# Revisiting Discriminative vs. Generative Classifiers: Theory and Implications

Chenyu Zheng [1] Guoqiang Wu [2] Fan Bao [3] Yue Cao [4] Chongxuan Li [1] Jun Zhu [3]

## Abstract

A large-scale deep model pre-trained on massive labeled or unlabeled data transfers well to downstream tasks. *Linear evaluation* freezes parameters in the pre-trained model and trains a linear classifier separately, which is efficient and attractive for transfer. However, little work has investigated the classifier in linear evaluation except for the default logistic regression. Inspired by the statistical efficiency of naïve Bayes, the paper revisits the classical topic on *discriminative vs. generative classifiers* (Ng & Jordan, 2001). Theoretically, the paper considers the surrogate loss instead of the zero-one loss in analyses and generalizes the classical results from binary cases to multiclass ones. We show that, under mild assumptions, multiclass naïve Bayes requires $O(\log n)$ samples to approach its asymptotic error while the corresponding multiclass logistic regression requires $O(n)$ samples, where $n$ is the feature dimension. To establish it, we present a *multiclass $\mathcal{H}$-consistency bound* framework and an explicit bound for logistic loss, which are of independent interests. Simulation results on a mixture of Gaussian validate our theoretical findings. Experiments on various pre-trained deep vision models show that naïve Bayes consistently converges faster as the number of data increases. Besides, naïve Bayes shows promise in few-shot cases and we observe the "two regimes" phenomenon in pre-trained supervised models. Our code is available at *https://github.com/ML-GSAI/Revisiting-Dis-vs-Gen-Classifiers*.

## 1. Introduction

Deep representation learning has achieved great success in many fields such as computer vision (Ren et al., 2015; He et al., 2017; Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Chen & He, 2021; Grill et al., 2020; He et al., 2022), natural language processing (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020) and cross-modal learning (Radford et al., 2021) over the past few years. The common paradigm behind them is to (pre-)train a large-scale model on an enormous amount of labeled or unlabeled data and transfer it to downstream tasks. During the transfer, *linear evaluation* (Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Grill et al., 2020; Radford et al., 2021) freezes all parameters in the pre-trained model and learns a linear classifier separately. Theoretically, it is validated by the (approximate) linear separability of the representations extracted by pre-trained models (Saunshi et al., 2019; Lee et al., 2021; Tosh et al., 2021; HaoChen et al., 2021). Practically, linear evaluation is an efficient and attractive alternative to fine-tuning, considering the extremely large and continually growing size of modern pre-trained models.

Although new algorithms and models for deep pre-training emerge in endlessly, little work has investigated the classifier except for the default logistic regression. Directly inspired by the classical work (Efron, 1975; Ng & Jordan, 2001) (detailed in Section 2) on the statistical efficiency of generative linear classifiers (e.g. naïve Bayes), we revisit the discriminative vs. generative linear classifiers in the context of deep representation learning.

In Section 3, we improve the classical theory (Ng & Jordan, 2001) in two aspects for subsequent analysis in deep representation learning. First, we characterize asymptotic behaviors of both multiclass naïve Bayes and logistic regression, generalizing the results in binary classification (Ng & Jordan, 2001). Second, in logistic regression, we consider the practically used surrogate loss in our analysis instead of directly optimizing the zero-one loss as assumed in (Ng & Jordan, 2001). To establish it, we introduce a general *multiclass $\mathcal{H}$-consistency bound* framework upon recent advances (Awasthi et al., 2022a) and a nontrivial explicit bound for multiclass logistic regression, which are of independent interests. We prove that for a fixed number of classes, the number of samples required to approach the

[1] Gaoling School of AI, Renmin University of China; Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China [2] School of Software, Shandong University [3] Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua-Huawei Joint Center for AI, BNRist Center, THBI Lab, Tsinghua University [4] Beijing Academy of Artificial Intelligence. Correspondence to: Chongxuan Li <chongxuanli@ruc.edu.cn>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

corresponding optimal classifier is $O(\log n)$ and $O(n)$ for naïve Bayes and logistic regression respectively, where $n$ is the feature dimension. We conduct synthetic experiments with tractable $\mathcal{H}$-optimal classifiers to validate our theory.

In Section 4, we discuss the implications of our theory in the linear evaluation of pre-trained deep models. We first analyze the main assumptions in our theory upon deep representations. We then perform extensive experiments on CIFAR10 and CIFAR100 datasets with various representative pre-trained vision models (He et al., 2016; Dosovitskiy et al., 2021; Chen et al., 2020d;c; Radford et al., 2021; Xie et al., 2022; He et al., 2022), which are trained in supervised or self-supervised manners. The results show that naïve Bayes consistently converges faster as the number of data increases in all settings, which agrees with our theory. Besides, naïve Bayes shows promise in few-shot cases and we observe the "two regimes" phonomenan (Ng & Jordan, 2001) in models pre-trained in a supervised manner, suggesting a distinction between the representations learned by supervised and self-supervised approaches.

## 2. Preliminaries

In this section, we present notations and preliminaries on discriminative vs. generative classifiers and $\mathcal{H}$-consistency.

Let lower, boldface lower and capital case letters denote scalers (e.g., a), vectors (e.g., $\boldsymbol{a}$), and matrices (e.g., $\boldsymbol{A}$), respectively. For a matrix $\boldsymbol{A}$, $\boldsymbol{A}_i$ and $A_{ij}$ denote its $i$-th row and $(i,j)$-th element. For a vector $\boldsymbol{a}$, $a_i$ denotes its $i$-th element. Similarly, for a vector function $\boldsymbol{f}$, $f_i(\boldsymbol{x})$ denotes the $i$-th element of $\boldsymbol{f}(\boldsymbol{x})$. We do not distinguish constants and random variables in notations if there is no confusion. We denote the KL divergence between distributions $p$ and $q$ by $D(p\|q)$. We use $\mathbb{E}$, $\mathbb{V}$, $\Delta_k$ to represent expectation, variance, and $k$-dimensional possibility simplex, respectively.

Let $\mathcal{X}$ denote the domain set and $\mathcal{Y} = \{1, \dots, K\}$ denote the label set, where $K$ is the number of classes. For simplicity, we assume $\mathcal{X} = \{0,1\}^n$ when inputs are discrete and $\mathcal{X} = [0,1]^n$ otherwise, where $n$ is the feature dimension. Note that our analysis can be easily extended to the general case with any bounded features. Let $\mathcal{H}$ be a hypothesis set of functions mapping from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}^K$. The prediction associated by a hypothesis $\boldsymbol{h} \in \mathcal{H}$ and $\boldsymbol{x} \in \mathcal{X}$ is $\operatorname{argmax}_{y\in\mathcal{Y}} h_y(\boldsymbol{x})$. In the main paper, we focus on the family of constrained linear hypotheses $\mathcal{H}_{lin} = \{\boldsymbol{x} \to \boldsymbol{h}(\boldsymbol{x}) : h_y(\boldsymbol{x}) = \langle \boldsymbol{w}_y, \boldsymbol{x}\rangle + b_y, \|\boldsymbol{w}_y\|_2 \le W, |b_y| \le B, y \in \mathcal{Y}\}$, where $W, B \in \mathbb{R}^+$. We also denote the hypothesis set of all measurable functions by $\mathcal{H}_{all}$. Given a hypothesis set $\mathcal{H}$ and distribution $\mathcal{D}$, the generalization error and minimal generalization error of a hypothesis $\boldsymbol{h}$ with respect to the loss function $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ are defined as $R_\ell(\boldsymbol{h}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(\boldsymbol{h}(\boldsymbol{x}), y)]$ and $R^*_{\ell,\mathcal{H}} = \inf_{\boldsymbol{h}\in\mathcal{H}} R_\ell(\boldsymbol{h})$.

### 2.1. Discriminative vs. Generative Classifiers

$K$-class logistic regression is parameterized by $[\boldsymbol{w}_1, \dots, \boldsymbol{w}_K, \boldsymbol{b}]$, where $\boldsymbol{w}_i \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^K$. Its prediction is given by $\operatorname{argmax}_{y\in\mathcal{Y}}(\langle \boldsymbol{w}_y, \boldsymbol{x}\rangle + b_y)$.

It's well known that the generative counterpart of the logistic regression is naïve Bayes (with some constraints presented later) (Ng & Jordan, 2001; Rubinstein & Hastie, 1997). When inputs are discrete, a naïve Bayes classifier uses a training set with $m$ i.i.d examples to calculate the empirical conditional distributions $\hat{p}(x_i|y)$ and empirical marginal distribution $\hat{p}(y)$ as follows:

$$\hat{p}(x_i = 1|y = k) = \frac{\#\{x_i = 1, y = k\} + \alpha}{\#\{y = k\} + K\alpha}, \quad (1)$$

$$\hat{p}(y = k) = \frac{\#\{y = k\} + \alpha}{m + K\alpha}, \quad (2)$$

where $\#\{\cdot\}$ is the counting function and $\alpha$ is a positive Laplace smoothing parameter. Corresponding population versions are denoted by $p(x_i|y)$ and $p(y)$ respectively. In case of continuous inputs, we let $\hat{p}(x_i|y = k)$ be a univariate Gaussian distribution with parameters $\hat{\mu}_{ki}$ and $\hat{\sigma}_i^2$. We note that $\hat{\sigma}_i^2$s do not depend on $y$ to keep the linearity of its decision boundary, otherwise logistic regression and naïve Bayes are no longer a fair discriminative-generative pair (Xue & Titterington, 2008). They are calculated as the empirical version of $\mu_{ki} = \mathbb{E}[x_i|y = k]$ and $\sigma_i^2 = \mathbb{E}_y[\mathbb{V}(x_i|y)]$.

Ng & Jordan (2001) proved that in binary classification, logistic regression enjoys a lower asymptotic error but approaches it much slower (w.r.t. the sample size) than naïve Bayes. The theory explains the *"two regimes"* (Ng & Jordan, 2001) phenomenon in practice. In particular, naïve Bayes generalizes better with limited data. However, the multiclass case has not been investigated yet, which is the main focus of this paper. Besides, prior work (Ng & Jordan, 2001) assumes that the zero-one loss can be directly optimized in logistic regression, which is impractical. To weaken the assumption, we introduce tools from *$\mathcal{H}$-consistency*.

### 2.2. $\mathcal{H}$-consistency

$\mathcal{H}$-consistency (Long & Servedio, 2013) analyzes the relationship between the estimation error of zero-one loss w.r.t. a hypothesis class $\mathcal{H}$ and that of a surrogate loss. It includes the classical Bayes consistency (Zhang, 2004b; Bartlett et al., 2006; Tewari & Bartlett, 2007) as a special case by setting $\mathcal{H}$ to $\mathcal{H}_{all}$. In this paper, we analyze the linear discriminative vs. generative classifiers upon recent advances on $\mathcal{H}$-consistency bounds (Awasthi et al., 2022a).

We first introduce some notations. We denote by $\boldsymbol{p}(\boldsymbol{x})$ the conditional distribution of $Y$ given $\boldsymbol{x}$, i.e., $p_y(\boldsymbol{x}) = \mathbb{P}(Y = y|X = \boldsymbol{x})$. We define the conditional risk as $\mathcal{C}_\ell(\boldsymbol{h}, \boldsymbol{x}) = \sum_{y=1}^K p_y(\boldsymbol{x})\ell(\boldsymbol{h}(\boldsymbol{x}), y)$, and note that generalization error

$R_\ell(\boldsymbol{h})$ can be rewritten as $\mathbb{E}_{\boldsymbol{x}}[\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x})]$. We also define its infimum $\mathcal{C}^*_{\ell,\mathcal{H}}(\boldsymbol{x}) = \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x})$ and the gap between them $\Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}) = \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x}) - \mathcal{C}^*_{\ell,\mathcal{H}}(\boldsymbol{x})$. A key quantity appears in our bounds is $M_{\ell,\mathcal{H}} = R^*_{\ell,\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}(\mathcal{C}^*_{\ell,\mathcal{H}}(\boldsymbol{x}))$, which is difficult to estimate (Awasthi et al., 2022a), but can be bounded by the approximate error. In addition, for any $\boldsymbol{p}$ in probability simplex $\Delta_K$, we can define $\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) = \sum_{y=1}^K p_y \ell(\boldsymbol{h}(\boldsymbol{x}),y)$ and $\Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) = \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})$. Furthermore, we define the $\epsilon$-regret of $t$ as $\langle t\rangle_\epsilon = t\mathbb{1}_{t>\epsilon}$.

The general $\mathcal{H}$-consistency bound (Awasthi et al., 2022a) for two loss functions $\ell_1$ and $\ell_2$ is defined as follows.

**Definition 2.1** ($\mathcal{H}$-consistency bound)**.** *$\mathcal{H}$-consistency bound is in the following form that holds for all $\boldsymbol{h}\in\mathcal{H}$, $\mathcal{D}\in\mathcal{P}$ and some non-decreasing function $f:\mathbb{R}_+\to\mathbb{R}_+$:*

$$R_{\ell_2}(\boldsymbol{h}) - R^*_{\ell_2,\mathcal{H}} \le f(R_{\ell_1}(\boldsymbol{h}) - R^*_{\ell_1,\mathcal{H}}). \qquad (3)$$

*If $\mathcal{P}$ is composed of all distributions over $\mathcal{X}\times\mathcal{Y}$, we call it a distribution-independent bound.*

Note that it covers the classical Bayes consistency bounds (Bartlett et al., 2006) by setting $\mathcal{H}=\mathcal{H}_{all}$. When $\ell_1$ is logistic loss $\ell_{log}$ and $\ell_2$ is zero-one loss $\ell_{0-1}$, Awasthi et al. (2022a) proved the following $\mathcal{H}$-consistency bound w.r.t. the bounded linear hypotheses.

**Theorem 2.1** ($\mathcal{H}$-consistency bound for binary logistic loss and zero-one loss, Appendix K.1.2 (Awasthi et al., 2022a))**.** *Given binary linear hypothesis set $\mathcal{H} = \{\boldsymbol{x}\to\langle\boldsymbol{w},\boldsymbol{x}\rangle + b : \|\boldsymbol{w}\|_2 \le W, |b|\le B\}$, if $R_{\ell_{log}}(h) - R^*_{\ell_{log},\mathcal{H}} + M_{\ell_{log},\mathcal{H}} \le \frac{1}{2}(\frac{e^B-1}{e^B+1})^2$, then it holds for any distribution that $R_{\ell_{0-1}}(h) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}} \le \sqrt{2}(R_{\ell_{log}}(h) - R^*_{\ell_{log},\mathcal{H}} + M_{\ell_{log},\mathcal{H}})^{\frac{1}{2}}$.*

To the best of our knowledge, there is no $\mathcal{H}$-consistency bound for logistic loss and zero-one loss in multiclass classification[1]. In this paper, we extend the binary framework (Awasthi et al., 2022a) to multiclass cases and derive an explicit bound for logistic loss.

## 3. Theory

In this section, we present our main theoretical results in Section 3.1: Under some mild assumptions, for any fixed class number $K$, the number of training samples required by naïve Bayes to approach its asymptotic error is $O(\log n)$ (Theorem 3.2), and that of logistic regression is $O(n)$ (Theorem 3.4). To establish it, we propose a general multiclass $\mathcal{H}$-consistency framework (Theorem 3.5) and a nontrivial multiclass $\mathcal{H}$-consistency bound for logistic loss and zero-one loss (Theorem 3.3) in Section 3.2. Notably, our theory includes the analysis for $K=2$ in Appendix B as a special case.

### 3.1. On Multiclass Discriminative vs. Generative Linear Classifiers

Let $\boldsymbol{h}_{Dis,m}$ and $\boldsymbol{h}_{Gen,m}$ denote the hypothesis returned by multiclass logistic regression and naïve Bayes with $m$ $i.i.d$ samples, respectively. Let $\boldsymbol{h}_{Dis,\infty}$ and $\boldsymbol{h}_{Gen,\infty}$ be the corresponding asymptotic version. We are interested in comparing the statistical efficiency of naïve Bayes and logistic regression (Ng & Jordan, 2001). Formally, we need to bound $R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty})$ and $R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty})$ respectively.

**Naïve Bayes.** Notably, the solution of Naïve Bayes is in a closed-form, as presented in Eq. (1&2). Therefore, we can characterize the gap between parameters in $\boldsymbol{h}_{Gen,m}$ and $\boldsymbol{h}_{Gen,\infty}$ to bound $R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty})$, similarly to the binary case (Ng & Jordan, 2001).

We make two mild assumptions about the data distribution similar to Ng & Jordan (2001). We avoid trivial cases where $p(y=k)=1$ or $p(y=k)=0$ for some $k$ in Assumption 3.1 and assume that the conditional distribution of $\boldsymbol{x}$ given $y$ can not be too concentrated in Assumption 3.2.

**Assumption 3.1.** *For some fixed $\rho_1\in(0,\frac{1}{2}]$, we have that $\rho_1\le p(y=k)\le 1-\rho_1$ for all $k\in\mathcal{Y}$.*

**Assumption 3.2.** *For some fixed $\rho_2\in(0,\frac{1}{2}]$, $\rho_2\le p(x_i=1|y=k)\le 1-\rho_2$ for all $i,k$ in the discrete case, and $\sigma_i^2\ge\rho_2$ for all $i$ in the continuous case.*

In practice, most deep learning work considers the balanced case where $\rho_1=\frac{1}{K}$ (Deng et al., 2009). Empirically, we found that $\rho_2\in[10^{-5},10^{-2}]$ on the features extracted by representative pre-trained vision models in Section 4. For clarity, we denote $\rho_0=\min\{\rho_1,\rho_2\}$ throughout the paper. We now define two key quantities in our proof as follows.

**Definition 3.1** (Pair activation function of naïve Bayes)**.** *For every $k_1,k_2\in\mathcal{Y}$, we define the pair activation function $\Delta a_{Gen}(\boldsymbol{x},k_1,k_2)$ as*

$$\Delta a_{Gen}(\boldsymbol{x},k_1,k_2) = a_{Gen}(\boldsymbol{x},k_1) - a_{Gen}(\boldsymbol{x},k_2), \qquad (4)$$

*where $a_{Gen}(\boldsymbol{x},k) = \sum_{i=1}^n \log\hat{p}(x_i|y=k) + \log\hat{p}(y=k)$.*

The paired activation function is important because it connects the estimated parameters and predictions of the hypothesis. For instance, $\Delta a_{Gen}(\boldsymbol{x},k_1,k_2)>0$ means that $\boldsymbol{x}$ is more likely to be predicted as an instance of class $k_1$ than class $k_2$. We can easily bound the gap between the parameters in $\boldsymbol{h}_{Gen,m}$ and $\boldsymbol{h}_{Gen,\infty}$ by standard concentration inequalities. To bound $R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty})$ as presented in Theorem 3.1, we further upper bound the probability of getting "bad training samples", which are pre-

[1] Most recently, the concurrent and independent work of Mao et al. (2023) also studies this problem and obtains similar results to ours.

dicted as different classes with high probability by $\boldsymbol{h}_{Gen,m}$ and $\boldsymbol{h}_{Gen,\infty}$, via the following $\widetilde{G}(\tau)$.

**Definition 3.2.** *We define the function $\widetilde{G}(\tau)$ as follows:*

$$\widetilde{G}(\tau) = \max_{k_1,k_2} \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}}(|\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)| \leq \tau n).$$

**Theorem 3.1** (Proof in Appendix D.1)**.** *Suppose that Assumption 3.1 and 3.2 are valid. Then with probability at least $1-\delta$:*

$$R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) \leq R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty}) + \frac{K(K-1)}{2}\Bigg(\widetilde{G}\big(O(\sqrt{\frac{1}{m}\log(\frac{n}{\delta})})\big) + \delta\Bigg).$$

The core of Theorem 3.1 is the $\widetilde{G}(\tau)$, which must be small when $\tau$ is small in order to obtain meaningful bound about $R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty})$. It holds under the following assumptions, similarly to Ng & Jordan (2001).

**Assumption 3.3.** *For all $k_1, k_2 (k_1 \neq k_2)$ and $k \in \mathcal{Y}$, it holds that $|\sum_{i=1}^{n}(D(p(x_i|y=k)\|p(x_i|y=k_1)) - D(p(x_i|y=k)\|p(x_i|y=k_2)))| = \beta_{k_1,k_2,k}n = \Omega(n)$.*

**Assumption 3.4.** *For all $k_1, k_2 (k_1 \neq k_2)$ and $k \in \mathcal{Y}$, it holds that $\mathbb{V}_{\boldsymbol{x}}[\sum_{i=1}^{n}\log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}|y=k] = \alpha_{k_1,k_2,k}n = O(n^r)$ for any $r \in [1,2)$.*

Intuitively, Assumption 3.3 requires that $\Omega(1)$ fraction of features distinct for any two different classes. Assumption 3.4 is more technical. In fact, it is derived when we attempt to bound $\widetilde{G}(\tau)$ via Chebyshev's inequality[2]. We empirically analyze both assumptions in Section 4. Proposition 3.1 presents a meaningful bound for $\widetilde{G}(\tau)$, which is followed by the main result of naïve Bayes in Theorem 3.2.

**Proposition 3.1** (Proof in Appendix D.2)**.** *Suppose that Assumption 3.1, 3.3 and 3.4 hold, then $\widetilde{G}(\tau)$ is polynomially small in $n$:*

$$\widetilde{G}(\tau) \leq \frac{\alpha}{(\tau-\zeta)^2 n},$$

*where $\alpha = \max_{k_1,k_2,k}\alpha_{k_1,k_2,k} = O(n^{r-1})$, $\mathbb{E}_{\boldsymbol{x}}[\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)|y=k] = \zeta_{k_1,k_2,k}n$, $\zeta = \min_{k_1,k_2,k}|\zeta_{k_1,k_2,k}| = \Omega(1)$ and $\tau < \zeta$.*

**Theorem 3.2** (Results for naïve Bayes, proof in Appendix D.3)**.** *Suppose the precondition of Proposition 3.1 holds. Then, it suffices to pick $m = O(\log n)$ training samples such that $R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) \leq R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty}) + \epsilon_0$ hold with probability $1-\delta_0$, for any $\epsilon_0 \in (0,1)$ and $\delta_0 \in (0, \frac{\epsilon_0}{K^2}]$.*

**Logistic Regression.** To directly compare with naïve Bayes, we aim to bound $R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty})$. However, the optimization of logistic regression does not have an analytic form, making the proof idea of naïve Bayes infeasible. Besides, Ng & Jordan (2001) proves the bound by directly optimizing the zero-one loss, which is impractical. Instead, we present a bound considering the surrogate logistic loss in this paper. To establish it, we exploit recent advances on $\mathcal{H}$-consistency bound (Awasthi et al., 2022a) as detailed in Defition 2.1. It is worth discussing an alternative approach based on *Bayes consistency bounds* (Bartlett et al., 2006). For a direct comparison with naïve Bayes, we care about the asymptotic error in $\mathcal{H}_{lin}$ instead of $\mathcal{H}_{all}$. Therefore, a $\mathcal{H}$-consistency bound is more natural and potentially tighter than a Bayes consistency bound. In fact, existing Bayes consistency bounds (Bartlett et al., 2006) are special cases of the $\mathcal{H}$-consistency bounds (Awasthi et al., 2022a).

Note that the binary $\mathcal{H}$-consistency bound (Awasthi et al., 2022a) in Theorem 2.1 does not directly apply to multiclass cases. We generalize the binary framework (Awasthi et al., 2022a) to multiclass cases and prove an explicit $\mathcal{H}$-consistency bound for logistic loss. We present the bound in Theorem 3.3 and defer the establishment to Section 3.2.

**Theorem 3.3** ($\mathcal{H}$-consistency bound for multiclass logistic loss and zero-one loss, proof in Appendix E.4)**.** *If $R_{\ell_{log}}(\boldsymbol{h}) - R^*_{\ell_{log},\mathcal{H}_{lin}} + M_{\ell_{log},\mathcal{H}_{lin}} \leq \frac{1}{2}(\frac{e^{2B}-1}{e^{2B}+K-1})^2$, then for any distribution satisfiying $\max_y p_y(\boldsymbol{x}) - \min_y p_y(\boldsymbol{x}) \leq \frac{e^{2B}-1}{e^{2B}+K-1}$ for all $\boldsymbol{x}$, it holds that $R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}_{lin}} + M_{\ell_{0-1},\mathcal{H}_{lin}} \leq \sqrt{2}(R_{\ell_{log}}(\boldsymbol{h}) - R^*_{\ell_{log},\mathcal{H}_{lin}} + M_{\ell_{log},\mathcal{H}_{lin}})^{\frac{1}{2}}$.*

Note that $R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty}) = R^*_{\ell_{log},\mathcal{H}_{lin}}$ by the definition. Besides, when $B \to +\infty$, we have $\frac{e^{2B}-1}{e^{2B}+K-1} \to 1$, and Theorem 3.3 holds for all distribution. Theorem 3.3 provides a tool to analyze the asymptotic behavior of multiclass logistic regression considering the surrogate loss. According to it, we need to bound the gap $R_{\ell_{log}}(\boldsymbol{h}_{Dis,m}) - R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty})$ and $M_{\ell_{log},\mathcal{H}_{lin}}$ to guarantee a small $R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty})$. The following Proposition characterizes $R_{\ell_{log}}(\boldsymbol{h}_{Dis,m}) - R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty})$ by Radmancher complexity (Bartlett et al., 2002; Mohri et al., 2018) and a contraction lemma (Maurer, 2016).

**Proposition 3.2** (Proof in appendix D.4)**.** *For any fixed $\delta_0 \in (0,1)$, with probability at least $1-\delta_0$, the following holds:*

$$R_{\ell_{log}}(\boldsymbol{h}_{Dis,m}) \leq R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty}) + O(\sqrt{\frac{K^3 n}{m}}).$$

$M_{\ell,\mathcal{H}}$ is a constant determined by the hypothesis set $\mathcal{H}$, loss function $\ell$, and data distribution $\mathcal{D}$. Its value is difficult to estimate directly (Awasthi et al., 2022a). However, according to the definition, $M_{\ell,\mathcal{H}}$ can be bounded by the corresponding approximate error. Prior works (Saunshi et al., 2019; Lee et al., 2021; Tosh et al., 2021; HaoChen et al., 2021) prove the (approximate) linear separability of the representations extracted by deep pre-trained models, suggesting a

[2] Indeed, if the naïve Bayes assumption really holds, we can obtain a stronger guarantee for $\widetilde{G}(\tau)$ by using Chernoff's bound. We put the result in Proposition C.2.

small approximation error for the logistic loss. Therefore, we make the following assumption, which is validatable in the context of linear evaluation of deep models.

**Assumption 3.5.** *The approximate error of the logistic loss is bounded by a small constant $\nu < \frac{1}{2}(\frac{e^{2B}-1}{e^{2B}+K-1})^2$. Namely, $\operatorname{argmin}_{\boldsymbol{h}\in\mathcal{H}_{lin}} R_{\ell_{log}}(\boldsymbol{h}) - \operatorname{argmin}_{\boldsymbol{h}\in\mathcal{H}_{all}} R_{\ell_{log}}(\boldsymbol{h}) \leq \nu$, which implies that $M_{\ell_{log},\mathcal{H}_{lin}} \leq \nu$.*

We characterize the number of samples required to approach the asymptotic error for logistic regression in Theorem 3.4 by combining Proposition 3.2 and Theorem 3.3.

**Theorem 3.4** (Results for multiclass logistic regression, proof in appendix D.5)**.** *Suppose that Assumption 3.5 holds. Then, it suffices to pick $m = O(n)$ training samples such that $R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,m}) \leq R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty}) + \epsilon_0$ hold with probability $1-\delta_0$, for any fixed $\epsilon_0 \in [\sqrt{2\nu}, \frac{e^{2B}-1}{e^{2B}+K-1}]$ and $\delta_0 \in (0,1)$.*

Notably, according to the multiclass fundamental theorem (Theorem 29.3 of Shalev-Shwartz & Ben-David (2014)), the sample complexity of $\mathcal{H}_{lin}$ for any algorithm is $\Omega(n)$ because the Natarajan dimension for $\mathcal{H}_{lin}$ is $\Omega(Kn)$, indicating the upper bound in Thereom 3.2 is tight with respect to the dimension $n$.

Theorem 3.2 and Theorem 3.4 show that the $O(n)$ vs. $O(\log(n))$ result (Ng & Jordan, 2001) still holds in multiclass cases, which suggests that naïve Bayes is possibly better than logistic regression when the sample size is limited. We validate our theory on a mixture of Gaussian distribution, as presented in Figuire 1. For a fixed feature dimension $n$, we increase the number of samples $m$ until the two models approach the corresponding asymptotic error, which is tractable in the experiment. Detailed configurations of the experiments and additional results are presented in Appendix H.

### 3.2. Multiclass $\mathcal{H}$-consistency Framework

We now present the general multiclass $\mathcal{H}$-consistency bound framework and prove the explicit bound for the logistic loss in Theorem 3.3, which are of independent interest. Similarly to the binary case (Awasthi et al., 2022a), we first introduce the following general multiclass $\mathcal{H}$-consistency bound between any target loss $\ell_2$ and surrogate loss $\ell_1$.

**Proposition 3.3** (Distribution-dependent convex bound, proof in Appendix E.1)**.** *For a fixed distribution, if there exists a convex function $g : \mathbb{R}_+ \to \mathbb{R}$ with $g(0) \geq 0$ and $\epsilon \geq 0$, and the following holds for any $\boldsymbol{h} \in \mathcal{H}$ and $\boldsymbol{x} \in \mathcal{X}$:*

$$g(\langle \Delta\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})\rangle_\epsilon) \leq \Delta\mathcal{C}_{\ell_1,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}). \tag{5}$$

*Then it holds for all $\boldsymbol{h} \in \mathcal{H}$ that*

$$\begin{aligned} &g(R_{\ell_2}(\boldsymbol{h}) - R^*_{\ell_2,\mathcal{H}} + M_{\ell_2,\mathcal{H}}) \\ &\leq R_{\ell_1}(\boldsymbol{h}) - R^*_{\ell_1,\mathcal{H}} + M_{\ell_1,\mathcal{H}} + \max(g(0), g(\epsilon)). \end{aligned} \tag{6}$$

We present the concave counterpart of it as Proposition C.1 of Appendix C. For simplicity, we fix the target loss $\ell_2$ as the zero-one loss in the following. Note that Proposition 3.3 is distribution-dependent while an asymptotically distribution-independent version is necessary for our analysis in Section 3.1. To this end, we introduce a tool called *multiclass $\mathcal{H}$-estimation error transformation*.

**Definition 3.3** (Multiclass $\mathcal{H}$-estimation error transformation)**.** *The multiclass $\mathcal{H}$-estimation error transformation of a surrogate loss $\ell$ is defined on $t \in [0,1]$ as $\mathcal{J}_\ell(t) = \inf_{\hat{y}\in\mathcal{Y},\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t),\boldsymbol{x}\in\mathcal{X},\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})$. Here $\mathcal{H}_{\hat{y}}(\boldsymbol{x}) \coloneqq \{\boldsymbol{h} \in \mathcal{H} : \operatorname{argmax}_{y\in\mathcal{Y}} h_y(\boldsymbol{x}) = \hat{y}\}$ is a collection of hypotheses that predicts $\boldsymbol{x}$ as class $\hat{y}$. $\mathcal{P}_{\hat{y}}(t) \coloneqq \{\boldsymbol{p} \in \Delta_K : \max_y p_y - p_{\hat{y}} = t\}$ is a subset of $K$-dimensional simplex indexed by classes and the gap between the max component and class-indexed component of $\boldsymbol{p}$.*

$\mathcal{J}_\ell(t)$ in Defition 3.3 is carefully derived such that plugging it to the right-hand side of Eq. (5) provides a sufficient condition such that Eq. (6) holds for any $\boldsymbol{h}, \boldsymbol{x}$, and $\boldsymbol{p}$ (i.e., distribution-independent). It is worth noting that the condition is actually necessary as well under further assumptions, as presented later in Theorem 3.6. Defition 3.3 generalizes the binary freamwork (Awasthi et al., 2022a) by optimizing $\boldsymbol{p}$ in a collection of subsets $\mathcal{P}_{\hat{y}}(t)$ to handle multiclass cases. Built upon Defition 3.3, we establish the multiclass distribution-independent bound for zero-one loss as follows.

**Theorem 3.5** (Distribution-independent convex $\ell_{0-1}$ bound, proof in Appendix E.2)**.** *Suppose that $\mathcal{H}$ satisfies that $\{\operatorname{argmax}_{y\in\mathcal{Y}} h_y(\boldsymbol{x}) : \boldsymbol{h} \in \mathcal{H}\} = \{1,\ldots,K\}$ for any $\boldsymbol{x} \in \mathcal{X}$. If there exists a convex function $g : \mathbb{R}_+ \to \mathbb{R}$ with $g(0) = 0$ and $g(t) \leq \mathcal{J}_\ell(t)$. Then it holds for any $\boldsymbol{h} \in \mathcal{H}$ and any distribution $\mathcal{D}$ that*

$$g(R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}}) \leq R_\ell(\boldsymbol{h}) - R^*_{\ell,\mathcal{H}} + M_{\ell,\mathcal{H}}.$$

We present the concave counterpart of it as Theorem C.1 in Appendix C. This theorem holds for any hypothesis set $\mathcal{H}$ that can divide any sample $\boldsymbol{x}$ into any category, including the linear hypothesis set and hypotheses of neural network. Notably, our multiclass $\mathcal{H}$-consistency result degenerates to the binary one exactly (Awasthi et al., 2022a) with $K = 2$. In addition, we note that if $\mathcal{J}_\ell(t)$ is convex and $\mathcal{J}_\ell(0) = 0$, then $\mathcal{J}_\ell$ satisfies the condition of $g$ in Theorem 3.5. In fact, it leads to the tightest multiclass $\mathcal{H}$-consistency bound.

**Theorem 3.6** (Tightness, proof in Appendix E.3)**.** *If $\mathcal{J}_\ell(t)$ is convex with $\mathcal{J}_\ell(0) = 0$, then for any $t \in [0,1]$ and $\delta > 0$, there exist a distribution $\mathcal{D}$ and a hypothesis $h \in \mathcal{H}$ such that $R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}} = t$ and $\mathcal{J}_\ell(t) \leq R_\ell(\boldsymbol{h}) - R^*_{\ell,\mathcal{H}} + M_{\ell,\mathcal{H}} \leq \mathcal{J}_\ell(t) + \delta$.*

To establish our main result in Section 3.1, we have presented an asymptotically distribution-independent multiclass $\mathcal{H}$-consistency bound for the logistic loss in an explicit

*Table 1.* Analysis of assumptions on CIFAR10 training dataset.

| Method | Backbone | Pre-training data | $\rho_0$ | $\beta$ | $\alpha$ |
|---|---|---|---|---|---|
| ViT (Dosovitskiy et al., 2021) | ViT-B/16 | Image-label | 2.80E-3 | 0.004 | 690 |
| ResNet (He et al., 2016) | ResNet50 | Image-label | 1.70E-3 | 0.06 | 11516 |
| CLIP (Radford et al., 2021) | ResNet50 | Image-text | 4.78E-3 | 0.203 | 6383 |
| MoCov2 (Chen et al., 2020d) | ResNet50 | Image | 5.03E-5 | 0.005 | 26640 |
| SimCLRv2 (Chen et al., 2020c) | ResNet50 | Image | 3.74E-5 | 0.01 | 2490 |
| MAE (He et al., 2022) | ViT-B/16 | Image | 6.37E-3 | 0.032 | 6919 |
| SimMIM (Xie et al., 2022) | ViT-B/16 | Image | 7.86E-3 | 0.002 | 5201 |

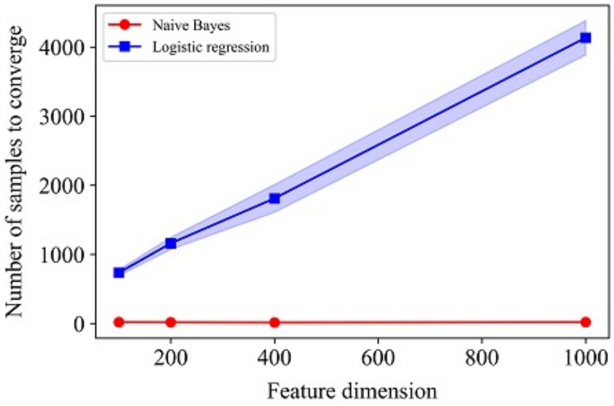


*Figure 1.* Multiclass ($K = 5$) simulation results. Empirically, logistic regression and naïve Bayes require $O(n)$ and $O(\log n)$ samples to approach the corresponding asymptotic error respectively. Error bars show the variance estimated by 5 runs.

form in Theorem 3.3. We mention that the proof of Theorem 3.3 is nontrivial because $\mathcal{J}_\ell(t)$ in the multiclass case involves a much more complex optimization problem than that in the binary case (Awasthi et al., 2022a).

The proposed framework is not limited to the linear hypothesis class and the logistic loss. In particular, we present a similar result for the hypothesis class of one-hidden-layer neural networks in Theorem C.2 of Appendix C. Besides, the general bound in Theorem 3.5 and the proof idea of Theorem 3.3 are applicable to hinge loss, exponential loss, $\rho$-margin loss, and so on, which are left for future work. Furthermore, the analysis idea can be used to obtain multiclass Bayes consistency bounds by setting $\mathcal{H}$ to $\mathcal{H}_{all}$.

## 4. Implications in Deep Learning

In this section, we discuss the implications of our theoretical results in the linear evaluation of pre-trained deep neural networks. First, as presented in Section 4.1, we empirically analyze the main assumptions of our theory in various deep vision models (Dosovitskiy et al., 2021; He et al., 2016; Radford et al., 2021; Chen et al., 2020d;c; He et al., 2022; Xie et al., 2022). Second, we systematically compare logistic regression and naïve Bayes on the CIFAR10 and CIFAR100 datasets (Krizhevsky et al., 2009) with various models and sample sizes in Section 4.2. Naïve Bayes always converges much faster, which agrees with our theory. The “two regimes” phenomenon (Ng & Jordan, 2001) almost happens with models pre-trained in a supervised manner (Dosovitskiy et al., 2021; He et al., 2016), which is analyzed in detail in Section 4.3. Details of experiments can be found in Appendix I.

### 4.1. Analyzing the Assumptions

We empirically analyze and discuss the main assumptions made in Section 3 on the CIFAR10 dataset. The results are summarized in Table 1. We emphasize that the concrete values of the quantities in the table won’t affect the asymptotic analyses in Section 3, i.e., $O(\log n)$ results for naïve Bayes, but may affect its performance given a fixed data size.

We consider linear evaluation for transfer learning on top of pre-trained models, whose parameters are frozen. Therefore, it is valid to assume that the features extracted on the target dataset satisfy the $i.i.d.$ assumption.

#### 4.1.1. Assumption 3.1 and 3.2

Assumption 3.1 holds naturally because the CIFAR10 dataset is class-balanced. For Assumption 3.2, we calculate the $\hat{\sigma_i}^2$ for each dimension of the training representations as approximations for $\sigma_i^2$. We present $\rho_0 = \min(\min_i \hat{\sigma_i}^2, \frac{1}{10})$ in Table 1, and Figure 5 in Appendix I.3 plots the histogram of $\hat{\sigma_i}^2$. Assumption 3.2 holds for all models.

#### 4.1.2. Assumption 3.3 and 3.4

It is hard to directly validate the two assumptions in practice. Nevertheless, we estimate $\beta_{k_1,k_2,k}$ and $\alpha_{k_1,k_2,k}$ for all $k_1, k_2 (k_1 \neq k_2)$ and $k \in \mathcal{Y}$ in different models for a comparison. We note that $\beta_{k_1,k_2,k} = \zeta_{k_1,k_2,k}$ in our experiments, because the CIFAR10 dataset is class-balanced. We report the estimated $\beta = \zeta = \min_{k1,k_2,k} |\beta_{k_1,k_2,k}|$ and

*Table 2.* Convergence comparison between multiclass logistic regression and naïve Bayes. "NB faster" means naïve Bayes approaches its asymptotic error faster.

| Method | Visual results | NB faster/ Two regimes | |
|---|---|---|---|
| | | CIFAR10 | CIFAR100 |
| ViT | Figure 8 | √ / √ | √ / √ |
| ResNet | Figure 9 | √ / √ | √ / √ |
| CLIP | Figure 10 | √ / √ | √ / √ |
| MoCov2 | Figure 11 | √ / × | √ / × |
| SimCLRv2 | Figure 12 | √ / × | √ / √ |
| MAE | Figure 13 | √ / √ | √ / × |
| SimMIM | Figure 14 | √ / × | √ / × |

$\alpha = \max_{k1,k_2,k} \alpha_{k_1,k_2,k}$ in Table 1. We also present the histograms of $|\beta_{k_1,k_2,k}|$ and $\alpha_{k_1,k_2,k}$ in Figure 6 and Figure 7 of Appendix I.3, respectively.

#### 4.1.3. Assumption 3.5

Assumption 3.5 is hard to validate in practice because the Bayes-optimal classifier is unknown. However, recent theoretical results in prior works (Saunshi et al., 2019; Lee et al., 2021; Tosh et al., 2021; HaoChen et al., 2021) suggest that it holds when the number of samples for pre-training is sufficiently large.

### 4.2. Empirical Results in Deep Learning

We systematically compare logistic regression and naïve Bayes on the CIFAR10 and CIFAR100 datasets in various models, which are trained on image-label pairs (Dosovitskiy et al., 2021; He et al., 2016), image-text pairs (Radford et al., 2021), or pure images (Chen et al., 2020d;c; He et al., 2022; Xie et al., 2022).

For a fair comparison, we keep the linear evaluation setting in (Radford et al., 2021) throughout the experiments. Specially, we train the logistic regression using scikit-learn's (Pedregosa et al., 2011) L-BFGS implementation, with a maximum of 1000 iterations. We adjust the weight of $\ell_2$ regularization of logistic regression carefully to reproduce the results reported in (Radford et al., 2021) on both datasets with full training data. We then adjust the number of training samples $m$ gradually. For each $m$, we obtain training samples randomly 5 times and record the mean test error of two models.

We plot the convergence curves in all settings in Appendix I.4, which are linked in Table 2. Notably, naïve Bayes approaches its asymptotic error much faster than logistic regression in all settings, like that presented in Figure 2, which is consistent with our theoretical results.

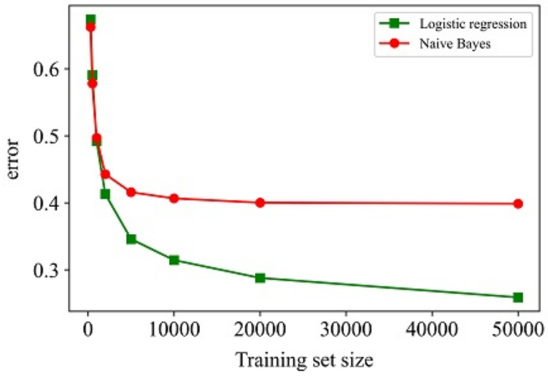


*Figure 2.* Comparison between naïve Bayes and logistic regression with the features extracted by ResNet on the CIFAR100 dataset. Naïve Bayes approaches its asymptotic error much faster.

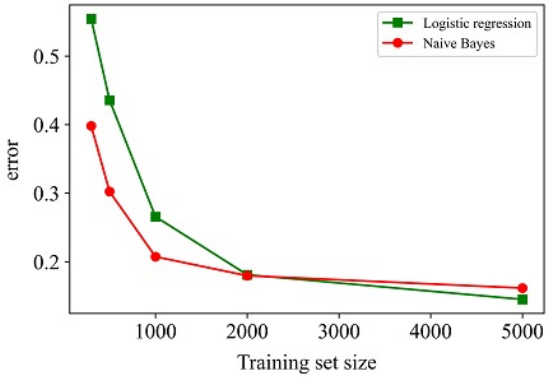


*Figure 3.* Comparison between naïve Bayes and logistic regression with the features extracted by ViT on the CIFAR100 dataset. The "two regimes" phenomenon is observed.

### 4.3. On the "Two Regimes" Phenomenon

Ng & Jordan (2001) suggests that there can often be two regimes of performance between naïve Bayes and logistic regression, that is, though logistic regression enjoys lower asymptotic error, naïve Bayes performs better with smaller training sets because of its fast convergence rate. They observed this phenomenon on many datasets from the UCI Machine Learning repository (Dua & Graff, 2017). These classical datasets are small and the features are mostly low-dimensional. However, nowadays, people prefer to obtain representations by using deep neural networks pre-trained by massive data. The occurrence of the "two regimes" phenomenon in this new setting has not been investigated yet.

We summarize the occurrence of the "two regimes" phenomenon in Table 2. The "two regimes" phenomenon occurs in half of our experiments, which suggests that naïve Bayes still shows promise when the training data is limited. We present a typical case in Figure 3 and see Appendix I.4

for complete results. Interestingly, the "two regimes" phenomenon almost happens when the deep vision model is pre-trained in a supervised manner (ViT, ResNet, and CLIP), which suggests a distinction between representations learned by supervised learning and self-supervised learning.

We conjecture that representations learned by supervised methods could have some better properties to make naïve Bayes converges faster than that learned by self-supervised methods. As validated in Section 4.2, though our theory could only prove the fast convergence rate of naïve Bayes, it does help us to understand this distinction to some extent. Combining the values presented in Table 1, we can get some preliminary results.

*Representations learned by supervised methods could be more robust for each dimension.* As shown in Table 1, features learned by supervised methods (ViT, ResNet, CLIP) tend to have larger $\rho_0$. In other words, these representations tend to have larger in-class variance $\sigma_i^2$ than others. Intuitively, it suggests that data in each dimension could be more robust to relieve the over-fitting and boost naïve Bayes learning better in the few-shot case. Besides, according to Eq. (7-8) in Appendix D.1 and the derivation in Appendix D.3, a larger $\rho_0$ implies faster convergence in a $1/\rho_0^2$ order, which explains it in a certain sense.

*Representations learned by supervised methods could be more separable between different categories.* From Table 1, representations learned by supervised methods (ResNet, CLIP) are inclined to have larger $\beta$ than others. Namely, there exists more distinction between the distributions of samples in different classes, which are easier to predict. In addition, by our derivation in Appendix D.3, a larger $\beta$ implies faster convergence in a $1/\beta^2$ order, which agrees with our observation.

# 5. Related Work

## 5.1. Deep Representative Learning

Deep representation learning aims to learn representations on the raw unlabeled data and transfer them to the downstream tasks. It has made remarkable progress in various machine learning fields (Ren et al., 2015; He et al., 2017; Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Chen & He, 2021; Grill et al., 2020; He et al., 2022; Xie et al., 2022; Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020). In particular, the promise of *linear evaluation* (Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Grill et al., 2020; Radford et al., 2021) suggests that representations extracted by pre-trained models are near to linear separable. Besides, the performance of such representations in linear evaluation is guaranteed in recent theoretical works (Saunshi et al., 2019; Lee et al., 2021; Tosh et al., 2021; HaoChen et al., 2021). All of these empirical and theoretical works encourage us to rethink the role of linear classifiers.

## 5.2. Discriminative vs. Generative Learning

Comparing discriminative with generative classifiers has long been an interesting topic (Efron, 1975; Rubinstein & Hastie, 1997; Ng & Jordan, 2001). Efron (1975) compared the logistic regression and normal discriminant analysis and claimed that the latter is only slightly more efficient. Ng & Jordan (2001) simplified the normal discriminant analysis to naïve Bayes and concluded that the discriminative model has lower asymptotic error while the generative classifier may approach its higher asymptotic error much faster. Ng & Jordan (2001) assume that one can directly optimize on zero-one loss. Instead, we weaken the assumption and introduce the theoretical tools from $\mathcal{H}$-consistency to obtain more reliable results.

## 5.3. $\mathcal{H}$-consistency

Most machine learning algorithms depend on optimizing a surrogate loss function rather than the target loss function. To find the favorable property of surrogate loss, consistency has been studied broadly in the last two decades. Classical Bayes consistency (Zhang, 2004a;b; Bartlett et al., 2006; Tewari & Bartlett, 2007) analyzes the relationship between the excess error of zero-one loss and that of a surrogate loss. Instead, $\mathcal{H}$-consistency (Long & Servedio, 2013) considers the estimation error w.r.t. a hypothesis set $\mathcal{H}$. It includes the classical Bayes consistency as a special case by setting $\mathcal{H}$ to $\mathcal{H}_{all}$. Most recently, Awasthi et al. (2022a) proposed a novel and solid framework named $\mathcal{H}$-consistency bounds, which consider the upper bounds on the target estimation error expressed by surrogate estimation error.

We proposed a novel multiclass $\mathcal{H}$-consistency framework, which includes the framework in (Awasthi et al., 2022a) as a special case. We notice that the independent work of (Awasthi et al., 2022b) also proposed a multiclass $\mathcal{H}$-consistency framework from the same general theorem (Proposition 3.3). We highlight the following comparison that distinguishes our work. First, the proof ideas are totally different. In particular, we directly generalize the binary framework in (Awasthi et al., 2022a) to the multiclass case in Theorem 3.5, which is general and tight (Theorem 3.6). In contrast, Awasthi et al. (2022b) argues that generalizing the binary framework is nontrivial and instead provides a case-by-case analysis for different losses, which does not enjoy the tightness guarantee. Second, we provide an explicit bound for logistic loss (Theorem 3.3), which is necessary for our subsequent analysis, while it is unclear how to derive such a bound by the prior work (Awasthi et al., 2022b).

**Concurrent work.** The concurrent and independent work of Mao et al. (2023) also obtains $\mathcal{H}$-consistency bounds of the multiclass logistic loss under a little stronger assumption.

The multiclass $\mathcal{H}$-estimation error transformation $\mathcal{J}_\ell(t)$ derived by them (Theorem 1 of Mao et al. (2023)) is actually the same as ours in Theorem 3.3, and their bounds also enjoy the tightness guarantee. However, they assume that the hypothesis set $\mathcal{H}$ is complete, that is, $\{h_y(\boldsymbol{x}) : \boldsymbol{h} \in \mathcal{H}\} = \mathbb{R}$ for any $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$, which does not hold for bounded linear hypotheses ($W, B < +\infty$) considered by this paper.

## 6. Conclusion

We revisit the classical topic of discriminative vs. generative classifiers (Ng & Jordan, 2001). Specially, we weaken the assumption in the previous work and extend the analysis to multiclass cases. As result, under some assumptions, we prove that multiclass naïve Bayes requires $O(\log n)$ samples to approach its asymptotic error while the logistic regression needs $O(n)$ samples. Technically, we proposed a multiclass $\mathcal{H}$-consistency framework, which is of independent interest. Experiments with various pre-trained deep vision models verify our theory and show the potential of the generative linear head in the few-shot cases. Finally, our experiments suggest differences between representations learned by supervised and self-supervised methods.

**Social Impact:** This is mainly theoretical work and we do not see a direct social impact of our theory. The experiments on Naïve Bayes may benefit applications with a few training data such as medical analysis.

## Acknowledgements

This work was supported by NSF of China (NO. 62076145, 62206159); Beijing Outstanding Young Scientist Program (NO. BJJWZYJH012019100020098); Shandong Provincial Natural Science Foundation (NO. ZR2022QF117); Major Innovation & Planning Interdisciplinary Platform for the “Double-First Class" Initiative, Renmin University of China; the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (22XNKJ13); the Fundamental Research Funds of Shandong University. C. Li was also sponsored by Beijing Nova Program.

## References

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. $\mathcal{H}$-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, volume 162, pp. 1117–1174, 2022a.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. Multi-class $\mathcal{H}$-consistency bounds. In *Advances in Neural Information Processing Systems*, 2022b.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Localized rademacher complexities. In Kivinen, J. and Sloan, R. H. (eds.), *Computational Learning Theory*, volume 2375, pp. 44–58, 2002.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020a.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, volume 119, pp. 1597–1607, 2020b.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, 2020c.

Chen, X. and He, K. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Chen, X., Fan, H., Girshick, R. B., and He, K. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020d.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Efron, B. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, pp. 5000–5011, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. B. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15979–15988, 2022.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, Canadian Institute for Advanced Research, Toronto, ON, Canada, 2009.

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 309–323, 2021.

Long, P. and Servedio, R. Consistency versus realizable h-consistency for multiclass classification. In *International Conference on Machine Learning*, pp. 801–809, 2013.

Mao, A., Mohri, M., and Zhong, Y. Cross-entropy loss functions: Theoretical analysis and applications. *CoRR*, abs/2304.07288, 2023.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17, 2016.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Ng, A. Y. and Jordan, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, pp. 841–848, 2001.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing System*, pp. 8024–8035, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Rubinstein, Y. D. and Hastie, T. Discriminative vs informative learning. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pp. 49–53, 1997.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, volume 97, pp. 5628–5637, 2019.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Tewari, A. and Bartlett, P. L. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(5), 2007.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, volume 132, pp. 1179–1206, 2021.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: a simple framework for masked image modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9643–9653, 2022.

Xue, J. and Titterington, D. M. Comment on "on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". *Neural Process. Lett.*, 28(3): 169–187, 2008.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004a.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004b.

## Contents of Appendix

# A. Detailed Notations and Definitions

Let lower, boldface lower and capital case letters denote scalers (e.g., a), vectors (e.g., $\boldsymbol{a}$), and matrices (e.g., $\boldsymbol{A}$) respectively. For a matrix $\boldsymbol{A}$, $\boldsymbol{A}_i$ and $A_{ij}$ denote its $i$-th row and $(i,j)$-th element. For a vector $\boldsymbol{a}$, $a_i$ denotes its $i$-th element. Similarly, for a vector function $\boldsymbol{f}$, $f_i(\boldsymbol{x})$ denotes the $i$-th element of $\boldsymbol{f}(\boldsymbol{x})$. Let $\mathcal{X}$ denote the domain set and $\mathcal{Y}$ denote the label set. For simplicity, we assume $\mathcal{X} = \{0,1\}^n$ when inputs are discrete and $\mathcal{X} = [0,1]^n$ otherwise, where $n$ is the feature dimension. Let $\mathcal{Y} = \{0,1\}$ be the binary label space and $\mathcal{Y} = \{1,\dots,K\}$ be the multiclass label space, where $K$ is the number of classes. $\mathcal{D}$ denotes the distribution on $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{P}$ denotes set of distribution. We denote the KL Divergence between two distributions $p$ and $q$ by $D(p\|q)$. We use $\mathbb{E}$ and $\mathbb{V}$ to represent expectation and variance, respectively.

For the binary case, let $\mathcal{H}$ be a hypothesis set of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. The prediction associated by a hypothesis $h \in \mathcal{H}$ and $\boldsymbol{x} \in \mathcal{X}$ is $\operatorname{sign}(h(x))$. In this paper, we mainly focus on the family of constrained binary linear hypotheses $\mathcal{H}_{lin} = \{x \to \boldsymbol{w}^T x + b : \|\boldsymbol{w}\|_2 \le W, |b| \le B\}$, where $W, B \in \mathbb{R}^+$. The generalization error and minimal generalization error of a hypothesis $h$ w.r.t. the loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ are defined as $R_\ell(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(h(\boldsymbol{x}),y))]$ and $R^*_{\ell,\mathcal{H}} = \inf_{h\in\mathcal{H}} R_\ell(h)$, where $\mathcal{H}$ is a hypothesis set and $\mathcal{D}$ is data distribution. We denote the empirical generalization error by $\hat{R}_\ell(h)$. Furthermore, given a family of functions $\mathcal{G}$ mapping from $\mathcal{Z}$ to $\mathbb{R}$, the empirical Rademacher complexity of $\mathcal{G}$ for a sample $S = (z_1,\dots,z_m)$ is defined by $\hat{\mathcal{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma[\frac{1}{m}\sup_{g\in\mathcal{G}}\sum_{i=1}^m \sigma_i g(z_i)]$, where $\sigma = (\sigma_1,\dots,\sigma_m)$ is a vector of $i.i.d.$ independent uniform random variables taking values in $\{-1,+1\}$. The Rademacher complexity of $\mathcal{G}$ is defined as $\mathcal{R}_m(\mathcal{G}) = \mathbb{E}_S[\hat{\mathcal{R}}_S(\mathcal{G})]$.

Notations listed in the following will be useful to analyze the $\mathcal{H}$-consistency bounds. For binary label space, let $\eta(x)$ denote the conditional distribution $\mathbb{P}(Y=1|X=x)$ and $\Delta\eta(x)$ the $\eta(x) - \frac{1}{2}$. We rewrite the generalization error as $R_\ell(h) = \mathbb{E}_{\boldsymbol{x}}[\mathcal{C}_\ell(h,x)]$, where $\mathcal{C}_\ell(h,x) = \eta(x)\ell(h,(x,1)) + (1-\eta(x))\ell(h,(x,0))$ is called as conditional risk. We can also define the minimal conditional risk as $\mathcal{C}^*_{\ell,\mathcal{H}}(x) = \inf_{h\in\mathcal{H}} \mathcal{C}_\ell(h,x)$. We use the shorthand for the gap $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x) = \mathcal{C}_\ell(h,x) - \mathcal{C}^*_{\ell,\mathcal{H}}(x)$ and conditional $\epsilon$-regret of $\ell$ $\langle\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)\rangle_\epsilon = \Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)\mathbb{1}_{\mathcal{C}_{\ell,\mathcal{H}}(h,x)>\epsilon}$. For any $t \in [0,1]$, we also define $\mathcal{C}_\ell(h,x,t) = t\ell(h,(x,1)) + (1-t)\ell(h,(x,0))$ and $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x,t) = \mathcal{C}_\ell(h,x,t) - \inf_{h\in\mathcal{H}} \mathcal{C}_\ell(h,x,t)$. It is worthwhile to note that a key quantity appears in the article is the $M_{\ell,\mathcal{H}} = R^*_{\ell,\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}(\mathcal{C}^*_{\ell,\mathcal{H}}(x))$, which is hard to estimate.

For the multiclass case, let $\mathcal{H}$ be a hypothesis set of functions mapping from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}^K$. The prediction associated by a hypothesis $\boldsymbol{h} \in \mathcal{H}$ and $\boldsymbol{x} \in \mathcal{X}$ is $\operatorname{argmax}_{y\in\mathcal{Y}} h_y(\boldsymbol{x})$. In the main paper, we mainly focus on the family of constrained linear hypotheses $\mathcal{H}_{lin} = \{x \to \boldsymbol{h}(x) : h_y(x) = \boldsymbol{w}_y^T x + b_y, \|\boldsymbol{w}_y\|_2 \le W, |b_y| \le B, y \in \mathcal{Y}\}$, where $W, B \in \mathbb{R}^+$. We also give $\mathcal{H}$-consistency bound for family of one-hidden-layer neural network hypotheses with ReLU activation function $(\cdot)_+$ $\mathcal{H}_{NN} = \{\boldsymbol{x} \to \boldsymbol{h}(\boldsymbol{x}) : h_y(\boldsymbol{x}) = \sum_{j=1}^n U_{yj}(\langle\boldsymbol{w}_j,\boldsymbol{x}\rangle + b)_+\}$, where $\boldsymbol{U} \in \mathbb{R}^{K\times n}$, $\boldsymbol{w}_j \in \mathbb{R}^n$ and $b \in \mathbb{R}$. The generalization error and minimal generalization error of a hypothesis $\boldsymbol{h}$ w.r.t. the loss function $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ are defined as $R_\ell(\boldsymbol{h}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(\boldsymbol{h}(\boldsymbol{x}),y)]$ and $R^*_{\ell,\mathcal{H}} = \inf_{\boldsymbol{h}\in\mathcal{H}} R_\ell(\boldsymbol{h})$, where $\mathcal{H}$ is a hypothesis set and $\mathcal{D}$ is data distribution. We denote by $\boldsymbol{p}(\boldsymbol{x})$ the conditional distribution of $y$ when given $\boldsymbol{x}$, i.e., $p_y(\boldsymbol{x}) = \mathbb{P}(Y=y|X=\boldsymbol{x})$. Similarly to the binary classification, we have $\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x}) = \sum_{y=1}^K p_y(\boldsymbol{x})\ell(\boldsymbol{h}(\boldsymbol{x}),y)$, $\mathcal{C}^*_{\ell,\mathcal{H}}(\boldsymbol{x}) = \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x})$, $\Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}) = \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x}) - \mathcal{C}^*_{\ell,\mathcal{H}}(\boldsymbol{x})$ and $M_{\ell,\mathcal{H}} = R^*_{\ell,\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}(\mathcal{C}^*_{\ell,\mathcal{H}}(\boldsymbol{x}))$. Furthermore, for any $\boldsymbol{p}$ in probability simplex $\Delta_K$, we can define $\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) = \sum_{y=1}^K p_y\ell(\boldsymbol{h}(\boldsymbol{x}),y)$ and $\Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) = \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})$.

# B. On Binary Discriminative vs. Generative Linear Classifiers

In this section, we focus on the binary case and obtain results that are similar to (Ng & Jordan, 2001), under weaker assumptions. Let $h_{Gen,m}$ and $h_{Dis,m}$ be logistic regression and naïve Bayes trained with $m$ $i.i.d$ samples, $h_{Gen,\infty}$ and $h_{Dis,\infty}$ be their asymptotic/population versions. Proofs of this section can be found in Appendix F.

We will compare the sample complexity of logistic regression with that of naïve Bayes. Consider optimizing the practicable logistic loss rather than zero-one loss, the estimation error of the logistic regression can be bounded by making use of the definition of Rademacher complexity from classical statistical learning techniques.

**Proposition B.1** (Proof in Appendix F.1)**.** *With a high probability of at least* $1-\delta_0$*, the following holds*

$$R_{\ell_{log}}(h_{Dis,m}) \le R_{\ell_{log}}(h_{Dis,\infty}) + O(\sqrt{\frac{n}{m}}).$$

Theorem 2.1 means that we can bound the estimation error of the zero-one loss by the estimation error of the logistic loss, which makes it possible to obtain an upper bound of the sample complexity with respect to zero-one loss.

**Theorem B.1** (Proof in Appendix F.2)**.** *Suppose that Assumption 3.5 is valid. Then, it suffices to pick $m = O(n)$ training samples such that $R_{\ell_{0-1}}(h_{Dis,m}) \leq R_{\ell_{0-1}}(h_{Dis,\infty}) + \epsilon_0$ hold with probability $1 - \delta_0$, for any $\epsilon_0 \in [\sqrt{2\nu}, \frac{e^B - 1}{e^B + 1}]$ and $\delta_0 \in (0, 1)$.*

By further using the Theorem 9.3 in (Shalev-Shwartz & Ben-David, 2014) and binary $\mathcal{H}$-consistency bound Theorem 2.1, which states that for $n$-dimension logistic regression, it needs at least $\Omega(n)$ training samples to guarantee the estimation error is small enough with high probability, we know the result in Theorem B.1 is tight.

In the rest of this subsection, we will discuss the sample complexity of naïve Bayes. The sketch of proofs has been adopted by (Ng & Jordan, 2001). However, their results are somewhat ambiguous and without detailed derivation, which is very important to the extended analysis in Section 3.1 for multiclass classification. Thus, we present the proof for completeness.

**Definition B.1.** *We define the $G(\tau)$ which will be useful to bound the generalization error of binary naïve Bayes as*

$$G(\tau) = \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}}(|\Delta a_{Gen,\infty}(\boldsymbol{x}, 1, 0)| \leq \tau n).$$

**Theorem B.2** (Proof in Appendix F.3)**.** *Suppose that Assumption 3.1 and 3.2 hold. Then with probability at least $1 - \delta$:*

$$R_{\ell_{0-1}}(h_{Gen,m}) \leq R_{\ell_{0-1}}(h_{Gen,\infty}) + G\big(O(\sqrt{\frac{1}{m}\log(\frac{n}{\delta})})\big) + \delta.$$

The key quantity in this Theorem is the $G(\tau)$ , which must be small when $\tau$ is small in order to bound $R_{\ell_{0-1}}(h_{Gen,m}) - R_{\ell_{0-1}}(h_{Gen,\infty})$. This property holds when we introduce the Assumption B.1 and B.2.

**Assumption B.1.** *For $k_1, k_2 \in \{0, 1\} (k_1 \neq k_2)$, it holds that $\sum_{i=1}^n D(p(x_i|y = k_1)\|p(x_i|y = k_2)) = \beta_{k_1,k_2} n = \Omega(n)$.*

It means that samples from different classes ($y = 0$ and $y = 1$) should have different distributions on at least $\Omega(1)$ fraction of their features.

**Assumption B.2.** *For all $k \in \{0, 1\}$, it holds that $\mathbb{V}_{\boldsymbol{x}}[\sum_{i=1}^n \log \frac{p(x_i|y=1)}{p(x_i|y=0)}|y = k] = \alpha_k n = O(n^r)$, where $r < 2$..*

**Proposition B.2** (Proof in Appendix F.4)**.** *Suppose that Assumption 3.1, B.1 and B.2 hold, then $G(\tau)$ is polynomially small in $n$:*

$$G(\tau) \leq \frac{\alpha}{(\tau - \zeta)^2 n},$$

*where $\alpha = \min_k |\alpha_k| = O(n^{r-1})$, $\mathbb{E}_{\boldsymbol{x}}[\Delta a_{Gen,\infty}(\boldsymbol{x}, 1, 0)|y = k] = \zeta_k n$, $\zeta = \min_k |\zeta_k| = \Omega(1)$ and $\tau < \zeta$.*

Indeed, if the naïve Bayes assumption really holds, that is, feature values are independent given the label, we can obtain a much stronger guarantee for $G(\tau)$.

**Proposition B.3** (Proof in Appendix F.5)**.** *Suppose that Assumption 3.1, 3.2, B.1 and the naïve Bayes assumption hold, then $G(\tau)$ is exponentially small in $n$, that is,*

$$G(\tau) \leq \exp -O((\tau - \beta)^2 n),$$

*where $\mathbb{E}_{\boldsymbol{x}}[\Delta a_{Gen,\infty}(\boldsymbol{x}, 1, 0)|y = k] = \zeta_k n$, $\zeta = \min_k |\zeta_k| = \Omega(1)$ and $\tau < \zeta$.*

Using the results from Theorem B.2, we can obtain the sample complexity of naïve Bayes as follows.

**Theorem B.3** (Proof in Appendix F.6)**.** *Suppose that either precondition of Proposition B.2 or Proposition B.3 holds. Then, it suffices to pick $m = O(\log n)$ training samples such that $R_{\ell_{0-1}}(h_{Gen,m}) \leq R_{\ell_{0-1}}(h_{Gen,\infty}) + \epsilon_0$ hold with probability $1 - \delta_0$, for any $\epsilon_0 \in (0, 1)$ and $\delta_0 \in (0, \frac{\epsilon_0}{2}]$.*

Compare Corollary B.1 with B.3, we revisit the results in (Ng & Jordan, 2001). But we highlight that our results are obtained based on different assumptions and novel $\mathcal{H}$-consistency bound.

## C. Deferred Results

Proofs of results in this section can be found in Section G.

**Proposition C.1** (Distribution-dependent concave bound, proof in G.1)**.** *For a fixed distribution, if there exists a concave function $s : \mathbb{R}_+ \to \mathbb{R}$ and $\epsilon \geq 0$ such that the following holds for any $\boldsymbol{h} \in \mathcal{H}$ and $x \in \mathcal{X}$:*

$$\langle \Delta\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})\rangle_\epsilon \leq s(\Delta\mathcal{C}_{\ell_1,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})).$$

*Then it holds for all $\mathbf{h} \in \mathcal{H}$ that*

$$R_{\ell_2}(\boldsymbol{h}) - R^*_{\ell_2,\mathcal{H}} + M_{\ell_2,\mathcal{H}} \leq s(R_{\ell_1}(\boldsymbol{h}) - R^*_{\ell_1,\mathcal{H}} + M_{\ell_1,\mathcal{H}}) + \epsilon.$$

**Theorem C.1** (Distribution-independent concave $\ell_{0-1}$ bound, proof in Appendix G.2)**.** *Suppose that $\mathcal{H}$ satisfies that $\{\mathrm{argmax}_{y\in\mathcal{Y}} h_y(\boldsymbol{x}) : \boldsymbol{h} \in \mathcal{H}\} = \{1,\dots,K\}$ for any $\boldsymbol{x} \in \mathcal{X}$. If there exists a non-decreasing concave function $s : \mathbb{R}_+ \to \mathbb{R}_+$ with $t \leq s(\mathcal{J}_\ell(t))$. Then it holds for all $\mathbf{h} \in \mathcal{H}$ and any distribution $\mathcal{D}$ that*

$$R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}} \leq s(R_{\ell(\boldsymbol{h})} - R^*_{\ell,\mathcal{H}} + M_{\ell,\mathcal{H}}).$$

**Theorem C.2** (Multiclass $\mathcal{H}$-consistency bound for $\ell_{log}$ with one-hidden-layer neural network, proof in Appendix G.3)**.** *Given family of one-hidden-layer neural network hypotheses with ReLU activation function $(\cdot)_+$ $\mathcal{H}_{NN} = \{\boldsymbol{x} \to \boldsymbol{h}(\boldsymbol{x}) : h_y(\boldsymbol{x}) = \sum_{j=1}^n U_{yj}((\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle + b)_+\}$, where $\boldsymbol{U} \in \mathbb{R}^{K\times n}$, $\boldsymbol{w}_j \in \mathbb{R}^n$ and $b \in \mathbb{R}$, then it holds for any distribution that $R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}_{NN}} + M_{\ell_{0-1},\mathcal{H}_{NN}} \leq \sqrt{2}(R_{\ell_{log}}(\boldsymbol{h}) - R^*_{\ell_{log},\mathcal{H}_{NN}} + M_{\ell_{log},\mathcal{H}_{NN}})^{\frac{1}{2}}$.*

**Proposition C.2** (Proof in Appendix G.4)**.** *Suppose that Assumption 3.1, 3.2,3.3 and naïve Bayes assumption hold, then $\widetilde{G}(\tau)$ is exponentially small in $n$:*

$$\widetilde{G}(\tau) \leq \exp -O((\tau - \zeta)^2 n),$$

*where $\mathbb{E}_{\boldsymbol{x}}[\Delta a_{Gen,\infty}(\boldsymbol{x}, k_1, k_2) | y = k] = \zeta_{k_1,k_2,k} n$, $\zeta = \min_{k_1,k_2,k} |\zeta_{k_1,k_2,k}| = \Omega(1)$ and $\tau < \zeta$.*

# D. Proofs of Section 3.1

## D.1. Proof of Theorem 3.1

The proof is very similar to the proof of binary case (Theorem B.2). Similarly, there are some lemmas to bound the $|\Delta a_{Gen}(\boldsymbol{x}, k_1, k_2) - \Delta a_{Gen,\infty}(\boldsymbol{x}, k_1, k_2)|$ with high probability.

**Lemma D.1.** *In case of discrete inputs, and suppose that Assumption 3.2 holds, then with probability at least $1 - \delta$, for every fixed $k_1, k_2$ the following holds:*

$$|\Delta a_{Gen}(\boldsymbol{x}, k_1, k_2) - \Delta a_{Gen,\infty}(\boldsymbol{x}, k_1, k_2)| \leq \frac{4(n+1)}{\rho_0}\sqrt{\frac{1}{\rho_0 m}\log(\frac{2(4n+2)}{\delta})} = O\big(n\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}\big). \tag{7}$$

*Proof.* The proof is almost the same as the proof of the binary case (Lemma F.5). Just replace the label $\{0, 1\}$ with $\{k_1, k_2\}$ and notice that $|\log \hat{p}(y = k_1) - \log p(y = k_1)| \leq \epsilon$ no longer implies that $|\log \hat{p}(y = k_2) - \log p(y = k_2)| \leq \epsilon$. □

**Lemma D.2.** *In case of continuous inputs, and suppose that Assumption 3.2 holds, then with probability at least $1 - \delta$, the following holds:*

$$\begin{aligned} |\Delta a_{Gen}(\boldsymbol{x}, k_1, k_2) - \Delta a_{Gen,\infty}(\boldsymbol{x}, k_1, k_2)| &\leq 4(\frac{n}{3\rho_0}(\frac{4}{\rho_0^2} + \frac{3}{\rho_0} + \sqrt{\frac{2}{\rho_0}}) + \frac{1}{\rho_0})\sqrt{\frac{1}{\rho_0 m}\log(\frac{2(5n+2)}{\delta})} \quad (8) \\ &= O\big(n\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}\big). \quad (9) \end{aligned}$$

*Proof.* The proof is almost the same as the proof of the binary case (Lemma F.7). Just replace the label $\{0, 1\}$ with $\{k_1, k_2\}$ and notice that $|\log \hat{p}(y = k_1) - \log p(y = k_1)| \leq \epsilon$ no longer implies that $|\log \hat{p}(y = k_2) - \log p(y = k_2)| \leq \epsilon$. □

Based on Lemma D.1 and D.2, we are ready to prove Theorem 3.1.

*Proof.* Let $\delta$ and $\epsilon = O\big(n\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}\big)$ are what claimed in the Lemma D.1 for discrete case and Lemma D.2 for the continuous case.We calculate the $|R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty})|$ for multiclass naïve Bayes as follows:

$$
\begin{aligned}
&|R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,m}) - R_{\ell_{0-1}}(\boldsymbol{h}_{Gen,\infty})| \\
&= |\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell_{0-1}(\boldsymbol{h}_{Gen,m},(\boldsymbol{x},y)) - \ell_{0-1}(\boldsymbol{h}_{Gen,\infty},(\boldsymbol{x},y))]| \\
&\le \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}|\ell_{0-1}(\boldsymbol{h}_{Gen,m},(\boldsymbol{x},y)) - \ell_{0-1}(\boldsymbol{h}_{Gen,\infty},(\boldsymbol{x},y))| \\
&= \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}}(\operatorname*{argmax}_k a_{Gen}(\boldsymbol{x},k) \neq \operatorname*{argmax}_k a_{Gen,\infty}(\boldsymbol{x},k)) \\
&\le \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}}(\cup_{k_1,k_2}\Delta a_{Gen}(\boldsymbol{x},k_1,k_2)\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2) < 0) \\
&\le \sum_{k_1\neq k_2}\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}}(\Delta a_{Gen}(\boldsymbol{x},k_1,k_2)\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2) < 0) \\
&\le \frac{K(K-1)}{2}\max_{k_1,k_2}\mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}}(\Delta a_{Gen}(\boldsymbol{x},k_1,k_2)\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2) < 0) \\
&\le \frac{K(K-1)}{2}\max_{k_1,k_2}\big(\mathbb{P}(\Delta a_{Gen}(\boldsymbol{x},k_1,k_2)\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2) < 0 \| \Delta a_{Gen}(\boldsymbol{x},k_1,k_2) - \Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)| \le \epsilon) + \delta\big) \\
&\le \frac{K(K-1)}{2}\max_{k_1,k_2}\Big(\mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)| \le O\Big(n\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}\Big))) + \delta\Big) \\
&= \frac{K(K-1)}{2}\bigg(\widetilde{G}\Big(O(\sqrt{\frac{1}{m}\log(\frac{n}{\delta})})\Big) + \delta\bigg).
\end{aligned}
$$

The proof of Theorem 3.1 is complete. □

### D.2. Proof of Proposition 3.1

The following lemma states that the expectation of $\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)$ condition on $y$ is always large, which is essential to the proof of Proposition 3.1.

**Lemma D.3.** *Suppose that Assumption 3.3 holds, then for every $k_1, k_2$ and $k \in \mathcal{Y}$, it holds that $|\mathbb{E}_{\boldsymbol{x}}[\sum_{i=1}^n \log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}|y = k]| = \Omega(n)$, which implies that $|\mathbb{E}[\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)|y = k]| = \Omega(n)$.*

*Proof.* We calculate $|\mathbb{E}_{\boldsymbol{x}}[\sum_{i=1}^n \log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}|y = k]|$ directly:

$$
\begin{aligned}
&|\mathbb{E}_{\boldsymbol{x}}[\sum_{i=1}^n \log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}|y=k]| \\
&= |\sum_{i=1}^n \mathbb{E}_{x_i}[\log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}|y=k]| \\
&= |\sum_{i=1}^n \sum_{x_i}(p(x_i|y=k)\log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)})| \\
&= |\sum_{i=1}^n \sum_{x_i}(p(x_i|y=k)\log\frac{p(x_i|y=k)}{p(x_i|y=k_2)} - p(x_i|y=k)\log\frac{p(x_i|y=k)}{p(x_i|y=k_1)})| \\
&= |\sum_{i=1}^n (D(p(x_i|y=k)\|p(x_i|y=k_2)) - D(p(x_i|y=k)\|p(x_i|y=k_1)))| \\
&= \beta_{k_2,k_1,k}n = \Omega(n). \qquad \text{(Assumption 3.3)}
\end{aligned}
$$

Furthermore, we can obtain

$$
\begin{aligned}
&|\mathbb{E}_{\boldsymbol{x}}[\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)|y=k]| \\
&=|\mathbb{E}_{\boldsymbol{x}}[\sum_{i=1}^{n}\log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}+\log\frac{p(y=k_1)}{p(y=k_2)}|y=k]| \\
&=|\sum_{i=1}^{n}\mathbb{E}_{x_i}[\log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}|y=k]+\log\frac{p(y=k_1)}{p(y=k_2)}| \\
&\geq \beta_{k_2,k_1,k}n-|\log\frac{p(y=k_1)}{p(y=k_2)}| \\
&\geq \beta_{k_2,k_1,k}n-|\log\frac{\rho_0}{1-\rho_0}| && \text{(Assumption 3.1)} \\
&=\Omega(n),
\end{aligned}
$$

which implies that $|\mathbb{E}_{\boldsymbol{x}}[\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)|y=k]|=\Omega(n)$. Then the lemma is proved. □

Built upon Lemma D.3, we prove Theorem B.2 as follows.

*Proof.* For $k_1,k_2$ and $k$ which satisfies $\zeta_{k_1,k_2,k}>0$, to bound $\mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)|\leq \tau n|y=k)$ with $\tau\in(0,\zeta)$. we can write:

$$
\begin{aligned}
&\mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)|\leq \tau n|y=k) \\
&\leq \mathbb{P}(\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)\leq \tau n|y=k) \\
&=\mathbb{P}(\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)-\zeta_{k_1,k_2,k}n\leq \tau n-\zeta_{k_1,k_2,k}n|y=k) \\
&=\mathbb{P}(\sum_{i=1}^{n}\log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}-\mathbb{E}_{\boldsymbol{x}}(\sum_{i=1}^{n}\log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)})\leq(\tau-\zeta_{k_1,k_2,k})n|y=k) \\
&\leq \mathbb{P}(|\sum_{i=1}^{n}\log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}-\mathbb{E}_{\boldsymbol{x}}(\sum_{i=1}^{n}\log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)})|\geq(\zeta_{k_1,k_2,k}-\tau)n|y=k) \\
&\leq \frac{\mathbb{V}[\sum_{i=1}^{n}\log\frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}|y=k]}{(\tau-\zeta_{k_1,k_2,k})^2n^2} && \text{(Chebyshev inequality)} \\
&=\frac{\alpha_{k_1,k_2,k}n}{(\tau-\zeta_{k_1,k_2,k})^2n^2} && \text{(Assumption 3.4)} \\
&=\frac{\alpha_{k_1,k_2,k}}{(\tau-\zeta_{k_1,k_2,k})^2n}.
\end{aligned}
$$

Similar to the above discussion, we have $\mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)|\leq \tau n|y=k)\leq\frac{\alpha_{k_1,k_2,k}}{(\tau-|\zeta_{k_1,k_2,k}|)^2n}$ for $k_1,k_2$ and $k$ which satisfies $\zeta_{k_1,k_2,k}<0$. Finally, we can conclude that:

$$
\begin{aligned}
\widetilde{G}(\tau)&=\max_{k_1,k_2}\sum_{k=1}^{K}p(y=k)\mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x},k_1,k_2)|\leq\tau n|y=k) \\
&\leq\max_{k_1,k_2}\sum_{k=1}^{K}p(y=k)\frac{\alpha_{k_1,k_2,k}}{(\tau-|\zeta_{k_1,k_2,k}|)^2n} \\
&\leq\max_{k_1,k_2}\frac{\max_k\alpha_{k_1,k_2,k}}{(\tau-\min_k|\zeta_{k_1,k_2,k}|)^2n} \\
&=\frac{\max_{k_1,k_2,k}\alpha_{k_1,k_2,k}}{(\tau-\min_{k_1,k_2,k}|\zeta_{k_1,k_2,k}|)^2n}=\frac{\alpha}{(\tau-\zeta)^2n}.
\end{aligned}
$$

□

### D.3. Proof of Theorem 3.2

*Proof.* In the case that precondition of Proposition 3.1 holds, combining Theorem 3.1 and Proposition 3.1, we know that there exist positive $c = \Theta(1)$ and large enough $m$ such that when $c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta$, with probability at least $1 - \delta$, we have

$$\begin{aligned}
R_{\ell_{0-1}}(h_{Gen,m}) &\leq R_{\ell_{0-1}}(h_{Gen,\infty}) + \frac{K(K-1)}{2}\left(\frac{\alpha}{(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n} + \delta\right) \\
&\leq R_{\ell_{0-1}}(h_{Gen,\infty}) + \frac{K^2}{2}\left(\frac{\alpha}{(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n} + \delta\right).
\end{aligned}$$

For fixed $\epsilon_0 \in (0,1)$, the logical relations listed in the following is correct:

$$\begin{aligned}
& R_{\ell_{0-1}}(h_{Gen,m}) \leq R_{\ell_{0-1}}(h_{Gen,\infty}) + \epsilon_0 \text{ with probability at least 1 - } \delta \\
& \Leftarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < 1 \wedge \frac{K^2}{2}\left(\frac{\alpha}{(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n} + \delta\right) \leq \epsilon_0 \\
& \Leftrightarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < 1 \wedge \frac{\alpha}{(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n} \leq \frac{2\epsilon_0}{K^2} - \delta \\
& \Leftrightarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < 1 \wedge \frac{2\epsilon_0}{K^2} - \delta > 0 \wedge (c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 \geq \frac{\alpha}{(\frac{2\epsilon_0}{K^2} - \delta)n} \\
& \Leftrightarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < \frac{2\epsilon_0}{K^2} \wedge (c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 \geq \frac{\alpha}{(\frac{2\epsilon_0}{K^2} - \delta)n} \\
& \Leftrightarrow 0 < \delta < \frac{2\epsilon_0}{K^2} \wedge \zeta - c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} \geq \sqrt{\frac{\alpha}{(\frac{2\epsilon_0}{K^2} - \delta)n}} \\
& \Leftrightarrow 0 < \delta < \frac{2\epsilon_0}{K^2} \wedge \zeta - \sqrt{\frac{\alpha}{(\frac{2\epsilon_0}{K^2} - \delta)n}} > 0 \wedge (\zeta - \sqrt{\frac{\alpha}{(\frac{2\epsilon_0}{K^2} - \delta)n}})^2 \geq c^2\frac{1}{m}\log(\frac{n}{\delta}) \\
& \Leftarrow 0 < \delta < \frac{2\epsilon_0}{K^2} - \frac{\alpha}{\zeta^2 n} \wedge \frac{2\epsilon_0}{K^2} - \frac{\alpha}{\zeta^2 n} > 0 \wedge m \geq \frac{c^2}{(\zeta - \sqrt{\frac{\alpha}{(\frac{2\epsilon_0}{K^2} - \delta)n}})^2}\log(\frac{n}{\delta}) \\
& \Leftarrow 0 < \delta \leq \frac{\epsilon_0}{K^2} \wedge \frac{2\epsilon_0}{K^2} - \frac{\alpha}{\zeta^2 n} > \frac{\epsilon_0}{K^2} \wedge m \geq \frac{c^2}{(\zeta - \sqrt{\frac{\alpha}{(\frac{2\epsilon_0}{K^2} - \delta)n}})^2}\log(\frac{n}{\delta}) \\
& \Leftrightarrow 0 < \delta \leq \frac{\epsilon_0}{K^2} \wedge K < \sqrt{\frac{\epsilon_0\zeta^2 n}{\alpha}} \wedge m \geq \frac{c^2}{(\zeta - \sqrt{\frac{\alpha}{(\frac{2\epsilon_0}{K^2} - \delta)n}})^2}\log(\frac{n}{\delta}) \\
& \Leftarrow 0 < \delta \leq \frac{\epsilon_0}{K^2} \wedge 2K < \sqrt{\frac{\epsilon_0\zeta^2 n}{\alpha}} \wedge m \geq \frac{c^2}{(\zeta - \sqrt{\frac{\alpha K^2}{\epsilon_0 n}})^2}\log(\frac{n}{\delta}) \\
& \Leftarrow 0 < \delta \leq \frac{\epsilon_0}{K^2} \wedge 2K < \sqrt{\frac{\epsilon_0\zeta^2 n}{\alpha}} \wedge m \geq \frac{c^2}{(\zeta - \frac{\zeta}{2})^2}\log(\frac{n}{\delta}) \\
& \Leftarrow 0 < \delta \leq \frac{\epsilon_0}{K^2} \wedge 2K < \sqrt{\frac{\epsilon_0\zeta^2 n}{\alpha}} \wedge m = O(\log(n)).
\end{aligned}$$

We note that in the case that precondition of Proposition C.2 holds, the $O(\log(n))$ result is correct as well. Combining Theorem 3.1 and Proposition C.2, we know that there exist positive $b, c = \Theta(1)$ and large enough $m$ such that when

$c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta$, with probability at least $1-\delta$, we have

$$\begin{aligned} R_{\ell_{0-1}}(h_{Gen,m}) &\leq R_{\ell_{0-1}}(h_{Gen,\infty}) + \frac{K(K-1)}{2}\left(\exp(-b(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}-\zeta)^2 n) + \delta\right) \\ &\leq R_{\ell_{0-1}}(h_{Gen,\infty}) + \frac{K^2}{2}\left(\exp(-b(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}-\zeta)^2 n) + \delta\right). \end{aligned}$$

For fixed $\epsilon_0 \in (0,1)$, the logical relations listed in the following is correct:

$$\begin{aligned}
&R_{\ell_{0-1}}(h_{Gen,m}) \leq R_{\ell_{0-1}}(h_{Gen,\infty}) + \epsilon_0 \text{ with probability at least 1 - } \delta \\
&\Leftarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \land 0 < \delta < 1 \land \frac{K^2}{2}\left(\exp(-b(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}-\zeta)^2 n) + \delta\right) \leq \epsilon_0 \\
&\Leftrightarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \land 0 < \delta < 1 \land \exp(-b(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}-\zeta)^2 n) \leq \frac{2\epsilon_0}{K^2} - \delta \\
&\Leftrightarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \land 0 < \delta < 1 \land \frac{2\epsilon_0}{K^2} - \delta > 0 \land -b(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}-\zeta)^2 n \leq \log(\frac{2\epsilon_0}{K^2} - \delta) \\
&\Leftrightarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \land 0 < \delta < \frac{2\epsilon_0}{K^2} \land (c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}-\zeta)^2 \geq \frac{1}{bn}\log(\frac{1}{\frac{2\epsilon_0}{K^2}-\delta}) \\
&\Leftarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \land 0 < \delta < \frac{2\epsilon_0}{K^2} \land \zeta - c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} \geq \sqrt{\frac{1}{bn}\log(\frac{1}{\frac{2\epsilon_0}{K^2}-\delta})} \\
&\Leftrightarrow 0 < \delta < \frac{2\epsilon_0}{K^2} \land \zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\frac{2\epsilon_0}{K^2}-\delta})} \geq c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} \\
&\Leftrightarrow 0 < \delta < \frac{2\epsilon_0}{K^2} \land \zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\frac{2\epsilon_0}{K^2}-\delta})} > 0 \land (\zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\frac{2\epsilon_0}{K^2}-\delta})})^2 \geq c^2\frac{1}{m}\log(\frac{n}{\delta}) \\
&\Leftarrow 0 < \delta < \frac{2\epsilon_0}{K^2} - \exp(-b\zeta^2 n) \land \frac{2\epsilon_0}{K^2} - \exp(-b\zeta^2 n) > 0 \land m \geq \frac{c^2}{(\zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\frac{2\epsilon_0}{K^2}-\delta})})^2}\log(\frac{n}{\delta}) \\
&\Leftarrow 0 < \delta \leq \frac{\epsilon_0}{K^2} \land \frac{2\epsilon_0}{K^2} - \exp(-b\zeta^2 n) > \frac{\epsilon_0}{K^2} \land m \geq \frac{c^2}{(\zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\frac{2\epsilon_0}{K^2}-\delta})})^2}\log(\frac{n}{\delta}) \\
&\Leftrightarrow 0 < \delta \leq \frac{\epsilon_0}{K^2} \land K < \sqrt{\epsilon_0}\exp(\frac{bn\zeta^2}{2}) \land m \geq \frac{c^2}{(\zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\frac{2\epsilon_0}{K^2}-\delta})})^2}\log(\frac{n}{\delta}) \\
&\Leftarrow 0 < \delta \leq \frac{\epsilon_0}{K^2} \land 2K < \sqrt{\epsilon_0}\exp(\frac{bn\zeta^2}{2}) \land m \geq \frac{c^2}{(\zeta - \sqrt{\frac{1}{bn}\log(\frac{K^2}{\epsilon_0})})^2}\log(\frac{K^2 n}{\epsilon_0}) \\
&\Leftarrow 0 < \delta \leq \frac{\epsilon_0}{K^2} \land 2K < \sqrt{\epsilon_0}\exp(\frac{bn\zeta^2}{2}) \land m \geq \frac{c^2}{\zeta^2(1 - \frac{\log(K^2/\epsilon_0)}{\log(4K^2/\epsilon_0)})^2}\log(\frac{K^2 n}{\epsilon_0}) \\
&\Leftrightarrow 0 < \delta \leq \frac{\epsilon_0}{K^2} \land 2K < \sqrt{\epsilon_0}\exp(\frac{bn\zeta^2}{2}) \land m = O(\log(n)).
\end{aligned}$$

□

### D.4. Proof of Proposition 3.2

We first present the following lemmas to show Proposition 3.2.

**Lemma D.4** ((Mohri et al., 2018), Theorem 3.3)**.** *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0,c]$. Then, for any $\delta$*

*$> 0$, with probability at least $1 - \delta$ over the draw of an $i.i.d.$ sample $S$ of size $m$, the following holds for all $g \in \mathcal{G}$:*

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(\mathcal{G}) + c\sqrt{\frac{1}{2m}\log(\frac{2}{\delta})},$$

*where $\mathcal{R}_m(\mathcal{G})$ is the Rademacher complexity of $\mathcal{G}$.*

**Lemma D.5** ((Maurer, 2016), Corollary 4)**.** *Let $\mathcal{X}$ be any set, $S = (\boldsymbol{x}_1, ..., \boldsymbol{x}_m) \in \mathcal{X}^m$, $\sigma_1, \ldots, \sigma_m$ be Rademacher random variables, $\mathcal{H}$ be a class of functions $\boldsymbol{h} : \mathcal{X} \to \ell_2$ and let $\Phi : \ell_2 \to \mathbb{R}$ have Lipschitz norm $L$, where $\ell_2$ is Hilbert space of square summable sequences of real numbers. Then we have*

$$\mathcal{R}_m(\Phi \circ \mathcal{H}) = \frac{1}{m}\mathbb{E}_{S,\sigma}\sup_{\boldsymbol{h}}\sum_{i}\sigma_i\Phi(\boldsymbol{h}(\boldsymbol{x}_i)) \leq \sqrt{2}L\frac{1}{m}\mathbb{E}_{S,\sigma}\sup_{\boldsymbol{h}}\sum_{i,k}\sigma_{ik}h_k(\boldsymbol{x}_i).$$

*where $\sigma_{ik}$ is an independent doubly indexed Rademacher sequence and $h_k(x_i)$ is the $k$-th component of $\boldsymbol{h}(x_i)$.*

**Lemma D.6.** *Let $\mathcal{X} = [0,1]^n$, $S = (\boldsymbol{x}_1, ..., \boldsymbol{x}_m) \in \mathcal{X}^m$, $\mathcal{H} = \{\boldsymbol{x} \to \boldsymbol{h}(\boldsymbol{x}) : h_y(\boldsymbol{x}) = \boldsymbol{w}_y^T x + b_y, \|\boldsymbol{w}_y\|_2 \leq W, |b_y| \leq B, y \in \mathcal{Y}\}$ and $\sigma_{ik}$ be independent doubly indexed Rademacher sequence. Then we have*

$$\frac{1}{m}\mathbb{E}_{S,\sigma}\sup_{\boldsymbol{h}}\sum_{i,k}\sigma_{ik}h_k(\boldsymbol{x}_i) \leq WK\sqrt{\frac{n}{m}}.$$

*Proof.*

$$\begin{aligned}
\frac{1}{m}\mathbb{E}_{S,\sigma}\sup_{\boldsymbol{h}}\sum_{i,k}\sigma_{ik}h_k(\boldsymbol{x}_i) &= \frac{1}{m}\mathbb{E}_{S,\sigma}\sup_{\boldsymbol{h}}\sum_{i,k}\sigma_{ik}(\langle\boldsymbol{w}_k, \boldsymbol{x}_i\rangle + b_k) \\
&= \frac{1}{m}\mathbb{E}_{S,\sigma}\sup_{\boldsymbol{h}}\sum_{i,k}\sigma_{ik}\langle\boldsymbol{w}_k, \boldsymbol{x}_i\rangle \\
&\leq \sum_{k=1}^{K}\frac{1}{m}\mathbb{E}_{S,\sigma}\sup_{h_k}\sum_{i=1}^{m}\sigma_{ik}\langle\boldsymbol{w}_k, \boldsymbol{x}_i\rangle \\
&= \sum_{k=1}^{K}\frac{1}{m}\mathbb{E}_{S,\sigma}\sup_{h_k}\langle\boldsymbol{w}_k, \sum_{i=1}^{m}\sigma_{ik}\boldsymbol{x}_i\rangle \\
&\leq \sum_{k=1}^{K}\frac{1}{m}\mathbb{E}_{S,\sigma}\sup_{h_k}\|\boldsymbol{w}_k\|_2\|\sum_{i=1}^{m}\sigma_{ik}\boldsymbol{x}_i\|_2 \\
&\leq \sum_{k=1}^{K}\frac{W}{m}\mathbb{E}_{S,\sigma}\|\sum_{i=1}^{m}\sigma_{ik}\boldsymbol{x}_i\|_2 \\
&\leq \sum_{k=1}^{K}\frac{W}{m}\sqrt{\mathbb{E}_{S,\sigma}\|\sum_{i=1}^{m}\sigma_{ik}\boldsymbol{x}_i\|_2^2} \\
&= \sum_{k=1}^{K}\frac{W}{m}\sqrt{\sum_{i=1}^{m}\|\boldsymbol{x}_i\|_2^2} \\
&\leq \frac{WK}{m}\sqrt{m \times n} = WK\sqrt{\frac{n}{m}}
\end{aligned}$$

□

**Lemma D.7.** *Let $\mathcal{X} = [0,1]^n$, $\mathcal{Y} = \{1, \ldots, K\}$, $S = ((\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_m, y_m)) \in (\mathcal{X}, \mathcal{Y})^m$, $\mathcal{H} = \{\boldsymbol{x} \to \boldsymbol{h}(\boldsymbol{x}) : h_y(\boldsymbol{x}) = \boldsymbol{w}_y^T\boldsymbol{x} + b_y, \|\boldsymbol{w}_y\|_2 \leq W, |b_y| \leq B, y \in \mathcal{Y}\}$ and $\Pi_1(\mathcal{H}) = \{(\boldsymbol{x}, y) \to h_y(\boldsymbol{x}), y \in \mathcal{Y}, \boldsymbol{h} \in \mathcal{H}\}$. Then we have*

$$\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^{m}\sigma_i h_{y_i}(\boldsymbol{x}_i)] \leq K\mathcal{R}_m(\Pi_1(\mathcal{H})).$$

*Proof.*

$$
\begin{aligned}
&\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^{m}\sigma_i h_{y_i}(\boldsymbol{x}_i)]\\
&=\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^{m}\sigma_i\sum_{y\in\mathcal{Y}}h_y(\boldsymbol{x}_i)\mathbb{1}_{y_i=y}]\\
&\leq\frac{1}{m}\sum_{y\in\mathcal{Y}}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^{m}\sigma_i h_y(\boldsymbol{x}_i)\mathbb{1}_{y_i=y}]\\
&=\sum_{y\in\mathcal{Y}}\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^{m}\sigma_i h_y(\boldsymbol{x}_i)(\frac{\epsilon_i}{2}+\frac{1}{2})] && (\epsilon_i=2\times\mathbb{1}_{y_i=y}-1\in\{-1,+1\})\\
&\leq\sum_{y\in\mathcal{Y}}\frac{1}{2m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^{m}\sigma_i h_y(\boldsymbol{x}_i)\epsilon_i]+\frac{1}{2m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^{m}\sigma_i h_y(\boldsymbol{x}_i)]\\
&=\sum_{y\in\mathcal{Y}}\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^{m}\sigma_i h_y(\boldsymbol{x}_i)]\\
&\leq K\mathcal{R}_m(\Pi_1(\mathcal{H})).
\end{aligned}
$$

□

**Lemma D.8.** *Let $\mathcal{X}=[0,1]^n$, $\mathcal{Y}=\{1,\ldots,K\}$, $S=((\boldsymbol{x}_1,y_1),...,(\boldsymbol{x}_m,y_m))\in(\mathcal{X},\mathcal{Y})^m$, $\mathcal{H}=\{x\to\boldsymbol{h}(\boldsymbol{x}):h_y(\boldsymbol{x})=\boldsymbol{w}_y^T\boldsymbol{x}+b_y,\|\boldsymbol{w}_y\|_2\leq W,|b_y|\leq B,y\in\mathcal{Y}\}$ and $\Pi_1(\mathcal{H})=\{(\boldsymbol{x},y)\to h_y(\boldsymbol{x}),y\in\mathcal{Y},\boldsymbol{h}\in\mathcal{H}\}$. Then we have*

$$\mathcal{R}_m(\Pi_1(\mathcal{H}))\leq W\sqrt{\frac{n}{m}}.$$

*Proof.*

$$
\begin{aligned}
\mathcal{R}_m(\Pi_1(\mathcal{H}))&=\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h},y}\sum_{i=1}^{m}\sigma_i h_y(\boldsymbol{x}_i)]=\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h},y}\sum_{i=1}^{m}\sigma_i(\langle\boldsymbol{w}_y,\boldsymbol{x}_i\rangle+b_y)]\\
&=\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h},y}\sum_{i=1}^{m}\sigma_i\langle\boldsymbol{w}_y,\boldsymbol{x}_i\rangle]=\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h},y}\langle\boldsymbol{w}_y,\sum_{i=1}^{m}\sigma_i\boldsymbol{x}_i\rangle]\\
&\leq\frac{W}{m}\mathbb{E}_{S,\sigma}[\|\sum_{i=1}^{m}\sigma_i\boldsymbol{x}_i\|_2]\leq\frac{W}{m}\sqrt{\mathbb{E}_{S,\sigma}[\|\sum_{i=1}^{m}\sigma_i\boldsymbol{x}_i\|_2^2]}\\
&=\frac{W}{m}\sqrt{\sum_{i=1}^{m}\|\boldsymbol{x}_i\|_2^2}\leq\frac{W}{m}\sqrt{m\times n}=W\sqrt{\frac{n}{m}}.
\end{aligned}
$$

□

**Lemma D.9** ((Shalev-Shwartz & Ben-David, 2014), Lemma B.6)**.** *Let $Z_1,\ldots,Z_m$ be a sequence of i.i.d. random variables and let $\bar{Z}=\frac{1}{m}\sum_{i=1}^{m}Z_i$. Assume that $\mathbb{E}[\bar{Z}]=\mu$ and $P[a\leq Z_i\leq b]=1$ for every $i$. Then, for any $\epsilon>0$:*

$$\mathbb{P}[|\bar{Z}-\mu|>\epsilon]\leq 2\exp(-\frac{2m\epsilon^2}{(b-a)^2}).$$

Based on the above lemmas, we now prove Proposition 3.2 as follows.

*Proof.* We first rewrite $R_{\ell_{log}}(\boldsymbol{h}_{Dis,m})-R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty})$.

$$
\begin{aligned}
&R_{\ell_{log}}(\boldsymbol{h}_{Dis,m})-R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty})\\
&=R_{\ell_{log}}(\boldsymbol{h}_{Dis,m})-\hat{R}_{\ell_{log},S}(\boldsymbol{h}_{Dis,m})+\hat{R}_{\ell_{log},S}(\boldsymbol{h}_{Dis,m})-\hat{R}_{\ell_{log},S}(\boldsymbol{h}_{Dis,\infty})+\hat{R}_{\ell_{log},S}(\boldsymbol{h}_{Dis,\infty})-R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty})\\
&\leq(R_{\ell_{log}}(\boldsymbol{h}_{Dis,m})-\hat{R}_{\ell_{log},S}(\boldsymbol{h}_{Dis,m}))+(\hat{R}_{\ell_{log},S}(\boldsymbol{h}^*)-R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty})).
\end{aligned}
$$

We consider the first summand now. By Lemma D.4, with probability of at least $1-\delta$, we have:

$$\begin{aligned} &R_{\ell_{log}}(\boldsymbol{h}_{Dis,m}) - \hat{R}_{\ell_{log},S}(\boldsymbol{h}_{Dis,m}) \\ &\leq 2\mathcal{R}_m(\ell_{log}\circ(\mathcal{H},\mathcal{Y})) + \log(1+(K-1)\exp 2(W\sqrt{n}+B))\sqrt{\frac{1}{2m}\log(\frac{2}{\delta})} \end{aligned}$$

We define $\Pi_1(\mathcal{H}) = \{(\boldsymbol{x},y)\to h_y(\boldsymbol{x}), y\in\mathcal{Y}, \boldsymbol{h}\in\mathcal{H}\}$ and $\Phi = \{\boldsymbol{h}\to\log(\sum_{y=1}^K\exp(h_y)), \boldsymbol{h}\in\mathcal{H}(\boldsymbol{x})\subseteq\mathbb{R}^K\}$. We can bound $\mathcal{R}_m(\ell_{log}\circ(\mathcal{H},\mathcal{Y}))$ as follows:

$$\begin{aligned} \mathcal{R}_m(\ell_{log}\circ(\mathcal{H},\mathcal{Y})) &= \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^m\sigma_i\ell_{log}(\boldsymbol{h}(\boldsymbol{x}_i),y_i)] \\ &= \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^m\sigma_i(\log(\sum_{i=1}^K\exp(h_k(\boldsymbol{x}_i))) - h_{y_i}(\boldsymbol{x}_i))] \\ &\leq \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^m\sigma_i\log(\sum_{i=1}^K\exp(h_k(\boldsymbol{x}_i)))] + \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^m\sigma_i h_{y_i}(\boldsymbol{x}_i)] \\ &= \mathcal{R}_m(\Phi\circ\mathcal{H}) + \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^m\sigma_i h_{y_i}(\boldsymbol{x}_i)]. \end{aligned}$$

We will bound $\mathcal{R}_m(\Phi\circ\mathcal{H})$ by using Lemma D.5. Before that, we note $\Phi$ has Lipschitz norm $\sqrt{K}$. Because $\frac{\partial\Phi}{\partial h_i}\leq 1$ for any $i\in\{1,\ldots,K\}$. Then, for any $\boldsymbol{h}_1,\boldsymbol{h}_2\in\mathbb{R}^K$, we have

$$|\Phi(\boldsymbol{h}_1)-\Phi(\boldsymbol{h}_2)| \leq |\sum_{k=1}^K|\boldsymbol{h}_{1k}-\boldsymbol{h}_{2k}|| \leq \sqrt{K}\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2.$$

Then we can bound $\mathcal{R}_m(\Phi\circ\mathcal{H})$ as the following

$$\begin{aligned} \mathcal{R}_m(\Phi\circ\mathcal{H}) &\leq \sqrt{2}\sqrt{K}\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i,k}\sigma_{ik}h_k(\boldsymbol{x}_i)] && \text{(by Lemma D.5)} \\ &\leq \sqrt{2K}WK\sqrt{\frac{n}{m}} = W\sqrt{\frac{2K^3n}{m}}. && \text{(by Lemma D.6)} \end{aligned}$$

We can also bound $\frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^m\sigma_i h_{y_i}(\boldsymbol{x}_i)]$ as follows

$$\begin{aligned} \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_{\boldsymbol{h}}\sum_{i=1}^m\sigma_i h_{y_i}(\boldsymbol{x}_i)] &\leq K\mathcal{R}_m(\Pi_1(\mathcal{H})) && \text{(by Lemma D.7)} \\ &\leq KW\sqrt{\frac{n}{m}} = W\sqrt{\frac{K^2n}{m}}. && \text{(by Lemma D.8)} \end{aligned}$$

Therefore, we can obtain

$$R_{\ell_{log}}(\boldsymbol{h}_{Dis,m}) - \hat{R}_{\ell_{log},S}(\boldsymbol{h}_{Dis,m}) \leq 2W(\sqrt{\frac{2K^3n}{m}}+\sqrt{\frac{K^2n}{m}}) + \log(1+(K-1)\exp 2(W\sqrt{n}+B))\sqrt{\frac{1}{2m}\log(\frac{2}{\delta})}. \tag{10}$$

For the second summand, we use the fact that $R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty})$ does not depend on $\mathcal{S}$; hence by Lemma D.9, we obtain its bound:

$$\mathbb{P}(|\hat{R}_{\ell_{log},S}(\boldsymbol{h}_{Dis,\infty}) - R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty})| > \epsilon) \leq 2\exp(-\frac{2m\epsilon^2}{(c-0)^2}) = 2\exp(-\frac{2m\epsilon^2}{c^2}),$$

where $c = \log(1+(K-1)\exp 2(W\sqrt{n}+B))$. It implies that with the probability of at least $1-\delta$, we have:

$$\hat{R}_{\ell_{log},S}(\boldsymbol{h}_{Dis,\infty}) - R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty}) \leq c\sqrt{\frac{1}{2m}\log(\frac{2}{\delta})}. \tag{11}$$

At last, we make use of the union bound for Eq. (10) and (11) to get the final result. With probability at least $1 - \delta$, the following holds:

$$
\begin{aligned}
& R_{\ell_{log}}(\boldsymbol{h}_{Dis,m}) - R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty}) \\
& \le 2W(\sqrt{\frac{2K^3 n}{m}} + \sqrt{\frac{K^2 n}{m}}) + \log(1 + (K-1)\exp 2(W\sqrt{n} + B))\sqrt{\frac{1}{2m}\log(\frac{4}{\delta})} + c\sqrt{\frac{1}{2m}\log(\frac{4}{\delta})} \\
& = 2W(\sqrt{\frac{2K^3 n}{m}} + \sqrt{\frac{K^2 n}{m}}) + 2\log(1 + (K-1)\exp 2(W\sqrt{n} + B))\sqrt{\frac{1}{2m}\log(\frac{4}{\delta})} \\
& = O(\sqrt{\frac{K^3 n}{m}}).
\end{aligned}
$$

Therefore, for $R_{\ell_{log}}(h_{Dis,m}) \le R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty}) + \epsilon_0$ to hold with high probability $1 - \delta_0$ (here, $\epsilon_0$ and $\delta_0$ are some fixed constant), it suffices to pick $m = O(n)$ samples. □

### D.5. Proof of Theorem 3.4

*Proof.* By Theorem 3.3 we know that for $R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,m}) \le R_{\ell_{0-1}}(\boldsymbol{h}_{Dis,\infty}) + \epsilon_0$, it is sufficient to ensure that $R_{\ell_{log}}(\boldsymbol{h}_{Dis,m}) \le R_{\ell_{log}}(\boldsymbol{h}_{Dis,\infty}) + \frac{1}{2}\epsilon_0^2$. Then by Proposition 3.2, it suffices to sample $m = O(\frac{K^3 n}{\epsilon_0^4}) = O(K^3 n)$. □

## E. Proofs of Section 3.2

### E.1. Proof of Proposition 3.3

*Proof.* Fix $\boldsymbol{h} \in \mathcal{H}$, because $g(\langle \Delta\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})\rangle_\epsilon) < \Delta\mathcal{C}_{\ell_1,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})$ for all $x \in \mathcal{X}$, we have:

$$
\begin{aligned}
& g(R_{\ell_2}(\boldsymbol{h}) - R^*_{\ell_2,\mathcal{H}} + M_{\ell_2,\mathcal{H}}) \\
& = g(\mathbb{E}_{\boldsymbol{x}}[\mathcal{C}_{\ell_2}(\boldsymbol{h},\boldsymbol{x})] - R^*_{\ell_2,\mathcal{H}} + R^*_{\ell_2,\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[\mathcal{C}^*_{\ell_2,\mathcal{H}}(\boldsymbol{x})]) \\
& = g(\mathbb{E}_{\boldsymbol{x}}[\mathcal{C}_{\ell_2}(\boldsymbol{h},\boldsymbol{x}) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(\boldsymbol{x})]) \\
& \le \mathbb{E}_{\boldsymbol{x}}[g(\mathcal{C}_{\ell_2}(\boldsymbol{h},\boldsymbol{x}) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(\boldsymbol{x}))] && \text{(Jensen's inequality)} \\
& = \mathbb{E}_{\boldsymbol{x}}[g(\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}))] \\
& = \mathbb{E}_{\boldsymbol{x}}[g(\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})\mathbb{1}_{\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})>\epsilon} + \Delta\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})\mathbb{1}_{\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})\le\epsilon})] \\
& \le \mathbb{E}_{\boldsymbol{x}}[g(\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})\mathbb{1}_{\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})>\epsilon}) + g(\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})\mathbb{1}_{\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})\le\epsilon})] && (g(0) \ge 0) \\
& \le \mathbb{E}_{\boldsymbol{x}}[\Delta\mathcal{C}_{\ell_1,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}) + \sup_{t\in[0,\epsilon]} g(t)] && \text{(assumption)} \\
& = R_{\ell_1}(\boldsymbol{h}) - R^*_{\ell_1,\mathcal{H}} + M_{\ell_1,\mathcal{H}} + \max(g(0), g(\epsilon)). && (g \text{ is convex})
\end{aligned}
$$

□

### E.2. Proof of Theorem 3.5

**Lemma E.1** (Character of conditional $\epsilon$-regret for $\ell_{0-1}$)**.** *Suppose that $\mathcal{H}$ satisfies that $\{\operatorname{argmax}_{y\in\mathcal{Y}} h_y(\boldsymbol{x}) : \boldsymbol{h} \in \mathcal{H}\} = \{1, \ldots, K\}$ for any $\boldsymbol{x} \in \mathcal{X}$, then the minimal conditional zero-one loss $\ell_{0-1}$ is*

$$\mathcal{C}^*_{\ell_{0-1},\mathcal{H}}(\boldsymbol{x}) = \mathcal{C}^*_{\ell_{0-1},\mathcal{H}_{all}}(\boldsymbol{x}) = 1 - \max_y p_y(\boldsymbol{x}).$$

*Furthermore, the conditional $\epsilon$-regret for $\ell_{0-1}$ can be characterized as*

$$\langle \Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})\rangle_\epsilon = \langle \max_y p_y(\boldsymbol{x}) - p_{\hat{y}}(\boldsymbol{x})\rangle_\epsilon,$$

*where $\hat{y} = \operatorname{argmax}_{y\in\mathcal{Y}} h_y(\boldsymbol{x})$.*

*Proof.* By the definition of $\mathcal{C}_{\ell_{0-1}}(\boldsymbol{h}, \boldsymbol{x})$, we have:

$$\mathcal{C}_{\ell_{0-1}}(\boldsymbol{h}, \boldsymbol{x}) = \sum_{y=1}^{K} p_y(\boldsymbol{x}) \ell_{0-1}(\boldsymbol{h}(\boldsymbol{x}), y) = \sum_{y=1}^{K} p_y(\boldsymbol{x}) \mathbb{1}_{\hat{y} \neq y}.$$

By the assumption, we know that there exists $\boldsymbol{h}^* \in \mathcal{H}$ which satisfies $\operatorname{argmax}_{y \in \mathcal{Y}} h_y^*(\boldsymbol{x}) = \operatorname{argmax}_{y \in \mathcal{Y}} p_y(\boldsymbol{x})$. Therefore, we have

$$\mathcal{C}^*_{\ell_{0-1}, \mathcal{H}}(\boldsymbol{x}) = \inf_{\boldsymbol{h} \in \mathcal{H}} \mathcal{C}_{\ell_{0-1}}(\boldsymbol{h}, \boldsymbol{x}) = \mathcal{C}_{\ell_{0-1}}(\boldsymbol{h}^*, \boldsymbol{x}) = 1 - \max_y p_y(\boldsymbol{x}).$$

Then we can find the characteristic of conditional $\epsilon$-regret for $\ell_{0-1}$ as follows:

$$\begin{aligned}
\Delta \mathcal{C}_{\ell_{0-1}, \mathcal{H}}(\boldsymbol{h}, \boldsymbol{x}) &= \mathcal{C}_{\ell_{0-1}}(\boldsymbol{h}, \boldsymbol{x}) - \mathcal{C}^*_{\ell_{0-1}, \mathcal{H}}(\boldsymbol{x}) \\
&= \sum_{y=1}^{K} p_y(\boldsymbol{x}) \mathbb{1}_{\hat{y} \neq y} - (1 - \max_y p_y(\boldsymbol{x})) \\
&= \sum_{y \neq \hat{y}} p_y(\boldsymbol{x}) - \sum_{y \neq y_{max}} p_y(\boldsymbol{x}) \\
&= \max_y p_y(\boldsymbol{x}) - p_{\hat{y}}(\boldsymbol{x}).
\end{aligned}$$

□

**Lemma E.2** (Distribution-dependent convex $\ell_{0-1}$ bound)**.** *Suppose that $\mathcal{H}$ satisfies that $\{\operatorname{argmax}_{y \in \mathcal{Y}} h_y(\boldsymbol{x}) : \boldsymbol{h} \in \mathcal{H}\} = \{1, \ldots, K\}$ for any $\boldsymbol{x} \in \mathcal{X}$, and there exists a convex function $g : \mathbb{R}_+ \to \mathbb{R}$ with $g(0) = 0$ and $\epsilon \geq 0$ that the following holds for any $\hat{y} \in \mathcal{Y}$, $x \in \mathcal{X}$ and $\boldsymbol{h} \in \mathcal{H}_{\hat{y}}(\boldsymbol{x})$:*

$$g(\langle \max_y p_y(\boldsymbol{x}) - p_{\hat{y}}(\boldsymbol{x}) \rangle_\epsilon) \leq \inf_{\boldsymbol{h} \in \mathcal{H}_{\hat{y}}(\boldsymbol{x})} \Delta \mathcal{C}_{\ell, \mathcal{H}}(\boldsymbol{h}, \boldsymbol{x}).$$

*Then it holds for all $\boldsymbol{h} \in \mathcal{H}$ that*

$$g(R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1}, \mathcal{H}} + M_{\ell_{0-1}, \mathcal{H}}) \leq R_{\ell(\boldsymbol{h})} - R^*_{\ell, \mathcal{H}} + M_{\ell, \mathcal{H}} + \max(0, g(\epsilon)).$$

*Proof.* For any $\boldsymbol{x}_0 \in \mathcal{X}$ and $\boldsymbol{h}_0 \in \mathcal{H}$, let $\hat{y}$ be the index of the largest element of $\boldsymbol{h}_0(\boldsymbol{x})$. Then by the precondition, we have

$$g(\langle \Delta \mathcal{C}_{\ell_{0-1}, \mathcal{H}}(\boldsymbol{h}_0, \boldsymbol{x}_0) \rangle_\epsilon) = g(\langle \max_y p_y(\boldsymbol{x}_0) - p_{\hat{y}}(\boldsymbol{x}_0) \rangle_\epsilon) \leq \inf_{\boldsymbol{h} \in \mathcal{H}_{\hat{y}}(\boldsymbol{x}_0)} \Delta \mathcal{C}_{\ell, \mathcal{H}}(\boldsymbol{h}, \boldsymbol{x}_0) \leq \Delta \mathcal{C}_{\ell, \mathcal{H}}(\boldsymbol{h}_0, \boldsymbol{x}_0).$$

Combining the condition in Proposition 3.3 we can see that this lemma is correct. □

Built upon the above lemmas, we can prove Theorem 3.5 as follows.

*Proof.* For any $\boldsymbol{x}_0 \in \mathcal{X}$, $\boldsymbol{p}(\boldsymbol{x}_0) \in \Delta_K$, $\hat{y}_0 \in \mathcal{Y}$, and $\boldsymbol{h} \in \mathcal{H}_{\hat{y}_0}(\boldsymbol{x}_0)$, we can write:

$$\begin{aligned}
& g(\max_y p_y(\boldsymbol{x}_0) - p_{\hat{y}_0}(\boldsymbol{x}_0)) \\
& \leq \inf_{\hat{y} \in \mathcal{Y}, x \in \mathcal{X}, \boldsymbol{h} \in \mathcal{H}_{\hat{y}}(\boldsymbol{x}), \boldsymbol{p} \in \mathcal{P}_{\hat{y}}(\max_y p_y(\boldsymbol{x}_0) - p_{\hat{y}_0}(\boldsymbol{x}_0))} \Delta \mathcal{C}_{\ell, \mathcal{H}}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{p}) && \text{(Assumption)} \\
& \leq \inf_{x \in \mathcal{X}, \boldsymbol{h} \in \mathcal{H}_{\hat{y}_0}(\boldsymbol{x}), \boldsymbol{p} \in \mathcal{P}_{\hat{y}_0}(\max_y p_y(\boldsymbol{x}_0) - p_{\hat{y}_0}(\boldsymbol{x}_0))} \Delta \mathcal{C}_{\ell, \mathcal{H}}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{p}) \\
& \leq \inf_{x \in \mathcal{X}, \boldsymbol{h} \in \mathcal{H}_{\hat{y}_0}(\boldsymbol{x})} \Delta \mathcal{C}_{\ell, \mathcal{H}}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{p}(\boldsymbol{x}_0)) \\
& \leq \inf_{\boldsymbol{h} \in \mathcal{H}_{\hat{y}_0}(\boldsymbol{x}_0)} \Delta \mathcal{C}_{\ell, \mathcal{H}}(\boldsymbol{h}, \boldsymbol{x}_0, \boldsymbol{p}(\boldsymbol{x}_0)) = \inf_{\boldsymbol{h} \in \mathcal{H}_{\hat{y}_0}(\boldsymbol{x}_0)} \Delta \mathcal{C}_{\ell, \mathcal{H}}(\boldsymbol{h}, \boldsymbol{x}_0).
\end{aligned}$$

Combining the result of Lemma E.2, we conclude the proof of Theorem 3.5.

□

### E.3. Proofs of Theorem 3.6

*Proof.* By Theorem 3.5, if $\mathcal{J}_\ell(t)$ is convex with $\mathcal{J}_\ell(0) = 0$, the first inequality holds. For any $t \in [0,1]$, denote that the solution of $\inf_{\hat{y}\in\mathcal{Y},\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t),\boldsymbol{x}\in\mathcal{X},\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})$ by $\boldsymbol{x}^*, \boldsymbol{h}^*, \boldsymbol{p}^*, \hat{y}^*$. We then consider the distribution that is supported on the single point $\boldsymbol{x}_0 = \boldsymbol{x}^*$ and satisfy that $\boldsymbol{p}(\boldsymbol{x}_0) = \boldsymbol{p}^*$. Thus,

$$\inf_{\hat{y}\in\mathcal{Y},\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t),\boldsymbol{x}\in\mathcal{X},\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) = \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}^*}(\boldsymbol{x}_0)} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}_0,\boldsymbol{p}(\boldsymbol{x}_0)) = \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}^*}(\boldsymbol{x}_0)} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}_0).$$

For any $\delta > 0$, take $\boldsymbol{h}_0 \in \mathcal{H}$ such that $\boldsymbol{h}_0 \in \mathcal{H}_{\hat{y}^*}(\boldsymbol{x}_0)$ and

$$\Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h}_0,\boldsymbol{x}_0) \le \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}^*}(\boldsymbol{x}_0)} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}_0) + \delta = \inf_{\hat{y}\in\mathcal{Y},\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t),\boldsymbol{x}\in\mathcal{X},\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) + \delta.$$

Then, we have

$$\begin{aligned}
R_{\ell_{0-1}}(\boldsymbol{h}_0) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}} &= R_{\ell_{0-1}}(\boldsymbol{h}_0) - \mathbb{E}_{\boldsymbol{x}}[\mathcal{C}^*_{\ell_{0-1},\mathcal{H}}(\boldsymbol{x})] \\
&= \Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(\boldsymbol{h}_0,\boldsymbol{x}_0) \\
&= \max_y \boldsymbol{p}_y(\boldsymbol{x}_0) - \boldsymbol{p}_{\hat{y}^*}(\boldsymbol{x}_0) \\
&= t, \\
R_\ell(\boldsymbol{h}_0) - R^*_{\ell,\mathcal{H}} + M_{\ell,\mathcal{H}} &= R_\ell(\boldsymbol{h}_0) - \mathbb{E}_{\boldsymbol{x}}[\mathcal{C}^*_{\ell,\mathcal{H}}(\boldsymbol{x})] \\
&= \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h}_0,\boldsymbol{x}_0) \\
&\le \inf_{\hat{y},\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t),\boldsymbol{x}\in\mathcal{X},\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) + \delta \\
&= \mathcal{J}_\ell(t) + \delta,
\end{aligned}$$

which completes the proof. □

### E.4. Proof of Theorem 3.3

To prove the Theorem 3.3, we first list the following lemmas.

**Lemma E.3** (Convexity of $\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})$)**.** *$\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) = \sum_{y=1}^K p_y(-h_y + \log(\sum_{j=1}^K \exp(h_j))$ is convex with respect to $\boldsymbol{h}$.*

*Proof.* For any fixed $\boldsymbol{x}$ and $\boldsymbol{p}$, we have

$$\frac{\partial\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})}{\partial h_i} = -p_i + \frac{\exp(h_i)}{\sum_{k=1}^K \exp(h_k)}.$$

Let $A_{ij} = \frac{\partial^2\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})}{\partial h_i \partial h_j}$, we have

$$A_{ij} = \begin{cases} -\frac{\exp(h_i)\exp(h_j)}{(\sum_{k=1}^K \exp(h_k))^2}, & i \neq j, \\ \frac{\exp(h_j)\sum_{k=1,k\neq j}^K \exp(h_k)}{(\sum_{k=1}^K \exp(h_k))^2}, & i = j. \end{cases}$$

To prove $\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})$ is convex with respect to $\boldsymbol{h}$, it's sufficient to show that $\boldsymbol{A}$ is positive semidefinite, which equals to $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \ge 0$ for any $\boldsymbol{x} \in \mathbb{R}^n$. We can calculate it as follows:

$$\begin{aligned}
\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} &= A_{11}x_1^2 + \cdots + A_{nn}x_n^2 + \sum_{i\neq j} A_{ij}x_ix_j \\
&= \frac{1}{(\sum_{k=1}^K \exp(h_k))^2}\Big[\sum_{i=1}^K \exp(h_i)x_i^2\Big(\sum_{k=1,k\neq i}^K \exp(h_k)\Big) - \sum_{i=1}^K \exp(h_i)x_i\Big(\sum_{j=1,j\neq i}^K \exp(h_j)x_j\Big)\Big] \\
&= \frac{1}{(\sum_{k=1}^K \exp(h_k))^2}\Big[\sum_{i=1}^K \exp(h_i)x_i\Big(\sum_{j=1,j\neq i}^K \exp(h_j)(x_i - x_j)\Big)\Big] \\
&= \frac{1}{(\sum_{k=1}^K \exp(h_k))^2}\Big[\sum_{i<j} \exp(h_i)\exp(h_j)(x_i - x_j)^2\Big] \ge 0.
\end{aligned}$$

which proves this lemma. □

**Lemma E.4** (Property of $M_{\ell_{0-1},\mathcal{H}}$)**.** *Suppose that $\mathcal{H}$ satisfies that $\{\operatorname{argmax}_{y\in\mathcal{Y}} h_y(\boldsymbol{x}) : \boldsymbol{h} \in \mathcal{H}\} = \{1,\dots,K\}$ for any $\boldsymbol{x} \in \mathcal{X}$. Then $M_{\ell_{0-1},\mathcal{H}}$ coincides with the approximation error $R^*_{\ell_{0-1},\mathcal{H}} - R^*_{\ell_{0-1},\mathcal{H}_{all}}$.*

*Proof.*

$$\begin{aligned}
M_{\ell_{0-1},\mathcal{H}} &= R^*_{\ell_{0-1},\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[\mathcal{C}^*_{\ell_{0-1},\mathcal{H}}(\boldsymbol{x})] \\
&= R^*_{\ell_{0-1},\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[\inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_{\ell_{0-1}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}(\boldsymbol{x})] \\
&= R^*_{\ell_{0-1},\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[1 - \max_y p_y(\boldsymbol{x})] && \text{(Lemma E.1)} \\
&= R^*_{\ell_{0-1},\mathcal{H}} - R^*_{\ell_{0-1},\mathcal{H}_{all}}.
\end{aligned}$$

□

**Lemma E.5.** *Given $\boldsymbol{x}$, and $\boldsymbol{p} \in \Delta_K$, the following statements are equivalent:*

*(1)Optimation problem (14) can reach the global optimum,*

*(2)*$\max_y p_y - \min_y p_y \le \frac{\exp(W\|\boldsymbol{x}\|+B)-\exp(-(W\|\boldsymbol{x}\|+B))}{\exp(W\|\boldsymbol{x}\|+B)+(K-1)\exp(-(W\|\boldsymbol{x}\|+B))}$.

*Proof.* First, we prove that (1) implies (2). By the solutions of KKT conditions in (16), (1) means that $\exists \boldsymbol{h} \in \mathbb{R}^K$, $|h_i| \le W\|x\| + B$, and $p_i = \frac{\exp(h_i)}{\sum_{j=1}^K \exp(h_j)}$. We can directly write

$$\begin{aligned}
\max_y p_y - \min_y p_y &= \frac{\max_y \exp(h_y) - \min_y \exp(h_y)}{\sum_{j=1}^K \exp(h_j)} \\
&\le \frac{\max_y \exp(h_y) - \min_y \exp(h_y)}{\max_y \exp(h_y) + (K-1)\min_y \exp(h_y)} \\
&\le \frac{\exp(W\|\boldsymbol{x}\| + B) - \min_y \exp(h_y)}{\exp(W\|\boldsymbol{x}\| + B) + (K-1)\min_y \exp(h_y)} && \text{(increasing w.r.t } \max_y \exp(h_y)) \\
&\le \frac{\exp(W\|\boldsymbol{x}\| + B) - \exp(-(W\|\boldsymbol{x}\| + B))}{\exp(W\|\boldsymbol{x}\| + B) + (K-1)\exp(-(W\|\boldsymbol{x}\| + B))} && \text{(decreasing w.r.t } \min_y \exp(h_y))
\end{aligned}$$

Second, we prove that (2) implies (1). We suppose that if (1) does not hold, then in this case, let $\mathcal{I}_2 = \{y : p_y > \frac{\exp(W\|\boldsymbol{x}\|+B)}{\sum_{k=1}^K \exp(h_k^*)}\}$, $\mathcal{I}_3 = \{y : p_y < \frac{\exp(-W\|\boldsymbol{x}\|-B)}{\sum_{k=1}^K \exp(h_k^*)}\}$ and $\mathcal{I}_1 = \{1,\dots,K\} - \mathcal{I}_2 - \mathcal{I}_3$. We note that $\#\mathcal{I}_2, \#\mathcal{I}_3 > 0$. By (15) we know that

$$\sum_{i=1}^K -p_i + \frac{\exp(h_i^*)}{\sum_{k=1}^K \exp(h_k^*)} + \lambda_i^* - \mu_i^* = \sum_{i=1}^K (\lambda_i^* - \mu_i^*) = 0 \tag{12}$$

Because either $\mathcal{I}_2$ or $\mathcal{I}_3$ must be non-empty. We assume that $\mathcal{I}_2$ is not empty, then there exists $y_1$ such that $\lambda^*_{y_1} - \mu^*_{y_1} = \lambda^*_{y_1} > 0$. To make (12) hold, there should exists $y_2$ such that $\lambda^*_{y_2} - \mu^*_{y_2} < 0$, which implies that $y_2 \in \mathcal{I}_3$. Thus, for any $\boldsymbol{h}$, we have $\max_y p_y \ge \frac{\exp(W\|x\|+B)}{\sum_{i=j}^K \exp(h_j)}$ and $\min_y p_y \le \frac{\exp(-(W\|x\|+B))}{\sum_{i=j}^K \exp(h_j)}$. Then $\max_y p_y - \min_y p_y \ge \frac{\exp(W\|x\|+B)-\exp(-(W\|x\|+B))}{\sum_{i=j}^K \exp(h_j)}$ for any $\boldsymbol{h}$. Therefore, $\max_y p_y - \min_y p_y \ge \frac{\exp(W\|x\|+B)-\exp(-(W\|x\|+B))}{\exp(W\|\boldsymbol{x}\|+B)+(K-1)\exp(-(W\|\boldsymbol{x}\|+B))}$, which leads to a confliction. □

**Lemma E.6.** *For any $\hat{y} \ne y_{max}$, it holds that $\inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) \ge -(p_{max} + p_{\hat{y}})\log(\frac{p_{max}+p_{\hat{y}}}{2}) - \sum_{y\notin\{y_{max},\hat{y}\}} p_y \log(p_y)$, where $y_{max} = \operatorname{argmax}_{y\in\mathcal{Y}} p_y$ and $\hat{y} = \operatorname{argmax}_{y\in\mathcal{Y}} h_y(\boldsymbol{x})$.*

*Proof.* For all $\boldsymbol{h} \in \mathcal{H}$ and $\boldsymbol{x} \in \mathcal{X}$, we have

$$\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) = \sum_{y=1}^K p_y \ell(y,\boldsymbol{h}(\boldsymbol{x})) = \sum_{y=1}^K p_y(-h_y + \log(\sum_{j=1}^K \exp(h_j))),$$

where we use $h_j$ to denote $h_j(\boldsymbol{x})$ for simplicity. To get the $\inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}))$, we consider the following problem

$$\begin{aligned}
&\min_{\boldsymbol{h}\in\mathcal{H}(\boldsymbol{x})} \sum_{y=1}^{K} p_y(-h_y + \log(\sum_{j=1}^{K} \exp(h_j))), \\
&s.t. \quad \begin{cases} h_i - (W\|\boldsymbol{x}\| + B) \le 0, & \forall i, \\ -h_i - (W\|\boldsymbol{x}\| + B) \le 0, & \forall i, \\ h_i - h_{\hat{y}} \le 0, & \forall i \ne \hat{y}. \end{cases}
\end{aligned}$$

We drop some constraints, and consider another problem, whose optimum is lower than the above:

$$\begin{aligned}
&\min_{\boldsymbol{h}\in\mathcal{H}(\boldsymbol{x})} \sum_{y=1}^{K} p_y(-h_y + \log(\sum_{j=1}^{K} \exp(h_j))), \\
&s.t. \quad h_i - h_{\hat{y}} \le 0, \forall i \ne \hat{y}.
\end{aligned}$$

Due to the convexity of $\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}))$ by Lemma E.3, we could write its KKT conditions to obtain the necessary conditions to reach the optimum. They are listed as follows:

$$\begin{cases}
h_i - h_{\hat{y}} \le 0, & \forall i \ne \hat{y}, \\
\lambda_i^* \ge 0, & \forall i \ne \hat{y}, \\
\lambda_i^*(h_i - h_{\hat{y}}) = 0, & \forall i \ne \hat{y}, \\
-p_i + \dfrac{\exp(h_i^*)}{\sum_{k=1}^{K}\exp(h_k^*)} + \lambda_i^* = 0, & \forall i \ne \hat{y}, \\
-p_{\hat{y}} + \dfrac{\exp(h_{\hat{y}}^*)}{\sum_{k=1}^{K}\exp(h_k^*)} - \sum_{i\ne\hat{y}}\lambda_i^* = 0. &
\end{cases} \tag{13}$$

If $\lambda_i^* = 0$ for all $i \in \{1,\dots,K\}$, then $h_i^* = \log(p_i(\sum_{k=1}^{K}\exp(h_k^*)))$. It means that $h_{y_{max}}^* \ge h_{\hat{y}}^*$, which conflicts with the precondition that $\hat{y} \ne y_{max}$. Thus, there exists a $y_m \ne \hat{y}$, $\lambda_{y_m}^* = 0$, and $h_{y_m}^* = h_{\hat{y}}$. It implies that

$$\begin{cases}
h_i^* = \log(p_i \sum_{k=1}^{K}\exp(h_k^*)), & i \notin \{y_m, \hat{y}\}, \\
h_{\hat{y}}^* = h_{y_m}^*, & \\
\dfrac{\exp(h_{y_m}^*) + \exp(h_{\hat{y}}^*)}{\sum_{k=1}^{K}\exp(h_k^*)} = p_{y_m} + p_{\hat{y}}. &
\end{cases}$$

Then we have $h_{\hat{y}}^* = h_{y_m}^* = \log(\frac{p_{y_m}+p_{\hat{y}}}{2}\sum_{k=1}^{K}\exp(h_k^*))$. If $y_m \ne y_{max}$, then $h_{\hat{y}}^* = \log(\frac{p_{y_m}+p_{\hat{y}}}{2}\sum_{k=1}^{K}\exp(h_k^*)) < \log(p_{max}\sum_{k=1}^{K}\exp(h_k^*)) = h_{y_{max}}^*$, which conflicts with $\hat{y} \ne y_{max}$. Thus, we conclude that $y_m = y_{max}$. We define

$s = \exp(h^*_{\hat{y}}) = \exp(h^*_{y_{max}})$ for simplicity, then we can obtain that

$$
\begin{aligned}
&\inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) = \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \sum_{y=1}^{K} p_y \ell(\boldsymbol{h}(\boldsymbol{x}), y) \\
&= \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \sum_{y=1}^{K} p_y(-h_y + \log(\sum_{j=1}^{K} \exp{(h_j)})) \\
&\geq p_{y_{max}}(-h^*_{y_{max}} + \log(\sum_{j=1}^{K} \exp{(h^*_j)})) + p_{\hat{y}}(-h^*_{\hat{y}} + \log(\sum_{j=1}^{K} \exp{(h^*_j)})) + \sum_{y\notin\{y_{max},\hat{y}\}} p_y(-h^*_y + \log(\sum_{j=1}^{K} \exp{(h^*_j)})) \\
&= -(p_{y_{max}} + p_{\hat{y}})(\log(s) - \log(\sum_{j=1}^{K} \exp{(h^*_j)})) - \sum_{y\notin\{y_{max},\hat{y}\}} p_y \log(p_y) \\
&= -(p_{y_{max}} + p_{\hat{y}}) \log(\frac{s}{\sum_{j=1}^{K} \exp{(h^*_j)}}) - \sum_{y\notin\{y_{max},\hat{y}\}} p_y \log(p_y) \\
&= -(p_{y_{max}} + p_{\hat{y}}) \log(\frac{\exp{(h^*_{y_{max}})} + \exp{(h^*_{\hat{y}})}}{2\sum_{k=1}^{K} \exp{(h^*_k)}}) - \sum_{y\notin\{y_{max},\hat{y}\}} p_y \log(p_y) \\
&= -(p_{y_{max}} + p_{\hat{y}}) \log(\frac{p_{y_{max}} + p_{\hat{y}}}{2}) - \sum_{y\notin\{y_{max},\hat{y}\}} p_y \log(p_y),
\end{aligned}
$$

which completes the proof. □

**Lemma E.7** (Technical lemma 1)**.** *For all $x \in [0, 1-t]$ and fixed $t \in \mathbb{R}_+$, it holds that $-(2x+t)\log(\frac{2x+t}{2}) + (x+t)\log(x+t) + x\log(x) \geq u(1-t) = -(2-t)\log(\frac{2-t}{2}) + (1-t)\log(1-t)$.*

*Proof.* We first prove that $u(x) = -(2x+t)\log(\frac{2x+t}{2}) + (x+t)\log(x+t) + t\log(t)$ is decreasing on $x \in [0,1]$, which could be obtained by $\frac{du}{dx} = \log(\frac{4x(x+t)}{(2x+t)^2}) \leq 0$. Thus we have $u(x) \geq u(1-t) = -(2-t)\log(\frac{2-t}{2}) + (1-t)\log(1-t)$, which complete the proof. □

**Lemma E.8** (Technical lemma 2)**.** *For all $t \in [0,1]$, it holds that $\frac{1+t}{2}\log(1+t) + \frac{1-t}{2}\log(1-t) \geq \frac{t^2}{2}$.*

*Proof.* We define $u(t) = \frac{1+t}{2}\log(1+t) + \frac{1-t}{2}\log(1-t) - \frac{t^2}{2}$. Then we calculate $\frac{du}{dt} = \frac{1}{2}\log(\frac{1+t}{1-t}) - t$ and $\frac{d^2u}{dt^2} = \frac{1}{1-t^2} - 1 \geq 0$. We have $\frac{du}{dt} = \frac{1}{2}\log(\frac{1+t}{1-t}) - t \geq \frac{1}{2}\log(\frac{1+0}{1-0}) - 0 = 0$. Thus, $u(x)$ is increasing on $[0,1]$ and $u(x) \geq u(0) = 0$, which proves the lemma. □

We now are ready to prove the Theorem 3.3 as follows.

*Proof.* We first rewrite the $\mathcal{J}_\ell(t)$ as follows.

$$
\begin{aligned}
\mathcal{J}_\ell(t) &= \inf_{\hat{y}\in\mathcal{Y},\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t),\boldsymbol{x}\in\mathcal{X},\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) \\
&= \inf_{\hat{y}\in\mathcal{Y}} \inf_{\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t)} \inf_{\boldsymbol{x}\in\mathcal{X}} \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \Delta\mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) \\
&= \inf_{\hat{y}\in\mathcal{Y}} \inf_{\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t)} \inf_{\boldsymbol{x}\in\mathcal{X}} \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} (\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) \\
&= \inf_{\hat{y}\in\mathcal{Y}} \inf_{\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t)} \inf_{\boldsymbol{x}\in\mathcal{X}} (\inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})).
\end{aligned}
$$

For all $\boldsymbol{h} \in \mathcal{H}$ and $\boldsymbol{x} \in \mathcal{X}$, we have

$$
\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) = \sum_{y=1}^{K} p_y \ell(y, \boldsymbol{h}(x)) = \sum_{y=1}^{K} p_y(-h_y + \log(\sum_{j=1}^{K} \exp{(h_j)})).
$$

To get the $\inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}))$, we consider the following problem

$$\begin{aligned} &\min_{\boldsymbol{h}} \sum_{y=1}^{K} p_y(-h_y + \log(\sum_{j=1}^{K}\exp(h_j))), \\ &s.t. \quad \begin{cases} h_i - (W\|\boldsymbol{x}\| + B) \le 0, \forall i, \\ -h_i - (W\|\boldsymbol{x}\| + B) \le 0, \forall i. \end{cases} \end{aligned} \tag{14}$$

By Lemma E.3, we know that this problem is convex, we can make use of KKT conditions (Boyd et al., 2004) to find the points that are primal and dual optimal, which can be listed as follows

$$\begin{cases} h_i^* - (W\|\boldsymbol{x}\| + B) \le 0, & i = 1,\dots,K, \\ -h_i^* - (W\|\boldsymbol{x}\| + B) \le 0, & i = 1,\dots,K, \\ \lambda_i^* \ge 0, \mu_i^* \ge 0, & i = 1,\dots,K, \\ \lambda_i^*(h_i^* - W\|\boldsymbol{x}\| - B) = 0, & i = 1,\dots,K, \\ \mu_i^*(-h_i^* - W\|\boldsymbol{x}\| - B) = 0, & i = 1,\dots,K, \\ -p_i + \dfrac{\exp(h_i^*)}{\sum_{k=1}^K \exp(h_k^*)} + \lambda_i^* - \mu_i^* = 0, & i = 1,\dots,K. \end{cases} \tag{15}$$

It implies that

$$\begin{cases} h_i^* = W\|\boldsymbol{x}\| + B, & p_i \ge \dfrac{\exp(W\|\boldsymbol{x}\| + B)}{\sum_{k=1}^K \exp(h_k^*)}, \\ h_i^* = -(W\|\boldsymbol{x}\| + B), & p_i \le \dfrac{\exp(-(W\|\boldsymbol{x}\| + B))}{\sum_{k=1}^K \exp(h_k^*)}, \\ h_i^* = \log(p_i \sum_{k=1}^{K} \exp(h_k^*)), & \text{otherwise.} \end{cases} \tag{16}$$

By the precondition of Theorem 3.3, we have

$$t = p_{max} - p_{\hat{y}} \le p_{max} - p_{min} \le \frac{\exp(B) - \exp(-B)}{\exp(B) + (K-1)\exp(-B)} \le \frac{\exp(W\|\boldsymbol{x}\| + B) - \exp(-(W\|\boldsymbol{x}\| + B))}{\exp(W\|\boldsymbol{x}\| + B) + (K-1)\exp(-(W\|\boldsymbol{x}\| + B))}. \tag{17}$$

In addition, in this case, by Lemma E.5, the global optimum could be reached, so we can omit the boundary situation

$$\inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) = \sum_{y=1}^{K} p_y(-h_y^* + \log(\sum_{j=1}^{K}\exp(h_j^*))) = -\sum_{y=1}^{K} p_y \log(p_y),$$

which is the entropy of distribution $\boldsymbol{p}$. Denote the index of the largest element of $\boldsymbol{p}$ by $y_{max}$. When $t > 0$, because $\boldsymbol{p}_{y_{max}} - \boldsymbol{p}_{\hat{y}} = t > 0$, then $y_{max} \ne \hat{y}$. By Lemma E.6, we know that

$$\inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) \ge -(p_{y_{max}} + p_{\hat{y}})\log(\frac{p_{y_{max}} + p_{\hat{y}}}{2}) - \sum_{y\notin\{y_{max},\hat{y}\}} p_y \log(p_y).$$

Then we have

$$\inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) \ge -(p_{y_{max}} + p_{\hat{y}})\log(\frac{p_{y_{max}} + p_{\hat{y}}}{2}) + p_{y_{max}}\log(p_{y_{max}}) + p_{\hat{y}}\log(p_{\hat{y}}). \tag{18}$$

To make (17) holds for all $\boldsymbol{x} \in \mathcal{X}$, we need $t \le \min_{\boldsymbol{x}} \frac{\exp(W\|\boldsymbol{x}\|+B)-\exp(-(W\|\boldsymbol{x}\|+B))}{\exp(W\|\boldsymbol{x}\|+B)+(K-1)\exp(-(W\|\boldsymbol{x}\|+B))} = \frac{\exp(B)-\exp(-B)}{\exp(B)+(K-1)\exp(-B)}$. Then, in this case, we can take infimum with regard to $\boldsymbol{x}$ as follows.

$$\inf_{\boldsymbol{x}\in\mathcal{X}}(\inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})}(\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}))) \ge -(p_{y_{max}} + p_{\hat{y}})\log(\frac{p_{y_{max}} + p_{\hat{y}}}{2}) + p_{y_{max}}\log(p_{y_{max}}) + p_{p_{\hat{y}}}\log(p_{p_{\hat{y}}}).$$

Now, we meet the following problem

$$\min_{\boldsymbol{p}} \; -(p_{y_{max}} + p_{\hat{y}}) \log(\frac{p_{y_{max}} + p_{\hat{y}}}{2}) + p_{y_{max}} \log(p_{y_{max}}) + p_{\hat{y}} \log(p_{\hat{y}}),$$
$$s.t. \quad \begin{cases} p_{y_{max}} - p_{\hat{y}} = t, \forall i. \\ \sum_{i=1}^{K} p_i = 1, \\ p_i \geq 0, \forall i, \end{cases}$$

which is equivalent to find the minimum of $-(2p_{\hat{y}} + t)\log(\frac{2p_{\hat{y}}+t}{2}) + (p_{\hat{y}} + t)\log((p_{\hat{y}} + t)) + p_{\hat{y}}\log(p_{\hat{y}})$ when $p_{\hat{y}} \in [0, \frac{1-t}{2}]$. By Lemma E.7, we know it is $\frac{1+t}{2}\log(1+t) + \frac{1-t}{2}\log(1-t)$. Thus

$$\begin{aligned} \mathcal{J}_\ell(t) &= \inf_{\hat{y}\in\mathcal{Y}} \inf_{\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t)} \inf_{\boldsymbol{x}\in\mathcal{X}} \big( \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) \big) \\ &\geq \inf_{\hat{y}\in\mathcal{Y}} -(2-t)\log(\frac{2-t}{2}) + (1-t)\log(1-t) \\ &= \frac{1+t}{2}\log(1+t) + \frac{1-t}{2}\log(1-t) \\ &\geq \frac{t^2}{2}. && \text{(Lemma E.8)} \end{aligned}$$

It is worthwhile to note that when the number of classes is 2, then the derivations and results above coincide with that in binary case (Awasthi et al., 2022a). Let $g(t) = \frac{t^2}{2}$ in Theorem 3.5, we have

$$\frac{1}{2}(R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}})^2 \leq R_{\ell_{log}}(\boldsymbol{h}) - R^*_{\ell_{log},\mathcal{H}} + M_{\ell_{log},\mathcal{H}},$$

which implies

$$R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}} \leq \sqrt{2}(R_{\ell_{log}}(\boldsymbol{h}) - R^*_{\ell_{log},\mathcal{H}} + M_{\ell_{log},\mathcal{H}})^{\frac{1}{2}},$$

when $R_{\ell_{log}}(\boldsymbol{h}) - R^*_{\ell_{log},\mathcal{H}} + M_{\ell_{log},\mathcal{H}} \leq \frac{1}{2}(\frac{\exp(2B)-1}{\exp(2B)+K-1})^2$. By Lemma E.4, we have $M_{\ell_{0-1},\mathcal{H}}$ coincides with the approximation error $R^*_{\ell_{0-1},\mathcal{H}} - R^*_{\ell_{0-1},\mathcal{H}_{all}}$. We also note that $M_{\ell_{log},\mathcal{H}}$ coincides with $R^*_{\ell_{log},\mathcal{H}} - R^*_{\ell_{log},\mathcal{H}_{all}}$ because

$$\begin{aligned} M_{\ell_{log},\mathcal{H}} &= R^*_{\ell_{log},\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[\mathcal{C}^*_{\ell_{log},\mathcal{H}}(\boldsymbol{x})] \\ &= R^*_{\ell_{log},\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[\inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_{\ell_{log}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}(\boldsymbol{x}))] \\ &= R^*_{\ell_{log},\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[-\sum_{y=1}^{K} p_y(\boldsymbol{x})\log(p_y(\boldsymbol{x}))] \\ &= R^*_{\ell_{log},\mathcal{H}} - R^*_{\ell_{log},\mathcal{H}_{all}}. \end{aligned}$$

Finally, we can conclude that

$$R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} \leq R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}} \leq \sqrt{2}(R_{\ell_{log}}(\boldsymbol{h}) - R^*_{\ell_{log},\mathcal{H}_{all}})^{\frac{1}{2}}.$$

□

## F. Proofs of Appendix B

### F.1. Proof of Proposition B.1

We first present the following lemmas to show Proposition B.1.

**Lemma F.1** ((Mohri et al., 2018), Lemma 5.7, Talagrand's lemma)**.** *Let $\Phi$ be L-Lipschitz functions from $\mathbb{R} \to \mathbb{R}$ and $\sigma_1, \ldots, \sigma_m$ be Rademacher random variables. Then, for any hypothesis set $\mathcal{H}$ of real-valued functions, the following inequality holds:*

$$\mathcal{R}_m(\Phi \circ \mathcal{H}) \leq L\mathcal{R}_m(\mathcal{H}).$$

**Lemma F.2** (Rademacher complexity of constrained linear hypotheses)**.** *Let $S = \{x_1, \dots, x_m\}$ where $x_i \in [0,1]$ for all $i \in \{1, \dots, m\}$ and $\mathcal{H} = \{x \to \langle w, x \rangle + b : \|w\|_2 \leq W, |b| \leq B\}$. Then, the Rademacher complexity of $\mathcal{H}$ can be bounded as follows:*

$$\mathcal{R}_m(\mathcal{H}) \leq W\sqrt{\frac{n}{m}}.$$

*Proof.*

$$\begin{aligned}
\mathcal{R}_m(\mathcal{H}) &= \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_h \sum_{i=1}^m \sigma_i h(x_i)] = \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_h \sum_{i=1}^m \sigma_i(\langle w, x_i \rangle + b)] \\
&= \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_h \sum_{i=1}^m \sigma_i \langle w, x_i \rangle + b\sum_{i=1}^m \sigma_i] \leq \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_w \sum_{i=1}^m \sigma_i \langle w, x_i \rangle + \sup_b b\sum_{i=1}^m \sigma_i] \\
&= \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_w \sum_{i=1}^m \sigma_i \langle w, x_i \rangle] = \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_w \langle w, \sum_{i=1}^m \sigma_i x_i \rangle] \\
&\leq \frac{W}{m}\mathbb{E}_{S,\sigma}[\| \sum_{i=1}^m \sigma_i x_i \|_2] \leq \frac{W}{m}\sqrt{\mathbb{E}_{S,\sigma}[\| \sum_{i=1}^m \sigma_i x_i \|_2^2]} \\
&= \frac{W}{m}\sqrt{\mathbb{E}_{S,\sigma}[\sum_{i,j=1}^m \sigma_i \sigma_j \langle x_i, x_j \rangle]} = \frac{W}{m}\sqrt{\sum_{i=1}^m \|x_i\|_2^2} \leq \frac{W}{m}\sqrt{m \times n} \\
&= W\sqrt{\frac{n}{m}}.
\end{aligned}$$

□

**Lemma F.3** (Rademacher complexity of $\widetilde{\mathcal{H}}$)**.** *Let $\widetilde{\mathcal{H}} = \{z = (\boldsymbol{x}, y) \to yh(x) : h \in \mathcal{H}\}$. Then, the Rademacher complexity of $\widetilde{\mathcal{H}}$ satisfies:*

$$\mathcal{R}_m(\widetilde{\mathcal{H}}) = \mathcal{R}_m(\mathcal{H}).$$

*Proof.*

$$\begin{aligned}
\mathcal{R}_m(\widetilde{\mathcal{H}}) &= \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_h \sum_{i=1}^m \sigma_i y_i h(x_i)] \\
&= \frac{1}{2m}\mathbb{E}_{S,\sigma}[\sup_h \sum_{i=1}^m \sigma_i (2y_i - 1) h(x_i) + \sum_{i=1}^m \sigma_i h(x_i)] \\
&= \frac{1}{2m}\mathbb{E}_{S,\sigma}[\sup_h \sum_{i=1}^m \sigma_i h(x_i) + \sum_{i=1}^m \sigma_i h(x_i)] && (2y_i - 1 \in \{-1, +1\}) \\
&= \frac{1}{m}\mathbb{E}_{S,\sigma}[\sup_h \sum_{i=1}^m \sigma_i h(x_i)] \\
&= \mathcal{R}_m(\mathcal{H}).
\end{aligned}$$

□

We now prove Proposition B.1 by using the above lemmas.

*Proof.* We first rewrite the $R_{\ell_{log}}(h_{Dis,m}) - R_{\ell_{log}}(h_{Dis,\infty})$.

$$\begin{aligned}
&R_{\ell_{log}}(h_{Dis,m}) - R_{\ell_{log}}(h_{Dis,\infty}) \\
&= R_{\ell_{log}}(h_{Dis,m}) - \hat{R}_{\ell_{log},S}(h_{Dis,m}) + \hat{R}_{\ell_{log},S}(h_{Dis,m}) - \hat{R}_{\ell_{log},S}(h_{Dis,\infty}) + \hat{R}_{\ell_{log},S}(h_{Dis,\infty}) - R_{\ell_{log}}(h_{Dis,\infty}) \\
&\leq (R_{\ell_{log}}(h_{Dis,m}) - \hat{R}_{\ell_{log},S}(h_{Dis,m})) + (\hat{R}_{\ell_{log},S}(h_{Dis,\infty}) - R_{\ell_{log}}(h_{Dis,\infty}).
\end{aligned}$$

The first summand on the right-hand side can be bounded by making use of Lemma D.4,F.1,F.2 and F.3 in sequence. Let $\widetilde{\mathcal{H}} = \{z = (\boldsymbol{x}, y) \to yh(x) : h \in \mathcal{H}\}$ and $\Phi = \{\ell_{log} \circ \widetilde{h} : \widetilde{h} \in \widetilde{\mathcal{H}}\}$ With probability of at least $1 - \delta$, we have:

$$\begin{aligned}
&R_{\ell_{log}}(h_{Dis,m}) - \hat{R}_{\ell_{log},S}(h_{Dis,m}) \\
&\leq 2\mathcal{R}_m(\ell_{log} \circ \widetilde{\mathcal{H}}) + \log(1 + \exp{(W\sqrt{n} + B)})\sqrt{\frac{1}{2m}\log(\frac{2}{\delta})} && (\ell_{log} \circ \widetilde{\mathcal{H}} \text{ is bounded, Lemma D.4}) \\
&\leq 2\mathcal{R}_m(\widetilde{\mathcal{H}}) + \log(1 + \exp{(W\sqrt{n} + B)})\sqrt{\frac{1}{2m}\log(\frac{2}{\delta})} && (\ell_{log} \text{ is 1-Lipschitz, Lemma F.1}) \\
&= 2\mathcal{R}_m(\mathcal{H}) + \log(1 + \exp{(W\sqrt{n} + B)})\sqrt{\frac{1}{2m}\log(\frac{2}{\delta})} && (\text{by Lemma F.3}) \\
&\leq 2W\sqrt{\frac{n}{m}} + \log(1 + \exp{(W\sqrt{n} + B)})\sqrt{\frac{1}{2m}\log(\frac{2}{\delta})} && (\text{by Lemma F.2}).
\end{aligned}$$

For the second summand, we use the fact that $R_{\ell_{log}}(h_{Dis,\infty})$ does not depend on sampled training dataset $S$; hence by Lemma D.9, we obtain its bound:

$$\mathbb{P}(|\hat{R}_{\ell_{log},S}(h_{Dis,\infty}) - R_{\ell_{log}}(h_{Dis,\infty})| > \epsilon) \leq 2\exp(-\frac{2m\epsilon^2}{(c-0)^2}) = 2\exp(-\frac{2m\epsilon^2}{c^2}),$$

where $c = \log(1 + \exp{(W\sqrt{n} + B)})$. It implies that with the probability of at least $1 - \delta$, we have:

$$\hat{R}_{\ell_{log},S}(h_{Dis,\infty}) - R_{\ell_{log}}(h_{Dis,\infty}) \leq c\sqrt{\frac{1}{2m}\log(\frac{2}{\delta})}.$$

At last, we use the union bound to get the final result. With probability at least $1 - \delta$, the following holds:

$$\begin{aligned}
&R_{\ell_{log}}(h_{Dis,m}) - R_{\ell_{log}}(h_{Dis,\infty}) \\
&\leq 2W\sqrt{\frac{n}{m}} + \log(1 + \exp{(W\sqrt{n} + B)})\sqrt{\frac{1}{2m}\log(\frac{4}{\delta})} + c\sqrt{\frac{1}{2m}\log(\frac{4}{\delta})} \\
&= 2W\sqrt{\frac{n}{m}} + (c + \log(1 + \exp{(W\sqrt{n} + B)}))\sqrt{\frac{1}{2m}\log(\frac{4}{\delta})} \\
&= 2W\sqrt{\frac{n}{m}} + +2\log(1 + \exp{(W\sqrt{n} + B)})\sqrt{\frac{1}{2m}\log(\frac{4}{\delta})} \\
&= O(\sqrt{\frac{n}{m}}).
\end{aligned}$$

Therefore, for $R_{\ell_{log}}(h_{Dis,m}) \leq R_{\ell_{log}}(h_{Dis,\infty}) + \epsilon_0$ to hold with high probability $1 - \delta_0$ (here, $\epsilon_0$ and $\delta_0$ are some fixed constant in $[0, 1]$), it suffices to pick $m = O(n)$ samples. □

### F.2. Proof of Theorem B.1

*Proof.* By Theorem 2.1 we know that for $R_{\ell_{0-1}}(h_{Dis,m}) \leq R_{\ell_{log}}(h_{Dis,\infty}) + \epsilon_0$, it is sufficient to ensure that $R_{log}(h_{Dis,m}) \leq R_{\ell_{log}}(h_{Dis,\infty}) + \frac{1}{2}\epsilon_0^2$. Then by Proposition B.1, it suffices to sample $m = O(\frac{n}{\epsilon_0^4}) = O(n)$. □

### F.3. Proof of Theorem B.2

To show Theorem B.2, we first present the following lemmas.

**Lemma F.4.** *In terms of binary naïve Bayes, let any $\epsilon, \delta > 0$ and any Laplace smoothing parameter $\alpha \geq 0$ be fixed. Assume that Assumption 3.1 holds. Let $m = O((\frac{1}{\epsilon^2})log(\frac{n}{\delta}))$, then with the probability of at least $1 - \delta$:*

1. *In case of discrete inputs, $|\hat{p}(x_i|y = k) - p(x_i|y = k)| \leq \epsilon$ and $|\hat{p}(y = k) - p(y = k)| \leq \epsilon$ for all $i \in \{1, \ldots n\}$ and $k \in \{0, 1\}$.*

2. *In case of continuous inputs,* $|\hat{\mu}_{ki} - \mu_{ki}| \le \epsilon$, $|\hat{\sigma}_i^2 - \sigma_i^2| \le \epsilon$ *and* $|\hat{p}(y = k) - p(y = k)| \le \epsilon$ *for all* $i \in \{1, \dots n\}$ *and* $k \in \{0, 1\}$.

*Proof.* First, we consider the discrete case, and let $\alpha = 0$ for now. Let $\epsilon \le \rho_0/2$. By the Lemma D.9, with probability at least $1 - \delta_1 = 1 - 2\exp(-2m\epsilon^2)$ we have $|\hat{p}(y = k) - p(y = k)| \le \epsilon$. It implies that $\hat{p}(y = k) \ge p(y = k) - \epsilon \ge \rho_0 - \epsilon = \gamma = \Omega(1)$. So $\#\{y = k\} \ge \gamma m$ with probability at least $1 - \delta_1$.To bound the $|\hat{p}(x_i|y = k) - p(x_i|y = k)|$, for fixed $i, k$, the following holds:

$$\begin{aligned}
&\mathbb{P}[|\hat{p}(x_i|y = k) - p(x_i|y = k)| > \epsilon] \\
&= \mathbb{P}(|\hat{p}(x_i|y = k) - p(x_i|y = k)| > \epsilon | \#\{y = k\} \ge \gamma m)\mathbb{P}(\#\{y = k\} \ge \gamma m) \\
&+ \mathbb{P}(|\hat{p}(x_i|y = k) - p(x_i|y = k)| > \epsilon | \#\{y = k\} < \gamma m)\mathbb{P}(\#\{y = k\} < \gamma m) \\
&\le 2\exp(-2\epsilon^2 \#\{y = k\})|_{\#\{y=k\} \ge \gamma m} + \delta_1 \\
&\le 2\exp(-2\epsilon^2 \gamma m) + \delta_1 = \delta_2.
\end{aligned}$$

Then we use the union bound to get the first result on the condition that $\alpha = 0$:

$$\begin{aligned}
&\mathbb{P}(\cup_{k=0}^1 (|\hat{p}(y = k) - p(y = k)| > \epsilon) \cup (\cup_{i=1}^n \cup_{k=0}^1 |\hat{p}(x_i|y = k) - p(x_i|y = k)| > \epsilon)) \\
&= \mathbb{P}((|\hat{p}(y = 1) - p(y = 1)| > \epsilon) \cup (\cup_{i=1}^n \cup_{k=0}^1 |\hat{p}(x_i|y = k) - p(x_i|y = k)| > \epsilon)) \\
&\le \delta_1 + 2n\delta_2 \\
&= 2\exp(-2m\epsilon^2) + 2n(2\exp(-2\epsilon^2\gamma m) + \delta_1) \\
&= (2n + 1)2\exp(-2m\epsilon^2) + 2n \times 2\exp(-2\epsilon^2\gamma m) \\
&\le 2(4n + 1)\exp(-2\gamma m\epsilon^2).
\end{aligned}$$

Therefore, for Lemma F.4.1 to hold with probability at least $1 - \delta$, it suffices to pick $m$ samples that

$$m = \frac{1}{2\gamma\epsilon^2}\log(\frac{2(4n + 1)}{\delta}) \le \frac{1}{\rho_0\epsilon^2}\log(\frac{2(4n + 1)}{\delta}) = O(\frac{1}{\epsilon^2}\log(\frac{n}{\delta})).$$

Second, we consider the discrete case, and let $\alpha > 0$. To bound $|\hat{p}(y = k) - p(y = k)|$, we calculate it based on the above condition as follows:

$$\begin{aligned}
&\mathbb{P}(|\hat{p}(y = k) - p(y = k)| > \epsilon) \\
&= \mathbb{P}(|\hat{p}(y = k) - \hat{p}(y = k)|_{\alpha=0} + \hat{p}(y = k)|_{\alpha=0} - p(y = k)| > \epsilon) \\
&\le \mathbb{P}(|\hat{p}(y = k) - \hat{p}(y = k)|_{\alpha=0}| + |\hat{p}(y = k)|_{\alpha=0} - p(y = k)| > \epsilon),
\end{aligned}$$

where the $|\hat{p}(y = k)|_{\alpha=0} - p(y = k)|$ has been discussed above, so we only need to bound $|\hat{p}(y = k) - \hat{p}(y = k)|_{\alpha=0}|$. We have,

$$|\hat{p}(y = k) - \hat{p}(y = k)|_{\alpha=0}| = |\frac{\#\{y = k\} + \alpha}{m + 2\alpha} - \frac{\#\{y = k\}}{m}| = |\frac{\alpha(m - \#\{y = k\})}{m(m + 2\alpha)}| = O(\frac{1}{m}).$$

So

$$\begin{aligned}
\mathbb{P}(|\hat{p}(y = k) - p(y = k)| > \epsilon) &\le \mathbb{P}(|\hat{p}(y = k)|_{\alpha=0} - p(y = k)| > \epsilon - O(\frac{1}{m})) \\
&\le 2\exp(-2m(\epsilon - O(\frac{1}{m}))^2) = \delta_1.
\end{aligned}$$

In the same way, we can write

$$|\hat{p}(x_i|y = k) - \hat{p}(x_i|y = k)|_{\alpha=0}| = \frac{\alpha(\#\{y = k\} - |\#\{x_i, y = k\})}{\#\{y = k\}(\#\{y = k\} + 2\alpha)}| = O(\frac{1}{\#\{y = k\}}) = O(\frac{1}{m}),$$

and

$$\begin{aligned}
\mathbb{P}(|\hat{p}(x_i|y = k) - p(x_i|y = k)| > \epsilon) &\le \mathbb{P}(|\hat{p}(x_i|y = k)|_{\alpha=0} - p(x_i|y = k)| > \epsilon - O(\frac{1}{m})) \\
&\le \delta_1 + 2\exp(-2\gamma m(\epsilon - O(\frac{1}{m}))^2).
\end{aligned}$$

Besides,

$$m = \frac{1}{2\gamma(\epsilon - O(\frac{1}{m}))^2}\log(\frac{2(4n+1)}{\delta}) \le \frac{1}{\rho_0(\epsilon - O(\frac{1}{m}))^2}\log(\frac{2(4n+1)}{\delta}) = O(\frac{1}{\epsilon^2}\log(\frac{n}{\delta})).$$

In the following proofs, we will not consider Laplace smoothing anymore due to its small influence on the results.

Third, we consider the continuous case. In the same way as discrete case, with probability at least $1-\delta_1 = 1-2\exp(-2m\epsilon^2)$ we have $|\hat{p}(y=k) - p(y=k)| \le \epsilon$, and $\#\{y=k\} \ge \gamma m$. We only need to bound $|\hat{\mu}_{ki} - \mu_{ki}|$ and $|\hat{\sigma}_i^2 - \sigma_i^2|$. Fix $i, k$, the following holds:

$$\begin{aligned}\mathbb{P}[|\hat{\mu}_{ki} - \mu_{ki}| > \epsilon] &= \mathbb{P}(|\hat{\mu}_{ki} - \mu_{ki}| > \epsilon | \#\{y=k\} \ge \gamma m)\mathbb{P}(\#\{y=k\} \ge \gamma m) \\ &\quad + \mathbb{P}(|\hat{\mu}_{ki} - \mu_{ki}| > \epsilon | \#\{y=k\} < \gamma m)\mathbb{P}(\#\{y=k\} < \gamma m) \\ &\le 2\exp(-2\epsilon^2\#\{y=k\})|_{\#\{y=k\}\ge\gamma m} + \delta_1 \\ &\le 2\exp(-2\epsilon^2\gamma m) + \delta_1 = \delta_2,\end{aligned}$$

where the first inequality use the fact that $x_i \in [0,1]$. For $|\hat{\sigma}_i^2 - \sigma_i^2|$, because $(x_i|_{y=k} - \mu_{ki})^2 \in [0,1]$, by Lemma D.9, we can write:

$$\mathbb{P}[|\hat{\sigma}_i^2 - \sigma_i^2| > \epsilon] \le 2\exp(-2m\epsilon^2) = \delta_3.$$

Finally, we use the union bound to get the result for the continuous case:

$$\begin{aligned}&\mathbb{P}((|\hat{p}(y=k) - p(y=k)| > \epsilon) \cup (\cup_{i=1}^n (|\hat{\sigma}_i^2 - \sigma_i^2| > \epsilon) \cup (\cup_{k=1}^2 |\hat{\mu}_{ki} - \mu_{ki}| > \epsilon))) \\ &\le \delta_1 + n(2\delta_2 + \delta_3) \\ &= (3n+1)2\exp(-2m\epsilon^2) + 2n \times 2\exp(-2\epsilon^2\gamma m) \\ &\le 2(5n+1)\exp(-2\epsilon^2\gamma m).\end{aligned}$$

Thus, for Lemma F.4.2 to hold with probability at least $1-\delta$, it suffices to pick m samples which satisfies

$$m = \frac{1}{2\gamma\epsilon^2}\log(\frac{2(5n+1)}{\delta}) \le \frac{1}{\rho_0\epsilon^2}\log(\frac{2(5n+1)}{\delta}) = O(\frac{1}{\epsilon^2}\log(\frac{n}{\delta})).$$

The proposition's proof is complete. □

**Lemma F.5.** *In case of discrete inputs, and suppose that Assumption 3.2 holds, then with probability at least $1-\delta$, the following holds*

$$|\Delta a_{Gen}(\boldsymbol{x},1,0) - \Delta a_{Gen,\infty}(\boldsymbol{x},1,0)| \le \frac{4(n+1)}{\rho_0}\sqrt{\frac{1}{\rho_0 m}\log(\frac{2(4n+1)}{\delta})} = O\big(n\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}\big).$$

*Proof.* By the derivation of Lemma F.4, let $\epsilon < \rho_0/2$, then with probability at least $1-\delta = 1-2(4n+1)\exp(-2\gamma m\epsilon^2)$, where $\gamma = \rho_0 - \epsilon$ the following holds:

$$\begin{aligned}&|\Delta a_{Gen}(\boldsymbol{x},1,0) - \Delta a_{Gen,\infty}(\boldsymbol{x},1,0)| \\ &= |\sum_{i=1}^n \log\frac{\hat{p}(x_i|y=1)p(x_i|y=0)}{\hat{p}(x_i|y=0)p(x_i|y=1)} + \log\frac{\hat{p}(y=1)p(y=0)}{\hat{p}(y=0)p(y=1)}| \\ &= |\sum_{i=1}^n(\log\hat{p}(x_i|y=1) - \log p(x_i|y=1)) + \sum_{i=1}^n(\log p(x_i|y=0) - \log\hat{p}(x_i|y=0)) \\ &\quad + \log\hat{p}(y=1) - \log p(y=1) + \log p(y=0) - \log\hat{p}(y=0)| \\ &\le \sum_{i=1}^n|\log\hat{p}(x_i|y=1) - \log p(x_i|y=1)| + \sum_{i=1}^n|\log p(x_i|y=0) - \log\hat{p}(x_i|y=0)| \\ &\quad + |\log\hat{p}(y=1) - \log p(y=1)| + |\log p(y=0) - \log\hat{p}(y=0)| \\ &\le \frac{1}{\gamma}\big(\sum_{i=1}^n\epsilon + \sum_{i=1}^n\epsilon + \epsilon + \epsilon\big) \le \frac{4(n+1)}{\rho_0}\epsilon.\end{aligned}$$

The penultimate inequality makes use of Lemma F.4 and the concavity of $\log()$ together. Replace $\epsilon$ with the expressions with respect to $\delta$, we can write:

$$
\begin{aligned}
|\Delta a_{Gen}(\boldsymbol{x},1,0) - \Delta a_{Gen,\infty}(\boldsymbol{x},1,0)| &\leq \frac{4(n+1)}{\rho_0}\sqrt{\frac{1}{2\gamma m}\log(\frac{2(4n+1)}{\delta})} \\
&\leq \frac{4(n+1)}{\rho_0}\sqrt{\frac{1}{\rho_0 m}\log(\frac{2(4n+1)}{\delta})} = O\big(n\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}\big).
\end{aligned}
$$

□

**Lemma F.6.** *Let $\epsilon < \rho_0/2$, assume that Assumption 3.2 holds, $|\hat{\sigma}_i^2 - \sigma_i^2| \leq \epsilon$ and $|\hat{\mu}_{ki} - \mu_{ki}| \leq \epsilon$ for all $i, k$. Then we have:*

$$
|\sigma_i\hat{\mu}_{ki} - \hat{\sigma}_i\mu_{ki}| \leq (1 + \frac{2}{3\rho_0})\epsilon.
$$

*Proof.* On the one hand, we can write:

$$
\sigma_i\hat{\mu}_{ki} - \hat{\sigma}_i\mu_{ki} \leq \sigma_i(\mu_{ki} + \epsilon) - \hat{\sigma}_i\mu_{ki} = (\sigma_i - \hat{\sigma}_i)\mu_{ki} + \epsilon\sigma_i.
$$

On the other hand, we have:

$$
\sigma_i\hat{\mu}_{ki} - \hat{\sigma}_i\mu_{ki} \geq \sigma_i(\mu_{ki} - \epsilon) - \hat{\sigma}_i\mu_{ki} = (\sigma_i - \hat{\sigma}_i)\mu_{ki} - \epsilon\sigma_i.
$$

We conclude that:

$$
\begin{aligned}
|\sigma_i\hat{\mu}_{ki} - \hat{\sigma}_i\mu_{ki}| &\leq |(\sigma_i - \hat{\sigma}_i)\mu_{ki}| + |\epsilon\sigma_i| \leq |\sigma_i - \hat{\sigma}_i| + \epsilon \\
&= |\frac{\sigma_i^2 - \hat{\sigma}_i^2}{\sigma_i + \hat{\sigma}_i}| + \epsilon \leq \frac{\epsilon}{\rho_0 + \rho_0 - \epsilon} + \epsilon \leq (1 + \frac{2}{3\rho_0})\epsilon.
\end{aligned}
$$

□

**Lemma F.7.** *In case of continuous inputs, and suppose that Assumption 3.2 holds, then with probability at least $1 - \delta$, the following holds:*

$$
|\Delta a_{Gen}(\boldsymbol{x},1,0) - \Delta a_{Gen,\infty}(\boldsymbol{x},1,0)| \leq 4(\frac{n}{3\rho_0}(\frac{4}{\rho_0^2} + \frac{3}{\rho_0} + \sqrt{\frac{2}{\rho_0}}) + \frac{1}{\rho_0})\sqrt{\frac{1}{\rho_0 m}\log(\frac{2(5n+1)}{\delta})} = O\big(n\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}\big).
$$

*Proof.* The following holds:

$$
\begin{aligned}
&|\Delta a_{Gen}(\boldsymbol{x},1,0) - \Delta a_{Gen,\infty}(\boldsymbol{x},1,0)| \\
&= |\sum_{i=1}^{n}\log\frac{\hat{p}(x_i|y=1)p(x_i|y=0)}{\hat{p}(x_i|y=0)p(x_i|y=1)} + \log\frac{\hat{p}(y=1)p(y=0)}{\hat{p}(y=0)p(y=1)}| \\
&= |\sum_{i=1}^{n}(\log\hat{p}(x_i|y=1) - \log p(x_i|y=1)) + \sum_{i=1}^{n}(\log p(x_i|y=0) - \log\hat{p}(x_i|y=0)) \\
&\quad + \log\hat{p}(y=1) - \log p(y=1) + \log p(y=0) - \log\hat{p}(y=0)| \\
&\leq \sum_{i=1}^{n}\sum_{k=0}^{1}|\log\hat{p}(x_i|y=k) - \log p(x_i|y=k)| + \sum_{k=0}^{1}|\log\hat{p}(y=k) - \log p(y=k)|.
\end{aligned}
$$

To bound $|\log\hat{p}(x_i|y=k) - \log p(x_i|y=k)|$, let $\epsilon < \rho_0/2$, then by Lemma F.4, with probability at least $1 - \delta = 1 - 2(5n +$

$1)\exp(-2\epsilon^2\gamma m)$, where $\gamma = \rho_0 - \epsilon$, we can write:

$$\begin{aligned}
&|\log \hat{p}(x_i|y=k) - \log p(x_i|y=k)| \\
&= |\log(\sigma_i) - \log(\hat{\sigma}_i) + \frac{1}{2\hat{\sigma}_i^2\sigma_i^2}(\hat{\sigma}_i^2(x_i-\mu_{ki})^2 - \sigma_i^2(x_i-\hat{\mu}_{ki})^2)| \\
&\leq |\log(\sigma_i) - \log(\hat{\sigma}_i)| + \frac{1}{2\hat{\sigma}_i^2\sigma_i^2}|\hat{\sigma}_i^2(x_i-\mu_{ki})^2 - \sigma_i^2(x_i-\hat{\mu}_{ki})^2| \\
&\leq \frac{1}{\min(\sigma_i,\hat{\sigma}_i)}|\sigma_i - \hat{\sigma}_i| + \frac{1}{2\rho_0\hat{\sigma}_i^2}|\hat{\sigma}_i(x_i-\mu_{ki}) + \sigma_i(x_i-\hat{\mu}_{ki})||\hat{\sigma}_i(x_i-\mu_{ki}) - \sigma_i(x_i-\hat{\mu}_{ki})| \\
&\leq \frac{1}{\min(\sigma_i,\hat{\sigma}_i)}|\sigma_i - \hat{\sigma}_i| + \frac{1}{\rho_0\hat{\sigma}_i^2}|\hat{\sigma}_i(x_i-\mu_{ki}) - \sigma_i(x_i-\hat{\mu}_{ki})| \\
&\leq \frac{1}{\min(\sigma_i,\hat{\sigma}_i)}|\sigma_i - \hat{\sigma}_i| + \frac{1}{\rho_0\hat{\sigma}_i^2}(|\hat{\sigma}_i - \sigma_i| + |\sigma_i\hat{\mu}_{ki} - \hat{\sigma}_i\mu_{ki}|) \\
&\leq \frac{1}{\sqrt{\gamma}}\frac{2\epsilon}{3\rho_0} + \frac{1}{\rho_0\gamma}(\frac{2\epsilon}{3\rho_0} + (1+\frac{2}{3\rho_0})\epsilon) \\
&\leq \sqrt{\frac{2}{\rho_0}}\frac{2\epsilon}{3\rho_0} + \frac{2}{\rho_0^2}(\frac{2\epsilon}{3\rho_0} + (1+\frac{2}{3\rho_0})\epsilon) = \frac{2}{3\rho_0}(\frac{4}{\rho_0^2} + \frac{3}{\rho_0} + \sqrt{\frac{2}{\rho_0}})\epsilon.
\end{aligned}$$

The last two inequalities make use of Lemma F.4 the concavity of $\log()$ together. At the same time, we have:

$$|\log \hat{p}(y=k) - \log p(y=k)| \leq \frac{1}{\gamma}|\hat{p}(y=k) - p(y=k)| \leq \frac{2}{\rho_0}|\hat{p}(y=k) - p(y=k)| \leq \frac{2}{\rho_0}\epsilon.$$

At last, combining the above findings and replace $\epsilon$ with the expressions with respect to $\delta$, we can get:

$$\begin{aligned}
|\Delta a_{Gen}(\boldsymbol{x},1,0) - \Delta a_{Gen,\infty}(\boldsymbol{x},1,0)| &\leq 2n\frac{2}{3\rho_0}(\frac{4}{\rho_0^2} + \frac{3}{\rho_0} + \sqrt{\frac{2}{\rho_0}})\epsilon + 2\frac{2}{\rho_0}\epsilon \\
&= 4(\frac{n}{3\rho_0}(\frac{4}{\rho_0^2} + \frac{3}{\rho_0} + \sqrt{\frac{2}{\rho_0}}) + \frac{1}{\rho_0})\sqrt{\frac{1}{\rho_0 m}\log(\frac{2(5n+1)}{\delta})} \\
&= O\big(n\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}\big).
\end{aligned}$$

□

Now, we are ready to prove Theorem B.2.

*Proof.* Let $\epsilon = O\big(n\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}\big)$ which are claimed in the Lemma F.5 for discrete case and Lemma F.7 for continuous case. Then we simplify the $|R_{\ell_{0-1}}(h_{Gen,m}) - R_{\ell_{0-1}}(h_{Gen,\infty})|$ as follows:

$$\begin{aligned}
&|R_{\ell_{0-1}}(h_{Gen,m}) - R_{\ell_{0-1}}(h_{Gen,\infty})| \\
&= |\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell_{0-1}(h_{Gen,m},(\boldsymbol{x},y)) - \ell_{0-1}(h_{Gen,\infty},(\boldsymbol{x},y))]| \\
&\leq \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}|\ell_{0-1}(h_{Gen,m},(x,y)) - \ell_{0-1}(h_{Gen,\infty},(\boldsymbol{x},y))| \\
&= \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}}(h_{Gen,m}(x) \neq h_{Gen,\infty}(x)) \\
&= \big(\mathbb{P}(h_{Gen,m}(x) \neq h_{Gen,\infty}(x)||\Delta a_{Gen}(\boldsymbol{x},1,0) - a_{Gen,\infty}(x)| \leq \epsilon)\mathbb{P}(|\Delta a_{Gen}(\boldsymbol{x},1,0) - a_{Gen,\infty}(x)| \leq \epsilon) \\
&\quad + \mathbb{P}(h_{Gen,m}(x) \neq h_{Gen,\infty}(x)||\Delta a_{Gen}(\boldsymbol{x},1,0) - a_{Gen,\infty}(x)| > \epsilon)\mathbb{P}(|\Delta a_{Gen}(\boldsymbol{x},1,0) - a_{Gen,\infty}(x)| > \epsilon)\big) \\
&\leq \mathbb{P}(h_{Gen,m}(x) \neq h_{Gen,\infty}(x)||\Delta a_{Gen}(\boldsymbol{x},1,0) - a_{Gen,\infty}(x)| \leq \epsilon) + \delta \\
&\leq \mathbb{P}(|a_{Gen,\infty}(x)| \leq O\big(n\sqrt{\frac{1}{m}\log(\frac{n}{\delta})}\big)) + \delta \\
&= G\big(O(\sqrt{\frac{1}{m}\log(\frac{n}{\delta})})\big) + \delta.
\end{aligned}$$

□

### F.4. Proof of Proposition B.2

**Lemma F.8.** *Suppose that Assumption B.1 holds, then* $\mathbb{E}[\Delta a_{Gen}(\boldsymbol{x},1,0)|y=1]=\Omega(n)$*, and* $\mathbb{E}[-\Delta a_{Gen}(\boldsymbol{x},1,0)|y=0]=\Omega(n)$*.*

*Proof.* We calculate $\mathbb{E}[\Delta a_{Gen}(\boldsymbol{x},1,0)|y=1]$ straightly:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\Delta a_{Gen}(\boldsymbol{x},1,0)|y=1] &= \mathbb{E}_{\boldsymbol{x}}[\sum_{i=1}^{n}\log\frac{p(x_i|y=1)}{p(x_i|y=0)}+\log\frac{p(y=1)}{p(y=0)}|y=1]\\ &= \sum_{i=1}^{n}\mathbb{E}_{x_i}[\log\frac{p(x_i|y=1)}{p(x_i|y=0)}|y=1]+\log\frac{p(y=1)}{p(y=0)}.\end{aligned}$$

We note that $\mathbb{E}_{x_i}[\log\frac{p(x_i|y=1)}{p(x_i|y=0)}|y=1]$ is the KL Divergence $D(p(x_i|y=1)\|p(x_i|y=0))$. It is nonnegative and equals 0 if and only if $p(x_i|y=1)=p(x_i|y=0)$ for all $x_i\in\mathcal{X}_i$ ($\{0,1\}$ in case of discrete inputs and $[0,1]$ in case of continuous inputs). By assumption B.1, we obtain that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\Delta a_{Gen}(\boldsymbol{x},1,0)|y=1] &= \sum_{i=1}^{n}D(p(x_i|y=1)\|p(x_i|y=0))+\log\frac{p(y=1)}{p(y=0)}\\ &= \beta_{1,0}n+\log\frac{p(y=1)}{p(y=0)}\\ &\geq \beta_{1,0}n+\log(\frac{\rho_0}{1-\rho_0}),\end{aligned}$$

which implies that $\mathbb{E}[\Delta a_{Gen}(\boldsymbol{x},1,0)|y=1]=\Omega(n)$. In the same way, we can know that $\mathbb{E}[-\Delta a_{Gen}(\boldsymbol{x},1,0)|y=0]=\Omega(n)$ as well. Then the proposition has been proved. □

Based on Lemma F.8, we can prove Proposition B.2.

*Proof.* For convenience, we denote $\mathbb{E}[\Delta a_{Gen}(\boldsymbol{x},1,0)|y=k]$ by $\zeta_k$. To bound $G(\tau)|_{y=1}=\mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x},1,0)|\leq\tau n|y=1)$, the following holds:

$$\begin{aligned}&\mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x},1,0)|\leq\tau n|y=1)\\ &\leq\mathbb{P}(\Delta a_{Gen,\infty}(\boldsymbol{x},1,0)\leq\tau n|y=1)\\ &=\mathbb{P}(\Delta a_{Gen,\infty}(\boldsymbol{x},1,0)-\zeta_1 n\leq\tau n-\zeta_1 n|y=1)\\ &=\mathbb{P}(\sum_{i=1}^{n}\log\frac{p(x_i|y=1)}{p(x_i|y=0)}-\mathbb{E}_{\boldsymbol{x}}(\sum_{i=1}^{n}\log\frac{p(x_i|y=1)}{p(x_i|y=0)})\leq(\tau-\zeta_1)n|y=1)\\ &=\mathbb{P}(|\sum_{i=1}^{n}\log\frac{p(x_i|y=1)}{p(x_i|y=0)}-\mathbb{E}_{\boldsymbol{x}}(\sum_{i=1}^{n}\log\frac{p(x_i|y=1)}{p(x_i|y=0)})|\geq(\zeta_1-\tau)n|y=1)\\ &\leq\frac{\mathbb{V}[\sum_{i=1}^{n}\log\frac{p(x_i|y=1)}{p(x_i|y=0)}|y=1]}{(\tau-\zeta_1)^2n^2} && \text{(Chebyshev inequality)}\\ &=\frac{\alpha_1 n}{(\tau-\zeta_1)^2n^2} && \text{(Assumption B.2)}\\ &=\frac{\alpha_1}{(\tau-\zeta_1)^2n}.\end{aligned}$$

Similar to the above discussion, we have: $G(\tau)|_{y=0}\leq\frac{\alpha_0}{(\tau-|\zeta_0|)^2n}$. Finally, we can conclude that:

$$\begin{aligned}\widetilde{G}(\tau)&=p(y=1)G(\tau)|_{y=1}+p(y=0)G(\tau)|_{y=0}\\ &\leq p(y=1)\frac{\alpha_1}{(\tau-\zeta_1)^2n}+p(y=0)\frac{\alpha_0}{(\tau-|\zeta_0|)^2n}\\ &\leq\frac{\alpha}{(\tau-\zeta)^2n}.\end{aligned}$$

□

### F.5. Proof of Proposition B.3

*Proof.* Based on the results from Lemma F.8, we first consider the discrete condition and the event that a test sample $\boldsymbol{x}$ with label 1. To bound $G(\tau)|_{y=1} = \mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x},1,0)| \leq \tau n | y=1)$, the following holds:

$$\begin{aligned} G(\tau)|_{y=1} &= \mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x},1,0)| \leq \tau n | y=1) \\ &\leq \mathbb{P}(\Delta a_{Gen,\infty}(\boldsymbol{x},1,0) \leq \tau n | y=1) \\ &= \mathbb{P}(\Delta a_{Gen,\infty}(\boldsymbol{x},1,0) - \zeta_1 n \leq \tau n - \zeta_1 n | y=1) \\ &= \mathbb{P}(\sum_{i=1}^{n} \log \frac{p(x_i|y=1)}{p(x_i|y=0)} - \mathbb{E}_{\boldsymbol{x}}(\sum_{i=1}^{n} \log \frac{p(x_i|y=1)}{p(x_i|y=0)}) \leq (\tau - \zeta_1) n | y=1) \\ &\leq \exp(-\frac{2(\tau-\zeta_1)^2 n^2}{n(\log \frac{1-\rho_0}{\rho_0} - \log \frac{\rho_0}{1-\rho_0})^2}) = \exp\Big(-\frac{(\tau-\zeta_1)^2 n}{2(\log \frac{1-\rho_0}{\rho_0})^2}\Big). \qquad \text{(by Lemma D.9)} \end{aligned}$$

Similar to the above discussion, we have: $G(\tau)|_{y=0} \leq \exp\Big(-\frac{(\tau-\zeta_2)^2 n}{2(\log \frac{1-\rho_0}{\rho_0})^2}\Big)$. Finally, we can conclude that:

$$\begin{aligned} G(\tau) &= p(y=1)G(\tau)|_{y=1} + p(y=0)G(\tau)|_{y=0} \\ &\leq p(y=1)\exp\Big(-\frac{(\tau-\zeta_1)^2 n}{2(\log \frac{1-\rho_0}{\rho_0})^2}\Big) + p(y=0)\exp\Big(-\frac{(\tau-\zeta_2)^2 n}{2(\log \frac{1-\rho_0}{\rho_0})^2}\Big) \\ &\leq \exp\Big(-\frac{(\tau-\zeta)^2 n}{2(\log \frac{1-\rho_0}{\rho_0})^2}\Big) = \exp\left(-O((\tau-\zeta)^2 n)\right). \end{aligned}$$

Second, we consider the continuous case, the only difference from the discrete case is that the range of $\log \frac{p(x_i|y=1)}{p(x_i|y=0)}$. For all $i$, it satisfies:

$$\begin{aligned} |\log \frac{p(x_i|y=1)}{p(x_i|y=0)}| &= |\log \frac{\frac{1}{\sqrt{2\pi}\sigma_i}\exp(-\frac{(x_i-\mu_{1i})^2}{2\sigma_i^2})}{\frac{1}{\sqrt{2\pi}\sigma_i}\exp(-\frac{(x_i-\mu_{0i})^2}{2\sigma_i^2})}| \\ &= |\frac{\mu_{1i}-\mu_{0i}}{\sigma_i^2}x_i + \frac{\mu_{0i}^2-\mu_{1i}^2}{2\sigma_i^2}| \\ &\leq |\frac{\mu_{1i}-\mu_{0i}}{\sigma_i^2}x_i| + |\frac{(\mu_{0i}-\mu_{1i})(\mu_{0i}+\mu_{1i})}{2\sigma_i^2}| \\ &\leq \frac{1}{\rho_0} + \frac{2}{2\rho_0} = \frac{2}{\rho_0}. \end{aligned}$$

So we can get:

$$G(\tau) \leq \exp\Big(-\frac{2(\tau-\zeta)^2 n}{(\frac{4}{\rho_0})^2}\Big) = \exp\left(-O((\tau-\zeta)^2 n)\right).$$

□

### F.6. Proof of TheoremB.3

*Proof.* In the case that precondition of Proposition B.2 holds, combining Theorem B.2 and Proposition B.2, we know that there exist positive $c = \Theta(1)$ such that when $c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta$, with probability at least $1-\delta$, we have

$$R_{\ell_{0-1}}(h_{Gen,m}) \leq R_{\ell_{0-1}}(h_{Gen,\infty}) + \frac{\alpha}{(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n} + \delta.$$

For fixed $\epsilon_0 \in (0,1)$, the logical relations listed in the following is correct:

$$
\begin{aligned}
& R_{\ell_{0-1}}(h_{Gen,m}) \le R_{\ell_{0-1}}(h_{Gen,\infty}) + \epsilon_0 \text{ with probability at least 1 - } \delta \\
& \Leftarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < 1 \wedge \frac{\alpha}{(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n} + \delta \le \epsilon_0 \\
& \Leftrightarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < 1 \wedge \frac{\alpha}{(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n} \le \epsilon_0 - \delta \\
& \Leftrightarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < 1 \wedge \epsilon_0 - \delta > 0 \wedge (c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 \ge \frac{\alpha}{(\epsilon_0 - \delta)n} \\
& \Leftrightarrow c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < \epsilon_0 \wedge (c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 \ge \frac{\alpha}{(\epsilon_0 - \delta)n} \\
& \Leftrightarrow 0 < \delta < \epsilon_0 \wedge \zeta - c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} \ge \sqrt{\frac{\alpha}{(\epsilon_0 - \delta)n}} \\
& \Leftrightarrow 0 < \delta < \epsilon_0 \wedge \zeta - \sqrt{\frac{\alpha}{(\epsilon_0 - \delta)n}} > 0 \wedge (\zeta - \sqrt{\frac{\alpha}{(\epsilon_0 - \delta)n}})^2 \ge c^2 \frac{1}{m}\log(\frac{n}{\delta}) \\
& \Leftarrow 0 < \delta < \epsilon_0 - \frac{\alpha}{\zeta^2 n} \wedge \epsilon_0 - \frac{\alpha}{\zeta^2 n} > 0 \wedge m \ge \frac{c^2}{(\zeta - \sqrt{\frac{\alpha}{(\epsilon_0-\delta)n}})^2}\log(\frac{n}{\delta}) \\
& \Leftarrow 0 < \delta \le \frac{\epsilon_0}{2} \wedge \epsilon_0 - \frac{\alpha}{\zeta^2 n} > \frac{\epsilon_0}{2} \wedge m \ge \frac{c^2}{(\zeta - \sqrt{\frac{\alpha}{(\epsilon_0-\delta)n}})^2}\log(\frac{n}{\delta}) \\
& \Leftrightarrow 0 < \delta \le \frac{\epsilon_0}{2} \wedge n < \frac{2\alpha}{\epsilon_0\zeta^2} \wedge m \ge \frac{c^2}{(\zeta - \sqrt{\frac{\alpha}{(\epsilon_0-\delta)n}})^2}\log(\frac{n}{\delta}) \\
& \Leftarrow 0 < \delta \le \frac{\epsilon_0}{2} \wedge n < 2\frac{2\alpha}{\epsilon_0\zeta^2} \wedge m \ge \frac{c^2}{(\zeta - \sqrt{\frac{2\alpha}{\epsilon_0 n}})^2}\log(\frac{n}{\delta}) \\
& \Leftarrow 0 < \delta \le \frac{\epsilon_0}{2} \wedge n < \frac{4\alpha}{\epsilon_0\zeta^2} \wedge m \ge \frac{c^2}{(\zeta - \sqrt{\frac{1}{2}}\zeta)^2}\log(\frac{n}{\delta}) \\
& \Leftarrow 0 < \delta \le \frac{\epsilon_0}{2} \wedge n < \frac{4\alpha}{\epsilon_0\zeta^2} \wedge m = O(\log(n)).
\end{aligned}
$$

In the case that precondition of Proposition B.3 holds, then by Theorem B.2 and Proposition B.3, we know that there exist positive constant $b, c = \Theta(1)$ such that when $c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta$, with probability at least $1 - \delta$, we have

$$
R_{\ell_{0-1}}(h_{Gen,m}) \le R_{\ell_{0-1}}(h_{Gen,\infty}) + \exp(-b(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n) + \delta.
$$

For fixed $\epsilon_0 \in (0,1)$, the logical relations listed in the following is correct:

$$
\begin{aligned}
& R_{\ell_{0-1}}(h_{Gen,m}) \leq R_{\ell_{0-1}}(h_{Gen,\infty}) + \epsilon_0 \text{ with probability at least 1 - } \delta \\
\Leftarrow\ & c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < 1 \wedge \exp(-b(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n) + \delta \leq \epsilon_0 \\
\Leftrightarrow\ & c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < 1 \wedge \exp(-b(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n) \leq \epsilon_0 - \delta \\
\Leftrightarrow\ & c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < 1 \wedge \epsilon_0 - \delta > 0 \wedge -b(c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 n \leq \log(\epsilon_0 - \delta) \\
\Leftrightarrow\ & c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < \epsilon_0 \wedge (c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} - \zeta)^2 \geq \frac{1}{bn}\log(\frac{1}{\epsilon_0 - \delta}) \\
\Leftarrow\ & c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} < \zeta \wedge 0 < \delta < \epsilon_0 \wedge \zeta - c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} \geq \sqrt{\frac{1}{bn}\log(\frac{1}{\epsilon_0 - \delta})} \\
\Leftrightarrow\ & 0 < \delta < \epsilon_0 \wedge \zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\epsilon_0 - \delta})} \geq c\sqrt{\frac{1}{m}\log(\frac{n}{\delta})} \\
\Leftarrow\ & 0 < \delta < \epsilon_0 \wedge \zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\epsilon_0 - \delta})} > 0 \wedge (\zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\epsilon_0 - \delta})})^2 \geq c^2 \frac{1}{m}\log(\frac{n}{\delta}) \\
\Leftarrow\ & 0 < \delta < \epsilon_0 - \exp(-b\zeta^2 n) \wedge \epsilon_0 - \exp(-b\zeta^2 n) > 0 \wedge m \geq \frac{c^2}{(\zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\epsilon_0 - \delta})})^2}\log(\frac{n}{\delta}) \\
\Leftarrow\ & 0 < \delta \leq \frac{\epsilon_0}{2} \wedge \epsilon_0 - \exp(-b\zeta^2 n) > \frac{\epsilon_0}{2} \wedge m \geq \frac{c^2}{(\zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\epsilon_0 - \delta})})^2}\log(\frac{n}{\delta}) \\
\Leftrightarrow\ & 0 < \delta \leq \frac{\epsilon_0}{2} \wedge \epsilon_0 \exp(b\zeta^2 n) > 2 \wedge m \geq \frac{c^2}{(\zeta - \sqrt{\frac{1}{bn}\log(\frac{1}{\epsilon_0 - \delta})})^2}\log(\frac{n}{\delta}) \\
\Leftarrow\ & 0 < \delta \leq \frac{\epsilon_0}{2} \wedge \epsilon_0 \exp(b\zeta^2 n) > 3 \wedge m \geq \frac{c^2}{(\zeta - \sqrt{\frac{1}{bn}\log(\frac{2}{\epsilon_0})})^2}\log(\frac{2n}{\epsilon_0}) \\
\Leftarrow\ & 0 < \delta \leq \frac{\epsilon_0}{2} \wedge \epsilon_0 \exp(b\zeta^2 n) > 3 \wedge m \geq \frac{c^2}{\zeta^2(1 - \frac{\log(2/\epsilon_0)}{\log(3/\epsilon_0)})^2}\log(\frac{2n}{\epsilon_0}) \\
\Leftrightarrow\ & 0 < \delta \leq \frac{\epsilon_0}{2} \wedge \epsilon_0 \exp(b\zeta^2 n) > 3 \wedge m = O(\log(n)).
\end{aligned}
$$

□

# G. Proofs of Appendix C

## G.1. Proof of Proposition C.1

*Proof.* Because $\langle \Delta \mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}) \rangle_\epsilon < s(\Delta \mathcal{C}_{\ell_1,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}))$ for all $x \in \mathcal{X}$, we have:

$$
\begin{aligned}
& R_{\ell_2}(\boldsymbol{h}) - R^*_{\ell_2,\mathcal{H}} + M_{\ell_2,\mathcal{H}} \\
& = \mathbb{E}_{\boldsymbol{x}}[\mathcal{C}_{\ell_2}(\boldsymbol{h},\boldsymbol{x})] - R^*_{\ell_2,\mathcal{H}} + R^*_{\ell_2,\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[\mathcal{C}^*_{\ell_2,\mathcal{H}}(\boldsymbol{x})] && \text{(by definition)} \\
& = \mathbb{E}_{\boldsymbol{x}}[\mathcal{C}_{\ell_2}(\boldsymbol{h},\boldsymbol{x}) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(\boldsymbol{x})] \\
& = \mathbb{E}_{\boldsymbol{x}}[\Delta \mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})] \\
& = \mathbb{E}_{\boldsymbol{x}}[\Delta \mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}) \mathbb{1}_{\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}) > \epsilon} + \Delta \mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}) \mathbb{1}_{\mathcal{C}_{\ell_2,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}) \le \epsilon}] \\
& \le \mathbb{E}_{\boldsymbol{x}}[s(\Delta \mathcal{C}_{\ell_1,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}))] + \epsilon \\
& \le s(\mathbb{E}_{\boldsymbol{x}}[\Delta \mathcal{C}_{\ell_1,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})]) + \epsilon && \text{(Jensen's inequality)} \\
& = s(R_{\ell_1}(\boldsymbol{h}) - R^*_{\ell_1,\mathcal{H}} + M_{\ell_1,\mathcal{H}}) + \epsilon.
\end{aligned}
$$

□

## G.2. Proof of Theorem C.1

**Lemma G.1** (Distribution-dependent concave $\ell_{0-1}$ bound)**.** *Suppose that $\mathcal{H}$ satisfies that $\{\operatorname{argmax}_{y \in \mathcal{Y}} h_y(\boldsymbol{x}) : \boldsymbol{h} \in \mathcal{H}\} = \{1, \ldots, K\}$ for any $\boldsymbol{x} \in \mathcal{X}$, and there exists a non-decreasing concave function $s : \mathbb{R}_+ \to \mathbb{R}_+$ and $\epsilon \ge 0$ that the following holds for any $\hat{y} \in \mathcal{Y}$, $x \in \mathcal{X}$ and $\boldsymbol{h} \in \mathcal{H}_{\hat{y}}(\boldsymbol{x})$:*

$$
\langle \max_y p_y(\boldsymbol{x}) - p_{\hat{y}}(\boldsymbol{x}) \rangle_\epsilon \le s(\inf_{\boldsymbol{h} \in \mathcal{H}_{\hat{y}}(\boldsymbol{x})} \Delta \mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x})).
$$

*Then it holds for all $\boldsymbol{h} \in \mathcal{H}$ that*

$$
R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}} \le s(R_{\ell(\boldsymbol{h})} - R^*_{\ell,\mathcal{H}} + M_{\ell,\mathcal{H}}) + \epsilon.
$$

*Proof.* For any $\boldsymbol{x}_0 \in \mathcal{X}$ and $\boldsymbol{h}_0 \in \mathcal{H}$, let $\hat{y}$ be the index of the largest element of $\boldsymbol{h}_0(\boldsymbol{x})$. Then by the precondition, we have

$$
\langle \Delta \mathcal{C}_{\ell_{0-1},\mathcal{H}}(\boldsymbol{h}_0,\boldsymbol{x}_0) \rangle_\epsilon = \langle \max_y p_y(\boldsymbol{x}_0) - p_{\hat{y}}(\boldsymbol{x}_0) \rangle_\epsilon \le s(\inf_{\boldsymbol{h} \in \mathcal{H}_{\hat{y}}(\boldsymbol{x}_0)} \Delta \mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}_0)) \le s(\Delta \mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h}_0,\boldsymbol{x}_0)).
$$

where we use the assumption that $s$ is non-decreasing. Combining the condition in Proposition C.1 we can conclude the proof. □

Built upon Lemma G.1, we can prove Theorem C.1 as follows.

*Proof.* For any $\boldsymbol{x}_0 \in \mathcal{X}$, $\boldsymbol{p}(\boldsymbol{x}_0) \in \Delta_K$, $\hat{y}_0 \in \mathcal{Y}$, and $\boldsymbol{h} \in \mathcal{H}_{\hat{y}_0}(\boldsymbol{x}_0)$, we can write:

$$
\begin{aligned}
& \max_y p_y(\boldsymbol{x}_0) - p_{\hat{y}_0}(\boldsymbol{x}_0) \\
& \le s(\inf_{\hat{y} \in \mathcal{Y}, x \in \mathcal{X}, \boldsymbol{h} \in \mathcal{H}_{\hat{y}}(\boldsymbol{x}), \boldsymbol{p} \in \mathcal{P}_{\hat{y}}(\max_y p_y(\boldsymbol{x}_0) - p_{\hat{y}_0}(\boldsymbol{x}_0))} \Delta \mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) && \text{(Assumption)} \\
& \le s(\inf_{x \in \mathcal{X}, \boldsymbol{h} \in \mathcal{H}_{\hat{y}_0}(\boldsymbol{x}), \boldsymbol{p} \in \mathcal{P}_{\hat{y}_0}(\max_y p_y(\boldsymbol{x}_0) - p_{\hat{y}_0}(\boldsymbol{x}_0))} \Delta \mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) \\
& \le s(\inf_{x \in \mathcal{X}, \boldsymbol{h} \in \mathcal{H}_{\hat{y}_0}(\boldsymbol{x})} \Delta \mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}(\boldsymbol{x}_0))) \\
& \le s(\inf_{\boldsymbol{h} \in \mathcal{H}_{\hat{y}_0}(\boldsymbol{x}_0)} \Delta \mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}_0,\boldsymbol{p}(\boldsymbol{x}_0))) \\
& = s(\inf_{\boldsymbol{h} \in \mathcal{H}_{\hat{y}_0}(\boldsymbol{x}_0)} \Delta \mathcal{C}_{\ell,\mathcal{H}}(\boldsymbol{h},\boldsymbol{x}_0)).
\end{aligned}
$$

Combining the result of Lemma G.1 we can prove Theorem C.1. □

### G.3. Proofs of Theorem C.2

*Proof.* The proof is essentially the same as that of Theorem 3.3. We use $\mathcal{H}, \ell$ to replace the $\mathcal{H}_{NN}, \ell_{log}$ in the following proof, which will not bring ambiguity. We can rewrite the $\mathcal{J}_\ell(t)$ as follows:

$$\mathcal{J}_\ell(t) = \inf_{\hat{y}\in\mathcal{Y}} \inf_{\boldsymbol{p}\in\mathcal{P}_{\hat{y}}(t)} \inf_{\boldsymbol{x}\in\mathcal{X}} \big( \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})\big).$$

For all $\boldsymbol{h} \in \mathcal{H}$ and $\boldsymbol{x} \in \mathcal{X}$, we have

$$\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) = \sum_{y=1}^{K} p_y \ell(y, \boldsymbol{h}(x)) = \sum_{y=1}^{K} p_y(-h_y + \log(\sum_{j=1}^{K} \exp(h_j))).$$

To get the $\inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}))$, we consider the following problem

$$\min_{\boldsymbol{h}} \sum_{y=1}^{K} p_y(-h_y + \log(\sum_{j=1}^{K} \exp(h_j))).$$

By Lemma E.3, we know that this problem is convex, we can make use of KKT conditions (Boyd et al., 2004) to find the points that are primal and dual optimal, which can be written as follows

$$-p_i + \frac{\exp(h_i^*)}{\sum_{k=1}^{K} \exp(h_k^*)} = 0 \quad i = 1, \ldots, K. \tag{19}$$

It implies that $h_i^* = \log(p_i \sum_{k=1}^{K} \exp(h_k^*)$. Thus, we have

$$\inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) = \sum_{y=1}^{K} p_y(-h_y^* + \log(\sum_{j=1}^{K} \exp(h_j^*))) = -\sum_{y=1}^{K} p_y \log(p_y).$$

which is the entropy of distribution $\boldsymbol{p}$. By Lemma E.6, we know that

$$\inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})) \geq -(p_{max} + p_{\hat{y}}) \log(\frac{p_{max} + p_{\hat{y}}}{2}) - \sum_{y\notin\{y_{max},\hat{y}\}} p_y \log(p_y).$$

Then we have

$$\inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) \geq -(p_{max} + p_{\hat{y}}) \log(\frac{p_{max} + p_{\hat{y}}}{2}) + p_{y_{max}} \log(p_{y_{max}}) + p_{\hat{y}} \log(p_{\hat{y}}),$$

and

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \big( \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} (\mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}))\big) \geq -(p_{y_{max}} + p_{\hat{y}}) \log(\frac{p_{y_{max}} + p_{\hat{y}}}{2}) + p_{y_{max}} \log(p_{y_{max}}) + p_{p_{\hat{y}}} \log(p_{p_{\hat{y}}}).$$

Now, we meet the following problem

$$\begin{aligned} \min_{\boldsymbol{p}} \;& -(p_{y_{max}} + p_{\hat{y}}) \log(\frac{p_{y_{max}} + p_{\hat{y}}}{2}) + p_{y_{max}} \log(p_{y_{max}}) + p_{\hat{y}} \log(p_{\hat{y}}) \\ s.t. \;& \begin{cases} p_{y_{max}} - p_{\hat{y}} = t, \forall i, \\ \sum_{i=1}^{K} p_i = 1, \\ p_i \geq 0, \forall i, \end{cases} \end{aligned}$$

which is equivalent to find the minimum of $-(2p_{\hat{y}} + t) \log(\frac{2p_{\hat{y}}+t}{2}) + (p_{\hat{y}} + t) \log((p_{\hat{y}} + t)) + p_{\hat{y}} \log(p_{\hat{y}})$ when $p_{\hat{y}} \in [0, \frac{1-t}{2}]$. By Lemma E.7, we know it is $\frac{1+t}{2} \log(1 + t) + \frac{1-t}{2} \log(1 - t)$. Thus,

$$\begin{aligned} \mathcal{J}_\ell(t) &= \inf_{\hat{y}\neq y_{max}} \inf_{\boldsymbol{p}\in\{\boldsymbol{p}:\boldsymbol{p}\in\Delta_k, p_{max}-p_{\hat{y}}=t\}} \inf_{\boldsymbol{x}\in\mathcal{X}} \big( \inf_{\boldsymbol{h}\in\mathcal{H}_{\hat{y}}(\boldsymbol{x})} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p}) - \inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_\ell(\boldsymbol{h},\boldsymbol{x},\boldsymbol{p})\big) \\ &\geq \inf_{\hat{y}\neq y_{max}} -(2-t) \log(\frac{2-t}{2}) + (1-t) \log(1-t) \\ &= \frac{1+t}{2} \log(1+t) + \frac{1-t}{2} \log(1-t) \\ &\geq \frac{t^2}{2}. && \text{(Lemma E.8)} \end{aligned}$$

Let $g(t) = \frac{t^2}{4}$ in Theorem 3.5, we have

$$\frac{1}{2}(R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}})^2 \le R_{\ell_{log}}(\boldsymbol{h}) - R^*_{\ell_{log},\mathcal{H}} + M_{\ell_{log},\mathcal{H}},$$

which implies

$$R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}} \le \sqrt{2}(R_{\ell_{log}}(\boldsymbol{h}) - R^*_{\ell_{log},\mathcal{H}} + M_{\ell_{log},\mathcal{H}})^{\frac{1}{2}}.$$

By Lemma E.4, we have $M_{\ell_{0-1},\mathcal{H}}$ coincides with the approximation error $R^*_{\ell_{0-1},\mathcal{H}} - R^*_{\ell_{0-1},\mathcal{H}_{all}}$. We also note that $M_{\ell_{log},\mathcal{H}}$ coincides with $R^*_{\ell_{log},\mathcal{H}} - R^*_{\ell_{log},\mathcal{H}_{all}}$ because

$$\begin{aligned}
M_{\ell_{log},\mathcal{H}} &= R^*_{\ell_{log},\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[\mathcal{C}^*_{\ell_{log},\mathcal{H}}(\boldsymbol{x})] \\
&= R^*_{\ell_{log},\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[\inf_{\boldsymbol{h}\in\mathcal{H}} \mathcal{C}_{\ell_{log}}(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{p}(\boldsymbol{x}))] \\
&= R^*_{\ell_{log},\mathcal{H}} - \mathbb{E}_{\boldsymbol{x}}[-\sum_{y=1}^{K} p_y(\boldsymbol{x}) \log(p_y(\boldsymbol{x}))] \\
&= R^*_{\ell_{log},\mathcal{H}} - R^*_{\ell_{log},\mathcal{H}_{all}}.
\end{aligned}$$

Finally, we can conclude that

$$R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} \le R_{\ell_{0-1}}(\boldsymbol{h}) - R^*_{\ell_{0-1},\mathcal{H}} + M_{\ell_{0-1},\mathcal{H}} \le \sqrt{2}(R_{\ell_{log}}(\boldsymbol{h}) - R^*_{\ell_{log},\mathcal{H}_{all}})^{\frac{1}{2}}.$$

□

### G.4. Proof of Proposition C.2

*Proof.* Based on the results of Lemma D.3, for $k_1, k_2$ and $k$ which satisfies $\zeta_{k_1,k_2,k} > 0$, to bound $\mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x}, k_1, k_2)| \le \tau n | y = k)$, we can write:

$$\begin{aligned}
&\mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x}, k_1, k_2)| \le \tau n | y = k) \\
&\le \mathbb{P}(\Delta a_{Gen,\infty}(\boldsymbol{x}, k_1, k_2) \le \tau n | y = k) \\
&= \mathbb{P}(\Delta a_{Gen,\infty}(\boldsymbol{x}, k_1, k_2) - \zeta_{k_1,k_2,k} n \le \tau n - \zeta_{k_1,k_2,k} n | y = k) \\
&= \mathbb{P}(\sum_{i=1}^{n} \log \frac{p(x_i|y=k_1)}{p(x_i|y=k_2)} - \mathbb{E}_{\boldsymbol{x}}(\sum_{i=1}^{n} \log \frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}) \le (\tau - \zeta_{k_1,k_2,k}) n | y = k) \\
&\le \exp(-\frac{2(\tau - \zeta_{k_1,k_2,k})^2 n^2}{n(\log \frac{1-\rho_0}{\rho_0} - \log \frac{\rho_0}{1-\rho_0})^2}) = \exp(-\frac{(\tau - \zeta_{k_1,k_2,k})^2 n}{2(\log \frac{1-\rho_0}{\rho_0})^2}). \qquad \text{(Assumption 3.2 and Lemma D.9)}
\end{aligned}$$

Similar to the above discussion, we have $\mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x}, k_1, k_2)| \le \tau n | y = k) \le \exp(-\frac{(\tau - |\zeta_{k_1,k_2,k}|)^2 n}{2(\log \frac{1-\rho_0}{\rho_0})^2})$ for $k_1, k_2$ and $k$ which satisfies $\beta_{k_1,k_2,k} < 0$. Finally, we can conclude that:

$$\begin{aligned}
\widetilde{G}(\tau) &= \max_{k_1,k_2} \sum_{k=1}^{K} p(y=k) \mathbb{P}(|\Delta a_{Gen,\infty}(\boldsymbol{x}, k_1, k_2)| \le \tau n | y = k) \\
&\le \max_{k_1,k_2} \sum_{k=1}^{K} p(y=k) \exp(-\frac{(\tau - |\zeta_{k_1,k_2,k}|)^2 n}{2(\log \frac{1-\rho_0}{\rho_0})^2}) \\
&\le \max_{k_1,k_2} \exp(-\frac{(\tau - \min_k |\zeta_{k_1,k_2,k}|)^2 n}{2(\log \frac{1-\rho_0}{\rho_0})^2}) \\
&= \exp(-\frac{(\tau - \zeta)^2 n}{2(\log \frac{1-\rho_0}{\rho_0})^2}) = \exp(-O((\tau - \zeta)^2 n)).
\end{aligned}$$

Second, we consider the continuous case, the only difference from the discrete case is that the range of $\log \frac{p(x_i|y=k_1)}{p(x_i|y=k_2)}$ is $[-\frac{2}{\rho_0}, \frac{2}{\rho_0}]$. So we can get:

$$\widetilde{G}(\tau) \leq \exp\big(-\frac{(\tau-\zeta)^2 n}{2(\frac{4}{\rho_0})^2}\big) = \exp\left(-O((\tau-\zeta)^2 n)\right).$$

The proof is complete. □

## H. Details of Simulation Experiments

### H.1. Implementation of Logistic Regression

We train the logistic regression using scikit-learn's (Pedregosa et al., 2011) L-BFGS implementation, with a maximum of 1000 iterations. The weight of $\ell_2$ regularization of logistic regression is fixed as 1. All experiments are done on a single GeForce RTX 3090 GPU.

### H.2. Sythentic Dataset

We construct a simulated multiclass balanced mixture Gaussian distribution dataset, which also satisfies all assumptions. The simulated data distribution satisfies $p(x|y=1) \sim \mathcal{N}(x;\{-1\}^n, diag\{\{n\}^{\frac{n}{2}},\{1\}^{\frac{n}{2}}\})$ and $p(x|y=k) \sim \mathcal{N}(x;\{2^{k-2}\}^n, diag\{\{n\}^{\frac{n}{2}},\{1\}^{\frac{n}{2}}\})$ for $k>1$, where $\mathcal{N}$ is Gaussian distribution, $diag(\boldsymbol{a})$ means a matrix whose diagonal is $\boldsymbol{a}$, and $\{a\}^n$ means a vector whose length is $n$ and all its elements are $a$.

### H.3. Discussion about the synthetic dataset

First, we note that the optimal classifier is a linear function, which means that Assumption 3.5 is valid with $\nu=0$.

**Binary case**. The data distribution satisfies $p(x|y=0) \sim \mathcal{N}(x;\{-1\}^n, diag\{\{n\}^{\frac{n}{2}},\{1\}^{\frac{n}{2}}\})$ and $p(x|y=1) \sim \mathcal{N}(\{1\}^n, diag\{\{n\}^{\frac{n}{2}},\{1\}^{\frac{n}{2}}\})$. The boundary of Bayes classifier $\Delta a_{Gen}(\boldsymbol{x},1,0)$ can be calculated as follows:

$$\begin{aligned}\Delta a_{Gen}(\boldsymbol{x},1,0) &= \sum_{i=1}^{n} \log \frac{\frac{1}{\sqrt{2\pi}\sigma_i}\exp(-\frac{(x_i-\mu_{1i})^2}{2\sigma_i^2})}{\frac{1}{\sqrt{2\pi}\sigma_i}\exp(-\frac{(x_i-\mu_{0i})^2}{2\sigma_i^2})} + \log\frac{q}{1-q}\\ &= \sum_{i=1}^{n} \frac{\mu_{1i}-\mu_{0i}}{\sigma_i^2}x_i + \sum_{i=1}^{n}\frac{\mu_{0i}^2-\mu_{1i}^2}{2\sigma_i^2} + \log\frac{q}{1-q}\\ &= \sum_{i=1}^{n/2}\frac{2}{n}x_i + \sum_{i=\frac{n}{2}+1}^{n} 2x_i.\end{aligned}$$

It is a linear function. In addition, the Bayes error $BE$ can be obtained as follows.

$$\begin{aligned}BE &= \frac{1}{(2\pi)^{\frac{n}{2}}(n^{\frac{n}{2}})^{\frac{1}{2}}}\int_{\sum_{i=1}^{n/2}\frac{2}{n}x_i+\sum_{i=\frac{n}{2}+1}^{n}2x_i<0} \exp(-\frac{1}{2}(\sum_{i=1}^{n/2}\frac{(x_i-1)^2}{n} + \sum_{i=\frac{n}{2}+1}^{n}(x_i-1)^2))dx_1\ldots dx_n\\ &= \frac{1}{(2\pi)^{\frac{n}{2}}(n^{\frac{n}{2}})^{\frac{1}{2}}}\int_{\sum_{i=1}^{n/2}\frac{2}{n}y_i+\sum_{i=\frac{n}{2}+1}^{n}2y_i<-(n+1)} \exp(-\frac{1}{2}(\sum_{i=1}^{n/2}\frac{y_i^2}{n} + \sum_{i=\frac{n}{2}+1}^{n}y_i^2))dy_1\ldots dy_n\\ &= \frac{1}{(2\pi)^{\frac{n}{2}}}\int_{\sum_{i=1}^{n/2}\frac{2}{\sqrt{n}}z_i+\sum_{i=\frac{n}{2}+1}^{n}2z_i<-(n+1)} \exp(-\frac{1}{2}(\sum_{i=1}^{n}z_i^2))dz_1\ldots dz_n\\ &= \int_{\sum_{i=1}^{\frac{n}{2}}\frac{\boldsymbol{z_i}}{\sqrt{n}}+\sum_{i=\frac{n}{2}+1}^{n}\boldsymbol{z_i}+\frac{n+1}{2}<0}\mathcal{N}(\boldsymbol{z};0,\boldsymbol{I})d\boldsymbol{z},\end{aligned}$$

which approaches 0 quickly as $n$ increases and can be approximated by the Monte Caro method efficiently.

**Multiclass case**. The boundary of Bayes classifier $a(\boldsymbol{x}, k_1, k_2)$ for class $k_1 = 1$ and $k_2$ can be calculated as follows:

$$\begin{aligned}\Delta a_{Gen}(\boldsymbol{x}, k_1, k_2) &= \sum_{i=1}^{n} \log \frac{\frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(x_i-\mu_{k_1 i})^2}{2\sigma_i^2})}{\frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(x_i-\mu_{k_2 i})^2}{2\sigma_i^2})} + \log \frac{q_{k_1}}{q_{k_2}} \\ &= \sum_{i=1}^{n} \frac{\mu_{k_1 i} - \mu_{k_2 i}}{\sigma_i^2} x_i + \sum_{i=1}^{n} \frac{\mu_{k_2 i}^2 - \mu_{k_1 i}^2}{2\sigma_i^2} + \log \frac{q_{k_1}}{q_{k_2}} \\ &= \sum_{i=1}^{n/2} \frac{-1 - 2^{k_2-2}}{n} x_i + \sum_{i=\frac{n}{2}+1}^{n} (-1 + 2^{k_2-2}) x_i.\end{aligned}$$

In addition, the boundary of Bayes classifier $a(\boldsymbol{x}, k_1, k_2)$ for class $k_1 \neq 1$ and $k_2 \neq 1$ can be calculated as follows:

$$\begin{aligned}\Delta a_{Gen}(\boldsymbol{x}, k_1, k_2) &= \sum_{i=1}^{n} \log \frac{\frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(x_i-\mu_{k_1 i})^2}{2\sigma_i^2})}{\frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(x_i-\mu_{k_2 i})^2}{2\sigma_i^2})} + \log \frac{q_{k_1}}{q_{k_2}} \\ &= \sum_{i=1}^{n} \frac{\mu_{k_1 i} - \mu_{k_2 i}}{\sigma_i^2} x_i + \sum_{i=1}^{n} \frac{\mu_{k_2 i}^2 - \mu_{k_1 i}^2}{2\sigma_i^2} + \log \frac{q_{k_1}}{q_{k_2}} \\ &= \sum_{i=1}^{n/2} \frac{2^{k_1-2} - 2^{k_2-2}}{n} x_i + \sum_{i=\frac{n}{2}+1}^{n} (2^{k_1-2} - 2^{k_2-2}) x_i + (4^{k_1-2} - 4^{k_2-2}) \frac{n+1}{4}\end{aligned}$$

The Bayes error is not easy to obtain in an analytic version. However, the test error can decrease to less than $10^{-4}$ in our multiclass experiments, so we set 0 as the estimated asymptotic error.

Second, Assumption 3.3 holds in this case, that is, for all $k_1, k_2 (k_1 \neq k_2)$ and $k \in \mathcal{Y}$, it holds that $|\sum_{i=1}^{n} (D(p(x_i|y = k) \| p(x_i|y = k_1)) - D(p(x_i|y = k) \| p(x_i|y = k_2)))| = \beta_{k_1,k_2,k} n = \Omega(n)$. For all $k_1, k_2 (k_1 \neq k_2) \in \mathcal{Y}$, we have

$$\begin{aligned}\sum_{i=1}^{n} D(p(x_i|y = k_1) \| p(x_i|y = k_2)) &= \sum_{i=1}^{n} D(p(x_i|y = k_2) \| p(x_i|y = k_1)) \\ &= \sum_{i=1}^{n} \frac{(\mu_{k_2 i} - \mu_{k_1 i})^2}{2\sigma_i^2} = \sum_{i=1}^{n/2} \frac{(2^{k_1-2} - 2^{k_2-2})^2}{2n} + \sum_{i=\frac{n}{2}+1}^{n} (2^{k_1-2} - 2^{k_2-2})^2 \\ &= \frac{(2^{k_1-2} - 2^{k_2-2})^2}{4} + \frac{n}{2} (2^{k_1-2} - 2^{k_2-2})^2.\end{aligned}$$

So we have

$$\begin{aligned}&|\sum_{i=1}^{n} (D(p(x_i|y = k) \| p(x_i|y = k_1)) - D(p(x_i|y = k) \| p(x_i|y = k_2)))| \\ &= |\frac{(2^{k_1-2} - 2^{k-2})^2}{4} + \frac{n}{2} (2^{k_1-2} - 2^{k-2})^2 - \frac{(2^{k_2-2} - 2^{k-2})^2}{4} - \frac{n}{2} (2^{k_2-2} - 2^{k-2})^2| = O(n).\end{aligned}$$

Third, Assumption 3.4 holds as well. This can be obtained by the property of conditional independence directly, that is, for all $k_1, k_2 (k_1 \neq k_2)$ and $k \in \mathcal{Y}$, it holds that $\mathbb{V}_{\boldsymbol{x}}[\sum_{i=1}^{n} \log \frac{p(x_i|y=k_1)}{p(x_i|y=k_2)} | y = k] = \sum_{i=1}^{n} \mathbb{V}_{\boldsymbol{x}}[\log \frac{p(x_i|y=k_1)}{p(x_i|y=k_2)} | y = k] = O(n)$.

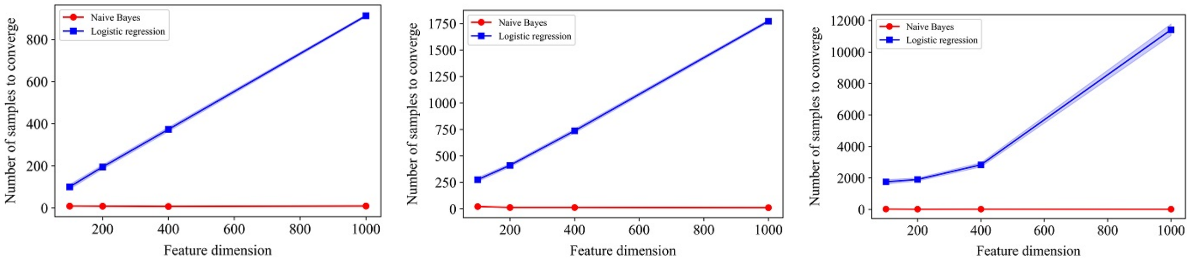
Finally, we note that we can directly scale this dataset because scaling will not influence the establishment of the above assumptions. In our multiclass experiments ($K > 2$), we scale the dataset to boost logistic regression converging faster. The scale function we use is $\boldsymbol{f}(\boldsymbol{x}) = \frac{\boldsymbol{x}}{2^{K-3}} - 1$, which can make the mean of each class to $[-1, 1]$.

### H.4. The number of samples required to converge

For a fixed $K$, we traversal $n$ from 100 to 1000 gradually. For each selected $n$, we randomly generate $1 \times 10^4$ samples as a test set. We increase the training dataset size $m$ gradually until the errors of two classifiers approach their asymptotic error. Specially, we conduct 5 random repeats to keep the stability of our results. We record the training set size $m_{conv}$ when the gap between the error and the estimation of asymptotic error is less than $\epsilon_0 = 0.01$ for the first time.

### H.5. Additional Results of Simulations

We present results with $K = 2, 3, 7$ here. Consistently, logistic regression and naïve Bayes require $O(n)$ and $O(\log n)$ samples to approach the estimated asymptotic error respectively. Error bars represent the variance estimated by 5 runs.



(a) $K = 2$ (b) $K = 3$ (c) $K = 7$

*Figure 4.* Additional results of simulations with $K = 2, 3, 7$.

## I. Details of Deep Learning Experiments

### I.1. Models

**ViT**. We include ViT-B/16 (Dosovitskiy et al., 2021) checkpoint pretrained on the ImageNet-21k dataset (Deng et al., 2009).

**ResNet**. We add the ResNet50 checkpoint released by Pytorch (Paszke et al., 2019).

**CLIP image encoder**. We use the image encoder released by CLIP (Radford et al., 2021) project with ResNet50 backbone.

**MoCov2**. We include the MoCov2 (Chen et al., 2020d) checkpoint trained with 800 epochs on the ImageNet dataset. The backbone is ResNet50.

**SimCLRv2**. The SimCLRv2 (Chen et al., 2020c) project released various pre-trained and fine-tuned models. We use the pretrain-only checkpoint with selective Kernels. The backbone is ResNet50.

**MAE**. We adopt pre-trained checkpoint in (He et al., 2022). The backbone is ViT-B/16.

**SimMIM**.We use the checkpoint pre-trained on the ImageNet-1K dataset with 800 epochs released in (Xie et al., 2022). The backbone is ViT-B/16.

The used codes and their licenses are listed as follows.

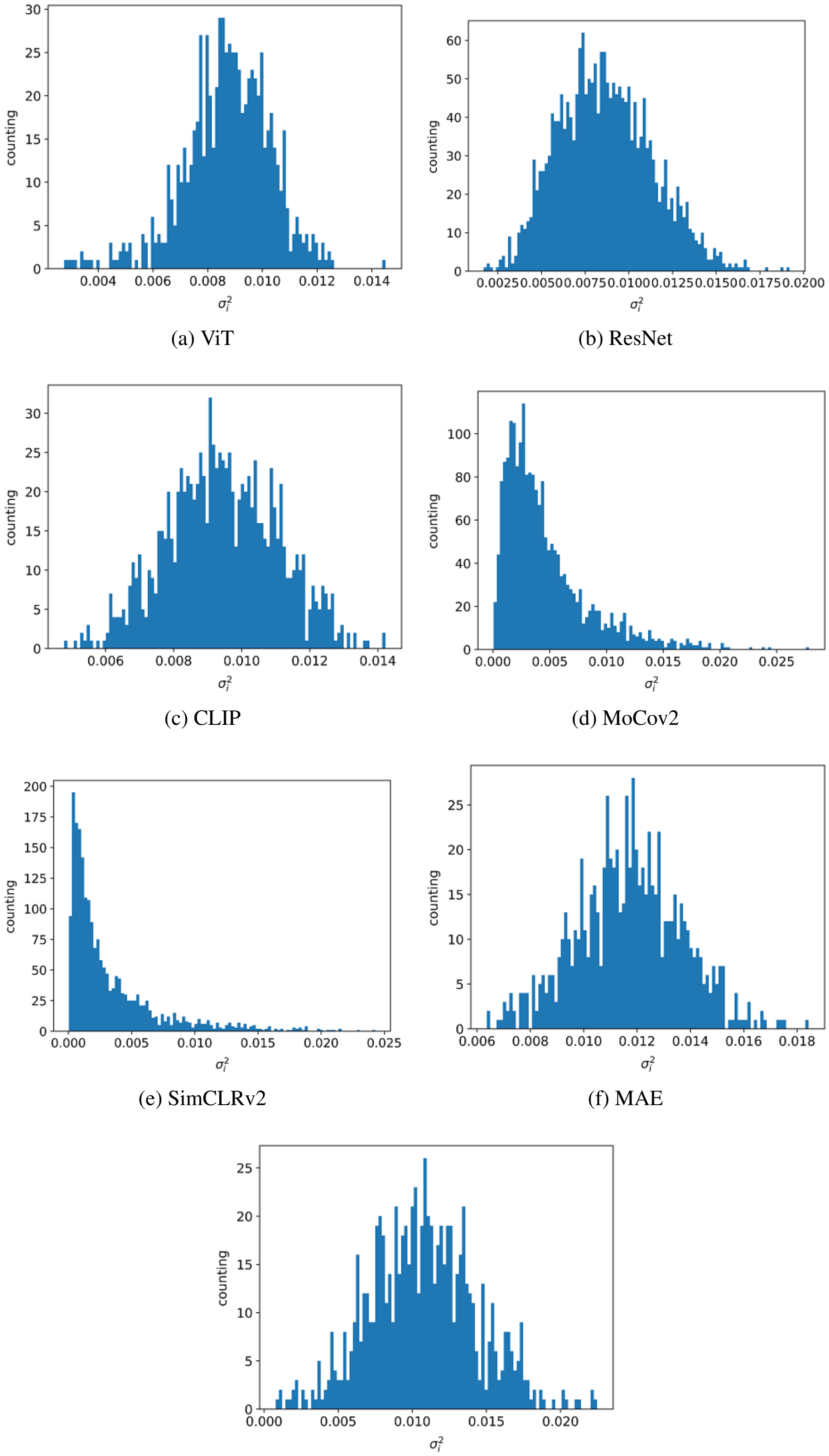
*Table 3.* The used codes and licenses.

| URL | citations | License |
|---|---|---|
| https://github.com/google-research/vision_transformer | (Dosovitskiy et al., 2021) | Apache-2.0 License |
| https://github.com/pytorch/pytorch | (Paszke et al., 2019) | License |
| https://github.com/openai/CLIP | (Radford et al., 2021) | MIT License |
| https://github.com/facebookresearch/moco | (Chen et al., 2020d) | MIT License |
| https://github.com/google-research/simclr | (Chen et al., 2020c) | Apache-2.0 License |
| https://github.com/Separius/SimCLRv2-Pytorch | - | GPL-3.0 license |
| https://github.com/facebookresearch/mae | (He et al., 2022) | License |
| https://github.com/microsoft/SimMIM | (Xie et al., 2022) | MIT License |
| https://github.com/scikit-learn/scikit-learn | (Pedregosa et al., 2011) | BSD-3-Clause License |

### I.2. Feature preprocessing

For the reason that our theory assumes that $\mathcal{X} = [0,1]^n$, we scale each dimension of features to $[0,1]$. It is implemented by using the MinMaxScaler supported in scikit-learn's (Pedregosa et al., 2011). Empirically, we note this transformation will not influence the happening of the "two regimes" phenomenon in practice.

### I.3. Additional Results of Validating the Assumptions



(a) ViT

(b) ResNet

(c) CLIP

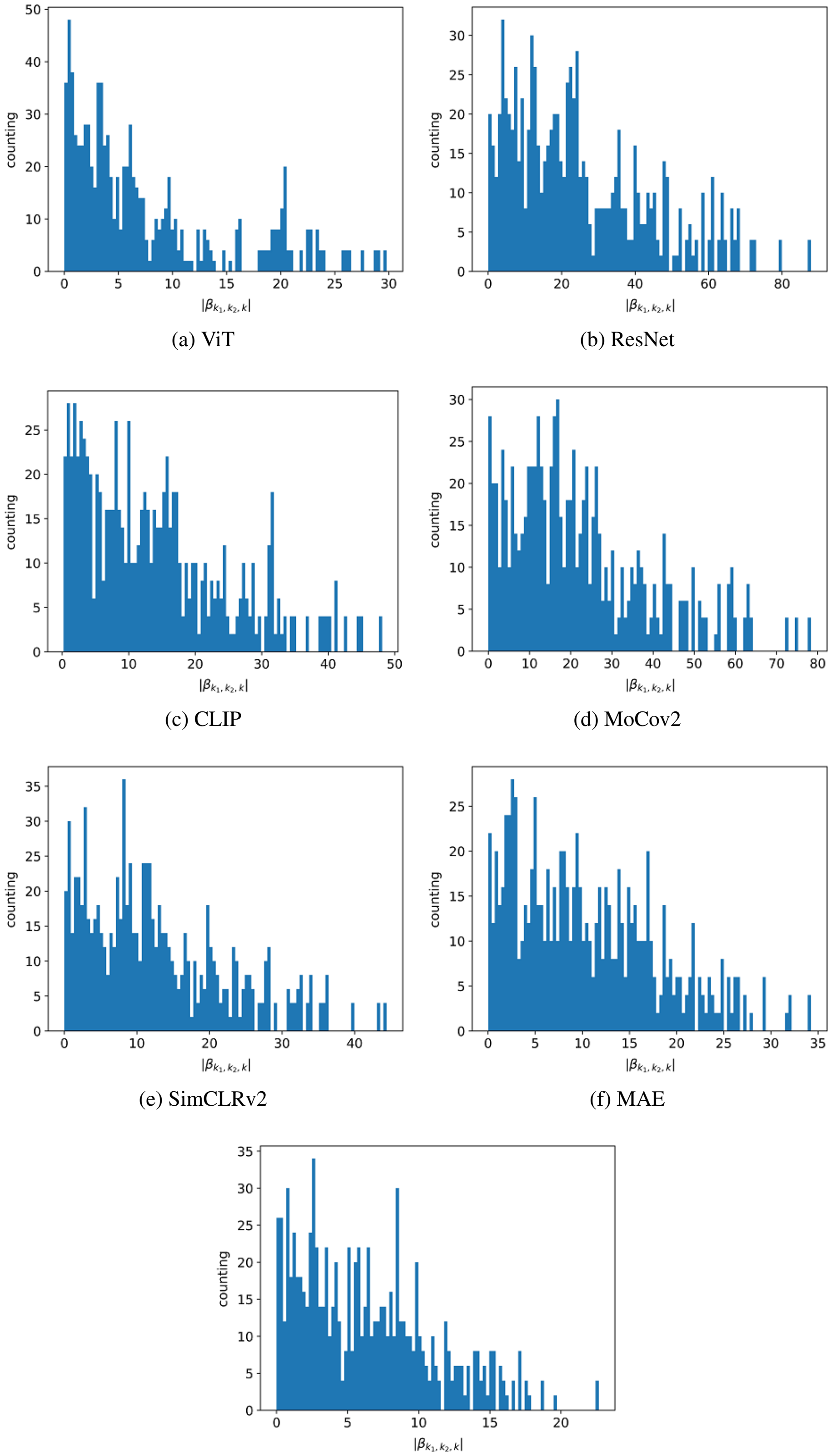
(d) MoCov2

(e) SimCLRv2

(f) MAE

(g) SimMIM

*Figure 5.* Distribution histogram of $\sigma_i^2$



(a) ViT

(b) ResNet

(c) CLIP

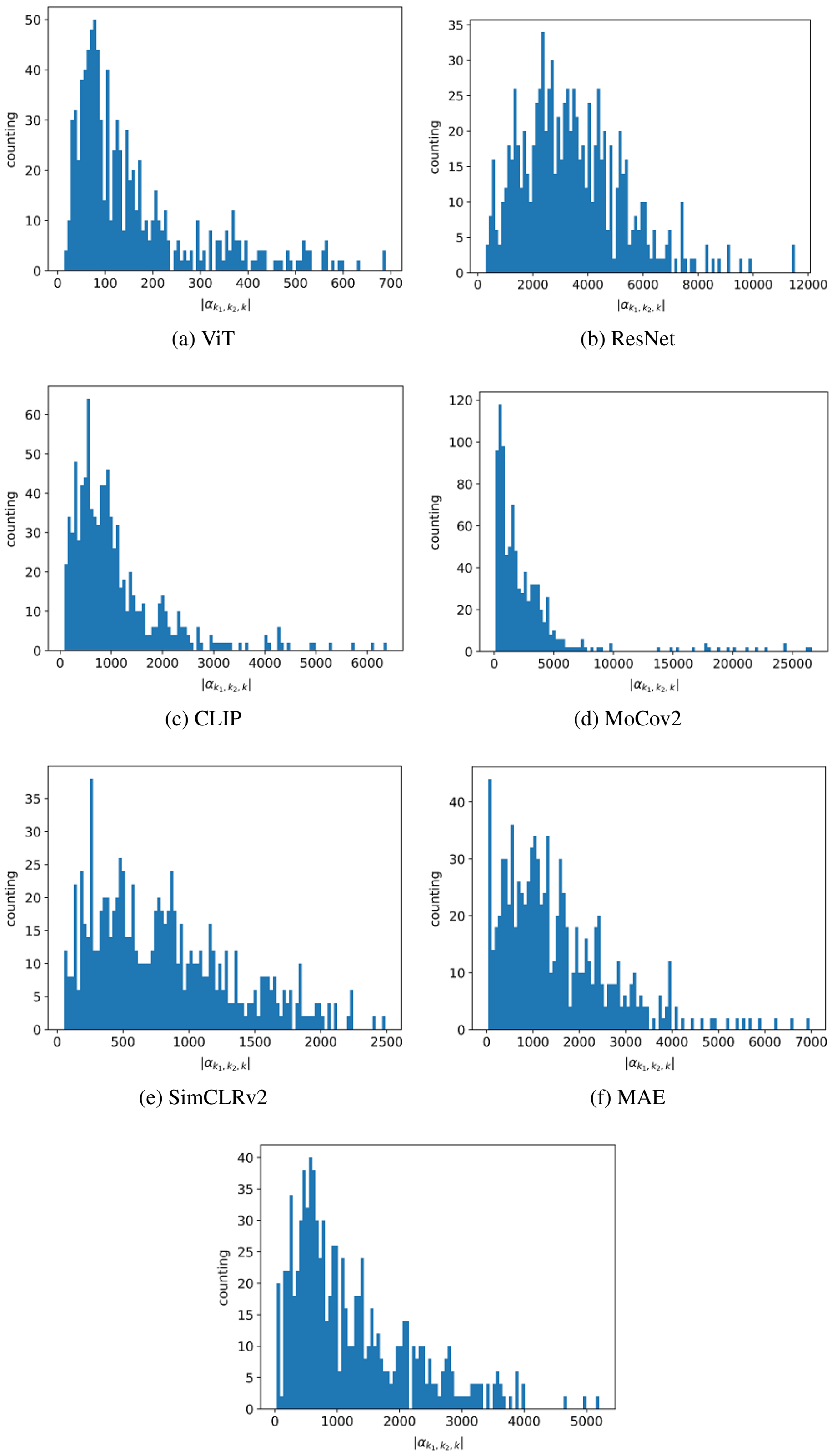
(d) MoCov2

(e) SimCLRv2

(f) MAE

(g) SimMIM

*Figure 6.* Distribution histogram of $|\beta_{k_1,k_2,k}|$.



(a) ViT

(b) ResNet

(c) CLIP

(d) MoCov2

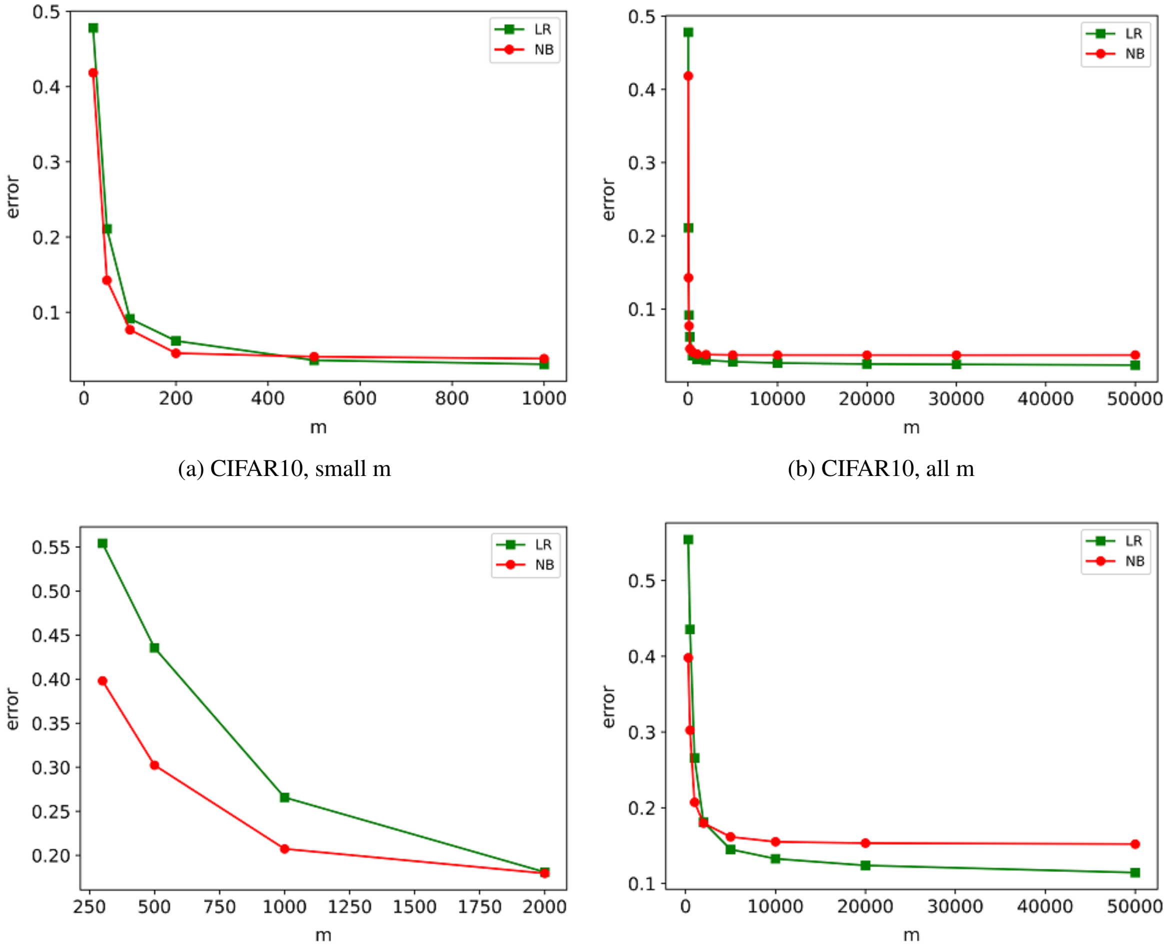
(e) SimCLRv2

(f) MAE

(g) SimMIM

*Figure 7.* Distribution histogram of $\alpha_{k_1,k_2,k}$.

### I.4. Additional Results of Deep Learning



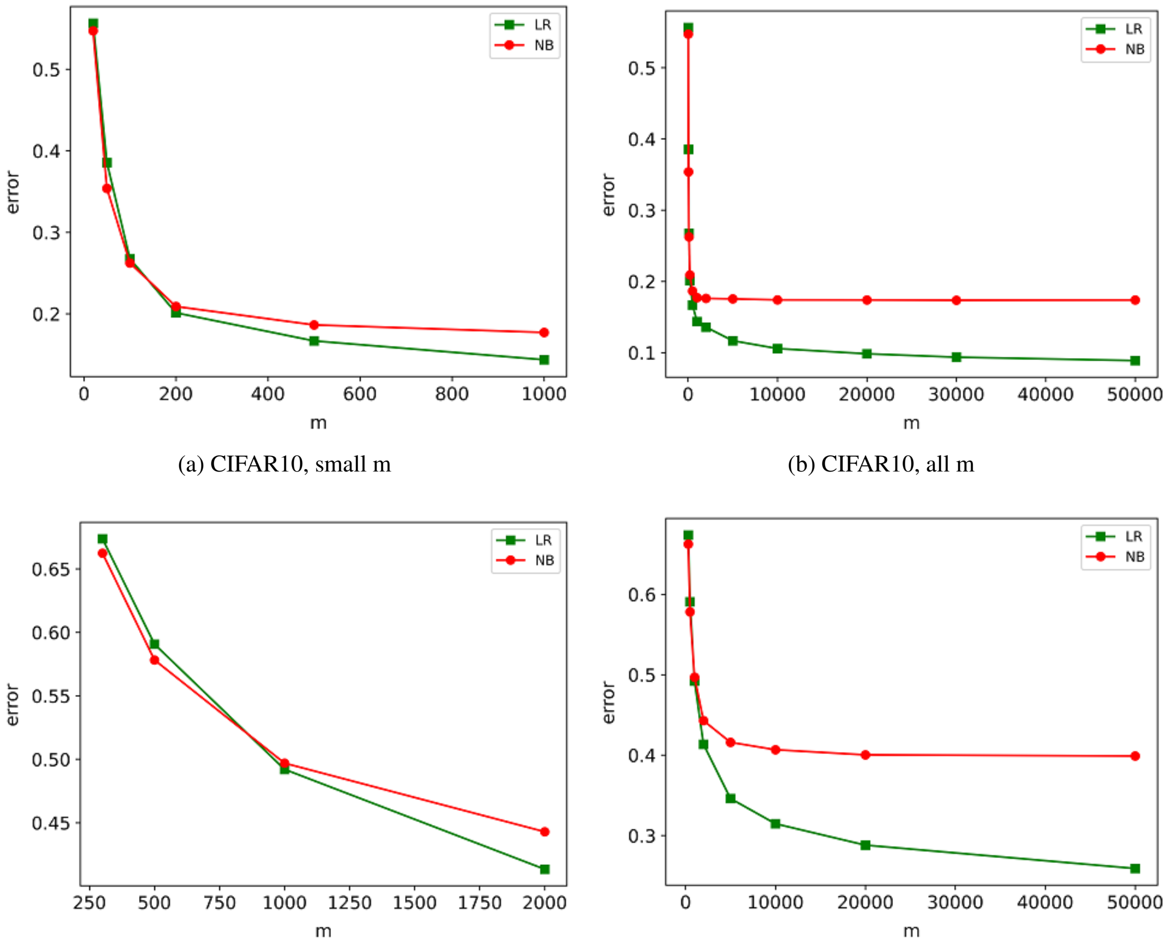
(a) CIFAR10, small m

(b) CIFAR10, all m

(c) CIFAR100, small m

(d) CIFAR100, all m

*Figure 8.* Comparison between naïve Bayes and logistic regression trained on features extracted by ViT.



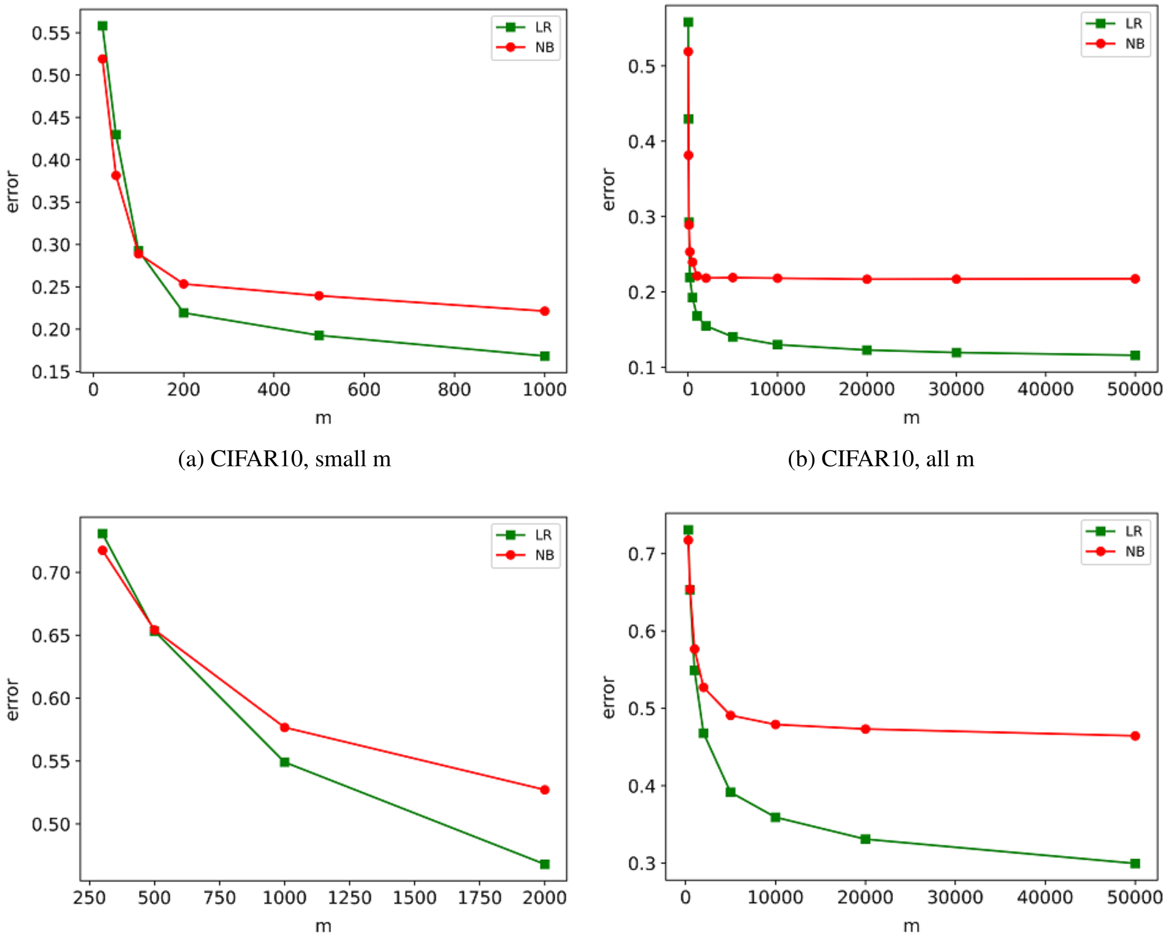
(a) CIFAR10, small m

(b) CIFAR10, all m

(c) CIFAR100, small m

(d) CIFAR100, all m

*Figure 9.* Comparison between naïve Bayes and logistic regression trained on features extracted by ResNet50.



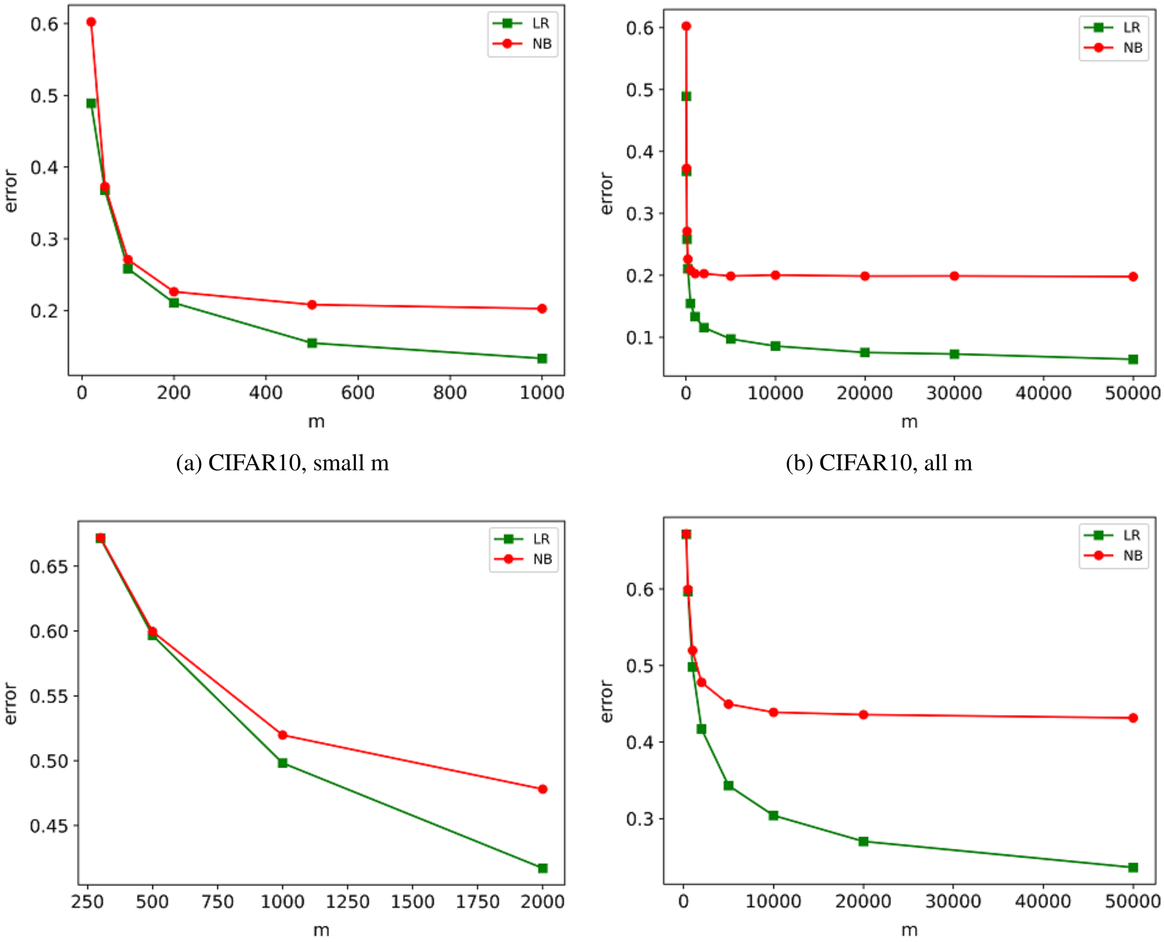
(a) CIFAR10, small m

(b) CIFAR10, all m

(c) CIFAR100, small m

(d) CIFAR100, all m

*Figure 10.* Comparison between naïve Bayes and logistic regression trained on features extracted by CLIP.



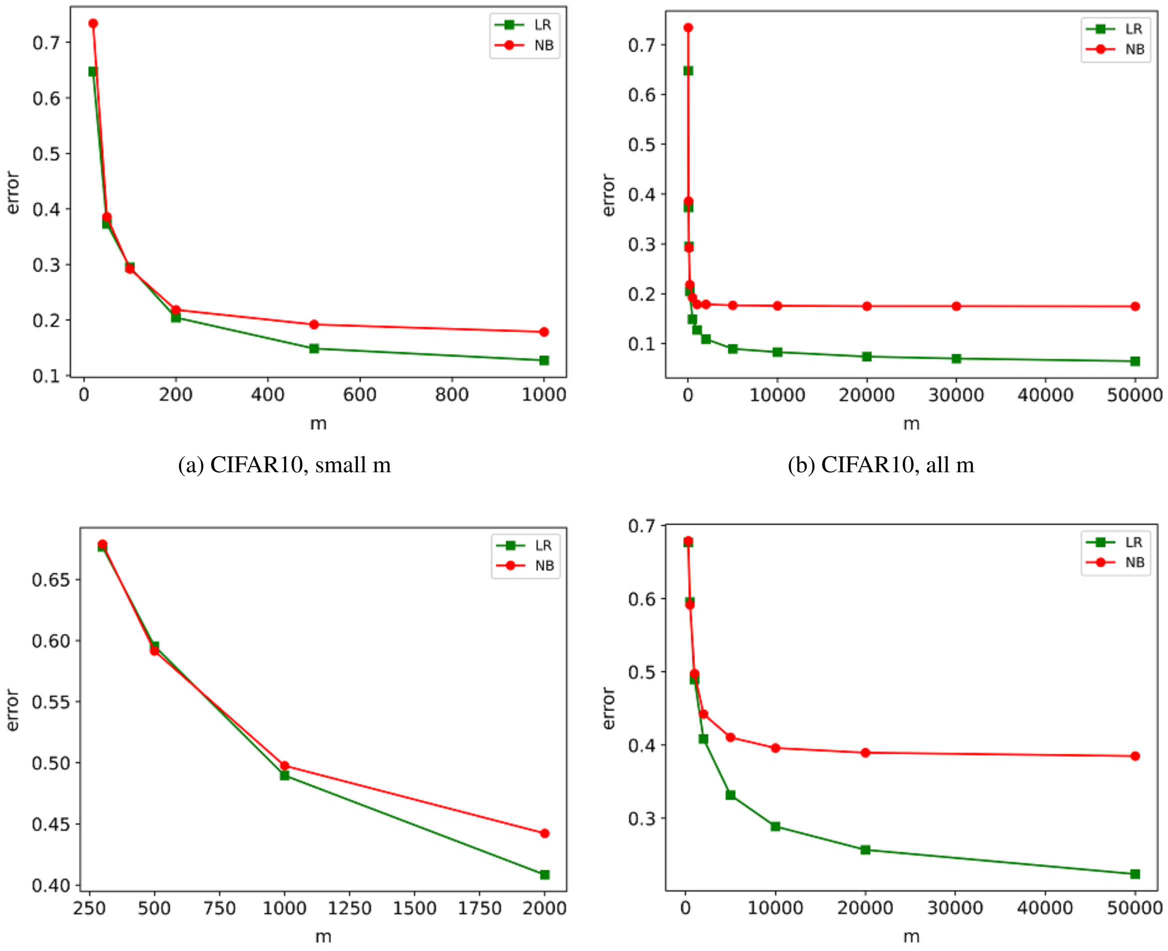
(a) CIFAR10, small m

(b) CIFAR10, all m

(c) CIFAR100, small m

(d) CIFAR100, all m

*Figure 11.* Comparison between naïve Bayes and logistic regression trained on features extracted by MoCov2.



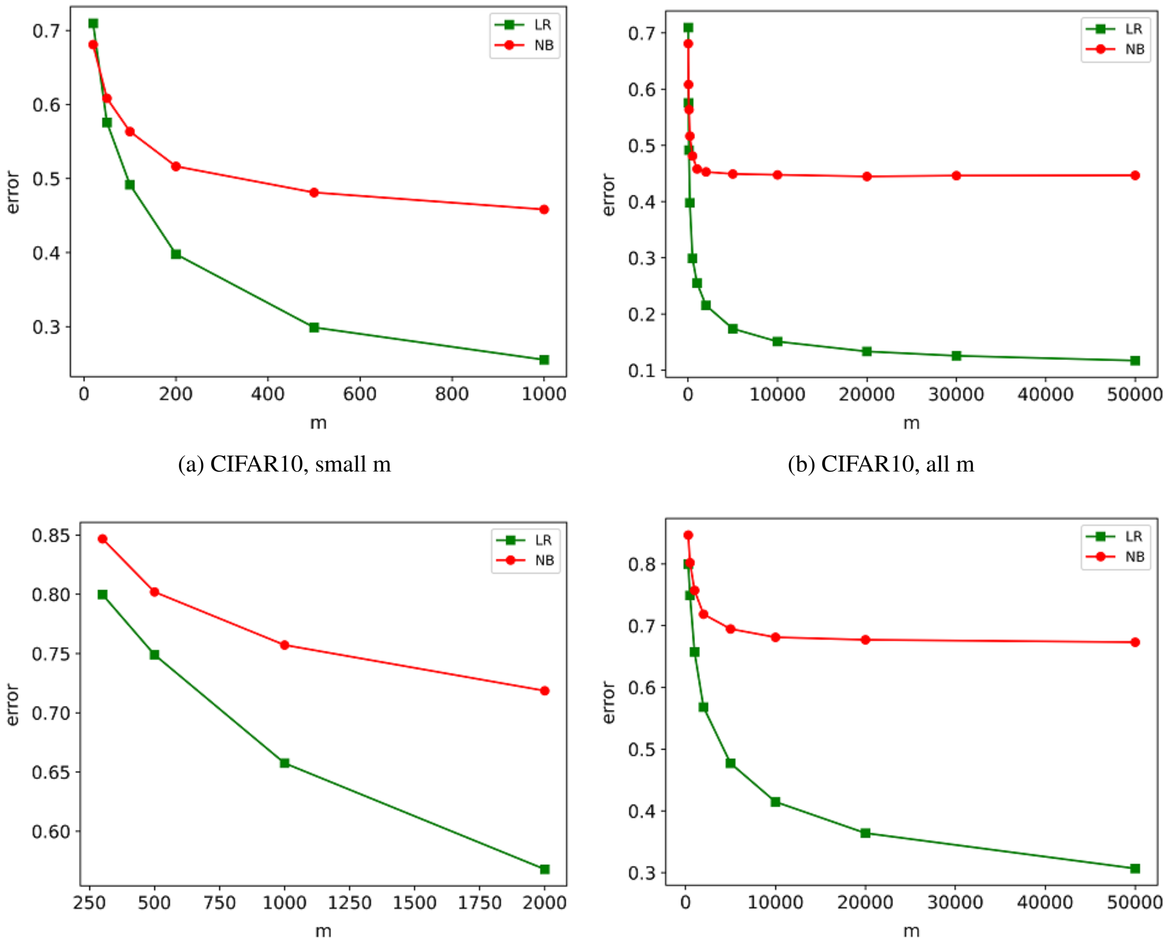
(a) CIFAR10, small m

(b) CIFAR10, all m

(c) CIFAR100, small m

(d) CIFAR100, all m

*Figure 12.* Comparison between naïve Bayes and logistic regression trained on features extracted by SimCLRv2.



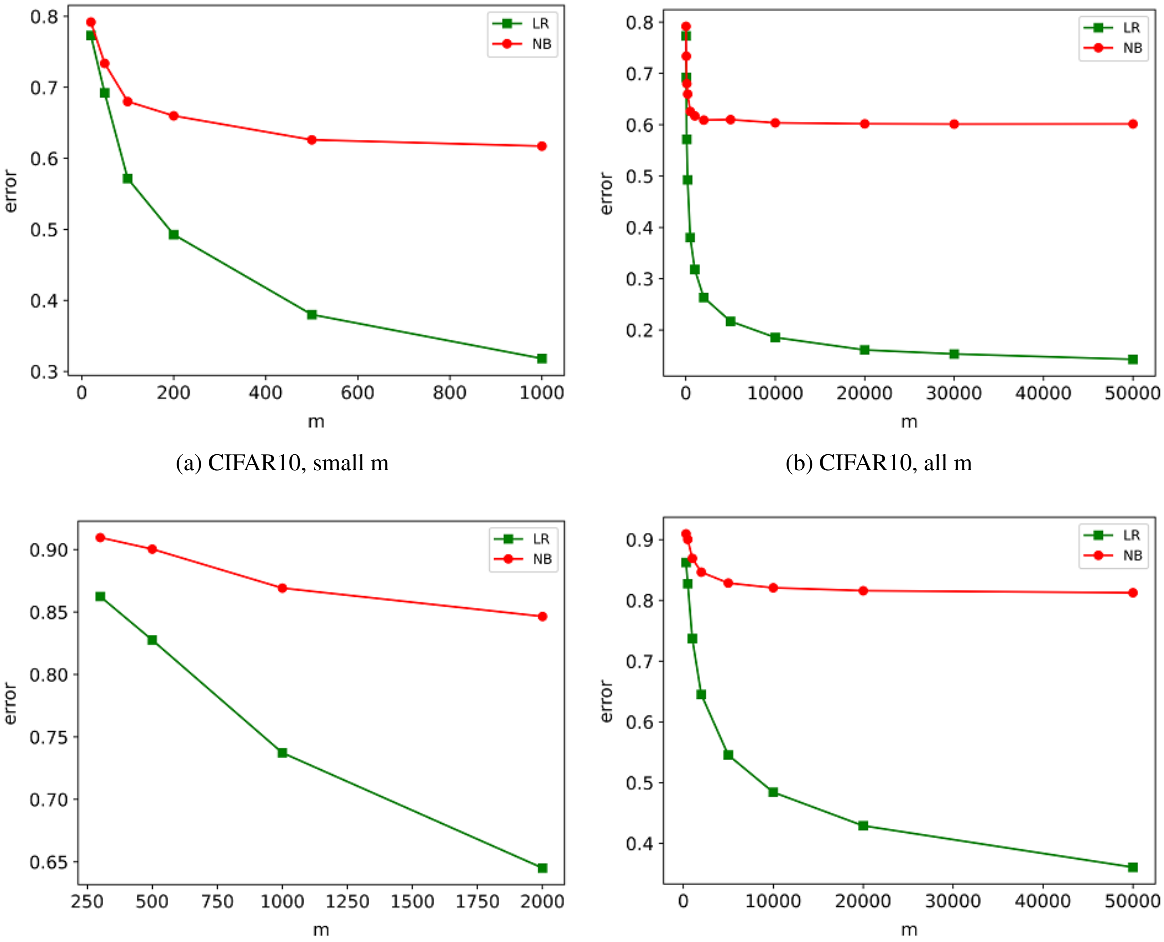
(a) CIFAR10, small m

(b) CIFAR10, all m

(c) CIFAR100, small m

(d) CIFAR100, all m

*Figure 13.* Comparison between naïve Bayes and logistic regression trained on features extracted by MAE.



(a) CIFAR10, small m

(b) CIFAR10, all m

(c) CIFAR100, small m

(d) CIFAR100, all m

*Figure 14.* Comparison between naïve Bayes and logistic regression trained on features extracted by SimMIM.