Provably Neural Active Learning Succeeds via Prioritizing Perplexing Samples

Dake Bu [1] **Wei Huang*** [2] **Taiji Suzuki** [3 2] **Ji Cheng** [1] **Qingfu Zhang** [1] **Zhiqiang Xu** [4] **Hau-San Wong*** [1]

Abstract

Neural Network-based active learning (NAL) is a cost-effective data selection technique that utilizes neural networks to select and train on a small subset of samples. While existing work successfully develops various effective or theory-justified NAL algorithms, the understanding of the two commonly used query criteria of NAL: uncertainty-based and diversity-based, remains in its infancy. In this work, we try to move one step forward by offering a unified explanation for the success of both query criteria-based NAL from a feature learning view. Specifically, we consider a feature-noise data model comprising easy-to-learn or hard-to-learn features disrupted by noise, and conduct analysis over 2-layer NN-based NALs in the pool-based scenario. We provably show that both uncertainty-based and diversity-based NAL are inherently amenable to one and the same principle, i.e., striving to prioritize samples that contain yet-to-be-learned features. We further prove that this shared principle is the key to their success-achieve small test error within a small labeled set. Contrastingly, the strategy-free passive learning exhibits a large test error due to the inadequate learning of yet-to-be-learned features, necessitating resort to a significantly larger label complexity for a sufficient test error reduction. Experimental results validate our findings.

1. Introduction

In the deep learning era, we witness the power of neural networks in representation learning. It is also well-known that their success relies on a substantial amount of data and extensive labeling efforts. On the other hand, active learning offers various approaches to select a small subset of unlabeled samples from a large pool of data for labeling and training, while achieving comparable generalization performance to learning on the entire dataset (Settles, 2009; Aggarwal et al., 2014). To enjoy the best of both worlds, people combine neural networks with active learning, giving rise to Neural Network-based Active Learning (NAL) or Deep Active Learning (DAL), such that over-parameterized neural models can work with limited size of labeled data. As summarized in Takezoe et al. (2023), NAL/DAL incorporates two primary criteria for querying (selecting) unlabeled samples: uncertainty-based (Roth and Small, 2006; Joshi et al., 2009) and diversity-based (Sener and Savarese, 2018; Gissin and Shalev-Shwartz, 2019). Also, some studies leverage both criteria to design NAL algorithms (Yin et al., 2017; Shui et al., 2020).

Notably, while various NAL algorithms, based on two query criteria, have achieved significant empirical success, they often come without provable performance guarantees. To overcome this limitation, recent studies (Gu et al., 2014; Gu, 2014; Wang et al., 2022a) came up with theory-driven NAL algorithms. These studies reformulate the problem into a subset selection problem or multi-armed bandit problem, and then utilize theoretical analysis techniques to guarantee the test performance. However, the internal mechanism remains not well understood on why the two widely used query criteria in the NAL family work so well, which naturally leads us to the following questions.

Essential Questions

1. What is the theoretical rationale behind the success of the two query criteria-based NAL algorithms, namely uncertainty-based and diversity-based?
2. Whether and how do the two query criteria of NAL connect to each other intrinsically?

1.1. Our Contribution

To answer the above questions, in this work, we delve into the **feature learning dynamics** of NAL algorithms. To start with, we draw inspiration from the data models in Zou et al. (2023a); Allen-Zhu and Li (2023); Lu et al. (2023) that consist of multiple task-related feature patches and noise

[1]Department of Computer Science, City University of Hong Kong, Hong Kong SAR [2]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan [3]Department of Mathematical Informatics, the University of Tokyo, Tokyo, Japan [4]Mohamed bin Zayed University of Artificial Intelligence, Masdar, United Arab Emirates. Correspondence to: Wei Huang < wei.huang.vr@riken.jp>, Hau-San Wong < cshswong@cityu.edu.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

patches with varying strengths and frequencies, similar to what is observed in real-world imbalanced datasets, and conjecture that successful NAL algorithms are able to ensure adequate learning of all types of task-related features.

In this spirit, we adopt a multi-view feature-noise data model that comprises two main components: i) easy-to-learn (i.e., strong & common) features or hard-to-learn (i.e., weak & rare) features, and ii) noise. In Figure 1, the easy-to-learn features are exemplified by the frontal male lions with brown fur in the first row, given their common and easily identifiable lion traits, while lions in all the other rows can be characterized as the hard-to-learn features since they exhibit distinctive poses, colors, ages, races, fur patterns, and even heterogeneity. Hard-to-learn features are less common in the dataset and correspond to weakly recognizable lion traits, compared to the easy-to-learn features.

Figure 1. Lions in real-world dataset.

Under our data model, we reformulate two representative NAL algorithms, i.e., Uncertainty Sampling and Diversity Sampling, in a pool-based setting, corresponding to two query criteria, respectively. Both are built upon a two-layer ReLU convolutional neural network, and trained by gradient descent. In accordance with the principle of each approach family (Takezoe et al., 2023), the proposed Uncertainty Sampling queries based on the lowest confidence (Lewis and Catlett, 1994), and Diversity Sampling queries based on the largest distance between feature representations of unlabeled samples in the pool and those of labeled data (Sener and Savarese, 2018).

Over our data and algorithm models, our theory sheds light on the benefits of the two primary query criteria in the NAL family. Surprisingly, our analysis unveils that the success of both criteria-based NAL stems from their inherent shared principle, leading to a unified view. Specifically, we make the following contributions in this work.

- We offer valuable insights that from a feature learning view, the two query criteria-based NAL can be **unified** as one family. We provably show that the two query criteria-based NAL share the same working principle, i.e., prioritizing **perplexing samples**-samples with yet-to-be-learned features. Our analysis reveals that in our scenario, those yet-to-be-learned features are actually those weak & rare features.
- We elucidate a marked disparity in the generalization capabilities between passive learning and NAL algorithms. Our analysis suggest that, both NAL algorithms can learn weak & rare features adequately via prioritizing **perplexing samples**, and thus achieve a small test error. Contrastingly, the strategy-free passive learning exhibits a large test error. The disparity can be intensified in some out-of-distribution cases. Our experimental study corroborates this finding.
- We further uncover why and to what extent the two query criteria can alleviate labelling effort. The key lies in NAL's ability to effectively query **perplexing samples** in the training distribution. But in contrast, we find that the strategy-free passive learning requires a significantly larger label complexity to adequately learn all types of features.

Perplexing Samples

Samples in the sampling pool that possess yet-to-be-learned features. We prove that both Uncertainty Sampling and Diversity Sampling inherently strive to query them.

1.2. Related Work

Neural Active Learning. Neural Network-based Active Learning (NAL) is one of the core data selection automation techniques in the field of Data-centric approaches for AutoML and Computer Version. As summarized in recent surveys (Zhan et al., 2021; 2022; Takezoe et al., 2023), there are two main query criteria: uncertainty-based, which chooses samples that the neural models feel most uncertain about (Seung et al., 1992; Lewis and Catlett, 1994; Roth and Small, 2006; Joshi et al., 2009; Houlsby et al., 2011; Cai et al., 2013; Yang and Loog, 2016; Kampffmeyer et al., 2016; Gal et al., 2017; Wang et al., 2022b; Kye et al., 2023; Duan et al., 2024; Cho et al., 2024) and diversity(representative)-based that selects samples that diverse from labeled set in the feature space (Stark et al., 2015; Du et al., 2015; Wang et al., 2016; Sener and Savarese, 2018; Gissin and Shalev-Shwartz, 2019; Sinha et al., 2019; Shui et al., 2020). Also, many works combine the two query criteria into the sampling (querying) strategy through weighted-sum optimization (Yin et al., 2017) or two-stage optimization (Ash et al., 2020; Zhdanov, 2019; Shui et al., 2020). In addi-

tion, to develop reliable algorithms, several design methods with theoretical guarantees, including theories such as VC bound (Balcan et al., 2006; Zhu and Nowak, 2022), Logistic Bound (Gu et al., 2014), Rademacher Complexity (Gu, 2014; Shui et al., 2020), and Neural Tangent Kernel (Wang et al., 2021; Mohamadi et al., 2022; Kong et al., 2022; Wang et al., 2022a; Wen et al., 2023). However, despite the development of numerous effective and theory-justified algorithms, the existing studies have not yet offered a comprehensive explanation for the underlying mechanisms of the two query criteria widely applied in NAL. Largely different from prior work, our work pioneeringly explore the theoretical aspect of the two criteria, via studying the **feature learning dynamic** in NAL algorithms.

Feature Learning in Learning Theory. Recent years witness an extensive body of research in learning theory on structured data from the perspective of feature learning (Li and Liang, 2018; Karp et al., 2021; Allen-Zhu and Li, 2023; Chen et al., 2022; 2023a;b;c;d; Zou et al., 2023b; Li et al., 2023; Kou et al., 2023a;c; Huang et al., 2023a;c; Chidambaram et al., 2023; Deng et al., 2023). The essence of this line-of-research is to explicitly study the learning progress of features and memorization degree of noise under different data and algorithm scenarios, which serves as an intermediate proxy to examine the convergence of training and 0-1 loss. Specifically, Cao et al. (2022a) demonstrate the occurrence of *benign overfitting* in Convolutional Neural Network over linearly separable data under distinct conditions. Subsequently, Kou et al. (2023b) conduct similar results with ReLU activation, Meng et al. (2023) further derive results over XOR data, Zou et al. (2023a) reveal the benefits of Mixup training over linearly separable data with common and rare features, and Lu et al. (2023) explore the phenomenon of *benign oscillation* over linearly separable data with common & weak and rare & strong features. Our work extends the line of research by investigating the rationale behind the two primary criteria in NAL family, over both linearly and non-linearly separable data scenarios that include common & strong and rare & weak features. Our study focuses on characterizing the **feature learning dynamics** in NAL algorithms and providing a mathematical explanation for the benefits and inner relationship of the two primary query criteria of NAL.

2. Problem Settings

Notations. For l_p norm we utilize $\|\cdot\|_p$ to denote its computation. Considering two series a_n and b_n, we denote $a_n = O(b_n)$ if there exists positive constant $C > 0$ and $N > 0$ such that for all $n \geq N$, $|a_n| \leq C|b_n|$. Similarly, we denote $a_n = \Omega(b_n)$ if $b_n = O(a_n)$ holds, $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ both hold, $c_n = O(a_n, b_n)$ if $c_n = O(\min\{a_n, b_n\})$ holds and $c_n = \Omega(a_n, b_n)$ if $c_n = \Omega(\max\{a_n, b_n\})$ holds. To omit logarithmic terms, we apply the notations $\widetilde{O}(\cdot), \widetilde{\Omega}(\cdot)$, and $\widetilde{\Theta}(\cdot)$. Our $\mathbb{1}(\cdot)$ is to denote the indicator variable of an event. We say $y = \text{poly}(a_1, \ldots, a_k)$ if $y = O\left(\max\{a_1, \ldots, a_k\}^D\right)$ for some $D > 0$, and $b = \text{polylog}(a)$ if $b = \text{poly}(\log(a))$.

2.1. Data Distribution

In this study, our focus is on the pool-based selective sampling scenario, where the algorithms initially train the model using an initial labeled set and subsequently query a single batch of unlabeled samples from a large sampling pool. Then the algorithms would retrain the model again with fresh initialization. We denote the size of the initial labeled set as n_0, the querying (sampling) size for all querying algorithms as n^* ($n^* = \Omega(n_0) > n_0$), and the size of the labeled set after querying as $n_1 = n_0 + n^*$. We also define $\widetilde{n}$ as the maximum size of the labeled set after querying, such that $n_1 \leq \widetilde{n}$. Moreover, we have the initial labeled set represented as $\mathcal{D}_{n_0} := \{\mathbf{x}^{(i)}\}_{i=1}^{n_0}$, and the sampling pool denoted as $\mathcal{P}$. Both of them are synthesized from the same data distribution $\mathcal{D}$, which is specified as follows.

Definition 2.1. Let $\boldsymbol{\mu}_1 \perp \boldsymbol{\mu}_2 \in \mathbb{R}^d$ be two fixed feature vectors. Each data point $(\mathbf{x}, y)$, where $\mathbf{x}$ contains two patches as $\mathbf{x}=[\mathbf{x}_1^T, \mathbf{x}_2^T]^T \in \mathbb{R}^{2d}$ and $y \in \{-1, 1\}$ are generated from the distribution $\mathcal{D}$:

- The ground truth label y is synthesized from a Rademacher distribution.
- **Noise Patch.** One patch of $\mathbf{x}$ is selected as a noise patch $\boldsymbol{\xi}$, synthesized from Gaussian distribution $N(\mathbf{0}, \sigma_p^2 \cdot \mathbf{I})$.
- **Feature Patch.** For a feeble p satisfying $p < 0.5$, the remaining patch of $\mathbf{x}$ is selected as label-related feature patch, and with high probability (1-p) the feature patch is a strong feature $y \cdot \boldsymbol{\mu}_1$, while only with probability p the feature patch is a weak feature $y \cdot \boldsymbol{\mu}_2$.

We assume the following about the feature norms: [1]: $\forall l \in \{1, 2\}, \|\boldsymbol{\mu}_l\|_2^2 = \Omega(\sigma_p^2 \log(n_0/\delta), \widetilde{n}^{-1} d\sigma_p^4)$, $\|\boldsymbol{\mu}_1\|_2^4 = \Omega(\sigma_p^4 d n_0^{-1})$ and $\|\boldsymbol{\mu}_2\|_2^4 = O(\sigma_p^4 d n_0^{-1})$.

This feature-noise data model captures the structure of an image, as depicted in Figure 1, by incorporating task-oriented distinctive patterns (features) and background patterns (noise) with different frequencies and strengths. Same

[1]The choices of $\|\boldsymbol{\mu}_l\|$ aim to prevent learning of features completely disrupted by noise, while amplifying the distinguishability of the strong feature patch compared to the weak one. Our theory allows for a broader range of parameter settings (see Appendix D.3 for general cases), but for the sake of simplicity in presentation, we here chose a feasible one.

as the patches setting in Zou et al. (2023a); Allen-Zhu and Li (2023); Lu et al. (2023), the weak feature patches are orthogonal to the strong feature patches in our setting, which is reasonable since the rare features appear largely different to the common ones. Worth noting that this type of data setting is common in the widely-recognized feature learning line-of-research (Allen-Zhu and Li, 2023; Cao et al., 2022a; Kou et al., 2023b; Zou et al., 2023a; Meng et al., 2023). Allen-Zhu and Li (2023) justify this type of data settings as plausible theoretical setups by highlighting the common occurrence of multiple one-task-oriented features in the latent space of Resnet, as shown in their Figure 2-4, 9. Furthermore, recent empirical and theoretical studies indicate the orthogonal nature of different features within the latent space of ViT and LLM (Yamagiwa et al., 2023; Jiang et al., 2024). To extend our contributions to more practical scenarios, we also conduct rigorous study and draw similar theoretical findings over a non-linearly separable, non-orthogonal data distribution - the XOR data defined in Definition C.2 - and obtained similar results.

2.2. Querying Algorithms

Neural Setting. This work considers a two-layer ReLU CNN adopted in Kou et al. (2023b); Meng et al. (2023); Kou et al. (2023c); Chen et al. (2023d) as the base neural network for querying algorithms. The CNN function $f(\mathbf{W}, \mathbf{x})$ is defined as $\sum_{j=\pm 1} j \cdot F_j(\mathbf{W}, \mathbf{x})$, with $F_j(\mathbf{W}, \mathbf{x})$ as

$$F_j(\mathbf{W}, \mathbf{x}) = \frac{1}{m} \sum_{r=1}^{m} \left[\sigma\left(\langle \mathbf{w}_{j,r}, y \cdot \boldsymbol{\mu} \rangle\right) + \sigma\left(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi} \rangle\right)\right].$$

where the second layer is fixed as $\pm 1/m$, m is the number of neurons, $\sigma(z) = \max\{z, 0\}$ is ReLU function, $\mathbf{w}_{j,r} \in \mathbb{R}^d$ denotes the weights of the r-th neuron of F_j, $\mathbf{W}_j \in \mathbb{R}^{m \times d}$ collects the weights in F_j and $\mathbf{W}$ collects all weights.

Training Setting. We utilize gradient descent to train the neural model. Denote n as the size of current labeled training set, denoted as $\mathcal{D} = \left\{\left(\mathbf{x}^{(i)}, y_i\right)\right\}_{i=1}^{n}$ generated from $\mathcal{D}$ over $\mathbf{x} \times y$. We apply the empirical logist loss:

$$L_S(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left[y_i \cdot f\left(\mathbf{W}, \mathbf{x}^{(i)}\right)\right], \tag{1}$$

where $\ell(z) = \log(1 + \exp(-z))$. The gradient update of the filters in the first layer can be written as follows:

$$\begin{aligned} \mathbf{w}_{j,r}^{(t+1)} &= \mathbf{w}_{j,r}^{(t)} - \eta \cdot \nabla_{\mathbf{w}_{j,r}} L_S\left(\mathbf{W}^{(t)}\right) \\ &= \mathbf{w}_{j,r}^{(t)} - \frac{\eta}{nm} \sum_{i=1}^{n} \ell_i'^{(t)} \cdot \sigma'\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle\right) \cdot j y_i \boldsymbol{\xi}_i \\ &- \frac{\eta}{nm} \sum_{l=1}^{2} \sum_{i \in U^l} \ell_i^{(t)} \cdot \sigma'^{(t)}\left(\left\langle \mathbf{w}_{j,r}^{(t)}, y_i \boldsymbol{\mu}_l \right\rangle\right) \cdot j \boldsymbol{\mu}_l, \end{aligned} \tag{2}$$

where $U^l = \{\mathbf{x} \in \mathcal{D} \mid \mathbf{x}_{\text{signal part}} = \boldsymbol{\mu}_l\}$ denote as the set of indices of $\mathcal{D}$ where the data's feature patch is $\boldsymbol{\mu}_l$, $\ell_i'^{(t)}$ denotes $\ell'\left[y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}^{(i)})\right]$. The initial values of all elements in $\mathbf{W}^{(0)}$ are generated from independent and identically distributed (i.i.d.) Gaussian distributions with mean 0 and variance σ_0^2. The querying algorithms would have the neural models retrained after a single querying with the same model initialization.

Querying Setting. During the querying stage, all the querying algorithms select n^* new unlabeled samples from $\mathcal{P}$, where the pool size $|\mathcal{P}|$ satisfies $|\mathcal{P}| = \Omega(p^{-1}\sigma_p^4 d \|\boldsymbol{\mu}_2\|_2^{-4}, p^{-1}\log(1/\delta))$[2]. The three querying algorithms differentiate from each other by their own sampling rules as below:

- **Random Sampling** (strategy-free passive learning) randomly selects n^* new samples from $\mathcal{P}$.
- **Uncertainty Sampling** (uncertainty-based NAL) selects top n^* new samples from $\mathcal{P}$ based on the lowest Confidence Score at time step t. The Confidence Score $C(\mathbf{W}, \mathbf{x})$ measures the model's confidence in predicting the label of sample $\mathbf{x}$, defined as below:

$$C(\mathbf{W}, \mathbf{x}) = \max\left\{\frac{1}{1 + \exp(-y \cdot f(\mathbf{W}, \mathbf{x}))}, 1 - \frac{1}{1 + \exp(-y \cdot f(\mathbf{W}, \mathbf{x}))}\right\},$$

  which represents the probability of the predicted label y of logistic loss. In our scenario, the proposed Uncertainty Sampling is actually equivalent to many well-known uncertainty-based approaches such as Least Confidence (Lewis and Catlett, 1994), Margin Roth and Small (2006), and Entropy methods (Joshi et al., 2009), as discussed in Lemma F.5 in Appendix F.2.
- **Diversity Sampling** (diversity-based NAL) selects the top n^* new samples from $\mathcal{P}$ based on the highest Feature Distance at time step t. The Feature Distance $D(\mathbf{W}, \mathbf{x} \mid \mathcal{D}_{n_0})$ measures the l_p distance between sample $\mathbf{x}$ and $\mathcal{D}_{n_0}$ in feature space, specified as:

$$D(\mathbf{W}, \mathbf{x} \mid \mathcal{D}_{n_0}) = \|\mathbf{Z}(\mathbf{x}, t) - \mathop{\mathbb{E}}_{\mathbf{x}^{(i)} \in \mathcal{D}_{n_0}} \mathbf{Z}(\mathbf{x}^{(i)}, t)\|_p,$$

  where the $\mathbf{Z}(\mathbf{x}, t)$ is defined as the sum of feature maps in the feature space of CNN:

$$\mathbf{Z}(\mathbf{x}, t) = \sum_j (\sigma(\langle \mathbf{W}_j^{(t)}, \mathbf{x}_1 \rangle)) + \sigma(\langle \mathbf{W}_j^{(t)}, \mathbf{x}_2 \rangle)).$$

  Specifically, Lemma 4.2 reveals that in our scenario, the proposed Diversity Sampling is equivalent for all

[2]The choice on $|\mathcal{P}|$ is to ensure the sufficient presence of weak features in $\mathcal{P}$.

Algorithm 1 Querying Algorithms

Require: Neural Network $f(\cdot;\cdot)$, initial labeled set $\mathcal{D}_{n_0} := \{\mathbf{x}^{(i)}\}_{i=1}^{n_0} \subseteq \mathcal{D}$, sampling pool $\mathcal{P} \subseteq \mathcal{D}$, test distribution $\mathcal{D}^*$, sample size $n^* = \widetilde{n} - n_0$, σ_0, T

Initialize Neural Network $f(\mathbf{W}^{(0)};\cdot)$
for $t \leftarrow 1$ to T **do**
  Train Neural Network over $\mathcal{D}_{n_0}$ by $L_S(\mathbf{W})$
end for
Querying: Sample n^* new samples from $\mathcal{P}$ based on particular rules. New samples $\mathcal{D}_{n^*}$ are labeled by oracle and included to the new labeled set $\mathcal{D}_{n_1} := \mathcal{D}_{n_0} \cup \mathcal{D}_{n^*}$
Initialize Neural Network $f(\mathbf{W}^{(0)};\cdot)$
for $t \leftarrow 1$ to T **do**
  Train Neural Network over $\mathcal{D}_{n_1}$ by $L_S(\mathbf{W})$
end for
Test performance of Neural Network $f(\mathbf{W}^{(T)};\cdot)$ over $\mathcal{D}^*$ and obtain $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(T)}\right)$
return $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(T)}\right)$

values of p within the range of $[1, \infty)$. This implies that our metric can be various distance measure, including Euclidean, Manhattan, or Minkowski distance.

The newly acquired samples are provided to an oracle to obtain their ground truth labels, which are then added to the training set. The whole procedure of the three querying algorithms are shown in Algorithm 1.

Testing Setting. The model performances at initial stage (before querying) and stage after querying are all measured by test error on a test distribution $\mathcal{D}^*$:

$$L_{\mathcal{D}^*}^{0-1}(\mathbf{W}) := \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*}[y \cdot f(\mathbf{W}, \mathbf{x}) < 0]. \tag{3}$$

It is important to note that $\mathcal{D}^*$ shares the same definition as stated in Definition 2.1. However, it can have any occurrence probability of the weak feature, denoted as p^*, without the limitation of $p^* < 0.5$ compared to the training distribution. Also, the test loss is defined as :

$$L_{\mathcal{D}^*}(\mathbf{W}) := \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}^*} \ell[y \cdot f(\mathbf{W}, \mathbf{x})].$$

3. Theoretical Results

For both the initialization stage and the second stage, we consider the learning period $0 \le t \le T^*$, where $T^* = \eta^{-1}$ poly $\left(\varepsilon^{-1}, d, n_0, m\right) \ge \widetilde{\Omega}\left(\eta^{-1}\varepsilon^{-1}mn_0d^{-1}\sigma_p^{-2}\right)$ is the maximum admissible iterations for the initial stage. The following provides our main theories over linearly separable data. For non-linear XOR data, please refer to our similar theoretical results in Appendix C.

We first adopt *signal-noise decomposition* techniques in Cao et al. (2022a) over $\mathbf{w}_{j,r}^{(t)}$. By the update rule in (2), we can derive that there exist unique coefficients $\gamma_{j,r,l}^{(t)}$ and $\rho_{j,r,i}^{(t)}$ such that

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \sum_{l=1}^{2} \gamma_{j,r,l}^{(t)} \cdot \frac{\boldsymbol{\mu}_l}{\|\boldsymbol{\mu}_l\|_2^2} + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} \tag{4}$$

The normalization factors $\|\boldsymbol{\mu}_l\|_2^{-2}$ and $\|\boldsymbol{\xi}_i\|_2^{-2}$ leads to $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_l \rangle \approx \gamma_{j,r,l}^{(t)}, \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \approx \rho_{j,r,i}^{(t)}$. Importantly, $\gamma_{j,r,l}^{(t)}$ characterizes **the learning progress** of feature $\boldsymbol{\mu}_l$, and $\rho_{j,r,i}^{(t)}$ characterizes the degree of noise memorization. Geometrically, the $\gamma_{j,r,l}$ indicates how well the model filters integrate the low-dimensional patterns of the task-oriented features in its latent projection space, and $\rho_{j,r,i}$ quantifies the extent to which model filters memorize the high-dimensional complex noise. Then, by conducting a scale analysis of the two coefficients, we can then assess the cases where models mainly focus on capturing underlying patterns while avoiding excessive fitting of noise, which we refer to as *benign overfitting*. Additionally, this analysis helps us identify situations of *harmful overfitting*, where the models become overly complex, primarily memorizing noise and leading to poor generalization on new, unseen data.

Our findings then reveal that in our case, both the two heuristic NAL methods inherently amenable to query those data with yet-to-be-learned features (i.e., features that model exhibits low $\gamma_{j,r,l}$). Consequently, the NNs are enabled to sufficiently learn all types of features, and then exhibit *benign overfitting* even in the case where the label complexity is quite limited.

To present our findings, we make the following assumptions.

Condition 3.1. *Suppose that:*

1. *The initial training size n_0, the maximum admissible size after querying $\widetilde{n}$, and the width of neural network m satisfy $n_0 = \Omega(\log(m/\delta), p^{-1}\log(1/\delta))$, $\widetilde{n} = O(p^{-1}\sigma_p^4 d\|\boldsymbol{\mu}_2\|_2^{-4})$, $m = \Omega(\log(n_0/\delta))$.*
2. *Dimension d is sufficiently large: $\forall l \in \{1,2\}$, $d = \Omega(\widetilde{n}\sigma_p^{-2}\|\boldsymbol{\mu}_l\|_2^2 \log(T^*), \widetilde{n}^2 \log(\widetilde{n}m/\delta)(\log(T^*))^2)$.*
3. *The standard deviation of Gaussian initialization σ_0 is appropriately chosen such that $\forall l \in \{1,2\}$, $\sigma_0 = O(\|\boldsymbol{\mu}_l\|_2^{-1}(\log(m/\delta)^{-1/2}), \sigma_p^{-1}d^{-1}\widetilde{n}^{1/2})$. The learning rate of all algorithms η satisfies that $\eta = O(\sigma_p^{-2}d^{-1}\widetilde{n}, \sigma_p^{-2}d^{-3/2}\widetilde{n}^2 m(\log(\widetilde{n}/\delta))^{1/2})$.*

The condition on n_0 is to guarantee there exists enough strong features in the initial training set with probability at least $1 - O(e^{-n_0 p})$, while the condition on $\widetilde{n}$ prevents the final training size from being too large, even for passive learning to perform well with considerable chance. The

requirement on d ensures the problem is in a sufficiently overparameterized setting, as in prior works (Chatterji and Long, 2021; Cao et al., 2022a; Frei et al., 2022; Kou et al., 2023b; Lu et al., 2023; Chidambaram et al., 2023). The conditions on σ_0 and η guarantee that gradient descent can effectively minimize the empirical loss. A detailed discussions over parameter settings are provided in Appendix B.

The following results illustrate the presence of *benign overfitting* (i.e., small training loss and small test error) as well as *harmful overfitting* (i.e., small training loss but large test error) in the three querying algorithms.

Proposition 3.2. ***(Before Querying)*** *At the initial stage before querying, $\forall \varepsilon > 0$, under Condition 3.1, with probability at least $1-\delta$, there exists $t = \widetilde{O}\left(\eta^{-1}\varepsilon^{-1}mn_0d^{-1}\sigma_p^{-2}\right)$, the followings hold for all of the three querying algorithms:*

1. *The training loss converges to ε, i.e., $L_S\left(\mathbf{W}^{(t)}\right) \leq \varepsilon$.*
2. *The test error remains at constant level, i.e., $L^{0-1}_{\mathcal{D}^*}\left(\mathbf{W}^{(t)}\right) = \Theta(1) \geq 0.12 \cdot p^*$.*

Proposition 3.2 outlines the scenarios of *harmful overfitting* for all algorithms at the initial stage, which is not a surprise since the initial size n_0 is limited and always insufficient for adequate learning. Subsequently, the following lemma uncovers a crucial finding regarding the querying stage.

Proposition 3.3. ***(Querying Stage)*** *During Querying, under the same conditions as Proposition 3.2, if[3] $\|\boldsymbol{\mu}_1\|_2^2 - \|\boldsymbol{\mu}_2\|_2^2 = \Omega({\sigma_p}^2(dn_0^{-1}\log(m/\delta'))^{1/2})$, with probability at least $1-\Theta(\delta+\delta')$, both Uncertainty Sampling and Diversity Sampling pick n^* samples that exhibit lowest $\mathop{\mathbb{E}}_{j,r}\gamma_{j,r,l}^{(t)}$.*

Proposition 3.3 provides a unifying insight that both NAL algorithms prioritize **perplexing samples**-samples that exhibit a lack of learning progress (measured by $\mathop{\mathbb{E}}_{j,r}\gamma_{j,r,l}^{(t)}$). Lemma 4.1 indicates that these **perplexing samples** here are essentially samples that contain weak & rare features. We discuss the nature of these **perplexing samples** in general cases in Appendix D.3. Our inference process for the following theorem reveals that the ability to prioritize these samples is the main contributor to the success of both NAL algorithms.

Theorem 3.4. ***(After Querying)*** *If the sampling size n^* of the three querying algorithms satisfies $C_1\sigma_p^4 d\|\boldsymbol{\mu}_2\|_2^{-4} - pn_0/2 \leq n^* = \Theta(\widetilde{n}-n_0) \leq \widetilde{n}-n_0$, where C_1 is some positive constant. Then for $\forall\varepsilon > 0$, under the same conditions as Proposition 3.3, with probability more than 1 - $\Theta(\delta+\delta')$, $\exists t = \widetilde{O}\left(\eta^{-1}\varepsilon^{-1}m(n_0+n^*)d^{-1}\sigma_p^{-2}\right)$ such that:*

- *For all of the three querying algorithms, the training loss converges to ε, i.e., $L_S\left(\mathbf{W}^{(t)}\right) \leq \varepsilon$.*
- ***Uncertainty Sampling** and **Diversity Sampling** algorithms have small test error: $L^{0-1}_{\mathcal{D}^*}\left(\mathbf{W}^{(t)}\right) \leq \exp(\Theta\left(\frac{-\widetilde{n}\|\boldsymbol{\mu}_l\|_2^4}{\sigma_p^4 d}\right)), l \in \{1,2\}$.*
- ***Random Sampling** algorithm would remain constant order test error: $L^{0-1}_{\mathcal{D}^*}\left(\mathbf{W}^{(t)}\right) = \Theta(1) \geq 0.12 \cdot p^*$.*

Theorem 3.4 implies that NAL algorithms achieve *benign overfitting*, whereas the passive learning remains *harmful overfitting*. It worth noting that as p^* increases, the test error of Random Sampling tends to explode, especially in out-of-distribution scenarios where $p^* > 0.5 > p$. In contrast, Uncertainty Sampling and Diversity Sampling consistently achieve low test errors regardless of the value of p^*, which highlights the superiority of Uncertainty Sampling and Diversity Sampling over Random Sampling.

Given that strategy-free passive learning can also adequately learn all types of features with ample data, the following corollary aim to show the extent to which NAL algorithms alleviate the burden of labeling.

Corollary 3.5. ***(Label Complexity)*** *Under the same conditions as stated in Theorem 3.4, with a probability of at least $1-\Theta(\delta+\delta')$, we observe distinct label complexities for strategy-free passive learning and NAL algorithms in achieving Bayes-optimal generalization ability:*

- *For a fully trained neural model, the label complexity n_{CNN} requires $\Omega(p^{-1}\sigma_p^2 d\|\boldsymbol{\mu}_2\|_2^{-4})$.*
- *For two NAL algorithms, the maximum label complexity $\widetilde{n}$ only requires $\Omega(\sigma_p^2 d\|\boldsymbol{\mu}_2\|_2^{-4})$.*

This corollary suggests that NAL algorithms can significantly reduce labeling effort, approximately on the order of $\Theta(p^{-1})$. This holds true even without the requirement of disparity between feature norms, as demonstrated in Appendix D.3. Hence, we can conclude that NAL algorithms are effective in minimizing labeling effort, particularly in imbalanced data scenarios where the degree of discrimination or rarity varies within the data. In collaboration with Proposition 3.3 and Theorem 3.4, the essence lies in NAL's capability to effectively grasp yet-to-be-learned features.

4. Proof Sketch

In this section, we provide an overview of the proof outlines for our theory over linearly separable data. Here we denote n as the number of training data in current labeled set, which is n_0 at initial stage and n_1 after sampling (querying). For

[3] We can relax the requirement for the discrepancy of feature norms, as discussed in Appendix D.3. The specific choice made in our presentation was for the sake of simplicity and clarity.

$s \in \{1,2\}, l \in \{1,2\}$, the notations of $n_{s,l}$ represent the number of feature $\boldsymbol{\mu}_l$ at the initial stage $s = 1$ and stage after querying $s = 2$. And for notation simplicity we denote τ_1 and τ_2 as the proportion of data with strong and weak feature in current dataset.

Here are the main challenges we faced and the techniques we used to address them:

- The synthesis of $\mathcal{D}_{n_0}$, $\mathcal{P}$, and the final labeled set obtained through Random Sampling require sequential martingale-type subset generations from distribution $\mathcal{D}$, which poses a big challenge to our analysis. Our solution was to treat the results as independent binomial random variables, which allow us to conduct a reliable analysis with high-probability results by leveraging the properties of binomial tails.
- During querying, NAL algorithms need to query the samples with the lowest Confidence Score or the highest Feature Distance from the entire sampling pool $\mathcal{P}$. This involves $|\mathcal{P}|(|\mathcal{P}|-1)/2$ comparison operations. To better scrutinize the sampling dynamics, we defined two full orders and conducted an order-dependent querying analysis to examine the high probability events via combinatorial analysis.
- Depicting the generalization capability of three different querying algorithms along the whole process was a big challenge. We addressed this by proposing a label complexity-based test error analysis regime, which allowed us to incorporate different scenarios into a single inferential process.

4.1. Feature Learning and Noise Memorization Analysis

Leverage the inductive techniques adopted in many works (Cao et al., 2022a; Kou et al., 2023b; Meng et al., 2023; Kou et al., 2023c; Chen et al., 2023d), we can in our case study the coefficient scales.

Lemma 4.1. *Under Condition 3.1, there exists $T_1 = \Theta(\eta^{-1} n m \sigma_p^2 d^{-1})$, for $t \in [T_1, T^*]$ we have the following hold for $\forall j \in \{\pm 1\}, r \in [m]$ and $l \in \{1,2\}$:*

- $\sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \mathbb{1}(\rho_{j,r,i}^{(t)} > 0) = \Omega(n)$,
- $\gamma_{j,r,l}^{(t)} = \Theta\left(\tau_l n \cdot \sigma_p^{-2} d^{-1} \|\boldsymbol{\mu}_l\|_2^2\right)$.

It is evident that there is a noticeable disparity in the learning efficiency of features, as $\rho_{j,r,i}^{(t)}$ is directly proportional to both the data proportion τ_l and the feature norms $\|\boldsymbol{\mu}_l\|_2$. Furthermore, according to Lemma G.3, we can model the data synthesis from $\mathcal{D}$ as a binomial variable. This allows effective control over the probability tails, resulting in $\tau_2 = \Theta(p)$ and $\tau_1 = \Theta(1-p)$. Thus, we can conclude that the **perplexing samples** are actually those $\boldsymbol{\mu}_2$-equipped samples. Subsequently, we can now examine the querying stage closely.

4.2. Order-dependent Sampling (Querying) Analysis

To rigorously analyze the statistics of the querying stage, we define two orders, namely Uncertainty Order $\preceq_C^{(t)}$ and Diversity Order $\preceq_D^{(t)}$. For $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{P}$, we have $\mathbf{x}' \preceq_C^{(t)} \mathbf{x}$ if $C\left(\mathbf{W}^{(t)}, \mathbf{x}'\right) \geq C\left(\mathbf{W}^{(t)}, \mathbf{x}\right)$, and $\mathbf{x}' \preceq_D^{(t)} \mathbf{x}$ if $D\left(\mathbf{W}^{(t)}, \mathbf{x}' \mid \mathcal{D}_n\right) \leq D\left(\mathbf{W}^{(t)}, \mathbf{x} \mid \mathcal{D}_n\right), \forall p \in [1, \infty)$. Specifically, if the Confidence Score of all elements in a set $\mathbf{X}$ at time step t are all less than those in the set $\mathbf{X}'$, we utilize the same notation to describe the Uncertainty Order between sets: $\mathbf{X} \preceq_C^{(t)} \mathbf{X}'$. Similarly, we also have set-level notation for $\preceq_D^{(t)}$. The detailed definitions are delayed to Appendix F.

The following lemma presents our important findings when examining the two orders of samples.

Lemma 4.2. *Under the same conditions in Proposition 3.3, for $\mathbf{x}, \mathbf{x}' \in \mathcal{P}$, denote $\boldsymbol{\mu}_{l_{\mathbf{x}}}, \boldsymbol{\mu}_{l_{\mathbf{x}'}}$ as the feature patches in $\mathbf{x}$ and $\mathbf{x}'$ separately, where $l_{\mathbf{x}}, l_{\mathbf{x}'} \in \{1,2\}$, it holds that*

- $\mathbf{x}' \preceq_C^{(t)} \mathbf{x}$ *has a sufficient event that*

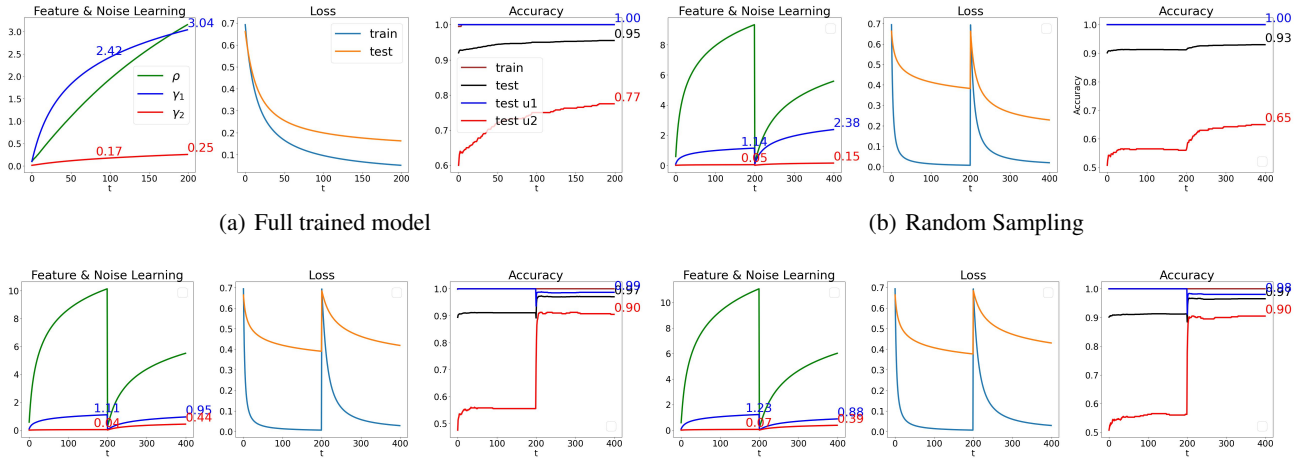
$$\{ \underbrace{\Theta(\mathop{\mathbb{E}}_{r}(\gamma_{y',r,l_{\mathbf{x}'}})) - \Theta(\mathop{\mathbb{E}}_{r}(\gamma_{y,r,l_{\mathbf{x}}}))}_{\text{Learning Progress Disparity: Feature in } \mathbf{x} \text{ vs. Feature in } \mathbf{x}'} > \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\}\}. \tag{5}$$

- $\mathbf{x}' \preceq_D^{(t)} \mathbf{x}$ *has a sufficient event that*

$$\{ \underbrace{|\Theta(\mathop{\mathbb{E}}_{r}(\gamma_{y,r,l_{\mathbf{x}}})) - \sum_l \tau_l \cdot \Theta(\mathop{\mathbb{E}}_{i_l \in U_0^l, r}(\gamma_{y_{i_l},r,l}))|}_{\text{Learning Progress Disparity: Feature in } \mathbf{x} \text{ vs. Features in Initial Set}} - \underbrace{|\Theta(\mathop{\mathbb{E}}_{r}(\gamma_{y',r,l_{\mathbf{x}'}})) - \sum_l \tau_l \cdot \Theta(\mathop{\mathbb{E}}_{i_l \in U_0^l, r}(\gamma_{y_{i_l},r,l}))|}_{\text{Learning Progress Disparity: Feature in } \mathbf{x}' \text{ vs. Features in Initial Set}} > \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\}\}, \tag{6}$$

where $U_0^l = \{\mathbf{x} \in \mathcal{D}_0 \mid \mathbf{x}_{\text{signal part}} = \boldsymbol{\mu}_l\}$.

Remark 4.3. This lemma demonstrate that Uncertainty Sampling holds the comparisons of the model's learning progress of features in $\mathcal{P}$, as shown in (5). On the other hand, Diversity Sampling cares the comparisons of the disparity between model's learning progress of samples and the labeled training set, as shown in (6).



(a) Full trained model

(b) Random Sampling

(c) Uncertainty Sampling

(d) Diversity Sampling

Figure 2. Learning/memorization progress of features and noise (γ_l represents $\max_{j,k} \gamma_{j,k,l}^{(t)}$, and ρ represents $\max_{j,k,i} \rho_{j,k,i}^{(t)}$, train/test losses, and test accuracy of the full-trained model and the three querying algorithms, with $T^* = 200$, $d = 2000$, $\|\boldsymbol{\mu}_1\| = 9$, $p = p^* = 0.2$, $\|\boldsymbol{\mu}_2\| = 3$, $n_{CNN} = 200$, $n_0 = 10$, $n^* = 30$ and $|\mathcal{P}| = 190$.

We note that (6) is irrelevant to the l_p distance measure metric (i.e., $\forall p \in [1, \infty)$). This is because we can eliminate the scaling term $m^{\frac{1}{p}}$ at two sides of the inequality when examining the probability lower bound (see more details in Appendix G.4). Based on Lemma 4.1, the event (5) and event (6) could be all simplified to the following shared sufficient event

$$\{\Theta(\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,l_{\mathbf{x}'}})) - \Theta(\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,l_{\mathbf{x}}})) > \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\}\}.$$

This implies that both the event $\{\mathbf{x}' \preceq_C^{(t)} \mathbf{x}\}$ and the event $\{\mathbf{x}' \preceq_D^{(t)} \mathbf{x}\}$ have a common occurrence where the model's learning of $\boldsymbol{\mu}_{l_{\mathbf{x}}}$ is considerably worse compared to its learning of $\boldsymbol{\mu}_{l_{\mathbf{x}'}}$. Based on this observation and Lemma 4.1, we can deduce the following lemma with some effort.

Lemma 4.4. *Under the same conditions as Proposition 3.3, denoting $\mathbf{X}_{\mathcal{P}}^1 \subsetneqq \mathcal{P}$ as the collection of all the data points with strong feature $\boldsymbol{\mu}_1$ in $\mathcal{P}$, and $\mathbf{X}_{\mathcal{P}}^2 \subsetneqq \mathcal{P}$ as the collection of data points with weak feature $\boldsymbol{\mu}_2$, we have the conclusion that with probability more than 1-$\Theta(\delta')$, $\mathbf{X}_{\mathcal{P}}^1 \preceq_C^{(t)} \mathbf{X}_{\mathcal{P}}^2$ and $\mathbf{X}_{\mathcal{P}}^1 \preceq_D^{(t)} \mathbf{X}_{\mathcal{P}}^2$ ($\forall p \in [1, \infty)$) both hold.*

This lemma directly implies the result in Proposition 3.3.

4.3. Label Complexity-based Test Error Analysis

To assess the generalization ability of all the three querying algorithms before and after querying, we establish a comprehensive analysis regime that examines the impact of label complexity for each type of feature on the test error, via a single inferential process. Specifically, We introduce the following lemma, employing a standard proving technique utilized in prior research (Chatterji and Long, 2021; Frei et al., 2022; Kou et al., 2023b; Meng et al., 2023).

Lemma 4.5. *Under Condition 3.1, $\forall \varepsilon > 0$, $\exists\ t = \widetilde{O}\left(\eta^{-1}\varepsilon^{-1}mn_0d^{-1}\sigma_p^{-2}\right)$, we have the following two situations before and after querying (i.e., $\forall s \in \{0, 1\}$) for three quering algorithms:*

- *The training loss converges to ε, i.e., $L_S\left(\mathbf{W}^{(t)}\right) \leq \varepsilon$.*
- *If $\forall l \in \{1, 2\}, n_{s,l} \geq C_1\sigma_p^4 d\|\boldsymbol{\mu}_l\|_2^{-4}$ holds, the test error achieves Bayes-optimal: $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \leq p_1^* \cdot \exp\left(\frac{-n_{s,1}\|\boldsymbol{\mu}_1\|_2^4}{C_3\sigma_p^4 d}\right) + p_2^* \cdot \exp\left(\frac{-n_{s,2}\|\boldsymbol{\mu}_2\|_2^4}{C_4\sigma_p^4 d}\right.$.*
- *If $\exists l' \in \{1, 2\}, n_{s,l'} \leq C_2\sigma_p^4 d\|\boldsymbol{\mu}_{l'}\|_2^{-4}$ holds, the test error stays constant-level: $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \geq 0.12 \cdot p_{l'}^*$.*

Here p_l^* denotes the occurrence probability of feature $\boldsymbol{\mu}_l$, C_1, C_2, C_3 and C_4 are some positive constants.

By Condition 3.1, along with the findings from Lemma 4.4 and Lemma 4.5, we can deduce that only the two NAL algorithms are able to obtain ample $\boldsymbol{\mu}_2$ for adequate learning after querying, which support the results in Proposition 3.2 and Theorem 3.4. Also, by Lemma G.3 and Lemma 4.5, Random Sampling necessitates a label complexity that is approximately $\Theta(p^{-1})$ times larger to sufficiently learn $\boldsymbol{\mu}_2$. This finding aligns with the conclusions in Corollary 3.5.

5. Experiments

In this section, we demonstrate the validity of our theoretical analysis through simulations. The experiments regarding the

theories of XOR data as well as other data settings are also conducted, please refer to Appendix E for further details.

Here we generate synthetic data exactly following Definition 2.1. Specifically, we let the dimensionality as $d = 2000$, and strengths of the strong and weak feature as $\|\boldsymbol{\mu}_1\|_2 = 9$ and $\|\boldsymbol{\mu}_2\|_2 = 3$, respectively. For the occurrence probability, we let $p = p^* = 0.2$. For size setting of data, we let the n_{CNN}=200, n_0=10, $n^* = 30$ and $\hat{n} = 40$, and set $|\mathcal{P}| = 190$. For model initialization, we let $\sigma_p = 1$ and $\sigma_0 = 0.01$. The parameters are initialized using the default method in PyTorch, and the models are trained using gradient descent with a learning rate of 0.1 for 200 iterations at the initial stage and the stage after sampling. All the data points are generated beforehand and shared by all the algorithms, thus the results are fairly comparable.

Figure 2 illustrates the effectiveness of both Uncertainty Sampling and Diversity Sampling in comparison to Random Sampling and full-trained ReLU CNN model with ample quantity of training samples. It's evident that the learning of weak & rare feature (quantified by γ_2) in hard-to-learn samples are significantly poorer than strong & common feature (quantified by γ_1) in easy-to-learn samples at the initial stage. After querying, we see explicitly that both the NAL algorithms learn the weak & rare feature well and achieve comparable test performance compared to full trained model after querying. In contrast, Random Sampling continues to exhibit limited learning progress of weak features and results in poor test accuracy. The results verify Proposition 3.2 and Theorem 3.4. Illustrations of the querying stage details are deferred to Appendix E.1.

6. Potential Extension and Implication for Practical NALs

In this section, we first explore the potential extensions of our findings to broader theoretical realm, then elaborate on the practical implications derived from our theories.

Potential Extension to Multi-round NALs. The intrinsic principle we uncovered underlying both NAL methods is not tied to the single-round setting, and a fine-grained analysis can be conducted on complex iterative processes, as discussed in Appendix D.5.

Potential Extension to Broader NALs: BADGE (Ash et al., 2020) as an Exemplar. The key idea behind BADGE is to prioritize samples exhibiting large and diverse gradients. Our analysis reveals that such samples in our context tend to have smaller-scale latent representations ($\gamma_{j,r,l}$ is smaller) or more diverse gradient directions (many diverging $\gamma_{j,r,l}$) due to the non-increasing nature of the logistic loss function. These characteristics align with the cases described in Lemma 4.2, which in our context refers to samples with lower $\gamma_{j,r,l}$ that correspond to yet-to-be-learned features. Therefore, BADGE is well-grounded in the principles uncovered by our theoretical analysis. A more detailed discussion is provided in Appendix D.2.

Potential Extension to Examine Criteria Preference. Our results of test error is based on the conditions that there is a clear learning progress disparity between different task-oriented features, under which we see that both NALs inherently favour samples with yet-to-be-learnt features. However, when this disparity does not hold prominently as dicussed in Appendix D.3.2, the behaviors of uncertainty-based and diversity-based sampling may diverge. For example, uncertainty sampling can more precisely prioritize samples with underexplored features when label budgets are not highly constrained. Conversely, diversity sampling may be preferred when label complexity is very limited, as it can enhance the model's ability to capture diverse low-dimensional patterns. This argument is consistent with the claim in recent survey (Zhan et al., 2021). Our theory also suggests that when the "easiness" of learning various task-oriented features is balanced, uniform random sampling may suffice, without clear advantages for NALs. Additionally, in scenarios of active fine-tuning where the task objective changes, the task-oriented representation could shift, reducing the effectiveness of NAL methods that leverage prior neural representations for sampling. In such cases, random sampling may already be a satisfactory choice. A refined discussion is in Appendix D.4.

Practical Lessons from Our Theoretical Results. Our theoretical analysis yields several important practical insights, as detailed in Appendix D.6. First, we find that NALs have the potential to surpass the performance of fully-trained neural networks. As corroborated by the results in Lu et al. (2023), NALs can more effectively balance the learning progress across features with different lengths. Additionally, our work suggests that techniques capable of capturing the meaningful orthogonal components of a NN's features or gradients, such as ICA (Yamagiwa et al., 2023), could help identify samples underrepresented in NN's latent space. State-of-the-art methods like BADGE (Ash et al., 2020) leverages this idea upon the gradient components.

7. Conclusion

In this work, we theoretically demystify and unify the primary query criteria-based NAL methods. We prove they inherently prioritize **perplexing samples** - those with yet-to-be-learned features. This ensures adequate learning of all feature types, underpinning their strong generalization with limited labeled data. Future work can extend our theory to other complex NAL scenarios, such as multi-model committee and stream-based sampling. Additionally, the potential extensions and implications discussed in Section 6 represent valuable directions for further fine-grained exploration.

Acknowledgements

We thank the anonymous reviewers for their instrumental comments. DB and HW are supported in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11206622). WH was partially supported by JSPS KAKENHI (24K20848). TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2115, JPMJCR2015).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Charu C Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S Yu Philip. Active learning: A survey. In *Data classification*, pages 599–634. Chapman and Hall/CRC, 2014.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 242–252. PMLR, 2019.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 37932–37946. Curran Associates, Inc., 2022.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine Learning*, volume 148, pages 65–72, 2006.

Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 51–60, 2013.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep ReLU networks. *arXiv preprint arkiv: 1902.01384*, 2019b.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arkiv: 1912.01198*, 2020.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 25237–25250, 2022a.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *arXiv preprint arkiv: 2104.13628*, 2022b.

Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

Jinghui Chen, Yuan Cao, and Quanquan Gu. Benign overfitting in adversarially robust linear classification. In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216, pages 313–323, 2023a.

Yilan Chen, Wei Huang, Lam Nguyen, and Tsui-Wei Weng. On the equivalence between neural network and support vector machine. In *Advances in Neural Information Processing Systems*, volume 34, pages 23478–23490. Curran Associates, Inc., 2021a.

Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 13363–13373. Curran Associates, Inc., 2020.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep ReLU networks? *arXiv preprint arkiv: 1911.12360*, 2021b.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23049–23062, 2022.

Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in CLIP. *arXiv preprint arkiv: 2310.00927*, 2023c.

Zixiang Chen, Junkai Zhang, Yiwen Kou, Xiangning Chen, Cho-Jui Hsieh, and Quanquan Gu. Why does sharpness-aware minimization generalize better than SGD? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023d.

Muthu Chidambaram, Xiang Wang, Chenwei Wu, and Rong Ge. Provably learning diverse features in multi-view data with midpoint mixup. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 5563–5599, 2023.

Seong Jin Cho, Gwangsu Kim, Junghyun Lee, Jinwoo Shin, and Chang D. Yoo. Querying easily flip-flopped samples for deep active learning. *arXiv preprint arkiv:2401.09787*, 2024.

Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arkiv: 1810.08693*, 2023.

Bo Du, Zengmao Wang, Lefei Zhang, Liangpei Zhang, Wei Liu, Jialie Shen, and Dacheng Tao. Exploring representativeness and informativeness for active learning. *IEEE transactions on cybernetics*, 47(1):14–26, 2015.

Ruxiao Duan, Brian Caffo, Harrison X. Bai, Haris I. Sair, and Craig Jones. Evidential uncertainty quantification: A variance-based perspective. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2132–2141, January 2024.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 2668–2703, 2022.

Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *Journal of Machine Learning Research*, 24(303):1–49, 2023.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1183–1192, 2017.

Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. *arXiv preprint arkiv: 1907.06347*, 2019.

Quanquan Gu. *online and Active Learning of Big networks: Theory and Algorithms*. Dissertation, University of Illinois at Urbana-Champaign, Urbana, IL, 09 2014.

Quanquan Gu, Tong Zhang, and Jiawei Han. Batch-mode active learning via error bound minimization. In *Uncertainty in Artificial Intelligence - Proceedings of the 30th Conference, UAI 2014*, pages 300–309, 2014.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arkiv: 1112.5745*, 2011.

Wei Huang, Weitao Du, Richard Yi Da Xu, and Chunrui Liu. Implicit bias of deep linear networks in the large learning rate phase. *arXiv preprint arkiv: 2011.12547*, 2020.

Wei Huang, Weitao Du, and Richard Yi Da Xu. On the neural tangent kernel of deep networks with orthogonal initialization. *arXiv preprint arkiv: 2004.05867*, 2021.

Wei Huang, Yayong Li, Weitao Du, Jie Yin, Richard Yi Da Xu, Ling Chen, and Miao Zhang. Towards deepening Graph neural networks: A gntk-based optimization perspective. *arXiv preprint arkiv: 2103.03113*, 2022.

Wei Huang, Yuan Cao, Haonan Wang, Xin Cao, and Taiji Suzuki. Graph neural networks provably benefit from structural information: A feature learning perspective. *arXiv preprint arkiv: 2306.13926*, 2023a.

Wei Huang, Chunrui Liu, Yilan Chen, Richard Yi Da Xu, Miao Zhang, and Tsui-Wei Weng. Analyzing deep PAC-Bayesian learning with neural tangent kernel: Convergence, analytic generalization bound, and efficient hyperparameter selection. *Transactions on Machine Learning Research*, 2023b. ISSN 2835-8856.

Wei Huang, Ye Shi, Zhongyi Cai, and Taiji Suzuki. Understanding convergence and generalization in federated learning through feature learning theory. In *The Twelfth International Conference on Learning Representations*, 2023c.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arkiv: 1806.07572*, 2020.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language lodels. *arXiv preprint arkiv: 2403.03867*, 2024.

Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, 2009.

Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional [n]eural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.

Stefani Karp, Ezra Winston, Yuanzhi Li, and Aarti Singh. local signal adaptivity: Provable feature learning in neural networks beyond kernels. In *Advances in Neural Information Processing Systems*, volume 34, pages 24883–24897, 2021.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. *arXiv preprint arkiv: 2402.01258*, 2024.

Juno Kim, Kakei Yamamoto, Kazusato Oko, Zhuoran Yang, and Taiji Suzuki. Symmetric mean-field langevin dynamics for distributional minimax problems. In *The Twelfth International Conference on Learning Representations*, 2024.

Seo Taek Kong, Soomin Jeon, Dongbin Na, Jaewon Lee, Hong-Seok Lee, and Kyu-Hwan Jung. A neural preconditioning active learning algorithm to reduce label complexity. In *Advances in Neural Information Processing Systems*, volume 35, pages 32842–32853, 2022.

Yiwen Kou, Zixiang Chen, Yuan Cao, and Quanquan Gu. How does semi-supervised learning with pseudo-labelers work? a case study. In *The Eleventh International Conference on Learning Representations*, 2023a.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer ReLU convolutional neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 17615–17659, 2023b.

Yiwen Kou, Zixiang Chen, and Quanquan Gu. Implicit bias of gradient descent for two-layer ReLU and leaky ReLU networks on nearly-orthogonal data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023c.

Seong Min Kye, Kwanghee Choi, Hyeongmin Byun, and Buru Chang. TiDAL: Learning training dynamics for active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22335–22345, October 2023.

David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning Proceedings 1994*, pages 148–156. Morgan Kaufmann, 1994.

Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *The Eleventh International Conference on Learning Representations*, 2023.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Miao Lu, Beining Wu, Xiaodong Yang, and Difan Zou. Benign oscillation of stochastic gradient descent with large learning rate. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 2388–2464. PMLR, 25–28 Jun 2019.

Xuran Meng, Difan Zou, and Yuan Cao. Benign overfitting in two-layer ReLU convolutional neural networks for XOR data. *arXiv preprint arkiv: 2310.01975*, 2023.

Mohamad Amin Mohamadi, Wonho Bae, and Danica J. Sutherland. Making look-ahead active learning strategies feasible with neural tangent kernels. In *Advances in Neural Information Processing Systems*, volume 35, pages 12542–12553, 2022.

Atsushi Nitanda. Improved particle approximation error for mean field neural networks. *arXiv preprint arXiv:2405.15767*, 2024.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Particle dual averaging: optimization of mean field neural networks with global convergence rate analysis. In *Advances in Neural Information Processing Systems*, volume 34, pages 19608–19621, 2021.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 9741–9757. PMLR, 2022.

Atsushi Nitanda, Kazusato Oko, Denny Wu, Nobuhito Takenouchi, and Taiji Suzuki. Primal and dual analysis of entropic fictitious play for finite-sum problems. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 26266–26282. PMLR, 2023a.

Atsushi Nitanda, Kazusato Oko, Denny Wu, Nobuhito Takenouchi, and Taiji Suzuki. Primal and dual analysis of entropic fictitious play for finite-sum problems. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 26266–26282. PMLR, 2023b.

Atsushi Nitanda, Kazusato Oko, Taiji Suzuki, and Denny Wu. Improved statistical and computational complexity of the mean-field langevin dynamics under structured data. In *The Twelfth International Conference on Learning Representations*, 2024.

Kazusato Oko, Taiji Suzuki, Atsushi Nitanda, and Denny Wu. Particle stochastic dual coordinate ascent: Exponential convergent algorithm for mean field neural network optimization. In *International Conference on Learning Representations*, 2022.

Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *Machine Learning: ECML 2006*, pages 413–424, 2006.

Grant M. Rotskoff and Eric Vanden-Eijnden. Trainability and Accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Burr Settles. Active learning literature survey. Technical Report TR1648, University of Wisconsin-Madison Department of Computer Sciences, 2009.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational Learning theory*, pages 287–294, 1992.

Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. Deep active learning: unified and principled method for query and training. In *International Conference on Artificial Intelligence and Statistics*, pages 1308–1318, 2020.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Fabian Stark, Caner Hazırbas, Rudolph Triebel, and Daniel Cremers. CAPTCHA recognition with active deep learning. In *Workshop new challenges in Neural computation*, volume 2015, page 94, 2015.

Taiji Suzuki, Atsushi Nitanda, and Denny Wu. Uniform-in-time propagation of chaos for the mean-field gradient langevin dynamics. In *The Eleventh International Conference on Learning Representations*, 2023a.

Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field langevin dynamics: Time-space discretization, stochastic gradient, and variance reduction. In *Advances in Neural Information Processing Systems*, volume 36, pages 15545–15577. Curran Associates, Inc., 2023b.

Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. Feature learning via mean-field langevin dynamics: Classifying sparse parities and beyond. In *Advances in Neural Information Processing Systems*, volume 36, pages 34536–34556. Curran Associates, Inc., 2023c.

Rinyoichi Takezoe, Xu Liu, Shunan Mao, Marco Tianyu Chen, Zhanpeng Feng, Shiliang Zhang, and Xiaoyu Wang. Deep active learning for computer vision: Past and future. *APSIPA Transactions on Signal and Information Processing*, 12(1):–, 2023.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *arXiv preprint arkiv: 2305.16380*, 2023.

Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. JoMA: Demystifying multilayer transformers via joint dynamics of MLP and attention. *arXiv preprint arkiv: 2310.00535*, 2024.

Roman Vershynin. *High-dimensional Probability: An Introduction with Applications in Data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic Viewpoint*, volume 48. Cambridge university press, 2019.

Haonan Wang, Wei Huang, Ziwei Wu, Hanghang Tong, Andrew J Margenot, and Jingrui He. Deep active learning by leveraging training dynamics. In *Advances in Neural Information Processing Systems*, volume 35, pages 25171–25184, 2022a.

Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

Tianyang Wang, Xingjian Li, Pengkun Yang, Guosheng Hu, Xiangrui Zeng, Siyu Huang, Cheng-Zhong Xu, and Min Xu. Boosting active learning via improving test performance. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, pages 8566–8574, 2022b.

Zhilei Wang, Pranjal Awasthi, Christoph Dann, Ayush Sekhari, and Claudio Gentile. Neural active learning with performance guarantees. In *Advances in Neural Information Processing Systems*, volume 34, pages 7510–7521, 2021.

Ziting Wen, Oscar Pizarro, and Stefan Williams. NTKCPL: Active learning on top of self-supervised model by estimating true coverage. *arXiv preprint arkiv: 2306.04099*, 2023.

Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. Discovering universal geometry in embeddings with ica. *arXiv preprint arkiv: 2305.13175*, 2023.

Greg Yang and Edward J. Hu. Feature learning in infinite-width neural networks. *arXiv preprint arkiv: 2011.14522*, 2022.

Yazhou Yang and Marco Loog. Active learning using uncertainty information. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2646–2651, 2016.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Changchang Yin, Buyue Qian, Shilei Cao, Xiaoyu Li, Jishang Wei, Qinghua Zheng, and Ian Davidson. Deep similarity-based batch mode active learning with exploration-exploitation. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 575–584, 2017.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D. Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. *arXiv preprint arkiv: 2306.01129*, 2023.

Xueying Zhan, Huan Liu, Qing Li, and Antoni B. Chan. A comparative survey: Benchmarking for pool-based active learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4679–4686, 2021.

Xueying Zhan, Qingzhong Wang, Kuan hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B. Chan. A comparative survey of deep active learning. *arXiv preprint arkiv: 2203.13450*, 2022.

Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arkiv: 1901.05954*, 2019.

Yinglun Zhu and Robert Nowak. Active learning with neural networks: Insights from nonparametric statistics. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 142–155. Curran Associates, Inc., 2022.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109:467–492, 2020.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. The benefits of mixup for feature learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 43423–43479, 2023a.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization. In *The Eleventh International Conference on Learning Representations*, 2023b.

A. Additional Related Work: Theory of Feature Learning in Overparameterized Neural Network

The rapid progress of deep neural networks has prompted growing interest in understanding their underlying theoretical principles, particularly regarding the optimization and generalization properties of overparameterized models. A key development in this area is the study of the Neural Tangent Kernel (NTK) (Jacot et al., 2020; Chen et al., 2020; Cao and Gu, 2019a;b; Cao et al., 2020; Allen-Zhu et al., 2019; Chen et al., 2021b; Zou et al., 2020; Huang et al., 2020; Chen et al., 2021a; Huang et al., 2021; 2022; 2023b; Yang and Hu, 2022). This has provided powerful insights into the training dynamics of wide neural networks, revealing that their behavior in the ℓ_2-loss setting closely mirrors the function approximation in reproducing kernel Hilbert spaces (RKHS), where the kernel is associated with the network architecture. However, instead of *feature learning*, this line of research suggest that the parameter update dynamics can be approximated by the first-order Taylor expansion at initialization, where the NN with *wide enough width* can effectively perform linear regression over a prescribed feature map, which cannot characterize the NN's ability to perform *feature learning* (Yang and Hu, 2022).

In parallel, an active research direction is the analysis of NN under mean-field regime (Mei et al., 2018; 2019), which allows the network parameters to evolve away from the initialization, thereby enabling *feature learning* for various target functions (Ba et al., 2022; Suzuki et al., 2023c). Recently, Mean-Field Langevin Dynamics (MFLD) has attracted increased attention, where Gaussian noise is added to the gradient to encourage "exploration" (Mei et al., 2018; Suzuki et al., 2023b). This framework lifts the learning of finite-width neural networks to an infinite-dimensional optimization problem in the space of probability measures, and by exploiting the convexity of the loss function in this measure space, MFLD can achieve near-optimal global convergence under gradient-based optimization (Nitanda and Suzuki, 2017; Mei et al., 2018; Nitanda et al., 2021; 2022; 2023a;b; 2024; Oko et al., 2022; Otto and Villani, 2000; Rotskoff and Vanden-Eijnden, 2018; Sirignano and Spiliopoulos, 2020; Suzuki et al., 2023a;c;b; Nitanda, 2024; Kim et al., 2024; Kim and Suzuki, 2024). Despite the remarkable ability of NNs under the MFLD regime to learn complex "features", their superior performance still requires a large width at the order of $e^{O(d)}$ (Suzuki et al., 2023c). Moreover, the optimization behavior of MFLD differs from the widely-applied SGD-based neural network algorithms, leaving the real-world *feature learning* phenomenon of commonly-utilized deep learning algorithms largely unexplained.

To overcome the technical challenges and shed light on the practical *feature learning* observed in GD/SGD-based learning algorithms, the seminal work by Allen-Zhu and Li (2023) takes a step forward. It first attempted to explain the observed success of ensemble methods in deep learning by adopting the NTK framework, but recognized the limitations of this approach. To tackle this challenge and fill the understanding gap, Allen-Zhu and Li (2023) considers a multi-view data model, which is a more complex version of the data model examined in the main body of our work. Allen-Zhu and Li (2023) justify this multi-view data model as plausible theoretical setups by empirically demonstrating the common occurrence of multiple one-task-oriented features in the latent space of ResNet, as shown in their Figures 2-4 and 9. Given the plausibility and suitability of this data setting for theoretical investigations of *feature learning dynamics*, a considerable body of research has delved into examining the capabilities of different learning algorithms under different structured conditions (Li and Liang, 2018; Karp et al., 2021; Yehudai and Shamir, 2019; Cao et al., 2022b; Chen et al., 2022; 2023b;c;a;d; Zou et al., 2023b; Li et al., 2023; Kou et al., 2023b;a;c; Meng et al., 2023; Huang et al., 2023a;c; Chidambaram et al., 2023; Deng et al., 2023; Frei et al., 2023; Tian et al., 2023; 2024). Notably, the width requirement for this line of research is considerably weaker compared to the NTK and MFLD regimes, which allows for a more fine-grained analysis of *feature learning dynamics* based on inner product-based feature direction reconstruction.

We believe our work extend this line of research by showing that the two primary criteria-based NALs are inherently prioritizing those underrepresented samples with yet-to-be-learned features. We hope this insight can help the community gain a deeper understanding of the heuristic NAL methods, and develop new principled approaches that can alleviate the data hungriness of deep learning.

B. Discussions on the Parameter Settings

In this section, we motivate the settings of our systems and dicuss the consequences of violating the requirements.

B.1. Choice of Systems

We would like to motivate our choice of systems in detail as below.

- **The system of learning dynamic:** $d, n, m, ||\boldsymbol{\mu}||, \sigma_0, \eta$**.** The choice of d, n, m aligns with the feature learning line of

research (Li and Liang, 2018; Karp et al., 2021; Frei et al., 2022; Chen et al., 2022; Allen-Zhu and Li, 2023; Chen et al., 2023b;c;a;d; Zou et al., 2023b; Li et al., 2023; Kou et al., 2023a; Huang et al., 2023a; Kou et al., 2023c; Chidambaram et al., 2023; Deng et al., 2023; Huang et al., 2023c), with the aim of ensuring the learning problem is in a small but sufficiently overparameterized regime where the benign overfitting - overparamiterized NN can generalize well when trained to convergence - could occur. This phenomenon is non-trivial against prior belief that overfit is always harmful-the greater the capacity of a model to fit data distribution, the worse the model's test results will be. The system chosen allows for analysis of learning progress of features, as the weak requirement on network width m allows us to conduct a fine-grained analysis based on inner product arguments (i.e., scale analysis of γ, ρ), which fundamentally differs from the NTK line of research (Jacot et al., 2020) that requires an infinitely wide network to perform linear regression over a prescribed feature map, rather than learning the features themselves. Moreover, this system ensures a small Signal-to-Noise Ratio (SNR), under which the memorization of noise would become the primary contributor to the volume of the NN's weight matrices, allowing a more balanced and controllable coefficient updates (Kou et al., 2023b; Meng et al., 2023).

- **The system of sampling dynamic:** $\widetilde{n}, n_0, n^*, p, |\mathcal{P}|, ||\boldsymbol{\mu}_1||, ||\boldsymbol{\mu}_2||, \sigma_p$**.** The choice of this system is to (i) avoid the cases where all sampling methods would succeed or fail simultaneously, and (ii) ensure there is a marked learning progress disparity between well-learned and yet-to-be-learned features within the initial stage. The reason to maintain these conditions is to help reveal the underlying rationale behind NAL. It's worth noting that we also provide discussions in Appendix D.3 on the general settings beyond the specific system chosen in the main body of the work. In these broader scenarios, there might be various patterns in the learning progress of features.

In all, albeit the two systems interact and operate together, they have distinct tasks. The first system is tailored to the non-trivial learning problem at hand. Meanwhile, the choice of the second system aims to help reveal the non-trivial connections between the two NAL methods, by closely tracking the learning progress of task-oriented features after sampling.

B.2. Consequences of Violating System Requirements

The following outlines the consequences that may arise where the requirements over the systems are violated:

1. The choice of d. The large d technically ensures the per-sample loss contributions are in a controllable order during training, preventing any individual's noise from exerting outsized influence on the dynamics. When d decreases with respect to n, m, the control of the order over $< \frac{\boldsymbol{\mu}_l}{||\boldsymbol{\mu}_l||}, \frac{\boldsymbol{\xi}_i}{||\boldsymbol{\xi}_i||} >, < \frac{\boldsymbol{\xi}_i}{||\boldsymbol{\xi}_i||}, \frac{\boldsymbol{\xi}_j}{||\boldsymbol{\xi}_j||} >, < \mathbf{w}_{j,r}^{(0)}, \frac{\boldsymbol{\mu}_l}{||\boldsymbol{\mu}_l||} >, < \mathbf{w}_{j,r}^{(0)}, \frac{\boldsymbol{\xi}_i}{||\boldsymbol{\xi}_i||} >$ $, \forall l, i \neq j$ no longer hold with high probability as listed in Appendix G.1, and our technical results on training convergence can not be assured to hold with high chance. Also, a small d leads to a large Signal-to-Noise Ratio (SNR), where the memorization of noise is no longer the dominant factor in the NN's weight matrix volume. This makes the *automatic balance of coefficient updates* techniques in Kou et al. (2023b); Meng et al. (2023) cannot hold, which serves as a convenient lever to observe the bounds on the coefficients and matrix volume update.

2. The choices of occurrence probability p, initial size n_0, query size n^*, pool size $|\mathcal{P}|$, feature norm $\|\boldsymbol{\mu}_l\|$ jointly determine the sampling results.
   - Combinations of $p, \|\boldsymbol{\mu}_l\|$ reflect the diverse "easiness" to learn particular features, leading to varied sampling scenarios as discussed in Appendix D.3.2.
   - As p, n_0 and n^* increase, the chance of getting all features well-learned goes up, reducing NAL's advantage over random sampling as discussed in Appendix D.3.2.
   - Lower p values (e.g. $p < 0.5$) allow NAL to better alleviate labeling efforts by prioritizing the samples with yet-to-be-learned features, but if $p \to 0$ or $|\mathcal{P}|$ decreases, there might be few yet-to-be-learned features in the pool, limiting NAL's ability to select enough of them to ensure sufficient learning, as discussed in Appendix D.3.2.
   - Smaller n_0 may limit the learning of all features at initial stage, and all sampling methods might behave similarly since all types of features require further learning as discussed in Appendix D.3.2. Decreases in n_0, n^*, and $|\mathcal{P}|$ would make it challenging to reliably control the proportions of samples as in Lemma G.3.

3. The choices of σ_0 and η aim to ensure effective optimization via GD. As σ_0 grows, the model has a stronger "belief" that is harder to change. While analysis under larger η is also doable (Lu et al., 2023), a small η is preferred to better present our main findings.

Amidst parameter variations, we believe our findings are non-trivial.

C. Theoretical Results: XOR data version

In a similar vein to the theoretical results on linearly separable data, we now present a theory specifically tailored for XOR data. The purpose or effect of each result is similar to those obtained for linearly separable data, so we will omit the detailed description of each result. The experiments and proofs can be found in Appendix E.3 and Appendix H.

Definition C.1. (Meng et al., 2023) Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$ with $\mathbf{a} \perp \mathbf{b}$ be two fixed vectors. For $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\bar{y} \in \{\pm 1\}$, we say that $\boldsymbol{\mu}$ and $\bar{y}$ are jointly generated from distribution $\mathcal{D}_{\mathrm{XOR}}(\mathbf{a}, \mathbf{b})$ if the pair $(\boldsymbol{\mu}, \bar{y})$ is randomly and uniformly drawn from the set $\{(\mathbf{a}+\mathbf{b}, +1), (-\mathbf{a}-\mathbf{b}, +1), (\mathbf{a}-\mathbf{b}, -1), (-\mathbf{a}+\mathbf{b}, -1)\}$.

Definition C.2. For $l \in \{1, 2\}$, let $\{\mathbf{a}_1, \mathbf{b}_1\} \perp \{\mathbf{a}_2, \mathbf{b}_2\} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$, with $\mathbf{a}_l \perp \mathbf{b}_l$ be two pair of fixed vectors satisfying $\|\mathbf{a}_l\|^2 + \|\mathbf{b}_l\|^2 = \|\boldsymbol{\mu}_l\|_2^2$, where $\|\boldsymbol{\mu}_l\|_2$ represents feature strength. Then each data point $(\mathbf{x}, y)$ with $\mathbf{x} = \left[\mathbf{x}^{(1)\top}, \mathbf{x}^{(2)\top}\right]^\top \in \mathbb{R}^{2d}$ and $y \in \{\pm 1\}$ is generated from $\mathcal{D}$ as follows:

- **Feature Patch.** For a feeble p satisfying $p < 0.5$, one patch of $\mathbf{x}$ is randomly selected as feature patch, and with high probability (1-p) the feature patch $\mathbf{x}_1$ is easy-to-learn feature $\boldsymbol{\mu}_1$, while only with probability p the feature patch is hard-to-learn feature $\boldsymbol{\mu}_2$. $\boldsymbol{\mu}_l \in \mathbb{R}^d$ and $\bar{y} \in \{\pm 1\}$ are jointly generated from $\mathcal{D}_{\mathrm{XOR}}(\mathbf{a}_l, \mathbf{b}_l)$.
- **Noise Patch.** The other patch of $\mathbf{x}$ is assigned as a randomly generated Gaussian vector $\boldsymbol{\xi} \sim N\left(\mathbf{0}, \sigma_p^2 \cdot \left(\mathbf{I} - \sum_l \left(\mathbf{a}_l \mathbf{a}_l^\top / \|\mathbf{a}_l\|^2 - \mathbf{b}_l \mathbf{b}_l^\top / \|\mathbf{b}_l\|^2\right)\right)\right)$.
- The ground truth label y is synthesized from a Rademacher distribution.

Here we assume the two types of feature differ: $(1-p)\|\boldsymbol{\mu}_1\|_2^4 = \Omega(\sigma_p^4 d n_0^{-1})$ and $p\|\boldsymbol{\mu}_2\|_2^4 = O(\sigma_p^4 d n_0^{-1})$. Also, we assume the noise cannot completely disturb the learning of features: $\widetilde{n}\|\boldsymbol{\mu}_l\|_2^4 = \Omega(\sigma_p^4 d), l \in \{1, 2\}$.

For $(\mathbf{x}, y) \sim \mathcal{D}$ in Definition C.2, it's safe to say that:

$$(\mathbf{x}, y) \overset{d}{=} (-\mathbf{x}, y), \text{ and therefore } \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}(y \cdot \langle \boldsymbol{\theta}, \mathbf{x} \rangle > 0) = 1/2 \text{ for any } \boldsymbol{\theta} \in \mathbb{R}^{2d}.$$

In other words, all linear predictors will provably fail to learn the XOR-type data $\mathcal{D}$.

Condition C.3. *For certain $\varepsilon, \delta > 0$, suppose that*

1. *The initial training size n_0, the maximum admissible size after querying $\widetilde{n}$, and the width of neural network m satisfy $n_0 = \Omega(\log(m/\delta), p^{-1}\log(1/\delta))$, $\widetilde{n} = O(p^{-1}\sigma_p^4 d\|\boldsymbol{\mu}_2\|_2^{-4})$, $m = \Omega(\log(\widetilde{n}/\delta))$.*
2. *The dimension d satisfies: $d = \widetilde{\Omega}\left(\widetilde{n}^2, \widetilde{n}\|\boldsymbol{\mu}_l\|_2^2 \sigma_p^{-2}\right) \cdot \mathrm{polylog}(1/\varepsilon) \cdot \mathrm{polylog}(1/\delta)$, for $l \in \{1, 2\}$.*
3. *Random initialization scale σ_0 satisfies: $\sigma_0 \leq \widetilde{O}\left(\min\left\{\sqrt{n_0}/\left(\sigma_p d\right), n_0\|\boldsymbol{\mu}_l\|_2/\left(\sigma_p^2 d\right)\right\}\right)$, for $l \in \{1, 2\}$, the learning rate η satisfies: $\eta = \widetilde{O}\left(\left[\max\left\{\sigma_p^2 d^{3/2}/\left(n_0^2\sqrt{m}\right), \sigma_p^2 d/(n_0 m)\right]^{-1}\right)\right.$.*
4. *The angle θ between $\mathbf{a}_l + \mathbf{b}_l$ and $\mathbf{a}_l - \mathbf{b}_l$ satisfies $\cos\theta < 1/2$, for $\forall l \in \{1, 2\}$.*

Proposition C.4. ***(Before Querying)*** *For any $\varepsilon, \delta > 0$, if Condition C.3 holds, when the probability of the appearance of weak feature in each data point generated from the testing distribution $\mathcal{D}^*$ is p^*, then with probability at least $1 - 2\delta$, the following results hold at a certain $t = \Omega\left(n_0 m/\left(\eta\varepsilon\sigma_p^2 d\right)\right)$:*

- *The training loss converges below ε, i.e., $L_S\left(\mathbf{W}^{(t)}\right) \leq \varepsilon$.*
- *The test error achieve sub-optimal constant-level $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \geq p^* \cdot 0.12$.*

Proposition C.5. ***(Querying Stage)*** *During Querying, under the same conditions as Proposition C.4, if $(1-p)\|\boldsymbol{\mu}_1\|_2^2 - p\|\boldsymbol{\mu}_2\|_2^2 = \Omega({\sigma_p}^2 d^{1/2} n_0^{-1/2}(\log(m/\delta'))^{1/2})$ and the size of the sampling pool $|\mathcal{P}|$ is adequately substantial, satisfying: $|\mathcal{P}| = \Omega(p^{-1}\sigma_p^4 d\|\boldsymbol{\mu}_2\|_2^{-4}, p^{-1}\log(1/\delta))$, then with probability at least $1 - \Theta(\delta + \delta')$, both Uncertainty Sampling and Diversity Sampling pick samples with hard-to-learn features $\boldsymbol{\mu}_2$ in $\mathcal{P}$.*

Theorem C.6. ***(After Querying)*** *If the sampling size n^* of the two types of Sampling algorithm satisfies $\frac{\hat{C}_1\sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4} - \frac{pn_0}{2} \leq n^* = \Theta(\widetilde{n} - n_0) \leq \widetilde{n} - n_0$, where $\hat{C}_1$ is some positive constant, under the same conditions as Proposition C.5, the $\mathcal{D}^*$ and p^* follows the same definitions in Proposition C.4, then with probability at least 1 - $\Theta(\delta + \delta')$, we have the following results hold at a certain $t = \Omega\left((n_0 + n^*)m/\left(\eta\varepsilon\sigma_p^2 d\right)\right)$:*

- *For both the Random Sampling method and Uncertainty Sampling method, the training loss converges to ε, i.e., $L_S\left(\mathbf{W}^{(t)}\right) \leq \varepsilon$.*
- ***Uncertainty Sampling** and **Diversity Sampling** algorithms both have negligible test error: $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \leq \exp(\Theta\left(\frac{-\widetilde{n}\|\boldsymbol{\mu}_l\|_2^4}{\sigma_p^4 d}\right)), l \in \{1, 2\}$.*
- ***Random Sampling** algorithm would remain the sub-optimal constant-level test error: $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \geq p^* \cdot 0.12$.*

D. Discussions over General Scenarios

Our findings align with the concept of "Active Learning," where models resemble students (models) actively selecting valuable practice questions (samples) to prepare for exams (tasks). Students prioritize perplexing questions based on high uncertainty of their answers, or rare knowledge points (features), in order to enhance their understanding of yet-to-be-mastered (lack of learning progress) knowledge points (features) in test questions. Similar to students, for most black-box deep neural models, the "learning progress" of particular "feature" is not readily available for algorithm developer due to their inherent opacity. From a feature learning view, that's why NAL algorithms need to indirectly prioritize those yet-to-be-learned features, since this is the key for their good generalization ability and achieve *benign overfitting*. Our study shows that uncertainty-based and diversity-based NAL inherently strive to prioritize yet-to-be-learned feature-assisted samples (i.e., **perplexing samples**) via different comparisons in a heuristic manner. We believe future work can figure out if developed interpretable models (Yu et al., 2023) reduced labelling efforts by prioritizing **perplexing samples**.

Below, we present several discussions regarding general scenarios and the potential wider applicability of our theorems, beyond the specific conditions considered in the main body of our work. It is important to note that our point-mass querying approach and one-round querying settings were adopted to better unveil the inherent principle of query criteria-based NAL algorithms in a rigorous manner, albeit other complex NAL algorithms may be better suited for real-world complex data distribution and corresponding tasks. Note that our multiple task-oriented feature-noise data modellings follow the modellings in Allen-Zhu and Li (2023); Chen et al. (2022; 2023b;c;a;d); Zou et al. (2023b); Li et al. (2023); Kou et al. (2023a;c), which empirically mirror the latent representation of models like Resnet (Allen-Zhu and Li, 2023) or transformer (Yamagiwa et al., 2023; Jiang et al., 2024).

D.1. Discussion of the Role of Benign Oscillation

In the work by Lu et al. (2023), they analyze the role of a large learning rate in the context of feature learning. Their data modeling includes weak features present in each data point, strong features present in a small fraction of data points, and noise. Although our work differs in terms of the data modeling and analysis framework, we might also observe the impact of a large learning rate. In Figures 2, 5, and 7, we can see that Uncertainty Sampling and Diversity Sampling algorithms empirically outperform the fully-trained model. Drawing insights from the results in Lu et al. (2023), we attribute this phenomenon to the large learning rate, which drives the model to be trained to focus more on weak and rare features. It is worth noting that although our training loss does not exhibit the *benign oscillation* phenomenon mentioned in Lu et al. (2023), this probably could be due to the difference in optimization algorithms (GD with logistic loss in our work versus SGD with square loss in Lu et al. (2023)).

D.2. Potential Extension over State-of-arts and Criteria-combined NALs: BADGE as an Exampler

We believe our analysis can indeed be extended to reveal the success of methods like BADGE (Ash et al., 2020) that combine uncertainty and diversity criteria. We show they share a common principle of prioritizing samples with yet-to-be-learned features. Like the inner product arguments in prior theoretical results (Li and Liang, 2018; Karp et al., 2021; Allen-Zhu and Li, 2023; Chen et al., 2022; 2023b;c;a;d; Zou et al., 2023b; Li et al., 2023; Kou et al., 2023a; Huang et al., 2023a; Kou et al.,

2023c; Chidambaram et al., 2023; Deng et al., 2023; Huang et al., 2023c), our theory characterizes learning progress via the coefficients $\gamma_{j,r,l}$, which high-levelly represent how well the NN has integrated low-dimensional task-oriented patterns into its latent space. We believe the underlying principle of BADGE (Ash et al., 2020) aligns well with this view:

- **Core idea of BADGE.** The key idea behind BADGE is to query samples that exhibit large and diverse gradients within a single batch, achieved through k-MEANS ++ or k-DPP in the pseudo gradient space.
- **Connection between gradient and latent space of NN.** Since our analysis utilizes the well-applied non-increasing logistic loss, the smaller the magnitude of the latent representation, the larger the magnitude of the gradient embedding will be. Additionally, the diversity of the latent vectors' directions will be preserved in the gradient space. Based on Lemma G.15, we see that the rows of the latent representations are roughly of the order as $\gamma_{j,r,l}^{(t)}$.
- **BADGE also prioritizes samples with yet-to-be-learned feature.** We now know the BADGE tends to prioritize samples with smaller scale latent representations (smaller $\gamma_{j,r,l}$) or more diverse directions (many diverging $\gamma_{j,r,l}$). These samples correspond to the cases described in Lemma 4.2, which in our context refers to samples with lower $\gamma_{j,r,l}$ that have yet-to-be-learned features.

Therefore, we claim that BADGE, in the context of our analysis regime, can be explained as a well-motivated NAL method. The key reason is that the two core ideas of BADGE align with the shared underlying rationale of NAL that we has uncovered. One of our future work would serve to give a fine-grained analysis of the success factors behind BADGE, and we also believe our theoretical framework has the potential to extend to the understanding of some other state-of-the-art methods.

D.3. Extension over Data Distribution under Other Conditions

The theory presented in our main study focuses on a data model that includes weak and rare features, strong and common features, and noise. This setting is motivated by real-world imbalanced datasets, as illustrated in Figure 1. However, **thanks to our general analysis framework**, we can also discuss more general scenarios with broader conditions. In the following sections, we first discuss a theory version that **relaxes the conditions on feature norms**. This case suggests that rare features may also possess sufficiently discriminative label-related features, such as Simba in the last row of Figure 1, even though they are rare occurrences in the overall data distribution. Secondly, we introduce **a more general theoretical results**. While our discussions below focused on results for linearly separable data, we assert that the same results hold for non-linearly separable XOR data, as the requirements for the parameters are indeed similar. The proofs of all results in this section can be readily obtained based on our results in Appendix G.4, H.3 , G.5 or H.4.

To start, we present the condition-relaxed versions of Proposition G.16, which describe the order situation of samples in $\mathcal{P}$ under relaxed conditions. Here we denote τ_l as the proportion of $\boldsymbol{\mu}_l$-equipped data in $\mathcal{D}_{n_0}$.

Proposition D.1. ***(Proposition G.16 with relaxed conditions on feature norms)*** *Under Condition 3.1, there exist* $t = \widetilde{O}\left(\eta^{-1}\varepsilon^{-1}mnd^{-1}\sigma_p^{-2}\right)$ *that for* $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{P} \subsetneq \mathcal{D}$ *where* $\mathbf{x}$ *contains feature patch* $y \cdot \boldsymbol{\mu}_2$ *while* $\mathbf{x}'$ *contains feature patch* $y' \cdot \boldsymbol{\mu}_1$*, with probability at least* $1 - 8m\exp\left\{-\Theta\left(\left[\tau_1 \|\boldsymbol{\mu}_1\|_2^2 - \tau_2 \|\boldsymbol{\mu}_2\|_2^2\right]^2/(\sigma_p^4 d/n_0)\right)\right\}$*, we have* $\mathbf{x}' \preceq^{(t)} \mathbf{x}$.

Proof of Proposition D.1. See the proving process of Proposition G.16.

This theorem serve as the key to analysis of the querying statistics, as samples with the lower $\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,l})$ are **perplexing samples**. Based on the coefficient scale presented in Lemma G.14, we can obtain the probability lower bound for $\mathbf{x}' \preceq^{(t)} \mathbf{x}$, which is

$$P(\mathbf{x}' \preceq^{(t)} \mathbf{x}) \geq 1 - 8m\exp\left\{-\Theta\left(\frac{\left[\tau_1 \|\boldsymbol{\mu}_1\|_2^2 - \tau_2 \|\boldsymbol{\mu}_2\|_2^2\right]^2}{\sigma_p^4 d/n_0}\right)\right\}. \tag{7}$$

Thus we can conclude that **perplexing samples** are samples with lower $\tau_l \|\boldsymbol{\mu}_l\|_2^2$. We then can relax the conditions on feature norms by imposing specific conditions on p. Additionally, we can relax both conditions on feature norms and conditions on p to consider a more general case. The upcoming sections will discuss these scenarios in detail.

D.3.1. Case 1: Relaxed Conditions on Feature Norms

In the main body of our work, we have the conditions on feature norms: $\|\boldsymbol{\mu}_1\|_2^4 = \Omega(\sigma_p^4 d n_0^{-1})$, $\|\boldsymbol{\mu}_2\|_2^4 = O(\sigma_p^4 d n_0^{-1})$ and $\|\boldsymbol{\mu}_1\|_2^2 - \|\boldsymbol{\mu}_2\|_2^2 = \Omega({\sigma_p}^2 d^{1/2} n_0^{-1/2}(\log(m/\delta'))^{1/2})$ for the ease of presentations. In this section we provide a theory version that relaxes these requirements (i.e., no discrepancy in terms of feature norms). The essence is that we can impose stricter assumptions on p to ensure there exists a learning progress disparity between the two features. Despite this, the inherent principle of the two-criteria-based NAL approach would still drive the algorithms to preferentially query the samples containing the yet-to-be-learned features. The rigorous rationale behind these will be thoroughly explored in Appendix G.3 and Appendix G.5. Here, we can leverage the deduction results in Appendix G.3, Appendix G.4 and Appendix G.5 to readily form the following results.

Definition D.2. (**Definition with relaxed conditions on feature norms**) Let $\boldsymbol{\mu}_1 \perp \boldsymbol{\mu}_2 \in \mathbb{R}^d$ be two fixed feature vectors. Each data point $(\mathbf{x}, y)$, where $\mathbf{x}$ contains two patches as $\mathbf{x}$=$[\mathbf{x}_1^T, \mathbf{x}_2^T]^T \in \mathbb{R}^{2d}$ and $y \in \{-1, 1\}$ are generated from the distribution $\mathcal{D}$:

- The ground truth label y is synthesized from a Rademacher distribution.
- **Noise Patch.** One patch of $\mathbf{x}$ is selected as a noise patch $\boldsymbol{\xi}$, synthesized from Gaussian distribution $N(\mathbf{0}, \sigma_p^2 \cdot \mathbf{I})$.
- **Feature Patch.** For a feeble p satisfying $p < O(n_0 \sigma_p^4 d \|\boldsymbol{\mu}_2\|_2^{-4}), (\|\boldsymbol{\mu}_1\|_2^2 + \|\boldsymbol{\mu}_2\|_2^2)^{-1}(\|\boldsymbol{\mu}_1\|_2^2 + {\sigma_p}^2 d^{1/2} n_0^{-1/2}(\log(8m/\delta'))^{1/2})$, the remaining patch of $\mathbf{x}$ is selected as label-related feature patch, and with high probability (1-p) the feature patch is a common feature $y \cdot \boldsymbol{\mu}_1$, while only with probability p the feature patch is a rare feature $y \cdot \boldsymbol{\mu}_2$.

Here we only require that the learning of features would not completely disturbed by noise: $\forall l \in \{1, 2\}, \|\boldsymbol{\mu}_l\|_2^2 = \Omega(\sigma_p^2 \log(n_0/\delta), n_0^{-1} d \sigma_p^4)$.

The specific condition on the occurrence probability p serves two purposes. Firstly, it ensures that strategy-free passive learning cannot sample enough rare data to adequately learn the rare label-related feature $\boldsymbol{\mu}_2$, as observed in the real-world scenario depicted in Figure 1. Secondly, it helps distinguish the learning progress between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

We can prove that three querying algorithms still exhibit *harmful overfitting* at the initial stage.

Proposition D.3. ***(Before Querying)*** *At the initial stage before querying, $\forall \varepsilon > 0$, under Condition 3.1, with probability at least $1 - \delta$, there exists $t = \widetilde{O}\left(\eta^{-1}\varepsilon^{-1} m n_0 d^{-1} \sigma_p^{-2}\right)$, the followings hold for all of the three querying algorithms:*

1. *The training loss converges to ε, i.e., $L_S\left(\mathbf{W}^{(t)}\right) \leq \varepsilon$.*
2. *The test error remains at constant level, i.e., $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) = \Theta(1) \geq 0.12 \cdot p^*$.*

Then, we can still have a look on the querying stage based on the techniques in Appendix G.4.

Proposition D.4. ***(Querying Stage)*** *During Querying, under the same conditions as Proposition D.3, then with probability at least $1 - \Theta(\delta + \delta')$, Uncertainty Sampling and Diversity Sampling would all pick n^* samples that models exhibit lowest $\mathop{\mathbb{E}}_{j,r} \gamma_{j,r,l}^{(t)}$ (i.e., **perplexing samples**). Moreover, those **perplexing samples** are samples with rare feature $\boldsymbol{\mu}_2$.*

Similar to the theories presented in the main body of our study, we can establish the following theorem.

Theorem D.5. ***(After Querying)*** *If the sampling size n^* of the three querying algorithms satisfies $C_1 \sigma_p^4 d \|\boldsymbol{\mu}_2\|_2^{-4} - p n_0/2 \leq n^* = \Theta(\widetilde{n} - n_0) \leq \widetilde{n} - n_0$, where C_1 is some positive constant. Then for $\forall \varepsilon > 0$, under the same conditions as Proposition 3.3, with probability more than 1 - $\Theta(\delta + \delta')$, there exists $t = \widetilde{O}\left(\eta^{-1}\varepsilon^{-1} m (n_0 + n^*) d^{-1} \sigma_p^{-2}\right)$ such that:*

- *For all of the three querying algorithms, the training loss converges to ε, i.e., $L_S\left(\mathbf{W}^{(t)}\right) \leq \varepsilon$.*
- ***Uncertainty Sampling** and **Diversity Sampling** algorithms have negligible near Bayes-optimal test error: $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \leq \exp(\Theta\left(\frac{-\widetilde{n}\|\boldsymbol{\mu}_l\|_2^4}{\sigma_p^4 d}\right)), l \in \{1, 2\}$.*
- ***Random Sampling** algorithm would remain constant order test error: $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) = \Theta(1) \geq 0.12 \cdot p^*$.*

D.3.2. Case 2: Flexible Cases

Indeed, we can relax both the conditions on feature norms and the conditions on p to explore more general cases. By (7), if $\tau_1 \|\boldsymbol{\mu}_1\|_2^2 \approx \tau_2 \|\boldsymbol{\mu}_2\|_2^2$, the learning progress of the two types of features would be alike (i.e., $\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,1}) \approx \mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,2})$), and we cannot clearly observe which type of feature-equipped samples are likely to be queried. Thanks to our sample-complexity analysis regimes in Appendix G.5, we can clearly examine two scenarios at the initial stage based on (G.3) and Lemma G.21:

- *Benign Overfitting*: if $\tau_l \|\boldsymbol{\mu}_l\|_2^4 \geq 2C_1\sigma_p^4 d n_0^{-1}$, the learning of $\boldsymbol{\mu}_l$-equipped data would be adequate, and the test error of algorithms achieve Bayes-optimal.
- *Harmful Overfitting*: if $\tau_l \|\boldsymbol{\mu}_l\|_2^4 \leq 2C_2/3\sigma_p^4 d n_0^{-1}$, the learning of $\boldsymbol{\mu}_l$-equipped data would be inadequate, and the test error of algorithms remains constant level.

Then, we can list some cases with certain p ($\tau_2 = \Theta(p)$ by Lemma G.3) , $\|\boldsymbol{\mu}_l\|_2, l \in \{1, 2\}$ in our analysis regime:

1. When the learning of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are all adequate, we can conclude that n_0 is already sufficient for training in this case.
2. When the learning of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are all inadequate at the initial stage, all querying algorithms (i.e., Random Sampling, Uncertainty Sampling and Diversity Sampling) can help leverage learning of features. While our theory indicates that the two NAL algorithms would tend to prioritize samples with comparatively poorer learned feature (i.e., $\{\boldsymbol{\mu}_l \mid \tau_l \|\boldsymbol{\mu}_l\|_2^4 = \min(\tau_1 \|\boldsymbol{\mu}_1\|_2^4, \tau_2 \|\boldsymbol{\mu}_2\|_2^4)\}$), the difference in generalization ability between Random Sampling and the two NAL algorithms would depend on certain parameters (i.e., $p, n^*, |\mathcal{P}|, \|\boldsymbol{\mu}_1\|_2, \|\boldsymbol{\mu}_2\|_2$).
3. When the learning of $\boldsymbol{\mu}_{l_1}$ is adequate while the learning of $\boldsymbol{\mu}_{l_2}$ is inadequate ($l_1 \neq l_2 \in \{1, 2\}$), we have the following cases based on our theory:
   - If $\tau_{l_1} \|\boldsymbol{\mu}_{l_1}\|_2^2 \approx \tau_{l_2} \|\boldsymbol{\mu}_{l_2}\|_2^2$, the prioritization by two NAL algorithms is not obvious, and they would perform similarly to Random Sampling.
   - If $\tau_{l_1} \|\boldsymbol{\mu}_{l_1}\|_2^2 > \tau_{l_2} \|\boldsymbol{\mu}_{l_2}\|_2^2$, two NAL algorithms would tend to prioritize **perplexing samples** (i.e., samples with $\boldsymbol{\mu}_{l_2}$), and their prioritization has lower probability bound in (7). Meanwhile, the difference in generalization ability between Random Sampling and the two NAL algorithms would depend on certain parameters (i.e., $p, n^*, |\mathcal{P}|, \|\boldsymbol{\mu}_1\|_2, \|\boldsymbol{\mu}_2\|_2$). Specifically, under Condition 3.1, Definition 2.1 and Definition D.2 provide two parameter settings satisfying $\tau_{l_1} \|\boldsymbol{\mu}_{l_1}\|_2^2 - \tau_{l_2} \|\boldsymbol{\mu}_{l_2}\|_2^2 = \Omega({\sigma_p}^2 d^{1/2} n_0^{-1/2} (\log(m/\delta'))^{1/2})$, where the two NAL algorithms succeed while Random Sampling fails (i.e., Theorem 3.4 and Theorem D.5). Other general scenarios can also be rigorously analyzed with the prioritization probability lower bound in (7) and permutation probability.
4. Other cases would be similar to the second or third case (i.e., where $\exists l \in \{1, 2\}, 2C_2/3\sigma_p^4 d n_0^{-1} \leq \tau_l \|\boldsymbol{\mu}_l\|_2^4 \leq 2C_1\sigma_p^4 d n_0^{-1}$).

In real-world scenarios, the pool-based setting often resembles a wide range of flexible cases. From the perspective of feature learning, our theoretical observations suggest that the occurrence probability and strength of different task-specific features can profoundly impact the efficiency of NAL algorithms.

D.4. Cases of Criteria Preference

Our work has uncovered a non-trivial connection between the two query criteria-based NAL methods. Specifically, they share a sufficient condition - which we also called it as the shared principle - that is vital to the success of NAL methods, which holds when the learning progress of the well-learned features greatly surpasses the learning of the yet-to-be-learned features to a certain degree

$$\underbrace{\Theta(\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,1})) - \Theta(\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,2}))}_{\text{Learning Progress Disparity: well-learned Feature } vs. \text{yet-to-be-learned Feature}} > \max_{j,r,l} |< \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l >|.$$

However, as discussed in Appendix D.3.2 above, when this shared sufficient condition (or principle) does not hold, the behaviors of the two heuristic criteria-based sampling methods may differ.

Cases favoring uncertainty-based Sampling. Specifically, when the label budget is not highly limited and there is sufficient opportunity to capture all feature types, uncertainty-based sampling may be preferred. Our analysis shows that compared to uncertainty sampling, diversity sampling has a stricter requirement, with a less than 1 scalar $(\tau_1 - \tau_2)$ in the left side of inequalities (37) and (70), versus (31) and (64) for uncertainty sampling. This allows uncertainty sampling to more precisely prioritize samples with yet-to-be-learned features, more easily ensuring adequate learning across all feature types.

Cases favoring diversity-based Sampling. However, when label complexity is quite limited (as per Appendix D.3.2) where all task-oriented features require further labelling budget, we may favor diversity-based sampling. Despite all sampling algorithms increasing test accuracy by addressing insufficient learning of certain features, diversity sampling's efficiency in obtaining diverse features could enhance the model's ability to grasp diverse low-dimensional patterns. This in turn could enrich generalization, even when the test distribution differs from training.

Our statements here align with discussions in the recent survey (Zhan et al., 2021). We believe this nuanced perspective deserves further exploration.

Cases favoring Strategy-free Random Sampling. As discussed in Appendix D.3.2, our theory suggests that when $\tau_1\|\boldsymbol{\mu}_1\|^2 \approx \tau_2\|\boldsymbol{\mu}_2\|^2$ where τ_l denotes the proportion of $\boldsymbol{\mu}_l$ in training set, it indicates a balanced "easiness" to learn multiple task-oriented features. In such cases, the learning progress of these features tends to be similar, and the prioritization by NAL methods may not be clearly evident. In other words, if there is no distinct gap between well-learned and yet-to-be-learned features, uniform sampling might be sufficient, and the advantage of NAL methods only emerges when there is a clear distinction of "learning easiness" among various task-oriented feature categories.

Additionally, when it comes to the scenarios of active fine-tuning, where the task objective is heavily or slightly changing. In such situations, the task-oriented low-dimensional patterns may shift, and the model's optimal representation could differ from before. As a result, NAL methods that leverage prior neural representations for sampling may not be as effective, and uniform sampling could be a satisfactory choice.

D.5. Discussions of Multi-round NALs

Our theory suggests that the core principle underlying both NAL methods is their tendency to prioritize the selection of samples containing yet-to-be-learned features. This fundamental characteristic is not inherently tied to the single-round setting, but rather reflects an intrinsic property of the two primary criteria-based NAL family.

In the multi-round iterative process, the learning progress of different features may diverge across rounds and potentially align with the various cases discussed in Appendix D.3.2. However, we expect the NAL methods to continue performing well due to their innate focus on prioritizing the selection of samples containing yet-to-be-learned features.

D.6. Discussions of Practical Lessons of our Results

Here are some key takeaways of our theory:

- **Potential of NAL to surpass fully-trained NN.** As discussed in Appendix D.1, and corroborated by the results in Lu et al. (2023), fully-trained neural networks tend to learn hard-to-learn features in an inefficient manner, as they place disproportionate emphasis on the easy-to-learn ones. In contrast, our analysis suggests that the NAL approach prioritizes samples with low $\gamma_{j,r,l}$, making it more likely to achieve a balanced rise in $\gamma_{j,r,1}$ and $\gamma_{j,r,2}$ during the new round of training. This implies that NAL has a better chance of ensuring sufficient learning of all features within a certain number of iterations, compared to fully-trained neural networks. This conclusion is partially validated by the empirical results presented in our Figures 2, 5, and 7, where the NALs outperform the neural networks. In real-world settings, we conjecture that NAL might have this potential when the neural network is sufficiently overparameterized and has the capacity to capture all relevant patterns of the problem instances within limited iterations.

- **Care orthogonal components of features or gradients.** Our theory suggests that if techniques can be adopted to capture the meaningful orthogonal components of a neural network's features or gradients (e.g., using ICA (Yamagiwa et al., 2023)), then the samples with low-magnitude latent feature components or high-magnitude gradient components might align with the perplexing samples in our work. This is because our theory indicates that yet-to-be-learned

features are often underrepresented in the neural network's latent space, and if the loss is non-increasing, the length in the latent space might be inversely proportional to the length in the corresponding gradient space. Notably, existing state-of-the-art methods such as BADGE (Ash et al., 2020) also leverage a similar idea with respect to the gradient component of the last layer.

- **Incorporate Signal-to-Noise Ratio (SNR) Measurement.** Our discussions in Appendix D.3 denote that the perplexing samples are often characterized by their rarity and low SNR (the scale ratio between feature and noise). Techniques, whether learnable or unlearnable, that can accurately or approximately measure the SNR of multiple task-oriented features in a NN's latent space may help develop a principled NAL approach, and for specific tasks and datasets, it may be feasible to develop such task-oriented SNR measurement methods.

E. Additional Experiments

E.1. Sampling Information of Main Results

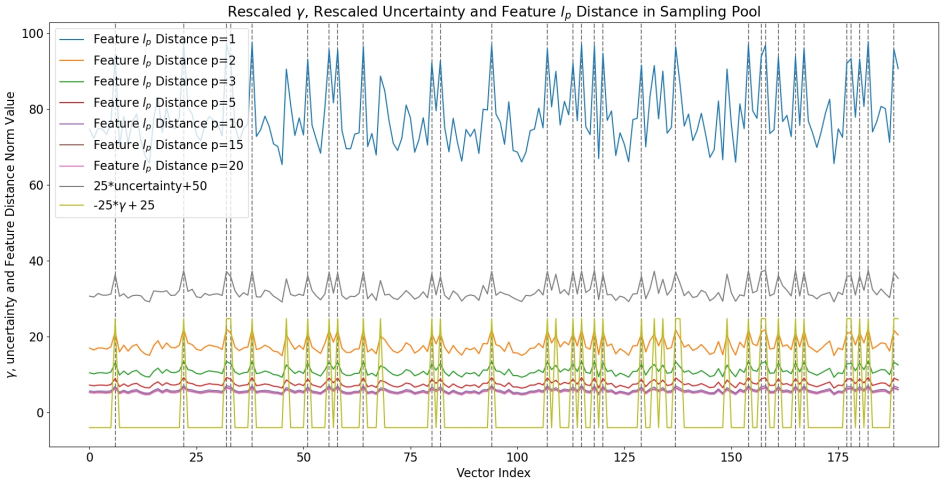
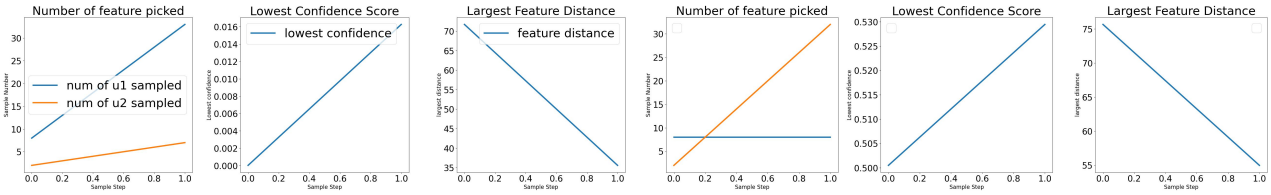
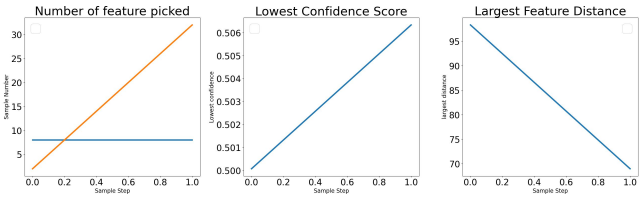


Figure 3. Rescaled γ ($\gamma = \mathbb{E}\gamma_{j,k,l}^{(t)}$), Uncertainty (i.e., $-$Confidence Score) and Feature Distance (with various p of l_p norm) of the samples in sampling pool $\mathcal{P}$, where γ represents the learning progress of feature in particular sample. The dashed line in the graph represents the top 30 samples with the highest Feature Distance.



(a) Random Sampling

(b) Uncertainty Sampling

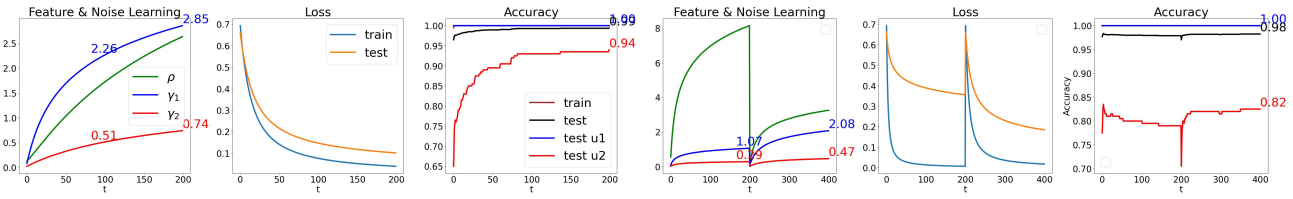


(c) Diversity Sampling

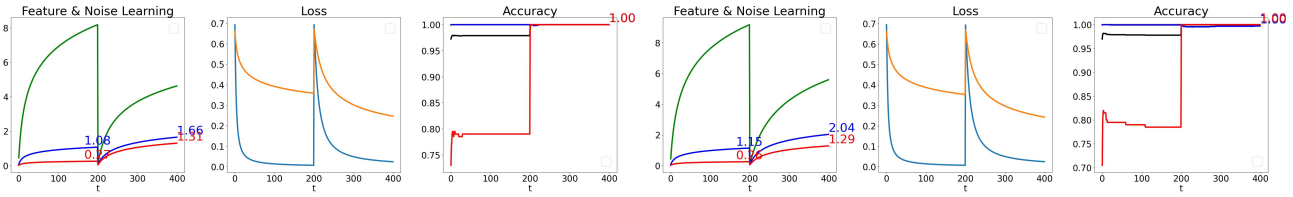
Figure 4. Comparison of querying information between two NAL algorithms, illustrating training size changes in labeled data sets, Confidence Score, and Feature Distance before and after querying.

Here we give more visualized details of the querying stage. The parameter settings are the same in Section 5. Figure 3 visualized the rescaled $\mathop{\mathbb{E}}_{j,k,l} \gamma_{j,k,l}$, uncertainty(-Confidence Score) and Feature Distance of each samples in the unlabeled sampling pool $\mathcal{P}$, where the dash line corresponds to the top n^* samples based on Diversity Order. It's obvious that regardless of the value p, the Uncertainty Order and Diversity Order of samples remain the same, and corresponds to the order of $\mathop{\mathbb{E}}_{j,k,l} \gamma_{j,k,l}$. This validates our unification claims in Proposition 3.3, and Lemma 4.4. Figure 4 makes it clear that the two NAL algorithms successfully obtain those hard-to-learn samples, while Random Sampling hardly obtain hard-to-learn samples as it selects samples in a random manner.

E.2. Experiments: Data Model under Other Conditions

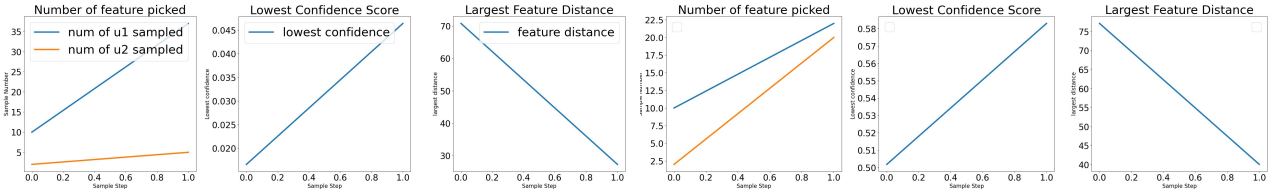


(a) Full trained model (b) Random Sampling

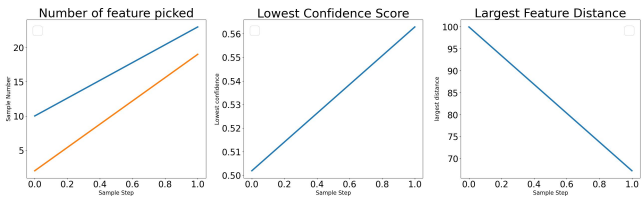


(c) Uncertainty Sampling (d) Diversity Sampling

Figure 5. Learning/memorization progress of features and noise (γ_l represents $\max_{j,k} \gamma_{j,k,l}^{(t)}$, and ρ represents $\max_{j,k,i} j, k, i^{(t)}$), train/test losses, and test accuracy of the full-trained model and the three querying algorithms, with $T^* = 200$, $d = 2000$, $\|\boldsymbol{\mu}_1\| = 8$, $p = p^* = 0.1$, $\|\boldsymbol{\mu}_2\| = 8$, $n_{CNN} = 200$, $n_0 = 10$, $n^* = 30$ and $|\mathcal{P}| = 190$.



(a) Random Sampling (b) Uncertainty Sampling

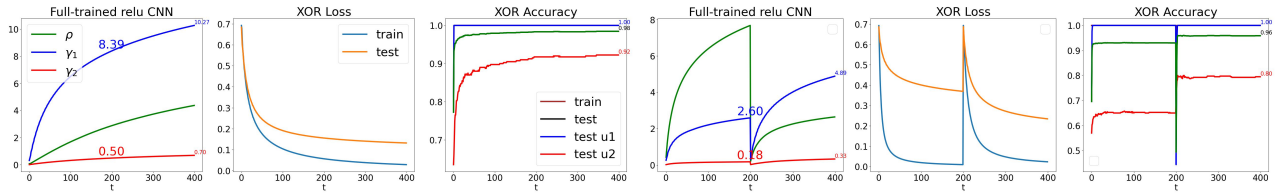


(c) Diversity Sampling

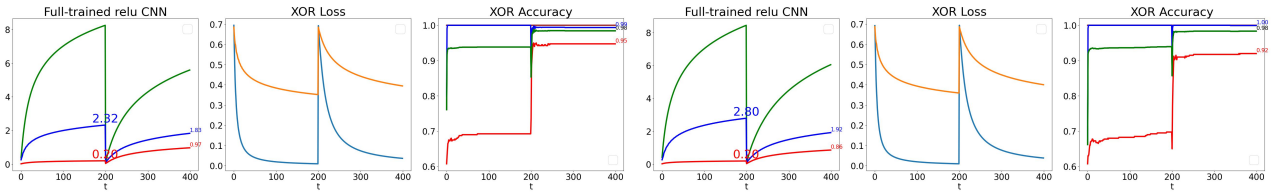
Figure 6. Comparison of querying information between two NAL algorithms, illustrating training size changes in labeled data sets, Confidence Score, and Feature Distance before and after querying. ($T^* = 200$, $d = 2000$, $\|\boldsymbol{\mu}_1\| = 9$, $p = p^* = 0.2$, $\|\boldsymbol{\mu}_2\| = 3$, $n_{CNN} = 200$, $n_0 = 10$, $n^* = 30$ and $|\mathcal{P}| = 190$)

We investigate the scenario where the strengths (i.e., feature norms) of different features do not vary significantly, as discussed in the main body of our work. Specifically, we set them as the same: $\|\boldsymbol{\mu}_1\|_2 = \|\boldsymbol{\mu}_2\|_2 = 8$. Other parameters are listed as the following: $T^* = 200$, $p = p^* = 0.1$, $d = 2000$, $n_{CNN} = 200$, $n_0 = 10$, $n^* = 30$, $|\mathcal{P}| = 190$, $\sigma_p = 1$ and $\sigma_0 = 0.01$. In this case, where $\tau_1\|\boldsymbol{\mu}_1\| < \tau_2\|\boldsymbol{\mu}_2\|$, the **perplexing samples** are those samples equipped with $\boldsymbol{\mu}_2$. It is worth noting that our chosen value of $p = 0.1$ is not small enough to satisfy the condition in Definition D.2. Instead, our parameter setting falls under the second bullet point of the third case discussed in Appendix D.3.2. Figure 5 demonstrates the success of both NAL algorithms, while Figure 6 illustrates the sample information. It is clear that both NAL algorithms prioritize the **perplexing samples** more effectively than Random Sampling, resulting in a lower test error rate.

E.3. Experiments: XOR Data Versions

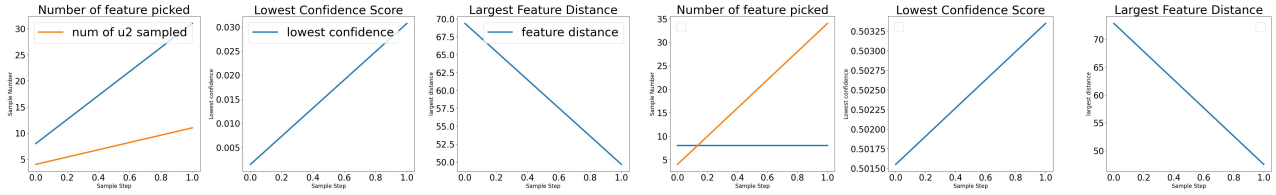


(a) Full-trained 2 layer CNN (b) Random Sampling



(c) Uncertainty Sampling (d) Diversity Sampling

Figure 7. Learning/memorization progress of features and noise (γ_l represents $\max_{j,k}\{\gamma_{j,k,\mathbf{u}_l}^{(t)}, \gamma_{j,k,\mathbf{v}_l}^{(t)}\}$, and ρ represents $\max_{j,k,i} \rho_{j,k,i}^{(t)}$), train/test losses, and test accuracy of the full-trained model and the three querying algorithms, with $\cos\theta = 0.4, T^* = 200$, $d = 2000$, $\|\boldsymbol{\mu}_1\| = 20$, $p = p^* = 0.2$, $\|\boldsymbol{\mu}_2\| = 6$, $n_{CNN} = 200$, $n_0 = 10$, $n^* = 30$ and $|\mathcal{P}| = 190$.



(a) Random Sampling (b) Uncertainty Sampling

(c) Diversity Sampling

Figure 8. Comparison of querying information between two NAL algorithms over XOR data, illustrating training size changes in labeled data sets, Confidence Score, and Feature Distance before and after querying. ($\cos\theta = 0.4, T^* = 200$, $d = 2000$, $\|\boldsymbol{\mu}_1\| = 20$, $p = p^* = 0.2$, $\|\boldsymbol{\mu}_2\| = 6$, $n_{CNN} = 200$, $n_0 = 10$, $n^* = 30$ and $|\mathcal{P}| = 190$)

We also conduct experiments on XOR data. We set the parameters as: $\cos\theta = 0.4, T^* = 200$, $d = 2000$, $\|\boldsymbol{\mu}_1\| = 20$, $p = p^* = 0.2$, $\|\boldsymbol{\mu}_2\| = 6$, $n_{CNN} = 200$, $n_0 = 10$, $n^* = 30$ and $|\mathcal{P}| = 190$. Figure 7 and Figure 8 clearly demonstrate that the two NAL algorithms succeed via prioritizing **perplexing samples**-samples with $\boldsymbol{\mu}_2$ features.

F. Details of Querying Algorithms

F.1. 2-layer ReLU CNN

We adopted the 2-layer ReLU CNN, which is representative for non-linear neural models. Also, this neural setting makes both the model's uncertainty towards samples and the latent feature representation available, paving the way to design NAL algorithms based on this neural settings. The first layer of the model is composed of $2m$ neurons/filters, with m positive and m negative, each of which is applied separately to the two patches $\mathbf{x}_1$ and $\mathbf{x}_2$, with a ReLU function $\sigma(z) = \max\{0, z\}$. Specifically, the parameters of the second pooling layer are set to $+\frac{1}{m}$ and $-\frac{1}{m}$ respectively. The network can thus be expressed as $f(\mathbf{W}, \mathbf{x}) = F_{+1}(\mathbf{W}_{+1}, \mathbf{x}) - F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$, where the partial network functions for positive and negative neurons/filters. For $j \in \{+1, -1\}$, $F_j(\mathbf{W}_j, \mathbf{x})$ is defined as follows:

$$\begin{aligned} F_j(\mathbf{W}_j, \mathbf{x}) &= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_1\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_2\rangle)] \\ &= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{j,r}, y\cdot\boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}\rangle)]. \end{aligned} \tag{8}$$

We denotes $\mathbf{w}_{j,r} \in \mathbb{R}^d$ as the weight vector for the r-th neuron/filter in $\mathbf{W}_j$, where $\mathbf{W}_j$ is the aggregate of model weights associated with F_j filters. We use $\mathbf{W}$ to denote the aggregate of all model weights. Without loss of generality, we let the derivative of the ReLU function at 0 is equal to 1, denoted as $\sigma'(0) = 1$.

F.2. Score and Order of Samples

We claim that the following definitions and lemmas hold for both linearly s

Definition F.1. (Confidence Score) The Confidence Score $C\left(\mathbf{W}^{(t)}, \mathbf{x}\right)$ is defined as below:

$$C\left(\mathbf{W}^{(t)}, \mathbf{x}\right) = \max\left\{\frac{1}{1+\exp\{-y\cdot f\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\}}, 1 - \frac{1}{1+\exp\{-y\cdot f\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\}}\right\} \tag{9}$$

The Confidence Score $C\left(\mathbf{W}^{(t)}, \mathbf{x}\right)$ represents the probability of the predicted label y of logistic loss.

Definition F.2. (Uncertainty Order) We denote the sampling pool as $\mathcal{P}$ that $\mathcal{P} \subsetneq \mathcal{D}$. For $t > 0$, $\forall \mathbf{x}$ and $\mathbf{x}' \in \mathcal{P}$, we define the Uncertainty Order $\prec_C^{(t)}$ and $\preceq_C^{(t)}$, which denote the order of the model's uncertainty upon its prediction upon $\mathbf{x}$ and $\mathbf{x}'$ at the time step t:

$$\begin{aligned} \mathbf{x} \prec_C^{(t)} \mathbf{x}' \text{ if } C\left(\mathbf{W}^{(t)}, \mathbf{x}\right) > C\left(\mathbf{W}^{(t)}, \mathbf{x}'\right), \\ \mathbf{x} \preceq_C^{(t)} \mathbf{x}' \text{ if } C\left(\mathbf{W}^{(t)}, \mathbf{x}\right) \geq C\left(\mathbf{W}^{(t)}, \mathbf{x}'\right). \end{aligned} \tag{10}$$

We say the model uncertainty at time step t upon $\mathbf{x}$ is less than $\mathbf{x}'$ if $\mathbf{x} \prec_C^{(t)} \mathbf{x}'$. Specifically, if the model's uncertainty towards its predictions upon all elements in a set $\mathbf{X}$ at time step t are all less than those in the set $\mathbf{X}'$, we utilize the same notation to describe the Uncertainty Order at time step t between sets: $\mathbf{X} \prec_C^{(t)} \mathbf{X}'$.

Lemma F.3. *The Uncertainty Order is a full order. In addition, for $\forall \mathbf{x}$ and $\mathbf{x}' \in \mathcal{P}$, at $t > 0$ we have:*

$$\mathbf{x} \preceq_C^{(t)} \mathbf{x}' \Leftrightarrow \left|f\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right| \geq \left|f\left(\mathbf{W}^{(t)}, \mathbf{x}'\right)\right| \tag{11}$$

Proof.

$$\begin{aligned}
&\mathbf{x} \preceq_C^{(t)} \mathbf{x}' \Leftrightarrow C\left(\mathbf{W}^{(t)}, \mathbf{x}\right) \geq C\left(\mathbf{W}^{(t)}, \mathbf{x}'\right) \\
&\Leftrightarrow \frac{1}{1+\exp\{|f(\mathbf{W}, \mathbf{x})|\}} \geq \frac{1}{1+\exp\left\{-\left|f\left(\mathbf{W}^{(t)}, \mathbf{x}'\right)\right|\right\}} \\
&\Leftrightarrow \left|f\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right| \geq \left|f\left(\mathbf{W}^{(t)}, \mathbf{x}'\right)\right|
\end{aligned}$$

□

As one can always get $f\left(\mathbf{W}^{(t)}, \mathbf{x}\right) \in \mathbb{R}$ by a given $\mathbf{x}$ at time step t, the Uncertainty Order is a full order.

In Lemma F.5, we will show that sampling based on the Uncertainty Order is equivalent to various typical sampling methods based on the score functions defined in many typical Model Uncertainty-based Approaches, such as Least Confidence (Lewis and Catlett, 1994), Margin Roth and Small (2006) and Entropy (Joshi et al., 2009) methods under our data model scenario, thus it's representative to the main idea of the approaches family while elegant.

Definition F.4. The following are the definitions of the score functions of LeastConf (Lewis and Catlett, 1994), Margin Roth and Small (2006) and Entropy (Joshi et al., 2009).

- Least Confidence selects data points whose predicted label y have the lowest posterior probability, so the score function of LeastConf is:
$$Score(\mathbf{W}^{(t)}, \mathbf{x}) = -P(y|\mathbf{x}, \mathbf{W}^{(t)}), \tag{12}$$
- The score function of Margin is:
$$Score(\mathbf{W}^{(t)}, \mathbf{x}) = -[p(y|\mathbf{x}, \mathbf{W}^{(t)}) - P(-y|\mathbf{x}, \mathbf{W}^{(t)})], \tag{13}$$
- The score function of Entropy is:
$$Score(\mathbf{W}^{(t)}, \mathbf{x}) = -[P(y|\mathbf{x}, \mathbf{W}^{(t)}) \log P(y|\mathbf{x}, \mathbf{W}^{(t)}) + P(-y|\mathbf{x}, \mathbf{W}^{(t)}) \log P(-y|\mathbf{x}, \mathbf{W}^{(t)})], \tag{14}$$

Lemma F.5. *Sampling based on the score functions defined in (12), (13) and (14) are equivalent to sampling based on the Confidence Order in Definition F.2.*

Proof. By definitions, $C\left(\mathbf{W}^{(t)}, \mathbf{x}\right) = P(y|\mathbf{x}, \mathbf{W}^{(t)}) = -Score(\mathbf{W}^{(t)}, \mathbf{x})$, showing the equivalence of LeastConf methods and ours. Then by Lemma F.3 and the property: $P(-y|\mathbf{x}, \mathbf{W}^{(t)}) = 1 - C\left(\mathbf{W}^{(t)}, \mathbf{x}\right)$, it's easy to verify that $\left|f\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right| \propto C\left(\mathbf{W}^{(t)}, \mathbf{x}\right) \propto [C\left(\mathbf{W}^{(t)}, \mathbf{x}\right) - (1 - C\left(\mathbf{W}^{(t)}, \mathbf{x}\right))]$, and $\left|f\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right| \propto C\left(\mathbf{W}^{(t)}, \mathbf{x}\right) \propto [C\left(\mathbf{W}^{(t)}, \mathbf{x}\right) \log C\left(\mathbf{W}^{(t)}, \mathbf{x}\right) + (1 - C\left(\mathbf{W}^{(t)}, \mathbf{x}\right)) \log(1 - C\left(\mathbf{W}^{(t)}, \mathbf{x}\right))]$. Therefore, the priority order of the samples based on those score functions are the same as the Uncertainty Order, thus the proof is completed. □

Definition F.6. (Feature Distance) The latent feature representation of a sample $\mathbf{x}$=$[\mathbf{x}_1^T, \mathbf{x}_2^T]^T$ in the latent feature space $\mathcal{Z} \subseteq \mathbb{R}^m$ of our ReLU CNN at timestep t is:

$$\mathbf{Z}(\mathbf{x}, t) = \sum_j (\sigma(\langle \mathbf{W}_j^{(t)}, \mathbf{x}_1 \rangle)) + \sigma(\langle \mathbf{W}_j^{(t)}, \mathbf{x}_2 \rangle))$$

Apparently $\mathbf{Z}(\mathbf{x}, t) \in \mathbb{R}^m$. The Feature Distance is measured by the l_p ($p \in [1, \infty)$) distance between sample's feature representation and the average feature representation of the current labeled set $\mathcal{D}_n := \{\mathbf{x}^{(i)}\}_{i=1}^n$:

$$D\left(\mathbf{W}^{(t)}, \mathbf{x} \mid \mathcal{D}_n\right) = \|\mathbf{Z}(\mathbf{x}, t) - \mathop{\mathbb{E}}_{\mathbf{x}^{(i)} \in \mathcal{D}_n} \mathbf{Z}(\mathbf{x}^{(i)}, t)\|_p \tag{15}$$

Definition F.7. (Diversity Order) Similar to Definition F.2, we defined Diversity Order $\prec_D^{(t)}$, $\preceq_D^{(t)}$ based on Feature Distance $D\left(\mathbf{W}^{(t)}, \mathbf{x} \mid \mathcal{D}_n\right)$. Borrowing the same notations in Definition F.2, we have:

$$\begin{aligned}
&\mathbf{x} \prec_D^{(t)} \mathbf{x}' \text{ if } D\left(\mathbf{W}^{(t)}, \mathbf{x} \mid \mathcal{D}_n\right) < D\left(\mathbf{W}^{(t)}, \mathbf{x}' \mid \mathcal{D}_n\right), \\
&\mathbf{x} \preceq_D^{(t)} \mathbf{x}' \text{ if } D\left(\mathbf{W}^{(t)}, \mathbf{x} \mid \mathcal{D}_n\right) \leq D\left(\mathbf{W}^{(t)}, \mathbf{x}' \mid \mathcal{D}_n\right).
\end{aligned} \tag{16}$$

Along with Definition F.2, we also have set-level notations such that $\mathbf{X} \prec_D^{(t)} (\preceq_D^{(t)}) \mathbf{X}'$. Based on the triangle inequality for the l_p norm and (15), we can easily draw the conclusion that the Diversity Order is also a full order. Furthermore, in the case that both $\mathbf{x} \prec_C^{(t)} (\preceq_C^{(t)}) \mathbf{x}'$ and $\mathbf{x} \prec_D^{(t)} (\preceq_D^{(t)}) \mathbf{x}', \forall p \in [1, \infty)$ hold, we denote the order relationship using $\prec^{(t)}$ $(\preceq^{(t)})$, such that $\mathbf{x} \prec^{(t)}$ $(\preceq^{(t)})$ $\mathbf{x}'$.

G. Proofs of Main Results

In this section, we denote n as the number of training data in current labeled training set, which is n_0 at initial stage and n_1 after sampling (querying). Besides, we denote the proportion of easy-to-learn data in current labeled set as τ_1, and utilize τ_2 to represent the proportion of hard-to-learn data in current labeled set for notation simplicity. Notably, we can use the same techniques in Cao et al. (2022a); Kou et al. (2023b); Meng et al. (2023); Lu et al. (2023) to achieve some statistical outcomes that are not directly related to our main contribution, we exclude the proof details for those outcomes. Instead, our focus is on providing comprehensive proofs of our primary contribution.

G.1. Preliminary Lemmas

The following lemmas give finite-sample concentration results to characterize the statistical properties of the random elements involved in our problem, and hold both under the linearly separable data and XOR data (i.e., $\boldsymbol{\mu}_l \in \{\boldsymbol{\mu}_l, \mathbf{u}_l, \mathbf{v}_l\}, \forall l\{1, 2\}$).

Lemma G.1. *Suppose that $\delta > 0$ and $d = \Omega(\log(\frac{6n}{\delta}))$. Then with probability at least $1 - \delta$,*

$$\frac{\sigma_p^2 d}{2} \leq \|\boldsymbol{\xi}_i\|_2^2 \leq 3\frac{\sigma_p^2 d}{2},$$
$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \leq 2\sigma_p^2 \cdot \sqrt{d \log\left(\frac{6n^2}{\delta}\right)},$$
$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\mu}_l \rangle| \leq \|\boldsymbol{\mu}_l\|_2 \sigma_p \cdot \sqrt{2\log(\frac{12n}{\delta})}$$

for all $i, i' \in [n], l \in \{1, 2\}$.

Proof of Lemma G.1. The proof can be found in Lemma B.2 in Cao et al. (2022a), Lemma B.4 in Kou et al. (2023b), Lemma B.3 in Meng et al. (2023) or Lemma A.3 in Lu et al. (2023).

Lemma G.2. *Suppose that $\delta > 0, d = \Omega(\log(\frac{mn}{\delta}))$, and $m = \Omega(\log(\frac{1}{\delta}))$. Then with probability at least $1 - \delta$,*

$$\frac{\sigma_0^2 d}{2} \leq \|\mathbf{w}_{j,r}^{(0)}\|_2^2 \leq 3\frac{\sigma_0^2 d}{2},$$
$$\left|\left\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_l \right\rangle\right| \leq \sqrt{2\log(\frac{16m}{\delta})} \cdot \sigma_0 \|\boldsymbol{\mu}_l\|_2,$$
$$\left|\left\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \right\rangle\right| \leq 2\sqrt{\log\frac{16mn}{\delta}} \cdot \sigma_0 \sigma_p \sqrt{d}$$

for all $r \in [m], j \in \{\pm 1\}, l \in \{1, 2\}$ and $i \in [n]$. Moreover,

$$\frac{\sigma_0 \|\boldsymbol{\mu}_l\|_2}{2} \leq \max_{r \in [m]} j \cdot \left\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_l \right\rangle \leq \sqrt{2\log(\frac{16m}{\delta})} \cdot \sigma_0 \|\boldsymbol{\mu}_l\|_2,$$
$$\frac{\sigma_0 \sigma_p \sqrt{d}}{4} \leq \max_{r \in [m]} j \cdot \left\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \right\rangle \leq 2\sqrt{\log\frac{16mn}{\delta}} \cdot \sigma_0 \sigma_p \sqrt{d}$$

for all $j \in \{\pm 1\}, l \in \{1, 2\}$ and $i \in [n]$.

Proof of Lemma G.2. The proof can be found in Lemma B.3 in Cao et al. (2022a), Lemma B.5 in Kou et al. (2023b), Lemma B.4 in Meng et al. (2023) or Lemma A.4 and Lemma C.1 in Lu et al. (2023).

Next, we utilize the property of binomial tails to examine the proportion of hard-to-learn data within the subsets generated from the data distribution $\mathcal{D}$ (i.e., the initial labeled set $\mathcal{D}_{n_0} := \{\mathbf{x}^{(i)}\}_{i=1}^{n_0} \subseteq \mathcal{D}$, the sampling pool $\mathcal{P} \subseteq \mathcal{D}$, and the final labeled set $\mathcal{D}_{n_1}^{(random)} := \{\mathbf{x}^{(random)(i)}\}_{i=1}^{n_1} \subseteq \mathcal{D}$ obtained through Random Sampling).

Lemma G.3. *Suppose that $\delta > 0$, $n_0, \tilde{n}, |P| = \Omega\left(\frac{1-p}{p}\log\left(\frac{1}{\delta}\right)\right)$, then for $n \in \{n_0, |P|, n_1\}$. Denote $n_p \leq n$ as the number of hard-to-learn data among n, then with probability at least $1-\delta$. We have*

$$\frac{1}{2}p \cdot n \leqslant n_p \leqslant \frac{3}{2}p \cdot n \tag{17}$$

proof of Lemma G.3. We can see n_p as a binomial random variable with probability p and number of experiments n. By Exercise 2.9.(a) in Wainwright (2019), we have

$$P\left(\frac{pn}{2} \leq n_p \leq \frac{3pn}{2}\right) \geqslant 1 - 2e^{-nD\left(\frac{p}{2}\|p\right)}$$

where the quantity $D(\delta\|\alpha)$ for $\forall \delta, \alpha \in \left(0, \frac{1}{2}\right]$ is defined as

$$D(\delta\|\alpha) := \delta \log\left(\frac{\delta}{\alpha}\right) + (1-\delta)\log\left(\frac{1-\delta}{1-\alpha}\right).$$

Since $\frac{p}{2} < p$. By Exercise 2.9.(b) in Wainwright (2019), we can obtain $P\left(\frac{pn}{2} \leq n_p \leq \frac{3pn}{2}\right) \geq 1-\delta$ directly by Hoeffding Inequality.

Remark G.4. It is important to note that the generation of $\mathcal{D}_{n_0}$ and $\mathcal{P}$ through sampling from $\mathcal{D}$ is independent. However, the generation of $\mathcal{D}_{n_1}^{(random)}$ is based on $\mathcal{D}_{n_0}$ and $\mathcal{P}$. In our analysis, instead of considering martingale with the perspective of conditional probability, we consider the overall process of the labeled set obtained by Random Sampling, where $\mathcal{D}_{n_1}^{(random)}$ is directly sampled from $\mathcal{D}$.

G.2. Coefficient Ratio and Scale Analysis

In this section, we provide lemmas that characterize the behavior of coefficients under gradient descent. Subsequently, we establish the scale of the coefficients in the training dynamics. It's worth noting that in this section we assume the results in Appendix G.1 all hold with high probability.

Definition G.5. (Equivalent techniques to Definition 4.1 in Cao et al. (2022a), Definition 5.1 in Kou et al. (2023b)) Denote $\mathbf{w}_{j,r}^{(t)}$ for $j \in \{\pm 1\}, r \in [m]$ as the convolution neurons/filters at the tth timestep of gradient descent, then there exist unique coefficients $\gamma_{j,r,l}^{(t)}$ and $\rho_{j,r,i}^{(t)}$ such that

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \sum_{l=1}^{2} \gamma_{j,r,l}^{(t)} \cdot \frac{\boldsymbol{\mu}_l}{\|\boldsymbol{\mu}_l\|_2^2} + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2}$$

Further denote $\bar{\rho}_{j,r,i}^{(t)}$ as $\rho_{j,r,i}^{(t)} \mathbb{1}\left(\rho_{j,r,i}^{(t)} \geq 0\right)$, $\underline{\rho}_{j,r,i}^{(t)}$ as $\rho_{j,r,i}^{(t)} \mathbb{1}\left(\rho_{j,r,i}^{(t)} \leq 0\right)$. Then:

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \sum_{l=1}^{2} \cdot \gamma_{j,r,l}^{(t)} \cdot \frac{\boldsymbol{\mu}_l}{\|\boldsymbol{\mu}_l\|_2^2} + \sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2}. \tag{18}$$

We denote $U_l = \left\{i \in [n] : \mathbf{x}^{(i)} = [y_i \cdot \boldsymbol{\mu}_l, \boldsymbol{\xi}_i]\right\}$, for $l \in \{1, 2\}$. The following lemma presents the update rule of coefficients.

Lemma G.6. *The coefficients $\gamma_{j,r,l}^{(t)}, \bar{\rho}_{j,r,i}^{(t)}, \underline{\rho}_{j,r,i}^{(t)}$ defined in Definition G.5 satisfy the following iterative equations:*

$$\begin{aligned}
&\gamma_{j,r,l}^{(0)}, \bar{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0, \\
&\gamma_{j,r,l}^{(t+1)} = \gamma_{j,r,l}^{(t)} - \frac{\eta}{nm} \cdot \sum_{i \in U_l} \ell_i'^{(t)} \sigma' \left(\left\langle \mathbf{w}_{j,r}^{(t)}, y_i \cdot \boldsymbol{\mu}_l \right\rangle \right) \cdot \|\boldsymbol{\mu}_l\|_2^2, \\
&\bar{\rho}_{j,r,i}^{(t+1)} = \bar{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma' \left(\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle \right) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}\left(y_i = j\right), \\
&\underline{\rho}_{j,r,i}^{(t+1)} = \underline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma' \left(\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle \right) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}\left(y_i = -j\right),
\end{aligned}$$

for all $r \in [m], j \in \{\pm 1\}, l \in \{1, 2\}$ and $i \in [n]$.

Remark G.7. This lemma serves as a cornerstone in our analysis of dynamics. Originally, the study of neural network dynamics under gradient descent required us to track the variations in weights. However, this Lemma enables us to view these dynamics from a new perspective, focusing on two distinct elements: feature learning (represented by $\gamma_{j,r,l}^{(t+1)}$) and noise memorization (represented by $\rho_{j,r,i}^{(t+1)}$). We can easily observe that the $\gamma_{j,r,l}^{(t)}$ is strictly increasing since $\ell_i'^{(t)}$ is strictly negative.

Proof of Lemma G.6. Applying the gradient descent rule in (2), we get

$$\begin{aligned}
\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(0)} &- \frac{\eta}{nm} \sum_{s=0}^{t} \sum_{i=1}^{n} \ell_i'^{(s)} \cdot \sigma' \left(\left\langle \mathbf{w}_{j,r}^{(s)}, \boldsymbol{\xi}_i \right\rangle \right) \cdot j y_i \boldsymbol{\xi}_i \\
&- \frac{\eta}{nm} \sum_{s=0}^{t} \sum_{i=1}^{n} \ell_i'^{(s)} \cdot \sigma' \left(\left\langle \mathbf{w}_{j,r}^{(s)}, y_i \boldsymbol{\mu}_l \right\rangle \right) \cdot j \boldsymbol{\mu}_l.
\end{aligned}$$

Based on the definition of $\gamma_{j,r,l}^{(t)}$ and $\rho_{j,r,i}^{(t)}$, we consider $\gamma_{j,r,l}^{(0)}, \bar{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0$ and

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \sum_{l=1}^{2} \gamma_{j,r,l}^{(t)} \cdot \|\boldsymbol{\mu}_l\|_2^{-2} \cdot \boldsymbol{\mu}_l + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i.$$

Note that $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\xi}_i$ are linearly independent with probability 1, thus we have the following unique representation

$$\begin{aligned}
\gamma_{j,r,l}^{(t)} &= -\frac{\eta}{nm} \sum_{s=0}^{t} \sum_{i \in U_l} \ell_i'^{(s)} \cdot \sigma' \left(\left\langle \mathbf{w}_{j,r}^{(s)}, y_i \boldsymbol{\mu}_l \right\rangle \right) \cdot \|\boldsymbol{\mu}_l\|_2^2, \\
\rho_{j,r,i}^{(t)} &= -\frac{\eta}{nm} \sum_{s=0}^{t} \ell_i'^{(s)} \cdot \sigma' \left(\left\langle \mathbf{w}_{j,r}^{(s)}, \boldsymbol{\xi}_i \right\rangle \right) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot j y_i.
\end{aligned}$$

Recall $U_l = \left\{ i \in [n] : \mathbf{x}^{(i)} = [y_i \cdot \boldsymbol{\mu}_l, \boldsymbol{\xi}_i] \right\}$, we have

$$\gamma_{j,r,l}^{(t)} = -\frac{\eta}{nm} \sum_{s=0}^{t} \sum_{i \in U_l} \ell_i'^{(s)} \cdot \sigma' \left(\left\langle \mathbf{w}_{j,r}^{(s)}, y_i \boldsymbol{\mu}_l \right\rangle \right) \cdot \|\boldsymbol{\mu}_l\|_2^2. \tag{19}$$

Now with the notation $\bar{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}\left(\rho_{j,r,i}^{(t)} \geq 0\right), \underline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}\left(\rho_{j,r,i}^{(t)} \leq 0\right)$ and the fact $\ell_i'^{(s)} < 0$, we get

$$\bar{\rho}_{j,r,i}^{(t)} = -\frac{\eta}{nm} \sum_{s=0}^{t} \ell_i'^{(s)} \cdot \sigma' \left(\left\langle \mathbf{w}_{j,r}^{(s)}, \boldsymbol{\xi}_i \right\rangle \right) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}\left(y_i = j\right), \tag{20}$$

$$\underline{\rho}_{j,r,i}^{(t)} = \frac{\eta}{nm} \sum_{s=0}^{t} \ell_i'^{(s)} \cdot \sigma' \left(\left\langle \mathbf{w}_{j,r}^{(s)}, \boldsymbol{\xi}_i \right\rangle \right) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}\left(y_i = -j\right). \tag{21}$$

The proof is completed.

Remark G.8. The proof strategy employed in this study follows the study of feature learning analysis techniques in Cao et al. (2022a); Kou et al. (2023b); Meng et al. (2023). However, our decomposition considers two task-specific features with different proportion. This disparity would finally lead to distinct learning efficiency among samples, as well as different generalization ability.

Next, we're dedicated to explore range scale evolution of the coefficients in the signal-noise decomposition. Let $T^* = \eta^{-1}$ poly $\left(\varepsilon^{-1}, d, n, m\right)$ be the maximum admissible iteration. Denote

$$\begin{aligned} &\alpha := 4\log\left(T^*\right), \\ &\beta := 2\max_{l,i,j,r}\left\{\left|\left\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_l\right\rangle\right|, \left|\left\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\right\rangle\right|\right\}, \\ &\mathrm{SNR}_l := \frac{\|\boldsymbol{\mu}_l\|_2}{\sigma_p\sqrt{d}}. \end{aligned} \tag{22}$$

By Lemma G.2, β can be bounded by $4\sigma_0 \cdot \max\left\{\sqrt{\log\frac{16mn}{\delta}} \cdot \sigma_p\sqrt{d}, \sqrt{\log(\frac{16m}{\delta})} \cdot \|\boldsymbol{\mu}_l\|_2\right\}$. Under Condition 3.1, it is straightforward to verify the following inequality with a large constant C:

$$\max_l\left\{\beta, \mathrm{SNR}_l\sqrt{\frac{32\log(\frac{12n}{\delta})}{d}}n\alpha, 5\sqrt{\frac{\log\left(\frac{6n^2}{\delta}\right)}{d}}n\alpha\right\} \leq \frac{1}{12}. \tag{23}$$

We then assert the following proposition hold for the entire training period. This proposition serves to show the evolution scale of the coefficients.

Proposition G.9. *Under Condition 3.1, for $0 \leq t \leq T^*$, there exists a positive constant C' such that*

$$\begin{aligned} &0 \leq \gamma_{j,r,l}^{(t)} \leq C' \cdot \tau_l n \cdot \mathrm{SNR}_l^2 \cdot \alpha \\ &0 \leq \bar{\rho}_{j,r,i}^{(t)} \leq \alpha, \\ &0 \geq \underline{\rho}_{j,r,i}^{(t)} \geq -\beta - 10\sqrt{\frac{\log\left(\frac{6n^2}{\delta}\right)}{d}}n\alpha \geq -\alpha, \end{aligned} \tag{24}$$

for all $j \in \{\pm 1\}, r \in [m], l \in \{1, 2\}$ and $i \in [n]$.

Remark G.10. Our results resemble those in the study of feature learning of CNN (Cao et al., 2022a; Kou et al., 2023b; Meng et al., 2023; Lu et al., 2023). However, the scale of our learning progress coefficient $\gamma_{j,r,l}^{(t)}$ depends on its corresponding feature proportion and strength in the labeled data distribution, which will significantly impact the learning process of specific type of data.

Proof of Proposition G.9. See Proposition C.2. and Proposition C.8. in Kou et al. (2023b) or Proposition C.2 and Proposition C.8 in Meng et al. (2023) for a proof. Regardless of the variations in data settings, obtaining the result through inductive techniques is readily feasible.

Based on Proposition G.9, we can analyze the convergence of the training dynamics via identifying the degree of feature learning and noise memorization in the following section.

G.3. Feature Learning and Noise Memorization Analysis

In this section, we adopt a two-stage analysis to evaluate the evolution of the coefficients. In the first stage, the loss function's derivative remains nearly constant due to the small weight initialization. However, in the subsequent stage, the derivative of the loss function becomes non-constant, requiring a careful analysis to address this change. We will see that the scale differences in the first stage remain the same. Worth noting that the results in this section are based on the previous results in Appendix G.2 holding with high probability.

G.3.1. First Stage: Feature Learning versus Noise Memorization

Lemma G.11. *There exist*

$$T_1 = C_3\eta^{-1}nm\sigma_p^{-2}d^{-1}, T_2 = C_4\eta^{-1}nm\sigma_p^{-2}d^{-1}$$

where $C_3 = \Theta(1)$ is a large constant and $C_4 = \Theta(1)$ is a small constant, such that

- $\max_{j,r}\gamma_{j,r,l}^{(t)} = O(\tau_l n\cdot \mathrm{SNR}_l^2)$ *, for all* $0\le t\le T_1, l\in\{1,2\}$.
- $\min_{j,r}\gamma_{j,r,l}^{(t)} = \Omega(\tau_l n\cdot \mathrm{SNR}_l^2)$*, for all* $t\ge T_2, l\in\{1,2\}$.
- $\bar{\rho}_{j,r^*,i}^{(T_1)}\ge 2$*, for any* $r^*\in S_i^{(0)} = \left\{r\in[m]:\left\langle \mathbf{w}_{y_i,r}^{(0)},\boldsymbol{\xi}_i\right\rangle>0\right\}, j\in\{\pm1\}$ *and* $i\in[n]$ *with* $y_i=j$.
- $\max_{j,r,i}\left|\underline{\rho}_{j,r,i}^{(t)}\right| = \max\left\{O\left(\sqrt{\log(\frac{mn}{\delta})}\cdot\sigma_0\sigma_p\sqrt{d}\right), O\left(n\sqrt{\log(\frac{n}{\delta})}\log\left(T^*\right)/\sqrt{d}\right)\right\}$*, for all* $0\le t\le T_1$.
- $\max_{j,r}\bar{\rho}_{j,r,i}^{(T_1)} = O(1)$*, for all* $i\in[n]$.

Proof of Lemma G.11. See Lemma D.1. in Kou et al. (2023b) or Lemma D.1, Proposition D.2-D.4 in Meng et al. (2023) for a proof.

G.3.2. Second Stage: Convergence of Training Error

At the end of the first stage, we have the following feature-to-noise decomposition:

$$\mathbf{w}_{j,r}^{(T_1)} = \mathbf{w}_{j,r}^{(0)} + j\cdot\sum_{l=1}^{2}\gamma_{j,r,l}^{(T_1)}\cdot\frac{\boldsymbol{\mu}_l}{\|\boldsymbol{\mu}_l\|_2^2} + \sum_{i=1}^{n}\bar{\rho}_{j,r,i}^{(T_1)}\cdot\frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} + \sum_{i=1}^{n}\underline{\rho}_{j,r,i}^{(T_1)}\cdot\frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2}$$

for $j\in[\pm1]$ and $r\in[m]$. Applying the results we obtain in the first stage, we have the following property holds at the beginning of this stage:

- $\gamma_{j,r,l}^{(T_1)} = \tau_l n\cdot\mathrm{SNR}_l^2)$ for any $j\in\{\pm1\}, r\in[m]$.
- $\bar{\rho}_{j,r^*,i}^{(T_1)}\ge 2$ for any $r^*\in S_i^{(0)} = \left\{r\in[m]:\left\langle \mathbf{w}_{y_i,r}^{(0)},\boldsymbol{\xi}_i\right\rangle>0\right\}, j\in\{\pm1\}$ and $i\in[n]$ with $y_i=j$.
- $\max_{j,r,i}\left|\underline{\rho}_{j,r,i}^{(T_1)}\right| = \max\left\{O\left(\sqrt{\log(\frac{mn}{\delta})}\cdot\sigma_0\sigma_p\sqrt{d}\right), O\left(n\sqrt{\log(\frac{n}{\delta})}\log\left(T^*\right)/\sqrt{d}\right)\right\}$.

Following the technique in Cao et al. (2022a), now we choose $\mathbf{W}^*$ as follows

$$\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^{(0)} + 5\log(\frac{2}{\varepsilon})\left[\sum_{i=1}^{n}\mathbb{1}\left(j=y_i\right)\cdot\frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2}\right].$$

Lemma G.12. *Under Condition 3.1, we have*

$$\max_{j,r,i}\left|\underline{\rho}_{j,r,i}^{(t)}\right| = \max\left\{O\left(\sqrt{\log(\frac{mn}{\delta})}\cdot\sigma_0\sigma_p\sqrt{d}\right), O\left(n\sqrt{\log(\frac{n}{\delta})}\log\left(T^*\right)/\sqrt{d}\right)\right\},$$

for all $T_1\le t\le T^*$. Besides,

$$\frac{1}{t-T_1+1}\sum_{s=T_1}^{t}L_S\left(\mathbf{W}^{(s)}\right)\le\frac{\left\|\mathbf{W}^{(T_1)}-\mathbf{W}^*\right\|_F^2}{\eta\left(t-T_1+1\right)}+\varepsilon$$

for all $T_1\le t\le T^*$. Therefore, we can find an iterate with training loss smaller than 2ε within $T = T_1 + \left|\left\|\mathbf{W}^{(T_1)}-\mathbf{W}^*\right\|_F^2/(\eta\varepsilon)\right| = T_1 + \widetilde{O}\left(\eta^{-1}\varepsilon^{-1}mnd^{-1}\sigma_p^{-2}\right)$ iterations.

Proof of Lemma G.12. See Lemma D.5 in Cao et al. (2022a) or Lemma D.6. in Kou et al. (2023b) for a proof.

Worth noting that since the n could be n_0 or n_1 and the τ_l could be any real number denoting the proportion of specific types of data in the labeled set, we have successfully concluded the proof of training loss convergence for all three querying algorithms. The following lemma characterized the feature-to-noise ratio during the whole duration.

Lemma G.13. *Under Condition 3.1, we have*

$$\sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)} / \gamma_{j',r',l}^{(t)} = \Theta\left(\tau_l^{-1} \cdot \mathrm{SNR}_l^{-2}\right)$$

for all $j, j' \in \{\pm 1\}, r, r' \in [m], l \in \{1, 2\}$ and $0 \le t \le T^*$.

Proof of Lemma G.13. See Lemma D.7. in Kou et al. (2023b) or Proposition C.8 in Meng et al. (2023) for a proof.

Now we can summarize current results into the following lemma.

Lemma G.14. *(Formal restatement of Lemma 4.1) Under Condition 3.1, there exists $T_1 = \Theta(\eta^{-1} n m \sigma_p^2 d^{-1})$, for $t \in [T_1, T^*]$ we have the following hold:*

- $\gamma_{j,r,l}^{(t)} = \Theta\left(\frac{\tau_l \|\boldsymbol{\mu}_l\|_2^2}{d\sigma_p^2}\right) \sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)}$, *for all $j \in \{\pm 1\}, r \in [m]$ and $l \in \{1, 2\}$ (from Lemma G.13).*
- $\sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)} = \Omega(n) = O(n \log(T^*)) = \widetilde{\Theta}(n)$, *for all $j \in \{\pm 1\}, r \in [m]$ and $l \in \{1, 2\}$ (from Proposition G.9 and Lemma G.11).*
- $\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(t)}| = \max\{O(\sigma_0 \sigma_p \sqrt{d} \cdot \sqrt{\log(\frac{mn}{\delta})}), O(\sqrt{\log(\frac{n}{\delta})} \log(T^*) \cdot n/\sqrt{d})\}$, *for all $j \in \{\pm 1\}, r \in [m]$ and $l \in \{1, 2\}$ (from Lemma G.12).*

Lemma G.15. *Under Condition 3.1, there exists $t = \widetilde{O}\left(\eta^{-1} \varepsilon^{-1} m n d^{-1} \sigma_p^{-2}\right)$, we have:*

$$\begin{aligned} \left\|\mathbf{w}_{j,r}^{(t)}\right\|_2 &\le \Theta\left(\sigma_p^{-1} d^{-\frac{1}{2}} n^{\frac{1}{2}}\right), \\ \left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_l \right\rangle &= \Theta\left(\gamma_{y,r,l}^{(t)}\right), \\ \left\langle \mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu}_l \right\rangle &= -\Theta\left(\gamma_{-y,r,l}^{(t)}\right) < 0. \end{aligned} \tag{25}$$

for all $j \in \{\pm 1\}, r \in [m]$ and $l \in \{1, 2\}$.

Proof of Lemma G.15. Recall the signal-noise decomposition of $\mathbf{w}_{j,r}^{(t)}$:

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \sum_{l=1}^{2} \gamma_{j,r,l}^{(t)} \cdot \frac{\boldsymbol{\mu}_l}{\|\boldsymbol{\mu}_l\|_2^2} + \sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2}.$$

For $l \in \{1, 2\}$, we can bound the inner product with $j = y$:

$$\begin{aligned} \left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_l \right\rangle =& \left\langle \mathbf{w}_{y,r}^{(0)}, y\boldsymbol{\mu}_l \right\rangle + \gamma_{y,r,l}^{(t)} + \sum_{i=1}^{n} \bar{\rho}_{y,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \langle \boldsymbol{\xi}_i, y\boldsymbol{\mu}_l \rangle + \sum_{i=1}^{n} \underline{\rho}_{y,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \langle \boldsymbol{\xi}_i, y\boldsymbol{\mu}_l \rangle \\ \ge& \gamma_{y,r,l}^{(t)} - \sqrt{2\log(\frac{16m}{\delta})} \cdot \sigma_0 \|\boldsymbol{\mu}_l\|_2 - \sqrt{2\log(\frac{12n}{\delta})} \cdot \sigma_p \|\boldsymbol{\mu}_l\|_2 \cdot \left(\frac{\sigma_p^2 d}{2}\right)^{-1} \left[\sum_{i=1}^{n} \bar{\rho}_{y,r,i}^{(t)} + \sum_{i=1}^{n} |\underline{\rho}_{y,r,i}^{(t)}|\right] \\ =& \gamma_{y,r,l}^{(t)} - \Theta\left(\sqrt{\log(\frac{m}{\delta})} \sigma_0 \|\boldsymbol{\mu}_l\|_2\right) - \Theta\left(\sqrt{\log(\frac{n}{\delta})} \cdot (\sigma_p d)^{-1} \|\boldsymbol{\mu}_l\|_2\right) \cdot \Theta\left(\mathrm{SNR}_l^{-2}\right) \cdot \gamma_{y,r,l}^{(t)} \\ =& \left[1 - \Theta\left(\sqrt{\log(\frac{n}{\delta})} \cdot \sigma_p / \|\boldsymbol{\mu}_l\|_2\right)\right] \gamma_{y,r,l}^{(t)} - \Theta\left(\sqrt{\log(\frac{m}{\delta})} (\sigma_p d)^{-1} \sqrt{n} \|\boldsymbol{\mu}_l\|_2\right) \\ =& \Theta\left(\gamma_{y,r,l}^{(t)}\right), \end{aligned} \tag{26}$$

where the inequality is justified by Lemma G.1 and Lemma G.2. The second equality is obtained by substituting the coefficient scales in G.14. The third equality follows from the condition $\sigma_0 \leq C^{-1}(\sigma_p d)^{-1}\sqrt{n}$ in Condition 3.1 and the feature-to-noise ratio $\mathrm{SNR}_l = \frac{\|\boldsymbol{\mu}_l\|_2}{\sigma_p\sqrt{d}}$. For the fourth equality, it should be noted that $\gamma_{j,r,l}^{(t)} = \Omega(\tau_l n \cdot \mathrm{SNR}_l^2)$, and also $\sqrt{\log(\frac{n}{\delta})} \cdot \frac{\sigma_p}{\|\boldsymbol{\mu}_l\|_2} \leq 1/\sqrt{C}$ and $\sqrt{\log(\frac{m}{\delta})}(\sigma_p d)^{-1}\frac{\sqrt{n}\|\boldsymbol{\mu}_l\|_2}{\tau_l n \cdot \mathrm{SNR}_l^2} = \sqrt{\log(\frac{m}{\delta})}\frac{\sigma_p}{\tau_l\sqrt{n}\|\boldsymbol{\mu}_l\|_2} \leq \sqrt{\log(\frac{m}{\delta})/n} \cdot 1/(\sqrt{C\log(\frac{n}{\delta})}) \leq 1/(C\sqrt{\log(\frac{n}{\delta})})$, which holds due to $\|\boldsymbol{\mu}_l\|_2^2 \geq C \cdot \sigma_p^2 \log(\frac{n}{\delta})$ and $n \geq C\log(\frac{m}{\delta})$ in Condition 3.1. Therefore, for a sufficiently large constant C, the equality holds. Moreover, we can deduce in a similar manner that

$$
\begin{aligned}
\left\langle \mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu}_l \right\rangle &= \left\langle \mathbf{w}_{-y,r}^{(0)}, y\boldsymbol{\mu}_l \right\rangle - \gamma_{-y,r,l}^{(t)} + \sum_{i=1}^{n} \bar{\rho}_{-y,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \langle \boldsymbol{\xi}_i, -y\boldsymbol{\mu}_l \rangle + \sum_{i=1}^{n} \underline{\rho}_{-y,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \langle \boldsymbol{\xi}_i, y\boldsymbol{\mu}_l \rangle \\
&\leq -\gamma_{-y,r,l}^{(t)} + \sqrt{2\log(\frac{16m}{\delta})} \cdot \sigma_0 \|\boldsymbol{\mu}_l\|_2 + \sqrt{2\log(\frac{12n}{\delta})} \cdot \sigma_p \|\boldsymbol{\mu}_l\|_2 \cdot (\frac{\sigma_p^2 d}{2})^{-1} [\sum_{i=1}^{n} \bar{\rho}_{-y,r,i}^{(t)} + \sum_{i=1}^{n} |\underline{\rho}_{-y,r,i}^{(t)}|] \\
&= -\Theta\left(\gamma_{-y,r,l}^{(t)}\right) < 0.
\end{aligned} \tag{27}
$$

Next, we seek to upper bound $\|\mathbf{w}_{j,r}^{(t)}\|_2$. The techniques are similar to Proposition D.5 in Meng et al. (2023). We first tackle the noise term in the decomposition, namely:

$$
\begin{aligned}
&\left\| \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} \right\|_2^2 \\
=& \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} + 2 \sum_{1 \leq i_1 < i_2 \leq n} \rho_{j,r,i_1}^{(t)} \rho_{j,r,i_2}^{(t)} \cdot \frac{\langle \boldsymbol{\xi}_{i_1}, \boldsymbol{\xi}_{i_2} \rangle}{\|\boldsymbol{\xi}_{i_1}\|_2^2 \cdot \|\boldsymbol{\xi}_{i_2}\|_2^2} \\
\leq& 4\sigma_p^{-2} d^{-1} \sum_{i=1}^{n} {\rho_{j,r,i}^{(t)}}^2 + 2 \sum_{1 \leq i_1 < i_2 \leq n} \rho_{j,r,i_1}^{(t)} \rho_{j,r,i_2}^{(t)} \cdot \left(16\sigma_p^{-4} d^{-2}\right) \cdot \left(2\sigma_p^2 \sqrt{d \log\left(\frac{6n^2}{\delta}\right)}\right) \\
=& 4\sigma_p^{-2} d^{-1} \sum_{i=1}^{n} {\rho_{j,r,i}^{(t)}}^2 + 32\sigma_p^{-2} d^{-3/2} \sqrt{\log\left(\frac{6n^2}{\delta}\right)} \left[\left(\sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \right)^2 - \sum_{i=1}^{n} {\rho_{j,r,i}^{(t)}}^2 \right] \\
=& \Theta\left(\sigma_p^{-2} d^{-1}\right) \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} + \widetilde{\Theta}\left(\sigma_p^{-2} d^{-3/2}\right) \left(\sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \right)^2 \\
\leq& \left[\Theta\left(\sigma_p^{-2} d^{-1} n^{-1}\right) + \widetilde{\Theta}\left(\sigma_p^{-2} d^{-3/2}\right) \right] \left(\sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)} + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \right)^2 \\
=& \Theta\left(\sigma_p^{-2} d^{-1} n^{-1}\right) \left(\sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)} \right)^2,
\end{aligned} \tag{28}
$$

where the first inequality is by Lemma G.1; the second inequality is by the Cauchy Schwartz Inequality on $(\sum_{i=1}^{n} \rho_{j,r,i}^{(t)})^2$. We can then upper bound the $\|\mathbf{w}_{j,r}^{(t)}\|_2$ as:

$$
\begin{aligned}
\|\mathbf{w}_{j,r}^{(t)}\|_2 &\leq \left\|\mathbf{w}_{j,r}^{(0)}\right\|_2 + \sum_{l=1}^{2} \frac{\gamma_{j,r,l}^{(t)}}{\|\boldsymbol{\mu}_l\|_2} + \left\| \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} \right\|_2 \\
&\leq \left\|\mathbf{w}_{j,r}^{(0)}\right\|_2 + \sum_{l=1}^{2} \frac{\gamma_{j,r,l}^{(t)}}{\|\boldsymbol{\mu}_l\|_2} + \Theta\left(\sigma_p^{-1} d^{-1/2} n^{-1/2}\right) \cdot \sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)} \\
&= \Theta\left(\sigma_p^{-1} d^{-1/2} n^{-1/2}\right) \cdot \sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)},
\end{aligned} \tag{29}
$$

where the first inequality is due to the triangle inequality, the second inequality is by (28), and the third equality is due to the following comparisons:

$$\frac{\frac{\gamma_{j,r,l}^{(t)}}{\|\boldsymbol{\mu}_l\|_2}}{\Theta\left(\sigma_p^{-1}d^{-1/2}n^{-1/2}\right)\cdot\sum_{i=1}^n\bar{\rho}_{j,r,i}^{(t)}}=\Theta\left(\sigma_p d^{1/2}n^{1/2}\|\boldsymbol{\mu}_l\|_2^{-1}\,\mathrm{SNR}_l^2\right)=\Theta\left(\sigma_p^{-1}d^{-1/2}n^{1/2}\|\boldsymbol{\mu}_l\|_2\right)=O(1),$$

which is by the coefficient scales in Lemma G.14, the coefficient order $\frac{\sum_{i=1}^n\bar{\rho}_{j,r,i}^{(t)}}{\gamma_{j,r,l}^{(t)}}=\Theta\left(\mathrm{SNR}_l^{-2}\right)$, and the d condition in Condition 3.1; and also we have:

$$\frac{\left\|\mathbf{w}_{j,r}^{(0)}\right\|_2}{\Theta\left(\sigma_p^{-1}d^{-1/2}n^{-1/2}\right)\cdot\sum_{i=1}^n\bar{\rho}_{j,r,i}^{(t)}}=\frac{\Theta\left(\sigma_0\sqrt{d}\right)}{\Theta\left(\sigma_p^{-1}d^{-1/2}n^{-1/2}\right)\cdot\sum_{i=1}^n\bar{\rho}_{j,r,i}^{(t)}}=O\left(\sigma_0\sigma_p dn^{-1/2}\right)=O(1),$$

which is by the coefficient scales in Lemma G.14, and the condition for σ_0 in Condition 3.1. Apply the coefficient order $\sum_{i=1}^n\bar{\rho}_{j,r,i}^{(t)}=\Omega(n)$ to (29), we directly have $\left\|\mathbf{w}_{j,r}^{(t)}\right\|_2\leq\Theta\left(\sigma_p^{-1}d^{-\frac{1}{2}}n^{\frac{1}{2}}\right)$.

G.4. Order-dependent Sampling (Querying) Analysis

Based on the scale of $\mathbf{w}_{j,r}^{(t)}$ and the inner product between it and features, we can now characterize the querying situation of two query criteria-based NAL methods. First, to address the issue of $\Theta(|\mathcal{P}|^2)$ comparisons in $\mathcal{P}$, we employ a full-order-based technique. We introduce the concepts of Uncertainty Order and Diversity Order in Appendix F.2. Subsequently, we delve into the order of the samples in $\mathcal{P}$ in the following proposition.

Proposition G.16. *Under the same conditions of Proposition 3.3, there exist $t=\widetilde{O}\left(\eta^{-1}\varepsilon^{-1}mnd^{-1}\sigma_p^{-2}\right)$ that for $\forall\mathbf{x},\mathbf{x}'\in\mathcal{P}\subsetneq\mathcal{D}$ where $\mathbf{x}$ contains weak feature patch while $\mathbf{x}'$ contains strong feature patch, with probability at least 1-δ', we have $\mathbf{x}'\preceq^{(t)}\mathbf{x}$.*

Proof of Proposition G.16. Firstly, suggest $\mathbf{x}=[y\cdot\boldsymbol{\mu}_2,\mathbf{z}_2],\mathbf{x}'=[y'\cdot\boldsymbol{\mu}_1,\mathbf{z}_1]$, where $\mathbf{z}_1,\mathbf{z}_2\sim N(\mathbf{0},\sigma_p^2\cdot\mathbf{I})$:

$$f\left(\mathbf{W}^{(t)},\mathbf{x}\right)=\sum_{j,r}\frac{j}{m}\left[\sigma\left(\left\langle\mathbf{w}_{j,r}^{(t)},y\boldsymbol{\mu}_2\right\rangle\right)+\sigma\left(\left\langle\mathbf{w}_{j,r}^{(t)},\mathbf{z}_2\right\rangle\right)\right],$$
$$f\left(\mathbf{W}^{(t)},\mathbf{x}'\right)=\sum_{j,r}\frac{j}{m}\left[\sigma\left(\left\langle\mathbf{w}_{j,r}^{(t)},y'\boldsymbol{\mu}_1\right\rangle\right)+\sigma\left(\left\langle\mathbf{w}_{j,r}^{(t)},\mathbf{z}_1\right\rangle\right)\right].$$

By (11) in Lemma F.3 and (16) in DefinitionF.7, we have the following

$$\mathbf{x}'\preceq_C^{(t)}\mathbf{x}\Leftrightarrow\underbrace{\left|f\left(\mathbf{W}^{(t)},\mathbf{x}\right)\right|<\left|f\left(\mathbf{W}^{(t)},\mathbf{x}'\right)\right|}_{\Omega_C},$$
$$\mathbf{x}'\preceq_D^{(t)}\mathbf{x}\Leftrightarrow\underbrace{D\left(\mathbf{W}^{(t)},\mathbf{x},p\mid\mathcal{D}_{n_0}\right)>D\left(\mathbf{W}^{(t)},\mathbf{x}',p\mid\mathcal{D}_{n_0}\right)}_{\Omega_D},$$
$$\mathbf{x}'\preceq^{(t)}\mathbf{x}\Leftrightarrow\underbrace{\{\Omega_C\cap\Omega_D,\forall p\in[1,\infty)\}}_{\Omega}$$

Denote $\sum_j j\cdot\sigma\left(\left\langle\mathbf{w}_{j,r}^{(t)},\mathbf{z}_1\right\rangle\right),\sum_j j\cdot\sigma\left(\left\langle\mathbf{w}_{j,r}^{(t)},\mathbf{z}_2\right\rangle\right)$ as $g_r(\mathbf{z}_1),g_r(\mathbf{z}_2)$ respectively, Notice that for $\mathbf{z}\sim N(\mathbf{0},\sigma_p^2\cdot\mathbf{I})$:

$$\begin{aligned}\left\langle\mathbf{w}_{j,r}^{(t)},\mathbf{z}\right\rangle&\sim\mathcal{N}\left(0,\left\|\mathbf{w}_{j,r}^{(t)}\right\|_2^2\sigma_p^2\cdot\mathbf{I}\right),\\ \sigma(\left\langle\mathbf{w}_{j,r}^{(t)},\mathbf{z}\right\rangle)&\sim\mathcal{N}^R\left(0,\left\|\mathbf{w}_{j,r}^{(t)}\right\|_2^2\sigma_p^2\cdot\mathbf{I}\right).\end{aligned}\tag{30}$$

Then:

$$\begin{aligned} P(\Omega_C) &= P(\left|f\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right| < \left|f\left(\mathbf{W}^{(t)}, \mathbf{x}'\right)\right|) \\ &\geq P(\sum_l (\sum_r |g_r(\mathbf{z}_l)|) < \sum_r (\Theta(\gamma_{y',r,1}) - \Theta(\gamma_{y,r,2}))) \\ &\geq P(m \cdot \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < m(\Theta(\mathop{\mathbb{E}}_r(\gamma_{y',r,1})) - \Theta(\mathop{\mathbb{E}}_r(\gamma_{y,r,2})))) \\ &= P(\underbrace{\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < \Theta((\mathop{\mathbb{E}}_r(\gamma_{y',r,1}) - \mathop{\mathbb{E}}_r(\gamma_{y,r,2})))}_{\Omega_\gamma}. \end{aligned} \tag{31}$$

The second inequality is by triangle inequality and (25) in Lemma G.15; the third inequality is by Lemma G.14.

For Ω_D, denoting $U_0^l = \{\mathbf{x} \in \mathcal{D}_0 \mid \mathbf{x}_{\text{signal part}} = \boldsymbol{\mu}_l\}$ as the set of indices of $\mathcal{D}_0$ where the data's feature patch is $\boldsymbol{\mu}_l$, We then take a look at the rth row of the Feature Distance $\mathbf{Z}(\mathbf{x}, t)$, which we denote as $\mathbf{Z}_r(\mathbf{x}, t)$:

$$\begin{aligned} \mathbf{Z}_r(\mathbf{x}, t) &= \sum_j \left(\sigma\left(\langle \mathbf{w}_{j,r}, y \cdot \boldsymbol{\mu}_2 \rangle\right) + \sigma\left(\langle \mathbf{w}_{j,r}, \mathbf{z}_r \rangle\right)\right) \\ &= \Theta\left(\gamma_{y,r,2}\right) + g_r(\mathbf{z}_2) \end{aligned} \tag{32}$$

$$\begin{aligned} \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)}, t)}{n_0} &= \sum_{i,j} \frac{\sigma\left(\left\langle \mathbf{w}_{j,r}, y_i \cdot \boldsymbol{\mu}^{(i)} \right\rangle\right) + \sigma\left(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_i \rangle\right)}{n_0} \\ &= \frac{\left[\sum_l \tau_l \cdot n_0 \cdot \mathop{\mathbb{E}}_{i_l \in U_0^l} \Theta(\gamma_{y_{i_l},r,l}) + \sum_i \sum_j \Theta\left(\bar{\rho}_{j,r,i}\right)\right]}{n_0} \end{aligned} \tag{33}$$

Let (32) - (33), we have:

$$\mathbf{Z}_r(\mathbf{x}, t) - \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)}, t)}{n_0} = \Theta(\gamma_{y,r,2}) + g_r(\mathbf{z}_2) - \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)}, t)}{n_0} \tag{34}$$

Now we can estimate $D\left(\mathbf{W}^{(t)}, \mathbf{x}, p \mid \mathcal{D}_{n_0}\right)$:

$$\begin{aligned} D\left(\mathbf{W}^{(t)}, \mathbf{x}, p \mid \mathcal{D}_{n_0}\right) &= \|\mathbf{Z}(\mathbf{x}, t) - \sum_{i=1}^{n_0} \frac{\mathbf{Z}(\mathbf{x}^{(i)}, t)}{n_0}\|_p \\ &= \left(\sum_r |\mathbf{Z}_r(\mathbf{x}, t) - \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)}, t)}{n_0}|^p\right)^{\frac{1}{p}} \\ &= \left(\sum_r |\Theta(\gamma_{y,r,2}) + g_r(\mathbf{z}_2) - \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)}, t)}{n_0}|^p\right)^{\frac{1}{p}} \end{aligned} \tag{35}$$

Similarly, the $D\left(\mathbf{W}^{(t)}, \mathbf{x}', p \mid \mathcal{D}_{n_0}\right)$ could be written as:

$$D\left(\mathbf{W}^{(t)}, \mathbf{x}', p \mid \mathcal{D}_{n_0}\right) = \left(\sum_r |\Theta(\gamma_{y,r,1}) + g_r(\mathbf{z}_1) - \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)}, t)}{n_0}|^p\right)^{\frac{1}{p}} \tag{36}$$

To compare $D\left(\mathbf{W}^{(t)}, \mathbf{x}, p \mid \mathcal{D}_{n_0}\right)$ and $D\left(\mathbf{W}^{(t)}, \mathbf{x}', p \mid \mathcal{D}_{n_0}\right)$, we first see that both expressions in the r-th filter owns

$$-\sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)}, t)}{n_0} = -\sum_l \tau_l \cdot \Theta(\mathop{\mathbb{E}}_{i_l \in U_0^l}(\gamma_{y_{i_l},r,l})) - n_0^{-1} \sum_i \sum_j \Theta\left(\bar{\rho}_{j,r,i}\right).$$

By Condition 3.1, we see that $\sigma_p^2 d/(n_0 \|\boldsymbol{\mu}_1\|_2^2) = \Omega(\log(T^*))$. We see that as T^* is the substantially large maximum admissible iterations, collaborating with (25), (33) and (30), it holds that the order of $n_0^{-1} \sum_{i,j} \sigma\left(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_i \rangle\right) =$

$n_0^{-1}\sum_i\sum_j \Theta\left(\bar{\rho}_{j,r,i}\right)$ in $\sum_i \dfrac{\mathbf{Z}_r(\mathbf{x}^{(i)},t)}{n_0}$ is indeed can dominate $n_0^{-1}\sum_{i,j}\sigma\left(\left\langle \mathbf{w}_{j,r}, y_i\cdot\boldsymbol{\mu}^{(i)}\right\rangle\right) = \sum_l \tau_l\cdot\Theta(\mathop{\mathbb{E}}_{i_l\in U_0^l}(\gamma_{y_{i_l},r,l}))$, $\Theta(\gamma_{y,r,1})$ and $g_r(\mathbf{z}_1)$. As $\sum_i \dfrac{\mathbf{Z}_r(\mathbf{x}^{(i)},t)}{n_0}$ is shared by both $D\left(\mathbf{W}^{(t)},\mathbf{x},p \mid \mathcal{D}_{n_0}\right)$ and $D\left(\mathbf{W}^{(t)},\mathbf{x}',p \mid \mathcal{D}_{n_0}\right)$ in the r-th filter, a sufficient event for $D\left(\mathbf{W}^{(t)},\mathbf{x},p \mid \mathcal{D}_{n_0}\right) > D\left(\mathbf{W}^{(t)},\mathbf{x}',p \mid \mathcal{D}_{n_0}\right)$ is that for $\forall r\in[m]$, we have

$$|\sum_l \tau_l\cdot\Theta(\mathop{\mathbb{E}}_{i_l\in U_0^l}(\gamma_{y_{i_l},r,l})) - \Theta(\gamma_{y,r,2}) - g_r(\mathbf{z}_2)| > |\max\{\sum_l \tau_l\cdot\Theta(\mathop{\mathbb{E}}_{i_l\in U_0^l}(\gamma_{y_{i_l},r,l})) - \Theta(\gamma_{y,r,1}) - g_r(\mathbf{z}_1), 0\}|.$$

Utilizing those results, we now could estimate the chance of event Ω_D:

$$\begin{aligned}
P(\Omega_D) &= P(D\left(\mathbf{W}^{(t)},\mathbf{x},p \mid \mathcal{D}_{n_0}\right) > D\left(\mathbf{W}^{(t)},\mathbf{x}',p \mid \mathcal{D}_{n_0}\right)) \\
&\geq P(m^{\frac{1}{p}}\sum_l(\max_r|g_r(\mathbf{z}_l)|) < m^{\frac{1}{p}}(|\Theta(\mathop{\mathbb{E}}_r(\gamma_{y,r,2})) - \sum_l \tau_l\cdot\Theta(\mathop{\mathbb{E}}_{i_l\in U_0^l,r}(\gamma_{y_{i_l},r,l}))| \\
&\qquad\qquad - |\Theta(\mathop{\mathbb{E}}_r(\gamma_{y,r,1})) - \sum_l \tau_l\cdot\Theta(\mathop{\mathbb{E}}_{i_l\in U_0^l,r}(\gamma_{y_{i_l},r,l}))|) \\
&\geq P(m^{\frac{1}{p}}\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)},\mathbf{z}_l\right\rangle\right|\} < m^{\frac{1}{p}}\left((\tau_1-\tau_2)\Theta(\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,1})) - (\tau_1-\tau_2)\Theta(\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,2}))\right) \\
&= P(m^{\frac{1}{p}}\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)},\mathbf{z}_l\right\rangle\right|\} < m^{\frac{1}{p}}\Theta(\frac{\tau_1(\tau_1-\tau_2)\|\boldsymbol{\mu}_1\|_2^2 - \tau_2(\tau_1-\tau_2)\|\boldsymbol{\mu}_2\|_2^2}{\sigma_p^2 d/n_0})) \\
&= P(m^{\frac{1}{p}}\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)},\mathbf{z}_l\right\rangle\right|\} < m^{\frac{1}{p}}\Theta(\mathop{\mathbb{E}}_r(\gamma_{y',r,1}) - \mathop{\mathbb{E}}_r(\gamma_{y,r,2}))) \\
&= P(\underbrace{\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)},\mathbf{z}_l\right\rangle\right|\} < \Theta((\mathop{\mathbb{E}}_r(\gamma_{y',r,1}) - \mathop{\mathbb{E}}_r(\gamma_{y,r,2})))}_{\Omega_\gamma}),
\end{aligned} \tag{37}$$

where the first inequality is by Lemma G.14, triangle inequality, (25), (35) and (36); The forth equality is by (30). Easy to see that if $p=\infty$, the third equality would be zero, thus our condition $p<\infty$ avoid this case. Now we take a look at the event Ω_γ:

$$\begin{aligned}
P(\Omega_\gamma) &= P(\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)},\mathbf{z}_l\right\rangle\right|\} < \Theta((\mathop{\mathbb{E}}_r(\gamma_{y',r,1}) - \mathop{\mathbb{E}}_r(\gamma_{y,r,2}))) \\
&= P(\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)},\mathbf{z}_l\right\rangle\right|\} < \Theta\left(\frac{\left[\tau_1\|\boldsymbol{\mu}_1\|_2^2 - \tau_2\|\boldsymbol{\mu}_2\|_2^2\right]}{\sigma_p^2 d/n_0}\right)) \\
&\geq P(\bigcup_{j,r,l}\underbrace{\{\left|\left\langle \mathbf{w}_{j,r}^{(t)},\mathbf{z}_l\right\rangle - 0\right| < \Theta\left(\frac{\left[\tau_1\|\boldsymbol{\mu}_1\|_2^2 - \tau_2\|\boldsymbol{\mu}_2\|_2^2\right]}{\sigma_p^2 d/n_0}\right)\}}_{\hat{\Omega}_{j,r,l}}) \\
&= \sum_{j,r,l} P(\hat{\Omega}_{j,r,l}),
\end{aligned} \tag{38}$$

where the second equality is by the first inference statement of Lemma G.14; the third inequality is by the equivalence property of the union by events; the last equality is by the Union Rule. Then, by Gaussian tail bound, we have:

$$P(\hat{\Omega}_{j,r,l}) \geq 1 - 2\exp\left\{-\Theta\left(\frac{\left[\tau_1\|\boldsymbol{\mu}_1\|_2^2 - \tau_2\|\boldsymbol{\mu}_2\|_2^2\right]^2}{\sigma_p^6 d^2/n_0^2\left\|w_{j,r}^{(t)}\right\|_2^2}\right)\right\}$$

Finally, with conditions on $\|\boldsymbol{\mu}_1\|_2^2 - \|\boldsymbol{\mu}_2\|_2^2$ in Proposition 3.3, Lemma G.3, (25) in Lemma G.15 and union bound, we have

the conclusion for event Ω:

$$\Rightarrow P(\Omega) \geq P(\Omega_\gamma) \geqslant 1 - 8m \exp\left\{-\Theta\left(\frac{\left[\tau_1 \|\boldsymbol{\mu}_1\|_2^2 - \tau_2 \|\boldsymbol{\mu}_2\|_2^2\right]^2}{\sigma_p^4 d/n_0}\right)\right\} \tag{39}$$
$$\geqslant 1 - \delta',$$

for $\forall p \in [1, \infty)$.

Remark G.17. We can observe that the Uncertainty Order and Diversity Order of samples rely heavily on the model's learning progresss upon them. By Lemma G.14, the learning progresss of samples depend heavily on the feature strength $\|\boldsymbol{\mu}_l\|_2$ and data proportion τ_l. That is to say, in our case, the **perplexing samples** are the samples containing weak feature $\boldsymbol{\mu}_2$. In the next section, we would show that the number of those **perplexing samples** in the labeled set after querying would determine the algorithm's generalization ability.

From the above proving process, we can deduce some important findings, which can be summarized in the following lemmas.

The following lemma shows that Uncertainty Sampling and Diversity Sampling correspond to different comparisons on the model's learning progress over samples in $\mathcal{P}$.

Lemma G.18. *(Restatement of Lemma 4.2) Under the same conditions in Proposition 3.3, with the same notations in Proposition G.16, there exists certain constants $c_1, c_2, c_3, c_4, c_5, c_6 > 0$, such that*

- $\mathbf{x} \preceq_C^{(t)} \mathbf{x}'$ *has a sufficient event that*

$$\{c_1 \mathop{\mathbb{E}}_r(\gamma_{y',r,1}) - c_2 \mathop{\mathbb{E}}_r(\gamma_{y,r,2}) > \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\}\}, \tag{40}$$

  among which the left side of the inequality corresponds to the comparison of learning progress of samples with different type of feature patch.

- $\mathbf{x} \preceq_D^{(t)} \mathbf{x}', \forall p \in [1, \infty)$ *has a sufficient event that*

$$\{|c_3 \mathop{\mathbb{E}}_r(\gamma_{y,r,2}) - c_4 \sum_l \tau_l \cdot \mathop{\mathbb{E}}_{i_l \in U_0^l, r}(\gamma_{y_{i_l},r,l})| - |c_5 \mathop{\mathbb{E}}_r(\gamma_{y',r,1}) - c_6 \sum_l \tau_l \cdot \mathop{\mathbb{E}}_{i_l \in U_0^l, r}(\gamma_{y_{i_l},r,l})| > \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\}\}, \tag{41}$$

  among which the left side of the inequality corresponds to the comparison of the disparity between learning toward samples and labeled training set.

Proof of Lemma G.18. The first bullet point can be easily derived from (31), while the second bullet point is readily apparent from (35), (36), and (37).

During the proving process of Proposition G.16, it is observed that for any $p \in [1, \infty)$, there exists a shared sufficient event for (40) and (41). This implies that it is also a shared sufficient event for the events Ω_C and Ω_D, denoted as Ω_γ:

$$\Omega_\gamma := \{\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < \Theta((\mathop{\mathbb{E}}_r(\gamma_{y',r,1}) - \mathop{\mathbb{E}}_r(\gamma_{y,r,2}))\}.$$

By the first inference statement of Lemma G.14, we have

$$\Omega_\gamma = \{\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < \Theta((\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,1}) - \mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,2}))\}. \tag{42}$$

Therefore, we can conclude that the significant difference in the model's learning of the feature $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is what causes the sufficient event for both event Ω_C and Ω_D. By (39), we have:

$$P(\Omega_\gamma) \geq 1 - 8m \exp\left\{-\Theta\left(\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,1}) - \mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,2})\right)\right\}. \tag{43}$$

Based on Lemma G.14, we see that the $\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,1})$ is significant larger than $\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,2})$ under our conditions, which causes the sufficient event Ω_γ.

Based on the above results, we can have a look on the overall order situation of the sampling pool $\mathcal{P}$.

Lemma G.19. *(Restatement of Lemma 4.4) Under Condition 3.1, when the results of Proposition 3.2 and Proposition G.16 hold at the initial stage and querying stage at a certain $t \leq T^*$, denoting $\mathbf{X}_{\mathcal{P}}^1 \subsetneqq \mathcal{P}$ as the collection of all the data points with strong feature $\boldsymbol{\mu}_1$ in $\mathcal{P}$, and $\mathbf{X}_{\mathcal{P}}^2 \subsetneqq \mathcal{P}$ as the collection of data points with weak feature $\boldsymbol{\mu}_2$, we have the conclusion that with probability more than 1-$\Theta(\delta')$, $\mathbf{X}_{\mathcal{P}}^1 \prec^{(t)} \mathbf{X}_{\mathcal{P}}^2$ holds.*

proof of Lemma G.19. By Proposition G.16, $\forall \mathbf{x}' \in \mathbf{X}_{\mathcal{P}}^1$, and $\forall \mathbf{x} \in \mathbf{X}_{\mathcal{P}}^2$, $\mathbf{x}' \prec^{(t)} \mathbf{x}$ with at least probability δ'. It's natural to see comparing every pairs in $\mathbf{X}_{\mathcal{P}}^1$ and $\mathbf{X}_{\mathcal{P}}^2$ as independent random events. Then given a certain $\mathbf{x}' \in \mathbf{X}_{\mathcal{P}}^1$, the chance that $\forall \mathbf{x} \in \mathbf{X}_{\mathcal{P}}^2$ satisfies $\mathbf{x}' \prec^{(t)} \mathbf{x}$ is $\Theta((1-\delta')^{|\mathbf{X}_{\mathcal{P}}^2|})$, therefore, for $\forall \mathbf{x}' \in \mathbf{X}_{\mathcal{P}}^1$, the chance is $\Theta((1-\delta')^{|\mathbf{X}_{\mathcal{P}}^2| \cdot |\mathbf{X}_{\mathcal{P}}^1|}) = \Theta((1-\delta')^{p(1-p)|\mathcal{P}|^2}) = 1 - \Theta(\delta')$ as $\delta' \ll 1$.

Based on Lemma G.19 and (42), we directly have the following lemma demonstrate that both NAL algorithms would all prioritize those poor learning samples.

Lemma G.20. *(Restatement of Proposition 3.3) Under the same conditions in Proposition 3.2, the Uncertainty Order and Diversity Order of the samples $[(y \cdot \boldsymbol{\mu}_l)^T, \xi^T]^T$ in sampling pool $\mathcal{P}$ follows the order of $\mathop{\mathbb{E}}_{j,r} \gamma_{j,r,l}^{(t)}$.*

G.5. Label Complexity-based Test Error Analysis

In this section, we suggest the results in the previous sections all hold with high probability. With the results of the final scale of the coefficients as well as the order situation of the data in sampling pool $\mathcal{P}$, we can now take a look on the test error upper and lower bound under distinct conditions before and after querying.

Lemma G.21. *(Partial restatement of Lemma 4.5) Under Condition 3.1, for a test set $\mathcal{D}^* \subseteq \mathcal{D}^*$ with occurrence probability p^* of the $\boldsymbol{\mu}_2$-equipped data, then $\exists\, t = \widetilde{O}\left(\eta^{-1}\varepsilon^{-1} m n_0 d^{-1} \sigma_p^{-2}\right)$, we have the following two situations before and after querying (i.e., $\forall s \in \{0,1\}$):*

- *If $\forall l \in \{1,2\}, n_{s,l} \geq \dfrac{C_1 \sigma_p^4 d}{\|\boldsymbol{\mu}_l\|_2^4}$ holds, we have the test error:*

$$L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \leq (1-p^*) \cdot \exp\left(\frac{-n_{s,1}\|\boldsymbol{\mu}_1\|_2^4}{C_3 \sigma_p^4 d}\right) + p^* \cdot \exp\left(\frac{-n_{s,2}\|\boldsymbol{\mu}_2\|_2^4}{C_4 \sigma_p^4 d}\right). \tag{44}$$

- *If $\exists l' \in \{1,2\} n_{s,l'} \leq \dfrac{C_2 \sigma_p^4 d}{\|\boldsymbol{\mu}_{l'}\|_2^4}$ holds, where C_1 is from Condition 3.1, we have the test error*

$$L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \geq 0.12 \cdot p_{l'}^*. \tag{45}$$

Here $p_{l'}^*$ denotes the occurrence probability of feature $\boldsymbol{\mu}_{l'}$, C_1, C_2, C_3 and C_4 are some positive constants.

Proof of Lemma G.21. Recall the test error definition and consider the proportion of different type of data in the testing set $\mathcal{D}^*$, we have:

$$\begin{aligned} L_{\mathcal{D}^*}^{0-1}(\mathbf{W}) &= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*}[y \cdot f(\mathbf{W},\mathbf{x}) < 0] \\ &= (1-p^*) \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\boldsymbol{\mu}_1}^*}[y \cdot f(\mathbf{W},\mathbf{x}) < 0] + p^* \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\boldsymbol{\mu}_2}^*}[y \cdot f(\mathbf{W},\mathbf{x}) < 0], \end{aligned} \tag{46}$$

where $\mathcal{D}_{\boldsymbol{\mu}_1}^*$ and $\mathcal{D}_{\boldsymbol{\mu}_2}^*$ denotes the collection of data points in $\mathcal{D}$ containing feature $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively.

First, we seek to prove the first bullet point. We utilize the techniques similar to the proofs of Theorem 1 in Chatterji and Long (2021), Lemma 3 in Frei et al. (2022), Theorem E.1 in Kou et al. (2023b) and Theorem 3.2 in Meng et al. (2023). Denote the feature patch in $\mathbf{x}$ as $\boldsymbol{\mu}_{l_x}$ ($l_x \in \{1,2\}$), we first take a look at the product

$$\begin{aligned} y \cdot f\left(\mathbf{W}^{(t)}, \mathbf{x}\right) &= \frac{1}{m}\sum_{j,r} yj\left[\sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right) + \sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right] \\ &= \frac{1}{m}\sum_{r}\left[\sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right) + \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right] - \frac{1}{m}\sum_{r}\left[\sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right) + \sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right] \\ &\leq \frac{1}{m}\left[\sum_{r}\sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right) - \sum_{r}\sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right]. \end{aligned} \tag{47}$$

Denote $g(\boldsymbol{\xi})$ as $\sum_r \sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)$. Since $\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle \sim \mathcal{N}\left(0, \left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2^2 \sigma_p^2\right)$, we can get

$$\mathbb{E}g(\boldsymbol{\xi}) = \sum_{r=1}^m \mathbb{E}\sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) = \sum_{r=1}^m \frac{\left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2 \sigma_p}{\sqrt{2\pi}} = \frac{\sigma_p}{\sqrt{2\pi}} \sum_{r=1}^m \left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2. \tag{48}$$

Then we can obtain the following test error upper bound on $\mathcal{D}_{\boldsymbol{\mu}_{l_x}}^*$ by adding $\mathbb{E}g(\boldsymbol{\xi})$ and $\frac{\sigma_p}{\sqrt{2\pi}} \sum_{r=1}^m \left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2$ at two sides of the inequality:

$$\begin{aligned}\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\boldsymbol{\mu}_{l_x}}^*}\left(yf\left(\boldsymbol{W}^{(t)}, \mathbf{x}\right) \le 0\right) &\le \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left(\sum_r \sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) \ge \sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right)\right) \\ &= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left(g(\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\xi}) \ge \sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right) - \frac{\sigma_p}{\sqrt{2\pi}} \sum_{r=1}^m \left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2\right).\end{aligned} \tag{49}$$

By the results in Lemma G.14 and Lemma G.15, we take a look at the comparison of the two terms at the right side of the inequality:

$$\frac{\sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right)}{\sigma_p \sum_{r=1}^m \left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2} \ge \frac{\Theta\left(\sum_r \gamma_{y,r,l_x}^{(t)}\right)}{\Theta\left(d^{-1/2} n_s^{-1/2}\right) \cdot \sum_{r,i} \bar{\rho}_{-y,r,i}^{(t)}} = \Theta\left(\tau_{l_x} d^{1/2} n_s^{1/2} \operatorname{SNR}_{l_x}^2\right) = \Theta\left(\tau_{l_x} n_s^{1/2} \|\boldsymbol{\mu}_{l_x}\|_2^2 / (\sigma_p^2 d^{1/2})\right), \tag{50}$$

where τ_{l_x} denotes the proportion of feature $\boldsymbol{\mu}_{l_x}$ in current training data set (before or after querying). Worth noting that we have assumption in the first bullet that $\forall l \in \{1,2\}, n_{s,l} \ge \frac{C_1 \sigma_p^4 d}{\|\boldsymbol{\mu}_l\|_2^4}$, which means $n_{1,l_x} \|\boldsymbol{\mu}_1\|_2^4 \ge 2C_1 \sigma_p^4 d, \forall l_x \in \{1,2\}$. Since C_1 is a sufficiently large constant, it directly follows that

$$\sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right) - \frac{\sigma_p}{\sqrt{2\pi}} \sum_{r=1}^m \left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2 > 0.$$

By Theorem 5.2.2 in Vershynin (2018), we know that for any $x \ge 0$, the following holds

$$P(g(\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\xi}) \ge x) \le \exp\left(-\frac{cx^2}{\sigma_p^2 \|g\|_{\text{Lip}}^2}\right), \tag{51}$$

where c is a constant. To calculate the Lipschitz norm, we have

$$\begin{aligned}|g(\boldsymbol{\xi}) - g(\boldsymbol{\xi}')| &= \left|\sum_{r=1}^m \sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) - \sum_{r=1}^m \sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}'\right\rangle\right)\right| \\ &\le \sum_{r=1}^m \left|\sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) - \sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}'\right\rangle\right)\right| \\ &\le \sum_{r=1}^m \left|\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi} - \boldsymbol{\xi}'\right\rangle\right| \\ &\le \sum_{r=1}^m \left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2 \cdot \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2,\end{aligned}$$

where the first inequality is by triangle inequality; the second inequality is by the property of ReLU; the last inequality is by Cauchy Schwartz Inequality. Therefore, we have

$$\|g\|_{\text{Lip}} \le \sum_{r=1}^m \left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2. \tag{52}$$

Utilize (51) and (52) in (49), we have

$$\begin{aligned}
\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*_{\boldsymbol{\mu}_{l_x}}}\left(yf\left(\boldsymbol{W}^{(t)},\mathbf{x}\right)\leq 0\right) &\leq \exp\left[-\frac{c\left(\sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right)-\left(\frac{\sigma_p}{\sqrt{2\pi}}\right)\sum_{r=1}^m\left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2\right)^2}{\sigma_p^2\left(\sum_{r=1}^m\left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2\right)^2}\right] \\
&= \exp\left[-c\left(\frac{\sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right)}{\sigma_p\sum_{r=1}^m\|\mathbf{w}_{-y,r}^{(t)}\|_2}-\frac{1}{\sqrt{2\pi}}\right)^2\right] \\
&\leq \exp(c/2\pi)\exp\left(-0.5c\left(\frac{\sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right)}{\sigma_p\sum_{r=1}^m\left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2}\right)^2\right),
\end{aligned} \tag{53}$$

where the third inequality is by $(s-t)^2\geq s^2/2-t^2, \forall s,t\geq 0$. And then by (50) and (53), we can have

$$\begin{aligned}
\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*_{\boldsymbol{\mu}_{l_x}}}\left(yf\left(\boldsymbol{W}^{(t)},\mathbf{x}\right)\leq 0\right) &\leq \exp(c/2\pi)\exp\left(-0.5c\left(\frac{\sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}_{l_x}\right\rangle\right)}{\sigma_p\sum_{r=1}^m\left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2}\right)^2\right) \\
&= \exp\left(\frac{c}{2\pi}-\frac{\tau_{l_x}n_{s,l_x}\|\boldsymbol{\mu}_{l_x}\|_2^4}{C\sigma_p^4 d}\right) \\
&= \exp\left(\frac{c}{2\pi}-\frac{n_{s,l_x}\|\boldsymbol{\mu}_{l_x}\|_2^4}{C_{l_x}\sigma_p^4 d}\right) \\
&\leq \exp\left(-\frac{n_{s,l_x}\|\boldsymbol{\mu}_{l_x}\|_2^4}{2C_{l_x}\sigma_p^4 d}\right)
\end{aligned} \tag{54}$$

where $C_{l_x}=C/\tau_{lx}=O(1)$; the last inequality holds if we choose $C_1\geq cC_{l_x}/\pi$, for any $l_x\in\{1,2\}$. If we choose C_3 as $2C_{l_1}$ and C_4 as $2C_{l_2}$, by (46) and (54) we have

$$L^{0-1}_{\mathcal{D}^*}\left(\mathbf{W}^{(t)}\right)\leq(1-p^*)\cdot\exp\left(\frac{-n_{s,1}\|\boldsymbol{\mu}_1\|_2^4}{C_3\sigma_p^4 d}\right)+p^*\cdot\exp\left(\frac{-n_{s,2}\|\boldsymbol{\mu}_2\|_2^4}{C_4\sigma_p^4 d}\right).$$

Next, we serve to prove the second bullet point. We utilize the pigeonhole principle technique in Kou et al. (2023b); Meng et al. (2023), which is based on the following two lemmas.

Lemma G.22. *For $t\in[T_1,T^*]$, denote $g(\boldsymbol{\xi})=\sum_{j,r}\sigma\left(\left\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}\right\rangle\right)$. There exists a fixed vector $\mathbf{v}_l$ with $\|\mathbf{v}_l\|_2\leq 0.02\sigma_p$ and constant C_6 such that*

$$\sum_{j'\in\{\pm1\}}\left[g\left(j'\boldsymbol{\xi}+\mathbf{v}_l\right)-g\left(j'\boldsymbol{\xi}\right)\right]\geq 4C_6\max_{j,l}\left\{\sum_r\gamma_{j,r,l}^{(t)}\right\},$$

for all $\boldsymbol{\xi}\in\mathbb{R}^d$.

Proof of Lemma G.22. See Lemma 5.8 in Kou et al. (2023b) or Theorem 3.2 in Meng et al. (2023) for a proof, where we utilize a large enough C_2 in the condition given in the second bullet point ($n_{s,l'}\leq\frac{C_2\sigma_p^4 d}{\|\boldsymbol{\mu}_{l'}\|_2^4}$) to control the norm of $\mathbf{v}_l$.

Lemma G.23. *(Proposition 2.1 in Devroye et al. (2023)). The TV distance between $\mathcal{N}\left(0,\sigma_p^2\mathbf{I}_d\right)$ and $\mathcal{N}\left(\mathbf{v}_l,\sigma_p^2\mathbf{I}_d\right)$ is smaller than $\|\mathbf{v}_l\|_2/2\sigma_p$.*

Proof of Lemma G.23. See Proposition 2.1 in Devroye et al. (2023) for a proof.

Now we take a look at $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right)$, by (46) we have:

$$\begin{aligned}
L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) &= \tau_1^* \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*_{\boldsymbol{\mu}_1}}\left[y\cdot f(\mathbf{W},\mathbf{x})<0\right] + \tau_2^*) \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\boldsymbol{\mu}_2}}\left[y\cdot f(\mathbf{W},\mathbf{x})<0\right] \\
&\geq \tau_{l'}^* \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*_{\boldsymbol{\mu}_{l'}}}\left[y\cdot f(\mathbf{W},\mathbf{x})<0\right] \\
&= \tau_{l'}^* \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*_{\boldsymbol{\mu}_{l'}}}\Big(\sum_r \sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)},\boldsymbol{\xi}\right\rangle\right) - \sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)},\boldsymbol{\xi}\right\rangle\right) \\
&\qquad\qquad \geq \sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)},y\boldsymbol{\mu}_{l'}\right\rangle\right) - \sum_r \sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)},y\boldsymbol{\mu}_{l'}\right\rangle\right)\Big) \\
&\geq 0.5\tau_{l'}^* \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*_{\boldsymbol{\mu}_{l'}}}\left(\left|\sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}\right\rangle\right) - \sum_r \sigma\left(\left\langle \mathbf{w}_{-1,r}^{(t)},\boldsymbol{\xi}\right\rangle\right)\right| \geq C_6 \max\left\{\sum_r \gamma_{1,r,l'}^{(t)}, \sum_r \gamma_{-1,r,l'}^{(t)}\right\}\right) \\
&= 0.5\tau_{l'}^* \cdot P(\Omega_{\boldsymbol{\xi}}),
\end{aligned} \tag{55}$$

where $\Omega_{\boldsymbol{\xi}} := \left\{\boldsymbol{\xi} \big| |g(\boldsymbol{\xi})| \geq C_6 \max\left\{\sum_r \gamma_{1,r,l'}^{(t)}, \sum_r \gamma_{-1,r,l'}^{(t)}\right\}\right\}$. The last inequality holds since we can always have a corresponding y to make a wrong prediction if given $\boldsymbol{\xi}$, the $\left|\sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}\right\rangle\right) - \sum_r \sigma\left(\left\langle \mathbf{w}_{-1,r}^{(t)},\boldsymbol{\xi}\right\rangle\right)\right|$ is large enough.

Next, we seek a lower bound of $P(\Omega_{\boldsymbol{\xi}})$. By Lemma G.22, we have that $\sum_j [g(j\boldsymbol{\xi}+\mathbf{v}_l) - g(j\boldsymbol{\xi})] \geq 4C_6 \max_{j,l}\left\{\sum_r \gamma_{j,r,l}^{(t)}\right\}$. Then by pigeon's hole principle, there must exist one of the $\boldsymbol{\xi}, \boldsymbol{\xi}+\mathbf{v}_l$, $-\boldsymbol{\xi}, -\boldsymbol{\xi}+\mathbf{v}_l$ belongs $\Omega_{\boldsymbol{\xi}}$. So we have proved that $\Omega_{\boldsymbol{\xi}} \cup -\Omega_{\boldsymbol{\xi}} \cup \Omega_{\boldsymbol{\xi}} - \{\mathbf{v}_l\} \cup -\Omega_{\boldsymbol{\xi}} - \{\mathbf{v}_l\} = \mathbb{R}^d$. Therefore at least one of $P(\Omega_{\boldsymbol{\xi}}), P(-\Omega_{\boldsymbol{\xi}}), P(\Omega_{\boldsymbol{\xi}} - \{\mathbf{v}_l\}), P(\Omega_{\boldsymbol{\xi}} - \{\mathbf{v}_l\}), P(-\Omega_{\boldsymbol{\xi}} - \{\mathbf{v}_l\})$ is greater than 0.25. By the definition of TV distance, we have:

$$\begin{aligned}
|P(\Omega_{\boldsymbol{\xi}}) - P(\Omega_{\boldsymbol{\xi}} - \mathbf{v}_l)| &= \left|\mathbb{P}_{\boldsymbol{\xi}\sim\mathcal{N}\left(0,\sigma_p^2\mathbf{I}_d\right)}(\boldsymbol{\xi}\in\Omega_{\boldsymbol{\xi}}) - \mathbb{P}_{\boldsymbol{\xi}\sim\mathcal{N}\left(\mathbf{v}_l,\sigma_p^2\mathbf{I}_d\right)}(\boldsymbol{\xi}\in\Omega_{\boldsymbol{\xi}})\right| \\
&\leq \mathrm{TV}\left(\mathcal{N}\left(0,\sigma_p^2\mathbf{I}_d\right), \mathcal{N}\left(\mathbf{v}_l,\sigma_p^2\mathbf{I}_d\right)\right) \\
&\leq \frac{\|\mathbf{v}_l\|_2}{2\sigma_p} \\
&\leq 0.02.
\end{aligned}$$

Also, notice that $P(-\Omega_{\boldsymbol{\xi}}) = P(\Omega_{\boldsymbol{\xi}})$, we have $4P(\Omega_{\boldsymbol{\xi}}) \geq 1 - 2\cdot 0.02$. Thus $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \geq 0.5\tau_{l'}^* \cdot 0.24 = 0.12 \cdot \tau_{l'}^*$. The proofs complete.

Based on Lemma G.21, our focus is to verify whether the NAL algorithms satisfy the condition stated in the first bullet point. On the other hand, it is highly likely that Random Sampling fulfills the condition stated in the second bullet point. The following proposition validates this intuition.

Proposition G.24. *When Lemma G.19 holds, and the sampling size of algorithm satisfies* $\frac{C_1\sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4} - \frac{pn_0}{2} \leq n^* = \Theta(\widetilde{n} - n_0) \leq \widetilde{n} - n_0$*, we have the following:*

- *The number of data with strong feature patch* $n_{s,1}$ *satisfies* $n_{s,1} \geq \frac{C_1\sigma_p^4 d}{\|\boldsymbol{\mu}_1\|_2^4}, \forall s \in \{0,1\}$.
- *The number of data with weak feature patch* $n_{s,2}$ *before querying and after* ***Random Sampling*** *satisfies* $n_{s,2} \leq \frac{C_2\sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4}, \forall s \in \{0,1\}$.
- *The total number of data with weak feature patch* $n_{1,2}$ *after* ***Uncertainty Sampling*** *and* ***Diversity Sampling*** *satisfies* $n_{1,2} \geq \frac{C_1\sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4}$.

For the sake of coherence, here C_1 *and* C_2 *are some constants shared with Theorem 3.4 and Lemma 4.5.*

Proof of Proposition G.24. By conditions in Definition 2.1, we have $(1-\frac{3}{2}p)n_0 \geq \frac{C_1\sigma_p^4 d}{\|\boldsymbol{\mu}_1\|_2^4}$ for a large constant C_1. Then by plugging the results of n_p for n_0 in Lemma G.3, as well as the definition of $n_{s,l}$, we have

$$n_{1,1} \geq n_{0,1} \geq (1-\frac{3}{2}p)n_0 \geq \frac{C_1\sigma_p^4 d}{\|\boldsymbol{\mu}_1\|_2^4}.$$

For the second bullet, by Lemma G.3, Lemma G.19 and conditions $n^* \geq \frac{C_1\sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4} - \frac{pn_0}{2}$, we have:

$$n_{1,2} \geq \frac{pn_0}{2} + n^* \geq \frac{C_1\sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4}$$

Besides, by Lemma G.3 and the condition $\widetilde{n} \leq \frac{2C_2\sigma_p^4 d}{3p\|\boldsymbol{\mu}_2\|_2^4}$, the third bullet holds straightforwardly.

By the result of Lemma G.21 and Proposition G.24, the results of Proposition 3.2 and Theorem 3.4 holds directly.

Lemma G.25. *(Restatement of Corollary 3.5) Under the same conditions as stated in Theorem 3.4, with a probability of at least $1-\Theta(\delta+\delta')$, we observe distinct label complexities for traditional 2-layer ReLU CNN and NAL algorithms in achieving Bayes-optimal generalization ability:*

- *For a fully trained neural model, the label complexity n_{CNN} requires $\Omega(p^{-1}\sigma_p^2 d\|\boldsymbol{\mu}_2\|_2^{-4})$.*
- *For two NAL algorithms, the maximum label complexity $\widetilde{n}$ only requires $\Omega(\sigma_p^2 d\|\boldsymbol{\mu}_2\|_2^{-4})$.*

Proof of Lemma G.25. According to Lemma G.21, to adequately learn the signal $\boldsymbol{\mu}_l$ for any $l \in \{1,2\}$, one needs at least $\hat{C}1\sigma_p^4 d\|\boldsymbol{\mu}_l\|_2^{-4}$. Since the occurrence probability of $\boldsymbol{\mu}_2$ is low (p), Random Sampling without any strategy requires a label complexity of at least $\Omega(p^{-1}\sigma_p^2 d\|\boldsymbol{\mu}_2\|_2^{-4})$ to capture sufficient instances of $\boldsymbol{\mu}_2$ from the training distribution. On the other hand, by leveraging the insights from Lemma G.19 and Lemma G.20, both Uncertainty Sampling and Diversity Sampling can effectively query yet-to-be-learned **perplexing samples**, which are typically samples associated with $\boldsymbol{\mu}_2$ by Lemma G.14. Hence, both querying algorithms only require a label complexity of $\Omega(\sigma_p^2 d\|\boldsymbol{\mu}_2\|_2^{-4})$.

H. Proofs of Main Results: XOR data version

In this section, we first introduce some notations. We denote n as the number of training data in the current labeled training set, which is initially n_0 and becomes n_1 after sampling (querying). We define $\mathbf{u}_l = \mathbf{a}_l + \mathbf{b}_l$ and $\mathbf{v}_l = \mathbf{a}_l - \mathbf{b}_l$. The proportion of easy-to-learn data $\boldsymbol{\mu}_1 = \pm(\mathbf{a}_1 \pm \mathbf{b}_1)$ in the current labeled set is denoted as τ_1, while τ_2 represents the proportion of hard-to-learn data $\boldsymbol{\mu}_2 = \pm(\mathbf{a}_2 \pm \mathbf{b}_2)$. In a manner similar to the proofs provided in Appendix G, in this section we utilize the techniques employed in Kou et al. (2023b); Meng et al. (2023) to obtain results that are not directly related to our main contribution. For the sake of brevity, we omit most of the proof details of those outcomes, as our setting aligns with the one considered in (Meng et al., 2023), despite the fact that we examine multiple task-oriented features. Instead, our focus is on providing comprehensive proofs of our primary contributions.

First, we claim that all preliminary Lemmas in Appendix G.1 hold with high probability. It is evident from Definition 8 that $F_{+1}(\mathbf{W}_{+1}, \mathbf{x})$ always contributes to the prediction of class $+1$, while $F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$ always contributes to the prediction of class -1. Therefore, the jobs of $\mathbf{w}_{+1,r}$ and $\mathbf{w}_{-1,r}$ are learning $\pm\mathbf{u}$ and $\pm\mathbf{v}$ respectively. Then, similar to (G.5), we take a look at the coefficient updates with *signal-noise decomposition* techniques, specified as the following.

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + \sum_{l=1}^{2}\gamma_{j,r,\mathbf{u}_l}^{(t)} \cdot \frac{j\cdot\mathbf{u}_l}{\|\mathbf{u}_l\|_2^2} - \sum_{l=1}^{2}\gamma_{j,r,\mathbf{v}_l}^{(t)} \cdot \frac{j\cdot\mathbf{v}_l}{\|\mathbf{v}_l\|_2^2} + \sum_{i=1}^{n}\bar{\rho}_{j,r,i}^{(t)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} + \sum_{i=1}^{n}\underline{\rho}_{j,r,i}^{(t)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2}, \tag{56}$$

where we denote $\bar{\rho}_{j,r,i}^{(t)}$ as $\rho_{j,r,i}^{(t)}\mathbb{1}\left(\rho_{j,r,i}^{(t)} \geq 0\right)$, $\underline{\rho}_{j,r,i}^{(t)}$ as $\rho_{j,r,i}^{(t)}\mathbb{1}\left(\rho_{j,r,i}^{(t)} \leq 0\right)$. Here $\gamma_{j,r,\mathbf{u}_l}^{(t)}$ are mainly contributed by $F_{+1}(\mathbf{W}_{+1}, \mathbf{x})$, and $\gamma_{\pm1,r,\mathbf{u}_l}^{(t)} \approx \left\langle \mathbf{w}_{j,r}^{(t)}, \pm\mathbf{u}\right\rangle$. Similarly $\gamma_{j,r,\mathbf{v}_l}^{(t)}$ are mainly contributed by $F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$, and $\gamma_{\pm1,r,\mathbf{v}_l}^{(t)} \approx \left\langle \mathbf{w}_{j,r}^{(t)}, \pm\mathbf{v}\right\rangle$. Worth noting that $j \in \{\pm1\}$ here denote the signal of $\mathbf{u}_l$ and $\mathbf{v}_l$, but not the signal of $F_{j'}(\mathbf{W}_{j'}, \mathbf{x}), j' \in \{\pm1\}$.

Specifically, the update rule can be written as:

$$\begin{aligned}\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} &- \frac{\eta j}{nm} \sum_{i \in S_{+\mathbf{u}_l,+1} \cup S_{-\mathbf{u}_l,-1}} \ell_i^{(t)} \mathbb{1}\left\{\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_i \right\rangle > 0\right\} \mathbf{u}_l + \frac{\eta j}{nm} \sum_{i \in S_{-\mathbf{u}_l,+1} \cup S_{+\mathbf{u}_l,-1}} \ell_i'^{(t)} \mathbb{1}\left\{\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_i \right\rangle > 0\right\} \mathbf{u}_l \\ &+ \frac{\eta j}{nm} \sum_{i \in S_{+\mathbf{v}_l,-1} \cup S_{-\mathbf{v}_l,+1}} \ell_i'^{(t)} \mathbb{1}\left\{\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_i \right\rangle > 0\right\} \mathbf{v}_l - \frac{\eta j}{nm} \sum_{i \in S_{-\mathbf{v}_l,-1} \cup S_{+\mathbf{v}_l,+1}} \ell_i^{(t)} \mathbb{1}\left\{\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_i \right\rangle > 0\right\} \mathbf{v}_l \\ &- \frac{\eta}{nm} \sum_{i=1}^{n} \ell_i'^{(t)} (j y_i) \mathbb{1}\left\{\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \right\rangle > 0\right\} \boldsymbol{\xi}_i,\end{aligned} \tag{57}$$

where $S_{\boldsymbol{\mu},j} = \{i \in [n], \boldsymbol{\mu}_i = \boldsymbol{\mu}, y_i = j\}$. Here $\boldsymbol{\mu} \in \{\pm\mathbf{u}_1, \pm\mathbf{u}_2, \pm\mathbf{v}_1, \pm\mathbf{v}_2\}, j \in \{\pm 1\}$, and we let $\boldsymbol{\mu}_i$ represents the feature in $\mathbf{x}_i$ and $\boldsymbol{\xi}_i$ represents the noise in $\mathbf{x}_i$.

The following lemma shows that a specific discrete process can be bounded by its continuous counterpart, which would be useful in bounding the coefficient $\sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)}$ and the derivative of loss function.

Lemma H.1. *(Lemma C.1 in Meng et al. (2023)) Suppose that a sequence $a_t, t \geq 0$ follows the iterative formula*

$$a_{t+1} = a_t + \frac{c}{1 + b e^{a_t}},$$

for some $1 \geq c \geq 0$ and $b \geq 0$. Then it holds that

$$x_t \leq a_t \leq \frac{c}{1 + b e^{a_0}} + x_t$$

for all $t \geq 0$. Here, x_t is the unique solution of

$$x_t + b e^{x_t} = ct + a_0 + b e^{a_0}.$$

H.1. Coefficient Ratio and Scale Analysis: XOR data version

Similar to the processes in Appendix G, we assume the results in the previous section hold with high probability. Meanwhile, let $T^* = \eta^{-1} \text{ poly}\left(\varepsilon^{-1}, d, n, m\right)$ be the maximum admissible iteration. We adopt similar notations as those in (22):

$$\begin{aligned}&\alpha := 4 \log\left(T^*\right), \\ &\beta := 2 \max_{l,i,j,r} \left\{ \left| \left\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_l \right\rangle \right|, \left| \left\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \right\rangle \right| \right\}, \\ &\mathrm{SNR}_l := \frac{\|\boldsymbol{\mu}_l\|_2}{\sigma_p \sqrt{d}}, \\ &\kappa = 56 \sqrt{\frac{\log\left(6n^2/\delta\right)}{d}} n \log\left(T^*\right) + 10 \sqrt{\log(16mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d} + \sum_{l=1}^{2} 64 \tau_l n \cdot \mathrm{SNR}_l^2 \log\left(T^*\right).\end{aligned} \tag{58}$$

Then, similar to our results in Proposition G.9, we here also have the coefficient scale as below.

Proposition H.2. *If Condition C.3 holds, then for any $0 \leq t \leq T^*, j \in \{\pm 1\}, r \in [m]$ and $i \in [n]$, it holds that*

$$\begin{aligned}&0 \leq |\left\langle \mathbf{w}_{+1,r}^{(t)}, \mathbf{u}_l \right\rangle|, |\left\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v}_l \right\rangle| = \Theta(\gamma_{j,r,\mathbf{u}_l}^{(t)}), \Theta(\gamma_{j,r,\mathbf{v}_l}^{(t)}) \leq 32 \tau_l n \cdot \mathrm{SNR}_l^2\, \alpha, \\ &0 \leq \bar{\rho}_{j,r,i}^{(t)} \leq 4\alpha, \quad 0 \geq \underline{\rho}_{j,r,i}^{(t)} \geq -\beta - 32 \sqrt{\frac{\log\left(6n^2/\delta\right)}{d}} n\alpha, \\ &-\frac{\kappa}{2} + \frac{1}{m} \sum_{r=1}^{m} \bar{\rho}_{y_i,r,i}^{(t)} \leq y_i f\left(\mathbf{W}^{(t)}, \mathbf{x}_i\right) \leq \frac{\kappa}{2} + \frac{1}{m} \sum_{r=1}^{m} \bar{\rho}_{y_i,r,i}^{(t)}.\end{aligned}$$

Moreover, define $\bar{c} = \dfrac{2\eta\sigma_p^2 d}{nm}, \underline{c} = \dfrac{\eta\sigma_p^2 d}{3nm}, \bar{b} = e^{-\kappa}$ and $\underline{b} = e^{\kappa}$, and let $\bar{x}_t, \underline{x}_t$ be the unique solution of

$$\begin{aligned}\bar{x}_t + \bar{b} e^{\bar{x}_t} &= \bar{c} t + \bar{b}, \\ \underline{x}_t + \underline{b} e^{\underline{x}_t} &= \underline{c} t + \underline{b},\end{aligned}$$

it holds that

$$\underline{x}_t \leq \frac{1}{m}\sum_{r=1}^{m}\bar{\rho}_{y_i,r,i}^{(t)} \leq \bar{x}_t + \bar{c}/(1+\bar{b}), \quad \log\left(\frac{\eta\sigma_p^2 d}{8nm}t + 2/3\right) \leq \bar{x}_t, \underline{x}_t \leq \log\left(\frac{2\eta\sigma_p^2 d}{nm}t + 1\right) \tag{59}$$

for all $r \in [m]$ and $i \in [n]$.

Proof of Proposition H.2. Please refer to Proposition C.2, Proposition C.8 and Lemma C.9 in Meng et al. (2023) for a proof. Regardless of the variations in data settings, it is feasible to obtain the result through inductive techniques (Cao et al., 2022a; Frei et al., 2022; Kou et al., 2023b; Lu et al., 2023).

Building upon Proposition H.2, we can further analyze the convergence of the training dynamics by examining the extent of feature learning and noise memorization in the subsequent section.

H.2. Feature Learning and Noise Memorization Analysis: XOR data version

Similar to Lemma G.13 and Lemma G.15 for linearly separable data, we can also determine the scale of coefficients and inner products as follows.

Proposition H.3. *Under Condition C.3, the following points hold ($n > n_0$) for $\forall l \in \{1, 2\}$:*

1. *For any $r \in [m]$, $\left\langle \mathbf{w}_{+1,r}^{(t)}, \mathbf{u}_l \right\rangle$ (or $\left\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v}_l \right\rangle$) increases if $\left\langle \mathbf{w}_{+1,r}^{(0)}, \mathbf{u}_l \right\rangle > 0$(or $\left\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v}_l \right\rangle < 0$), $\left\langle \mathbf{w}_{+1,r}^{(t)}, \mathbf{u}_l \right\rangle$ (or $\left\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v}_l \right\rangle$) decreases if $\left\langle \mathbf{w}_{+1,r}^{(0)}, \mathbf{u}_l \right\rangle < 0$ (or $\left\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v}_l \right\rangle$) > 0. Moreover, it holds that*

$$\begin{aligned}
&\gamma_{j,r,\mathbf{u}_l}^{(t)}, \gamma_{j,r,\mathbf{v}_l}^{(t)} = \Theta(\frac{\tau_l n\|\boldsymbol{\mu}_l\|_2^2}{\sigma_p^2 d}\cdot\log\left(\frac{\eta\sigma_p^2 dt}{nm}\right)), |\left\langle \mathbf{w}_{+1,r}^{(t)}, \mathbf{u}_l \right\rangle|, |\left\langle \mathbf{w}_{-1,r}^{(t)} = \Theta(\frac{\tau_l n\|\boldsymbol{\mu}_l\|_2^2}{\sigma_p^2 d}\cdot\log\left(\frac{\eta\sigma_p^2 dt}{nm}\right)), \mathbf{v}_l \right\rangle|, \\
&|\left\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{u}_l \right\rangle| \leq |\left\langle \mathbf{w}_{-1,r}^{(0)}, \mathbf{u}_l \right\rangle| + \eta\|\boldsymbol{\mu}_l\|_2^2/m, |\left\langle \mathbf{w}_{+1,r}^{(t)}, \mathbf{v}_l \right\rangle| \leq |\left\langle \mathbf{w}_{+1,r}^{(0)}, \mathbf{v}_l \right\rangle| + \eta\|\boldsymbol{\mu}_l\|_2^2/m.
\end{aligned} \tag{60}$$

2. *Let $\underline{x}_t$ defined in Proposition H.2, we have*

$$\Omega(n) \leq \frac{n}{5}\cdot(\bar{x}_{t-1} - \bar{x}_1) \leq \sum_{i=1}^{n}\bar{\rho}_{j,r,i}^{(t)} \leq 3n\underline{x}_t \leq 3n\cdot\log\left(\frac{2\eta\sigma_p^2 d}{nm}t + 1\right) = \Theta(n\cdot\log\left(\frac{\eta\sigma_p^2 dt}{nm}\right)), \tag{61}$$

for all $t \in [T^*]$ and $r \in [m]$. Moreover, we have:

$$\sum_{i=1}^{n}\bar{\rho}_{j,r,i}^{(t)}/\gamma_{\boldsymbol{\mu}_l,j',r',l}^{(t)} = \Theta\left(\tau_l^{-1}\cdot\mathrm{SNR}_l^{-2}\right) = \sum_{i=1}^{n}\bar{\rho}_{j,r,i}^{(t)}/|\left\langle \mathbf{w}_{\pm1,r'}^{(t)}, \boldsymbol{\mu}_l \right\rangle|,$$

for all $j, j' \in \{\pm1\}, r, r' \in [m]$.

3. *For $t = \Omega\left(nm/\left(\eta\sigma_p^2 d\right)\right)$, the bound for $\left\|\mathbf{w}_{j,r}^{(t)}\right\|_2$ is given by:*

$$\left\|\mathbf{w}_{j,r}^{(t)}\right\|_2 = \Theta\left(\sigma_p^{-1}d^{-1/2}n^{1/2}\cdot\log\left(\frac{\eta\sigma_p^2 dt}{nm}\right)\right). \tag{62}$$

Proof of Proposition H.2. The basic techniques are the same as Lemma G.13 and Lemma G.15 despite variation in data settings. Please refer to Proposition 4.2, Proposition D.3-5 in Meng et al. (2023) for a comprehensive proof.

H.3. Order-dependent Sampling (Querying) Analysis: XOR data version

Based on the scale of $\mathbf{w}_{j,r}^{(t)}$ and the inner product between it and features, we can now characterize the querying situation of the two NAL methods based on the query criteria. Similar to the order-dependent analysis techniques utilized in Appendix G.4, we employ a full-order-based technique to tackle the problem of $\Theta(|\mathcal{P}|^2)$ comparisons in $\mathcal{P}$. The concepts of Uncertainty Order and Diversity Order are introduced in Appendix F.2. We then proceed to examine the order of the samples in $\mathcal{P}$ in the following proposition.

Proposition H.4. *Under the same conditions of Proposition C.5, there exist* $t = \widetilde{O}\left(\eta^{-1}\varepsilon^{-1}mnd^{-1}\sigma_p^{-2}\right)$ *that for* $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{P} \subsetneq \mathcal{D}$ *where* $\mathbf{x}$ *contains hard-to-learn feature patch while* $\mathbf{x}'$ *contains easy-to-learn feature patch, with probability at least 1-*δ'*, we have* $\mathbf{x}' \preceq^{(t)} \mathbf{x}$.

Proof of Proposition H.4. Firstly, suggest $\mathbf{x} = [y \cdot \boldsymbol{\mu}_2, \mathbf{z}_2], \mathbf{x}' = [y' \cdot \boldsymbol{\mu}_1, \mathbf{z}_1]$, where $\boldsymbol{\mu}_1 \in \{\mathbf{u}_1, \mathbf{v}_1\}, \boldsymbol{\mu}_2 \in \{\mathbf{u}_2, \mathbf{v}_2\}, y, y' \in [\pm 1], \mathbf{z}_1, \mathbf{z}_2 \sim N(\mathbf{0}, \sigma_p^2 \cdot \mathbf{I})$:

$$f\left(\mathbf{W}^{(t)}, \mathbf{x}\right) = \sum_{j,r} \frac{j}{m}\left[\sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)}, y\boldsymbol{\mu}_2\right\rangle\right) + \sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_2\right\rangle\right)\right],$$
$$f\left(\mathbf{W}^{(t)}, \mathbf{x}'\right) = \sum_{j,r} \frac{j}{m}\left[\sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)}, y'\boldsymbol{\mu}_1\right\rangle\right) + \sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_1\right\rangle\right)\right].$$

By (11) in Lemma F.3 and (16) in Definition F.7, we have the following

$$\mathbf{x}' \preceq_C^{(t)} \mathbf{x} \Leftrightarrow \underbrace{\left|f\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right| < \left|f\left(\mathbf{W}^{(t)}, \mathbf{x}'\right)\right|}_{\Omega_C},$$
$$\mathbf{x}' \preceq_D^{(t)} \mathbf{x} \Leftrightarrow \underbrace{D\left(\mathbf{W}^{(t)}, \mathbf{x}, p \mid \mathcal{D}_{n_0}\right) > D\left(\mathbf{W}^{(t)}, \mathbf{x}', p \mid \mathcal{D}_{n_0}\right)}_{\Omega_D},$$
$$\mathbf{x}' \preceq^{(t)} \mathbf{x} \Leftrightarrow \underbrace{\{\Omega_C \cap \Omega_D, \forall p \in [1, \infty)\}}_{\Omega}$$

Denote $\sum_j j \cdot \sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_1\right\rangle\right), \sum_j j \cdot \sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_2\right\rangle\right)$ as $g_r(\mathbf{z}_1), g_r(\mathbf{z}_2)$ respectively, Notice that for $\mathbf{z} \sim N(\mathbf{0}, \sigma_p^2 \cdot \mathbf{I})$:

$$\begin{aligned} \left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}\right\rangle &\sim \mathcal{N}\left(0, \left\|\mathbf{w}_{j,r}^{(t)}\right\|_2^2 \sigma_p^2 \cdot \mathbf{I}\right), \\ \sigma(\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}\right\rangle) &\sim \mathcal{N}^R\left(0, \left\|\mathbf{w}_{j,r}^{(t)}\right\|_2^2 \sigma_p^2 \cdot \mathbf{I}\right). \end{aligned} \tag{63}$$

Then:

$$\begin{aligned} P(\Omega_C) &= P(\left|f\left(\mathbf{W}^{(t)}, \mathbf{x}\right)\right| < \left|f\left(\mathbf{W}^{(t)}, \mathbf{x}'\right)\right|) \\ &\geq P(\sum_l (\sum_r |g_r(\mathbf{z}_l)|) < \sum_r (\Theta(\gamma_{y',r,\boldsymbol{\mu}_1}) - \Theta(\gamma_{y,r,\boldsymbol{\mu}_2}))) \\ &\geq P(m \cdot \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l\right\rangle\right|\} < m(\Theta(\mathop{\mathbb{E}}_r(\gamma_{y',r,\boldsymbol{\mu}_1})) - \Theta(\mathop{\mathbb{E}}_r(\gamma_{y,r,\boldsymbol{\mu}_2})))) \\ &= P(\underbrace{\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l\right\rangle\right|\} < \Theta((\mathop{\mathbb{E}}_r(\gamma_{y',r,\boldsymbol{\mu}_1}) - \mathop{\mathbb{E}}_r(\gamma_{y,r,\boldsymbol{\mu}_2})))}_{\Omega_\gamma}. \end{aligned} \tag{64}$$

The second inequality is by triangle inequality and (60) in Proposition H.3; the third inequality is by (63).

For Ω_D, denoting $U_0^l = \{\mathbf{x} \in \mathcal{D}_0 \mid \mathbf{x}_{\text{signal part}} = \boldsymbol{\mu}_l\}$ as the set of indices of $\mathcal{D}_0$ where the data's feature patch is $\boldsymbol{\mu}_l$, We then take a look at the rth row of the Feature Distance $\mathbf{Z}(\mathbf{x}, t)$, which we denote as $\mathbf{Z}_r(\mathbf{x}, t)$:

$$\begin{aligned} \mathbf{Z}_r(\mathbf{x}, t) &= \sum_j \left(\sigma\left(\langle \mathbf{w}_{j,r}, y \cdot \boldsymbol{\mu}_2\rangle\right) + \sigma\left(\langle \mathbf{w}_{j,r}, \mathbf{z}_r\rangle\right)\right) \\ &= \Theta\left(\gamma_{y,r,\boldsymbol{\mu}_2}\right) + g_r(\mathbf{z}_2) \end{aligned} \tag{65}$$

$$\begin{aligned} \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)}, t)}{n_0} &= \sum_{i,j} \frac{\sigma\left(\left\langle \mathbf{w}_{j,r}, y_i \cdot \boldsymbol{\mu}^{(i)}\right\rangle\right) + \sigma\left(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_i\rangle\right)}{n_0} \\ &= \frac{\left[\sum_l \tau_l \cdot n_0 \cdot \mathop{\mathbb{E}}_{i_l \in U_0^l} \Theta(\gamma_{y_{i_l},r,\boldsymbol{\mu}_l}) + \sum_i \sum_j \Theta\left(\bar{\rho}_{j,r,i}\right)\right]}{n_0} \end{aligned} \tag{66}$$

Let (65) - (66), we have:

$$\mathbf{Z}_r(\mathbf{x},t) - \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)},t)}{n_0} = \Theta(\gamma_{y,r,\boldsymbol{\mu}_2}) + g_r(\mathbf{z}_2) - \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)},t)}{n_0} \tag{67}$$

Now we can estimate $D\left(\mathbf{W}^{(t)},\mathbf{x},p \mid \mathcal{D}_{n_0}\right)$:

$$\begin{aligned} D\left(\mathbf{W}^{(t)},\mathbf{x},p \mid \mathcal{D}_{n_0}\right) &= \|\mathbf{Z}(\mathbf{x},t) - \sum_{i=1}^{n_0} \frac{\mathbf{Z}(\mathbf{x}^{(i)},t)}{n_0}\|_p \\ &= \left(\sum_r |\mathbf{Z}_r(\mathbf{x},t) - \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)},t)}{n_0}|^p\right)^{\frac{1}{p}} \\ &= \left(\sum_r |\Theta(\gamma_{y,r,\boldsymbol{\mu}_2}) + g_r(\mathbf{z}_2) - \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)},t)}{n_0}|^p\right)^{\frac{1}{p}} \end{aligned} \tag{68}$$

Similarly, the $D\left(\mathbf{W}^{(t)},\mathbf{x}',p \mid \mathcal{D}_{n_0}\right)$ could be written as:

$$D\left(\mathbf{W}^{(t)},\mathbf{x}',p \mid \mathcal{D}_{n_0}\right) = \left(\sum_r |\Theta(\gamma_{y,r,\boldsymbol{\mu}_1}) + g_r(\mathbf{z}_1) - \sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)},t)}{n_0}|^p\right)^{\frac{1}{p}} \tag{69}$$

To compare $D\left(\mathbf{W}^{(t)},\mathbf{x},p \mid \mathcal{D}_{n_0}\right)$ and $D\left(\mathbf{W}^{(t)},\mathbf{x}',p \mid \mathcal{D}_{n_0}\right)$, we first see that both expressions in the r-th filter owns

$$-\sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)},t)}{n_0} = -\sum_l \tau_l \cdot \Theta(\mathop{\mathbb{E}}_{i_l \in U_0^l}(\gamma_{y_{i_l},r,\boldsymbol{\mu}_l})) - n_0^{-1}\sum_i\sum_j \Theta\left(\bar{\rho}_{j,r,i}\right).$$

By Condition C.3, it holds that $\sigma_p^2 d/(n_0\|\boldsymbol{\mu}_1\|_2^2) = \Omega(\log(T^*))$. We see that as T^* is the substantially large maximum admissible iterations, collaborating with (60), (66) and (63), it holds that the order of $n_0^{-1}\sum_{i,j}\sigma\left(\langle \mathbf{w}_{j,r},\boldsymbol{\xi}_i\rangle\right) = n_0^{-1}\sum_i\sum_j \Theta\left(\bar{\rho}_{j,r,i}\right)$ in $\sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)},t)}{n_0}$ is indeed can dominate $n_0^{-1}\sum_{i,j}\sigma\left(\left\langle \mathbf{w}_{j,r}, y_i\cdot\boldsymbol{\mu}^{(i)}\right\rangle\right) = \sum_l \tau_l \cdot \Theta(\mathop{\mathbb{E}}_{i_l \in U_0^l}(\gamma_{y_{i_l},r,\boldsymbol{\mu}_l}))$, $\Theta(\gamma_{y,r,\boldsymbol{\mu}_1})$ and $g_r(\mathbf{z}_1)$. As $\sum_i \frac{\mathbf{Z}_r(\mathbf{x}^{(i)},t)}{n_0}$ is shared by both $D\left(\mathbf{W}^{(t)},\mathbf{x},p \mid \mathcal{D}_{n_0}\right)$ and $D\left(\mathbf{W}^{(t)},\mathbf{x}',p \mid \mathcal{D}_{n_0}\right)$ in the r-th filter, a sufficient event for $D\left(\mathbf{W}^{(t)},\mathbf{x},p \mid \mathcal{D}_{n_0}\right) > D\left(\mathbf{W}^{(t)},\mathbf{x}',p \mid \mathcal{D}_{n_0}\right)$ is that for $\forall r \in [m]$, it holds that

$$|\sum_l \tau_l \cdot \Theta(\mathop{\mathbb{E}}_{i_l \in U_0^l}(\gamma_{y_{i_l},r,\boldsymbol{\mu}_l})) - \Theta(\gamma_{y,r,\boldsymbol{\mu}_2}) - g_r(\mathbf{z}_2)| > |\max\{\sum_l \tau_l \cdot \Theta(\mathop{\mathbb{E}}_{i_l \in U_0^l}(\gamma_{y_{i_l},r,\boldsymbol{\mu}_l})) - \Theta(\gamma_{y,r,1}) - g_r(\mathbf{z}_1), 0\}|.$$

Utilizing those results, we now could estimate the chance of event Ω_D:

$$
\begin{aligned}
P(\Omega_D) &= P(D\left(\mathbf{W}^{(t)}, \mathbf{x}, p \mid \mathcal{D}_{n_0}\right) > D\left(\mathbf{W}^{(t)}, \mathbf{x}', p \mid \mathcal{D}_{n_0}\right)) \\
&\geq P(m^{\frac{1}{p}} \sum_l (\max_r |g_r(\mathbf{z}_l)|) < m^{\frac{1}{p}} (|\Theta(\mathop{\mathbb{E}}_r(\gamma_{y,r,\boldsymbol{\mu}_2})) - \sum_l \tau_l \cdot \Theta(\mathop{\mathbb{E}}_{i_l \in U_0^l, r}(\gamma_{y_{i_l},r,\boldsymbol{\mu}_l}))| \\
&\qquad\qquad - |\Theta(\mathop{\mathbb{E}}_r(\gamma_{y,r,\boldsymbol{\mu}_1})) - \sum_l \tau_l \cdot \Theta(\mathop{\mathbb{E}}_{i_l \in U_0^l, r}(\gamma_{y_{i_l},r,\boldsymbol{\mu}_l}))|) \\
&\geq P(m^{\frac{1}{p}} \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < m^{\frac{1}{p}} \left((\tau_1 - \tau_2)\Theta(\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,\boldsymbol{\mu}_1})) - (\tau_1 - \tau_2)\Theta(\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,\boldsymbol{\mu}_2}))\right) \\
&= P(m^{\frac{1}{p}} \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < m^{\frac{1}{p}} \Theta(\frac{\tau_1(\tau_1 - \tau_2)\|\boldsymbol{\mu}_1\|_2^2 - \tau_2(\tau_1 - \tau_2)\|\boldsymbol{\mu}_2\|_2^2}{\sigma_p^2 d/n_0}) \cdot \log\left(\frac{\eta \sigma_p^2 dt}{nm}\right)) \\
&= P(m^{\frac{1}{p}} \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < m^{\frac{1}{p}} \Theta(\mathop{\mathbb{E}}_r(\gamma_{y',r,\boldsymbol{\mu}_1}) - \mathop{\mathbb{E}}_r(\gamma_{y,r,\boldsymbol{\mu}_2}))) \\
&= P(\underbrace{\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < \Theta((\mathop{\mathbb{E}}_r(\gamma_{y',r,\boldsymbol{\mu}_1}) - \mathop{\mathbb{E}}_r(\gamma_{y,r,\boldsymbol{\mu}_2})))}_{\Omega_\gamma},
\end{aligned}
\tag{70}
$$

where the first inequality is by triangle inequality, (68) and (69); The forth equality is by (63). Easy to see that if $p = \infty$, the third equality would be zero, thus our condition $p < \infty$ avoid this case. Now we take a look at the event Ω_γ:

$$
\begin{aligned}
P(\Omega_\gamma) &= P(\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < \Theta((\mathop{\mathbb{E}}_r(\gamma_{y',r,\boldsymbol{\mu}_1}) - \mathop{\mathbb{E}}_r(\gamma_{y,r,\boldsymbol{\mu}_2}))) \\
&= P(\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < \Theta\left(\frac{\left[\tau_1 \|\boldsymbol{\mu}_1\|_2^2 - \tau_2 \|\boldsymbol{\mu}_2\|_2^2\right]}{\sigma_p^2 d/n_0} \cdot \log\left(\frac{\eta \sigma_p^2 dt}{nm}\right)\right)) \\
&\geq P(\bigcup_{j,r,l} \underbrace{\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle - 0\right| < \Theta\left(\frac{\left[\tau_1 \|\boldsymbol{\mu}_1\|_2^2 - \tau_2 \|\boldsymbol{\mu}_2\|_2^2\right]}{\sigma_p^2 d/n_0} \cdot \log\left(\frac{\eta \sigma_p^2 dt}{nm}\right)\right)\}}_{\hat{\Omega}_{j,r,l}}) \\
&= \sum_{j,r,l} P(\hat{\Omega}_{j,r,l}),
\end{aligned}
\tag{71}
$$

where the second equality is by the first inference statement of Lemma G.14; the third inequality is by the equivalence property of the union by events; the last equality is by the Union Rule. Then, by Gaussian tail bound, we have:

$$
P(\hat{\Omega}_{j,r,l}) \geq 1 - 2\exp\left\{-\Theta\left(\frac{\left[\tau_1 \|\boldsymbol{\mu}_1\|_2^2 - \tau_2 \|\boldsymbol{\mu}_2\|_2^2\right]^2}{\sigma_p^6 d^2/n_0^2 \left\|w_{j,r}^{(t)}\right\|_2^2} \cdot \log\left(\frac{\eta \sigma_p^2 dt}{nm}\right)\right)\right\}
$$

Finally, with conditions in Proposition C.5, Lemma G.3, Proposition H.3 and union bound, we have the conclusion for event Ω:

$$
\begin{aligned}
\Rightarrow P(\Omega) \geq P(\Omega_\gamma) &\geqslant 1 - 8m\exp\left\{-\Theta\left(\frac{\left[\tau_1 \|\boldsymbol{\mu}_1\|_2^2 - \tau_2 \|\boldsymbol{\mu}_2\|_2^2\right]^2}{\sigma_p^4 d/n_0}\right)\right\} \\
&\geqslant 1 - \delta',
\end{aligned}
\tag{72}
$$

for $\forall p \in [1, \infty)$.

Remark H.5. The proof process is nearly identical to that of the linearly separable case (i.e., the proof of Proposition G.16). The only differences lie in the scale of $\|w_{j,r}^{(t)}\|_2$ and $\gamma_{\pm 1,r,\boldsymbol{\mu}}$, but the conditions required are the same.

Similar to Lemma G.18 in Appendix G.4, we have the following lemma.

Lemma H.6. *Under the same conditions in Proposition 3.3, with the same notations in Proposition H.4, there exists certain constants $c_1, c_2, c_3, c_4, c_5, c_6 > 0$, such that*

- $\mathbf{x} \preceq_C^{(t)} \mathbf{x}'$ *has a sufficient event that*

$$\{c_1 \mathop{\mathbb{E}}_r(\gamma_{y',r,\boldsymbol{\mu}_1}) - c_2 \mathop{\mathbb{E}}_r(\gamma_{y,r,\boldsymbol{\mu}_2}) > \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\}\}, \tag{73}$$

  among which the left side of the inequality corresponds to the comparison of learning progress of samples with different type of feature patch.

- $\mathbf{x} \preceq_D^{(t)} \mathbf{x}', \forall p \in [1, \infty)$ *has a sufficient event that*

$$\{|c_3 \mathop{\mathbb{E}}_r(\gamma_{y,r,\boldsymbol{\mu}_2}) - c_4 \sum_l \tau_l \cdot \mathop{\mathbb{E}}_{i_l \in U_0^l, r}(\gamma_{y_{i_l},r,\boldsymbol{\mu}_l})| - |c_5 \mathop{\mathbb{E}}_r(\gamma_{y',r,\boldsymbol{\mu}_1}) - c_6 \sum_l \tau_l \cdot \mathop{\mathbb{E}}_{i_l \in U_0^l, r}(\gamma_{y_{i_l},r,\boldsymbol{\mu}_l})| > \max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\}\}, \tag{74}$$

  among which the left side of the inequality corresponds to the comparison of the disparity between learning toward samples and labeled training set.

Proof of Lemma H.6. The first bullet point can be easily derived from (64), while the second bullet point is readily apparent from (68), (69), and (70).

Similar to the discussions in Appendix G.4, it is observed that for any $p \in [1, \infty)$, there exists a shared sufficient event for (73) and (74). This implies that it is also a shared sufficient event for the events Ω_C and Ω_D, denoted as Ω_γ:

$$\Omega_\gamma := \{\max_{j,r,l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < \Theta((\mathop{\mathbb{E}}_r(\gamma_{y',r,\boldsymbol{\mu}_1}) - \mathop{\mathbb{E}}_r(\gamma_{y,r,\boldsymbol{\mu}_2}))\}.$$

By the first inference statement of Proposition H.3, we have

$$\Omega_\gamma = \{\max_{j,r,\boldsymbol{\mu}_l}\{\left|\left\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{z}_l \right\rangle\right|\} < \Theta((\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,\boldsymbol{\mu}_1}) - \mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,\boldsymbol{\mu}_2}))\}. \tag{75}$$

Therefore, we can conclude that the significant difference in the model's learning of the feature $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is what causes the sufficient event for both event Ω_C and Ω_D. By (72), we have:

$$P(\Omega_\gamma) \geq 1 - 8m \exp\left\{-\Theta\left(\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,\boldsymbol{\mu}_1}) - \mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,\boldsymbol{\mu}_2})\right)\right\}. \tag{76}$$

Based on Proposition H.3, we see that the $\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,\boldsymbol{\mu}_1})$ is significant larger than $\mathop{\mathbb{E}}_{j,r}(\gamma_{j,r,\boldsymbol{\mu}_2})$ under our conditions, which causes the sufficient event Ω_γ.

Similar to Lemma 4.4 for linearly separable XOR data, we also have conclusions regarding the order of pool for XOR data.

Lemma H.7. *Under Condition C.3, when the results of Proposition 3.2 and Proposition H.4 hold at the initial stage and querying stage at a certain $t \leq T^*$, denoting $\mathbf{X}_\mathcal{P}^1 \subsetneq \mathcal{P}$ as the collection of all the data points with strong feature $\boldsymbol{\mu}_1$ in $\mathcal{P}$, and $\mathbf{X}_\mathcal{P}^2 \subsetneq \mathcal{P}$ as the collection of data points with weak feature $\boldsymbol{\mu}_2$, we have the conclusion that with probability more than 1-$\Theta(\delta')$, $\mathbf{X}_\mathcal{P}^1 \prec^{(t)} \mathbf{X}_\mathcal{P}^2$ holds.*

proof of Lemma H.7. See Lemma G.19 for a proof.

Similar to Lemma G.20, we directly have the following lemma demonstrate that both NAL algorithms would all prioritize those **perplexing samples**.

Lemma H.8. *(Formal Restatement of Proposition C.5) Under the same conditions in Proposition C.5, the Uncertainty Order and Diversity Order of the samples $[(y \cdot \boldsymbol{\mu}_l)^T, \xi^T]^T$ in sampling pool $\mathcal{P}$ follows the order of $\mathop{\mathbb{E}}_{j,k,l} \gamma_{j,k,\boldsymbol{\mu}_l}^{(t)}$.*

H.4. Label Complexity-based Test Error Analysis: XOR data version

The underlying philosophy in this section is the same as that in Appendix G.5 for the theory regarding linearly separable data. We propose that the results obtained in the previous section hold with high probability. By considering the scale of the coefficients, inner products, and the order of the data in the sampling pool $\mathcal{P}$, we can now examine the upper and lower bounds of the test error under different conditions before and after querying.

Lemma H.9. *Under Condition C.3, for a test set $\mathcal{D}^* \subseteq \mathcal{D}^*$ with occurrence probability p^* of the $\boldsymbol{\mu}_2$-equipped data, then $\exists\, t = \widetilde{O}\left(\eta^{-1}\varepsilon^{-1}m\widetilde{n}d^{-1}\sigma_p^{-2}\right)$, we have the following two situations before and after querying (i.e., $\forall s \in \{0,1\}$):*

- *For $t = \Omega\left(\widetilde{n}m/\left(\eta\sigma_p^2 d\varepsilon\right)\right)$, the training loss converges $L_S\left(\mathbf{W}^{(t)}\right) \leq \varepsilon$.*
- *If $\forall l \in \{1,2\}, n_{s,l} \geq \dfrac{\hat{C}_1\sigma_p^4 d}{\|\boldsymbol{\mu}_l\|_2^4}$ holds, we have the test error:*

$$L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \leq (1-p^*)\cdot\exp\left(\frac{-n_{s,1}\|\boldsymbol{\mu}_1\|_2^4}{\hat{C}_3\sigma_p^4 d}\right) + p^*\cdot\exp\left(\frac{-n_{s,2}\|\boldsymbol{\mu}_2\|_2^4}{\hat{C}_4\sigma_p^4 d}\right). \tag{77}$$

- *If $\exists l' \in \{1,2\} n_{s,l'} \leq \dfrac{\hat{C}_2\sigma_p^4 d}{\|\boldsymbol{\mu}_{l'}\|_2^4}$ holds, where $\hat{C}_1$ is from Condition 3.1, we have the test error*

$$L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \geq 0.12\cdot\tau_{l'}^*. \tag{78}$$

Here $\tau_{l'}^*$ denotes the occurrence probability of feature $\boldsymbol{\mu}_{l'}$, $\hat{C}_1$, $\hat{C}_2$, $\hat{C}_3$ and $\hat{C}_4$ are some positive constants.

Proof of Lemma H.9. The proof flow follows Theorem 3.2 in Meng et al. (2023) despite that we consider two features. For the training convergence, by Proposition H.2 we have

$$\begin{aligned} y_i f\left(\mathbf{W}^{(t)}, \mathbf{x}_i\right) &\geq -\frac{\kappa}{2} + \frac{1}{m}\sum_{r=1}^{m}\bar{\rho}_{y_i,r,i}^{(t)} \\ &\geq -\frac{\kappa}{2} + \underline{x}_t \\ &\geq -\kappa + \log\left(\Theta(\frac{\eta\sigma_p^2 d}{n_s m})t + \frac{2}{3}\right). \end{aligned}$$

Recall κ is defined in (58). Here, the first inequality is by the conclusion in Proposition H.2 and the second inequality is by (59) Proposition H.2, and last inequality are by (59). Then we have

$$L\left(\mathbf{W}^{(t)}\right) \leq \log\left(1 + \exp\{\kappa\}/\left(\Theta(\frac{\eta\sigma_p^2 d}{n_s m})t + \frac{2}{3}\right)\right) \leq \frac{e^\kappa}{\Theta(\frac{\eta\sigma_p^2 d}{n_s m})t + \frac{2}{3}} \leq \frac{e^\kappa}{2/\varepsilon + \frac{2}{3}} \leq \varepsilon$$

The last inequality is by $\log(1+x) \leq x$, $t \geq \Omega\left(\frac{\widetilde{n}m}{\eta\sigma_p^2 d\varepsilon}\right)$ and $exp\{\kappa\} \leq 1.5$.

For evaluating test error, same as techniques in Lemma G.21, we have

$$\begin{aligned} L_{\mathcal{D}^*}^{0-1}(\mathbf{W}) &= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*}[y\cdot f(\mathbf{W},\mathbf{x}) < 0] \\ &= (1-p^*)\cdot\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*_{\boldsymbol{\mu}_1}}[y\cdot f(\mathbf{W},\mathbf{x}) < 0] + p^*\cdot\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*_{\boldsymbol{\mu}_2}}[y\cdot f(\mathbf{W},\mathbf{x}) < 0], \end{aligned} \tag{79}$$

where $\mathcal{D}^*_{\boldsymbol{\mu}_1}$ and $\mathcal{D}^*_{\boldsymbol{\mu}_2}$ denotes the collection of data points in $\mathcal{D}$ containing feature $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively. Notably, $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}^*_{\boldsymbol{\mu}_l}}[y\cdot f(\mathbf{W},\mathbf{x}) < 0]$ is equal to

$$\sum_{\boldsymbol{\mu}\in\{\pm\mathbf{u}_l,\pm\mathbf{v}_l\}} P\left(yf\left(\mathbf{W}^{(t)},\mathbf{x}\right) > 0 \mid \mathbf{x}_{\text{signal part}} = \boldsymbol{\mu}\right)\cdot\frac{1}{4},$$

then without loss of generality, we can only investigate

$$P\left(1 \cdot f\left(\mathbf{W}^{(t)}, \mathbf{x}\right) > 0 \mid \mathbf{x}_{\text{signal part}} = \mathbf{u}_l\right), \forall l \in \{1, 2\}$$

and the proofs for other cases (i.e., $\boldsymbol{\mu} \in \{-\mathbf{u}_1, -\mathbf{u}_2, \pm\mathbf{v}_1, \pm\mathbf{v}_2\}$) are the same. Denote the feature patch in $\mathbf{x}$ as $\mathbf{u}_{l_x}$ ($l_x \in \{1, 2\}$), when $\mathbf{x} = \left(\mathbf{u}_{l_x}^\top, \boldsymbol{\xi}^\top\right)^\top$, the true label $y = +1$. Considering this case, we have

$$\begin{aligned} 1 \cdot f\left(\mathbf{W}^{(t)}, \mathbf{x}\right) &= \frac{1}{m}\sum_{r=1}^{m} F_{+1,r}\left(\mathbf{W}^{(t)}, \mathbf{u}_{l_x}\right) + F_{+1,r}\left(\mathbf{W}^{(t)}, \boldsymbol{\xi}\right) - \frac{1}{m}\sum_{r=1}^{m}\left(F_{-1,r}\left(\mathbf{W}^{(t)}, \mathbf{u}_{l_x}\right) + F_{-1,r}\left(\mathbf{W}^{(t)}, \boldsymbol{\xi}\right)\right) \\ &\leq \frac{1}{m}\left[\sum_r \sigma\left(\left\langle \mathbf{w}_{+1,r}^{(t)}, \mathbf{u}_{l_x}\right\rangle\right) - \sum_r \sigma\left(\left\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right]. \end{aligned}$$

Then we can adopt the exact same techniques in Lemma G.21. Recall $g(\boldsymbol{\xi})$ is denoted as $\sum_r \sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)$, also (48):

$$\mathbb{E}g(\boldsymbol{\xi}) = \sum_{r=1}^{m} \mathbb{E}\sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) = \sum_{r=1}^{m} \frac{\left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2 \sigma_p}{\sqrt{2\pi}} = \frac{\sigma_p}{\sqrt{2\pi}} \sum_{r=1}^{m} \left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2. \tag{80}$$

Then we can obtain the following test error upper bound on $\mathcal{D}^*_{\mathbf{u}_{l_x}}$ by adding $\mathbb{E}g(\boldsymbol{\xi})$ and $\frac{\sigma_p}{\sqrt{2\pi}}\sum_{r=1}^{m}\left\|\mathbf{w}_{-y,r}^{(t)}\right\|_2$ at two sides of the inequality:

$$\begin{aligned} \mathbb{P}_{(\mathbf{x},+1)\sim\mathcal{D}^*_{\mathbf{u}_{l_x}}}\left(1 \cdot f\left(\boldsymbol{W}^{(t)}, \mathbf{x}\right) \leq 0\right) &\leq P\left(\sum_r \sigma\left(\left\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) \geq \sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}_{l_x}\right\rangle\right)\right) \\ &= P\left(g(\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\xi}) \geq \sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}_{l_x}\right\rangle\right) - \frac{\sigma_p}{\sqrt{2\pi}}\sum_{r=1}^{m}\left\|\mathbf{w}_{-1,r}^{(t)}\right\|_2\right). \end{aligned} \tag{81}$$

By the results in Proposition H.3, we take a look at the comparison of the two terms at the right side of the inequality:

$$\frac{\sum_r \sigma\left(\left\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}_{l_x}\right\rangle\right)}{\sigma_p \sum_{r=1}^{m}\left\|\mathbf{w}_{-1,r}^{(t)}\right\|_2} \geq \frac{\Theta\left(\sum_r \gamma_{1,r,\mathbf{u}_{l_x}}^{(t)}\right)}{\Theta\left(d^{-1/2}n_s^{-1/2}\right) \cdot \sum_{r,i} \bar{\rho}_{-1,r,i}^{(t)}} = \Theta\left(\tau_{l_x} d^{1/2} n_s^{1/2} \operatorname{SNR}_{l_x}^2\right) = \Theta\left(\tau_{l_x} n_s^{1/2} \|\mathbf{u}_{l_x}\|_2^2 / (\sigma_p^2 d^{1/2})\right), \tag{82}$$

where τ_{l_x} denotes the proportion of feature $\mathbf{u}_{l_x}$ in current training data set (before or after querying). Worth noting that we have assumption in the first bullet that $\forall l \in \{1, 2\}, n_{s,l} \geq \frac{\hat{C}_1 \sigma_p^4 d}{\|\mathbf{u}_l\|_2^4}$, which means $n_{1,l_x}\|\mathbf{u}_1\|_2^4 \geq 2\hat{C}_1\sigma_p^4 d, \forall l_x \in \{1, 2\}$. Since $\hat{C}_1$ is a sufficiently large constant, it directly follows that

$$\sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}_{l_x}\right\rangle\right) - \frac{\sigma_p}{\sqrt{2\pi}}\sum_{r=1}^{m}\left\|\mathbf{w}_{-1,r}^{(t)}\right\|_2 > 0.$$

Same as (83), we adopt the techniques of Theorem 5.2.2 in Vershynin (2018):

$$P(g(\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\xi}) \geq x) \leq \exp\left(-\frac{cx^2}{\sigma_p^2 \|g\|_{\text{Lip}}^2}\right), \tag{83}$$

where c is a constant. To calculate the Lipschitz norm, we have

$$\begin{aligned}
|g(\boldsymbol{\xi}) - g(\boldsymbol{\xi}')| &= \left|\sum_{r=1}^{m} \sigma\left(\left\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) - \sum_{r=1}^{m} \sigma\left(\left\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}'\right\rangle\right)\right| \\
&\leq \sum_{r=1}^{m} \left|\sigma\left(\left\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) - \sigma\left(\left\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}'\right\rangle\right)\right| \\
&\leq \sum_{r=1}^{m} \left|\left\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi} - \boldsymbol{\xi}'\right\rangle\right| \\
&\leq \sum_{r=1}^{m} \left\|\mathbf{w}_{-1,r}^{(t)}\right\|_2 \cdot \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2 ,
\end{aligned}$$

where the first inequality is by triangle inequality; the second inequality is by the property of ReLU; the last inequality is by Cauchy Schwartz Inequality. Therefore, we have

$$\|g\|_{\text{Lip}} \leq \sum_{r=1}^{m} \left\|\mathbf{w}_{-1,r}^{(t)}\right\|_2 . \tag{84}$$

Utilize (83) and (84) in (81), we have

$$\begin{aligned}
\mathbb{P}_{(\mathbf{x},+1)\sim\mathcal{D}^*_{\mathbf{u}_{l_x}}}\left(1 \cdot f\left(\boldsymbol{W}^{(t)}, \mathbf{x}\right) \leq 0\right) &\leq \exp\left[-\frac{c\left(\sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}_{l_x}\right\rangle\right) - \left(\frac{\sigma_p}{\sqrt{2\pi}}\right)\sum_{r=1}^{m}\left\|\mathbf{w}_{-1,r}^{(t)}\right\|_2\right)^2}{\sigma_p^2\left(\sum_{r=1}^{m}\left\|\mathbf{w}_{-1,r}^{(t)}\right\|_2\right)^2}\right] \\
&= \exp\left[-c\left(\frac{\sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}_{l_x}\right\rangle\right)}{\sigma_p \sum_{r=1}^{m}\|\mathbf{w}_{-1,r}^{(t)}\|_2} - \frac{1}{\sqrt{2\pi}}\right)^2\right] \\
&\leq \exp(c/2\pi)\exp\left(-0.5c\left(\frac{\sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}_{l_x}\right\rangle\right)}{\sigma_p \sum_{r=1}^{m}\left\|\mathbf{w}_{-1,r}^{(t)}\right\|_2}\right)^2\right),
\end{aligned} \tag{85}$$

where the third inequality is by $(s-t)^2 \geq s^2/2 - t^2, \forall s, t \geq 0$. And then by (82) and (85), we can have

$$\begin{aligned}
\mathbb{P}_{(\mathbf{x},+1)\sim\mathcal{D}^*_{\mathbf{u}_{l_x}}}\left(1 \cdot f\left(\boldsymbol{W}^{(t)}, \mathbf{x}\right) \leq 0\right) &\leq \exp(c/2\pi)\exp\left(-0.5c\left(\frac{\sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}_{l_x}\right\rangle\right)}{\sigma_p \sum_{r=1}^{m}\left\|\mathbf{w}_{-1,r}^{(t)}\right\|_2}\right)^2\right) \\
&= \exp\left(\frac{c}{2\pi} - \frac{\tau_{l_x} n_{s,l_x}\|\mathbf{u}_{l_x}\|_2^4}{\hat{C}\sigma_p^4 d}\right) \\
&= \exp\left(\frac{c}{2\pi} - \frac{n_{s,l_x}\|\mathbf{u}_{l_x}\|_2^4}{\hat{C}_{l_x}\sigma_p^4 d}\right) \\
&\leq \exp\left(-\frac{n_{s,l_x}\|\mathbf{u}_{l_x}\|_2^4}{2\hat{C}_{l_x}\sigma_p^4 d}\right)
\end{aligned} \tag{86}$$

where $\hat{C}_{l_x} = \hat{C}/\tau_{lx} = O(1)$; the last inequality holds if we choose $\hat{C}_1 \geq c\hat{C}_{l_x}/\pi$, for any $l_x \in \{1, 2\}$. If we choose $\hat{C}_3$ as $2\hat{C}_{l_1}$ and $\hat{C}_4$ as $2\hat{C}_{l_2}$, by (79) and (86) we have

$$L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \leq (1-p^*) \cdot \exp\left(\frac{-n_{s,1}\|\mathbf{u}_1\|_2^4}{\hat{C}_3\sigma_p^4 d}\right) + p^* \cdot \exp\left(\frac{-n_{s,2}\|\mathbf{u}_2\|_2^4}{\hat{C}_4\sigma_p^4 d}\right).$$

Next, we serve to prove the test error upper bound. Same as the proof in Lemma G.21, we utilize the pigeonhole principle technique in Kou et al. (2023b); Meng et al. (2023), which is based on the following two lemmas.

Lemma H.10. *For $t \in [T_1, T^*]$, denote $g(\boldsymbol{\xi}) = \sum_{j,r} \sigma\left(\left\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \right\rangle\right)$. There exists a fixed vector $\mathbf{v}_l$ with $\|\mathbf{v}_l\|_2 \leq 0.01\sigma_p$ and constant $\hat{C}_6$ such that*

$$\sum_{j' \in \{\pm 1\}} \left[g\left(j'\boldsymbol{\xi} + \mathbf{v}_l\right) - g\left(j'\boldsymbol{\xi}\right)\right] \geq 4\hat{C}_6 \max_{j,l} \left\{\sum_r \gamma_{j,r,\boldsymbol{\mu}_l}^{(t)}\right\},$$

for all $\boldsymbol{\xi} \in \mathbb{R}^d$.

Proof of Lemma H.10. See Lemma 5.8 in Kou et al. (2023b) or Theorem 3.2 in Meng et al. (2023) for a proof, where we utilize a large enough $\hat{C}_2$ in the condition given in the second bullet point ($n_{s,l'} \leq \dfrac{\hat{C}_2 \sigma_p^4 d}{\|\boldsymbol{\mu}_{l'}\|_2^4}$) to control the norm of $\mathbf{v}_l$.

Lemma H.11. *(Proposition 2.1 in Devroye et al. (2023)). The TV distance between $\mathcal{N}\left(0, \sigma_p^2 \mathbf{I}_d\right)$ and $\mathcal{N}\left(\mathbf{v}_l, \sigma_p^2 \mathbf{I}_d\right)$ is smaller than $\|\mathbf{v}_l\|_2 / 2\sigma_p$.*

Proof of Lemma H.11. See Proposition 2.1 in Devroye et al. (2023) for a proof.

Now we take a look at $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right)$, by (79) we have:

$$\begin{aligned}
L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) &= \tau_1^* \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\boldsymbol{\mu}_1}^*}\left[y \cdot f(\mathbf{W}, \mathbf{x}) < 0\right] + \tau_2^* \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\boldsymbol{\mu}_2}}\left[y \cdot f(\mathbf{W}, \mathbf{x}) < 0\right] \\
&\geq \tau_{l'}^* \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\boldsymbol{\mu}_{l'}}^*}\left[y \cdot f(\mathbf{W}, \mathbf{x}) < 0\right] \\
&\geq 0.5\tau_{l'}^* \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\boldsymbol{\mu}_{l'}}^*}\Bigg(\left|\sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) - \sum_r \sigma\left(\left\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right| \\
&\qquad\qquad \geq \hat{C}_6 \max\left\{\sum_r \gamma_{1,r,\boldsymbol{\mu}_{l'}}^{(t)}, \sum_r \gamma_{-1,r,\boldsymbol{\mu}_{l'}}^{(t)}\right\}\Bigg) \\
&= 0.5\tau_{l'}^* \cdot P(\Omega_{\boldsymbol{\xi}}),
\end{aligned} \tag{87}$$

where $\Omega_{\boldsymbol{\xi}} := \left\{\boldsymbol{\xi} \middle| |g(\boldsymbol{\xi})| \geq \hat{C}_6 \max\left\{\sum_r \gamma_{1,r,\boldsymbol{\mu}_{l'}}^{(t)}, \sum_r \gamma_{-1,r,\boldsymbol{\mu}_{l'}}^{(t)}\right\}\right\}$. The last inequality holds since we can always have a corresponding y to make a wrong prediction if given $\boldsymbol{\xi}$, the $\left|\sum_r \sigma\left(\left\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right) - \sum_r \sigma\left(\left\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}\right\rangle\right)\right|$ is large enough.

Next, we seek a lower bound of $P(\Omega_{\boldsymbol{\xi}})$. By Lemma H.10, we have that $\sum_j [g(j\boldsymbol{\xi} + \mathbf{v}_l) - g(j\boldsymbol{\xi})] \geq 4\hat{C}_6 \max_{j,l}\left\{\sum_r \gamma_{j,r,\boldsymbol{\mu}_l}^{(t)}\right\}$. Then by pigeon's hole principle, there must exist one of the $\boldsymbol{\xi}, \boldsymbol{\xi} + \mathbf{v}_l, -\boldsymbol{\xi}, -\boldsymbol{\xi} + \mathbf{v}_l$ belongs $\Omega_{\boldsymbol{\xi}}$. So we have proved that $\Omega_{\boldsymbol{\xi}} \cup -\Omega_{\boldsymbol{\xi}} \cup \Omega_{\boldsymbol{\xi}} - \{\mathbf{v}_l\} \cup -\Omega_{\boldsymbol{\xi}} - \{\mathbf{v}_l\} = \mathbb{R}^d$. Therefore at least one of $P(\Omega_{\boldsymbol{\xi}}), P(-\Omega_{\boldsymbol{\xi}}), P(\Omega_{\boldsymbol{\xi}} - \{\mathbf{v}_l\}), P(\Omega_{\boldsymbol{\xi}} - \{\mathbf{v}_l\}), P(-\Omega_{\boldsymbol{\xi}} - \{\mathbf{v}_l\})$ is greater than 0.25. By the definition of TV distance, we have:

$$\begin{aligned}
|P(\Omega_{\boldsymbol{\xi}}) - P(\Omega_{\boldsymbol{\xi}} - \mathbf{v}_l)| &= \left|\mathbb{P}_{\boldsymbol{\xi}\sim\mathcal{N}\left(0,\sigma_p^2\mathbf{I}_d\right)}(\boldsymbol{\xi} \in \Omega_{\boldsymbol{\xi}}) - \mathbb{P}_{\boldsymbol{\xi}\sim\mathcal{N}\left(\mathbf{v}_l,\sigma_p^2\mathbf{I}_d\right)}(\boldsymbol{\xi} \in \Omega_{\boldsymbol{\xi}})\right| \\
&\leq \mathrm{TV}\left(\mathcal{N}\left(0, \sigma_p^2 \mathbf{I}_d\right), \mathcal{N}\left(\mathbf{v}_l, \sigma_p^2 \mathbf{I}_d\right)\right) \\
&\leq \frac{\|\mathbf{v}_l\|_2}{2\sigma_p} \\
&\leq 0.02.
\end{aligned}$$

Also, notice that $P(-\Omega_{\boldsymbol{\xi}}) = P(\Omega_{\boldsymbol{\xi}})$, we have $4P(\Omega_{\boldsymbol{\xi}}) \geq 1 - 2 \cdot 0.02$. Thus $L_{\mathcal{D}^*}^{0-1}\left(\mathbf{W}^{(t)}\right) \geq 0.5\tau_{l'}^* \cdot 0.24 = 0.12 \cdot \tau_{l'}^*$. The proofs of Lemma H.9 complete.

Similar to the proof process in Appendix G.5, our main focus is to verify whether the NAL algorithms satisfy the condition stated in the first bullet point of Lemma H.9. Conversely, it is highly probable that Random Sampling satisfies the condition stated in the second bullet point. The following proposition validates this intuition.

Proposition H.12. *When Lemma H.7 holds, and the sampling size of algorithm satisfies $\dfrac{\hat{C}_1 \sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4} - \dfrac{pn_0}{2} \leq n^* = \Theta(\widetilde{n} - n_0) \leq \widetilde{n} - n_0$, we have the following:*

- *The number of data with strong feature patch $n_{s,1}$ satisfies* $n_{s,1} \geq \dfrac{\hat{C}_1 \sigma_p^4 d}{\|\boldsymbol{\mu}_1\|_2^4}, \forall s \in \{0,1\}$.
- *The number of data with weak feature patch $n_{s,2}$ before querying and after* ***Random Sampling*** *satisfies* $n_{s,2} \leq \dfrac{\hat{C}_2 \sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4}, \forall s \in \{0,1\}$.
- *The total number of data with weak feature patch $n_{1,2}$ after* ***Uncertainty Sampling*** *and* ***Diversity Sampling*** *satisfies* $n_{1,2} \geq \dfrac{\hat{C}_1 \sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4}$.

For the sake of coherence, here $\hat{C}_1$ and $\hat{C}_2$ are some constants shared with Theorem C.6.

Proof of Proposition H.12. According to the conditions stated in Definition C.1, we have $(1 - \frac{3}{2}p)n_0 \geq \frac{\hat{C}_1 \sigma_p^4 d}{\|\boldsymbol{\mu}_1\|_2^4}$ for a large constant $\hat{C}_1$. By substituting the results of n_p for n_0 from Lemma G.3, as well as the definition of $n_{s,l}$, we obtain the following:

$$n_{1,1} \geq n_{0,1} \geq (1 - \frac{3}{2}p)n_0 \geq \frac{\hat{C}_1 \sigma_p^4 d}{\|\boldsymbol{\mu}_1\|_2^4}.$$

For the second bullet, by Lemma G.3, Lemma H.7 and conditions $n^* \geq \dfrac{\hat{C}_1 \sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4} - \dfrac{pn_0}{2}$, we have:

$$n_{1,2} \geq \frac{pn_0}{2} + n^* \geq \frac{\hat{C}_1 \sigma_p^4 d}{\|\boldsymbol{\mu}_2\|_2^4}$$

Furthermore, by using Lemma G.3 and the condition $\widetilde{n} \leq \dfrac{2\hat{C}_2 \sigma_p^4 d}{3p\|\boldsymbol{\mu}_2\|_2^4}$, the third bullet point is satisfied straightforwardly.

Based on the results of Lemma H.9 and Proposition H.12, the conclusions of Proposition C.4 and Theorem C.6 follow directly.

I. Attribution of Lion Images

In Figure 1, a collection of various lion images found on Google is presented. Due to the challenge of accurately determining the copyright attribution of these images, specific acknowledgments to individual websites or sources cannot be provided here. However, we express our gratitude to all creators, and sincerely hope that they do not find any offense in the use of their work for illustrative purposes in our paper.