Improved Balanced Classification with Theoretically Grounded Loss Functions

Corinna Cortes
Google Research
New York, NY 10011
corinna@google.com

Mehryar Mohri
Google Research & CIMS
New York, NY 10011
mohri@google.com

Yutao Zhong
Google Research
New York, NY 10011
yutaozhong@google.com

Abstract

The *balanced loss* is a widely adopted objective for multi-class classification under class imbalance. By assigning equal importance to all classes, regardless of their frequency, it promotes fairness and ensures that minority classes are not overlooked. However, directly minimizing the balanced classification loss is typically intractable, which makes the design of effective surrogate losses a central question. This paper introduces and studies two advanced surrogate loss families: Generalized Logit-Adjusted (GLA) loss functions and Generalized Class-Aware weighted (GCA) losses. GLA losses generalize Logit-Adjusted losses, which shift logits based on class priors, to the broader general cross-entropy loss family. GCA loss functions extend the standard class-weighted losses, which scale losses inversely by class frequency, by incorporating class-dependent confidence margins and extending them to the general cross-entropy family. We present a comprehensive theoretical analysis of consistency for both loss families. We show that GLA losses are Bayes-consistent, but only $\mathcal{H}$-consistent for complete (i.e., unbounded) hypothesis sets. Moreover, their $\mathcal{H}$-consistency bounds depend inversely on the minimum class probability, scaling at least as $1/\mathsf{p}_{\min}$. In contrast, GCA losses are $\mathcal{H}$-consistent for any hypothesis set that is bounded or complete, with $\mathcal{H}$-consistency bounds that scale more favorably as $1/\sqrt{\mathsf{p}_{\min}}$, offering significantly stronger theoretical guarantees in imbalanced settings. We report the results of experiments demonstrating that, empirically, both the GCA losses with calibrated class-dependent confidence margins and GLA losses can greatly outperform straightforward class-weighted losses as well as the LA losses. GLA generally performs slightly better in common benchmarks, whereas GCA exhibits a slight edge in highly imbalanced settings. Thus, we advocate for both GLA and GCA losses as principled, theoretically sound, and state-of-the-art surrogates for balanced classification under class imbalance.

1 Introduction

Class imbalance is a prevalent challenge in real-world multi-class classification problems. Applications such as medical diagnosis, fraud detection, and rare event prediction often involve highly skewed label distributions, where a small subset of classes dominate the data, while others, sometimes the most critical, are heavily underrepresented. Standard training objectives, such as minimizing the unweighted cross-entropy loss, tend to be biased toward majority classes, leading to poor performance on minority classes and undermining the fairness, soundness and reliability of learned models.

To address this issue, a widely studied approach is to minimize the *balanced loss*, which assigns equal importance to all classes regardless of their frequency in the training data [Chan and Stolfo, 1998, Brodersen et al., 2010, Kotlowski et al., 2011, Menon et al., 2013, Cao et al., 2019, Menon

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

et al., 2021, Cui et al., 2019]. This promotes fairness by equalizing performance across demographic groups [Khalili et al., 2023, Hardt et al., 2016] and ensures that minority classes are not overlooked in long-tailed datasets [Feldman, 2020, Zhang et al., 2023] (see Appendix A). It is also crucial in federated learning, where data imbalances across clients can lead to biased models that favor heavy users [Li et al., 2021, McMahan et al., 2017, Mohri et al., 2019]. By reweighting the loss contributions from different classes, the balanced loss promotes equitable treatment of all labels and has been shown to better align with metrics such as balanced accuracy and macro-F1. However, directly optimizing the balanced classification loss is typically intractable in practice. Thus, the design of effective surrogate losses that are tractable to optimize is a central challenge in imbalanced learning.

This paper introduces and studies two families of surrogate losses: Generalized Logit-Adjusted (GLA) loss functions and Generalized Class-Aware weighted (GCA) losses. GLA losses generalize Logit-Adjusted losses [Menon et al., 2021], which shift logits based on class priors, to the broader general cross-entropy loss family [Mao et al., 2023f]. GCA loss functions extend the standard class-weighted losses, which scale losses inversely by class frequency, by incorporating class-dependent confidence margins and extending them to the general cross-entropy family.

We present a comprehensive theoretical analysis of their consistency. We show that GLA losses are Bayes-consistent [Zhang, 2004a, Bartlett et al., 2006, Zhang, 2004b, Tewari and Bartlett, 2007, Steinwart, 2007], but only $\mathcal{H}$-consistent [Awasthi et al., 2022a,b, Mao et al., 2023f,b] for complete (i.e., unbounded) hypotheses. Moreover, their $\mathcal{H}$-consistency bounds depend inversely on the minimum class probability, $\mathsf{p}_{\min}$, scaling at least as $1/\mathsf{p}_{\min}$. In contrast, GCA losses are $\mathcal{H}$-consistent for any hypothesis set that is bounded or complete, with $\mathcal{H}$-consistency bounds that scale more favorably as $1/\sqrt{\mathsf{p}_{\min}}$, offering significantly stronger theoretical guarantees in imbalanced settings.

We also report the results of experiments demonstrating that, empirically, both the GCA losses with calibrated class-dependent confidence margins and GLA losses comfortably outperform straightforward class-weighted losses as well as the LA losses. GLA generally performs slightly better in common benchmarks, whereas GCA exhibits a slight edge in highly imbalanced settings.

Taken together, our results establish GLA and GCA losses as theoretically grounded and practically effective classification algorithms for tackling class imbalance in multi-class learning. Their complementary strengths make them well-suited for a wide range of real-world applications where fairness across classes is paramount.

The rest of this paper is structured as follows. Section 3 reviews fundamental concepts related to class imbalance in multi-class classification, introduces the balanced loss (Section 3.1), discusses existing surrogate losses (Section 3.2), and highlights the limitations of current approaches (Section 3.3). Section 4 introduces two novel surrogate loss families: Generalized Logit-Adjusted (GLA) (Section 4.1) and Generalized Class-Aware weighted (GCA) losses (Section 4.2). A comprehensive theoretical analysis of their consistency and margin bounds is provided in Section 5 and Appendix B. Finally, Section 6 reports empirical results on CIFAR-10, CIFAR-100, and Tiny ImageNet, demonstrating the effectiveness of our algorithms, which are based on the minimization of these loss functions.

2 Preliminaries

Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = [n] \coloneqq \{1, \ldots, n\}$ represent the set of n possible labels. We consider a data distribution $\mathcal{D}$ over the combined input-label space $\mathcal{X} \times \mathcal{Y}$. Our hypothesis set, denoted by $\mathcal{H}$, consists of functions that map an input-label pair (x, y) to a real-valued score, $h \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. We denote by $\mathsf{p}(x)$ the marginal probability density of an input x, and by $\mathsf{p}(y)$ the marginal probability of a class label y. The minimum class marginal is defined as $\mathsf{p}_{\min} = \min_{y \in \mathcal{Y}} \mathsf{p}(y)$. The conditional distributions $\mathsf{p}(x \mid y)$ and $\mathsf{p}(y \mid x)$ represent the probability of input x given label y, and label y given input x, respectively.

Let $\mathcal{H}_{\text{all}}$ denote the set of all measurable functions, and a $\ell \colon \mathcal{H}_{\text{all}} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ the loss function adopted to penalize inaccurate predictions. Then, the *generalization error* of a hypothesis $h \in \mathcal{H}$ is defined as its expected loss: $\mathcal{R}_\ell(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h, x, y)]$. The lowest possible generalization error achievable within the hypothesis set $\mathcal{H}$ is the *best-in-class generalization error*, $\mathcal{R}^*_\ell(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{R}_\ell(h)$.

For any input $x \in \mathcal{X}$, a hypothesis $h \in \mathcal{H}$ assigns a predicted label $\mathsf{h}(x)$ by selecting the class with the highest score: $\mathsf{h}(x) = \operatorname{argmax}_{y \in \mathcal{Y}} h(x, y)$ (ties are broken by choosing the highest index). The

standard *zero-one loss function* for multi-class classification is defined as $\ell_{0-1}(h, x, y) \coloneqq \mathbb{1}_{\mathsf{h}(x)\neq y}$, which is 1 if the prediction is incorrect and 0 otherwise.

The *margin* $\rho_h(x, y)$ for a predictor $h \in \mathcal{H}$ on a labeled example (x, y) measures the confidence of the correct prediction: $\rho_h(x, y) = h(x, y) - \max_{y' \neq y} h(x, y')$. This is the difference between the score of the true label y and the highest score among all other labels y'.

The generalization error of a hypothesis h can also be expressed as the expectation of the *conditional error* over the input x: $\mathcal{R}_\ell(h) = \mathbb{E}_x[\mathcal{C}_\ell(h, x)]$, where $\mathcal{C}_\ell(h, x) = \sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)\ell(h, x, y)$. The *best-in-class conditional error* is $\mathcal{C}^*_\ell(\mathcal{H}, x) = \inf_{h\in\mathcal{H}} \mathcal{C}_\ell(h, x)$. The difference, $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x) = \mathcal{C}_\ell(h, x) - \mathcal{C}^*_\ell(\mathcal{H}, x)$, is termed the *conditional regret* for the loss function ℓ. These concepts and definitions are useful in our analysis of the consistency of loss functions.

3 Background and Related Work

We first review fundamental concepts related to class imbalance in multi-class classification, introduce the balanced loss, discuss existing surrogate losses, and highlight the limitations of current approaches.

3.1 Class Imbalance and Balanced Loss

Class imbalance in multi-class settings arises when the label distribution $\mathsf{p}(y)$ is highly skewed, with some classes (often referred to as "tail" labels) having much lower probabilities of occurrence compared to others (the "head" or majority classes). In such cases, many recent studies [Chan and Stolfo, 1998, Brodersen et al., 2010, Kotlowski et al., 2011, Menon et al., 2013, Cao et al., 2019, Menon et al., 2021, Cui et al., 2019] suggest that the balanced loss (ℓ_{BAL}) is a more appropriate loss function than the standard zero-one loss. The balanced loss assigns equal importance to all classes, irrespective of their frequency, and is thus viewed as promoting fairness by equalizing performance across demographic groups [Khalili et al., 2023, Hardt et al., 2016, Conitzer et al., 2019] and ensuring minority classes are not overlooked in long-tailed datasets [Feldman, 2020, Zhang et al., 2023] (see Appendix A). It is also crucial in federated learning, where data imbalances across clients can lead to biased models that favor majority users [Li et al., 2021, McMahan et al., 2017, Mohri et al., 2019].

The balanced loss reduces the influence of class imbalances by averaging the per-class loss by weighting the error for each example (h, x, y) by the inverse of the probability of the true class $\mathsf{p}(y)$:

$$\ell_{\mathrm{BAL}}(h, x, y) = \frac{1_{\mathsf{h}(x)\neq y}}{\mathsf{p}(y)}. \tag{1}$$

The following lemma characterizes the best-in-class conditional error and the corresponding conditional regret for the balanced loss. For any input $x \in \mathcal{X}$, we denote by $\mathsf{H}(x)$ the set of labels that can be predicted by hypotheses in $\mathcal{H}$ for that input: $\mathsf{H}(\mathsf{x}) = \{\mathsf{h}(\mathsf{x}) : \mathsf{h} \in \mathcal{H}\}$. The proof of Lemma 1 is provided in Appendix D.

Lemma 1. *For any $x \in \mathcal{X}$, the best-in-class conditional error and the conditional regret for ℓ_{BAL} can be expressed as follows:*

$$\mathcal{C}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}, x) = \sum_{y\in\mathcal{Y}} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} - \max_{y\in\mathsf{H}(\mathsf{x})} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} \qquad \Delta\mathcal{C}_{\ell_{\mathrm{BAL}},\mathcal{H}}(h, x) = \max_{y\in\mathsf{H}(\mathsf{x})} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} - \frac{\mathsf{p}(\mathsf{h}(x)) \mid x)}{\mathsf{p}(\mathsf{h}(x))}.$$

3.2 Existing Surrogate Losses for Balanced Learning

Several surrogate losses have been proposed for optimizing the balanced loss. Here, we review two prominent Bayes-consistent examples:

Class-Weighted Cross-Entropy: A common strategy is to use the class-weighted cross-entropy loss [Xie and Manski, 1989, Morik et al., 1999], which adjusts the standard cross-entropy loss by weighting each example inversely proportional to its class frequency $\mathsf{p}(y)$:

$$\ell_{\mathrm{WCE}}(h, x, y) = -\frac{1}{\mathsf{p}(y)} \log\left(\frac{e^{h(x,y)}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')}}\right). \tag{2}$$

As pointed by [Byrd and Lipton, 2019], the limitation of ℓ_{WCE} is that in separable cases, class-weighted cross-entropy may still yield solutions with zero training loss that do not adjust decision

boundaries meaningfully toward minority or majority classes. This is because class weighting does not influence the classifier once perfect separation is achieved. As a result, the method fails to address imbalance in such regimes.

Logit-Adjusted (LA) Losses: More recently, Menon et al. [2021] introduced Logit-Adjusted (LA) losses. These losses modify the logits (outputs before softmax) based on class priors, typically by adding a term $\tau \log(\mathsf{p}(y))$ with $\tau > 0$:

$$\ell_{\mathrm{LA}}(h, x, y) = -\log\left(\frac{e^{h(x,y)+\tau \log(\mathsf{p}(y))}}{\sum_{y' \in \mathcal{Y}} e^{h(x,y')+\tau \log(\mathsf{p}(y'))}}\right). \tag{3}$$

As we will show in Section 5, ℓ_{LA} is not Bayes-consistent for the balanced loss when $\tau \neq 1$.

A detailed discussion of other approaches for handling class imbalance, including alternative loss weighting schemes [Cui et al., 2019, Fan et al., 2017, Jamal et al., 2020, Wang et al., 2023, 2025, Li et al., 2025], margin modifications [Masnadi-Shirazi and Vasconcelos, 2010, Iranmehr et al., 2019, Zhang et al., 2017, Cao et al., 2019, Tan et al., 2020, Jiawei et al., 2020], data augmentation and sampling techniques [Kubat and Matwin, 1997, Wallace et al., 2011, Chawla et al., 2002, Yin et al., 2018], threshold adjustments [Fawcett and Provost, 1996, Provost, 2000, Maloof, 2003, King and Zeng, 2001, Collell et al., 2016, Menon et al., 2021, Zhu et al., 2023], and weight normalization methods [Zhang et al., 2019a, Kim and Kim, 2019, Kang et al., 2020] is included in Appendix A.

3.3 Limitations of Existing Approaches

Despite their usefulness, existing surrogate losses and related methods admit some limitations. *Class-weighted cross-entropy* often has a minimal effect in settings where data is easily separable. In such cases, solutions that achieve zero training loss (perfect separation) remain optimal even with class weighting, failing to shift decision boundaries effectively towards dominant classes as might be desired [Byrd and Lipton, 2019]. *Logit-Adjusted (LA) losses*, as we will demonstrate in Section 5, are not Bayes-consistent for the balanced loss when the temperature parameter $\tau \neq 1$. Consequently, optimal tuning of τ often lacks a theoretical guarantee, and the method itself offers limited flexibility. *Other margin modification techniques* [e.g., Cao et al., 2019, Tan et al., 2020] may not be Bayes-consistent for the balanced loss, even in simpler binary classification problems [Menon et al., 2021]. *The drawbacks of other strategies* beyond direct loss modification, such as weight normalization, have also been previously noted [Menon et al., 2021].

4 Surrogate Loss Families

This section generalizes two surrogate loss families designed for learning with class imbalance: Generalized Logit-Adjusted (GLA) loss functions and Generalized Class-Aware weighted (GCA) losses. Both families are derived from the general cross-entropy (GCE) framework [Mao et al., 2023f]. For any $(h, x, y) \in \mathcal{H} \times \mathcal{X} \times \mathcal{Y}$, the GCE loss is defined as:

$$\ell_{\mathrm{GCE}}(h, x, y) = \Psi^q\left(\frac{e^{h(x,y)}}{\sum_{y' \in \mathcal{Y}} e^{h(x,y')}}\right), \quad \text{with} \quad \Psi^q(t) = \begin{cases} -\log(t) & \text{if } q = 0 \\ \frac{1}{q}(1 - t^q) & \text{if } q \in (0, \infty). \end{cases}$$

Specific choices of q recover well-known loss functions: $q = 0$ yields the *logistic loss* (or standard cross-entropy) [Verhulst, 1838, 1845, Berkson, 1944, 1951]; $q \in (0, 1)$ gives the *generalized cross-entropy loss* notable for its robustness to label noise [Zhang and Sabuncu, 2018]; and $q = 1$ corresponds to the *mean absolute error loss* [Ghosh et al., 2017].

4.1 Generalized Logit-Adjusted (GLA) Losses

A *Generalized Logit-Adjusted (GLA) Loss* modifies the logits within the GCE family by incorporating a class-prior-based bias term, $\log(\mathsf{p}(y))/(1-q)$:

$$\ell_{\mathrm{GLA}}(h, x, y) = \Psi^q\left(\frac{e^{h(x,y)+\frac{\log(\mathsf{p}(y))}{1-q}}}{\sum_{y' \in \mathcal{Y}} e^{h(x,y')+\frac{\log(\mathsf{p}(y'))}{1-q}}}\right), \tag{4}$$

The GLA loss family generalizes the Logit-Adjusted (LA) loss with $\tau = 1$. Specifically, when $q = 0$, Eq. (4) recovers the LA loss with $\tau = 1$ previously defined in Eq. (3). Thus, GLA extends the concept

of logit adjustment to the broader GCE family. As will be detailed in Section 5.2, GLA losses are Bayes-consistent for any $q \in [0, 1)$, offering greater flexibility compared to the original LA loss (whose limitations were discussed in Section 3.3).

The term inside the Ψ^q function in Eq. (4) can be rewritten to highlight its behavior:

$$\frac{e^{h(x,y)+\frac{\log(\mathsf{p}(y))}{1-q}}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')+\frac{\log(\mathsf{p}(y'))}{1-q}}} = \frac{e^{h(x,y)} \cdot \mathsf{p}(y)^{\frac{1}{1-q}}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')} \cdot \mathsf{p}(y')^{\frac{1}{1-q}}} = \frac{1}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')-h(x,y)} \cdot \left(\frac{\mathsf{p}(y')}{\mathsf{p}(y)}\right)^{\frac{1}{1-q}}}.$$

In this formulation, the term $\left(\mathsf{p}(y')/\mathsf{p}(y)\right)^{\frac{1}{1-q}}$ acts as a weighting factor in the denominator, effectively creating a pairwise label margin adjustment that depends on the relative frequencies of class y (the true class) and other classes y'. This mechanism encourages a larger separation (margin) when y is a rare class (low $\mathsf{p}(y)$) and y' is a dominant class (high $\mathsf{p}(y')$) and reduces the risk that scores for dominant classes overshadow those for rare classes.

4.2 Generalized Class-Aware (GCA) Losses

A *Generalized Class-Aware (GCA) loss* introduces class sensitivity by inversely weighting the GCE loss by class frequency $\mathsf{p}(y)$ and incorporating class-dependent confidence margins ρ_y:

$$\ell_{\mathrm{GCA}}(h, x, y) = \frac{1}{\mathsf{p}(y)} \Psi^q\left(\frac{e^{h(x,y)/\rho_y}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')/\rho_y}}\right), \tag{5}$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)$ is a vector of positive confidence margin parameters for each class. The GCA formulation encompasses standard class-weighting as a special case. For instance, the class-weighted cross-entropy loss (Eq. (2)) is recovered when $q = 0$ and all confidence margins ρ_y are set to 1. If all $\rho_y = 1$, Eq. (5) simplifies to: $\ell_{\mathrm{GCA}}(h, x, y) = \frac{1}{\mathsf{p}(y)} \Psi^q\left(\frac{e^{h(x,y)}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')}}\right)$, thereby extending the class-weighted cross-entropy concept to the entire GCE family. The motivation for using the inverse of the prior in GCA remains the same for $q \neq 1$ as for $q = 1$. The parameter q simply specifies a particular loss within the generalized cross-entropy family, applicable in both standard and imbalanced settings. The inverse of the prior is used to align with the definition of the balanced loss, which reduces the influence of class imbalance by reweighting each example's error accordingly. This ensures that GCA losses benefit from consistency guarantees with respect to the balanced loss.

The introduction of distinct confidence margin parameters $\boldsymbol{\rho}$ is a key aspect of GCA losses. These parameters allow for fine-tuned adjustments to the decision boundaries. By applying class-specific scaling with factors related to ρ_y to the logit differences $[h(x, y) - h(x, y')]$-terms that inherently represent margins, the GCA loss (through an effective transformation to $(h(x, y) - h(x, y'))/\rho_y$) can more effectively separate dominant and rare classes, as such transformation modulates how confidently each class needs to be separated. Such margin adjustments, as highlighted by recent work of Cortes et al. [2025], play a crucial role in effectively shifting decision boundaries across classes and mitigating imbalance. This, in turn, addresses the limitations of simpler class-weighting schemes mentioned in Section 3.3.

Note that while the ρ_k values can be treated as tunable hyperparameters and freely tuned via cross-validation, the search can be effectively guided by focusing on vectors $[\rho_k]_k$ near $[m_k^{1/3}]_k$, where m_k denotes the number of samples in class k, as suggested by Cortes et al. [2025] and followed in our experiments. A similar derivation to theirs, adaptable to our setting, shows these values are theoretically optimal in a separable case, providing justification and guidance for selecting ρ_k for GCA losses. Empirically, we also found GCA losses to be robust to variations in ρ_k around these values. Consequently, while ρ_k can be tuned, the default choice of $m_k^{1/3}$ performs well. When the number of classes n is large, the search space can be further reduced by assigning identical ρ_k values to underrepresented classes and reserving distinct values for the most frequent ones.

For fixed hyperparameters, the computational cost of GLA and GCA losses is comparable to that of standard neural networks trained with cross-entropy loss (that is, logistic loss with softmax) and to that of the baselines. Our loss functions are adapted from the general cross-entropy family and both share similar convergence behavior and remain practical when optimized with commonly used optimizers such as SGD, Adam, and AdaGrad. While our methods introduce additional hyperparameters, namely

ρ_k and q in GCA losses and q in GLA losses, the value of ρ_k has a default choice (as discussed above), and q serves a similar role to hyperparameters in the baseline methods listed in Table 1 in Section 6, many of which also involve at least one extra tunable parameter.

5 Theoretical Analysis

In this section, we leverage Lemma 1 to present a comprehensive theoretical analysis of the consistency for the two proposed surrogate loss families: Generalized Logit-Adjusted (GLA) losses and Generalized Class-Aware (GCA) losses.

5.1 Consistency Notions

A critical characteristic of a surrogate loss function ℓ_A, used in place of a target loss function ℓ_B, is its *Bayes-consistency* [Steinwart, 2007]. This property ensures that if a sequence of predictor $\{h_n\}_{n\in\mathbb{N}}$ within $\mathcal{H}_{\text{all}}$ (the set of all measurable functions) asymptotically minimizes the surrogate loss ℓ_A, it will also asymptotically minimize the target loss ℓ_B. Formally: $\lim_{n\to+\infty} \mathcal{R}_{\ell_A}(h_n) = \mathcal{R}^*_{\ell_A}(\mathcal{H}_{\text{all}}) \Rightarrow \lim_{n\to+\infty} \mathcal{R}_{\ell_B}(h_n) = \mathcal{R}^*_{\ell_B}(\mathcal{H}_{\text{all}})$. However, Bayes-consistency is an asymptotic concept and is defined only for the comprehensive class of all measurable functions $\mathcal{H}_{\text{all}}$. A more practically relevant and informative concept is that of $\mathcal{H}$-*consistency bounds*. These bounds are non-asymptotic and tailored to a specific hypothesis class $\mathcal{H}$ [Awasthi et al., 2022a,b, 2021a,b, 2023a,b, Mao et al., 2023a,b,c,d,e,f, 2024a,b,c,d,e,f,g, Mohri et al., 2024, Cortes et al., 2024, Mao et al., 2025a,b, Mao, 2025, Zhong, 2025]). In the realizable setting, these bounds take the form:

$$\forall h \in \mathcal{H}, \quad \mathcal{R}_{\ell_B}(h) - \mathcal{R}^*_{\ell_B}(\mathcal{H}) \leq \Gamma\Big(\mathcal{R}_{\ell_A}(h) - \mathcal{R}^*_{\ell_A}(\mathcal{H})\Big).$$

Here, Γ is a non-increasing concave function such that $\Gamma(0) = 0$. In the more general non-realizable setting, the bound is augmented by a *minimizability gap*, $\mathcal{M}_\ell(\mathcal{H}) = \mathcal{R}^*_\ell(\mathcal{H}) - \mathbb{E}_x[\mathcal{C}^*_\ell(\mathcal{H}, x)]$. This gap quantifies the difference between the best-in-class error and the expected best-in-class conditional error. The augmented bound is:

$$\mathcal{R}_{\ell_B}(h) - \mathcal{R}^*_{\ell_B}(\mathcal{H}) + \mathcal{M}_{\ell_B}(\mathcal{H}) \leq \Gamma\Big(\mathcal{R}_{\ell_A}(h) - \mathcal{R}^*_{\ell_A}(\mathcal{H}) + \mathcal{M}_{\ell_A}(\mathcal{H})\Big).$$

As demonstrated by Mao et al. [2024h], the minimizability gap is always non-negative and is bounded above by the approximation error $\mathcal{A}_\ell(\mathcal{H}) = \mathcal{R}^*_\ell(\mathcal{H}) - \mathcal{R}^*_\ell(\mathcal{H}_{\text{all}})$, i.e., $0 \leq \mathcal{M}_\ell(\mathcal{H}) \leq \mathcal{A}_\ell(\mathcal{H})$. The minimizability gap becomes zero when $\mathcal{H} = \mathcal{H}_{\text{all}}$ or, more generally, when the approximation error $\mathcal{A}_\ell(\mathcal{H}) = 0$. In other cases, it is typically non-zero and offers a more refined measure than the approximation error. In particular, $\mathcal{H}$-consistency bounds imply Bayes-consistency when $\mathcal{H} = \mathcal{H}_{\text{all}}$ and generally provide stronger and more applicable guarantees.

5.2 GLA Losses

We now analyze the consistency properties of the GLA loss family. We establish that the LA loss is only Bayes-consistent for $\tau = 1$.

Bayes-Consistency. It is known that the Logit-Adjusted (LA) loss is Bayes-consistent with respect to the balanced loss when its temperature parameter is set to one, $\tau = 1$ [Menon et al., 2021]. We begin by establishing a negative result: this consistency does not extend to other values of τ.

Theorem 2. *When $\tau \neq 1$, the LA loss ℓ_{LA} is not Bayes-consistent with respect to the balanced loss ℓ_{BAL}.*

The proof, which involves characterizing the Bayes classifiers for both the LA loss and the balanced loss, is detailed in Appendix F. In contrast, the following result establishes the Bayes-consistency of the GLA loss with respect to the balanced loss for any $q \in [0, 1)$.

Theorem 3. *For any $q \in [0, 1)$, the GLA Loss ℓ_{GLA} is Bayes-consistent with respect to the balanced loss ℓ_{BAL}.*

The proof, provided in Appendix G, characterizes the Bayes classifiers for the GLA loss. Note that Theorem 3 recovers the Bayes-consistency of the LA loss (when $q = 0$) as a special case, consistent with [Menon et al., 2021].

$\mathcal{H}$-Consistency Bounds. We first present a counter-example (Figure 1) demonstrating that even when

$\tau = 1$ (that is, for the standard LA loss, which is GLA with $q = 0$), ℓ_{LA} is not $\mathcal{H}$-consistent with respect to the balanced loss ℓ_{BAL} for certain bounded hypothesis sets. In this example, considering a two-dimensional distribution where $x_1 \sim U[0,1]$ and $x_2 \mid x_1 \sim \mathcal{N}(yx_1, x_1^2)$, with y following a Bernoulli distribution ($\mathbb{P}(+1) = \frac{1}{8}$), if the hypothesis set consists of linear models with bounded weights, specifically $\{(x,y) \mapsto w_y \cdot x : \|w_y\| = 100\}$, the best-in-class classifier for both the balanced loss and a GCA loss is $x_2 = 0$. However, the best-in-class classifier for the LA loss (with $\tau = 1$) differs and is not parallel to $x_2 = 0$. This implies that the LA loss with $\tau = 1$ is not $\mathcal{H}$-consistent for this bounded hypothesis set.

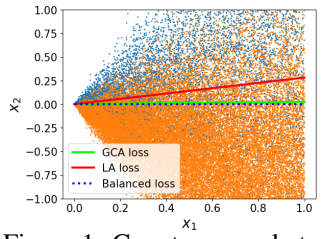

Figure 1: Counterexample to the $\mathcal{H}$-consistency of ℓ_{LA} for bounded hypothesis sets.

This counterexample shows that GLA losses do not guarantee $\mathcal{H}$-consistency for bounded hypothesis sets. The following theorem establishes that the GLA loss ℓ_{GLA} is $\mathcal{H}$-consistent with respect to the balanced loss ℓ_{BAL} if the hypothesis set $\mathcal{H}$ is *complete* that is, for every instance $x \in \mathcal{X}$, the scoring vectors spanned by $\mathcal{H}$ cover the entire space $\mathbb{R}^n$: $\{h(x,\cdot) : h \in \mathcal{H}\} = \mathbb{R}^n$. Naturally, bounded hypothesis sets cannot satisfy this condition. Note that a complete set can be a strict subset of $\mathcal{H}_{\mathrm{all}}$. For example, linear models with unbounded weights are complete, yet they do not equal $\mathcal{H}_{\mathrm{all}}$. Note, the same positive result does not hold for LA losses with general τs. Being not Bayes-consistent, LA losses are not $\mathcal{H}$-consistent for complete hypothesis sets.

Theorem 4. *Assume that $\mathcal{H}$ is complete. Then, for any $q \in [0,1)$, the following $\mathcal{H}$-consistency bound holds for the GLA loss ℓ_{GLA}:*

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \Gamma\Big(\mathcal{R}_{\ell_{\mathrm{GLA}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{GLA}}}(\mathcal{H})\Big),$$

where $\Gamma(t) = \frac{\sqrt{2t}}{\mathsf{p}_{\min}}$ for $q = 0$, and $\Gamma(t) = \frac{\sqrt{2t}}{(\mathsf{p}_{\min})^{\frac{1}{1-q}}(1-q)^{\frac{1}{2}}}$ for $q \in (0,1)$. In the special case where the approximation error $\mathcal{A}_{\ell_{\mathrm{GLA}}}(\mathcal{H}) = 0$, the bound simplifies to:

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \Gamma\Big(\mathcal{R}_{\ell_{\mathrm{GLA}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H})\Big),$$

The proof, presented in Appendix H, consists of first defining a Gibbs distribution induced by h and next of applying a Pinsker-type inequality. Our technique is novel: it constructively upper-bounds the conditional regret of the balanced loss by that of the GLA loss, leveraging Lemma 1. Remarkably, when $q = 0$, Theorem 4 yields $\mathcal{H}$-consistency guarantees for the LA loss with $\tau = 1$ under the completeness assumption, a significantly stronger guarantee that the previously established Bayes-consistency result of Menon et al. [2021]. The $\mathcal{H}$-consistency bounds for GLA losses depend inversely on the minimum class probability, scaling as $1/\mathsf{p}_{\min}$ when $q = 0$ and, more generally, as $(1/\mathsf{p}_{\min})^{\frac{1}{1-q}}$ when $q \in (0,1)$,

5.3 GCA Losses

This section presents consistency guarantees for GCA losses. We define a hypothesis set $\mathcal{H}$ as *regular* if, for any $x \in \mathcal{X}$, the predictions made by the hypotheses in $\mathcal{H}$ cover the complete set of n possible classification labels: $\mathsf{H}(\mathsf{x}) = \{\mathsf{h}(\mathsf{x}) : \mathsf{h} \in \mathcal{H}\} = [\mathsf{n}]$. Widely used hypothesis sets, such as linear models, neural network families, as well as the family of all measurable functions, are all regular. In particular, every complete hypothesis set is regular, while regularity alone is a much weaker yet natural assumption in practice.

The following theorem shows that for a regular hypothesis set, if a GCE loss ℓ_{GCE} is $\mathcal{H}$-consistent with respect to ℓ_{0-1} then its corresponding GCA loss ℓ_{GCA} (Eq. (5)) is also $\mathcal{H}$-consistent with respect to the balanced loss ℓ_{BAL} (Eq. (1)). For simplicity, we assume $\rho_y = 1$ for all y throughout this section.

Theorem 5. *Let $\mathcal{H}$ be a regular hypothesis set and ℓ_{GCE} a GCE loss. Assume that there exists a function $\Gamma(t) = \beta t^{\alpha}$ for some $\alpha \in (0,1]$ and $\beta > 0$, such that the following $\mathcal{H}$-consistency bound holds for all $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H}) \leq \Gamma\Big(\mathcal{R}_{\ell_{\mathrm{GCE}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GCE}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{GCE}}}(\mathcal{H})\Big).$$

Then, the following $\mathcal{H}$-consistency bound holds for ℓ_{GCA} with respect to ℓ_{BAL} for all $h \in \mathcal{H}$ and any distribution:

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \overline{\Gamma}\Big(\mathcal{R}_{\ell_{\mathrm{GCA}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GCA}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{GCA}}}(\mathcal{H})\Big),$$

where $\overline{\Gamma}(t) = \beta\left(\frac{1}{\mathsf{p}_{\min}}\right)^{1-\alpha} t^{\alpha}$. *In the special case where the approximation error* $\mathcal{A}_{\ell_{\mathrm{GCA}}}(\mathcal{H}) = 0$, *this bound simplifies to:*

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^{*}_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \overline{\Gamma}\Big(\mathcal{R}_{\ell_{\mathrm{GCA}}}(h) - \mathcal{R}^{*}_{\ell_{\mathrm{GCA}}}(\mathcal{H})\Big).$$

The proof is provided in Appendix E, where we constructively define new conditional probabilities $\mathsf{q}(y \mid x)$ along with a normalization factor $Z(x) = \sum_{y \in \mathcal{Y}} \frac{\mathsf{p}(y|x)}{\mathsf{p}(y)} \leq \frac{1}{\mathsf{p}_{\min}}$. These probabilities transform the conditional regret of the balanced loss and the GCA loss into the conditional regrets of the zero-one loss and the GCE loss, respectively, under the newly defined distribution.

When $\mathcal{A}_{\ell_{\mathrm{GCA}}}(\mathcal{H}) = 0$, the $\mathcal{H}$-consistency bound guarantees that if the surrogate estimation error $\mathcal{R}_{\ell_{\mathrm{GCA}}}(h) - \mathcal{R}^{*}_{\ell_{\mathrm{GCA}}}(\mathcal{H})$ is optimized up to ϵ, the estimation error for the balanced loss, $\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^{*}_{\ell_{\mathrm{BAL}}}(\mathcal{H})$, is upper-bounded by $\Gamma(\epsilon)$. For common choices of Ψ in ℓ_{GCA}, Mao et al. [2023f,b] show that Γ takes specific forms: for $\Psi(t) = -\log(t)$, $\Gamma(t) = \sqrt{2t}$ (so $\alpha = 1/2$ and $\beta = \sqrt{2}$); for $\Psi(t) = \frac{1}{q}(1 - t^{q})$ with $q \in (0, 1)$, $\Gamma(t) = \sqrt{2n^{q}t}$ (so $\alpha = 1/2$ and $\beta = \sqrt{2n^{q}}$). This leads to the following corollary for GCA losses:

Corollary 6. *Under the assumptions of Theorem 5, for all* $h \in \mathcal{H}$ *and any distribution, the following* $\mathcal{H}$*-consistency bound holds for* ℓ_{GCA} *with respect to* ℓ_{BAL}*:*

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^{*}_{\ell_{\mathrm{BAL}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \overline{\Gamma}\Big(\mathcal{R}_{\ell_{\mathrm{GCA}}}(h) - \mathcal{R}^{*}_{\ell_{\mathrm{GCA}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{GCA}}}(\mathcal{H})\Big),$$

where $\overline{\Gamma}(t) = \frac{\sqrt{2t}}{\sqrt{\mathsf{p}_{\min}}}$ *for* $\Psi(t) = -\log(t)$ *and* $\overline{\Gamma}(t) = \frac{\sqrt{2n^{q}t}}{\sqrt{\mathsf{p}_{\min}}}$ *for* $\Psi(t) = \frac{1}{q}(1 - t^{q})$ *with* $q \in (0, 1)$. *In the special case where the approximation error* $\mathcal{A}_{\ell_{\mathrm{GCA}}}(\mathcal{H}) = 0$, *this bound simplifies to:*

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^{*}_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \overline{\Gamma}\Big(\mathcal{R}_{\ell_{\mathrm{GCA}}}(h) - \mathcal{R}^{*}_{\ell_{\mathrm{GCA}}}(\mathcal{H})\Big),$$

If $\mathcal{H} = \mathcal{H}_{\mathrm{all}}$, taking the limit on both sides implies the Bayes-consistency of these GCA losses ℓ_{GCA} with respect to the balanced loss ℓ_{BAL}. More generally, Corollary 6 demonstrates that ℓ_{GCA} admits an excess error bound relative to ℓ_{BAL} if ℓ_{GCE} has such a bound relative to ℓ_{0-1}.

Mao et al. [2023f] and Mao et al. [2023b] showed that loss functions belonging to the widely used general cross-entropy (GCE) family (including logistic loss) admit $\mathcal{H}$-consistency bounds with respect to the multi-class zero-one loss ℓ_{0-1} when the hypothesis set is complete and *bounded*, respectively. We say a hypothesis set $\mathcal{H}$ is *bounded* if $\mathcal{H} = \{h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \mid h(\cdot, y) \in \mathcal{F},\ \forall y \in \mathcal{Y}\}$, where $\mathcal{F}$ is a family of real-valued functions f satisfying $|f(x)| \leq \Lambda(x)$ for all $x \in \mathcal{X}$, and all values in $[-\Lambda(x), +\Lambda(x)]$ are attainable. Here, $\Lambda(x) > 0$ is a fixed function on $\mathcal{X}$. Boundedness also implies regularity. Thus, a key advantage of GCA losses is their general $\mathcal{H}$-consistency: they are $\mathcal{H}$-consistent for any hypothesis set that is bounded or complete. Furthermore, their consistency bounds exhibit an improved scaling with the minimum class probability, $1/\sqrt{\mathsf{p}_{\min}}$. This contrasts favorably with GLA losses, offering potentially stronger theoretical support in highly imbalanced settings.

Comparison and Discussion. Our theoretical analysis reveals distinct characteristics for the two loss families: GLA losses are Bayes-consistent (for $q \in [0, 1)$). However, their $\mathcal{H}$-consistency requires the hypothesis set $\mathcal{H}$ to be complete (and thus unbounded). The corresponding bounds depend on the minimum class probability $\mathsf{p}_{\min}$, scaling as $1/\mathsf{p}_{\min}$ (for $q = 0$) or less favorably as $(1/\mathsf{p}_{\min})^{\frac{1}{1-q}}$ (for $q \in (0, 1)$). In contrast, GCA losses demonstrate $\mathcal{H}$-consistency for any hypothesis set that is bounded or complete. Their $\mathcal{H}$-consistency bounds scale more favorably with the minimum class probability, as $1/\sqrt{\mathsf{p}_{\min}}$. This suggests GCA losses offer stronger theoretical guarantees, particularly in settings with significant class imbalance or when using more restricted hypothesis sets.

The trade-offs between these theoretical properties and empirical performance are important. As we will show in the next experimental section (Section 6), GLA losses often achieve slightly better empirical results on common benchmarks. Conversely, GCA losses tend to have an edge in highly imbalanced scenarios. This empirical behavior aligns with our theoretical findings: GLA losses may be preferred for moderately imbalanced scenarios when using expressive, potentially unbounded hypothesis sets where their specific form of logit adjustment is beneficial; GCA losses are theoretically better-suited for highly imbalanced settings due to their favorable consistency scaling and applicability to a wider range of hypothesis sets. For bounded hypothesis sets where GLA's $\mathcal{H}$-consistency is not guaranteed, GCA is the theoretically preferred option.

The assumptions in this section primarily concern properties of the hypothesis set. These are standard and typically satisfied in practice. Most natural hypothesis sets, such as linear models, neural networks, and the set of all measurable functions, are regular, meaning they produce predictions across all n classes. Whether a hypothesis set is bounded or complete depends on the modeling choice (e.g., bounded weights in linear models). Importantly, our results do not assume any specific data distribution and hold for arbitrary distributions, including those arising in real-world settings.

Compared to the previous work [Cortes et al., 2025], the key difference is that IMMAX [Cortes et al., 2025] is designed for optimizing the standard multi-class 0-1 loss under imbalanced data, whereas the proposed GCE and GCA losses are designed to optimize the balanced loss. As a result, IMMAX enjoys consistency with respect to the standard 0-1 loss, while GCE and GCA are consistent with respect to the balanced loss, a property most existing surrogate losses lack, as discussed in Section 3.3.

Appendix B further provides margin bounds for both the GCA and GLA losses in the more general cost-sensitive multi-class classification setting. We show that both losses benefit from margin guarantees, with more favorable bounds for GCA losses, as the GLA bounds depend on $1/\mathsf{p}_{\min}$.

Theoretical novelty. Classical margin bounds have been extensively studied (see, for example [Koltchinskii and Panchenko, 2000, 2002, Schapire et al., 1997, Cortes et al., 2021, Mohri et al., 2018]). In particular, Mohri et al. [2018] derived margin bounds for standard multi-class classification. In contrast, we derive new margin bounds for cost-sensitive classification, a setting that introduces additional complexity due to the presence of instance-dependent cost functions. This requires the development of new proof techniques, including the derivation of an upper bound on the loss function expressed in terms of a margin loss and a maximum operator, along with an analysis of the Rademacher complexity of this maximum term via the vector contraction lemma. Moreover, in addition to the resulting margin bounds for GCA loss functions, our margin bounds for GLA loss functions are non-trivial and require a specific and entirely new analysis (Appendix B.2). Mao et al. [2023f,b] studied $\mathcal{H}$-consistency bounds for loss functions in the general cross-entropy (GCE) family with respect to the standard zero-one loss. In contrast, our work establishes $\mathcal{H}$-consistency bounds for the proposed GCA and GLA losses with respect to the balanced loss, where both the surrogate and target losses are more complex. This required several novel technical contributions, including a characterization of the conditional regret of the balanced loss, the use of Gibbs distributions and Pinsker-type inequalities for analyzing GLA losses, and a reduction of the conditional regrets of the balanced and GCA losses to those of the zero-one and GCE losses under a newly defined distribution.

6 Experiments

This section details the empirical evaluation of our proposed Generalized Logit-Adjusted (GLA) and Generalized Class-Aware (GCA) loss functions. We compare their effectiveness in minimizing the balanced loss against several baseline methods on the CIFAR-10, CIFAR-100 [Krizhevsky, 2009], and Tiny ImageNet [Le and Yang, 2015] datasets with respectively 10, 100 and 200 classes. To simulate class imbalance, we reduced the percentage of examples per class identically in both training and test sets, following exactly the protocol in [Menon et al., 2021]. Two types of imbalance were considered: Long-tailed imbalance where class sample sizes decrease exponentially across sorted classes [Cui et al., 2019], and Step imbalance where minority classes share one sample size, and majority classes share another, creating a distinct two-group split [Buda et al., 2018]. The severity of imbalance is quantified by the imbalance ratio, $\rho = \frac{\max_{k=1}^{n} m_k}{\min_{k=1}^{n} m_k}$, where m_k is the number of samples in class k. We evaluated performance at $\rho = 100$ (C), following Menon et al. [2021], and at a more extreme setting of $\rho = 1000$ (M).

Our experimental setup, including training procedures and neural network architectures, strictly followed Menon et al. [2021]. We used a ResNet-32 architecture with ReLU activations [He et al., 2016]. Standard data augmentation techniques were applied: for CIFAR-10 and CIFAR-100, this involved 4-pixel padding followed by 32×32 random crops and random horizontal flips; for Tiny ImageNet, 8-pixel padding was used, followed by 64×64 random crops. All models were trained for 200 epochs using Stochastic Gradient Descent (SGD) with Nesterov momentum [Nesterov, 1983]. We used a a batch size of $1,024$, a weight decay of 1×10^{-3}, and a cosine decay learning rate schedule [Loshchilov and Hutter, 2016] without restarts, with an initial learning rate of 0.2.

Table 1: Balanced error of ResNet-32 on *long-tailed* (left) and *step-imbalanced* (right) imbalanced CIFAR-10, CIFAR-100 and Tiny ImageNet; means ± standard deviations over 5 runs. Note, we are reporting total error and not dividing by number of classes. Imbalance ratios $\rho = 1000$ (M), 100 (C).

Method	ρ	CIFAR-10	CIFAR-100	Tiny I.Net
CE		2.46 ± 0.09	38.45 ± 0.37	70.23 ± 0.38
WCE		2.52 ± 0.17	39.89 ± 0.76	75.89 ± 0.67
LA ($\tau = 1$)		2.18 ± 0.18	35.92 ± 0.47	67.17 ± 0.49
EQUAL		2.38 ± 0.07	37.33 ± 0.36	68.44 ± 0.72
CB	M	2.58 ± 0.03	41.46 ± 0.41	80.22 ± 0.59
FOCAL		2.43 ± 0.10	38.02 ± 0.54	69.13 ± 0.83
LDAM		2.39 ± 0.08	37.39 ± 0.36	68.27 ± 0.81
GCA		**2.02 ± 0.15**	**33.17 ± 0.57**	**64.88 ± 0.66**
GLA		2.04 ± 0.15	33.99 ± 0.52	65.57 ± 0.27
CE		2.72 ± 0.02	61.53 ± 0.29	106.93 ± 0.89
WCE		2.80 ± 0.08	62.20 ± 0.57	112.50 ± 0.97
LA ($\tau = 1$)		2.23 ± 0.08	56.23 ± 0.21	102.81 ± 0.89
EQUAL		2.60 ± 0.08	57.25 ± 0.40	104.91 ± 0.84
CB	C	2.76 ± 0.04	61.55 ± 0.28	115.22 ± 0.71
FOCAL		2.70 ± 0.06	61.21 ± 0.24	105.47 ± 0.59
LDAM		2.66 ± 0.08	60.37 ± 0.60	103.99 ± 0.58
GCA		2.19 ± 0.08	54.02 ± 0.38	101.34 ± 0.81
GLA		**2.07 ± 0.06**	**53.68 ± 0.76**	**100.70 ± 0.83**

Method	ρ	CIFAR-10	CIFAR-100	Tiny I.Net
CE		6.33 ± 0.01	12.47 ± 0.12	39.41 ± 0.40
WCE		6.44 ± 0.02	13.66 ± 0.45	39.28 ± 0.31
LA ($\tau = 1$)		5.54 ± 0.48	11.42 ± 0.33	37.44 ± 0.25
EQUAL		5.89 ± 0.24	12.24 ± 0.20	38.43 ± 0.44
CB	M	6.38 ± 0.01	14.96 ± 0.32	47.35 ± 0.73
FOCAL		6.35 ± 0.01	12.25 ± 0.17	39.21 ± 0.31
LDAM		6.34 ± 0.01	12.30 ± 0.11	38.21 ± 0.27
GCA		**5.35 ± 0.02**	**10.43 ± 0.15**	**36.32 ± 0.32**
GLA		5.39 ± 0.02	10.58 ± 0.19	36.57 ± 0.35
CE		3.66 ± 0.15	60.16 ± 0.09	39.68 ± 0.25
WCE		3.68 ± 0.11	61.40 ± 0.51	43.68 ± 0.42
LA ($\tau = 1$)		2.70 ± 0.12	55.43 ± 0.63	38.42 ± 0.14
EQUAL		3.18 ± 0.12	57.73 ± 0.54	38.91 ± 0.20
CB	C	3.81 ± 0.02	66.41 ± 0.11	50.51 ± 0.45
FOCAL		3.60 ± 0.11	60.06 ± 0.13	39.63 ± 0.27
LDAM		3.41 ± 0.10	58.95 ± 0.11	38.67 ± 0.19
GCA		2.57 ± 0.04	53.85 ± 0.47	37.59 ± 0.43
GLA		**2.48 ± 0.11**	**52.70 ± 0.15**	**36.71 ± 0.33**

We compared our GLA and GCA losses against a suite of widely used baseline methods: standard cross-entropy (CE) loss, class-weighted cross-entropy (WCE) loss [Xie and Manski, 1989, Morik et al., 1999], Logit Adjusted (LA) loss [Menon et al., 2021], Equalization (EQUAL) loss [Tan et al., 2020], Class-Balanced (CB) loss [Cui et al., 2019], FOCAL loss [Ross and Dollár, 2017] and the LDAM loss [Cao et al., 2019]. For all methods, including our GLA and GCA losses, we tune the hyperparameters using a validation set held out separately from the training set. For the parameter q in both GLA and GCA, we selected values from $\{0.0, 0.1, \ldots, 0.9\}$, which are standard choices within the general cross-entropy family. Its performance depends on dataset imbalance (e.g., long-tailed vs. step imbalance). Further details about the experiments including baselines are provided in Appendix C. Performance was primarily evaluated using the balanced error on the imbalanced test sets (i.e., the average of the balanced loss over the test data). Results were averaged over five independent runs, and we report means and standard deviations. Table 1 presents the balanced error for ResNet-32 on long-tailed and step-imbalanced versions of CIFAR-10, CIFAR-100, and Tiny ImageNet.

The results in Table 1 highlight that both our proposed GCA losses and GLA losses generally outperform key baselines such as class-weighted cross-entropy (WCE) and Logit-Adjusted (LA) losses across the tested datasets and imbalance types. This demonstrates the efficacy of our novel loss formulations in achieving better balanced error, indicating improved fairness and accuracy on minority classes. Comparing our two proposed families, GLA losses often achieve the best overall results on several benchmarks, particularly under moderate imbalance ($\rho = 100$). However, GCA losses in accordance with its better $1/\sqrt{\mathsf{p}_{\min}}$ bound tend to exhibit an advantage in settings with high class imbalance ($\rho = 1000$).

The strong performance of GCA losses, especially their edge in highly imbalanced scenarios ($\rho = 1000$), underscores the impact of using class-dependent confidence margins. These margins allow GCA to adapt more effectively to severe skews in data distribution compared to simpler weighting or logit adjustment techniques. The performance difference observed between $\rho = 100$ and $\rho = 1000$ across all methods, and particularly the relative strengths of GLA and GCA, highlights the sensitivity of these approaches to the severity of class imbalance.

7 Conclusion

We introduced two novel families of surrogate losses, GLA and GCA losses, for balanced multi-class classification under class imbalance. Both are principled extensions of widely used loss designs, and our theoretical analysis establishes their consistency properties, highlighting the more favorable $\mathcal{H}$-consistency bounds of GCA losses in imbalanced regimes. Empirically, both loss families outperform existing baselines, with GLA performing better in common benchmarks and GCA offering an edge in highly imbalanced settings. These results position GLA and GCA losses as state-of-the-art surrogates for balanced classification, bridging the gap between fairness, consistency, and practical performance. The extension of these surrogate loss families to structured prediction or multi-label classification could significantly broaden their impact. Finally, refining consistency bounds under realistic hypothesis classes and leveraging recent enhanced $\mathcal{H}$-consistency bounds could provide deeper insights into the behavior of these and related loss functions in balanced learning settings.

References

P. Awasthi, N. Frank, A. Mao, M. Mohri, and Y. Zhong. Calibration and consistency of adversarial surrogate losses. In *Advances in Neural Information Processing Systems*, pages 9804–9815, 2021a.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. A finer calibration analysis for adversarial robustness. *arXiv preprint arXiv:2105.01550*, 2021b.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pages 1117–1174, 2022a.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. Multi-class H-consistency bounds. In *Advances in Neural Information Processing Systems*, pages 782–795, 2022b.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. Theoretically grounded loss functions and algorithms for adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 10077–10094, 2023a.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. DC-programming for neural network optimizations. *Journal of Global Optimization*, 2023b.

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

P. L. Bartlett, D. J. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. *CoRR*, abs/1706.08498, 2017.

J. Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39:357–365, 1944.

J. Berkson. Why I prefer logits to probits. *Biometrics*, 7(4):327–339, 1951.

K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *International Conference on Pattern Recognition*, pages 3121–3124, 2010.

M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

J. Byrd and Z. Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881, 2019.

K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

C. Cardie and N. Nowe. Improving minority class prediction using case-specific feature weights. In *International Conference on Machine Learning*, pages 57–65, 1997.

P. K. Chan and S. J. Stolfo. Learning with non-uniform class and cost distributions: Effects and a distributed multi-classifier approach. In *Workshop Notes KDD-98 Workshop on Distributed Data Mining*, pages 1–9, 1998.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

G. Collell, D. Prelec, and K. R. Patil. Reviving threshold-moving: a simple plug-in bagging ensemble for binary and multiclass imbalanced data. *CoRR*, abs/1606.08698, 2016.

V. Conitzer, R. Freeman, N. Shah, and J. W. Vaughan. Group fairness for the allocation of indivisible goods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1853–1860, 2019.

C. Cortes, V. Kuznetsov, M. Mohri, and S. Yang. Structured prediction theory based on factor graph complexity. In *Advances in Neural Information Processing Systems*, 2016.

C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, and S. Yang. AdaNet: Adaptive structural learning of artificial neural networks. In *International Conference on Machine Learning*, pages 874–883, 2017.

C. Cortes, M. Mohri, and A. T. Suresh. Relative deviation margin bounds. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2122–2131. PMLR, 2021. URL http://proceedings.mlr.press/v139/cortes21a.html.

C. Cortes, A. Mao, C. Mohri, M. Mohri, and Y. Zhong. Cardinality-aware set prediction and top-k classification. In *Advances in Neural Information Processing Systems*, 2024.

C. Cortes, A. Mao, M. Mohri, and Y. Zhong. Balancing the scales: A theoretical and algorithmic framework for learning from imbalanced data. In *International Conference on Machine Learning*, 2025.

J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia. Parametric contrastive learning. In *International Conference on Computer Vision*, 2021.

J. Cui, S. Liu, Z. Tian, Z. Zhong, and J. Jia. Reslt: Residual learning for long-tailed recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

C. Du, Y. Han, and G. Huang. Simpro: A simple probabilistic framework towards realistic long-tailed semi-supervised learning. In *International Conference on Machine Learning*, 2024.

C. Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, 2001.

A. Estabrooks, T. Jo, and N. Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36, 2004.

Y. Fan, S. Lyu, Y. Ying, and B. Hu. Learning with average top-k loss. In *Advances in Neural Information Processing Systems*, pages 497–505, 2017.

T. Fawcett and F. Provost. Combining data mining and machine learning for effective user profiling. In *International Conference on Knowledge Discovery and Data Mining*, pages 8–13, 1996.

V. Feldman. Does learning require memorization? a short tale about a long tail. In *Symposium on Theory of Computing*, pages 954–959, 2020.

M. Gabidolla, A. Zharmagambetov, and M. Á. Carreira-Perpiñán. Beyond the ROC curve: Classification trees using cost-optimal curves, with application to imbalanced datasets. In *International Conference on Machine Learning*, 2024.

J. Gao, H. Zhao, Z. Li, and D. Guo. Enhancing minority classes by mixing: an adaptative optimal transport approach for long-tailed classification. In *Advances in Neural Information Processing Systems*, 2023.

J. Gao, H. Zhao, D. dan Guo, and H. Zha. Distribution alignment optimization through neural collapse for long-tailed classification. In *International Conference on Machine Learning*, 2024.

A. Ghosh, H. Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

B. Han. Wrapped cauchy distributed angular softmax for long-tailed visual recognition. In *International Conference on Machine Learning*, pages 12368–12388, 2023.

H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, pages 878–887, 2005.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

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

Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

A. Iranmehr, H. Masnadi-Shirazi, and N. Vasconcelos. Cost-sensitive support vector machines. *Neurocomputing*, 343:50–64, 2019.

M. A. Jamal, M. Brown, M.-H. Yang, L. Wang, and B. Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020.

R. Jiawei, C. Yu, X. Ma, H. Zhao, S. Yi, et al. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, pages 4175–4186, 2020.

B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021.

T. Kasarla, G. Burghouts, M. Van Spengler, E. Van Der Pol, R. Cucchiara, and P. Mettes. Maximum class separation as inductive bias in one matrix. In *Advances in Neural Information Processing Systems*, pages 19553–19566, 2022.

M. M. Khalili, X. Zhang, and M. Abroshan. Loss balancing for fair supervised learning. In *International Conference on Machine Learning*, pages 16271–16290, 2023.

S. Khan, M. Hayat, S. W. Zamir, J. Shen, and L. Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 103–112, 2019.

B. Kim and J. Kim. Adjusting decision boundary for class imbalanced learning, 2019.

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163, 2001.

G. R. Kini, O. Paraskevas, S. Oymak, and C. Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. In *Advances in Neural Information Processing Systems*, volume 34, pages 18970–18983, 2021.

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–459. Birkhäuser, 2000.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

W. Kotlowski, K. Dembczynski, and E. Hüllermeier. Bipartite ranking through minimization of univariate loss. In *International Conference on Machine Learning*, pages 1113–1120, 2011.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Toronto University, 2009.

M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *International Conference on Machine Learning*, 1997.

Y. Le and X. Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

F. Li, Q. Xu, S. Bao, Z. Yang, R. Cong, X. Cao, and Q. Huang. Size-invariance matters: Rethinking metrics and losses for imbalanced multi-object salient object detection. In *International Conference on Machine Learning*, 2024a.

L. Li, X.-C. Li, H.-J. Ye, and D.-C. Zhan. Enhancing class-imbalanced learning with pre-trained guidance through class-conditional knowledge distillation. In *International Conference on Machine Learning*, pages 28204–28221, 2024b.

M. Li, X. Zhang, C. Thrampoulidis, J. Chen, and S. Oymak. Autobalance: Optimized loss functions for imbalanced data. In *Advances in Neural Information Processing Systems*, pages 3163–3177, 2021.

S. Li, Q. Xu, Z. Yang, Z. Wang, L. Zhang, X. Cao, and Q. Huang. Focal-sam: Focal sharpness-aware minimization for long-tailed classification. In *International Conference on Machine Learning*, 2025.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision*, pages 2980–2988, 2017.

L. Liu, S. He, A. Ming, R. Xie, and H. Ma. Elta: An enhancer against long-tail for aesthetics-oriented models. In *International Conference on Machine Learning*, 2024.

X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 39(2):539–550, 2008.

Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

E. Loffredo, M. Pastore, S. Cocco, and R. Monasson. Restoring balance: principled under/oversampling of data for optimal classification. In *International Conference on Machine Learning*, pages 32643–32670, 2024.

I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

M. A. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML 2003 Workshop on Learning from Imbalanced Datasets*, 2003.

A. Mao. *Theory and Algorithms for Learning with Multi-Class Abstention and Multi-Expert Deferral*. PhD thesis, New York University, 2025.

A. Mao, C. Mohri, M. Mohri, and Y. Zhong. Two-stage learning to defer with multiple experts. In *Advances in Neural Information Processing Systems*, 2023a.

A. Mao, M. Mohri, and Y. Zhong. H-consistency bounds: Characterization and extensions. In *Advances in Neural Information Processing Systems*, 2023b.

A. Mao, M. Mohri, and Y. Zhong. H-consistency bounds for pairwise misranking loss surrogates. In *International Conference on Machine learning*, 2023c.

A. Mao, M. Mohri, and Y. Zhong. Ranking with abstention. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023d.

A. Mao, M. Mohri, and Y. Zhong. Structured prediction with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, 2023e.

A. Mao, M. Mohri, and Y. Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning*, 2023f.

A. Mao, M. Mohri, and Y. Zhong. Principled approaches for learning to defer with multiple experts. In *International Symposium on Artificial Intelligence and Mathematics*, 2024a.

A. Mao, M. Mohri, and Y. Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, 2024b.

A. Mao, M. Mohri, and Y. Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In *International Conference on Artificial Intelligence and Statistics*, 2024c.

A. Mao, M. Mohri, and Y. Zhong. H-consistency guarantees for regression. In *International Conference on Machine Learning*, pages 34712–34737, 2024d.

A. Mao, M. Mohri, and Y. Zhong. Multi-label learning with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, 2024e.

A. Mao, M. Mohri, and Y. Zhong. Realizable H-consistent and Bayes-consistent loss functions for learning to defer. In *Advances in Neural Information Processing Systems*, 2024f.

A. Mao, M. Mohri, and Y. Zhong. Regression with multi-expert deferral. In *International Conference on Machine Learning*, pages 34738–34759, 2024g.

A. Mao, M. Mohri, and Y. Zhong. A universal growth rate for learning with smooth surrogate losses. In *Advances in Neural Information Processing Systems*, 2024h.

A. Mao, M. Mohri, and Y. Zhong. Mastering multiple-expert routing: Realizable H-consistency and strong guarantees for learning to defer. In *International Conference on Machine Learning*, 2025a.

A. Mao, M. Mohri, and Y. Zhong. Principled algorithms for optimizing generalized metrics in binary classification. In *International Conference on Machine Learning*, 2025b.

H. Masnadi-Shirazi and N. Vasconcelos. Risk minimization, probability elicitation, and cost-sensitive SVMs. In *International Conference on Machine Learning*, page 759–766, 2010.

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

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

L. Meng, X. Dai, J. Yang, D. Chen, Y. Chen, M. Liu, Y.-L. Chen, Z. Wu, L. Yuan, and Y.-G. Jiang. Learning from rich semantics and coarse locations for long-tailed object detection. In *Advances in Neural Information Processing Systems*, 2023.

A. Menon, H. Narasimhan, S. Agarwal, and S. Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pages 603–611, 2013.

A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

C. Mohri, D. Andor, E. Choi, M. Collins, A. Mao, and Y. Zhong. Learning to reject with a fixed predictor: Application to decontextualization. In *International Conference on Learning Representations*, 2024.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018.

M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Mchine Learning*, pages 4615–4625, 2019.

K. Morik, P. Brockhausen, and T. Joachims. Combining statistical learning with a knowledge-based approach-a case study in intensive care monitoring. In *International Conference on Machine Learning*, pages 268–277, 1999.

Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. akad. nauk Sssr*, 269:543–547, 1983.

B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. *CoRR*, abs/1503.00036, 2015.

F. Provost. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI-2000 Workshop on Imbalanced Data Sets*, 2000.

X. Qiao and Y. Liu. Adaptive weighted learning for unbalanced multicategory classification. *Biometrics*, 65:159–68, 2008.

A. E. Rastegin. Bounds of the pinsker and fannes types on the tsallis relative entropy. *Mathematical Physics, Analysis and Geometry*, 16(3):213–228, 2013.

T.-Y. Ross and G. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017.

R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of ICML*, pages 322–330, 1997.

J.-X. Shi, T. Wei, Y. Xiang, and Y.-F. Li. How re-sampling helps for long-tail learning? In *Advances in Neural Information Processing Systems*, 2023.

J.-X. Shi, T. Wei, Z. Zhou, J.-J. Shao, X.-Y. Han, and Y.-F. Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *International Conference on Machine Learning*, 2024.

I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

M.-K. Suh and S.-W. Seo. Long-tailed recognition by mutual information maximization between latent features and ground-truth labels. In *International Conference on Machine Learning*, pages 32770–32782, 2023.

Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007.

J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, and J. Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11662–11671, 2020.

K. Tang, J. Huang, and H. Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems*, 2020.

A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007.

J. Tian, Y.-C. Liu, N. Glaser, Y.-C. Hsu, and Z. Kira. Posterior re-calibration for imbalanced datasets. In *Advances in Neural Information Processing Systems*, 2020.

J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. Experimental perspectives on learning from imbalanced data. In *International Conference on Machine Learning*, 2007.

P. F. Verhulst. Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et physique*, 10:113–121, 1838.

P. F. Verhulst. Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 18:1–42, 1845.

B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. Class imbalance, redux. In *International Conference on Data Mining*, pages 754–763, 2011.

H. Wang, M. Xia, Y. Li, Y. Mao, L. Feng, G. Chen, and J. Zhao. Solar: Sinkhorn label refinery for imbalanced partial-label learning. In *Advances in Neural Information Processing Systems*, pages 8104–8117, 2022.

J. Wang, T. Lukasiewicz, X. Hu, J. Cai, and Z. Xu. Rsg: A simple but effective module for learning imbalanced datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3784–3793, 2021a.

X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021b.

Z. Wang, Q. Xu, Z. Yang, Y. He, X. Cao, and Q. Huang. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. In *Advances in Neural Information Processing Systems*, pages 48417–48430, 2023.

Z. Wang, Q. Xu, Z. Yang, Z. Xu, L. Zhang, X. Cao, and Q. Huang. A unified perspective for loss-oriented imbalanced learning via localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

T. Wei, Z. Mao, Z.-H. Zhou, Y. Wan, and M.-L. Zhang. Learning label shift correction for test-agnostic long-tailed recognition. In *International Conference on Machine Learning*, 2024.

L. Xiang, G. Ding, and J. Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263, 2020.

Z. Xiao, Z. Chen, S. Liu, H. Wang, Y. Feng, J. Hao, J. T. Zhou, J. Wu, H. Yang, and Z. Liu. Fedgrab: Federated long-tailed learning with self-adjusting gradient balancer. In *Advances in Neural Information Processing Systems*, 2023.

Y. Xie and C. F. Manski. The logit model and response-based samples. *Sociological Methods & Research*, 17(3):283–302, 1989.

Y. Yang and Z. Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Advances in Neural Information Processing Systems*, 2020.

Y. Yang, S. Chen, X. Li, L. Xie, Z. Lin, and D. Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *Advances in Neural Information Processing Systems*, pages 37991–38002, 2022.

Z. Yang, Q. Xu, Z. Wang, S. Li, B. Han, S. Bao, X. Cao, and Q. Huang. Harnessing hierarchical label distribution variations in test agnostic long-tail recognition. In *International Conference on Machine Learning*, 2024.

H.-J. Ye, H.-Y. Chen, D.-C. Zhan, and W.-L. Chao. Identifying and compensating for feature deviation in imbalanced deep learning, 2020.

X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Feature transfer learning for deep face recognition with long-tail data. *CoRR*, abs/1803.09014, 2018.

J. Zhang, L. Liu, P. Wang, and C. Shen. To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions, 2019a.

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

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

X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao. Range loss for deep face recognition with long-tailed training data. In *International Conference on Computer Vision*, pages 5409–5418, 2017.

Y. Zhang, P. Zhao, J. Cao, W. Ma, J. Huang, Q. Wu, and M. Tan. Online adaptive asymmetric active learning for budgeted imbalanced data. In *SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2768–2777, 2018.

Y. Zhang, P. Zhao, S. Niu, Q. Wu, J. Cao, J. Huang, and M. Tan. Online adaptive asymmetric active learning with limited budgets. *IEEE Transactions on Knowledge and Data Engineering*, 2019b.

Y. Zhang, B. Hooi, L. Hong, and J. Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Advances in Neural Information Processing Systems*, pages 34077–34090, 2022.

Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023.

Z. Zhang and T. Pfister. Learning fast sample re-weighting without reward data. In *International Conference on Computer Vision*, 2021.

Z. Zhang and M. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, 2018.

P. Zhao, Y. Zhang, M. Wu, S. C. Hoi, M. Tan, and J. Huang. Adaptive cost-sensitive online classification. *IEEE Transactions on Knowledge and Data Engineering*, 31(2):214–228, 2018.

Y. Zhong. *Fundamental Novel Consistency Theory: H-Consistency Bounds*. PhD thesis, New York University, 2025.

Z. Zhong, J. Cui, S. Liu, and J. Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2020.

Z.-H. Zhou and X.-Y. Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2005.

B. Zhu, K. Tang, Q. Sun, and H. Zhang. Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models. In *Advances in Neural Information Processing Systems*, pages 64663–64680, 2023.

M. Zhu, C. Fan, H. Chen, Y. Liu, W. Mao, X. Xu, and C. Shen. Generative active learning for long-tailed instance segmentation. In *International Conference on Machine Learning*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 1. The paper contains both the theory and experiments described.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: These results position GLA and GCA losses as state-of-the-art surrogates for balanced classification, bridging the gap between fairness, consistency, and practical performance. The extension of these surrogate loss families to structured prediction or multi-label classification could significantly broaden their impact. Finally, refining consistency bounds under realistic hypothesis classes and leveraging recent enhanced $\mathcal{H}$-consistency bounds could provide deeper insights into the behavior of these and related loss functions in balanced learning settings.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We have sections that introduces formalism and explains terminology. Every proof lists conditions. See Section 4, Section 5, Appendix B, Appendix D, Appendix E, Appendix F, Appendix G, Appendix H, and Appendix I.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We build on other people's approaches, explain our methodology and provide full experimental details in Section 6 and Appendix C.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: See Section 6 and Appendix C.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 6 and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Model training was performed using hardware accelerators providing the equivalent computational power of 64 GPUs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Balanced loss promotes fairness by equalizing performance across demographic groups and ensures that minority classes are not overlooked in long-tailed datasets. It is also crucial in federated learning, where data imbalances across clients can lead to biased models that favor heavy users. This represents a broader impact of balanced multi-class classification under class imbalance.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Section 6. Each dataset is licensed under CC-BY 4.0.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Contents of Appendix

A Related work

Class imbalance is a prevalent challenge in real-world multi-class classification problems [Cui et al., 2019, Fawcett and Provost, 1996, Kang et al., 2021, Kubat and Matwin, 1997, Lewis and Gale, 1994, Liu et al., 2019, Menon et al., 2021]. Applications such as medical diagnosis, fraud detection, and rare event prediction often involve highly skewed label distributions, where a small subset of classes dominate the data, while others, sometimes the most critical, are heavily underrepresented. Standard training objectives, such as minimizing the unweighted cross-entropy loss, tend to be biased toward majority classes, leading to poor performance on minority classes and undermining the fairness, soundness and reliability of learned models.

The extensive literature on class imbalance has yielded a diverse array of techniques [Cardie and Nowe, 1997, Chawla et al., 2002, He and Garcia, 2009, Kubat and Matwin, 1997, Wallace et al., 2011]. Due to space constraints, a comprehensive review of every method is infeasible. Instead, we will categorize and discuss several major strategic directions, referring the reader to recent surveys, such as Zhang et al. [2023], for a more exhaustive treatment. These strategies can be broadly grouped as follows:

1. Data-Level Approaches These methods aim to directly modify the training dataset's class distribution to create a more balanced representation.

- **Re-sampling Techniques:** This is the most traditional approach, involving either oversampling the minority classes (e.g., by duplicating instances or more advanced interpolation) or undersampling the majority classes (by removing instances) [Kubat and Matwin, 1997, Wallace et al., 2011].
- **Synthetic Data Generation:** More sophisticated methods generate new synthetic samples for minority classes. SMOTE (Synthetic Minority Over-sampling Technique) and its variants are prominent examples [Chawla et al., 2002, Han et al., 2005, Qiao and Liu, 2008].
- **Advanced Data Augmentation:** Recent works explore targeted data augmentation strategies to enhance minority class representation, sometimes using generative models or optimal transport principles (e.g., [Gao et al., 2023, Liu et al., 2024, Wang et al., 2021a, Zhu et al., 2024]). While these methods can improve minority class recognition, oversampling may lead to overfitting, undersampling can discard valuable data, and the effectiveness of synthetic data depends heavily on the generation quality [Estabrooks et al., 2004, Liu et al., 2008, Shi et al., 2023, Zhang and Pfister, 2021].

2. Algorithm-Level Cost-Sensitive Learning This category focuses on modifying the learning algorithm to treat classes differently, typically by assigning higher misclassification costs to errors on minority classes.

- **Class Re-Weighting:** A common implementation involves incorporating class weights directly into the loss function, where weights are often inversely proportional to class frequencies or based on concepts like the "effective number of samples" [Cui et al., 2019]. Examples include weighted versions of Softmax or the 0/1 loss [Gabidolla et al., 2024, Morik et al., 1999, Xie and Manski, 1989].
- **Cost-Sensitive Classifiers:** Some learning algorithms, like SVMs, have explicit cost-sensitive formulations [Iranmehr et al., 2019, Masnadi-Shirazi and Vasconcelos, 2010]. Many other methods adapt standard learners to be cost-aware [Elkan, 2001, Fan et al., 2017, Jamal et al., 2020, Sun et al., 2007, Wang et al., 2022, Suh and Seo, 2023, Wang et al., 2023, 2025, Li et al., 2025, Xiao et al., 2023, Zhang et al., 2018, 2019b, 2022, Zhao et al., 2018, Zhou and Liu, 2005]. Cost-sensitive methods offer a principled way to emphasize underrepresented classes. While they can be viewed as algorithmically achieving effects similar to re-sampling, they avoid explicit data duplication or removal. However, their success often hinges on the appropriate selection of costs/weights, and they may not fundamentally alter the decision boundaries if the classes are inherently hard to separate or if the chosen weights are not optimal [Van Hulse et al., 2007].

3. Loss Function and Logit Adjustment This broad category involves designing or modifying loss functions to be more robust to class imbalance or to directly optimize for balanced performance metrics.

- **Modulating Sample Contributions:** Some losses dynamically adjust the contribution of each sample to the total loss based on its difficulty or class. The Focal loss [Lin et al., 2017], for instance, down-weights well-classified (often majority class) examples, allowing the model to focus on hard, minority examples.
- **Margin-Based Modifications:** Several approaches aim to enforce larger decision margins for minority classes or between specific class pairs. Examples include LDAM [Cao et al., 2019], Equalization loss (ESQL) [Tan et al., 2020], and Balanced Softmax [Jiawei et al., 2020]. LADE [Hong et al., 2021] also explores disentangling label distributions.
- **Direct Logit Adjustments:** This sub-group directly modifies the logits (pre-softmax outputs) of the model, often by adding class-specific biases. The Logit Adjustment (LA) method by Menon et al. [2021], Khan et al. [2019] and related techniques like UNO-IC [Tian et al., 2020, Wei et al., 2024] and LSC [Wei et al., 2024] fall here. Menon et al. [2021] showed that a specific form of logit adjustment can achieve Bayes-consistency for the balanced error. Other works explore multiplicative logit modifications [Ye et al., 2020] or combinations of additive and multiplicative changes, like the Vector-Scaling loss [Kini et al., 2021], though multiplicative changes can sometimes be seen as equivalent to input feature re-normalization. To capture how these modified loss functions handle different classes, Wang et al. [2023] proposed a novel technique named data-dependent contraction. Wang et al. [2025] showed that the additive and multiplicative logit modifications essentially correspond to different local calibration assumptions. These methods directly influence the optimization landscape and decision boundaries but may introduce new hyperparameters requiring careful tuning.

4. Representation Learning for Imbalanced Data Instead of (or in addition to) modifying data or loss functions, these techniques focus on learning feature representations that are inherently more robust to class imbalance or that better highlight minority class characteristics.

- Examples include OLTR [Liu et al., 2019], PaCo [Cui et al., 2021], DisA [Gao et al., 2024], and other recent methods focused on semantic richness or distribution alignment (e.g., RichSem [Meng et al., 2023], RBL [Meng et al., 2023], WCDAS [Han, 2023]). Learning discriminative and balanced representations is a fundamental goal, and these methods often aim to decouple feature learning from classifier training to some extent.

5. Decoupled Training and Post-Hoc Adjustments This strategy involves separating the learning process into stages or applying corrections after an initial model has been trained.

- **Decoupled Training:** Representation learning and classifier training are often performed separately. For example, a model might first be trained with instance-balanced sampling or a standard loss, and then the classifier head is fine-tuned using a class-balanced approach (e.g., Decouple-IB-CRT [Kang et al., 2020], CB-CRT [Kang et al., 2020], SR-CRT [Kang et al., 2020], PB-CRT [Kang et al., 2020], MiSLAS [Zhong et al., 2021]). Weight normalization techniques [Kim and Kim, 2019, Kang et al., 2020, Zhang et al., 2019a] also often fall under this paradigm.
- **Post-Hoc Correction:** These methods adjust the outputs or decision thresholds of a pre-trained classifier to improve performance on imbalanced data, without retraining the model [Collell et al., 2016, Fawcett and Provost, 1996, Zhu et al., 2023]. These approaches offer flexibility and can be applied to existing models, but post-hoc methods may not achieve the same level of performance as methods that incorporate imbalance considerations throughout training.

6. Ensemble Learning Approaches Ensemble methods combine multiple classifiers to achieve better predictive performance than any single constituent classifier. For imbalanced learning, ensembles are often constructed by training base learners on different re-sampled versions of the data or by using different cost-sensitive strategies for each member.

- Examples include BBN [Zhou et al., 2020], LFME [Xiang et al., 2020], RIDE [Wang et al., 2021b], ResLT [Cui et al., 2022], SADE [Zhang et al., 2022], and DirMixE [Yang et al., 2024]. Ensembles are often robust but can increase computational expense and reduce model interpretability.

7. Other Notable Strategies The field also includes various other specialized techniques:

- **Transfer Learning:** Leveraging knowledge from related tasks or datasets can help, especially for data-scarce minority classes (e.g., SSP [Yang and Xu, 2020]).
- **Specialized Classifier Design:** Some works focus on designing classifier architectures or objective functions specifically robust to long tails or confounding factors (e.g., De-confound [Tang et al., 2020], [Kasarla et al., 2022, Yang et al., 2022], LIFT [Shi et al., 2024], SimPro [Du et al., 2024]).
- **Metric-Focused Optimization:** Recent studies also analyze the asymptotic performance of classifiers under different metrics on imbalanced data [Loffredo et al., 2024] or develop size-invariant metrics for specific tasks like salient object detection [Li et al., 2024a]. Information and data augmentation via distillation have also been explored [Li et al., 2024b].

This categorization highlights the multifaceted nature of addressing class imbalance. Our work contributes to the area of loss function and logit adjustment, aiming for theoretically grounded and empirically effective solutions. For further details on the landscape of imbalanced learning, we again refer the reader to comprehensive surveys like Zhang et al. [2023].

B Margin bounds

This section provides a margin-based theoretical analysis of cost-sensitive multi-class classification. We derive margin bounds for both the GCA and GLA families. The analysis for the GLA family is more complex, and the resulting bound is generally less favorable, with a dependence on $1/\mathsf{p}_{\min}$.

The proof involves the derivation of an upper bound on the cost-sensitive zero-one loss function expressed in terms of a margin loss and a maximum operator, along with an analysis of the Rademacher complexity of this maximum term via the vector contraction lemma. Moreover, our margin bounds for GLA loss functions are non-trivial and require a specific and entirely new analysis (Appendix B.2).

B.1 Theoretical analysis

Let $h\colon \mathcal{X}\times[n] \to \mathbb{R}$ be scoring function belonging to the hypothesis set $\mathcal{H}$. We define the cost-sensitive zero-one loss function L as follows: for all $(h, x, y) \in \mathcal{H} \times \mathcal{X} \times [n]$,

$$\mathsf{L}(h, x, k) = c(x, y)\, 1_{\mathsf{h}(x)\neq y},$$

where $c(x, y)$ is a non-negative cost that is upper bounded by $\overline{C}$. Note that ℓ_{BAL} is a special case of L.

A. Cost-sensitive margin loss functions. We first introduce new cost-sensitive margin loss functions which will play a central role in our derivation of margin-based guarantees for cost-sensitive learning.

Let $\Phi_\rho\colon u \mapsto \min(1, \max(0, 1 - u/\rho))$ denote the ρ-margin loss function. We can upper-bound the cost-sensitive zero-one loss function L as follows:

$$\begin{aligned}
\mathsf{L}(h, x, y) &\leq c(x, y)\Phi_\rho(\rho_h(x, y)) \\
&= c(x, y)\Phi_\rho\Big(h(x, y) - \max_{y'\neq y} h(x, y')\Big) \\
&\leq c(x, y)\Phi_\rho(h(x, y) - h(x, \mathsf{h}(x))) \\
&= c(x, y) \max_{y'\in[n]} \{\Phi_\rho(h(x, y) - h(x, y'))\}.
\end{aligned}$$

The second inequality follows from the fact that when $y = \mathsf{h}(x)$ we have $h(x, y) = h(x, \mathsf{h}(x)) \geq \max_{y'\neq y} h(x, y')$. Otherwise, for $y \neq \mathsf{h}(x)$, the runner-up prediction satisfies $\operatorname{argmax}_{y'\neq y} h(x, y') = \mathsf{h}(x)$.

The analysis above motivates the definition of the *cost-sensitive margin loss function* as the function $\mathsf{L}_\rho\colon \mathcal{H}_{\mathrm{all}} \times \mathcal{X} \times [n] \to \mathbb{R}$, defined as follows, for any fixed $\rho > 0$:

$$\mathsf{L}_\rho(h, x, y) = c(x, y) \max_{y'\in[n]} \{\Phi_\rho(h(x, y) - h(x, y'))\}.$$

B. Margin bounds. We now establish a general margin-based generalization bound, which serves as the foundation for deriving new algorithms for cost-sensitive classification.

Given a sample $S = (x_1, \ldots, x_m)$ and a hypothesis h, the *empirical cost-sensitive margin loss* is defined by $\widehat{\mathcal{R}}_{S,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathsf{L}_\rho(h, x_i, y_i)$ and the *empirical GCA loss* is defined by $\widehat{\mathcal{R}}_{S,\ell_{\mathrm{GCA}}}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell_{\mathrm{GCA}}(h, x_i, y_i)$. The empirical Rademacher complexity of $\mathcal{H}$ for a sample S is defined as:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\sup_{h\in\mathcal{H}}\left\{\sum_{i=1}^{m}\sum_{y=1}^{n}\epsilon_{iy}h(x_i, y)\right\}\right],$$

where $\boldsymbol{\epsilon} = (\epsilon_{iy})_{i,y}$ represents a matrix of independent Rademacher variables ϵ_{iy}s, each uniformly distributed over $\{-1, +1\}$. For any integer $m \geq 1$, the Rademacher complexity of $\mathcal{H}$ is the expectation of $\widehat{\mathfrak{R}}_S(\mathcal{H})$ over all samples S of size m: $\mathfrak{R}_m(\mathcal{H}) = \mathbb{E}_{S\sim\mathcal{D}^m}\left[\widehat{\mathfrak{R}}_S(\mathcal{H})\right]$.

Using these notions of complexity, we prove the following cost-sensitive margin-based guarantees.

Theorem 7 (Margin bound for cost-sensitive classification)**.** *Let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}\times[n]$ to $\mathbb{R}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following inequalities holds for all $h \in \mathcal{H}$:*

$$\mathcal{R}_{\mathsf{L}}(h) \leq \widehat{\mathcal{R}}_{S,\rho}(h) + 4\overline{C}\sqrt{2n}\,\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$
$$\mathcal{R}_{\mathsf{L}}(h) \leq \widehat{\mathcal{R}}_{S,\rho}(h) + 4\overline{C}\sqrt{2n}\,\widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

The proof is included in Appendix I. These bounds can be generalized to hold uniformly for all $\rho \in (0, 1]$, at the cost of additional $\log\log$-terms, using standard proof techniques [Mohri et al., 2018, Theorem 5.9]. As with standard margin bounds, these learning guarantees suggest a trade-off: Increasing ρ reduces the complexity term (second term) but simultaneously increases the empirical cost-sensitive margin loss, $\widehat{\mathcal{R}}_{S,\rho}(h)$ (first term), by imposing stricter confidence margin requirements. Thus, if h maintains a low empirical cost-sensitive margin loss even with a relatively large ρ value, it admits a strong generalization error guarantee. Using the fact that $\mathsf{L}_\rho(h)$ is upper bounded by $\ell_{\mathrm{GCA}}(h/\rho)$, where $c(x, y) = \frac{1}{\mathsf{p}(y)} \leq \frac{1}{\mathsf{p}_{\min}} = \overline{C}$, we derive the margin bounds for GCA losses below.

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) \leq \widehat{\mathcal{R}}_{S,\ell_{\mathrm{GCA}}}(h/\rho) + \frac{4}{\mathsf{p}_{\min}}\sqrt{2n}\,\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$
$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) \leq \widehat{\mathcal{R}}_{S,\ell_{\mathrm{GCA}}}(h/\rho) + \frac{4}{\mathsf{p}_{\min}}\sqrt{2n}\,\widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

B.2 Margin bounds for GLA lossess

The previous section established margin bounds for GCA losses by leveraging their class-weighted structure. In contrast, deriving analogous bounds for GLA losses is non-trivial due to their different formulation, which involves shifting logits based on class priors. To address this, we will rely on a non-trivial inequality presented in Lemma 8.

Given a sample $S = (x_1, \ldots, x_m)$ and a hypothesis h, the *empirical GLA loss* is defined by $\widehat{\mathcal{R}}_{S,\ell_{\mathrm{GLA}}}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell_{\mathrm{GLA}}(h, x_i, y_i)$. For simplicity, our analysis focuses on the GLA loss with $q = 0$. A similar line of reasoning allows for the extension of this proof to the general case where $q \in (0, 1)$. In our setting of the balanced loss, the costs only depend on y with $c(y) = 1/\mathsf{p}(y)$. Our analysis holds for arbitrary such y-dependent costs. Let $c_{\max}$ denote an upper bound $c_{\min}$ a lower bound on the costs. Define $C_{\max} = \frac{c_{\max}}{\log\left[1+\frac{c_{\min}}{c_{\max}}\right]}$. Then, for any $y, y' \in \mathcal{Y}$, the following holds:

$$\frac{c(y)}{\log\left[1 + \frac{c(y)}{c(y')}\right]} \leq \frac{c(y)}{\log\left[1 + \frac{c_{\min}}{c_{\max}}\right]} \leq C_{\max}.$$

Thus, for any $\rho > 0$ and $y, y' \in \mathcal{Y}$, we have (see illustration in Figure 2 and proof of Lemma 8)

$$c(y)\Phi_\rho(v) \leq C_{\max}\log\left[1 + \frac{c(y)}{c(y')}\exp\left(-\frac{v}{\rho}\right)\right].$$

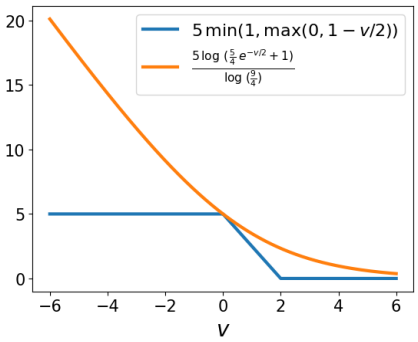

Figure 2: Illustration of the bound in the proof of the margin loss for ℓ_{LA}.

Using the monotonicity of the logarithm and upper-bounding a maximum of non-negative terms by a sum yields the following any $\rho > 0$ and any $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$\begin{aligned}
c(y) \max_{y' \neq y} \{\Phi_\rho(h(x,y) - h(x,y'))\} &\leq C_{\max} \max_{y' \neq y} \left\{ \log\left[1 + \frac{c(y)}{c(y')} \exp\left(\frac{h(x,y') - h(x,y)}{\rho}\right)\right]\right\} \\
&\leq C_{\max} \left\{ \log\left[1 + \max_{y' \neq y} \frac{c(y)}{c(y')} \exp\left(\frac{h(x,y') - h(x,y)}{\rho}\right)\right]\right\} \\
&\leq C_{\max} \left\{ \log\left[1 + \sum_{y' \neq y} \frac{c(y)}{c(y')} \exp\left(\frac{h(x,y') - h(x,y)}{\rho}\right)\right]\right\} \\
&= C_{\max} \left\{ \log\left[\sum_{y' \in \mathcal{Y}} \frac{c(y)}{c(y')} \exp\left(\frac{h(x,y') - h(x,y)}{\rho}\right)\right]\right\} \\
&= C_{\max}\, \ell_{\mathrm{GLA}}(h/\rho, x, y).
\end{aligned}$$

Thus, this yields the margin-based bounds for the GLA loss below. In our setting, $c_{\min} = 1 \leq \frac{1}{\mathsf{p}(y)}$ and $c_{\max} = \frac{1}{\mathsf{p}_{\min}}$. Thus, as with the $\mathcal{H}$-consistency guarantees, the margin bounds here depend on $\frac{1}{\mathsf{p}_{\min}}$.

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) \leq \frac{1}{\mathsf{p}_{\min} \log[1 + \mathsf{p}_{\min}]} \widehat{\mathcal{R}}_{S, \ell_{\mathrm{GLA}}}(h/\rho) + \frac{4}{\mathsf{p}_{\min}} \sqrt{2n}\, \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) \leq \frac{1}{\mathsf{p}_{\min} \log[1 + \mathsf{p}_{\min}]} \widehat{\mathcal{R}}_{S, \ell_{\mathrm{GLA}}}(h/\rho) + \frac{4}{\mathsf{p}_{\min}} \sqrt{2n}\, \widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Lemma 8. *For any $\rho > 0$ and $y, y' \in \mathcal{Y}$, we have*

$$c(y)\Phi_\rho(v) \leq C_{\max} \log\left[1 + \frac{c(y)}{c(y')} \exp\left(-\frac{v}{\rho}\right)\right],$$

for every $v \in \mathbb{R}$.

Proof. Fix labels y, y' and a margin value $v \in \mathbb{R}$. Write $a = \dfrac{c(y)}{c(y')}$, $t = \dfrac{v}{\rho}$, and recall the bounds $c_{\min} \leq c(y), c(y') \leq c_{\max}$. By definition, $C_{\max} = \frac{c_{\max}}{\log\left[1 + \frac{c_{\min}}{c_{\max}}\right]}$. Then, for any $y, y' \in \mathcal{Y}$, the following holds:

$$\frac{c(y)}{\log\left[1 + \frac{c(y)}{c(y')}\right]} \leq \frac{c(y)}{\log\left[1 + \frac{c_{\min}}{c_{\max}}\right]} \leq C_{\max}. \tag{6}$$

Next, we analyze case by case.

(i) $v \leq 0$ ($t \leq 0$). Then $\Phi_\rho(t) = 1$ and $\exp(-t) \geq 1$, using (6) gives

$$C_{\max} \log\left(1 + a e^{-t}\right) \geq C_{\max} \log(1 + a) \geq c(y) = c(y)\Phi_\rho(t).$$

(ii) $0 \le v \le \rho$ $(0 \le t \le 1)$. Define $h(t) = \log(1 + ae^{-t}) - (1-t)\log(1+a)$. Since $h'(t) = -\dfrac{ae^{-t}}{1+ae^{-t}} + \log(1+a) \ge -\dfrac{a}{1+a} + \log(1+a) \ge 0$ (because $\log(1+u) \ge u/(1+u)$ for all $u \ge 0$), we have $h(t) \ge h(0) = 0$; hence

$$(1-t)\log(1+a) \le \log\bigl(1+ae^{-t}\bigr).$$

Multiplying by $C_{\max}$ and using (6) gives

$$C_{\max}\log\bigl(1+ae^{-t}\bigr) \ge C_{\max}\log(1+a)(1-t) \ge c(y)(1-t) = c(y)\Phi_\rho(v).$$

(iii) $v \ge \rho$ $(t \ge 1)$. Then $\Phi_\rho(v) = 0$ and the desired inequality is trivial because the right–hand side is non-negative.

In conclusion, all three cases yield

$$c(y)\Phi_\rho(v) \le C_{\max}\log\left[1 + \frac{c(y)}{c(y')}\exp\left(-\frac{v}{\rho}\right)\right],$$

for every $v \in \mathbb{R}$. This completes the proof. □

B.3 Algorithms

The margin guarantees established in the previous section provide a foundation for developing new algorithms. We begin by deriving a more explicit learning guarantee within a broad framework, which we then use to define a general cost-sensitive learning algorithm.

A. Explicit upper bounds. To make these guarantees more explicit, we introduce the following setup. Given a feature mapping $\Phi\colon \mathcal{X} \times [n] \to \mathbb{R}^d$, we can identify $\mathcal{X} \times [n]$ with a subset of $\mathbb{R}^d$, with $\|\Psi(x,y)\| \le \mathsf{X}_y$ for all $x \in \mathcal{X}$ and $\mathsf{X} = \max_{y\in[n]} \mathsf{X}_y$, for some norm $\|\cdot\|$. We assume $\mathcal{H}$ is given by $\mathcal{H} = \bigl\{h \in \overline{\mathcal{H}}\colon \|h\|_* \le \overline{\mathsf{H}}\bigr\}$, for some appropriate norm $\|\cdot\|_*$ on some space $\overline{\mathcal{H}}$ and $\overline{\mathsf{H}} > 0$. This formulation covers a wide range of hypothesis sets, including linear, kernel-based, and neural network models. In particular, it captures the settings of neural networks with weight matrices constrained by a Frobenius norm bound [Cortes et al., 2017, Neyshabur et al., 2015] or a spectral norm complexity constraint relative to reference weight matrices [Bartlett et al., 2017]. In all of these cases, the empirical Rademacher complexity can be upper bounded as follows:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \le \frac{\sqrt{n}\,\mathsf{H}}{m}\sqrt{\sum_{j=1}^{n} m_j \mathsf{X}_j^2} \le \frac{\sqrt{n}\,\mathsf{H}\mathsf{X}}{\sqrt{m}},$$

where the complexity term H depends on $\overline{\mathsf{H}}$. Combining this upper bound with Theorem 7 yields the following more explicit guarantee.

Corollary 9. *Fix $\rho = [\rho_k]_{k\in[n]}$, then, for any $\delta > 0$, with probability at least $1-\delta$ over the choice of a sample S of size m, the following holds for any $f \in \mathcal{H}$:*

$$\mathcal{R}_{\mathsf{L}}(h) \le \widehat{\mathcal{R}}_{S,\rho}(h) + \frac{4\overline{C}\sqrt{2}n\mathsf{H}}{m}\sqrt{\sum_{j=1}^{n} m_j \mathsf{X}_j^2} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

As with Theorem 7, this bound can be generalized to hold uniformly for all $\rho \in (0,1]$, at the cost of additional $\log\log$-terms. This generalized guarantee provides a basis for designing algorithms choosing $h \in \mathcal{H}$ and ρ to minimize the bound.

Let Ψ be a decreasing convex function such that $\Phi_\rho(x) \le \Psi\left(\frac{x}{\rho}\right)$ for all $x \in \mathbb{R}$ and $\rho > 0$. Ψ may be the hinge loss, $\Psi(x) = \max(0, 1-x)$, or any member of the broad family of composition-sum (comp-sum) losses [Mao et al., 2023f] defined by $\Psi(x) = \Phi^\tau(e^{-x})$, with Φ^τ for $\tau \ge 0$ given by

$$\Phi^\tau(u) = \begin{cases} \frac{1}{1-\tau}\bigl((1+u)^{1-\tau} - 1\bigr) & \tau \ge 0, \tau \ne 1 \\ \log(1+u) & \tau = 1, \end{cases}$$

for all $u \ge 0$. This family includes the logistic loss ($\tau = 1$) and the exponential loss ($\tau = 0$). Using the fact that $\Phi_\rho(t) \le \Psi\left(\frac{t}{\rho}\right)$, the guarantee of Corollary 9 and its generalization to a uniform bound

can be expressed as: for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in \mathcal{H}$, where the last term accounts for the log-log terms and the δ-confidence term

$$\mathcal{R}_{\mathsf{L}}(h) \leq \frac{1}{m}\left[\sum_{i=1}^{m} c(x_i, y_i) \max_{y' \in [n]} \left\{\Psi\left(\frac{h(x_i,y_i) - h(x_i,y')}{\rho}\right)\right\}\right] + \frac{4\overline{C}\sqrt{2}n\mathsf{H}}{m}\sqrt{\sum_{j=1}^{n} m_j \mathsf{X}_j^2} + O\left(\frac{1}{\sqrt{m}}\right).$$

B. General cost-sensitive algorithm. The bound leads to the following regularization-based algorithm:

$$\min_{h \in \overline{\mathcal{H}}} \lambda \|h\|^2 + \frac{1}{m}\left[\sum_{i=1}^{m} c(x_i, y_i) \max_{y' \in [n]} \left\{\Psi\left(\frac{h(x_i,y_i) - h(x_i,y')}{\rho}\right)\right\}\right],$$

where λ and ρs are selected via cross-validation. This is equivalent to minimizing the following surrogate loss:

$$\ell(h, x, y) = c(x, y) \max_{y' \in [n]} \left\{\Psi\left(\frac{h(x,y) - h(x,y')}{\rho}\right)\right\} \tag{7}$$

The preceding derivation shows that this form can be further upper-bounded by both GCA and GLA losses. Consequently, both loss families benefit from margin guarantees, though GCA losses achieve more favorable bounds due to the GLA bounds' dependence on $1/\mathsf{p}_{\min}$.

C Experimental details

This appendix provides supplementary information regarding the experimental setup discussed in Section 6. We first present the precise mathematical formulations for our algorithms and all baseline loss functions used in the comparative analysis. Then, we detail the specific hyperparameter ranges explored during the cross-validation process for each method.

Since our work focuses on principled surrogate losses for imbalanced data, our experiments aimed for a direct comparison with existing losses in their basic forms. We excluded common enhancements from data modification or optimization strategies to isolate the performance of the loss functions.

C.1 Loss function formulations

Let m_k be the number of samples in class k, and m be the total number of samples. Below are the definitions of the loss functions optimized by our algorithms and those optimized by the baseline methods. For any triplet (h, x, y), where h is the hypothesis, x is the input, and y is the true label from a set of n classes:

- **Cross-Entropy (CE) Loss:**

$$\ell_{\text{CE}}(h, x, y) = -\log\left(\frac{e^{h(x,y)}}{\sum_{j=1}^{n} e^{h(x,j)}}\right).$$

- **Class-Weighted Cross-Entropy (WCE) loss** [Morik et al., 1999, Xie and Manski, 1989]:

$$\ell_{\text{WCE}}(h, x, y) = -\frac{m}{m_y}\log\left(\frac{e^{h(x,y)}}{\sum_{j=1}^{n} e^{h(x,j)}}\right).$$

- **Logit Adjusted (LA) Loss** ($\tau = 1$) [Menon et al., 2021]:

$$\ell_{\text{LA}}(h, x, y) = -\log\left(\frac{e^{h(x,y)+\log(m_y)}}{\sum_{j=1}^{n} e^{h(x,j)+\log(m_j)}}\right).$$

- **Equalization (EQUAL) Loss** [Tan et al., 2020]:

$$\ell_{\text{EQUAL}}(h, x, y) = -\log\left(\frac{e^{h(x,y)}}{\sum_{j=1}^{n} w_j e^{h(x,j)}}\right),$$

with weight $w_j = 1 - \beta \mathbf{1}_{\{\frac{m_j}{m} < \lambda\}} \mathbf{1}_{\{y \neq j\}}$, where $\beta \sim \text{Bernoulli}(p)$, and $1 > p > 0$, $1 > \lambda > 0$ are hyperparameters.

- **Class-Balanced (CB) Loss** [Cui et al., 2019]:

$$\ell_{\text{CB}}(h, x, y) = -\frac{1-\gamma}{1-\gamma^{\frac{m_y}{m}}} \log\left(\frac{e^{h(x,y)}}{\sum_{j=1}^{n} e^{h(x,j)}}\right),$$

  where $1 > \gamma > 0$ is a hyperparameter.

- **FOCAL Loss** [Ross and Dollár, 2017]:

$$\ell_{\text{FOCAL}}(h, x, y) = -\left(1 - \frac{e^{h(x,y)}}{\sum_{j=1}^{n} e^{h(x,j)}}\right)^{\gamma} \log\left(\frac{e^{h(x,y)}}{\sum_{j=1}^{n} e^{h(x,j)}}\right),$$

  where $\gamma \geq 0$ is a hyperparameter.

- **LDAM Loss** [Cao et al., 2019]:

$$\ell_{\text{LDAM}}(h, x, y) = -\log\left(\frac{e^{h(x,y)-\Delta_y}}{e^{h(x,y)-\Delta_y} + \sum_{j \neq y} e^{h(x,j)}}\right),$$

  where $\Delta_j = \frac{C}{m_j^{\frac{1}{4}}}$ for $j \in [n]$, and $C > 0$ is a hyperparameter.

- **Generalized Class-Aware (GCA) Loss**:

$$\ell_{\text{GCA}}(h, x, y) = \frac{m}{m_y} \Psi^q\left(\frac{e^{h(x,y)/\rho_y}}{\sum_{y' \in \mathcal{Y}} e^{h(x,y')/\rho_y}}\right),$$

  where $q \in [0, 1)$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)$ is a vector of positive parameters for each class.

- **Generalized Logit-Adjusted (GLA) Loss**:

$$\ell_{\text{GLA}}(h, x, y) = \Psi^q\left(\frac{e^{h(x,y)+\frac{\log(m_y/m)}{1-q}}}{\sum_{y' \in \mathcal{Y}} e^{h(x,y')+\frac{\log\left(m_{y'}/m\right)}{1-q}}}\right),$$

  where $q \in [0, 1)$ is a hyperparameter.

C.2 Hyperparameter search protocol

As stated in Section 6, all hyperparameters for the baseline methods and our algorithms were optimized via cross-validation. The search ranges for each tunable parameter were as follows:

- **CE Loss, WCE Loss:** These methods do not have tunable hyperparameters beyond standard optimization settings.
- **LA Loss:** We fixed the hyperparameter $\tau = 1$ as the algorithm is only Bayes-consistent for that value.
- **EQUAL Loss:** p was selected from $\{0.1, 0.2, \ldots, 0.9\}$, and λ was selected from $\{0.176, 0.5, 0.8, 1.5, 1.76, 2.0, 3.0, 5.0\} \times 10^{-3}$ by following Tan et al. [2020],
- **CB Loss:** γ was selected from $\{0.1, 0.2, \ldots, 0.9, 0.99, 0.999, 0.9999\}$ by following Cui et al. [2019],
- **FOCAL Loss:** γ was selected from $\{1.0, 1.5, \ldots, 10.0\}$ and $\{0.0, 0.1, \ldots, 0.9\}$ by following Ross and Dollár [2017].
- **LDAM Loss:** C was selected from $\{10^{-4}, \ldots, 10^{4}\}$ and $\{5 \times 10^{-4}, \ldots, 5 \times 10^{3}\}$ by following Cao et al. [2019].
- **GCA Loss:** $\boldsymbol{\rho}$ was chosen as $\left(\frac{m_1^{1/3}}{\sum_{k=1}^{n} m_k^{1/3}}, \ldots, \frac{m_n^{1/3}}{\sum_{k=1}^{n} m_k^{1/3}}\right)$ by following Cortes et al. [2025]. q was selected from $\{0.0, 0.1, \ldots, 0.9\}$.
- **GLA Loss:** q was selected from $\{0.0, 0.1, \ldots, 0.9\}$.

D Conditional regret for the balanced loss: proof of Lemma 1

Lemma 1. *For any $x \in \mathcal{X}$, the best-in-class conditional error and the conditional regret for ℓ_{BAL} can be expressed as follows:*

$$\mathcal{C}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}, x) = \sum_{y \in \mathcal{Y}} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} - \max_{y \in \mathsf{H}(\mathsf{x})} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} \quad \Delta\mathcal{C}_{\ell_{\mathrm{BAL}}, \mathcal{H}}(h, x) = \max_{y \in \mathsf{H}(\mathsf{x})} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} - \frac{\mathsf{p}(\mathsf{h}(x)) \mid x)}{\mathsf{p}(\mathsf{h}(x))}.$$

Proof. By the definition and Bayes' theorem, the conditional error can be expressed as follows:

$$\begin{aligned}\mathcal{C}_{\ell_{\mathrm{BAL}}}(h, x) &= \sum_{y \in \mathcal{Y}} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} 1_{\mathsf{h}(x) \neq y} \\ &= \sum_{y \in \mathcal{Y}} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} - \frac{\mathsf{p}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))}.\end{aligned}$$

Since $\{\mathsf{h}(x) \colon h \in \mathcal{H}\} = \mathsf{H}(\mathsf{x})$, the best-in-class conditional error can be expressed as follows:

$$\mathcal{C}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}, x) = \sum_{y \in \mathcal{Y}} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} - \max_{y \in \mathsf{H}(\mathsf{x})} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)},$$

which proves the first part of the lemma. This leads to

$$\Delta\mathcal{C}_{\ell_{\mathrm{BAL}}, \mathcal{H}}(h, x) = \mathcal{C}_{\ell_{\mathrm{BAL}}}(h, x) - \mathcal{C}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}, x) = \max_{y \in \mathsf{H}(\mathsf{x})} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} - \frac{\mathsf{p}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))},$$

which proves the second part of the lemma. □

E $\mathcal{H}$-Consistency for the GCA losses: proof of Theorem 5

Theorem 5. *Let $\mathcal{H}$ be a regular hypothesis set and ℓ_{GCE} a GCE loss. Assume that there exists a function $\Gamma(t) = \beta t^\alpha$ for some $\alpha \in (0, 1]$ and $\beta > 0$, such that the following $\mathcal{H}$-consistency bound holds for all $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H}) \leq \Gamma\Big(\mathcal{R}_{\ell_{\mathrm{GCE}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GCE}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{GCE}}}(\mathcal{H})\Big).$$

Then, the following $\mathcal{H}$-consistency bound holds for ℓ_{GCA} with respect to ℓ_{BAL} for all $h \in \mathcal{H}$ and any distribution:

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \overline{\Gamma}\Big(\mathcal{R}_{\ell_{\mathrm{GCA}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GCA}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{GCA}}}(\mathcal{H})\Big),$$

where $\overline{\Gamma}(t) = \beta\left(\frac{1}{\mathsf{p}_{\min}}\right)^{1-\alpha} t^\alpha$. In the special case where the approximation error $\mathcal{A}_{\ell_{\mathrm{GCA}}}(\mathcal{H}) = 0$, this bound simplifies to:

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \overline{\Gamma}\Big(\mathcal{R}_{\ell_{\mathrm{GCA}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GCA}}}(\mathcal{H})\Big).$$

Proof. The proof involves a reduction of the conditional regrets of the balanced and GCA losses to those of the zero-one and GCE losses under a newly defined distribution and the use of known $\mathcal{H}$-consistency bounds for GCE losses. We define a new conditional probability $\mathsf{q}(y \mid x)$ as $\mathsf{q}(y \mid x) = \frac{\mathsf{p}(y|x)}{\mathsf{p}(y)} \frac{1}{Z(x)}$, where $Z(x) = \sum_{y \in \mathcal{Y}} \frac{\mathsf{p}(y|x)}{\mathsf{p}(y)} \leq \frac{1}{\mathsf{p}_{\min}}$ is the normalization factor. By Lemma 1, the

conditional regret of ℓ_{BAL} can be expressed and upper-bounded as follows:

$$
\begin{aligned}
\Delta\mathcal{C}_{\ell_{\mathrm{BAL}},\mathcal{H}}(h,x) &= \max_{y\in\mathcal{Y}} \frac{\mathsf{p}(y\mid x)}{\mathsf{p}(y)} - \frac{\mathsf{p}(\mathsf{h}(x))\mid x)}{\mathsf{p}(\mathsf{h}(x))}\\
&= Z(x)\left(\max_{y\in\mathcal{Y}} \mathsf{q}(y\mid x) - \mathsf{q}(x\mid \mathsf{h}(x))\right)\\
&= Z(x)\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)\\
&\leq Z(x)\Gamma(\Delta\mathcal{C}_{\ell_{\mathrm{GCE}},\mathcal{H}}(h,x)) && (\mathcal{H}\text{-consistency bound of } \ell_{\mathrm{GCE}})\\
&= Z(x)\Gamma\left(\sum_{y\in\mathcal{Y}} \mathsf{q}(y\mid x)\ell_{\mathrm{GCE}}(h,x,y) - \inf_{h\in\mathcal{Y}}\sum_{y\in\mathcal{Y}} \mathsf{q}(y\mid x)\ell_{\mathrm{GCE}}(h,x,y)\right)\\
&= Z(x)\Gamma\left(\frac{1}{Z(x)}\left(\sum_{y\in\mathcal{Y}} \frac{\mathsf{p}(y\mid x)}{\mathsf{p}(y)}\ell_{\mathrm{GCE}}(h,x,y) - \inf_{h\in\mathcal{Y}}\sum_{y\in\mathcal{Y}} \frac{\mathsf{p}(y\mid x)}{\mathsf{p}(y)}\ell_{\mathrm{GCE}}(h,x,y)\right)\right)\\
&= Z(x)\Gamma\left(\frac{1}{Z(x)}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y\mid x)\ell_{\mathrm{GCA}}(h,x,y) - \inf_{h\in\mathcal{Y}}\sum_{y\in\mathcal{Y}} \mathsf{p}(y\mid x)\ell_{\mathrm{GCA}}(h,x,y)\right)\right)\\
&= Z(x)\Gamma\left(\frac{1}{Z(x)}\Delta\mathcal{C}_{\ell_{\mathrm{GCA}},\mathcal{H}}(h,x)\right)\\
&= \beta\, Z(x)^{1-\alpha}\Delta\mathcal{C}_{\ell_{\mathrm{GCA}},\mathcal{H}}(h,x)^{\alpha}\\
&\leq \beta\left(\frac{1}{p_{\min}}\right)^{1-\alpha}\Delta\mathcal{C}_{\ell_{\mathrm{GCA}},\mathcal{H}}(h,x)^{\alpha}
\end{aligned}
$$

Thus, taking expectations gives:

$$
\begin{aligned}
\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{BAL}}}(\mathcal{H}) &= \mathop{\mathbb{E}}_{x}[\Delta\mathcal{C}_{\ell_{\mathrm{BAL}},\mathcal{H}}(h,x)]\\
&\leq \mathop{\mathbb{E}}_{x}\left[\overline{\Gamma}(\Delta\mathcal{C}_{\ell_{\mathrm{GCA}},\mathcal{H}}(h,x))\right]\\
&\leq \overline{\Gamma}\left(\mathop{\mathbb{E}}_{x}[\Delta\mathcal{C}_{\ell_{\mathrm{GCA}},\mathcal{H}}(h,x)]\right)\\
&\qquad\qquad \text{(concavity of } \Gamma \text{ and Jensen's ineq.)}\\
&= \Gamma\left(\mathcal{R}_{\ell_{\mathrm{GCA}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GCA}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{GCA}}}(\mathcal{H})\right),
\end{aligned}
$$

where $\overline{\Gamma}(t) = \beta\left(\frac{1}{\mathsf{p}_{\min}}\right)^{1-\alpha} t^{\alpha}$. This concludes the first part of the proof. The second part follows directly from the fact that the minimizability gap $\mathcal{M}_{\ell_{\mathrm{GCA}}}(\mathcal{H})$ vanishes when the approximation error, $\mathcal{A}_{\ell_{\mathrm{GCA}}}(\mathcal{H})$, is zero. This concludes the first part of the proof. The second part follows directly using the fact that when the approximation error is zero: $\mathcal{A}_{\ell_{\mathrm{GCA}}}(\mathcal{H}) = 0$, the minimizability gap $\mathcal{M}_{\ell_{\mathrm{GCA}}}(\mathcal{H})$ vanishes. □

Note that, for simplicity, we assumed $\rho_y = 1$ for all y in Theorem 5 and its proof. To handle varying values of ρ_y, we can directly extend the $\mathcal{H}$-consistency bounds for the general cross-entropy (GCE) family, as derived in [Mao et al., 2023f,b], to the setting where GCE uses distinct ρ_y values. We can then similarly show that these extended bounds for the GCE family can be transformed into bounds for the GCA losses.

F Negative results for the LA losses: proof of Theorem 2

Theorem 2. *When $\tau \neq 1$, the LA loss ℓ_{LA} is not Bayes-consistent with respect to the balanced loss ℓ_{BAL}.*

Proof. The Bayes classifier h^*_{LA} of the LA loss satisfies the following condition:

$$
\frac{e^{h^*_{\mathrm{LA}}(x,y)+\tau\log(\mathsf{p}(y))}}{\sum_{y'\in\mathcal{Y}} e^{h^*_{\mathrm{LA}}(x,y')+\tau\log(\mathsf{p}(y'))}} = \mathsf{p}(y\mid x)
$$

By rearranging the terms, we have

$$\begin{aligned} e^{h^*_{\mathrm{LA}}(x,y)} &= \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)^\tau} \sum_{y' \in \mathcal{Y}} e^{h^*_{\mathrm{LA}}(x,y') + \tau \log(\mathsf{p}(y'))} \\ &= \frac{\mathsf{p}(x \mid y)\mathsf{p}(y)}{\mathsf{p}(y)^\tau \mathsf{p}(x)} \sum_{y' \in \mathcal{Y}} e^{h^*_{\mathrm{LA}}(x,y') + \tau \log(\mathsf{p}(y'))} && \text{(Bayes' theorem)} \\ &= \frac{\mathsf{p}(x \mid y)\mathsf{p}(y)^{1-\tau}}{\mathsf{p}(x)} \sum_{y' \in \mathcal{Y}} e^{h^*_{\mathrm{LA}}(x,y') + \tau \log(\mathsf{p}(y'))}. \end{aligned}$$

Thus, since the term $\frac{\sum_{y' \in \mathcal{Y}} e^{h^*_{\mathrm{LA}}(x,y') + \tau \log(\mathsf{p}(y'))}}{\mathsf{p}(x)}$ does not depend on y, we obtain

$$\mathsf{h}^*_{\mathrm{LA}}(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} h^*_{\mathrm{LA}}(x,y) = \operatorname*{argmax}_{y \in \mathcal{Y}} e^{h^*_{\mathrm{LA}}(x,y)} = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathsf{p}(x \mid y)\mathsf{p}(y)^{1-\tau}.$$

By Lemma 1, we know that the Bayes classifier h^*_{bal} of the Balanced loss satisfies that

$$\mathsf{h}^*_{\mathrm{bal}} = \operatorname*{argmax}_{y \in \mathcal{Y}} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathsf{p}(x \mid y).$$

Therefore, for any $\tau \neq 1$, there exists a distribution such that $\mathsf{h}^*_{\mathrm{LA}}(x) \neq \mathsf{h}^*_{\mathrm{bal}}$. This implies that when $\tau \neq 1$, the LA loss ℓ_{LA} is not Bayes-consistent with respect to the balanced loss ℓ_{BAL}. □

G Bayes-Consistency for the GLA losses: proof of Theorem 3

Theorem 3. *For any $q \in [0,1)$, the GLA Loss ℓ_{GLA} is Bayes-consistent with respect to the balanced loss ℓ_{BAL}.*

Proof. The Bayes classifier h^*_{GLA} of the GLA loss satisfies the following condition:

$$\frac{e^{h^*_{\mathrm{GLA}}(x,y) + \frac{\log(\mathsf{p}(y))}{1-q}}}{\sum_{y' \in \mathcal{Y}} e^{h^*_{\mathrm{GLA}}(x,y') + \frac{\log(\mathsf{p}(y'))}{1-q}}} = \frac{(\mathsf{p}(y \mid x))^{\frac{1}{1-q}}}{\sum_{y' \in \mathcal{Y}} (\mathsf{p}(y' \mid x))^{\frac{1}{1-q}}}$$

By rearranging the terms, we have

$$\begin{aligned} e^{h^*_{\mathrm{GLA}}(x,y)} &= \frac{(\mathsf{p}(y \mid x))^{\frac{1}{1-q}}}{(\mathsf{p}(y))^{\frac{1}{1-q}}} \frac{\sum_{y' \in \mathcal{Y}} e^{h^*_{\mathrm{GLA}}(x,y') + \frac{\log(\mathsf{p}(y'))}{1-q}}}{\sum_{y' \in \mathcal{Y}} (\mathsf{p}(y' \mid x))^{\frac{1}{1-q}}} \\ &= \left(\frac{\mathsf{p}(x \mid y)}{\mathsf{p}(x)}\right)^{\frac{1}{1-q}} \frac{\sum_{y' \in \mathcal{Y}} e^{h^*_{\mathrm{GLA}}(x,y') + \frac{\log(\mathsf{p}(y'))}{1-q}}}{\sum_{y' \in \mathcal{Y}} (\mathsf{p}(y' \mid x))^{\frac{1}{1-q}}}. && \text{(Bayes' theorem)} \end{aligned}$$

Thus, since the term $\frac{\sum_{y' \in \mathcal{Y}} e^{h^*_{\mathrm{GLA}}(x,y') + \frac{\log(\mathsf{p}(y'))}{1-q}}}{\sum_{y' \in \mathcal{Y}} (\mathsf{p}(y'|x))^{\frac{1}{1-q}}}$ does not depend on y, we obtain

$$\mathsf{h}^*_{\mathrm{GLA}}(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} h^*_{\mathrm{GLA}}(x,y) = \operatorname*{argmax}_{y \in \mathcal{Y}} e^{h^*_{\mathrm{GLA}}(x,y)} = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathsf{p}(x \mid y).$$

By Lemma 1, we know that the Bayes classifier h^*_{bal} of the Balanced loss satisfies that

$$\mathsf{h}^*_{\mathrm{bal}} = \operatorname*{argmax}_{y \in \mathcal{Y}} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathsf{p}(x \mid y).$$

Therefore, we have $\mathsf{h}^*_{\mathrm{GLA}}(x) = \mathsf{h}^*_{\mathrm{bal}}$. This implies that the GLA loss ℓ_{GLA} is Bayes-consistent with respect to the balanced loss ℓ_{BAL}. □

H $\mathcal{H}$-Consistency for the GLA losses: proof of Theorem 4

Theorem 4. *Assume that $\mathcal{H}$ is complete. Then, for any $q \in [0,1)$, the following $\mathcal{H}$-consistency bound holds for the GLA loss ℓ_{GLA}:*

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \Gamma\Big(\mathcal{R}_{\ell_{\mathrm{GLA}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{GLA}}}(\mathcal{H})\Big),$$

where $\Gamma(t) = \frac{\sqrt{2t}}{\mathsf{p}_{\min}}$ for $q = 0$, and $\Gamma(t) = \frac{\sqrt{2t}}{(\mathsf{p}_{\min})^{\frac{1}{1-q}}(1-q)^{\frac{1}{2}}}$ for $q \in (0,1)$. In the special case where the approximation error $\mathcal{A}_{\ell_{\mathrm{GLA}}}(\mathcal{H}) = 0$, the bound simplifies to:

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \Gamma\Big(\mathcal{R}_{\ell_{\mathrm{GLA}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H})\Big),$$

Proof. The proof involves a characterization of the conditional regret of the balanced loss and the use of Gibbs distributions and Pinsker-type inequalities for analyzing GLA losses.

By Lemma 1, for complete hypothesis sets, the conditional regret of the balanced loss can be expressed as follows:

$$\Delta\mathcal{C}_{\ell_{\mathrm{BAL}},\mathcal{H}}(h,x) = \max_{y\in\mathcal{Y}} \frac{\mathsf{p}(y \mid x)}{\mathsf{p}(y)} - \frac{\mathsf{p}(\mathsf{h}(x)) \mid x)}{\mathsf{p}(\mathsf{h}(x))}.$$

Let $\mathsf{y}(x) = \operatorname{argmax}_{y\in\mathcal{Y}} \frac{\mathsf{p}(y|x)}{\mathsf{p}(y)}$, where we choose the label with the same deterministic strategy for breaking ties as that of $\mathsf{h}(x) = \operatorname{argmax}_{y\in\mathcal{Y}} h(x,y)$. We analyze by cases.

Case I: $q = 0$. In this case, the conditional regret for the GLA loss can be written as

$$\mathcal{C}_{\ell_{\mathrm{GLA}}}(h,x)) = -\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x) \log\left(\frac{e^{h(x,y)+\log(\mathsf{p}(y))}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')+\log(\mathsf{p}(y'))}}\right) = -\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x) \log\big(\overline{\mathcal{S}}(x,y)\big)$$

where we let $\overline{\mathcal{S}}(x,y) = \frac{e^{\overline{h}(x,y)}}{\sum_{y'\in\mathcal{Y}} e^{\overline{h}(x,y')}} \in [0,1]$ for any $y \in \mathcal{Y}$ with $\overline{h}(x,y) = h(x,y) + \log(\mathsf{p}(y))$ and the constraint that $\sum_{y\in\mathcal{Y}} \overline{\mathcal{S}}(x,y) = 1$. Note that $\overline{\mathcal{S}}$ can be viewed as a Gibbs distribution induced by h with prior $\mathsf{p}(y)$. Leveraging the facts that $\overline{\mathcal{S}}$ is a surjection and $\mathcal{H}$ is complete, minimizing over $\overline{\mathcal{S}}$, we know that $\mathcal{C}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H},x)$ has the following form:

$$\mathcal{C}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H},x) = -\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x) \log(\mathsf{p}(y \mid x)).$$

Thus, we obtain

$$\begin{aligned}
\Delta\mathcal{C}_{\ell_{\mathrm{GLA}}\mathcal{H}}(h,x) &= \mathcal{C}_{\ell_{\mathrm{GLA}}}(h,x) - \mathcal{C}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H},x)\\
&= \sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x) \log(\mathsf{p}(y \mid x)) - \sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x) \log(\mathcal{S}(x,y))\\
&= \sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x) \log(\mathsf{p}(y \mid x)) - \sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x) \log\left(\frac{e^{h(x,y)+\log(\mathsf{p}(y))}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')+\log(\mathsf{p}(y'))}}\right)\\
&= \sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x) \log\left(\mathsf{p}(y \mid x) \frac{\sum_{y'\in\mathcal{Y}} e^{h(x,y')+\log(\mathsf{p}(y'))}}{e^{h(x,y)+\log(\mathsf{p}(y))}}\right)\\
&= \mathsf{D}\big(\mathsf{p}(\cdot \mid x) \| \overline{S}(x,\cdot)\big)
\end{aligned}$$

where $\mathsf{D}(\mathsf{p}\|\mathsf{q})$ is the relative entropy of two distributions p and q. Consider the case where $\mathsf{y}(x) \neq \mathsf{h}(x)$. Then, by Pinsker's inequality [Mohri et al., 2018, Proposition E.7], we have

$$\begin{aligned}
&\Delta\mathcal{C}_{\ell_{\mathrm{GLA}}\mathcal{H}}(h,x)\\
&= \mathsf{D}\big(\mathsf{p}(\cdot \mid x) \| \overline{S}(x,\cdot)\big)\\
&\geq \frac{1}{2}\big\|\mathsf{p}(\cdot \mid x) - \overline{S}(x,\cdot)\big\|_1^2 && \text{(Pinsker's inequality)}\\
&\geq \frac{1}{2}\Big(\big|\mathsf{p}(\mathsf{y}(x) \mid x) - \overline{S}(x,\mathsf{y}(x))\big| + \big|\mathsf{p}(\mathsf{h}(x) \mid x) - \overline{S}(x,\mathsf{h}(x))\big|\Big)^2\\
&= \frac{1}{2}\left(\mathsf{p}(\mathsf{y}(x))\left|\frac{\mathsf{p}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))} - \frac{\overline{S}(x,\mathsf{y}(x))}{\mathsf{p}(\mathsf{y}(x))}\right| + \mathsf{p}(\mathsf{h}(x))\left|\frac{\mathsf{p}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))} - \frac{\overline{S}(x,\mathsf{h}(x))}{\mathsf{p}(\mathsf{h}(x))}\right|\right)^2.
\end{aligned}$$

Plugging the expression of π_h, we have

$$
\begin{aligned}
&\Delta\mathcal{C}_{\ell_{\mathrm{GLA}}\mathcal{H}}(h,x)\\
&\geq \frac{1}{2}\left(\mathsf{p}(\mathsf{y}(x))\left|\frac{\mathsf{p}(\mathsf{y}(x)\mid x)}{\mathsf{p}(\mathsf{y}(x))}-\frac{e^{h(x,\mathsf{y}(x))}}{\sum_{y'\in\mathcal{Y}}e^{h(x,y')+\log(\mathsf{p}(y'))}}\right|+\mathsf{p}(\mathsf{h}(x))\left|\frac{\mathsf{p}(\mathsf{h}(x)\mid x)}{\mathsf{p}(\mathsf{h}(x))}-\frac{e^{h(x,\mathsf{h}(x))}}{\sum_{y'\in\mathcal{Y}}e^{h(x,y')+\log(\mathsf{p}(y'))}}\right|\right)^2\\
&\geq \frac{(\mathsf{p}_{\min})^2}{2}\left(\left|\frac{\mathsf{p}(\mathsf{y}(x)\mid x)}{\mathsf{p}(\mathsf{y}(x))}-\frac{e^{h(x,\mathsf{y}(x))}}{\sum_{y'\in\mathcal{Y}}e^{h(x,y')+\log(\mathsf{p}(y'))}}\right|+\left|\frac{\mathsf{p}(\mathsf{h}(x)\mid x)}{\mathsf{p}(\mathsf{h}(x))}-\frac{e^{h(x,\mathsf{h}(x))}}{\sum_{y'\in\mathcal{Y}}e^{h(x,y')+\log(\mathsf{p}(y'))}}\right|\right)^2\\
&\geq \frac{(\mathsf{p}_{\min})^2}{2}\left|\frac{\mathsf{p}(\mathsf{y}(x)\mid x)}{\mathsf{p}(\mathsf{y}(x))}-\frac{\mathsf{p}(\mathsf{h}(x)\mid x)}{\mathsf{p}(\mathsf{h}(x))}+\frac{e^{h(x,\mathsf{h}(x))}}{\sum_{y'\in\mathcal{Y}}e^{h(x,y')+\log(\mathsf{p}(y'))}}-\frac{e^{h(x,\mathsf{y}(x))}}{\sum_{y'\in\mathcal{Y}}e^{h(x,y')+\log(\mathsf{p}(y'))}}\right|^2\\
&\qquad\qquad (|a|+|b|\geq|a-b|)\\
&\geq \frac{(\mathsf{p}_{\min})^2}{2}\left|\frac{\mathsf{p}(\mathsf{y}(x)\mid x)}{\mathsf{p}(\mathsf{y}(x))}-\frac{\mathsf{p}(\mathsf{h}(x)\mid x)}{\mathsf{p}(\mathsf{h}(x))}\right|^2\\
&\left(\tfrac{\mathsf{p}(\mathsf{y}(x)|x)}{\mathsf{p}(\mathsf{y}(x))}-\tfrac{\mathsf{p}(\mathsf{h}(x)|x)}{\mathsf{p}(\mathsf{h}(x))}\geq 0 \text{ and } \tfrac{e^{h(x,\mathsf{h}(x))}}{\sum_{y'\in\mathcal{Y}}e^{h(x,y')+\log(\mathsf{p}(y'))}}-\tfrac{e^{h(x,\mathsf{y}(x))}}{\sum_{y'\in\mathcal{Y}}e^{h(x,y')+\log(\mathsf{p}(y'))}}\geq 0 \text{ by def. of } \mathsf{y}(x) \text{ and } \mathsf{h}(x)\right)\\
&= \frac{(\mathsf{p}_{\min})^2}{2}(\Delta\mathcal{C}_{\ell_{\mathrm{BAL}},\mathcal{H}}(h,x))^2.
\end{aligned}
$$

Then, by taking the expectation on both sides and using the Jensen's inequality, we obtain

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h)-\mathcal{R}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H})+\mathcal{M}_{\ell_{\mathrm{BAL}}}(\mathcal{H})\leq\Gamma\Big(\mathcal{R}_{\ell_{\mathrm{GLA}}}(h)-\mathcal{R}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H})+\mathcal{M}_{\ell_{\mathrm{GLA}}}(\mathcal{H})\Big),$$

where $\Gamma(t)=\frac{\sqrt{2t}}{\mathsf{p}_{\min}}$.

Case II: $q\in(0,1)$. In this case, the conditional regret for the GLA loss can be written as

$$\mathcal{C}_{\ell_{\mathrm{GLA}}}(h,x))=-\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)\Psi^q\left(\frac{e^{h(x,y)+\frac{\log(\mathsf{p}(y))}{1-q}}}{\sum_{y'\in\mathcal{Y}}e^{h(x,y')+\frac{\log(\mathsf{p}(y'))}{1-q}}}\right)=-\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)\Psi^q\big(\overline{\mathsf{S}}(x,y)\big)$$

where we let $\overline{\mathsf{S}}(x,y)=\frac{e^{\overline{h}(x,y)}}{\sum_{y'\in\mathcal{Y}}e^{\overline{h}(x,y')}}\in[0,1]$ for any $y\in\mathcal{Y}$ with $\overline{h}(x,y)=h(x,y)+\frac{\log(\mathsf{p}(y))}{1-q}$ and the constraint that $\sum_{y\in\mathcal{Y}}\overline{\mathsf{S}}(x,y)=1$. Note that $\overline{\mathsf{S}}$ can be viewed as a Gibbs distribution induced by h. Leveraging the facts that $\overline{\mathsf{S}}$ is a surjection and $\mathcal{H}$ is complete, minimizing over $\overline{\mathsf{S}}$, we know that $\mathcal{C}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H},x)$ has the following form:

$$\mathcal{C}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H},x)=\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)\Psi^q\left(\frac{\mathsf{p}(y\mid x)^{\frac{1}{1-q}}}{\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)^{\frac{1}{1-q}}}\right)=\frac{1}{q}\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)\left(1-\left(\frac{\mathsf{p}(y\mid x)^{\frac{1}{1-q}}}{\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)^{\frac{1}{1-q}}}\right)^q\right).$$

Thus, we obtain

$$
\begin{aligned}
&\Delta\mathcal{C}_{\ell_{\mathrm{GLA}}\mathcal{H}}(h,x)\\
&=\mathcal{C}_{\ell_{\mathrm{GLA}}}(h,x)-\mathcal{C}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H},x)\\
&=\frac{1}{q}\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)\big(1-\overline{S}(x,y)^q\big)-\frac{1}{q}\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)\left(1-\left(\frac{\mathsf{p}(y\mid x)^{\frac{1}{1-q}}}{\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)^{\frac{1}{1-q}}}\right)^q\right)\\
&=\frac{\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)\left(\left(\frac{\mathsf{p}(y|x)^{\frac{1}{1-q}}}{\sum_{y\in\mathcal{Y}}\mathsf{p}(y|x)^{\frac{1}{1-q}}}\right)^q-\overline{S}(x,y)^q\right)}{q}\\
&=\left(\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)^{\frac{1}{1-q}}\right)^{1-q}\frac{\left(1-\sum_{y\in\mathcal{Y}}\left(\frac{\mathsf{p}(y|x)^{\frac{1}{1-q}}}{\sum_{y\in\mathcal{Y}}\mathsf{p}(y|x)^{\frac{1}{1-q}}}\right)^{1-q}\overline{S}(x,y)^q\right)}{q}\\
&=\left(\sum_{y\in\mathcal{Y}}\mathsf{p}(y\mid x)^{\frac{1}{1-q}}\right)^{1-q}\mathsf{T}_{1-q}\big(\mathsf{s}(\cdot\mid x)\|\overline{S}(x,\cdot)\big)
\end{aligned}
$$

where $\mathsf{T}_q(\mathsf{p}\|\mathsf{q})$ denotes the Tsallis relative entropy of order q between the distributions p and q, and $\mathsf{s}(y \mid x) = \frac{\mathsf{p}(y|x)^{\frac{1}{1-q}}}{\sum_{y\in\mathcal{Y}} \mathsf{p}(y|x)^{\frac{1}{1-q}}}$. Consider the case where $\mathsf{y}(x) \neq \mathsf{h}(x)$. Then, by a Pinsker-type inequality [Rastegin, 2013, Eq. (4.13)], we have

$$
\begin{aligned}
&\Delta\mathcal{C}_{\ell_{\mathrm{GLA}}\mathcal{H}}(h,x) \\
&= \left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{1-q} \mathsf{T}_{1-q}\big(\mathsf{s}(\cdot \mid x)\|\overline{S}(x,\cdot)\big) \\
&\geq \frac{1-q}{2}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{1-q} \big\|\mathsf{s}(\cdot \mid x) - \overline{S}(x,\cdot)\big\|_1^2 \\
&\qquad\qquad \text{(Pinsker-type inequality [Rastegin, 2013, Eq. (4.13)])} \\
&\geq \frac{1-q}{2}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{1-q} \Big(\big|\mathsf{s}(\mathsf{y}(x) \mid x) - \overline{S}(x,\mathsf{y}(x))\big| + \big|\mathsf{s}(\mathsf{h}(x) \mid x) - \overline{S}(x,\mathsf{h}(x))\big|\Big)^2 \\
&= \frac{1-q}{2}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{1-q} \\
&\qquad \times \left(\mathsf{p}(\mathsf{y}(x))^{\frac{1}{1-q}}\left|\frac{\mathsf{s}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))^{\frac{1}{1-q}}} - \frac{\overline{S}(x,\mathsf{y}(x))}{\mathsf{p}(\mathsf{y}(x))^{\frac{1}{1-q}}}\right| + \mathsf{p}(\mathsf{h}(x))^{\frac{1}{1-q}}\left|\frac{\mathsf{s}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))^{\frac{1}{1-q}}} - \frac{\overline{S}(x,\mathsf{h}(x))}{\mathsf{p}(\mathsf{h}(x))^{\frac{1}{1-q}}}\right|\right)^2 \\
&\geq \frac{1-q}{2}(\mathsf{p}_{\min})^{\frac{2}{1-q}}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{1-q} \\
&\qquad \times \left(\left|\frac{\mathsf{s}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))^{\frac{1}{1-q}}} - \frac{\mathsf{s}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))^{\frac{1}{1-q}}} + \frac{\overline{S}(x,\mathsf{h}(x))}{\mathsf{p}(\mathsf{h}(x))^{\frac{1}{1-q}}} - \frac{\overline{S}(x,\mathsf{y}(x))}{\mathsf{p}(\mathsf{y}(x))^{\frac{1}{1-q}}}\right|\right)^2 \qquad (|a|+|b| \geq |a-b|) \\
&\geq \frac{1-q}{2}(\mathsf{p}_{\min})^{\frac{2}{1-q}}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{1-q}\left|\frac{\mathsf{s}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))^{\frac{1}{1-q}}} - \frac{\mathsf{s}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))^{\frac{1}{1-q}}}\right|^2. \\
&\qquad \left(\tfrac{\mathsf{s}(\mathsf{y}(x)|x)}{\mathsf{p}(\mathsf{y}(x))^{\frac{1}{1-q}}} - \tfrac{\mathsf{s}(\mathsf{h}(x)|x)}{\mathsf{p}(\mathsf{h}(x))^{\frac{1}{1-q}}} \geq 0 \text{ and } \tfrac{\overline{S}(x,\mathsf{h}(x))}{\mathsf{p}(\mathsf{h}(x))^{\frac{1}{1-q}}} - \tfrac{\overline{S}(x,\mathsf{y}(x))}{\mathsf{p}(\mathsf{y}(x))^{\frac{1}{1-q}}} \geq 0 \text{ by def. of } \mathsf{y}(x) \text{ and } \mathsf{h}(x)\right)
\end{aligned}
$$

Plugging the expression of $\mathsf{s}(\cdot \mid x)$, we have

$$
\begin{aligned}
&\Delta\mathcal{C}_{\ell_{\mathrm{GLA}}\mathcal{H}}(h,x) \\
&\geq \frac{1-q}{2}(\mathsf{p}_{\min})^{\frac{2}{1-q}}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{1-q}\left|\frac{\mathsf{s}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))^{\frac{1}{1-q}}} - \frac{\mathsf{s}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))^{\frac{1}{1-q}}}\right|^2 \\
&= \frac{1-q}{2}(\mathsf{p}_{\min})^{\frac{2}{1-q}}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{-q-1}\left|\left(\frac{\mathsf{p}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))}\right)^{\frac{1}{1-q}} - \left(\frac{\mathsf{p}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))}\right)^{\frac{1}{1-q}}\right|^2 \\
&\leq \frac{1-q}{2}(\mathsf{p}_{\min})^{\frac{2}{1-q}}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{-q-1}\left|\frac{\mathsf{p}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))}\left(\frac{\mathsf{p}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))}\right)^{\frac{q}{1-q}} - \frac{\mathsf{p}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))}\left(\frac{\mathsf{p}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))}\right)^{\frac{q}{1-q}}\right|^2 \\
&\geq \frac{1-q}{2}(\mathsf{p}_{\min})^{\frac{2}{1-q}}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{-q-1}\left|\frac{\mathsf{p}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))}\left(\frac{\mathsf{p}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))}\right)^{\frac{q}{1-q}} - \frac{\mathsf{p}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))}\left(\frac{\mathsf{p}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))}\right)^{\frac{q}{1-q}}\right|^2 \\
&\geq \frac{1-q}{2}(\mathsf{p}_{\min})^{\frac{2}{1-q}}\left(\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}}\right)^{-q-1}\left(\frac{\mathsf{p}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))}\right)^{\frac{2q}{1-q}}\left|\frac{\mathsf{p}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))} - \frac{\mathsf{p}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))}\right|^2.
\end{aligned}
$$

Next, using $\sum_{y\in\mathcal{Y}} \mathsf{p}(y \mid x)^{\frac{1}{1-q}} = \|p(\cdot \mid x)\|_{\frac{1}{1-q}}^{\frac{1}{1-q}} \leq \|p(\cdot \mid x)\|_1^{\frac{1}{1-q}} = 1$ and $\frac{\mathsf{p}(\mathsf{y}(x)|x)}{\mathsf{p}(\mathsf{y}(x))} = \max_{y\in\mathcal{Y}} \frac{\mathsf{p}(y|x)}{\mathsf{p}(y)} \geq 1$, we can write:

$$\begin{aligned}\Delta\mathcal{C}_{\ell_{\mathrm{GLA}},\mathcal{H}}(h,x) &\geq \frac{1-q}{2}(\mathsf{p}_{\min})^{\frac{2}{1-q}}\left|\frac{\mathsf{p}(\mathsf{y}(x) \mid x)}{\mathsf{p}(\mathsf{y}(x))} - \frac{\mathsf{p}(\mathsf{h}(x) \mid x)}{\mathsf{p}(\mathsf{h}(x))}\right|^2 \\ &= \frac{1-q}{2}(\mathsf{p}_{\min})^{\frac{2}{1-q}}(\Delta\mathcal{C}_{\ell_{\mathrm{BAL}},\mathcal{H}}(h,x))^2.\end{aligned}$$

Then, by taking the expectation on both sides and using the Jensen's inequality, we obtain

$$\mathcal{R}_{\ell_{\mathrm{BAL}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{BAL}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{BAL}}}(\mathcal{H}) \leq \Gamma\Big(\mathcal{R}_{\ell_{\mathrm{GLA}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{GLA}}}(\mathcal{H}) + \mathcal{M}_{\ell_{\mathrm{GLA}}}(\mathcal{H})\Big),$$

where $\Gamma(t) = \frac{\sqrt{2t}}{(\mathsf{p}_{\min})^{\frac{1}{1-q}}(1-q)^{\frac{1}{2}}}$. This concludes the first part of the proof. The second part follows directly using the fact that when the approximation error is zero: $\mathcal{A}_{\ell_{\mathrm{GLA}}}(\mathcal{H}) = 0$, the minimizability gap $\mathcal{M}_{\ell_{\mathrm{GLA}}}(\mathcal{H})$ vanishes. □

I Margin bound: proof of Theorem 7

Theorem 7 (Margin bound for cost-sensitive classification)**.** *Let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}\times[n]$ to $\mathbb{R}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following inequalities holds for all $h \in \mathcal{H}$:*

$$\mathcal{R}_{\mathsf{L}}(h) \leq \widehat{\mathcal{R}}_{S,\rho}(h) + 4\overline{C}\sqrt{2n}\,\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$\mathcal{R}_{\mathsf{L}}(h) \leq \widehat{\mathcal{R}}_{S,\rho}(h) + 4\overline{C}\sqrt{2n}\,\widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

Proof. Consider the family of functions taking values in $[0,1]$:

$$\mathcal{H}' = \{z = (x,y) \mapsto \mathsf{L}_\rho(h,x,y) \colon h \in \mathcal{H}\}.$$

By [Mohri et al., 2018, Theorem 3.3], with probability at least $1-\delta$, for all $g \in \mathcal{H}'$,

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^m g(z_i) + 2\widehat{\mathfrak{R}}_S(\mathcal{H}') + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}},$$

and thus, for all $h \in \mathcal{H}$,

$$\mathbb{E}[\mathsf{L}_\rho(h,x,y)] \leq \widehat{\mathcal{R}}_{S,\rho}(h) + 2\widehat{\mathfrak{R}}_S(\mathcal{H}') + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

Since $\mathcal{R}_{\mathsf{L}}(h) \leq \mathcal{R}_{\mathsf{L}_\rho}(h) = \mathbb{E}[\mathsf{L}_\rho(h,x,y)]$, we have

$$\mathcal{R}_{\mathsf{L}}(h) \leq \widehat{\mathcal{R}}_{S,\rho}(h) + 2\widehat{\mathfrak{R}}_S(\mathcal{H}') + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

Fix h, (x_i,y_i) and $\rho > 0$, define Ψ as follows:

$$\Psi([h(x_i,y)]_{y\in[n]}) = c(x_i,y_i)\max_{y'\in[n]}\{\Phi_\rho(h(x_i,y_i) - h(x_i,y'))\}.$$

Then, by the sub-additivity of the maximum operator, we can write for any $f, \widetilde{f} \in \mathcal{H}$:

$$\begin{aligned}&\Psi([h(x_i,y)]_{y\in[n]}) - \Psi([\widetilde{h}(x_i,y)]_{y\in[n]}) \\ &\leq c(x_i,y_i)\max_{y'\in[n]}\{\Phi_\rho(h(x_i,y_i) - h(x_i,y'))\} - c(x_i,y_i)\max_{y'\in[n]}\{\Phi_\rho(\widetilde{h}(x_i,y_i) - \widetilde{h}(x_i,y'))\} \\ &\leq \frac{2c(x_i,y_i)}{\rho}\{\|[h(x_i,y) - \widetilde{h}(x_i,y)]_{y\in[n]}\|_1\} \qquad \text{(by } \tfrac{1}{\rho}\text{-Lipschitzness of } \Phi_\rho) \\ &\leq \frac{2\overline{C}\sqrt{n}}{\rho}\|[h(x_i,y) - \widetilde{h}(x_i,y)]_{y\in[n]}\|_2.\end{aligned}$$

Thus, Ψ is $\frac{2\sqrt{n}}{\rho}$-Lipschitz with respect to the $\|\cdot\|_2$ norm. Thus, by the vector contraction lemma [Maurer, 2016, Cortes et al., 2016], $\widehat{\mathfrak{R}}_S(\mathcal{H}')$ can be bounded as follows:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}') \leq 2\overline{C}\sqrt{2n}\,\widehat{\mathfrak{R}}_S(\mathcal{H}).$$

This proves the second inequality. The first inequality, can be derived in the same way by using the first inequality of [Mohri et al., 2018, Theorem 3.3]. □