Proceedings of Machine Learning Research vol 195:1–23, 2023 36th Annual Conference on Learning Theory

Proper Losses, Moduli of Convexity, and Surrogate Regret Bounds

Han Bao BAO@I.KYOTO-U.AC.JP
Kyoto University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Proper losses (or proper scoring rules) have been used for over half a century to elicit users' subjective probability from the observations. In the recent machine learning community, we often tackle downstream tasks such as classification and bipartite ranking with the elicited probabilities. Here, we engage in assessing the quality of the elicited probabilities with different proper losses, which can be characterized by surrogate regret bounds to describe the convergence speed of an estimated probability to the optimal one when optimizing a proper loss. This work contributes to a sharp analysis of surrogate regret bounds in two ways. First, we provide general surrogate regret bounds for proper losses measured by the L^1 distance. This abstraction eschews a tailor-made analysis of each downstream task and delineates how universally a loss function operates. Our analysis relies on a classical mathematical tool known as the moduli of convexity, which is of independent interest per se. Second, we evaluate the surrogate regret bounds with polynomials to identify the quantitative convergence rate. These devices enable us to compare different losses, with which we confirm that the lower bound of the surrogate regret bounds is $\Omega(\varepsilon^{1/2})$ for popular loss functions.

1. Introduction

Proper losses (Buja et al., 2005) are loss functions to measure the discrepancy between a probabilistic prediction and an expected outcome, and have been studied in the field of statistics (Shuford et al., 1966; Savage, 1971; Schervish, 1989; Gneiting and Raftery, 2007) and machine learning (Masnadi-Shirazi and Vasconcelos, 2009; Reid and Williamson, 2009b, 2010; Agarwal, 2014). The celebrated log loss belongs to this class corresponding to the Shannon entropy. In this view, proper losses are a generalization of the log loss with respect to the Shannon entropy, often referred to as *generalized entropies*. The elicited probability via proper loss minimization is then used for downstream tasks such as classification (by choosing the most plausible outcome) and bipartite ranking (by scoring more plausible items higher) (Narasimhan and Agarwal, 2013). Therefore, we are interested in quantitatively characterizing the performance of the elicit probability under a downstream task. A *surrogate regret bound* is useful in this regard, where the suboptimality of a given probability (called *regret*) under a downstream task is upper bounded by the regret of a proper loss.

In the previous literature, surrogate regret bounds have been derived for the binary classification problem (Reid and Williamson, 2009b), bipartite ranking (Agarwal, 2014), and property elicitation (Agarwal and Agarwal, 2015). These works leverage the one-to-one correspondence between a proper loss and a generalized entropy (Savage, 1971), and a regret bound of a proper loss is written by the corresponding generalized entropy (as we will review in Section 2). In a different line, Frongillo and Waggoner (2021) recently showed that surrogate regret bounds of binary classification cannot be faster than the square-root rate when a loss function is sufficiently smooth, and highlighted polyhedral losses (Finocchiaro et al., 2019) achieving the linear regret rate. However, these works have derived a surrogate regret bound for each downstream problem independently, and moreover, the bounds are expressed in terms of generalized entropies and are not straightforwardly comparable with each other.

In this work, we aim at providing a unified framework to derive surrogate regret bounds and quantify their polynomial rate. To unify surrogate regret bounds of different downstream tasks, we focus on regret bounds of the L^1 *distance*, taking the following form:

$$|\eta - \hat{\eta}| \le \psi(R_\ell(\eta, \hat{\eta})), \tag{1}$$

where η and $\hat{\eta}$ are true and estimated probabilities, respectively, R_ℓ is the regret of a proper loss ℓ, and ψ is the regret rate function. The formal definitions will be stated later. We refer to Eq. (1) as the L^1 *regret bounds*. We consider the L^1 regret bound because the regrets of many downstream tasks can be upper bounded by $|\eta - \hat{\eta}|$ (Reid and Williamson, 2009b; Menon et al., 2013; Agarwal, 2014). Then, our first main theorem (Theorem 6) shows that the rate function ψ can be expressed by the *moduli of convexity* (Figiel, 1976), which describe curvature information of convex functions, of a generalized entropy. Our second main theorem (Theorem 10) yields the polynomial regret rate such that $\psi(\varepsilon) = \Omega(\varepsilon^r)$ and $\psi(\varepsilon) = O(\varepsilon^R)$ for some r and R. These polynomial rates are far more interpretable than the original ψ and can be used to compare the convergence speed of different proper losses. With these main results, we show that $\psi(\varepsilon) = \Omega(\sqrt{\varepsilon})$ for a number of loss functions (Proposition 11). We believe that these results altogether support the design of a new loss function from the perspective of statistical estimation.

Related work. Moduli of convexity are key components in this paper. The moduli have been initially introduced to study and generalize uniformly convex Banach spaces (Figiel, 1976) and uniformly convex functions later (Borwein et al., 2009). The moduli have been applied to rate estimation of the Bregman divergences (Sprung, 2019), whose idea is partially related to the current paper, but we are concerned with the Jensen–Bregman divergence (see Section 3.1). Several works on statistics and machine learning leveraged the moduli to analyze estimation errors (Mendelson, 2002; Bartlett et al., 2006). The polynomial evaluation of the moduli have been provided for L^p space (Hanner, 1956) and Orlicz space (Hudzik, 1991). The derivation of the polynomial rates (including our results) often relies on the Simonenko index (Simonenko, 1964). Ishige et al. (2022) studied a hierarchy of generalized convex functions via an analogous analysis. Mey and Loog (2021) showed error analysis of class probability estimation by leveraging moduli of continuity, which is a closely related yet different notion from the moduli of convexity.

Despite that the connection between moduli of convexity and loss functions has not been pointed out so far, Zhang (2004, Theorem 2.1) derived surrogate regret bounds for binary classification based on a similar idea to leverage the L^1 distance. Later, Steinwart (2007), Osokin et al. (2017), and Bao et al. (2020) have introduced a notion called *calibration function* to characterize surrogate regret bounds for classification problems, which is akin to moduli of convexity in this paper. We leverage moduli of convexity to point out a systematic way to derive surrogate regret bounds for various downstream tasks (concerning binary outcomes) beyond classification problems.

2. Review of loss function structures

In this section, we introduce the basics of loss functions and their structure.

Notation. The extended real line is denoted by $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$. We occasionally use notation $\mathbb{R}_{\ge 0} := [0, \infty)$ and $\overline{\mathbb{R}}_{\ge 0} := [0, \infty]$. In this paper, we mostly focus on a (convex) function over one-dimensional space $\mathcal{X} \subseteq \mathbb{R}$. The domain of a function $f : \mathcal{X} \to \overline{\mathbb{R}}$ is denoted by $\mathrm{dom}(f) := \{x \in \mathcal{X} \mid f(x) < \infty\}$. We write the convex conjugate of f by $f^\star(\eta) := \sup_{x \in \mathrm{dom}(f)} \eta x - f(x)$. The convex biconjugate of f is denoted by $f^{\star\star}$. For a convex function $f : \mathcal{X} \to \overline{\mathbb{R}}$, the associated Bregman divergence (Bregman, 1967) is defined as

$$B_f(x \,\|\, x_0) := f(x) - f(x_0) - f'(x_0)(x - x_0)$$

for any $x, x_0 \in \mathcal{X}$. Here, f' is the right derivative of f, which always exists for convex functions. We do not mention the differentiability to avoid unnecessary technicality unless otherwise noted.[1]

Below, some additional notations are introduced. We write $[x]_+ := \max\{x, 0\}$. For the sign of real values $\mathrm{sign}(x) \in \{0, 1\}$, we adopt the convention $\mathrm{sign}(0) = 0$. The one-dimensional probability simplex is denoted by $\triangle := \left\{\boldsymbol{\eta} \in \mathbb{R}^2_{\geq 0} \;\middle|\; \langle \boldsymbol{\eta}, \mathbf{1} \rangle = 1\right\} \cong [0, 1]$. The Iverson bracket $[\![A]\!]$ is 1 if the predicate A is true otherwise 0. We employ the *infinitesimal* asymptotic notations O, Θ, and Ω. For example, $f(\varepsilon) = O(g(\varepsilon))$ should be understood as $\limsup_{\varepsilon \to 0+} f(\varepsilon)/g(\varepsilon) < \infty$. The other asymptotic notations are interpreted in the same way.

Problem setup. Let $(\mathcal{X}, \mathscr{X}, \mu)$ be a probability space that represents the input space. Throughout this paper, we concentrate on supervised learning associated with the binary outcome space $\mathcal{Y} := \{0, 1\}$. We usually suppose that an underlying probability distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$ generates both i.i.d. training and test samples; nonetheless, the main scope of this paper is a relationship between loss functions and hence most of the discussions involve test samples only. In binary supervised learning, the fundamental target is the class probability $\eta(\mathbf{x}) := \mathbb{P}(Y = 1 \mid X = \mathbf{x}) \in \triangle$ because it is amenable to future changes of a decision rule in, e.g., class-imbalanced problems (Kotłowski and Dembczyński, 2016) and classification with rejection (Charoenphakdee et al., 2021). Occasionally, one may seek problem-specific decision rules such as the Bayes classifier $f^*(\mathbf{x}) := \mathrm{sign}(\eta(\mathbf{x}) - \frac{1}{2})$ and the Bayes scorer for bipartite ranking $s^*(\mathbf{x}) = \iota(\eta(\mathbf{x}))$ for some monotone function ι (Menon and Williamson, 2014). These decision rules are typically "irreversible" transforms of η and therefore class probability estimation is an essential step in supervised learning. Despite that one may do better estimation by directly solving the decision problem of interest (Frongillo and Waggoner, 2021), we shed light on class probability estimation as it serves as a flexible and versatile component in decision-making. We measure the quality of estimated probabilities with the L^1 distance.

Proper losses. Proper losses are typical choices for class probability estimation, which is introduced from now on based on the presentation of Reid and Williamson (2010). Let $\ell : \mathcal{Y} \times \triangle \to \overline{\mathbb{R}}_{\geq 0}$ be a loss function, where $\ell(y, \hat{\eta})$ assesses a probability estimate $\hat{\eta} \in \triangle$ for a binary label $y \in \mathcal{Y}$. Define the *pointwise* risk of a pointwise probability estimate $\hat{\eta} \in \triangle$ at a ground-truth $\eta \in \triangle$ as

$$L_\ell(\eta, \hat{\eta}) := \mathop{\mathbb{E}}_{Y \sim \eta} \ell(Y, \hat{\eta}) = \eta \ell(1, \hat{\eta}) + (1 - \eta)\ell(0, \hat{\eta}),$$

which is the Bernoulli average of a proper loss with mean η. The *pointwise* Bayes risk is the minimal achievable value of $L_\ell(\eta, \cdot)$, denoted as $\underline{L}_\ell(\eta) := \inf_{\hat{\eta} \in \triangle} L_\ell(\eta, \hat{\eta})$. A loss function ℓ is said to be *proper* if $\underline{L}_\ell(\eta) = L_\ell(\eta, \eta)$ for all $\eta \in \triangle$, which is a minimum necessary condition as a reasonable loss. A *strictly proper* loss is a proper loss whose pointwise risk $L_\ell(\eta, \hat{\eta})$ is uniquely minimized at $\hat{\eta} = \eta$. As we will see later, $\underline{L}_\ell$ forms a rich structure of proper losses. The pointwise Bayes risk $\underline{L}_\ell$

[1] Liese and Vajda (2006) discussed the first- and second-order differentiability for non-smooth functions when discussing the Bregman divergences.

is evidently concave over $\triangle$ because it is the pointwise infimum of linear functions (Agarwal, 2014, Lemma 1).

We remark that the *full* risk for a (μ-measurable) class probability estimator $\hat{\eta} : \mathcal{X} \to \triangle$ is recovered from the pointwise risk as $\mathbb{L}_\ell[\eta, \hat{\eta}] \coloneqq \mathbb{E}_{X\sim\mu} L_\ell(\eta(X), \hat{\eta}(X))$, with the (full) Bayes risk $\underline{\mathbb{L}}_\ell[\eta] \coloneqq \inf_{\hat{\eta}\in\triangle^\mathcal{X}} \mathbb{L}_\ell[\eta, \hat{\eta}] = \mathbb{E}_{X\sim\mu} \underline{L}_\ell(\eta(X))$.[2] The full Bayes risk determines the difficulty of a task. As seen here, we use η for both pointwise and full estimators with abuse of notation but it should be clear from the context.

The pointwise Bayes risk $\underline{L}_\ell$, sometimes referred to as the *generalized entropy*, induces much of the structure of a proper loss. Before showing them, an additional technical condition is introduced. We say a proper loss ℓ is *regular* if $\ell(y, \hat{\eta}) \in \mathbb{R}_{\geq 0}$ for all $y \in \mathcal{Y}$ and $\hat{\eta} \in \triangle$ except possibly that $\ell(0, 1) = \infty$ or $\ell(1, 0) = \infty$ (Savage, 1971; Gneiting and Raftery, 2007; Agarwal, 2014). With this condition, a proper loss is connected to the Bregman divergence.

Theorem 1 (Savage (1971)) *A regular loss $\ell : \mathcal{Y} \times \triangle \to \overline{\mathbb{R}}_{\geq 0}$ is proper if and only if for all $\eta, \hat{\eta} \in \triangle$,*

$$L_\ell(\eta, \hat{\eta}) = \underline{L}_\ell(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'_\ell(\hat{\eta}). \tag{2}$$

*In addition, the (pointwise) ℓ-*regret *is equal to the Bregman divergence generated with $-\underline{L}_\ell$:*

$$R_\ell(\eta, \hat{\eta}) \coloneqq L_\ell(\eta, \hat{\eta}) - \underline{L}_\ell(\eta) = B_{-\underline{L}_\ell}(\eta \,\|\, \hat{\eta}). \tag{3}$$

The regularity is needed to ensure Eq. (2) at the endpoints $\eta, \hat{\eta} \in \{0, 1\}$. The regret measures the suboptimality of an estimate $\hat{\eta}$ from η, following the geometry induced by $-\underline{L}_\ell$.

It is further known that the strict properness of ℓ can be tested with the generalized entropy.

Theorem 2 (Theorem 4 in Agarwal (2014)) *ℓ is strictly proper if and only if $\underline{L}_\ell$ is strictly concave.*

From these statements, we see that a generalized entropy is a powerful device to characterize a proper loss. Given Theorem 1, it is possible to generate a proper loss from a concave function.

Corollary 3 *Given a concave function $H : \triangle \to \mathbb{R}_{\geq 0}$, define a binary loss $\ell_H : \mathcal{Y} \times \triangle \to \overline{\mathbb{R}}_{\geq 0}$ as follows: for all $\hat{\eta} \in \triangle$,*

$$\ell_H(1, \hat{\eta}) \coloneqq H(\hat{\eta}) + (1 - \hat{\eta})H'(\hat{\eta}), \qquad \ell_H(0, \hat{\eta}) \coloneqq H(\hat{\eta}) - \hat{\eta}H'(\hat{\eta}), \tag{4}$$

where $H'(\hat{\eta})$ is the right derivative of H at $\hat{\eta}$. Then, ℓ_H is a regular proper loss.

Later, we discuss properties of proper losses by focusing on a concave function H merely, but we remark that the associated proper loss can be recovered from H by Corollary 3.

When optimizing a proper loss in the machine learning pipeline, a number of previous works operate with an *inverse link function*, which translates a real-valued prediction into a probabilistic prediction $\hat{\eta} \in \triangle$. Specifically, Buja et al. (2005) and Reid and Williamson (2010) emphasized the benefit of *canonical* links because the composition of a proper loss and the corresponding canonical link is convex. In addition, Agarwal et al. (2014) discussed the design of a canonical proper loss for a given link function and Bao and Sugiyama (2021) drew its connection to the *Fenchel–Young losses* (Blondel et al., 2020). In this work, we leave a choice of link functions out of account and devote ourselves to solely analyzing the convergence of probabilistic predictions.

[2] The infimum is taken over all measurable functions. The last identity requires suitable measurability conditions on ℓ, but we do not discuss it in detail. From Steinwart (2007, Theorem 3.2), it is sufficient to have $\underline{L}_\ell(\eta) < \infty$ $(\forall \eta \in \triangle)$.

Regret. Given a joint distribution $\mathbb{P}$, we seek a class probability estimator $\hat{\eta} \in \triangle^{\mathcal{X}}$ that is as close to the optimal class probability function $\eta \in \triangle^{\mathcal{X}}$ as possible. The suboptimality is measured by the (full) ℓ-*regret* associated with a proper loss ℓ:

$$\mathrm{Reg}_\ell[\eta, \hat{\eta}] \coloneqq \mathbb{L}_\ell[\eta, \hat{\eta}] - \underline{\mathbb{L}}_\ell[\eta] = \mathop{\mathbb{E}}_{X \sim \mu} R_\ell(\eta(X), \hat{\eta}(X)). \tag{5}$$

We mainly analyze the pointwise regret R_ℓ as it is free from the variational problem.

A class probability estimator is used for downstream decision-making. We specifically pay attention to binary classification and bipartite ranking. The goal of binary classification is to correctly predict binary label $y \in \mathcal{Y}$, where a classifier $\mathbf{x} \mapsto \mathrm{sign}(\hat{\eta}(\mathbf{x}) - \frac{1}{2})$ built on $\hat{\eta} \in \triangle^{\mathcal{X}}$ is evaluated by the following *0-1 regret* (Reid and Williamson, 2009b, Lemma 8):

$$\mathrm{Reg}_{01}[\eta, \hat{\eta}] \coloneqq \mathop{\mathbb{E}}_{X \sim \mu} R_{01}(\eta(X), \hat{\eta}(X)); \quad R_{01}(\eta, \hat{\eta}) \coloneqq \left|\eta - \tfrac{1}{2}\right| \left[\!\left[\min\{\eta, \hat{\eta}\} \le \tfrac{1}{2} < \max\{\eta, \hat{\eta}\}\right]\!\right]. \tag{6}$$

The goal of bipartite ranking is to predict higher scores for inputs labeled with $y = 1$ than ones with $y = 0$. A class probability estimator $\hat{\eta} \in \triangle^{\mathcal{X}}$ directly used as a scorer is evaluated by the following *ranking regret* (Clémençon et al., 2008, Example 1): for $\pi \coloneqq \mathbb{P}(Y = 1)$,

$$\begin{aligned} \mathrm{Reg}_{\mathrm{rank}}[\eta, \hat{\eta}] &\coloneqq \tfrac{1}{2\pi(1-\pi)} \mathop{\mathbb{E}}_{X, X' \sim \mu^2} R_{\mathrm{rank}}(\eta(X), \eta(X'), \hat{\eta}(X), \hat{\eta}(X')), \\ R_{\mathrm{rank}}(\eta, \eta', \hat{\eta}, \hat{\eta}') &\coloneqq |\eta - \eta'| \left\{ [\![(\hat{\eta} - \hat{\eta}')(\eta - \eta') < 0]\!] + \tfrac{1}{2} [\![\hat{\eta} = \hat{\eta}']\!] \right\}. \end{aligned} \tag{7}$$

To see downstream optimality for $\hat{\eta}$ obtained via proper loss minimization, surrogate regret bounds $R_{01}(\eta, \hat{\eta}) \le \psi_{01}(R_\ell(\eta, \hat{\eta}))$ and $R_{\mathrm{rank}}(\eta, \eta', \hat{\eta}, \hat{\eta}') \le \psi_{\mathrm{rank}}(R_\ell(\eta, \hat{\eta})) + \psi_{\mathrm{rank}}(R_\ell(\eta', \hat{\eta}'))$ are informative. Nevertheless, deriving regret bounds for each downstream task may obfuscate how *universally* a proper loss ℓ behaves well. We instead seek an upper bound of the L^1 distance $|\eta - \hat{\eta}|$ by $R_\ell(\eta, \hat{\eta})$ as an intermediate step because the downstream regrets can be bounded by the L^1 distance such as $R_{01}(\eta, \hat{\eta}) \le |\eta - \hat{\eta}|$ (Menon et al., 2013, Lemma 4) and $R_{\mathrm{rank}}(\eta, \eta', \hat{\eta}, \hat{\eta}') \le |\eta - \hat{\eta}| + |\eta' - \hat{\eta}'|$ (Agarwal, 2014, Corollary 12). For this reason, this work primarily focuses on the L^1 regret bound $|\eta - \hat{\eta}| \le \psi(R_\ell(\eta, \hat{\eta}))$ in Eq. (1) and attempts to derive the rate function ψ.

Previously, Agarwal (2014) proved regret bounds with a loss function class called strongly proper losses: a proper loss $\ell : \mathcal{Y} \times \triangle \to \overline{\mathbb{R}}_{\ge 0}$ is said to be λ-*strongly proper* for $\lambda > 0$ if for all $\eta, \hat{\eta} \in \triangle$,

$$L_\ell(\eta, \hat{\eta}) - \underline{L}_\ell(\eta) \ge \tfrac{\lambda}{2}(\eta - \hat{\eta})^2. \tag{8}$$

Interestingly, a strongly proper loss is tightly related to the strong concavity of a generalized entropy: for $\lambda > 0$, a regular proper loss ℓ is λ-strongly proper if and only if $\underline{L}_\ell$ is λ-strongly concave (Agarwal, 2014, Theorem 10). Popular loss functions such as log loss and exponential loss are strongly proper. The strong properness immediately induces the L^1 regret bound $|\eta - \hat{\eta}| \le \sqrt{2R_\ell(\eta, \hat{\eta})/\lambda}$, and thus $\psi(\varepsilon) = O(\sqrt{\varepsilon})$. This result has been used to derive regret bounds for bipartite ranking (Agarwal, 2014), imbalanced classification (Kotłowski and Dembczyński, 2016), and noisy label classification (Zhang et al., 2021). Our work essentially extends their direction beyond strongly proper losses.

3. Main results

This section explicates our main results: the L^1 regret upper bounds (Theorem 6) and polynomial evaluation of regret rate functions (Theorem 10). These results yield a finer hierarchy of proper losses. The proofs of other technical statements are deferred to Appendix A.

3.1. Preparation: Moduli of convexity

Before the presentation, a key device for analysis, *modulus of convexity*, is introduced.

Definition 4 (Modulus of convexity) *Given a proper convex function $f : \triangle \to \overline{\mathbb{R}}$, the* modulus of convexity *of f is the function $\delta_f : [0,1] \to \overline{\mathbb{R}}_{\geq 0}$ satisfying*

$$\delta_f(\varepsilon) := \inf \left\{ \frac{f(\eta) + f(\eta')}{2} - f\left(\frac{\eta + \eta'}{2}\right) \,\middle|\, \eta, \eta' \in \triangle, |\eta - \eta'| \geq \varepsilon \right\}. \tag{9}$$

Definition 4 follows Borwein et al. (2009), and is the infimum of the Jensen–Bregman divergence generated by f (Nielsen and Boltz, 2011): $J_f(\eta \,\|\, \eta') := \frac{f(\eta)+f(\eta')}{2} - f(\frac{\eta+\eta'}{2})$. As the Jensen–Bregman divergence has a connection to the Bregman divergence such as $J_f(\eta \,\|\, \eta') = \frac{1}{2}\{B_f(\eta \,\|\, \frac{\eta+\eta'}{2}) + B_f(\eta' \,\|\, \frac{\eta+\eta'}{2})\}$, some works define moduli based on the Bregman divergence (Sprung, 2019). The Jensen–Bregman divergence is more convenient for our purpose because it admits an upper bound with the total variation (Lin, 1991), moreover, it is expressed without the derivative of the generator f.

The moduli of convexity are non-negative, nondecreasing, and satisfy $\delta_f(0) = 0$. These basic properties can be confirmed by Jensen's inequality and the monotonicity of the domain of the infimum. Thanks to these properties, the biconjugate $\delta_f^{\star\star}$ is a proper convex function. Note that the modulus $\delta_f(\varepsilon)$ itself is not necessarily convex in ε.[3] The moduli have richer information about the convexity of a function, which we state below for the sake of completeness.

Proposition 5 *Suppose that $f : \triangle \to \mathbb{R}$ is convex and lower semicontinuous. Then, f is strictly convex if and only if $\delta_f(\varepsilon) > 0$ for all $\varepsilon \in (0,1]$. Further, f is strongly convex if and only if there exists $\kappa > 0$ such that $\delta_f(\varepsilon) \geq \kappa\varepsilon^2$ for all $\varepsilon \in [0,1]$.*

3.2. Main result 1: Regret bounds

The first main result is the L^1 regret upper bounds of proper losses. This result leads to a unified proof of the 0-1 regret bounds via the Bregman divergence (Reid and Williamson, 2009b, Theorem 3), the 0-1 regret bounds via strongly proper losses (Menon et al., 2013, Lemma 4), and the ranking regret bounds via strongly proper losses (Agarwal, 2014, Theorem 13).

Theorem 6 (Regret upper bounds) *For a regular proper loss $\ell : \mathcal{Y} \times \triangle \to \overline{\mathbb{R}}_{\geq 0}$, the following inequality holds for all $\eta, \hat{\eta} \in \triangle$:*

$$\delta_{-\underline{L}_\ell}(|\eta - \hat{\eta}|) \leq \frac{1}{2} R_\ell(\eta, \hat{\eta}). \tag{10}$$

The inequality is tight. If $\delta_{-\underline{L}_\ell}^{\star\star}$ is invertible, then $|\eta - \hat{\eta}| \leq (\delta_{-\underline{L}_\ell}^{\star\star})^{-1}(\frac{1}{2}R_\ell(\eta, \hat{\eta}))$ for all $\eta, \hat{\eta} \in \triangle$.

Proof By the definition of $\delta_{-\underline{L}_\ell}$, we have $J_{-\underline{L}_\ell}(\eta \,\|\, \hat{\eta}) \geq \delta_{-\underline{L}_\ell}(|\eta - \hat{\eta}|)$. In addition, Theorem 1 ensures $R_\ell(\eta, \hat{\eta}) = B_{-\underline{L}_\ell}(\eta \,\|\, \hat{\eta})$. To show Eq. (10), the remaining piece is to show $\frac{1}{2}B_{-\underline{L}_\ell}(\eta \,\|\, \hat{\eta}) \geq J_{-\underline{L}_\ell}(\eta \,\|\, \hat{\eta})$ for all $\eta, \hat{\eta} \in \triangle$. Figure 1 graphically illustrates the relationship between $B_{-\underline{L}_\ell}(\eta \,\|\, \hat{\eta})$ and $J_{-\underline{L}_\ell}(\eta \,\|\, \hat{\eta})$. The convexity of $-\underline{L}_\ell$ implies

$$-\underline{L}_\ell\left(\frac{\eta + \hat{\eta}}{2}\right) \geq -\underline{L}_\ell(\hat{\eta}) - \underline{L}_\ell'(\hat{\eta})\left(\frac{\eta + \hat{\eta}}{2} - \hat{\eta}\right) = -\underline{L}_\ell(\hat{\eta}) - \frac{1}{2}\underline{L}_\ell'(\hat{\eta})(\eta - \hat{\eta}).$$

[3] A counterexample is the (negative) norm entropies with, e.g., $\alpha = 3$ (see Section 4).

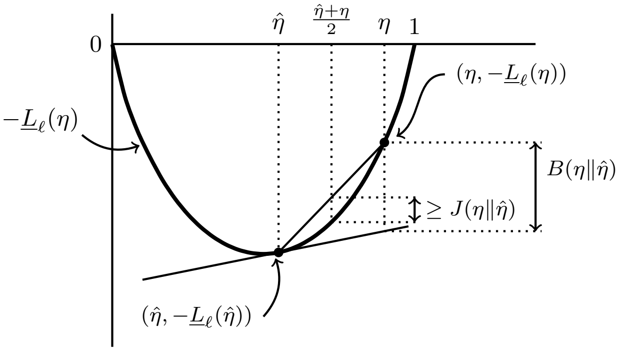
$\hat\eta$
$\frac{\hat\eta+\eta}{2}$
η
1
0
$(\eta, -\underline{L}_\ell(\eta))$
$-\underline{L}_\ell(\eta)$
$B(\eta\|\hat\eta)$
$\geq J(\eta\|\hat\eta)$
$(\hat\eta, -\underline{L}_\ell(\hat\eta))$

Figure 1: Relationship between the Jensen–Bregman divergence $J_{-\underline{L}_\ell}(\eta \,\|\, \hat\eta)$ and Bregman divergence $B_{-\underline{L}_\ell}(\eta \,\|\, \hat\eta)$. $J_{-\underline{L}_\ell}(\eta \,\|\, \hat\eta)$ is the gap between the secant and $-\underline{L}_\ell(\eta)$ at the midpoint $\frac{\hat\eta+\eta}{2}$, which is majorized by $B_{-\underline{L}_\ell}(\eta \,\|\, \hat\eta)$, the first-order approximation error of $-\underline{L}_\ell(\hat\eta)$ at η.

This indicates the desiderata:

$$J_{-\underline{L}_\ell}(\eta \,\|\, \hat\eta) \leq -\frac{\underline{L}_\ell(\eta) + \underline{L}_\ell(\hat\eta)}{2} + \underline{L}_\ell\left(\frac{\eta+\hat\eta}{2}\right) \leq \frac{1}{2}\underbrace{\left\{-\underline{L}_\ell(\eta) + \underline{L}_\ell(\hat\eta) + \underline{L}'_\ell(\hat\eta)(\eta-\hat\eta)\right\}}_{=B_{-\underline{L}_\ell}(\eta \,\|\, \hat\eta)}.$$

The tightness of the inequality is easy to see by choosing $\hat\eta = \eta$.

■

Therefore, we emphasize that the L^1 regret bound (1) is dominated by $\psi = (\delta^{\star\star}_{-\underline{L}_\ell})^{-1}$. Theorem 6 can be viewed as a generalized Pinsker's inequality (Reid and Williamson, 2009a) in that the L^1 distance $|\eta - \hat\eta|$ is bounded by the Bregman divergence $R_\ell(\eta, \hat\eta)$, which generalizes the Kullback–Leibler divergence. It does not only assess the quality of class probability estimation in terms of the L^1 distance but also recover the several existing regret bounds for specific downstream tasks.

Corollary 7 (0-1 regret bounds) *For a regular proper loss ℓ with invertible $\delta^{\star\star}_{-\underline{L}_\ell}$, the 0-1 regret is bounded as follows:* $\mathrm{Reg}_{01}[\eta, \hat\eta] \leq (\delta^{\star\star}_{-\underline{L}_\ell})^{-1}\left(\frac{1}{2}\mathrm{Reg}_\ell[\eta, \hat\eta]\right)$ *for all* $\eta, \hat\eta \in \triangle^{\mathcal{X}}$.

Corollary 8 (Ranking regret bounds) *For a regular proper loss ℓ with invertible $\delta^{\star\star}_{-\underline{L}_\ell}$, the ranking regret is bounded as follows:* $\mathrm{Reg}_{\mathrm{rank}}[\eta, \hat\eta] \leq \frac{1}{\pi(1-\pi)}(\delta^{\star\star}_{-\underline{L}_\ell})^{-1}\left(\frac{1}{2}\mathrm{Reg}_\ell[\eta, \hat\eta]\right)$ *for all* $\eta, \hat\eta \in \triangle^{\mathcal{X}}$, *where* $\pi := \mathbb{P}(Y = 1)$.

Corollary 7 is akin to the 0-1 regret bounds of Reid and Williamson (2011, Corollary 27) with a symmetric $-\underline{L}_\ell$,[4] and Corollary 8 is an extension of Agarwal (2014, Theorem 13) beyond strongly proper losses. Thus, we see that the modulus of convexity of a generalized entropy $-\underline{L}_\ell$ enables us to derive the regret bounds in a comprehensive way.

[4] We say a generalized entropy $\underline{L}_\ell$ is *symmetric* when $\underline{L}_\ell(\eta) = \underline{L}_\ell(1-\eta)$.

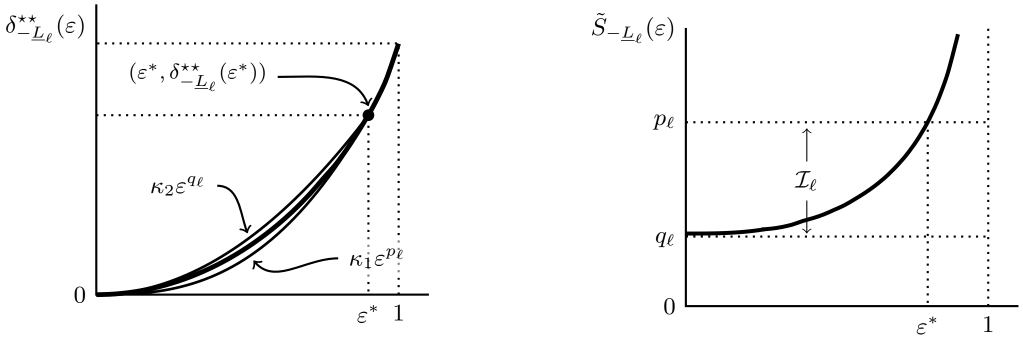

Figure 2: **(Left)** The modulus of convexity $\delta^{\star\star}_{-\underline{L}_\ell}$ and the corresponding rate evaluation $\kappa_1\varepsilon^{p_\ell} \le \delta^{\star\star}_{-\underline{L}_\ell}(\varepsilon) \le \kappa_2\varepsilon^{q_\ell}$ (Theorem 10). **(Right)** The Simonenko order function $\tilde{S}_{-\underline{L}_\ell}$ and the corresponding upper/lower order p_ℓ and q_ℓ.

3.3. Main result 2: Polynomial rate evaluation

The second main result is the polynomial rate evaluation of the regret bounds. The regret bounds obtained in Section 3.2 are expressed by $(\delta^{\star\star}_{-\underline{L}_\ell})^{-1}$, which would not immediately tell us much how fast the convergences are. Ideally, the polynomial rate evaluation of $(\delta^{\star\star}_{-\underline{L}_\ell})^{-1}$ (equivalently, the rate of $\delta^{\star\star}_{-\underline{L}_\ell}$) would provide us much more insights into loss comparison. More specifically, the polynomial rate of $\delta^{\star\star}_{-\underline{L}_\ell}$ close to the origin $\varepsilon = 0$ is of the most importance because it governs the convergence speed around the optimum.

Hence, we first define a concept to characterize the polynomial rate of a given function. We call it the *Simonenko order function* as it owes a lot to the earlier studies on the relevant concept called the Simonenko indices (Simonenko, 1964; Hudzik, 1991; Fiorenza and Krbec, 1997).

Definition 9 *For a non-negative, nondecreasing, and convex $f : [0, C] \to \overline{\mathbb{R}}_{\ge 0}$ for some $C > 0$ satisfying $f(t) = 0$ if and only if $t = 0$,* Simonenko order function *$S_f : (0, C] \to \overline{\mathbb{R}}$ is defined as follows:*

$$S_f(t) := \frac{tf'(t)}{f(t)}. \tag{11}$$

Owing to the Simonenko order function, we are ready to evaluate the polynomial rate of $\delta^{\star\star}_{-\underline{L}_\ell}$. For simplicity, we use an auxiliary notation $\tilde{S}_H := S_{\delta^{\star\star}_H}$ specifically for the Simonenko order function of the biconjugated moduli. In the definition of the Simonenko order function of $\delta^{\star\star}_H$, $(\delta^{\star\star}_H)'$ should be understood as the right derivative, which always exists for convex function $\delta^{\star\star}_H$.

Theorem 10 (Polynomial rate evaluation) *Let $\ell : \mathcal{Y} \times \triangle \to \overline{\mathbb{R}}_{\ge 0}$ be a regular and strictly proper loss. For a fixed $\varepsilon_* \in (0, 1]$, define the interval $\mathcal{I}_\ell := \tilde{S}_{-\underline{L}_\ell}((0, \varepsilon_*])$, $p_\ell := \sup \mathcal{I}_\ell$, and $q_\ell := \inf \mathcal{I}_\ell$. Then, the following inequalities hold: for all $\varepsilon \in [0, \varepsilon_*]$,*

$$\kappa_1\varepsilon^{p_\ell} \le \delta^{\star\star}_{-\underline{L}_\ell}(\varepsilon) \le \kappa_2\varepsilon^{q_\ell}, \qquad \text{where } \kappa_1 := \frac{\delta^{\star\star}_{-\underline{L}_\ell}(\varepsilon_*)}{\varepsilon_*^{p_\ell}} \text{ and } \kappa_2 := \frac{\delta^{\star\star}_{-\underline{L}_\ell}(\varepsilon_*)}{\varepsilon_*^{q_\ell}}. \tag{12}$$

Proof In this proof, we simply write δ instead of $\delta^{\star\star}_{-\underline{L}_\ell}$. The Simonenko order function is well-defined for δ because ℓ is strictly proper, which implies $\delta(\varepsilon) > 0$ for $\varepsilon > 0$ (Proposition 5).

The following proof idea is based on Hudzik (1991). As the two inequalities in Eq. (12) can be proven essentially in the same manner, we only show the proof for $\kappa_1\varepsilon^{p_\ell} \le \delta(\varepsilon)$. By the definition of p_ℓ, we have $p_\ell = \sup_{t\in(0,\varepsilon_*]} \frac{t\delta'(t)}{\delta(t)} \ge \frac{t\delta'(t)}{\delta(t)}$ for all $t \in (0, \varepsilon_*]$. Then, for all $\varepsilon \in (0, \varepsilon_*]$,

$$p_\ell \int_\varepsilon^{\varepsilon_*} \frac{\mathrm{d}t}{t} \ge \int_\varepsilon^{\varepsilon_*} \frac{\delta'(t)}{\delta(t)} \mathrm{d}t.$$

Here, the integrals should be understood as the Henstock–Kurzweil integral (Bartle, 2001), which recovers the Riemann integral for Riemann integrable functions. We need this technicality to handle the (possibly) noncontinuous integrand $\delta'(t)/\delta(t)$. This integrand is the right derivative of the primitive function $\ln\delta(t)$ and hence $\ln\delta(t)$ has the derivative $\delta'(t)/\delta(t)$ all but countably many points on $(0, \varepsilon_*)$. This is because proper convex δ is differentiable almost everywhere in $(0, \varepsilon_*)$ (Rockafellar, 1970, Theorem 25.5). Then, by the generalized second fundamental theorem of calculus (Bartle, 2001, Theorem 4.7), the definite integral of $\delta'(t)/\delta(t)$ is $\ln\delta(\varepsilon_*) - \ln\delta(\varepsilon)$. Hence, we obtain $p_\ell(\ln\varepsilon_* - \ln\varepsilon) \ge \ln\delta(\varepsilon_*) - \ln\delta(\varepsilon)$, from which we can show the desired inequality. ■

From Theorem 10, we have $(\delta_H^{\star\star})^{-1}(\varepsilon) = \Omega(\varepsilon^{1/p_\ell})$ and $(\delta_H^{\star\star})^{-1}(\varepsilon) = O(\varepsilon^{1/q_\ell})$. This implies that smaller p_ℓ and q_ℓ induce a faster regret rate. Additionally, the polynomial rate evaluation becomes better as the interval $\mathcal{I}_\ell$ becomes smaller. When the upper and lower rates match ($p_\ell = q_\ell$), the inequality (12) is tight: $\delta_{-\underline{L}_\ell}^{\star\star}(\varepsilon) = \kappa_1\varepsilon^{p_\ell}$. In a nutshell, the regret convergence rate of different proper losses can be compared by looking at the interval $\mathcal{I}_\ell$. Figure 2 illustrates the polynomial rate evaluation of $\delta_{-\underline{L}_\ell}^{\star\star}$ (Left) and the relationship between $\tilde{S}_{-\underline{L}_\ell}$ and $\mathcal{I}_\ell$ (Right).

Note that, though the analysis of Theorem 10 is specialized for $\delta_{-\underline{L}_\ell}^{\star\star}$, it should apply to broader contexts. Examples of the Simonenko order functions are shown in Section 4.

4. Examples of loss functions

We summarize examples of losses in terms of their associated (negative) Bayes risks $H := -\underline{L}_\ell$ below. A few examples are borrowed from Blondel et al. (2020). The detailed computations are shown in Appendix B.

Shannon entropy. Consider $H(\eta) = \eta\ln\eta + (1-\eta)\ln(1-\eta)$. The associated proper loss is the *log loss* (Buja et al., 2005), a common loss function in the maximum likelihood and logistic regression. Then, $\delta_H^{\star\star} \equiv \delta_H^{(1)}$ and $\tilde{S}_H \equiv S_H^{(1)}$, where

$$\delta_H^{(1)}(\varepsilon) := H\left(\frac{1+\varepsilon}{2}\right) - H\left(\frac{1}{2}\right) \quad \text{and} \quad S_H^{(1)}(\varepsilon) := \frac{\varepsilon H'(\frac{1+\varepsilon}{2})}{2\{H(\frac{1+\varepsilon}{2}) - H(\frac{1}{2})\}}. \tag{13}$$

The explicit forms are obtained as follows:

$$\delta_H^{\star\star}(\varepsilon) = \left(\frac{1+\varepsilon}{2}\right)\ln\left(\frac{1+\varepsilon}{2}\right) + \left(\frac{1-\varepsilon}{2}\right)\ln\left(\frac{1-\varepsilon}{2}\right) + \ln 2 \;\text{ and }\; \tilde{S}_H(\varepsilon) = \frac{\varepsilon\ln\left(\frac{1+\varepsilon}{1-\varepsilon}\right)}{2\delta_H^{\star\star}(\varepsilon)}.$$

Note that Reid and Williamson (2011, Corollary 27) showed the 0-1 regret bound

$$\psi_{\mathrm{RW}}(R_{01}(\eta,\hat{\eta})) \le R_\ell(\eta,\hat{\eta}) \quad \text{where} \quad \psi_{\mathrm{RW}}(\varepsilon) = H\left(\frac{1}{2}+\varepsilon\right) - H\left(\frac{1}{2}\right), \tag{14}$$

for symmetric entropy H, where ψ_{RW} is the same as $\delta_H^{(1)}$ up to scale. Despite $\delta_H^{\star\star} \equiv \delta_H^{(1)}$ when H is the Shannon entropy, $\delta_H^{\star\star}(\varepsilon) \le \delta_H^{(1)}(\varepsilon)$ in general and hence the 0-1 regret bound with ψ_{RW} is

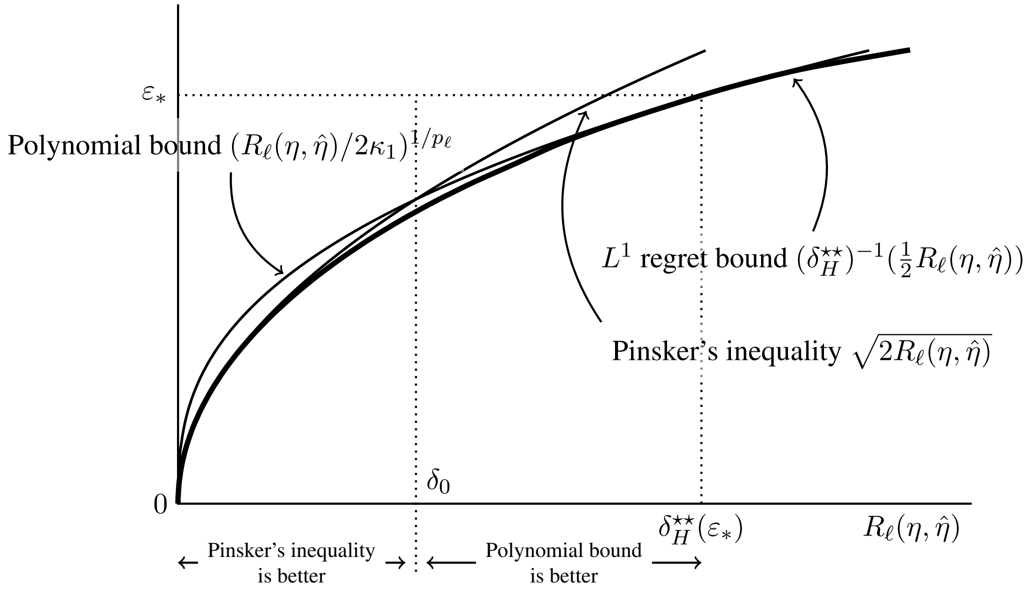

Figure 3: For the L^1 regret bound $(\delta_H^{\star\star})^{-1}(\frac{1}{2}R_\ell(\eta,\hat{\eta}))$ (Theorem 6, shown by the bold line) with the log loss ℓ, we show the two upper bounds: the polynomial bound $(R_\ell(\eta,\hat{\eta})/2\kappa_1)^{1/p_\ell}$ (Theorem 10) and Pinsker's inequality $\sqrt{2R_\ell(\eta,\hat{\eta})}$. δ_0 is the intersecting point of the two bounds. When $0 < R_\ell(\eta,\hat{\eta}) < \delta_0$, Pinsker's inequality is smaller, and when $\delta_0 < R_\ell(\eta,\hat{\eta}) \le \delta_H^{\star\star}(\varepsilon_*)$, the polynomial bound is smaller. Although Pinsker's inequality is better around the origin, it becomes loose around $(\delta_H^{\star\star}(\varepsilon_*), \varepsilon_*)$. Note that the polynomial bound does not necessarily upper bound $(\delta_H^{\star\star})^{-1}$ when $R_\ell(\eta,\hat{\eta}) > \delta_H^{\star\star}(\varepsilon_*)$ for general entropy H (while it still does for the negative Shannon entropy H).

occasionally better than our Corollary 7 because they directly bound the 0-1 regret while we bound the L^1 regret as an intermediate milestone. This bypass is necessary to treat the 0-1 and ranking regret bounds in a unified way.

From Theorems 6 and 10, we obtain $|\eta - \hat{\eta}| \le (\delta_H^{\star\star})^{-1}(\frac{1}{2}R_\ell(\eta,\hat{\eta})) \le \{R_\ell(\eta,\hat{\eta})/2\kappa_1\}^{1/p_\ell}$. Remark that the rightmost $\{R_\ell(\eta,\hat{\eta})/2\kappa_1\}^{1/p_\ell}$ is slightly worse than the rate of Pinsker's inequality $|\eta - \hat{\eta}| \le \sqrt{2R_\ell(\eta,\hat{\eta})}$ (and the regret bound via strongly proper losses (Agarwal, 2014)) around the origin, attributed to $p_\ell > 2$ (see Figure 4). This is because the polynomial rate bound (12) pays an extra price to tightly bound the entire range $R_\ell(\eta,\hat{\eta}) \in [0, \delta_H^{\star\star}(\varepsilon_*)]$. This is important to finely characterize the polynomial rate of $\delta_H^{\star\star}$ around the origin. As $\varepsilon_* \to 0$, the polynomial rate p_ℓ approaches to the rate of Pinsker's inequality asymptotically. See Figure 3 for the comparison.

Boosting/exponential loss. Consider $H(\eta) = -2\sqrt{\eta(1-\eta)}$, which is the associated Bayes risk with the *exponential loss* (Buja et al., 2005; Agarwal, 2014), being used with AdaBoost. Then,

$$\delta_H^{\star\star}(\varepsilon) = \delta_H^{(1)}(\varepsilon) = 1 - \sqrt{1-\varepsilon^2} \quad \text{and} \quad \tilde{S}_H(\varepsilon) = S_H^{(1)}(\varepsilon) = \frac{\varepsilon^2}{\sqrt{1-\varepsilon^2}(1-\sqrt{1-\varepsilon^2})}.$$

We refer to this H as the exponential entropy for convenience.

Norm entropies. Consider $H(\eta) = \{\eta^\alpha + (1-\eta)^\alpha\}^{1/\alpha} - 1$ for $\alpha \ge 2$, which is the α-norm of the probability vector $[\eta; 1-\eta]$. It is also known as the pseudo-spherical entropy, with the *pseudo-spherical loss* being associated (Gneiting and Raftery, 2007). When $\alpha = 2$, this recovers the spherical loss (Agarwal, 2014). In this case, $\delta_H^{\star\star} \equiv \delta_H^{(2)}$ and $\tilde{S}_H \equiv S_H^{(2)}$, where

$$\delta_H^{(2)}(\varepsilon) := \frac{H(0)+H(\varepsilon)}{2} - H\left(\frac{\varepsilon}{2}\right) \quad \text{and} \quad S_H^{(2)}(\varepsilon) := \frac{\varepsilon\{H'(\varepsilon)-H'(\frac{\varepsilon}{2})\}}{H(0)+H(\varepsilon)-2H(\frac{\varepsilon}{2})}. \tag{15}$$

The modulus in Eq. (15) is evidence that the L^1 regret bound is not necessarily $\delta_H^{(1)}$ in general unlike the 0-1 regret bound of Reid and Williamson (2011) mentioned above. We refer to the 2-norm entropy as the spherical entropy for convenience. Note that the norm entropies with $\alpha > 2$ induces strictly but not strongly proper losses (cf. Agarwal (2014, Theorem 10)).

Squared norm entropies. Consider $H(\eta) = \frac{1}{2}\{\eta^\alpha + (1-\eta)^\alpha\}^{2/\alpha} - \frac{1}{2}$ for $\alpha > 1$, which is the squared α-norm of the probability vector $[\eta; 1-\eta]$. The associated modulus and Simonenko order function entail two modes depending on the value of α:

$$\delta_H^{\star\star}(\varepsilon) = \begin{cases} \delta_H^{(2)}(\varepsilon) & \text{if } \alpha \ge 2 \\ \delta_H^{(1)}(\varepsilon) & \text{if } 1 < \alpha \le 2 \end{cases} \quad \text{and} \quad \tilde{S}_H(\varepsilon) = \begin{cases} S_H^{(2)}(\varepsilon) & \text{if } \alpha \ge 2 \\ S_H^{(1)}(\varepsilon) & \text{if } 1 < \alpha \le 2 \end{cases}.$$

When $\alpha = 2$, this recovers the *Gini index* and *squared loss* (Brier score) as the generalized entropy and proper loss, respectively (Buja et al., 2005), where $\delta_H^{\star\star}(\varepsilon) = \frac{\varepsilon^2}{4}$ and $\tilde{S}_H(\varepsilon) = 2$.

Polynomial entropies. To interpolate between the (negative) entropy corresponding to the 0-1 (and hinge) loss $-\min\{\eta, 1-\eta\}$ (Buja et al., 2005; Masnadi-Shirazi and Vasconcelos, 2009) and the Gini index, we consider $H(\eta) = \left|\eta - \frac{1}{2}\right|^\alpha - \frac{1}{2^\alpha}$ for $\alpha > 1$. The two entropies are interpolated with $\alpha = 1$ (0-1) and $\alpha = 2$ (Gini). In this case,

$$\delta_H^{\star\star}(\varepsilon) = \begin{cases} \delta_H^{(1)}(\varepsilon) = \left(\frac{\varepsilon}{2}\right)^\alpha & \text{if } \alpha > 2 \\ \delta_H^{(2)}(\varepsilon) & \text{if } 1 < \alpha \le 2 \end{cases} \quad \text{and} \quad \tilde{S}_H(\varepsilon) = \begin{cases} S_H^{(1)}(\varepsilon) = \alpha & \text{if } \alpha > 2 \\ S_H^{(2)}(\varepsilon) & \text{if } 1 < \alpha \le 2 \end{cases}.$$

When $\alpha = 1$, Theorem 10 is no longer applicable because the polynomial 1-entropy is not strictly convex, The corresponding modulus is $\delta_H^{\star\star}(\varepsilon) = [\varepsilon - \frac{1}{2}]_+$, for which the only valid polynomial lower bound is $0 = \kappa\varepsilon^\infty$ (for any $\kappa > 0$), so we informally interpret the order as ∞ in this case. Note that the polynomial entropies with $\alpha > 2$ induces strictly but not strongly proper losses (cf. Agarwal (2014, Theorem 10)).

5. Discussion

Hierarchy of proper losses. The polynomial rate evaluation in Theorem 10 provides us a quantitative comparison of the L^1 regret bounds. Specifically, for a fixed ε_*, two proper losses ℓ_1 and ℓ_2 can be compared as follows: if $p_{\ell_1} \le p_{\ell_2}$, ℓ_1 admits the faster L^1 regret upper rate than that of ℓ_2. This comparison eventually constitutes a *loss hierarchy* of proper losses. See Figure 4 (for $\varepsilon_* = 0.5$). The Shannon entropy (associated with the log loss), exponential loss, and squared α-norm entropies admit the L^1 regret lower rates $\Omega(\varepsilon^{1/2})$ ($q_\ell = 2$), while the upper rates $O(\varepsilon^{1/p_\ell})$ differ among loss functions and the squared 2-norm entropy has the tight upper rate $O(\varepsilon^{1/2})$. Hence, we can argue that the squared 2-norm entropy entails the fastest regret rate, followed by the Shannon entropy and exponential loss.

Nonetheless, please be aware that the final learning performance is not solely dominated by surrogate regret—indeed, the sample complexity and optimization governs as well. We believe that surrogate regret bounds provided in this work may cooperate with standard sample complexity and optimization analyses to characterizes the overall learning performance.

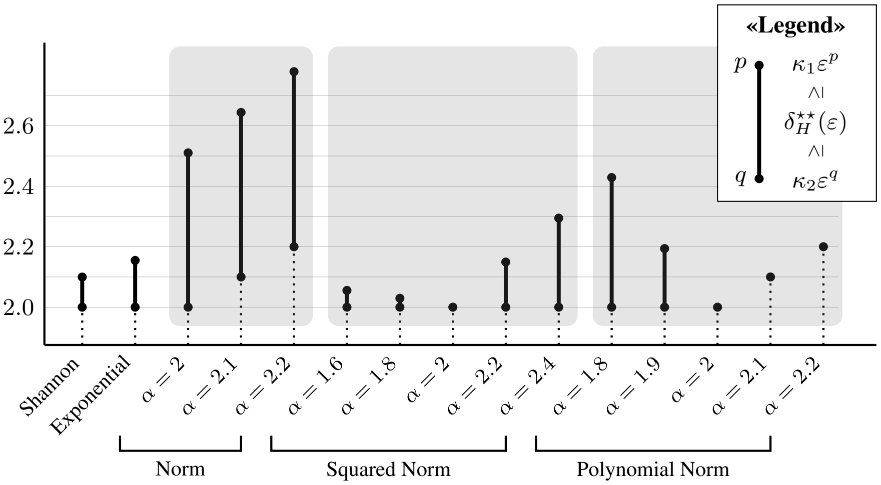

Figure 4: Rate comparison of different losses (associated with the corresponding entropies) for $\varepsilon_* = 0.5$. For each loss, the upper and lower ends indicate the orders p and q, respectively. Each interval illustrates $\mathcal{I}_\ell$, corresponding to the rate evaluation $\kappa_1\varepsilon^p \le \delta_H^{\star\star}(\varepsilon) \le \kappa_2\varepsilon^q$ for $\varepsilon \in [0, \varepsilon_*]$ (Theorem 10). The smaller p and q are, the polynomial rate of the modulus $\delta_H^{\star\star}$ is better. In addition, the polynomial rate evaluation becomes nearly tight as p and q become closer.

Pitfall of polyhedral losses. Frongillo and Waggoner (2021) claimed that polyhedral losses (including the hinge loss) can be good alternatives to smooth losses thanks to the linear 0-1 regret rate. Despite its merit, the hinge loss may not be a sensible choice when the L^1 regret is to be bounded.

The (negative) Bayes risk associated with the hinge loss is $H(\eta) = -\min\{\eta, 1-\eta\}$ (Masnadi-Shirazi and Vasconcelos, 2009), for which the modulus of convexity is $\delta_H^{\star\star}(\varepsilon) = [\varepsilon - \frac{1}{2}]_+$. Consequently, for $\eta, \hat{\eta} \in \triangle$ such that $|\eta - \hat{\eta}| \le \frac{1}{2}$, we cannot obtain the L^1 regret bound via Theorem 6.[5] Its tightness asserts that the hinge loss has trouble estimating η in the sense of the L^1 regret. Although this does not contradict the linear 0-1 regret rate, we conjecture that the ranking regret of the hinge loss is vacuous in light of the relationship between the ranking regret and the L^1 regret (Narasimhan and Agarwal, 2013). This discussion applies to any non-strictly proper losses because the biconjugated modulus $\delta_H^{\star\star}$ of a non-strictly convex entropy H entails a positive ε_0 with $\delta_H^{\star\star}(\varepsilon_0) = 0$ (Proposition 5).

Limitation of this work. We are mainly concerned with the L^1 regret because of its versatility to relate proper losses to many other tasks. However, the L^1 regret bounds may be suboptimal when some specific tasks such as binary classification are of our interest.

As an example, recap the polynomial entropy $H(\eta) = |\eta - \frac{1}{2}|^\alpha - \frac{1}{2^\alpha}$. When $1 < \alpha \le 2$, the modulus $\delta_H^{\star\star}$ changes the mode to $\delta_H^{(2)}$, for which a quadratic lower bound $(\delta_H^{(2)})^{-1}(\varepsilon) = \Omega(\varepsilon^{1/2})$ always exists (see Appendix B.5). By contrast, the direct 0-1 regret rate ψ_{RW} by Reid and Williamson (2011) is characterized by the mode $\delta_H^{(1)}$ (see Eq. (14)), which is α-polynomial $\delta_H^{(1)}(\varepsilon) = (\varepsilon/2)^\alpha$. Hence, the 0-1 regret rate $(\delta_H^{(1)})^{-1}(\varepsilon) = \Theta(\varepsilon^{1/\alpha})$ is faster than the L^1 regret rate $(\delta_H^{(2)})^{-1}(\varepsilon) = \Omega(\varepsilon^{1/2})$ when $1 < \alpha < 2$, which means that the moduli may fail to capture the exact 0-1 regret rate in some cases.

[5] By taking the inverse of $\delta_H^{\star\star}$, the L^1 regret bound informally reads $|\eta - \hat{\eta}| \le \frac{1}{2}R_\ell(\eta, \hat{\eta}) + \frac{1}{2}$, which is vacuous.

The L^1 regret rate is no faster than $\Omega(\varepsilon^{1/2})$ with many entropies shown in Section 4. Consequently, the 0-1 regret bound obtained via the L^1 regret bound cannot escape from sacrificing the optimality in this case.

Proposition 11 *For H being any of the Shannon entropy, exponential entropy, spherical entropy, squared α-norm entropies with $\alpha > 1$, and α-polynomial entropies with $\alpha > 1$, $\lim_{\varepsilon \to 0+} \tilde{S}_H(\varepsilon) \geq 2$.*

Refer to each derivation in Appendix B for the proof. Proposition 11 implies that the L^1 regret rate of the corresponding ℓ is $(\delta_H^{\star\star})^{-1}(\varepsilon) = \Omega(\varepsilon^{1/2})$. This lower bound is achieved by the squared 2-norm entropy. We can see this from Figure 4, where the upper order p_ℓ of the squared 2-norm entropy is $p_\ell = 2$, matching the lower order $q_\ell = 2$. By contrast, other entropies such as the Shannon entropy induce the suboptimal regret upper bounds, as long as $\varepsilon_* > 0$. We conjecture that Proposition 11 holds under a broader class of generalized entropies (including the α-norm entropies with $\alpha > 2$), and leave it for future work.

Acknowledgment

We thank an anonymous reviewer for an idea to improve the constant factor in Theorem 6.

References

Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *Proceedings of the 28th Conference on Learning Theory*, pages 4–22, 2015.

Arpit Agarwal, Harikrishna Narasimhan, Shivaram Kalyanakrishnan, and Shivani Agarwal. GEV-canonical regression for accurate binary class probability estimation when one class is rare. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1989–1997, 2014.

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(1):1653–1674, 2014.

Han Bao and Masashi Sugiyama. Fenchel–Young losses with skewed entropies for class-posterior probability estimation. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 1648–1656, 2021.

Han Bao, Clay Scott, and Masashi Sugiyama. Calibrated surrogate losses for adversarially robust classification. In *Proceedings of the 30th Conference on Learning Theory*, pages 408–451. PMLR, 2020.

Robert Gardner Bartle. *A Modern Theory of Integration*, volume 32. American Mathematical Society, 2001.

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

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

Jonathan Borwein, A Guirao, Petr Hájek, and Jon Vanderwerff. Uniformly convex functions on Banach spaces. *Proceedings of the American Mathematical Society*, 137(3):1081–1091, 2009.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Technical Report*, 2005.

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1507–1517. PMLR, 2021.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Tadeusz Figiel. On the moduli of convexity and smoothness. *Studia Mathematica*, 56(2):121–155, 1976.

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. In *Advances in Neural Information Processing Systems 33*, volume 32, pages 10781–10791, 2019.

Alberto Fiorenza and Miroslav Krbec. Indices of Orlicz spaces and some applications. *Commentationes Mathematicae Universitatis Carolinae*, 38(3):433–451, 1997.

Rafael Frongillo and Bo Waggoner. Surrogate regret bounds for polyhedral losses. In *Advances in Neural Information Processing Systems 34*, 2021.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Olof Hanner. On the uniform convexity of L^p and l^p. *Arkiv för Matematik*, 3(3):239–244, 1956.

Henryk Hudzik. Lower and upper estimations of the modulus of convexity in some Orlicz spaces. *Archiv der Mathematik*, 57(1):80–87, 1991.

Kazuhiro Ishige, Paolo Salani, and Asuka Takatsu. Hierarchy of deformations in concavity. *Information Geometry*, pages 1–19, 2022.

Wojciech Kotłowski and Krzysztof Dembczyński. Surrogate regret bounds for generalized classification performance metrics. In *Proceedings of the 8th Asian Conference on Machine Learning*, pages 301–316, 2016.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and SavageBoost. In *Advances in Neural Information Processing Systems 22*, pages 1049–1056, 2009.

Shahar Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceeddings of the 30th International Conference on Machine Learning*, pages 603–611. PMLR, 2013.

Aditya Krishna Menon and Robert C Williamson. Bayes-optimal scorers for bipartite ranking. In *Proceedings of the 27th Conference on Learning Theory*, pages 68–106. PMLR, 2014.

Alexander Mey and Marco Loog. Consistency and finite sample behavior of binary class probability estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8967–8974, 2021.

Harikrishna Narasimhan and Shivani Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Advances in Neural Information Processing Systems 26*, pages 2913–2921, 2013.

Frank Nielsen and Sylvain Boltz. The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466, 2011.

Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems 31*, pages 302–313, 2017.

Mark D Reid and Robert C Williamson. Generalised Pinsker inequalities. In *Proceedings of the 22nd International Conference on Machine Learning*, 2009a.

Mark D Reid and Robert C Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the 26th International Conference on Machine Learning*, pages 897–904, 2009b.

Mark D Reid and Robert C Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Mark D Reid and Robert C Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011.

R Tyrrell Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1970.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Mark J Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989.

Emir H Shuford, Arthur Albert, and H Edward Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966.

Igor Borisovich Simonenko. Interpolation and extrapolation of linear operators in Orlicz spaces. *Matematicheskii Sbornik*, 105(4):536–553, 1964.

Benjamin Sprung. Upper and lower bounds for the Bregman divergence. *Journal of Inequalities and Applications*, 2019(1):1–12, 2019.

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

Mingyuan Zhang, Jane Lee, and Shivani Agarwal. Learning from noisy labels with no change to the training process. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12468–12478. PMLR, 2021.

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

Appendix A. Proofs

A.1. Proof of Proposition 5

Proof If f is strictly convex, then the strict inequality $\frac{f(\eta)+f(\eta')}{2} > f(\frac{\eta+\eta'}{2})$ holds for all η, η' with $\eta \neq \eta'$. Fix $\varepsilon \in (0, 1]$, then the extreme value theorem ensures the existence of minimizers η_*, η'_* with $|\eta_* - \eta'_*| \geq \varepsilon$ such that $\frac{f(\eta)+f(\eta')}{2} - f(\frac{\eta+\eta'}{2}) \geq \frac{f(\eta_*)+f(\eta'_*)}{2} - f(\frac{\eta_*+\eta'_*}{2}) > 0$ for all η, η' with $|\eta - \eta'| \geq \varepsilon$. Note that the extreme value theorem can be invoked because f is lower semicontinuous (and so is the objective to be minimized in δ_f) and the domain $\{\eta, \eta' \in \triangle \mid |\eta - \eta'| \geq \varepsilon\}$ is compact. Thus, $\delta_f(\varepsilon) > 0$.

Conversely, $\delta_f(\varepsilon) > 0$ for all $\varepsilon \in (0, 1]$ implies that $\frac{f(\eta)+f(\eta')}{2} - f(\frac{\eta+\eta'}{2}) > 0$ for all η, η' with $\eta \neq \eta'$. For $t = \frac{1}{2}$, it is trivial to see $f(t\eta + (1-t)\eta') < tf(\eta) + (1-t)f(\eta')$. For $t \in (\frac{1}{2}, 1)$, we can assume $\eta < \eta'$ and hence $t\eta + (1-t)\eta' < \frac{\eta+\eta'}{2}$ without loss of generality. Then, for $\tilde{t} := 2t - 1 \in (0, 1)$,

$$\begin{aligned} f(t\eta + (1-t)\eta') &= f\left(\tilde{t}\eta + (1-\tilde{t})\frac{\eta+\eta'}{2}\right) \\ &\leq \tilde{t}f(\eta) + (1-\tilde{t})f\left(\frac{\eta+\eta'}{2}\right) \\ &< \tilde{t}f(\eta) + (1-\tilde{t})\frac{f(\eta)+f(\eta')}{2} \\ &= tf(\eta) + (1-t)f(\eta'), \end{aligned}$$

which implies the strict convexity of f. For $t \in (0, \frac{1}{2})$, we can show the strict inequality similarly. Hence, $\delta_f(\varepsilon) > 0$ implies the strict convexity of f.

If f is strongly convex with parameter $\tilde{\kappa} > 0$, the inequality $\frac{f(\eta)+f(\eta')}{2} - f(\frac{\eta+\eta'}{2}) \geq \frac{\tilde{\kappa}}{8}|\eta - \eta'|^2$ holds for all η, η'. Given $\varepsilon \in [0, 1]$, we can show the existence of minimizers η_*, η'_* with $|\eta_* - \eta'_*| \geq \varepsilon$ such that $\delta_f(\varepsilon) \geq \frac{f(\eta_*)+f(\eta'_*)}{2} - f(\frac{\eta_*+\eta'_*}{2}) \geq \frac{\tilde{\kappa}}{8}\varepsilon^2$ by using the same argument as in the strict convexity case.

Conversely, suppose that $\delta_f(\varepsilon) \geq \kappa\varepsilon^2$ for all $\varepsilon \in [0,1]$. As $\delta_f(\varepsilon)$ is nondecreasing in ε, the modulus of convexity indeed satisfies $\delta_f(\varepsilon) = \inf\left\{\frac{f(\eta)+f(\eta')}{2} - f(\frac{\eta+\eta'}{2}) \,\middle|\, |\eta-\eta'| = \varepsilon\right\}$. Then, the quadratic lower bound of $\delta_f(\varepsilon)$ implies that $\frac{f(\eta)+f(\eta')}{2} - f(\frac{\eta+\eta'}{2}) \geq \kappa|\eta-\eta'|^2$ for all η, η'. Using this inequality and the same argument as in the strict convexity case, we can show

$$tf(\eta) + (1-t)f(\eta') - f(t\eta + (1-t)\eta') \geq \frac{1}{2}\tilde{\kappa}t(1-t)|\eta-\eta'|^2$$

with some $\tilde{\kappa} > 0$ for all η, η', and $t \in [0,1]$. Hence, $\delta_f(\varepsilon) > 0$ implies the strong convexity of f. ■

A.2. Proof of Corollary 7

Proof By beginning from Eq. (6),

$$R_{01}(\eta,\hat{\eta}) = \left|\eta - \tfrac{1}{2}\right| \left[\!\left[\min\{\eta,\hat{\eta}\} \leq \tfrac{1}{2} < \max\{\eta,\hat{\eta}\}\right]\!\right] \leq |\eta - \hat{\eta}|,$$

where the inequality can be seen as follows: if $\eta > \hat{\eta}$,

$$\begin{aligned}\min\{\eta,\hat{\eta}\} \leq \tfrac{1}{2} < \max\{\eta,\hat{\eta}\} &\implies \hat{\eta} \leq \tfrac{1}{2} < \eta \\ &\implies \left|\eta - \tfrac{1}{2}\right| \leq |\eta - \hat{\eta}|,\end{aligned}$$

and it is proven similarly otherwise. Thus, by invoking Theorem 6,

$$\begin{aligned}\mathrm{Reg}_{01}[\eta,\hat{\eta}] &= \mathop{\mathbb{E}}_{X\sim\mu} R_{01}(\eta(X),\hat{\eta}(X)) \\ &\leq \mathop{\mathbb{E}}_{X\sim\mu}\Big[|\eta(X) - \hat{\eta}(X)|\Big] \\ &\leq \mathop{\mathbb{E}}_{X\sim\mu}\Big[(\delta^{\star\star}_{-\underline{L}_\ell})^{-1}\big(\tfrac{1}{2}R_\ell(\eta(X),\hat{\eta}(X))\big)\Big] \\ &\leq (\delta^{\star\star}_{-\underline{L}_\ell})^{-1}\Big(\tfrac{1}{2}\mathop{\mathbb{E}}_{X\sim\mu} R_\ell(\eta(X),\hat{\eta}(X))\Big) \\ &= (\delta^{\star\star}_{-\underline{L}_\ell})^{-1}\big(\tfrac{1}{2}\mathrm{Reg}_\ell[\eta,\hat{\eta}]\big),\end{aligned}$$

where the last inequality follows from Jensen's inequality. Note that the invertible convex function $\delta^{\star\star}_{-\underline{L}_\ell}$ always has the concave inverse $(\delta^{\star\star}_{-\underline{L}_\ell})^{-1}$. ■

A.3. Proof of Corollary 8

Proof The proof mostly follows Agarwal (2014, Corollary 12), but is shown subsequently for the sake of completeness. By beginning from Eq. (7),

$$\begin{aligned}R_{\mathrm{rank}}(\eta,\eta',\hat{\eta},\hat{\eta}') &= |\eta-\eta'|\left\{[\![(\hat{\eta}-\hat{\eta}')(\eta-\eta') < 0]\!] + \frac{1}{2}[\![\hat{\eta} = \hat{\eta}']\!]\right\} \\ &\leq |\eta-\eta'|[\![(\hat{\eta}-\hat{\eta}')(\eta-\eta') \leq 0]\!] \\ &\leq |\hat{\eta}-\eta| + |\hat{\eta}'-\eta'|,\end{aligned}$$

where the last inequality can be seen as follows: if $\eta > \eta'$,

$$\begin{aligned}(\hat{\eta} - \hat{\eta}')(\eta - \eta') \leq 0 &\implies \hat{\eta} \leq \hat{\eta}' \\ &\implies \eta - \eta' \leq (\eta - \hat{\eta}) + (\hat{\eta}' - \eta') \\ &\implies |\eta - \eta'| \leq |\eta - \hat{\eta}| + |\hat{\eta}' - \eta'|,\end{aligned}$$

and it is proven similarly otherwise. Thus, by invoking Theorem 6,

$$\begin{aligned}\mathrm{Reg}_{\mathrm{rank}}[\eta, \hat{\eta}] &= \frac{1}{2\pi(1-\pi)} \mathop{\mathbb{E}}_{X,X' \sim \mu^2} R_{\mathrm{rank}}(\eta(X), \eta(X'), \hat{\eta}(X), \hat{\eta}(X')) \\ &\leq \frac{1}{2\pi(1-\pi)} \mathop{\mathbb{E}}_{X,X' \sim \mu^2} \left[|\hat{\eta}(X) - \eta(X)| + |\hat{\eta}(X') - \eta(X')| \right] \\ &= \frac{1}{\pi(1-\pi)} \mathop{\mathbb{E}}_{X \sim \mu} \left[|\hat{\eta}(X) - \eta(X)| \right] \\ &\leq \frac{1}{\pi(1-\pi)} \mathop{\mathbb{E}}_{X \sim \mu} \left[(\delta^{\star\star}_{-\underline{L}_\ell})^{-1} \big(\tfrac{1}{2} R_\ell(\eta(X), \hat{\eta}(X)) \big) \right] \\ &\leq \frac{1}{\pi(1-\pi)} (\delta^{\star\star}_{-\underline{L}_\ell})^{-1} \Big(\tfrac{1}{2} \mathop{\mathbb{E}}_{X \sim \mu} R_\ell(\eta(X), \hat{\eta}(X)) \Big) \\ &= \frac{1}{\pi(1-\pi)} (\delta^{\star\star}_{-\underline{L}_\ell})^{-1} \big(\tfrac{1}{2} \mathrm{Reg}_\ell[\eta, \hat{\eta}] \big),\end{aligned}$$

where the last inequality follows from Jensen's inequality. Note that the invertible convex function $\delta^{\star\star}_{-\underline{L}_\ell}$ always has the concave inverse $(\delta^{\star\star}_{-\underline{L}_\ell})^{-1}$. ■

Appendix B. Additional derivations for examples

B.1. Shannon entropy

Here, we discuss the Shannon entropy $H(\eta) = \eta \ln \eta + (1-\eta) \ln(1-\eta)$. The modulus is computed as follows:

$$\delta_H(\varepsilon) = \inf_{\eta \in [\frac{\varepsilon}{2}, 1 - \frac{\varepsilon}{2}]} \left\{ \frac{H\left(\eta - \frac{\varepsilon}{2}\right) + H\left(\eta + \frac{\varepsilon}{2}\right)}{2} - H(\eta) \right\},$$

where we write the objective as $G(\eta)$. The first and second derivatives of G are computed as follows:

$$\begin{aligned}G'(\eta) &= \frac{1}{2} \ln\left(\frac{\eta - \frac{\varepsilon}{2}}{1 + \frac{\varepsilon}{2} - \eta} \right) + \frac{1}{2} \ln\left(\frac{\eta + \frac{\varepsilon}{2}}{1 - \frac{\varepsilon}{2} - \eta} \right) - \ln\left(\frac{\eta}{1 - \eta} \right), \\ G''(\eta) &= \underbrace{\frac{1}{2(\eta - \frac{\varepsilon}{2})} + \frac{1}{2(\eta + \frac{\varepsilon}{2})} - \frac{1}{\eta}}_{\geq 0} + \underbrace{\frac{1}{2(1 + \frac{\varepsilon}{2} - \eta)} + \frac{1}{2(1 - \frac{\varepsilon}{2} - \eta)} - \frac{1}{1 - \eta}}_{\geq 0} \\ &\geq 0,\end{aligned}$$

where the inequalities can be seen from Jensen's inequality. Hence, G is convex and $\eta = \frac{1}{2}$ is the minimizer, so the modulus and order function are computed as follows:

$$\begin{aligned}
\delta_H(\varepsilon) &= G\left(\tfrac{1}{2}\right) = \delta_H^{(1)}(\varepsilon) = \left(\frac{1+\varepsilon}{2}\right) \ln\left(\frac{1+\varepsilon}{2}\right) + \left(\frac{1-\varepsilon}{2}\right) \ln\left(\frac{1-\varepsilon}{2}\right) + \ln 2,\\
\tilde{S}_H(\varepsilon) &= \frac{\varepsilon \ln\left(\frac{1+\varepsilon}{1-\varepsilon}\right)}{(1+\varepsilon)\ln(\frac{1+\varepsilon}{2}) + (1-\varepsilon)\ln(\frac{1-\varepsilon}{2}) + 2\ln 2}.
\end{aligned}$$

Further, $\delta_H^{\star\star} \equiv \delta_H$ is witnessed. To compute the limit $\varepsilon \to 0+$, we invoke l'Hôpital's rule twice:

$$\begin{aligned}
\lim_{\varepsilon\to 0+} \tilde{S}_H(\varepsilon) &= \lim_{\varepsilon\to 0+} \frac{\ln(1+\varepsilon) - \ln(1-\varepsilon) + \frac{\varepsilon}{1+\varepsilon} + \frac{\varepsilon}{1-\varepsilon}}{\ln(\frac{1+\varepsilon}{2}) - \ln(\frac{1-\varepsilon}{2})}\\
&= \lim_{\varepsilon\to 0+} \frac{\frac{1}{1+\varepsilon} + \frac{1}{1-\varepsilon} + \frac{1}{(1+\varepsilon)^2} + \frac{1}{(1-\varepsilon)^2}}{\frac{1}{1+\varepsilon} + \frac{1}{1-\varepsilon}}\\
&= 2.
\end{aligned}$$

B.2. Exponential entropy

Here, we discuss the exponential entropy $H(\eta) = -2\sqrt{\eta(1-\eta)}$. The modulus is computed as follows:

$$\delta_H(\varepsilon) = \inf_{\eta\in[\frac{\varepsilon}{2}, 1-\frac{\varepsilon}{2}]} \left\{ \frac{H\left(\eta - \frac{\varepsilon}{2}\right) + H\left(\eta + \frac{\varepsilon}{2}\right)}{2} - H(\eta) \right\},$$

where we write the objective as $G(\eta)$. The first and second derivatives of G are computed as follows:

$$\begin{aligned}
G'(\eta) &= \frac{\eta - \frac{1+\varepsilon}{2}}{2\sqrt{(\eta - \frac{\varepsilon}{2})(1 + \frac{\varepsilon}{2} - \eta)}} + \frac{\eta - \frac{1-\varepsilon}{2}}{2\sqrt{(\eta + \frac{\varepsilon}{2})(1 - \frac{\varepsilon}{2} - \eta)}} - \frac{\eta - \frac{1}{2}}{\sqrt{\eta(1-\eta)}},\\
G''(\eta) &= \frac{1}{8((\eta - \frac{\varepsilon}{2})(1 + \frac{\varepsilon}{2} - \eta))^{3/2}} + \frac{1}{8((\eta + \frac{\varepsilon}{2})(1 - \frac{\varepsilon}{2} - \eta))^{3/2}} - \frac{1}{4(\eta(1-\eta))^{3/2}} \geq 0,
\end{aligned}$$

where the last inequality is attributed to Jensen's inequality. Hence, G is convex and $\eta = \frac{1}{2}$ is the minimizer, so the modulus and order function are computed as follows:

$$\begin{aligned}
\delta_H(\varepsilon) &= G(\tfrac{1}{2}) = \delta_H^{(1)}(\varepsilon) = 1 - \sqrt{1-\varepsilon^2},\\
\tilde{S}_H(\varepsilon) &= \frac{\varepsilon^2}{\sqrt{1-\varepsilon^2}(1 - \sqrt{1-\varepsilon^2})}.
\end{aligned}$$

Further, $\delta_H^{\star\star} \equiv \delta_H$ is witnessed. To compute the limit $\varepsilon \to 0+$, we invoke l'Hôpital's rule:

$$\lim_{\varepsilon\to 0+} \tilde{S}_H(\varepsilon) = \lim_{\varepsilon\to 0+} \frac{2\varepsilon}{\frac{-\varepsilon}{\sqrt{1-\varepsilon^2}} + 2\varepsilon} = 2.$$

B.3. Norm entropies

Here, we discuss the norm entropies $H(\eta) = \{\eta^\alpha + (1-\eta)^\alpha\}^{1/\alpha} - 1$ for $\alpha \geq 2$. Due to the symmetry of H about $\eta = \frac{1}{2}$, the modulus is computed as follows:

$$\delta_H(\eta) = \inf_{\eta\in[\frac{\varepsilon}{2}, \frac{1}{2}]} \left\{ \frac{H\left(\eta - \frac{\varepsilon}{2}\right) + H\left(\eta + \frac{\varepsilon}{2}\right)}{2} - H(\eta) \right\},$$

where we write the objective as $G(\eta)$. The first derivative of G is computed as follows:

$$G'(\eta) = \frac{\tilde{G}(\eta - \frac{\varepsilon}{2}) + \tilde{G}(\eta + \frac{\varepsilon}{2})}{2} - \tilde{G}(\eta) \geq 0$$

where $\tilde{G}(\eta) := \{\eta^\alpha + (1-\eta)^\alpha\}^{\frac{1}{\alpha}-1}\left\{\eta^{\alpha-1} - (1-\eta)^{\alpha-1}\right\}$, and the inequality holds due to the convexity of $\tilde{G}$ for $\eta \in [\frac{\varepsilon}{2}, \frac{1}{2}]$. Hence, G is nondecreasing and minimized at $\eta = \frac{\varepsilon}{2}$, so the modulus and order function are computed as follows:

$$\begin{aligned} \delta_H(\varepsilon) &= G\left(\frac{\varepsilon}{2}\right) = \delta_H^{(2)}(\varepsilon) = \frac{H(\varepsilon)}{2} - H\left(\frac{\varepsilon}{2}\right), \\ \tilde{S}_H(\varepsilon) &= \frac{\varepsilon\{H'(\varepsilon) - H'(\frac{\varepsilon}{2})\}}{H(\varepsilon) - 2H(\frac{\varepsilon}{2})}. \end{aligned}$$

Further, $\delta_H^{\star\star} \equiv \delta_H$ is witnessed. For the norm entropies, we only compute the limit of the Simonenko order function when $\alpha = 2$. First, we simplify the higher-order derivatives of H:

$$\begin{aligned} H(\eta) &= \|\eta\|_2 - 1, \\ H'(\eta) &= (\|\eta\|_2)' = (2\eta - 1)\|\eta\|_2^{-1}, \\ H''(\eta) &= \|\eta\|_2^{-3}, \\ H'''(\eta) &= 3(1-2\eta)\|\eta\|_2^{-5}, \end{aligned}$$

where $\|\eta\|_2$ is a shorthand for $\sqrt{\eta^2 + (1-\eta)^2}$. By using them, the limit at $\varepsilon \to 0+$ can be computed as follows:

$$\lim_{\varepsilon \to 0+} \tilde{S}_H(\varepsilon) = 2 + \lim_{\varepsilon \to 0+} \frac{\varepsilon\left\{H'''(\varepsilon) - \frac{1}{4}H'''\left(\frac{\varepsilon}{2}\right)\right\}}{H''(\varepsilon) - \frac{1}{2}H''\left(\frac{\varepsilon}{2}\right)} = 2 + \lim_{\varepsilon \to 0+} \frac{\varepsilon\{3(1-2\varepsilon) + \frac{3}{4}(\varepsilon - 1)\}}{\|\varepsilon\|_2^2(1 - \frac{1}{2})} = 2,$$

where l’Hôpital’s rule is invoked twice at the first identity.

B.4. Squared norm entropies

Here, we discuss the squared norm entropies $H(\eta) = \frac{1}{2}\{\eta^\alpha + (1-\eta)^\alpha\}^{2/\alpha} - \frac{1}{2}$ for $\alpha > 1$. Due to the symmetry of H about $\eta = \frac{1}{2}$, the modulus is computed as follows:

$$\delta_H(\eta) = \inf_{\eta \in [\frac{\varepsilon}{2}, \frac{1}{2}]} \left\{ \frac{H\left(\eta - \frac{\varepsilon}{2}\right) + H\left(\eta + \frac{\varepsilon}{2}\right)}{2} - H\left(\eta\right) \right\},$$

where we write the objective as $G(\eta)$. The first derivative of G is computed as follows:

$$G'(\eta) = \frac{\tilde{G}(\eta - \frac{\varepsilon}{2}) + \tilde{G}(\eta + \frac{\varepsilon}{2})}{2} - \tilde{G}(\eta),$$

where $\tilde{G}(\eta) := \{\eta^\alpha + (1-\eta)^\alpha\}^{\frac{2}{\alpha}-1}\{\eta^{\alpha-1} - (1-\eta)^{\alpha-1}\}$.

When $1 < \alpha \leq 2$, $G'(\eta) \leq 0$ holds for $\eta \in [\frac{\varepsilon}{2}, \frac{1}{2}]$ because of the concavity of $\tilde{G}$. This indicates that G is nonincreasing and hence minimized at $\eta = \frac{1}{2}$, so the modulus and order function are computed as follows:

$$\delta_H(\varepsilon) = G\left(\frac{1}{2}\right) = \delta_H^{(1)}(\varepsilon) = H\left(\frac{1+\varepsilon}{2}\right) - H\left(\frac{1}{2}\right),$$

$$\tilde{S}_H(\varepsilon) = \frac{\varepsilon H'(\frac{1+\varepsilon}{2})}{2\{H(\frac{1+\varepsilon}{2}) - H(\frac{1}{2})\}}.$$

Further, $\delta_H^{\star\star} \equiv \delta_H$ is witnessed. To compute the limit at $\varepsilon \to 0+$, we compute the higher-order derivatives of H first:

$$\begin{aligned}
H(\eta) &= \tfrac{1}{2}\|\eta\|_\alpha^2 - \tfrac{1}{2}, \\
H'(\eta) &= \|\eta\|_\alpha\{\eta^{\alpha-1} - (1-\eta)^{\alpha-1}\}, \\
H''(\eta) &= (2-\alpha)\|\eta\|_\alpha^{2(1-\alpha)}\{\eta^{\alpha-1} - (1-\eta)^{\alpha-1}\}^2 + (\alpha-1)\|\eta\|_\alpha^{2-\alpha}\{\eta^{\alpha-2} + (1-\eta)^{\alpha-2}\}, \\
H'''(\eta) &= 2(2-\alpha)(1-\alpha)\|\eta\|_\alpha^{2-3\alpha}\{\eta^{\alpha-1} - (1-\eta)^{\alpha-1}\}^3 \\
&\quad + 3(2-\alpha)(\alpha-1)\|\eta\|_\alpha^{2(1-\alpha)}\{\eta^{\alpha-1} - (1-\eta)^{\alpha-1}\}\{\eta^{\alpha-2} + (1-\eta)^{\alpha-2}\} \\
&\quad + (\alpha-1)(\alpha-2)\|\eta\|_\alpha^{2-\alpha}\{\eta^{\alpha-3} - (1-\eta)^{\alpha-3}\},
\end{aligned}$$

where $\|\eta\|_\alpha$ is a shorthand for $\{\eta^\alpha + (1-\eta)^\alpha\}^{1/\alpha}$. Then, the limit of the order function can be computed as follows:

$$\lim_{\varepsilon \to 0+} \tilde{S}_H(\varepsilon) = 2 + \lim_{\varepsilon \to 0+} \frac{\frac{1}{4}\varepsilon H'''(\frac{1+\varepsilon}{2})}{\frac{1}{2}H''(\frac{1+\varepsilon}{2})} = 2,$$

where l'Hôpital's rule is used twice at the first identity and the second identity is attributed to $H''(\frac{1}{2}) = (\alpha-1)2^{\frac{1+2\alpha-\alpha^2}{\alpha}} > 0$ and $H'''(\frac{1}{2}) = 0$.

When $2 < \alpha$, $G'(\eta) \geq 0$ holds for $\eta \in [\frac{\varepsilon}{2}, \frac{1}{2}]$ because of the convexity of $\tilde{G}$. This indicates that G is nondecreasing and hence minimized at $\eta = \frac{\varepsilon}{2}$, so the modulus and order function are computed as follows:

$$\delta_H(\varepsilon) = G\left(\frac{\varepsilon}{2}\right) = \delta_H^{(2)}(\varepsilon) = \frac{H(\varepsilon)}{2} - H\left(\frac{\varepsilon}{2}\right),$$

$$\tilde{S}_H(\varepsilon) = \frac{\varepsilon\{H'(\varepsilon) - H'(\frac{\varepsilon}{2})\}}{H(\varepsilon) - 2H(\frac{\varepsilon}{2})}.$$

Further, $\delta_H^{\star\star} \equiv \delta_H$ is witnessed. The limit of the order function can be computed as follows:

$$\lim_{\varepsilon \to 0+} \tilde{S}_H(\varepsilon) = 2 + \lim_{\varepsilon \to 0+} \frac{\varepsilon\{H'''(\varepsilon) - \frac{1}{4}H'''(\frac{\varepsilon}{2})\}}{H''(\varepsilon) - \frac{1}{2}H''(\frac{\varepsilon}{2})} = 2,$$

where l'Hôpital's rule is used twice at the first identity and the second identity is attributed to $H''(0) = 1$ and $H'''(0) = 0$.

B.5. Polynomial entropies

Here, we discuss the polynomial entropies $H(\eta) = \left|\eta - \frac{1}{2}\right|^\alpha - \frac{1}{2^\alpha}$ for $\alpha > 1$. When $1 < \alpha \leq 2$, the modulus is computed as follows:

$$\delta_H(\varepsilon) = \inf_{\eta \in [\frac{\varepsilon}{2}, 1-\frac{\varepsilon}{2}]} \left\{ \frac{\left|\eta - \frac{1+\varepsilon}{2}\right|^\alpha + \left|\eta - \frac{1-\varepsilon}{2}\right|^\alpha}{2} - \left|\eta - \frac{1}{2}\right|^\alpha \right\},$$

where we write the objective as $G(\eta)$. As G is symmetric about $\eta = \frac{1}{2}$, we only need to consider $\eta \in [\frac{\varepsilon}{2}, \frac{1}{2}]$. When $\frac{\varepsilon}{2} \leq \eta \leq \frac{1-\varepsilon}{2}$, the right derivative of G is

$$\frac{G'(\eta)}{\alpha} = -\frac{\left(\frac{1+\varepsilon}{2} - \eta\right)^{\alpha-1} + \left(\frac{1-\varepsilon}{2} - \eta\right)^{\alpha-1}}{2} + \left(\frac{1}{2} - \eta\right)^{\alpha-1} \geq 0,$$

where the inequality can be seen from Jensen's inequality being applied to a concave function $\eta \mapsto \left(\frac{1}{2} - \eta\right)^{\alpha-1}$. When $\frac{1-\varepsilon}{2} < \eta \leq \frac{1}{2}$, the right derivative of G can be evaluated as

$$\frac{G'(\eta)}{\alpha} = \frac{-\left(\frac{1+\varepsilon}{2} - \eta\right)^{\alpha-1} + \left(\eta - \frac{1-\varepsilon}{2}\right)^{\alpha-1}}{2} + \left(\frac{1}{2} - \eta\right)^{\alpha-1} =: \tilde{G}(\eta).$$

If $\alpha = 2$, $\tilde{G}(\eta) = 0$. Otherwise, the second derivative of $\tilde{G}$ is evaluated as

$$\begin{aligned}\frac{\tilde{G}''(\eta)}{(\alpha-1)(\alpha-2)} &= \frac{-\left(\frac{1+\varepsilon}{2} - \eta\right)^{\alpha-3} + \left(\eta - \frac{1-\varepsilon}{2}\right)^{\alpha-3}}{2} + \left(\frac{1}{2} - \eta\right)^{\alpha-3} \\ &\geq \frac{-\left(\frac{1+\varepsilon}{2} - \eta\right)^{\alpha-3} + \left(\eta - \frac{1-\varepsilon}{2}\right)^{\alpha-3}}{2} \\ &\geq 0,\end{aligned}$$

from which we see $\tilde{G}''(\eta) \leq 0$ and $\tilde{G}$ is concave in $\eta \in (\frac{1-\varepsilon}{2}, \frac{1}{2}]$. Then,

$$G'(\eta) \geq \min\left\{G'\left(\frac{1-\varepsilon}{2}\right), G'\left(\frac{1}{2}\right)\right\} = 0.$$

Hence, G is nondecreasing for $\frac{\varepsilon}{2} \leq \eta \leq \frac{1}{2}$ and minimized at $\eta = \frac{\varepsilon}{2}$, which implies that $\delta_H(\varepsilon) = G(\varepsilon/2) = \delta_H^{(2)}(\varepsilon)$:

$$\delta_H(\varepsilon) = \frac{|2\varepsilon - 1|^\alpha - 2|\varepsilon - 1|^\alpha + 1}{2^{\alpha+1}}.$$

Further, $\delta_H^{\star\star} \equiv \delta_H$ is witnessed. The Simonenko order function is computed as follows: for $\varepsilon < \frac{1}{2}$,

$$\tilde{S}_H(\varepsilon) = \frac{2\alpha\varepsilon\{(1-\varepsilon)^{\alpha-1} - (1-2\varepsilon)^{\alpha-1}\}}{(1-2\varepsilon)^\alpha - 2(1-\varepsilon)^\alpha + 1}.$$

When $\alpha = 2$, $\tilde{S}_H(\varepsilon) = 2$. Below, we compute the limit at $\varepsilon \to 0+$ for $1 < \alpha < 2$, by invoking l'Hôpital's rule twice:

$$\begin{aligned}&\lim_{\varepsilon \to 0+} \tilde{S}_H(\varepsilon) \\ &= \lim_{\varepsilon \to 0+} \frac{2\alpha\{(1-\varepsilon)^{\alpha-1} - (1-2\varepsilon)^{\alpha-1}\} + 2\alpha\varepsilon \cdot (\alpha-1)\{-(1-\varepsilon)^{\alpha-2} + 2(1-2\varepsilon)^{\alpha-2}\}}{\alpha\{-2(1-2\varepsilon)^{\alpha-1} + 2(1-\varepsilon)^{\alpha-1}\}} \\ &= \lim_{\varepsilon \to 0+} \frac{(1-\varepsilon)^{\alpha-1} - (1-2\varepsilon)^{\alpha-1} + (\alpha-1)\varepsilon\{-(1-\varepsilon)^{\alpha-2} + 2(1-2\varepsilon)^{\alpha-2}\}}{-(1-2\varepsilon)^{\alpha-1} + (1-\varepsilon)^{\alpha-1}} \\ &= \lim_{\varepsilon \to 0+} \frac{1}{2(1-2\varepsilon)^{\alpha-2} - (1-\varepsilon)^{\alpha-2}} \big\{ -(1-\varepsilon)^{\alpha-2} + 2(1-2\varepsilon)^{\alpha-2} \\ &\qquad\qquad + (\alpha-2)\varepsilon \left\{(1-\varepsilon)^{\alpha-3} - 4(1-2\varepsilon)^{\alpha-3}\right\} \\ &\qquad\qquad - (1-\varepsilon)^{\alpha-2} + 2(1-2\varepsilon)^{\alpha-2}\big\} \\ &= 2,\end{aligned}$$

where l'Hôpital's rule is invoked at the first and third identities.

When $\alpha > 2$, we compute the modulus in the same way as the case $1 < \alpha \leq 2$. Again, G is symmetric about $\eta = \frac{1}{2}$, so we only need to consider $\eta \in [\frac{\varepsilon}{2}, \frac{1}{2}]$. We can confirm G is nonincreasing in this range similarly, so the modulus is computed as follows:

$$\delta_H(\varepsilon) = G\left(\frac{1}{2}\right) = \delta_H^{(1)}(\varepsilon) = \left(\frac{\varepsilon}{2}\right)^{\alpha} \quad \text{and} \quad \delta_H^{\star\star} \equiv \delta_H.$$

The Simonenko order function is computed as follows:

$$\tilde{S}_H(\varepsilon) = \frac{\varepsilon \cdot \frac{\alpha}{2}\left(\frac{\varepsilon}{2}\right)^{\alpha-1}}{\left(\frac{\varepsilon}{2}\right)^{\alpha}} = \alpha.$$

Its infinitesimal limit is α as well.