z}_{k,j}^{(1)}+S_{i,t,k}^{\setminus j}+b_i)\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\leq -b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\right] &(146)\\
&\quad+\mathbb{E}\left[\delta\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq b_i}\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\right]\\
&=J_1+J_2+J_3+J_4 &(147)
\end{aligned}$$

Here we introduce J_1, J_2, J_3, J_4 as follows

$$
\begin{aligned}
J_1 &= \mathbb{E}\left[\alpha_{i,j}\cdot \bar{z}_{k,j}\bar{z}^{(1)}_{k,j}\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(2)}_k\rangle\geq b_i}\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(1)}_k\rangle\geq b_i}\right] \\
&\quad + \mathbb{E}\left[\alpha_{i,j}\cdot \bar{z}_{k,j}\bar{z}^{(1)}_{k,j}\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(2)}_k\rangle\leq -b_i}\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(1)}_k\rangle\leq -b_i}\right] \qquad (148)\\
J_2 &= \mathbb{E}\left[(S^{\setminus j}_{i,t,k}-b_i)\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(2)}_k\rangle\geq b_i}\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(1)}_k\rangle\geq b_i}\cdot \bar{z}^{(1)}_{k,j}\right] \\
&\quad + \mathbb{E}\left[(S^{\setminus j}_{i,t,k}+b_i)\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(2)}_k\rangle\leq -b_i}\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(1)}_k\rangle\leq -b_i}\cdot \bar{z}^{(1)}_{k,j}\right] \qquad (149)\\
J_3 &= \mathbb{E}\left[(\alpha_j\bar{z}^{(1)}_{k,j}+S^{\setminus j}_{i,t,k}-b_i)\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(2)}_k\rangle\geq b_i}\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(1)}_k\rangle\leq -b_i}\cdot \bar{z}^{(1)}_{k,j}\right] \\
&\quad + \mathbb{E}\left[(\alpha_j\bar{z}^{(1)}_{k,j}+S^{\setminus j}_{i,t,k}+b_i)\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(2)}_k\rangle\leq -b_i}\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(1)}_k\rangle\geq b_i}\cdot \bar{z}^{(1)}_{k,j}\right] \qquad (150)\\
J_4 &= \mathbb{E}\left[\delta\cdot \mathbb{1}_{|\langle w_i,\bar{x}^{(2)}_k\rangle|\geq b_i}\cdot \mathbb{1}_{|\langle w_i,\bar{x}^{(1)}_k\rangle|\geq b_i}\cdot \bar{z}^{(1)}_{k,j}\right] \qquad (151)
\end{aligned}
$$

Starting from analysing the terms in the J_2

$$
\begin{aligned}
J_2 =& \mathbb{E}\left[(S^{\setminus j}_{i,t,k}-b_i)\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(2)}_k\rangle\geq b_i}\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(1)}_k\rangle\geq b_i}\cdot \bar{z}^{(1)}_{k,j}\right] + \mathbb{E}\left[(S^{\setminus j}_{i,t,k}+b_i)\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(2)}_k\rangle\leq -b_i}\cdot \mathbb{1}_{\langle w_i,\bar{x}^{(1)}_k\rangle\leq -b_i}\cdot \bar{z}^{(1)}_{k,j}\right] \qquad (152)\\
&= \mathbb{E}\left[(S^{\setminus j}_{i,t,k}-b_i)\cdot \mathbb{1}_{\langle w_i,\bar{x}_k\rangle\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}^{(1)}_k\rangle|}\cdot \bar{z}^{(1)}_{k,j}\right] + \mathbb{E}\left[(S^{\setminus j}_{i,t,k}+b_i)\cdot \mathbb{1}_{\langle w_i,\bar{x}_k\rangle\leq -b_i-|\langle w_i,\bar{x}_k-\bar{x}^{(1)}_k\rangle|}\cdot \bar{z}^{(1)}_{k,j}\right] \\
&= J_{21}+J_{22} \qquad (153)\\
J_{21} &= \mathbb{E}\left[(S^{\setminus j}_{i,t,k}-b_i)\cdot \mathbb{1}_{\langle w_i,\bar{x}_k\rangle\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}^{(1)}_k\rangle|}\cdot \bar{z}^{(1)}_{k,j}\right] \qquad (154)\\
&= \mathbb{E}\left[(S^{\setminus j}_{i,t,k}-b_i)\cdot \mathbb{1}_{S^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z}_{k,j}\geq b_i+|\bar{S}^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z'}_{k,j'}|}\cdot \bar{z}^{(1)}_{k,j}\right] \\
&= \mathbb{E}\left[(S^{\setminus j}_{i,t,k}-b_i)\cdot |\bar{z}^{(1)}_{k,j}|\left(\mathbb{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}+|\bar{S}^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z'}_{k,j'}|}-\mathbb{1}_{S^{\setminus j}_{i,t,k}\geq b_i+\alpha_{i,j}\bar{z}_{k,j}+|\bar{S}^{\setminus j}_{i,t,k}-\alpha_{i,j}\bar{z'}_{k,j'}|}\right)\mathbb{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right] \\
&= \mathbb{E}\left[(S^{\setminus j}_{i,t,k}-b_i)\cdot |\bar{z}^{(1)}_{k,j}|\mathbb{1}_{S^{\setminus j}_{i,t,k}-b_i-A_k\in[-|\alpha_{i,j}\bar{z}_{k,j}-A'_k|,|\alpha_{i,j}\bar{z}_{k,j}-A'_k|]}\mathbb{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right] \\
J_{21} &= \mathbb{E}\left[(S^{\setminus j}_{i,t,k}-b_i+A_k-A_k)\cdot |\bar{z}^{(1)}_{k,j}|\mathbb{1}_{S^{\setminus j}_{i,t,k}-b_i-A_k\in[-|\alpha_{i,j}\bar{z}_{k,j}-A'_k|,|\alpha_{i,j}\bar{z}_{k,j}-A'_k|]}\mathbb{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right] \qquad (155)\\
&\leq \mathbb{E}\left[|S^{\setminus j}_{i,t,k}-b_i-A_k|\cdot |\bar{z}^{(1)}_{k,j}|\mathbb{1}_{S^{\setminus j}_{i,t,k}-b_i-A_k\in[-|\alpha_{i,j}\bar{z}_{k,j}-A'_k|,|\alpha_{i,j}\bar{z}_{k,j}-A'_k|]}\mathbb{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right] \\
&\quad + \mathbb{E}\left[(|\bar{S}^{\setminus j}_{i,t,k}|+|\alpha_{i,j}||\bar{z'}_{k,j}|)\cdot |\bar{z}^{(1)}_{k,j}|\mathbb{1}_{S^{\setminus j}_{i,t,k}-b_i-A_k\in[-|\alpha_{i,j}\bar{z}_{k,j}-A'_k|,|\alpha_{i,j}\bar{z}_{k,j}-A'_k|]}\mathbb{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right] \\
&\leq \mathbb{E}\left[(|\alpha_{i,j}||\bar{z}_{k,j}|+2|\alpha_{i,j}||\bar{z'}_{k,j}|)|\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right] \\
&\quad + \mathbb{E}\left[|\bar{S}^{\setminus j}_{i,t,k}||\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k}-b_i-A_k\in[-|\alpha_{i,j}\bar{z}_{k,j}-A'_k|,|\alpha_{i,j}\bar{z}_{k,j}-A'_k|]}\mathbf{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right] \\
&= \mathbb{E}\left[(|\alpha_{i,j}||\bar{z}_{k,j}|+2|\alpha_{i,j}||\bar{z'}_{k,j}|)|\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right] \\
&\quad + \sqrt{\mathbb{E}\left[|\bar{S}^{\setminus j}_{i,t,k}||\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right]^2}\cdot\sqrt{\mathbb{E}\left[\mathbf{1}_{S^{\setminus j}_{i,t,k}-b_i-A_k\in[-|\alpha_{i,j}\bar{z}_{k,j}-A'_k|,|\alpha_{i,j}\bar{z}_{k,j}-A'_k|]}\right]} \\
&= \mathbb{E}\left[(|\alpha_{i,j}||\bar{z}_{k,j}|+2|\alpha_{i,j}||\bar{z'}_{k,j}|)|\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right] \\
&\quad + \sqrt{\mathbb{E}\left[|\bar{S}^{\setminus j}_{i,t,k}||\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}\right]^2}\cdot\sqrt{\mathbb{E}\left[\frac{O(\alpha_{i,j}|\bar{z}_{k,j}|+\alpha_{i,j}|\bar{z'}_{k,j'}|)^2}{\langle w,\bar{\xi}_k\rangle^2}\right]} \qquad (156)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[(|\alpha_{i,j}||\bar{z}_{k,j}| + 2|\alpha_{i,j}||\bar{z'}_{k,j}|)|\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}\right] \\
&\quad + \sqrt{\mathbb{E}\left[O(\alpha_{i,j}|\bar{z}_{k,j}| + \alpha_{i,j}|\bar{z'}_{k,j}|)^2\right]|\bar{z}^{(1)}_{k,j}|^2}\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}} \\
&\leq \alpha_{i,j}\mathbb{E}\left[(|\bar{z}_{k,j}| + 2|\bar{z'}_{k,j}|)|\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}\right] \\
&\quad + \sqrt{\mathbb{E}\left[(2\alpha^2_{i,j}|\bar{z}_{k,j}|^2 + 2\alpha^2_{i,j}|\bar{z'}_{k,j}|^2)|\bar{z}^{(1)}_{k,j}|^2\right]}\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}} \\
&\leq \alpha_{i,j}\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}\right] + 2\alpha_{i,j}\mathbb{E}\left[|\bar{z'}_{k,j}||\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}\right] \\
&\quad + \left(\sqrt{\mathbb{E}\left[(2\alpha^2_{i,j}|\bar{z}_{k,j}|^2|\bar{z}^{(1)}_{k,j}|^2\right]} + \sqrt{\mathbb{E}\left[2\alpha^2_{i,j}|\bar{z'}_{k,j}|^2|\bar{z}^{(1)}_{k,j}|^2\right]}\right)\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}} \\
&\leq \alpha_{i,j}\Big(\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}\right] + \sqrt{O(\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}^{(1)}_{k,j}|\right]^2)}\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}}\Big) \\
&\quad + \alpha_{i,j}\Big(\mathbb{E}\left[2|\bar{z'}_{k,j}||\bar{z}^{(1)}_{k,j}|\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}\right] + \sqrt{O(\mathbb{E}\left[|\bar{z'}_{k,j}||\bar{z}^{(1)}_{k,j}|\right]^2)}\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}}\Big)
\end{aligned} \tag{157}
$$

$$
J_{21} \leq \alpha_{i,j}\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}^{(1)}_{k,j}|\right]\left[Pr(S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}) + O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}}\Big)\right] \tag{158}
$$

$$
+ \alpha_{i,j}\mathbb{E}\left[|\bar{z'}_{k,j}||\bar{z}^{(1)}_{k,j}|\right]\left[Pr(S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}) + O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{S^{\setminus j}_{i,t,k} \geq b_i - \alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}}\Big)\right] \tag{159}
$$

$$
J_{21} \leq \alpha_{i,j}\Big(\alpha_1\rho_1^2 O\Big(\frac{\log\log d}{d}\Big) + 2\frac{(1-\alpha_1\rho_1)}{(K+1)}\frac{O(\log\log d)}{d}\Big)\Big[Pr(E^1_k) + O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{E^1_k}]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}}\Big)\Big] \quad \text{For } j \in M^1 \tag{160}
$$

$$
J_{21} \leq \alpha_{i,j}\Big(\alpha_2\rho_2^2 + \frac{(1-\alpha_2\rho_2)}{2(K+1)}\Big)\frac{O(\log\log d)}{d}\Big)\Big[Pr(E^1_k) + O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{E^1_k}]}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}}\Big)\Big] \quad \text{For } j \in M^2 \tag{161}
$$

Define

$$
A_k = \frac{1}{2}\Big[|\bar{S}^{\setminus j}_{i,t,k} + \alpha_{i,j}\bar{z'}_{k,j'}| + |\bar{S}^{\setminus j}_{i,t,k} - \alpha_{i,j}\bar{z'}_{k,j'}|\Big], \quad A'_k = \frac{1}{2}\Big[|\bar{S}^{\setminus j}_{i,t,k} + \alpha_{i,j}\bar{z'}_{k,j'}| - |\bar{S}^{\setminus j}_{i,t,k} - \alpha_{i,j}\bar{z'}_{k,j'}|\Big] \tag{162}
$$

$$
\begin{aligned}
S^{\setminus j}_{i,t,k} - b_i - A_k \in [-|\alpha_{i,j}\bar{z}_{k,j} - A'_k|, |\alpha_{i,j}\bar{z}_{k,j} - A'_k|] &\subseteq S^{\setminus j}_{i,t,k} - |\bar{S}^{\setminus j}_{i,t,k}| - |\alpha_{i,j}\bar{z'}_{k,j'}| \in [-|\alpha_{i,j}\bar{z}_{k,j}| - |\alpha_{i,j}\bar{z'}_{k,j'}| \\
&\quad , |\alpha_{i,j}\bar{z'}_{k,j'}| + |\alpha_{i,j}\bar{z}_{k,j}|] \\
&\subseteq S^{\setminus j}_{i,t,k} - |\bar{S}^{\setminus j}_{i,t,k}| \in [-|\alpha_{i,j}\bar{z}_{k,j}| - |\alpha_{i,j}\bar{z'}_{k,j'}| + |\alpha_{i,j}\bar{z'}_{k,j'}| \\
&\quad , |\alpha_{i,j}\bar{z'}_{k,j'}| + |\alpha_{i,j}\bar{z}_{k,j}| + |\alpha_{i,j}\bar{z'}_{k,j'}|] \\
&\subseteq S^{\setminus j}_{i,t,k} - |\bar{S}^{\setminus j}_{i,t,k}| \in [-|\alpha_{i,j}\bar{z}_{k,j}|, |\alpha_{i,j}\bar{z}_{k,j} + 2\alpha_{i,j}\bar{z'}_{k,j'}|] \\
&\subseteq S^{\setminus j}_{i,t,k} - |\bar{S}^{\setminus j}_{i,t,k}| \in [-|\alpha_{i,j}\bar{z}_{k,j} + \alpha_{i,j}\bar{z'}_{k,j'}|, |\alpha_{i,j}\bar{z}_{k,j} + \alpha_{i,j}\bar{z'}_{k,j'}|]
\end{aligned}
$$

$$
S^{\setminus j}_{i,t,k} - b_i - A_k \in [-|\alpha_{i,j}\bar{z}_{k,j} - A'_k|, |\alpha_{i,j}\bar{z}_{k,j} - A'_k|] \subseteq \langle w^{(t)}_i, \bar{\xi}_k\rangle - |\langle w^{(t)}_i, \bar{\xi}^{(2)}_k - \bar{\xi}^{(1)}_k\rangle| \in [-|\alpha_{i,j}\bar{z}_{k,j} + \alpha_{i,j}\bar{z'}_{k,j'}|, |\alpha_{i,j}\bar{z}_{k,j} + \alpha_{i,j}\bar{z'}_{k,j'}|] \tag{163}
$$

We have $|A_k| \leq |\bar{S}^{\setminus j}_{i,t,k}| + |\alpha_{i,j}\bar{z'}_{k,j'}|$ and similarly $|A'_k| \leq |\alpha_{i,j}\bar{z'}_{k,j'}|$ Now using chernoff bound for Gaussian Concentration

$$\mathbb{E}\left[|\mathbb{1}_{S^{\setminus j}-b_i-Z\in[-[\alpha_j z_j-Z'],[\alpha_j z_j-Z']]}\right] = \left[|\mathbb{1}_{\langle w_i^{(t)},\bar{\xi}_k\rangle-|\langle w_i^{(t)},\bar{\xi}_k^{(2)}-\bar{\xi}_k^{(1)}\rangle|\in[-|\alpha_{i,j}\bar{z}_{k,j}+\alpha_{i,j}\bar{z'}_{k,j'}|,|\alpha_{i,j}\bar{z}_{k,j}+\alpha_{i,j}\bar{z'}_{k,j'}|]}\right]$$

$$\leq \mathbb{E}\left[\frac{O(|\alpha_{i,j}||\bar{z}_{k,j}|+|\alpha_{i,j}||\bar{z'}_{k,j'}|)}{Var(\langle w,\bar{\xi}_k\rangle)}\right] \tag{164}$$

$$\leq \mathbb{E}\left[\frac{O(\alpha_{i,j}|\bar{z}_{k,j}|+\alpha_{i,j}|\bar{z'}_{k,j'}|)^2}{\langle w,\bar{\xi}_k\rangle^2}\right] \tag{165}$$

Next, we bound the J_3,We have $\langle w_i,\bar{x}_k^{(2)}\rangle \geq b_i, \langle w_i,\bar{x}_k^{(1)}\rangle \leq -b_i$ we can write $\langle w_i,\bar{x}_k^{(2)}\rangle \geq b_i, \langle w_i,\bar{x}_k^{(1)}\rangle \leq -b_i$as $\bar{S}^{\setminus j}_{i,t,k}+\bar{\alpha}_{i,j}\bar{z'}_{k,j'} \geq b_i + |S^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z}_{k,j}|$

$$J_3 = \mathbb{E}\left[(\alpha_j\bar{z}_{k,j}^{(1)}+S^{\setminus j}_{i,t,k}-b_i)\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\leq -b_i}\cdot\bar{z}_{k,j}^{(1)}\right] \tag{166}$$

$$+\mathbb{E}\left[(\alpha_j\bar{z}_{k,j}^{(1)}+S^{\setminus j}_{i,t,k}+b_i)\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\leq -b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\right]$$

$$= J_{31}+J_{32} \tag{167}$$

$$J_{31} = \mathbb{E}\left[(\alpha_j\bar{z}_{k,j}^{(1)}+S^{\setminus j}_{i,t,k}-b_i)\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\leq -b_i}\cdot\bar{z}_{k,j}^{(1)}\right]$$

$$= \mathbb{E}\left[(\alpha_j\bar{z}_{k,j}^{(1)}+S^{\setminus j}_{i,t,k}-b_i)\cdot\mathbb{1}_{\bar{S}^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z'}_{k,j}\geq b_i+|S^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z}_{k,j}|}\cdot\bar{z}_{k,j}^{(1)}\right]$$

$$= \mathbb{E}\Big[(\alpha_j\bar{z}_{k,j}^{(1)}+S^{\setminus j}_{i,t,k}-b_i)\cdot(\mathbb{1}_{\bar{S}^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z'}_{k,j}\geq b_i+|S^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z}_{k,j}|} \tag{168}$$

$$-\mathbb{1}_{\bar{S}^{\setminus j}_{i,t,k}-\alpha_{i,j}\bar{z'}_{k,j}\geq b_i+|S^{\setminus j}_{i,t,k}-\alpha_{i,j}\bar{z}_{k,j}|})\cdot\bar{z}_{k,j}^{(1)}\Big]$$

$$\leq \mathbb{E}\Big[|\bar{S}^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z'}_{k,j}|\cdot\mathbb{1}_{\bar{S}^{\setminus j}_{i,t,k}\in[b_i-\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z}_{k,j}|,b_i+\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}-\alpha_{i,j}\bar{z}_{k,j}|]}\cdot \tag{169}$$

$$\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}\bar{z}_{k,j}^{(1)}\Big]$$

$$-\mathbb{E}\left[b_i\cdot\mathbb{1}_{\bar{S}^{\setminus j}_{i,t,k}\in[b_i-\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z}_{k,j}|,b_i+\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}-\alpha_{i,j}\bar{z}_{k,j}|]}\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}\bar{z}_{k,j}^{(1)}\right]$$

$$\leq \mathbb{E}\left[|\bar{S}^{\setminus j}_{i,t,k}|\cdot\mathbb{1}_{\bar{S}^{\setminus j}_{i,t,k}\in[b_i-\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z}_{k,j}|,b_i+\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}-\alpha_{i,j}\bar{z}_{k,j}|]}\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}\bar{z}_{k,j}^{(1)}\right]$$

$$+\mathbb{E}\left[|\alpha_{i,j}\bar{z'}_{k,j}|\cdot\mathbb{1}_{\bar{S}^{\setminus j}_{i,t,k}\in[b_i-\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z}_{k,j}|,b_i+\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}-\alpha_{i,j}\bar{z}_{k,j}|]}\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}\bar{z}_{k,j}^{(1)}\right]$$

$$-\mathbb{E}\left[b_i\cdot\mathbb{1}_{\bar{S}^{\setminus j}_{i,t,k}\in[b_i-\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z}_{k,j}|,b_i+\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}-\alpha_{i,j}\bar{z}_{k,j}|]}\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}\bar{z}_{k,j}^{(1)}\right]$$

$$\leq \sqrt{\mathbb{E}\left[|\bar{S}^{\setminus j}_{i,t,k}|\cdot\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}\bar{z}_{k,j}^{(1)}\right]^2}\sqrt{\mathbb{E}\left[\mathbb{1}_{\bar{S}^{\setminus j}_{i,t,k}\in[b_i-\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}+\alpha_{i,j}\bar{z}_{k,j}|,b_i+\alpha_{i,j}\bar{z'}_{k,j}+|S^{\setminus j}_{i,t,k}-\alpha_{i,j}\bar{z}_{k,j}|]}\right]^2}$$

$$+\quad\mathbb{E}\left[|\alpha_{i,j}\bar{z'}_{k,j}|\cdot\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}\bar{z}_{k,j}^{(1)}\right]$$

$$\leq \alpha_{i,j}\mathbb{E}\left[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]\cdot Pr(\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j})+\alpha_{i,j}\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)$$

$$+\alpha_{i,j}\mathbb{E}\left[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)$$

$$\leq \alpha_{i,j}\mathbb{E}\left[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]\cdot\Big[Pr(\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j})+O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\Big]$$

$$+\alpha_{i,j}\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2\mathbf{1}_{\bar{S}^{\setminus j}_{i,t,k}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)$$

For $j \in [d]$

$$J_{31} \leq \alpha_{i,j}\Big(\frac{1-\alpha_1\rho_1}{K+1}\Big)\cdot\frac{O(\log\log d)}{d}Pr(E_k^2)$$
$$+\alpha_{i,j}\big(\alpha_1\rho_1^2+\frac{1-\alpha_1\rho_1}{K+1}\big)\frac{O(\log\log d)}{d}\cdot O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2\mathbf{1}_{\bar{S}_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big) \tag{170}$$

Consider only the randomness

$$\langle w_i^{(t)},\bar{\xi}_k^{(2)}-\bar{\xi}_k^{(1)}\rangle \in [-\bar{\alpha}_{i,j}\bar{z'}_{k,j'}+|\alpha_{i,j}\bar{z}_{k,j}+\langle w_i^{(t)},\bar{\xi}_k\rangle|,\ \bar{\alpha}_{i,j}\bar{z'}_{k,j'}+|\langle w_i^{(t)},\bar{\xi}_k\rangle-\alpha_{i,j}\bar{z}_{k,j}|]$$

$$\begin{aligned}
&\mathbb{E}\Big[\delta\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq b_i}\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\Big]\\
&=\mathbb{E}\Big[(\bar{S}_{i,t,k}^{\setminus j}+\alpha_{i,j}\bar{z'}_{k,j})\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(2)}\rangle|\geq b_i}\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\Big]\\
&=\mathbb{E}\Big[\alpha_{i,j}\bar{z'}_{k,j}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\Big]+\mathbb{E}\Big[\alpha_{i,j}\bar{z'}_{k,j}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\leq -b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\leq -b_i}\cdot\bar{z}_{k,j}^{(1)}\Big]\\
&\quad+\mathbb{E}\Big[\bar{S}_{i,t,k}^{\setminus j}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\Big]+\mathbb{E}\Big[(\bar{S}_{i,t,k}^{\setminus j}+\alpha_{i,j}\bar{z'}_{k,j})\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\leq -b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\Big]\\
&\quad+\mathbb{E}\Big[(\bar{S}_{i,t,k}^{\setminus j}+\alpha_{i,j}\bar{z'}_{k,j})\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\leq -b_i}\cdot\bar{z}_{k,j}^{(1)}\Big]+\mathbb{E}\Big[\bar{S}_{i,t,k}^{\setminus j}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\leq -b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\leq -b_i}\cdot\bar{z}_{k,j}^{(1)}\Big]\\
&\leq\mathbb{E}\Big[\alpha_{i,j}\bar{z'}_{k,j}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\geq b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\geq b_i}\cdot\bar{z}_{k,j}^{(1)}\Big]+\mathbb{E}\Big[\alpha_{i,j}\bar{z'}_{k,j}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(2)}\rangle\leq -b_i}\cdot\mathbb{1}_{\langle w_i,\bar{x}_k^{(1)}\rangle\leq -b_i}\cdot\bar{z}_{k,j}^{(1)}\Big]\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2\mathbf{1}_{S_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)+\alpha_{i,j}\mathbb{E}\Big[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2\mathbf{1}_{S_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]\cdot\Big[Pr(\bar{S}_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j})+O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2\mathbf{1}_{\bar{S}_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\Big]\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2\mathbf{1}_{\bar{S}_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\\
&\leq\alpha_{i,j}\mathbb{E}\Big[\bar{z'}_{k,j}\bar{z}_{k,j}^{(1)}\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\Big]+\alpha_{i,j}\mathbb{E}\Big[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbf{1}_{E_k^1}+\mathbf{1}_{E_k^2})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]\Big[O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbf{1}_{E_k^1}+\mathbf{1}_{E_k^2})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)+Pr(\bar{S}_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j})\Big]
\end{aligned}$$

For all the neurons

$$
\begin{aligned}
&\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot \bar{z}_{k,j}^{(1)}\Big]\\
&\leq \alpha_{i,j}\mathbb{E}\Big[\bar{z}_{k,j}\bar{z}_{k,j}^{(1)}\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\Big]+\alpha_{i,j}\mathbb{E}\Big[\bar{z'}_{k,j}\bar{z}_{k,j}^{(1)}\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\Big]\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]\Big[Pr(S_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z}_{k,j})+O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2\mathbf{1}_{S_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\Big]\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]\Big[Pr(S_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z}_{k,j})+O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2\mathbf{1}_{S_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\Big]\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]\cdot\Big[Pr(\bar{S}_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j})+O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2\mathbf{1}_{\bar{S}_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\Big]\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2\mathbf{1}_{\bar{S}_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j}}]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)+\alpha_{i,j}\mathbb{E}\Big[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbf{1}_{E_k^1}+\mathbf{1}_{E_k^2})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]\Big[O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbf{1}_{E_k^1}+\mathbf{1}_{E_k^2})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)+Pr(\bar{S}_{i,t,k}^{\setminus j}\geq b_i-\alpha_{i,j}\bar{z'}_{k,j})\Big]\\
&\leq \alpha_{i,j}\mathbb{E}\Big[\bar{z}_{k,j}^{(2)}\bar{z}_{k,j}^{(1)}\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\Big]+\alpha_{i,j}\mathbb{E}\Big[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]\Big[\Pr(E_k^1)+O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbf{1}_{E_k^1}+\mathbf{1}_{E_k^2})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\Big]\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]\Big[\Pr(E_k^1)+2\Pr(E_k^2)+O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbf{1}_{E_k^1}+\mathbf{1}_{E_k^2})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\Big]\\
&\leq \alpha_{i,j}\mathbb{E}\Big[\bar{z}_{k,j}^{(2)}\bar{z}_{k,j}^{(1)}\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\Big]+\alpha_{i,j}\mathbb{E}\Big[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]\Big[\Pr(E_k^1)+O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbf{1}_{E_k^1}+\mathbf{1}_{E_k^2})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\Big]\\
&\quad+\alpha_{i,j}\mathbb{E}\Big[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\Big]\Big[\Pr(E_k^1)+2\Pr(E_k^2)+O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbf{1}_{E_k^1}+\mathbf{1}_{E_k^2})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\Big]
\end{aligned}
$$

For $j\in M^1$

$$
\begin{aligned}
\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot \bar{z}_{k,j}^{(1)}\Big]&\leq \alpha_{i,j}\alpha_1\rho_1^2O\Big(\frac{\log\log d}{d}\Big)\Big(Pr(|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|)\pm O(L_1+L_2)\Big)\\
&\quad\pm\alpha_{i,j}\frac{1-\alpha_1\rho_1}{(K+1)}\cdot\frac{O(\log\log d)}{d}O\big(L_2+L_4+L_1\big)\qquad(171)
\end{aligned}
$$

For $j\in M^2$

$$
\begin{aligned}
\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot \bar{z}_{k,j}^{(1)}\Big]&\leq \alpha_{i,j}\alpha_2\rho_2^2O\Big(\frac{\log\log d}{d}\Big)\Big(Pr(|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|)\pm O(L_1+L_2)\Big)\\
&\quad\pm\alpha_{i,j}\frac{1-\alpha_2\rho_2}{(K+1)}\cdot\frac{O(\log\log d)}{d}O\big(L_2+L_4+L_1\big)\qquad(172)
\end{aligned}
$$

Proof of Lemma G.2 (b)

For the lucky neurones $\alpha_{i,j} > b_i^{(t)}$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[h_i(\bar{x}_k^{(2)})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot \bar{z}_{k,j}^{(1)}\right]$$

$$= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(S_{i,t,k}^{\setminus j} - b_i + \alpha_{i,j}\bar{z}_{k,j} + \bar{S}_{i,t,k}^{\setminus j} + \alpha_{i,j}\bar{z'}_{k,j})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot \bar{z}_{k,j}^{(1)}\right] \tag{173}$$

$$= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(\alpha_{i,j}\bar{z}_{k,j} + \alpha_{i,j}\bar{z'}_{k,j} - b_i^{(t)})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\cdot \bar{z}_{k,j}^{(1)}\right] \tag{174}$$

$$+ \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[S_{i,t,k}^{\setminus j}\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\cdot \bar{z}_{k,j}^{(1)}\right]$$

$$+ \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\bar{S}_{i,t,k}^{\setminus j}\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\cdot \bar{z}_{k,j}^{(1)}\right] \tag{175}$$

$$+ \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(S_{i,t,k}^{\setminus j} + \alpha_{i,j}\bar{z}_{k,j})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|\geq b_i+|\langle w_i,\bar{x}_k\rangle|}\cdot \bar{z}_{k,j}^{(1)}\right]$$

$$+ \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(\bar{S}_{i,t,k}^{\setminus j} + \alpha_{i,j}\bar{z'}_{k,j})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|\geq b_i+|\langle w_i,\bar{x}_k\rangle|}\cdot \bar{z}_{k,j}^{(1)}\right]$$

and we have $1 \leq \mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}}$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[S_{i,t,k}^{\setminus j}\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\cdot \bar{z}_{k,j}^{(1)}\right]$$

$$= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[S_{i,t,k}^{\setminus j}\bar{z}_{k,j}^{(1)}\cdot \mathbb{1}_{S_{i,t,k}^{\setminus j}\in[b_i-\alpha_{i,j}\bar{z}_{k,j}+|\bar{S}_{i,t,k}^{\setminus j}+\alpha_{i,j}\bar{z'}_{k,j}|,b_i+\alpha_{i,j}\bar{z}_{k,j}+|\bar{S}_{i,t,k}^{\setminus j}-\alpha_{i,j}\bar{z'}_{k,j}|]}\right] \tag{176}$$

$$= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[|S_{i,t,k}^{\setminus j}||\bar{z}_{k,j}^{(1)}|(\mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}}\mathbb{1}_{E_{3k}^C})\cdot \mathbb{1}_{S_{i,t,k}^{\setminus j}\in[b_i-\alpha_{i,j}\bar{z}_{k,j}+|\bar{S}_{i,t,k}^{\setminus j}+\alpha_{i,j}\bar{z'}_{k,j}|,b_i+\alpha_{i,j}\bar{z}_{k,j}+|\bar{S}_{i,t,k}^{\setminus j}-\alpha_{i,j}\bar{z'}_{k,j}|]}\right]$$

$$\leq \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(|\alpha_{i,j}||\bar{z}_{k,j}| + |\bar{S}_{i,t,k}^{\setminus j}| + |\alpha_{i,j}||\bar{z'}_{k,j}|)|\bar{z}_{k,j}^{(1)}|\cdot \mathbb{1}_{S_{i,t,k}^{\setminus j}\in[b_i-\alpha_{i,j}\bar{z}_{k,j}+|\bar{S}_{i,t,k}^{\setminus j}+\alpha_{i,j}\bar{z'}_{k,j}|,b_i+\alpha_{i,j}\bar{z}_{k,j}+|\bar{S}_{i,t,k}^{\setminus j}-\alpha_{i,j}\bar{z'}_{k,j}|]}\right]$$
$$\times (\mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}}\mathbb{1}_{E_{3k}^C})$$

$$\leq \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(|\alpha_{i,j}||\bar{z}_{k,j}| + |\alpha_{i,j}||\bar{z'}_{k,j}|)|\bar{z}_{k,j}^{(1)}|\mathbb{1}_{E_{3k}}\right]$$
$$+ \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[|\bar{S}_{i,t,k}^{\setminus j}||\bar{z}_{k,j}^{(1)}|(\mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}})\cdot \mathbb{1}_{S_{i,t,k}^{\setminus j}\in[b_i-\alpha_{i,j}\bar{z}_{k,j}+|\bar{S}_{i,t,k}^{\setminus j}+\alpha_{i,j}\bar{z'}_{k,j}|,b_i+\alpha_{i,j}\bar{z}_{k,j}+|\bar{S}_{i,t,k}^{\setminus j}-\alpha_{i,j}\bar{z'}_{k,j}|]}\right]$$

$$\leq \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(|\alpha_{i,j}||\bar{z}_{k,j}| + |\alpha_{i,j}||\bar{z'}_{k,j}|)|\bar{z}_{k,j}^{(1)}|\mathbb{1}_{E_{3k}}\right] + \sqrt{\mathbb{E}\left[O(\alpha_{i,j}|\bar{z}_{k,j}| + \alpha_{i,j}|\bar{z'}_{k,j}|)^2\right]|\bar{z}_{k,j}^{(1)}|^2}\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}$$

$$\leq \alpha_{i,j}\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}| + |\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]\cdot Pr(E_{3k}) + (\alpha_{i,j}\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}| + |\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\right])O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)$$

$$\leq \alpha_{i,j}\Big(\alpha_1\rho_1^2 O\Big(\frac{\log\log d}{d}\Big) + \frac{(1-\alpha_1\rho_1)}{(K+1)}\frac{O(\log\log d)}{d}\Big)\Big[Pr(E_{3k}) + O(L_3)\Big] \quad \text{For } j \in M^1 \tag{177}$$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[h_i(\bar{x}_k^{(2)})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k^{(1)}\rangle|\geq b_i}\cdot \bar{z}_{k,j}^{(1)}\right]$$

$$\leq \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(\alpha_{i,j}\bar{z}_{k,j} + \alpha_{i,j}\bar{z'}_{k,j} - b_i^{(t)})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\cdot \bar{z}_{k,j}^{(1)}\right] + \alpha_{i,j}\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}| + |\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]$$
$$\times \left[\Pr(E_{3k}) + O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\right] + \alpha_{i,j}\mathbb{E}\left[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]\cdot\left[2\Pr(E_{3k}) + O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\right]$$
$$+ \alpha_{i,j}\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)$$

$$\leq \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(\alpha_{i,j}\bar{z}_{k,j} + \alpha_{i,j}\bar{z'}_{k,j} - b_i^{(t)})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\cdot \bar{z}_{k,j}^{(1)}\right] + \alpha_{i,j}\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]$$
$$\times \left[\Pr(E_{3k}) + O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\right] + \alpha_{i,j}\mathbb{E}\left[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]\cdot\left[3\Pr(E_{3k}) + O\Big(\sqrt{\frac{\mathbb{E}[|\bar{S}_{i,t,k}^{\setminus j}|^2(\mathbb{1}_{E_{3k}} + \mathbb{1}_{E_{4k}})]}{\mathbb{E}[\langle w,\bar{\xi}_k\rangle^2]}}\Big)\right]$$

$$\leq \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(\alpha_{i,j}\bar{z}_{k,j} + \alpha_{i,j}\bar{z'}_{k,j} - b_i^{(t)})\cdot \mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\cdot \bar{z}_{k,j}^{(1)}\right] + \alpha_{i,j}\mathbb{E}\left[|\bar{z}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]\Big(L_5 + O(L_3)\Big)$$
$$+ \alpha_{i,j}\mathbb{E}\left[|\bar{z'}_{k,j}||\bar{z}_{k,j}^{(1)}|\right]\cdot\Big(3L_5 + O(L_3)\Big)$$

For $j \in M^1$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)}) \cdot \mathbb{1}_{|\langle w_i, \bar{x}_k^{(1)}\rangle| \geq b_i} \cdot \bar{z}_{k,j}^{(1)}\Big] \leq \alpha_{i,j}\Big(\alpha_1\rho_1^2 + \frac{1-\alpha_1\rho_1}{K+1}\Big)O\Big(\frac{\log\log d}{d}\Big)\Big[Pr(|\langle w_i, \bar{x}_k\rangle| \geq b_i + |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|) \pm O\Big(L_5 + L_3\Big)\Big] \pm \alpha_{i,j}\frac{1}{(K+1)}\frac{O(\log\log d)}{d}\Big)O\Big(L_5 + L_3\Big) \tag{178}$$

Lemma G.3. *Gradient of the dense siganals for $i \in [m]$ and $jin[d]$ for the current iteration t, For $M_j \in M$ and $j \in [d]$*

$$\Big|\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i, \bar{x}_k^{(1)}\rangle| \geq b_i}\langle\bar{\xi}_k^{(1)}, M_j\rangle\Big]\Big| \leq O\Big(\frac{\|w_i^{(t)}\|_2}{d^2(K+1)^{1/2}}\Big)\max\{\Pr(|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i), \tilde{O}\Big(\frac{1}{\sqrt{d}}\Big)\} \tag{179}$$

For $M_j \in M^\perp$ and $j \in [d_1 - d]$

$$\Big|\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i, \bar{x}_k^{(1)}\rangle| \geq b_i}\langle\bar{\xi}_k^{(1)}, M_j\rangle\Big]\Big| \leq O\Big(\frac{\|w_i^{(t)}\|_2}{dd_1(K+1)^{1/2}}\Big)\max\{\Pr(|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i), \tilde{O}\Big(\frac{1}{\sqrt{d}}\Big)\} \tag{180}$$

Proof of Lemma G.3
First we can decompose the noise in the first view $\bar{\xi}_k^{(1)}$ into two orthogonal parts by projecting towards the $M_j, j \in [d_1]$ and its' orthogonal direction $(I - M_j M_j^\top)$.

$$\bar{\xi}_k^{(1)} = (I - M_j M_j^\top)\bar{\xi}_k^{(1)} + \langle\bar{\xi}_k^{(1)}, M_j\rangle M_j$$

$$\bar{x}_k^{(1)} = \bar{x'}_{1k} + \langle\bar{\xi}_k^{(1)}, M_j\rangle M_j, \text{ where } \bar{x'}_{1k} = M\bar{z}_k^{(1)} + (I - M_j M_j^\top)\bar{\xi}_k^{(1)}$$

$$\begin{aligned}
&\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)}) \cdot \mathbb{1}_{|\langle w_i, \bar{x}_k^{(1)}\rangle| \geq b_i} \cdot \langle\bar{\xi}_k^{(1)}, M_j\rangle\Big] \\
&= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[(\langle w_i^{(t)}, \bar{x}_k^{(2)}\rangle - b_i) \cdot \mathbb{1}_{|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i} \cdot \mathbb{1}_{|\langle w_i, \bar{x}_k^{(1)}\rangle| \geq b_i}\langle\bar{\xi}_k^{(1)}, M_j\rangle\Big] \\
&= \mathbb{E}_{\tilde{x'}_{1k},\mathcal{B}}\Big[(\langle w_i^{(t)}, \bar{x}_k^{(2)}\rangle - b_i) \cdot \Big(\mathbb{1}_{\langle w_i^{(t)}, \tilde{x'}_{1k}\rangle + \alpha_{i,j}|\langle\bar{\xi}_k^{(1)}, M_j\rangle| \geq b_i} - \mathbb{1}_{\langle w_i^{(t)}, \tilde{x'}_{1k}\rangle - \alpha_{i,j}|\langle\bar{\xi}_k^{(1)}, M_j\rangle| \geq b_i}\Big)\mathbb{1}_{|\langle w_i, \bar{x}_k^{(1)}\rangle| \geq b_i}\langle\bar{\xi}_k^{(1)}, M_j\rangle\Big] \\
&= \mathbb{E}_{\tilde{x'}_{1k},\mathcal{B}}\Big[(\langle w_i^{(t)}, \bar{x}_k^{(2)}\rangle - b_i)\mathbb{1}_{|\langle w_i, \bar{x}_k^{(1)}\rangle| \geq b_i}\langle\bar{\xi}_k^{(1)}, M_j\rangle\Big] \cdot \\
&\quad \mathbb{E}_{\tilde{x'}_{1k},\mathcal{B}}\Big[\Big(\mathbb{1}_{\langle w_i^{(t)}, \tilde{x'}_{1k}\rangle + \alpha_{i,j}|\langle\bar{\xi}_k^{(1)}, M_j\rangle| \geq b_i} - \mathbb{1}_{\langle w_i^{(t)}, \tilde{x'}_{1k}\rangle - \alpha_{i,j}|\langle\bar{\xi}_k^{(1)}, M_j\rangle| \geq b_i}\Big)\Big] \\
&\leq \mathbb{E}\left[\alpha_{i,j}^2|\langle\bar{\xi}_k^{(1)}, M_j\rangle|\langle\bar{\xi}_k^{(1)}, M_j\rangle^2\frac{\langle w_i^{(t)}, \bar{x}_k^{(2)}\rangle - b_i}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}\mathbb{1}_{|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i}\right]
\end{aligned}$$

For $M_j \in M$ and $j \in [d]$ in $\alpha_{i,j} = \langle w_i^{(t)}, M_j\rangle$ we have $|\langle w_i^{(t)}, M_j\rangle| \leq O\Big(\frac{\|w_i^{(t)}\|_2}{\sqrt{d}}\Big)$ and we can derive that $\mathbb{E}[|\langle\bar{\xi}_k^{(1)}, M_j\rangle|\langle\bar{\xi}_k^{(1)}, M_j\rangle^2] = O(1/(d(K+1))^{3/2})$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)}) \cdot \mathbb{1}_{|\langle w_i, \bar{x}_k^{(1)}\rangle| \geq b_i} \cdot \langle\bar{\xi}_k^{(1)}, M_j\rangle\Big] \leq \mathbb{E}\left[\alpha_{i,j}^2|\langle\bar{\xi}_k^{(1)}, M_j\rangle|\langle\bar{\xi}_k^{(1)}, M_j\rangle^2\frac{\langle w_i^{(t)}, \bar{x}_k^{(2)}\rangle - b_i}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]}\mathbb{1}_{|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i}\right] \tag{181}$$

$$\begin{aligned}
&\leq \mathbb{E}\left[O\Big(\frac{\|w_i^{(t)}\|_2^2}{d}\Big)\frac{1}{(d(K+1))^{3/2}}\right]\frac{d(K+1)}{\|w_i^{(t)}\|_2^2}\|w_i^{(t)}\|_2\Pr(|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i) \\
&\quad \times \max\{\tilde{O}\Big(\frac{1}{\sqrt{d}}\Big), \tilde{O}\Big(\frac{1}{d}\Big)\} \\
&\leq O\Big(\frac{\|w_i^{(t)}\|_2}{d^2(K+1)^{1/2}}\Big)\max\{\Pr(|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i), \tilde{O}\Big(\frac{1}{\sqrt{d}}\Big)\}
\end{aligned} \tag{182}$$

For $j \in [d_1 - d]$ dense features $M^\perp$ and , for $i \in [m]$, we have $|\langle w_i^{(t)}, M_j^\perp \rangle| \leq O\Big(\frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}}\Big)$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[h_i(\bar{x}_k^{(2)}) \cdot \mathbb{1}_{|\langle w_i, \bar{x}_k^{(1)}\rangle| \geq b_i} \cdot \langle \bar{\xi}_k^{(1)}, M_j \rangle\Big]] \leq \mathbb{E}\left[\alpha_{i,j}^2 |\langle \bar{\xi}_k^{(1)}, M_j\rangle| \langle \bar{\xi}_k^{(1)}, M_j\rangle^2 \frac{\langle w_i^{(t)}, \bar{x}_k^{(2)}\rangle - b_i}{\mathbb{E}[\langle w, \bar{\xi}_k\rangle^2]} \mathbb{1}_{|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i}\right] \tag{183}$$

$$\begin{aligned} &\leq \mathbb{E}\left[O\Big(\frac{\|w_i^{(t)}\|_2^2}{d_1}\Big)\frac{1}{(d(K+1))^{3/2}}\right]\frac{d(K+1)}{\|w_i^{(t)}\|_2^2}\|w_i^{(t)}\|_2 \Pr(|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i) \\ &\quad \times \max\{\tilde{O}\Big(\frac{1}{\sqrt{d}}\Big), \tilde{O}\Big(\frac{1}{d}\Big)\} \\ &\leq O\Big(\frac{\|w_i^{(t)}\|_2}{dd_1(K+1)^{1/2}}\Big) \cdot \max\{\Pr(|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i), \tilde{O}\Big(\frac{1}{\sqrt{d}}\Big)\} \end{aligned} \tag{184}$$

we can compute the bound of $\langle w_i^{(t)}, \bar{x}_k^{(2)}\rangle - b_i$ as

$$\langle w_i^{(t)}, \bar{x}_k^{(2)}\rangle - b_i \leq \sum_{j \in \mathcal{S}_i} |\langle w_i^{(t)}, M_j\rangle| \bar{z}_{k,j}^{(2)} + \sum_{j \notin \mathcal{S}_i} |\langle w_i^{(t)}, M_j\rangle| \bar{z}_{k,j}^{(2)} + |\langle w_i^{(t)}, \bar{\xi}_k^{(2)}\rangle|$$

Lemma G.4. *Let $i \in [m]$ suppose the following holds,*

- $\langle w_i^{(t)}, M_j\rangle^2 \geq \Omega(\frac{(b_i^{(t)})^2}{\rho^2})$ *for at most O(1) many $j \in [d]$]*
- $\langle w_i^{(t)}, M_j\rangle^2 \geq \Omega(\frac{(b_i^{(t)}/\rho)^2}{\sqrt{\log d}})$ *for at most $O(e^{-\Omega(\alpha\sqrt{\log d})}d)$*
- $\|w_i^{(t)}\|_2^2 \leq O(\frac{d(b_i^{(t)}/\rho)^2}{\alpha \log(d)})$

Then for λ

$$\Pr(|\langle w_i^{(t)}, \bar{x}_k\rangle| \geq \lambda b_i^{(t)}) \leq e^{-\Omega(\log^{1/4} d)}$$

Proof of Lemma G.4
Our proof is very similar to the proof in Lemma B.5 in[ref]. Here we define two sets $H_j := \{j \in [d] : \langle w_i^{(t)}, M_j\rangle^2 \geq \Omega(\frac{(b_i^{(t)})^2}{\rho^2})\}$ with $|H_j| = O(1)$ and $B_j := \{\langle w_i^{(t)}, M_j\rangle^2 \geq \Omega(\frac{(b_i^{(t)}/\rho)^2}{\sqrt{\log d}})\}$ with $|B_j| = O(e^{-\Omega(\alpha\sqrt{\log d})}d$.
We have $\Pr(\bar{z}_{k,j} \neq 0) = (K(2-\alpha_1\rho_1)+2)\frac{O(\log\log d)}{d}$ and $K = O(\log d)$. Since our goal is to show the iwth small probability,we can define the bad event as the union of the above two events E_{bad}.

$$E_{bad} = \{\exists j \in H_j : \bar{z}_{k,j} \neq 0\} \cup \{\exists j \in B_j \backslash H_j : \bar{z}_{k,j} \neq 0\}$$

probability of occurance.

$$\begin{aligned} \Pr(E_{bad}) &\leq \sum_{j \in B_j} \Pr(\bar{z}_{k,j} \neq 0) + \sum_{j \in H_j} \Pr(\bar{z}_{k,j} \neq 0) \leq (|B_j| + |H_j|)\Pr(\bar{z}_{k,j} \neq 0) \\ &\leq (K(2-\alpha_1\rho_1)+2)\frac{O(\log\log d)}{d} \times O(de^{-\Omega(\sqrt{\log d})}) \leq e^{-\Omega(\log^{1/4} d)} \end{aligned} \tag{185}$$

Hence we can see that the chance of appearing these event is very small. We define the good event E_{bad}^c occur w.h.p. $1 - e^{-\Omega(\log^{1/4} d)}$.

$$Y = \sum_{j \in [d]} \underbrace{\langle w_i^{(t)}, M_j\rangle z_{k,j}}_{X_j} \tag{186}$$

Since $\mathbb{E}[z_{k,j}] = 0$, $\mathbb{E}[X_j] = 0$ and $|X_j| \leq \Omega(\frac{(b_i^{(t)})^2}{\log^{1/4} d})$.Therefore, Variance proxy V,

$$V = \sum_{j \in [d]} \mathbb{E}[\langle w_i^{(t)}, M_j\rangle^2 z_{k,j}^2]$$

we have $\langle w_i^{(t)}, M_j\rangle^2 \leq O(\|w_i^{(t)}\|_2^2)$ implies

$$V \leq O(\|w_i^{(t)}\|_2^2) \sum_{j \in [d]} \mathbb{E}[z_{k,j}^2] \leq O(\frac{d(b_i^{(t)}/\rho_1)^2}{\alpha_1 \log(d)})(\alpha_1\rho_1^2 + \frac{2}{K+1})O(\frac{\log\log d}{d}) = O\Big(\frac{(b_i^{(t)})^2}{\log d}\Big) \tag{187}$$

From applying the Bernstien concentration we can derive

$$\Pr(|Y| \geq \lambda b_i^{(t)} | E_{bad}^c) \leq 2exp(-\Omega\Big(\frac{\lambda^2 (b_i^{(t)})^2}{V}\Big) \leq e^{-\Omega(\lambda \log d)} \leq e^{-\Omega(\log^{1/4} d)}$$

$$\Pr(|Y| \geq \lambda b_i^{(t)}) = \Pr(|Y| \geq \lambda b_i^{(t)} | E_{bad}) \Pr(E_{bad}) + \Pr(|Y| \leq \lambda b_i^{(t)} | E_{bad}^c) \Pr(E_{bad}^c)$$

Since $\Pr(|Y| \geq \lambda b_i^{(t)} | E_{bad}) \approx 1$ we can simplify

$$\Pr(|Y| \geq \lambda b_i^{(t)}) \leq \Pr(E_{bad}) + \Pr(|Y| \geq \lambda b_i^{(t)} | E_{bad}^c) \Pr(E_{bad}^c) \leq e^{-\Omega(\log^{1/4} d)}$$

Similiarly we can prove that for the noise term,

$$\Pr(|\langle w_i^{(t)}, \bar{\xi}_k \rangle| \geq \lambda b_i^{(t)}) \leq e^{-\Omega(\log^{1/4} d)} \tag{188}$$

Finally we can derive

$$\Pr(|\langle w_i^{(t)}, \bar{x}_k \rangle| \geq \lambda b_i^{(t)}) \leq e^{-\Omega(\log^{1/4} d)} \tag{189}$$

Lemma G.5. *Let $i \in [m]$ suppose the following holds,and if there exists $\mathcal{S}_i \subseteq [d]$ and $|\mathcal{S}_i| = O(1)$*

- $\langle w_i^{(t)}, M_j \rangle^2 \leq \Omega(\frac{(b_i^{(t)}/\rho_1)^2}{\text{polylog} d})$ *for all $j \notin \mathcal{S}_i$*
- $\|w_i^{(t)}\|_2^2 \leq O(\frac{d(b_i^{(t)}/\rho_1)^2}{\alpha_1 \text{polylog} d})$

Then for small λ we can show

$$\Pr\Big(|\sum_{j \notin \mathcal{S}_i} \langle w_i^{(t)}, M_j \rangle z_{k,j} + \langle w_i^{(t)}, \bar{\xi}_k \rangle| \geq \lambda b_i^{(t)}\Big) \leq e^{-\Omega(\log^2 d)} \tag{190}$$

The proof is exactly similar to the proof of the Lemma H.6.

Lemma G.6. *Here we compute the parameters used in above proof. From Lemma E.2 we have*

- $|\bar{S}_{i,t,k}^{\setminus j}| \leq \tilde{O}\Big(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}}\Big)$
- $Pr(E_k^1), Pr(E_k^2) \leq e^{-\Omega(log^{1/4}(d))}$ *when* $|\alpha_{i,j}| \leq (1 - c_0/2) b_i$
- $Pr(E_{3k}), Pr(E_{4k}) \leq e^{-\Omega(log^{1/4}(d))}$ *when* $|\alpha_{i,j}| \geq (1 + c_0/2) b_i$

Proof of Lemma G.6

From the Lemma E.2 we have the bound for $\langle w_i^{(t)}, M_j \bar{z}_{vk,j} \rangle$ with high probability $\Pr\Big(\langle w_i,\ M\bar{z}_{2k} \rangle \geq \|w_i^{(t)}\|_2 \max_{j \in [d]} \|M_j\|_\infty \bar{z}_{k,j}^{(2)} \log d\Big) \leq e^{-\Omega(\log^2 d)}$. In here also we follow the same argument that we used in the proof of the Lemma A.1 for the $\max_{j \in [d]} \bar{z}_{k,j}^{(2)}$.

$$\Pr\Big(\langle w_i,\ M\bar{z}_{vk} \rangle \geq \|w_i^{(t)}\|_2 \max_{j \in [d]} \|M_j\|_\infty z_{vk,j} \log d\Big) \leq e^{-\Omega(\log^2 d)}$$

$$Pr(|\sum_{j \in [d]} \langle w_i^{(t)}, M_j \rangle (\bar{z}_{k,j} \pm \bar{z'}_{k,j})| \geq \|w_i^{(t)}\|_2 \max_{j \in [d]} \|M_j\|_\infty z_{vk,j} \log d) \leq e^{-\Omega(\log^2 d)}$$

$$Pr(|\sum_{j \in [d]} \langle w_i^{(t)}, M_j \rangle| (|\bar{z}_{k,j}| + |\bar{z'}_{k,j}|) \geq \|w_i^{(t)}\|_2 \max_{j \in [d]} \|M_j\|_\infty z_{vk,j} \log d) \leq e^{-\Omega(\log^2 d)}$$

$$|\langle w_i^{(t)}, M_j \rangle| (|\bar{z}_{k,j}| + |\bar{z'}_{k,j}|) \leq \tilde{O}\Big(\frac{\|w_i^{(t)}\|_2 \rho}{\sqrt{d_1}}\Big) \text{ w.h.p.} 1 - e^{-\Omega(\log^2 d)}$$

$$|\langle w_i^{(t)}, M_j \rangle| |\bar{z'}_{k,j}| \leq \tilde{O}\Big(\frac{\|w_i^{(t)}\|_2 \rho}{\sqrt{d_1}}\Big)$$

From our definitions, we have $\bar{S}_{i,t,k}^{\setminus j} = \frac{S_{i,t,2k}^{\setminus j} - S_{i,t,1k}^{\setminus j}}{2} \implies \langle w_i^{(t)}, \frac{\bar{x}_k^{(2)\setminus j} - \bar{x}_k^{(1)\setminus j}}{2} \rangle$

$$|\bar{S}_{i,t,k}^{\setminus j}| \leq \tilde{O}\Big(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}}\Big)$$

We have $\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|](\mathbb{1}_{E^1_k}+\mathbb{1}_{E^2_k}) \le \sqrt{\mathbb{E}[|\bar{S}^{\setminus j}_{i,t,k}|^2](\mathbb{1}_{E^1_k}+\mathbb{1}_{E^2_k})}$ for any random variable

$$\sqrt{\mathbb{E}[|\bar{S}^{\setminus j}|^2](\mathbb{1}_{E^1_k}+\mathbb{1}_{E^2_k})} \le \sqrt{\tilde{O}(\frac{\|w_i^{(t)}\|_2^2}{d(K+1)})(Pr(E^1_k)+Pr(E^1_k))}$$

when $|\alpha_{i,j}| \le (1-c_0/2)b_i$ Let $S_{i,t,k}$ be sub-Gaussian with variance proxy σ.Then, for any $\epsilon > 0, \epsilon = c_0 b_i/2$

$$Pr(S^{\setminus j}_{i,t,k} \ge b_i - \alpha_{i,j}) = Pr(S^{\setminus j}_{i,t,k} \ge c_0 b_i/2) \le e^{-(c_0 b_i)^2/8\sigma^2}$$

$$Pr(S^{\setminus j}_{i,t,k} \ge c_0 b_i/2) \le e^{-(c_0 b_i)^2/8\sigma^2}$$

$$\frac{b_i^2}{8\sigma^2} \ge \Omega(\log^{1/4}(d)) => b_i \ge \sigma\Omega(\log^{1/8}(d))$$

It is given that in Lemma B4 $Pr(E^1_k) \le e^{-\Omega(log^{1/4}(d))}$

$$e^{-\Omega(log^{1/4}(d))} < \frac{1}{\log^k(d)} \quad \text{Since } log^{1/4}(d) > k\log(\log(d))$$

We relax the bound considering the rate of decay to decrease the computational complexity of our bound. For a large d, $e^{-\Omega(log^{1/4}(d))}$ decay faster than $1/polylog(d)$. we re define the probability bound

$$Pr(E^1_k), Pr(E^2_k) < \frac{1}{poly\log(d)}$$

$$\sqrt{\tilde{O}(\frac{\|w_i^{(t)}\|_2^2}{d(K+1)})(Pr(E^1_k)+Pr(E^1_k))} \le \sqrt{\tilde{O}(\frac{\|w_i^{(t)}\|_2^2}{d(K+1)})\frac{1}{poly\log(d)}} \le \frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}}\frac{1}{poly\log(d)}$$

a) When $i \in \mathcal{M}_j^{1*}$ and $\bar{z}_{k,j} \ne 0$,$|\alpha_{i,j}| \ge (1+c_0/2)b_i$,we can show that $\{|\langle w_i, \bar{x}_k^{\setminus j}\rangle| \le \alpha_{i,j}\bar{z}_{k,j} - b_i - |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|\} \subseteq \{|\langle w_i, \bar{x}_k\rangle| \ge b_i + |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|\}$.This implies that

$$\Pr(|\langle w_i, \bar{x}_k\rangle| \ge b_i + |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|) \ge 1 - \Pr(|\langle w_i, \bar{x}_k^{\setminus j}\rangle| \ge \alpha_{i,j}\bar{z}_{k,j} - b_i - |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|)$$

Then

$$\langle w_i, \bar{\bar{x}}_k^{(2)} - \bar{\bar{x}}_k^{(1)}\rangle = \langle w_i, M\bar{z'}_k\rangle + \langle w_i, \bar{\xi'}_k\rangle.$$

$$\Pr\big(|\langle w_i, \bar{\xi'}_k\rangle| \ge t\big) \le 2\exp\left(-c\,\frac{t^2}{\sigma_\xi^2\|w_i\|_2^2}\right).$$

Taking $t = C\,\|w_i^{(t)}\|_2\,\sigma_\xi\,(\log d)^{1/4}$, we obtain $|\langle w_i, \bar{\xi'}_k\rangle| \le C\,\|w_i\|_2\sigma_\xi(\log d)^{1/4}$with prob. $\ge 1 - e^{-\Omega(\log d)^{1/2}}$.

For the signal part of $\langle w_i, \bar{\bar{x}}_k^{(2)} - \bar{\bar{x}}_k^{(1)}\rangle$ we can find using the Hoeffding's inequality,$|\langle w_i, M\bar{z'}_k\rangle| \le C'\frac{\|w_i^{(t)}\|_2\rho}{\sqrt{d_1}}\log^{\frac{1}{4}} d$ with prob. $\ge 1 - e^{-\Omega(\log d)^{1/2}}$ Combined bound with probability at least

$$1 - e^{-\Omega((\log d)^{1/2})} - e^{-\Omega((\log d)^{1/2})} \;\ge\; 1 - \frac{1}{\text{polylog}(d)},$$

we have

$$|\langle w_i, \bar{\bar{x}}_k^{(2)} - \bar{\bar{x}}_k^{(1)}\rangle| \le C\,\|w_i^{(t)}\|_2\frac{(\log d)^{1/2}}{\sqrt{d(K+1)}}(\log d)^{1/4} \;+\; C'\frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}}\log^{\frac{1}{4}} d \;=:\; \Delta_i.$$

$$|\langle w_i, \bar{\bar{x}}_k^{(2)} - \bar{\bar{x}}_k^{(1)}\rangle| \;\le\; \Delta_i \quad \text{with probability } \ge 1 - \frac{1}{\text{polylog}(d)}.$$

interference control With probability at least $1 - e^{-\Omega(\log^{1/4} d)}$,

$$\langle w_i, \bar{x}_k^{\setminus j}\rangle \le \frac{c_0}{2}\,b_i.$$

augmentation difference control With probability at least $1 - e^{-\Omega(\log^{1/2} d)}$,

$$\big|\langle w_i, \bar{x}_k - \bar{\bar{x}}_k^{(1)}\rangle\big| \le \Delta_i, \qquad \Delta_i := O\Big(\|w_i\|_2\sigma_\xi\log^{1/4} d\Big).$$

We eliminate the bad event on the intersection of the events, we have

$$\langle w_i, \bar{x}_k^{\setminus j} \rangle \leq \frac{c_0}{2} b_i \quad \text{and} \quad \big|\langle w_i, \bar{x}_k - \bar{=}x_k^{(1)} \rangle\big| \leq \Delta_i.$$

Using the margin condition, it follows that

$$\alpha_{i,j}\bar{z}_{k,j} - b_i - \big|\langle w_i, \bar{x}_k - \bar{=}x_k^{(1)} \rangle\big| \geq \Big(1 + \frac{c_0}{2}\Big) b_i - b_i - \Delta_i \geq \frac{c_0}{2} b_i.$$

Therefore, from the contradiction we have ,

$$\langle w_i, \bar{x}_k^{\setminus j} \rangle \geq \alpha_{i,j}\bar{z}_{k,j} - b_i - \big|\langle w_i, \bar{x}_k - \bar{=}x_k^{(1)} \rangle\big|,$$

happens with a very low probability.Hence, by the union bound over the complements we conclude

$$\Pr\Big(|\langle w_i, \bar{x}_k^{\setminus j} \rangle| \geq \alpha_{i,j}\bar{z}_{k,j} - b_i - \big|\langle w_i, \bar{x}_k - \bar{=}x_k^{(1)} \rangle\big|\Big) \leq \frac{1}{\text{polylog}(d)}.$$

Therefore, with high probability

$$1 - Pr(|\langle w_i, \bar{x}_k^{\setminus j} \rangle| \geq \alpha_{i,j}\bar{z}_{k,j} - b_i - |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)} \rangle|) \geq 1 - \frac{1}{\text{polylog}(d)}.$$

$$\begin{aligned}
\mathbb{E}\left[\bar{z}_{k,j}^{(2)}\bar{z}_{k,j}^{(1)} \cdot \mathbb{1}_{|\langle w_i, \bar{x}_k\rangle| \geq b_i + |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|}\right] &= \Big(\frac{\rho K + 1}{K+1}\Big)^2 \frac{\alpha O(\log\log d)}{d}\Big(1 - \frac{1}{\text{polylog}(d)}\Big) \\
&= \Big(\alpha_1\rho_1^2 + \frac{1-\alpha\rho}{K+1}\Big)\frac{O(\log\log d)}{d}\Big(1 - \frac{1}{\text{polylog}(d)}\Big)
\end{aligned}$$

For $M_j \in M^1, i \in \mathcal{M}_j^{1*}$ at iteration $t+1$

$$\begin{aligned}
\langle w_i^{(t+1)}, M_j\rangle &= \langle w_i^{(t)}, M_j\rangle - \eta\langle \nabla_{w_i} \text{Obj}(f_{\mathcal{G}}^t), M_j\rangle \\
&= (1-\eta\lambda)\langle w_i^{(t)}, M_j\rangle + \mathbb{E}_{\bar{x}_k^{(1)}, \mathcal{B}}\left[(1 - \ell'_{p,t}(x_p^+, \mathcal{B})) \cdot h_{i,t}(\bar{z}_{k,j}^{(2)}) \cdot (\bar{z}_{k,j}^{(1)} + \langle \bar{\xi}_k^{(1)}, M_j\rangle)\right] \\
&\quad - \sum_{x_{n,s} \in \mathcal{B}_k^-} \mathbb{E}\left[\ell'_{s,t}(x_p, \mathcal{B}) \cdot h_{i,t}(x_{n,s}) \cdot (\bar{z}_{k,j}^{(1)} + \langle \bar{\xi}_k^{(1)}, M_j\rangle)\right] + \frac{\eta\|w_i^{(t)}\|_2}{\text{poly}(d_1)\sqrt{(K+1)}} \\
&= \langle w_i^{(t)}, M_j\rangle \left(1 - \eta\lambda + \eta\Big(\alpha_1\rho_1^2 + \frac{1-\alpha_1\rho_1}{K+1}\Big)\frac{O(\log\log d)}{d}\Big(1 - \frac{1}{\text{polylog } d}\Big)\right) - O\Big(\frac{\eta\langle w_i^{(t)}, M_j\rangle}{d \text{ polylog } d}\Big) \\
&\quad + \frac{\eta\|w_i^{(t)}\|_2}{\text{poly}(d_1)\sqrt{(K+1)}} \pm \tilde{O}\left(\frac{\eta\|w_i^{(t)}\|_2}{d^2\sqrt{K+1}}\right) \pm \tilde{O}\Big(\eta\frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d(K+1)}\|w_i^{(t)}\|_2\sqrt{\frac{\alpha_1\rho_1^2(K+1)\log\log d}{d_1}}\Big) \\
&\geq \langle w_i^{(t)}, M_j\rangle \left(1 - \eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}} + \eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big(1 - \frac{1}{\text{polylog } d}\Big)\right)
\end{aligned}$$

b) When $i \in \mathcal{M}_j^2$ and $\bar{z}_{k,j} \neq 0$,$|\alpha_{i,j}| \leq (1 - c_0/2)b_i$,we can show that $\{|\langle w_i, \bar{x}_k\rangle| \geq b_i + |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|\} \subseteq \{|\langle w_i, \bar{x}_k^{\setminus j}\rangle| \geq b_i - \alpha_{i,j}\bar{z}_{k,j} + |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|\}$.This implies that

$$\Pr(|\langle w_i, \bar{x}_k\rangle| \geq b_i + |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|) \leq \Pr(|\langle w_i, \bar{x}_k^{\setminus j}\rangle| \geq b_i - \alpha_{i,j}\bar{z}_{k,j} + |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|)$$

$$\Pr(|\langle w_i, \bar{x}_k^{\setminus j}\rangle| \geq b_i - \alpha_{i,j}\bar{z}_{k,j} + |\langle w_i, \bar{x}_k - \bar{x}_k^{(1)}\rangle|) \leq \frac{1}{\text{polylog}(d)}.$$

For $j \in [d]$ until $|\langle w_i^{(t)}, M_j\rangle| \geq \frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}}$ holds

$$\begin{aligned}
\langle w_i^{(t+1)}, M_j\rangle &= \langle w_i^{(t)}, M_j\rangle - \eta\langle \nabla_{w_i} \text{Obj}(f_{\mathcal{G}}^t), M_j\rangle \\
&= (1-\eta\lambda)\langle w_i^{(t)}, M_j\rangle + \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(1-\ell'_{p,t}(x_p^+,\mathcal{B}))\cdot h_{i,t}(\bar{z}_{k,j}^{(2)})\cdot(\bar{z}_{k,j}^{(1)} + \langle\bar{\xi}_k^{(1)}, M_j\rangle)\right] \\
&\quad - \sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}\left[\ell'_{s,t}(x_p,\mathcal{B})\cdot h_{i,t}(x_{n,s})\cdot(\bar{z}_{k,j}^{(1)} + \langle\bar{\xi}_k^{(1)}, M_j\rangle)\right] + \frac{\eta\|w_i^{(t)}\|_2}{\text{poly}(d_1)\sqrt{(K+1)}} \\
&= \langle w_i^{(t)}, M_j\rangle\left(1-\eta\lambda+\eta\alpha_2\rho_2^2\frac{O(\log\log d)}{d}\Big(\frac{1}{\text{polylog}\, d}\Big)\right) + \frac{\eta\|w_i^{(t)}\|_2}{\text{poly}(d_1)\sqrt{(K+1)}} \\
&\quad \pm\tilde{O}\left(\frac{\eta\|w_i^{(t)}\|_2}{d^2\sqrt{K+1}}\right) \pm \tilde{O}\Big(\eta\frac{\sum_{r=1}^m\|w_r^{(t)}\|_2^2}{\tau d(K+1)}\|w_i^{(t)}\|_2\sqrt{\alpha_2\rho_2^2(K+1)\log\log d}\Big) \\
&\leq \langle w_i^{(t)}, M_j\rangle\left(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}} + O(\eta'\alpha_2\rho_2^2\sqrt{\frac{K+1}{\alpha_1\rho_1^2}})\frac{1}{d\text{polylog}\, d}\right) + \tilde{O}\left(\frac{\eta'\|w_i^{(t)}\|_2}{d^2}\sqrt{\frac{\alpha_2\rho_2^2}{\alpha_1\rho_1^2}}\right) \\
&\leq \langle w_i^{(t)}, M_j\rangle\left(1+O(\eta'\alpha_2\rho_2^2\sqrt{\frac{K+1}{\alpha_1\rho_1^2}})\frac{1}{d\text{polylog}\, d}\right) + \tilde{O}\left(\frac{\eta'\|w_i^{(t)}\|_2}{d^2}\sqrt{\frac{\alpha_2\rho_2^2}{\alpha_1\rho_1^2}}\right)
\end{aligned}$$

Since our bias is grown by following update rule $b_i^{(t+1)} \geq b_i^{(t)}(1+\frac{\eta}{d})$. From the Lemma for $i \notin \mathcal{M}_j^{1*}$ we have $|\langle w_i^{(t+1)}, M_j\rangle| \leq (1-c_0)b_i^{(t+1)}$ and then we can derive from above training step and the bias growth formula $\frac{|\langle w_i^{(t+1)}, M_j\rangle|}{|\langle w_i^{(t)}, M_j\rangle|} \leq \frac{b_i^{(t+1)}}{b_i^{(t)}}$.At iteration t,

$$\frac{|\langle w_i^{(t+1)}, M_j\rangle|}{b_i^{(t+1)}} \leq \frac{|\langle w_i^{(t)}, M_j\rangle|}{b_i^{(t)}} \leq (1-c_0) \text{ if } |\langle w_i^{(t)}, M_j\rangle| \geq \frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}}$$

Similarly, for the case where $|\langle w_i^{(t)}, M_j\rangle| \leq \frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}}$we can say that

$$|\langle w_i^{(t+1)}, M_j\rangle| \leq \frac{\|w_i^{(t)}\|_2}{\sqrt{d_1}} \leq (1-c_0)b_i^{(t+1)}$$

These arguments ensure that with the bias growth, feature alignment with unlucky neurones maintain a small alignment may not misclassified with the alignments of the lucky neurones while in the training process.
From above we have,

$$|\langle w_i^{(t+1)}, M_j\rangle| \leq |\langle w_i^{(t)}, M_j\rangle|(1-\eta\lambda) + \tilde{O}\left(\frac{\eta\|w_i^{(t)}\|_2}{d^2\sqrt{K+1}}\right)$$

Similarly,

$$|\langle w_i^{(t+1)}, M_j\rangle| \geq |\langle w_i^{(t)}, M_j\rangle|(1-\eta\lambda) - \tilde{O}\left(\frac{\eta\|w_i^{(t)}\|_2}{d^2\sqrt{K+1}}\right)$$

For $M_j \in M^1, i \notin \mathcal{M}_j$ at iteration $t+1$

$$\begin{aligned}
\langle w_i^{(t+1)}, M_j\rangle &= \langle w_i^{(t)}, M_j\rangle - \eta\langle \nabla_{w_i} \text{Obj}(f_{\mathcal{G}}^t), M_j\rangle \\
&= (1-\eta\lambda)\langle w_i^{(t)}, M_j\rangle + \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(1-\ell'_{p,t}(x_p^+,\mathcal{B}))\cdot h_{i,t}(\bar{z}_{k,j}^{(2)})\cdot(\bar{z}_{k,j}^{(1)} + \langle\bar{\xi}_k^{(1)}, M_j\rangle)\right] \\
&\quad - \sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}\left[\ell'_{s,t}(x_p,\mathcal{B})\cdot h_{i,t}(x_{n,s})\cdot(\bar{z}_{k,j}^{(1)} + \langle\bar{\xi}_k^{(1)}, M_j\rangle)\right] + \frac{\eta\|w_i^{(t)}\|_2}{\text{poly}(d_1)\sqrt{(K+1)}} \\
&= \langle w_i^{(t)}, M_j\rangle\left(1-\eta\lambda+\eta\Big(\alpha_1\rho_1^2 + \frac{1-\alpha_1\rho_1}{K+1}\Big)O\Big(\frac{1}{d\,\text{polylog}\, d}\Big)\right) \\
&\pm\tilde{O}\Big(\eta\frac{\sum_{r=1}^m\|w_r^{(t)}\|_2^2}{\tau d(K+1)}\|w_i^{(t)}\|_2\sqrt{\alpha_1\rho_1^2(K+1)\log\log d}\Big) \quad + \frac{\eta\|w_i^{(t)}\|_2}{\text{poly}(d_1)\sqrt{(K+1)}} \pm \tilde{O}\left(\frac{\eta\|w_i^{(t)}\|_2}{d^2\sqrt{K+1}}\right) \\
&\leq \langle w_i^{(t)}, M_j\rangle\left(1+O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\,\text{polylog}\, d}\Big)\right) + \tilde{O}\left(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2}d^2}\right)
\end{aligned}$$

c) Since the proof of the part (c) is simila to the proof of Lemma B.2 (f), we did not include it in here.
d) Next, the training dynamics of the noise features. We use Lemma D.2 to compute dynamics

$$
\begin{aligned}
\langle w_i^{(t+1)}, M_j^{\perp}\rangle &= \langle w_i^{(t)}, M_j^{\perp}\rangle - \langle \nabla_{w_i}\mathrm{Obj}(f_{\mathcal{G}}^t), M_j^{\perp}\rangle \\
&= (1-\eta\lambda)\langle w_i^{(t)}, M_j^{\perp}\rangle + \eta\mathbb{E}_{x_p^+, x_p^{++}}\Big[(1-\ell'_{p,t}(x_p^+, B))\cdot h_{i,t}(x_p^{++})\cdot\langle \nabla_{w_i} h_{i,t}(x_p^+), M_j^{\perp}\rangle\Big] \\
&\quad - \eta\sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}\Big[\ell'_{s,t}(x_p, B)\cdot h_{i,t}(x_{n,s})\cdot\langle\nabla_{w_i}h(x_p^+), M_j^{\perp}\rangle\Big] + \frac{\eta\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)\sqrt{(K+1)}} \\
&= (1-\eta\lambda)\langle w_i^{(t)}, M_j^{\perp}\rangle \pm \tilde{O}\left(\frac{\eta\|w_i^{(t)}\|_2}{dd_1\sqrt{K+1}}\right)\max\{\Pr(|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i), \frac{1}{\sqrt{d}}\} \pm \eta O\Big(\frac{\sum_{r=1}^m \|w_r^{(t)}\|_2^2}{\tau d_1}\|w_i^{(t)}\|_2\Big) \\
&\quad + \frac{\eta\|w_i^{(t)}\|_2}{\mathrm{poly}(d_1)\sqrt{(K+1)}} \\
\langle w_i^{(t+1)}, M_j^{\perp}\rangle &\leq \langle w_i^{(t)}, M_j^{\perp}\rangle + O\Big(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2}dd_1}e^{-\Omega(\log^{1/4} d)}\Big)
\end{aligned}
$$

Let's take $\epsilon := O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\ \mathrm{polylog}\ d}\Big), \epsilon_1 = \eta'\sqrt{(K+1)\alpha_1\rho_1^2}\cdot\frac{O(\log\log d)}{d}\Big(1-\frac{1}{\mathrm{polylog}\ d}\Big)$

$$
\begin{aligned}
&\langle w_i^{(t+1)}, M_j\rangle^2 + \sum_{j'\in[d], j\neq j}\langle w_i^{(t+1)}, M_{j'}\rangle^2 + \sum_{k\in[d_1]\setminus[d]}\langle w_i^{(t+1)}, M_k^{\perp}\rangle^2 \\
&\leq \sum_{j'\in[d], j'\neq j}\langle w_i^{(t)}, M_{j'}\rangle^2(1+\epsilon)^2 + \sum_{k\in[d_1]\setminus[d]}\langle w_i^{(t)}, M_j^{\perp}\rangle^2(1+\epsilon)^2 + O\Big(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2}dd_1}\frac{1}{\mathrm{polylog}\, d}\Big)^2 + \langle w_i^{(t+1)}, M_j\rangle^2 \\
&\leq \|w^{(t)}\|_2^2(1+\epsilon)^2 - \langle w_i^{(t)}, M_j\rangle^2(1+\epsilon)^2 + \langle w_i^{(t+1)}, M_j\rangle^2 + O\Big(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2}dd_1}\frac{1}{\mathrm{polylog}\, d}\Big)^2 \\
&\leq \|w^{(t)}\|_2^2\Big((1+\epsilon)^2 + O\Big(\frac{\eta'}{\sqrt{\alpha_1\rho_1^2}dd_1}\frac{1}{\mathrm{polylog}\, d}\Big)^2\Big) + \langle w_i^{(t+1)}, M_j\rangle^2 - \langle w_i^{(t)}, M_j\rangle^2(1+\epsilon)^2 \\
&\leq \|w^{(t)}\|_2^2\Big((1+\epsilon)^2 + O\Big(\frac{\eta'}{\sqrt{\alpha_1\rho_1^2}dd_1}\frac{1}{\mathrm{polylog}\, d}\Big)^2\Big) + \Big(\frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle} - 1 - \epsilon\Big)\Big(\frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle} + 1 + \epsilon\Big)\langle w_i^{(t)}, M_j\rangle^2
\end{aligned}
$$

Since we know $\langle w_i^{(t)}, M_j\rangle^2 \leq \|w^{(t)}\|_2^2$

$$
\begin{aligned}
\frac{\|w^{(t+1)}\|_2^2}{\|w^{(t)}\|_2^2} &\leq \Big((1+\epsilon)^2 + O\Big(\frac{\eta'}{\sqrt{\alpha_1\rho_1^2}dd_1}\frac{1}{\mathrm{polylog}\, d}\Big)^2\Big) + \Big(\frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle}\Big)^2 - (1+\epsilon)^2 \\
\frac{\|w^{(t+1)}\|_2^2}{\|w^{(t)}\|_2^2} &\leq O\Big(\frac{\eta'}{\sqrt{\alpha_1\rho_1^2}dd_1}\frac{1}{\mathrm{polylog}\, d}\Big)^2 + \Big(\frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle}\Big)^2 \\
\frac{\|w^{(t+1)}\|_2}{\|w^{(t)}\|_2} &\leq \frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle}
\end{aligned}
$$

The last inequality we can derive since we know the $O\Big(\frac{\eta'}{\sqrt{\alpha_1\rho_1^2}dd_1}\frac{1}{\mathrm{polylog}\, d}\Big)$ is very samll compared to the $\frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle}$

Now we need to prove the bound of the trained unlucky neurone alignment wih the features $|\langle w_i^{(t)}, M_j\rangle| \leq O(\frac{\|w_I^{(t)}\|_2}{\sqrt{d}})$

$$\begin{aligned}
\|M^1 M^{1\top} w_i^{(t+1)}\|_2 &\geq \|M^1 M^{1\top} w_i^{(T_1)}\|_2 \cdot (1-\eta\lambda)^{(t-T_1)} - \sum_{j\in[d]} \tilde{O}\Big(\frac{(t-T_1+1)\eta \max_{T_1\leq t'\leq t+1}\|w_i^{(t')}\|_2}{d^2\sqrt{K+1}}\Big) \\
&\geq \|M^1 M^{1\top} w_i^{(T_1)}\|_2 \cdot (1-\eta\lambda)^{(t-T_1)} - O\left(\frac{\max_{T_1\leq t'\leq t+1}\|w_i^{(t')}\|_2}{\alpha_1\rho_1^2(K+1)}\right) \\
&\geq \|M^1 M^{1\top} w_i^{(T_1)}\|_2 \cdot (1-\eta\lambda)^{(t-T_1)} - O\left(\frac{\|w_i^{(T_1)}\|_2}{\alpha_1\rho_1^2\sqrt{K+1}}\right) \\
&\text{From the Theorem we have } t \geq O\Big(\frac{d\log d}{\eta'\alpha_1\rho_1^2\sqrt{(K+1)}C_\eta \log\log d}\Big) \\
&\geq \|M^1 M^{1\top} w_i^{(T_1)}\|_2 \cdot \left(1 - O\Big(\frac{d\log d}{\alpha_1\rho_1^2\log\log d}\Big)\lambda\right) - O\left(\frac{\|w_i^{(T_1)}\|_2}{\alpha_1\rho_1^2\sqrt{K+1}}\right) \\
&\geq \|M^1 M^{1\top} w_i^{(T_1)}\|_2 \cdot \Big(1 - o(1)\Big)
\end{aligned}$$

From the Theorem ,we have $|\langle w_i^{(t)}, M_j\rangle| \leq \sqrt{\frac{c_1\log d}{d}}\|M^1 M^{1\top} w_i^{(t)}\|$ and it holds for the iteration t as,

$$\begin{aligned}
|\langle w_i^{(t+1)}, M_j\rangle| &\leq \langle w_i^{(t)}, M_j\rangle \left(1 + O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\,\mathrm{polylog}\,d}\Big)\right) + \tilde{O}\left(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2 d^2}}\right) \\
&\leq \langle w_i^{(t)}, M_j\rangle \left(1 + O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\,\mathrm{polylog}\,d}\Big)\right)^{(t-T_1)} + \tilde{O}\left(\frac{\eta'(t+1-T_1)\|w_i^{(T_1)}\|_2}{\sqrt{\alpha_1\rho_1^2 d^2}}\right) \\
&\leq \sqrt{\frac{c_1\log d}{d}}\|M^1 M^{1\top} w_i^{(T_1)}\|_2(1+o(1)) \\
&\leq \sqrt{\frac{c_1\log d}{d}}\|M^1 M^{1\top} w_i^{(t+1)}\| \\
&\leq O\Big(\sqrt{\frac{c_1\log d}{d}}\Big)\|w_i^{(t+1)}\|
\end{aligned}$$

For the dense features, we can follow the similar approach

$$\begin{aligned}
\langle w_i^{(t+1)}, M_j^\perp\rangle &\leq \langle w_i^{(t)}, M_j^\perp\rangle + O\Big(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2 d d_1}} e^{-\Omega(\log^{1/4} d)}\Big) \\
&\leq \langle w_i^{(T_1)}, M_j^\perp\rangle + \sum_{t'=T_1}^{t} O\Big(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2 d d_1}} e^{-\Omega(\log^{1/4} d)}\Big)
\end{aligned}$$

From $\frac{\|w^{(t+1)}\|_2^2}{\|w^{(t)}\|_2^2} \leq \frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle}$ we can show $\|w^{(t)}\|_2 \leq \|w^{(t+1)}\|_2 \left(1 + O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\,\mathrm{polylog}\,d}\Big)\right)$

$$\begin{aligned}
&\leq \Big[\sqrt{\frac{c_1\log d}{d_1}}\|w^{(T_1+1)}\|_2 + \sum_{t'=T_1}^{t} O\Big(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2 d d_1}} e^{-\Omega(\log^{1/4} d)}\Big)\Big]\left(1 + O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\,\mathrm{polylog}\,d}\Big)\right) \\
&\leq \Big[\sqrt{\frac{c_1\log d}{d_1}}\|w^{(t+1)}\|_2 + (t-T_1+1)O\Big(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2 d d_1}} e^{-\Omega(\log^{1/4} d)}\Big)\Big]\left(1 + O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\,\mathrm{polylog}\,d}\Big)\right)^{(t-T_1+1)} \\
&\leq O\Big(\sqrt{\frac{\log d}{d_1}}\Big)\|w^{(t+1)}\|_2
\end{aligned}$$

In the last inequality we have used the property of $e^{-\Omega(\log^{1/4} d)} \leq \frac{d_1\log\log d}{\sqrt{\log d}}$ for $t < T_2 = O\Big(\frac{d\log d}{\eta C_1\log\log d}\Big)$.
From the previous derivations we can bound

$$\|M^1 M^{1\top} w_i^{(t+1)}\|_2 \geq \|M^1 M^{1\top} w_i^{(t)}\|_2 \cdot \left(1 - \eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}})\right) - O\left(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2 d^{3/2}}} e^{\Omega(\log^{1/4} d)}\right)$$

$$\|M^{\perp}(M^{\perp})^{\top}w_i^{(t+1)}\|_2 \geq \|M^{\perp}(M^{\perp})^{\top}w_i^{(t)}\|_2 \cdot \left(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}})\right) - O\left(\frac{\eta'\|w_i^{(t)}\|_2}{d\sqrt{\alpha_1\rho_1^2 d_1}}e^{\Omega(\log^{1/4} d)}\right)$$

The trained weight lower bound can derive by combining those two as

$$\|w_i^{(t+1)}\|_2^2 \geq \|w_i^{(t)}\|_2^2(1-\eta\lambda)^2 - O\left(\frac{\eta'\|w_i^{(t)}\|_2}{\sqrt{\alpha_1\rho_1^2 d^{3/2}}}e^{\Omega(\log^{1/4} d)}\right) \geq \|w_i^{(t)}\|_2^2\Big(1-\frac{O(\eta')}{\sqrt{\alpha_1\rho_1^2 d^{3/2}}\text{polylog } d}\Big)$$

Further,from our previous results we can show that, the lucky neurones $i \in \mathcal{M}_j^{1*}$ shows an exponential growth with the aligned dictionary components.

$$|\langle w_i^{(t)}, M_j\rangle| \geq |\langle w_i^{(T_1)}, M_j\rangle| \left(1+\Omega\Big(\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)\right)^{(t-T_1)}$$

We define the time $T_2 > T_1$ such that $T_2 := \{t \in \mathbb{N}, \exists i \in [m] \text{ s.t } \|w_i^{(t)}\|_2^2 \geq d\|w_i^{(T_1)}\|_2^2\}$

$$|\langle w_i^{(t)}, M_j\rangle| \geq O\Big(\sqrt{\frac{\log d}{d}}\Big)\|w^{(T_1)}\|_2 \left(1+\Omega\Big(\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)\right)^{(t-T_1)}$$

To satisfy the definition of T_2 we need to have

$$\begin{aligned}
\left(1+\Omega\Big(\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}\Big)\right)^{(t-T_1)} &\geq \frac{d}{\sqrt{\log d}} \\
t - T_1 &\geq O\Big(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\log\log d}\Big) \\
t &= O\Big(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\log\log d}\Big)
\end{aligned}$$

$$\begin{aligned}
&\sum_{\substack{j'\neq j\\ j'\in[d]}}\langle w_i^{(t)}, M_{j'}\rangle^2 + \sum_{j'\in[d_1]\setminus d}\langle w_i^{(t)}, M_{j'}^{\perp}\rangle^2 \\
&\leq \|w^{(T_1)}\|_2^2\left(1+O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\,\text{polylog } d}\Big)\right)^{2(t-T_1)} + (t-T_1)\max_{t'<t} O\left(\frac{\eta'\|w_i^{(t')}\|_2}{d^{3/2}\sqrt{\alpha_1\rho_1^2}}e^{\Omega(\log^{1/4} d)}\right) \\
&\leq \|w^{(T_1)}\|_2^2(1+o(1)) + O(\frac{1}{\alpha_1\rho_1^2\sqrt{d(K+1)}}0\|w^{(T_1)}\|_2^2) \\
&\leq O(\|w^{(T_1)}\|_2^2)
\end{aligned}$$

Further, we compute the time to reach $\|w_i^{(t)}\|_2^2 \geq d\|w_i^{(T_1)}\|_2^2$ for the upper bound of the lucky neurones as

$$\begin{aligned}
\langle w_i^{(t)}, M_j\rangle &\leq \langle w_i^{(T_1)}, M_j\rangle \left(1-O(\eta'\lambda C_\eta\sqrt{K+1}) + \eta'\alpha_1\rho_1^2 C_\eta\sqrt{K+1}\frac{O(\log\log d)}{d}\Big(1-\frac{1}{\text{polylog } d}\Big)\right)^{(t-T_1)} \\
&\leq O\Big(\sqrt{\frac{\log d}{d}}\Big)\|w_i^{(t)}\|_2\left(1+\eta'\alpha_1\rho_1^2 C_\eta\sqrt{K+1}\frac{O(\log\log d)}{d}\Big(1-\frac{1}{\text{polylog } d}\Big)\right)^{(t-T_1)} \\
\sqrt{d} &\leq O\Big(\sqrt{\frac{\log d}{d}}\Big)\left(1+\eta'\alpha_1\rho_1^2 C_\eta\sqrt{K+1}\frac{O(\log\log d)}{d}\Big(1-\frac{1}{\text{polylog } d}\Big)\right)^{(t-T_1)} \\
t - T_1 &\geq \frac{d\log(\Omega(d/\sqrt{\log d}))}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\log\log d(1-o(1))}
\end{aligned}$$

From the upper bound of the lucky neuron alignment with $M_j$where $\langle w_i^{(t)}, M_j\rangle, j \in [d], i \in \mathcal{M}_j^{1*}$ provides the information about the time start the training process of the M_js very closed to the $w_i^{(t)}, i \in \mathcal{M}_j^{1*}$ to reach $d\|w_i^{(T_1)}\|_2^2$. Similarly the lower bound shows the time (T_2) to satisfy the condition $\|w_i^{(t)}\|_2^2 \geq d\|w_i^{(T_1)}\|_2^2$ by weakly align M_j s in $i \in \mathcal{M}_j^{1*}$.Hence at this T_2, we can conclude that all the dictionary columns in $\mathcal{M}_j^{1*}$ is completely learned by $w_i^{(t)}$. We have $T_2 = O\Big(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\log\log d}\Big)$.

Theorem G.7. *For each neuron $i \in [m]$, at iteration $t = T_2$ following conditions holds.*

a) For $j \in [d]$, if $i \in \mathcal{M}_j^{1*}$,$|\langle w_i^{(T_2)}, M_j\rangle| \geq \Omega(g)\|w_i^{(T_2)}\|_2$ where $g \geq 1$

b) $b_i^{(T_2)} \geq \frac{\alpha\|w_i^{(T_2)}\|_2}{\sqrt{d}} poly\log(d)$

Proof of Theorem G.7
In the second stage defined as $t \in \{T_1, T_2\}$and the neuron $i \in [m]$ satisfy $\|w_i^{(T_2)}\|_2^2 \geq d\|w_i^{(T_1)}\|_2^2$. Here we are interested a time $T_1' = \max\{\Omega\Big(\frac{d\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}C_z\log\log d}\Big), T_2\}$. We define the set of dictionary columns learned by the neuron $i \in [m]$ as $\mathcal{S}_i \in [d], |\mathcal{S}_i| = O(1)$. From the Theorem we have at $t = T_1, \|MM^\top w_i^{(T_1)}\|_2^2 \geq \|w_i^{(T_1)}\|_2^2$

When$\|w_i^{(t)}\|_2^2 \leq (g+1)\|w_i^{(T_1)}\|_2^2, g \geq 1$

$$\begin{aligned}\sum_{j'\notin\mathcal{S}_i, j\in[d]}\langle w_i^{(T_1')}, M_{j'}\rangle^2 &\leq \sum_{j'\notin\mathcal{S}_i, j'\in[d]}\langle w_i^{(T_1)}, M_{j'}\rangle^2\left(1+O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\,\mathrm{polylog}\,d}\Big)\right)^{2(T_1'-T_1)} + \left(\frac{\eta' C_\eta\|w_i^{(t)}\|_2^2}{(\alpha_1\rho_1^2)^2(K+1)d^2}\right)\\ &\leq \|MM^\top w_i^{(T_1)}\|_2^2(1+o(1))\end{aligned}$$

For $j' \in [d_1]\backslash[d]$

$$\begin{aligned}\sum_{j'\in[d_1]\backslash[d]}\langle w_i^{(T_1')}, M_{j'}^\perp\rangle^2 &\leq \sum_{j'\in[d_1]\backslash[d]}\langle w_i^{(T_1)}, M_{j'}^\perp\rangle^2 + O\Big(\frac{\eta^2(T_1'-T_1)\max_{t'\in[T_1,T_1']}\|w_i^{(t)}\|_2^2}{d^2 d_1}e^{-\Omega(\log^{1/2}d)}\Big)\\ &\leq \|M^\perp(M^\perp)^\top w_i^{(T_1)}\|_2^2 + O\Big(\frac{\eta'\|w_i^{(T_1)}\|_2^2}{d_1}\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\frac{1}{polylog(d)}\Big)\\ &\leq \|M^\perp(M^\perp)^\top w_i^{(T_1)}\|_2^2(1+o(1))\end{aligned}$$

Now for the lucky neuron $i \in \mathcal{M}_j^{1*}$, there exist $t \leq T_2$ such that $\|w_i^{(T_2)}\|_2^2 \geq (g+1)\|w_i^{(T_1)}\|_2^2, g \geq 1$ we have ,

$$\begin{aligned}\langle w_i^{(T_1')}, M_j\rangle^2 &\geq (g+1)\|w_i^{(T_1)}\|_2^2 - \sum_{j\notin\mathcal{S}_i, j\in[d]}\langle w_i^{(T_1')}, M_j\rangle^2 - \sum_{j\in[d_1]\backslash[d]}\langle w_i^{(T_1')}, M_j^\perp\rangle^2\\ &\geq (g+1)\|w_i^{(T_1)}\|_2^2 - \|MM^\top w_i^{(T_1)}\|_2^2(1+o(1)) - \|M^\perp(M^\perp)^\top w_i^{(T_1)}\|_2^2(1+o(1))\\ &\geq \|w_i^{(T_1)}\|_2^2(g-o(1))\\ &\geq \Omega(g)\|w_i^{(T_1)}\|_2^2\end{aligned}$$

Meanwhile, we observe the bias growth at T_1' as

$$b_i^{(T_1')} \geq b_i^{(T_1)}\Big(1+\frac{\eta'}{d}\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}\Big)^{(T_1'-T_1)} \geq b_i^{(T_1)} poly\log(d) \geq \frac{\|w_i^{(t)}\|_2}{\sqrt{d}} poly\log(d)$$

To derive the above last inequality we use for the large d,$(1+\frac{\eta}{d})^T \approx e^{(T\eta/d)} \approx (\log d)^{O(1)} = polylog(d)$
When$\|w_i^{(t)}\|_2^2 \geq (g+1)\|w_i^{(T_1)}\|_2^2, g \geq 1$ and if $i \in \mathcal{M}_j^{1*}$ we have

$$\langle w_i^{(t+1)}, M_j\rangle^2 \geq \langle w_i^{(t)}, M_j\rangle^2\left(1+\Omega(\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d})\Big(1-\frac{1}{\mathrm{polylog}\,d}\Big)\right)^2$$

$$\sum_{j'\in[d], j'\notin j}\langle w_i^{(t+1)}, M_{j'}\rangle^2 \leq \sum_{j'\in[d], j'\notin j}\langle w_i^{(t)}, M_{j'}\rangle^2\left(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\,\mathrm{polylog}\,d}\Big)\right)^2$$

$$\sum_{j'\in[d_1]\backslash[d]}\langle w_i^{(t+1)}, M_{j'}^\perp\rangle^2 \leq \sum_{j'\in[d_1]\backslash[d]}\langle w_i^{(t)}, M_{j'}^\perp\rangle^2 + O\Big(\eta'\frac{\|w_i^{(t)}\|_2}{dd_1}\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}e^{-\Omega(\log^{1/4}d)}\Big)$$

This implies that

$$\sum_{j'\in[d_1]\backslash[d]}\langle w_i^{(t+1)}, M_{j'}^\perp\rangle^2 \leq \sum_{j'\in[d_1]\backslash[d]}\langle w_i^{(t)}, M_{j'}^\perp\rangle^2\left(1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}+O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\,\mathrm{polylog}\,d}\Big)\right)^2$$

Define $\epsilon_1 = \Omega(\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d}), \epsilon = O\Big(\frac{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}{d\ \mathrm{polylog}\ d}\Big)$

$$
\begin{aligned}
\|w_i^{(t+1)}\|_2^2 &\le \langle w_i^{(t+1)}, M_j\rangle^2 + (\sum_{j'\in[d],j'\notin j}\langle w_i^{(t)}, M_{j'}\rangle^2 + \sum_{j'\in[d_1]\setminus[d]}\langle w_i^{(t)}, M_{j'}^{\perp}\rangle^2)(1+\epsilon)^2 \\
&\le \langle w_i^{(t+1)}, M_j\rangle^2 + (\frac{1+\epsilon}{1+\epsilon_1})^2(1+\epsilon_1)^2(\sum_{j'\in[d],j'\notin j}\langle w_i^{(t)}, M_{j'}\rangle^2 + \sum_{j'\in[d_1]\setminus[d]}\langle w_i^{(t)}, M_{j'}^{\perp}\rangle^2) \\
&\le \langle w_i^{(t+1)}, M_j\rangle^2 + \Big(\frac{1+\epsilon}{1+\epsilon_1}\Big)^2\Big(\frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle}\Big)^2(\sum_{j'\in[d],j'\notin j}\langle w_i^{(t)}, M_{j'}\rangle^2 + \sum_{j'\in[d_1]\setminus[d]}\langle w_i^{(t)}, M_{j'}^{\perp}\rangle^2) \\
&\le \Big(\frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle}\Big)^2\Big[\langle w_i^{(t)}, M_j\rangle^2 + \sum_{j'\in[d],j'\notin j}\langle w_i^{(t)}, M_{j'}\rangle^2 + \sum_{j'\in[d_1]\setminus[d]}\langle w_i^{(t)}, M_{j'}^{\perp}\rangle^2\Big] \\
\|w_i^{(t+1)}\|_2^2 &\le \Big(\frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle}\Big)^2\|w_i^{(t)}\|_2^2
\end{aligned}
$$

This implies that

$$
|\langle w_i^{(t+1)}, M_j\rangle| \ge \langle w_i^{(t)}, M_j\rangle\frac{\|w_i^{(t+1)}\|_2}{\|w_i^{(t)}\|_2} \ge \Omega(g)\|w_i^{(t+1)}\|_2
$$

We have derived for the bias growth in prvious section $\frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle} \le \frac{b_i^{(t+1)}}{b_i^{(t)}}$

$$
\begin{aligned}
\frac{\|w_i^{(t+1)}\|_2}{\|w_i^{(t)}\|_2} \le \frac{\langle w_i^{(t+1)}, M_j\rangle}{\langle w_i^{(t)}, M_j\rangle} &\le \frac{b_i^{(t+1)}}{b_i^{(t)}} \\
b_i^{(t+1)} &\ge \frac{\|w_i^{(t+1)}\|_2}{\sqrt{d}} poly\log(d)
\end{aligned}
$$

For $i \in \mathcal{M}_j^{1*}$

$$
\langle w_i^{(t+1)}, M_j\rangle \ge \langle w_i^{(T_1)}, M_j\rangle\left(1+\Omega(\eta'\sqrt{\alpha_1\rho_1^2(K+1)}\frac{\log\log d}{d})\Big(1-\frac{1}{\mathrm{polylog}\ d}\Big)\right)^{(t-T_1)}
$$

From the Theorem we have $\langle w_i^{(T_1)}, M_j\rangle \ge \Omega(b_i^{(t)})$ and can ensure this bias growth may not impact to the learning of lucky neurones. We will have for $t - T_1 = T_1' = O(d/\eta'\sqrt{\alpha_1\rho_1^2(K+1)})$

$$
\langle w_i^{(T_1+T_1')}, M_j\rangle \ge \langle w_i^{(T_1)}, M_j\rangle polylog(d) \ge b_i^{(T_1)} poly\log(d)
$$

H CONVERGENCE

In the Convergence analysis section,we will demonstrate that the neurones will converge to sparse solutions as long as they are sparsely activated,while guaranteeing sparse representations.The formulation of our convergence theorem is given below.

Theorem H.1. *We have the following results at $t \ge T_2$*

(a) For $j \in [d]$ and $i \in \mathcal{M}_j^{1\star}$,then $|\langle w_i^{(t)}, M_j\rangle| \in [\frac{\tau}{\Xi_2\alpha_1\rho^2}, O(1)]$

(b) For $i \notin \mathcal{M}_j$, then $|\langle w_i^{(t)}, M_j\rangle| \le O(\frac{1}{\sqrt{d(K+1)}})\|w_i^{(t)}\|_2$

(c) For $j \in [d'] \setminus [d]$ and $i \in \mathcal{M}_j^2$,then $|\langle w_i^{(t)}, M_j\rangle| \le O(\frac{\alpha_2\rho_2^2}{\alpha_1\rho_1^2}\frac{1}{\sqrt{d(K+1)}})\|w_i^{(t)}\|_2$

(d) For all neurons and $j \in [d_1] \setminus [d+d']$, it holds $|\langle w_i^{(t)}, M_j^{\perp}\rangle| \le O(\frac{1}{\sqrt{d_1(K+1)}})\|w_i^{(t)}\|_2$

(e) At the iteration $T \in [\frac{d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}, \frac{d^2}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}]$ *We can guarantee the contrastive loss convergence*

$$\frac{1}{T}\sum_{t=T_3}^{T_3+T-1}\mathbb{E}\big[L(f_{\mathcal{G}}, \bar{x}_k^{(1)}, \bar{x}_k^{(2)}\mathfrak{B}_k^-)\big] \le O\Big(\frac{1}{\alpha_1\rho_1^2(K+1)\cdot\log d}\Big)$$

We introduce the following induction hypothesis which holds throughout the stage iii to support to the proof of the Theorem.

Induction Hypothesis H.2. For all $t \ge T_2$:

1. If $i \in M_j^\star$, then $|\langle w_i^{(t)}, M_j\rangle| \ge \Omega(1)\|w_i^{(t)}\|_2$;
2. For $i \in [m]$, we have $\|w_i^{(t)}\|_2 \le O(1)$;
3. For each $j \in [d]$, $\mathfrak{F}_j^{(t)} := \sum_{i\in M_j}\langle w_i^{(t)}, M_j\rangle^2 \le O(\frac{\tau}{\alpha_1\rho_1^2}\log d)$;
4. Let $j \in [d]$ and $i \in M_j^\star$, there exists $C = \Theta(1)$ such that $|\langle w_i^{(t)}, M_j\rangle| \ge C\max_{i'\in M_j}|\langle w_{i'}^{(t)}, M_j\rangle|$;
5. For $i \notin M_j$, it holds $|\langle w_i^{(t)}, M_j\rangle| \le O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}\Xi_2^5})$;
6. For any $i \in [m]$ and any $j \in [d_1]\setminus[d]$, it holds $|\langle w_i^{(t)}, M_j^\perp\rangle| \le O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d_1(K+1)}\Xi_2^5})$;
7. The bias $b_i^{(t)} \ge \frac{\mathrm{polylog}(d)}{\sqrt{d}}\|w_i^{(t)}\|_2$.

Definition H.3. For all $i \in [m], j \in [d]$ we enlarge our initial expression for the gradient computed as the sum of $\Psi_{i,j}^{(t)}, \Phi_{i,j}^{(t)}$, and $\xi_{i,j}^{(t)}$ defined below.

Gradient of the loss function is

$$\begin{aligned}\langle\nabla_{w_i}L(f_{\mathcal{G}}^t), M_j\rangle &= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})h_{i,t}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}h_{i,t}(x_{n,s})\right)\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\ge b_i^{(t)}}\langle\bar{x}_k^{(1)}, M_j\rangle\right]\\ &= \Psi_{i,j}^{(t)} + \Phi_{i,j}^{(t)} + \xi_{i,j}^{(t)}\end{aligned}$$

Now we can write the update rule by substituting the above notations as

$$|\langle w_i^{(t+1)}, M_j\rangle| = (1-\eta\lambda)|\langle w_i^{(t)}, M_j\rangle| - \langle\nabla_{w_i}L(f_{\mathcal{G}}^t), M_j\rangle = (1-\eta\lambda)|\langle w_i^{(t)}, M_j\rangle| - \Psi_{i,j}^{(t)} - \Phi_{i,j}^{(t)} - \xi_{i,j}^{(t)}$$

Here we define the $\Psi_{i,j}^{(t)}, \Phi_{i,j}^{(t)}$, and $\xi_{i,j}^{(t)}$ which valid for each node of the graph views

$$\begin{aligned}\psi_{i,j}^{(t)}(x) &= \Big(\langle w_i^{(t)}, M_j\rangle\bar{z}_{k,j}^{(2)} - b_i^{(t)}\Big)\mathbb{1}_{\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle > b_i^{(t)}} - \Big(\langle w_i^{(t)}, M_j\rangle\bar{z}_{k,j}^{(2)} + b_i^{(t)}\Big)\mathbb{1}_{\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle < -b_i^{(t)}}\\ \phi_{i,j}^{(t)}(x) &= \langle w_i^{(t)}, \bar{x}_k^{(2)\setminus j}\rangle\mathbb{1}_{\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle > b_i^{(t)}} - \langle w_i^{(t)}, \bar{x}_k^{(2)\setminus j}\rangle\mathbb{1}_{\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle < -b_i^{(t)}}\\ \Psi_{i,j}^{(t)} &:= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot\psi_{i,j}^{(t)}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot\psi_{i,j}^{(t)}(x_{n,s})\right)\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\ge b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right]\\ \Phi_{i,j}^{(t)} &:= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot\phi_{i,j}^{(t)}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot\phi_{i,j}^{(t)}(x_{n,s})\right)\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\ge b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right]\\ \xi_{1,i,j}^{(t)} &:= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot h_{i,t}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot h_{i,t}(x_{n,s})\right)\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\ge b_i^{(t)}}\langle M_j, \bar{\xi}_k^{(1)}\rangle\right]\end{aligned}$$

UPPER AND LOWER BOUNDS OF THE $\langle \nabla_{w_i} L(f^t_{\mathcal{G}}), M_j \rangle$

Lemma H.4. *For $j \in [d]$ and $i \in \mathcal{M}_j^{1*}$ suppose the Induction Hypothesis hold at iteration t, and there exist a constant H_1 such that$\sum_{i \in \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle|^2 \leq H_1 \frac{\tau}{\alpha_1 \rho_1^2} \log d$,we can derive the upper bound of $\Psi_{i,j}^{(t)}$*

$$\Psi_{i,j}^{(t)} Sign(\langle w_i^{(t)}, M_j \rangle) \geq \frac{\alpha\rho}{\text{polylog}\, d}\Big(1 - O\Big(\rho^2 \cdot \frac{\log\log d}{\Xi_2^3}\Big)\Big) \cdot |\langle w_i^{(t)}, M_j \rangle - b_{1i}^{(t)}| \mathbb{E}_{\bar{x}_k^{(1)}, \mathcal{B}}[\bar{z}_j]$$

Proof of Lemma H.4
we start the proof by decomposing the $\Psi_{i,j}^{(t)}$ in to two parts

$$\begin{aligned}
\Psi_{i,j}^{(t)} &:= \mathbb{E}_{\bar{x}_k^{(1)}, \mathcal{B}}\left[(1 - \ell'_{p,t}) \cdot \psi_{i,j}^{(t)}(\bar{x}_k^{(2)}) \mathbb{1}_{\left|\langle w_i^{(t)}, \bar{x}_k^{(1)} \rangle\right| \geq b_i^{(t)}} \bar{z}_{k,j}^{(1)}\right] + \mathbb{E}_{\bar{x}_k^{(1)}, \mathcal{B}}\left[\sum_{x_{n,s} \in \mathcal{B}_k^-} \ell'_{s,t} \cdot \psi_{i,j}^{(t)}(x_{n,s}) \mathbb{1}_{\left|\langle w_i^{(t)}, \bar{x}_k^{(1)} \rangle\right| \geq b_i^{(t)}} \bar{z}_{k,j}^{(1)}\right] \\
&= \Psi_{1,i,j}^{(t)} + \Psi_{2,i,j}^{(t)}
\end{aligned}$$

We first bound the $\Psi_{2,i,j}^{(t)}$

$$\begin{aligned}
\Psi_{2,i,j}^{(t)} = &\sum_{x_{n,s} \in \mathcal{B}_k^-} \mathbb{E}_{\bar{x}_k^{(1)}, \mathcal{B}}\left[(\ell'_{s,t}(\bar{x}_k^{(1)}, \mathcal{B}) - \ell'_{s,t}(\bar{x}_k^{(1)\setminus j}, \mathcal{B})) \cdot \psi_{i,j}^{(t)}(x_{n,s}) \mathbb{1}_{\left|\langle w_i^{(t)}, \bar{x}_k^{(1)} \rangle\right| \geq b_i^{(t)}} \bar{z}_{k,j}^{(1)}\right] \\
&+ \sum_{x_{n,s} \in \mathcal{B}_k^-} \mathbb{E}_{\bar{x}_k^{(1)}, \mathcal{B}}\left[\ell'_{s,t}(\bar{x}_k^{(1)\setminus j}, \mathcal{B}) \cdot \psi_{i,j}^{(t)}(x_{n,s}) \mathbb{1}_{\left|\langle w_i^{(t)}, \bar{x}_k^{(1)} \rangle\right| \geq b_i^{(t)}} \bar{z}_{k,j}^{(1)}\right]
\end{aligned}$$

We apply Newton-Leibniz formula to bound the first term by defining $\bar{x}_k^{(1)}(\nu) = \bar{x}_k^{(1)\setminus j} + \zeta M \bar{z}_{k,j}^{(1)}$

$$\begin{aligned}
\ell'_{s,t}(\bar{x}_k^{(1)}, \mathcal{B}) - \ell'_{s,t}(\bar{x}_k^{(1)\setminus j}, \mathcal{B}) &= \int_0^1 \frac{\partial}{\partial \zeta}(\ell'_{s,t}(\bar{x}_k^{(1)\setminus j} + \zeta M \bar{z}_{k,j}^{(1)}, \mathcal{B}))\partial\zeta \\
\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta), \mathcal{B}) &= \frac{e^{\text{Sim}_{f^t_{\mathcal{G}}}(\bar{x}_k^{(1)}(\zeta), x_{n,s})/\tau}}{\sum_{x \in B} e^{\text{Sim}_{f^t_{\mathcal{G}}}(\bar{x}_k^{(1)}(\tau), x)/\tau}} = \frac{e^{\langle f(\bar{x}_k^{(1)\setminus j}) + \zeta(f(\bar{x}_k^{(1)}) - f(\bar{x}_k^{(1)\setminus j})), f(\bar{x}_{n,s})\rangle/\tau}}{\sum_{x \in B} e^{\langle f(\bar{x}_k^{(1)\setminus j}) + \zeta(f(\bar{x}_k^{(1)}) - f(\bar{x}_k^{(1)\setminus j})), x\rangle/\tau}} \\
\ell'_{p,t}(\bar{x}_k^{(1)}(\zeta), \mathcal{B}) &= \frac{e^{\text{Sim}_{f^t_{\mathcal{G}}}(\bar{x}_k^{(1)}(\zeta), \bar{x}_k^{(1)})/\tau}}{\sum_{x \in B} e^{\text{Sim}_{f^t_{\mathcal{G}}}(\bar{x}_k^{(1)}(\tau), x)/\tau}} = \frac{e^{\langle f(\bar{x}_k^{(1)\setminus j}) + \zeta(f(\bar{x}_k^{(1)}) - f(\bar{x}_k^{(1)\setminus j})), f(\bar{x}_k^{(1)})\rangle/\tau}}{\sum_{x \in B} e^{\langle f(\bar{x}_k^{(1)\setminus j}) + \zeta(f(\bar{x}_k^{(1)}) - f(\bar{x}_k^{(1)\setminus j})), x\rangle/\tau}} \\
\frac{\partial}{\partial \zeta}(\ell'_{s,t}(\bar{x}_k^{(1)\setminus j} + \zeta M \bar{z}_{k,j}^{(1)}, \mathcal{B})) &= \ell'_{s,t}(\bar{x}_k^{(1)}(\zeta), \mathcal{B})(1 - \ell'_{s,t}(\bar{x}_k^{(1)}(\zeta), \mathcal{B}))\frac{1}{\tau}\langle f(\bar{x}_k^{(1)}) - f(\bar{x}_k^{(1)\setminus j}), f(x_{n,s})\rangle \\
&\quad - \frac{1}{\tau} \sum_{x_{n,u} \in \mathcal{N} \setminus \{x_{n,s}\}} \ell'_{s,t}(\bar{x}_k^{(1)}(\zeta), \mathcal{B}) \ell'_{u,t}(\bar{x}_k^{(1)}(\zeta), \mathcal{B}) \langle f(\bar{x}_k^{(1)}) - f(\bar{x}_k^{(1)\setminus j}), f(x_{n,u})\rangle \\
&\quad - \frac{1}{\tau} \ell'_{p,t}(\bar{x}_k^{(1)}(\zeta), \mathcal{B}) \ell'_{s,t}(\bar{x}_k^{(1)}(\zeta), \mathcal{B}) \langle f(\bar{x}_k^{(1)}) - f(\bar{x}_k^{(1)\setminus j}), f(\bar{x}_k^{(2)})\rangle
\end{aligned}$$

Now we apply the above expansion to our original expression.

$$
\begin{aligned}
&\sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[(\ell'_{s,t}(\bar{x}_k^{(1)},\mathcal{B})-\ell'_{s,t}(\bar{x}_k^{(1)\setminus j},\mathcal{B}))\cdot\psi_{i,j}^{(t)}(x_{n,s})\mathbf{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\Big]\\
&=\frac{1}{\tau}\sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\int_0^1\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})(1-\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B}))\\
&\qquad\langle f(\bar{x}_k^{(1)})-f(\bar{x}_k^{(1)\setminus j}),f(x_{n,s})\rangle\partial\zeta\cdot\psi_{i,j}^{(t)}(x_{n,s})\mathbf{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\Bigg]\\
&\quad-\frac{1}{\tau}\sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\int_0^1\sum_{x_{n,u}\in\mathcal{N}\setminus\{x_{n,s}\}}\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\ell'_{u,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\\
&\qquad\langle f(\bar{x}_k^{(1)})-f(\bar{x}_k^{(1)\setminus j}),f(x_{n,u})\rangle\partial\zeta\psi_{i,j}^{(t)}(x_{n,s})\mathbf{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\Bigg]\\
&\quad-\frac{1}{\tau}\sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\int_0^1\ell'_{p,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\\
&\qquad\langle f(\bar{x}_k^{(1)})-f(\bar{x}_k^{(1)\setminus j}),f(\bar{x}_k^{(2)})\rangle\partial\zeta\cdot\psi_{i,j}^{(t)}(x_{n,s})\mathbf{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\Bigg]\\
&=B_1+B_2+B_3
\end{aligned}
$$

From $B_1+B_2+B_3$, aplying the independance of the samples in $\mathcal{N}$ and the triangular inequality. Moreover, we can show $(1-\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B}))+\sum_{x_{n,u}\in\mathcal{N}\setminus\{x_{n,s}\}}\ell'_{u,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})=2-2\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B}))-\ell'_{p,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B}))$. Since we know $0\leq\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})),\ell'_{p,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B}))\leq 1$, we have $2-2\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B}))-\ell'_{p,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B}))\leq 2$.

$$
\begin{aligned}
B_1+B_2&=\frac{|\mathcal{N}|}{\tau}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\Big(\int_0^1\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})(1-\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B}))+\sum_{x_{n,u}\in\mathcal{N}\setminus\{x_{n,s}\}}\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\ell'_{u,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\Big)\partial\zeta\\
&\quad\times\max_{x_{n,u}\in\mathcal{N}\setminus\{x_{n,s}\}}|\langle f(\bar{x}_k^{(1)})-f(\bar{x}_k^{(1)\setminus j}),f(x_{n,s})\rangle||\psi_{i,j}^{(t)}(x_{n,s})|\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}|\bar{z}_{k,j}^{(1)}|\Bigg]\\
&\leq\frac{|\mathcal{N}|}{\tau}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\int_0^1\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\partial\zeta\max_{x\in\mathcal{B}}\Big(\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle||h_{i,t}(x)||\Big)|\psi_{i,j}^{(t)}(x_{n,s})||\bar{z}_{k,j}^{(1)}|^2\Bigg]\\
&\quad+\frac{|\mathcal{N}|}{\tau}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\Big(\int_0^1\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\partial\zeta(\max_{i\notin\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|\tilde{O}(\Xi_2))|\psi_{i,j}^{(t)}(x_{n,s})||\bar{z}_{k,j}^{(1)}|^2\Bigg]
\end{aligned}
$$

we have $\sum_{i\in[m]}\mathbb{1}_{h_{i,t}(x)\neq 0}\leq\tilde{O}(\Xi_2)$

$$
\begin{aligned}
|\langle f(\bar{x}_k^{(1)})-f(\bar{x}_k^{(1)\setminus j}),f(x)\rangle|&\leq\sum_{i\in[m]}|\langle w_i^{(t)},M_j\rangle||h_{i,t}(x)||\bar{z}_{k,j}^{(1)}|\\
&=\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle||h_{i,t}(x)||\bar{z}_{k,j}^{(1)}|+\max_{i\notin\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|\tilde{O}(\Xi_2)|\bar{z}_{k,j}^{(1)}|
\end{aligned}
$$

We can expand the above terms to find the high probability bound applying the **Lemma A2** to $h_{i,t}(x_{n,s})$

$$
\sum_{i\in[m]}|\langle w_i^{(t)},M_j\rangle||h_{i,t}(x)|=\sum_{j'\in[d]}\sum_{i\in\mathcal{M}_{j'}}|\langle w_i^{(t)},M_j\rangle||\langle w_i^{(t)},M_{j'}\rangle||\bar{z}_{n,s,j'}|+\sum_{i\in[m]}|\langle w_i^{(t)},M_j\rangle|\cdot O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}})
$$

For each j' we partition the index set $\mathcal{M}_{j'}$ into two disjoint sets those neurons that are also associated with feature j,$(\mathcal{M}_{j'} \cap \mathcal{M}_j)$and those that are not $(\mathcal{M}_{j'} \backslash \mathcal{M}_j)$

$$
\begin{aligned}
&= \sum_{j' \in [d]} \sum_{i \in \mathcal{M}_{j'} \cap \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle| |\langle w_i^{(t)}, M_{j'} \rangle| |\bar{z}_{n,s,j'}| + \sum_{j' \in [d]} \sum_{i \in \mathcal{M}_{j'} \backslash \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle| |\langle w_i^{(t)}, M_{j'} \rangle| |\bar{z}_{n,s,j'}| \\
&\quad + \sum_{i \in [m]} |\langle w_i^{(t)}, M_j \rangle| \cdot O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}}) \\
&= \sum_{i \in \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle| |\langle w_i^{(t)}, M_j \rangle| |\bar{z}_{n,s,j}| + \sum_{j' \neq j} \sum_{i \in \mathcal{M}_{j'} \cap \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle| |\langle w_i^{(t)}, M_{j'} \rangle| |\bar{z}_{n,s,j'}| \\
&\quad + \sum_{j' \in [d]} \sum_{i \in \mathcal{M}_{j'} \backslash \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle| |\langle w_i^{(t)}, M_{j'} \rangle| |\bar{z}_{n,s,j'}| + \sum_{i \in [m]} |\langle w_i^{(t)}, M_j \rangle| \cdot O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}})
\end{aligned}
$$

We define the set of common features in $\mathcal{M}_j$ and $\mathcal{M}_{j'}$ of neuron $i \in [m]$ as $\Gamma_j := \{j' \neq j, j' \in [d], \mathcal{M}_{j'} \cap \mathcal{M}_j \neq \emptyset\}$ such that $|\Gamma_j| \leq O(\log d)$. We know if $j' \notin \Gamma_j$, $\mathcal{M}_{j'} \cap \mathcal{M}_j = \emptyset$. Then we can derive the bound by assuming the uniformity of the overlap of the sets $\mathcal{M}_{j'}$ and $\mathcal{M}_j$ with $|\mathcal{M}_j| = O(\Xi_2)$. For each $j \in [d]$ we have $\sum_{i \in \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle|^2 \leq H_1 \frac{\tau}{\alpha_1 \rho_1^2} \log d$

$$
\begin{aligned}
\left| \sum_{j' \neq j} \sum_{i \in \mathcal{M}_{j'} \cap \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle| |\langle w_i^{(t)}, M_{j'} \rangle| |\bar{z}_{n,s,j'}| \right| &\leq \sum_{j' \in \Gamma_j} \left[\Big(\sum_{i \in \mathcal{M}_j \cap \mathcal{M}_{j'}} |\langle w_i^{(t)}, M_{j'} \rangle|^2 \Big)^{1/2} \Big(\sum_{i \in \mathcal{M}_j \cap \mathcal{M}_{j'}} |\langle w_i^{(t)}, M_j \rangle|^2 \Big)^{1/2} |\bar{z}_{n,s,j'}| \right] \\
&\leq \sum_{j' \in \Gamma_j} \left[\frac{|\mathcal{M}_j \cap \mathcal{M}_{j'}|}{|\mathcal{M}_j|} \Big(\sum_{i \in \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle|^2 \Big)^{\frac{1}{2}} \Big(\sum_{i \in \mathcal{M}_{j'}} |\langle w_i^{(t)}, M_{j'} \rangle|^2 \Big)^{\frac{1}{2}} |\bar{z}_{n,s,j'}| \right] \\
&\leq O\Big(\frac{\log d}{\Xi_2} \Big) \sum_{j' \in \Gamma_j} \Big[\sum_{i \in \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle|^2 |\bar{z}_{n,s,j'}| \Big] \\
&\leq O\Big(\frac{\log d}{\Xi_2} \Big) \sum_{j' \in \Gamma_j} \Big[H_1 \frac{\tau}{\alpha_1 \rho_1^2} \log d \cdot |\bar{z}_{n,s,j'}| \Big] \\
&\leq O\Big(\frac{H_1 \tau \log^3 d}{\Xi_2 \alpha_1 \rho_1^2 (K+1)} \Big) \\
&\leq O\Big(\frac{\tau}{\alpha_1 \rho_1^2 (K+1) \log d} \Big)
\end{aligned}
$$

For the second term of the above can be bounded. Here we are interested in the neurons $i \in \mathcal{M}_{j'} \backslash \mathcal{M}_j$. In other words $i \notin \mathcal{M}_j$; we have proved in past stages that for $i \notin \mathcal{M}_j$, $|\langle w_i^{(t)}, M_j \rangle| \leq O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d}})$ and $\|w_i^{(t)}\|_2 \leq O(1)$

$$
\begin{aligned}
\sum_{j' \in [d]} \sum_{i \in \mathcal{M}_{j'} \backslash \mathcal{M}_j} |\langle w_i^{(t)}, M_j \rangle| |\langle w_i^{(t)}, M_{j'} \rangle| |\bar{z}_{n,s,j'}| &\leq O\Big(\frac{\|w_i^{(t)}\|_2}{\sqrt{d}} \Big) \sum_{j' \in [d]} \sum_{i \in \mathcal{M}_{j'} \backslash \mathcal{M}_j} |\langle w_i^{(t)}, M_{j'} \rangle| |\bar{z}_{n,s,j'}| \\
&\leq \sum_{j' \in [d]} \sum_{i \in \mathcal{M}_{j'}} |\langle w_i^{(t)}, M_j \rangle| |\langle w_i^{(t)}, M_{j'} \rangle| |\bar{z}_{n,s,j'}| \\
&\leq O\left(\frac{1}{\sqrt{d}} \right) \sum_{j' \in [d]} |\mathcal{M}_{j'}| \max_{i \in \mathcal{M}_{j'}} |\langle w_i^{(t)}, M_{j'} \rangle| |\bar{z}_{n,s,j'}| \\
&\leq O\left(\frac{\Xi_2}{\sqrt{d}} \right) \sum_{j' \in [d]} |\bar{z}_{n,s,j'}|
\end{aligned}
$$

Combining the both of above

$$\begin{aligned}\sum_{i\in[m]}|\langle w_i^{(t)},M_j\rangle||h_{i,t}(x)| &\le \sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle||\langle w_i^{(t)},M_j\rangle||\bar{z}_{n,s,j}| + O\left(\frac{\Xi_2}{\sqrt{d}}\right)\sum_{j'\in[d]}|\bar{z}_{n,s,j'}| \\ &\quad + O\Big(\frac{\tau}{\alpha_1\rho_1^2(K+1)\log d}\Big) + \sum_{i\in[m]}|\langle w_i^{(t)},M_j\rangle|\cdot O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}}) \\ &\le \sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle||\langle w_i^{(t)},M_j\rangle||\bar{z}_{n,s,j}| + O\Big(\frac{\tau}{\alpha_1\rho_1^2(K+1)\log d}\Big)\end{aligned}$$

To remove the dependance of the j of the $x_{n,s}$ we define a set $\mathcal{B}^{\setminus s}=\{\bar{x}_k^{(2)}\cup\mathcal{N}\setminus x_{n,s}\},\mathcal{B}'=\{\mathcal{B}^{\setminus s}\cup x_{n,s}^{\setminus j}\}$.
From the fact of $\sum_{x_{n,s}\in\mathcal{N}}\ell'_{s,t}(\bar{x}_k^{(1)}(1),\mathcal{B}')\le 1$ and max

$$\frac{\ell'_{s,t}(\bar{x}_k^{(1)}(1),\mathcal{B})}{\ell'_{s,t}(\bar{x}_k^{(1)}(1),\mathcal{B}')}\le\frac{e^{\langle f(\bar{x}_k^{(1)}),f(x_{n,s})\rangle/\tau}}{e^{\langle f(\bar{x}_k^{(1)}),f(x_{n,s}^{\setminus j})\rangle/\tau}}\frac{\sum_{x\in B'}e^{\langle f(\bar{x}_k^{(1)}),f(x)\rangle/\tau}}{\sum_{x\in B}e^{\langle f(\bar{x}_k^{(1)}),f(x)\rangle/\tau}}\le e^{\langle f(\bar{x}_k^{(1)}),f(x_{n,s})-f(x_{n,s}^{\setminus j})\rangle/\tau}\le e^{H_1\log d+O(\frac{1}{(K+1)\log d})}=o(d/\Xi_2^5)$$

$$\begin{aligned}B_1 &\le \frac{|\mathcal{N}|}{\tau}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\int_0^1\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\partial\zeta\max_{x\in\mathcal{B}}\Big(\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle||h_{i,t}(x)|\Big)|\psi_{i,j}^{(t)}(x_{n,s})||\bar{z}_{k,j}^{(1)}|^2\right] \\ &\le \frac{1}{\tau}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\sum_{x_{n,s}\in\mathcal{N}}\int_0^1\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\partial\zeta\Big(\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|^2|\bar{z}_{k,j}^{(1)}|+O\Big(\frac{\tau}{\alpha_1\rho_1^2(K+1)\log d}\Big)\Big)|\psi_{i,j}^{(t)}(x_{n,s})||\bar{z}_{k,j}^{(1)}|^2\right] \\ &\le \frac{1}{\tau}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\sum_{x_{n,s}\in\mathcal{N}}\int_0^1\ell'_{s,t}(\bar{x}_k^{(1)}(1),\mathcal{B}')\partial\zeta\Big(\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|^2|\bar{z}_{k,j}^{(1)}|+O\Big(\frac{\tau}{\alpha_1\rho_1^2(K+1)\log d}\Big)\Big)|\psi_{i,j}^{(t)}(x_{n,s})||\bar{z}_{k,j}^{(1)}|^2\right] \\ &\le \frac{O(d)}{\tau\Xi_2^5}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\int_0^1\sum_{x_{n,s}\in\mathcal{N}}\ell'_{s,t}(\bar{x}_k^{(1)}(1),\mathcal{B}')|\psi_{i,j}^{(t)}(x_{n,s})|\partial\zeta\Big(\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|^2|\bar{z}_{k,j}^{(1)}|+O\Big(\frac{\tau}{\alpha_1\rho_1^2(K+1)\log d}\Big)\Big)|\bar{z}_{k,j}^{(1)}|^2\right] \\ &\le \frac{O(d)}{\tau\Xi_2^5}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[|\psi_{i,j}^{(t)}(x_{n,s})|\Big(\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|^2|\bar{z}_{k,j}^{(1)}|+O\Big(\frac{\tau}{\alpha_1\rho_1^2(K+1)\log d}\Big)\Big)|\bar{z}_{k,j}^{(1)}|^2\right] \\ &\le O\Big(\frac{\log\log d}{\tau\Xi_2^5}\Big)\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\Big|\langle w_i^{(t)},M_j\rangle z_{n,s,j}-b_i^{(t)}\Big|\mathbb{1}_{\{|\langle w_i^{(t)},x_{n,s}\rangle|\ge b_i^{(t)}\}} \\ &\qquad\cdot\Big(\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|^2\rho+O\Big(\frac{\tau}{\alpha_1\rho_1^2(K+1)\log d}\Big)\Big)\alpha_1\rho_1^2\;\Big|\;|\bar{z}_{k,j}^{(1)}|\ne 0\Bigg] \\ &\le \alpha\rho^3O\Big(\Xi_2\cdot\frac{\log\log d}{\Xi_2^5}\Big)\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[|\langle w_i^{(t)},M_j\rangle z_{n,s,j}-b_i^{(t)}|\Big] \\ &\le \alpha^2\rho^4O\Big(\frac{\log\log d}{\Xi_2^3}\Big)|\langle w_i^{(t)},M_j\rangle-b_{1i}^{(t)}|\Pr(|\bar{z}_{n,s,j}|\ne 0)\text{ where }b_i^{(t)}=\rho b_{1i}^{(t)}\end{aligned}$$

We have our second term $B_2=\max_{i\notin\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|\tilde{O}(\Xi_2)\frac{|\mathcal{N}|}{\tau}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[\Big(\int_0^1\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\partial\zeta|\psi_{i,j}^{(t)}(x_{n,s})||\bar{z}_{k,j}^{(1)}|^2\Big]$ and by applying for $i\notin\mathcal{M}_j$, $|\langle w_i^{(t)},M_j\rangle|\le O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d}})$ and $\|w_i^{(t)}\|_2\le O(1)$ we can simplify the expression written for B_2 as $\frac{1}{\tau\sqrt{d}}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[\Big(\int_0^1\sum_{x_{n,s}\in\mathcal{N}}\ell'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\partial\zeta|\psi_{i,j}^{(t)}(x_{n,s})||\bar{z}_{k,j}^{(1)}|^2\Big]\le\frac{1}{\tau\sqrt{d}}\tilde{O}(\frac{\alpha_1\rho_1^2}{d})|\langle w_i^{(t)},M_j\rangle\rho-b_i^{(t)}|$. From the definition $\ell'_{s,t}(\bar{x}_k^{(1)\setminus j},\mathcal{B})$ is independent from $\bar{z}_{k,j}^{(1)}$. Further, we observed that $\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\ge b_i^{(t)}}=\mathbb{1}_{\bar{z}_{k,j}^{(1)}\ne 0}$ with high probability. However,$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\ge b_i^{(t)}}\bar{z}_{k,j}^{(1)}\Big]$

is zero due to symmetry of the $\bar{z}_k^{(1)}$. For $|\mathfrak{B}_k^-| = ploy(d)$

$$\sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\ell'_{s,t}(\bar{x}_k^{(1)\setminus j},\mathcal{B})\cdot\psi_{i,j}^{(t)}(x_{n,s})\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right] = \sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\ell'_{s,t}(\bar{x}_k^{(1)\setminus j},\mathcal{B})\cdot\psi_{i,j}^{(t)}(x_{n,s})\right]\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\mathbb{1}_{\mathcal{G}_j}\bar{z}_{k,j}^{(1)}\right]$$
$$+\sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\ell'_{s,t}(\bar{x}_k^{(1)\setminus j},\mathcal{B})\cdot\psi_{i,j}^{(t)}(x_{n,s})\mathbb{1}_{\mathcal{G}_j^c}\bar{z}_{k,j}^{(1)}\right]$$

$$\sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[|\ell'_{s,t}(\bar{x}_k^{(1)\setminus j},\mathcal{B})\cdot\psi_{i,j}^{(t)}(x_{n,s})\mathbb{1}_{\mathcal{G}_j^c}\bar{z}_{k,j}^{(1)}|\right] \leq |\mathfrak{B}_k^-|.|\psi_{i,j}^{(t)}(x_{n,s})|.\rho.e^{-\Omega(\log^2 d)}$$
$$\leq ploy(d)|\psi_{i,j}^{(t)}(x_{n,s})|.\rho.e^{-\Omega(\log^2 d)}$$
$$\leq \frac{1}{ploy(d)^{\Omega(\log d)}}$$
$$\Psi_{2,i,j}^{(t)} \leq \alpha^2\rho^4 O\Big(\frac{\log\log d}{\Xi_2^3}\Big)|\langle w_i^{(t)},M_j\rangle - b_{1i}^{(t)}|\Pr(|\bar{z}_{n,s,j}|\neq 0)$$

We define an event $\mathcal{G}_j = \{j\in[d], \mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}} = \mathbb{1}_{\bar{z}_{k,j}^{(1)}\neq 0}\}$the neurones fires on positive sample if and only if the j^{th} cordinate is non zero w.h.p $1-e^{-\Omega(\log^2 d)}$.

Next we focus on the the first term of $\Phi_{i,j}^{(t)}$is $\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(1-\ell'_{p,t})\cdot\psi_{i,j}^{(t)}(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right]$ following the similar method used for the second term. Since we know from the softmax score, $(1-\ell'_{p,t})\geq 0$. If we can find a set of sparse features which fires the neuron i with high probability.

Define a set of sparse features $D_j = \{j'\neq j, \bar{z}_{k,j'}\neq 0\}$ which activate neurone i with high probability.When D_j holds and $\langle w_i^{(t)},M_j\rangle \geq 0$ we have $\psi_{i,j}^{(t)}(\bar{x}_k^{(2)}) = (\langle w_i^{(t)},M_j\rangle\bar{z}_{k,j}^{(2)} - b_i^{(t)})$ and then Sign$(\psi_{i,j}^{(t)}(\bar{x}_k^{(2)}))$ = Sign$(\langle w_i^{(t)},M_j\rangle\bar{z}_{k,j}^{(2)})$.

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(1-\ell'_{p,t})\cdot\psi_{i,j}^{(t)}(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\big|D_j\right]\geq 0$$

Similarly, when $\langle w_i^{(t)},M_j\rangle\leq 0$ only the direction of the above inequality reversed.

When D_j^c happens we have only few sparse features which activates the neuron i. From the results of the previous derivations it is easy to derive that

$$|\langle f(\bar{x}_k^{(1)}),f(\bar{x}_k^{(2)})\rangle| \leq \sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|^2|\bar{z}_{k,j}^{(1)}||\bar{z}_{k,j}^{(2)}| + O(\Xi_2)\max_{i\notin\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle| \leq \sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|^2|\bar{z}_{k,j}^{(1)}||\bar{z}_{k,j}^{(2)}| + O(\frac{\Xi_2}{\sqrt{d_1}})$$

Under the condition of D_j^c holds, we compute the bound of the difference of the inner products with j^{th} component and without j^{th} component.

$$|\langle f(\bar{x}_k^{(1)}),f(\bar{x}_k^{(2)})\rangle - \langle f(\bar{x}_k^{(1)\setminus j}),f(\bar{x}_k^{(2)\setminus j})\rangle| \leq \sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle|^2|\bar{z}_{k,j}^{(1)}||\bar{z}_{k,j}^{(2)}| + O(\frac{\Xi_2}{\sqrt{d_1}})$$

As in the previous section, we need to make an independent of the j^{th} coordinate from the $\ell'_{p,t}(\bar{x}_k^{(1)},\mathcal{B})$.Hence we define the set $\mathcal{B}'_p := \{\mathfrak{B}_k^-\cup\bar{x}_k^{(2)\setminus j}\}$ we can derive as same as the previous case $\frac{\ell'_{p,t}(\bar{x}_k^{(1)},\mathcal{B})}{\ell'_{p,t}(\bar{x}_k^{(1)},\mathcal{B}'_p)}\leq O(d/\Xi_2^5)$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\ell'_{p,t}(\bar{x}_k^{(1)},\mathcal{B})\cdot\psi_{i,j}^{(t)}(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\big|D_j^c\right] \leq \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[\frac{\ell'_{p,t}(\bar{x}_k^{(1)},\mathcal{B})}{\ell'_{p,t}(\bar{x}_k^{(1)},\mathcal{B}'_p)}\cdot\ell'_{p,t}(\bar{x}_k^{(1)},\mathcal{B}'_p)\cdot\psi_{i,j}^{(t)}(\bar{x}_k^{(2)})$$
$$\cdot\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\big|D_j^c\Big]$$
$$\leq \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[O(L)\cdot\ell'_{p,t}(\bar{x}_k^{(1)},\mathcal{B}'_p)\cdot\psi_{i,j}^{(t)}(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\big|D_j^c\right]$$
$$\leq O(\alpha\rho/\Xi_2^5)|\langle w_i^{(t)},M_j\rangle - b_{1i}^{(t)}|\Pr(|\bar{z}_{k,j}^{(2)}|\neq 0)$$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\psi_{i,j}^{(t)}(\bar{x}_k^{(2)})\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right] = \alpha\rho\cdot Sign(\langle w_i^{(t)},M_j\rangle)\cdot|\langle w_i^{(t)},M_j\rangle - b_{1i}^{(t)}|\cdot\frac{1}{\text{polylog }d}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}[\bar{z}_{k,j}^{(1)}]$$

By combining all above we have for the j^{th} component of the sparse vector of any sample in $\mathcal{B}$ as z_j

$$\begin{aligned}
\Psi_{1,i,j}^{(t)} &\leq O(\alpha\rho/\Xi_2^5)|\langle w_i^{(t)}, M_j\rangle - b_{1i}^{(t)}|\Pr(|\bar{z}_{k,j}^{(2)}| \neq 0) - \alpha\rho \cdot Sign(\langle w_i^{(t)}, M_j\rangle) \cdot |\langle w_i^{(t)}, M_j\rangle - b_{1i}^{(t)}| \\
&\quad \cdot \frac{1}{\text{polylog } d}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}[\bar{z}_{k,j}^{(1)}] \\
\Psi_{i,j}^{(t)} Sign(\langle w_i^{(t)}, M_j\rangle) &= \Psi_{1,i,j}^{(t)} Sign(\langle w_i^{(t)}, M_j\rangle) + \Psi_{2,i,j}^{(t)} Sign(\langle w_i^{(t)}, M_j\rangle) \\
&\geq \alpha\rho\frac{\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}[\bar{z}_j]}{\text{polylog } d} \cdot |\langle w_i^{(t)}, M_j\rangle - b_{1i}^{(t)}| - \alpha\rho^3 O\Big(\Xi_2 \cdot \frac{\log\log d}{\Xi_2^5}\Big)|\langle w_i^{(t)}, M_j\rangle - b_{1i}^{(t)}|\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}[\bar{z}_j] \\
&\geq \Big(-\alpha\rho^3 O\Big(\Xi_2 \cdot \frac{\log\log d}{\Xi_2^5}\Big) + \alpha\rho\frac{1}{\text{polylog } d}\Big) \cdot |\langle w_i^{(t)}, M_j\rangle - b_{1i}^{(t)}|\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}[\bar{z}_j] \\
\Psi_{i,j}^{(t)} Sign(\langle w_i^{(t)}, M_j\rangle) &\geq \frac{\alpha\rho}{\text{polylog } d}\Big(1 - O\Big(\rho^2 \cdot \frac{\log\log d}{\Xi_2^3}\Big)\Big) \cdot |\langle w_i^{(t)}, M_j\rangle - b_{1i}^{(t)}|\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}[\bar{z}_j]
\end{aligned}$$

Lemma H.5. *For $j \in [d]$ and $i \in \mathcal{M}_j^{1*}$ suppose the Induction Hypothesis hold at iteration t, and there exist a constant H_2 such that$\sum_{i\in\mathcal{M}_j} |\langle w_i^{(t)}, M_j\rangle|^2 \geq H_2 \frac{\tau}{\alpha_1\rho_1^2}\log d$,we can derive the upper bound of $\Psi_{i,j}^{(t)}$*

$$\Psi_{i,j}^{(t)} \leq O\Big(\frac{1}{\text{poly}d}\Big)|\langle w_i^{(t)}, M_j\rangle|$$

Proof of Lemma H.5
From the expansion of the gradient we have

$$\begin{aligned}
\Psi_{i,j}^{(t)} &:= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t}) \cdot \psi_{i,j}^{(t)}(\bar{z}_{k,j}) + \sum_{x_{n,s}\in\mathcal{B}_k^-} \ell'_{s,t} \cdot \psi_{i,j}^{(t)}(x_{n,s})\right)\mathbb{1}_{\bar{z}_{k,j}\neq 0}\bar{z}_{k,j}\right] \\
I_1 &= \mathbb{E}_{\mathfrak{B}_k^-}\left[\left(\ell'_{p,t} \cdot \psi_{i,j}^{(t)}(\bar{z}_{k,j}) + \sum_{x_{n,s}\in\mathcal{B}_k^-} \ell'_{s,t} \cdot \psi_{i,j}^{(t)}(x_{n,s})\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}}\right)\bar{z}_{k,j}\right] \\
&= \psi_{i,j}^{(t)}(\bar{z}_{k,j})\mathbb{E}_{\mathfrak{B}_k^-}\left[\frac{e^{\langle f(\bar{x}_k^{(1)}, f(\bar{x}_k^{(2)})\rangle/\tau} + \sum_{x_{n,s}\in\mathcal{B}_k^-} e^{\langle f(\bar{x}_k^{(1)}, f(x_{n,s})\rangle/\tau}\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}}}{e^{\langle f(\bar{x}_k^{(1)}, f(x_{n,s})\rangle/\tau} + \sum\limits_{x\in B\setminus\{x_{n,s}\}} e^{\langle f(\bar{x}_k^{(1)}, f(\bar{x})\rangle/\tau}}\right]\bar{z}_{k,j} \\
&= \psi_{i,j}^{(t)}(\bar{z}_{k,j})\mathbb{E}_{\mathfrak{B}_k^-}\left[\frac{e^{\langle f(\bar{x}_k^{(1)}, f(\bar{x}_k^{(2)})\rangle/\tau} + |\mathfrak{B}_k^-|e^{\langle f(\bar{x}_k^{(1)}, f(x_{n,s})\rangle/\tau}\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}}}{e^{\langle f(\bar{x}_k^{(1)}, f(x_{n,s})\rangle/\tau} + \sum\limits_{x\in B\setminus\{x_{n,s}\}} e^{\langle f(\bar{x}_k^{(1)}, f(\bar{x})\rangle/\tau}}\right]\bar{z}_{k,j} \\
I_2 &= \mathbb{E}_{\mathfrak{B}_k^-}\left[\left(\sum_{x_{n,s}\in\mathcal{B}_k^-} \ell'_{s,t} \cdot \psi_{i,j}^{(t)}(x_{n,s})\mathbb{1}_{\bar{z}_{k,j}\neq\bar{z}_{n,s,j}}\right)\bar{z}_{k,j}\right]
\end{aligned}$$

Combining above we can write the original expression as

$$\Psi_{i,j}^{(t)} := \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Big[\Big(I_1 + I_2 - \psi_{i,j}^{(t)}(\bar{z}_{k,j})\Big)\mathbb{1}_{\bar{z}_{k,j}\neq 0}\bar{z}_{k,j}\Big]$$

In the second step of I_1 used the condition $\bar{z}_{k,j} = \bar{z}_{n,s,j}$ and for I_2 we used the condition of $\bar{z}_{k,j} \neq \bar{z}_{n,s,j}$. We let $V = e^{\langle f(\bar{x}_k^{(1)}, f(\bar{x}_k^{(2)})\rangle/\tau}, X = e^{\langle f(\bar{x}_k^{(1)}, f(x_{n,s})\rangle/\tau}, Y = \sum_{x\in\mathfrak{B}_k^-\setminus\{x_{n,s}\}} e^{\langle f(\bar{x}_k^{(1)}, f(\bar{x})\rangle/\tau}$, and $Y_n = e^{\langle f(\bar{x}_k^{(1)}, f(\bar{x}_n)\rangle/\tau}$. Here we get the expectaion over negative samples $\mathfrak{B}_k^- = x_{n,s}, \{x_{n,u}\}_{u\neq s}$. From the low of total expectation we have $\mathbb{E}_{\mathfrak{B}_k^-}[f(x_{n,s}, \{x_{n,u}\}_{u\neq s})] = \mathbb{E}_{x_{n,s}}[\mathbb{E}_{x_{n,u}}[f(x_{n,s}, \{x_{n,u}\}_{u\neq s})|x_{n,s}]]$ and most of these neagative samples are i.i.d

$$
\begin{aligned}
\frac{I_1}{\psi_{i,j}^{(t)}(\bar{z}_{k,j})} &= \mathbb{E}_{\mathfrak{B}_k^-}\left[\frac{V+|\mathfrak{B}_k^-|e^{\langle f(\bar{x}_k^{(1)}, f(x_{n,s})\rangle/\tau}\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}}}{e^{\langle f(\bar{x}_k^{(1)}, f(x_{n,s})\rangle/\tau}+V+\sum_{x\in\mathfrak{B}_k^-\setminus\{x_{n,s}\}}e^{\langle f(\bar{x}_k^{(1)}, f(\bar{x})\rangle/\tau}}\right]\bar{z}_{k,j} \\
&= \mathbb{E}_{x_{n,s}}\left[\frac{\frac{1}{|\mathfrak{B}_k^-|}V+X\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}}}{\frac{1}{|\mathfrak{B}_k^-|}X+\frac{1}{|\mathfrak{B}_k^-|}V+\frac{1}{|\mathfrak{B}_k^-|}\mathbb{E}_{x_{n,u}}[Y|x_{n,s}]}\right]\bar{z}_{k,j} \\
&= \mathbb{E}_{x_{n,s}}\left[\frac{\frac{1}{|\mathfrak{B}_k^-|}V+X\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}}}{\frac{1}{|\mathfrak{B}_k^-|}X+\frac{1}{|\mathfrak{B}_k^-|}V+\frac{\mathfrak{B}_k^--1}{|\mathfrak{B}_k^-|}\mathbb{E}_{x_n}[Y_n]}\right]\bar{z}_{k,j} \\
&X_n = Y_n\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}} \\
&= \mathbb{E}_{x_{n,s}}\left[\frac{\frac{1}{|\mathfrak{B}_k^-|}V+X_n}{\frac{1}{|\mathfrak{B}_k^-|}X_n+\frac{1}{|\mathfrak{B}_k^-|}V+\frac{\mathfrak{B}_k^--1}{|\mathfrak{B}_k^-|}\Big(1+\frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}\Big)\mathbb{E}_{x_n}[X_n]}\right]\bar{z}_{k,j} \\
&= 1-\frac{1}{|\mathfrak{B}_k^-|}\mathbb{E}_{x_{n,s}}\left[\frac{\mathbb{E}[X_n]}{\frac{1}{|\mathfrak{B}_k^-|}X_n+\frac{1}{|\mathfrak{B}_k^-|}V+\frac{\mathfrak{B}_k^--1}{|\mathfrak{B}_k^-|}\Big(1+\frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}\Big)\mathbb{E}[X_n]}\right]\bar{z}_{k,j} \\
&\geq 1-\frac{1}{|\mathfrak{B}_k^-|}\left[\frac{\mathbb{E}[X_n]}{\frac{\mathfrak{B}_k^--1}{|\mathfrak{B}_k^-|}\Big(1+\frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}\Big)\mathbb{E}[X_n]}\right]\bar{z}_{k,j}\ \text{From}X_n+V>0 \\
&\geq 1-\frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}
\end{aligned}
$$

As same as the previous parts, we can lower bound the gap between the precence and absense of the j^{th} component as,

$$
|\langle f(\bar{x}_k^{(1)}), f(\bar{x}_n)\rangle - \langle f(\bar{x}_k^{(1)}), f(\bar{x}_n^{\setminus j})\rangle|/\tau \geq \frac{1}{\tau}\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)}, M_j\rangle|^2|\bar{z}_{k,j}^{(1)}||\bar{z}_{n,j}| - A_1 \geq \frac{H_2\log d}{\alpha_1\rho_1^2(K+1)} - O\Big(\frac{1}{\alpha_1\rho_1^2(K+1)\log d}\Big)
$$

Here $H_2 = O(1)$ is a large constant which is large enough to be $\frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}$ to imply below

$$
\begin{aligned}
e^{\langle f(\bar{x}_k^{(1)}), f(\bar{x}_n^{\setminus j})\rangle/\tau}\mathbb{1}_{\bar{z}_{k,j}\neq\bar{z}_{n,s,j}} &\leq d^{-H_2}e^{\langle f(\bar{x}_k^{(1)}), f(\bar{x}_n)\rangle/\tau}\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}} \\
\mathbb{E}_{x_n}[e^{\langle f(\bar{x}_k^{(1)}), f(\bar{x}_n^{\setminus j})\rangle/\tau}\mathbb{1}_{\bar{z}_{k,j}\neq\bar{z}_{n,s,j}}] &\leq \frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}\mathbb{E}_{x_n}[e^{\langle f(\bar{x}_k^{(1)}), f(\bar{x}_n)\rangle/\tau}\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}}] \\
\mathbb{E}_{x_n}[Y_n(\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}}+\mathbb{1}_{\bar{z}_{k,j}\neq\bar{z}_{n,s,j}})] &\leq \Big(1+\frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}\Big)\mathbb{E}_{x_n}[Y_n\mathbb{1}_{\bar{z}_{k,j}=\bar{z}_{n,s,j}}]
\end{aligned}
$$

Next we need to bound the I_2 for the condition of $\bar{z}_{k,j} \neq \bar{z}_{n,s,j}$. We know our $\bar{z}_{k,j}, \bar{z}_{n,s,j} \in \{\pm\frac{\rho K+1}{K+1}, \pm\frac{1}{K+1}, 0\}$.Then for I_2,

$$
\begin{aligned}
F &= \{\pm\frac{\rho K+1}{K+1}, \pm\frac{1}{K+1}, 0\} \\
I_2 &= \mathbb{E}_{\mathfrak{B}_k^-}\left[\left(\sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot\psi_{i,j}^{(t)}(x_{n,s})\mathbb{1}_{\bar{z}_{k,j}\neq\bar{z}_{n,s,j}}\right)\bar{z}_{k,j}\right]
\end{aligned}
$$

We can expand the indicator

$$\begin{aligned}
\mathbb{1}_{\bar{z}_{k,j}\neq\bar{z}_{n,s,j}} &= \sum_{f\in F\setminus\bar{z}_{k,j}} \mathbb{1}_{\{\bar{z}_{n,s,j}=f\}} \\
I_2 &= \sum_{f\in F\setminus\bar{z}_{k,j}} \psi_{i,j}^{(t)}(f)\mathbb{E}_{\mathfrak{B}_k^-}\left[\left(\sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot\mathbb{1}_{\{\bar{z}_{n,s,j}=f\}}\right)\bar{z}_{k,j}\right] \\
\sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot\mathbb{1}_{\{\bar{z}_{n,s,j}=f\}} &= \mathbb{E}_{\mathfrak{B}_k^-}\left[\frac{\sum_{x_{n,s}\in\mathcal{B}_k^-} e^{\langle f(\bar{x}_k^{(1)}),f(\bar{x}_{n,s})\rangle/\tau}\mathbb{1}_{\bar{z}_{k,j}\neq\bar{z}_{n,s,j}}}{e^{\langle f(\bar{x}_k^{(1)}),f(\bar{x}_k^{(2)})\rangle/\tau}+\sum_{x\in\mathfrak{B}_k^-} e^{\langle f(\bar{x}_k^{(1)}),f(\bar{x})\rangle/\tau}}\right] \\
&= \mathbb{E}_{x_{n,s}}\left[\frac{X_n}{\frac{1}{|\mathfrak{B}_k^-|}X_n+\frac{1}{|\mathfrak{B}_k^-|}V+\frac{\mathfrak{B}_k^- -1}{|\mathfrak{B}_k^-|}\Big(1+\frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}\Big)\mathbb{E}_{x_n}[X_n]}\right]\bar{z}_{k,j} \\
&\leq \frac{1}{\alpha_1\rho_1^2\mathrm{poly}d} \\
I_2 &\leq \frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}\sum_{f\neq\bar{z}_{k,j}}\psi_{i,j}^{(t)}(f) \\
|I_2| &\leq \frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}\max_{f\in F}|\psi_{i,j}^{(t)}(f)| \\
\Psi_{i,j}^{(t)} &:= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\Big(I_1+I_2-\psi_{i,j}^{(t)}(\bar{z}_{k,j})\Big)\mathbb{1}_{\bar{z}_{k,j}\neq 0}\bar{z}_{k,j}\right] \\
&\geq -\frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\psi_{i,j}^{(t)}(\bar{z}_{k,j})\mathbb{1}_{\bar{z}_{k,j}\neq 0}\bar{z}_{k,j}\right]-\frac{1}{\alpha_1\rho_1^2\mathrm{poly}d}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\max_{f\in F}|\psi_{i,j}^{(t)}(f)|\mathbb{1}_{\bar{z}_{k,j}\neq 0}\bar{z}_{k,j}\right] \\
-\Psi_{i,j}^{(t)} &\leq O\Big(\frac{1}{\mathrm{poly}d}\Big)|\langle w_i^{(t)},M_j\rangle|
\end{aligned}$$

For all $i\in\mathcal{M}_j$ and there is a constant $C=O(1)$ we have $|\langle w_i^{(t)},M_j\rangle|\leq C\max_{i\in\mathcal{M}_j^{1*}}|\langle w_i^{(t)},M_j\rangle|$ and $\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle^2| \leq C\sum_{i\in\mathcal{M}_j}\max_{i\in\mathcal{M}_j^{1*}}|\langle w_i^{(t)},M_j\rangle|^2 \leq |\mathcal{M}_j|^2\max_{i\in\mathcal{M}_j^{1*}}|\langle w_i^{(t)},M_j\rangle|^2 \leq O(d^{4\gamma c0}\sigma_0^2\log d)d^{O(C_z)}$.Then we have $\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)},M_j\rangle^2| \leq O(\mathrm{polylog}\quad d\log d) \leq O(\frac{\tau}{\alpha_1\rho_1^2}\log d)$.Where $\tau=(\alpha_1\rho_1^2+\frac{1-\alpha\rho}{K+1})\mathrm{polylog}\,d$

Lemma H.6. *Let $j\in[d]$ and $i\in[m]$ when the induction hypothesis holds we have*

$$\xi_{i,j}^{(t)}\leq O(\frac{\Xi_2^2\|w_i^{(t)}\|_2}{\tau d^2(K+1)})\max_{i'\in[m]}\Big(|\langle w_{i'}^{(t)},M_j\rangle|+|\frac{\|w_{i'}^{(t)}\|_2}{\sqrt{d_1}}\Big))$$

When $j\in[d_1]\backslash[d]$ we need to change the $|\langle w_{i'}^{(t)},M_j\rangle|\frac{1}{K+1}$ with $|\langle w_{i'}^{(t)},M_j^{\perp}\rangle|$

Proof of Lemma H.6
We are going to bound $\xi_{1,i,j}^{(t)}$

$$\begin{aligned}
\xi_{1,i,j}^{(t)} &:= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot h_{i,t}(\bar{x}_k^{(2)})+\sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot h_{i,t}(x_{n,s})\right)\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\langle M_j,\bar{\xi}_k^{(1)}\rangle\right] \\
&= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot h_{i,t}(x_{n,s})\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\langle M_j,\bar{\xi}_k^{(1)}\rangle\right]-\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[(1-\ell'_{p,t})\cdot h_{i,t}(\bar{x}_k^{(2)})\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\langle M_j,\bar{\xi}_k^{(1)}\rangle\right] \\
&= \xi_{1,1,i,j}^{(t)}-\xi_{1,2,i,j}^{(t)} \\
|\xi_{1,i,j}^{(t)}| &\leq |\xi_{1,1,i,j}^{(t)}|+|\xi_{1,2,i,j}^{(t)}|
\end{aligned}$$

From the Lemma G.4 we can bound $\xi^{(t)}_{1,2,i,j} \leq O\Big(\frac{\|w_i^{(t)}\|_2}{d^2(K+1)\sqrt{d_1}}\Big) \cdot \max\{\Pr(|\langle w_i, \bar{x}_k^{(2)}\rangle| \geq b_i), \tilde{O}\Big(\frac{1}{\sqrt{d}}\Big)\}$. We define the noise free augmented view $\bar{x}_k^{(1)+} := M\bar{z}_k^{(1)}$ and then we can write $\bar{x}_k^{(1)}(\zeta) = \bar{x}_k^{(1)++}\zeta\bar{\xi}_k^{(1)}$. The logit $\ell'_{s,t}(\bar{x}_k^{(1)}, \mathcal{B})$ related to $\ell'_{s,t}(\bar{x}_k^{(1)+}, \mathcal{B})$ via a path integral along the direction of the spurious noise

$$\ell'_{s,t}(\bar{x}_k^{(1)}, \mathcal{B}) - \ell'_{s,t}(\bar{x}_k^{(1)+}, \mathcal{B}) = \int_0^1 \frac{d}{d\zeta}\ell'_{s,t}(\bar{x}_{1k}^{+} + \zeta.\bar{\xi}_k^{(1)}, \mathcal{B})\partial\zeta$$

$$\begin{aligned}
\xi^{(t)}_{1,1,i,j} &= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\sum_{x_{n,s}\in\mathcal{B}_k^-} \ell'_{s,t}\cdot h_{i,t}(x_{n,s})\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\langle M_j, \bar{\xi}_k^{(1)}\rangle\right] \\
&= \sum_{x_{n,s}\in\mathcal{B}_k^-} \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\ell'_{s,t}\cdot h_{i,t}(x_{n,s})\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\langle M_j, \bar{\xi}_k^{(1)}\rangle\right] \\
&= \frac{1}{\tau}\sum_{x_{n,s}\in\mathcal{B}_k^-} \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\int_0^1 \Big(\hat{\ell}'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\big(1-\hat{\ell}'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\big) - \sum_{x_{n,u}\in\mathfrak{B}_k^-\setminus\{x_{n,s}\}} \hat{\ell}'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\,\hat{\ell}'_{u,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\Big)\partial\zeta \\
&\quad \times \big\langle f(\bar{x}_k^{(1)}(\zeta)) - f(\bar{x}_k^{(1)+}), f(x_{n,s})\big\rangle \cdot h_{i,t}(x_{n,s})\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\langle M_j, \bar{\xi}_k^{(1)}\rangle\Bigg] \\
&\leq \frac{1}{\tau}\sum_{x_{n,s}\in\mathcal{B}_k^-} \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\int_0^1 \Big(\hat{\ell}'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\big(1-\hat{\ell}'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\big) - \sum_{x_{n,u}\in\mathfrak{B}_k^-\setminus\{x_{n,s}\}} \hat{\ell}'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\,\hat{\ell}'_{u,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\Big)\partial\zeta \\
&\quad \times \max_{x_{n,u}\in\mathfrak{B}_k^-} |\big\langle f(\bar{x}_k^{(1)}(\zeta)) - f(\bar{x}_k^{(1)+}), f(x_{n,s})\big\rangle| \cdot |h_{i,t}(x_{n,s})|\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\langle M_j, \bar{\xi}_k^{(1)}\rangle\Bigg] \\
&\leq O(\Xi_2)\frac{1}{\tau}\sum_{x_{n,s}\in\mathcal{B}_k^-} \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\int_0^1 \Big(\hat{\ell}'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\,\partial\zeta \times \max_{i\in[m]} |\big\langle f(\bar{x}_k^{(1)}(\zeta)) - f(\bar{x}_k^{(1)+})\big\rangle| \\
&\quad \cdot |h_{i,t}(x_{n,s})|\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\langle M_j, \bar{\xi}_k^{(1)}\rangle\Bigg] \\
&\leq O(\Xi_2)\frac{1}{\tau}\sum_{x_{n,s}\in\mathcal{B}_k^-} \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\Bigg[\int_0^1 \Big(\hat{\ell}'_{s,t}(\bar{x}_k^{(1)}(\zeta),\mathcal{B})\,\partial\zeta \times \max_{i\in[m]} |\langle w_i^{(t)}, \bar{\xi}_k^{(1)}\rangle \cdot \langle M_j, \bar{\xi}_k^{(1)}\rangle| \cdot |h_{i,t}(x_{n,s})|\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\Bigg] \\
&\leq O(\frac{\Xi_2^2\|w_i^{(t)}\|_2}{\tau})\Big(\max_{i'\in[m]} |\langle w_{i'}^{(t)}, M_j\rangle| \cdot \frac{1}{d(K+1)}\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\Big) \\
&\leq O(\frac{\Xi_2^2\|w_i^{(t)}\|_2}{\tau d^2(K+1)}) \max_{i'\in[m]} |\langle w_{i'}^{(t)}, M_j\rangle|
\end{aligned}$$

Combining both of the terms

$$\xi^{(t)}_{i,j} \leq O(\frac{\Xi_2^2\|w_i^{(t)}\|_2}{\tau d^2(K+1)}) \max_{i'\in[m]}\Big(|\langle w_{i'}^{(t)}, M_j\rangle| + |\frac{\|w_{i'}^{(t)}\|_2}{\sqrt{d_1}}\Big)$$

Here we use the fact of $\sum_{i\in[m]} \mathbb{1}_{h_{i,t}(x)\neq 0} \leq \tilde{O}(\Xi_2)$ and Lemma A2 we have $|h_{i,t}(x_{n,s})| \leq O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}})$. For the $|\bar{z}_{k,j}^{(1)} - \bar{z}_{k,j}^{(2)}|$, when we get the difference of the two views they are mainly different from the rare features aggregate in $M^{(2)}$ with $|\bar{z}_{k,j}^{(1)} - \bar{z}_{k,j}^{(2)}| = \frac{1}{K+1}$ and the firing probability $\Pr(\langle w_i^{(t)}, \bar{x}_k^{(1)}\rangle| \geq b_i^{(t)}) = O(\frac{1}{d})$.

Lemma H.7. *(Derivation of upper bound of $\Phi^{(t)}$ from $\Psi^{(t)}$). Let $j \in [d]$ and $i \in \mathcal{M}_j$. Suppose Induction Hypothesis hold for all iteration prior to $t \in [\frac{d^{1.01}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}, \frac{d^{1.99}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}]$ and after T_2, and also we suppose that for some $t' = \Theta(T_2)$ we have $\tilde{\mathfrak{F}}_l^{(t')} = \Omega(\tau\log d)$, and alll $\in [d]$ then*

- *for iteration $t \in [\frac{d^{1.01}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}, \frac{d^{1.49}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}]$:*

$$\Phi_{i,j}^{(t)} \leq \widetilde{O}\left(\frac{\Xi_2}{d^{3/2}\sqrt{\alpha_1\rho_1^2(K+1)}}\right)\|w_i^{(t)}\|_2$$

- *for iteration* $t \in [\frac{d^{1.49}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}, \frac{d^{1.99}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}]$:

$$\Phi_{i,j}^{(t)} \leq \widetilde{O}\left(\frac{\Xi_2}{d^2\sqrt{\alpha_1\rho_1^2(K+1)}}\right)\|w_i^{(t)}\|_2$$

Proof of Lemma H.7
Define the parameter $\rho_{i,j}^{(t)} = (\langle w_i^{(t)}, \bar{x}_k\rangle - \langle w_i^{(t)}, Mz\rangle)\mathbb{1}_{z_j\neq 0}$. Let G denote the intersection of the high-probability structural events ensuring sparsity of $\bar{z}$,w.h.p.$|\{j' \in [d] : \bar{z}_{k,j'} \neq 0\}|, |\{j' \in [d] : \bar{z}_{n,s,j'} \neq 0\}| = \tilde{O}(1)$ and the probability of the firing the neurones gives

$$\Pr(G^c) \leq e^{-\Omega(\log^2 d)}$$

$$\begin{aligned}
\Phi_{i,j}^{(t)} &:= \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot\phi_{i,j}^{(t)}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot\phi_{i,j}^{(t)}(x_{n,s})\right)\mathbb{1}_{\left|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle\right|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right] \\
&= \sum_{j'\in[d],j'\neq j}\langle w_i^{(t)}, M_{j'}\rangle\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot\bar{z}_{2k,j'}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot z_{n,s,j}\right)\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right] \\
&\quad + \mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot\rho_{i,j}^{(t)}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot\rho_{i,j}^{(t)}(x_{n,s})\right)\mathbb{1}_{\|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right] + \frac{1}{d^{\Omega(\log d)}} \\
&= H_1 + H_2 + \frac{1}{d^{\Omega(\log d)}}
\end{aligned}$$

From the proofs of the previous Lemmas, We can bound the H_2 as $|H_2| \leq O(\frac{\Xi_2^2\|w_i^{(t)}\|_2}{\tau d^2\sqrt{(K+1)}})$.From the induction hypothesis we have $\max_{j'\neq j}|\langle w_i^{(t)}, M_{j'}\rangle| \leq O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)\Xi_2^5}})$and the saprsity of $\bar{z}_k$, we can bound the H_1

$$\begin{aligned}
H_1 &= O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)}\Xi_2^5})\sum_{j'\in[d],j'\neq j}\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot\bar{z}_{2k,j'}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot z_{n,s,j}\right)\mathbb{1}_{|\langle w_i^{(t)},\bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right] \\
&\leq O(\frac{\|w_i^{(t)}\|_2}{d^{3/2}\sqrt{(K+1)}\Xi_2^5})
\end{aligned}$$

Then we can obtain the bound

$$\Phi_{i,j}^{(t)} \leq H_1 + H_2 + \frac{1}{d^{\Omega(\log d)}} \leq O\Big(\frac{\|w_i^{(t)}\|_2}{d^{3/2}\sqrt{(K+1)}}\Big)$$

We establish a connection between $\Phi_{i,j}^{(t)}$ and $\Psi_{i,j}^{(t)}$. Initially we assumme that for all $j' \neq j, j' \in [d]$ from using the induction $\Psi_{i,j}^{(t)}$ in Lemma E2 as $\frac{|\Psi_{i,j}^{(t)}|}{|\langle w_i^{(t)}, M_{j'}\rangle|} \leq O\Big(\frac{1}{\text{poly}d}\Big) \leq \Omega\Big(\frac{\Xi_2}{\eta'(K+1)\sqrt{\alpha_1\rho_1^2 dt}}\Big)$holds for all iteration $t \leq \frac{d\text{poly}\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}$Now suppose that at some iteration $t \geq \frac{d\text{poly}\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}$ there exist a feature $j' \neq j, j' \in [d]$, and lucky neurones $i \in \mathcal{M}_j^{1*}$ which can exceed the bound above.

$$\frac{|\Psi_{i,j}^{(t)}|}{|\langle w_i^{(t)}, M_{j'}\rangle|} \geq \Omega\Big(\frac{\Xi_2}{\eta'(K+1)\sqrt{\alpha_1\rho_1^2 dt}}\Big)$$

From the definition of $\Psi_{i,j}^{(t)}$ above we have the bound below

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot\bar{z}_{2k,j'} + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot z_{n,s,j}\right)\bar{z}_{k,j}^{(1)}\right] \geq \Omega\Big(\frac{\Xi_2}{\eta'(K+1)\sqrt{\alpha_1\rho_1^2 dt}}\Big)$$

Now we define a safe limit for the growth of $\Psi_{i,j}^{(t)}$ when learning features $j' \neq j$. We use the energy of the alignment of the feature j with the lucky neurons as the bound of this growth.

Assume when this energy $\mathfrak{F}_j^{(t)} = \Delta$ the ratio is bounded by $\frac{|\Psi_{i,j}^{(t)}|}{|\langle w_i^{(t)}, M_{j'}\rangle|} \leq O\Big(\frac{\sqrt{\Xi_2}\tau\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)dt}}\Big)$ When t grows,weights are training with the features, hence this bound become time dependent variable. Moreover, the role of this Δ is controlling the value of the $\Psi_{i,j}^{(t)}$.

From the Lemma E1 and E.2 above, there exist a constant $\delta > \Omega(1)$ such that,if $\mathfrak{F}_j^{(t)} \geq \Delta - \delta\frac{\tau}{\alpha_1\rho_1^2}\log d$ it is lower than the best and at time t_1, the growth of feature j is already close to the threshold Δ. Therefore $\Psi_{i,j}^{(t)}$ starts to growth. However this growth cannot continue indefinitely.Consider the time interval $t' = O\Big(\frac{t}{\sqrt{\Xi_2\alpha_1\rho_1^2(K+1)}}\Big)$For each $i' \in \mathcal{M}_{j'}$, the weight update satisfied

$$|\langle w_i^{(t'+t1)}, M_j\rangle| \geq |\langle w_i^{(t)}, M_j\rangle|(1-\eta\lambda)^{t'} - \sum_t^{t+t'-1}\langle\nabla_{w_i}L(f_{\mathcal{G}}^t), M_j\rangle \geq |\langle w_i^{(t)}, M_j\rangle|(1-\eta\lambda)^{t'} + \sum_t^{t+t'-1}\Psi_{i,j}^{(t)} - t'\Phi_{i,j}^{(t)}$$

$$\geq |\langle w_i^{(t)}, M_j\rangle|(1-\eta\lambda)^{t'} + \sum_t^{t+t'-1}\Psi_{i,j}^{(t)} - t'O\Big(\frac{\|w_i^{(t)}\|_2}{d^{3/2}}\Big)$$

Now we build the contradiction argument here that, if the $\frac{\Psi_{i,j}^{(t)}}{|\langle w_i^{(t)}, M_{j'}\rangle|}$remained "large" throughout the entire window$[t1, t'+t1]$, then the feature growth would satisfy $\mathfrak{F}_j^{(t+t')} \geq \mathfrak{F}_j^{(t)} + t'O\Big(\frac{\sqrt{\Xi_2}\tau\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)dt}}\Big) \geq \Delta + \delta\frac{\tau}{\alpha_1\rho_1^2}\log d$ will large enough $\sum_t^{t+t'-1}\Psi_{i,j}^{(t)}$ to force $\mathfrak{F}_j^{(t)} \geq \Delta + t'O(\frac{\sqrt{\Xi_2}\tau\log d}{\sqrt{\alpha_1\rho_1^2(K+1)d\eta t}})$.This will violates the definition of the threshold of the $\mathfrak{F}_j^{(t)}$ is Δ. After exceed the $\mathfrak{F}_j^{(t)}$ than Δ ,$\Psi_{i,j}^{(t)}$ start to shrink again. Hence, there must exist some $t* \in [t1, t'+t1]$ such that$|\Psi_{i,j}^{(t)}| \leq \Omega\Big(\frac{\Xi_2}{C_\eta\eta'(K+1)\sqrt{\alpha_1\rho_1^2dt}}\Big)\|w_i^{(t)}\|_2$.

From above we can argue that $|\Psi_{i,j}^{(t)}| \leq \Omega\Big(\frac{\Xi_2}{\eta'(K+1)\sqrt{\alpha_1\rho_1^2dt}}\Big)\|w_i^{(t)}\|_2$ at all $t \in [\frac{d\text{poly}\log d}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}, \frac{d^{1.5}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}]$. This gives the

$\Omega(\frac{\Xi_2}{d^{1.5}\sqrt{(K+1)}})\|w_i^{(t*)}\|_2 \geq |\Psi_{i,j}^{(t)}| \geq \Omega(\frac{\Xi_2}{d^2\sqrt{(K+1)}})\|w_i^{(t*)}\|_2$

For the remaining iterations $t \geq \frac{d^{1.49}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}$,when $j' \in \mathcal{S}_i$ we have the highest alignment with the M_j^1 and the neurones almost activated when $\bar{z}_{k,j}^{(1)} \neq 0$ with the probability $\leq \tilde{O}(\frac{1}{d})$ and the first term of H_1 bounded as $O(\frac{1}{d^2})\|w_i^{(t*)}\|_2$

$$H_1 = \sum_{j'\in\mathcal{S}_i, j'\neq j}\langle w_i^{(t)}, M_{j'}\rangle\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot\bar{z}_{2k,j'}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot z_{n,s,j}\right)\mathbb{1}_{|\langle w_i^{(t)}, \bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right]$$

$$+ \sum_{j'\notin\mathcal{S}_i, j'\neq j}\langle w_i^{(t)}, M_{j'}\rangle\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot\bar{z}_{2k,j'}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot z_{n,s,j}\right)\mathbb{1}_{|\langle w_i^{(t)}, \bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right]$$

$$\mathbb{E}_{\bar{x}_k^{(1)},\mathcal{B}}\left[\left(-(1-\ell'_{p,t})\cdot\bar{z}_{2k,j'}(\bar{x}_k^{(2)}) + \sum_{x_{n,s}\in\mathcal{B}_k^-}\ell'_{s,t}\cdot z_{n,s,j}\right)\mathbb{1}_{|\langle w_i^{(t)}, \bar{x}_k^{(1)}\rangle|\geq b_i^{(t)}}\bar{z}_{k,j}^{(1)}\right] \leq \tilde{O}\Big(\max\Big\{\frac{\sqrt{\Xi_2}}{\eta'\sqrt{dt}}, \frac{\Xi_2}{d^2}\Big\}\Big)\|w_i^{(t*)}\|_2\sqrt{\frac{\alpha_1\rho_1^2}{K+1}}$$

Proof of Induction Hypothesis H.2
Next we are going to prove the induction hypothesis holds for $t = T_2$.From the lemma D3 and D.4 we have w.h.p for $i \notin \mathcal{M}_j$

$$\Pr(E_k^1) \leq \frac{1}{d^{\Omega(\log d)}}, \Pr(E_k^2) \leq \frac{1}{d^{\Omega(\log d)}} \text{ this implies that } L_1, L_2 \leq \frac{1}{d^{\Omega(\log d)}}$$

Furthermore we have

$$\mathbb{E}\left[\bar{z}_{k,j}^{(2)}\bar{z}_{k,j}^{(1)}\cdot\mathbb{1}_{|\langle w_i,\bar{x}_k\rangle|\geq b_i+|\langle w_i,\bar{x}_k-\bar{x}_k^{(1)}\rangle|}\right] \leq \frac{1-\alpha\rho}{K+1}\cdot\frac{O(\log\log d)}{d}\frac{1}{d^{\Omega(\log d)}} \leq \frac{1-\alpha\rho}{K+1}\cdot\frac{1}{d^{\Omega(\log d)}}$$

From Lemma H.4 , Lemma H.5 and Lemma H.6 we have the calculations of the bounds of $\Psi_{i,j}^{(t)}$, $\Phi_{i,j}^{(t)}$, and $\xi_{i,j}^{(t)}$ Considering that we can obtain the following

$$\begin{aligned}
\langle w_i^{(t+1)}, M_j\rangle &= \langle w_i^{(t)}, M_j\rangle - \eta\langle \nabla_{w_i}\text{Obj}(f_{\mathcal{G}}^t), M_j\rangle \pm \eta\frac{\|w_i^{(t)}\|_2}{\text{poly}(d_1)} \\
&= (1-\eta\lambda)\langle w_i^{(t)}, M_j\rangle + \eta\mathbb{E}_{x_p^+, x_p^{++}}\left[(1-\ell'_{p,t}(x_p^+, B))\cdot h_{i,t}(x_p^{++})\cdot\langle\nabla_{w_i}h_{i,t}(x_p^+), M_j\rangle\right] \\
&\quad - \eta\sum_{x_{n,s}\in\mathcal{B}_k^-}\mathbb{E}\left[\ell'_{s,t}(x_p, B)\cdot h_{i,t}(x_{n,s})\cdot\langle\nabla_{w_i}h(x_p^+), M_j\rangle\right] \pm \eta\frac{\|w_i^{(t)}\|_2}{\text{poly}(d_1)} \\
&= (1-\eta\lambda)\langle w_i^{(t)}, M_j\rangle - \Psi_{i,j}^{(t)} - \Phi_{i,j}^{(t)} - \xi_{i,j}^{(t)} + \eta\frac{\|w_i^{(t)}\|_2}{\text{poly}(d_1)} \\
|\langle w_i^{(t+1)}, M_j\rangle| &\le (1-\eta\lambda+\frac{\eta\rho}{d^2})|\langle w_i^{(t)}, M_j\rangle| + O(\frac{\Xi_2^2\|w_i^{(t)}\|_2}{d^2\sqrt{K+1}}) + \eta\frac{\|w_i^{(t)}\|_2}{\text{poly}(d_1)}
\end{aligned}$$

For all $i\notin\mathcal{M}_j$ we can choose η and λ such that $\lambda-\frac{\rho}{d^2}>0$. We can select $\eta=\eta'\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}$, where keeping $\eta'\approx\frac{1}{\text{poly}d}$ and $\lambda\in[\frac{1}{d^{1.01}},\frac{1}{d^{1.49}}]$.From the Lemmas F.4,F.5, and F.6

$$\begin{aligned}
|\langle w_i^{(t)}, M_j\rangle| &\le (1-\eta'\lambda\sqrt{\frac{K+1}{\alpha_1\rho_1^2}} + \frac{\eta'}{d^2}\sqrt{\frac{K+1}{\alpha_1\rho_1^2}})|\langle w_i^{(t-1)}, M_j\rangle| + O(\frac{\eta'\Xi_2^2\|w_i^{(t)}\|_2}{d^2(K+1)}\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}) \\
&\le \Big(1-\eta'\sqrt{\frac{K+1}{\alpha_1\rho_1^2}}(\lambda-\frac{1}{d^2})\Big)|\langle w_i^{(t-1)}, M_j\rangle| + O(\frac{\eta'\Xi_2^2\|w_i^{(t)}\|_2}{d^2\sqrt{\alpha_1\rho_1^2(K+1)}}) \le O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d(K+1)\Xi_2^5}})
\end{aligned}$$

Now we begin to prove for all $i\in[m]$, we have $\max_{j\in[d_1]\setminus[d]}|\langle w_i^{(t)}, M_j^\perp\rangle| \le O(\frac{\|w_i^{(t)}\|_2}{\sqrt{d_1(K+1)\Xi_2^5}})$, and we have for $j\in[d_1]\setminus[d], |\Phi_{i,j}^{(t)}|=0$. At the iteration $t=T_2$. We apply this to the SGD update

$$\begin{aligned}
|\langle w_i^{(t+1)}, M_j^\perp\rangle| &\le (1-\eta\lambda)|\langle w_i^{(t)}, M_j^\perp\rangle| + |\Psi_{i,j}^{(t)}| + |\Phi_{i,j}^{(t)}| + |\xi_{i,j}^{(t)}| + \eta\frac{\|w_i^{(t)}\|_2}{\text{poly}(d_1)} \\
&\le (1-\eta\lambda)|\langle w_i^{(t)}, M_j^\perp\rangle| + O(\eta\frac{\Xi_2^2\|w_i^{(t)}\|_2}{d^2\sqrt{d_1}(K+1)}) + \eta\frac{\|w_i^{(t)}\|_2}{\text{poly}(d_1)} \\
&\le (1-\eta'\sqrt{\frac{K+1}{\alpha_1\rho_1^2}})|\langle w_i^{(t)}, M_j^\perp\rangle| + O(\frac{\eta'\Xi_2^2\|w_i^{(t)}\|_2}{d^2\sqrt{d_1}\sqrt{\alpha_1\rho_1^2(K+1)}}) + \eta\frac{\|w_i^{(t)}\|_2}{\text{poly}(d_1)}
\end{aligned}$$

From the Lemma H.7 we can obtained the last inequality

$$\begin{aligned}
\xi_{i,j}^{(t)} &\le O(\frac{\Xi_2^2\|w_i^{(t)}\|_2}{\tau d^2(K+1)})\max_{i'\in[m]}\Big(|\langle w_{i'}^{(t)}, M_j^\perp\rangle| + |\frac{\|w_{i'}^{(t)}\|_2}{\sqrt{d_1}}\Big) \\
&\le O(\frac{\Xi_2^2\|w_i^{(t)}\|_2}{d^2\sqrt{d_1}(K+1)})
\end{aligned}$$

After above proven we can show that

$$\begin{aligned}
\|w_i^{(t)}\|_2^2 &= \sum_{j\in\mathcal{S}_i}\langle w_i^{(t)}, M_j\rangle^2 + \sum_{j\notin\mathcal{S}_i}\langle w_i^{(t)}, M_j\rangle^2 + \sum_{j\in[d_1]\setminus[d]}\langle w_i^{(t)}, M_j^\perp\rangle^2 \\
&\le \sum_{j\in\mathcal{S}_i}\langle w_i^{(t)}, M_j\rangle^2 + O(d).O(\frac{\|w_i^{(t)}\|_2^2}{d(K+1)\Xi_2^{10}}) + O(d_1).O(\frac{\|w_i^{(t)}\|_2^2}{d_1(K+1)\Xi_2^{10}}) \\
&\le \sum_{j\in\mathcal{S}_i}\langle w_i^{(t)}, M_j\rangle^2 + o(\frac{\|w_i^{(t)}\|_2^2}{\Xi_2^{10}(K+1)})
\end{aligned}$$

Since we have $\mathfrak{F}_j^t\le O(\tau\log d)$, we can conclude that from the above derivation for all $i\in[m]$

$$\|w_i^{(t)}\|_2^2 \le O(1)$$

Next we prove $\mathfrak{F}_j^{(t)} \leq O(\tau \log d)$ for each $j \in [d]$ and $i \in \mathcal{M}_j$

$$\begin{aligned}\langle w_i^{(t+1)}, M_j\rangle &= (1-\eta\lambda)\langle w_i^{(t)}, M_j\rangle + \Psi_{i,j}^{(t)} + \Phi_{i,j}^{(t)} + \xi_{i,j}^{(t)} + \eta\frac{\|w_i^{(t)}\|_2}{\text{poly}(d_1)} \\ &\leq (1-\eta\lambda+\frac{\eta}{\text{poly}(d)})\langle w_i^{(t)}, M_j\rangle + O(\frac{\Xi_2^2\|w_i^{(t)}\|_2}{d^2\sqrt{K+1}}) \\ \langle w_i^{(t+1)}, M_j\rangle &\leq (1-\eta'\sqrt{\frac{K+1}{\alpha_1\rho_1^2}} + \frac{\eta'}{\text{poly}(d)}\sqrt{\frac{K+1}{\alpha_1\rho_1^2}})\langle w_i^{(t)}, M_j\rangle\end{aligned}$$

Since from the values chosen for $\eta' \approx \frac{1}{\text{poly}d}$ and $\lambda \in [\frac{1}{d^{1.01}}, \frac{1}{d^{1.49}}]$ we can conclude that $\lambda > \frac{1}{\text{poly}(d)}$. This ensures that $\langle w_i^{(t+1)}, M_j\rangle \leq \langle w_i^{(t)}, M_j\rangle$.This proves $\mathfrak{F}_j^{(t+1)} \leq \mathfrak{F}_j^{(t)}$ which means $\mathfrak{F}_j^t$ has the upper bound at iteration t.

Now we prove the $|\langle w_i^{(t)}, M_j\rangle| \geq C \max_{i' \in \mathcal{M}_j} |\langle w_i^{(t)}, M_j\rangle|$ for $j \in [d]$. When $t \in [T', \frac{d^{1.5}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}]$ from the place we bound the H_1

$$\left|\frac{|\Psi_{i,j}^{(t)}|}{|\langle w_i^{(t)}, M_j\rangle|} - \frac{|\Psi_{i',j}^{(t)}|}{|\langle w_{i'}^{(t)}, M_j\rangle|}\right| \leq O\Big(\frac{\Xi_2}{\eta'\alpha_1\rho_1^2(K+1)\sqrt{dt}}\Big)$$

We can substitute $\tilde{\Psi}_{i,j}^{(t)} = \frac{|\Psi_{i,j}^{(t)}|}{|\langle w_i^{(t)}, M_{j'}\rangle|}$.First, at iteration, $t \in [T_1', \frac{d^{1.49}}{C_\eta\eta'\alpha_1\rho_1^2\sqrt{K+1}}]$ we can argue that $|\tilde{\Psi}_{i,j}^{(t)} - \lambda| \leq O\Big(\frac{\Xi_2}{\eta t}\Big)$

$$\begin{aligned}\frac{|\langle w_i^{(t+1)}, M_j\rangle|}{|\langle w_{i'}^{(t+1)}, M_j\rangle|} &= \frac{(1-\eta\lambda)|\langle w_i^{(t)}, M_j\rangle| + \eta|\tilde{\Psi}_{i,j}^{(t)}| \pm O(\frac{\sqrt{\Xi_2}}{\sqrt{(K+1)dt}})\|w_i^{(t)}\|_2}{(1-\eta\lambda)|\langle w_{i'}^{(t)}, M_j\rangle| + \eta|\tilde{\Psi}_{i',j}^{(t)}| \pm O(\frac{\sqrt{\Xi_2}}{\sqrt{(K+1)dt}})\|w_{i'}^{(t)}\|_2} \\ &= \frac{(1-\eta\lambda+\eta\tilde{\Psi}_{i,j}^{(t)})|\langle w_i^{(t)}, M_j\rangle| \pm O(\frac{\sqrt{\Xi_2}}{\sqrt{(K+1)dt}})\|w_i^{(t)}\|_2}{(1-\eta\lambda+\eta\tilde{\Psi}_{i',j}^{(t)})|\langle w_{i'}^{(t)}, M_j\rangle| \pm O(\frac{\sqrt{\Xi_2}}{\sqrt{(K+1)dt}})\|w_{i'}^{(t)}\|_2} \\ &\geq \frac{|\langle w_i^{(t)}, M_j\rangle|(1+\eta(\tilde{\Psi}_{i,j}^{(t)}-\lambda)(1-\frac{\Xi_2}{\sqrt{\alpha_1\rho_1^2(K+1)d}}))}{|\langle w_{i'}^{(t)}, M_j\rangle|(1+\eta(\tilde{\Psi}_{i',j}^{(t)}-\lambda)(1+\frac{\Xi_2}{\sqrt{\alpha_1\rho_1^2(K+1)d}}))} \\ &\geq \frac{|\langle w_i^{(t)}, M_j\rangle|(1-\frac{\Xi_2}{\sqrt{\alpha_1\rho_1^2(K+1)d}})}{|\langle w_{i'}^{(t)}, M_j\rangle|(1+\frac{\Xi_2}{\sqrt{\alpha_1\rho_1^2(K+1)d}})} \\ \frac{|\langle w_i^{(t+1)}, M_j\rangle|}{|\langle w_{i'}^{(t+1)}, M_j\rangle|} &\geq (1-\frac{\Xi_2}{t\sqrt{\alpha_1\rho_1^2(K+1)d}})\frac{|\langle w_i^{(t)}, M_j\rangle|}{|\langle w_{i'}^{(t)}, M_j\rangle|} \\ &\geq \prod_{s=T_1'}^{t-1}(1-\frac{\Xi_2}{s\sqrt{\alpha_1\rho_1^2(K+1)d}})\frac{|\langle w_i^{(t)}, M_j\rangle|}{|\langle w_{i'}^{(t)}, M_j\rangle|}\end{aligned}$$

We follow the similar argument as above for the iterations $t \in [\frac{d^{1.49}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}, \frac{d^{1.99}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}}]$. Only difference is the update step $|\langle w_i^{(t+1)}, M_j\rangle| = (1-\eta\lambda)|\langle w_i^{(t)}, M_j\rangle| + \eta|\tilde{\Psi}_{i,j}^{(t)}| \pm O(\frac{\Xi_2}{d^2\sqrt{\alpha_1\rho_1^2(K+1)}})$

Definition H.8. We define "The optimal feature map $f_{t,\theta,i}$" as $\theta_i^* = \frac{\sqrt{\tau\kappa}}{|\mathcal{M}_j^{1*}|} \cdot sign(\langle w_i^{(T_2)}, M_j\rangle)$if $i \in \mathcal{M}_j$ and $\theta_i^* = 0, i \notin \mathcal{M}_j$ for $\kappa > 0$. We define the pseudo loss function as

$$\tilde{L}(f_{t,\theta^*}, f_{\mathcal{G}}^t) := \mathbb{E}\left[-\tau\log\left(\frac{e^{\langle f_{t,\theta^*}(\bar{x}_k^{(1)}), f_{\mathcal{G}}^t(\bar{x}_k^{(2)})\rangle/\tau}}{\sum_{x\in B} e^{\langle f_{t,\theta^*}(\bar{x}_k^{(1)}), f_{\mathcal{G}}^t(x)\rangle/\tau}}\right)\right]$$

$$\frac{1}{\tau}\sum_{i\in[m]} f_{t,\theta,i}(\bar{x}_k^{(1)})\Big[h_{i,t}(\bar{x}_{n,s}) - h_{i,t}(\bar{x}_k^{(2)})\Big] = \frac{1}{\tau}\sum_{i\in\mathcal{M}_j} f_{t,\theta,i}(\bar{x}_k^{(1)})\Big[h_{i,t}(\bar{x}_{n,s}) - h_{i,t}(\bar{x}_k^{(2)})\Big] + 0$$

$$\leq \frac{1}{\tau}\frac{\sqrt{\tau}\kappa}{|\mathcal{M}_j^{1*}|}\sum_{i\in\mathcal{M}_j}\langle\theta_i^*, M_j\rangle\langle w_i^{(t)}, M_j\rangle\bar{z}_{k,j}^{(1)}(\bar{z}_{n,s,j} - \bar{z}_{k,j}^{(2)}) + O(\frac{1}{\alpha_1\rho_1^2(K+1)\log d})$$

$$\leq \frac{1}{\tau}\frac{\sqrt{\tau}\kappa}{|\mathcal{M}_j^{1*}|}\Big[\sum_{i\in\mathcal{M}_j}\langle w_i^{(t)}, M_j\rangle\cdot sign(\langle w_i^{(T_2)}, M_j\rangle)\bar{z}_{k,j}^{(1)}(\bar{z}_{n,s,j} - \bar{z}_{k,j}^{(2)})\Big]$$

$$+ O(\frac{1}{\alpha_1\rho_1^2(K+1)\log d})$$

$$\leq -\frac{1}{\tau}\frac{\sqrt{\tau}\kappa}{|\mathcal{M}_j^{1*}|}\Big[\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)}, M_j\rangle|\bar{z}_{k,j}^{(1)}(\bar{z}_{k,j}^{(2)} - \bar{z}_{n,s,j})\Big] + O(\frac{1}{\alpha_1\rho_1^2(K+1)\log d})$$

By choosing $\kappa = O(\Xi_2)$

$$\leq -\frac{1}{\sqrt{\tau}}\Big[\sum_{i\in\mathcal{M}_j}|\langle w_i^{(t)}, M_j\rangle|\bar{z}_{k,j}^{(1)}(\bar{z}_{k,j}^{(2)} - \bar{z}_{n,s,j})\Big] + O(\frac{1}{\alpha_1\rho_1^2(K+1)\log d})$$

We define a set $G_k, k \in N$ for the positive node set

$$G_k := \{j \in [d] : \bar{z}_{k,j}^{(1)}, \bar{z}_{k,j}^{(2)} \neq 0, sign(\bar{z}_{k,j}^{(1)}) = sign(\bar{z}_{k,j}^{(2)})\}$$

$$|G_k| = \sum_{j\in G_k}\mathbb{1}_{\bar{z}_{k,j}^{(1)}\bar{z}_{k,j}^{(2)}\neq 0} = \Omega(\log\log d)$$

Then for every $j \in G_k$,

$$\frac{1}{(K+1)^2} \leq \bar{z}_{k,j}^{(1)}\bar{z}_{k,j}^{(2)} \leq \frac{(\rho K+1)^2}{(K+1)^2}$$

For negative samples we define the bad overlap count as

$$C_s := \sum_{j\in G_k}\mathbb{1}\{sign(\bar{z}_{n,s,j}) = sign(\bar{z}_{k,j}^{(1)}), \bar{z}_{n,s,j} \neq 0\}$$

Next we need to find the probability of occurring this "bad event" $E_j := \{sign(\bar{z}_{n,s,j}) = sign(\bar{z}_{k,j}^{(1)}), \bar{z}_{n,s,j} \neq 0\}$. We have $\Pr(z_{n,s,j} \neq 0) = \frac{O(\log\log d)}{d}$ and

$$p_{bad} = \Pr(z_{n,s,j} \neq 0).\Pr\Big(sign(\bar{z}_{n,s,j}) = sign(\bar{z}_{k,j}^{(1)})|z_{n,s,j} \neq 0\Big)$$

Now we need to find the size of the subset of G_k which satisfies the conditions of the "bad event" with probability. We can consider this $E_j \sim Bern(p_{bad})$ and $C_s \sim Bin(|G_k|, p_{bad})$. Next we assume $Cs \geq m|G_k|$ for $m \in (0,1)$, from the chernoff bound for the sum of independent bernouili random variables

$$\Pr(C_s \geq m|G_k|) \leq (\frac{e\mu}{m|G_k|})^{m|G_k|} \leq \Big(\frac{ep_{bad}|G_k|}{m|G_k|}\Big)^{m|G_k|} \leq \Big(\frac{c\log\log d}{d}\Big)^{m|G_k|} \approx d^{-\Omega(\log\log d)}$$

In our data model we have $|\mathfrak{B}_k^-| = \text{polylog}d$ number of negative samples then $\Pr(\exists s \in [\mathfrak{B}_k^-] : C_s \geq m|G_k|) \leq \mathfrak{B}_k^- \cdot d^{-\Omega(\log\log d)} \approx d^{-\Omega(\log\log d)}$.Thus w.h.p for all $s \in [\mathfrak{B}_k^-], C_s \leq m|G_k|$.This finalizes the our proof by proving

$$\sum_{j\in G_k}\mathbb{1}_{\bar{z}_{k,j}^{(1)}=\bar{z}_{n,s,j}} \leq \sum_{j\in G_k}\mathbb{1}_{\bar{z}_{k,j}^{(1)}\bar{z}_{k,j}^{(2)}\neq 0}$$

Therefore, we can conclude that w.h.p, $\bar{z}_{k,j}^{(1)}(\bar{z}_{k,j}^{(2)} - \bar{z}_{n,s,j})$ is always positive

$$\begin{aligned}
\widetilde{\mathcal{L}}(f_{\mathcal{G}}^t, \theta^\star, f_{\mathcal{G}}^t) &\leq \left(1 - \frac{1}{\text{polylog}(d)}\right) \mathbb{E}\Big[\log\Big(1 + \sum_{x_{n,s} \in \mathcal{B}_k^-} e^{\langle f_{t,\theta^*}(\bar{x}_k^{(1)}), f_t(x_{n,s})\rangle/\tau - \langle f_{t,\theta^*}(\bar{x}_k^{(1)}), f_t(\bar{x}_k^{(2)})\rangle/\tau}\Big) | |\bar{z}_{k,j}^{(1)}| \neq 0\Big] \\
&\quad + \Pr(|\bar{z}_{k,j}^{(1)}| = 0) \cdot O(\log|\mathcal{B}|) \\
&\leq \left(1 - \frac{1}{\text{polylog}(d)}\right) \mathbb{E}\left[\log\left(\sum_{x \in \mathcal{B}} e^{\left(\frac{1}{\tau}\sum_{i \in \mathcal{M}_j *} f_{t,\theta,i}(\bar{x}_k^{(1)})\left[h_{i,t}(x_{n,s}) - h_{i,t}(\bar{x}_k^{(2)})\right]\right)}\right) \,\middle|\, |\bar{z}_{k,j}^{(1)}| \neq 0\right] \\
&\quad + \frac{1}{\text{polylog}(d)} \cdot O(\log|\mathcal{B}|) \\
&\leq \left(1 - \frac{1}{\text{polylog}(d)}\right) \mathbb{E}\left[\log\left(\sum_{x \in \mathcal{B}} e^{\left(-\frac{1}{\sqrt{\tau}}\left[\sum_{i \in \mathcal{M}_j} |\langle w_i^{(t)}, M_j\rangle| \bar{z}_{k,j}^{(1)}(\bar{z}_{k,j}^{(2)} - \bar{z}_{n,s,j})\right] + O(\frac{1}{\alpha_1 \rho_1^2 (K+1)\log d})\right)}\right) \,\middle|\, |\bar{z}_{k,j}^{(1)}| \neq 0\right] \\
&\quad + \frac{1}{\text{polylog}(d)} \cdot O(\log|\mathcal{B}|) \\
&\leq O\Big(\frac{1}{\alpha_1 \rho_1^2 (K+1) \cdot \log d}\Big)
\end{aligned}$$

H.1 Proof of Theorem 3.1

Theorem 3.1 can be derived by the Induction Hypothesis H.2 (1),(2),(3), and (4). Also Theorem 3.1 coinsides with the Theorem H.1 (e). Hense we give the complete proof of the Theorem 3.1 as follows

Proof of Theorem 3.1
From the Induction Hypothesis H.2 we have,

- For $i \in \mathcal{M}_j^{1*}$ $|\langle w_i^{(t)}, M_j\rangle| \geq \Omega(1)\|w_i^{(t)}\|_2$
- For $i \in [m]$ $\|w_i^{(t)}\|_2 \leq O(1)$
- For $j \in [d]$ and $i \in \mathcal{M}_j^{1*}$ we have $|\langle w_i^{(t)}, M_j\rangle| \geq C_1 \max_{i' \in \mathcal{M}_j} |\langle w_{i'}^{(t)}, M_j\rangle|$
- From Lemma H.4 we have $\mathfrak{F}_j^{(t)} \geq O(\frac{\tau}{\alpha_1 \rho_1^2} \log d)$

First derive the upper bound using the Induction Hypothesis For $i \in \mathcal{M}_j^{1*}$ we can write

$$|\langle w_i^{(t)}, M_j\rangle| \leq \|w_i^{(t)}\|_2 \|M_j\|_2 \leq O(1)$$

Let's take $\max_{i' \in \mathcal{M}_j} |\langle w_{i'}^{(t)}, M_j\rangle| = a$ and From the definition of the $\mathfrak{F}_j^{(t)}$ we can write

$$|\mathcal{M}_j| \cdot a^2 \geq O\frac{\tau}{\alpha_1 \rho_1^2} \log d)$$

$$a \geq O(\sqrt{\frac{\tau \log d}{\Xi_2 \alpha_1 \rho_1^2}}) \geq O(\frac{\tau}{\Xi_2 \alpha_1 \rho_1^2})$$

We have for $i \in \mathcal{M}_j^{1*}$, $|\langle w_i^{(t)}, M_j\rangle| \in [\frac{\tau}{\Xi_2 \alpha_1 \rho_1^2}, O(1)]$ for the iteration $t \in [\frac{d}{\eta' \sqrt{\alpha_1 \rho_1^2 (K+1)}} \frac{d^2}{\eta' \sqrt{\alpha_1 \rho_1^2 (K+1)}}]$. Following features will learn through contrastive learning at iteration $t \in [\frac{d}{\eta' \sqrt{\alpha_1 \rho_1^2 (K+1)}}, \frac{d^2}{\eta' \sqrt{\alpha_1 \rho_1^2 (K+1)}}]$

$$w_i^{(t)} = \sum_{j \in \mathcal{I}_i} A_{ij} M_j + \sum_{j \notin \mathcal{I}_i} B_{ij} M_j + \sum_{j \in [d']} B'_{ij} M_j + \sum_{j \in [d_1 - (d+d'))]} C_{ij} M_j^\perp$$

Where $A_{ij} \in [\frac{\tau}{\Xi_2 \alpha_1 \rho_1^2}, O(\tau)], B_{ij} \leq o(\frac{1}{\sqrt{d(K+1)}})\|w_i^{(t)}\|_2, B'_{ij} \leq o(\frac{\alpha_2 \rho_2^2}{\alpha_1 \rho_1^2} \frac{1}{\sqrt{d(K+1)}})\|w_i^{(t)}\|_2, C_{ij} \leq o(\frac{1}{\sqrt{d_1(K+1)}})\|w_i^{(t)}\|_2$

Next, we compute the upper bound of $|\mathcal{S}_i|$. for $\gamma \in (0, 0.01)$

$$\mathbb{E}[|\mathcal{I}_i|] = \frac{1}{m}\sum_{i=1}^{m}|\mathcal{I}_i| = \frac{1}{m}\sum_{j=1}^{d}|\mathcal{M}_j^1| \le \frac{1}{m}\cdot d\cdot O(d^{2\gamma c_0})$$
$$= O(d^{2\gamma c_0 - 0.01}) \tag{191}$$
$$\le o(m/d) \tag{192}$$

Let $\delta = 0.01 - 2\gamma c_0 > 0$, so that $\mathbb{E}[|\mathcal{I}_i|] = O(d^{-\delta})$. For any fixed integer $r \ge 1$, using the independence of the initialization coordinates and a union bound over all r-subsets of $[d]$,

$$\Pr(|\mathcal{I}_i| \ge r) \le \binom{d}{r}\left(\frac{\mu_i}{d}\right)^r \le \left(\frac{e\mu_i}{r}\right)^r = O(d^{-r\delta}).$$

Taking a union bound over all $m = d^{1.01}$ neurons gives

$$\Pr\left[\max_{i\in[m]}|\mathcal{I}_i| \ge r\right] \le O(d^{1.01-r\delta}).$$

Therefore, choosing any constant $r > \frac{1.01+c}{0.01-2\gamma c_0}$ yields $\Pr\left[\max_{i\in[m]}|\mathcal{I}_i| \le r\right] \le 1 - d^{c}$Hence, $|\mathcal{I}_i| = O(1)$ for every $i \in [m]$ with high probability.

Let $g_t := \nabla_w Obj(f_{\mathcal{G}}^t), \qquad e_t := O(\frac{1}{\mathrm{poly}(d_1)\sqrt{(K+1)}})$ Weight update of the model is $w^{(t+1)} = w^{(t)} - \eta g_t + e_t$.

Now we define the error to the comparator $\theta^\star$ by $u_t := w^{(t)} - \theta^\star$. Then we can write $u_{t+1} = u_t - \eta g_t + e_t$..Taking the squared Frobenius norm

$$\|u_{t+1}\|_F^2 = \|u_t - \eta g_t + e_t\|_F^2.$$
$$\|u_{t+1}\|_F^2 = \|u_t\|_F^2 + \eta^2\|g_t\|_F^2 + \|e_t\|_F^2 - 2\eta\langle g_t, u_t\rangle + 2\langle u_t, e_t\rangle - 2\eta\langle g_t, e_t\rangle.$$

Rearrange to isolate the inner product:

$$2\eta\langle g_t, u_t\rangle = \eta^2\|g_t\|_F^2 + \|u_t\|_F^2 - \|u_{t+1}\|_F^2 + \|e_t\|_F^2 + 2\langle u_t, e_t\rangle - 2\eta\langle g_t, e_t\rangle.$$

$$\eta\langle\nabla_w Obj(f_{\mathcal{G}}^t),\, w^{(t)} - \theta^\star\rangle = \frac{\eta^2}{2}\|\nabla Obj(f_{\mathcal{G}}^t)\|_F^2 + \frac{1}{2}\|w^{(t)} - \theta^\star\|_F^2 - \frac{1}{2}\|w^{(t+1)} - \theta^\star\|_F^2 + \frac{\eta^2}{\mathrm{poly}(d_1)(K+1)},$$

Considering the convexity of the $\widetilde{Obj}_t(w^{(t)}$ we can derive

$$\langle\nabla_\theta\widetilde{Obj}_t(w^{(t)}),\, w^{(t)} - \theta^\star\rangle \ge \widetilde{Obj}_t(w^{(t)}) - \widetilde{Obj}_t(\theta^\star).$$

$$\eta\big(\widetilde{Obj}_t(w^{(t)}) - \widetilde{Obj}_t(\theta^\star)\big) \le \frac{\eta^2}{2}\|\nabla Obj(f_{\mathcal{G}}^t)\|_F^2 + \frac{1}{2}\|w^{(t)} - \theta^\star\|_F^2 - \frac{1}{2}\|w^{(t+1)} - \theta^\star\|_F^2 + \frac{\eta^2}{\mathrm{poly}(d_1)(K+1)}.$$

$$\eta'\sqrt{\frac{\alpha\rho^2}{K+1}}\big(\widetilde{Obj}_t(w^{(t)}) - \widetilde{Obj}_t(\theta^\star)\big) \le \frac{(K+1)}{\alpha_1\rho_1^2\mathrm{poly}(d)} + \frac{1}{2}\|w^{(t)} - \theta^\star\|_F^2 - \frac{1}{2}\|w^{(t+1)} - \theta^\star\|_F^2 + \frac{C_\eta^2\eta'^2}{\alpha_1\rho_1^2\mathrm{poly}(d_1)}.$$

$$\eta'\sqrt{\frac{\alpha\rho^2}{K+1}}\sum_{t=t_0}^{t_0+T-1}\big(\widetilde{Obj}_t(w^{(t)}) - \widetilde{Obj}_t(\theta^\star)\big) \le \sum_{t=t_0}^{T-1}\left[\frac{1}{2}\|w^{(t)} - \theta^\star\|_F^2 - \frac{1}{2}\|w^{(t+1)} - \theta^\star\|_F^2\right] + \frac{T\eta'^2}{\alpha_1\rho_1^2\mathrm{poly}(d_1)}.$$

The middle sum by telescopes expansion for $t = T_3, T_3+1, \ldots, T_3+T-1$, where $t_0 = T_3$ and $T_4 = T_3 + T$

$$\sum_{t=T_3}^{T_3+T-1}\left[\frac{1}{2}\|w^{(t)} - \theta^\star\|_F^2 - \frac{1}{2}\|w^{(t+1)} - \theta^\star\|_F^2\right] = \frac{1}{2}\|w^{(t_0)} - \theta^\star\|_F^2 - \frac{1}{2}\|w^{(T)} - \theta^\star\|_F^2 \le \frac{1}{2}\|w^{(t_0)} - \theta^\star\|_F^2.$$

$$\eta'\sqrt{\frac{\alpha\rho^2}{K+1}}\sum_{t=T_3}^{T_3+T-1}\big(\widetilde{Obj}_t(w^{(t)}) - \widetilde{Obj}_t(\theta^\star)\big) \le \frac{1}{2}\|w^{(t_0)} - \theta^\star\|_F^2 + \frac{TC_\eta^2\eta'^2}{\mathrm{poly}(d_1)}$$

$$\frac{1}{T}\sum_{t=T_3}^{T_3+T-1}\big(\widetilde{Obj}_t(w^{(t)}) - \widetilde{Obj}_t(\theta^\star)\big) \le \frac{\|w^{(t_0)} - \theta^\star\|_F^2}{2\eta' T}\sqrt{\frac{\alpha_1\rho_1^2}{K+1}} + \frac{\eta'}{\sqrt{\alpha_1\rho_1^2(K+1)}\mathrm{poly}(d_1)}$$
$$\le \frac{\|w^{(t_0)} - \theta^\star\|_F^2}{2\eta' T}\sqrt{\frac{\alpha_1\rho_1^2}{K+1}}$$
$$\le O\Big(\frac{\tau m\Xi_2}{\eta' T\sqrt{\alpha_1\rho_1^2(K+1)}}\Big) \text{ For } T \ge \frac{m\Xi_2^{10}}{\eta'\sqrt{\alpha_1\rho_1^2(K+1)}} \text{ Condition proved}$$

From the results of the Induction hypothesis we have For $i \in [m]$ $\|w_i^{(t)}\|_2 \leq O(1)$ and the definition of the optimal weight we can argue that $\sum_{i\in[m]} \|w_i^{(t_0)}\|_2^2 \leq O(m)$, $\sum_{i\in[m]} \|\theta_i^\star\|_2^2 = \sum_{j\in[d]} \sum_{i\in\mathcal{M}_j^{1*}} \frac{\tau\kappa^2}{\alpha_1\rho_1^2|\mathcal{M}_j^{1*}|^2} \leq O\Big(\frac{\tau\Xi_2^2 d}{\alpha_1\rho_1^2\Xi_1}\Big) \leq O\Big(\tau d^{1.01}\Xi_2/\alpha_1\rho_1^2\Big) \leq O(\tau m\Xi_2/\alpha_1\rho_1^2)$.Here we use $m = d^{1.01}$ and we can compute $\|w^{(t_0)} - \theta^\star\|_F^2 \leq 2\sum_{i\in[m]} \|w_i^{(t_0)}\|_2^2 + 2\sum_{i\in[m]} \|\theta_i^\star\|_2^2 \leq O(m) + O(\tau m\Xi_2/\alpha_1\rho_1^2)$.

I Downstream- Task Training Convergence

The downstream logistic objective is convex in the linear classifier. To parallel the online-learning argument used for the contrastive objective, we define a downstream pseudo-objective $Obj_T^{dwn}(\mu)$. For each node v, the learned representation of the graph centered at v is$f(x_v)$. The downstream label is $y_v = sign\big(\langle u^\star, z_v^{(1)}\rangle\big)$ where $u^\star \in \mathbb{R}^{d/2}$ is the ground-truth downstream task vector in the latent feature space and $z_v^{(1)} \in \mathbb{R}^{d/2}$ is the first half of the latent sparse vector of node v.

For downstream linear classifier $U \in \mathbb{R}^m$,downstream learning rate η_d, and downstream regularizer parameter λ_d define the downstream pseudo-objective

$$Obj_T^{dwn}(U) = \mathbb{E}_v\big[\log\big(1 + exp\big(-y_v\langle U, f(x_v)\rangle\big)\big)\big] + \frac{\lambda_d}{2}\|U\|_2^2$$

Theorem I.1 (Downstream training convergence)**.** *Let $\lambda_d = O(1/d)$ and $\eta_d = O(1/\operatorname{poly}(d))$. For $S\eta_d = O(1)$, the downstream classifier satisfies*

$$\frac{1}{S}\sum_{s=0}^{S-1} \mathrm{Obj}_T^{\mathrm{dwn}}(U^{(s)}) \leq \left(\frac{1}{\log d}\right)^{O\left(\sqrt{\frac{\alpha_1\rho_1^2(K+1)}{\log\log d}}\right)} \tag{193}$$

I.1 Proof of Theorem 3.2

Suppose $U \in \mathbb{R}^m$ is the classification layer weights. We train the downstream classifier by applying gradient descent

$$U^{(s+1)} = U^{(s)} - \eta_{dwn}\nabla Obj_T^{dwn}(U^{(s)})$$

Let $\tilde{U} \in \mathbb{R}^m$ be any comparison classifier defined as below for the selected neuron $s(j), \mathcal{N}_{s(j)} = \{j\}$.

$$\tilde{U} := \begin{cases} \dfrac{u_j^\star\sqrt{K+1}}{A_{s(j),j}|\mathcal{M}_j^{1*}|} & \text{if } i = s(j) \text{ for some } j \in [d/2] \\ 0 & \text{Otherwise} \end{cases}$$

Then the standard gradient-descent gives,

$$\eta_d\big(Obj_T^{dwn}(U^{(t)}) - Obj_T^{dwn}(\tilde{U})\big) \leq \frac{\eta_d^2}{2}\|\nabla Obj_T^{dwn}(U^{(s)})\|_2^2 + \frac{1}{2}\|U^{(t)} - \tilde{U}\|_2^2 - \frac{1}{2}\|U^{(t+1)} - \tilde{U}\|_2^2 \tag{194}$$

$$\eta_d\sum_{s=0}^{S-1}\big(Obj_T^{dwn}(U^{(s)}) - Obj_T^{dwn}(\tilde{u})\big) \leq \frac{\eta_d^2}{2}\sum_{s=0}^{S-1}\|\nabla Obj_T^{dwn}(U^{(s)})\|_2^2 + \frac{1}{2}\|U^{(0)} - \tilde{U}\|_2^2 \tag{195}$$

$$\frac{1}{S}\sum_{s=0}^{S-1} Obj_T^{dwn}(U^{(s)}) \leq Obj_T^{dwn}(\tilde{u}) + \frac{\eta_d}{2S}\sum_{s=0}^{S-1}\|\nabla Obj_T^{dwn}(U^{(s)})\|_2^2 + \frac{\|U^{(0)} - \tilde{U}\|_2^2}{2S\eta_d} \tag{196}$$

Assume at the initialization of the linear layer weights, $U^{(0)} = 0$ then,

$$\|U^{(0)} - \tilde{U}\|_2^2 = \|\tilde{U}\|_2^2 \tag{197}$$

$$\|\tilde{U}\|_2 \leq O\Big(\frac{\Xi_2\alpha_1\rho_1^2\sqrt{K+1}\|u^\star\|_2}{\tau}\Big) \tag{198}$$

$$\begin{aligned}
\langle \tilde{U}, f_{\mathcal{G}}^t(x_v)\rangle &= \sum_{i=1}^m \tilde{U}_i h_{T,i}(x_v) = \sum_{i=1}^m \tilde{U}_i\big(\langle w_i^{(T)}, Mz_v\rangle + \langle w_i^{(T)}, \xi_v\rangle\big) \\
&= \sum_{i\in\mathcal{M}_j^{1*}} \tilde{U}_i\Big(\sum_{j\in\mathcal{S}_i} A_{s(j),j} z_{v,j} + \sum_{j\notin\mathcal{S}_i,\, j\in[d]} B_{s(j),j} z_{v,j} + \mathfrak{r}_v\Big) \\
&= \sum_{j=1}^{d/2} \sum_{i\in\mathcal{M}_j^{1*}} \frac{u_j^\star\sqrt{K+1}}{A_{s(j),j}|\mathcal{M}_j^{1*}|}\Big(A_{s(j),j} z_{v,j}^{(1)} + \sum_{k\notin[d/2],\, k\neq j} B_{s(k),k} z_{v,k}^{(1)} + \mathfrak{r}_{v,j}\Big) \\
&= \sum_{j=1}^{d/2} u_j^* z_{v,j}^{(1)}\sqrt{K+1} + \sum_{j=1}^{d/2} u_j^* \frac{\sqrt{K+1}}{|M_j^*|} \sum_{i\in M_j^*} \frac{\sum_{k\in[d/2]\, k\neq j} B_{ik} z_{v,k}^{(1)} + r_{v,i}}{A_{ij}} \\
&= \sqrt{K+1}(\langle u^\star, z_v^{(1)}\rangle + \langle u^\star, \mathfrak{e}_v\rangle) \\
y_v\langle u^*, f_{\mathcal{G}}^t(x_v)\rangle &= \sqrt{K+1}(y_v\langle u^\star, z_v^{(1)}\rangle + y_v\langle u^\star, \mathfrak{e}_v\rangle) \geq \sqrt{K+1}(|\langle w^\star, z_v^{(1)}\rangle| - |\langle u^\star, \mathfrak{e}_v\rangle|) \\
&\geq \sqrt{K+1}\Big(O\Big(\sqrt{\alpha_1\rho_1^2 + \frac{1}{K+1}\log\log d}\Big) - \frac{\Xi_2}{(1+\frac{1}{\alpha_1\rho_1^2(K+1)})} o\Big(\sqrt{\frac{\log\log d}{d(K+1)^2}}\Big)\Big) \\
&\geq O(\sqrt{\alpha_1\rho_1^2(K+1)\log\log d}) - o(\frac{1}{\sqrt{(K+1)d}}) \qquad (199)
\end{aligned}$$

we assume $|u_j^\star| = O(1)$ for all $j \in [d]$ since our contrastive learned weight consists of label-relevant features in the first half of d. Hence $|\langle w^\star, z_v\rangle| = O\Big(\sqrt{\alpha_1\rho_1^2 + \frac{1}{K+1}\log\log d}\Big)$.For the logistic loss of the downstream task $Obj_T^{dwn}(u^*) = \mathbb{E}_v\big[\log\big(1 + e^{-y_v\langle u^*, f_{\mathcal{G}}^t(x_v)\rangle}\big)\big] + \frac{\lambda_d}{2}\|u^*\|_2^2 \leq \Big(\frac{1}{\log d}\Big)^{O(\sqrt{\frac{\alpha_1\rho_1^2(K+1)}{\log\log d}})} + O\Big(\frac{\lambda_d\Xi_2^2\alpha^2\rho^4\|w^\star\|_2^2(K+1)}{\tau^2\Xi_1^2}\Big)$

Before we compute the bound of $\|\nabla \text{Obj}_T^{\text{dwn}}(U^{(s)})\|_2$. We need to derive the gradient of the loss function $\ell_v(\mu) = \log\big(1+\exp(-y_v\langle\mu, f_{\mathcal{G}}^t(x_v)\rangle)\big)$ as follows $\nabla\ell_v(\mu) = -\frac{y_v f_{\mathcal{G}}^t(x_v)}{1+\exp(y_v\langle\mu, f_{\mathcal{G}}^t(x_v)\rangle)}$. Since $0 \leq \frac{1}{1+\exp(y_v\langle\mu, f_{\mathcal{G}}^t(x_v)\rangle)} \leq 1$ we have$\|\nabla\ell_v(\mu)\|_2 \leq \|f_{\mathcal{G}}^t(x_v)\|_2$.

$$\begin{aligned}
\|f_{\mathcal{G}}^t(x_v)\|_2^2 &= \sum_{i=1}^m h_{T,i}(x_v)^2 \qquad (200) \\
&= \sum_{i=1}^m \Big(\sum_{j\in\mathcal{S}_i} A_{ij} z_{v,j} + \sum_{\substack{j\notin\mathcal{S}_i \\ j\in[d/2]}} B_{ij} z_{v,j} + \sum_{\substack{j\notin\mathcal{S}_i \\ j\in[d]\setminus[d/2]}} B'_{ij} z_{v,j} + \sum_{j\in[d_1]\setminus[d]} C_{ij} p_{v,j}\Big)^2 \qquad (201)
\end{aligned}$$

We have $|\sum_{j\in\mathcal{S}_i} A_{ij} z_{v,j}| \leq \sum_{j\in\mathcal{S}_i} |A_{ij}||z_{v,j}| \leq O(\tau\rho)$ we have $|\mathcal{S}_i| = O(1)$. We neglect the remaining terms since they are small.

$$\begin{aligned}
&\leq O(m\tau^2\rho^2) \\
\|\nabla \text{Obj}_T^{\text{dwn}}(\mu)\|_2 &= \|\mathbb{E}_v[\nabla\ell_v(\mu)] + \lambda_{dwn}\mu\|_2 \leq \|f_{\mathcal{G}}^t(x_v)\|_2 + \lambda_{dwn}\|\mu\|_2 \qquad (202) \\
&\leq O(\sqrt{m}\tau\rho) + \lambda\|U^{(s)}\|_2 \qquad (203)
\end{aligned}$$

$$\frac{1}{S}\sum_{s=0}^{S-1} Obj_T^{dwn}(U^{(s)}) \leq \Big(\frac{1}{\log d}\Big)^{O(\sqrt{\frac{\alpha_1\rho_1^2(K+1)}{\log\log d}})} + O\Big(\frac{\Xi_2^2\|w^\star\|_2^2}{\text{poly}\log d}(\lambda_d + \frac{1}{S\eta_{dwn}})\Big) + \frac{1}{S}O\Big(\eta_{dwn}m\tau^2\rho^2 + \eta_{dwn}\lambda_{dwn}^2\|U^{(s)}\|_2^2\Big)$$

J ALGORITHM

Algorithm 1 Theory-Guided Neighborhood Augmentation (TGNA)

Require: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, node features X, similarity threshold γ, candidate ratios $\mathcal{R}$, number of random trials R

Compute the average degree of $\mathcal{G}$ and denote it by $2K$.
for each $v_k \in \mathcal{V}$ **do**
  Compute $s_{kj} = \frac{\langle x_k, x_j \rangle}{\|x_k\|_2 \|x_j\|_2}$ for all $v_j \in \mathcal{N}(k)$.
  $\mathcal{N}_k^+ \leftarrow \{v_j \in \mathcal{N}(k) : s_{kj} \geq \gamma\}$ and $\mathcal{N}_k^- \leftarrow \{v_j \in \mathcal{N}(k) : s_{kj} < \gamma\}$.
end for
$\widehat{\mathcal{S}}_{\textbf{TGNA}} \leftarrow 0$.
for each $\rho \in \mathcal{R}$ **do**
  Define

$$\mathcal{V}_{\mathrm{anc}}(\rho) \leftarrow \left\{ v_k \in \mathcal{V} : |\mathcal{N}_k^+| \geq 2\lfloor \rho K \rfloor,\ |\mathcal{N}_k^-| \geq 2\lfloor (1-\rho) K \rfloor \right\}.$$

  for $r = 1, \ldots, R$ **do**
    for each $v_k \in \mathcal{V}_{\mathrm{anc}}(\rho)$ **do**
      Randomly sample two disjoint subsets $\widetilde{\mathcal{N}}_k^{A+}, \widetilde{\mathcal{N}}_k^{B+} \subseteq \mathcal{N}_k^+$, each of size $\lfloor \rho K \rfloor$.
      Randomly sample two disjoint subsets $\widetilde{\mathcal{N}}_k^{A-}, \widetilde{\mathcal{N}}_k^{B-} \subseteq \mathcal{N}_k^-$, each of size $\lfloor (1-\rho) K \rfloor$.
      $\widetilde{\mathcal{N}}_k^A \leftarrow \widetilde{\mathcal{N}}_k^{A+} \cup \widetilde{\mathcal{N}}_k^{A-}$, and $\widetilde{\mathcal{N}}_k^B \leftarrow \widetilde{\mathcal{N}}_k^{B+} \cup \widetilde{\mathcal{N}}_k^{B-}$.
    end for
    Construct $\mathcal{G}^A$ and $\mathcal{G}^B$ from $\{\widetilde{\mathcal{N}}_k^A\}_{v_k \in \mathcal{V}_{\mathrm{anc}}(\rho)}$ and $\{\widetilde{\mathcal{N}}_k^B\}_{v_k \in \mathcal{V}_{\mathrm{anc}}(\rho)}$, respectively.
    Compute

$$\widehat{\alpha} \leftarrow \frac{1}{2|\mathcal{V}_{\mathrm{anc}}(\rho)|} \sum_{q \in \{A,B\}} \sum_{v_k \in \mathcal{V}_{\mathrm{anc}}(\rho)} \frac{1}{|\widetilde{\mathcal{N}}_k^{q+}|} \sum_{v_j \in \widetilde{\mathcal{N}}_k^{q+}} s_{kj}.$$

    Compute

$$\widehat{\rho} \leftarrow \frac{1}{2|\mathcal{V}_{\mathrm{anc}}(\rho)|} \sum_{q \in \{A,B\}} \sum_{v_k \in \mathcal{V}_{\mathrm{anc}}(\rho)} \frac{|\widetilde{\mathcal{N}}_k^{q+}|}{|\widetilde{\mathcal{N}}_k^q|},$$

$$\widehat{K} \leftarrow \frac{1}{2|\mathcal{V}_{\mathrm{anc}}(\rho)|} \sum_{q \in \{A,B\}} \sum_{v_k \in \mathcal{V}_{\mathrm{anc}}(\rho)} |\widetilde{\mathcal{N}}_k^q|.$$

    $S \leftarrow \widehat{\alpha} \widehat{\rho}^2 (\widehat{K} + 1)$.
    if $S > \widehat{\mathcal{S}}_{\textbf{TGNA}}$ **then**
      $\widehat{\mathcal{S}}_{\textbf{TGNA}} \leftarrow S$.
      $(\mathcal{G}^{A\star}, \mathcal{G}^{B\star}, \mathcal{V}_{\mathrm{anc}}^\star) \leftarrow (\mathcal{G}^A, \mathcal{G}^B, \mathcal{V}_{\mathrm{anc}}(\rho))$.
    end if
  end for
end for
for each $v_k \in \mathcal{V}_{\mathrm{anc}}^\star$ **do**
  $\mathcal{B}_k^- \leftarrow \mathcal{V}_{\mathrm{anc}}^\star \setminus \{v_k\}$.
end for
return $\mathcal{G}^{A\star}, \mathcal{G}^{B\star}, \mathcal{V}_{\mathrm{anc}}^\star, \{\mathcal{B}_k^-\}_{v_k \in \mathcal{V}_{\mathrm{anc}}^\star}$.

K More Experiments

Synthetic data

Evolution of Feature Alignment . Figures 3 and 4 together illustrate how the learned representation becomes increasingly concentrated on the desired feature block M^1. As the semantic signal strength $\sqrt{\alpha\rho^2(K+1)}$ increases, the final projection onto M^1 rapidly approaches one, while the projections onto M^2 and $M^\perp$ decrease toward zero Figure 3. Figure 4 further shows that this separation emerges early in training: the M^1-projection increases sharply and stabilizes near one, whereas the M^2- and $M^\perp$-projections quickly decay. Together, these results are consistent with our theoretical learning dynamics, where stronger semantic consistency preferentially amplifies the desired M^1-features, suppresses competing and spurious directions, and ultimately stabilizes the representation around M^1.

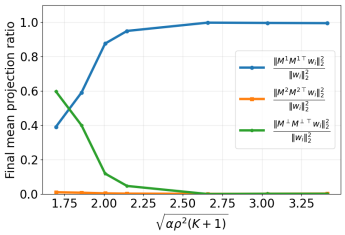


Figure 3: Mean projection ratio with semantic similarity

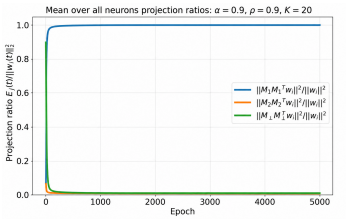


Figure 4: Mean projection ratios with Epochs

Downstream Classification Performance Figures 5and 6 together demonstrate the quality of the learned representations for downstream classification. As the semantic signal strength $\sqrt{\alpha\rho^2(K+1)}$ increases, the downstream classification accuracy improves substantially and approaches nearly perfect performance at larger values (Figure 5). This trend is consistent with the feature-alignment results in Figure 3, where stronger semantic similarity leads to greater concentration on the desired M^1-features. Figure 6 further shows that both downstream training and testing accuracies rise rapidly with training epochs and remain close to one, with only a small gap between them. Together, these results indicate that the GCL encoder learns task-relevant and generalizable representations that can be effectively exploited by a simple downstream classifier.

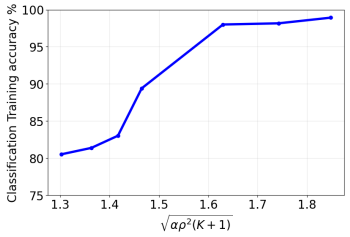


Figure 5: Classification Accuracy with semantic similarity.

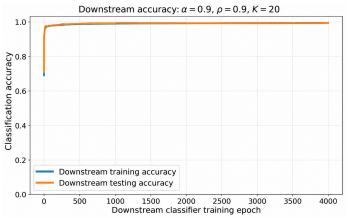


Figure 6: Classification Accuracies with Epochs.

Emergence and Stability of Lucky Neurons. Figure 7 tracks the number of lucky neurons aligned with M^1 throughout training. The number increases rapidly at the beginning and then remains nearly constant for the rest of the optimization process. This supports the theoretical argument that lucky neurons emerge during the early learning stage and retain their feature advantage, rather than being repeatedly replaced as training proceeds.

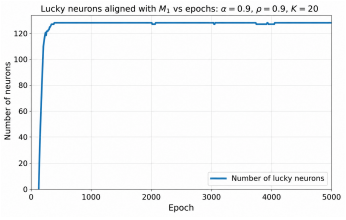


Figure 7: Number of Lucky neurons with Epochs.

REAL DATA

Figures 8, 9, 10 show the relationship between classification error and the estimated semantic signal strength $\sqrt{\hat{\alpha}\hat{\rho}^2(\hat{K}+1)}$ across PubMed, AmazonPhoto, and AmazonComputers.Although the magnitude and smoothness of this tendency differ between datasets and GCL backbones, the classification error generally tends to decrease as the semantic signal gets stronger. While AmazonComputers exhibits a lesser but still generally declining pattern for numerous techniques, PubMed and AmazonPhoto demonstrate the most consistent declines. These findings support our theoretical hypothesis that more informative graph representations arise from stronger neighborhood-level semantic consistency, which in turn improves downstream classification accuracy.

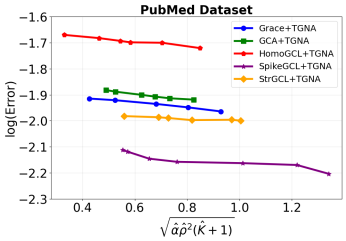


Figure 8: Classification error versus $\sqrt{\hat{\alpha}\hat{\rho}^2(\hat{K}+1)}$ on PubMed.

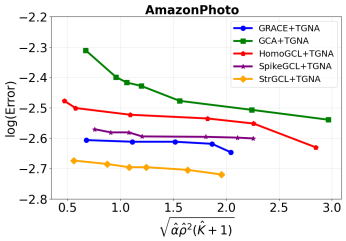


Figure 9: Classification error versus $\sqrt{\hat{\alpha}\hat{\rho}^2(\hat{K}+1)}$ on AmazonPhoto.

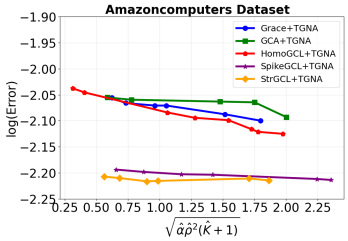


Figure 10: Classification error versus $\sqrt{\hat{\alpha}\hat{\rho}^2(\hat{K}+1)}$ on AmazonComputers.