Multi-Class $\mathcal{H}$-Consistency Bounds

Pranjal Awasthi
Google Research
New York, NY 10011
pranjalawasthi@google.com

Anqi Mao
Courant Institute
New York, NY 10012
aqmao@cims.nyu.edu

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

Yutao Zhong
Courant Institute
New York, NY 10012
yutao@cims.nyu.edu

Abstract

We present an extensive study of $\mathcal{H}$-*consistency bounds* for multi-class classification. These are upper bounds on the target loss estimation error of a predictor in a hypothesis set $\mathcal{H}$, expressed in terms of the surrogate loss estimation error of that predictor. They are stronger and more significant guarantees than Bayes-consistency, $\mathcal{H}$-calibration or $\mathcal{H}$-consistency, and more informative than excess error bounds derived for $\mathcal{H}$ being the family of all measurable functions. We give a series of new $\mathcal{H}$-consistency bounds for surrogate multi-class losses, including max losses, sum losses, and constrained losses, both in the non-adversarial and adversarial cases, and for different differentiable or convex auxiliary functions used. We also prove that no non-trivial $\mathcal{H}$-consistency bound can be given in some cases. To our knowledge, these are the first $\mathcal{H}$-consistency bounds proven for the multi-class setting. Our proof techniques are also novel and likely to be useful in the analysis of other such guarantees.

1 Introduction

The loss functions optimized by learning algorithms are often distinct from the original one specified for a task. This is typically because optimizing the original loss is computationally intractable or because it does not admit some favorable properties of differentiability or smoothness. As an example, the loss function minimized by the support vector machine (SVM) algorithm is the hinge loss (Cortes and Vapnik, 1995) or the one associated to AdaBoost is the exponential loss (Schapire and Freund, 2012), both distinct from the binary classification loss used as a benchmark in applications. But, what learning guarantees can we rely on when using a surrogate loss? This is a fundamental question in learning theory that directly relates to the design of algorithms.

The standard property of *Bayes-consistency*, which has been shown to hold for several surrogate losses (Zhang, 2004a,b; Bartlett, Jordan, and McAuliffe, 2006; Tewari and Bartlett, 2007; Steinwart, 2007), does not supply a sufficient guarantee, since it only ensures that, *asymptotically*, near optimal minimizers of the surrogate excess loss nearly optimally minimize the target excess error. Moreover, this asymptotic property only holds for the full family of measurable functions, which of course is distinct from the more restricted hypothesis set used by a learning algorithm. In fact, it has been shown by Long and Servedio (2013), both theoretically and empirically, that for some hypothesis sets and distributions, the expected error of an algorithm minimizing a Bayes-consistent loss is bounded below by a positive constant, while that of an algorithm minimizing an inconsistent loss goes to zero.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

This suggests that a hypothesis set-dependent notion of $\mathcal{H}$-consistency is more pertinent to the study of consistency for learning (Long and Servedio, 2013), which has been used by Kuznetsov et al. (2014); Cortes et al. (2016a,b) and Zhang and Agarwal (2020) and more generally by Awasthi, Frank, Mao, Mohri, and Zhong (2021a) in an extensive study of both binary classification and *adversarial* binary classification losses, as defined in (Goodfellow et al., 2014; Madry et al., 2017; Tsipras et al., 2018; Carlini and Wagner, 2017). Nevertheless, $\mathcal{H}$-consistency remains an asymptotic property and does not provide guarantees for approximate surrogate loss minimizers that rely on finite samples.

Awasthi, Mao, Mohri, and Zhong (2022) recently presented a series of results providing $\mathcal{H}$-*consistency bounds* in binary classification. These are upper bounds on the target loss estimation error of a predictor in a hypothesis set $\mathcal{H}$, expressed in terms of the surrogate loss estimation error of that predictor. These guarantees are significantly stronger than the $\mathcal{H}$-calibration or $\mathcal{H}$-consistency properties studied by Awasthi et al. (2021a). They are also more informative than similar excess error bounds derived in the literature, which correspond to the special case where $\mathcal{H}$ is the family of all measurable functions (Zhang, 2004a; Bartlett et al., 2006; Mohri et al., 2018). Combining $\mathcal{H}$-consistency bounds with existing surrogate loss estimation bounds directly yields finite sample bounds on the estimation error for the original loss. See Appendix C for a more detailed discussion.

This paper presents an extensive study of $\mathcal{H}$-consistency bounds for multi-class classification. We show in Section 4.1 that, in general, no non-trivial $\mathcal{H}$-consistency bounds can be derived for multi-class *max losses* such as those of Crammer and Singer (2001), when used with a convex loss auxiliary function such as the hinge loss. On the positive side, we prove multi-class $\mathcal{H}$-consistency bounds for max losses under a realizability assumption and give multi-class $\mathcal{H}$-consistency bounds using as an auxiliary function the ρ-margin loss, without requiring a realizability assumption. For *sum losses*, that is multi-class losses such as that of Weston and Watkins (1998), we give a series of results, including a negative result when using as auxiliary function the hinge-loss, and $\mathcal{H}$-consistency bounds when using the exponential loss, the squared hinge-loss, and the ρ-margin loss (Section 4.2). We also present a series of results for the so-called *constrained losses*, such as the loss function adopted by Lee et al. (2004) in the analysis of multi-class SVM. Here, we prove multi-class $\mathcal{H}$-consistency bounds when using as an auxiliary function the hinge-loss, the squared hinge-loss, the exponential loss, and the ρ-margin loss (Section 4.3). We further give multi-class *adversarial* $\mathcal{H}$-consistency bounds for all three of the general multi-class losses just mentioned (max losses, sum losses and constrained losses) in Section 5.

We are not aware of any prior $\mathcal{H}$-consistency bound derived in the multi-class setting, even in the special case of $\mathcal{H}$ being the family of all measurable functions, whether in the non-adversarial or adversarial setting. All of our results are novel, including our proof techniques. Our results are given for the hypothesis set $\mathcal{H}$ being the family of all measurable functions, the family of linear functions, or the family of one-hidden-layer ReLU neural networks. The binary classification results of Awasthi et al. (2022) do not readily extend to the multi-class setting since the study of calibration and conditional risk is more complex, the form of the surrogate losses is more diverse, and in general the analysis is more involved and requires entirely novel proof techniques in the multi-class setting (see Section 3 for a more detailed discussion of this point).

We give a detailed discussion of related work in Appendix A. We start with the introduction of several multi-class definitions, as well as key concepts and definitions related to the study of $\mathcal{H}$-consistency bounds (Section 2).

2 Preliminaries

We consider the familiar multi-class classification scenario with $c \geq 2$ classes. We denote by $\mathcal{X}$ the input space and by $\mathcal{Y} = \{1, \ldots, c\}$ the set of classes or categories. Let $\mathcal{H}$ be a hypothesis set of functions mapping from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. The label $\mathsf{h}(x)$ associated by a hypothesis $h \in \mathcal{H}$ to $x \in \mathcal{X}$ is the one with the largest score: $\mathsf{h}(x) = \operatorname{argmax}_{y \in \mathcal{Y}} h(x, y)$ with an arbitrary but fixed deterministic strategy used for breaking ties. For simplicity, we fix that strategy to be the one selecting the label with the highest index under the natural ordering of labels. See Appendix B for a more detailed discussion of this choice.

The *margin* $\rho_h(x, y)$ of a hypothesis $h \in \mathcal{H}$ for a labeled example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is defined by

$$\rho_h(x, y) = h(x, y) - \max_{y' \neq y} h(x, y'),$$

that is the difference between the score assigned to (x, y) and that of the runner-up. Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and a loss function $\ell: \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the *generalization error* of a hypothesis $h \in \mathcal{H}$ and the *minimal generalization error* are defined as follows:

$$\mathcal{R}_\ell(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h, x, y)] \quad \text{and} \quad \mathcal{R}^*_{\ell,\mathcal{H}} = \inf_{h\in\mathcal{H}} \mathcal{R}_\ell(h).$$

The goal in multi-class classification is to select a hypothesis $h \in \mathcal{H}$ with small generalization error with respect to the multi-class $0/1$ loss defined, for any $h \in \mathcal{H}$, by $\ell_{0-1}(h, x, y) = \mathbb{1}_{\mathsf{h}(x)\neq y}$. In the adversarial scenario, the goal is to select a hypothesis $h \in \mathcal{H}$ with small *adversarial generalization error* defined, for any $\gamma \in (0, 1)$ and $p \in [1, +\infty]$, by $\mathcal{R}_{\ell_\gamma}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_\gamma(h, x, y)]$, where

$$\ell_\gamma(h, x, y) = \sup_{x':\|x-x'\|_p\leq\gamma} \mathbb{1}_{\rho_h(x',y)\leq 0} = \mathbb{1}_{\inf_{x':\|x-x'\|_p\leq\gamma} \rho_h(x',y)\leq 0},$$

is the adversarial multi-class $0/1$ loss. More generally, the *adversarial generalization error* and *minimal adversarial generalization error* for a loss function $\ell(h, x, y)$ are defined as follows:

$$\mathcal{R}_{\widetilde{\ell}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\widetilde{\ell}(h, x, y)\right] \quad \text{and} \quad \mathcal{R}^*_{\widetilde{\ell},\mathcal{H}} = \inf_{h\in\mathcal{H}} \mathcal{R}_{\widetilde{\ell}}(h),$$

where $\widetilde{\ell}(h, x, y) = \sup_{x':\|x-x'\|_p\leq\gamma} \ell(h, x', y)$ is the supremum-based counterpart of ℓ.

For a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, we define, for any $x \in \mathcal{X}$, $p(x) = (p(x, 1), \ldots, p(x, c))$, where $p(x, y) = \mathcal{D}(Y = y | X = x)$ is the conditional probability of $Y = y$ given $X = x$. We can then write the generalization error as $\mathcal{R}_\ell(h) = \mathbb{E}_X[\mathcal{C}_\ell(h, x)]$, where $\mathcal{C}_\ell(h, x)$ is the *conditional ℓ-risk* defined by $\mathcal{C}_\ell(h, x) = \sum_{y\in\mathcal{Y}} p(x, y)\ell(h, x, y)$. We will denote by $\mathcal{P}$ a set of distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and by $\mathcal{P}_{\text{all}}$ the set of all such distributions. For convenience, we define $y_{\max}$ by $y_{\max} = \operatorname{argmax}_{y\in\mathcal{Y}} p(x, y)$. When there is a tie, we pick the label with the highest index under the natural ordering of labels.

The *minimal conditional ℓ-risk* is denoted by $\mathcal{C}^*_{\ell,\mathcal{H}}(x) = \inf_{h\in\mathcal{H}} \mathcal{C}_\ell(h, x)$. We also use the following shorthand for the gap $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x) = \mathcal{C}_\ell(h, x) - \mathcal{C}^*_{\ell,\mathcal{H}}(x)$ and call $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x)\mathbb{1}_{\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)>\epsilon}$ the *conditional ϵ-regret* for ℓ. For convenience, we also define, for any vector $\tau = (\tau_1, \ldots, \tau_c)$ in the probability simplex of $\mathbb{R}^c$, $\mathcal{C}_\ell(h, x, \tau) = \sum_{y\in\mathcal{Y}} \tau_y \ell(h, x, y)$, $\mathcal{C}^*_{\ell,\mathcal{H}}(x, \tau) = \inf_{h\in\mathcal{H}} \mathcal{C}_\ell(h, x, \tau)$ and $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x, \tau) = \mathcal{C}_\ell(h, x, \tau) - \mathcal{C}^*_{\ell,\mathcal{H}}(x, \tau)$. Thus, we have $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x, p(x)) = \Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x)$. For any $\epsilon > 0$, we will denote by $[t]_\epsilon$ the ϵ-truncation of $t \in \mathbb{R}$ defined by $t\mathbb{1}_{t>\epsilon}$. Thus, the conditional ϵ-regret can be rewritten as $[\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x)]_\epsilon$.

For a hypothesis set $\mathcal{H}$ and distribution $\mathcal{D}$, we also define the $(\ell, \mathcal{H})$-*minimizability gap* as $\mathcal{M}_{\ell,\mathcal{H}} = \mathcal{R}^*_{\ell,\mathcal{H}} - \mathbb{E}_X\left[\mathcal{C}^*_{\ell,\mathcal{H}}(x)\right]$, that is the difference between the best-in class error and the expectation of the minimal conditional ℓ-risk. This is a key quantity appearing in our bounds that we cannot hope to estimate or minimize. Its value only depends on the distribution $\mathcal{D}$ and the hypothesis set $\mathcal{H}$. As an example, when $\mathcal{H}$ is the family of all measurable functions, then the minimizability gap for the multi-class $0/1$ loss is zero for any distribution $\mathcal{D}$.

3 General theorems

The general form of the $\mathcal{H}$-*consistency bounds* that we are seeking for a surrogate loss ℓ_1 of a target loss ℓ_2 is $\mathcal{R}_{\ell_2}(h) - \mathcal{R}^*_{\ell_2,\mathcal{H}} \leq f(\mathcal{R}_{\ell_1}(h) - \mathcal{R}^*_{\ell_1,\mathcal{H}})$ for all $h \in \mathcal{H}$, for some non-decreasing function f. To derive such bounds for surrogate multi-class losses, we draw on the following two general theorems, which show that, under some conditions, the target loss estimation error can be bounded by some functional form of the surrogate loss estimation error involving minimizability gaps.

Theorem 1 (**Distribution-dependent Ψ-bound**). *Assume that there exists a convex function $\Psi: \mathbb{R}_+ \to \mathbb{R}$ with $\Psi(0) \geq 0$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$ and $\mathcal{D} \in \mathcal{P}$: $\Psi\left([\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h, x)]_\epsilon\right) \leq \Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h, x)$. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D} \in \mathcal{P}$,*

$$\Psi\left(\mathcal{R}_{\ell_2}(h) - \mathcal{R}^*_{\ell_2,\mathcal{H}} + \mathcal{M}_{\ell_2,\mathcal{H}}\right) \leq \mathcal{R}_{\ell_1}(h) - \mathcal{R}^*_{\ell_1,\mathcal{H}} + \mathcal{M}_{\ell_1,\mathcal{H}} + \max\{\Psi(0), \Psi(\epsilon)\}.$$

Theorem 2 (**Distribution-dependent Γ-bound**). *Assume that there exists a concave function $\Gamma: \mathbb{R}_+ \to \mathbb{R}$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$ and $\mathcal{D} \in \mathcal{P}$: $[\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h, x)]_\epsilon \leq \Gamma(\Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h, x))$. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D} \in \mathcal{P}$,*

$$\mathcal{R}_{\ell_2}(h) - \mathcal{R}^*_{\ell_2,\mathcal{H}} \leq \Gamma\left(\mathcal{R}_{\ell_1}(h) - \mathcal{R}^*_{\ell_1,\mathcal{H}} + \mathcal{M}_{\ell_1,\mathcal{H}}\right) - \mathcal{M}_{\ell_2,\mathcal{H}} + \epsilon.$$

The theorems show that, to derive such bounds for a specific hypothesis set and a set of distributions, it suffices to verify that for the same hypothesis set and set of distributions, the conditional ϵ-regret for the target loss can be upper bounded with the same functional form of the gap between the conditional risk and minimal conditional risk of the surrogate loss. These results are similar to their binary classification counterparts due to Awasthi et al. (2022). In particular, the conditional ℓ-risk $\mathcal{C}_\ell(h, x)$ in our theorems is the multi-class generalization of their binary definition. The proofs are similar and are included in Appendix E for completeness.

For a given hypothesis set $\mathcal{H}$, the resulting bounds suggest three key ingredients for the choice of a surrogate loss: (1) the functional form of the $\mathcal{H}$-consistency bound, which is specified by the function Ψ or Γ; (2) the smoothness of the loss and more generally its optimization virtues, as needed for the minimization of $\mathcal{R}_{\ell_1}(h) - \mathcal{R}^*_{\ell_1,\mathcal{H}}$; (3) and the approximation properties of the surrogate loss function which determine the value of the minimizability gap $\mathcal{M}_{\ell_1,\mathcal{H}}$. Our quantitative $\mathcal{H}$-consistency bounds can help select the most favorable surrogate loss function among surrogate losses with good optimization merits and comparable approximation properties.

In Section 4 and Section 5, we will apply Theorem 1 and Theorem 2 to the analysis of multi-class loss functions and hypothesis sets widely used in practice. Here, we wish to first comment on the novelty of our results and proof techniques. Let us emphasize that although the general tools of Theorems 1 and 2 are the multi-class generalization of that in (Awasthi et al., 2022), the binary classification results of Awasthi et al. (2022) do not readily extend to the multi-class setting. This is true, even in the classical study of Bayes-consistency, where the multi-class setting (Tewari and Bartlett, 2007) does not readily follow the binary case (Bartlett et al., 2006) and required an alternative analysis and new proofs. Note that, additionally, in the multi-class setting, surrogate losses are more diverse: we will distinguish max losses, sum losses, and constrained losses and present an analysis for each loss family with various auxiliary functions for each (see Section 4).

Proof techniques. More specifically, the need for novel proof techniques stems from the following. To use Theorem 1 and Theorem 2, we need to find Ψ and Γ such that the inequality conditions in these theorems hold. This requires us to characterize the conditional risk and the minimal conditional risk of the multi-class zero-one loss function and the corresponding ones for diverse surrogate loss functions in both the non-adversarial and adversarial scenario. Unlike the binary case, such a characterization in the multi-class setting is very difficult. For example, for the constrained loss, solving the minimal conditional risk given a hypothesis set is equivalent to solving a c-dimensional constrained optimization problem, which does not admit an analytical expression. In contrast, in the binary case, solving the minimal conditional risk is equivalent to solving a minimization problem for a univariate function and the needed function Ψ can be characterized explicitly by the $\mathcal{H}$-estimation error transformation, as shown in (Awasthi et al., 2022). Unfortunately, such binary classification transformation tools cannot be adapted to the multi-class setting. Instead, in our proof for the multi-class setting, we adopt a new idea that avoids directly characterizing the explicit expression of the minimal conditional risk.

For example, for the constrained loss, we leverage the condition of (Lee et al., 2004) that the scores sum to zero, and appropriately choose a hypothesis $\overline{h}$ that differs from h only by its scores for $\mathsf{h}(x)$ and $y_{\max}$ (see Appendix K). Then, we can upper bound the minimal conditional risk by the conditional risk of $\overline{h}$ without having to derive the closed form expression of the minimal conditional risk. Therefore, the conditional regret of the surrogate loss can be lower bounded by that of the zero-one loss with an appropriate function Ψ. To the best of our knowledge, this proof idea and technique are entirely novel. We believe that they can be used for the analysis of other multi-class surrogate losses. Furthermore, all of our multi-class $\mathcal{H}$-consistency results are new. Likewise, our proofs of the $\mathcal{H}$-consistency bounds for sum losses for the squared hinge loss and exponential loss use similarly a new technique and idea, and so does the proof for the ρ-margin loss. Furthermore, we also present an analysis of the adversarial scenario (see Section 5), for which the multi-class proofs are also novel. Finally, our bounds in the multi-class setting are more general: for $c = 2$, we recover the binary classification bounds of (Awasthi et al., 2022). Thus, our bounds benefit from the same tightness guarantees shown by (Awasthi et al., 2022). A further analysis of the tightness of our guarantees in the multi-class setting is left to future work.

4 $\mathcal{H}$-consistency bounds

In this section, we discuss $\mathcal{H}$-consistency bounds in the non-adversarial scenario where the target loss ℓ_2 is ℓ_{0-1}, the multi-class $0/1$ loss. The lemma stated next characterizes the minimal conditional ℓ_{0-1}-risk and the corresponding conditional ϵ-regret, which will be helpful for instantiating Theorems 1 and 2 in the non-adversarial scenario. For any $x \in \mathcal{X}$, we will denote, by $\mathsf{H}(x)$ the set of labels generated by hypotheses in $\mathcal{H}$: $\mathsf{H}(x) = \{\mathsf{h}(x) : h \in \mathcal{H}\}$.

Lemma 3. *For any $x \in \mathcal{X}$, the minimal conditional ℓ_{0-1}-risk and the conditional ϵ-regret for ℓ_{0-1} can be expressed as follows:*

$$\mathcal{C}^*_{\ell_{0-1},\mathcal{H}}(x) = 1 - \max_{y \in \mathsf{H}(x)} p(x,y)$$

$$\left[\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)\right]_\epsilon = \left[\max_{y \in \mathsf{H}(x)} p(x,y) - p(x,\mathsf{h}(x))\right]_\epsilon.$$

The proof of Lemma 3 is given in Appendix F. By Lemma 3, Theorems 1 and 2 can be instantiated as Theorems 4 and 5 in the non-adversarial scenario as follows, where $\mathcal{H}$-consistency bounds are provided between the multi-class $0/1$ loss and a surrogate loss ℓ.

Theorem 4 (**Non-adversarial distribution-dependent Ψ-bound**). *Assume that there exists a convex function $\Psi : \mathbb{R}_+ \to \mathbb{R}$ with $\Psi(0) \geq 0$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$ and $\mathcal{D} \in \mathcal{P}$:*

$$\Psi\left(\left[\max_{y \in \mathsf{H}(x)} p(x,y) - p(x,\mathsf{h}(x))\right]_\epsilon\right) \leq \Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x). \tag{1}$$

Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D} \in \mathcal{P}$, we have

$$\Psi\left(\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} + \mathcal{M}_{\ell_{0-1},\mathcal{H}}\right) \leq \mathcal{R}_\ell(h) - \mathcal{R}^*_{\ell,\mathcal{H}} + \mathcal{M}_{\ell,\mathcal{H}} + \max\{\Psi(0), \Psi(\epsilon)\}. \tag{2}$$

Theorem 5 (**Non-adversarial distribution-dependent Γ-bound**). *Assume that there exists a concave function $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$ and $\mathcal{D} \in \mathcal{P}$:*

$$\left[\max_{y \in \mathsf{H}(x)} p(x,y) - p(x,\mathsf{h}(x))\right]_\epsilon \leq \Gamma(\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)). \tag{3}$$

Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D} \in \mathcal{P}$, we have

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \Gamma\left(\mathcal{R}_\ell(h) - \mathcal{R}^*_{\ell,\mathcal{H}} + \mathcal{M}_{\ell,\mathcal{H}}\right) - \mathcal{M}_{\ell_{0-1},\mathcal{H}} + \epsilon. \tag{4}$$

In the following, we will apply Theorems 4 and 5 to study the $\mathcal{H}$-consistency bounds for different families of multi-class losses parameterized by various auxiliary functions, for several general hypothesis sets. It is worth emphasizing that the form of the surrogate losses is more diverse in the multi-class setting and each case requires a careful analysis and that the techniques used in the binary case (Awasthi et al., 2022) do not apply and cannot be readily extended to our case.

Hypothesis sets. Let $B^d_p(r) = \left\{z \in \mathbb{R}^d \mid \|z\|_p \leq r\right\}$ denote the d-dimensional ℓ_p-ball with radius r, with $p \in [1, +\infty]$. Without loss of generality, in the following, we choose $\mathcal{X} = B^d_p(1)$. Let $p, q \in [1, +\infty]$ be conjugate indices, that is $\frac{1}{p} + \frac{1}{q} = 1$. In the following, we will specifically study three families: the family of all measurable functions $\mathcal{H}_{\text{all}}$, the family of linear hypotheses

$$\mathcal{H}_{\text{lin}} = \left\{(x,y) \mapsto w_y \cdot x + b_y \mid \|w_y\|_q \leq W, |b_y| \leq B\right\},$$

and that of one-hidden-layer ReLU networks defined by the following, where $(\cdot)_+ = \max(\cdot, 0)$:

$$\mathcal{H}_{\text{NN}} = \left\{(x,y) \mapsto \sum_{j=1}^n u_{y,j}(w_{y,j} \cdot x + b_{y,j})_+ \mid \|u_y\|_1 \leq \Lambda, \|w_{y,j}\|_q \leq W, |b_{y,j}| \leq B\right\}.$$

Multi-class loss families. We will study three broad families of multi-class loss functions: *max losses*, *sum losses* and *constrained losses*, each parameterized by an auxiliary function Φ on $\mathbb{R}$, assumed to be non-increasing and non-negative. In particular, we will consider the following

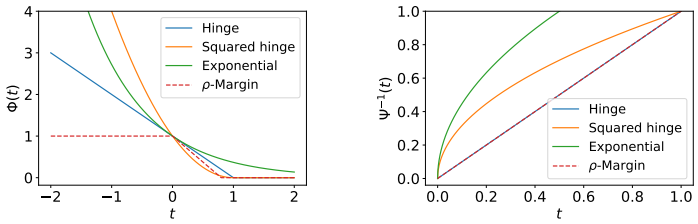

Figure 1: Left: auxiliary functions with $\rho = 0.8$. Right: $\mathcal{H}$-consistency dependence between ℓ_{0-1} and Φ^{cstnd} with $\rho = 0.8$.

common auxiliary functions: the hinge loss $\Phi_{\mathrm{hinge}}(t) = \max\{0, 1-t\}$, the squared hinge loss $\Phi_{\mathrm{sq-hinge}}(t) = \max\{0, 1-t\}^2$, the exponential loss $\Phi_{\exp}(t) = e^{-t}$, and the ρ-margin loss $\Phi_\rho(t) = \min\{\max\{0, 1-t/\rho\}, 1\}$. Note that the first three auxiliary functions are convex, while the last one is not. Figure 1 shows plots of these auxiliary functions.

We will say that a hypothesis set $\mathcal{H}$ is *symmetric* if there exists a family $\mathcal{F}$ of functions f mapping from $\mathcal{X}$ to $\mathbb{R}$ such that $\{[h(x,1), \ldots, h(x,c)] \colon h \in \mathcal{H}\} = \{[f_1(x), \ldots, f_c(x)] \colon f_1, \ldots, f_c \in \mathcal{F}\}$ and $|\{f(x) \colon f \in \mathcal{F}\}| \geq 2$ for any $x \in \mathcal{X}$. The hypothesis sets defined above ($\mathcal{H}_{\mathrm{all}}$, $\mathcal{H}_{\mathrm{lin}}$ and $\mathcal{H}_{\mathrm{NN}}$) are all symmetric. Note that for a symmetric hypothesis set $\mathcal{H}$, we have $\mathsf{H}(x) = \mathcal{Y}$.

We will say that a hypothesis set $\mathcal{H}$ is *complete* if the set of scores it generates spans $\mathbb{R}$, that is, $\{h(x,y) \colon h \in \mathcal{H}\} = \mathbb{R}$, for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$. The hypothesis sets defined above, $\mathcal{H}_{\mathrm{all}}$, $\mathcal{H}_{\mathrm{lin}}$ and $\mathcal{H}_{\mathrm{NN}}$ with $B = +\infty$ are all complete.

4.1 Max losses

In this section, we discuss guarantees for *max losses*, that is loss functions that can be defined by the application of an auxiliary function Φ to the margin $\rho_h(x,y)$, as in (Crammer and Singer, 2001):

$$\forall (x,y) \in \mathcal{X} \times \mathcal{Y}, \quad \Phi^{\max}(h,x,y) = \max_{y' \neq y} \Phi(h(x,y) - h(x,y')) = \Phi(\rho_h(x,y)). \tag{5}$$

i) Negative results. We first give negative results showing that max losses $\Phi^{\max}(h,x,y)$ with convex and non-increasing auxiliary functions Φ do not admit useful $\mathcal{H}$-consistency bounds for multi-class classification ($c > 2$). The proof is given in Appendix G.

Theorem 6 (**Negative results for convex Φ**). *Assume that $c > 2$. Suppose that Φ is convex and non-increasing, and $\mathcal{H}$ satisfies there exist $x \in \mathcal{X}$ and $h \in \mathcal{H}$ such that $|\mathsf{H}(x)| \geq 2$ and $h(x,y)$ are equal for all $y \in \mathcal{Y}$. If for a non-decreasing function $f \colon \mathbb{R}_+ \to \mathbb{R}_+$, the following $\mathcal{H}$-consistency bound holds for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$:*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq f\Big(\mathcal{R}_{\Phi^{\max}}(h) - \mathcal{R}^*_{\Phi^{\max},\mathcal{H}}\Big), \tag{6}$$

then, f is lower bounded by $\frac{1}{2}$.

The condition on the hypothesis set in Theorem 6 is very general and all symmetric hypothesis sets verify the condition, e.g. $\mathcal{H}_{\mathrm{all}}$, $\mathcal{H}_{\mathrm{lin}}$ and $\mathcal{H}_{\mathrm{NN}}$. It is also worth pointing out that when $c = 2$, that is, in binary classification, Theorem 6 does not hold. Indeed, Awasthi et al. (2022) present a series of results providing $\mathcal{H}$-consistency bounds for convex Φ in the binary case. In the proof, we make use of the assumption that $c > 2$ and thus are able to take a probability vector $p(x)$ whose dimension is at least three, which is crucial for the proof.

ii) Positive results without distributional assumptions. On the positive side, the max loss with the non-convex auxiliary function $\Phi = \Phi_\rho$ admits $\mathcal{H}$-consistency bounds.

Theorem 7 (**$\mathcal{H}$-consistency bound of $\Phi^{\max}_\rho$**). *Suppose that $\mathcal{H}$ is symmetric. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \frac{\mathcal{R}_{\Phi^{\max}_\rho}(h) - \mathcal{R}^*_{\Phi^{\max}_\rho,\mathcal{H}} + \mathcal{M}_{\Phi^{\max}_\rho,\mathcal{H}}}{\min\Big\{1, \frac{\inf_{x \in \mathcal{X}} \sup_{h \in \mathcal{H}} \rho_h(x,\mathsf{h}(x))}{\rho}\Big\}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}. \tag{7}$$

See Appendix G for the proof. Theorem 7 is very powerful since it only requires $\mathcal{H}$ to be symmetric. We can use it to derive $\mathcal{H}$-consistency bounds for $\Phi^{\max}_\rho$ with common symmetric hypothesis sets

Table 1: $\mathcal{H}$-consistency bounds for $\Phi_\rho^{\max}$ with common symmetric hypothesis sets.

Hypothesis set	$\mathcal{H}$-consistency bound of $\Phi_\rho^{\max}$ (Corollaries 18, 19 and 20)
$\mathcal{H}_{\text{all}}$	$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{all}}} \leq \mathcal{R}_{\Phi_\rho^{\max}}(h) - \mathcal{R}^*_{\Phi_\rho^{\max},\mathcal{H}_{\text{all}}}$
$\mathcal{H}_{\text{lin}}$	$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{lin}}} \leq \frac{\mathcal{R}_{\Phi_\rho^{\max}}(h) - \mathcal{R}^*_{\Phi_\rho^{\max},\mathcal{H}_{\text{lin}}} + \mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}_{\text{lin}}}}{\min\left\{1, \frac{2B}{\rho}\right\}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{lin}}}$
$\mathcal{H}_{\text{NN}}$	$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{NN}}} \leq \frac{\mathcal{R}_{\Phi_\rho^{\max}}(h) - \mathcal{R}^*_{\Phi_\rho^{\max},\mathcal{H}_{\text{NN}}} + \mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}_{\text{NN}}}}{\min\left\{1, \frac{2\Lambda B}{\rho}\right\}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{NN}}}$

such as $\mathcal{H}_{\text{all}}$, $\mathcal{H}_{\text{lin}}$ and $\mathcal{H}_{\text{NN}}$, as summarized in Table 1. The proofs with corresponding summarized Corollaries 18, 19 and 20 are included in Appendix H. In the proofs, we characterize the term $\inf_{x\in\mathcal{X}} \sup_{h\in\mathcal{H}} \rho_h(x, \mathsf{h}(x))$ for each hypothesis set.

Note that by Theorem 6, there is no useful $\mathcal{H}$-consistency bound for the max loss with $\Phi = \Phi_{\text{hinge}}$, $\Phi_{\text{sq-hinge}}$ or Φ_{exp} in these cases. However, under the realizability assumption (Definition 8), we will show that such bounds hold.

iii) Positive results with realizable distributions. We consider the $\mathcal{H}$-realizability condition (Long and Servedio, 2013; Kuznetsov et al., 2014; Cortes et al., 2016a,b; Zhang and Agarwal, 2020; Awasthi et al., 2021a) which is defined as follows.

Definition 8 ($\mathcal{H}$**-realizability**). *A distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ is $\mathcal{H}$-realizable if it labels points according to a deterministic model in $\mathcal{H}$, i.e., if $\exists h \in \mathcal{H}$ such that $\mathbb{P}_{(x,y)\sim\mathcal{D}}(\rho_h(x,y) > 0) = 1$.*

Theorem 9 (**Realizable $\mathcal{H}$-consistency bound of $\Phi^{\max}$**). *Suppose that $\mathcal{H}$ is symmetric and complete, and Φ is non-increasing and satisfies that $\lim_{t\to+\infty}\Phi(t) = 0$. Then, for any hypothesis $h\in\mathcal{H}$ and any $\mathcal{H}$-realizable distribution $\mathcal{D}$, we have*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi^{\max}}(h) - \mathcal{R}^*_{\Phi^{\max},\mathcal{H}} + \mathcal{M}_{\Phi^{\max},\mathcal{H}}. \tag{8}$$

See Appendix G for the proof. Long and Servedio (2013, Theorem 9) show that $\Phi^{\max}_{\text{hinge}}$ is realizable $\mathcal{H}$-consistent for any symmetric hypothesis set $\mathcal{H}$ that is closed under scaling. Since for any $\mathcal{H}$-realizable distribution, the assumption that $\mathcal{H}$ is closed under scaling implies that $\mathcal{H}$ is complete and $\mathcal{M}_{\Phi^{\max},\mathcal{H}} = 0$, Theorem 9 also yields a quantitative relationship in that case that is stronger than the asymptotic consistency property of that previous work.

4.2 Sum losses

In this section, we discuss guarantees for *sum losses*, that is loss functions defined via a sum, as in (Weston and Watkins, 1998):

$$\Phi^{\text{sum}}(h, x, y) = \sum_{y'\neq y} \Phi(h(x,y) - h(x,y')). \tag{9}$$

i) Negative results. We first give a negative result showing that when using as auxiliary function the hinge-loss, the sum loss cannot benefit from any useful $\mathcal{H}$-consistency guarantee. The proof is deferred to Appendix J.

Theorem 10 (**Negative results for hinge loss**). *Assume that $c > 2$. Suppose that $\mathcal{H}$ is symmetric and complete. If for a non-decreasing function $f\colon \mathbb{R}_+ \to \mathbb{R}_+$, the following $\mathcal{H}$-consistency bound holds for any hypothesis $h\in\mathcal{H}$ and any distribution $\mathcal{D}$:*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq f\left(\mathcal{R}_{\Phi^{\text{sum}}_{\text{hinge}}}(h) - \mathcal{R}^*_{\Phi^{\text{sum}}_{\text{hinge}},\mathcal{H}}\right), \tag{10}$$

then, f is lower bounded by $\frac{1}{6}$.

ii) Positive results. We then complement this negative result with positive results when using the exponential loss, the squared hinge-loss, and the ρ-margin loss, as summarized in Table 2. The proofs with corresponding summarized Theorems 22, 23 and 24 are included in Appendix J for completeness. For Φ^{sum}_ρ, the symmetry and completeness assumption can be relaxed to symmetry and the condition that for any $x\in\mathcal{X}$, there exists a hypothesis $h\in\mathcal{H}$ such that $|h(x,i) - h(x,j)| \geq \rho$ for any $i\neq j\in\mathcal{Y}$, as shown in Theorem 24. In the proof, we introduce an auxiliary Lemma 21 in Appendix I, which would be helpful for lower bounding the conditional regret of Φ^{sum}_ρ with that of the multi-class $0/1$ loss.

Table 2: $\mathcal{H}$-consistency bounds for sum losses with symmetric and complete hypothesis sets.

Sum loss	$\mathcal{H}$-consistency bound (Theorems 22, 23 and 24)
$\Phi^{\text{sum}}_{\text{sq-hinge}}$	$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \left(\mathcal{R}_{\Phi^{\text{sum}}_{\text{sq-hinge}}}(h) - \mathcal{R}^*_{\Phi^{\text{sum}}_{\text{sq-hinge}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{sum}}_{\text{sq-hinge}},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}$
$\Phi^{\text{sum}}_{\text{exp}}$	$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \sqrt{2}\left(\mathcal{R}_{\Phi^{\text{sum}}_{\text{exp}}}(h) - \mathcal{R}^*_{\Phi^{\text{sum}}_{\text{exp}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{sum}}_{\text{exp}},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}$
Φ^{sum}_{ρ}	$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi^{\text{sum}}_{\rho}}(h) - \mathcal{R}^*_{\Phi^{\text{sum}}_{\rho},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{sum}}_{\rho},\mathcal{H}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}$

Table 3: $\mathcal{H}$-consistency bounds for constrained losses with symmetric and complete hypothesis sets.

Constrained loss	$\mathcal{H}$-consistency bound (Theorems 25, 26, 27 and 28)
$\Phi^{\text{cstnd}}_{\text{hinge}}$	$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi^{\text{cstnd}}_{\text{hinge}}}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_{\text{hinge}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_{\text{hinge}},\mathcal{H}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}$
$\Phi^{\text{cstnd}}_{\text{sq-hinge}}$	$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \left(\mathcal{R}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}}}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_{\text{sq-hinge}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}$
$\Phi^{\text{cstnd}}_{\text{exp}}$	$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \sqrt{2}\left(\mathcal{R}_{\Phi^{\text{cstnd}}_{\text{exp}}}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_{\text{exp}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_{\text{exp}},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}$
$\Phi^{\text{cstnd}}_{\rho}$	$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi^{\text{cstnd}}_{\rho}}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_{\rho},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_{\rho},\mathcal{H}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}$

4.3 Constrained losses

In this section, we discuss guarantees for *constrained loss*, that is loss functions defined via a constraint, as in (Lee et al., 2004):

$$\Phi^{\text{cstnd}}(h, x, y) = \sum_{y' \neq y} \Phi(-h(x, y')) \tag{11}$$

with the constraint that $\sum_{y \in \mathcal{Y}} h(x, y) = 0$. We present a series of positive results by proving multi-class $\mathcal{H}$-consistency bounds when using as an auxiliary function the hinge-loss, the squared hinge-loss, the exponential loss, and the ρ-margin loss, as summarized in Table 3. As with the binary case (Awasthi et al., 2022), the bound admits a linear dependency for $\Phi^{\text{cstnd}}_{\text{hinge}}$ and $\Phi^{\text{cstnd}}_{\rho}$, in contrast with a square-root dependency for $\Phi^{\text{cstnd}}_{\text{sq-hinge}}$ and $\Phi^{\text{cstnd}}_{\text{exp}}$, as illustrated in Figure 1. The proofs with corresponding summarized Theorems 25, 26, 27 and 28 are included in Appendix K for completeness. For $\Phi^{\text{cstnd}}_{\rho}$, the symmetric and complete assumption can be relaxed to be symmetric and satisfy that for any $x \in \mathcal{X}$, there exists a hypothesis $h \in \mathcal{H}$ such that $h(x, y) \leq -\rho$ for any $y \neq y_{\max}$, as shown in Theorem 28.

The main idea of the proofs in this section is to leverage the constraint condition of Lee et al. (2004) that the scores sum to zero, and appropriately choose a hypothesis $\overline{h}$ that differs from h only by its scores for $\mathsf{h}(x)$ and $y_{\max}$. We can then upper bound the minimal conditional risk by the conditional risk of $\overline{h}$, without having to derive the closed form expression of the minimal conditional risk.

As shown by Steinwart (2007, Theorem 3.2), for the family of all measurable functions, the min-imizability gaps vanish: $\mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{all}}} = \mathcal{M}_{\Phi^{\text{sum}},\mathcal{H}_{\text{all}}} = \mathcal{M}_{\Phi^{\text{cstnd}},\mathcal{H}_{\text{all}}} = 0$, for $\Phi = \Phi_{\text{hinge}}, \Phi_{\text{sq-hinge}}, \Phi_{\text{exp}}$ and Φ_{ρ}. Therefore, when $\mathcal{H} = \mathcal{H}_{\text{all}}$, our quantitative bounds in Table 2 and Table 3 imply the asymptotic consistency results of those multi-class losses in (Tewari and Bartlett, 2007), which shows that our results are stronger and more significant. We also provide bounds for multi-class losses using a non-convex auxiliary function, which are not studied in the previous work.

5 Adversarial $\mathcal{H}$-consistency bounds

In this section, we analyze multi-class $\mathcal{H}$-consistency bounds in the adversarial scenario ($\ell_2 = \ell_\gamma$).

For any $x \in \mathcal{X}$, we denote by $\mathcal{H}_\gamma(x)$ the set of hypotheses h with a positive margin on the ball of radius γ around x, $\mathcal{H}_\gamma(x) = \left\{h \in \mathcal{H} : \inf_{x' : \|x - x'\|_p \leq \gamma} \rho_h(x', \mathsf{h}(x)) > 0\right\}$, and by $\mathsf{H}_\gamma(x)$ the set of labels generated by these hypotheses, $\mathsf{H}_\gamma(x) = \{\mathsf{h}(x) : h \in \mathcal{H}_\gamma(x)\}$. When $\mathcal{H}$ is symmetric, we have $\mathsf{H}_\gamma(x) = \mathcal{Y}$ iff $\mathcal{H}_\gamma(x) \neq \varnothing$. The following lemma characterizes the conditional ϵ-regret for

adversarial $0/1$ loss, which will be helpful for applying Theorem 1 and Theorem 2 to the adversarial scenario.

Lemma 11. *For any $x \in \mathcal{X}$, the minimal conditional ℓ_γ-risk and the conditional ϵ-regret for ℓ_γ can be expressed as follows:*

$$\mathcal{C}^*_{\ell_\gamma,\mathcal{H}}(x) = 1 - \max_{y \in \mathsf{H}_\gamma(x)} p(x,y) \mathbb{1}_{\mathcal{H}_\gamma(x) \neq \varnothing}$$

$$\left[\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x)\right]_\epsilon = \begin{cases} \left[\max_{y \in \mathsf{H}_\gamma(x)} p(x,y) - p(x,\mathsf{h}(x)) \mathbb{1}_{h \in \mathcal{H}_\gamma(x)}\right]_\epsilon & \text{if } \mathcal{H}_\gamma(x) \neq \varnothing \\ 0 & \text{otherwise.} \end{cases}$$

The proof of Lemma 11 is presented in Appendix F. By Lemma 11, Theorems 1 and 2 can be instantiated as Theorems 12 and 13 in the adversarial scenario as follows, where $\mathcal{H}$-consistency bounds are provided between the adversarial multi-class $0/1$ loss and a surrogate loss ℓ.

Theorem 12 (**Adversarial distribution-dependent Ψ-bound**). *Assume that there exists a convex function $\Psi\colon \mathbb{R}_+ \to \mathbb{R}$ with $\Psi(0) = 0$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \{x \in \mathcal{X} : \mathcal{H}_\gamma(x) \neq \varnothing\}$ and $\mathcal{D} \in \mathcal{P}$:*

$$\Psi\left(\left[\max_{y \in \mathsf{H}_\gamma(x)} p(x,y) - p(x,\mathsf{h}(x)) \mathbb{1}_{h \in \mathcal{H}_\gamma(x)}\right]_\epsilon\right) \leq \Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x). \tag{12}$$

Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D} \in \mathcal{P}$, we have

$$\Psi\left(\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} + \mathcal{M}_{\ell_\gamma,\mathcal{H}}\right) \leq \mathcal{R}_\ell(h) - \mathcal{R}^*_{\ell,\mathcal{H}} + \mathcal{M}_{\ell,\mathcal{H}} + \max\{0, \Psi(\epsilon)\}. \tag{13}$$

Theorem 13 (**Adversarial distribution-dependent Γ-bound**). *Assume that there exists a non-negative concave function $\Gamma\colon \mathbb{R}_+ \to \mathbb{R}$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \{x \in \mathcal{X} : \mathcal{H}_\gamma(x) \neq \varnothing\}$ and $\mathcal{D} \in \mathcal{P}$:*

$$\left[\max_{y \in \mathsf{H}_\gamma(x)} p(x,y) - p(x,\mathsf{h}(x)) \mathbb{1}_{h \in \mathcal{H}_\gamma(x)}\right]_\epsilon \leq \Gamma(\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)). \tag{14}$$

Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D} \in \mathcal{P}$, we have

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq \Gamma\left(\mathcal{R}_\ell(h) - \mathcal{R}^*_{\ell,\mathcal{H}} + \mathcal{M}_{\ell,\mathcal{H}}\right) - \mathcal{M}_{\ell_\gamma,\mathcal{H}} + \epsilon. \tag{15}$$

Next, we will apply Theorem 12 and Theorem 13 to study various hypothesis sets and adversarial surrogate loss functions in Sections 5.1 for negative results and Section 5.2, 5.3, and 5.4 for positive results. A careful analysis is presented in each case (see Appendix L, M, N and O).

5.1 Negative results for adversarial robustness

The following result rules out the $\mathcal{H}$-consistency guarantee of multi-class losses with a convex auxiliary function, which are commonly used in practice. The proof is given in Appendix L.

Theorem 14 (**Negative results for convex functions**). *Fix $c = 2$. Suppose that Φ is convex and non-increasing, and $\mathcal{H}$ contains 0 and satisfies the condition that there exists $x \in \mathcal{X}$ such that $\mathcal{H}_\gamma(x) \neq \varnothing$. If for a non-decreasing function $f\colon \mathbb{R}_+ \to \mathbb{R}_+$, the following $\mathcal{H}$-consistency bound holds for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$:*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq f\left(\mathcal{R}_{\widetilde{\ell}}(h) - \mathcal{R}^*_{\widetilde{\ell},\mathcal{H}}\right), \tag{16}$$

then, f is lower bounded by $\frac{1}{2}$, for $\widetilde{\ell} = \widetilde{\Phi}^{\max}$, $\widetilde{\Phi}^{\text{sum}}$ and $\widetilde{\Phi}^{\text{cstnd}}$.

Instead, we show in Sections 5.2, 5.3, and 5.4 that the max, sum and constrained losses using as auxiliary function the non-convex ρ-margin loss admit favorable $\mathcal{H}$-consistency bounds in the multi-class setting, thereby significantly generalizing the binary counterpart in (Awasthi et al., 2022).

5.2 Adversarial max losses

We first consider the adversarial max loss $\widetilde{\Phi}^{\max}$ defined as the supremum based counterpart of (5):

$$\widetilde{\Phi}^{\max}(h,x,y) = \sup_{x' : \|x - x'\|_p \leq \gamma} \Phi(\rho_h(x',y)). \tag{17}$$

For the adversarial max loss with $\Phi = \Phi_\rho$, we can obtain $\mathcal{H}$-consistency bounds as follows.

Theorem 15 ($\mathcal{H}$**-consistency bound of** $\widetilde{\Phi}_\rho^{\max}$)**.** *Suppose that* $\mathcal{H}$ *is symmetric. Then, for any hypothesis* $h \in \mathcal{H}$ *and any distribution* $\mathcal{D}$*, we have*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq \frac{\mathcal{R}_{\widetilde{\Phi}_\rho^{\max}}(h) - \mathcal{R}^*_{\widetilde{\Phi}_\rho^{\max},\mathcal{H}} + \mathcal{M}_{\widetilde{\Phi}_\rho^{\max},\mathcal{H}}}{\min\left\{1, \frac{\inf_{x\in\{x\in\mathcal{X}:\mathcal{H}_\gamma(x)\neq\varnothing\}} \sup_{h\in\mathcal{H}_\gamma(x)} \inf_{x':\|x-x'\|_p\leq\gamma} \rho_h(x',h(x))}{\rho}\right\}} - \mathcal{M}_{\ell_\gamma,\mathcal{H}}. \tag{18}$$

5.3 Adversarial sum losses

Next, we consider the adversarial sum loss $\widetilde{\Phi}^{\mathrm{sum}}$ defined as the supremum based counterpart of (9):

$$\widetilde{\Phi}^{\mathrm{sum}}(h,x,y) = \sup_{x':\|x-x'\|_p\leq\gamma} \sum_{y'\neq y} \Phi(h(x',y) - h(x',y')). \tag{19}$$

Using the auxiliary Lemma 21 in Appendix I, we can obtain the $\mathcal{H}$-consistency bound of $\widetilde{\Phi}_\rho^{\mathrm{sum}}$.

Theorem 16 ($\mathcal{H}$**-consistency bound of** $\widetilde{\Phi}_\rho^{\mathrm{sum}}$)**.** *Assume that* $\mathcal{H}$ *is symmetric and that for any* $x \in \mathcal{X}$*, there exists a hypothesis* $h \in \mathcal{H}$ *inducing the same ordering of the labels for any* $x' \in \left\{x' : \|x-x'\|_p \leq \gamma\right\}$ *and such that* $\inf_{x':\|x-x'\|_p\leq\gamma} |h(x',i) - h(x',j)| \geq \rho$ *for any* $i \neq j \in \mathcal{Y}$*. Then, for any hypothesis* $h \in \mathcal{H}$ *and any distribution* $\mathcal{D}$*, the following inequality holds:*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq \mathcal{R}_{\widetilde{\Phi}_\rho^{\mathrm{sum}}}(h) - \mathcal{R}^*_{\widetilde{\Phi}_\rho^{\mathrm{sum}},\mathcal{H}} + \mathcal{M}_{\widetilde{\Phi}_\rho^{\mathrm{sum}},\mathcal{H}} - \mathcal{M}_{\ell_\gamma,\mathcal{H}}. \tag{20}$$

5.4 Adversarial constrained loss

Similarly, we define the adversarial constrained loss $\widetilde{\Phi}^{\mathrm{cstnd}}$ as supremum based counterpart of (11):

$$\widetilde{\Phi}^{\mathrm{cstnd}}(h,x,y) = \sup_{x':\|x-x'\|_p\leq\gamma} \sum_{y'\neq y} \Phi(-h(x',y')) \tag{21}$$

with the constraint that $\sum_{y\in\mathcal{Y}} h(x,y) = 0$. For the adversarial constrained loss with $\Phi = \Phi_\rho$, we can obtain the $\mathcal{H}$-consistency bound of $\widetilde{\Phi}_\rho^{\mathrm{cstnd}}$ as follows.

Theorem 17 ($\mathcal{H}$**-consistency bound of** $\widetilde{\Phi}_\rho^{\mathrm{cstnd}}$)**.** *Suppose that* $\mathcal{H}$ *is symmetric and satisfies that for any* $x \in \mathcal{X}$*, there exists a hypothesis* $h \in \mathcal{H}$ *with the constraint* $\sum_{y\in\mathcal{Y}} h(x,y) = 0$ *such that* $\sup_{x':\|x-x'\|_p\leq\gamma} h(x',y) \leq -\rho$ *for any* $y \neq y_{\max}$*. Then, for any hypothesis* $h \in \mathcal{H}$ *and any distribution,*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq \mathcal{R}_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}}}(h) - \mathcal{R}^*_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}},\mathcal{H}} + \mathcal{M}_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}},\mathcal{H}} - \mathcal{M}_{\ell_\gamma,\mathcal{H}}. \tag{22}$$

The proofs of Theorems 15, 16 and 17 are included in Appendix M, N and O respectively. These results are significant since they apply to general hypothesis sets. In particular, symmetric hypothesis sets $\mathcal{H}_{\mathrm{all}}$, $\mathcal{H}_{\mathrm{lin}}$ and $\mathcal{H}_{\mathrm{NN}}$ with $B = +\infty$ all verify the conditions of those theorems. When $B < +\infty$, the conditions in Theorems 16 and 17 can still be verified with a suitable choice of ρ, where we can consider the hypotheses such that $w_y = 0$ in $\mathcal{H}_{\mathrm{lin}}$ and $\mathcal{H}_{\mathrm{NN}}$, while Theorem 15 holds for any $\rho > 0$.

6 Conclusion

We presented a comprehensive study of $\mathcal{H}$-consistency bounds for multi-class classification, including the analysis of the three most commonly used families of multi-class surrogate losses (max losses, sum losses and constrained losses) and including the study of surrogate losses for the adversarial robustness. Our theoretical analysis helps determine which surrogate losses admit a favorable guarantee for a given hypothesis set $\mathcal{H}$. Our bounds can help guide the design of multi-class classification algorithms for both the adversarial and non-adversarial settings. They also help compare different surrogate losses for the same setting and the same hypothesis set. Of course, in addition to the functional form of the $\mathcal{H}$-consistency bound, the approximation property of a surrogate loss function combined with the hypothesis set plays an important role.

References

A. Agarwal and S. Agarwal. On consistent surrogate risk minimization and property elicitation. In *Conference on Learning Theory*, pages 4–22, 2015.

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. $\mathcal{H}$-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pages 1117–1174, 2022.

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

H. Bao, C. Scott, and M. Sugiyama. Calibrated surrogate losses for adversarially robust classification. In *Conference on Learning Theory*, pages 408–451, 2020.

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.

M. Blondel. Structured prediction with projection oracles. In *Advances in neural information processing systems*, 2019.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

D.-R. Chen and T. Sun. Consistency of multiclass empirical risk minimization methods based on convex loss. *Journal of Machine Learning Research*, 7:2435–2447, 2006.

D.-R. Chen and D.-H. Xiang. The consistency of multicategory support vector machines. *Advances in Computational Mathematics*, 24(1):155–169, 2006.

C. Ciliberto, L. Rosasco, and A. Rudi. A consistent regularization approach for structured prediction. In *Advances in neural information processing systems*, 2016.

C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *Algorithmic Learning Theory*, pages 67–82, 2016a.

C. Cortes, G. DeSalvo, and M. Mohri. Boosting with abstention. In *Advances in Neural Information Processing Systems*, pages 1660–1668, 2016b.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

K. Dembczynski, W. Kotlowski, and E. Hüllermeier. Consistent multilabel ranking through univariate losses. *arXiv preprint arXiv:1206.6401*, 2012.

U. Dogan, T. Glasmachers, and C. Igel. A unified view on multi-class support vector classification. *Journal of Machine Learning Research*, 17:1–32, 2016.

W. Gao and Z.-H. Zhou. On the consistency of multi-label learning. In *Conference on learning theory*, pages 341–358, 2011.

W. Gao and Z.-H. Zhou. On the consistency of auc pairwise optimization. In *International Joint Conference on Artificial Intelligence*, 2015.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

V. Kuznetsov, M. Mohri, and U. Syed. Multi-class deep boosting. In *Advances in Neural Information Processing Systems*, pages 2501–2509, 2014.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

Y. Liu. Fisher consistency of multicategory support vector machines. In *Artificial intelligence and statistics*, pages 291–298, 2007.

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

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

H. Narasimhan, H. Ramaswamy, A. Saha, and S. Agarwal. Consistent multiclass algorithms for complex performance measures. In *International Conference on Machine Learning*, pages 2398–2407, 2015.

A. Osokin, F. Bach, and S. Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, 2017.

F. Pedregosa, F. Bach, and A. Gramfort. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18:1–35, 2017.

B. Á. Pires and C. Szepesvári. Multiclass classification calibration functions. *arXiv preprint arXiv:1609.06385*, 2016.

B. A. Pires, C. Szepesvari, and M. Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In *International Conference on Machine Learning*, pages 1391–1399, 2013.

H. G. Ramaswamy and S. Agarwal. Classification calibration dimension for general multiclass losses. In *Advances in Neural Information Processing Systems*, 2012.

H. G. Ramaswamy and S. Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17(1):397–441, 2016.

H. G. Ramaswamy, S. Agarwal, and A. Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems*, 2013.

H. G. Ramaswamy, A. Tewari, and S. Agarwal. Consistent algorithms for multiclass classification with a reject option. *arXiv preprint arXiv:1505.04137*, 2015.

P. Ravikumar, A. Tewari, and E. Yang. On NDCG consistency of listwise ranking methods. In *International Conference on Artificial Intelligence and Statistics*, pages 618–626, 2011.

R. E. Schapire and Y. Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.

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

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

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

K. Uematsu and Y. Lee. On theoretically optimal ranking functions in bipartite ranking. *Journal of the American Statistical Association*, 112(519):1311–1322, 2017.

J. Weston and C. Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

R. C. Williamson, E. Vernet, and M. D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17:1–52, 2016.

M. Zhang and S. Agarwal. Bayes consistency vs. H-consistency: The interplay between surrogate loss functions and the scoring function class. In *Advances in Neural Information Processing Systems*, pages 16927–16936, 2020.

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.

Contents of Appendix

A Related work

The notions of Bayes-consistency (also known as consistency) and calibration have been well studied not only with respect to the binary zero-one loss (Zhang, 2004a; Bartlett et al., 2006; Steinwart, 2007; Mohri et al., 2018), but also with respect to the multi-class zero-one loss (Zhang, 2004b; Tewari and Bartlett, 2007), the general multi-class losses (Ramaswamy and Agarwal, 2012; Narasimhan et al., 2015; Ramaswamy and Agarwal, 2016), the multi-class SVMs (Chen and Sun, 2006; Chen and Xiang, 2006; Liu, 2007; Dogan et al., 2016), the multi-label losses (Gao and Zhou, 2011; Dembczynski et al., 2012), the losses with a reject option (Ramaswamy et al., 2015), the ranking losses (Ravikumar et al., 2011; Ramaswamy et al., 2013; Gao and Zhou, 2015; Uematsu and Lee, 2017), the cost sensitive losses (Pires et al., 2013; Pires and Szepesvári, 2016), the structured losses (Ciliberto et al., 2016; Osokin et al., 2017; Blondel, 2019), the proper losses (Agarwal and Agarwal, 2015; Williamson et al., 2016) and the losses of ordinal regression (Pedregosa et al., 2017).

Bayes-consistency only holds for the full family of measurable functions, which of course is distinct from the more restricted hypothesis set used by a learning algorithm. Therefore, a hypothesis set-dependent notion of $\mathcal{H}$-consistency has been proposed by Long and Servedio (2013) in the realizable setting, used by Zhang and Agarwal (2020) for linear models, and generalized by Kuznetsov et al. (2014) to the structured prediction case. Long and Servedio (2013) showed that there exists a case where a Bayes-consistent loss is not $\mathcal{H}$-consistent while inconsistent losses can be $\mathcal{H}$-consistent. Zhang and Agarwal (2020) further investigated the phenomenon in (Long and Servedio, 2013) and showed that the situation of losses that are not $\mathcal{H}$-consistent with linear models can be remedied by carefully choosing a larger piecewise linear hypothesis set. Kuznetsov et al. (2014) proved positive results for the $\mathcal{H}$-consistency of several multi-class ensemble algorithms, as an extension of $\mathcal{H}$-consistency results in (Long and Servedio, 2013).

Recently, the notions of $\mathcal{H}$-calibration and $\mathcal{H}$-consistency have been used by Bao et al. (2020); Awasthi et al. (2021a) in the study of adversarial binary classification losses, as defined in (Goodfellow et al., 2014; Madry et al., 2017; Tsipras et al., 2018; Carlini and Wagner, 2017; Awasthi et al., 2023). The calibration and consistency of adversarial losses present new challenges and require more careful analysis. The work of Bao et al. (2020) showed that for the linear hypothesis set, convex margin based losses are not calibrated with respect to the adversarial $0/1$ loss. Instead, they proposed a class of non-convex losses that could be calibrated under some necessary and sufficient conditions. The work of Awasthi et al. (2021a) generalized the results in (Bao et al., 2020) to the nonlinear hypothesis sets. They also pointed out that $\mathcal{H}$-calibration and $\mathcal{H}$-consistency are not equivalent in the adversarial scenario by showing that no continuous surrogates can be $\mathcal{H}$-consistent with linear models. They further provided sufficient conditions guaranteeing $\mathcal{H}$-consistency for $\mathcal{H}$-calibrated surrogates.

Most recently, Awasthi et al. (2022) presented a series of results providing $\mathcal{H}$-consistency bounds in binary classification, for both the adversarial and non-adversarial settings. These guarantees are significantly stronger than the $\mathcal{H}$-calibration or $\mathcal{H}$-consistency properties studied by Awasthi et al. (2021a,b). They are also more informative than similar excess error bounds derived in the literature, which correspond to the special case where $\mathcal{H}$ is the family of all measurable functions (Zhang, 2004a; Bartlett et al., 2006; Mohri et al., 2018). Our work significantly generalizes the results in (Awasthi et al., 2022) to the multi-class setting, in both the adversarial and non-adversarial scenarios, where the study of calibration and conditional risk is more complex, the form of the surrogate losses is more diverse, and in general the analysis is more involved and entirely novel proof techniques are required. As a by-product, our work contributes more significant results of consistency for the insufficiently understood setting of adversarial robustness.

B Discussion on multi-class $0/1$ loss

The multi-class $0/1$ loss can be defined in multiple ways, e.g. $\mathbb{1}_{\rho_h(x,y)\leq 0}$, $\mathbb{1}_{\rho_h(x,y)<0}$ and $\mathbb{1}_{\mathsf{h}(x)\neq y}$ where $\mathsf{h}(x) = \operatorname{argmax}_{y\in\mathcal{Y}} h(x,y)$ with an arbitrary but fixed deterministic strategy used for breaking ties. The counterparts of these three formulas in binary classification are $\mathbb{1}_{yh(x)\leq 0}$, $\mathbb{1}_{yh(x)<0}$ and $\mathbb{1}_{\operatorname{sgn}(h(x))\neq y}$ where $\operatorname{sgn}(0)$ is defined as $+1$ or -1. To be consistent with the literature on Bayes-consistency (Bartlett et al., 2006; Tewari and Bartlett, 2007), in this paper we adopt the last formula $\mathbb{1}_{\mathsf{h}(x)\neq y}$ of multi-class $0/1$ loss. Moreover, to be consistent with the binary case (Awasthi et al., 2022), we assume that in case of a tie, $\mathsf{h}(x)$ is defined as the label with the highest index under the natural ordering of labels. This assumption corresponds to the binary case where we always predict $+1$ in case of a tie, that is, the case where the binary $0/1$ loss is defined by $\mathbb{1}_{\operatorname{sgn}(h(x))\neq y}$ with $\operatorname{sgn}(0) = +1$, as in (Awasthi et al., 2022). Nevertheless, other deterministic strategies would lead to similar results.

C Discussion on finite sample bounds

Here, we discuss several ways to derive the finite sample bounds on the estimation error for the target $0/1$ loss. One can directly derive estimation error bounds for the $0/1$ loss, typically for Empirical Risk Minimization (ERM), e.g. $\mathcal{R}_{\ell_{0-1}}\big(h_S^{\mathrm{ERM}}\big) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}}$ with $h_S^{\mathrm{ERM}} = \operatorname{argmin}_{h\in\mathcal{H}} \widehat{\mathcal{R}}_S(h)$ can be upper-bounded using the standard generalization bounds, as shown in (Mohri et al., 2018). But, those bounds would not say anything about the use of a surrogate loss.

An alternative is to use the excess error bound for the target $0/1$ loss and split the excess error of the surrogate loss into an estimation term and an approximation term, i.e. for some function $f\colon \mathbb{R}_+ \to \mathbb{R}_+$, the following inequality holds:

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\mathrm{all}}} \leq f\Big(\mathcal{R}_{\ell_{\mathrm{sur}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{sur}},\mathcal{H}} + \mathcal{R}^*_{\ell_{\mathrm{sur}},\mathcal{H}} - \mathcal{R}^*_{\ell_{\mathrm{sur}},\mathcal{H}_{\mathrm{all}}}\Big).$$

Then, an estimation error bound for the surrogate loss can be used to upper bound $\mathcal{R}_{\ell_{\mathrm{sur}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{sur}},\mathcal{H}}$, as shown in (Bartlett et al., 2006). But, those bounds would not be an estimation error guarantee for the target loss ℓ_{0-1}.

Finally, using the $\mathcal{H}$-consistency bound proposed by Awasthi et al. (2022), that is, for some non-decreasing function $f\colon \mathbb{R}_+ \to \mathbb{R}_+$,

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq f\Big(\mathcal{R}_{\ell_{\mathrm{sur}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{sur}},\mathcal{H}}\Big),$$

we can directly derive the estimation error bound for the target $0/1$ loss by upper bounding $\mathcal{R}_{\ell_{\mathrm{sur}}}(h) - \mathcal{R}^*_{\ell_{\mathrm{sur}},\mathcal{H}}$ with the estimation error bound for the surrogate loss. In conclusion, the $\mathcal{H}$-consistency bound is a useful tool to derive non-trivial finite sample bounds on the estimation error for the target $0/1$ loss.

D Future work

While we presented a comprehensive study of $\mathcal{H}$-consistency bounds for surrogate losses in multi-class classification, which could help compare different surrogate losses for the same setting and the same hypothesis set, the optimization property of a surrogate loss function combined with the hypothesis set also plays an important role. Nevertheless, we believe our results in the paper can help guide the design of multi-class classification algorithms for both the adversarial and non-adversarial settings.

E General $\mathcal{H}$-consistency bounds

Theorem 1 (**Distribution-dependent Ψ-bound**). *Assume that there exists a convex function $\Psi\colon \mathbb{R}_+ \to \mathbb{R}$ with $\Psi(0) \geq 0$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$ and $\mathcal{D} \in \mathcal{P}$: $\Psi\Big(\big[\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\big]_\epsilon\Big) \leq \Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h,x)$. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D} \in \mathcal{P}$,*

$$\Psi\Big(\mathcal{R}_{\ell_2}(h) - \mathcal{R}^*_{\ell_2,\mathcal{H}} + \mathcal{M}_{\ell_2,\mathcal{H}}\Big) \leq \mathcal{R}_{\ell_1}(h) - \mathcal{R}^*_{\ell_1,\mathcal{H}} + \mathcal{M}_{\ell_1,\mathcal{H}} + \max\{\Psi(0), \Psi(\epsilon)\}.$$

Proof. For any $h \in \mathcal{H}$ and $\mathcal{D} \in \mathcal{P}$, since $\Psi\Big(\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\mathbb{1}_{\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)>\epsilon}\Big) \leq \Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h,x), \forall x \in \mathcal{X}$, we can write

$$\begin{aligned}
&\Psi\Big(\mathcal{R}_{\ell_2}(h) - \mathcal{R}^*_{\ell_2,\mathcal{H}} + \mathcal{M}_{\ell_2,\mathcal{H}}\Big)\\
&= \Psi\Big(\mathbb{E}_X\Big[\mathcal{C}_{\ell_2}(h,x) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(x)\Big]\Big)\\
&= \Psi\big(\mathbb{E}_X\big[\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\big]\big)\\
&\leq \mathbb{E}_X\big[\Psi\big(\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\big)\big] && \text{(Jensen's ineq.)}\\
&= \mathbb{E}_X\Big[\Psi\Big(\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\mathbb{1}_{\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)>\epsilon} + \Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\mathbb{1}_{\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\leq\epsilon}\Big)\Big]\\
&\leq \mathbb{E}_X\Big[\Psi\Big(\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\mathbb{1}_{\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)>\epsilon}\Big) + \Psi\Big(\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\mathbb{1}_{\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\leq\epsilon}\Big)\Big] && (\Psi(0) \geq 0)\\
&\leq \mathbb{E}_X\big[\Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h,x)\big] + \sup_{t\in[0,\epsilon]} \Psi(t) && \text{(assumption)}\\
&= \mathcal{R}_{\ell_1}(h) - \mathcal{R}^*_{\ell_1,\mathcal{H}} + \mathcal{M}_{\ell_1,\mathcal{H}} + \max\{\Psi(0), \Psi(\epsilon)\}, && \text{(convexity of } \Psi)
\end{aligned}$$

which proves the theorem. □

Theorem 2 (**Distribution-dependent Γ-bound**). *Assume that there exists a concave function $\Gamma\colon \mathbb{R}_+ \to \mathbb{R}$ and $\epsilon \geq 0$ such that the following holds for all $h \in \mathcal{H}$, $x \in \mathcal{X}$ and $\mathcal{D} \in \mathcal{P}$: $\big[\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\big]_\epsilon \leq \Gamma\big(\Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h,x)\big)$. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D} \in \mathcal{P}$,*

$$\mathcal{R}_{\ell_2}(h) - \mathcal{R}^*_{\ell_2,\mathcal{H}} \leq \Gamma\Big(\mathcal{R}_{\ell_1}(h) - \mathcal{R}^*_{\ell_1,\mathcal{H}} + \mathcal{M}_{\ell_1,\mathcal{H}}\Big) - \mathcal{M}_{\ell_2,\mathcal{H}} + \epsilon.$$

Proof. For any $h \in \mathcal{H}$ and $\mathcal{D} \in \mathcal{P}$, since $\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\mathbb{1}_{\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)>\epsilon} \leq \Gamma\big(\Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h,x)\big), \forall x \in \mathcal{X}$, we can write

$$\begin{aligned}
&\mathcal{R}_{\ell_2}(h) - \mathcal{R}^*_{\ell_2,\mathcal{H}} + \mathcal{M}_{\ell_2,\mathcal{H}}\\
&= \mathbb{E}_X\Big[\mathcal{C}_{\ell_2}(h,x) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(x)\Big]\\
&= \mathbb{E}_X\big[\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\big]\\
&= \mathbb{E}_X\Big[\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\mathbb{1}_{\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)>\epsilon} + \Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\mathbb{1}_{\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)\leq\epsilon}\Big]\\
&\leq \mathbb{E}_X\big[\Gamma\big(\Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h,x)\big)\big] + \epsilon && \text{(assumption)}\\
&\leq \Gamma\big(\mathbb{E}_X\big[\Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h,x)\big]\big) + \epsilon && \text{(concavity of } \Gamma)\\
&= \Gamma\Big(\mathcal{R}_{\ell_1}(h) - \mathcal{R}^*_{\ell_1,\mathcal{H}} + \mathcal{M}_{\ell_1,\mathcal{H}}\Big) + \epsilon,
\end{aligned}$$

which proves the theorem. □

F Non-adversarial and adversarial conditional regrets

Lemma 3. *For any $x \in \mathcal{X}$, the minimal conditional ℓ_{0-1}-risk and the conditional ϵ-regret for ℓ_{0-1} can be expressed as follows:*

$$\mathcal{C}^*_{\ell_{0-1},\mathcal{H}}(x) = 1 - \max_{y\in\mathsf{H}(x)} p(x,y)$$

$$\big[\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)\big]_\epsilon = \left[\max_{y\in\mathsf{H}(x)} p(x,y) - p(x,\mathsf{h}(x))\right]_\epsilon.$$

Proof. By the definition, the conditional ℓ_{0-1}-risk can be expressed as follows:

$$\mathcal{C}_{\ell_{0-1}}(h,x) = \sum_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathsf{h}(x)\neq y} = 1 - p(x,\mathsf{h}(x)). \tag{23}$$

Since $\{\mathsf{h}(x) : h \in \mathcal{H}\} = \mathsf{H}(x)$, the minimal conditional ℓ_{0-1}-risk can be expressed as follows:

$$\mathcal{C}^*_{\ell_{0-1},\mathcal{H}}(x) = 1 - \max_{y\in\mathsf{H}(x)} p(x,y),$$

which proves the first part of the lemma. By the definition,

$$\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) = \mathcal{C}_{\ell_{0-1}}(h,x) - \mathcal{C}^*_{\ell_{0-1},\mathcal{H}}(x) = \max_{y\in\mathsf{H}(x)} p(x,y) - p(x,\mathsf{h}(x)).$$

This leads to

$$\left[\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)\right]_\epsilon = \left[\max_{y\in\mathsf{H}(x)} p(x,y) - p(x,\mathsf{h}(x))\right]_\epsilon.$$

□

Lemma 11. *For any $x \in \mathcal{X}$, the minimal conditional ℓ_γ-risk and the conditional ϵ-regret for ℓ_γ can be expressed as follows:*

$$\mathcal{C}^*_{\ell_\gamma,\mathcal{H}}(x) = 1 - \max_{y\in\mathsf{H}_\gamma(x)} p(x,y)\mathbb{1}_{\mathcal{H}_\gamma(x)\neq\varnothing}$$

$$\left[\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x)\right]_\epsilon = \begin{cases}\left[\max_{y\in\mathsf{H}_\gamma(x)} p(x,y) - p(x,\mathsf{h}(x))\mathbb{1}_{h\in\mathcal{H}_\gamma(x)}\right]_\epsilon & \text{if } \mathcal{H}_\gamma(x)\neq\varnothing\\ 0 & \text{otherwise.}\end{cases}$$

Proof. By the definition, the conditional ℓ_γ-risk can be expressed as follows:

$$\mathcal{C}_{\ell_\gamma}(h,x) = \sum_{y\in\mathcal{Y}} p(x,y) \sup_{x':\|x-x'\|_p\leq\gamma} \mathbb{1}_{\rho_h(x',y)\leq 0} = \begin{cases}1 - p(x,\mathsf{h}(x)) & h\in\mathcal{H}_\gamma(x)\\ 1 & \text{otherwise.}\end{cases} \tag{24}$$

When $\mathcal{H}_\gamma(x) = \varnothing$, (24) implies that $\mathcal{C}^*_{\ell_\gamma,\mathcal{H}}(x) = 1$. When $\mathcal{H}_\gamma(x)\neq\varnothing$, $\mathsf{H}_\gamma(x)$ is also non-empty. By (24), $y\in\mathcal{Y}_\gamma(x)$ if and only if there exists $h\in\mathcal{H}_\gamma$ such that $\mathcal{C}_{\ell_\gamma}(h,x) = 1 - p(x,y)$. Therefore, the minimal conditional ℓ_γ-risk can be expressed as follows:

$$\mathcal{C}^*_{\ell_\gamma,\mathcal{H}}(x) = 1 - \max_{y\in\mathsf{H}_\gamma(x)} p(x,y)\mathbb{1}_{\mathcal{H}_\gamma(x)\neq\varnothing},$$

which proves the first part of lemma. When $\mathcal{H}_\gamma(x) = \varnothing$, $\mathcal{C}_{\ell_\gamma}(h,x)\equiv 1$, which implies that $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x)\equiv 0$. When $\mathcal{H}_\gamma(x)\neq\varnothing$, $\mathsf{H}_\gamma(x)$ is also non-empty, for $h\in\mathcal{H}_\gamma(x)$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x) = 1 - p(x,\mathsf{h}(x)) - \left(1 - \max_{y\in\mathsf{H}_\gamma(x)} p(x,y)\right) = \max_{y\in\mathsf{H}_\gamma(x)} p(x,y) - p(x,\mathsf{h}(x))$; for $h\notin\mathcal{H}_\gamma(x)$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x) = 1 - \left(1 - \max_{y\in\mathsf{H}_\gamma(x)} p(x,y)\right) = \max_{y\in\mathsf{H}_\gamma(x)} p(x,y)$. Therefore,

$$\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x) = \begin{cases}\max_{y\in\mathsf{H}_\gamma(x)} p(x,y) - p(x,\mathsf{h}(x))\mathbb{1}_{h\in\mathcal{H}_\gamma(x)} & \mathcal{H}_\gamma(x)\neq\varnothing\\ 0 & \text{otherwise.}\end{cases}$$

This leads to

$$\left[\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x)\right]_\epsilon = \begin{cases}\left[\max_{y\in\mathsf{H}_\gamma(x)} p(x,y) - p(x,\mathsf{h}(x))\mathbb{1}_{h\in\mathcal{H}_\gamma(x)}\right]_\epsilon & \mathcal{H}_\gamma(x)\neq\varnothing\\ 0 & \text{otherwise.}\end{cases}$$

□

G Proof of negative results and $\mathcal{H}$-consistency bounds for max losses $\Phi^{\max}$

Theorem 6 (Negative results for convex Φ). *Assume that $c > 2$. Suppose that Φ is convex and non-increasing, and $\mathcal{H}$ satisfies there exist $x\in\mathcal{X}$ and $h\in\mathcal{H}$ such that $|\mathsf{H}(x)|\geq 2$ and $h(x,y)$ are equal for all $y\in\mathcal{Y}$. If for a non-decreasing function $f:\mathbb{R}_+\to\mathbb{R}_+$, the following $\mathcal{H}$-consistency bound holds for any hypothesis $h\in\mathcal{H}$ and any distribution $\mathcal{D}$:*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq f\left(\mathcal{R}_{\Phi^{\max}}(h) - \mathcal{R}^*_{\Phi^{\max},\mathcal{H}}\right), \tag{6}$$

then, f is lower bounded by $\frac{1}{2}$.

Proof. Consider the distribution that supports on a singleton domain $\{x\}$ with x satisfying that $|\mathsf{H}(x)| \geq 2$. Take $y_1 \in \mathsf{H}(x)$ such that $y_1 \neq c$ and $y_2 \in \mathcal{Y}$ such that $y_2 \neq y_1$, $y_2 \neq c$. We define $p(x)$ as $p(x, y_1) = p(x, y_2) = \frac{1}{2}$ and $p(x, y) = 0$ for other $y \in \mathcal{Y}$. Let $h_0 \in \mathcal{H}$ such that $h_0(x, 1) = h_0(x, 2) = \ldots = h_0(x, c)$. By Lemma 3 and the fact that $y_1 \in \mathsf{H}(x)$, the minimal conditional ℓ_{0-1}-risk is

$$\mathcal{R}^*_{\ell_{0-1},\mathcal{H}} = \mathcal{C}^*_{\ell_{0-1},\mathcal{H}}(x) = 1 - \max_{y \in \mathsf{H}(x)} p(x, y) = 1 - p(x, y_1) = \frac{1}{2}.$$

For $h = h_0$, we have

$$\mathcal{R}_{\ell_{0-1}}(h_0) = \mathcal{C}_{\ell_{0-1}}(h_0, x) = \sum_{y \in \mathcal{Y}} p(x, y) \mathbb{1}_{\mathsf{h}_0(x) \neq y} = 1 - p(x, \mathsf{h}_0(x)) = 1 - p(x, c) = 1.$$

For the max loss, the conditional $\Phi^{\max}$-risk can be expressed as follows:

$$\mathcal{C}_{\Phi^{\max}}(h, x) = \sum_{y \in \mathcal{Y}} p(x, y) \Phi(\rho_h(x, y)) = \frac{1}{2}\Phi(\rho_h(x, y_1)) + \frac{1}{2}\Phi(\rho_h(x, y_2)).$$

If Φ is convex and non-increasing, we obtain for any $h \in \mathcal{H}$,

$$\begin{aligned}
\mathcal{R}_{\Phi^{\max}}(h) = \mathcal{C}_{\Phi^{\max}}(h, x) &= \frac{1}{2}\Phi(\rho_h(x, y_1)) + \frac{1}{2}\Phi(\rho_h(x, y_2)) \\
&\geq \Phi\left(\frac{1}{2}\rho_h(x, y_1) + \frac{1}{2}\rho_h(x, y_2)\right) && (\Phi \text{ is convex}) \\
&= \Phi\left(\frac{1}{2}\left(h(x, y_1) + h(x, y_2) - \max_{y \neq y_1} h(x, y) - \max_{y \neq y_2} h(x, y)\right)\right) \\
&\geq \Phi(0), && (\Phi \text{ is non-increasing})
\end{aligned}$$

where both equality can be achieved by h_0. Therefore,

$$\mathcal{R}^*_{\Phi^{\max},\mathcal{H}} = \mathcal{C}^*_{\Phi^{\max},\mathcal{H}}(x) = \mathcal{R}_{\Phi^{\max}}(h_0) = \Phi(0).$$

If (6) holds for some non-decreasing function f, then, we obtain for any $h \in \mathcal{H}$,

$$\mathcal{R}_{\ell_{0-1}}(h) - \frac{1}{2} \leq f(\mathcal{R}_{\Phi^{\max}}(h) - \Phi(0)).$$

Let $h = h_0$, then $f(0) \geq 1/2$. Since f is non-decreasing, for any $t \geq 0$, $f(t) \geq 1/2$. □

Theorem 7 ($\mathcal{H}$**-consistency bound of** $\Phi^{\max}_\rho$)**.** *Suppose that $\mathcal{H}$ is symmetric. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \frac{\mathcal{R}_{\Phi^{\max}_\rho}(h) - \mathcal{R}^*_{\Phi^{\max}_\rho,\mathcal{H}} + \mathcal{M}_{\Phi^{\max}_\rho,\mathcal{H}}}{\min\left\{1, \frac{\inf_{x \in \mathcal{X}} \sup_{h \in \mathcal{H}} \rho_h(x, \mathsf{h}(x))}{\rho}\right\}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}. \tag{7}$$

Proof. By the definition, the conditional $\Phi^{\max}_\rho$-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\Phi^{\max}_\rho}(h, x) &= \sum_{y \in \mathcal{Y}} p(x, y) \Phi_\rho(\rho_h(x, y)) \\
&= 1 - p(x, \mathsf{h}(x)) + \max\left\{0, 1 - \frac{\rho_h(x, \mathsf{h}(x))}{\rho}\right\} p(x, \mathsf{h}(x)) \\
&= 1 - \min\left\{1, \frac{\rho_h(x, \mathsf{h}(x))}{\rho}\right\} p(x, \mathsf{h}(x))
\end{aligned} \tag{25}$$

Since $\mathcal{H}$ is symmetric, for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$\sup_{h \in \{h \in \mathcal{H} : \mathsf{h}(x) = y\}} \rho_h(x, \mathsf{h}(x)) = \sup_{h \in \mathcal{H}} \rho_h(x, \mathsf{h}(x))$$

Therefore, the minimal conditional $\Phi^{\max}_\rho$-risk can be expressed as follows:

$$\mathcal{C}^*_{\Phi^{\max}_\rho,\mathcal{H}}(x) = 1 - \min\left\{1, \frac{\sup_{h \in \mathcal{H}} \rho_h(x, \mathsf{h}(x))}{\rho}\right\} \max_{y \in \mathcal{Y}} p(x, y).$$

By the definition and using the fact that $\mathsf{H}(x) = \mathcal{Y}$ when $\mathcal{H}$ is symmetric, we obtain

$$\begin{aligned}
\Delta\mathcal{C}_{\Phi_\rho^{\max},\mathcal{H}}(h,x) &= \mathcal{C}_{\Phi_\rho^{\max}}(h,x) - \mathcal{C}^*_{\Phi_\rho^{\max},\mathcal{H}}(x)\\
&= \min\left\{1, \frac{\sup_{h\in\mathcal{H}} \rho_h(x,\mathsf{h}(x))}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y) - \min\left\{1, \frac{\rho_h(x,\mathsf{h}(x))}{\rho}\right\} p(x,\mathsf{h}(x))\\
&\geq \min\left\{1, \frac{\sup_{h\in\mathcal{H}} \rho_h(x,\mathsf{h}(x))}{\rho}\right\} \left(\max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\right)\\
&\geq \min\left\{1, \frac{\sup_{h\in\mathcal{H}} \rho_h(x,\mathsf{h}(x))}{\rho}\right\} \Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) && (\mathsf{H}(x) = \mathcal{Y})\\
&\geq \min\left\{1, \frac{\sup_{h\in\mathcal{H}} \rho_h(x,\mathsf{h}(x))}{\rho}\right\} [\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)]_\epsilon && ([x]_\epsilon \leq x)\\
&\geq \min\left\{1, \frac{\inf_{x\in\mathcal{X}} \sup_{h\in\mathcal{H}} \rho_h(x,\mathsf{h}(x))}{\rho}\right\} [\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)]_\epsilon
\end{aligned}$$

for any $\epsilon \geq 0$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric hypothesis set, $\epsilon = 0$ and

$$\Psi(t) = \min\left\{1, \frac{\inf_{x\in\mathcal{X}} \sup_{h\in\mathcal{H}} \rho_h(x,\mathsf{h}(x))}{\rho}\right\} t$$

in Theorem 4, or, equivalently, $\Gamma(t) = \Psi^{-1}(t)$ in Theorem 5, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \frac{\mathcal{R}_{\Phi_\rho^{\max}}(h) - \mathcal{R}^*_{\Phi_\rho^{\max},\mathcal{H}} + \mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}}}{\min\left\{1, \frac{\inf_{x\in\mathcal{X}} \sup_{h\in\mathcal{H}} \rho_h(x,\mathsf{h}(x))}{\rho}\right\}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}.$$

□

Theorem 9 (Realizable $\mathcal{H}$-consistency bound of $\Phi^{\max}$). *Suppose that $\mathcal{H}$ is symmetric and complete, and Φ is non-increasing and satisfies that $\lim_{t\to+\infty} \Phi(t) = 0$. Then, for any hypothesis $h \in \mathcal{H}$ and any $\mathcal{H}$-realizable distribution $\mathcal{D}$, we have*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi^{\max}}(h) - \mathcal{R}^*_{\Phi^{\max},\mathcal{H}} + \mathcal{M}_{\Phi^{\max},\mathcal{H}}. \tag{8}$$

Proof. Under the $\mathcal{H}$-realizability assumption of distribution, for any $x \in \mathcal{X}$, there exists $y \in \mathcal{Y}$ such that $p(x,y) = 1$. Then, the conditional $\Phi^{\max}$-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\Phi^{\max}}(h,x) &= \sum_{y\in\mathcal{Y}} p(x,y)\Phi(\rho_h(x,y))\\
&= \Phi(\rho_h(x,y_{\max})).
\end{aligned} \tag{26}$$

Since $\mathcal{H}$ is symmetric and complete, there exists $h \in \mathcal{H}$ such that $\mathsf{h}(x) = y_{\max}$ and we have

$$\begin{aligned}
\sup_{h\in\{h\in\mathcal{H}:\mathsf{h}(x)=y_{\max}\}} \rho_h(x,\mathsf{h}(x)) &= \sup_{h\in\mathcal{H}} \rho_h(x,\mathsf{h}(x))\\
&= \sup_{h\in\mathcal{H}} \left(\max_{y\in\mathcal{Y}} h(x,y) - \max_{y\neq\mathsf{h}(x)} h(x,y)\right)\\
&= +\infty.
\end{aligned}$$

Thus, using the fact that $\lim_{t\to+\infty} \Phi(t) = 0$, the minimal conditional $\Phi^{\max}$-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}^*_{\Phi^{\max},\mathcal{H}}(x) &= \inf_{h\in\mathcal{H}} \mathcal{C}_{\Phi^{\max}}(h,x)\\
&= \inf_{h\in\mathcal{H}} \Phi(\rho_h(x,\mathsf{h}(x)))\\
&= \Phi\left(\sup_{h\in\mathcal{H}} \rho_h(x,\mathsf{h}(x))\right) && (\Phi \text{ is non-increasing})\\
&= 0 && (\lim_{t\to+\infty} \Phi(t) = 0)
\end{aligned}$$

By the definition and using the fact that $\mathsf{H}(x) = \mathcal{Y}$ when $\mathcal{H}$ is symmetric, we obtain

$$\begin{aligned}
\Delta\mathcal{C}_{\Phi^{\max},\mathcal{H}}(h,x) &= \mathcal{C}_{\Phi^{\max}}(h,x) - \mathcal{C}^*_{\Phi^{\max},\mathcal{H}}(x) \\
&= \Phi(\rho_h(x, y_{\max})) \\
&\geq \Phi(0)\mathbb{1}_{y_{\max}\neq \mathsf{h}(x)} && (\Phi \text{ is non-increasing}) \\
&\geq \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x)) \\
&= \Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) && (\text{by Lemma 3 and } \mathsf{H}(x) = \mathcal{Y}) \\
&\geq [\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)]_\epsilon && ([t]_\epsilon \leq t)
\end{aligned}$$

for any $\epsilon \geq 0$. Note that $\mathcal{M}_{\ell_{0-1},\mathcal{H}} = 0$ under the realizability assumption. Therefore, taking $\mathcal{P}$ be the set of $\mathcal{H}$-realizable distributions, $\mathcal{H}$ be the symmetric and complete hypothesis set, $\epsilon = 0$ and $\Psi(t) = t$ in Theorem 4, or, equivalently, $\Gamma(t) = t$ in Theorem 5, we obtain for any hypothesis $h \in \mathcal{H}$ and any $\mathcal{H}$-realizable distribution,

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi^{\max}}(h) - \mathcal{R}^*_{\Phi^{\max},\mathcal{H}} + \mathcal{M}_{\Phi^{\max},\mathcal{H}}$$

□

H Proof of $\mathcal{H}_{\text{all}}, \mathcal{H}_{\text{lin}}, \mathcal{H}_{\text{NN}}$-consistency bounds for max ρ-margin loss $\Phi^{\max}_\rho$

Corollary 18 ($\mathcal{H}_{\text{all}}$**-consistency bound of** $\Phi^{\max}_\rho$)**.** *For any hypothesis $h \in \mathcal{H}_{\text{all}}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{all}}} \leq \mathcal{R}_{\Phi^{\max}_\rho}(h) - \mathcal{R}^*_{\Phi^{\max}_\rho,\mathcal{H}_{\text{all}}}. \tag{27}$$

Proof. For $\mathcal{H} = \mathcal{H}_{\text{all}}$, we have for all $x \in \mathcal{X}$, $\sup_{h\in\mathcal{H}_{\text{all}}} \rho_h(x,\mathsf{h}(x)) > \rho$. Furthermore, as shown by Steinwart (2007, Theorem 3.2), the minimizability gaps $\mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{all}}} = \mathcal{M}_{\Phi^{\max}_\rho,\mathcal{H}_{\text{all}}} = 0$. Therefore, by Theorem 7, the $\mathcal{H}_{\text{all}}$-consistency bound of $\Phi^{\max}_\rho$ can be expressed as follows:

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{all}}} \leq \mathcal{R}_{\Phi^{\max}_\rho}(h) - \mathcal{R}^*_{\Phi^{\max}_\rho,\mathcal{H}_{\text{all}}}.$$

□

Corollary 19 ($\mathcal{H}_{\text{lin}}$**-consistency bound of** $\Phi^{\max}_\rho$)**.** *For any hypothesis $h \in \mathcal{H}_{\text{lin}}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{lin}}} \leq \frac{\mathcal{R}_{\Phi^{\max}_\rho}(h) - \mathcal{R}^*_{\Phi^{\max}_\rho,\mathcal{H}_{\text{lin}}} + \mathcal{M}_{\Phi^{\max}_\rho,\mathcal{H}_{\text{lin}}}}{\min\left\{1, \frac{2B}{\rho}\right\}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{lin}}}, \tag{28}$$

where $\mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{lin}}} = \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{lin}}} - \mathbb{E}_X[1 - \max_{y\in\mathcal{Y}} p(x,y)]$ *and* $\mathcal{M}_{\Phi^{\max}_\rho,\mathcal{H}_{\text{lin}}} = \mathcal{R}^*_{\Phi^{\max}_\rho,\mathcal{H}_{\text{lin}}} - \mathbb{E}_X\left[1 - \min\left\{1, \frac{2\left(W\|x\|_p + B\right)}{\rho}\right\}\max_{y\in\mathcal{Y}} p(x,y)\right]$.

Proof. For $\mathcal{H} = \mathcal{H}_{\text{lin}}$, we have for all $x \in \mathcal{X}$,

$$\begin{aligned}
\sup_{h\in\mathcal{H}_{\text{lin}}} \rho_h(x,\mathsf{h}(x)) &= \sup_{h\in\mathcal{H}_{\text{lin}}} \left(\max_{y\in\mathcal{Y}} h(x,y) - \max_{y\neq\mathsf{h}(x)} h(x,y)\right) \\
&= \max_{\|w\|_q\leq W, |b|\leq B} (w\cdot x + b) - \min_{\|w\|_q\leq W, |b|\leq B} (w\cdot x + b) \\
&= 2\left(W\|x\|_p + B\right)
\end{aligned} \tag{29}$$

Thus, $\inf_{x\in\mathcal{X}} \sup_{h\in\mathcal{H}_{\text{lin}}} \rho_h(x,\mathsf{h}(x)) = \inf_{x\in\mathcal{X}} 2\left(W\|x\|_p + B\right) = 2B$. Since $\mathcal{H} = \mathcal{H}_{\text{lin}}$ is symmetric, by lemma 3, we have

$$\mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{lin}}} = \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{lin}}} - \mathbb{E}_X\left[1 - \max_{y\in\mathcal{Y}} p(x,y)\right]. \tag{30}$$

By the definition, the conditional $\Phi_\rho^{\max}$-risk can be expressed as follows:

$$\begin{aligned}\mathcal{C}_{\Phi_\rho^{\max}}(h,x) &= \sum_{y\in\mathcal{Y}} p(x,y)\Phi_\rho(\rho_h(x,y))\\ &= 1 - p(x,\mathsf{h}(x)) + \max\left\{0, 1-\frac{\rho_h(x,\mathsf{h}(x))}{\rho}\right\} p(x,\mathsf{h}(x))\\ &= 1 - \min\left\{1, \frac{\rho_h(x,\mathsf{h}(x))}{\rho}\right\} p(x,\mathsf{h}(x))\end{aligned}$$

Since $\mathcal{H}_{\text{lin}}$ is symmetric, for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$\sup_{h\in\{h\in\mathcal{H}_{\text{lin}}:\mathsf{h}(x)=y\}} \rho_h(x,\mathsf{h}(x)) = \sup_{h\in\mathcal{H}_{\text{lin}}} \rho_h(x,\mathsf{h}(x)).$$

Thus, using (29), the minimal conditional $\Phi_\rho^{\max}$-risk can be expressed as follows:

$$\begin{aligned}\mathcal{C}^*_{\Phi_\rho^{\max},\mathcal{H}_{\text{lin}}}(x) &= 1 - \min\left\{1, \frac{\sup_{h\in\mathcal{H}_{\text{lin}}}\rho_h(x,\mathsf{h}(x))}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y)\\ &= 1 - \min\left\{1, \frac{2\left(W\|x\|_p + B\right)}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y) \qquad \text{(by (29))}\end{aligned}$$

Therefore, the $\left(\Phi_\rho^{\max}, \mathcal{H}_{\text{lin}}\right)$-minimizability gap is

$$\mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}_{\text{lin}}} = \mathcal{R}^*_{\Phi_\rho^{\max},\mathcal{H}_{\text{lin}}} - \mathbb{E}_X\left[1 - \min\left\{1, \frac{2\left(W\|x\|_p + B\right)}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y)\right]. \tag{31}$$

By Theorem 7, the $\mathcal{H}_{\text{lin}}$-consistency bound of $\Phi_\rho^{\max}$ can be expressed as follows:

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{lin}}} \leq \frac{\mathcal{R}_{\Phi_\rho^{\max}}(h) - \mathcal{R}^*_{\Phi_\rho^{\max},\mathcal{H}_{\text{lin}}} + \mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}_{\text{lin}}}}{\min\left\{1, \frac{2B}{\rho}\right\}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{lin}}}.$$

where $\mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{lin}}}$ and $\mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}_{\text{lin}}}$ are given by (30) and (31) respectively. □

Corollary 20 ($\mathcal{H}_{\text{NN}}$**-consistency bound of** $\Phi_\rho^{\max}$)**.** *For any hypothesis $h \in \mathcal{H}_{\text{NN}}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{NN}}} \leq \frac{\mathcal{R}_{\Phi_\rho^{\max}}(h) - \mathcal{R}^*_{\Phi_\rho^{\max},\mathcal{H}_{\text{NN}}} + \mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}_{\text{NN}}}}{\min\left\{1, \frac{2\Lambda B}{\rho}\right\}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{NN}}}, \tag{32}$$

where $\mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{NN}}} = \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{NN}}} - \mathbb{E}_X[1 - \max_{y\in\mathcal{Y}} p(x,y)]$ *and* $\mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}_{\text{NN}}} = \mathcal{R}^*_{\Phi_\rho^{\max},\mathcal{H}_{\text{NN}}} - \mathbb{E}_X\left[1 - \min\left\{1, \frac{2\Lambda\left(W\|x\|_p+B\right)}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y)\right]$.

Proof. For $\mathcal{H} = \mathcal{H}_{\text{NN}}$, we have for all $x \in \mathcal{X}$,

$$\begin{aligned}&\sup_{h\in\mathcal{H}_{\text{NN}}} \rho_h(x,\mathsf{h}(x)) = \sup_{h\in\mathcal{H}_{\text{NN}}} \left(\max_{y\in\mathcal{Y}} h(x,y) - \max_{y\neq\mathsf{h}(x)} h(x,y)\right)\\ &= \max_{\|u\|_1\leq\Lambda,\|w_j\|_q\leq W,|b_j|\leq B} \left(\sum_{j=1}^n u_j(w_j\cdot x + b_j)_+\right) - \min_{\|u\|_1\leq\Lambda,\|w_j\|_q\leq W,|b_j|\leq B} \left(\sum_{j=1}^n u_j(w_j\cdot x + b_j)_+\right) \qquad (33)\\ &= 2\Lambda\left(W\|x\|_p + B\right)\end{aligned}$$

Thus, $\inf_{x\in\mathcal{X}} \sup_{h\in\mathcal{H}_{\text{NN}}} \rho_h(x,\mathsf{h}(x)) = \inf_{x\in\mathcal{X}} 2\Lambda\left(W\|x\|_p + B\right) = 2\Lambda B$. Since $\mathcal{H} = \mathcal{H}_{\text{NN}}$ is symmetric, by lemma 3, we have

$$\mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{NN}}} = \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\text{NN}}} - \mathbb{E}_X\left[1 - \max_{y\in\mathcal{Y}} p(x,y)\right]. \tag{34}$$

By the definition, the conditional $\Phi_\rho^{\max}$-risk can be expressed as follows:

$$\begin{aligned}\mathcal{C}_{\Phi_\rho^{\max}}(h,x) &= \sum_{y\in\mathcal{Y}} p(x,y)\Phi_\rho(\rho_h(x,y))\\ &= 1 - p(x,\mathsf{h}(x)) + \max\left\{0, 1-\frac{\rho_h(x,\mathsf{h}(x))}{\rho}\right\} p(x,\mathsf{h}(x))\\ &= 1 - \min\left\{1, \frac{\rho_h(x,\mathsf{h}(x))}{\rho}\right\} p(x,\mathsf{h}(x))\end{aligned}$$

Since $\mathcal{H}_{\mathrm{NN}}$ is symmetric, for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$\sup_{h\in\{h\in\mathcal{H}_{\mathrm{NN}}:\mathsf{h}(x)=y\}} \rho_h(x,\mathsf{h}(x)) = \sup_{h\in\mathcal{H}_{\mathrm{NN}}} \rho_h(x,\mathsf{h}(x)).$$

Thus, using (33), the minimal conditional $\Phi_\rho^{\max}$-risk can be expressed as follows:

$$\begin{aligned}\mathcal{C}^*_{\Phi_\rho^{\max},\mathcal{H}_{\mathrm{NN}}}(x) &= 1 - \min\left\{1, \frac{\sup_{h\in\mathcal{H}_{\mathrm{NN}}}\rho_h(x,\mathsf{h}(x))}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y)\\ &= 1 - \min\left\{1, \frac{2\Lambda\left(W\|x\|_p + B\right)}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y) \qquad \text{(by (33))}\end{aligned}$$

Therefore, the $\left(\Phi_\rho^{\max}, \mathcal{H}_{\mathrm{NN}}\right)$-minimizability gap is

$$\mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}_{\mathrm{NN}}} = \mathcal{R}^*_{\Phi_\rho^{\max},\mathcal{H}_{\mathrm{NN}}} - \mathbb{E}_X\left[1 - \min\left\{1, \frac{2\Lambda\left(W\|x\|_p + B\right)}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y)\right]. \tag{35}$$

By Theorem 7, the $\mathcal{H}_{\mathrm{NN}}$-consistency bound of $\Phi_\rho^{\max}$ can be expressed as follows:

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}_{\mathrm{NN}}} \leq \frac{\mathcal{R}_{\Phi_\rho^{\max}}(h) - \mathcal{R}^*_{\Phi_\rho^{\max},\mathcal{H}_{\mathrm{NN}}} + \mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}_{\mathrm{NN}}}}{\min\left\{1, \frac{2\Lambda B}{\rho}\right\}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}_{\mathrm{NN}}}.$$

where $\mathcal{M}_{\ell_{0-1},\mathcal{H}_{\mathrm{NN}}}$ and $\mathcal{M}_{\Phi_\rho^{\max},\mathcal{H}_{\mathrm{NN}}}$ are given by (34) and (35) respectively. □

I Auxiliary Lemma for sum losses

Lemma 21. *Fix a vector $\tau = (\tau_1, \ldots, \tau_c)$ in the probability simplex of $\mathbb{R}^c$ and any real values $a_1 \leq a_2 \leq \cdots \leq a_c$ in increasing order. Then, for any permutation σ of the set $\{1, \ldots, c\}$,*

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \cdot \begin{bmatrix} \tau_{\sigma(1)} \\ \tau_{\sigma(2)} \\ \vdots \\ \tau_{\sigma(c)} \end{bmatrix} \leq \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \cdot \begin{bmatrix} \tau_{[1]} \\ \tau_{[2]} \\ \vdots \\ \tau_{[c]} \end{bmatrix},$$

where we define $\tau_{[1]}, \tau_{[2]}, \ldots, \tau_{[c]}$ by sorting the probabilities $\{\tau_y : y \in \{1, \ldots, c\}\}$ in increasing order.

Proof. For any permutation σ of the set $\{1, \ldots, c\}$, we prove by induction. At the first step, if $\sigma(c) = [c]$, then let $\sigma_1 = \sigma$. Otherwise, denote $k_1 \in \{1, \ldots, c-1\}$ such that $\sigma(k_1) = [c]$ and choose σ_1 to be the permutation that differs from σ only by permuting c and k_1. Thus,

$$\begin{aligned}\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \cdot \begin{bmatrix} \tau_{\sigma(1)} \\ \tau_{\sigma(2)} \\ \vdots \\ \tau_{\sigma(c)} \end{bmatrix} - \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \cdot \begin{bmatrix} \tau_{\sigma_1(1)} \\ \tau_{\sigma_1(2)} \\ \vdots \\ \tau_{\sigma_1(c)} \end{bmatrix} &= a_{k_1}\tau_{[c]} + a_c\tau_{\sigma(c)} - \left(a_{k_1}\tau_{\sigma(c)} + a_c\tau_{[c]}\right)\\ &= \left(a_{k_1} - a_c\right)\left(\tau_{[c]} - \tau_{\sigma(c)}\right) \leq 0.\end{aligned}$$

At the second step, if $\sigma_1(c-1) = [c-1]$, then let $\sigma_2 = \sigma_1$. Otherwise, denote $k_2 \in \{1, \ldots, c-2\}$ such that $\sigma_1(k_2) = [c-1]$ and choose σ_2 to be the permutation that differs from σ_1 only by permuting $c-1$ and k_2. Thus,

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \cdot \begin{bmatrix} \tau_{\sigma_1(1)} \\ \tau_{\sigma_1(2)} \\ \vdots \\ \tau_{\sigma_1(c)} \end{bmatrix} - \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \cdot \begin{bmatrix} \tau_{\sigma_2(1)} \\ \tau_{\sigma_2(2)} \\ \vdots \\ \tau_{\sigma_2(c)} \end{bmatrix} = (a_{k_2} - a_{c-1})\big(\tau_{[c-1]} - \tau_{\sigma_1(c-1)}\big) \leq 0.$$

And so on, at the nth step, if $\sigma_{n-1}(c-n+1) = [c-n+1]$, then let $\sigma_n = \sigma_{n-1}$. Otherwise, denote $k_n \in \{1, \ldots, c-n\}$ such that $\sigma_{n-1}(k_n) = [c-n+1]$ and choose σ_n to be the permutation that differs from σ_{n-1} only by permuting $c-n+1$ and k_n. We have

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \cdot \begin{bmatrix} \tau_{\sigma_{n-1}(1)} \\ \tau_{\sigma_{n-1}(2)} \\ \vdots \\ \tau_{\sigma_{n-1}(c)} \end{bmatrix} \leq \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \cdot \begin{bmatrix} \tau_{\sigma_n(1)} \\ \tau_{\sigma_n(2)} \\ \vdots \\ \tau_{\sigma_n(c)} \end{bmatrix}.$$

Finally, after c steps, we will obtain σ_c which satisfies $\sigma_c(y) = [y]$ for any $y \in \{1, \ldots, c\}$. Therefore, we obtain

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \cdot \begin{bmatrix} \tau_{\sigma(1)} \\ \tau_{\sigma(2)} \\ \vdots \\ \tau_{\sigma(c)} \end{bmatrix} \leq \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \begin{bmatrix} \tau_{\sigma_1(1)} \\ \tau_{\sigma_1(2)} \\ \vdots \\ \tau_{\sigma_1(c)} \end{bmatrix} \leq \ldots \leq \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \begin{bmatrix} \tau_{\sigma_n(1)} \\ \tau_{\sigma_n(2)} \\ \vdots \\ \tau_{\sigma_n(c)} \end{bmatrix} \leq \ldots \leq \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \cdot \begin{bmatrix} \tau_{[1]} \\ \tau_{[2]} \\ \vdots \\ \tau_{[c]} \end{bmatrix}$$

which proves the lemma. □

J Proof of negative and $\mathcal{H}$-consistency bounds for sum losses Φ^{sum}

By the definition, the conditional Φ^{sum}-risk can be expressed as follows:

$$\begin{aligned} \mathcal{C}_{\Phi^{\text{sum}}}(h, x) &= \sum_{y \in \mathcal{Y}} p(x, y) \sum_{y' \neq y} \Phi(h(x, y) - h(x, y')) \\ &= \sum_{y \in \mathcal{Y}} p(x, y) \sum_{y' \in \mathcal{Y}} \Phi(h(x, y) - h(x, y')) - \Phi(0) \end{aligned} \tag{36}$$

Theorem 10 (**Negative results for hinge loss**). *Assume that $c > 2$. Suppose that $\mathcal{H}$ is symmetric and complete. If for a non-decreasing function $f: \mathbb{R}_+ \to \mathbb{R}_+$, the following $\mathcal{H}$-consistency bound holds for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$:*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}, \mathcal{H}} \leq f\Big(\mathcal{R}_{\Phi^{\text{sum}}_{\text{hinge}}}(h) - \mathcal{R}^*_{\Phi^{\text{sum}}_{\text{hinge}}, \mathcal{H}}\Big), \tag{10}$$

then, f is lower bounded by $\frac{1}{6}$.

Proof. Consider the distribution that supports on a singleton domain $\{x\}$. We define $p(x)$ as $p(x, 1) = \frac{1}{2} - \epsilon$, $p(x, 2) = \frac{1}{3}$, $p(x, 3) = \frac{1}{6} + \epsilon$ and $p(x, y) = 0$ for other $y \in \mathcal{Y}$, where $0 < \epsilon < \frac{1}{6}$. Note $p(x, 1) > p(x, 2) > p(x, 3) > p(x, y) = 0$, $y \notin \{1, 2, 3\}$. Let $h_0 \in \mathcal{H}$ such that $h_0(x, 1) = 1$, $h_0(x, 2) = 1$, $h_0(x, 3) = 0$ and $h_0(x, y) = -1$ for other $y \in \mathcal{Y}$. By the completeness of $\mathcal{H}$, the hypothesis h is in $\mathcal{H}$. By Lemma 3 and the fact that $\mathsf{H}(x) = \mathcal{Y}$ when $\mathcal{H}$ is symmetric, the minimal conditional ℓ_{0-1}-risk is

$$\mathcal{R}^*_{\ell_{0-1}, \mathcal{H}} = \mathcal{C}^*_{\ell_{0-1}, \mathcal{H}}(x) = 1 - \max_{y \in \mathcal{Y}} p(x, y) = 1 - p(x, 1) = \frac{1}{2} + \epsilon.$$

For $h = h_0$, we have

$$\mathcal{R}_{\ell_{0-1}}(h_0) = \mathcal{C}_{\ell_{0-1}}(h_0, x) = \sum_{y \in \mathcal{Y}} p(x, y) \mathbb{1}_{\mathsf{h}_0(x) \neq y} = 1 - p(x, \mathsf{h}_0(x)) = 1 - p(x, 2) = \frac{2}{3}.$$

For the sum hinge loss, by (36), the conditional $\Phi_{\text{hinge}}^{\text{sum}}$-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\Phi_{\text{hinge}}^{\text{sum}}}(h,x) &= \sum_{y\in\mathcal{Y}} p(x,y) \sum_{y'\neq y} \max\{0, 1+h(x,y')-h(x,y)\} \\
&= \sum_{y\in\{1,2,3\}} p(x,y) \sum_{y'\neq y} \max\{0, 1+h(x,y')-h(x,y)\} \\
&\geq \sum_{y\in\{1,2,3\}} p(x,y) \sum_{y'\neq y, y'\in\{1,2,3\}} \max\{0, 1+h(x,y')-h(x,y)\} \\
&= \left(\frac{1}{2}-\epsilon\right)[\max\{0,1+h(x,2)-h(x,1)\}+\max\{0,1+h(x,3)-h(x,1)\}] \\
&+ \frac{1}{3}[\max\{0,1+h(x,1)-h(x,2)\}+\max\{0,1+h(x,3)-h(x,2)\}] \\
&+ \left(\frac{1}{6}+\epsilon\right)[\max\{0,1+h(x,1)-h(x,3)\}+\max\{0,1+h(x,2)-h(x,3)\}] \\
&= g(h).
\end{aligned}$$

Note $\mathcal{C}_{\Phi_{\text{hinge}}^{\text{sum}}}(h_0,x) = 3\epsilon+\frac{3}{2}$. Since $\frac{1}{2}-\epsilon > \frac{1}{3} > \frac{1}{6}+\epsilon$, by Lemma 21, we have

$$\inf_{h\in\mathcal{H}} g(h) = \inf_{h\in\mathcal{H}:\, h(x,1)\geq h(x,2)\geq h(x,3)} g(h).$$

When $h(x,1)\geq h(x,2)\geq h(x,3)$, $g(h)$ can be written as

$$\begin{aligned}
g(h) &= \left(\frac{1}{2}-\epsilon\right)[\max\{0,1+h(x,2)-h(x,1)\}+\max\{0,1+h(x,3)-h(x,1)\}] \\
&+ \frac{1}{3}[(1+h(x,1)-h(x,2))+\max\{0,1+h(x,3)-h(x,2)\}] \\
&+ \left(\frac{1}{6}+\epsilon\right)[(1+h(x,1)-h(x,3))+(1+h(x,2)-h(x,3))]
\end{aligned}$$

If $h(x,1)-h(x,2)>1$, define the hypothesis $\overline{h}\in\mathcal{H}$ by

$$\overline{h}(x,y) = \begin{cases} h(x,1)-\frac{h(x,1)-h(x,2)-1}{2}, & \text{if } y=1 \\ h(x,y) & \text{otherwise.} \end{cases}$$

By the completeness of $\mathcal{H}$ and some computation, the new hypothesis $\overline{h}$ is in $\mathcal{H}$ and satisfies that $g(\overline{h})<g(h)$. Similarly, if $h(x,2)-h(x,3)>1$, define the hypothesis $\overline{h}\in\mathcal{H}$ by

$$\overline{h}(x,y) = \begin{cases} h(x,2)-\frac{h(x,2)-h(x,3)-1}{2}, & \text{if } y=2 \\ h(x,y) & \text{otherwise.} \end{cases}$$

By the completeness of $\mathcal{H}$ and some computation, the new hypothesis $\overline{h}$ is in $\mathcal{H}$ and satisfies that $g(\overline{h})<g(h)$. Therefore,

$$\inf_{h\in\mathcal{H}} g(h) = \inf_{h\in\mathcal{H}:\, h(x,1)\geq h(x,2)\geq h(x,3)} g(h) = \inf_{h\in\mathcal{H}:\, h(x,1)\geq h(x,2)\geq h(x,3),\, h(x,1)-h(x,2)\leq 1,\, h(x_2)-h(x,3)\leq 1} g(h)$$

When $h(x,1)\geq h(x,2)\geq h(x,3)$, $h(x,1)-h(x,2)\leq 1$ and $h(x_2)-h(x,3)\leq 1$, $g(h)$ can be written as

$$\begin{aligned}
g(h) &= \left(\frac{1}{2}-\epsilon\right)[(1+h(x,2)-h(x,1))+\max\{0,1+h(x,3)-h(x,1)\}] \\
&+ \frac{1}{3}[(1+h(x,1)-h(x,2))+(1+h(x,3)-h(x,2))] \\
&+ \left(\frac{1}{6}+\epsilon\right)[(1+h(x,1)-h(x,3))+(1+h(x,2)-h(x,3))]
\end{aligned}$$

If $h(x,1) - h(x,3) > 1$, define the hypothesis $\overline{h} \in \mathcal{H}$ by

$$\overline{h}(x,y) = \begin{cases} h(x,1) - \frac{h(x,1)-h(x,3)-1}{2}, & \text{if } y = 1 \\ h(x,y) & \text{otherwise.} \end{cases}$$

By the completeness of $\mathcal{H}$ and some computation using the fact that $0 < \epsilon < \frac{1}{6}$, the new hypothesis $\overline{h}$ is in $\mathcal{H}$ and satisfies that $g(\overline{h}) < g(h)$. Therefore,

$$\begin{aligned} \inf_{h\in\mathcal{H}} g(h) &= \inf_{h\in\mathcal{H}: h(x,1)\geq h(x,2)\geq h(x,3),\, h(x,1)-h(x,2)\leq 1,\, h(x_2)-h(x,3)\leq 1,\, h(x,1)-h(x,3)\leq 1} g(h) \\ &= \inf_{h\in\mathcal{H}: h(x,1)\geq h(x,2)\geq h(x,3),\, h(x,1)-h(x,2)\leq 1,\, h(x_2)-h(x,3)\leq 1,\, h(x,1)-h(x,3)\leq 1} \left(3\epsilon - \frac{1}{2}\right)(h(x,1) - h(x,3)) + 2 \\ &= 3\epsilon + \frac{3}{2} \end{aligned}$$

Thus, we obtain for any $h \in \mathcal{H}$,

$$\mathcal{R}_{\Phi^{\text{sum}}_{\text{hinge}}}(h) = \mathcal{C}_{\Phi^{\text{sum}}_{\text{hinge}}}(h,x) \geq g(h) \geq 3\epsilon + \frac{3}{2} = \mathcal{C}_{\Phi^{\text{sum}}_{\text{hinge}}}(h_0,x)$$

Therefore,

$$\mathcal{R}^*_{\Phi^{\text{sum}}_{\text{hinge}},\mathcal{H}} = \mathcal{C}^*_{\Phi^{\text{sum}}_{\text{hinge}},\mathcal{H}}(x) = \mathcal{R}_{\Phi^{\text{sum}}_{\text{hinge}}}(h_0) = 3\epsilon + \frac{3}{2}.$$

If (10) holds for some non-decreasing function f, then, we obtain for any $h \in \mathcal{H}$,

$$\mathcal{R}_{\ell_{0-1}}(h) - \frac{1}{2} - \epsilon \leq f\left(\mathcal{R}_{\Phi^{\text{sum}}_{\text{hinge}}}(h) - \mathcal{R}_{\Phi^{\text{sum}}_{\text{hinge}}}(h_0)\right).$$

Let $h = h_0$, then $f(0) \geq 1/6 - \epsilon$. Since f is non-decreasing, for any $t \geq 0$ and $0 < \epsilon < \frac{1}{6}$, $f(t) \geq 1/6 - \epsilon$. Let $\epsilon \to 0$, we obtain that f is lower bounded by $\frac{1}{6}$. □

Theorem 22 ($\mathcal{H}$**-consistency bound of** $\Phi^{\text{sum}}_{\text{sq-hinge}}$)**.** *Suppose that $\mathcal{H}$ is symmetric and complete. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \left(\mathcal{R}_{\Phi^{\text{sum}}_{\text{sq-hinge}}}(h) - \mathcal{R}^*_{\Phi^{\text{sum}}_{\text{sq-hinge}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{sum}}_{\text{sq-hinge}},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}. \tag{37}$$

Proof. For the sum squared hinge loss $\Phi^{\text{sum}}_{\text{sq-hinge}}$, by (36), the conditional $\Phi^{\text{sum}}_{\text{sq-hinge}}$-risk can be expressed as follows:

$$\begin{aligned} &\mathcal{C}_{\Phi^{\text{sum}}_{\text{sq-hinge}}}(h,x) \\ &= \sum_{y\in\mathcal{Y}} p(x,y) \sum_{y'\neq y} \max\{0, 1 + h(x,y') - h(x,y)\}^2 \\ &= p(x,y_{\max}) \sum_{y'\neq y_{\max}} \max\{0, 1 + h(x,y') - h(x,y)\}^2 + \sum_{y\neq y_{\max}} p(x,y) \sum_{y'\neq y} \max\{0, 1 + h(x,y') - h(x,y)\}^2 \\ &= p(x,y_{\max}) \sum_{y'\neq y_{\max}} \max\{0, 1 + h(x,y') - h(x,y_{\max})\}^2 + \sum_{y\neq y_{\max}} p(x,y) \max\{0, 1 + h(x,y_{\max}) - h(x,y)\}^2 \\ &+ \sum_{y\neq y_{\max}} p(x,y) \sum_{y'\notin\{y_{\max},y\}} \max\{0, 1 + h(x,y') - h(x,y)\}^2 \end{aligned}$$

For any $h \in \mathcal{H}$, define the hypothesis $\overline{h}_\lambda \in \mathcal{H}$ by

$$\overline{h}_\lambda(x,y) = \begin{cases} h(x,y) & \text{if } y \neq y_{\max} \\ \lambda & \text{if } y = y_{\max} \end{cases}$$

for any $\lambda \in \mathbb{R}$. By the completeness of $\mathcal{H}$, the new hypothesis $\overline{h}_\lambda$ is in $\mathcal{H}$. Therefore, the minimal conditional $\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}}$-risk satisfies that for any $\lambda \in \mathbb{R}$,

$$
\begin{aligned}
&\mathcal{C}^*_{\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}},\mathcal{H}}(x) \leq \mathcal{C}_{\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}}}(\overline{h}_\lambda, x)\\
&= p(x, y_{\max}) \sum_{y' \neq y_{\max}} \max\{0, 1 + h(x, y') - \lambda\}^2 + \sum_{y \neq y_{\max}} p(x, y) \max\{0, 1 + \lambda - h(x, y)\}^2\\
&+ \sum_{y \neq y_{\max}} p(x, y) \sum_{y' \notin \{y_{\max}, y\}} \max\{0, 1 + h(x, y') - h(x, y)\}^2\\
&= \sum_{y \neq y_{\max}} \Big[p(x, y_{\max}) \max\{0, 1 + h(x, y) - \lambda\}^2 + p(x, y) \max\{0, 1 + \lambda - h(x, y)\}^2\Big]\\
&+ \sum_{y \neq y_{\max}} p(x, y) \sum_{y' \notin \{y_{\max}, y\}} \max\{0, 1 + h(x, y') - h(x, y)\}^2.
\end{aligned}
$$

Let $h \in \mathcal{H}$ be a hypothesis such that $\mathsf{h}(x) \neq y_{\max}$. By the definition and using the fact that $\mathsf{H}(x) = \mathcal{Y}$ when $\mathcal{H}$ is symmetric, we obtain

$$
\begin{aligned}
&\Delta\mathcal{C}_{\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}},\mathcal{H}}(h, x) = \mathcal{C}_{\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}}}(h, x) - \mathcal{C}^*_{\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}},\mathcal{H}}(x)\\
&\geq \mathcal{C}_{\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}}}(h, x) - \mathcal{C}_{\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}}}(\overline{h}_\lambda, x)\\
&\geq p(x, y_{\max}) \max\{0, 1 + h(x, \mathsf{h}(x)) - h(x, y_{\max})\}^2 + p(x, \mathsf{h}(x)) \max\{0, 1 + h(x, y_{\max}) - h(x, \mathsf{h}(x))\}^2\\
&- \frac{4p(x, y_{\max})p(x, \mathsf{h}(x))}{p(x, y_{\max} + p(x, \mathsf{h}(x))} && \text{(taking supremum with respect to } \lambda)\\
&\geq p(x, y_{\max}) + p(x, \mathsf{h}(x)) - \frac{4p(x, y_{\max})p(x, \mathsf{h}(x))}{p(x, y_{\max} + p(x, \mathsf{h}(x))} && (h(x, \mathsf{h}(x)) - h(x, y_{\max}) \geq 0)\\
&= \frac{(p(x, y_{\max}) - p(x, \mathsf{h}(x)))^2}{p(x, y_{\max} + p(x, \mathsf{h}(x))}\\
&\geq \left(\max_{y \in \mathcal{Y}} p(x, y) - p(x, \mathsf{h}(x))\right)^2 && (0 \leq p(x, y_{\max}) + p(x, \mathsf{h}(x)) \leq 1)\\
&= (\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h, x))^2 && \text{(by Lemma 3 and } \mathsf{H}(x) = \mathcal{Y})\\
&\geq \left([\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h, x)]_\epsilon\right)^2 && ([t]_\epsilon \leq t)
\end{aligned}
$$

for any $\epsilon \geq 0$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric and complete hypothesis set, $\epsilon = 0$ and $\Psi(t) = t^2$ in Theorem 4, or, equivalently, $\Gamma(t) = \sqrt{t}$ in Theorem 5, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$
\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \left(\mathcal{R}_{\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}}}(h) - \mathcal{R}^*_{\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}},\mathcal{H}} + \mathcal{M}_{\Phi^{\mathrm{sum}}_{\mathrm{sq-hinge}},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}.
$$

□

Theorem 23 ($\mathcal{H}$**-consistency bound of** $\Phi^{\mathrm{sum}}_{\mathrm{exp}}$)**.** *Suppose that $\mathcal{H}$ is symmetric and complete. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution,*

$$
\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \sqrt{2}\left(\mathcal{R}_{\Phi^{\mathrm{sum}}_{\mathrm{exp}}}(h) - \mathcal{R}^*_{\Phi^{\mathrm{sum}}_{\mathrm{exp}},\mathcal{H}} + \mathcal{M}_{\Phi^{\mathrm{sum}}_{\mathrm{exp}},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}. \tag{38}
$$

Proof. For the sum exponential loss $\Phi^{\mathrm{sum}}_{\mathrm{exp}}$, by (36), the conditional $\Phi^{\mathrm{sum}}_{\mathrm{exp}}$-risk can be expressed as follows:

$$
\begin{aligned}
&\mathcal{C}_{\Phi^{\mathrm{sum}}_{\mathrm{exp}}}(h, x) = \sum_{y \in \mathcal{Y}} p(x, y) \sum_{y' \neq y} \exp(h(x, y') - h(x, y))\\
&= p(x, y_{\max}) \sum_{y' \neq y_{\max}} \exp(h(x, y') - h(x, y_{\max})) + \sum_{y \neq y_{\max}} p(x, y) \sum_{y' \neq y} \exp(h(x, y') - h(x, y))\\
&= p(x, y_{\max}) \sum_{y' \neq y_{\max}} \exp(h(x, y') - h(x, y_{\max})) + \sum_{y \neq y_{\max}} p(x, y) \exp(h(x, y_{\max}) - h(x, y))\\
&+ \sum_{y \neq y_{\max}} p(x, y) \sum_{y' \notin \{y_{\max}, y\}} \exp(h(x, y') - h(x, y))
\end{aligned}
$$

For any $h \in \mathcal{H}$, define the hypothesis $\overline{h}_\lambda \in \mathcal{H}$ by

$$\overline{h}_\lambda(x,y) = \begin{cases} h(x,y) & \text{if } y \neq y_{\max} \\ \lambda & \text{if } y = y_{\max} \end{cases}$$

for any $\lambda \in \mathbb{R}$. By the completeness of $\mathcal{H}$, the new hypothesis $\overline{h}_\lambda$ is in $\mathcal{H}$. Therefore, the minimal conditional $\Phi_{\exp}^{\sum}$-risk satisfies that for any $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathcal{C}^*_{\Phi_{\exp}^{\sum},\mathcal{H}}(x) &\leq \mathcal{C}_{\Phi_{\exp}^{\sum}}(\overline{h}_\lambda, x) \\
&= p(x,y_{\max})e^{-\lambda} \sum_{y' \neq y_{\max}} e^{h(x,y')} + \sum_{y \neq y_{\max}} p(x,y)\exp(\lambda - h(x,y)) \\
&+ \sum_{y \neq y_{\max}} p(x,y) \sum_{y' \notin \{y_{\max},y\}} \exp(h(x,y') - h(x,y)) \\
&= \sum_{y \neq y_{\max}} \left[p(x,y_{\max})e^{h(x,y)}e^{-\lambda} + p(x,y)e^{-h(x,y)}e^{\lambda}\right] \\
&+ \sum_{y \neq y_{\max}} p(x,y) \sum_{y' \notin \{y_{\max},y\}} \exp(h(x,y') - h(x,y)).
\end{aligned}$$

Let $h \in \mathcal{H}$ be a hypothesis such that $\mathsf{h}(x) \neq y_{\max}$. By the definition and using the fact that $\mathsf{H}(x) = \mathcal{Y}$ when $\mathcal{H}$ is symmetric, we obtain

$$\begin{aligned}
&\Delta\mathcal{C}_{\Phi_{\exp}^{\sum},\mathcal{H}}(h,x) = \mathcal{C}_{\Phi_{\exp}^{\sum}}(h,x) - \mathcal{C}^*_{\Phi_{\exp}^{\sum},\mathcal{H}}(x) \\
&\geq \mathcal{C}_{\Phi_{\exp}^{\sum}}(h,x) - \mathcal{C}_{\Phi_{\exp}^{\sum}}(\overline{h}_\lambda, x) \\
&\geq \left(\sqrt{p(x,y_{\max})e^{h(x,\mathsf{h}(x))}e^{-h(x,y_{\max})}} - \sqrt{p(x,\mathsf{h}(x))e^{-h(x,\mathsf{h}(x))}e^{h(x,y_{\max})}}\right)^2 \\
&\qquad\qquad \text{(taking supremum with respect to } \lambda) \\
&\geq \left(\sqrt{p(x,y_{\max})} - \sqrt{p(x,\mathsf{h}(x))}\right)^2 \qquad (h(x,\mathsf{h}(x)) \geq h(x,y_{\max}) \text{ and } p(x,\mathsf{h}(x)) \leq p(x,y_{\max})) \\
&= \left(\frac{p(x,y_{\max}) - p(x,\mathsf{h}(x))}{\sqrt{p(x,\mathsf{h}(x))} + \sqrt{p(x,y_{\max})}}\right)^2 \\
&\geq \frac{1}{2}\left(\max_{y \in \mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\right)^2 \qquad (0 \leq p(x,y_{\max}) + p(x,\mathsf{h}(x)) \leq 1) \\
&= \frac{1}{2}\left(\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)\right)^2 \qquad \text{(by Lemma 3 and } \mathsf{H}(x) = \mathcal{Y}) \\
&\geq \frac{1}{2}\left([\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)]_\epsilon\right)^2 \qquad ([t]_\epsilon \leq t)
\end{aligned}$$

for any $\epsilon \geq 0$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric and complete hypothesis set, $\epsilon = 0$ and $\Psi(t) = \frac{t^2}{2}$ in Theorem 4, or, equivalently, $\Gamma(t) = \sqrt{2t}$ in Theorem 5, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \sqrt{2}\left(\mathcal{R}_{\Phi_{\exp}^{\sum}}(h) - \mathcal{R}^*_{\Phi_{\exp}^{\sum},\mathcal{H}} + \mathcal{M}_{\Phi_{\exp}^{\sum},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}.$$

□

Theorem 24 ($\mathcal{H}$**-consistency bound of** $\Phi_\rho^{\sum}$)**.** *Suppose that $\mathcal{H}$ is symmetric and satisfies that for any $x \in \mathcal{X}$, there exists a hypothesis $h \in \mathcal{H}$ such that $|h(x,i) - h(x,j)| \geq \rho$ for any $i \neq j \in \mathcal{Y}$. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi_\rho^{\sum}}(h) - \mathcal{R}^*_{\Phi_\rho^{\sum},\mathcal{H}} + \mathcal{M}_{\Phi_\rho^{\sum},\mathcal{H}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}. \tag{39}$$

Proof. For any $x \in \mathcal{X}$, we define $p_{[1]}(x), p_{[2]}(x), \ldots, p_{[c]}(x)$ by sorting the probabilities $\{p(x,y) : y \in \mathcal{Y}\}$ in increasing order. Similarly, for any $x \in \mathcal{X}$ and $h \in \mathcal{H}$, we define

$h(x,\{1\}_x), h(x,\{2\}_x), \ldots, h(x,\{c\}_x)$ by sorting the scores $\{h(x,y) : y \in \mathcal{Y}\}$ in increasing order. In particular, we have

$$h(x,\{1\}_x) = \min_{y\in\mathcal{Y}} h(x,y), \quad h(x,\{c\}_x) = \max_{y\in\mathcal{Y}} h(x,y), \quad h(x,\{i\}_x) \le h\big(x,\{j\}_j\big),\ \forall i \le j.$$

If there is a tie for the maximum, we pick the label with the highest index under the natural ordering of labels, i.e. $\{c\}_x = \mathsf{h}(x)$. By the definition, the conditional Φ_ρ^{sum}-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\Phi_\rho^{\mathrm{sum}}}(h,x) &= \sum_{y\in\mathcal{Y}} p(x,y) \sum_{y'\neq y} \Phi_\rho(h(x,y) - h(x,y')) \\
&= \sum_{i=1}^{c} p(x,\{i\}_x)\left[\sum_{j=1}^{i-1} \Phi_\rho(h(x,\{i\}_x) - h(x,\{j\}_x)) + \sum_{j=i+1}^{c} \Phi_\rho(h(x,\{i\}_x) - h(x,\{j\}_x))\right] \\
&= \sum_{i=1}^{c} p(x,\{i\}_x)\left[\sum_{j=1}^{i-1} \Phi_\rho(h(x,\{i\}_x) - h(x,\{j\}_x)) + c - i\right] && (\Phi_\rho(t) = 1 \text{ for } t \le 0)
\end{aligned}$$

By the assumption, there exists a hypotheses $h \in \mathcal{H}$ such that $|h(x,i) - h(x,j)| \ge \rho$ for any $i \neq j \in \mathcal{Y}$. Since $\mathcal{H}$ is symmetric, we can always choose h^* among these hypotheses such that h^* and $p(x)$ induce the same ordering of the labels, i.e. $p(x,\{k\}_x) = p_{[k]}(x)$ for any $k \in \mathcal{Y}$. Then, we have

$$\begin{aligned}
\mathcal{C}^*_{\Phi_\rho^{\mathrm{sum}},\mathcal{H}}(x) &\le \mathcal{C}_{\Phi_\rho^{\mathrm{sum}}}(h^*,x) \\
&= \sum_{i=1}^{c} p(x,\{i\}_x)\left[\sum_{j=1}^{i-1} \Phi_\rho(h^*(x,\{i\}_x) - h^*(x,\{j\}_x)) + c - i\right] \\
&= \sum_{i=1}^{c} p(x,\{i\}_x)(c-i) && (|h^*(x,i) - h^*(x,j)| \ge \rho \text{ for any } i \neq j \text{ and } \Phi_\rho(t) = 0,\ \forall t \ge \rho) \\
&= \sum_{i=1}^{c} p_{[i]}(x)(c-i) && (h^* \text{ and } p(x) \text{ induce the same ordering of the labels}) \\
&= c - \sum_{i=1}^{c} i\, p_{[i]}(x) && (\textstyle\sum_{i=1}^{c} p_{[i]}(x) = 1)
\end{aligned}$$

By the definition and using the fact that $\mathsf{H}(x) = \mathcal{Y}$ when $\mathcal{H}$ is symmetric, we obtain

$$\begin{aligned}
&\Delta\mathcal{C}_{\Phi_\rho^{\mathrm{sum}},\mathcal{H}}(h,x) \\
&= \mathcal{C}_{\Phi_\rho^{\mathrm{sum}}}(h,x) - \mathcal{C}^*_{\Phi_\rho^{\mathrm{sum}},\mathcal{H}}(x) \\
&= \sum_{i=1}^{c} p(x,\{i\}_x)\left[\sum_{j=1}^{i-1} \Phi_\rho(h(x,\{i\}_x) - h(x,\{j\}_x)) + c - i\right] - \left(c - \sum_{i=1}^{c} i\, p_{[i]}(x)\right) \\
&\ge \sum_{i=1}^{c} p(x,\{i\}_x)(c-i) - \left(c - \sum_{i=1}^{c} i\, p_{[i]}(x)\right) && (\Phi_\rho \ge 0) \\
&= \sum_{i=1}^{c} i\, p_{[i]}(x) - \sum_{i=1}^{c} i\, p(x,\{i\}_x) && (\textstyle\sum_{i=1}^{c} p(x,\{i\}) = 1) \\
&= \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x)) + \begin{bmatrix} c-1 \\ c-1 \\ c-2 \\ \vdots \\ 1 \end{bmatrix} \cdot \begin{bmatrix} p_{[c]}(x) \\ p_{[c-1]}(x) \\ p_{[c-2]}(x) \\ \vdots \\ p_{[1]}(x) \end{bmatrix} - \begin{bmatrix} c-1 \\ c-1 \\ c-2 \\ \vdots \\ 1 \end{bmatrix} \cdot \begin{bmatrix} p(x,\{c\}_x) \\ p(x,\{c-1\}_x) \\ p(x,\{c-2\}_x) \\ \vdots \\ p(x,\{1\}_x) \end{bmatrix} \\
&&& (p_{[c]}(x) = \max_{y\in\mathcal{Y}} p(x,y) \text{ and } \{c\}_x = \mathsf{h}(x)) \\
&\ge \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x)) && (\text{by Lemma 21}) \\
&= \Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) && (\text{by Lemma 3}) \\
&\ge \big[\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)\big]_\epsilon && ([t]_\epsilon \le t)
\end{aligned}$$

for any $\epsilon \ge 0$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric hypothesis set, $\epsilon = 0$ and $\Psi(t) = t$ in Theorem 12, or, equivalently, $\Gamma(t) = t$ in Theorem 5, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \le \mathcal{R}_{\Phi_\rho^{\mathrm{sum}}}(h) - \mathcal{R}^*_{\Phi_\rho^{\mathrm{sum}},\mathcal{H}} + \mathcal{M}_{\Phi_\rho^{\mathrm{sum}},\mathcal{H}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}.$$

□

K Proof of $\mathcal{H}$-consistency bounds for constrained losses Φ^{cstnd}

Recall that $\mathsf{h}(x)$ and $y_{\max}$ are defined by $\mathsf{h}(x) = \operatorname{argmax}_{y\in\mathcal{Y}} h(x,y)$ and $y_{\max} = \operatorname{argmax}_{y\in\mathcal{Y}} p(x,y)$. If there is a tie, we pick the label with the highest index under the natural ordering of labels. The main idea of the proofs in this section is to leverage the constraint condition of Lee et al. (2004) that the scores sum to zero, and appropriately choose a hypothesis $\overline{h}$ that differs from h only for its scores for $\mathsf{h}(x)$ and $y_{\max}$. Then, we can upper bound the minimal conditional risk by the conditional risk of $\overline{h}$ without requiring complicated computation of the minimal conditional risk. By the definition, the conditional Φ^{cstnd}-risk can be expressed as follows:

$$\begin{aligned}\mathcal{C}_{\Phi^{\text{cstnd}}}(h,x) &= \sum_{y\in\mathcal{Y}} p(x,y) \sum_{y'\neq y} \Phi(-h(x,y')) \\ &= \sum_{y\in\mathcal{Y}} \Phi(-h(x,y)) \sum_{y'\neq y} p(x,y') \\ &= \sum_{y\in\mathcal{Y}} (1-p(x,y))\Phi(-h(x,y))\end{aligned} \tag{40}$$

Theorem 25 ($\mathcal{H}$**-consistency bound of** $\Phi^{\text{cstnd}}_{\text{hinge}}$)**.** *Suppose that $\mathcal{H}$ is symmetric and complete. Then, for any hypothesis $h\in\mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi^{\text{cstnd}}_{\text{hinge}}}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_{\text{hinge}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_{\text{hinge}},\mathcal{H}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}. \tag{41}$$

Proof. For the constrained hinge loss $\Phi^{\text{cstnd}}_{\text{hinge}}$, by (40), the conditional $\Phi^{\text{cstnd}}_{\text{hinge}}$-risk can be expressed as follows:

$$\begin{aligned}&\mathcal{C}_{\Phi^{\text{cstnd}}_{\text{hinge}}}(h,x) = \sum_{y\in\mathcal{Y}} (1-p(x,y))\max\{0,1+h(x,y)\} \\ &= \sum_{y\in\{y_{\max},\mathsf{h}(x)\}} (1-p(x,y))\max\{0,1+h(x,y)\} + \sum_{y\notin\{y_{\max},\mathsf{h}(x)\}} (1-p(x,y))\max\{0,1+h(x,y)\}\end{aligned}$$

Let $h\in\mathcal{H}$ be a hypothesis such that $\mathsf{h}(x)\neq y_{\max}$. For any $x\in\mathcal{X}$, if $h(x,y_{\max})\leq -1$, define the hypothesis $\overline{h}\in\mathcal{H}$ by

$$\overline{h}(x,y) = \begin{cases} h(x,y) & \text{if } y\notin\{y_{\max},\mathsf{h}(x)\} \\ h(x,y_{\max}) & \text{if } y=\mathsf{h}(x) \\ h(x,\mathsf{h}(x)) & \text{if } y=y_{\max}. \end{cases}$$

Otherwise, define the hypothesis $\overline{h}\in\mathcal{H}$ by

$$\overline{h}(x,y) = \begin{cases} h(x,y) & \text{if } y\notin\{y_{\max},\mathsf{h}(x)\} \\ -1 & \text{if } y=\mathsf{h}(x) \\ h(x,y_{\max}) + h(x,\mathsf{h}(x)) + 1 & \text{if } y=y_{\max}. \end{cases}$$

By the completeness of $\mathcal{H}$, the new hypothesis $\overline{h}$ is in $\mathcal{H}$ and satisfies that $\sum_{y\in\mathcal{Y}}\overline{h}(x,y)=0$. Since $\sum_{y\in\mathcal{Y}} h(x,y)=0$, there must be non-negative scores. By definition of $\mathsf{h}(x)$ as a maximizer, we must thus have $h(x,\mathsf{h}(x))\geq 0$. Therefore, the minimal conditional $\Phi^{\text{cstnd}}_{\text{hinge}}$-risk satisfies:

$$\begin{aligned}&\mathcal{C}^*_{\Phi^{\text{cstnd}}_{\text{hinge}},\mathcal{H}}(x) \leq \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{hinge}}}(\overline{h},x) \\ &= \begin{cases} (1-p(x,y_{\max}))(1+h(x,\mathsf{h}(x))) + \sum_{y\notin\{y_{\max},\mathsf{h}(x)\}}(1-p(x,y))(1+h(x,y)) & \text{if } h(x,y_{\max})\leq -1 \\ (1-p(x,y_{\max}))(h(x,y_{\max})+h(x,\mathsf{h}(x))+2) + \sum_{y\notin\{y_{\max},\mathsf{h}(x)\}}(1-p(x,y))(1+h(x,y)) & \text{otherwise.}\end{cases}\end{aligned}$$

By the definition and using the fact that $\mathsf{H}(x)=\mathcal{Y}$ when $\mathcal{H}$ is symmetric, we obtain

$$\begin{aligned}\Delta\mathcal{C}_{\Phi^{\text{cstnd}}_{\text{hinge}},\mathcal{H}}(h,x) &= \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{hinge}}}(h,x) - \mathcal{C}^*_{\Phi^{\text{cstnd}}_{\text{hinge}},\mathcal{H}}(x) \\ &\geq \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{hinge}}}(h,x) - \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{hinge}}}(\overline{h},x) \\ &= (1+h(x,\mathsf{h}(x)))(p(x,y_{\max}) - p(x,\mathsf{h}(x))) \\ &\geq \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x)) && (h(x,\mathsf{h}(x))\geq 0) \\ &= \Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) && (\text{by Lemma 3 and } \mathsf{H}(x)=\mathcal{Y}) \\ &\geq \left[\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)\right]_\epsilon && (\left[t\right]_\epsilon \leq t)\end{aligned}$$

for any $\epsilon \geq 0$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric and complete hypothesis set, $\epsilon = 0$ and $\Psi(t) = t$ in Theorem 4, or, equivalently, $\Gamma(t) = t$ in Theorem 5, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi^{\text{cstnd}}_{\text{hinge}}}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_{\text{hinge}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_{\text{hinge}},\mathcal{H}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}.$$

□

Theorem 26 (**$\mathcal{H}$-consistency bound of $\Phi^{\text{cstnd}}_{\text{sq-hinge}}$**). *Suppose that $\mathcal{H}$ is symmetric and complete. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \left(\mathcal{R}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}}}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_{\text{sq-hinge}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}. \tag{42}$$

Proof. For the constrained squared hinge loss $\Phi^{\text{cstnd}}_{\text{sq-hinge}}$, by (40), the conditional $\Phi^{\text{cstnd}}_{\text{sq-hinge}}$-risk can be expressed as follows:

$$\begin{aligned}
&\mathcal{C}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}}}(h,x) = \sum_{y\in\mathcal{Y}}(1-p(x,y))\max\{0,1+h(x,y)\}^2\\
&= \sum_{y\in\{y_{\max},\mathsf{h}(x)\}}(1-p(x,y))\max\{0,1+h(x,y)\}^2 + \sum_{y\notin\{y_{\max},\mathsf{h}(x)\}}(1-p(x,y))\max\{0,1+h(x,y)\}^2
\end{aligned}$$

Let $h \in \mathcal{H}$ be a hypothesis such that $\mathsf{h}(x) \neq y_{\max}$. For any $x \in \mathcal{X}$, if $h(x, y_{\max}) \leq -1$, define the hypothesis $\overline{h} \in \mathcal{H}$ by

$$\overline{h}(x,y) = \begin{cases} h(x,y) & \text{if } y \notin \{y_{\max}, \mathsf{h}(x)\}\\ h(x,y_{\max}) & \text{if } y = \mathsf{h}(x)\\ h(x,\mathsf{h}(x)) & \text{if } y = y_{\max}. \end{cases}$$

Otherwise, define the hypothesis $\overline{h} \in \mathcal{H}$ by

$$\overline{h}(x,y) = \begin{cases} h(x,y) & \text{if } y \notin \{y_{\max}, \mathsf{h}(x)\}\\ \frac{1-p(x,y_{\max})}{2-p(x,y_{\max})-p(x,\mathsf{h}(x))}(2+h(x,y_{\max})+h(x,\mathsf{h}(x))) - 1 & \text{if } y = \mathsf{h}(x)\\ \frac{1-p(x,\mathsf{h}(x))}{2-p(x,y_{\max})-p(x,\mathsf{h}(x))}(2+h(x,y_{\max})+h(x,\mathsf{h}(x))) - 1 & \text{if } y = y_{\max}. \end{cases}$$

By the completeness of $\mathcal{H}$, the new hypothesis $\overline{h}$ is in $\mathcal{H}$ and satisfies that $\sum_{y\in\mathcal{Y}}\overline{h}(x,y) = 0$. Since $\sum_{y\in\mathcal{Y}} h(x,y) = 0$, there must be non-negative scores. By definition of $\mathsf{h}(x)$ as a maximizer, we must thus have $h(x,\mathsf{h}(x)) \geq 0$. Therefore, the minimal conditional $\Phi^{\text{cstnd}}_{\text{sq-hinge}}$-risk satisfies:

$$\begin{aligned}
&\mathcal{C}^*_{\Phi^{\text{cstnd}}_{\text{sq-hinge}},\mathcal{H}}(x) \leq \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}}}(\overline{h},x)\\
&= \begin{cases} (1-p(x,y_{\max}))(1+h(x,\mathsf{h}(x)))^2 + \sum_{y\notin\{y_{\max},\mathsf{h}(x)\}}(1-p(x,y))(1+h(x,y)) & \text{if } h(x,y_{\max}) \leq -1\\ \frac{(1-p(x,y_{\max}))(1-p(x,\mathsf{h}(x)))(2+h(x,y_{\max})+h(x,\mathsf{h}(x)))^2}{2-p(x,y_{\max})-p(x,y)} + \sum_{y\notin\{y_{\max},\mathsf{h}(x)\}}(1-p(x,y))(1+h(x,y)) & \text{otherwise.} \end{cases}
\end{aligned}$$

By the definition and using the fact that $\mathsf{H}(x) = \mathcal{Y}$ when $\mathcal{H}$ is symmetric, we obtain

$$\begin{aligned}
&\Delta\mathcal{C}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}},\mathcal{H}}(h,x) = \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}}}(h,x) - \mathcal{C}^*_{\Phi^{\text{cstnd}}_{\text{sq-hinge}},\mathcal{H}}(x)\\
&\geq \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}}}(h,x) - \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}}}(\overline{h},x)\\
&= \begin{cases} (1+h(x,\mathsf{h}(x)))^2(p(x,y_{\max}) - p(x,\mathsf{h}(x))) & \text{if } h(x,y_{\max}) \leq -1\\ g(1-p(x,y_{\max}), 1-p(x,\mathsf{h}(x)), 1+h(x,y_{\max}), 1+h(x,\mathsf{h}(x))) & \text{otherwise} \end{cases}\\
&\geq (1+h(x,\mathsf{h}(x)))^2\left(\max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\right)^2 && \text{(property of } g \text{ and } p(x,y_{\max}) \leq 1)\\
&\geq \left(\max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\right)^2 && (h(x,\mathsf{h}(x)) \geq 0)\\
&= \left(\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)\right)^2 && \text{(by Lemma 3 and } \mathsf{H}(x) = \mathcal{Y})\\
&\geq \left([\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)]_\epsilon\right)^2 && ([t]_\epsilon \leq t)
\end{aligned}$$

for any $\epsilon \geq 0$, where $g(x, y, \alpha, \beta) = \frac{x^2\alpha^2 + y^2\beta^2 - 2xy\alpha\beta}{x+y} \geq \beta^2(x-y)^2$ when $0 \leq x \leq y \leq 1$, $x + y \geq 1$ and $1 \leq \alpha \leq \beta$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric and complete hypothesis set, $\epsilon = 0$ and $\Psi(t) = t^2$ in Theorem 4, or, equivalently, $\Gamma(t) = \sqrt{t}$ in Theorem 5, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \left(\mathcal{R}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}}}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_{\text{sq-hinge}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_{\text{sq-hinge}},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}.$$

□

Theorem 27 ($\mathcal{H}$**-consistency bound of** $\Phi^{\text{cstnd}}_{\text{exp}}$)**.** *Suppose that $\mathcal{H}$ is symmetric and complete. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \sqrt{2}\left(\mathcal{R}_{\Phi^{\text{cstnd}}_{\text{exp}}}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_{\text{exp}},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_{\text{exp}},\mathcal{H}}\right)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}. \tag{43}$$

Proof. For the constrained exponential loss $\Phi^{\text{cstnd}}_{\text{exp}}$, by (40), the conditional $\Phi^{\text{cstnd}}_{\text{exp}}$-risk can be expressed as follows:

$$\begin{aligned}
&\mathcal{C}_{\Phi^{\text{cstnd}}_{\text{exp}}}(h, x) = \sum_{y \in \mathcal{Y}} (1 - p(x, y)) \exp(h(x, y)) \\
&= \sum_{y \in \{y_{\max}, \mathsf{h}(x)\}} (1 - p(x, y)) \exp(h(x, y)) + \sum_{y \notin \{y_{\max}, \mathsf{h}(x)\}} \exp(h(x, y))
\end{aligned}$$

Let $h \in \mathcal{H}$ be a hypothesis such that $\mathsf{h}(x) \neq y_{\max}$. For any $x \in \mathcal{X}$, define the hypothesis $\overline{h}_\mu \in \mathcal{H}$ by

$$\overline{h}_\mu(x, y) = \begin{cases} h(x, y) & \text{if } y \notin \{y_{\max}, \mathsf{h}(x)\} \\ h(x, y_{\max}) + \mu & \text{if } y = \mathsf{h}(x) \\ h(x, \mathsf{h}(x)) - \mu & \text{if } y = y_{\max} \end{cases}$$

for any $\mu \in \mathbb{R}$. By the completeness of $\mathcal{H}$, the new hypothesis $\overline{h}_\mu$ is in $\mathcal{H}$ and satisfies that $\sum_{y \in \mathcal{Y}} \overline{h}_\mu(x, y) = 0$. Since $\sum_{y \in \mathcal{Y}} h(x, y) = 0$, there must be non-negative scores. By definition of $\mathsf{h}(x)$ as a maximizer, we must thus have $h(x, \mathsf{h}(x)) \geq 0$. Therefore, the minimal conditional $\Phi^{\text{cstnd}}_{\text{exp}}$-risk satisfies that for any $\mu \in \mathbb{R}$,

$$\begin{aligned}
&\mathcal{C}^*_{\Phi^{\text{cstnd}}_{\text{exp}},\mathcal{H}}(x) \leq \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{exp}}}(\overline{h}_\mu, x) \\
&= (1 - p(x, y_{\max}))e^{h(x,\mathsf{h}(x)) - \mu} + (1 - p(x, \mathsf{h}(x)))e^{h(x,y_{\max}) + \mu} + \sum_{y \notin \{y_{\max}, \mathsf{h}(x)\}} (1 - p(x, y)) \exp(h(x, y)).
\end{aligned}$$

By the definition and using the fact that $\mathsf{H}(x) = \mathcal{Y}$ when $\mathcal{H}$ is symmetric, we obtain

$$\begin{aligned}
&\Delta\mathcal{C}_{\Phi^{\text{cstnd}}_{\text{exp}},\mathcal{H}}(h, x) = \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{exp}}}(h, x) - \mathcal{C}^*_{\Phi^{\text{cstnd}}_{\text{exp}},\mathcal{H}}(x) \\
&\geq \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{exp}}}(h, x) - \mathcal{C}_{\Phi^{\text{cstnd}}_{\text{exp}}}(\overline{h}_\mu, x) \\
&\geq \left(\sqrt{(1 - p(x, \mathsf{h}(x)))e^{h(x,\mathsf{h}(x))}} - \sqrt{(1 - p(x, y_{\max}))e^{h(x,y_{\max})}}\right)^2 \\
&\qquad\qquad \text{(taking supremum with respect to } \mu) \\
&\geq e^{h(x,\mathsf{h}(x))}\left(\sqrt{(1 - p(x, \mathsf{h}(x)))} - \sqrt{(1 - p(x, y_{\max}))}\right)^2 \\
&\qquad\qquad (e^{h(x,\mathsf{h}(x))} \geq e^{h(x,y_{\max})} \text{ and } p(x, \mathsf{h}(x)) \leq p(x, y_{\max})) \\
&\geq \left(\sqrt{(1 - p(x, \mathsf{h}(x)))} - \sqrt{(1 - p(x, y_{\max}))}\right)^2 \qquad (h(x, \mathsf{h}(x)) \geq 0) \\
&= \left(\frac{p(x, y_{\max}) - p(x, \mathsf{h}(x))}{\sqrt{(1 - p(x, \mathsf{h}(x)))} + \sqrt{(1 - p(x, y_{\max}))}}\right)^2 \\
&\geq \frac{1}{2}\left(\max_{y \in \mathcal{Y}} p(x, y) - p(x, \mathsf{h}(x))\right)^2 \qquad (0 \leq p(x, y_{\max}) + p(x, \mathsf{h}(x)) \leq 1) \\
&= \frac{1}{2}(\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h, x))^2 \qquad \text{(by Lemma 3 and } \mathsf{H}(x) = \mathcal{Y}) \\
&\geq \frac{1}{2}\left([\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h, x)]_\epsilon\right)^2 \qquad ([t]_\epsilon \leq t)
\end{aligned}$$

for any $\epsilon \geq 0$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric and complete hypothesis set, $\epsilon = 0$ and $\Psi(t) = \frac{t^2}{2}$ in Theorem 4, or, equivalently, $\Gamma(t) = \sqrt{2t}$ in Theorem 5, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \sqrt{2}\Big(\mathcal{R}_{\Phi^{\text{cstnd}}_{\exp}}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_{\exp},\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_{\exp},\mathcal{H}}\Big)^{\frac{1}{2}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}.$$

□

Theorem 28 ($\mathcal{H}$**-consistency bound of** Φ^{cstnd}_ρ)**.** *Suppose that $\mathcal{H}$ is symmetric and satisfies that for any $x \in \mathcal{X}$, there exists a hypothesis $h \in \mathcal{H}$ such that $h(x,y) \leq -\rho$ for any $y \neq y_{\max}$. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution,*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi^{\text{cstnd}}_\rho}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_\rho,\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_\rho,\mathcal{H}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}. \tag{44}$$

Proof. Since $\sum_{y\in\mathcal{Y}} h(x,y) = 0$, by definition of $\mathsf{h}(x)$ as a maximizer, we must thus have $h(x,\mathsf{h}(x)) \geq 0$. For the constrained ρ-margin loss Φ^{cstnd}_ρ, by (40), the conditional Φ^{cstnd}_ρ-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\Phi^{\text{cstnd}}_\rho}(h,x) &= \sum_{y\in\mathcal{Y}}(1-p(x,y))\min\left\{\max\left\{0, 1+\frac{h(x,y)}{\rho}\right\}, 1\right\}\\
&= \sum_{y\in\mathcal{Y}:h(x,y)\geq 0}(1-p(x,y)) + \sum_{y\in\mathcal{Y}:h(x,y)<0}(1-p(x,y))\max\left\{0, 1+\frac{h(x,y)}{\rho}\right\}\\
&\geq 1 - p(x,\mathsf{h}(x))\\
&\geq 1 - \max_{y\in\mathcal{Y}} p(x,y).
\end{aligned}$$

By the assumption, the equality can be achieved by some $h^*_\rho \in \mathcal{H}$ with the constraint $\sum_{y\in\mathcal{Y}} h(x,y) = 0$ such that $h^*_\rho(x,y) \leq -\rho$ for any $y \neq y_{\max}$ and $h^*_\rho(x,y_{\max}) = -\sum_{y'\neq y_{\max}} h^*_\rho(x,y') \geq 0$. Therefore, the minimal conditional Φ^{cstnd}_ρ-risk can be expressed as follows:

$$\mathcal{C}^*_{\Phi^{\text{cstnd}}_\rho,\mathcal{H}}(x) = 1 - \max_{y\in\mathcal{Y}} p(x,y).$$

By the definition and using the fact that $\mathsf{H}(x) = \mathcal{Y}$ when $\mathcal{H}$ is symmetric, we obtain

$$\begin{aligned}
&\Delta\mathcal{C}_{\Phi^{\text{cstnd}}_\rho,\mathcal{H}}(h,x) = \mathcal{C}_{\Phi^{\text{cstnd}}_\rho}(h,x) - \mathcal{C}^*_{\Phi^{\text{cstnd}}_\rho,\mathcal{H}}(x)\\
&= \sum_{y\in\mathcal{Y}:h(x,y)\geq 0}(1-p(x,y)) + \sum_{y\in\mathcal{Y}:h(x,y)<0}(1-p(x,y))\max\left\{0, 1+\frac{h(x,y)}{\rho}\right\} - \left(1-\max_{y\in\mathcal{Y}} p(x,y)\right)\\
&\geq 1 - p(x,\mathsf{h}(x)) - \left(1-\max_{y\in\mathcal{Y}} p(x,y)\right)\\
&= \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\\
&= \Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) && \text{(by Lemma 3 and } \mathsf{H}(x) = \mathcal{Y})\\
&\geq [\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x)]_\epsilon && ([t]_\epsilon \leq t)
\end{aligned}$$

for any $\epsilon \geq 0$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric hypothesis set, $\epsilon = 0$ and $\Psi(t) = t$ in Theorem 4, or, equivalently, $\Gamma(t) = t$ in Theorem 5, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} \leq \mathcal{R}_{\Phi^{\text{cstnd}}_\rho}(h) - \mathcal{R}^*_{\Phi^{\text{cstnd}}_\rho,\mathcal{H}} + \mathcal{M}_{\Phi^{\text{cstnd}}_\rho,\mathcal{H}} - \mathcal{M}_{\ell_{0-1},\mathcal{H}}.$$

□

L Proof of negative results for adversarial robustness

Theorem 14 (**Negative results for convex functions**)**.** *Fix $c = 2$. Suppose that Φ is convex and non-increasing, and $\mathcal{H}$ contains 0 and satisfies the condition that there exists $x \in \mathcal{X}$ such that $\mathcal{H}_\gamma(x) \neq \varnothing$.*

If for a non-decreasing function $f\colon \mathbb{R}_+ \to \mathbb{R}_+$, the following $\mathcal{H}$-consistency bound holds for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$:

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq f\Big(\mathcal{R}_{\widetilde{\ell}}(h) - \mathcal{R}^*_{\widetilde{\ell},\mathcal{H}}\Big), \tag{16}$$

then, f is lower bounded by $\frac{1}{2}$, for $\widetilde{\ell} = \widetilde{\Phi}^{\max}$, $\widetilde{\Phi}^{\mathrm{sum}}$ and $\widetilde{\Phi}^{\mathrm{cstnd}}$.

Proof. Consider the distribution that supports on a singleton domain $\{x\}$ with x satisfying that $\mathcal{H}_\gamma(x) \neq \varnothing$. When $\mathcal{H}_\gamma(x) \neq \varnothing$, $\mathsf{H}_\gamma(x)$ is also non-empty. Take $y_1 \in \mathcal{H}_\gamma(x)$ and let $y_2 \neq y_1$. We define $p(x)$ as $p(x, y_1) = p(x, y_2) = \frac{1}{2}$. Let $h_0 = 0 \in \mathcal{H}$. By Lemma 11 and the fact that $\mathcal{H}_\gamma(x) \neq \varnothing$ and $y_1 \in \mathsf{H}_\gamma(x)$, the minimal conditional ℓ_γ-risk is

$$\mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = \mathcal{C}^*_{\ell_\gamma,\mathcal{H}}(x) = 1 - \max_{y \in \mathsf{H}_\gamma(x)} p(x, y) = 1 - p(x, y_1) = \frac{1}{2}.$$

For $h = h_0$, we have

$$\mathcal{R}_{\ell_\gamma}(h_0) = \mathcal{C}_{\ell_\gamma}(h_0, x) = \sum_{y \in \mathcal{Y}} p(x, y) \sup_{x' : \|x - x'\|_p \leq \gamma} \mathbb{1}_{\rho_h(x', y) \leq 0} = 1.$$

For the adversarial max loss with non-increasing Φ, the conditional $\widetilde{\Phi}^{\max}$-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\widetilde{\Phi}^{\max}}(h, x) &= \sum_{y \in \mathcal{Y}} p(x, y) \sup_{x' : \|x - x'\|_p \leq \gamma} \Phi(\rho_h(x', y)) \\
&= \sum_{y \in \mathcal{Y}} p(x, y) \Phi\left(\inf_{x' : \|x - x'\|_p \leq \gamma} \rho_h(x', y)\right) \\
&= \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_1) - h(x', y_2))\right) + \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_2) - h(x', y_1))\right) \\
&= \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_1) - h(x', y_2))\right) + \frac{1}{2}\Phi\left(- \sup_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_1) - h(x', y_2))\right)
\end{aligned}$$

If Φ is convex and non-increasing, we obtain for any $h \in \mathcal{H}$,

$$\begin{aligned}
\mathcal{R}_{\widetilde{\Phi}^{\max}}(h) &= \mathcal{C}_{\widetilde{\Phi}^{\max}}(h, x) \\
&= \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_1) - h(x', y_2))\right) + \frac{1}{2}\Phi\left(\sup_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_1) - h(x', y_2))\right) \\
&\geq \Phi\left(\frac{1}{2}\inf_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_1) - h(x', y_2)) - \frac{1}{2}\sup_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_1) - h(x', y_2))\right) && (\Phi \text{ is convex}) \\
&\geq \Phi(0), && (\Phi \text{ is non-increasing})
\end{aligned}$$

where the equality can be achieved by h_0. Therefore,

$$\mathcal{R}^*_{\widetilde{\Phi}^{\max},\mathcal{H}} = \mathcal{C}^*_{\widetilde{\Phi}^{\max},\mathcal{H}}(x) = \mathcal{R}_{\widetilde{\Phi}^{\max}}(h_0) = \Phi(0).$$

If (16) holds for some non-decreasing function f and $\widetilde{\ell} = \widetilde{\Phi}^{\max}$, then, we obtain for any $h \in \mathcal{H}$,

$$\mathcal{R}_{\ell_\gamma}(h) - \frac{1}{2} \leq f\Big(\mathcal{R}_{\widetilde{\Phi}^{\max}}(h) - \Phi(0)\Big).$$

Let $h = h_0$, then $f(0) \geq 1/2$. Since f is non-decreasing, for any $t \geq 0$, $f(t) \geq 1/2$.

For the adversarial sum loss with non-increasing Φ, the conditional $\widetilde{\Phi}^{\mathrm{sum}}$-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\widetilde{\Phi}^{\mathrm{sum}}}(h, x) &= \sum_{y \in \mathcal{Y}} p(x, y) \sup_{x' : \|x - x'\|_p \leq \gamma} \sum_{y' \neq y} \Phi(h(x', y) - h(x', y')) \\
&= \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_1) - h(x', y_2))\right) + \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_2) - h(x', y_1))\right) \\
&= \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_1) - h(x', y_2))\right) + \frac{1}{2}\Phi\left(- \sup_{x' : \|x - x'\|_p \leq \gamma} (h(x', y_1) - h(x', y_2))\right)
\end{aligned}$$

If Φ is convex and non-increasing, we obtain for any $h \in \mathcal{H}$,

$$
\begin{aligned}
\mathcal{R}_{\widetilde{\Phi}^{\mathrm{sum}}}(h) &= \mathcal{C}_{\widetilde{\Phi}^{\mathrm{sum}}}(h, x) \\
&= \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \le \gamma} (h(x', y_1) - h(x', y_2))\right) + \frac{1}{2}\Phi\left(\sup_{x' : \|x - x'\|_p \le \gamma} (h(x', y_1) - h(x', y_2))\right) \\
&\ge \Phi\left(\frac{1}{2}\inf_{x' : \|x - x'\|_p \le \gamma} (h(x', y_1) - h(x', y_2)) - \frac{1}{2}\sup_{x' : \|x - x'\|_p \le \gamma} (h(x', y_1) - h(x', y_2))\right) \\
&\qquad\qquad (\Phi \text{ is convex}) \\
&\ge \Phi(0), \qquad\qquad (\Phi \text{ is non-increasing})
\end{aligned}
$$

where the equality can be achieved by h_0. Therefore,

$$
\mathcal{R}^*_{\widetilde{\Phi}^{\mathrm{sum}}, \mathcal{H}} = \mathcal{C}^*_{\widetilde{\Phi}^{\mathrm{sum}}, \mathcal{H}}(x) = \mathcal{R}_{\widetilde{\Phi}^{\mathrm{sum}}}(h_0) = \Phi(0).
$$

If (16) holds for some non-decreasing function f and $\widetilde{\ell} = \widetilde{\Phi}^{\mathrm{sum}}$, then, we obtain for any $h \in \mathcal{H}$,

$$
\mathcal{R}_{\ell_\gamma}(h) - \frac{1}{2} \le f\big(\mathcal{R}_{\widetilde{\Phi}^{\mathrm{sum}}}(h) - \Phi(0)\big).
$$

Let $h = h_0$, then $f(0) \ge 1/2$. Since f is non-decreasing, for any $t \ge 0$, $f(t) \ge 1/2$.

For the adversarial constrained loss with non-increasing Φ, using the fact that $h(x, y_1) + h(x, y_2) = 0$, the conditional $\widetilde{\Phi}^{\mathrm{cstnd}}$-risk can be expressed as follows:

$$
\begin{aligned}
\mathcal{C}_{\widetilde{\Phi}^{\mathrm{cstnd}}}(h, x) &= \sum_{y \in \mathcal{Y}} p(x, y) \sup_{x' : \|x - x'\|_p \le \gamma} \sum_{y' \ne y} \Phi(-h(x', y')) \\
&= \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \le \gamma} (-h(x', y_2))\right) + \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \le \gamma} (-h(x', y_1))\right) \\
&= \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \le \gamma} h(x', y_1)\right) + \frac{1}{2}\Phi\left(-\sup_{x' : \|x - x'\|_p \le \gamma} h(x', y_1)\right)
\end{aligned}
$$

If Φ is convex and non-increasing, we obtain for any $h \in \mathcal{H}$,

$$
\begin{aligned}
\mathcal{R}_{\widetilde{\Phi}^{\mathrm{cstnd}}}(h) &= \mathcal{C}_{\widetilde{\Phi}^{\mathrm{cstnd}}}(h, x) \\
&= \frac{1}{2}\Phi\left(\inf_{x' : \|x - x'\|_p \le \gamma} h(x', y_1)\right) + \frac{1}{2}\Phi\left(-\sup_{x' : \|x - x'\|_p \le \gamma} h(x', y_1)\right) \\
&\ge \Phi\left(\frac{1}{2}\inf_{x' : \|x - x'\|_p \le \gamma} h(x', y_1) - \frac{1}{2}\sup_{x' : \|x - x'\|_p \le \gamma} h(x', y_1)\right) \qquad (\Phi \text{ is convex}) \\
&\ge \Phi(0), \qquad\qquad (\Phi \text{ is non-increasing})
\end{aligned}
$$

where the equality can be achieved by h_0. Therefore,

$$
\mathcal{R}^*_{\widetilde{\Phi}^{\mathrm{cstnd}}, \mathcal{H}} = \mathcal{C}^*_{\widetilde{\Phi}^{\mathrm{cstnd}}, \mathcal{H}}(x) = \mathcal{R}_{\widetilde{\Phi}^{\mathrm{cstnd}}}(h_0) = \Phi(0).
$$

If (16) holds for some non-decreasing function f and $\widetilde{\ell} = \widetilde{\Phi}^{\mathrm{cstnd}}$, then, we obtain for any $h \in \mathcal{H}$,

$$
\mathcal{R}_{\ell_\gamma}(h) - \frac{1}{2} \le f\big(\mathcal{R}_{\widetilde{\Phi}^{\mathrm{cstnd}}}(h) - \Phi(0)\big).
$$

Let $h = h_0$, then $f(0) \ge 1/2$. Since f is non-decreasing, for any $t \ge 0$, $f(t) \ge 1/2$.

□

M Proof of $\mathcal{H}$-consistency bounds for adversarial max losses $\widetilde{\Phi}^{\mathrm{max}}$

Theorem 15 (**$\mathcal{H}$-consistency bound of $\widetilde{\Phi}_\rho^{\mathrm{max}}$**)**.** *Suppose that $\mathcal{H}$ is symmetric. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$, we have*

$$
\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma, \mathcal{H}} \le \frac{\mathcal{R}_{\widetilde{\Phi}_\rho^{\mathrm{max}}}(h) - \mathcal{R}^*_{\widetilde{\Phi}_\rho^{\mathrm{max}}, \mathcal{H}} + \mathcal{M}_{\widetilde{\Phi}_\rho^{\mathrm{max}}, \mathcal{H}}}{\min\left\{1, \frac{\inf_{x \in \{x \in \mathcal{X} : \mathcal{H}_\gamma(x) \ne \emptyset\}} \sup_{h \in \mathcal{H}_\gamma(x)} \inf_{x' : \|x - x'\|_p \le \gamma} \rho_h(x', \mathsf{h}(x))}{\rho}\right\}} - \mathcal{M}_{\ell_\gamma, \mathcal{H}}. \tag{18}
$$

Proof. By the definition, the conditional $\widetilde{\Phi}_\rho^{\max}$-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\widetilde{\Phi}_\rho^{\max}}(h,x) &= \sum_{y\in\mathcal{Y}} p(x,y) \sup_{x':\|x-x'\|_p\le\gamma} \Phi_\rho(\rho_h(x',y)) \\
&= \begin{cases} 1 - p(x,\mathsf{h}(x)) + \max\left\{0, 1 - \frac{\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} p(x,\mathsf{h}(x)) & h\in\mathcal{H}_\gamma(x) \\ 1 & \text{otherwise.} \end{cases} \\
&= \begin{cases} 1 - \min\left\{1, \frac{\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} p(x,\mathsf{h}(x)) & h\in\mathcal{H}_\gamma(x) \\ 1 & \text{otherwise.} \end{cases}
\end{aligned} \tag{45}$$

Since $\mathcal{H}$ is symmetric, for any $x\in\mathcal{X}$, either for any $y\in\mathcal{Y}$,

$$\sup_{h\in\{h\in\mathcal{H}_\gamma(x):\mathsf{h}(x)=y\}} \inf_{x':\|x-x'\|_p\le\gamma} \rho_h(x',\mathsf{h}(x)) = \sup_{h\in\mathcal{H}_\gamma(x)} \inf_{x':\|x-x'\|_p\le\gamma} \rho_h(x',\mathsf{h}(x))$$

or $\mathcal{H}_\gamma(x)=\varnothing$. When $\mathcal{H}_\gamma(x)=\varnothing$, (45) implies that $\mathcal{C}^*_{\widetilde{\Phi}_\rho^{\max},\mathcal{H}}(x)=1$. When $\mathcal{H}_\gamma(x)\neq\varnothing$,

$$\mathcal{C}^*_{\widetilde{\Phi}_\rho^{\max},\mathcal{H}}(x) = 1 - \min\left\{1, \frac{\sup_{h\in\mathcal{H}_\gamma(x)}\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y).$$

Therefore, the minimal conditional $\widetilde{\Phi}_\rho^{\max}$-risk can be expressed as follows:

$$\mathcal{C}^*_{\widetilde{\Phi}_\rho^{\max},\mathcal{H}}(x) = 1 - \min\left\{1, \frac{\sup_{h\in\mathcal{H}_\gamma(x)}\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y)\mathbb{1}_{\mathcal{H}_\gamma(x)\neq\varnothing}$$

When $\mathcal{H}_\gamma(x)=\varnothing$, $\mathcal{C}_{\widetilde{\Phi}_\rho^{\max}}(h,x)\equiv 1$, which implies that $\Delta\mathcal{C}_{\widetilde{\Phi}_\rho^{\max},\mathcal{H}}(h,x)\equiv 0$. When $\mathcal{H}_\gamma(x)\neq\varnothing$, using the fact that $\mathsf{H}_\gamma(x)=\mathcal{Y} \iff \mathcal{H}_\gamma(x)\neq\varnothing$ when $\mathcal{H}$ is symmetric,

$$\begin{aligned}
\Delta\mathcal{C}_{\widetilde{\Phi}_\rho^{\max},\mathcal{H}}(h,x) &= \min\left\{1, \frac{\sup_{h\in\mathcal{H}_\gamma(x)}\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} \max_{y\in\mathcal{Y}} p(x,y) \\
&\quad - \min\left\{1, \frac{\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} p(x,\mathsf{h}(x))\mathbb{1}_{h\in\mathcal{H}_\gamma(x)} \\
&\ge \min\left\{1, \frac{\sup_{h\in\mathcal{H}_\gamma(x)}\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} \left(\max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\mathbb{1}_{h\in\mathcal{H}_\gamma(x)}\right) \\
&= \min\left\{1, \frac{\sup_{h\in\mathcal{H}_\gamma(x)}\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} \Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x) \\
&\ge \min\left\{1, \frac{\sup_{h\in\mathcal{H}_\gamma(x)}\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} \left[\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x)\right]_\epsilon \\
&\ge \min\left\{1, \frac{\inf_{x\in\{x\in\mathcal{X}:\mathcal{H}_\gamma(x)\neq\varnothing\}}\sup_{h\in\mathcal{H}_\gamma(x)}\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} \left[\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x)\right]_\epsilon
\end{aligned}$$

for any $\epsilon\ge 0$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric hypothesis set, $\epsilon=0$ and

$$\Psi(t) = \min\left\{1, \frac{\inf_{x\in\{x\in\mathcal{X}:\mathcal{H}_\gamma(x)\neq\varnothing\}}\sup_{h\in\mathcal{H}_\gamma(x)}\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\} t$$

in Theorem 12, or, equivalently, $\Gamma(t)=\Psi^{-1}(t)$ in Theorem 13, we obtain for any hypothesis $h\in\mathcal{H}$ and any distribution,

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \le \frac{\mathcal{R}_{\widetilde{\Phi}_\rho^{\max}}(h) - \mathcal{R}^*_{\widetilde{\Phi}_\rho^{\max},\mathcal{H}} + \mathcal{M}_{\widetilde{\Phi}_\rho^{\max},\mathcal{H}}}{\min\left\{1, \frac{\inf_{x\in\{x\in\mathcal{X}:\mathcal{H}_\gamma(x)\neq\varnothing\}}\sup_{h\in\mathcal{H}_\gamma(x)}\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',\mathsf{h}(x))}{\rho}\right\}} - \mathcal{M}_{\ell_\gamma,\mathcal{H}}.$$

□

N Proof of $\mathcal{H}$-consistency bounds for adversarial sum losses $\widetilde{\Phi}^{\mathrm{sum}}$

Theorem 16 ($\mathcal{H}$**-consistency bound of** $\widetilde{\Phi}^{\mathrm{sum}}_{\rho}$**).** *Assume that $\mathcal{H}$ is symmetric and that for any $x \in \mathcal{X}$, there exists a hypothesis $h \in \mathcal{H}$ inducing the same ordering of the labels for any $x' \in \{x' : \|x - x'\|_p \leq \gamma\}$ and such that $\inf_{x' : \|x - x'\|_p \leq \gamma} |h(x', i) - h(x', j)| \geq \rho$ for any $i \neq j \in \mathcal{Y}$. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$, the following inequality holds:*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma, \mathcal{H}} \leq \mathcal{R}_{\widetilde{\Phi}^{\mathrm{sum}}_{\rho}}(h) - \mathcal{R}^*_{\widetilde{\Phi}^{\mathrm{sum}}_{\rho}, \mathcal{H}} + \mathcal{M}_{\widetilde{\Phi}^{\mathrm{sum}}_{\rho}, \mathcal{H}} - \mathcal{M}_{\ell_\gamma, \mathcal{H}}. \tag{20}$$

Proof. For any $x \in \mathcal{X}$, we define $p_{[1]}(x), p_{[2]}(x), \ldots, p_{[c]}(x)$ by sorting the probabilities $\{p(x, y) : y \in \mathcal{Y}\}$ in increasing order. Similarly, for any $x \in \mathcal{X}$ and $h \in \mathcal{H}$, we define $h(x, \{1\}_x), h(x, \{2\}_x), \ldots, h(x, \{c\}_x)$ by sorting the scores $\{h(x, y) : y \in \mathcal{Y}\}$ in increasing order. In particular, we have

$$h(x, \{1\}_x) = \min_{y \in \mathcal{Y}} h(x, y), \quad h(x, \{c\}_x) = \max_{y \in \mathcal{Y}} h(x, y), \quad h(x, \{i\}_x) \leq h\big(x, \{j\}_j\big), \ \forall i \leq j.$$

If there is a tie for the maximum, we pick the label with the highest index under the natural ordering of labels, i.e. $\{c\}_x = \mathsf{h}(x)$. By the definition, the conditional $\widetilde{\Phi}^{\mathrm{sum}}_{\rho}$-risk can be expressed as follows:

$$\begin{aligned}
&\mathcal{C}_{\widetilde{\Phi}^{\mathrm{sum}}_{\rho}}(h, x) = \sum_{y \in \mathcal{Y}} p(x, y) \sup_{x' : \|x - x'\|_p \leq \gamma} \sum_{y' \neq y} \Phi_\rho(h(x', y) - h(x', y')) \\
&= \sum_{y \in \mathcal{Y}} \sup_{x' : \|x - x'\|_p \leq \gamma} p(x, y) \sum_{y' \neq y} \Phi_\rho(h(x', y) - h(x', y')) \\
&= \sum_{i=1}^{c} \sup_{x' : \|x - x'\|_p \leq \gamma} p(x, \{i\}_{x'}) \left[\sum_{j=1}^{i-1} \Phi_\rho(h(x', \{i\}_{x'}) - h(x', \{j\}_{x'})) + \sum_{j=i+1}^{c} \Phi_\rho(h(x', \{i\}_{x'}) - h(x', \{j\}_{x'})) \right] \\
&= \sum_{i=1}^{c} \sup_{x' : \|x - x'\|_p \leq \gamma} p(x, \{i\}_{x'}) \left[\sum_{j=1}^{i-1} \Phi_\rho(h(x', \{i\}_{x'}) - h(x', \{j\}_{x'})) + c - i \right]
\end{aligned} \tag{46}$$

By the assumption, there exists a hypothesis $h \in \mathcal{H}$ inducing the same ordering of the labels for any $x' \in \{x' : \|x - x'\|_p \leq \gamma\}$ and such that $\inf_{x' : \|x - x'\|_p \leq \gamma} |h(x', i) - h(x', j)| \geq \rho$ for any $i \neq j \in \mathcal{Y}$, i.e. $\{k\}_{x'} = \{k\}_x$ for any $k \in \mathcal{Y}$ and $x' \in \{x' : \|x - x'\|_p \leq \gamma\}$. Since $\mathcal{H}$ is symmetric, we can always choose h^* among these hypotheses such that h^* and $p(x)$ induce the same ordering of the labels, i.e. $p(x, \{k\}_x) = p_{[k]}(x)$ for any $k \in \mathcal{Y}$. Then, by (46), we have

$$\begin{aligned}
\mathcal{C}^*_{\widetilde{\Phi}^{\mathrm{sum}}_{\rho}, \mathcal{H}}(x) &\leq \mathcal{C}_{\widetilde{\Phi}^{\mathrm{sum}}_{\rho}}(h^*, x) \\
&= \sum_{i=1}^{c} \sup_{x' : \|x - x'\|_p \leq \gamma} p(x, \{i\}_{x'}) \left[\sum_{j=1}^{i-1} \Phi_\rho(h^*(x', \{i\}_{x'}) - h^*(x', \{j\}_{x'})) + c - i \right] \quad \text{(by (46))} \\
&= \sum_{i=1}^{c} \sup_{x' : \|x - x'\|_p \leq \gamma} p(x, \{i\}_{x'})(c - i) \\
&\qquad (\inf_{x' : \|x - x'\|_p \leq \gamma} |h^*(x', i) - h^*(x', j)| \geq \rho \text{ for any } i \neq j \text{ and } \Phi_\rho(t) = 0, \ \forall t \geq \rho) \\
&= \sum_{i=1}^{c} p(x, \{i\}_x)(c - i) \\
&\qquad (h^* \text{ induces the same ordering of the labels for any } x' \in \{x' : \|x - x'\|_p \leq \gamma\}) \\
&= \sum_{i=1}^{c} p_{[i]}(x)(c - i) \qquad (h^* \text{ and } p(x) \text{ induce the same ordering of the labels}) \\
&= c - \sum_{i=1}^{c} i\, p_{[i]}(x) \qquad (\textstyle\sum_{i=1}^{c} p_{[i]}(x) = 1)
\end{aligned}$$

By the assumption, $\mathcal{H}_\gamma(x) \neq \varnothing$ and $\mathsf{H}_\gamma(x) = \mathcal{Y}$ since $\mathcal{H}$ is symmetric. Thus, for any $h \in \mathcal{H}$,

$$
\begin{aligned}
&\Delta\mathcal{C}_{\widetilde{\Phi}^{\mathrm{sum}}_\rho,\mathcal{H}}(h,x)\\
&= \mathcal{C}_{\widetilde{\Phi}^{\mathrm{sum}}_\rho}(h,x) - \mathcal{C}^*_{\widetilde{\Phi}^{\mathrm{sum}}_\rho,\mathcal{H}}(x)\\
&= \sum_{i=1}^{c} \sup_{x':\|x-x'\|_p\leq\gamma} p(x,\{i\}_{x'})\left[\sum_{j=1}^{i-1}\Phi_\rho(h(x',\{i\}_{x'}) - h(x',\{j\}_{x'})) + c - i\right] - \left(c - \sum_{i=1}^{c} i\, p_{[i]}(x)\right)\\
&\qquad\qquad (\Phi_\rho(t) = 1 \text{ for } t \leq 0)\\
&\geq p(x,\mathsf{h}(x))\mathbb{1}_{h\notin\mathcal{H}_\gamma(x)} + \sum_{i=1}^{c} \sup_{x':\|x-x'\|_p\leq\gamma} p(x,\{i\}_{x'})(c-i) - \left(c - \sum_{i=1}^{c} i\, p_{[i]}(x)\right) \qquad (\Phi_\rho \geq 0)\\
&\geq p(x,\mathsf{h}(x))\mathbb{1}_{h\notin\mathcal{H}_\gamma(x)} + \sum_{i=1}^{c} p(x,\{i\}_x)(c-i) - \left(c - \sum_{i=1}^{c} i\, p_{[i]}(x)\right) \quad \text{(lower bound the supremum)}\\
&= p(x,\mathsf{h}(x))\mathbb{1}_{h\notin\mathcal{H}_\gamma(x)} + \sum_{i=1}^{c} i\, p_{[i]}(x) - \sum_{i=1}^{c} i\, p(x,\{i\}_x) \qquad (\textstyle\sum_{i=1}^{c} p(x,\{i\}) = 1)\\
&= p(x,\mathsf{h}(x))\mathbb{1}_{h\notin\mathcal{H}_\gamma(x)} + \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x)) + \begin{bmatrix} c-1\\ c-1\\ c-2\\ \vdots\\ 1\end{bmatrix}\cdot\begin{bmatrix} p_{[c]}(x)\\ p_{[c-1]}(x)\\ p_{[c-2]}(x)\\ \vdots\\ p_{[1]}(x)\end{bmatrix} - \begin{bmatrix} c-1\\ c-1\\ c-2\\ \vdots\\ 1\end{bmatrix}\cdot\begin{bmatrix} p(x,\{c\}_x)\\ p(x,\{c-1\}_x)\\ p(x,\{c-2\}_x)\\ \vdots\\ p(x,\{1\}_x)\end{bmatrix}\\
&\qquad\qquad (p_{[c]}(x) = \max_{y\in\mathcal{Y}} p(x,y) \text{ and } \{c\}_x = \mathsf{h}(x))\\
&\geq p(x,\mathsf{h}(x))\mathbb{1}_{h\notin\mathcal{H}_\gamma(x)} + \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x)) \qquad \text{(by Lemma 21)}\\
&= \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\mathbb{1}_{h\in\mathcal{H}_\gamma(x)}\\
&= \Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x) \qquad \text{(by Lemma 11 and } \mathsf{H}_\gamma(x) = \mathcal{Y})\\
&\geq \left[\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x)\right]_\epsilon \qquad ([t]_\epsilon \leq t)
\end{aligned}
$$

for any $\epsilon \geq 0$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric hypothesis set, $\epsilon = 0$ and $\Psi(t) = t$ in Theorem 12, or, equivalently, $\Gamma(t) = t$ in Theorem 13, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$
\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq \mathcal{R}_{\widetilde{\Phi}^{\mathrm{sum}}_\rho}(h) - \mathcal{R}^*_{\widetilde{\Phi}^{\mathrm{sum}}_\rho,\mathcal{H}} + \mathcal{M}_{\widetilde{\Phi}^{\mathrm{sum}}_\rho,\mathcal{H}} - \mathcal{M}_{\ell_\gamma,\mathcal{H}}.
$$

□

O Proof of $\mathcal{H}$-consistency bounds for adversarial constrained losses $\widetilde{\Phi}^{\mathrm{cstnd}}$

Theorem 17 (**$\mathcal{H}$-consistency bound of $\widetilde{\Phi}^{\mathrm{cstnd}}_\rho$**). *Suppose that $\mathcal{H}$ is symmetric and satisfies that for any $x \in \mathcal{X}$, there exists a hypothesis $h \in \mathcal{H}$ with the constraint $\sum_{y\in\mathcal{Y}} h(x,y) = 0$ such that $\sup_{x':\|x-x'\|_p\leq\gamma} h(x',y) \leq -\rho$ for any $y \neq y_{\max}$. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution,*

$$
\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq \mathcal{R}_{\widetilde{\Phi}^{\mathrm{cstnd}}_\rho}(h) - \mathcal{R}^*_{\widetilde{\Phi}^{\mathrm{cstnd}}_\rho,\mathcal{H}} + \mathcal{M}_{\widetilde{\Phi}^{\mathrm{cstnd}}_\rho,\mathcal{H}} - \mathcal{M}_{\ell_\gamma,\mathcal{H}}. \tag{22}
$$

Proof. Define $y_{\max}$ by $y_{\max} = \operatorname{argmax}_{y\in\mathcal{Y}} p(x,y)$. If there is a tie, we pick the label with the highest index under the natural ordering of labels. Since $\sum_{y\in\mathcal{Y}} h(x,y) = 0$, by definition of $\mathsf{h}(x)$ as a maximizer, we must thus have $h(x,\mathsf{h}(x)) \geq 0$. By the definition, the conditional $\widetilde{\Phi}^{\mathrm{cstnd}}_\rho$-risk can be

expressed as follows:

$$
\begin{aligned}
&\mathcal{C}_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}}}(h,x) = \sum_{y\in\mathcal{Y}} p(x,y) \sup_{x':\|x-x'\|_p\le\gamma} \sum_{y'\neq y} \Phi_\rho(-h(x',y')) \\
&= \sum_{y\in\mathcal{Y}} \sup_{x':\|x-x'\|_p\le\gamma} p(x,y) \sum_{y'\neq y} \Phi_\rho(-h(x',y')) \\
&= \sum_{y\in\mathcal{Y}} \sup_{x':\|x-x'\|_p\le\gamma} p(x,y) \left[\sum_{y'\neq y: h(x',y')>0} \Phi_\rho(-h(x',y')) + \sum_{y'\neq y: h(x',y')\le 0} \Phi_\rho(-h(x',y')) \right] \\
&\ge \sum_{y\neq \mathsf{h}(x')} \sup_{x':\|x-x'\|_p\le\gamma} p(x,y) \\
&\ge 1 - \max_{y\in\mathcal{Y}} p(x,y). && (\Phi_\rho \ge 0 \text{ and } \Phi_\rho(t) = 1 \text{ for } t \le 0)
\end{aligned}
$$

By the assumption, the equality can be achieved by some $h_\rho^* \in \mathcal{H}$ with the constraint $\sum_{y\in\mathcal{Y}} h(x,y) = 0$ such that $\sup_{x':\|x-x'\|_p\le\gamma} h_\rho^*(x',y) \le -\rho$ for any $y \neq y_{\max}$ and $h_\rho^*(x',y_{\max}) = -\sum_{y'\neq y_{\max}} h_\rho^*(x',y')$ for any $x' \in \left\{x' : \|x-x'\|_p \le \gamma\right\}$. Therefore, the minimal conditional $\widetilde{\Phi}_\rho^{\mathrm{cstnd}}$-risk can be expressed as follows:

$$
\mathcal{C}^*_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}},\mathcal{H}}(x) = 1 - \max_{y\in\mathcal{Y}} p(x,y).
$$

By the assumption, $\mathcal{H}_\gamma(x) \neq \varnothing$ and $\mathsf{H}_\gamma(x) = \mathcal{Y}$ since $\mathcal{H}$ is symmetric. Thus, for any $h \in \mathcal{H}$ with the constraint that $\sum_{y\in\mathcal{Y}} h(x,y) = 0$,

$$
\begin{aligned}
&\Delta\mathcal{C}_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}},\mathcal{H}}(h,x) = \mathcal{C}_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}}}(h,x) - \mathcal{C}^*_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}},\mathcal{H}}(x) \\
&= \sum_{y\in\mathcal{Y}} \sup_{x':\|x-x'\|_p\le\gamma} p(x,y) \sum_{y'\neq y} \Phi_\rho(-h(x',y')) - \left(1 - \max_{y\in\mathcal{Y}} p(x,y)\right) \\
&\ge \sum_{y\in\mathcal{Y}} p(x,y) \sum_{y'\neq y} \Phi_\rho(-h(x,y')) - \left(1 - \max_{y\in\mathcal{Y}} p(x,y)\right) && \text{(lower bound the supremum)} \\
&= \sum_{y\in\mathcal{Y}} (1 - p(x,y)) \Phi_\rho(-h(x,y)) - \left(1 - \max_{y\in\mathcal{Y}} p(x,y)\right) && \text{(swap } y \text{ and } y') \\
&\ge p(x,\mathsf{h}(x)) \mathbb{1}_{h\notin\mathcal{H}_\gamma(x)} + 1 - p(x,\mathsf{h}(x)) - \left(1 - \max_{y\in\mathcal{Y}} p(x,y)\right) \\
&= \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x)) \mathbb{1}_{h\in\mathcal{H}_\gamma(x)} \\
&= \Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x) && \text{(by Lemma 11 and } \mathsf{H}_\gamma(x) = \mathcal{Y}) \\
&\ge \left[\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x)\right]_\epsilon && ([t]_\epsilon \le t)
\end{aligned}
$$

for any $\epsilon \ge 0$. Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric hypothesis set, $\epsilon = 0$ and $\Psi(t) = t$ in Theorem 12, or, equivalently, $\Gamma(t) = t$ in Theorem 13, we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$
\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \le \mathcal{R}_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}}}(h) - \mathcal{R}^*_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}},\mathcal{H}} + \mathcal{M}_{\widetilde{\Phi}_\rho^{\mathrm{cstnd}},\mathcal{H}} - \mathcal{M}_{\ell_\gamma,\mathcal{H}}.
$$

□