Cross-Entropy Loss Functions: Theoretical Analysis and Applications

Anqi Mao [1] **Mehryar Mohri** [2 1] **Yutao Zhong** [1]

Abstract

Cross-entropy is a widely used loss function in applications. It coincides with the logistic loss applied to the outputs of a neural network, when the softmax is used. But, what guarantees can we rely on when using cross-entropy as a surrogate loss? We present a theoretical analysis of a broad family of loss functions, *comp-sum losses*, that includes cross-entropy (or logistic loss), generalized cross-entropy, the mean absolute error and other cross-entropy-like loss functions. We give the first $\mathcal{H}$-consistency bounds for these loss functions. These are non-asymptotic guarantees that upper bound the zero-one loss estimation error in terms of the estimation error of a surrogate loss, for the specific hypothesis set $\mathcal{H}$ used. We further show that our bounds are *tight*. These bounds depend on quantities called *minimizability gaps*. To make them more explicit, we give a specific analysis of these gaps for comp-sum losses. We also introduce a new family of loss functions, *smooth adversarial comp-sum losses*, that are derived from their comp-sum counterparts by adding in a related smooth term. We show that these loss functions are beneficial in the adversarial setting by proving that they admit $\mathcal{H}$-consistency bounds. This leads to new adversarial robustness algorithms that consist of minimizing a regularized smooth adversarial comp-sum loss. While our main purpose is a theoretical analysis, we also present an extensive empirical analysis comparing comp-sum losses. We further report the results of a series of experiments demonstrating that our adversarial robustness algorithms outperform the current state-of-the-art, while also achieving a superior non-adversarial accuracy.

[1]Courant Institute of Mathematical Sciences, New York, NY; [2]Google Research, New York, NY. Correspondence to: Anqi Mao <aqmao@cims.nyu.edu>, Mehryar Mohri <mohri@google.com>, Yutao Zhong <yutao@cims.nyu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

Most current learning algorithms rely on minimizing the cross-entropy loss to achieve a good performance in a classification task. This is because directly minimizing the zero-one classification loss is computationally hard. But, what guarantees can we benefit from when using cross-entropy as a surrogate loss?

Cross-entropy coincides with the (multinomial) logistic loss applied to the outputs of a neural network, when the softmax is used. It is known that the logistic loss is Bayes consistent (Zhang, 2004a). Thus, asymptotically, a nearly optimal minimizer of the logistic loss over the family of all measurable functions is also a nearly optimal optimizer of the zero-one classification loss. However, this does not supply any information about learning with a typically restricted hypothesis set, which of course would not contain all measurable functions. It also provides no guarantee for approximate minimizers (non-asymptotic guarantee) since convergence could be arbitrarily slow. What non-asymptotic guarantees can we rely on when minimizing the logistic loss with a restricted hypothesis set, such as a family of neural networks?

Recent work by Awasthi, Mao, Mohri, and Zhong (2022b;a) introduced the notion of $\mathcal{H}$-*consistency bounds*. These are upper bounds on the zero-one estimation error of any predictor in a hypothesis set $\mathcal{H}$ in terms of its surrogate loss estimation error. Such guarantees are thus both non-asymptotic and hypothesis set-specific and therefore more informative than Bayes consistency guarantees (Zhang, 2004a; Bartlett et al., 2006; Steinwart, 2007; Tewari & Bartlett, 2007). For multi-class classification, the authors derived such guarantees for the so-called *sum-losses*, such as the sum-exponential loss function of Weston & Watkins (1998), and *constrained losses*, such as the loss function of Lee et al. (2004), where the scores (logits) must sum to zero. To the best of our knowledge, no such guarantee has been given for the more widely used logistic loss (or cross-entropy).

This paper presents the first $\mathcal{H}$-consistency bounds for the logistic loss, which can be used to derive directly guarantees for current algorithms used in the machine learning community. More generally, we will consider a broader

family of loss functions that we refer to as *comp-sum losses*, that is loss functions obtained by composition of a concave function, such as logarithm in the case of the logistic loss, with a sum of functions of differences of score, such as the negative exponential. We prove $\mathcal{H}$-consistency bounds for a wide family of comp-sum losses, which includes as special cases the logistic loss (Verhulst, 1838; 1845; Berkson, 1944; 1951), the *generalized cross-entropy loss* (Zhang & Sabuncu, 2018), and the *mean absolute error loss* (Ghosh et al., 2017). We further show that our bounds are *tight* and thus cannot be improved.

$\mathcal{H}$-consistency bounds are expressed in terms of a quantity called *minimizability gap*, which only depends on the loss function and the hypothesis set $\mathcal{H}$ used. It is the difference of the best-in class expected loss and the expected pointwise infimum of the loss. For the loss functions we consider, the minimizability gap vanishes when $\mathcal{H}$ is the full family of measurable functions. However, in general, the gap is non-zero and plays an important role, depending on the property of the loss function and the hypothesis set. Thus, to better understand $\mathcal{H}$-consistency bounds for comp-sum losses, we specifically analyze their minimizability gaps, which we use to compare their guarantees.

A recent challenge in the application of neural networks is their robustness to imperceptible perturbations (Szegedy et al., 2013). While neural networks trained on large datasets often achieve a remarkable performance (Sutskever et al., 2014; Krizhevsky et al., 2012), their accuracy remains substantially lower in the presence of such perturbations. One key issue in this scenario is the definition of a useful surrogate loss for the adversarial loss. To tackle this problem, we introduce a family of loss functions designed for adversarial robustness that we call *smooth adversarial comp-sum loss functions*. These are loss functions derived from their comp-sum counterparts by augmenting them with a natural smooth term. We show that these loss functions are beneficial in the adversarial setting by proving that they admit $\mathcal{H}$-consistency bounds. This leads to a family of algorithms for adversarial robustness that consist of minimizing a regularized smooth adversarial comp-sum loss.

While our main purpose is a theoretical analysis, we also present an extensive empirical analysis. We compare the empirical performance of comp-sum losses for different tasks and relate that to their theoretical properties. We further report the results of experiments with the CIFAR-10, CIFAR-100 and SVHN datasets comparing the performance of our algorithms based on smooth adversarial comp-sum losses with that of the state-of-the-art algorithm for this task TRADES (Zhang et al., 2019b). The results show that our adversarial algorithms outperform TRADES and also achieve a substantially better non-adversarial (clean) accuracy.

The rest of this paper is organized as follows. In Section 2, we introduce some basic concepts and definitions related to comp-sum loss functions. In Section 3, we present our $\mathcal{H}$-consistency bounds for comp-sum losses. We further carefully compare their minimizability gaps in Section 4. In Section 5, we define and motivate our smooth adversarial comp-sum losses, for which we prove $\mathcal{H}$-consistency bounds, and briefly discuss corresponding adversarial algorithms. In Section 6, we report the results of our experiments both to compare comp-sum losses in several tasks, and to compare the performance of our algorithms based on smooth adversarial comp-sum losses. In Section 7, we discuss avenues for future research. In Appendix A, we give a comprehensive discussion of related work.

2. Preliminaries

We consider the familiar multi-class classification setting and denote by $\mathcal{X}$ the input space, by $\mathcal{Y} = [n] = \{1, \ldots, n\}$ the set of classes or categories ($n \geq 2$) and by $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{Y}$.

We study general loss functions $\ell \colon \mathcal{H}_{\text{all}} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ where $\mathcal{H}_{\text{all}}$ is the family of all measurable functions $h \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. In particular, the zero-one classification loss is defined, for all $h \in \mathcal{H}_{\text{all}}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, by $\ell_{0-1}(h, x, y) = 1_{\mathsf{h}(x) \neq y}$, where $\mathsf{h}(x) = \operatorname{argmax}_{y \in \mathcal{Y}} h(x, y)$ with an arbitrary but fixed deterministic strategy used for breaking the ties. For simplicity, we fix that strategy to be the one selecting the label with the highest index under the natural ordering of labels.

We denote by $\mathcal{R}_\ell(h)$ the generalization error or expected loss of a hypothesis $h \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$: $\mathcal{R}_\ell(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h, x, y)]$. For a hypothesis set $\mathcal{H} \subseteq \mathcal{H}_{\text{all}}$ of functions mapping from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, $\mathcal{R}^*_\ell(\mathcal{H})$ denotes the best-in class expected loss: $\mathcal{R}^*_\ell(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{R}_\ell(h)$.

We will prove $\mathcal{H}$-consistency bounds, which are inequalities relating the zero-one classification estimation loss ℓ_{0-1} of any hypothesis $h \in \mathcal{H}$ to that of its surrogate loss ℓ (Awasthi, Mao, Mohri, and Zhong, 2022b;a). They take the following form: $\forall h \in \mathcal{H}, \mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}) \leq f(\mathcal{R}_\ell(h) - \mathcal{R}^*_\ell(\mathcal{H}))$, where f is a non-decreasing real-valued function. Thus, they show that the estimation zero-one loss of h, $\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H})$, is bounded by $f(\epsilon)$ when its surrogate estimation loss, $\mathcal{R}_\ell(h) - \mathcal{R}^*_\ell(\mathcal{H})$, is bounded by ϵ. These guarantees are thus non-asymptotic and depend on the hypothesis set $\mathcal{H}$ considered.

$\mathcal{H}$-consistency bounds are expressed in terms of a quantity depending on the hypothesis set $\mathcal{H}$ and the loss function ℓ called *minimizability gap*, and defined by $\mathcal{M}_\ell(\mathcal{H}) = \mathcal{R}^*_\ell(\mathcal{H}) - \mathbb{E}_x\big[\inf_{h \in \mathcal{H}} \mathbb{E}_y[\ell(h, X, y) \mid X = x]\big]$. By the super-additivity of the infimum, since $\mathcal{R}^*_\ell(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathbb{E}_x\big[\mathbb{E}_y[\ell(h, X, y) \mid X = x]\big]$, the minimizability gap is always non-negative. It measures the difference between the best-in-class expected loss and the expected infimum of

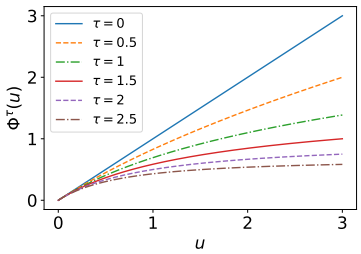

Figure 1. Function Φ^τ with different values of τ.

the pointwise expected loss. When the loss function ℓ only depends on $h(x,\cdot)$ for all h, x, and y, that is $\ell(h,x,y) = \Psi(h(x,1),\ldots,h(x,n),y)$, for some function Ψ, then it is not hard to show that the minimizability gap vanishes for the family of all measurable functions: $\mathcal{M}(\mathcal{H}_{\text{all}}) = 0$ (Steinwart, 2007)[lemma 2.5]. In general, however, the minimizabiliy gap is non-zero for a restricted hypothesis set $\mathcal{H}$ and is therefore important to analyze. Note that the minimizabiliy gap can be upper bounded by the approximation error $\mathcal{A}(\mathcal{H}) = \mathcal{R}^*_\ell(\mathcal{H}) - \mathbb{E}_x\big[\inf_{h\in\mathcal{H}_{\text{all}}} \mathbb{E}_y[\ell(h,X,y) \mid X = x]\big]$. It is however a finer quantity than the approximation error and can thus lead to more favorable guarantees.

Comp-sum losses. In this paper, we derive guarantees for *comp-sum losses*, a family of functions including the logistic loss that is defined via a composition of two functions Φ_1 and Φ_2:

$$\ell^{\text{comp}}_{\Phi_1[\Phi_2]}(h,x,y) = \Phi_1\left(\sum_{y'\neq y} \Phi_2(h(x,y) - h(x,y'))\right), \quad (1)$$

where Φ_2 is a non-increasing function upper bounding $\mathbb{1}_{u\leq 0}$ over $u \in \mathbb{R}$ and Φ_1 a non-decreasing auxiliary function. We will specifically consider $\Phi_2(u) = \exp(-u)$ as with the loss function related to AdaBoost (Freund & Schapire, 1997) and Φ_1 chosen out of the following family of functions Φ^τ, $\tau \geq 0$, defined for all $u \geq 0$ by

$$\Phi^\tau(u) = \begin{cases} \frac{1}{1-\tau}\big((1+u)^{1-\tau} - 1\big) & \tau \geq 0, \tau \neq 1 \\ \log(1+u) & \tau = 1. \end{cases} \quad (2)$$

Figure 1 shows the plot of function Φ^τ for different values of τ. Functions Φ^τ verify the following identities:

$$\frac{\partial \Phi^\tau}{\partial u}(u) = \frac{1}{(1+u)^\tau}, \quad \Phi^\tau(0) = 0. \quad (3)$$

In view of that, by l'Hôpital's rule, Φ^τ is continuous as a function of τ at $\tau = 1$. To simplify the notation, we will use ℓ^{comp}_τ as a short-hand for $\ell^{\text{comp}}_{\Phi_1[\Phi_2]}$ when $\Phi_1 = \Phi^\tau$ and $\Phi_2(u) = \exp(-u)$. $\ell^{\text{comp}}_\tau(h,x,y)$ can be expressed as follows for any h, x, y and $\tau \geq 0$:

$$\ell^{\text{comp}}_\tau(h,x,y) = \Phi^\tau\left(\sum_{y'\in\mathcal{Y}} e^{h(x,y')-h(x,y)} - 1\right)$$
$$= \begin{cases} \frac{1}{1-\tau}\left(\left[\sum_{y'\in\mathcal{Y}} e^{h(x,y')-h(x,y)}\right]^{1-\tau} - 1\right) & \tau \neq 1 \\ \log\left(\sum_{y'\in\mathcal{Y}} e^{h(x,y')-h(x,y)}\right) & \tau = 1. \end{cases} \quad (4)$$

When $\tau = 0$, ℓ^{comp}_τ coincides with the sum-exponential loss (Weston & Watkins, 1998; Awasthi et al., 2022a)

$$\ell^{\text{comp}}_{\tau=0}(h,x,y) = \sum_{y'\neq y} e^{h(x,y')-h(x,y)}.$$

When $\tau = 1$, it coincides with the (multinomial) logistic loss (Verhulst, 1838; 1845; Berkson, 1944; 1951):

$$\ell^{\text{comp}}_{\tau=1}(h,x,y) = -\log\left[\frac{e^{h(x,y)}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')}}\right].$$

For $1 < \tau < 2$, it matches the *generalized cross entropy loss* (Zhang & Sabuncu, 2018):

$$\ell^{\text{comp}}_{1<\tau<2}(h,x,y) = \frac{1}{\tau-1}\left[1 - \left[\frac{e^{h(x,y)}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')}}\right]^{\tau-1}\right],$$

for $\tau = 2$, the *mean absolute error loss* (Ghosh et al., 2017):

$$\ell^{\text{comp}}_{\tau=2}(h,x,y) = 1 - \frac{e^{h(x,y)}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')}}.$$

Since for any $\tau \geq 0$, $\frac{\partial \Phi^\tau}{\partial u}$ is non-increasing and satisfies $\frac{\partial \Phi^\tau}{\partial u}(0) = 1$, $\Phi^\tau(0) = 0$, for any $\tau \geq 0$, Φ^τ is concave, non-decreasing, differentiable, 1-Lipschitz, and satisfies that

$$\forall u \geq 0, \ \Phi^\tau(u) \leq u. \quad (5)$$

3. $\mathcal{H}$-Consistency Bounds for Comp-Sum Losses

In this section, we present and discuss $\mathcal{H}$-consistency bounds for comp-sum losses in the standard multi-class classification scenario. We say that a hypothesis set is *symmetric* when it does not depend on a specific ordering of the classes, that is, when there exists a family $\mathcal{F}$ of functions f mapping from $\mathcal{X}$ to $\mathbb{R}$ such that $\{[h(x,1),\ldots,h(x,c)]\colon h \in \mathcal{H}\} = \{[f_1(x),\ldots,f_c(x)]\colon f_1,\ldots,f_c \in \mathcal{F}\}$, for any $x \in \mathcal{X}$. We 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 widely used in practice are all symmetric and complete.

3.1. $\mathcal{H}$-Consistency Guarantees

The following holds for all comp-sum loss functions and all symmetric and complete hypothesis sets, which includes those typically considered in applications.

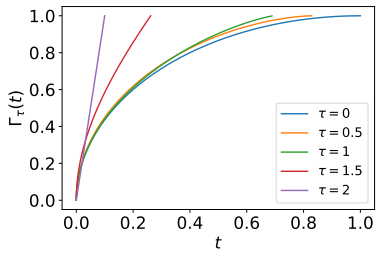

Figure 2. Function Γ_τ with different values of τ for $n = 10$.

Theorem 3.1 (**$\mathcal{H}$-consistency bounds for comp-sum losses**). *Assume that $\mathcal{H}$ is symmetric and complete. Then, for any $\tau \in [0, \infty)$ and any $h \in \mathcal{H}$, the following inequality holds:*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}) \leq \Gamma_\tau\Big(\mathcal{R}_{\ell^{\text{comp}}_\tau}(h) - \mathcal{R}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}) + \mathcal{M}_{\ell^{\text{comp}}_\tau}(\mathcal{H})\Big) - \mathcal{M}_{\ell_{0-1}}(\mathcal{H}),$$

where $\Gamma_\tau(t) = \mathcal{T}_\tau^{-1}(t)$ is the inverse of $\mathcal{H}$-consistency comp-sum transformation, defined for all $\beta \in [0,1]$ by $\mathcal{T}_\tau(\beta) =$

$$\begin{cases} \frac{2^{1-\tau}}{1-\tau}\left[1 - \left[\frac{(1+\beta)^{\frac{1}{2-\tau}} + (1-\beta)^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau}\right] & \tau \in [0,1) \\ \frac{1+\beta}{2}\log[1+\beta] + \frac{1-\beta}{2}\log[1-\beta] & \tau = 1 \\ \frac{1}{(\tau-1)n^{\tau-1}}\left[\left[\frac{(1+\beta)^{\frac{1}{2-\tau}} + (1-\beta)^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau} - 1\right] & \tau \in (1,2) \\ \frac{1}{(\tau-1)n^{\tau-1}}\,\beta & \tau \in [2, +\infty). \end{cases}$$

By l'Hôpital's rule, $\mathcal{T}_\tau$ is continuous as a function of τ at $\tau = 1$. Using the fact that $\lim_{x\to 0^+}\left(a^{\frac{1}{x}} + b^{\frac{1}{x}}\right)^x = \max\{a, b\}$, $\mathcal{T}_\tau$ is continuous as a function of τ at $\tau = 2$. Furthermore, for any $\tau \in [0, +\infty)$, $\mathcal{T}_\tau$ is a convex and increasing function, and satisfies that $\mathcal{T}_\tau(0) = 0$. Note that for the sum-exponential loss ($\tau = 0$) and logistic loss ($\tau = 1$), the expression $\mathcal{T}_\tau$ matches that of their binary $\mathcal{H}$-consistency estimation error transformation $1 - \sqrt{1-t^2}$ and $\frac{1+t}{2}\log(1+t) + \frac{1-t}{2}\log(1-t)$ in the binary classification setting (Awasthi et al., 2022b), which were proven to be tight. We will show that, for these loss functions and in this multi-class classification setting, $\mathcal{T}_\tau$s admit a tight functional forms as well. We illustrate the function Γ_τ with different values of τ in Figure 2.

By using Taylor expansion, $\mathcal{T}_\tau(\beta)$ can be lower bounded by its polynomial approximation with the tightest order as

$$\mathcal{T}_\tau(\beta) \geq \widetilde{\mathcal{T}}_\tau(\beta) = \begin{cases} \frac{\beta^2}{2^\tau(2-\tau)} & \tau \in [0,1) \\ \frac{\beta^2}{2n^{\tau-1}} & \tau \in [1,2) \\ \frac{\beta}{(\tau-1)n^{\tau-1}} & \tau \in [2, +\infty). \end{cases} \tag{6}$$

Accordingly, $\Gamma_\tau(t)$ can be upper bounded by the inverse of $\widetilde{\mathcal{T}}_\tau$, which is denoted by $\widetilde{\Gamma}_\tau(t) = \widetilde{\mathcal{T}}_\tau^{-1}(t)$, as shown below

$$\Gamma_\tau(t) \leq \widetilde{\Gamma}_\tau(t) = \begin{cases} \sqrt{2^\tau(2-\tau)t} & \tau \in [0,1) \\ \sqrt{2n^{\tau-1}t} & \tau \in [1,2) \\ (\tau-1)n^{\tau-1}t & \tau \in [2, +\infty). \end{cases} \tag{7}$$

A detailed derivation is given in Appendix C. The plots of function Γ_τ and their corresponding upper bound $\widetilde{\Gamma}_\tau$ ($n = 10$) are shown in Figure 3, for different values of τ; they illustrate the quality of the approximations via $\widetilde{\Gamma}_\tau$.

Recall that the minimizability gaps vanish when $\mathcal{H}$ is the family of all measurable functions or when $\mathcal{H}$ contains the Bayes predictor. In their absence, the theorem shows that if the estimation loss $(\mathcal{R}_{\ell^{\text{comp}}_\tau}(h) - \mathcal{R}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}))$ is reduced to ϵ, then, for $\tau \in [0, 2)$, in particular for the logistic loss ($\tau = 1$) and the generalized cross-entropy loss ($\tau \in (1, 2)$), modulo a multiplicative constant, the zero-one estimation loss $(\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}))$ is bounded by $\sqrt{\epsilon}$. For the logistic loss, the following guarantee holds for all $h \in \mathcal{H}$:

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}) \leq \sqrt{2\Big(\mathcal{R}_{\ell^{\text{comp}}_1}(h) - \mathcal{R}^*_{\ell^{\text{comp}}_1}(\mathcal{H})\Big)}.$$

The bound is even more favorable for the mean absolute error loss ($\tau = 2$) or for comp-sum losses ℓ^{comp}_τ with $\tau \in (2, +\infty)$ since in that case, modulo a multiplicative constant, the zero-one estimation loss $(\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}))$ is bounded by ϵ. In general, the minimizability gaps are not null however and, in addition to the functional form of Γ_τ, two other key features help compare comp-sum losses: (i) the magnitude of the minimizability gap $\mathcal{M}_{\ell^{\text{comp}}_\tau}(\mathcal{H})$; and (ii) the dependency of the multiplicative constant on the number of classes, which makes it less favorable for $\tau \in (1, +\infty)$. Thus, we will specifically further analyze the minimizability gaps in the next section (Section 4).

The proof of the theorem is given in Appendix B. It consists of using the general $\mathcal{H}$-consistency bound tools given by Awasthi et al. (2022b;a) and of analyzing the calibration gap of the loss function ℓ^{comp}_τ for different values of τ in order to lower bound it in terms of the zero-one loss calibration gap. As pointed out by Awasthi et al. (2022a), deriving such bounds is non-trivial in the multi-class classification setting. In the proof, we specifically choose auxiliary functions $\overline{h}_\mu$ target to the comp-sum losses, which satisfies the property $\sum_{y\in\mathcal{Y}} e^{h(x,y)} = \sum_{y\in\mathcal{Y}} e^{\overline{h}_\mu(x,y)}$. Using this property, we then establish several general lemmas that are applicable to any $\tau \in [0, \infty)$ and are helpful to lower bound the calibration gap of ℓ^{comp}_τ. This is significantly different from the proofs of Awasthi et al. (2022a) whose analysis depends on concrete loss functions case by case. Furthermore, our proof technique actually leads to the tightest bounds as shown below. Our proofs are novel and cover the full comp-sum loss family, which includes the logistic loss. Next, we further prove that the functional form of our bounds $\mathcal{H}$-consistency bounds cannot be improved.

Theorem 3.2 (**Tightness**). *Assume that $\mathcal{H}$ is symmetric and complete. Then, for any $\tau \in [0,1]$ and $\beta \in [0,1]$, there exist a distribution $\mathcal{D}$ and a hypothesis $h \in \mathcal{H}$ such that $\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} + \mathcal{M}_{\ell_{0-1},\mathcal{H}} = \beta$ and $\mathcal{R}_{\ell^{\text{comp}}_\tau}(h) - \mathcal{R}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}) + \mathcal{M}_{\ell^{\text{comp}}_\tau}(\mathcal{H}) = \mathcal{T}_\tau(\beta)$.*

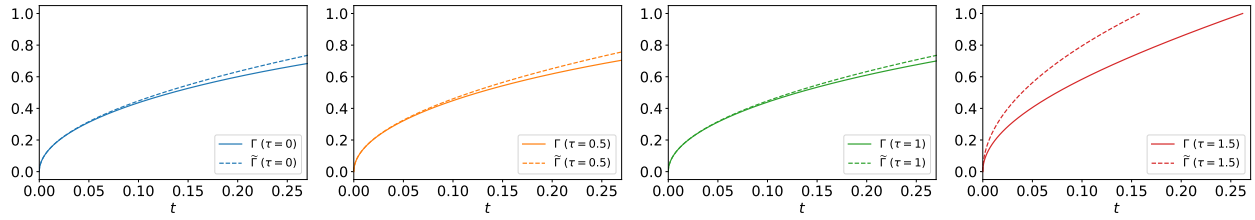

Figure 3. Function Γ_τ and its upper bound $\widetilde{\Gamma}_\tau$ with different values of τ and $n = 10$.

The proof is given in Appendix B. The theorem shows that the bounds given by the $\mathcal{H}$-consistency comp-sum transformation $\mathcal{T}_\tau$, or, equivalently, by its inverse Γ_τ in Theorem 3.1 is tight for any $\tau \in [0, 1]$, which includes as special cases the logistic loss ($\tau = 1$).

3.2. Learning Bounds

Our $\mathcal{H}$-consistency bounds can be used to derive zero-one learning bounds for a hypothesis set $\mathcal{H}$. For a sample size m, let $\mathfrak{R}^\tau_m(\mathcal{H})$ denote the Rademacher complexity of the family of functions $\{(x, y) \mapsto \ell^{\text{comp}}_\tau(h, x, y) : h \in \mathcal{H}\}$ and B_τ an upper bound on the loss ℓ^{comp}_τ.

Theorem 3.3. *With probability at least $1-\delta$ over the draw of a sample S from $\mathcal{D}^m$, the following zero-one loss estimation bound holds for an empirical minimizer $\widehat{h}_S \in \mathcal{H}$ of the comp-sum loss ℓ^{comp}_τ over S:*

$$\mathcal{R}_{\ell_{0-1}}(\widehat{h}_S) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}) \leq \Gamma_\tau\left(\mathcal{M}_{\ell^{\text{comp}}_\tau}(\mathcal{H}) + 4\mathfrak{R}^\tau_m(\mathcal{H}) + 2B_\tau\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right) - \mathcal{M}_{\ell_{0-1}}(\mathcal{H}).$$

The proof is given in Appendix G. To our knowledge, these are the first zero-one estimation loss guarantees for empirical minimizers of a comp-sum loss such as the logistic loss. Our previous comments about the properties of Γ_τ, in particular its functional form or its dependency on the number of classes n, similarly apply here. These are precise bounds that take into account the minimizability gaps.

4. Comparison of Minimizability Gaps

We now further analyze these quantities and make our guarantees even more explicit. Consider a composed loss function defined by $(\Phi_1 \circ \ell_2)(h, x, y)$, for all $h \in \mathcal{H}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, with Φ_1 concave and non-decreasing. Then, by Jensen's inequality, we can write:

$$\begin{aligned}\mathcal{R}^*_{\Phi_1\circ\ell_2}(\mathcal{H}) &= \inf_{h\in\mathcal{H}}\left\{\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\Phi_1\circ\ell_2)(h,x,y)]\right\}\\ &\leq \inf_{h\in\mathcal{H}}\left\{\Phi_1\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell_2(h,x,y)]\right)\right\}\\ &= \Phi_1\left(\inf_{h\in\mathcal{H}}\left\{\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell_2(h,x,y)]\right\}\right)\\ &= \Phi_1\left(\mathcal{R}^*_{\ell_2}(\mathcal{H})\right).\end{aligned} \tag{8}$$

Recall that the comp-sum losses ℓ^{comp}_τ can be written as $\ell^{\text{comp}}_\tau = \Phi^\tau \circ \ell^{\text{comp}}_{\tau=0}$, where $\ell^{\text{comp}}_{\tau=0}(h, x, y) = \sum_{y'\neq y} \exp(h(x, y') - h(x, y))$. Φ^τ is concave since we have $\frac{\partial^2\Phi^\tau}{\partial^2 u}(u) = \frac{-\tau}{(1+u)^{\tau+1}} \leq 0$ for all $\tau \geq 0$ and $u \geq 0$. Using these observations, the following results can be shown.

Theorem 4.1 (Characterization of minimizability gaps - stochastic case). *Assume that $\mathcal{H}$ is symmetric and complete. Then, for the comp-sum losses ℓ^{comp}_τ, the minimizability gaps can be upper bounded as follows:*

$$\mathcal{M}_{\ell^{\text{comp}}_\tau}(\mathcal{H}) \leq \Phi^\tau\left(\mathcal{R}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H})\right) - \mathop{\mathbb{E}}_x[\mathcal{C}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}, x)], \tag{9}$$

where $\mathcal{C}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}, x)$ is given by

$$\begin{cases}\frac{1}{1-\tau}\left(\left[\sum_{y\in\mathcal{Y}} p(x,y)^{\frac{1}{2-\tau}}\right]^{2-\tau} - 1\right) & \tau \geq 0, \tau \neq 1, \tau \neq 2\\ -\sum_{y\in\mathcal{Y}} p(x,y)\log[p(x,y)] & \tau = 1\\ 1 - \max_{y\in\mathcal{Y}} p(x,y) & \tau = 2.\end{cases} \tag{10}$$

Note that the expressions for $\mathcal{C}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}, x)$ in (10) can be formulated in terms of the $(2-\tau)$-Rényi entropy.

Theorem 4.2 (Characterization of minimizability gaps - deterministic case). *Assume that for any $x \in \mathcal{X}$, we have $\{(h(x, 1), \ldots, h(x, n)) : h \in \mathcal{H}\} = [-\Lambda, +\Lambda]^n$. Then, for comp-sum losses ℓ^{comp}_τ and any deterministic distribution, the minimizability gaps can be upper bounded as follows:*

$$\mathcal{M}_{\ell^{\text{comp}}_\tau}(\mathcal{H}) \leq \Phi^\tau\left(\mathcal{R}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H})\right) - \mathcal{C}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}, x), \tag{11}$$

where $\mathcal{C}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}, x)$ is given by

$$\begin{cases}\frac{1}{1-\tau}\left(\left[1 + e^{-2\Lambda}(n-1)\right]^{1-\tau} - 1\right) & \tau \geq 0, \tau \neq 1\\ \log\left[1 + e^{-2\Lambda}(n-1)\right] & \tau = 1.\end{cases} \tag{12}$$

The proofs of these theorems are given in Appendix D. Note that, when $\tau = 0$, $\Phi^\tau(u) = u$ gives the sum exponential loss $\Phi^\tau \circ \ell^{\text{comp}}_{\tau=0} = \ell^{\text{comp}}_{\tau=0}$. For deterministic distributions, by (12), we obtain $\mathcal{C}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H}, x) = e^{-2\Lambda}(n-1)$. Therefore, (12) can be rewritten as $\mathcal{C}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}, x) = \Phi^\tau\left(\mathcal{C}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H}, x)\right)$. Thus, inequality (11) can be rewritten as follows:

$$\mathcal{M}_{\ell^{\text{comp}}_\tau}(\mathcal{H}) \leq \Phi^\tau\left(\mathcal{R}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H})\right) - \Phi^\tau\left(\mathcal{C}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H}, x)\right). \tag{13}$$

We will denote the right-hand side by $\widetilde{\mathcal{M}}_{\ell_\tau^{\text{comp}}}(\mathcal{H})$, $\widetilde{\mathcal{M}}_{\ell_\tau^{\text{comp}}}(\mathcal{H}) = \Phi^\tau\Big(\mathcal{R}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H})\Big) - \Phi^\tau\Big(\mathcal{C}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H},x)\Big)$. Note that we always have $\mathcal{R}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H}) \geq \mathbb{E}_x\Big[\mathcal{C}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H},x)\Big]$. Here, $\mathbb{E}_x[\mathcal{C}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H},x)] = \mathcal{C}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H},x)$ since $\mathcal{C}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H},x)$ is independent of x as shown in (12). Then, (13) can be used to compare the minimizability gaps for different τ.

Lemma 4.3. *For any $u_1 \geq u_2 \geq 0$, $\Phi^\tau(u_1) - \Phi^\tau(u_2)$ is non-increasing with respect to τ.*

The proof is given in Appendix E. Lemma 4.3 implies that $\widetilde{\mathcal{M}}_{\ell_\tau^{\text{comp}}}(\mathcal{H})$ is a non-increasing function of τ. Thus, given a hypothesis set $\mathcal{H}$, we have:

$$\widetilde{\mathcal{M}}_{\ell_{\tau=0}}(\mathcal{H}) \geq \widetilde{\mathcal{M}}_{\ell_{\tau=1}}(\mathcal{H}) \geq \widetilde{\mathcal{M}}_{\ell_{1<\tau<2}}(\mathcal{H}) \geq \widetilde{\mathcal{M}}_{\ell_{\tau=2}}(\mathcal{H}). \quad (14)$$

By Section 2, these minimizability gaps specifically correspond to that of sum-exponential loss ($\tau = 0$), logistic loss ($\tau = 1$), generalized cross-entropy loss ($1 < \tau < 2$) and mean absolute error loss ($\tau = 2$) respectively. Note that for those loss functions, by Theorem 3.1, when the estimation error $\mathcal{R}_{\ell_\tau^{\text{comp}}}(h) - \mathcal{R}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H})$ is minimized to zero, the estimation error of zero-one classification loss is upper bounded by $\widetilde{\Gamma}_\tau(\mathcal{M}_{\ell_\tau})$. Therefore, (14) combined with the form of $\widetilde{\Gamma}_\tau$ helps compare the sum-exponential loss ($\tau = 0$), logistic loss ($\tau = 0$), generalized cross-entropy loss ($1 < \tau < 2$) and mean absolute error loss ($\tau = 2$) in practice. See Section 6.1 for a discussion of the empirical results in light of these theoretical findings.

5. Smooth Adversarial Comp-Sum Losses

A recent challenge in the application of neural networks is their robustness to imperceptible perturbations (Szegedy et al., 2013). While neural networks trained on large datasets have achieved breakthroughs in speech and visual recognition tasks in recent years (Sutskever et al., 2014; Krizhevsky et al., 2012), their accuracy remains substantially lower in the presence of such perturbations even for state-of-the-art robust algorithms. One key factor in the design of robust algorithms is the choice of the surrogate loss function used for training since directly optimizing the target adversarial zero-one loss with most hypothesis sets is NP-hard. To tackle this problem, we introduce a family of loss functions designed for adversarial robustness that we call *smooth adversarial comp-sum loss functions*. These are loss functions obtained by augmenting comp-sum losses with a natural corresponding smooth term. We show that these loss functions are beneficial in the adversarial setting by proving that they admit $\mathcal{H}$-consistency bounds. This leads to a family of algorithms for adversarial robustness that consist of minimizing a regularized smooth adversarial comp-sum loss.

5.1. Definition

In adversarial robustness, the target adversarial zero-one classification loss is defined as the worst loss incurred over an ℓ_p perturbation ball of x with perturbation size γ, $p \in [1, +\infty]$, $\mathsf{B}_p(x,\gamma) = \{x' : \|x - x'\|_p \leq \gamma\}$:

$$\ell_\gamma(h,x,y) = \sup_{x' \in \mathsf{B}_p(x,\gamma)} \ell_{0-1}(h,x',y).$$

We first introduce the adversarial comp-sum ρ-margin losses, which is defined as the supremum based counterpart of comp-sum losses (1) with $\Phi_1 = \Phi^\tau$ and $\Phi_2(u) = \Phi_\rho(u) = \min\Big\{\max\Big\{0, 1 - \frac{u}{\rho}\Big\}, 1\Big\}$, the ρ-margin loss function (see for example (Mohri et al., 2018)):

$$\widetilde{\ell}^{\text{comp}}_{\tau,\rho}(h,x,y) = \sup_{x' : \|x-x'\|_p \leq \gamma} \Phi^\tau\left(\sum_{y' \neq y} \Phi_\rho(h(x',y') - h(x',y))\right).$$

In the next section, we will show that $\widetilde{\ell}^{\text{comp}}_{\tau,\rho}$ admits an $\mathcal{H}$-consistency bound with respect to the adversarial zero-one loss ℓ_γ. Since Φ_ρ is not-convex, we will further derive the *smooth adversarial comp-sum loss* based on $\widetilde{\ell}^{\text{comp}}_{\tau,\rho}$, that has similar $\mathcal{H}$-consistency guarantees and is better to optimize. By the expression of the derivative of Φ^τ in (3), for all $\tau \geq 0$ and $u \geq 0$, we have $\Big|\frac{\partial \Phi^\tau}{\partial u}(u)\Big| = \frac{1}{(1+u)^\tau} \leq 1$, thus Φ^τ is 1-Lipschitz over $\mathbb{R}_+$. Define $\Delta_h(x,y,y') = h(x,y) - h(x,y')$ and let $\overline{\Delta}_h(x,y)$ denote the $(n-1)$-dimensional vector $(\Delta_h(x,y,1), \ldots, \Delta_h(x,y,y-1), \Delta_h(x,y,y+1), \ldots, \Delta_h(x,y,n))$. For any $\tau \geq 0$, since Φ^τ is 1-Lipschitz and non-decreasing, we have:

$$\widetilde{\ell}^{\text{comp}}_{\tau,\rho}(h,x,y) - \ell^{\text{comp}}_{\tau,\rho}(h,x,y) \leq \sup_{x' \in \mathsf{B}(x,\gamma)} \sum_{y' \neq y} \Phi_\rho(-\Delta_h(x',y,y')) - \Phi_\rho(-\Delta_h(x,y,y')).$$

Since $\Phi_\rho(u)$ is $\frac{1}{\rho}$-Lipschitz, by the Cauchy-Schwarz inequality, for any $\nu \geq \frac{\sqrt{n-1}}{\rho} \geq \frac{1}{\rho}$, we have

$$\begin{aligned}
&\widetilde{\ell}^{\text{comp}}_{\tau,\rho}(h,x,y) \\
&\leq \ell^{\text{comp}}_{\tau,\rho}(h,x,y) + \nu \sup_{x' \in \mathsf{B}(x,\gamma)} \left\|\overline{\Delta}_h(x',y) - \overline{\Delta}_h(x,y)\right\|_2 \\
&\leq \ell^{\text{comp}}_{\tau}\left(\frac{h}{\rho},x,y\right) + \nu \sup_{x' \in \mathsf{B}(x,\gamma)} \left\|\overline{\Delta}_h(x',y) - \overline{\Delta}_h(x,y)\right\|_2,
\end{aligned}$$

where we used the inequality $\exp(-u/\rho) \geq \Phi_\rho(u)$. We will refer to a loss function defined by the last expression as a *smooth adversarial comp-sum loss* and denote it by $\ell^{\text{comp}}_{\text{smooth}}$. In the next section, we will provide strong $\mathcal{H}$-consistency guarantees for $\ell^{\text{comp}}_{\text{smooth}}$.

5.2. Adversarial $\mathcal{H}$-Consistency Guarantees

To derive guarantees for our smooth adversarial comp-sum loss, we first prove an adversarial $\mathcal{H}$-consistency bound

for adversarial comp-sum ρ-margin losses $\widetilde{\ell}^{\text{comp}}_{\tau,\rho}$ for any symmetric and *locally ρ-consistent* hypothesis set.

Definition 5.1. We say that a hypothesis set $\mathcal{H}$ is *locally ρ-consistent* if for any $x \in \mathcal{X}$, there exists a hypothesis $h \in \mathcal{H}$ such that $\inf_{x':\|x-x'\|\leq\gamma} |h(x',i) - h(x',j)| \geq \rho > 0$ for any $i \neq j \in \mathcal{Y}$ and for any $x' \in \{x' : \|x - x'\| \leq \gamma\}$, $\{h(x',y) : y \in \mathcal{Y}\}$ has the same ordering.

Common hypothesis sets used in practice, such as the family of linear models, that of neural networks and of course that of all measurable functions are all locally ρ-consistent for some $\rho > 0$. The guarantees given in the following result are thus general and widely applicable.

Theorem 5.2 ($\mathcal{H}$**-consistency bound of** $\widetilde{\ell}^{\text{comp}}_{\tau,\rho}$**).** *Assume that $\mathcal{H}$ is symmetric and locally ρ-consistent. Then, for any choice of the hyperparameters $\tau, \rho > 0$, any hypothesis $h \in \mathcal{H}$, the following inequality holds:*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma}(\mathcal{H}) \leq \tag{15}$$
$$\Phi^\tau(1)\Big[\mathcal{R}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(h) - \mathcal{R}^*_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H})\Big] + \mathcal{M}_{\ell_\gamma}(\mathcal{H}).$$

The proof is given in Appendix F. Using the inequality $\ell^{\text{comp}}_{\text{smooth}} \geq \widetilde{\ell}^{\text{comp}}_{\tau,\rho}$ yields the following similar guarantees for smooth adversarial comp-sum loss under the same condition of hypothesis sets.

Corollary 5.3 (Guarantees for smooth adversarial comp-sum losses). *Assume that $\mathcal{H}$ is symmetric and locally ρ-consistent. Then, for any choice of the hyperparameters $\tau, \rho > 0$, any hypothesis $h \in \mathcal{H}$, the following inequality holds:*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma}(\mathcal{H}) \leq \tag{16}$$
$$\Phi^\tau(1)\Big[\mathcal{R}_{\ell^{\text{comp}}_{\text{smooth}}}(h) - \mathcal{R}^*_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H})\Big] + \mathcal{M}_{\ell_\gamma}(\mathcal{H}).$$

This is the first $\mathcal{H}$-consistency bound for the comp-sum loss in the adversarial robustness. As with the non-adversarial scenario in Section 3, the minimizability gaps appearing in those bounds in Theorem 5.2 and Corollary 5.3 actually equal to zero in most common cases. More precisely, Theorem 5.2 guarantees $\mathcal{H}$-consistency for distributions such that the minimizability gaps vanish:

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma}(\mathcal{H}) \leq \Phi^\tau(1)\Big[\mathcal{R}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(h) - \mathcal{R}^*_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H})\Big].$$

For $\tau \in [0, \infty)$ and $\rho > 0$, if the estimation loss $(\mathcal{R}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(h) - \mathcal{R}^*_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H}))$ is reduced to ϵ, then, the adversarial zero-one estimation loss $(\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma}(\mathcal{H}))$ is bounded by ϵ modulo a multiplicative constant. A similar guarantee applies to smooth adversarial comp-sum loss as well. These guarantees suggest an adversarial robustness algorithm that consists of minimizing a regularized empirical smooth adversarial comp-sum loss, $\ell^{\text{comp}}_{\text{smooth}}$. We call this algorithm ADV-COMP-SUM. In the next section, we report empirical results for ADV-COMP-SUM, demonstrating that it significantly outperform the current state-of-the-art loss/algorithm TRADES.

6. Experiments

We first report empirical results comparing the performance of comp-sum losses for different values of τ. Next, we report a series of empirical results comparing our adversarial robust algorithm ADV-COMP-SUM with several baselines.

6.1. Standard Multi-Class Classification

We compared comp-sum losses with different values of τ on CIFAR-10 and CIFAR-100 datasets (Krizhevsky, 2009). All models were trained via Stochastic Gradient Descent (SGD) with Nesterov momentum (Nesterov, 1983), batch size 1,024 and weight decay 1×10^{-4}. We used ResNet-34 and trained for 200 epochs using the cosine decay learning rate schedule (Loshchilov & Hutter, 2016) without restarts. The initial learning rate was selected from $\{0.01, 0.1, 1.0\}$; the best model is reported for each surrogate loss. We report the zero-one classification accuracy of the models and the standard deviation for three trials.

Table 1. Zero-one classification accuracy for comp-sum surrogates; mean ± standard deviation over three runs for different τ.

τ	0	0.5	1.0	1.5	2.0
CIFAR-10	87.37	90.28	92.59	92.03	90.35
±	0.57	0.10	0.10	0.08	0.24
CIFAR-100	57.87	65.52	70.93	69.87	8.99
±	0.60	0.34	0.34	0.39	0.98

Table 1 shows that on CIFAR-10 and CIFAR-100, the logistic loss ($\tau = 1$) outperforms the comp-sum loss ($\tau = 0.5$) and, by an even larger margin, the sum-exponential loss ($\tau = 0$). This is consistent with our theoretical analysis based on $\mathcal{H}$-consistency bounds in Theorem 3.1 since all three losses have the same square-root functional form and since, by Lemma 4.3 and (7), the magnitude of the minimizability gap decreases with τ.

Table 1 also shows that on CIFAR-10 and CIFAR-100, the logistic loss ($\tau = 1$) and the generalized cross-entropy loss ($\tau = 1.5$) achieve relatively close results that are clearly superior to that of mean absolute error loss ($\tau = 2$). This empirical observation agrees with our theoretically analysis based on their $\mathcal{H}$-consistency bounds (Theorem 3.1): by Lemma 4.3, the minimizability gap of $\tau = 1.5$ and $\tau = 2$ is smaller than that of $\tau = 1$; however, by (7), the dependency of the multiplicative constant on the number of classes appears for $\tau = 1.5$ in the form of $\sqrt{n}$, which makes the generalized cross-entropy loss less favorable, and for $\tau = 2$ in the form of n, which makes the mean absolute error loss least favorable. Another reason for the inferior performance of the mean absolute error loss ($\tau = 2$) is that, as observed in our experiments, it is difficult to optimize in practice,

using deep neural networks on complex datasets. This has also been previously reported by Zhang & Sabuncu (2018). In fact, the mean absolute error loss can be formulated as an ℓ_1-distance and is therefore not smooth; but it has the advantage of robustness, as shown in (Ghosh et al., 2017).

6.2. Adversarial Multi-Class Classification

Here, we report empirical results for our adversarial robustness algorithm ADV-COMP-SUM on CIFAR-10, CIFAR-100 (Krizhevsky, 2009) and SVHN (Netzer et al., 2011) datasets. No generated data or extra data was used.

Experimental settings. We followed exactly the experimental settings of Gowal et al. (2020) and adopted precisely the same training procedure and neural network architectures, which are WideResNet (WRN) (Zagoruyko & Komodakis, 2016) with SiLU activations (Hendrycks & Gimpel, 2016). Here, WRN-n-k denotes a residual network with n convolutional layers and a widening factor k. For CIFAR-10 and CIFAR-100, the simple data augmentations, 4-pixel padding with 32×32 random crops and random horizontal flips, were applied. We used 10-step Projected Gradient-Descent (PGD) with random starts to generate training attacks. All models were trained via Stochastic Gradient Descent (SGD) with Nesterov momentum (Nesterov, 1983), batch size $1{,}024$ and weight decay 5×10^{-4}. We trained for 400 epochs using the cosine decay learning rate schedule (Loshchilov & Hutter, 2016) without restarts. The initial learning rate is set to 0.4. We used model weight averaging (Izmailov et al., 2018) with decay rate 0.9975. For TRADES, we adopted exactly the same setup as Gowal et al. (2020). For our smooth adversarial comp-sum losses, we set both ρ and ν to 1 by default. In practice, they can be selected by cross-validation and that could potentially lead to better performance. The per-epoch computational cost of our method is similar to that of TRADES.

Evaluation. We used early stopping on a held-out validation set of $1{,}024$ samples by evaluating its robust accuracy throughout training with 40-step PGD on the margin loss, denoted by $\text{PGD}^{40}_{\text{margin}}$, and selecting the best check-point (Rice et al., 2020). We report the *clean accuracy*, that is the standard classification accuracy on the test set, and the robust accuracy with ℓ_∞-norm perturbations bounded by $\gamma = 8/255$ under PGD attack, measured by $\text{PGD}^{40}_{\text{margin}}$ on the full test set, as well as under AutoAttack (Croce & Hein, 2020) (https://github.com/fra31/auto-attack), the state-of-the-art attack for measuring empirically adversarial robustness. We averaged accuracies over three runs and report the standard deviation for both ADV-COMP-SUM and TRADES, reproducing the results reported for TRADES in (Gowal et al., 2020).

Results. Table 2 shows that ADV-COMP-SUM outperforms TRADES on CIFAR-10 for all the neural network architectures adopted (WRN-70-16, WRN-34-20 and WRN-28-10). Here, ADV-COMP-SUM was implemented with $\tau = 0.4$. Other common choices of τ yield similar results, including $\tau = 1$ (logistic loss). In all the settings, robust accuracy under AutoAttack is higher by at least 0.6% for ADV-COMP-SUM, by at least 1.36% under the $\text{PGD}^{40}_{\text{margin}}$ attack.

It is worth pointing out that the improvement in robustness accuracy for our models does not come at the expense of a worse clean accuracy than TRADES. In fact, ADV-COMP-SUM consistently outperforms TRADES for the clean accuracy as well. For the largest model WRN-70-16, the improvement is over 0.8%. For completeness, we also include in Table 2 the results for some other well-known adversarial defense models. ADV-COMP-SUM with the smallest model WRN-28-10 surpasses (Pang et al., 2020a; Rice et al., 2020; Qin et al., 2019). (Wu et al., 2020) is significantly outperformed by ADV-COMP-SUM with a slightly larger model WRN-34-20, by more than 1.2% in the robust accuracy and also in the clean accuracy.

To show the generality of our approach, we carried out experiments with other datasets, including CIFAR-100 and SVHN. For WRN-70-16 on CIFAR-100, ADV-COMP-SUM outperforms TRADES by 1.12% in the robust accuracy and 2.54% in the clean accuracy. For WRN-34-20 on SVHN, ADV-COMP-SUM also outperforms TRADES by 0.29% in the robust accuracy and 0.95% in the clean accuracy.

Let us underscore that outperforming the state-of-the-art results of Gowal et al. (2020) in the same scenario and without resorting to additional unlabeled data has turned out to be very challenging: despite the large research emphasis on this topic in the last several years and the many publications, none was reported to surpass that performance, using an alternative surrogate loss.

7. Discussion

Applications of $\mathcal{H}$-consistency bounds. Given a hypothesis set $\mathcal{H}$, our quantitative $\mathcal{H}$-consistency bounds can help select the most favorable surrogate loss, which depends on (i) the functional form of the $\mathcal{H}$-consistency bound: for instance, the bound for the mean absolute error loss exhibits a linear dependency, while that of the logistic loss and generalized cross-entropy losses exhibit a square-root dependency, resulting in a less favorable convergence rate; (ii) the smoothness of the loss and, more generally, its optimization properties; for example, the mean absolute error loss is less smooth than the logistic loss, and surrogate losses with more favorable bounds may lead to more challenging optimizations; in fact, the zero-one loss serves as its own surrogate with the tightest bound for any hypothesis set, but is known to result in NP-complete optimization problems for many common choices of $\mathcal{H}$; (iii) approximation properties

Table 2. Clean accuracy and robust accuracy under $\text{PGD}^{40}_{\text{margin}}$ and AutoAttack; mean ± standard deviation over three runs for both ADV-COMP-SUM and the state-of-the-art TRADES in (Gowal et al., 2020). Accuracies of some well-known adversarial defense models are included for completeness. ADV-COMP-SUM significantly outperforms TRADES for both robust and clean accuracy in all the settings.

Method	Dataset	Clean	$\text{PGD}^{40}_{\text{margin}}$	AutoAttack
Gowal et al. (2020) (WRN-70-16)	CIFAR-10	85.34 ± 0.04	57.90 ± 0.13	57.05 ± 0.17
ADV-COMP-SUM (WRN-70-16)		**86.16 ± 0.16**	**59.35 ± 0.07**	**57.77 ± 0.08**
Gowal et al. (2020) (WRN-34-20)		85.21 ± 0.16	57.54 ± 0.18	56.70 ± 0.14
ADV-COMP-SUM (WRN-34-20)		**85.59 ± 0.17**	**58.92 ± 0.06**	**57.41 ± 0.06**
Gowal et al. (2020) (WRN-28-10)		84.33 ± 0.18	55.92 ± 0.20	55.19 ± 0.23
ADV-COMP-SUM (WRN-28-10)		**84.50 ± 0.33**	**57.28 ± 0.05**	**55.79 ± 0.06**
Pang et al. (2020a) (WRN-34-20)		86.43	—	54.39
Rice et al. (2020) (WRN-34-20)		85.34	—	53.42
Wu et al. (2020) (WRN-34-10)		85.36	—	56.17
Qin et al. (2019) (WRN-40-8)		86.28	—	52.84
Gowal et al. (2020) (WRN-70-16)	CIFAR-100	60.56 ± 0.31	31.39 ± 0.19	29.93 ± 0.14
ADV-COMP-SUM (WRN-70-16)		**63.10 ± 0.24**	**33.76 ± 0.18**	**31.05 ± 0.15**
Gowal et al. (2020) (WRN-34-20)	SVHN	93.03 ± 0.13	61.01 ± 0.16	57.84 ± 0.19
ADV-COMP-SUM (WRN-34-20)		**93.98 ± 0.12**	**62.97 ± 0.05**	**58.13 ± 0.12**

of the surrogate loss function: for instance, given a choice of $\mathcal{H}$, the minimizability gap for a surrogate loss may be more or less favorable; (iv) the dependency of the multiplicative constant on the number of classes: for example, the linear dependency of n in the bound for the mean absolute error loss makes it less favorable than the logistic loss.

Another application is the derivation of generalization bounds for surrogate loss minimizers (see Theorem 3.3), expressed in terms of the quantities discussed above.

Concurrent work. The concurrent and independent study of Zheng et al. (2023) also provides an $\mathcal{H}$-consistency bound for the logistic loss. Their bound holds for the special case of $\mathcal{H}$ being a constrained linear hypothesis set, subject to an additional assumption on the distribution. In contrast, our bounds do not require any distributional assumption. However, it should be noted that our results are only applicable to complete hypothesis sets. In upcoming work, we present $\mathcal{H}$-consistency bounds for non-complete hypothesis sets and arbitrary distributions.

Future work. In addition to the extension to non-complete hypothesis sets just mentioned, it would be valuable to investigate the application or generalization of $\mathcal{H}$-consistency bounds in scenarios involving noisy labels (Ghosh et al., 2017; Zhang & Sabuncu, 2018). For comp-sum losses, this paper focuses on the case where Φ_2 is the exponential loss and Φ_1 is based on (2). This includes the cross-entropy loss (or logistic loss), generalized cross-entropy, the mean absolute error and other cross-entropy-like functions, which are the most widely used ones in the family of comp-sum losses. The study of other such loss functions and the comparison with other families of multi-class loss functions (Awasthi et al., 2022a) is left to the future work. Although our algorithm demonstrates improvements over the current state-of-the-art technique, adversarial robustness remains a challenging problem. A key issue seems to be that of generalization for complex families of neural networks (see for example (Awasthi, Frank, and Mohri, 2020)). A more detailed study of that problem might help enhance the performance of our algorithm. Finally, in addition to their immediate implications, our results and techniques have broader applications in analyzing surrogate losses and algorithms across different learning scenarios. For instance, they can be used in the context of ranking, as demonstrated in recent work by Mao, Mohri, and Zhong (2023). Furthermore, they can be extended to address the challenges of learning with abstention (Cortes, DeSalvo, and Mohri, 2016b;a). Additionally, our findings can be valuable in non-i.i.d. learning settings, such as drifting (Mohri & Medina, 2012) or time series prediction (Kuznetsov & Mohri, 2018; 2020).

8. Conclusion

We presented a detailed analysis of the theoretical properties of a family of surrogate losses that includes the logistic loss (or cross-entropy with the softmax). These are more precise and more informative guarantees than Bayes consistency since they are non-asymptotic and specific to the hypothesis set used. Our bounds are tight and can be made more explicit, when combined with our analysis of minimizability gaps. These inequalities can help compare different surrogate losses and evaluate their advantages in different scenarios. We showcased one application of this analysis by extending comp-sum losses to the adversarial robustness setting, which yields principled surrogate losses and algorithms for that scenario. We believe that our analysis can be helpful to the design of algorithms in many other scenarios.

Acknowledgements

Part of the work of A. Mao and Y. Zhong was done during their internship at Google Research.

References

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

Alayrac, J.-B., Uesato, J., Huang, P.-S., Fawzi, A., Stanforth, R., and Kohli, P. Are labels required for improving adversarial robustness? In *Advances in Neural Information Processing Systems*, 2019.

Andriushchenko, M. and Flammarion, N. Understanding and improving fast adversarial training. In *Advances in Neural Information Processing Systems*, pp. 16048–16059, 2020.

Ashtiani, H., Pathak, V., and Urner, R. Black-box certification and learning under adversarial perturbations. In *International Conference on Machine Learning*, pp. 388–398, 2020.

Attias, I. and Hanneke, S. Adversarially robust learning of real-valued functions. *arXiv preprint arXiv:2206.12977*, 2022.

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pp. 162–183, 2019.

Attias, I., Hanneke, S., and Mansour, Y. A characterization of semi-supervised adversarially robust pac learnability. In *Advances in Neural Information Processing Systems*, 2022a.

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for adversarially robust learning. *The Journal of Machine Learning Research*, 23(1):7897–7927, 2022b.

Awasthi, P., Dutta, A., and Vijayaraghavan, A. On robustness to adversarial examples and polynomial optimization. In *Advances in Neural Information Processing Systems*, pp. 13737–13747, 2019.

Awasthi, P., Frank, N., and Mohri, M. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pp. 431–441, 2020.

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

Awasthi, P., Frank, N., and Mohri, M. On the existence of the adversarial bayes classifier. In *Advances in Neural Information Processing Systems*, pp. 2978–2990, 2021b.

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

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. Multi-class $\mathcal{H}$-consistency bounds. In *Advances in neural information processing systems*, 2022a.

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

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

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

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

Bartlett, P., Bubeck, S., and Cherapanamjeri, Y. Adversarial examples in multi-layer random relu networks. In *Advances in Neural Information Processing Systems*, pp. 9241–9252, 2021.

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

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

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

Bhattacharjee, R., Jha, S., and Chaudhuri, K. Sample complexity of robust linear classification on separated data. In *International Conference on Machine Learning*, pp. 884–893, 2021.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402, 2013.

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

Bubeck, S. and Sellke, M. A universal law of robustness via isoperimetry. In *Advances in Neural Information Processing Systems*, 2021.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. Adversarial examples from cryptographic pseudo-random generators. *arXiv preprint arXiv:1811.06418*, 2018.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pp. 831–840, 2019.

Bubeck, S., Cherapanamjeri, Y., Gidel, G., and Tachet des Combes, R. A single gradient step finds adversarial examples on random two-layers neural networks. In *Advances in Neural Information Processing Systems*, pp. 10081–10091, 2021.

Cai, Q.-Z., Liu, C., and Song, D. Curriculum adversarial training. In *International Joint Conference on Artificial Intelligence*, pp. 3740–3747, 2018.

Cao, Y., Cai, T., Feng, L., Gu, L., Gu, J., An, B., Niu, G., and Sugiyama, M. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In *Advances in Neural Information Processing Systems*, pp. 521–534, 2022.

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

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. In *Advances in neural information processing systems*, 2019.

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

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

Cheng, M., Lei, Q., Chen, P. Y., Dhillon, I., and Hsieh, C. J. Cat: Customized adversarial training for improved robustness. In *International Joint Conference on Artificial Intelligence*, pp. 673–679, 2022.

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

Cortes, C., DeSalvo, G., and Mohri, M. Boosting with abstention. In *Advances in Neural Information Processing Systems*, 2016a.

Cortes, C., DeSalvo, G., and Mohri, M. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pp. 67–82, 2016b.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216, 2020.

Cullina, D., Bhagoji, A. N., and Mittal, P. Pac-learning in the presence of adversaries. *Advances in Neural Information Processing Systems*, 2018.

Dan, C., Wei, Y., and Ravikumar, P. Sharp statistical guarantees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pp. 2345–2355, 2020.

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

Diakonikolas, I., Kane, D. M., and Manurangsi, P. The complexity of adversarially robust proper learning of halfspaces with agnostic noise. *arXiv preprint arXiv:2007.15220*, 2020.

Ding, G. W., Sharma, Y., Lui, K. Y. C., and Huang, R. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2022.

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

Duchi, J. C., Mackey, L. W., and Jordan, M. I. On the consistency of ranking algorithms. In *International Conference on Machine Learning*, 2010.

Feige, U., Mansour, Y., and Schapire, R. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pp. 637–657, 2015.

Feige, U., Mansour, Y., and Schapire, R. E. Robust inference for multiclass classification. In *Algorithmic Learning Theory*, pp. 368–386, 2018.

Finocchiaro, J., Frongillo, R., and Waggoner, B. An embedding framework for consistent polyhedral surrogates. In *Advances in neural information processing systems*, 2019.

Finocchiaro, J., Frongillo, R. M., and Waggoner, B. An embedding framework for the design and analysis of consistent polyhedral surrogates. *arXiv preprint arXiv:2206.14707*, 2022.

Frank, N. and Niles-Weed, J. The adversarial consistency of surrogate risks for binary classification. *arXiv preprint arXiv:2305.09956*, 2023.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Frongillo, R. and Waggoner, B. Surrogate regret bounds for polyhedral losses. In *Advances in Neural Information Processing Systems*, pp. 21569–21580, 2021.

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

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

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

Goldblum, M., Fowl, L., Feizi, S., and Goldstein, T. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3996–4003, 2020.

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

Gowal, S., Qin, C., Uesato, J., Mann, T., and Kohli, P. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

Guo, J.-Q., Teng, M.-Z., Gao, W., and Zhou, Z.-H. Fast provably robust decision trees and boosting. In *International Conference on Machine Learning*, pp. 8127–8144, 2022.

Guo, M., Yang, Y., Xu, R., Liu, Z., and Lin, D. When nas meets robustness: In search of robust architectures against adversarial attacks. In *Conference on Computer Vision and Pattern Recognition*, pp. 631–640, 2020.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. P., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 876–885, 2018.

Jin, G., Yi, X., Huang, W., Schewe, S., and Huang, X. Enhancing adversarial training with second-order statistics of weights. In *Conference on Computer Vision and Pattern Recognition*, pp. 15273–15283, 2022.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Khim, J. and Loh, P.-L. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Kontorovich, A. and Attias, I. Fat-shattering dimension of k-fold maxima. *arXiv preprint arXiv:2110.04763*, 2021.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Kuznetsov, V. and Mohri, M. Theory and algorithms for forecasting time series. *CoRR*, abs/1803.05814, 2018.

Kuznetsov, V. and Mohri, M. Discrepancy-based theory and algorithms for forecasting non-stationary time series. *Annals of Mathematics and Artificial Intelligence*, 88(4): 367–399, 2020.

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

Lee, S., Lee, H., and Yoon, S. Adversarial vertex mixup: Toward better adversarially robust generalization. In *Conference on Computer Vision and Pattern Recognition*, pp. 272–281, 2020.

Lee, Y., Lin, Y., and Wahba, G. 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.

Levi, M., Attias, I., and Kontorovich, A. Domain invariant adversarial learning. *Transactions of Machine Learning Research*, 2022.

Li, J. D. and Telgarsky, M. On achieving optimal adversarial test error. In *International Conference on Learning Representations*, 2023.

Liu, A., Tang, S., Liu, X., Chen, X., Huang, L., Tu, Z., Song, D., and Tao, D. Towards defending multiple adversarial perturbations via gated batch normalization. *arXiv preprint arXiv:2012.01654*, 2020.

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

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

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

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

Mao, A., Mohri, M., and Zhong, Y. $\mathcal{H}$-consistency bounds for pairwise misranking loss surrogates. In *International conference on Machine learning*, 2023.

Meunier, L., Ettedgui, R., Pinot, R., Chevaleyre, Y., and Atif, J. Towards consistency in adversarial classification. In *Advances in Neural Information Processing Systems*, 2022.

Mohri, M. and Medina, A. M. New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory*, pp. 124–138, 2012.

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

Montasser, O., Hanneke, S., and Srebro, N. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pp. 2512–2530, 2019.

Montasser, O., Goel, S., Diakonikolas, I., and Srebro, N. Efficiently learning adversarially robust halfspaces with noise. In *International Conference on Machine Learning*, pp. 7010–7021, 2020a.

Montasser, O., Hanneke, S., and Srebro, N. Reducing adversarially robust learning to non-robust pac learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14626–14637, 2020b.

Montasser, O., Hanneke, S., and Srebro, N. Adversarially robust learning with unknown perturbation sets. In *Conference on Learning Theory*, pp. 3452–3482, 2021.

Montasser, O., Hanneke, S., and Srebro, N. Transductive robust learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 11461–11471, 2022.

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

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

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *Advances in Neural Information Processing Systems*, 2011.

Nueve, E. B., Frongillo, R., and Finocchiaro, J. J. The structured abstain problem and the lovász hinge. In *Conference on Learning Theory*, pp. 3718–3740, 2022.

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

Pang, T., Xu, K., Du, C., Chen, N., and Zhu, J. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pp. 4970–4979, 2019.

Pang, T., Yang, X., Dong, Y., Su, H., and Zhu, J. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020a.

Pang, T., Yang, X., Dong, Y., Xu, K., Zhu, J., and Su, H. Boosting adversarial training with hypersphere embedding. In *Advances in Neural Information Processing Systems*, pp. 7779–7792, 2020b.

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

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

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

Prabhu, V. U., Yap, D. A., Xu, J., and Whaley, J. Understanding adversarial robustness through loss landscape geometries. *arXiv preprint arXiv:1907.09061*, 2019.

Qian, Z., Zhang, S., Huang, K., Wang, Q., Zhang, R., and Yi, X. Improving model robustness with latent distribution locally and globally. *arXiv preprint arXiv:2107.04401*, 2021.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., and Kohli, P. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, 2019.

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

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

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

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

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

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021a.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. A. Data augmentation can improve robustness. In *Advances in Neural Information Processing Systems*, pp. 29935–29948, 2021b.

Rice, L., Wong, E., and Kolter, J. Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104, 2020.

Robey, A., Chamon, L., Pappas, G., Hassani, H., and Ribeiro, A. Adversarial robustness with semi-infinite constrained learning. In *Advances in Neural Information Processing Systems*, pp. 6198–6215, 2021.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in neural information processing systems*, 2018.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pp. 3353–3364, 2019.

Song, C., He, K., Wang, L., and Hopcroft, J. E. Improving the generalization of adversarial training with domain adaptation. In *International Conference on Learning Representations*, 2019.

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

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

Thilagar, A., Frongillo, R., Finocchiaro, J. J., and Goodwill, E. Consistent polyhedral surrogates for top-k classification and variants. In *International Conference on Machine Learning*, pp. 21329–21359, 2022.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

Tsai, Y.-L., Hsu, C.-Y., Yu, C.-M., and Chen, P.-Y. Formalizing generalization and robustness of neural networks to weight perturbations. In *International Conference on Learning Representations*, 2021.

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

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

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

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

Viallard, P., VIDOT, E. G., Habrard, A., and Morvant, E. A pac-bayes analysis of adversarial robustness. In *Advances in Neural Information Processing Systems*, pp. 14421–14433, 2021.

Wang, Y. and Scott, C. Weston-Watkins hinge loss and ordered partitions. In *Advances in neural information processing systems*, pp. 19873–19883, 2020.

Wang, Y. and Scott, C. D. On classification-calibration of gamma-phi losses. *arXiv preprint arXiv:2302.07321*, 2023.

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., and Gu, Q. On the convergence and robustness of adversarial training. In *International Conference on Machine Learning*, pp. 6586–6595, 2019.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

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

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

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, pp. 2958–2969, 2020.

Xiao, J., Fan, Y., Sun, R., Wang, J., and Luo, Z.-Q. Stability analysis and generalization bounds of adversarial training. In *Advances in Neural Information Processing Systems*, 2022.

Xie, C. and Yuille, A. Intriguing properties of adversarial training at scale. In *International Conference on Learning Representations*, 2020.

Xie, C., Wu, Y., Maaten, L. v. d., Yuille, A. L., and He, K. Feature denoising for improving adversarial robustness. In *Conference on computer vision and pattern recognition*, pp. 501–509, 2019.

Xing, Y., Zhang, R., and Cheng, G. Adversarially robust estimate and risk analysis in linear regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 514–522, 2021.

Yang, H., Zhang, J., Dong, H., Inkawhich, N., Gardner, A., Touchet, A., Wilkes, W., Berry, H., and Li, H. Dverge: diversifying vulnerabilities for enhanced robust generation of ensembles. *Advances in Neural Information Processing Systems*, pp. 5505–5515, 2020.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pp. 7085–7094, 2019.

Yu, F., Liu, C., Wang, Y., Zhao, L., and Chen, X. Interpreting adversarial robustness: A view from decision surface in input space. *arXiv preprint arXiv:1810.00144*, 2018.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J., and Wang, L. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., and Dong, B. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, 2019a.

Zhang, H. and Wang, J. Defense against adversarial attacks using feature scattering-based adversarial training. In *Advances in Neural Information Processing Systems*, 2019.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019b.

Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., and Kankanhalli, M. Attacks which do not kill training make adversarial learning stronger. In *International conference on machine learning*, pp. 11278–11287, 2020a.

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

Zhang, M., Ramaswamy, H. G., and Agarwal, S. Convex calibrated surrogates for the multi-label f-measure. In *International Conference on Machine Learning*, pp. 11246–11255, 2020b.

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

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

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, 2018.

Zheng, C., Wu, G., Bao, F., Cao, Y., Li, C., and Zhu, J. Revisiting discriminative vs. generative classifiers: Theory and implications. *arXiv preprint arXiv:2302.02334*, 2023.

Contents of Appendix

A. Related work

Consistency guarantees. The concept of Bayes consistency, or the related one of classification calibration (Zhang, 2004a; Bartlett et al., 2006; Steinwart, 2007; Mohri et al., 2018), have been extensively explored in a wide range of contexts, including multi-class classification (Zhang, 2004b; Chen & Sun, 2006; Chen & Xiang, 2006; Tewari & Bartlett, 2007; Liu, 2007; Dogan et al., 2016; Wang & Scott, 2020; Ramaswamy & Agarwal, 2012; Narasimhan et al., 2015; Agarwal & Agarwal, 2015; Williamson et al., 2016; Ramaswamy & Agarwal, 2016; Finocchiaro et al., 2019; Frongillo & Waggoner, 2021; Finocchiaro et al., 2022; Wang & Scott, 2023), multi-label classification (Gao & Zhou, 2011; Dembczynski et al., 2012; Zhang et al., 2020b), learning with rejection (Cortes et al., 2016b;a; Ramaswamy et al., 2015; Cao et al., 2022), ranking (Duchi et al., 2010; Ravikumar et al., 2011; Ramaswamy et al., 2013; Gao & Zhou, 2015; Uematsu & Lee, 2017; Mao et al., 2023), cost-sensitive classification (Pires et al., 2013; Pires & Szepesvári, 2016), structured prediction (Ciliberto et al., 2016; Osokin et al., 2017; Blondel, 2019), top-k classification (Thilagar et al., 2022), structured abstain problem (Nueve et al., 2022) and ordinal regression (Pedregosa et al., 2017). Bayes-consistency does not supply any information about learning with a typically restricted hypothesis set. Another line of research focuses on realizable $\mathcal{H}$-consistency guarantees (Long & Servedio, 2013; Zhang & Agarwal, 2020; Kuznetsov et al., 2014), which provides hypothesis set-specific consistency guarantees under the assumption that the underlying distribution is $\mathcal{H}$-realizable. However, none of these guarantees is informative for approximate minimizers (non-asymptotic guarantee) since convergence could be arbitrarily slow.

The concept of $\mathcal{H}$-consistency bounds was first introduced by Awasthi et al. (2022b) in binary classification and subsequently extended by Awasthi et al. (2022a) to the scenario of multi-class classification. Such guarantees are both non-asymptotic and hypothesis set-specific. This paper presents the first tight $\mathcal{H}$-consistency bounds for the *comp-sum losses*, that includes cross-entropy (or logistic loss), generalized cross-entropy, the mean absolute error and other loss cross-entropy-like functions.

The concurrent and independent study of Zheng et al. (2023) also provides an $\mathcal{H}$-consistency bound for the logistic loss. Their bound holds for the special case of $\mathcal{H}$ being a constrained linear hypothesis set, subject to an additional assumption on the distribution. In contrast, our bounds do not require any distributional assumption. However, it should be noted that our results are only applicable to complete hypothesis sets. In upcoming work, we present $\mathcal{H}$-consistency bounds for non-complete hypothesis sets and arbitrary distributions.

Adversarial robustness. From a theoretical perspective, there has been significant dedication towards providing guarantees for adversarial robustness (Szegedy et al., 2013; Biggio et al., 2013; Goodfellow et al., 2014; Madry et al., 2017; Carlini & Wagner, 2017), including PAC Learnability (Feige et al., 2015; 2018; Montasser et al., 2019; Attias & Hanneke, 2022; Ashtiani et al., 2020; Bhattacharjee et al., 2021; Cullina et al., 2018; Dan et al., 2020; Montasser et al., 2020a;b; Attias et al., 2022a; Diakonikolas et al., 2020; Montasser et al., 2021; Kontorovich & Attias, 2021; Montasser et al., 2022), robust generalization (Khim & Loh, 2018; Attias et al., 2019; Xing et al., 2021; Yin et al., 2019; Schmidt et al., 2018; Awasthi et al., 2020; Attias et al., 2022b; Xiao et al., 2022; Viallard et al., 2021; Bubeck & Sellke, 2021; Li & Telgarsky, 2023), adversarial examples (Bubeck et al., 2018; 2019; Bartlett et al., 2021; Bubeck et al., 2021), consistency guarantees and optimal adversarial classifiers (Bao et al., 2020; Awasthi et al., 2021a;c;b; 2022b;a; Meunier et al., 2022; Li & Telgarsky, 2023; Frank & Niles-Weed, 2023) and optimization (Awasthi et al., 2019; Robey et al., 2021; Awasthi et al., 2023a).

From an algorithmic standpoint, numerous defense strategies have been proposed historically, including adversarial surrogate loss functions (Kurakin et al., 2016; Madry et al., 2017; Tsipras et al., 2018; Kannan et al., 2018; Zhang et al., 2019b; Wang et al., 2020; Levi et al., 2022; Jin et al., 2022; Awasthi et al., 2023b), curriculum and adaptive attack methods (Cai et al., 2018; Wang et al., 2019; Zhang et al., 2020a; Ding et al., 2022; Cheng et al., 2022), efficient adversarial training (Shafahi et al., 2019; Zhang et al., 2019a; Wong et al., 2020; Andriushchenko & Flammarion, 2020), ensemble techniques (Tramèr et al., 2018; Pang et al., 2019; Yang et al., 2020; Guo et al., 2022), unlabeled data (Carmon et al., 2019; Alayrac et al., 2019; Zhai et al., 2019), data augmentation (Rebuffi et al., 2021a;b), neural network architectures (Xie & Yuille, 2020; Xie et al., 2019; Liu et al., 2020; Guo et al., 2020), weight averaging/perturbation (Gowal et al., 2020; Wu et al., 2020; Tsai et al., 2021; Yu et al., 2018; Prabhu et al., 2019) and other techniques (Zhang & Wang, 2019; Qin et al., 2019; Goldblum et al., 2020; Song et al., 2019; Pang et al., 2020b; Lee et al., 2020; Qian et al., 2021).

This paper introduces a new family of loss functions, *smooth adversarial comp-sum losses*, derived from their comp-sum counterparts by adding in a related smooth term. We show that these loss functions are beneficial in the adversarial setting by proving that they admit $\mathcal{H}$-consistency bounds. This leads to new adversarial robustness algorithms that consist of minimizing a regularized smooth adversarial comp-sum loss. We report the results of a series of experiments demonstrating that our algorithms outperform the current state-of-the-art, while also achieving a superior non-adversarial accuracy.

B. Proofs of $\mathcal{H}$-consistency bounds for comp-sum losses (Theorem 3.1) and tightness (Theorem 3.2)

To begin with the proof, we first introduce some notation. We denote by $p(x, y) = \mathcal{D}(Y = y \mid X = x)$ the conditional probability of $Y = y$ given $X = x$. The generalization error for a surrogate loss can be rewritten 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 denote by $\mathcal{C}^*_\ell(\mathcal{H}, x) = \inf_{h \in \mathcal{H}} \mathcal{C}_\ell(h, x)$ the minimal conditional ℓ-risk. Then, the minimizability gap can be rewritten as follows:

$$\mathcal{M}_\ell(\mathcal{H}) = \mathcal{R}^*_\ell(\mathcal{H}) - \mathbb{E}_X[\mathcal{C}^*_\ell(\mathcal{H}, x)].$$

We further refer to $\mathcal{C}_\ell(h, x) - \mathcal{C}^*_\ell(\mathcal{H}, x)$ as the calibration gap and denote it by $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x)$.

For any $h \in \mathcal{H}$ and $x \in \mathcal{X}$, by the symmetry and completeness of $\mathcal{H}$, we can always find a family of hypotheses $\left\{\overline{h}_\mu \colon \mu \in \mathbb{R}\right\} \subset \mathcal{H}$ such that $h_\mu(x, \cdot)$ take the following values:

$$\overline{h}_\mu(x, y) = \begin{cases} h(x, y) & \text{if } y \notin \{y_{\max}, \mathsf{h}(x)\} \\ \log(\exp[h(x, y_{\max})] + \mu) & \text{if } y = \mathsf{h}(x) \\ \log(\exp[h(x, \mathsf{h}(x))] - \mu) & \text{if } y = y_{\max}. \end{cases} \tag{17}$$

Note that the hypotheses $\overline{h}_\mu$ has the following property:

$$\sum_{y \in \mathcal{Y}} e^{h(x,y)} = \sum_{y \in \mathcal{Y}} e^{\overline{h}_\mu(x,y)}, \ \forall \mu \in \mathbb{R}. \tag{18}$$

Lemma B.1. *Assume that $\mathcal{H}$ is symmetric and complete. Then, for any $h \in \mathcal{X}$ and $x \in \mathcal{X}$, the following equality holds:*

$$\begin{aligned}
&\mathcal{C}_{\ell^{\text{comp}}_\tau}(h, x) - \inf_{\mu \in \mathbb{R}} \mathcal{C}_{\ell^{\text{comp}}_\tau}(\overline{h}_\mu, x) \\
&= \sup_{\mu \in \mathbb{R}} \left\{ p(x, y_{\max}) \left(\Phi^\tau \left(\frac{\sum_{y' \in \mathcal{Y}} e^{h(x,y')}}{e^{h(x,y_{\max})}} - 1 \right) - \Phi^\tau \left(\frac{\sum_{y' \in \mathcal{Y}} e^{h(x,y')}}{e^{h(x,\mathsf{h}(x))-\mu}} - 1 \right) \right) \right. \\
&\qquad \left. + p(x, \mathsf{h}(x)) \left(\Phi^\tau \left(\frac{\sum_{y' \in \mathcal{Y}} e^{h(x,y')}}{e^{h(x,\mathsf{h}(x))}} - 1 \right) - \Phi^\tau \left(\frac{\sum_{y' \in \mathcal{Y}} e^{h(x,y')}}{e^{h(x,y_{\max})+\mu}} - 1 \right) \right) \right\} \\
&= \begin{cases} \frac{1}{\tau - 1} \left[\sum_{y' \in \mathcal{Y}} e^{h(x,y')} \right]^{1-\tau} \left[\frac{\left(p(x,y_{\max})^{\frac{1}{2-\tau}} + p(x,\mathsf{h}(x))^{\frac{1}{2-\tau}} \right)^{2-\tau}}{\left(e^{h(x,y_{\max})} + e^{h(x,\mathsf{h}(x))} \right)^{1-\tau}} - \frac{p(x,y_{\max})}{\left(e^{h(x,y_{\max})} \right)^{1-\tau}} - \frac{p(x,\mathsf{h}(x))}{\left(e^{h(x,\mathsf{h}(x))} \right)^{1-\tau}} \right] & \tau \in [0, 2)/\{1\} \\ p(x, y_{\max}) \log \left[\frac{\left(e^{h(x,y_{\max})} + e^{h(x,\mathsf{h}(x))} \right) p(x,y_{\max})}{e^{h(x,y_{\max})} (p(x,y_{\max}) + p(x,\mathsf{h}(x)))} \right] + p(x, \mathsf{h}(x)) \log \left[\frac{\left(e^{h(x,y_{\max})} + e^{h(x,\mathsf{h}(x))} \right) p(x,\mathsf{h}(x))}{e^{h(x,\mathsf{h}(x))} (p(x,y_{\max}) + p(x,\mathsf{h}(x)))} \right] & \tau = 1 \\ \frac{1}{\tau - 1} \left[\sum_{y' \in \mathcal{Y}} e^{h(x,y')} \right]^{1-\tau} \left[\frac{p(x,y_{\max})}{\left(e^{h(x,y_{\max})} + e^{h(x,\mathsf{h}(x))} \right)^{1-\tau}} - \frac{p(x,y_{\max})}{\left(e^{h(x,y_{\max})} \right)^{1-\tau}} - \frac{p(x,\mathsf{h}(x))}{\left(e^{h(x,\mathsf{h}(x))} \right)^{1-\tau}} \right] & \tau \in [2, +\infty). \end{cases}
\end{aligned}$$

Proof. For the comp-sum loss ℓ^{comp}_τ, the conditional ℓ^{comp}_τ-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\ell^{\text{comp}}_\tau}(h, x) &= \sum_{y \in \mathcal{Y}} p(x, y)\ell^{\text{comp}}_\tau(h, x, y) \\
&= \sum_{y \in \mathcal{Y}} p(x, y)\Phi^\tau \left(\sum_{y' \in \mathcal{Y}} e^{h(x,y')-h(x,y)} - 1 \right) \\
&= p(x, y_{\max})\Phi^\tau \left(\sum_{y' \in \mathcal{Y}} e^{h(x,y')-h(x,y_{\max})} - 1 \right) + p(x, \mathsf{h}(x))\Phi^\tau \left(\sum_{y' \in \mathcal{Y}} e^{h(x,y')-h(x,\mathsf{h}(x))} - 1 \right) \\
&\quad + \sum_{y \notin \{y_{\max}, \mathsf{h}(x)\}} p(x, y)\Phi^\tau \left(\sum_{y' \in \mathcal{Y}} e^{h(x,y')-h(x,y)} - 1 \right).
\end{aligned}$$

Therefore, by (17) and (18), we obtain the first equality. The second equality can be obtained by taking the derivative with respect to μ. □

Lemma B.2. *Assume that $\mathcal{H}$ is symmetric and complete. Then, for any $h \in \mathcal{X}$ and $x \in \mathcal{X}$, the following equality holds*

$$\inf_{h\in\mathcal{H}}\left(\mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x) - \inf_{\mu\in\mathbb{R}} \mathcal{C}_{\ell_\tau^{\text{comp}}}(\overline{h}_\mu, x)\right)$$
$$= \begin{cases} \frac{2^{2-\tau}}{1-\tau}\left[\frac{p(x,y_{\max})+p(x,\mathsf{h}(x))}{2} - \left[\frac{p(x,y_{\max})^{\frac{1}{2-\tau}}+p(x,\mathsf{h}(x))^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau}\right] & \tau\in[0,1)\\ p(x,y_{\max})\log\left[\frac{2p(x,y_{\max})}{p(x,y_{\max})+p(x,\mathsf{h}(x))}\right] + p(x,\mathsf{h}(x))\log\left[\frac{2p(x,\mathsf{h}(x))}{p(x,y_{\max})+p(x,\mathsf{h}(x))}\right] & \tau=1\\ \frac{2}{(\tau-1)n^{\tau-1}}\left(\left[\frac{p(x,y_{\max})^{\frac{1}{2-\tau}}+p(x,\mathsf{h}(x))^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau} - \frac{p(x,y_{\max})+p(x,\mathsf{h}(x))}{2}\right) & \tau\in(1,2)\\ \frac{1}{(\tau-1)n^{\tau-1}}(p(x,y_{\max})-p(x,\mathsf{h}(x))) & \tau\in[2,+\infty). \end{cases}$$

Proof. By using Lemma B.1 and taking infimum with respect to $e^{h(x,1)},\ldots,e^{h(x,n)}$, the equality is proved directly. □

Let $\alpha = p(x,y_{\max}) + p(x,\mathsf{h}(x)) \in [0,1]$ and $\beta = p(x,y_{\max} - p(x,\mathsf{h}(x))) \in [0,1]$. Then, using the fact that $p(x,y_{\max}) = \frac{\alpha+\beta}{2}$ and $p(x,\mathsf{h}(x)) = \frac{\alpha-\beta}{2}$, we can rewrite $\inf_{h\in\mathcal{H}}\left(\mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x) - \inf_{\mu\in\mathbb{R}} \mathcal{C}_{\ell_\tau^{\text{comp}}}(\overline{h}_\mu,x)\right)$ as

$$\inf_{h\in\mathcal{H}}\left(\mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x) - \inf_{\mu\in\mathbb{R}} \mathcal{C}_{\ell_\tau^{\text{comp}}}(\overline{h}_\mu, x)\right) = \Psi_\tau(\alpha,\beta) = \begin{cases} \frac{2^{1-\tau}}{1-\tau}\left[\alpha - \left[\frac{(\alpha+\beta)^{\frac{1}{2-\tau}}+(\alpha-\beta)^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau}\right] & \tau\in[0,1)\\ \frac{\alpha+\beta}{2}\log\left[\frac{\alpha+\beta}{\alpha}\right] + \frac{\alpha-\beta}{2}\log\left[\frac{\alpha-\beta}{\alpha}\right] & \tau=1\\ \frac{1}{(\tau-1)n^{\tau-1}}\left(\left[\frac{(\alpha+\beta)^{\frac{1}{2-\tau}}+(\alpha-\beta)^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau} - \alpha\right) & \tau\in(1,2)\\ \frac{1}{(\tau-1)n^{\tau-1}}\beta & \tau\in[2,+\infty). \end{cases} \tag{19}$$

By taking the partial derivative of $\Psi_\tau(\alpha,\cdot)$ with respect to α and analyzing the minima, we obtain the following result.

Lemma B.3. *For any $\tau \in [0,+\infty)$ and $\alpha\in[0,1]$, the following inequality holds for any $\beta\in[0,1]$,*

$$\Psi_\tau(\alpha,\beta) \geq \Psi_\tau(1,\beta) = \mathcal{T}_\tau(\beta) = \begin{cases} \frac{2^{1-\tau}}{1-\tau}\left[1 - \left[\frac{(1+\beta)^{\frac{1}{2-\tau}}+(1-\beta)^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau}\right] & \tau\in[0,1)\\ \frac{1+\beta}{2}\log[1+\beta] + \frac{1-\beta}{2}\log[1-\beta] & \tau=1\\ \frac{1}{(\tau-1)n^{\tau-1}}\left[\left[\frac{(1+\beta)^{\frac{1}{2-\tau}}+(1-\beta)^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau} - 1\right] & \tau\in(1,2)\\ \frac{1}{(\tau-1)n^{\tau-1}}\beta & \tau\in[2,+\infty). \end{cases}$$

We denote by $\mathcal{T}_\tau(\beta) = \Psi_\tau(1,\beta)$ and call it the $\mathcal{H}$-consistency comp-sum transformation, and denote by Γ_τ the inverse of $\mathcal{T}_\tau$: $\Gamma_\tau(t) = \mathcal{T}_\tau^{-1}(t)$. We then present the proofs of Theorem 3.1 and Theorem 3.2 in the below.

Theorem 3.1 (**$\mathcal{H}$-consistency bounds for comp-sum losses**). *Assume that $\mathcal{H}$ is symmetric and complete. Then, for any $\tau\in[0,\infty)$ and any $h\in\mathcal{H}$, the following inequality holds:*

$$\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}) \leq \Gamma_\tau\left(\mathcal{R}_{\ell_\tau^{\text{comp}}}(h) - \mathcal{R}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}) + \mathcal{M}_{\ell_\tau^{\text{comp}}}(\mathcal{H})\right) - \mathcal{M}_{\ell_{0-1}}(\mathcal{H}),$$

where $\Gamma_\tau(t) = \mathcal{T}_\tau^{-1}(t)$ is the inverse of $\mathcal{H}$-consistency comp-sum transformation, defined for all $\beta\in[0,1]$ by $\mathcal{T}_\tau(\beta) =$

$$\begin{cases} \frac{2^{1-\tau}}{1-\tau}\left[1 - \left[\frac{(1+\beta)^{\frac{1}{2-\tau}}+(1-\beta)^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau}\right] & \tau\in[0,1)\\ \frac{1+\beta}{2}\log[1+\beta] + \frac{1-\beta}{2}\log[1-\beta] & \tau=1\\ \frac{1}{(\tau-1)n^{\tau-1}}\left[\left[\frac{(1+\beta)^{\frac{1}{2-\tau}}+(1-\beta)^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau} - 1\right] & \tau\in(1,2)\\ \frac{1}{(\tau-1)n^{\tau-1}}\beta & \tau\in[2,+\infty). \end{cases}$$

Proof. Using previous lemmas, we can lower bound the calibration gap of comp-sum losses as follows, for any $h \in \mathcal{H}$,

$$
\begin{aligned}
&\mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x) - \mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H},x)\\
&\geq \mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x) - \inf_{\mu\in\mathbb{R}} \mathcal{C}_{\ell_\tau^{\text{comp}}}(\overline{h}_\mu,x)\\
&\geq \inf_{h\in\mathcal{H}}\left(\mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x) - \inf_{\mu\in\mathbb{R}} \mathcal{C}_{\ell_\tau^{\text{comp}}}(\overline{h}_\mu,x)\right)\\
&= \begin{cases}
\frac{2^{2-\tau}}{1-\tau}\left[\frac{p(x,y_{\max})+p(x,\mathsf{h}(x))}{2} - \left[\frac{p(x,y_{\max})^{\frac{1}{2-\tau}}+p(x,\mathsf{h}(x))^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau}\right] & \tau\in[0,1)\\
p(x,y_{\max})\log\left[\frac{2p(x,y_{\max})}{p(x,y_{\max})+p(x,\mathsf{h}(x))}\right] + p(x,\mathsf{h}(x))\log\left[\frac{2p(x,\mathsf{h}(x))}{p(x,y_{\max})+p(x,\mathsf{h}(x))}\right] & \tau=1\\
\frac{2}{(\tau-1)n^{\tau-1}}\left(\left[\frac{p(x,y_{\max})^{\frac{1}{2-\tau}}+p(x,\mathsf{h}(x))^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau} - \frac{p(x,y_{\max})+p(x,\mathsf{h}(x))}{2}\right) & \tau\in(1,2)\\
\frac{1}{(\tau-1)n^{\tau-1}}(p(x,y_{\max})-p(x,\mathsf{h}(x))) & \tau\in[2,+\infty)
\end{cases} && \text{(By Lemma B.2)}\\
&\geq \mathcal{T}_\tau(p(x,y_{\max})-p(x,\mathsf{h}(x))) && \text{(By (19) and Lemma B.3)}\\
&= \mathcal{T}_\tau\left(\mathcal{C}_{\ell_{0-1}}(h,x) - \mathcal{C}^*_{\ell_{0-1}}(\mathcal{H},x)\right) && \text{(by (Awasthi et al., 2022a, Lemma 3))}
\end{aligned}
$$

Therefore, taking $\mathcal{P}$ be the set of all distributions, $\mathcal{H}$ be the symmetric and complete hypothesis set, $\epsilon = 0$ and $\Psi(\beta) = \mathcal{T}_\tau(\beta)$ in (Awasthi et al., 2022a, Theorem 4), or, equivalently, $\Gamma(t) = \Gamma_\tau(t)$ in (Awasthi et al., 2022a, Theorem 5), we obtain for any hypothesis $h \in \mathcal{H}$ and any distribution,

$$
\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}) \leq \Gamma_\tau\left(\mathcal{R}_{\ell_\tau^{\text{comp}}}(h) - \mathcal{R}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}) + \mathcal{M}_{\ell_\tau^{\text{comp}}}(\mathcal{H})\right) - \mathcal{M}_{\ell_{0-1}}(\mathcal{H}).
$$

□

Theorem 3.2 (Tightness). *Assume that $\mathcal{H}$ is symmetric and complete. Then, for any $\tau \in [0,1]$ and $\beta \in [0,1]$, there exist a distribution $\mathcal{D}$ and a hypothesis $h \in \mathcal{H}$ such that $\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} + \mathcal{M}_{\ell_{0-1},\mathcal{H}} = \beta$ and $\mathcal{R}_{\ell_\tau^{\text{comp}}}(h) - \mathcal{R}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}) + \mathcal{M}_{\ell_\tau^{\text{comp}}}(\mathcal{H}) = \mathcal{T}_\tau(\beta)$.*

Proof. For any $\beta \in [0,1]$, we consider the distribution that concentrates on a singleton $\{x_0\}$ and satisfies $p(x_0,1) = \frac{1+\beta}{2}$, $p(x_0,2) = \frac{1-\beta}{2}$, $p(x_0,y) = 0$, $3 \leq y \leq n$. We take $h_\tau \in \mathcal{H}$ such that $e^{h_\tau(x,1)} = e^{h_\tau(x,2)}$, $e^{h_\tau(x,y)} = 0$, $3 \leq y \leq n$. Then,

$$
\mathcal{R}_{\ell_{0-1}}(h_\tau) - \mathcal{R}^*_{\ell_{0-1},\mathcal{H}} + \mathcal{M}_{\ell_{0-1},\mathcal{H}} = \mathcal{R}_{\ell_{0-1}}(h_\tau) - \mathbb{E}_X\left[\mathcal{C}^*_{\ell_{0-1}}(\mathcal{H},x)\right] = \mathcal{C}_{\ell_{0-1}}(h_\tau,x_0) - \mathcal{C}^*_{\ell_{0-1}}(\mathcal{H},x_0) = \beta
$$

and for any $\tau \in [0,1]$,

$$
\begin{aligned}
&\mathcal{R}_{\ell_\tau^{\text{comp}}}(h_\tau) - \mathcal{R}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}) + \mathcal{M}_{\ell_\tau^{\text{comp}}}(\mathcal{H})\\
&= \mathcal{R}_{\ell_\tau^{\text{comp}}}(h_\tau) - \mathbb{E}_X\left[\mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H},x)\right]\\
&= \mathcal{C}_{\ell_\tau^{\text{comp}}}(h_\tau,x_0) - \mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H},x_0)\\
&= p(x_0,1)\ell_\tau^{\text{comp}}(h_\tau,x_0,1) + p(x_0,2)\ell_\tau^{\text{comp}}(h_\tau,x_0,2) - \inf_{h\in\mathcal{H}}\left[p(x_0,1)\ell_\tau^{\text{comp}}(h,x_0,1) + p(x_0,2)\ell_\tau^{\text{comp}}(h,x_0,2)\right]\\
&= \mathcal{T}_\tau(\beta), && \text{(By (4) and (10))}
\end{aligned}
$$

which completes the proof. □

C. Approximations of $\mathcal{T}_\tau$ and Γ_τ

In this section, we show how $\mathcal{T}_\tau$ can be lower bounded by its polynomial approximation $\widetilde{\mathcal{T}}_\tau$, and accordingly, Γ_τ can then be upper bounded by $\widetilde{\Gamma}_\tau = \widetilde{\mathcal{T}}_\tau^{-1}$. By analyzing the Taylor expansion, we obtain for any $\beta \in [-1,1]$,

$$
\begin{aligned}
&\left(\frac{(1+\beta)^r + (1-\beta)^r}{2}\right)^{\frac{1}{r}} \geq 1 + \frac{\beta^2}{2}\left(1-\frac{1}{r}\right), \text{ for all } r \geq 1\\
&\left(\frac{(1+\beta)^r + (1-\beta)^r}{2}\right)^{\frac{1}{r}} \leq 1 - \frac{\beta^2}{2}(1-r), \text{ for all } \frac{1}{2} \leq r \leq 1.
\end{aligned}
\tag{20}
$$

and

$$\frac{1+\beta}{2}\log[1+\beta]+\frac{1-\beta}{2}\log[1-\beta]\geq\frac{\beta^2}{2}. \tag{21}$$

For $\tau\in[0,1)$, we have

$$\begin{aligned}
\mathcal{T}_\tau(\beta)&=\frac{2^{1-\tau}}{1-\tau}\left[1-\left[\frac{(1+\beta)^{\frac{1}{2-\tau}}+(1-\beta)^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau}\right]\\
&\geq\frac{2^{1-\tau}}{1-\tau}\left[1-\left[1-\frac{\beta^2}{2}\frac{1-\tau}{2-\tau}\right]\right] && \text{(using (20) with } r=\tfrac{1}{2-\tau}\in\left[\tfrac{1}{2},1\right])\\
&=\frac{\beta^2}{2^\tau(2-\tau)}\\
&=\widetilde{\mathcal{T}}_\tau(\beta).
\end{aligned}$$

For $\tau\in(1,2)$, we have

$$\begin{aligned}
\mathcal{T}_\tau(\beta)&=\frac{1}{(\tau-1)n^{\tau-1}}\left[\left[\frac{(1+\beta)^{\frac{1}{2-\tau}}+(1-\beta)^{\frac{1}{2-\tau}}}{2}\right]^{2-\tau}-1\right]\\
&\geq\frac{1}{(\tau-1)n^{\tau-1}}\left[\left[1+\frac{\beta^2}{2}(\tau-1)\right]-1\right] && \text{(using (20) with } r=\tfrac{1}{2-\tau}\geq 1)\\
&=\frac{\beta^2}{2n^{\tau-1}}\\
&=\widetilde{\mathcal{T}}_\tau(\beta).
\end{aligned}$$

For $\tau=1$, we have

$$\begin{aligned}
\mathcal{T}_\tau(\beta)&=\frac{1+\beta}{2}\log[1+\beta]+\frac{1-\beta}{2}\log[1-\beta]\\
&\geq\frac{\beta^2}{2} && \text{(using (21))}\\
&=\widetilde{\mathcal{T}}_\tau(\beta).
\end{aligned}$$

For $\tau\geq 2$, $\mathcal{T}_\tau(\beta)=\frac{\beta}{(\tau-1)n^{\tau-1}}=\widetilde{\mathcal{T}}_\tau(\beta)$. Therefore, for any $\tau\in[0,+\infty)$,

$$\mathcal{T}_\tau(\beta)\geq\widetilde{\mathcal{T}}_\tau(\beta)=\begin{cases}\frac{\beta^2}{2^\tau(2-\tau)} & \tau\in[0,1)\\ \frac{\beta^2}{2n^{\tau-1}} & \tau\in[1,2)\\ \frac{\beta}{(\tau-1)n^{\tau-1}} & \tau\in[2,+\infty).\end{cases}$$

Furthermore, by using Taylor expansion, we have

$$\lim_{\beta\to 0^+}\frac{\widetilde{\mathcal{T}}_\tau(\beta)}{\mathcal{T}_\tau(\beta)}=c>0 \text{ for some constant } c>0.$$

Thus, the order of polynomials $\widetilde{\mathcal{T}}_\tau(\beta)$ is tightest. Since $\Gamma_\tau=\mathcal{T}_\tau^{-1}$ and $\widetilde{\Gamma}_\tau=\widetilde{\mathcal{T}}_\tau^{-1}$, we also obtain for any $\tau\in[0,+\infty)$,

$$\Gamma_\tau(t)\leq\widetilde{\Gamma}_\tau(t)=\widetilde{\mathcal{T}}_\tau^{-1}(t)=\begin{cases}\sqrt{2^\tau(2-\tau)t} & \tau\in[0,1)\\ \sqrt{2n^{\tau-1}t} & \tau\in[1,2)\\ (\tau-1)n^{\tau-1}t & \tau\in[2,+\infty).\end{cases}$$

D. Characterization of minimizability gaps (proofs of Theorem 4.1 and Theorem 4.2)

Theorem 4.1 (Characterization of minimizability gaps - stochastic case). *Assume that $\mathcal{H}$ is symmetric and complete. Then, for the comp-sum losses ℓ_τ^{comp}, the minimizability gaps can be upper bounded as follows:*

$$\mathcal{M}_{\ell_\tau^{\text{comp}}}(\mathcal{H}) \leq \Phi^\tau\left(\mathcal{R}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H})\right) - \mathop{\mathbb{E}}_x[\mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}, x)], \tag{9}$$

where $\mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}, x)$ is given by

$$\begin{cases} \frac{1}{1-\tau}\left(\left[\sum_{y\in\mathcal{Y}} p(x,y)^{\frac{1}{2-\tau}}\right]^{2-\tau} - 1\right) & \tau \geq 0, \tau \neq 1, \tau \neq 2 \\ -\sum_{y\in\mathcal{Y}} p(x,y)\log[p(x,y)] & \tau = 1 \\ 1 - \max_{y\in\mathcal{Y}} p(x,y) & \tau = 2. \end{cases} \tag{10}$$

Proof. Using the fact that Φ^τ is concave and non-decreasing, by (8), we can then upper bound the minimizability gaps for different $\tau \geq 0$ as follows,

$$\mathcal{M}_{\ell_\tau^{\text{comp}}}(\mathcal{H}) \leq \Phi_\tau\left(\mathcal{R}^*_{\ell^{\text{comp}}_{\tau=0}}(\mathcal{H})\right) - \mathop{\mathbb{E}}_x[\mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}, x)].$$

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

$$\mathcal{C}_{\ell_\tau^{\text{comp}}}(h, x) = \sum_{y\in\mathcal{Y}} p(x,y)\Phi_\tau\left(\sum_{y'\neq y} \exp(h(x,y') - h(x,y))\right). \tag{22}$$

Note that $\mathcal{C}_{\ell_\tau^{\text{comp}}}(h, x)$ is convex and differentiable with respect to $h(x,y)$s, by taking the partial derivative and using the derivative of Φ_τ given in (3), we obtain

$$\begin{aligned} &\frac{\partial \mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x)}{\partial h(x,y)} \\ &= p(x,y)\frac{\partial \Phi_\tau}{\partial u}\left(\sum_{y'\neq y} \exp(h(x,y') - h(x,y))\right)\left(-\sum_{y'\neq y} \exp(h(x,y') - h(x,y))\right) \\ &+ \sum_{y'\neq y} p(x,y')\frac{\partial \Phi_\tau}{\partial u}\left(\sum_{y''\neq y'} \exp(h(x,y'') - h(x,y'))\right)(\exp(h(x,y) - h(x,y'))) \\ &= p(x,y)\frac{-\sum_{y'\neq y}\exp(h(x,y') - h(x,y))}{\left[\sum_{y'\in\mathcal{Y}}\exp(h(x,y') - h(x,y))\right]^\tau} + \sum_{y'\neq y} p(x,y')\frac{\exp(h(x,y) - h(x,y'))}{\left[\sum_{y''\in\mathcal{Y}}\exp(h(x,y'') - h(x,y'))\right]^\tau} \end{aligned} \tag{23}$$

Let $\mathcal{S}(x,y) = \sum_{y'\in\mathcal{Y}}\exp(h(x,y') - h(x,y))$. Then, $\exp(h(x,y) - h(x,y')) = \frac{\mathcal{S}(x,y')}{\mathcal{S}(x,y)}$ and thus (23) can be written as

$$\frac{\partial \mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x)}{\partial h(x,y)} = p(x,y)\frac{-\mathcal{S}(x,y)+1}{\mathcal{S}(x,y)^\tau} + \sum_{y'\neq y} p(x,y')\frac{1}{\mathcal{S}(x,y')^{\tau-1}\mathcal{S}(x,y)} \tag{24}$$

It is straightforward to verify that

$$\mathcal{S}^*(x,y) = \begin{cases} \frac{\sum_{y'\in\mathcal{Y}} p(x,y')^{\frac{1}{2-\tau}}}{p(x,y)^{\frac{1}{2-\tau}}} & \tau \neq 2 \\ \frac{1}{\mathbb{1}_{y=\text{argmax}_{y'\in\mathcal{Y}}\, p(x,y')}} & \tau = 2 \end{cases} \tag{25}$$

satisfy

$$\frac{\partial \mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x)}{\partial h(x,y)} = 0, \forall y \in \mathcal{Y}.$$

When $\mathcal{H}$ is symmetric and complete, (25) can be attained by some $h^* \in \mathcal{H}$. Since $\mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x)$ is convex and differentiable with respect to $h(x,y)$s, we know that h^* achieves the minimum of $\mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x)$ within $\mathcal{H}$. Then,

$$\begin{aligned}
\mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H},x) &= \mathcal{C}_{\ell_\tau^{\text{comp}}}(h^*,x)\\
&= \sum_{y\in\mathcal{Y}} p(x,y)\Phi_\tau\left(\sum_{y'\neq y}\exp(h^*(x,y')-h^*(x,y))\right)\\
&= \sum_{y\in\mathcal{Y}} p(x,y)\Phi_\tau(\mathcal{S}^*(x,y)-1) && \text{(by the def. of } \mathcal{S}(x,y).)\\
&= \begin{cases}\sum_{y\in\mathcal{Y}} p(x,y)\frac{1}{1-\tau}\left(\mathcal{S}^*(x,y)^{1-\tau}-1\right) & \tau\geq 0,\tau\neq 1\\ \sum_{y\in\mathcal{Y}} p(x,y)\log[\mathcal{S}^*(x,y)] & \tau=1\end{cases} && \text{(by (4).)}\\
&= \begin{cases}\frac{1}{1-\tau}\left(\left[\sum_{y\in\mathcal{Y}} p(x,y)^{\frac{1}{2-\tau}}\right]^{2-\tau}-1\right) & \tau\geq 0,\tau\neq 1,\tau\neq 2\\ -\sum_{y\in\mathcal{Y}} p(x,y)\log[p(x,y)] & \tau=1\\ 1-\max_{y\in\mathcal{Y}} p(x,y) & \tau=2.\end{cases} && \text{(by (25))}
\end{aligned}$$

□

Theorem 4.2 (**Characterization of minimizability gaps - deterministic case**). *Assume that for any $x\in\mathcal{X}$, we have $\{(h(x,1),\ldots,h(x,n)) : h\in\mathcal{H}\} = [-\Lambda,+\Lambda]^n$. Then, for comp-sum losses ℓ_τ^{comp} and any deterministic distribution, the minimizability gaps can be upper bounded as follows:*

$$\mathcal{M}_{\ell_\tau^{\text{comp}}}(\mathcal{H}) \leq \Phi^\tau\left(\mathcal{R}^*_{\ell_{\tau=0}^{\text{comp}}}(\mathcal{H})\right) - \mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H},x), \tag{11}$$

where $\mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H},x)$ is given by

$$\begin{cases}\frac{1}{1-\tau}\left(\left[1+e^{-2\Lambda}(n-1)\right]^{1-\tau}-1\right) & \tau\geq 0,\tau\neq 1\\ \log\left[1+e^{-2\Lambda}(n-1)\right] & \tau=1.\end{cases} \tag{12}$$

Proof. Using the fact that Φ_τ is concave and non-decreasing, by (8), we can then upper bound the minimizability gaps for different $\tau\geq 0$ as follows,

$$\mathcal{M}_{\ell_\tau^{\text{comp}}}(\mathcal{H}) \leq \Phi_\tau\left(\mathcal{R}^*_{\ell_{\tau=0}^{\text{comp}}}(\mathcal{H})\right) - \mathop{\mathbb{E}}_x[\mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H},x)].$$

Let $y_{\max} = \operatorname{argmax} p(x,y)$. By definition, for any deterministic distribution, the conditional ℓ_τ^{comp}-risk can be expressed as follows:

$$\begin{aligned}
\mathcal{C}_{\ell_\tau^{\text{comp}}}(h,x) &= \Phi_\tau\left(\sum_{y'\neq y_{\max}}\exp(h(x,y')-h(x,y_{\max}))\right)\\
&= \begin{cases}\frac{1}{1-\tau}\left(\left(1+\frac{\sum_{y'\neq y_{\max}}\exp(h(x,y'))}{\exp(h(x,y_{\max}))}\right)^{1-\tau}-1\right)\\ \log\left(1+\frac{\sum_{y'\neq y_{\max}}\exp(h(x,y'))}{\exp(h(x,y_{\max}))}\right).\end{cases}
\end{aligned} \tag{26}$$

Since for any $\tau>0$, Φ_τ is increasing, under the assumption of $\mathcal{H}$, h^* that satisfies

$$h^*(x,y) = \begin{cases}\Lambda & y = y_{\max}\\ -\Lambda & \text{otherwise}\end{cases} \tag{27}$$

achieves the minimum of $\mathcal{C}_{\ell_\tau^{\text{comp}}}(h, x)$ within $\mathcal{H}$. Then,

$$\begin{aligned}
\mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}, x) &= \mathcal{C}_{\ell_\tau^{\text{comp}}}(h^*, x) \\
&= \begin{cases} \frac{1}{1-\tau}\left(\left(1 + \frac{\sum_{y' \neq y_{\max}} \exp(h^*(x,y'))}{\exp(h^*(x,y_{\max}))}\right)^{1-\tau} - 1\right) \\ \log\left(1 + \frac{\sum_{y' \neq y_{\max}} \exp(h^*(x,y'))}{\exp(h^*(x,y_{\max}))}\right). \end{cases} \\
&= \begin{cases} \frac{1}{1-\tau}\left(\left[1 + e^{-2\Lambda}(n-1)\right]^{1-\tau} - 1\right) & \tau \geq 0, \tau \neq 1 \\ \log\left[1 + e^{-2\Lambda}(n-1)\right] & \tau = 1. \end{cases} \qquad \text{(by (27))}
\end{aligned}$$

Since $\mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}, x)$ is independent of x, we have $\mathbb{E}_x[\mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}, x)] = \mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}, x)$ and thus

$$\mathcal{M}_{\ell_\tau^{\text{comp}}}(\mathcal{H}) \leq \Phi_\tau\left(\mathcal{R}^*_{\ell_{\tau=0}^{\text{comp}}}(\mathcal{H})\right) - \mathcal{C}^*_{\ell_\tau^{\text{comp}}}(\mathcal{H}, x),$$

which concludes the proof. □

E. Proof of Lemma 4.3

Lemma E.1. *For any $u_1 \geq u_2 \geq 0$, $\Phi^\tau(u_1) - \Phi^\tau(u_2)$ is non-increasing with respect to τ.*

Proof. For any $u_1 \geq u_2 \geq 0$ and $\tau \neq 1$, we have

$$\begin{aligned}
&\frac{\partial(\Phi_\tau(u_1) - \Phi_\tau(u_2))}{\partial \tau} \\
&= \frac{\left((1+u_1)^{1-\tau} - (1+u_2)^{1-\tau}\right)}{(1-\tau)^2} + \frac{1}{1-\tau}\left((1+u_2)^{1-\tau}\log(1+u_2) - (1+u_1)^{1-\tau}\log(1+u_1)\right) \\
&= \frac{g(u_1,\tau) - g(u_2,\tau)}{(1-\tau)^2}
\end{aligned}$$

where $g(t,\tau) = (1+t)^{1-\tau} - (1-\tau)(1+t)^{1-\tau}\log(1+t)$. By taking the partial derivative, we obtain for any $\tau \neq 1$ and $t \geq 0$,

$$\frac{\partial g}{\partial t} = -(1-\tau)^2(1+t)^\tau \log(1+t) \leq 0$$

Therefore, for any $u_1 \geq u_2 \geq 0$ and $\tau \neq 1$, $g(u_1,\tau) \leq g(u_2,\tau)$ and

$$\frac{\partial(\Phi_\tau(u_1) - \Phi_\tau(u_2))}{\partial \tau} \leq 0,$$

which implies that for any $u_1 \geq u_2 \geq 0$ and $\tau \neq 1$, $\Phi_\tau(u_1) - \Phi_\tau(u_2)$ is a non-increasing function of τ. Moreover, since for $x \geq 1$, $\frac{1}{\tau-1}\left(x^{\tau-1} - 1\right) \to \log(x)$ as $\tau \to 1$, we know that for any $u_1 \geq u_2 \geq 0$, $\Phi_\tau(u_1) - \Phi_\tau(u_2)$ is continuous with respect to $\tau = 1$. Therefore, we conclude that for any $u_1 \geq u_2 \geq 0$, $\Phi_\tau(u_1) - \Phi_\tau(u_2)$ is non-increasing with respect to τ. □

F. Proof of adversarial $\mathcal{H}$-consistency bound for adversarial comp-sum losses (Theorem 5.2)

Theorem 5.2 (**$\mathcal{H}$-consistency bound of $\widetilde{\ell}^{\text{comp}}_{\tau,\rho}$**). *Assume that $\mathcal{H}$ is symmetric and locally ρ-consistent. Then, for any choice of the hyperparameters $\tau, \rho > 0$, any hypothesis $h \in \mathcal{H}$, the following inequality holds:*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma}(\mathcal{H}) \leq \Phi^\tau(1)\left(\mathcal{R}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(h) - \mathcal{R}^*_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H})\right) - \mathcal{M}_{\ell_\gamma}(\mathcal{H}).$$

Proof. Let $\overline{\mathcal{H}}_\gamma(x) = \left\{h \in \mathcal{H} : \inf_{x' : \|x - x'\| \leq \gamma} \rho_h(x', \mathsf{h}(x)) > 0\right\}$ and $p(x) = (p(x,1), \ldots, p(x,c))$. For any $x \in \mathcal{X}$ and $h \in \mathcal{H}$, we define $h\left(x, \{1\}^h_x\right), h\left(x, \{2\}^h_x\right), \ldots, h\left(x, \{c\}^h_x\right)$ by sorting the scores $\{h(x,y) : y \in \mathcal{Y}\}$ in increasing order, and

$p_{[1]}(x), p_{[2]}(x), \ldots, p_{[c]}(x)$ by sorting the probabilities $\{p(x,y) : y \in \mathcal{Y}\}$ in increasing order. Note $\{c\}_x^h = \mathsf{h}(x)$. Since $\mathcal{H}$ is symmetric and locally ρ-consistent, for any $x \in \mathcal{X}$, there exists a hypothesis $h^* \in \mathcal{H}$ such that

$$\inf_{x' : \|x - x'\| \le \gamma} |h^*(x', i) - h^*(x', j)| \ge \rho, \forall i \ne j \in \mathcal{Y}$$
$$p(x, \{k\}_{x'}^{h^*}) = p_{[k]}(x), \forall x' \in \{x' : \|x - x'\| \le \gamma\}, \forall k \in \mathcal{Y}.$$

Then, we have

$$\begin{aligned}
&\mathcal{C}^*_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H}, x)\\
&\le \mathcal{C}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(h^*, x)\\
&= \sum_{y \in \mathcal{Y}} \sup_{x' : \|x - x'\| \le \gamma} p(x,y) \Phi^\tau\left(\sum_{y' \ne y} \Phi_\rho(h^*(x', y) - h^*(x', y'))\right)\\
&= \sum_{i=1}^{c} \sup_{x' : \|x - x'\| \le \gamma} p(x, \{i\}_{x'}^{h^*}) \Phi^\tau\left[\sum_{j=1}^{i-1} \Phi_\rho\Big(h^*(x', \{i\}_{x'}^{h^*}) - h^*(x', \{j\}_{x'}^{h^*})\Big) + \sum_{j=i+1}^{c} \Phi_\rho\Big(h^*(x', \{i\}_{x'}^{h^*}) - h^*(x', \{j\}_{x'}^{h^*})\Big)\right]\\
&= \sum_{i=1}^{c} \sup_{x' : \|x - x'\| \le \gamma} p(x, \{i\}_{x'}^{h^*}) \Phi^\tau\left[\sum_{j=1}^{i-1} \Phi_\rho\Big(h^*(x', \{i\}_{x'}^{h^*}) - h^*(x', \{j\}_{x'}^{h^*})\Big) + c - i\right] && (\Phi_\rho(t) = 1,\ \forall t \le 0)\\
&= \sum_{i=1}^{c} \sup_{x' : \|x - x'\| \le \gamma} p(x, \{i\}_{x'}^{h^*}) \Phi^\tau(c - i) && (\inf_{x' : \|x - x'\| \le \gamma} |h^*(x', i) - h^*(x', j)| \ge \rho \text{ for any } i \ne j \text{ and } \Phi_\rho(t) = 0,\ \forall t \ge \rho)\\
&= \sum_{i=1}^{c} p_{[i]}(x) \Phi^\tau(c - i). && (p(x, \{k\}_{x'}^{h^*}) = p_{[k]}(x), \forall x' \in \{x' : \|x - x'\| \le \gamma\}, \forall k \in \mathcal{Y})
\end{aligned}$$

Note $\overline{\mathcal{H}}_\gamma(x) \ne \varnothing$ under the assumption. Then, use the derivation above, we obtain

$$\begin{aligned}
&\Delta\mathcal{C}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}, \mathcal{H}}(h, x)\\
&= \sum_{i=1}^{c} \sup_{x' : \|x - x'\| \le \gamma} p(x, \{i\}_{x'}^{h}) \Phi^\tau\left[\sum_{j=1}^{i-1} \Phi_\rho\Big(h(x', \{i\}_{x'}^{h}) - h(x', \{j\}_{x'}^{h})\Big) + c - i\right] - \sum_{i=1}^{c} p_{[i]}(x) \Phi^\tau(c - i)\\
&\ge \Phi^\tau(1)\, p(x, \mathsf{h}(x)) \mathbb{1}_{h \notin \overline{\mathcal{H}}_\gamma(x)} + \sum_{i=1}^{c} \sup_{x' : \|x - x'\| \le \gamma} p(x, \{i\}_{x'}^{h}) \Phi^\tau(c - i) - \sum_{i=1}^{c} p_{[i]}(x) \Phi^\tau(c - i)\\
&\qquad\qquad (\Phi_\rho \text{ is non-negative and } \Phi^\tau \text{ is non-decreasing})\\
&\ge \Phi^\tau(1)\, p(x, \mathsf{h}(x)) \mathbb{1}_{h \notin \overline{\mathcal{H}}_\gamma(x)} + \sum_{i=1}^{c} p(x, \{i\}_{x}^{h}) \Phi^\tau(c - i) - \sum_{i=1}^{c} p_{[i]}(x) \Phi^\tau(c - i) \quad (\sup_{x' : \|x - x'\| \le \gamma} p(x, \{i\}_{x'}^{h}) \ge p(x, \{i\}_{x}^{h})\\
&= \Phi^\tau(1)\, p(x, \mathsf{h}(x)) \mathbb{1}_{h \notin \overline{\mathcal{H}}_\gamma(x)} + \Phi^\tau(1)\Big(\max_{y \in \mathcal{Y}} p(x,y) - p(x, \mathsf{h}(x))\Big)\\
&\qquad + \begin{bmatrix} \Phi^\tau(1)\\ \Phi^\tau(1)\\ \Phi^\tau(2)\\ \vdots\\ \Phi^\tau(c-1) \end{bmatrix} \cdot \begin{bmatrix} p(x, \{c\}_x^h)\\ p(x, \{c-1\}_x^h)\\ p(x, \{c-2\}_x^h)\\ \vdots\\ p(x, \{1\}_x^h) \end{bmatrix} - \begin{bmatrix} \Phi^\tau(1)\\ \Phi^\tau(1)\\ \Phi^\tau(2)\\ \vdots\\ \Phi^\tau(c-1) \end{bmatrix} \cdot \begin{bmatrix} p_{[c]}(x)\\ p_{[c-1]}(x)\\ p_{[c-2]}(x)\\ \vdots\\ p_{[1]}(x) \end{bmatrix}\\
&\qquad\qquad (p_{[c]}(x) = \max_{y \in \mathcal{Y}} p(x,y),\ \{c\}_x^h = \mathsf{h}(x) \text{ and } \Phi^\tau(0) = 0)\\
&\ge \Phi^\tau(1)\, p(x, \mathsf{h}(x)) \mathbb{1}_{h \notin \overline{\mathcal{H}}_\gamma(x)} + \Phi^\tau(1)\Big(\max_{y \in \mathcal{Y}} p(x,y) - p(x, \mathsf{h}(x))\Big)\\
&\qquad\qquad (\text{rearrangement inequality for } \Phi^\tau(1) \le \Phi^\tau(1) \le \Phi^\tau(2) \le \cdots \le \Phi^\tau(c-1) \text{ and } p_{[c]}(x) \ge \cdots \ge p_{[1]}(x))\\
&= \Phi^\tau(1)\Big(\max_{y \in \mathcal{Y}} p(x,y) - p(x, \mathsf{h}(x))\mathbb{1}_{h \in \overline{\mathcal{H}}_\gamma(x)}\Big)
\end{aligned}$$

for any $h \in \mathcal{H}$. Since $\mathcal{H}$ is symmetric and $\overline{\mathcal{H}}_\gamma(x) \neq \varnothing$, we have

$$\begin{aligned}
\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x) &= \mathcal{C}_{\ell_\gamma}(h,x) - \mathcal{C}^*_{\ell_\gamma}(\mathcal{H},x)\\
&= \sum_{y\in\mathcal{Y}} p(x,y) \sup_{x':\|x-x'\|\leq\gamma} \mathbb{1}_{\rho_h(x',y)\leq 0} - \inf_{h\in\mathcal{H}} \sum_{y\in\mathcal{Y}} p(x,y) \sup_{x':\|x-x'\|\leq\gamma} \mathbb{1}_{\rho_h(x',y)\leq 0}\\
&= (1-p(x,\mathsf{h}(x)))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)} + \mathbb{1}_{h\notin\overline{\mathcal{H}}_\gamma(x)} - \inf_{h\in\mathcal{H}}\Big[(1-p(x,\mathsf{h}(x)))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)} + \mathbb{1}_{h\notin\overline{\mathcal{H}}_\gamma(x)}\Big]\\
&= (1-p(x,\mathsf{h}(x)))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)} + \mathbb{1}_{h\notin\overline{\mathcal{H}}_\gamma(x)} - \Big(1-\max_{y\in\mathcal{Y}} p(x,y)\Big) \qquad (\mathcal{H} \text{ is symmetric and } \overline{\mathcal{H}}_\gamma(x)\neq\varnothing)\\
&= \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)}.
\end{aligned}$$

Therefore, by the definition, we obtain

$$\begin{aligned}
\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma}(\mathcal{H}) + \mathcal{M}_{\ell_\gamma}(\mathcal{H}) &= \mathbb{E}_X\big[\Delta\mathcal{C}_{\ell_\gamma}(h,x)\big]\\
&= \mathbb{E}_X\Big[\max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)}\Big]\\
&\leq \Phi^\tau(1)\,\mathbb{E}_X\Big[\Delta\mathcal{C}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho},\mathcal{H}}(h,x)\Big]\\
&= \Phi^\tau(1)\Big(\mathcal{R}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(h) - \mathcal{R}^*_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H})\Big),
\end{aligned}$$

which implies that

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma}(\mathcal{H}) \leq \Phi^\tau(1)\Big(\mathcal{R}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(h) - \mathcal{R}^*_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\ell}^{\text{comp}}_{\tau,\rho}}(\mathcal{H})\Big) - \mathcal{M}_{\ell_\gamma}(\mathcal{H}).$$

□

G. Learning bounds (proof of Theorem 3.3)

Theorem 3.3. *With probability at least $1-\delta$ over the draw of a sample S from $\mathcal{D}^m$, the following zero-one loss estimation bound holds for an empirical minimizer $\widehat{h}_S \in \mathcal{H}$ of the comp-sum loss ℓ^{comp}_τ over S:*

$$\mathcal{R}_{\ell_{0-1}}(\widehat{h}_S) - \mathcal{R}^*_{\ell_{0-1}}(\mathcal{H}) \leq \Gamma_\tau\left(\mathcal{M}_{\ell^{\text{comp}}_\tau}(\mathcal{H}) + 4\mathfrak{R}^\tau_m(\mathcal{H}) + 2B_\tau\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right) - \mathcal{M}_{\ell_{0-1}}(\mathcal{H}).$$

Proof. By the standard Rademacher complexity bounds (Mohri et al., 2018), the following holds with probability at least $1-\delta$ for all $h \in \mathcal{H}$:

$$\left|\mathcal{R}_{\ell^{\text{comp}}_\tau}(h) - \widehat{\mathcal{R}}_{\ell^{\text{comp}}_\tau,S}(h)\right| \leq 2\mathfrak{R}^\tau_m(\mathcal{H}) + B_\tau\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Fix $\epsilon > 0$. By the definition of the infimum, there exists $h^* \in \mathcal{H}$ such that $\mathcal{R}_{\ell^{\text{comp}}_\tau}(h^*) \leq \mathcal{R}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}) + \epsilon$. By definition of $\widehat{h}_S$, we have

$$\begin{aligned}
&\mathcal{R}_{\ell^{\text{comp}}_\tau}(\widehat{h}_S) - \mathcal{R}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H})\\
&= \mathcal{R}_{\ell^{\text{comp}}_\tau}(\widehat{h}_S) - \widehat{\mathcal{R}}_{\ell^{\text{comp}}_\tau,S}(\widehat{h}_S) + \widehat{\mathcal{R}}_{\ell^{\text{comp}}_\tau,S}(\widehat{h}_S) - \mathcal{R}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H})\\
&\leq \mathcal{R}_{\ell^{\text{comp}}_\tau}(\widehat{h}_S) - \widehat{\mathcal{R}}_{\ell^{\text{comp}}_\tau,S}(\widehat{h}_S) + \widehat{\mathcal{R}}_{\ell^{\text{comp}}_\tau,S}(h^*) - \mathcal{R}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H})\\
&\leq \mathcal{R}_{\ell^{\text{comp}}_\tau}(\widehat{h}_S) - \widehat{\mathcal{R}}_{\ell^{\text{comp}}_\tau,S}(\widehat{h}_S) + \widehat{\mathcal{R}}_{\ell^{\text{comp}}_\tau,S}(h^*) - \mathcal{R}^*_{\ell^{\text{comp}}_\tau}(h^*) + \epsilon\\
&\leq 2\left[2\mathfrak{R}^\tau_m(\mathcal{H}) + B_\tau\sqrt{\frac{\log(2/\delta)}{2m}}\right] + \epsilon.
\end{aligned}$$

Since the inequality holds for all $\epsilon > 0$, it implies:

$$\mathcal{R}_{\ell^{\text{comp}}_\tau}(\widehat{h}_S) - \mathcal{R}^*_{\ell^{\text{comp}}_\tau}(\mathcal{H}) \leq 4\mathfrak{R}^\tau_m(\mathcal{H}) + 2B_\tau\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Plugging in this inequality in the bound of Theorem 3.1 completes the proof. □