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

Anonymous Author(s)
Affiliation
Address
email

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, 2004b,a; Bartlett et al., 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.

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 Zhang and Agarwal (2020) and more generally by Awasthi et al. (2021) 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.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Awasthi et al. (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. (2021). 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, 2004b; 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 a detailed discussion in Appendix C.

This paper presents an extensive study of $\mathcal{H}$-consistency bounds for multi-class classification. 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.

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 a 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 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.

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 assume that in case there is a tie, $\mathsf{h}(x)$ is defined as the label with the highest index under the natural ordering of labels. See more discussion in Appendix B.

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

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

Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and a loss function $\ell \colon \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, 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 by $\ell_{0-1}(h, x, y) = \mathbb{1}_{\mathsf{h}(x) \neq y}$, for any $h \in \mathcal{H}$. In the

adversarial scenario, the goal is to select a hypothesis $h \in \mathcal{H}$ with small *adversarial generalization error* defined by $\mathcal{R}_{\ell_\gamma}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_\gamma(h,x,y)]$, where

$$\ell_\gamma(h,x,y) = \sup_{x':\|x-x'\|_p\le\gamma} \mathbb{1}_{\rho_h(x',y)\le 0} = \mathbb{1}_{\inf_{x':\|x-x'\|_p\le\gamma}\rho_h(x',y)\le 0}, \quad \gamma\in(0,1), \quad p\in[1,+\infty]$$

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:

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

where $\widetilde{\ell}(h,x,y) = \sup_{x':\|x-x'\|_p\le\gamma} \ell(h,x',y)$ is the supremum based counterpart of the original loss function $\ell(h,x,y)$. For a distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, we define $p(x) = (p(x,1),\cdots,p(x,c))$, 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 $\mathcal{C}_\ell(h,x) = \sum_{y\in\mathcal{Y}} p(x,y)\ell(h,x,y)$. 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 $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x) = \mathcal{C}_\ell(h,x) - \mathcal{C}^*_{\ell,\mathcal{H}}(x)$. We 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 a hypothesis set $\mathcal{H}$ and distribution $\mathcal{D}$, we define the $(\ell,\mathcal{H})$*-minimizability gap* as $\mathcal{M}_{\ell,\mathcal{H}} = \mathcal{R}^*_{\ell,\mathcal{H}} - \mathbb{E}_X\big[\mathcal{C}^*_{\ell,\mathcal{H}}(x)\big]$, which is a key quantity appearing in our bounds. For any $\epsilon>0$, we will denote by $[t]_\epsilon$ the ϵ-truncation of $t\in\mathbb{R}$ defined by $t\mathbb{1}_{t>\epsilon}$. As a result, the conditional ϵ-regret can be rewritten as $\big[\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)\big]_\epsilon$.

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.

3 General theorems

We will use the following two results similar to their binary classification counterparts due to Awasthi et al. (2022). The conditional ℓ-risk $\mathcal{C}_\ell(h,x)$ in our theorems is the multi-class generalization of that in (Awasthi et al., 2022). The proofs are similar and are included in Appendix E for completeness.

Theorem 1 (**Distribution-dependent Ψ-bound**). *Assume that there exists a convex function $\Psi:\mathbb{R}_+\to\mathbb{R}$ with $\Psi(0)\ge 0$ and $\epsilon\ge 0$ such that the following holds for all $h\in\mathcal{H}$, $x\in\mathcal{X}$ and $\mathcal{D}\in\mathcal{P}$:*

$$\Psi\Big([\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)]_\epsilon\Big) \le \Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h,x).$$

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

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

Theorem 2 (**Distribution-dependent Γ-bound**). *Assume that there exists a concave function $\Gamma:\mathbb{R}_+\to\mathbb{R}$ and $\epsilon\ge 0$ such that the following holds for all $h\in\mathcal{H}$, $x\in\mathcal{X}$ and $\mathcal{D}\in\mathcal{P}$:*

$$[\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x)]_\epsilon \le \Gamma(\Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h,x)).$$

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

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

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. Let $B^d_p(r)$ denote the d-dimensional ℓ_p-ball with radius r: $B^d_p(r) = \big\{z\in\mathbb{R}^d \mid \|z\|_p \le r\big\}$, where $\|z\|_p := \big[\sum_{i=1}^d |z_i|^p\big]^{\frac{1}{p}}$, $p\in[1,+\infty]$. 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 the family of all measurable functions $\mathcal{H}_{\text{all}}$, the family of linear hypotheses $\mathcal{H}_{\text{lin}} = \big\{(x,y)\mapsto w_y\cdot x + b_y \mid \|w_y\|_q \le W, |b_y|\le B\big\}$ and that of one-hidden-layer ReLU networks $\mathcal{H}_{\text{NN}} = \big\{(x,y)\mapsto \sum_{j=1}^n u_{y,j}(w_{y,j}\cdot x + b_{y,j})_+ \mid \|u_y\|_1 \le \Lambda, \|w_{y,j}\|_q \le W, |b_{y,j}|\le B\big\}$, where $(\cdot)_+ = \max(\cdot,0)$.

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 would be helpful for instantiating Theorems 1 and 2 in the non-adversarial scenario. For convenience, we will adopt the following notation:

$$\overline{\mathcal{Y}}(x) = \{y \in \mathcal{Y} : \exists h \in \mathcal{H} \text{ such that } \mathsf{h}(x) = y\}.$$

Lemma 3. *For any hypothesis set $\mathcal{H}$, the minimal conditional ℓ_{0-1}-risk can be expressed as*

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

The conditional ϵ-regret for ℓ_{0-1} can be characterized as

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

The proof of Lemma 3 is deferred to 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 $\Psi\colon \mathbb{R}_+ \to \mathbb{R}$ 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 $\mathcal{D} \in \mathcal{P}$:*

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

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

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

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

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

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

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

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

We call a multi-class surrogate loss ℓ *max loss* if it is of the form in (Crammer and Singer, 2001), *sum loss* if it is of the form in (Weston and Watkins, 1998), and *constrained loss* if it is of the form in (Lee et al., 2004) with rigorous definitions in Sections 4.1, 4.2 and 4.3 respectively. For each structure, we denote by Φ a non-increasing and non-negative auxiliary function on $\mathbb{R}$. In particular, we will consider the following common functions as auxiliary functions used to construct the surrogate loss: the hinge loss $\Phi_{\text{hinge}}(t) = \max\{0, 1-t\}$, the squared hinge loss $\Phi_{\text{sq-hinge}}(t) = \max\{0, 1-t\}^2$, the exponential loss $\Phi_{\exp}(t) = e^{-t}$, and the ρ-margin loss $\Phi_\rho(t) = \min\left\{\max\left\{0, 1 - \frac{t}{\rho}\right\}, 1\right\}$. Note that the first three auxiliary functions are convex, while the last one is not. See Figure 1 for the visualization of the auxiliary functions.

We will say that a hypothesis set $\mathcal{H}$ is *symmetric* if individual scoring functions $h(\cdot, y)$ are independently drawn from a family of real valued functions, that is, there exists a family $\mathcal{F}$ of functions f mapping from $\mathcal{X}$ to $\mathbb{R}$ such that $\{[h(x, 1), \ldots, h(x, c)]\colon h \in \mathcal{H}\} = \{[f_1(x), \ldots, f_c(x)] : f_1, \ldots, f_c \in \mathcal{F}\}$ and $|\{f(x)\colon f \in \mathcal{F}\}| \geq 2$ for any $x \in \mathcal{X}$. All the hypothesis sets defined above, $\mathcal{H}_{\text{all}}$, $\mathcal{H}_{\text{lin}}$ and $\mathcal{H}_{\text{NN}}$, are symmetric. Note that for a symmetric hypothesis set $\mathcal{H}$, $\overline{\mathcal{Y}}(x) = \mathcal{Y}$. We will say that a hypothesis set $\mathcal{H}$ is *complete* if the set of scores it generates spans $\mathbb{R}$, that is, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\{h(x, y)\colon h \in \mathcal{H}\} = \mathbb{R}$. The hypothesis sets defined above, $\mathcal{H}_{\text{all}}$, $\mathcal{H}_{\text{lin}}$ and $\mathcal{H}_{\text{NN}}$ with $B = +\infty$ are all complete.

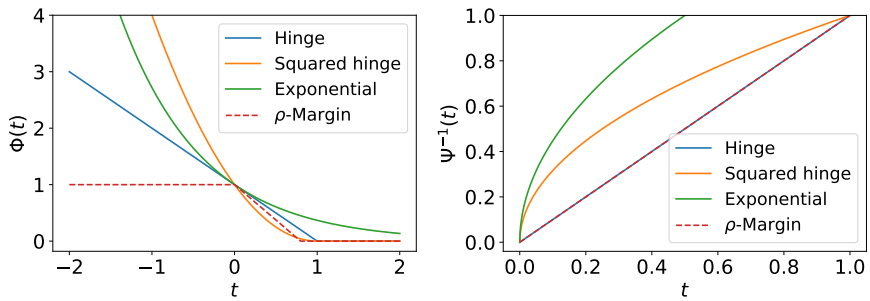

Figure 1: Left: auxiliary functions. Right: $\mathcal{H}$-consistency dependence between ℓ_{0-1} and Φ^{cstnd}.

4.1 Max losses

In this section, we discuss guarantees for *max loss* (Crammer and Singer, 2001), defined by

$$\Phi^{\max}(h, x, y) = \max_{y' \neq y} \Phi(h(x, y) - h(x, y')) = \Phi(\rho_h(x, y)). \tag{5}$$

4.1.1 Negative results

We first give negative results showing that the 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, that is, the case where $c > 2$. 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 $\left|\overline{\mathcal{Y}}(x)\right| \geq 2$ 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:*

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

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.

4.1.2 Positive results without distributional assumptions

On a positive side, the max loss with a non-convex auxiliary function $\Phi = \Phi_\rho$ presents intriguing possibility of $\mathcal{H}$-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,*

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

See Appendix G for the proof. Theorem 7 is very powerful since it only requires $\mathcal{H}$ to be symmetric. By applying Theorem 7, we can derive $\mathcal{H}$-consistency bounds of $\Phi^{\max}_\rho$ with common symmetric hypothesis sets, e.g. $\mathcal{H}_{\mathrm{all}}$, $\mathcal{H}_{\mathrm{lin}}$ and $\mathcal{H}_{\mathrm{NN}}$, as summarized in Table 1. The proofs with corresponding summarized Corollaries 18, 19 and 20 are included in Appendix H. In the proof, 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 are no useful $\mathcal{H}$-consistency bounds for the max loss with $\Phi = \Phi_{\mathrm{hinge}}$, $\Phi_{\mathrm{sq-hinge}}$ or Φ_{exp} in these cases. However, under the realizability assumption (Definition 8), we will show that such bounds exist.

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

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

4.1.3 Positive results with realizable distributions

We consider $\mathcal{H}$-realizability condition (Long and Servedio, 2013; Zhang and Agarwal, 2020; Awasthi et al., 2021) 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,*

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

See Appendix G for the proof. Long and Servedio (2013, Theorem 9) show that $\Phi_{\text{hinge}}^{\max}$ 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 exhibits a quantitative relationship in that case, which is stronger than the asymptotic consistency property in the previous work.

4.2 Sum losses

In this section, we discuss guarantees for *sum loss* (Weston and Watkins, 1998), defined by

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

4.2.1 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:*

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

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

4.2.2 Positive results

We then complement the negative result by showing that there exist non-trivial $\mathcal{H}$-consistency bounds 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_ρ^{sum}, 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 $|h(x,i) - h(x,j)| \geq \rho$ 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_ρ^{sum} with that of the multi-class $0/1$ loss.

Table 2: $\mathcal{H}$-consistency bounds of 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 of 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* (Lee et al., 2004), defined by

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

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 exhibits linear dependency for $\Phi^{\text{cstnd}}_{\text{hinge}}$ and $\Phi^{\text{cstnd}}_{\rho}$, while 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 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 having to derive the closed form of the minimal conditional risk.

As shown by Steinwart (2007, Theorem 3.2), the minimizability gaps $\mathcal{M}_{\ell_{0-1},\mathcal{H}_{\text{all}}} = \mathcal{M}_{\Phi^{\text{sum}},\mathcal{H}_{\text{all}}} = \mathcal{M}_{\Phi^{\text{cstnd}},\mathcal{H}_{\text{all}}} = 0$ with $\Phi = \Phi_{\text{hinge}}$, $\Phi_{\text{sq-hinge}}$, Φ_{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 discuses $\mathcal{H}$-consistency bounds in the adversarial scenario where $\ell_2 = \ell_\gamma$ is the adversarial multi-class $0/1$ loss. For convenience, we define $\overline{\mathcal{H}}_\gamma(x) = \left\{h \in \mathcal{H} : \inf_{x' : \|x - x'\|_p \leq \gamma} \rho_h(x', \mathsf{h}(x)) > 0\right\}$ and $\overline{\mathcal{Y}}_\gamma(x) = \left\{y \in \mathcal{Y} : \exists h \in \overline{\mathcal{H}}_\gamma(x) \text{ such that } \mathsf{h}(x) = y\right\}$. Note that $\overline{\mathcal{H}}_\gamma(x) = \varnothing \iff \overline{\mathcal{Y}}_\gamma(x) = \varnothing$. When the hypothesis set $\mathcal{H}$ is symmetric, $\overline{\mathcal{Y}}_\gamma(x) = \mathcal{Y}$ if and only if $\overline{\mathcal{H}}_\gamma(x) \neq \varnothing$. The lemma stated next 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 hypothesis set $\mathcal{H}$, the minimal conditional ℓ_γ-risk can be expressed as*

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

The conditional ϵ-regret for ℓ_γ can be characterized as

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

The proof of Lemma 11 is deferred to 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 $h \in \mathcal{H}$, $x \in \left\{x \in \mathcal{X} : \overline{\mathcal{H}}_\gamma(x) \neq \varnothing\right\}$ and $\mathcal{D} \in \mathcal{P}$:*

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

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

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

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 $h \in \mathcal{H}$, $x \in \left\{x \in \mathcal{X} : \overline{\mathcal{H}}_\gamma(x) \neq \varnothing\right\}$ and $\mathcal{D} \in \mathcal{P}$:*

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

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

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

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 robustness**). *Consider $c = 2$. Suppose that Φ is convex and non-increasing, and $\mathcal{H}$ contains 0 and satisfies 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:*

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

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, 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** $\widetilde{\Phi}^{\max}_\rho$)**.** *Suppose that $\mathcal{H}$ is symmetric. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution,*

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

5.3 Adversarial sum losses

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

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

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

Theorem 16 ($\mathcal{H}$**-consistency bound of** $\widetilde{\Phi}^{\text{sum}}_\rho$)**.** *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 $x' \in \left\{x' : \|x - x'\|_p \leq \gamma\right\}$ 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:*

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

5.4 Adversarial constrained loss

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

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

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 $\widetilde{\Phi}^{\text{cstnd}}_\rho$ as follows.

Theorem 17 ($\mathcal{H}$**-consistency bound of** $\widetilde{\Phi}^{\text{cstnd}}_\rho$)**.** *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,*

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

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 $\mathcal{H}_{\text{all}}$, $\mathcal{H}_{\text{lin}}$ and $\mathcal{H}_{\text{NN}}$ 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 $\mathcal{H}_{\text{lin}}$ and $\mathcal{H}_{\text{NN}}$., 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

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

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

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. (2022). $\mathcal{H}$-consistency estimation error of surrogate loss minimizers. In *International Conference on Machine Learning*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lee, Y., Lin, Y., and Wahba, G. (2004). 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.

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

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

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

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

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

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

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

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

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

Ramaswamy, H. G. and Agarwal, S. (2012). Classification calibration dimension for general multi-class losses. In *Advances in Neural Information Processing Systems*.

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

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

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

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

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

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

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

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

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

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

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

Zhang, M. and Agarwal, S. (2020). 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.

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

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See abstract and introduction.
   (b) Did you describe the limitations of your work? [Yes] See Appendix D.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We don't foresee any negative societal impact of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3, Section 4 and Section 5.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix E, F, G, H, I, J, K, L, M, N, O.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]