Journal of Machine Learning Research 26 (2025) 1-50 Submitted 7/24; Revised 10/25; Published 12/25

Proper losses regret at least 1/2-order

Han Bao BAO.HAN@ISM.AC.JP
The Institute of Statistical Mathematics

Asuka Takatsu ASUKA-TAKATSU@G.ECC.U-TOKYO.AC.JP
The University of Tokyo

Editor: Mehryar Mohri

Abstract

A fundamental challenge in machine learning is the choice of a loss as it characterizes our learning task, is minimized in the training phase, and serves as an evaluation criterion for estimators. Proper losses are commonly chosen, ensuring minimizers of the full risk match the true probability vector. Estimators induced from a proper loss are widely used to construct forecasters for downstream tasks such as classification and ranking. In this procedure, how does the forecaster based on the obtained estimator perform well under a given downstream task? This question is substantially relevant to the behavior of the p-norm between true probability and estimated vectors when the estimator is updated. In the proper loss framework, the suboptimality of the estimated probability vector from the true probability vector is measured by a surrogate regret. First, we analyze a surrogate regret and show that the *strict* properness of a loss is necessary and sufficient to establish a non-vacuous surrogate regret bound. Second, we tackle an important open question that the order of convergence in p-norm cannot be faster than the 1/2-order of surrogate regrets for a broad class of strictly proper losses. This implies that strongly proper losses asymptotically achieve the optimal convergence rate.

Keywords: loss functions, proper scoring rules, supervised learning, surrogate regret bounds, convex analysis

1. Introduction

Proper losses, also known as proper scoring rules, are measurements of the quality of a probabilistic prediction given a true probability vector [BSS05, GR07, RW10]. Intuitively, we say a loss is *proper* if the target probability vector is its minimizer, and *strictly proper* if the minimizer is unique, which are a basic property for a reasonable loss. Proper losses are prevailing in modern machine learning: for example, the cross-entropy loss popular in deep learning essentially corresponds to the log loss (or logarithmic score), and the Brier score is used for assessing model uncertainties [OFR$^+$19] [GB22]. As such, probabilistic estimators are obtained via proper loss minimization. It is common to post-process a minimizer of a proper loss for downstream tasks, such as classification (by choosing the most likely label), ranking (by giving ranking scores to each label [NA13]), F-measure optimization (by thresholding the estimated probability [KNRD14]), and probability calibration [KSFF17, BGHN23]. Here, we are interested in the predictive performance of post-processed estimators in downstream tasks. Given the true and estimated probability vectors $\mathbf{q}$ and $\widehat{\mathbf{q}}$, respectively, the *surrogate regret* $R(\mathbf{q}, \widehat{\mathbf{q}})$ (introduced in §3) measures the suboptimality of $\widehat{\mathbf{q}}$ from $\mathbf{q}$ in terms of a proper

loss. Can we relate the suboptimality of a forecaster for a downstream task to the surrogate regret?

Surrogate regret bounds relate the surrogate regret to the performance for downstream tasks, and have been derived for binary classification [Zha04, RW09], bipartite ranking [KDH11, Aga14], property elicitation [AA15], F-measure optimization [KD16, ZRA20], and learning with noisy labels [NDRT13, ZLA21], independently. Recently, a unified surrogate regret bound across different downstream tasks has been established [Bao23], where surrogate regret bounds are unified in terms of the 1-norm. This is based on the observation that the suboptimality of the aforementioned downstream tasks can be controlled by the 1-norm. However, the derived bound has been limited to the binary classification case, and it remains unclear when the surrogate regret bound is non-vacuous. A reasonable loss should entail a non-vacuous regret bound, which is crucial to tackling numerous downstream tasks simultaneously. Moreover, an important conjecture that the convergence rate of surrogate regret bounds cannot be faster than the 1/2-order has yet to be solved. This conjecture has a significant role in the choice of losses because the lower bound of the order of convergence contributes to delineating the optimality of a given proper loss.

In this article, we aim to study when the surrogate regret bounds are non-vacuous and how fast the order of convergence in terms of the p-norm. To this end, we analyze the p-norm bounds by the surrogate regret $R(\mathbf{q}, \widehat{\mathbf{q}})$ jointly with a rate function ψ in the following form by extending from the binary classification case [Bao23] to the multiclass classification case:

$$\|\mathbf{q} - \widehat{\mathbf{q}}\|_p \leqslant \psi(R(\mathbf{q}, \widehat{\mathbf{q}})). \tag{1}$$

After formalizing these notions in §3, we derive the surrogate regret bounds in §4. To derive bounds of the form (1), we introduce the *moduli of convexity* [Pol66] [Fig76], which describe the information of its second derivative of convex functions (on the probability simplex). The rate ψ in (1) can be characterized by the modulus of a Bregman generator function associated with a proper loss $\boldsymbol{\ell}$ (Theorem 10). We first show that the strict properness of a loss is necessary and sufficient to obtain a non-vacuous surrogate regret bound, or strictly increasing ψ, to put it differently (Theorem 8). Whereas it has been known that non-strictly proper losses can achieve non-vacuous bounds for classification [RW11, Corollary 27], our sufficiency result argues that the strict properness is a minimal requirement for an estimate to be non-vacuous in terms of the p-norm. As our second main result, we provide an affirmative answer to the above conjecture: the optimal rate $\psi(\rho)$ as $\rho \downarrow 0$ is $O(\rho^{1/2})$, for a broad class of proper losses (Theorem 15). This convergence rate has already been known for a restricted class of proper losses, known as strongly proper losses [Aga14]. Hence, our result ensures the asymptotic optimality of strongly proper losses. This gives an answer to the question, "Do we have an interesting loss that is strictly proper but not strongly proper?" [Bao22, §6.2.9]: there is no better proper loss outside of strongly proper losses, as long as we are concerned with the asymptotic rate of ψ in Eq. (1).

1.1 Organization

The organization of this article and our contributions are summarized as follows.

- §2: Notation and necessary backgrounds on convex analysis are summarized. We tailor subdifferentials for convex functions defined on the probability simplex.

- §3: Proper losses for multiclass classification are introduced. Theorem 1 characterizes the existence of minimizers for general losses, and Theorem 4 gives a self-contained and rigorous proof of the well-known representation of proper losses [Sav71].

- §4: Theorem 8 is our first result, proving that the strict properness of a loss is a necessary and sufficient condition for an associated surrogate regret bound to be non-vacuous. Theorem 10 extends surrogate regret bounds for binary classification [Bao23] to multiclass classification. This is achieved by extending the moduli of convexity to multivariate functions (Theorem 7).

  Then, the benefits of Theorem 10 are discussed in §4.3. In particular, we can obtain the p-norm bound in the form of (1), which can be used to control the performance of plug-in forecasters for downstream tasks such as multiclass classification, learning with noisy labels, and bipartite ranking.

- §5: We evaluate the rate $\psi(\rho)$ by power functions such as $\rho^{1/s} \lesssim \psi(\rho) \lesssim \rho^{1/S}$ for some constants $s, S > 0$,[1] which is based on the Simonenko order function previously adopted [Bao23]. Our second main result roughly shows that $s \geqslant 2$ (Theorem 15), establishing the asymptotic optimality $\psi(\rho) \gtrsim \rho^{1/2}$ of strongly proper losses.

- §6: Several examples of convex functions to generate proper losses are discussed.

2. Background

In this section, we summarize the notation and basic properties of convex functions.

2.1 Notation

Throughout this article, fix $N \in \mathbb{N}$ and $p \in [1, \infty]$. The Kronecker delta is denoted by δ_{ij}. For $k \in \mathbb{N}$, we set $[k] \coloneqq \{1, 2, \dots, k\}$. A vector is denoted by bold-face such as $\boldsymbol{\xi} \in \mathbb{R}^N$, and its n-th (scalar) component is written as non-bold ξ_n for each $n \in [N]$. The p-norm of $\boldsymbol{\xi} \in \mathbb{R}^N$ is denoted by $\|\boldsymbol{\xi}\|_p$. For a topology on $\mathbb{R}^N$, we refer to one induced from the 2-norm, but it makes no difference whichever norm we choose. Similarly, a convexity of a function on $(\mathbb{R}^N, \|\cdot\|_p)$ is determined independently of the choice of p. The standard inner product on $\mathbb{R}^N$ is denoted by $\langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle \coloneqq \sum_{n \in [N]} \xi_n \xi'_n$. We introduce the notation

$$\triangle^N \coloneqq \left\{ \mathbf{q} \in \mathbb{R}^N \;\middle|\; q_n \geqslant 0 \ (n \in [N]),\ \langle \mathbf{q}, \mathbf{1} \rangle = 1 \right\}, \quad \triangle^N_+ \coloneqq \left\{ \mathbf{q} \in \triangle^N \;\middle|\; q_n > 0 \ (n \in [N]) \right\},$$

where $\mathbf{1} \in \mathbb{R}^N$ is the vector with each component being one. For $\mathbf{q} \in \triangle^N$, we denote by $\operatorname{supp}(\mathbf{q})$ the *support* of $\mathbf{q}$, that is,

$$\operatorname{supp}(\mathbf{q}) \coloneqq \{ n \in [N] \mid q_n > 0 \}.$$

We adhere to the convention that

$$\pm\infty \leqslant \pm\infty, \ \ a \pm \infty = \pm\infty, \ \ b \cdot (\pm\infty) = \pm\infty, \ \ -b \cdot (\pm\infty) = \mp\infty, \ \ 0 \cdot (\pm\infty) = 0, \ \ \ln 0 = -\infty,$$

1. In our notation, $\psi_1 \lesssim \psi_2$ indicates the existence of an absolute constant $C > 0$ such that $C\psi_1 \leqslant \psi_2$.

for $a \in \mathbb{R}$ and $b > 0$. We use O and Ω to denote the *infinitesimal* asymptotic order. To be precise, for two functions ϕ, ψ defined around 0, $\phi(\varepsilon) = O(\psi(\varepsilon))$ and $\phi(\varepsilon) = \Omega(\psi(\varepsilon))$ as $\varepsilon \downarrow 0$ should be understood as

$$\limsup_{\varepsilon \downarrow 0} \left| \frac{\phi(\varepsilon)}{\psi(\varepsilon)} \right| < \infty \quad \text{and} \quad \liminf_{\varepsilon \downarrow 0} \left| \frac{\phi(\varepsilon)}{\psi(\varepsilon)} \right| > 0,$$

respectively.

2.2 Convex analysis

In this subsection, let $f : \mathbb{R}^N \to (-\infty, \infty]$ denote a proper convex function on $\mathbb{R}^N$. Its *effective domain* is defined by

$$\operatorname{dom} f \coloneqq \left\{ \boldsymbol{\xi} \in \mathbb{R}^N \;\middle|\; f(\boldsymbol{\xi}) < \infty \right\}.$$

For a convex set $S \subseteq \mathbb{R}^N$, a convex function $f : S \to (-\infty, \infty]$ is said *strongly convex* on S with respect to the p-norm if we have

$$(1-t)f(\boldsymbol{\xi}) + tf\big(\check{\boldsymbol{\xi}}\big) - f\big((1-t)\boldsymbol{\xi} + t\check{\boldsymbol{\xi}}\big) \geqslant \frac{1}{2}\kappa t(1-t)\big\|\boldsymbol{\xi} - \check{\boldsymbol{\xi}}\big\|_p^2 \quad \text{for } \boldsymbol{\xi}, \check{\boldsymbol{\xi}} \in S,\ t \in (0,1),$$

for some $\kappa > 0$. A vector $\mathbf{v} \in \mathbb{R}^N$ is called a *subgradient* of f at $\boldsymbol{\xi}^0 \in \mathbb{R}^N$ if

$$f(\boldsymbol{\xi}) \geqslant f(\boldsymbol{\xi}^0) + \langle \mathbf{v}, \boldsymbol{\xi} - \boldsymbol{\xi}^0 \rangle \quad \text{for all } \boldsymbol{\xi} \in \mathbb{R}^N. \tag{2}$$

We say that f is *subdifferentiable* at $\boldsymbol{\xi}^0$ if there exists a subgradient of f at $\boldsymbol{\xi}^0$.

Assume $\triangle^N \subseteq \operatorname{dom} f$. Then f is subdifferentiable at $\mathbf{q}^0 \in \triangle^N_+$ [Roc70, Theorem 23.4]. For $\mathbf{q}^0 \in \triangle^N$, we extended the set of subgradient of f at $\mathbf{q}^0$ as

$$\partial f(\mathbf{q}^0) = \left\{ \mathbf{v} \in [-\infty, \infty)^N \;\middle|\; f(\mathbf{q}) \geqslant f(\mathbf{q}^0) + \big\langle \mathbf{v}, \mathbf{q} - \mathbf{q}^0 \big\rangle \text{ holds for all } \mathbf{q} \in \triangle^N \right\}.$$

The notion of $\partial f(\mathbf{q}^0)$ differs from the usual one [Roc70, §23] due to the restriction of the domain of f from $\mathbb{R}^N$ to $\triangle^N$ and the extension of the range of $\mathbf{v}$ from $\mathbb{R}^N$ to $[-\infty, \infty)^N$. This relaxation is useful to accommodate regular losses (given in Theorem 3 later). Note that, for $\mathbf{v} \in \partial f(\mathbf{q}^0)$, $v_n = -\infty$ happens only for $n \notin \operatorname{supp}(\mathbf{q}^0)$. We will show that $\partial f(\mathbf{q}^0)$ is nonempty for any $\mathbf{q}^0 \in \triangle^N$ in Appendix A.

We call a map $\partial f : \triangle^N \to 2^{[-\infty,\infty)^N}$ the *subdifferential* of f.

Bregman divergence. Let us denote an arbitrary *selector* of $\partial f(\mathbf{q})$ by ∇f, that is, a map assigning to each point $\mathbf{q} \in \triangle^N$ an element in $\partial f(\mathbf{q})$. Since $\partial f(\mathbf{q}^0)$ consists of the gradient of f at $\mathbf{q}^0$ if f is differentiable at $\mathbf{q}^0$, it is consistent to use the notation ∇f for a selector. For $\mathbf{q}, \mathbf{q}^0 \in \triangle^N$, the associated *Bregman divergence* [Bre67] of $\mathbf{q}$ given $\mathbf{q}^0$ is defined by

$$B_{(f,\nabla f)}(\mathbf{q} \| \mathbf{q}^0) \coloneqq f(\mathbf{q}) - f(\mathbf{q}^0) - \big\langle \nabla f(\mathbf{q}^0), \mathbf{q} - \mathbf{q}^0 \big\rangle \in [0, \infty].$$

Here, f is called the *generator* of the Bregman divergence $B_{(f,\nabla f)}$.

Subgradient equivalence. Note that the definition of the subgradient introduced here restricts the domain to $\triangle^N$. Therefore, if $\mathbf{v} \in \partial f(\mathbf{q}^0)$ for $\mathbf{q}^0 \in \triangle^N$, then for $\gamma : \triangle^N \to \mathbb{R}$, we have

$$\begin{aligned} f(\mathbf{q}) &\geqslant f(\mathbf{q}^0) + \left\langle \mathbf{v}, \mathbf{q} - \mathbf{q}^0 \right\rangle \\ &= f(\mathbf{q}^0) + \left\langle \mathbf{v}, \mathbf{q} - \mathbf{q}^0 \right\rangle + \gamma(\mathbf{q}^0) \cdot \left\langle \mathbf{1}, \mathbf{q} - \mathbf{q}^0 \right\rangle \qquad (\text{because } \mathbf{q}, \mathbf{q}^0 \in \triangle^N) \\ &= f(\mathbf{q}^0) + \left\langle \mathbf{v} + \gamma(\mathbf{q}^0) \cdot \mathbf{1}, \mathbf{q} - \mathbf{q}^0 \right\rangle. \end{aligned}$$

Thus, for a selector ∇f of ∂f and a function γ on $\triangle^N$, $\nabla f + \gamma \mathbf{1}$ is also a selector of ∂f and

$$B_{(f,\nabla f)} = B_{(f,\nabla f + \gamma \mathbf{1})} \tag{3}$$

holds. Readers must distinguish this notion of equivalence from the equivalence of scoring rules [Daw07, §1.1].

3. Classification, proper losses, and Savage representation

After introducing the learning problem of multiclass classification, we discuss loss functions and their properties. We review the notion of proper losses and its connection to Bregman divergences. Although this connection is already known, we formalize it rigorously. In particular, we verify the existence of minimizers of the conditional risk and its measurable selection (Theorem 1), which have been implicitly used in previous literature without any proof.

3.1 Multiclass classification

We regard a Radon space $\mathcal{X}$ as an input space, that is, the set of possible observations, and $\mathcal{Y} \coloneqq [N]$ as a set of labels. The set $\triangle^N$ is identified with the space of all probability measures on $\mathcal{Y}$. We fix a probability measure ν on $\mathcal{X} \times \mathcal{Y}$ and denote by $\nu_{\mathcal{X}}$ the marginal of ν on $\mathcal{X}$, that is, $\nu(\mathcal{B} \times \mathcal{Y}) = \nu_{\mathcal{X}}(\mathcal{B})$ holds for any measurable set $\mathcal{B} \subseteq \mathcal{X}$. Then, by the disintegration theorem [DM78, Chapter III-70 and 72], there exists a Borel map $\mathbf{q} : \mathcal{X} \to \triangle^N$, uniquely defined $\nu_{\mathcal{X}}$-a.e., such that

$$\nu(\mathcal{B} \times \{y\}) = \int_{\mathcal{B}} q_y(\mathbf{x}) \mathrm{d}\nu_{\mathcal{X}}(\mathbf{x}) \quad \text{for a measurable set } \mathcal{B} \subseteq \mathcal{X} \text{ and } y \in \mathcal{Y}.$$

We deem $\mathbf{q}(\mathbf{x}) \in \triangle^N$ a true probability vector at the input $\mathbf{x}$ induced from ν. *Multiclass classification* is a task to learn a forecaster $y : \mathcal{X} \to \mathcal{Y}$ to predict the most likely label

$$y(\mathbf{x}) \in \operatorname*{arg\,max}_{y \in \mathcal{Y}} q_y(\mathbf{x}) \quad \text{for each } \mathbf{x} \in \mathcal{X}.$$

3.2 Proper losses

We continue to use the notation in the previous subsection. To elicit the true probability map $\mathbf{q} : \mathcal{X} \to \triangle^N$, we use a *loss* $\boldsymbol{\ell}$, which is a Borel map from $\triangle^N$ to $[0, \infty]^N$. Define the associated *full risk* by

$$\mathbb{L}[\widehat{\mathbf{q}}] \coloneqq \int_{\mathcal{X} \times \mathcal{Y}} \ell_y(\widehat{\mathbf{q}}(\mathbf{x})) \mathrm{d}\nu(\mathbf{x}, y) \quad \text{for a Borel map } \widehat{\mathbf{q}} : \mathcal{X} \to \triangle^N.$$

A minimizer of $\mathbb{L}$ among Borel maps $\widehat{\mathbf{q}} : \mathcal{X} \to \triangle^N$ is called an *estimator* of $\mathbf{q} : \mathcal{X} \to \triangle^N$. The choice of $\boldsymbol{\ell}$ directly affects the quality of an estimator. It is more intuitive to work on the conditional counterpart of the full risk instead. Let $\mathbf{q}, \widehat{\mathbf{q}} \in \triangle^N$ with slight abuse of notation. For a loss $\boldsymbol{\ell}$, the associated *conditional risk* of $\widehat{\mathbf{q}}$ given $\mathbf{q}$ and *conditional Bayes risk* of $\mathbf{q}$ are defined by

$$L(\mathbf{q}, \widehat{\mathbf{q}}) \coloneqq \langle \mathbf{q}, \boldsymbol{\ell}(\widehat{\mathbf{q}}) \rangle = \sum_{y \in \mathcal{Y}} q_y \ell_y(\widehat{\mathbf{q}}) \quad \text{and} \quad \underline{L}(\mathbf{q}) \coloneqq \inf_{\widehat{\mathbf{q}} \in \triangle^N} L(\mathbf{q}, \widehat{\mathbf{q}}),$$

respectively. Here, we regard $\mathbf{q} \in \triangle^N$ as a *true* probability vector and $\widehat{\mathbf{q}} \in \triangle^N$ as an *estimate*. The full risk is rewritten as

$$\mathbb{L}[\widehat{\mathbf{q}}] = \int_{\mathcal{X}} L(\mathbf{q}(\mathbf{x}), \widehat{\mathbf{q}}(\mathbf{x})) \mathrm{d}\nu_{\mathcal{X}}(\mathbf{x}).$$

Since the infimum of a family of linear functions is concave, $\underline{L}$ is concave on $\triangle^N$, consequently $\underline{L} \circ \widehat{\mathbf{q}} : \mathcal{X} \to [-\infty, \infty)$ is measurable on $\mathcal{X}$, which in turn shows

$$\mathbb{L}[\widehat{\mathbf{q}}] \geqslant \int_{\mathcal{X}} \underline{L}(\mathbf{q}(\mathbf{x})) \mathrm{d}\nu_{\mathcal{X}}(\mathbf{x}).$$

Thus, the minimization problem of the full risk is reduced to that of the conditional risk if the map $\mathcal{M} : \triangle^N \to 2^{\triangle^N}$ defined by

$$\mathcal{M}(\mathbf{q}) \coloneqq \operatorname*{arg\,min}_{\widehat{\mathbf{q}} \in \triangle^N} L(\mathbf{q}, \widehat{\mathbf{q}}) \quad \text{for } \mathbf{q} \in \triangle^N.$$

has a Borel selector. Note that $\mathcal{M}(\mathbf{q}) = \varnothing$ may happen, whose example is given in Appendix B.

Next, we show that $\mathcal{M}$ has a Borel selector for continuous $\boldsymbol{\ell}$. Although this fact is not directly relevant to our main topic, we detail it in this article because we are unaware of any previous literature formalizing it for losses defined on $\triangle^N$. For a different type of (margin-based) losses, the existence of a measurable full risk minimizer has been studied [Ste07, Theorem 3.2 (ii)]. The proof is deferred to Appendix E.

Lemma 1 *Suppose $\boldsymbol{\ell} : \triangle^N \to [0, \infty]^N$ is lower semi-continuous. Then, $\underline{L}(\mathbf{q}) \geqslant 0$ holds and $\mathcal{M}(\mathbf{q})$ is nonempty and closed for $\mathbf{q} \in \triangle^N$. Moreover, if $\boldsymbol{\ell}$ is continuous, then there exists a Borel selector of $\mathcal{M}$.*

Our proof is based on a measurable selection theorem, which can be extended beyond proper losses to work on margin-based losses. However, the proof crucially depends on the continuity of $\boldsymbol{\ell}$ to invoke the selection theorem. This point is restrictive than the previous result [Ste07, Theorem 3.2 (ii)], but we hope that the proof of Theorem 1 is more concise and provides an insight.

Since $\mathcal{M}(\mathbf{q})$ is ideally a singleton of $\mathbf{q}$, we consider such a class of losses.

Definition 2 (Proper losses [WM68]) *A loss $\boldsymbol{\ell} : \triangle^N \to [0, \infty]^N$ is said to be* proper *if $\mathbf{q} \in \mathcal{M}(\mathbf{q})$ holds for each $\mathbf{q} \in \triangle^N$. We say $\boldsymbol{\ell}$ is* strictly proper *if $\mathcal{M}(\mathbf{q}) = \{\mathbf{q}\}$ holds for each $\mathbf{q} \in \triangle^N$.*

For a proper loss $\boldsymbol{\ell}$, the conditional risk is minimized at the true probability vector, and the identity map on $\triangle^N$ becomes a Borel selector of $\mathcal{M}$. In this case, it follows that $\underline{L}(\mathbf{q}) = L(\mathbf{q}, \mathbf{q})$ for $\mathbf{q} \in \triangle^N$. Note that recently discovered *calm composite losses* are also valid loss functions on $\triangle^N$ but a selector of $\mathcal{M}$ can be nonlinear, which generalize proper losses [BC25].

3.3 Savage representation

We will see that a proper loss has a connection with a Bregman divergence under the regularity.

Definition 3 (Regular losses [GR07, Definition 1]) *A loss $\boldsymbol{\ell} : \triangle^N \to [0,\infty]^N$ is said to be* regular *if $\ell_y(\mathbf{q}) = \infty$ happens only for $y \notin \operatorname{supp}(\mathbf{q})$.*

In what follows, we consider a regular loss $\boldsymbol{\ell} : \triangle^N \to [0,\infty]^N$.

For a regular loss, we define the *surrogate regret* $R : \triangle^N \times \triangle^N \to [0,\infty]$ by

$$R(\mathbf{q}, \widehat{\mathbf{q}}) \coloneqq L(\mathbf{q}, \widehat{\mathbf{q}}) - \underline{L}(\mathbf{q}) \quad \text{for } \mathbf{q}, \widehat{\mathbf{q}} \in \triangle^N,$$

which measures the suboptimality of an estimate $\widehat{\mathbf{q}}$ given a true $\mathbf{q}$. The surrogate regret will be used to assess the performance of $\widehat{\mathbf{q}}$ across different downstream tasks in §4.3. We call R the *surrogate* regret since it is a proxy performance measure to downstream tasks.

Although the following property has been well known in literature [Sav71, §4] [HB71, Theorem 3.1] [GR07, Theorem 2] [WVR16, Proposition 7], we provide its proof to handle the regularity and subdifferentials carefully.[2] In essence, the negative Bayes risk $-\underline{L}$ behaves as a Bregman generator therein.

Proposition 4 (Savage representation [Sav71, §4]) *Let $\boldsymbol{\ell}$ be regular. Then, $\boldsymbol{\ell}$ is proper (resp. strictly proper) if and only if there exists a proper convex (resp. strictly convex) function f on $\mathbb{R}^N$ such that $\operatorname{dom} f = \triangle^N$ and, for all $\widehat{\mathbf{q}} \in \triangle^N$, there exists a subgradient $\widehat{\mathbf{v}} \in \partial f(\widehat{\mathbf{q}})$ satisfying*

$$L(\mathbf{q}, \widehat{\mathbf{q}}) = -f(\widehat{\mathbf{q}}) - \langle \widehat{\mathbf{v}}, \mathbf{q} - \widehat{\mathbf{q}} \rangle \quad \textit{for } \mathbf{q} \in \triangle^N. \tag{4}$$

Proof First, assume that $\boldsymbol{\ell}$ is proper. Then, $\underline{L}(\mathbf{q}) = L(\mathbf{q}, \mathbf{q}) \in \mathbb{R}$. Define $f : \mathbb{R}^N \to (-\infty, \infty]$ by

$$f(\boldsymbol{\xi}) \coloneqq \begin{cases} -\underline{L}(\boldsymbol{\xi}) & \text{if } \boldsymbol{\xi} \in \triangle^N, \\ \infty & \text{otherwise,} \end{cases} \tag{5}$$

then f is a proper convex function on $\mathbb{R}^N$ such that $\operatorname{dom} f \subseteq \triangle^N$. For $\mathbf{q}, \widehat{\mathbf{q}} \in \triangle^N$, we have

$$f(\mathbf{q}) = -L(\mathbf{q}, \mathbf{q}) \geqslant -L(\mathbf{q}, \widehat{\mathbf{q}}) = -L(\widehat{\mathbf{q}}, \widehat{\mathbf{q}}) + \langle -\boldsymbol{\ell}(\widehat{\mathbf{q}}), \mathbf{q} - \widehat{\mathbf{q}} \rangle = f(\widehat{\mathbf{q}}) + \langle -\boldsymbol{\ell}(\widehat{\mathbf{q}}), \mathbf{q} - \widehat{\mathbf{q}} \rangle, \tag{6}$$

where the inequality is thanks to the properness of $\boldsymbol{\ell}$. Thus, $-\boldsymbol{\ell}(\widehat{\mathbf{q}}) \in \partial f(\widehat{\mathbf{q}})$ and Eq. (4) hold for any $\widehat{\mathbf{q}} \in \triangle^N$.

Conversely, suppose that there is a proper convex function f on $\mathbb{R}^N$ such that $\operatorname{dom} f = \triangle^N$ and, for all $\widehat{\mathbf{q}} \in \triangle^N$, there exists $\widehat{\mathbf{v}} \in \partial f(\widehat{\mathbf{q}})$ satisfying Eq. (4). Then, for $\mathbf{q} \in \triangle^N$, we have $L(\mathbf{q}, \mathbf{q}) = -f(\mathbf{q})$ and

$$L(\mathbf{q}, \mathbf{q}) = -f(\mathbf{q}) \leqslant -f(\widehat{\mathbf{q}}) - \langle \widehat{\mathbf{v}}, \mathbf{q} - \widehat{\mathbf{q}} \rangle = L(\mathbf{q}, \widehat{\mathbf{q}}) \quad \text{for } \widehat{\mathbf{q}} \in \triangle^N,$$

in turn, $\boldsymbol{\ell}$ is proper.

2. None of the aforementioned previous literature dealt with subgradients whose elements possibly take $-\infty$, though subtle. We carefully extend subdifferentials in §2.2 and check the subgradient inequality (2) for such subgradients in the proof.

Next, we show the equivalence of the strict properness of $\boldsymbol{\ell}$ and the strict convexity of f on $\triangle^N$. Let f be a convex function on $\triangle^N$ such that (4) holds. On the one hand, if $\boldsymbol{\ell}$ is strictly proper, then we have

$$\begin{aligned} f((1-t)\mathbf{q}+t\mathbf{q}') &= L((1-t)\mathbf{q}+t\mathbf{q}', (1-t)\mathbf{q}+t\mathbf{q}') \\ &= (1-t)L(\mathbf{q}, (1-t)\mathbf{q}+t\mathbf{q}') + tL(\mathbf{q}', (1-t)\mathbf{q}+t\mathbf{q}') \\ &> (1-t)L(\mathbf{q},\mathbf{q}) + tL(\mathbf{q}',\mathbf{q}') \\ &= (1-t)f(\mathbf{q},\mathbf{q}) + tf(\mathbf{q}',\mathbf{q}'), \end{aligned}$$

for $\mathbf{q}, \mathbf{q}' \in \triangle^N$ and $t \in (0,1)$, proving the strict convexity of f on $\triangle^N$. On the other hand, if $\boldsymbol{\ell}$ is not strictly proper, then there exist distinct $\mathbf{q}, \mathbf{q}' \in \triangle^N$ such that

$$f(\mathbf{q}) = -L(\mathbf{q},\mathbf{q}) = -L(\mathbf{q},\mathbf{q}') = f(\mathbf{q}') + \left\langle \mathbf{v}', \mathbf{q}-\mathbf{q}' \right\rangle,$$

where $\mathbf{v}' \in \partial f(\mathbf{q}')$ satisfies Eq. (4) such that $L(\mathbf{q},\mathbf{q}') = -f(\mathbf{q}') - \langle \mathbf{v}', \mathbf{q}-\mathbf{q}' \rangle$. For any $t \in (0,1)$, we have

$$\begin{aligned} -L((1-t)\mathbf{q}+t\mathbf{q}', (1-t)\mathbf{q}+t\mathbf{q}') &= f((1-t)\mathbf{q}+t\mathbf{q}') \\ &\leqslant (1-t)f(\mathbf{q}) + tf(\mathbf{q}') \\ &= f(\mathbf{q}') + (1-t)\left\langle \mathbf{v}', \mathbf{q}-\mathbf{q}' \right\rangle \\ &= f(\mathbf{q}') + \left\langle \mathbf{v}', (1-t)\mathbf{q}+t\mathbf{q}'-\mathbf{q}' \right\rangle \\ &= -L((1-t)\mathbf{q}+t\mathbf{q}', (1-t)\mathbf{q}+t\mathbf{q}'). \end{aligned}$$

This yields $f((1-t)\mathbf{q}+t\mathbf{q}') = (1-t)f(\mathbf{q}) + f(\mathbf{q}')$, that is, f is not strictly convex on $\triangle^N$.

This completes the proof of the proposition. □

As a by-product of the proof of Theorem 4, we obtain the following property thanks to the construction (5).

Corollary 5 (Subgradient of conditional Bayes risk) *For a regular proper loss $\boldsymbol{\ell}$, define a proper convex function f on $\mathbb{R}^N$ by (5). Then, we have $-\boldsymbol{\ell}(\widehat{\mathbf{q}}) \in \partial f(\widehat{\mathbf{q}})$ for $\widehat{\mathbf{q}} \in \triangle^N$. In addition, $-\boldsymbol{\ell} + \gamma\mathbf{1}$ remains to be a selector of ∂f for any $\gamma : \triangle^N \to \mathbb{R}$.*

Let $\boldsymbol{\ell}$ be regular and proper. Then we observe from inequalities (3) and (6) that

$$R = B_{(f,\nabla f)} = B_{(f,-\boldsymbol{\ell}+\gamma\mathbf{1})} \tag{7}$$

for a function on γ on $\triangle^N$. This gives a closed form of a subgradient of $-\underline{L}$ and is of interest per se. Nevertheless, when one generates a proper loss from a proper convex function f, it is more standard via the Savage representation (4) by $\ell_y(\widehat{\mathbf{q}}) = L(\mathbf{e}_y, \widehat{\mathbf{q}})$, where $\mathbf{e}_y = [\delta_{1y} \ \cdots \ \delta_{Ny}]^\top$ is the standard basis encoding the label $y \in [N]$.

Initial examples. We quickly see some examples of multiclass proper losses to let readers familiarize with the definitions so far, and discuss more examples in §6.

The first example is the log loss, $\ell_y(\mathbf{q}) = -\ln q_y$ for $y \in [N]$. This possibly takes ∞ for $q_y = 0$, for which we need the regularity (see Theorem 3). The log loss corresponds to the Kullback–Leibler divergence and Shannon entropy

$$R(\mathbf{q},\widehat{\mathbf{q}}) = D_{\mathrm{KL}}(\mathbf{q}\|\widehat{\mathbf{q}}) \coloneqq \sum_{y\in[N]} q_y \ln\left(\frac{q_y}{\widehat{q}_y}\right), \quad \underline{L}(\mathbf{q}) = -\sum_{y\in[N]} q_y \ln q_y$$

as the regret and the conditional Bayes risk, respectively. Here, we have $\nabla \underline{L}(\mathbf{q}) = \boldsymbol{\ell}(\mathbf{q}) - \mathbf{1}$, which is equivalent to $\boldsymbol{\ell}(\mathbf{q}) \in \partial \underline{L}(\mathbf{q})$ (in Theorem 5; and §2.2 for equivalent subgradients). Since each ℓ_y depends solely on q_y, this type of loss functions is called *local*. Indeed, the log loss is the only local proper loss [PDL12]. The locality is considered to be a desirable property in terms of interpretability [Du21].

The second example is the Brier score [Bri50]:

$$\ell_y(\mathbf{q}) = \frac{1}{2} \sum_{y' \in [N]} (\delta_{yy'} - q_{y'})^2 = -q_y + \frac{1 + \|\mathbf{q}\|_2^2}{2}.$$

This is no longer local as $\ell_y(\mathbf{q})$ depends on $q_{y'}$ for $y' \neq y$. Interestingly, the lack of the locality has been reportedly relevant to the emergent ability of language models [DZDT24]. The Brier score is strictly proper associated with the following regret and conditional Bayes risk

$$R(\mathbf{q}, \widehat{\mathbf{q}}) = \frac{1}{2} \|\mathbf{q} - \widehat{\mathbf{q}}\|_2^2, \quad \underline{L}(\mathbf{q}) = \frac{1 - \|\mathbf{q}\|_2^2}{2},$$

which are the squared 2-norm distance and negative squared 2-norm, respectively. Here, we have $\nabla \underline{L}(\mathbf{q}) = -\mathbf{q}$, which is equivalent to $\boldsymbol{\ell}(\mathbf{q}) \in \partial \underline{L}(\mathbf{q})$ because $\boldsymbol{\ell}(\mathbf{q}) = -\mathbf{q} + \gamma(\mathbf{q}) \cdot \mathbf{1}$ with the choice $\gamma(\mathbf{q}) \coloneqq (1 + \|\mathbf{q}\|_2^2)/2$. Note that $\nabla \underline{L}(\mathbf{q}) = -\mathbf{q}$ is not proper when regarded as a loss function [GR07, §4.1]; for this reason, we must always interpret the formula $\boldsymbol{\ell} \in \partial \underline{L}$ in Theorem 5 under the subgradient equivalence.

3.4 Strongly proper losses

For $\kappa > 0$, a loss $\boldsymbol{\ell}$ is called *κ-strongly proper* if

$$R(\mathbf{q}, \widehat{\mathbf{q}}) = L(\mathbf{q}, \widehat{\mathbf{q}}) - \underline{L}(\mathbf{q}) \geqslant \frac{\kappa}{2} \|\mathbf{q} - \widehat{\mathbf{q}}\|_2^2 \quad \text{for } \mathbf{q}, \widehat{\mathbf{q}} \in \triangle^N. \tag{8}$$

Strongly proper losses have been introduced for $N = 2$ [Aga14] and for general $N \geqslant 3$ [ZLA21] to derive a surrogate regret bound in the form of (1). For example, the log loss is 1-strongly proper [ZLA21, Lemma 3]. For $N = 2$, $\boldsymbol{\ell}$ is regular and strongly proper if and only if its conditional Bayes risk $-\underline{L}$ is strongly convex (with respect to the 2-norm) [Aga14, Theorem 10]. As an immediate consequence of the inequality (8), we have the 1/2-order surrogate regret bounds for strongly proper losses:

$$\|\mathbf{q} - \widehat{\mathbf{q}}\|_2 \leqslant \sqrt{\frac{2}{\kappa} R(\mathbf{q}, \widehat{\mathbf{q}})} \qquad \text{for any } \mathbf{q}, \widehat{\mathbf{q}} \in \triangle^N. \tag{9}$$

Several binary losses are shown to be strongly proper [Aga14, Table 1]. Beyond binary losses, a similar bound to (9) has been known for *Fenchel–Young losses* [BMN20] (which is relevant to proper losses but defined over "dual" points of $\widehat{\mathbf{q}} \in \triangle^N$), but requires a restrictive condition on $-\underline{L}$, Legendre-type. See [Blo19, Lemma 3] and [SBTO24, footnote 8] for details. In the next section, we derive surrogate regret bounds for general multiclass proper losses.

4. Regret bounds: Necessity of strict properness

In this section, we first study the moduli of convexity in §4.1. Therein, we show that the strict convexity of a function is equivalent to the strict monotonicity of its modulus (Theorem 8),

which ensures that its surrogate regret bound is non-vacuous. Then, we show in §4.2 the surrogate regret bounds for general multiclass proper losses beyond strongly proper losses. In §4.3, we relate surrogate regret bounds to several downstream tasks. Readers who are interested in the benefits of surrogate regret bounds may refer to this section first.

4.1 Moduli of convexity

Before introducing the moduli of convexity, we study the *midpoint Jensen gap* of a convex function $f : \triangle^N \to \mathbb{R}$, which is defined by

$$J(\mathbf{q}, \breve{\mathbf{q}}) \coloneqq \frac{f(\mathbf{q}) + f(\breve{\mathbf{q}})}{2} - f\left(\frac{\mathbf{q} + \breve{\mathbf{q}}}{2}\right) \quad \text{for } \mathbf{q}, \breve{\mathbf{q}} \in \triangle^N.$$

The midpoint Jensen gap is nonnegative by the convexity of f on $\triangle^N$. The midpoint Jensen gap is invariant under adding an affine function, and so is the modulus of convexity (defined later). That is, the midpoint Jensen gaps of two convex functions $f : \triangle^N \to \mathbb{R}$ and $f_{\lambda,\mathbf{u}} : \triangle^N \to \mathbb{R}$ defined by

$$f_{\lambda,\mathbf{u}}(\mathbf{q}) \coloneqq f(\mathbf{q}) + \langle \mathbf{u}, \mathbf{q} \rangle + \lambda \quad \text{for } \mathbf{q} \in \triangle^N$$

are the same, for any $\mathbf{u} \in \mathbb{R}^N$ and $\lambda \in \mathbb{R}$. Moreover, we will show that for continuous convex functions $f, g : \triangle^N \to \mathbb{R}$, their midpoint Jensen gaps are the same if and only if $f - g$ is affine. This property is reminiscent of the condition for the payoff equivalence [McC56, Theorem 3] and the universal equivalence of surrogate losses [NWJ09, Theorem 3] [DKR18, Theorem 1]. The proof is deferred to Appendix E.

Proposition 6 (Uniqueness up to affine functions) *Let $f, g : \triangle^N \to \mathbb{R}$ be continuous convex functions. Then, their midpoint Jensen gaps are the same if and only if $f - g$ is affine.*

We extend the moduli of convexity defined on $(\triangle^2, \|\cdot\|_1)$ [Bao23, Definition 4] to $(\triangle^N, \|\cdot\|_p)$. Note that the diameter of $(\triangle^N, \|\cdot\|_p)$ is $2^{1/p}$.

Definition 7 (Modulus of convexity [BGHV09]) *For a convex function $f : \triangle^N \to \mathbb{R}$, its* modulus of convexity *of f with respect to the p-norm is the function $\omega : [0, 2^{1/p}] \to [0, \infty)$ defined by*

$$\omega(r) \coloneqq \inf \left\{ J(\mathbf{q}, \breve{\mathbf{q}}) \;\middle|\; \mathbf{q}, \breve{\mathbf{q}} \in \triangle^N \text{ with } \|\mathbf{q} - \breve{\mathbf{q}}\|_p \geqslant r \right\} \quad \text{for } r \in [0, 2^{1/p}].$$

While the notion of moduli of convexity dates back to the classical literature on optimization [Pol66] and Banach spaces [Fig76], we view this as the smallest possible Jensen–Bregman divergence [NB11] with the fixed p-norm distance. This idea is similar to variational problems for deriving tight Pinsker's inequalities [Vaj70] [FHT03].

We will show that the convexity and strict convexity of a function are translated to the monotonicity and strict monotonicity of its modulus, respectively. This is an important result throughout this article because the moduli of convexity characterize surrogate regret bounds, as we will see in §4.2 soon.

Theorem 8 (Monotonicity of modulus) *Let $f : \triangle^N \to \mathbb{R}$ be a convex function. Then, the following assertions hold.*

1. *The modulus ω is non-decreasing on $[0, 2^{1/p}]$ and $\omega(0) = 0$.*
2. *f is strictly convex on $\triangle^N$ if and only if ω is strictly monotone on $[0, 2^{1/p}]$.*
3. *f is strongly convex on $\triangle^N$ with respect to the p-norm if and only if there exists $\kappa > 0$ such that $\omega(r) \geqslant \kappa r^2$ on $r \in [0, 2^{1/p}]$.*

From Theorem 8, the strong convexity of f is equivalent to the quadratic bound $\omega(r) \gtrsim r^2$. Thus, the modulus of convexity quantifies the convexity of a function.

Before proving Theorem 8, we show a lemma that is repeatedly used in the rest of the article. This lemma guarantees that the infimum of the modulus of convexity ω is attainable, and the minimizer lies at the boundary of the constraint $\|\mathbf{q} - \breve{\mathbf{q}}\|_p \geqslant r$ in its definition. The proof is deferred to Appendix E.

Lemma 9 *Let $f : \triangle^N \to \mathbb{R}$ be a convex function. For $r \in [0, 2^{1/p}]$, there exist $\mathbf{q}^r, \breve{\mathbf{q}}^r \in \triangle^N$ such that $\omega(r) = J(\mathbf{q}^r, \breve{\mathbf{q}}^r)$ and $\|\mathbf{q}^r - \breve{\mathbf{q}}^r\|_p = r$.*

Proof of Theorem 8 Define

$$\mathcal{D}^N(r) \coloneqq \left\{(\mathbf{q}, \widehat{\mathbf{q}}) \in \triangle^N \times \triangle^N \;\middle|\; \|\mathbf{q} - \widehat{\mathbf{q}}\|_p \geqslant r\right\}.$$

For $r', r \in [0, 2^{1/p}]$ with $r' \leqslant r$, we observe from the monotonicity $\mathcal{D}^N(r) \subseteq \mathcal{D}^N(r')$ that $\omega(r') \leqslant \omega(r)$. It is easily seen that $J(\mathbf{q}, \mathbf{q}) = 0$ holds for any $\mathbf{q} \in \triangle^N$ hence $\omega(0) = 0$. Thus, the first assertion follows.

To show the second assertion, assume the strict convexity of f on $\triangle^N$. Let $r \in (0, 2^{1/p}]$. By Theorem 9, there exist $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ such that $\omega(r) = J(\mathbf{q}, \breve{\mathbf{q}})$ and $\|\mathbf{q} - \breve{\mathbf{q}}\|_p = r$. Define a curve $c : [0, 1] \to \triangle^N$ by

$$c(t) \coloneqq (1 - t)\mathbf{q} + t\breve{\mathbf{q}} \quad \text{for } t \in [0, 1].$$

Since $f \circ c$ is strictly convex on $[0, 1]$, we have the strict inequality

$$\frac{f(c(\tau)) - f(c(0))}{\tau} < \frac{f(c(1)) - f(c(1 - \tau))}{\tau} \quad \text{for } \tau \in (0, 1/2].$$

Consequently, we conclude

$$\omega((1 - 2\tau)r) \leqslant J(c(\tau), c(1 - \tau)) < J(c(0), c(1)) = \omega(r) \quad \text{for } \tau \in (0, 1/2],$$

that is, the strict monotonicity of ω on $[0, 2^{1/p}]$. Conversely, if f is not strictly convex on $\triangle^N$, there exist distinct $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ such that

$$f((1 - t)\mathbf{q} + t\breve{\mathbf{q}}) = (1 - t)f(\mathbf{q}) + tf(\breve{\mathbf{q}}) \quad \text{for } t \in [0, 1].$$

This leads to $J(\mathbf{q}, \breve{\mathbf{q}}) = 0$. Consequently, ω is not strictly increasing on $[0, \|\mathbf{q} - \breve{\mathbf{q}}\|_p]$.

To show the third assertion, assume that f is $\widetilde{\kappa}$-strongly convex on $\triangle^N$ with respect to the p-norm for some $\widetilde{\kappa} > 0$. By Theorem 9, there exist $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ with $\omega(r) = J(\mathbf{q}, \breve{\mathbf{q}})$ and $\|\mathbf{q} - \breve{\mathbf{q}}\|_p = r$. Then, the strong convexity of f with the choice $t = 1/2$ implies

$$\omega(r) = \frac{f(\mathbf{q}) + f(\breve{\mathbf{q}})}{2} - f\left(\frac{\mathbf{q} + \breve{\mathbf{q}}}{2}\right) \geqslant \frac{\widetilde{\kappa}}{8}\|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 = \frac{\widetilde{\kappa}}{8}r^2.$$

Conversely, suppose that $\omega(r) \geqslant \kappa r^2$ on $r \in [0, 2^{1/p}]$. This implies

$$J(\mathbf{q}, \breve{\mathbf{q}}) = \frac{f(\mathbf{q}) + f(\breve{\mathbf{q}})}{2} - f\left(\frac{\mathbf{q} + \breve{\mathbf{q}}}{2}\right) \geqslant \omega(\|\mathbf{q} - \breve{\mathbf{q}}\|_p) \geqslant \kappa \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 \quad \text{for any } \mathbf{q}, \breve{\mathbf{q}} \in \triangle^N. \tag{10}$$

Taking $t \in [1/2, 1)$, we have

$$\begin{aligned} f((1-t)\mathbf{q} + t\breve{\mathbf{q}}) &= f\left((1-2t)\mathbf{q} + 2t\,\frac{\mathbf{q} + \breve{\mathbf{q}}}{2}\right) \\ &\leqslant (1-2t) f(\mathbf{q}) + 2t f\left(\frac{\mathbf{q} + \breve{\mathbf{q}}}{2}\right) \\ &\leqslant (1-2t) f(\mathbf{q}) + 2t \left[\frac{f(\mathbf{q}) + f(\breve{\mathbf{q}})}{2} - \kappa \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2\right] \\ &= (1-t) f(\mathbf{q}) + t f(\breve{\mathbf{q}}) - 2\kappa t \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 \\ &\leqslant (1-t) f(\mathbf{q}) + t f(\breve{\mathbf{q}}) - 2\kappa t(1-t) \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2, \end{aligned}$$

where we used (10). Switching the roles of $\mathbf{q}$ and $\widehat{\mathbf{q}}$ yields the 4κ-strongly convexity of f with respect to the p-norm.

All in all, the proof of the theorem is achieved. □

Despite the simple proof, this will lead to the necessity and sufficiency for a surrogate regret bound being non-vacuous in §4.2, together with Theorem 10.

4.2 Surrogate regret bounds

Now, we give surrogate regret bounds with respect to the p-norm. The asymptotic behavior of a surrogate regret bound for a proper loss $\boldsymbol{\ell}$ is essentially governed by the modulus of convexity of (the negative of) its conditional Bayes risk $-\underline{L}$. This is an extension of surrogate regret bounds for binary classification [Bao23, Theorem 6] to multiclass classification. Its proof (shown below) is an immediate generalization from [Bao23].

Theorem 10 (Surrogate regret bounds) *Let $\boldsymbol{\ell} : \triangle^N \to [0, \infty]^N$ be a regular proper loss and $f : \mathbb{R}^N \to (-\infty, \infty]$ a proper convex function defined by* (5) *with its modulus of convexity ω. For $\mathbf{q}, \widehat{\mathbf{q}} \in \triangle^N$, it holds*

$$\omega(\|\mathbf{q} - \widehat{\mathbf{q}}\|_p) \leqslant \frac{1}{2} R(\mathbf{q}, \widehat{\mathbf{q}}), \tag{11}$$

with the equality if $\mathbf{q} = \widehat{\mathbf{q}}$. If $\boldsymbol{\ell}$ is strictly proper additionally, then the equality of (11) *holds if and only if $\mathbf{q} = \widehat{\mathbf{q}}$.*

Proof By the definition of ω together with (7), it is sufficient to show

$$J(\mathbf{q}, \widehat{\mathbf{q}}) \leqslant \frac{1}{2} B_{(f, -\boldsymbol{\ell})}(\mathbf{q} \| \widehat{\mathbf{q}}) \quad \text{for } \mathbf{q}, \widehat{\mathbf{q}} \in \triangle^N.$$

By Theorem 5, we have

$$f\left(\frac{\mathbf{q} + \widehat{\mathbf{q}}}{2}\right) \geqslant f(\widehat{\mathbf{q}}) + \left\langle -\boldsymbol{\ell}(\widehat{\mathbf{q}}), \frac{\mathbf{q} + \widehat{\mathbf{q}}}{2} - \widehat{\mathbf{q}} \right\rangle = f(\widehat{\mathbf{q}}) + \frac{1}{2} \langle -\boldsymbol{\ell}(\widehat{\mathbf{q}}), \mathbf{q} - \widehat{\mathbf{q}} \rangle,$$

which implies

$$J(\mathbf{q}, \widehat{\mathbf{q}}) \leqslant \frac{1}{2}\left[f(\mathbf{q}) - f(\widehat{\mathbf{q}}) - \langle -\boldsymbol{\ell}(\widehat{\mathbf{q}}), \mathbf{q} - \widehat{\mathbf{q}}\rangle\right] = \frac{1}{2} B_{(f,-\boldsymbol{\ell})}(\mathbf{q} \| \widehat{\mathbf{q}}).$$

The equality can be seen immediately by choosing $\widehat{\mathbf{q}} = \mathbf{q}$. If $\boldsymbol{\ell}$ is strictly proper, then ω is strictly monotone by Theorem 8 and f is strictly convex, which indicates that the equality of (11) yields $\mathbf{q} = \widehat{\mathbf{q}}$. □

Let us discuss when a proper loss entails a *non-vacuous* bound, which means that $\|\mathbf{q} - \widehat{\mathbf{q}}\|_p$ approaches zero whenever the surrogate regret $R(\mathbf{q}, \widehat{\mathbf{q}})$ goes to zero. If $\boldsymbol{\ell}$ is strictly proper, then ω is strictly increasing, which in turn has an inverse function ω^{-1} and leads (11) to

$$\|\mathbf{q} - \widehat{\mathbf{q}}\|_p \leqslant \begin{cases} \omega^{-1}\left(\frac{1}{2} R(\mathbf{q}, \widehat{\mathbf{q}})\right) & \text{if } \ \frac{1}{2} R(\mathbf{q}, \widehat{\mathbf{q}}) \leqslant \omega\left(2^{\frac{1}{p}}\right), \\ 2^{\frac{1}{p}} & \text{otherwise.} \end{cases} \tag{12}$$

The strict monotonicity of ω^{-1} is essential for non-vacuous bounds because we then have $\rho \downarrow 0$ if and only if $\omega^{-1}(\rho) \downarrow 0$. Otherwise, we cannot always expect that the estimate $\widehat{\mathbf{q}}$ approaches $\mathbf{q}$ even if the suboptimality $R(\mathbf{q}, \widehat{\mathbf{q}})$ vanishes. By Theorems 4 and 8, the *strict* properness of $\boldsymbol{\ell}$ is necessary and sufficient for the surrogate regret bound (11) being non-vacuous. This is why strict properness matters.

If $\boldsymbol{\ell}$ is strongly proper, the inequality (8) implies that the negative Bayes risk $-\underline{L}$ is strongly convex. Combining the p-norm bound (12) with Theorem 8, we recover the 1/2-regret bound (9) modulo a constant.

Comparison: Pinsker's inequality. For illustration, let us consider the log loss under the binary case $N = 2$, where we identify $[q\ 1-q] \in \triangle^2$ with $q \in [0, 1]$. Then, the generator function f defined by (5) is the negative binary Shannon entropy $f(q) = q \ln q + (1-q)\ln(1-q)$. Its modulus of convexity (with 1-norm) admits the following form:

$$\omega(r) = \frac{1}{2}\left[\left(1 + \frac{r}{2}\right)\ln\left(1 + \frac{r}{2}\right) + \left(1 - \frac{r}{2}\right)\ln\left(1 - \frac{r}{2}\right)\right].$$

Note that this form coincides with the calibration function of the logistic loss [Ste07, Table 1]. The p-norm bound gives

$$2^{\frac{1}{p}}|q - \widehat{q}| \leqslant \omega^{-1}\left(\frac{1}{2} D_{\mathrm{KL}}(q \| \widehat{q})\right) \quad \text{for } q, \widehat{q} \in [0, 1]. \tag{13}$$

Note $2^{1/p}|q - \widehat{q}| = \|[q\ 1-q]^\top - [\widehat{q}\ 1-\widehat{q}]^\top\|_p$. Moreover, we can verify $\omega(r) \geqslant r^2/2$, which gives $|q - \widehat{q}|^2 \lesssim D_{\mathrm{KL}}(q\|\widehat{q})$, namely, Pinsker's inequality. The p-norm bound (12) can be viewed as generalizing Pinsker's inequality by allowing other Bregman divergences in the upper bound and the p-norm distance in the lower bound.

Comparison: surrogate regret bounds for margin-based losses. Margin-based classification [WS24], where a learner acts on $\mathbb{R}^N$-valued margin instead of a probabilistic estimate $\widehat{\mathbf{q}} \in \triangle^N$, is commonly used. Let us consider binary classification based on the

binary margin $z \in \mathbb{R}$. Given a true probability $q \in [0,1]$ (identified with $[q\ 1-q]^\top \in \triangle^2$) and margin $z \in \mathbb{R}$, the classification performance is evaluated by the (conditional) 0-1 regret

$$\mathrm{Reg}_{01}(q,z) \coloneqq [q\mathbb{1}_{\{z\leqslant 0\}} + (1-q)\mathbb{1}_{\{z>0\}}] - \min\{q, 1-q\},$$

where q and $1-q$ indicates the class probabilities of $y=1$ and $y=2$, respectively.

In the binary case $N=2$, we often use a (symmetric) margin-based losses $\phi : \mathbb{R} \to [0,\infty]$ as a surrogate loss, which operates on the binary margin $z \in \mathbb{R}$: the logistic loss $\phi_{\log}(z) = \ln(1+\exp(-z))$ and the hinge loss $\phi_{\mathrm{hinge}}(z) = \max\{0, 1-z\}$ are common examples. The prediction performances of the binary margin z for $y=1$ and $y=2$ are evaluated by $\phi(z)$ and $\phi(-z)$, respectively. For binary margin-based losses, surrogate regret bounds with respect to Reg_{01} have been studied intensively [BJM06, Theorem 3]. Here, we compare the 0-1 regret bounds and the p-norm regret bounds (Theorem 10). Let us denote the conditional risk *of the binary margin* z and Bayes risk given the true probability q by

$$L^{\mathrm{mgn}}(q,z) \coloneqq q\phi(z) + (1-q)\phi(-z), \quad \underline{L}^{\mathrm{mgn}}(q) \coloneqq \inf_{z\in\mathbb{R}} L^{\mathrm{mgn}}(q,z),$$

respectively, and define $\psi : [0,1] \to [0,\infty)$ by

$$\psi(r) \coloneqq \inf\left\{ L^{\mathrm{mgn}}\left(\frac{1+r}{2}, z\right) \,\middle|\, z \in \mathbb{R}, z \leqslant 0 \right\} - \underline{L}^{\mathrm{mgn}}\left(\frac{1+r}{2}\right). \tag{14}$$

Then, for $(q,z) \in [0,1]\times\mathbb{R}$, we have

$$\psi(\mathrm{Reg}_{01}(q,z)) \leqslant L^{\mathrm{mgn}}(q,z) - \underline{L}^{\mathrm{mgn}}(q). \tag{15}$$

The function ψ is called *ψ-transform*, and later generalized as a calibration function [Ste07]. The ψ-transform and the modulus of convexity ω characterize the 0-1 regret bound (15) and the p-norm bound (11), respectively, and are closely related with each other. The modulus of convexity is defined (in Theorem 7) by the best possible Jensen gap $J(\mathbf{q}, \widehat{\mathbf{q}})$ such that the true probability $\mathbf{q}$ and estimated probability $\widehat{\mathbf{q}}$ are distant at least r in the sense of the p-norm. By contrast, the ψ-transform in (14) can be rewritten as follows:

$$\begin{aligned} \psi(r) &= \inf\left\{ L^{\mathrm{mgn}}\left(\frac{1+r}{2}, z\right) - \underline{L}^{\mathrm{mgn}}\left(\frac{1+r}{2}\right) \,\middle|\, z \in \mathbb{R}, z \leqslant 0 \right\} \\ &= \inf\left\{ L^{\mathrm{mgn}}(q,z) - \underline{L}^{\mathrm{mgn}}(q) \mid \mathrm{Reg}_{01}(q,z) \geqslant r \right\}, \end{aligned} \tag{16}$$

which is the best possible surrogate regret $L^{\mathrm{mgn}}(q,z) - \underline{L}^{\mathrm{mgn}}(q)$ such that the margin prediction z is suboptimal with respect to the true class probability q by the level r at least. Therefore, both ψ-transform and the modulus of convexity measures the best possible surrogate regret given a true probability and suboptimal prediction by the level r least. Interested readers can refer to [Ste07, §4.1] to find more details of (16).

Comparison: surrogate regret bounds for proper composite losses. Let us leave a remark on the existing surrogate regret bounds for proper *composite* losses. A proper composite loss [WVR16] are the composition of a proper loss $\ell : \triangle^N \to [0,\infty]^N$ and an invertible link function $\boldsymbol{\lambda} : \triangle^N \to \mathbb{R}^N$, $\boldsymbol{\ell} \circ \boldsymbol{\lambda}^{-1} : \mathbb{R}^N \to [0,\infty]^N$, so that the composite loss

can operate on a multiclass margin $\mathbf{z} \in \mathbb{R}^N$ directly. The cross-entropy loss is of this type, where $\boldsymbol{\ell}$ is the log loss and $\boldsymbol{\lambda}^{-1}$ is the softmax function:

$$\lambda_y(\mathbf{z}) = \frac{\exp(z_y)}{\sum_{i \in [N]} \exp(z_i)}.$$

Prior to this article, a surrogate regret bound similar to (12) has been derived for proper composite losses, with the moduli of *continuity* of the conditional risk $L(\mathbf{q}, \cdot)$ [ML21, Corollary 3]. While the relationship between the moduli of convexity of $-\underline{L}$ and the moduli of continuity of $L(\mathbf{q}, \cdot)$ has not been clear, $-\underline{L}$ suffices for deriving surrogate regret bounds because a surrogate regret is solely determined by $\underline{L}$ due to (7) and Theorem 5. Therefore, our surrogate regret bounds in Theorem 10 can be readily applied to proper composite losses. Moreover, the existing surrogate regret bounds [ML21, Corollary 3] has been limited to the binary case $N = 2$. Our Theorem 10 is more general therein.

4.3 Relating surrogate regret to downstream tasks

The upper bound for the p-norm (12) is useful for many scenarios to assess the predictive performance of plug-in forecasters, i.e., post-processed forecasters based on the estimate $\widehat{\mathbf{q}}$. Thus, we can regard the p-norm bound as a *versatile* surrogate regret bound across different downstream tasks. Subsequently, we provide several examples of downstream tasks to support this idea.

Task 1: multiclass classification. Let us consider multiclass classification based on the post-process approach. Given true and estimated probability vectors $\mathbf{q}, \widehat{\mathbf{q}} \in \triangle^N$, respectively, the plug-in forecaster based on the estimate $\widehat{\mathbf{q}}$ is given by $\widehat{y} \in \arg\max_{y \in \mathcal{Y}} \widehat{q}_y$, where the tie is broken arbitrarily. Define $\mathbf{L} \in \mathbb{R}^{N \times N}$ the 0-1 loss matrix by $L_{ij} \coloneqq 1 - \delta_{ij}$ for each $(i, j) \in [N]^2$, and $\mathbf{L}_y$ denotes the y-th column vector of $\mathbf{L}$. Here, the forecaster's suboptimality in multiclass classification is measured by the (conditional) *0-1 regret*

$$\begin{aligned}
\mathrm{Reg}_{01}(\mathbf{q}, \widehat{\mathbf{q}}) &\coloneqq \sum_{n \in \mathcal{Y}} q_n (1 - \delta_{n\widehat{y}}) - \min_{y \in \mathcal{Y}} \sum_{n \in \mathcal{Y}} q_n (1 - \delta_{ny}) \\
&= \sum_{n \in \mathcal{Y}} q_n \left(L_{n\widehat{y}} - \min_{y \in \mathcal{Y}} L_{ny} \right) \\
&= \max_{y \in \mathcal{Y}} \sum_{n \in \mathcal{Y}} q_n \left(L_{n\widehat{y}} - L_{ny} \right) \\
&= \max_{y \in \mathcal{Y}} \left\langle \mathbf{q}, \mathbf{L}_{\widehat{y}} - \mathbf{L}_y \right\rangle,
\end{aligned}$$

for $\mathbf{q}, \widehat{\mathbf{q}} \in \triangle^N$. Let p^* denote the Hölder conjugate of p. The 0-1 regret can be bounded as

$$\mathrm{Reg}_{01}(\mathbf{q}, \widehat{\mathbf{q}}) \leqslant \max_{y \in \mathcal{Y}} \left\langle \mathbf{q} - \widehat{\mathbf{q}}, \mathbf{L}_{\widehat{y}} - \mathbf{L}_y \right\rangle \leqslant \|\mathbf{q} - \widehat{\mathbf{q}}\|_p \max_{y \in \mathcal{Y}} \|\mathbf{L}_{\widehat{y}} - \mathbf{L}_y\|_{p^*} \leqslant 2^{1 - \frac{1}{p}} \|\mathbf{q} - \widehat{\mathbf{q}}\|_p,$$

where the first inequality holds because $\left\langle \widehat{\mathbf{q}}, \mathbf{L}_{\widehat{y}} - \mathbf{L}_y \right\rangle \leqslant 0$ for any $y \in \mathcal{Y}$ attributed to the construction of $\widehat{y}$, and the second inequality owes to Hölder's inequality. Eventually, the 0-1 regret is controlled by the surrogate regret $R(\mathbf{q}, \widehat{\mathbf{q}})$ via (12) if $\boldsymbol{\ell}$ is strictly proper, which relates

the estimation quality of $\widehat{\mathbf{q}}$ to the predictive performance of the post-processed forecaster via the p-norm.

In case of binary classification, a closely related surrogate regret bound was presented [MNAC13, Lemma 4] (but for class-imbalanced plug-in forecasters), which has been later used to control the 1-norm between the estimated and true class probabilities for F-measure optimization [KNRD14]. Our result allows to generalize them to the multiclass case.

Note, however, that more direct control of the 0-1 regret is possible by

$$\Psi(\mathrm{Reg}_{01}(\mathbf{q}, \widehat{\mathbf{q}})) \leqslant R(\mathbf{q}, \widehat{\mathbf{q}}), \qquad \text{where } \Psi(r) = \underline{L}\left(\frac{1}{2}\right) - \underline{L}\left(\frac{1}{2} + r\right),$$

which can be obtained via the second-order Taylor expansion of proper losses [RW11, Corollary 27]. In this case, we may obtain non-vacuous bounds even for non-strictly proper losses. Our Theorem 10 does not provide such a tailored bound for the 0-1 regret but "one-size-fits-all" bounds for multiple tasks so that we can control the performance of other downstream tasks, not only classification.

Task 2: learning with noisy labels. Let us consider multiclass classification with class-conditional label noises: a true label y is observed as $\widetilde{y}$ with probability $C_{y\widetilde{y}}$ with a row-stochastic noise matrix $\mathbf{C} \in [0,1]^{N \times N}$. In this scenario, our access is limited to the noisy target probability vector $\widetilde{\mathbf{q}} = \mathbf{C}^\top \mathbf{q}$, through which a noisy estimate $\widehat{\mathbf{q}}$ is obtained. By following the noise-correction strategy [ZLA21], the plug-in forecaster based on the noisy estimate $\widehat{\mathbf{q}}$ is given by $\breve{y} \in \arg\max_{y \in \mathcal{Y}} \breve{q}_y$, where $\breve{\mathbf{q}} \coloneqq (\mathbf{C}^\top)^{-1} \widehat{\mathbf{q}}$ (provided that $\mathbf{C}$ is invertible). Under this setup, the 0-1 regret of $\breve{\mathbf{q}}$ given $\mathbf{q}$ is bounded as follows:

$$\begin{aligned}
\mathrm{Reg}_{01}(\mathbf{q}, \breve{\mathbf{q}}) &= \max_{y \in \mathcal{Y}} \left\langle \mathbf{q}, \mathbf{L}_{\breve{y}} - \mathbf{L}_y \right\rangle \\
&\leqslant \max_{y \in \mathcal{Y}} \left\langle \mathbf{q} - \breve{\mathbf{q}}, \mathbf{L}_{\breve{y}} - \mathbf{L}_y \right\rangle = \max_{y \in \mathcal{Y}} \left\langle \widetilde{\mathbf{q}} - \widehat{\mathbf{q}}, \mathbf{C}^{-1}(\mathbf{L}_{\breve{y}} - \mathbf{L}_y) \right\rangle \\
&\leqslant \|\widetilde{\mathbf{q}} - \widehat{\mathbf{q}}\|_p \max_{y \in \mathcal{Y}} \|\mathbf{C}^{-1}(\mathbf{L}_{\breve{y}} - \mathbf{L}_y)\|_{p*},
\end{aligned}$$

where the first inequality holds because $\left\langle \breve{\mathbf{q}}, \mathbf{L}_{\breve{y}} - \mathbf{L}_y \right\rangle \leqslant 0$ for any $y \in \mathcal{Y}$ attributed to the construction of $\breve{y}$, and the second inequality owes to Hölder's inequality. The p-norm $\|\widetilde{\mathbf{q}} - \widehat{\mathbf{q}}\|_p$ can be minimized even with access to the noisy observations only, and the p-norm bound (12) controls this by the surrogate regret of a strictly proper loss. This is also an extension of the previous surrogate regret transfer bounds [ZLA21, Theorem 4] beyond strongly proper losses.

Task 3: bipartite ranking. Consider $N = 2$ and identify $\mathbf{q} = [q \ 1-q]^\top \in \triangle^2$ with the instance $q \in [0,1]$. Given two instances $q, q' \in [0,1]$, we are interested in giving estimates $\widehat{q}, \widehat{q}' \in [0,1]$ that yield a consistent ranking with (q, q'). In bipartite ranking, we use the estimates $(\widehat{q}, \widehat{q}')$ directly without any post process. The (conditional) *ranking regret* [CLV08] [NA13] is measured by

$$\mathrm{Reg}_{\mathrm{rank}}(q, q', \widehat{q}, \widehat{q}') \coloneqq |q - q'| \left[\mathbb{1}_{\{(\widehat{q} - \widehat{q}')(q - q') < 0\}} + \frac{1}{2} \mathbb{1}_{\{\widehat{q} = \widehat{q}'\}} \right],$$

where $\mathbb{1}_{\{A\}} = 1$ when the predicate A holds and 0 otherwise, and the first and second terms penalize an inconsistent ranking and tie, respectively. This can be immediately related to the 1-norm [Aga14]:

$$\mathrm{Reg}_{\mathrm{rank}}(q, q', \widehat{q}, \widehat{q}') \leqslant |q - \widehat{q}| + |q' - \widehat{q}'|,$$

where the bound (12) can be further applied. Thus, the ranking regret is controlled by the surrogate regret.[3]

Other benefits. In addition to the above examples, one can easily relate the p-norm and downstream tasks such as binary classification with generalized performance criteria [KD16, (9)], which we omit here. Another benefit of the p-norm bound (12) is that it relates a possibly non-metric R to the p-norm.

To conclude this section, we raise attention to the kinship between moduli of convexity and the known devices such as calibration functions [BJM06] [Ste07] [OBLJ17] [BSS20] [BSX$^+$22], comparison inequalities [MPRS12] [CRR20], and Fisher consistency bounds [AMMZ22a] [AMMZ22b] [MMZ23b]. In spite of the relevance, moduli of convexity are different in that these devices have been tailored for a specific target loss of each downstream task, whereas moduli are concerned with the p-norm. For recent developments, see [CMZ25]—the author group has extended to $\mathcal{H}$-consistency bounds accounting for hypothesis spaces. Interested readers may consult their prolific body of work on this topic.[4]

5. Lower bounds of surrogate regret order

We move on to the next main result: the surrogate regret order via the modulus of convexity cannot go beyond the square root. To this end, we first review the Simonenko order function and the strong convexity used to establish the main result, and then show the main result. In this section, let $f : \triangle^N \to \mathbb{R}$ be a convex function unless otherwise stated.

5.1 Power evaluation of moduli

To analyze how fast the surrogate regret bound (11) can be, we analyze the behavior of the modulus ω in terms of power functions. To this end, we introduce the order of ω below, which is well-defined since $\omega(r) > 0$ for $r \in (0, 2^{1/p}]$ from Theorem 8.

Definition 11 (Simonenko order function [Sim64]) *Let $f : \triangle^N \to \mathbb{R}$ be a strictly convex function and $\omega : [0, 2^{1/p}] \to [0, \infty)$ be the modulus of convexity of f. The* Simonenko order function $\sigma : (0, 2^{1/p}] \to [0, \infty]$ *(associated with ω) is defined by*

$$\sigma(r) := \frac{rD^-\omega(r)}{\omega(r)} \quad \textit{for } r \in \left(0, 2^{\frac{1}{p}}\right], \quad \textit{where} \quad D^-\omega(r) := \limsup_{\varepsilon \downarrow 0} \frac{\omega(r) - \omega(r-\varepsilon)}{\varepsilon}.$$

The quantity $D^-\omega$ is called the *upper left Dini derivative* of ω at r. If ω is differentiable at r, then $D^-\omega(r) = \omega'(r)$ holds. The Simonenko order function σ evaluates the order of ω. Note that the following result holds for general continuous functions beyond moduli of convexity, but we restrict ourselves to moduli of convexity for brevity.

3. While we consider the plug-in approach to bipartite ranking, a number of studies have considered the pairwise ranking approach [AGH$^+$05] [KDH11], taking the difference of two margin predictions optimized with a margin-based loss function. Interestingly, the surrogate regret bound unavoidably becomes vacuous when we use a restricted hypothesis space, as shown recently [MMZ23a].
4. We are grateful to a reviewer for alerting us to this remarkably recent body of work.

Proposition 12 (Power evaluations of moduli) *Let $f : \triangle^N \to \mathbb{R}$ be a strictly convex function. For a fixed $r_0 \in (0, 2^{1/p}]$, define $s, S \in [0, \infty]$ by*

$$s := \inf_{r \in (0, r_0]} \sigma(r), \quad S := \sup_{r \in (0, r_0]} \sigma(r),$$

and assume $S < \infty$. Then, the function $r \mapsto \omega(r) r^{-s}$ is non-decreasing on $(0, r_0)$ and the function $r \mapsto \omega(r) r^{-S}$ is non-increasing on $(0, r_0)$. Moreover, the following inequalities hold for any $r \in [0, r_0]$:

$$\left[\frac{\omega(r_0)}{r_0^S}\right] r^S \leqslant \omega(r) \leqslant \left[\frac{\omega(r_0)}{r_0^s}\right] r^s.$$

Roughly speaking, Theorem 12 provides us $r^S \lesssim \omega(r) \lesssim r^s$ as $r \downarrow 0$. This order evaluation is useful when analyzing the asymptotic convergence rate of the surrogate regret bound (11). Theorem 12 is easily proved when ω is differentiable. Indeed, if ω is differentiable, then it holds for $t \in (0, r_0)$ that

$$\frac{s}{t} = \frac{1}{t} \inf_{r \in (0, r_0]} \sigma(r) \leqslant \frac{\omega'(t)}{\omega(t)} \leqslant \frac{1}{t} \sup_{r \in (0, r_0]} \sigma(r) = \frac{S}{t}$$

by the definition s and S, and integrating it on $[r, r'] \subseteq [0, r_0]$ gives

$$s \ln \frac{r'}{r} \leqslant \int_r^{r'} \frac{\omega'(t)}{\omega(t)} \mathrm{d}t = \ln \frac{\omega(r')}{\omega(r)} \leqslant S \ln \frac{r'}{r},$$

which is equivalent to the desired monotonicity. Thus, the p-norm upper bound (12) is controlled by the rate $\omega^{-1}(\rho) = O(\rho^{1/S})$ as $\rho \downarrow 0$. Since we are interested in the behavior of $\|\mathbf{q} - \widehat{\mathbf{q}}\|_p$ when $\widehat{\mathbf{q}}$ is close to the minimizer of $R(\mathbf{q}, \cdot)$, we focus on the asymptotic behavior of the Simonenko order function as $r \downarrow 0$.

The complete proof of Theorem 8 for non-differentiable ω is deferred to Appendix C.

5.2 Strong convexity and its relation to moduli

For the asymptotic analysis of σ, we leverage strong convexity [Nes13, §2.1.3]. Herein, we define the strong convexity parameter for a convex function $f : \triangle^N \to \mathbb{R}$ and $t \in (0, 1)$ by

$$\kappa_p^{f,t} := \inf \left\{ \frac{2\left[(1-t)f(\mathbf{q}) + tf(\breve{\mathbf{q}}) - f((1-t)\mathbf{q} + t\breve{\mathbf{q}})\right]}{t(1-t)\|\mathbf{q} - \breve{\mathbf{q}}\|_p^2} \;\middle|\; \text{distinct } \mathbf{q}, \breve{\mathbf{q}} \in \triangle^N \right\},$$
$$\kappa_p^f := \inf_{t \in (0,1)} \kappa_p^{f,t}.$$

We observe from the convexity of f that $\kappa_p^f \in [0, \infty)$ and

$$f((1-t)\mathbf{q} + t\breve{\mathbf{q}}) \leqslant (1-t)f(\mathbf{q}) + tf(\breve{\mathbf{q}}) - \frac{\kappa_p^f}{2} t(1-t)\|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 \quad \text{for } \mathbf{q}, \breve{\mathbf{q}} \in \triangle^N \text{ and } t \in (0, 1).$$

When $p = 2$ and f is twice continuously differentiable over $\triangle_+^N$, $\operatorname{Hess} f$ (the Hessian of f) satisfies $\operatorname{Hess} f - \kappa_2^f \mathbf{I}_N \succeq \mathbf{O}$ [Nes13, Theorem 2.1.10], where $\mathbf{I}_N$ is the identity matrix.[5]

5. Remark that the strong convexity parameter depends on the underlying set where we take the infimum. Let us define the strong convexity parameter on $\mathbb{R}^N$ by replacing $\triangle^N$ with $\mathbb{R}^N$ and write $\bar{\kappa}_p^{f,t}$ instead of $\kappa_p^{f,t}$ for each $t \in (0, 1)$. By $\triangle^N \subseteq \mathbb{R}^N$, we observe that $\bar{\kappa}_p^{f,t} \leqslant \kappa_p^{f,t}$ for a function $f : \mathbb{R}^N \to (-\infty, \infty]$, where the equality does not necessarily hold.

To calculate κ_p^f, we only need to know $\kappa_p^{f,1/2}$. The proof is deferred to Appendix E.

Proposition 13 (Strong convexity parameter at midpoint) *For a continuous convex function $f : \triangle^N \to \mathbb{R}$, it holds $\kappa_p^f = \kappa_p^{f,1/2}$.*

The strong convexity parameter κ_p^f is equivalent up to constant across different $p \geqslant 1$, which can be seen as follows. For the midpoint Jensen gap J (defined in §4.1), since we have

$$\kappa_p^{f,\frac{1}{2}} = \inf \left\{ \frac{8J(\mathbf{q}, \breve{\mathbf{q}})}{\|\mathbf{q} - \breve{\mathbf{q}}\|_p^2} \;\middle|\; \text{distinct } \mathbf{q}, \breve{\mathbf{q}} \in \triangle^N \right\}, \tag{17}$$

the following bound holds:

$$\inf \left\{ \frac{\|\mathbf{q}\|_p^2}{\|\mathbf{q}\|_2^2} \;\middle|\; \mathbf{q} \in \triangle^N \right\} \leqslant \frac{\kappa_2^{f,\frac{1}{2}}}{\kappa_p^{f,\frac{1}{2}}} \leqslant \sup \left\{ \frac{\|\mathbf{q}\|_p^2}{\|\mathbf{q}\|_2^2} \;\middle|\; \mathbf{q} \in \triangle^N \right\}.$$

In particular, if $N = 2$, we have

$$\|\mathbf{q} - \breve{\mathbf{q}}\|_p = [(q_1 - \breve{q}_1)^p + (q_2 - \breve{q}_2)^p]^{\frac{1}{p}} = [(q_1 - \breve{q}_1)^p + (q_1 - \breve{q}_1)^p]^{\frac{1}{p}} = 2^{\frac{1}{p}} |q_1 - \breve{q}_1|$$

for $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^2$, and hence, $2\kappa_2^{f,1/2} = 2^{2/p} \kappa_p^{f,1/2}$. Therefore, κ_p^f remains the same up to constant regardless of the choice of $p \geqslant 1$.

We will use the following representation of $\kappa_p^{f,1/2} = \kappa_p^f$ (given by Theorem 13) repeatedly later, which follows by definition of the modulus of convexity (defined in Theorem 7) and (17):

$$\kappa_p^f = \inf \left\{ \frac{8\omega(r)}{r^2} \;\middle|\; r \in (0, 2^{\frac{1}{p}}] \right\}. \tag{18}$$

5.3 Asymptotic lower bound

The asymptotic behavior of the Simonenko order $\sigma(r)$ as $r \downarrow 0$ is controlled by the strong convexity parameter κ_p^f. We consider a "local" version of the strong convexity parameter.

Definition 14 (Local strong convexity modulus) *For a convex function $f : \triangle^N \to \mathbb{R}$, define $K_p^f : (0, 2^{1/p}] \to \mathbb{R}$ by*

$$K_p^f(r) := \frac{8\omega(r)}{r^2}.$$

This quantity is defined based on the alternative expression of the strong convexity parameter κ_p^f in (18). From the relationship (18), $K_p^f(r) \geqslant \kappa_p^f$ always holds on $r \in (0, 2^{1/p}]$. We show that K_p^f is lower semi-continuous and left-continuous (but not continuous in general without additional assumptions) in Appendix D.

Now, we analyze the asymptotic behavior of the moduli of convexity $\sigma(r)$ at $r \downarrow 0$ when the Bregman generator f is continuous on $\triangle^N$, which is our second main result. Therein, we assume the continuity of f to prevent f from being discontinuous on the $\triangle^N \backslash \triangle_+^N$.

Theorem 15 (Lower bound of order) *Let $f : \triangle^N \to \mathbb{R}$ be a continuous strictly convex function. Assume one of the following two conditions:*

(C1) $\kappa_p^f > 0$.

(C2) K_p^f *is continuous on* $(0, r_0]$ *for some* $r_0 \in (0, 2^{1/p}]$ *and* K_p^f *converges as* $r \downarrow 0$.

Then,

$$\limsup_{r\downarrow 0} \sigma(r) \geqslant 2. \tag{19}$$

Moreover, if we assume both conditions, then

$$\liminf_{r\downarrow 0} \sigma(r) \geqslant 2. \tag{20}$$

Let us discuss the applicability of Theorem 15. First, f is assumed to be strictly convex, which means that we deal with a strictly proper loss through the strictly convex negative Bayes risk $f = -\underline{L}$ in (5). In Theorem 15, we additionally require either (C1) of (C2). The condition (C1) assumes nothing else but the strong convexity of f. In other words, (C1) assumes that the underlying proper loss is strongly proper. Indeed, the strong convexity parameter $\kappa_p^f > 0$ is equivalent to the modulus κ defining strongly proper losses in (8). It is more interesting when (C1) does not hold but (C2) holds, where the underlying loss is strictly proper but no longer strongly proper. The continuity assumption of K_p^f is mild enough to cover many reasonable examples of the negative Bayes risk f, as we will see in §6.

It follows from Theorem 12 and Theorem 15 that the p-norm bound (12) is controlled by the rate of $\omega^{-1}(\rho)$ cannot be faster than $O(\rho^{1/2})$ for a strictly proper $\boldsymbol{\ell}$ satisfying either (C1) or (C2). To see this, we invoke Theorem 12 to observe that for a fixed $r_0 \in (0, 2^{1/p}]$,

$$\omega^{-1}(\rho) \leqslant r_0 \omega(r_0)^{-\frac{1}{S}} \cdot \rho^{\frac{1}{S}} \quad \text{for any } r \in [0, r_0] \text{ such that } \rho = \omega(r).$$

If the bound (19) holds, then we have

$$2 \leqslant \limsup_{r\downarrow 0} \sigma(r) \leqslant \sup_{r\in(0,r_0]} \sigma(r) = S,$$

which implies $\rho^{1/S} \geqslant \rho^{1/2}$ (for $\rho < 1$). Thus, we discern the optimal rate $\omega^{-1}(\rho) = O(\rho^{1/2})$ for $\rho \in [0, 1]$. This assures that strongly proper losses asymptotically achieve the optimal rate $O(\rho^{1/2})$ as seen in (9). Moreover, the optimal rate $\omega^{-1}(\rho) = O(\rho^{1/2})$ penetrates into a broad family of strictly proper losses.

To prove Theorem 15, we leverage the following lemma to locally control $\sigma(r)$, which is proven in Appendix E.

Lemma 16 *Let* $f : \triangle^N \to \mathbb{R}$ *be a continuous convex function. Then, for any* $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ *and* $r \in (0, 2^{1/p}]$*,*

$$\liminf_{r\downarrow 0} K_p^f(r) = \kappa_p^f, \quad J(\mathbf{q}, \breve{\mathbf{q}}) \geqslant \frac{\kappa_p^f}{8} \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2, \quad \textit{and} \quad D^-\omega(r) \geqslant \frac{\kappa_p^f}{4} r.$$

Proof of Theorem 15 We observe from the strict convexity of f and Theorem 8 that $K_p^f > 0$ on $r \in (0, 2^{1/p}]$. By assuming (C1) only, it follows from Theorem 16 that

$$\limsup_{r\downarrow 0} \sigma(r) = \limsup_{r\downarrow 0} \frac{r D^-\omega(r)}{\omega(r)} \geqslant \limsup_{r\downarrow 0} \frac{r \cdot \frac{\kappa_p^f}{4} r}{\frac{K_p^f(r)}{8} r^2} = \limsup_{r\downarrow 0} \frac{2\kappa_p^f}{K_p^f(r)} = 2. \tag{21}$$

In addition, assume (C2) together. Then, in the similar manner to Eq. (21), we have

$$\liminf_{r\downarrow 0}\sigma(r) \geqslant \liminf_{r\downarrow 0}\frac{2\kappa_p^f}{K_p^f(r)} = \lim_{r\downarrow 0}\frac{2\kappa_p^f}{K_p^f(r)} = 2.$$

Next, assume (C2) only, and $\kappa_p^f > 0$ does not hold. In this case, Theorem 16 indicates that $K_p^f(r) \downarrow 0$ as $r \downarrow 0$, from which with the intermediate value theorem, we can inductively define $(r_j)_{j\in\mathbb{N}} \subseteq (0, r_0]$ by

$$r_j := \inf\left\{ r \in (0, 2^{1/p}] \;\middle|\; K_p^f(r) \geqslant \frac{1}{2}K_p^f(r_{j-1}) \right\}.$$

Then, $(r_j)_{j\in\mathbb{N}}$ converges to 0 because $K_p^f > 0$ always holds on $r \in (0, 2^{1/p}]$ and hence $K_p^f(r) = 0$ if and only if $r = 0$. We see

$$K_p^f(r) < \frac{1}{2}K_p^f(r_{j-1}) = K_p^f(r_j) \quad \text{for } r \in (0, r_j].$$

This yields

$$\begin{aligned} D^-\omega(r_j) = \limsup_{\varepsilon\downarrow 0}\frac{\omega(r_j)-\omega(r_j-\varepsilon)}{\varepsilon} &= \limsup_{\varepsilon\downarrow 0}\frac{\frac{K_p^f(r_j)}{8}r_j^2 - \frac{K_p^f(r_j-\varepsilon)}{8}(r_j-\varepsilon)^2}{\varepsilon} \\ &\geqslant \limsup_{\varepsilon\downarrow 0}\frac{\frac{K_p^f(r_j)}{8}r_j^2 - \frac{K_p^f(r_j)}{8}(r_j-\varepsilon)^2}{\varepsilon} = \frac{K_p^f(r_j)}{4}r_j, \end{aligned}$$

which implies

$$\limsup_{r\downarrow 0}\sigma(r) \geqslant \limsup_{j\to\infty}\sigma(r_j) = \limsup_{j\to\infty}\frac{r_j D^-\omega(r_j)}{\omega(r_j)} \geqslant \limsup_{j\to\infty}\frac{r_j\cdot\frac{K_p^f(r_j)}{4}r_j}{\frac{K_p^f(r_j)}{8}r_j^2} = 2.$$

Thus the proof of Theorem 15 is completed. □

Lower bound of $\omega^{-1}(\rho)$. Theorem 12 also implies that for some $r_0 \in (0, 2^{1/p}]$,

$$\omega^{-1}(\rho) \geqslant [r_0\omega(r_0)^{-\frac{1}{s}}]\cdot\rho^{\frac{1}{s}} \quad \text{for } r \in [0, r_0] \text{ such that } \rho = \omega(r).$$

If the bound (20) holds with the *strict* inequality, then we can choose r_0 to satisfy

$$\frac{2+\varsigma}{2} \leqslant \inf_{r\in(0,r_0]}\sigma(r) = s \quad \text{for } \varsigma := \liminf_{r\downarrow 0}\sigma(r) > 2,$$

which implies $\rho^{1/s} \geqslant \rho^{2/(2+\varsigma)} > \rho^{1/2}$ (for $\rho < 1$). Thus, ω^{-1} admits the lower bound $\Omega(\rho^{1/2})$ for $\rho \in [0, 1]$.

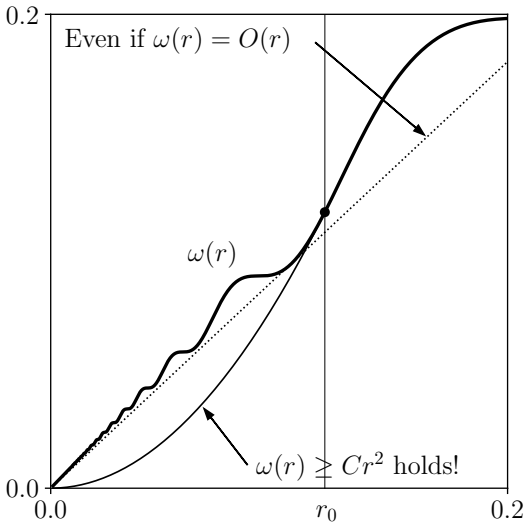

Figure 1: Illustration of $\omega(r) = r\sin\left(\frac{1}{r}\right) - \mathrm{Ci}\left(\frac{1}{r}\right) + r$. This ω is asymptotically linear, but can only have a slower finite-range bound $\omega(r) \gtrsim r^2$ than the linear rate.

Comparison with the known lower bound. A relevant lower bound $\omega^{-1}(\rho) = \Omega(\rho^{1/2})$ has been shown previously for margin-based losses [FW21, Theorem 4] [MMZ24, Theorem 4.2]. These lower bounds assume that a loss is strongly convex and has a locally Lipschitz gradient [FW21, Assumption 1] or a loss satisfies a relaxed version of the strict convexity [MMZ24, Theorem 4.2]. These conditions are assumed under the loss differentiability, whereas both our (C1) and (C2) do not need the differentiability of ℓ. Ergo, the differentiability assumption is lifted to show the optimality of $\omega^{-1}(\rho) = O(\rho^{1/2})$. To show $\omega^{-1}(\rho) = \Omega(\rho^{1/2})$ in our case, (C1) and (C2) coupled with the existence of the limit of $K_p^f(r)$ as $r \downarrow 0$ suffice.

As a side note, convolutional Fenchel–Young losses have been recently proposed to circumvent these commonly known squared-root lower bounds for convex smooth surrogate losses [CBFA25]. Their surrogate regret bounds are derived for the multiclass 0-1 loss (and its generalization), which is clearly different from our the p-norm distance. Thus, it does not contradict with the square-root lower bounds of Theorem 10.

Remark 17 *Our analysis with the Simonenko order evaluates the order of ω by a power function in the form of $r^S \lesssim \omega(r) \lesssim r^s$ for $r \in [0, 2^{1/p}]$. This is an evaluation for a* finite *range, which is more than an* asymptotic *evaluation. Despite its subtlety, it often matters, as seen in the following example:*

$$\omega(r) = r\sin\left(\frac{1}{r}\right) - \mathrm{Ci}\left(\frac{1}{r}\right) + r \quad \textit{for } r > 0, \quad \textit{where} \quad \mathrm{Ci}(z) := -\int_z^\infty t^{-1}\cos(t)\mathrm{d}t.$$

This ω is monotonically increasing and satisfies $\omega(r) = O(r)$ as $r \downarrow 0$. However, when it comes to the finite evaluation, we cannot go faster than $\omega(r) \gtrsim r^2$ just because this ω satisfies (C2) *and Theorem 15 implies* $\limsup_{r\downarrow 0} \sigma(r) \geqslant 2$. *Thus, the finite evaluation gives a better characterization when we assess the convergence rate of finitely large surrogate regret. See Fig. 1 to better understand the above example.*

Table 1: Examples of a convex function f. For ω, we show the expressions with $(N, p) = (2, 1)$.

	$f(\mathbf{q})$	α	Modulus $\omega(r)$	Loss $\boldsymbol{\ell}$
Shannon	$\langle \mathbf{q}, \ln \mathbf{q} \rangle$	—	$\frac{1}{2}[(1+\frac{r}{2})\ln(1+\frac{r}{2}) + (1-\frac{r}{2})\ln(1-\frac{r}{2})]$	Log
$\|\cdot\|_\alpha^2$	$\frac{\|\mathbf{q}\|_\alpha^2 - 1}{2}$	$1 < \alpha < 2$	$\frac{1}{8}\left\|\begin{bmatrix}1+r/2\\1-r/2\end{bmatrix}\right\|_\alpha^2 - 2^{2/\alpha-3}$	Brier ($\alpha = 2$)
		$2 \leqslant \alpha$	$\frac{1}{4}\left\|\begin{bmatrix}r/2\\1-r/2\end{bmatrix}\right\|_\alpha^2 - \frac{1}{8}\left\|\begin{bmatrix}r/2\\2-r/2\end{bmatrix}\right\|_\alpha^2 + \frac{1}{4}$	
$\|\cdot\|_\alpha$	$\frac{\|\mathbf{q}\|_\alpha - 1}{\alpha - 1}$	$1 < \alpha \leqslant 3/2$	$\frac{1}{2(\alpha-1)}\left(\left\|\begin{bmatrix}1+r/2\\1-r/2\end{bmatrix}\right\|_\alpha - 2^{1/\alpha}\right)$	Pseudo-spherical
		$3/2 < \alpha < 2$	(No closed-form in general)	
		$2 \leqslant \alpha$	$\frac{1}{2(\alpha-1)}\left(\left\|\begin{bmatrix}r/2\\1-r/2\end{bmatrix}\right\|_\alpha - \left\|\begin{bmatrix}r/2\\2-r/2\end{bmatrix}\right\|_\alpha + 1\right)$	
Tsallis	$\frac{\|\mathbf{q}\|_\alpha^\alpha - 1}{\alpha - 1}$	$1 < \alpha < 2$ $3 < \alpha$	$\frac{1}{2^\alpha(\alpha-1)}\left(\left\|\begin{bmatrix}1+r/2\\1-r/2\end{bmatrix}\right\|_\alpha^\alpha - 2\right)$	α-log
		$2 \leqslant \alpha \leqslant 3$	$\frac{1}{2(\alpha-1)}\left(\left\|\begin{bmatrix}r/2\\1-r/2\end{bmatrix}\right\|_\alpha^\alpha - \frac{1}{2^{\alpha-1}}\left\|\begin{bmatrix}r/2\\2-r/2\end{bmatrix}\right\|_\alpha^\alpha + 1\right)$	

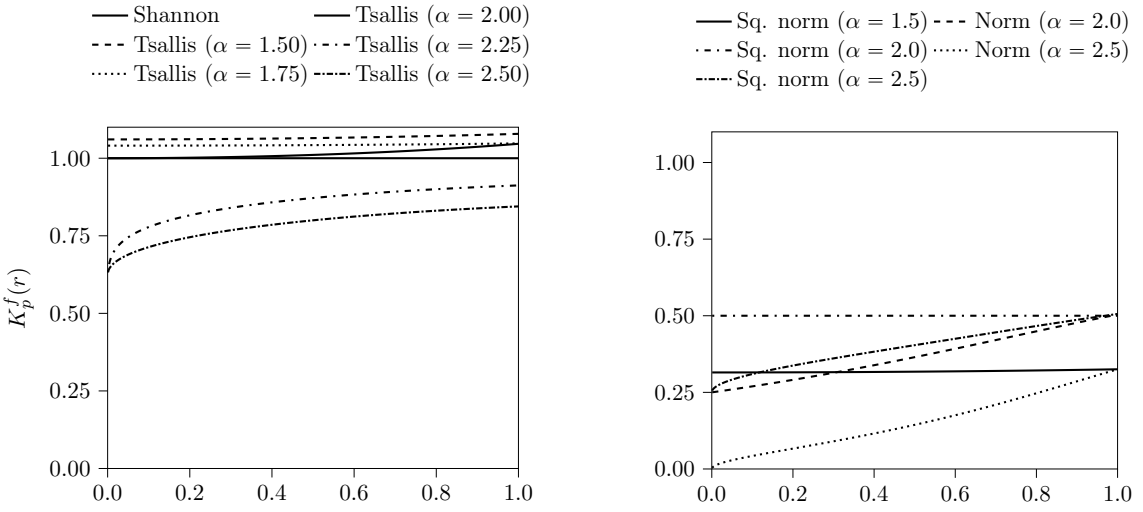

Figure 2: Numerical plots of $K_p^f(r) = 8\omega(r)/r^2$ for each f in Table 1.

6. Examples

We overview a couple of proper losses. Since one can generate a proper loss $\boldsymbol{\ell}$ from a convex function $f = -\underline{L}$ thanks to the Savage representation (4), we show examples in terms of the corresponding convex functions in Table 1. To facilitate closed-form solutions of ω, we restrict ourselves to $N = 2$ and $p = 1$. Table 1 lists several convex functions with their moduli, whose derivations are given previously [Bao23].

Log loss. If we choose

$$f(\mathbf{q}) = \langle \mathbf{q}, \ln \mathbf{q} \rangle = \sum_{n \in [N]} q_n \ln q_n,$$

where ln is applied in the element-wise manner, we can generate the log loss $\ell_y(\mathbf{q}) = -\ln q_y$. In the binary case, we have

$$\ell_1(\mathbf{q}) = -\ln q \quad \text{and} \quad \ell_2(\mathbf{q}) = -\ln(1-q) \quad \text{for } \mathbf{q} = [q \; 1-q]^\top \in \triangle^2.$$

As we saw in §3.3 and §4.2, the log loss is associated with the Kullback–Leibler divergence, and its surrogate regret bound slightly improves Pinsker's inequality yet is asymptotically equivalent.

To see the asymptotic speed of the 1-norm bound (12) for the log loss, we investigate the power evaluation of $\omega^{-1}(\rho) \lesssim \rho^{1/S}$ through Theorem 15, where S is given in Theorem 12. We can see that both (C1) and (C2) are satisfied in this case. Indeed, the continuity of K_p^f is obvious, and

$$\begin{aligned}
\lim_{r \downarrow 0} K_p^f(r) &= 4 \lim_{r \downarrow 0} \frac{(1+\frac{r}{2})\ln(1+\frac{r}{2}) + (1-\frac{r}{2})\ln(1-\frac{r}{2})}{r^2} \\
&= \lim_{r \downarrow 0} \left(\frac{1}{2+r} + \frac{1}{2-r} \right) \\
&= 1 < \infty,
\end{aligned}$$

where L'Hôpital's rule is used twice. See also Fig. 2 for the illustration of K_p^f. Thus, Theorem 15 yields $\limsup_{r \downarrow 0} \sigma(r) \geqslant 2$ and $S \geqslant 2$, which indicates that the polynomial rate of $\omega^{-1}(\rho)$ cannot be faster than $\rho^{1/2}$. The asymptotic lower bound of σ provided here is indeed tight, as we see $\sigma(r) \to 2$ for $r \downarrow 0$ [Bao23, Appendix B.1]. All in all, we asymptotically have the 1-norm bound

$$|q - \widehat{q}| \lesssim \sqrt{D_{\mathrm{KL}}(q \| \widehat{q})} \quad \text{for } q, \widehat{q} \in [0,1],$$

recovering Pinsker's inequality.

Squared norms. Consider the squared α-norms for $\alpha > 1$:

$$f(\mathbf{q}) = \frac{\|\mathbf{q}\|_\alpha^2 - 1}{2} = \frac{1}{2} \left(\sum_{n \in [N]} q_n^\alpha \right)^{\frac{2}{\alpha}} - \frac{1}{2}.$$

By the Savage representation (4), we can generate proper losses:

$$\ell_y(\mathbf{q}) = -\|\mathbf{q}\|_\alpha^{2-\alpha} q_y^{\alpha-1} + \frac{1 + \|\mathbf{q}\|_\alpha^2}{2}.$$

By plugging in $\alpha = 2$, the Brier score in §3.3 is recovered.

To see the asymptotic speed of the p-norm bound (12) for $(N, p) = (2, 1)$, we take the limit of K_p^f similarly to the log-loss case. By using the closed form of $\omega(r)$ provided in

Table 1, when $\alpha \in (1,2)$,

$$\begin{aligned}\lim_{r\downarrow 0} K_p^f(r) &= \lim_{r\downarrow 0} \frac{[(1+\frac{r}{2})^\alpha + (1-\frac{r}{2})^\alpha]^{\frac{2}{\alpha}} - 4^{\frac{1}{\alpha}}}{r^2} \\ &= \frac{1}{4}\lim_{r\downarrow 0}\Big\{(2-\alpha)[(1+r)^\alpha + (1-r)^\alpha]^{\frac{2}{\alpha}-2}[(1+r)^{\alpha-1} - (1-r)^{\alpha-1}]^2 \\ &\qquad +(\alpha-1)[(1+r)^\alpha + (1-r)^\alpha]^{\frac{2}{\alpha}-1}[(1+r)^{\alpha-2} + (1-r)^{\alpha-2}]\Big\} \\ &= (\alpha-1)2^{\frac{2}{\alpha}-2} < \infty,\end{aligned}$$

where L'Hôpital's rule is used twice. When $\alpha \geqslant 2$, we can similarly show the existence of the limit of K_p^f as $r \downarrow 0$. Thus, these losses satisfy both (C1) and (C2) of Theorem 15, which indicates that $\omega^{-1}(\rho)$ cannot be faster than $O(\rho^{1/2})$. Across different $\alpha > 1$, we have $\sigma(r) \to 2$ as $r \downarrow 0$ [Bao23, Figure 4], and thus the provided asymptotic lower bound of σ is tight.

Pseudo-spherical losses. Consider the α-norms for $\alpha > 1$:

$$f(\mathbf{q}) = \frac{\|\mathbf{q}\|_\alpha - 1}{\alpha - 1} = \frac{1}{\alpha-1}\left[\left(\sum_{n\in[N]} q_n^\alpha\right)^{\frac{1}{\alpha}} - 1\right],$$

which has been sometimes used as the α-norm information measure [BvdL80]. By the Savage representation (4), we can generate proper losses:

$$\ell_y(\mathbf{q}) = \frac{1}{\alpha-1}\left(1 - \frac{q_y^{\alpha-1}}{\|\mathbf{q}\|_\alpha^{\alpha-1}}\right),$$

which is called the *pseudo-spherical losses* [Goo71]. By plugging in $\alpha = 2$, the spherical loss is recovered. The associated Bregman divergence is

$$R(\mathbf{q},\widehat{\mathbf{q}}) = \frac{1}{\alpha-1}\left(\|\mathbf{q}\|_\alpha - \frac{\langle \mathbf{q}, \widehat{\mathbf{q}}^{\alpha-1}\rangle}{\|\widehat{\mathbf{q}}\|_\alpha^{\alpha-1}}\right) = \frac{1}{\alpha-1}\left(\|\mathbf{q}\|_\alpha - \frac{1}{\|\widehat{\mathbf{q}}\|_\alpha^{\alpha-1}}\sum_{n\in[N]} q_n \widehat{q}_n^{\alpha-1}\right).$$

Note that $\ln[\|\mathbf{q}\|_\alpha - (\alpha-1)R(\mathbf{q},\widehat{\mathbf{q}})]/(\alpha-1)$ can be identified with the (cross-entropy of) *gamma-divergences*, which is commonly used in robust regression [FE08] and inference with unnormalized models [KF15]. With the limit $\alpha \downarrow 1$, the pseudo-spherical loss approaches the log loss, and the associated Bregman divergence approaches to the Kullback–Leibler divergence, correspondingly.

We illustrate the case $N = 2$ and $\alpha \in (1, 3/2] \cup [2, \infty)$ since the modulus ω can be written analytically herein (shown in Table 1). When $1 < \alpha \leqslant 3/2$, by invoking L'Hôpital's rule

twice,

$$\begin{aligned}
\lim_{r\downarrow 0} K_p^f(r) &= \frac{4}{\alpha-1}\lim_{r\downarrow 0}\frac{\left[\left(1+\frac{r}{2}\right)^\alpha+\left(1-\frac{r}{2}\right)^\alpha\right]^{\frac{1}{\alpha}}-2^{\frac{1}{\alpha}}}{r^2}\\
&= \lim_{r\downarrow 0}\left\{\frac{1}{2}\left[\left(1+\frac{r}{2}\right)^\alpha+\left(1-\frac{r}{2}\right)^\alpha\right]^{\frac{1}{\alpha}-1}\left[\left(1+\frac{r}{2}\right)^{\alpha-2}+\left(1-\frac{r}{2}\right)^{\alpha-2}\right]\right.\\
&\qquad\left.-\frac{1}{\alpha}\left[\left(1+\frac{r}{2}\right)^\alpha+\left(1-\frac{r}{2}\right)^\alpha\right]^{\frac{1}{\alpha}-2}\left[\left(1+\frac{r}{2}\right)^{\alpha-1}-\left(1-\frac{r}{2}\right)^{\alpha-1}\right]\right\}\\
&= 2^{\frac{1}{\alpha}-1}\\
&< \infty.
\end{aligned}$$

When $\alpha \geqslant 2$, by invoking L'Hôpital's rule twice,

$$\begin{aligned}
&\lim_{r\downarrow 0} K_p^f(r)\\
&= \frac{4}{\alpha-1}\lim_{r\downarrow 0}\frac{\left[\left(\frac{r}{2}\right)^\alpha+\left(1-\frac{r}{2}\right)^\alpha\right]^{\frac{1}{\alpha}}-\left[\left(\frac{r}{2}\right)^\alpha+\left(2-\frac{r}{2}\right)^\alpha\right]^{\frac{1}{\alpha}}+1}{r^2}\\
&= \frac{1}{2}\lim_{r\downarrow 0}\left\{-\left[\left(\frac{r}{2}\right)^\alpha+\left(1-\frac{r}{2}\right)^\alpha\right]^{\frac{1-2\alpha}{\alpha}}\left[\left(\frac{r}{2}\right)^{\alpha-1}-\left(1-\frac{r}{2}\right)^{\alpha-1}\right]^2\right.\\
&\qquad+\left[\left(\frac{r}{2}\right)^\alpha+\left(1-\frac{r}{2}\right)^\alpha\right]^{\frac{1-\alpha}{\alpha}}\left[\left(\frac{r}{2}\right)^{\alpha-2}+\left(1-\frac{r}{2}\right)^{\alpha-2}\right]\\
&\qquad+\left[\left(\frac{r}{2}\right)^\alpha+\left(2-\frac{r}{2}\right)^\alpha\right]^{\frac{1-2\alpha}{\alpha}}\left[\left(\frac{r}{2}\right)^{\alpha-1}-\left(2-\frac{r}{2}\right)^{\alpha-1}\right]^2\\
&\qquad\left.-\left[\left(\frac{r}{2}\right)^\alpha+\left(2-\frac{r}{2}\right)^\alpha\right]^{\frac{1-\alpha}{\alpha}}\left[\left(\frac{r}{2}\right)^{\alpha-2}+\left(2-\frac{r}{2}\right)^{\alpha-2}\right]\right\}\\
&= \begin{cases}\frac{1}{4} & \text{if } \alpha=2\\ 0 & \text{if } \alpha>2\end{cases}\\
&< \infty.
\end{aligned}$$

Thus, these losses satisfy (C2) of Theorem 15, which indicates that $\omega^{-1}(\rho)$ cannot be faster than $O(\rho^{1/2})$. Compared with the log loss and the squared norms, the pseudo-spherical losses are more interesting examples for us because $K_p^f(r)$ is no longer always positive. Indeed, (C1) of Theorem 15 is satisfied when $\alpha \in (1, 3/2] \cup \{2\}$ but not satisfied when $\alpha > 2$. The previous $O(\rho^{1/2})$ lower bounds typically require the local strong convexity [FW21] [MMZ24], which is similar to (C2). Therefore, our Theorem 15 slightly lifted the assumptions, requiring (C2) solely. See Fig. 2 to confirm that K_p^f is indeed asymptotically vanishing with the case $\alpha = 2.5$, for which (C1) no longer holds.

Tsallis losses. Consider the negative Tsallis α-entropy

$$f(\mathbf{q}) = \frac{\|\mathbf{q}\|_\alpha^\alpha - 1}{\alpha - 1} = \frac{1}{\alpha-1}\left(\sum_{n\in[N]} q_n^\alpha - 1\right)$$

as a convex potential, for $\alpha > 1$. The Tsallis entropies generalize the Shannon entropy for non-extensive systems [Tsa88], and recovers the Shannon entropy at the limit $\alpha \downarrow 1$. By the Savage representation (4), we can generate proper losses:

$$\ell_y(\mathbf{q}) = -\frac{\alpha q_y^{\alpha-1} - 1}{\alpha - 1} + \|\mathbf{q}\|_\alpha^\alpha,$$

which recovers the log loss at the limit $\alpha \downarrow 1$. We call them the *α-log loss* for convenience. Note that this loss is slightly different from the α-loss $\ell_y(\mathbf{q}) = -\alpha(q_y^{1-1/\alpha} - 1)/(\alpha - 1)$ [SDSK19]. Indeed, the α-log loss is proper by its construction, while the α-loss is known to be improper [SN22]; despite that both of them approach the log loss at the same limit. The associated Bregman divergence to the α-log loss is

$$R(\mathbf{q}, \widehat{\mathbf{q}}) = \frac{\|\mathbf{q}\|_\alpha^\alpha - \alpha \left\langle \mathbf{q}, \widehat{\mathbf{q}}^{\alpha-1} \right\rangle + (\alpha - 1)\|\widehat{\mathbf{q}}\|_\alpha^\alpha}{\alpha - 1},$$

which is the *Tsallis divergence* [Daw07], and also corresponds to *density power divergence* (or the beta-divergence) [BHHJ98] up to constant, used in robust statistics. The Tsallis divergence interpolates the Kullback–Leibler divergence and the squared 2-norm distance at the limits of $\alpha \to 1$ and $\alpha \to 2$, respectively. Some literature opts another definition of the Tsallis divergence, defined by replacing ln in the Kullback–Leibler divergence with the α-logarithmic function [AWS19]—precisely, this another definition should be distinguished as the t-divergence [DQV11].

To see the asymptotic speed of the p-norm bound (12) for $(N, p) = (2, 1)$, we take the limit of K_p^f. When $\alpha \in (1, 2) \cap (3, \infty)$, the explicit form of the modulus ω in Table 1 yields

$$\begin{aligned}
\lim_{r \downarrow 0} K_p^f(r) &= \frac{8}{2^\alpha(\alpha - 1)} \lim_{r \downarrow 0} \frac{(1 + \frac{r}{2})^\alpha + (1 - \frac{r}{2})^\alpha - 2}{r^2} \\
&= \frac{2\alpha}{2^\alpha} \lim_{r \downarrow 0} \frac{(1 + \frac{r}{2})^{\alpha-2} + (1 - \frac{r}{2})^{\alpha-2}}{2} \\
&= \alpha 2^{1-\alpha} < \infty,
\end{aligned}$$

where the L'Hôpital's rule is invoked twice. When $2 \leqslant \alpha \leqslant 3$,

$$\begin{aligned}
\lim_{r \downarrow 0} K_p^f(r) &= \frac{4}{\alpha - 1} \lim_{r \downarrow 0} \frac{\left[\left(\frac{r}{2}\right)^\alpha + \left(1 - \frac{r}{2}\right)^\alpha\right] - 2^{1-\alpha}\left[\left(\frac{r}{2}\right)^\alpha + \left(2 - \frac{r}{2}\right)^\alpha\right] + 1}{r^2} \\
&= \frac{1}{2}\alpha \lim_{r \downarrow 0} \left\{ \left[\left(\frac{r}{2}\right)^{\alpha-2} + \left(1 - \frac{r}{2}\right)^{\alpha-2} \right] - 2^{1-\alpha} \left[\left(\frac{r}{2}\right)^{\alpha-2} + \left(2 - \frac{r}{2}\right)^{\alpha-2} \right] \right\} \\
&= \frac{\alpha}{4} < \infty,
\end{aligned}$$

where the L'Hôpital's rule is invoked twice. In either case, these losses satisfy both (C1) and (C2) of Theorem 15, which indicates that $\omega^{-1}(\rho)$ cannot be faster than $O(\rho^{1/2})$.

Non-differentiable generator. While all the above examples are generated by differentiable convex generator f, our Theorem 15 is applicable even to non-differentiable f, relaxing

the previous lower bounds on surrogate regret bounds [FW21] [MMZ24]. To demonstrate its full capacity, we artificially consider the following non-differentiable convex function:

$$f(\mathbf{q}) = \max_{n\in[N]} \left(q_n - \frac{2}{3}\right)^2 - \frac{1}{9} = \left(\min_{n\in[N]} q_n - \frac{2}{3}\right)^2 - \frac{4}{9},$$

which reduces to

$$f(q) = \begin{cases} \left(\frac{2}{3} - q\right)^2 - \frac{4}{9} & \text{if } q \in [0, \frac{1}{2}] \\ \left(q - \frac{1}{3}\right)^2 - \frac{4}{9} & \text{if } q \in (\frac{1}{2}, 1] \end{cases}$$

for $(N, p) = (2, 1)$. This is non-differentiable at $q = 1/2$. For $(N, p) = (2, 1)$, we can generate the binary loss by Savage representation (4) as follows: for $y = 1$,

$$\ell(q) = \begin{cases} q^2 - 2q + \frac{4}{3} & \text{if } q \in [0, \frac{1}{2}] \\ q^2 - 2q + 1 & \text{if } q \in (\frac{1}{2}, 1] \end{cases},$$

and $\ell(1 - q)$ for $y = 2$, which are discontinuous at $q = 1/2$ yet strictly proper due to the strict convexity of f. For this example, $\omega(r) = r^2/4$ holds for $r \in [0, 1/2]$, and hence

$$\lim_{r\downarrow 0} K_p^f(r) = \lim_{r\downarrow 0} \frac{8\omega(r)}{r^2} = 2 < \infty.$$

Thus, (C2) is satisfied.

When $N \geqslant 3$. Though deriving a closed form of ω for general $N > 2$ is challenging, we can delineate ω for the negative Shannon entropy with $p = 2$: for $r \in (0, 2^{1/2})$, define $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ by

$$\mathbf{q} = \begin{bmatrix} \frac{1 + 2^{-1/2}r}{2} & \frac{1 - 2^{-1/2}r}{2} & 0 & \dots & 0 \end{bmatrix}^\top \text{ and } \breve{\mathbf{q}} = \begin{bmatrix} \frac{1 - 2^{-1/2}r}{2} & \frac{1 + 2^{-1/2}r}{2} & 0 & \dots & 0 \end{bmatrix}^\top.$$

Then, $\|\mathbf{q} - \breve{\mathbf{q}}\|_2 = r$ and $\omega(r) = J(\mathbf{q}, \breve{\mathbf{q}})$. At this minimizer, ω can be written as

$$\omega(r) = \frac{1 + 2^{-1/2}r}{2} \ln \frac{1 + 2^{-1/2}r}{2} + \frac{1 - 2^{-1/2}r}{2} \ln \frac{1 - 2^{-1/2}r}{2} + \ln 2.$$

This ω (for $p = 2$) is akin to the form of ω shown in Table 1, which is for $(N, p) = (2, 1)$, with a slight difference in the scale. Its derivation is based on the method of Lagrange multipliers and deferred to Theorem 32, which is highly non-trivial and interesting in its own right.

To apply Theorem 15 for this example, let us confirm (C2) is satisfied by taking the asymptotic limit of $K_p^f(r)$. By invoking L'Hôpital's rule twice, we have

$$\begin{aligned} \lim_{r\downarrow 0} K_p^f(r) &= 8 \lim_{r\downarrow 0} \frac{\frac{1+2^{-1/2}r}{2} \ln \frac{1+2^{-1/2}r}{2} + \frac{1-2^{-1/2}r}{2} \ln \frac{1-2^{-1/2}r}{2} + \ln 2}{r^2} \\ &= 2 \lim_{r\downarrow 0} \frac{1}{(1 + 2^{-1/2}r)(1 - 2^{-1/2}r)} \\ &= 2 < \infty, \end{aligned}$$

which indicates (C2) is satisfied. Thus, $\omega^{-1}(\rho)$ cannot be faster than $O(\rho^{1/2})$.

7. Conclusion

In this work, we examine surrogate regret bounds on the p-norm, $\|\mathbf{q} - \widehat{\mathbf{q}}\|_p \lesssim \omega^{-1}(R(\mathbf{q}, \widehat{\mathbf{q}}))$, which measures predictive performances of plug-in forecasters under downstream tasks such as classification and ranking. A surrogate regret bound is characterized by the modulus of convexity ω associated with the Bregman generator $f = -\underline{L}$ for a given proper loss. First, we show that the existence of non-vacuous regret bounds is equivalent to the strict properness of losses. Then, we prove that the p-norm upper bound $\omega^{-1}(\rho)$ cannot be faster than the 1/2-order of surrogate regrets $O(\sqrt{\rho})$ for a wide range of strictly proper losses. Herein, the assumptions on loss functions are greatly relaxed so that we do not require the differentiability or the local strong convexity of loss functions anymore. We demonstrate that many loss functions such as the log loss, Brier loss, pseudo-spherical losses, and α-log losses satisfy the assumptions of our optimal-order argument.

As a side note, there is a fundamental relationship between a proper loss and a convex body [Wil14] [WC23]. Specifically, the Bayes risk $\underline{L}$ of a proper loss is the support function of the superprediction set (which is a convex body in $\mathbb{R}^N$) associated with the proper loss. Hence, we can work on a convex body instead of directly working on a proper loss. This perspective has been used to consider aggregating algorithms. In this connection, we studied the modulus of convexity of Bregman generators $f = -\underline{L}$, while the modulus of convexity of Banach spaces has been more commonly studied to measure the set curvature [Fig76]. We conjecture that the modulus of convexity of Bregman generators has a tight connection to the modulus of convexity of superprediction sets, which remains an interesting open question from the viewpoint of convex analysis.

This work is concerned with only *expected* surrogate regrets, induced from the full risk $\mathbb{L}[\widehat{\mathbf{q}}]$ in §3.2. Yet, its empirical estimation and optimization together play an important role in realistic learning scenarios. To take them into account, it is more standard to consider a learner acting on $\mathbb{R}^N$-valued margin, instead of $\triangle^N$-valued prediction as supposed in §3. Proper *composite* losses are common therein, where a link function connects a probabilistic report $\widehat{\mathbf{q}} \in \triangle^N$ to a real-valued report on $\mathbb{R}^N$. Hence, we can consider estimation and optimization of $\mathbb{R}^N$-valued functions. In spite of the scarcity, the estimation error rates of class probability models under the binary case $N = 2$ have been studied rigorously for linear hypotheses [TDS15], while the optimization error rates of proper composite losses by gradient descent under the binary case have been characterized recently [BST25]. We hope to thoroughly understand how a proper loss behaves by integrating surrogate regrets, estimation, and optimization all at once, and leave this for future work.

Acknowledgments

HB and AT are supported by JSPS Grant-in-Aid for Transformative Research Areas(A) (22A201). AT is supported by JSPS Grant-in-Aid for Scientific Research(C) (19K03494).

References

[AA15] 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.

[Aga14] Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(1):1653–1674, 2014.

[AGH+05] Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, Dan Roth, and Michael I. Jordan. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6(4):393–425, 2005.

[AMMZ22a] Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. $\mathcal{H}$-consistency bounds for surrogate loss minimizers. In *Proceedings of the 39th International Conference on Machine Learning*, pages 1117–1174, 2022.

[AMMZ22b] Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-class $\mathcal{H}$-consistency bounds. *Advances in Neural Information Processing Systems*, 35:782–795, 2022.

[AWS19] Ehsan Amid, Manfred K. Warmuth, and Sriram Srinivasan. Two-temperature logistic regression based on the Tsallis divergence. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 2388–2396, 2019.

[Bao22] Han Bao. *Excess Risk Transfer and Learning Problem Reduction towards Reliable Machine Learning*. PhD thesis, University of Tokyo, 2022.

[Bao23] Han Bao. Proper losses, moduli of convexity, and surrogate regret bounds. In *Proceedings of the 36th Conference on Learning Theory*, pages 525–547, 2023.

[BC25] Han Bao and Nontawat Charoenphakdee. Calm composite losses: Being improper yet proper composite. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, pages 2800–2808, 2025.

[BGHN23] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? *Advances in Neural Information Processing Systems*, 36:72071–72095, 2023.

[BGHV09] Jonathan Borwein, Antonio J. Guirao, Petr Hájek, and Jon Vanderwerff. Uniformly convex functions on Banach spaces. *Proceedings of the American Mathematical Society*, 137(3):1081–1091, 2009.

[BHHJ98] Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[BJM06] 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.

[Blo19] Mathieu Blondel. Structured prediction with projection oracles. *Advances in Neural Information Processing Systems*, 32:12145–12156, 2019.

[BMN20] Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with Fenchel–Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

[Bre67] 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.

[Bri50] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[BSS05] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Technical Report*, 2005.

[BSS20] 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, 2020.

[BST25] Han Bao, Shinsaku Sakaue, and Yuki Takezawa. Any-stepsize gradient descent for separable data under Fenchel–Young losses. *arXiv preprint arXiv:2502.04889*, 2025.

[BSX+22] Han Bao, Takuya Shimada, Liyuan Xu, Issei Sato, and Masashi Sugiyama. Pairwise supervision can provably elicit a decision boundary. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 2618–2640, 2022.

[BvdL80] Dick E. Boekee and Jan C. A. van der Lubbe. The R-norm information measure. *Information and Control*, 45(2):136–155, 1980.

[CBFA25] Yuzhou Cao, Han Bao, Lei Feng, and Bo An. Establishing linear surrogate regret bounds for convex smooth losses via convolutional Fenchel–Young losses. *Advances in Neural Information Processing Systems 33*, 2025.

[CLV08] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[CMZ25] Corinna Cortes, Mehryar Mohri, and Yutao Zhong. Improved balanced classification with theoretically grounded loss functions. *Advances in Neural Information Processing Systems 33*, 2025.

[CRR20] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 21(1):3852–3918, 2020.

[Daw07] A. Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

[DKR18] John Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275, 2018.

[DM78] Claude Dellacherie and Paul-André Meyer. *Probabilities and Potential*, volume 29 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam-New York, 1978.

[DQV11] Nan Ding, Yuan Qi, and Svn Vishwanathan. t-divergence based approximate inference. *Advances in Neural Information Processing Systems*, 24:1494–1502, 2011.

[Du21] Hailiang Du. Beyond strictly proper scoring rules: The importance of being local. *Weather and forecasting*, 36(2):457–468, 2021.

[DZDT24] Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. *Advances in Neural Information Processing Systems*, 37:53138–53167, 2024.

[FE08] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

[FHT03] Alexei A. Fedotov, Peter Harremoës, and Flemming Topsøe. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, 2003.

[Fig76] Tadeusz Figiel. On the moduli of convexity and smoothness. *Studia Mathematica*, 56(2):121–155, 1976.

[FW21] Rafael Frongillo and Bo Waggoner. Surrogate regret bounds for polyhedral losses. *Advances in Neural Information Processing Systems*, 34:21569–21580, 2021.

[GB22] Sebastian Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems*, 35:8618–8632, 2022.

[Goo71] I. J. Good. Comment on "measuring information and uncertainty" by robert j. buehler. *Foundations of Statistical Inference*, pages 337–339, 1971.

[GR07] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[HB71] Arlo D. Hendrickson and Robert J. Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42(6):1916–1921, 1971.

[KD16] 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.

[KDH11] Wojciech Kotłowski, Krzysztof Dembczyński, and Eyke Huellermeier. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1113–1120, 2011.

[KF15] Takafumi Kanamori and Hironori Fujisawa. Robust estimation under heavy contamination using unnormalized models. *Biometrika*, 102(3):559–572, 2015.

[KK96] Rangachary Kannan and Carole K. Krueger. *Advanced Analysis on the Real Line*. Universitext. Springer-Verlag, New York, 1996.

[KNRD14] Oluwasanmi O. Koyejo, Nagarajan Natarajan, Pradeep K. Ravikumar, and Inderjit S. Dhillon. Consistent binary classification with generalized performance metrics. *Advances in Neural Information Processing Systems*, 27:2744–2752, 2014.

[KRN65] Kazimierz Kuratowski and Czesław Ryll-Nardzewski. A general theorem on selectors. *Bulletin de l'Académie Polonaise des Sciences. Série des Sciences Mathématiques, Astronomiques et Physiques*, 13:397–403, 1965.

[KSFF17] Meelis Kull, Telmo S. Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 623–631, 2017.

[McC56] John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655, 1956.

[ML21] 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.

[MMZ23a] Anqi Mao, Mehryar Mohri, and Yutao Zhong. $\mathcal{H}$-consistency bounds for pairwise misranking loss surrogates. In *Proceedings of the 40th International Conference on Machine Learning*, pages 23743–23802, 2023.

[MMZ23b] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Structured prediction with stronger consistency guarantees. *Advances in Neural Information Processing Systems*, 36:46903–46937, 2023.

[MMZ24] Anqi Mao, Mehryar Mohri, and Yutao Zhong. A universal growth rate for learning with smooth surrogate losses. *Advances in Neural Information Processing Systems*, 37:41670–41708, 2024.

[MNAC13] Aditya K. 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, 2013.

[MPRS12] Youssef Mroueh, Tomaso Poggio, Lorenzo Rosasco, and Jean-Jeacques Slotine. Multiclass learning with simplex coding. *Advances in Neural Information Processing Systems*, 25:2789–2797, 2012.

[NA13] Harikrishna Narasimhan and Shivani Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. *Advances in Neural Information Processing Systems*, 26:2913–2921, 2013.

[NB11] Frank Nielsen and Sylvain Boltz. The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466, 2011.

[NDRT13] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in Neural Information Processing Systems*, 26:1196–1204, 2013.

[Nes13] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

[NWJ09] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. On surrogate loss functions and f-divergences. *The Annals of Statistics*, 37(2):876–904, 2009.

[OBLJ17] Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. *Advances in Neural Information Processing Systems*, 31:302–313, 2017.

[OFR+19] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32:13991–14002, 2019.

[OT13] Shin-Ichi Ohta and Asuka Takatsu. Displacement convexity of generalized relative entropies. II. *Communications in Analysis and Geometry*, 21(4):687–785, 2013.

[PDL12] Mattew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012.

[Pol66] Boris T. Polyak. Existence theorems and convergence of minimizing sequences for extremal problems with constraints. *Doklady Akademii Nauk*, 166(2):287–290, 1966.

[Roc70] R. Tyrrell Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1970.

[RW09] 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, 2009.

[RW10] Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

[RW11] Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011.

[Sav71] Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[SBTO24] Shinsaku Sakaue, Han Bao, Taira Tsuchiya, and Taihei Oki. Online structured prediction with Fenchel–Young losses and improved surrogate regret for online multiclass classification with logistic loss. In *Proceedings of the 37th Conference on Learning Theory*, pages 4458–4486, 2024.

[SDSK19] Tyler Sypherd, Mario Diaz, Lalitha Sankar, and Peter Kairouz. A tunable loss function for binary classification. In *2019 IEEE International Symposium on Information Theory*, pages 2479–2483. IEEE, 2019.

[Sim64] Igor B. Simonenko. Interpolation and extrapolation of linear operators in Orlicz spaces. *Matematicheskii Sbornik*, 105(4):536–553, 1964.

[SN22] Tyler Sypherd and Richard Nock. Being properly improper. In *Proceedings of the 39th International Conference on Machine Learning*, pages 20891–20932, 2022.

[Ste07] Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

[TDS15] Matus Telgarsky, Miroslav Dudik, and Robert Schapire. Convex risk minimization and conditional probability estimation. In *Proceedings of the 28th Conference on Learning Theory*, pages 1629–1682, 2015.

[Tsa88] Constantino Tsallis. Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

[Vaj70] Igor Vajda. Note on discrimination information and variation. *IEEE Transactions on Information Theory*, 16(6):771–773, 1970.

[WC23] Robert C. Williamson and Zac Cranko. The geometry and calculus of losses. *Journal of Machine Learning Research*, 24(342):1–72, 2023.

[Wil14] Robert C. Williamson. The geometry of losses. In *Proceedings of the 27th Conference on Learning Theory*, pages 1078–1108, 2014.

[WM68] Robert L. Winkler and Allan H. Murphy. "Good" probability assessors. *Journal of Applied Meteorology and Climatology*, 7(5):751–758, 1968.

[WS24] Yutong Wang and Clayton Scott. Unified binary and multiclass margin-based classification. *Journal of Machine Learning Research*, 25(143):1–51, 2024.

[WVR16] Robert C. Williamson, Elodie Vernet, and Mark D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17:1–52, 2016.

[Zha04] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

[ZLA21] 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, 2021.

[ZRA20] Mingyuan Zhang, Harish G. Ramaswamy, and Shivani Agarwal. Convex calibrated surrogates for the multi-label F-measure. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11246–11255, 2020.

Appendix A. Subgradient inequality

In this paper, we adopted a slightly non-conventional definition of subdifferentials ∂f in §2 to allow some elements of the subgradient to be $-\infty$.

Lemma 18 *Let $f : \mathbb{R}^N \to (-\infty, \infty]$ be a proper convex function such that $\triangle^N \subseteq \operatorname{dom} f$. For $\mathbf{q}^0 \in \triangle^N$, the set $\partial f(\mathbf{q}^0)$ is nonempty.*

Proof Fix $\mathbf{q}^0 \in \triangle^N$. If f is subdifferentiable at $\mathbf{q}^0$, then its subgradient belongs to $\partial f(\mathbf{q}^0)$ and the claim holds true. Thus we assume that f is not subdifferentiable at $\mathbf{q}^0$. Since f is subdifferentiable at $\mathbf{q}^0 \in \triangle^N_+$ [Roc70, Theorem 23.4], $I \coloneqq |\operatorname{supp}(\mathbf{q}^0)|$ satisfies $1 \leqslant I \leqslant N-1$. For $\boldsymbol{\eta} \in \mathbb{R}^I$, define $\boldsymbol{\xi}^{\boldsymbol{\eta}} \in \mathbb{R}^N$ by

$$\xi^{\boldsymbol{\eta}}_n \coloneqq \begin{cases} \eta_n & \text{if } n \in \operatorname{supp}(\mathbf{q}^0), \\ 0 & \text{if } n \notin \operatorname{supp}(\mathbf{q}^0), \end{cases}$$

and define a function $f_I : \mathbb{R}^I \to (-\infty, \infty]^N$ by

$$f_I(\boldsymbol{\eta}) \coloneqq f(\boldsymbol{\xi}^{\boldsymbol{\eta}}) \quad \text{for } \boldsymbol{\eta} \in \mathbb{R}^I.$$

Then, f_I is a proper convex function on $\mathbb{R}^I$ such that $\triangle^I \subseteq \operatorname{dom} f_I$, consequently, f_I is subdifferentiable at $\widehat{\boldsymbol{\eta}} \in \triangle^I_+$ [Roc70, Theorem 23.4]. For $\mathbf{q} \in \triangle^N$, define $\boldsymbol{\eta}^{\mathbf{q}} \in \mathbb{R}^I$ by $\eta^{\mathbf{q}}_n = q_n$ for $n \in \operatorname{supp}(\mathbf{q}^0)$. Then, for $\mathbf{q} \in \triangle^N$ with $\operatorname{supp}(\mathbf{q}) \subseteq \operatorname{supp}(\mathbf{q}^0)$, we have $\boldsymbol{\eta}^{\mathbf{q}} \in \triangle^I$ and $f_I(\boldsymbol{\eta}^{\mathbf{q}}) = f(\mathbf{q})$. Moreover, if $\mathbf{q} \in \triangle^N$ satisfies $\operatorname{supp}(\mathbf{q}) = \operatorname{supp}(\mathbf{q}^0)$, then $\boldsymbol{\eta}^{\mathbf{q}} \in \triangle^I_+$ and hence $\partial f_I(\boldsymbol{\eta}^{\mathbf{q}}) \neq \varnothing$. Choose $\mathbf{w} \in \partial f_I(\boldsymbol{\eta}^{\mathbf{q}^0})$ and define $\mathbf{v} \in [-\infty, \infty)^N$ by

$$v_n \coloneqq \begin{cases} w_n & \text{if } n \in \operatorname{supp}(\mathbf{q}^0), \\ -\infty & \text{if } n \notin \operatorname{supp}(\mathbf{q}^0). \end{cases}$$

From now on, we show that $\mathbf{v} \in \partial f(\mathbf{q}^0)$. For $\mathbf{q} \in \triangle^N$, if $q_n > 0$ holds for some $n \notin \operatorname{supp}(\mathbf{q}^0)$, then $\langle \mathbf{v}, \mathbf{q} - \mathbf{q}^0 \rangle = -\infty$ and (2) holds. On the other hand, if $q_n = 0$ for all $n \notin \operatorname{supp}(\mathbf{q}^0)$, then $\mathbf{q} \in \triangle^N$ with $\operatorname{supp}(\mathbf{q}) \subseteq \operatorname{supp}(\mathbf{q}^0)$ and

$$f(\mathbf{q}) = f_I(\boldsymbol{\eta}^{\mathbf{q}}) \geqslant f_I\left(\boldsymbol{\eta}^{\mathbf{q}^0}\right) + \left\langle \mathbf{w}, \boldsymbol{\eta}^{\mathbf{q}} - \boldsymbol{\eta}^{\mathbf{q}^0} \right\rangle = f(\mathbf{q}^0) + \langle \mathbf{v}, \mathbf{q} - \mathbf{q}^0 \rangle,$$

that is, (2) holds for $\mathbf{q} \in \triangle^N$. Thus, $\mathbf{v} \in \partial f(\mathbf{q}^0)$ follows. This completes the proof of the lemma. □

Appendix B. An example of empty $\mathcal{M}(\mathbf{q})$

If a loss $\boldsymbol{\ell}$ is not lower semi-continuous (as in Theorem 1), the set of its minimizers $\mathcal{M}(\mathbf{q})$ (introduced in §3.2) can be empty.

Consider the following example:

$$\ell_y(\widehat{\mathbf{q}}) = \begin{cases} 1 - \widehat{q}_y & \text{if } \widehat{q}_y \neq 1, \\ 1 & \text{if } \widehat{q}_y = 1. \end{cases}$$

Then,

$$L(\mathbf{e}_1, \widehat{\mathbf{q}}) = \ell_1(\widehat{\mathbf{q}}) = \begin{cases} 1 - \widehat{q}_1 & \text{if } \widehat{q}_1 \neq 1, \\ 1 & \text{if } \widehat{q}_1 = 1, \end{cases}$$

where $\mathbf{e}_1 \coloneqq [1, 0, \dots, 0]^\top \in \mathbb{R}^N$. In this case, we have

$$\inf_{\widehat{\mathbf{q}} \in \triangle^N} L(\mathbf{e}_1, \widehat{\mathbf{q}}) = 0,$$

but there does not exist $\widehat{\mathbf{q}} \in \triangle^N$ such that $L(\mathbf{e}_1, \widehat{\mathbf{q}}) = 0$. Thus, $\mathcal{M}(\mathbf{e}_1) = \varnothing$.

Appendix C. Proof of power evaluations without differentiability

In §5, we show power evaluations of the moduli with the differentiability of ω.

Proposition 19 (Power evaluations of moduli) *Let $f : \triangle^N \to \mathbb{R}$ be a strictly convex function. For a fixed $r_0 \in (0, 2^{1/p}]$, define $s, S \in [0, \infty]$ by*

$$s \coloneqq \inf_{r \in (0, r_0]} \sigma(r), \quad S \coloneqq \sup_{r \in (0, r_0]} \sigma(r),$$

and assume $S < \infty$. Then, the function $r \mapsto \omega(r) r^{-s}$ is non-decreasing on $(0, r_0)$ and the function $r \mapsto \omega(r) r^{-S}$ is non-increasing on $(0, r_0)$. Moreover, the following inequalities hold for any $r \in [0, r_0]$:

$$\left[\frac{\omega(r_0)}{r_0^S}\right] r^S \leqslant \omega(r) \leqslant \left[\frac{\omega(r_0)}{r_0^s}\right] r^s.$$

Proof We first show that $r \mapsto \omega(r) r^{-S}$ is non-increasing on $(0, r_0)$ in a similar way to the existing argument [OT13, Lemma 2.9]. For $r \in (0, r_0]$ and $\delta > 0$, there exists $\varepsilon_{r,\delta} > 0$ such that

$$\frac{r}{\omega(r)} \cdot \sup_{\varepsilon \in (0, \varepsilon_{r,\delta})} \frac{\omega(r) - \omega(r - \varepsilon)}{\varepsilon} \leqslant \sigma(r) + \frac{1}{2}\delta \tag{22}$$

by the definition of $D^- \omega$. Define

$$g(t) \coloneqq S + \frac{1}{2}\delta + \frac{1}{t}[(1-t)^{S+\delta} - 1] \quad \text{for } t \in (0, 1).$$

Then, g is continuous on $(0, 1)$ and

$$\lim_{t \downarrow 0} g(t) = S + \frac{1}{2}\delta - (S + \delta) = -\frac{1}{2}\delta < 0,$$

which implies the existence of $\tau \in (0,1)$ such that $g(t) < 0$ for $t \in (0,\tau)$. Then, for any $u \in (0,\varepsilon_{r,\delta}) \cap (0, r\tau)$,

$$\begin{aligned}
-\frac{r^{S+\delta}}{\omega(r)} + \frac{(r-u)^{S+\delta}}{\omega(r-u)} &= \frac{ur^{S+\delta-1}}{\omega(r-u)}\left[\frac{r}{\omega(r)}\frac{\omega(r)-\omega(r-u)}{u}\right] - \frac{r^{S+\delta}}{\omega(r-u)} + \frac{r^{S+\delta}(1-\frac{u}{r})^{S+\delta}}{\omega(r-u)} \\
&\leqslant \frac{ur^{S+\delta-1}}{\omega(r-u)}\left[\sigma(r)+\frac{1}{2}\delta\right] + \frac{ur^{S+\delta-1}}{\omega(r-u)} \cdot \frac{1}{\frac{u}{r}}\left[\left(1-\frac{u}{r}\right)^{S+\delta} - 1\right] \\
&\leqslant \frac{ur^{S+\delta-1}}{\omega(r-u)}\left[S+\frac{1}{2}\delta + g\left(\frac{u}{r}\right) - \left(S+\frac{1}{2}\delta\right)\right] \\
&= \frac{ur^{S+\delta-1}}{\omega(r-u)} \cdot g\left(\frac{u}{r}\right) \\
&< 0,
\end{aligned}$$

where the inequality (22) is used at the second line and the third line follows from the definition of S. Hence, we have

$$\frac{(r-u)^{S+\delta}}{\omega(r-u)} < \frac{r^{S+\delta}}{\omega(r)} \quad \text{for } r \in (0,r_0) \text{ and } u \in (0,\varepsilon_{r,\delta}) \cap (0,r\tau).$$

Letting $\delta \downarrow 0$, we conclude that $r \mapsto r^S/\omega(r)$ is non-decreasing on $(0,r_0)$. This is equivalent to that $r \mapsto \omega(r)r^{-S}$ is non-increasing on $(0,r_0)$.

Next, we show that $r \mapsto \omega(r)r^{-s}$ is non-decreasing on $(0,r_0)$. For $r \in (0,r_0]$, let $u \in (0,r)$. Since $\ln\omega$ is non-decreasing on $[r-u,r]$, it follows from the fundamental theorem of calculus [KK96, Theorem 1.3.1] that

$$\int_{r-u}^{r} D^-\ln\omega(r')\mathrm{d}r' \leqslant \ln\frac{\omega(r)}{\omega(r-u)}.$$

By $D^-\omega(r') \leqslant S < \infty$, we have

$$\lim_{\varepsilon\downarrow 0}\omega(r'-\varepsilon) = \omega(r') \quad \text{for } r' \in (0,r_0),$$

which yields

$$\lim_{\varepsilon\downarrow 0}\frac{\ln\omega(r')-\ln\omega(r'-\varepsilon)}{\omega(r')-\omega(r'-\varepsilon)} = \frac{1}{\omega(r')}$$

and

$$D^-\ln\omega(r') = \limsup_{\varepsilon\downarrow 0}\left[\frac{\ln\omega(r')-\ln\omega(r'-\varepsilon)}{\omega(r')-\omega(r'-\varepsilon)} \cdot \frac{\omega(r')-\omega(r'-\varepsilon)}{\varepsilon}\right] = \frac{1}{\omega(r')}D^-\omega(r') \geqslant \frac{s}{r'}.$$

Thus, we have

$$\int_{r-u}^{r} D^-\ln\omega(r')\mathrm{d}r' \geqslant \int_{r-u}^{r}\frac{s}{r'}\mathrm{d}r' = s\ln\frac{r}{r-u}.$$

These imply

$$\omega(r-u)\cdot(r-u)^{-s} \leqslant \omega(r)\cdot r^{-s},$$

that is, $r \mapsto \omega(r)r^{-s}$ is non-decreasing on $(0,r_0)$. □

Appendix D. Continuity property of local modulus of convexity

The local modulus of convexity defined in Theorem 14 naturally entails the following continuity property. We state it in the following lemma for the sake of completeness.

Lemma 20 *If $f : \triangle^N \to \mathbb{R}$ is continuous convex, then $K_p^f : (0, 2^{1/p}] \to \mathbb{R}$ is lower semi-continuous and left-continuous.*

To prove Theorem 20, we use the following supplement result.

Lemma 21 *If $f : \triangle^N \to \mathbb{R}$ is continuous convex, then for $r \in (0, 2^{1/p}]$ and $\tau \in (0, 1/2)$, we have*

$$\frac{K_p^f((1-2\tau)r)}{8}[(1-2\tau)r]^2 = \omega((1-2\tau)r) \leqslant \omega(r) - \frac{\kappa_p^f}{2}\tau(1-\tau)r^2 \leqslant \frac{K_p^f(r)}{8}r^2 - \frac{\kappa_p^f}{2}\tau(1-\tau)r^2.$$

Proof Choose distinct $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ satisfying

$$\|\mathbf{q} - \breve{\mathbf{q}}\|_p = r \quad \text{and} \quad \omega(r) = J(\mathbf{q}, \breve{\mathbf{q}}),$$

which exist thanks to Theorem 9 together with the convexity of f. For these $\mathbf{q}, \breve{\mathbf{q}}$, define

$$c(t) \coloneqq (1-t)\mathbf{q} + t\breve{\mathbf{q}} \quad \text{for } t \in [0,1].$$

For $\tau \in (0, 1/2)$, we have

$$\begin{aligned}
\frac{K_p^f((1-2\tau)r)}{8}[(1-2\tau)r]^2 &= \omega((1-2\tau)r) \\
&\leqslant J(c(\tau), c(1-\tau)) \\
&\leqslant J(c(0), c(1)) - \frac{\kappa_p^f}{2}\tau(1-\tau)r^2 \\
&= \omega(r) - \frac{\kappa_p^f}{2}\tau(1-\tau)r^2 \\
&\leqslant \frac{K_p^f(r)}{8}r^2 - \frac{\kappa_p^f}{2}\tau(1-\tau)r^2.
\end{aligned}$$

□

Proof of Theorem 20 Fix $r \in (0, 2^{1/p}]$. Let $(r_j)_{j\in\mathbb{N}} \subseteq (0, 2^{1/p}]$ be a sequence converging to r. For each $j \in \mathbb{N}$, there exist $\mathbf{q}^j, \breve{\mathbf{q}}^j \in \triangle^N$ satisfying

$$\|\mathbf{q}^j - \breve{\mathbf{q}}^j\|_p = r_j \quad \text{and} \quad \omega(r_j) = J(\mathbf{q}^j, \breve{\mathbf{q}}^j)$$

from Theorem 9. Define $c_j : [0,1] \to \triangle^N$ by

$$c_j(t) \coloneqq (1-t)\mathbf{q}^j + t\breve{\mathbf{q}}^j \quad \text{for } t \in [0,1].$$

By the Arzelá–Ascoli theorem, we can extract a subsequence $(c_{j_m})_{m\in\mathbb{N}}$ converging uniformly to some $c : [0,1] \to \triangle^N$ uniformly, where

$$c(t) = (1-t)c(0) + tc(1) \quad \text{for } t \in [0,1] \quad \text{and} \quad \|c(0) - c(1)\|_p = r$$

hold. Since f is continuous, we have

$$K_p^f(r) = \frac{8}{r^2}\omega(r) \leqslant \frac{8}{r^2}J(c(0),c(1)) \\ = \lim_{m\to\infty}\frac{8}{r_{j_m}^2}J(c_{j_m}(0),c_{j_m}(1)) = \lim_{m\to\infty}\frac{8}{r_{j_m}^2}\omega(r_{j_m}) = \lim_{m\to\infty}K_p^f(r_{j_m}).$$

Thus, $K_p^f : (0, 2^{1/p}] \to \mathbb{R}$ is lower semi-continuous.

Next, by Theorem 21, we have

$$\frac{K_p^f((1-2\tau)r)}{8}[(1-2\tau)r]^2 \leqslant \frac{K_p^f(r)}{8}r^2 - \frac{\kappa_p^f}{2}\tau(1-\tau)r^2$$

for $\tau \in (0, 1/2)$. Dividing by $[(1-2\tau)r]^2/8 \neq 0$ and then taking the limit yields

$$\limsup_{\tau\downarrow 0} K_p^f((1-2\tau)r) \leqslant K_p^f(r).$$

Together with the lower semi-continuity K_p^f, the left-continuity K_p^f is ensured.

This completes the proof of the lemma. □

Appendix E. Deferred proofs

Lemma 22 *Suppose $\boldsymbol{\ell} : \triangle^N \to [0,\infty]^N$ is lower semi-continuous. Then, $\underline{L}(\mathbf{q}) \geqslant 0$ holds and $\mathcal{M}(\mathbf{q})$ is nonempty and closed for $\mathbf{q} \in \triangle^N$. Moreover, if $\boldsymbol{\ell}$ is continuous, then there exists a Borel selector of $\mathcal{M}$.*

Proof For $\mathbf{q} \in \triangle^N$, the lower semi-continuity of $L(\mathbf{q},\cdot)$ follows from that of $\boldsymbol{\ell}$, which ensures the closedness of $\mathcal{M}(\mathbf{q})$. Moreover, by the extreme value theorem with the closedness of $\triangle^N$, we see $\mathcal{M}(\mathbf{q}) \neq \varnothing$ holds.

Assume the continuity of $\boldsymbol{\ell}$ and we show the existence of a Borel selector of $\mathcal{M}$. Note that the continuity of ℓ guarantees the continuity of $L(\mathbf{q},\cdot)$ on $\triangle^N$ for each $\mathbf{q} \in \triangle^N$. By the Kuratowski and Ryll-Nardzewski measurable selection theorem [KRN65, Main Theorem & Corollary 1], it is enough to show that

$$\mathcal{B}_{\mathcal{K}} \coloneqq \left\{\mathbf{q} \in \triangle^N \;\middle|\; \mathcal{M}(\mathbf{q}) \cap \mathcal{K} \neq \varnothing\right\}$$

is Borel for any closed set $\mathcal{K}$ in $\mathbb{R}^N$ with $\triangle^N \cap \mathcal{K} \neq \varnothing$.

Fix a compact set $\mathcal{K}$ in $\mathbb{R}^N$ with $\triangle^N \cap \mathcal{K} \neq \varnothing$. Since $\triangle^N \cap \mathcal{K}$ is a compact metric space hence separable, there exists a dense countable set $\left\{\mathbf{q}^j\right\}_{j\in\mathbb{N}}$ in $\triangle^N \cap \mathcal{K}$. For each $j \in \mathbb{N}$, define $d_j : \triangle^N \to \mathbb{R}$ by

$$d_j(\mathbf{q}) \coloneqq L(\mathbf{q},\mathbf{q}^j) - \inf_{\widehat{\mathbf{q}}\in\triangle^N} L(\mathbf{q},\widehat{\mathbf{q}}) \quad \text{for } \mathbf{q} \in \triangle^N,$$

which is lower semi-continuous, in particular, Borel on $\triangle^N$. Then,

$$\mathcal{B} \coloneqq \bigcap_{m\in\mathbb{N}}\bigcup_{j\in\mathbb{N}} d_j^{-1}([0,m^{-1}))$$

is Borel. We will show $\mathcal{B} = \mathcal{B}_{\mathcal{K}}$. For $\mathbf{q} \in \mathcal{B}_{\mathcal{K}}$, there exists $\widehat{\mathbf{q}} \in \mathcal{M}(\mathbf{q}) \cap \mathcal{K}$. By the continuity of $L(\mathbf{q}, \cdot)$ on $\triangle^N$, for each $m \in \mathbb{N}$, there exists $\delta_m > 0$ such that if $\mathbf{q}' \in \triangle^N$ satisfies $\|\mathbf{q}' - \widehat{\mathbf{q}}\|_2 < \delta_m$, then $0 \leqslant L(\mathbf{q}, \mathbf{q}') - L(\mathbf{q}, \widehat{\mathbf{q}}) < m^{-1}$. By the density of $\left\{\mathbf{q}^j\right\}_{j \in \mathbb{N}}$, there exists $j_m \in \mathbb{N}$ such that $\|\mathbf{q}^{j_m} - \widehat{\mathbf{q}}\|_2 < \delta_m$ and hence

$$d_{j_m}(\mathbf{q}) = L(\mathbf{q}, \mathbf{q}^{j_m}) - L(\mathbf{q}, \widehat{\mathbf{q}}) \in [0, m^{-1}),$$

which in turn implies $\mathbf{q} \in \mathcal{B}$. Conversely, for $\mathbf{q} \in \mathcal{B}$ and $m \in \mathbb{N}$, there exists $j_m \in \mathbb{N}$ such that

$$L(\mathbf{q}, \mathbf{q}^{j_m}) < \inf_{\widehat{\mathbf{q}} \in \triangle^N} L(\mathbf{q}, \widehat{\mathbf{q}}) + m^{-1}.$$

We extract a convergent subsequence of $(\mathbf{q}^{j_m})_{m \in \mathbb{N}}$ (not relabeled) with limit $\widehat{\mathbf{q}} \in \triangle^N \cap \mathcal{K}$. The continuity of $L(\mathbf{q}, \cdot)$ gives

$$L(\mathbf{q}, \widehat{\mathbf{q}}) = \lim_{m \to \infty} L(\mathbf{q}, \mathbf{q}^{j_m}) \leqslant \inf_{\widehat{\mathbf{q}}' \in \triangle^N} L(\mathbf{q}, \widehat{\mathbf{q}}'),$$

proving $\widehat{\mathbf{q}} \in \mathcal{M}(\mathbf{q})$ hence $\mathbf{q} \in \mathcal{B}_{\mathcal{K}}$. This completes the proof of the lemma. □

Proposition 23 (Uniqueness up to affine functions) *Let $f, g : \triangle^N \to \mathbb{R}$ be continuous convex functions. Then, their midpoint Jensen gaps are the same if and only if $f - g$ is affine.*

Proof We only show that $f - g$ is affine under the assumption that the midpoint Jensen gaps of f and g are the same since the converse implication is trivial. Hereafter, let us write the midpoint Jensen gaps of f and g by J_f and J_g, respectively.

Without loss of generality, we pick $\mathbf{q}^0 \in \triangle^N_+$ such that f and g are differentiable at $\mathbf{q}^0$ because a convex function is differentiable almost everywhere in the interior of its domain. Define

$$\mathbf{v}^0 \coloneqq \nabla f(\mathbf{q}^0) - \nabla g(\mathbf{q}^0) \quad \text{and} \quad \lambda \coloneqq f(\mathbf{q}^0) - g(\mathbf{q}^0).$$

Fix any $\mathbf{q} \in \triangle^N$ and set

$$h(t) \coloneqq f(\mathbf{q}^0 + t(\mathbf{q} - \mathbf{q}^0)) - g(\mathbf{q}^0 + t(\mathbf{q} - \mathbf{q}^0)) - \left\langle \mathbf{v}^0, t(\mathbf{q} - \mathbf{q}^0) \right\rangle - \lambda \quad \text{for } t \in [0, 1].$$

Then, $h : [0, 1] \to \mathbb{R}$ is continuous with $h(0) = 0$ and $h'(0) = 0$. With elementary algebra, we have

$$0 = J_f(\mathbf{q}^0, \mathbf{q}^0 + t(\mathbf{q} - \mathbf{q}^0)) - J_g(\mathbf{q}^0, \mathbf{q}^0 + t(\mathbf{q} - \mathbf{q}^0)) = \frac{1}{2} h(t) - h\left(\frac{t}{2}\right) \quad \text{for all } t \in [0, 1],$$

which implies

$$h\left(\frac{1}{2} t\right) = \frac{1}{2} h(t) \quad \text{for all } t \in [0, 1].$$

By invoking this relation recursively, we have $h(t) = 2^k h(2^{-k} t)$ for any $k \in \mathbb{N}$, which yields

$$h(1) = \lim_{k \to \infty} \frac{h(2^{-k}) - h(0)}{2^{-k}} = h'(0) = 0.$$

Consequently, we have

$$f(\mathbf{q}) = g(\mathbf{q}) + \left\langle \mathbf{v}^0, \mathbf{q} - \mathbf{q}^0 \right\rangle + \lambda.$$

Thus, we have shown that $f - g$ is affine. □

Lemma 24 *Let $f : \triangle^N \to \mathbb{R}$ be a convex function. For $r \in [0, 2^{1/p}]$, there exist $\mathbf{q}^r, \breve{\mathbf{q}}^r \in \triangle^N$ such that $\omega(r) = J(\mathbf{q}^r, \breve{\mathbf{q}}^r)$ and $\|\mathbf{q}^r - \breve{\mathbf{q}}^r\|_p = r$.*

Proof Let $r \in [0, 2^{1/p}]$. Define

$$\mathcal{D}^N(r) \coloneqq \left\{(\mathbf{q}, \breve{\mathbf{q}}) \in \triangle^N \times \triangle^N \;\middle|\; \|\mathbf{q} - \hat{\mathbf{q}}\|_p \geqslant r\right\}.$$

Since $\mathcal{D}^N(r)$ is compact and J is continuous on $\mathcal{D}^N(r)$, there is $(\mathbf{q}, \breve{\mathbf{q}}) \in \mathcal{D}^N(r)$ such that $\omega(r) = J(\mathbf{q}, \breve{\mathbf{q}})$. Define $c : [0,1] \to \triangle^N$ by

$$c(t) \coloneqq (1-t)\mathbf{q} + t\breve{\mathbf{q}} \quad \text{for } t \in [0,1].$$

In the case of $\|\mathbf{q} - \breve{\mathbf{q}}\|_p = r$, we can take $(\mathbf{q}^r, \breve{\mathbf{q}}^r) = (\mathbf{q}, \breve{\mathbf{q}})$, and the statement follows. Assume $\|\mathbf{q} - \breve{\mathbf{q}}\|_p > r$. Then, there exists $\tau \in (0, 1/2]$ such that

$$\|c(\tau) - c(1-\tau)\|_p = (1-2\tau)\|\mathbf{q} - \breve{\mathbf{q}}\|_p = r.$$

Since $f \circ c : [0,1] \to \mathbb{R}$ is convex, we have

$$\frac{f(c(\tau)) - f(c(0))}{\tau} \leqslant \frac{f(c(1)) - f(c(1-\tau))}{\tau}, \tag{23}$$

which is equivalent to

$$\begin{aligned} J(c(\tau), c(1-\tau)) &= \frac{f(c(\tau)) + f(c(1-\tau)}{2} - f(c(1/2)) \\ &\leqslant \frac{f(c(0)) + f(c(1)}{2} - f(c(1/2)) = J(\mathbf{q}, \breve{\mathbf{q}}). \end{aligned}$$

This yields $J(c(\tau), c(1-\tau)) = \omega(r)$, and hence, we can take $(\mathbf{q}^r, \breve{\mathbf{q}}^r) = (c(\tau), c(1-\tau))$. Thus, we have confirmed the statement. □

Proposition 25 (Strong convexity parameter at midpoint) *For a continuous convex function $f : \triangle^N \to \mathbb{R}$, it holds $\kappa_p^f = \kappa_p^{f,1/2}$.*

Proof By definition, $\kappa_p^f \leqslant \kappa_p^{f,1/2}$ trivially holds. We shall prove the converse inequality. Observe from the definition that

$$f\left(\frac{\mathbf{q} + \breve{\mathbf{q}}}{2}\right) \leqslant \frac{1}{2}f(\mathbf{q}) + \frac{1}{2}f(\breve{\mathbf{q}}) - \frac{\kappa_p^{f,\frac{1}{2}}}{8}\|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 \quad \text{for all } \mathbf{q}, \breve{\mathbf{q}} \in \triangle^N.$$

For $(i,j) \in \mathbb{N} \times \mathbb{Z}_{\geqslant 0}$, set

$$t_{i,j} \coloneqq 2^{-i}j.$$

Fix distinct $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ and define $c : [0,1] \to \triangle^N$ by

$$c(t) \coloneqq (1-t)\mathbf{q} + t\breve{\mathbf{q}} \quad \text{for } t \in [0,1].$$

We will show that

$$f(c(t_{i,j})) \leqslant (1-t_{i,j})f(c(0)) + t_{i,j}f(c(1)) - \frac{\kappa_p^{f,\frac{1}{2}}}{2}t_{i,j}(1-t_{i,j})\|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 \tag{24}$$

for $(i,j) \in \mathbb{N} \times \mathbb{Z}_{\geqslant 0}$ with $t_{i,j} \in [0,1]$ (namely, for $i \in \mathbb{N}$ and $0 \leqslant j \leqslant 2^i$) by induction on i. We immediately observe that (24) always holds for $t_{i,0} = 0$ and $t_{i,2^i} = 1$ regardless of i.

The inequality (24) trivially holds for $i = 1$. Assume that (24) holds for some $i \in \mathbb{N}$ and all j with $0 \leqslant j \leqslant 2^i$. Then, (24) also holds for $t_{i+1,2j} = t_{i,j}$ with $0 \leqslant j \leqslant 2^i$. For $0 \leqslant j \leqslant 2^i - 1$, we have

$$t_{i+1,2j+1} = \frac{t_{i+1,2j} + t_{i+1,2j+2}}{2} = \frac{t_{i,j} + t_{i,j+1}}{2}.$$

Define $c_{i,j} : [0,1] \to \triangle^N$ by

$$c_{i,j}(t) \coloneqq c((1-t) \cdot t_{i,j} + t \cdot t_{i,j+1}) \quad \text{for } t \in [0,1].$$

This implies

$$\begin{aligned}
&f(c(t_{i+1,2j+1})) \\
&\leqslant \frac{1}{2} f(c_{i,j}(0)) + \frac{1}{2} f(c_{i,j}(1)) - \frac{\kappa^{f,\frac{1}{2}}}{8} \|c_{i,j}(0) - c_{i,j}(1)\|_p^2 \\
&= \frac{1}{2} f(c_{i,j}(0)) + \frac{1}{2} f(c_{i,j}(1)) - \frac{\kappa_p^{f,\frac{1}{2}}}{8} \cdot 2^{-2i} \cdot \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 \\
&= \frac{1}{2} f(c(t_{i,j})) + \frac{1}{2} f(c(t_{i,j+1})) - \frac{\kappa_p^{f,\frac{1}{2}}}{8} \cdot 2^{-2i} \cdot \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 \\
&\leqslant \frac{1}{2}\left[(1-t_{i,j}) f(c(0)) + t_{i,j} f(c(1)) - \frac{\kappa_p^{f,\frac{1}{2}}}{2} t_{i,j}(1-t_{i,j}) \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2\right] \\
&\quad + \frac{1}{2}\left[(1-t_{i,j+1}) f(c(0)) + t_{i,j+1} f(c(1)) - \frac{\kappa_p^{f,\frac{1}{2}}}{2} t_{i,j+1}(1-t_{i,j+1}) \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2\right] \\
&\quad - \frac{\kappa_p^{f,\frac{1}{2}}}{8} \cdot 2^{-2i} \cdot \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 \\
&= (1 - t_{i+1,2j+1}) f(c(0)) + t_{i+1,2j+1} f(c(1)) - \frac{\kappa_p^{f,\frac{1}{2}}}{2} t_{i+1,2j+1}(1 - t_{i+1,2j+1}) \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2
\end{aligned}$$

as desired.

Because f is continuous on $\triangle^N$ and $\{t_{i,j} \in [0,1] \mid (i,j) \in \mathbb{N} \times \mathbb{Z}_{\geqslant 0}\}$ is dense in $[0,1]$, we find

$$f((1-t)\mathbf{q} + t\breve{\mathbf{q}}) \leqslant (1-t) f(\mathbf{q}) + t f(\breve{\mathbf{q}}) - \frac{\kappa_p^{f,\frac{1}{2}}}{2} t(1-t) \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 \quad \text{for } t \in (0,1),$$

which leads to

$$\kappa_p^{f,\frac{1}{2}} \leqslant \frac{2[(1-t) f(\mathbf{q}) + t f(\breve{\mathbf{q}}) - f((1-t)\mathbf{q} + t\breve{\mathbf{q}})]}{t(1-t)\|\mathbf{q} - \breve{\mathbf{q}}\|_p^2} \quad \text{for } t \in (0,1).$$

Since $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ are arbitrary, this ensures that $\kappa_p^{f,t} \geqslant \kappa_p^{f,1/2}$ for $t \in (0,1)$ and completes the proof of the proposition. □

Lemma 26 *Let $f : \triangle^N \to \mathbb{R}$ be a continuous convex function. Then, for any $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ and $r \in (0, 2^{1/p}]$,*

$$\liminf_{r \downarrow 0} K_p^f(r) = \kappa_p^f, \quad J(\mathbf{q}, \breve{\mathbf{q}}) \geqslant \frac{\kappa_p^f}{8} \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2, \quad \textit{and} \quad D^- \omega(r) \geqslant \frac{\kappa_p^f}{4} r.$$

Proof Assume that there exists $r_* \in (0, 2^{1/p}]$ such that $K_p^f(r_*) = \kappa_p^f$. We can see from Theorem 21 that

$$\frac{K_p^f((1-2\tau)r_*)}{8}[(1-2\tau)r_*]^2 \leqslant \frac{K_p^f(r_*)}{8} r_*^2 - \frac{\kappa_p^f}{2}\tau(1-\tau)r_*^2 = \frac{K_p^f(r_*)}{8}[(1-2\tau)r_*]^2$$

for $\tau \in (0, 1/2)$. Since $K_p^f(r) \geqslant \kappa_p^f$ for $r \in (0, 2^{1/p}]$ (by Theorem 14 and Eq. (18)), this implies $K_p^f(r) = \kappa_p^f$ for $r \in (0, r_*]$ hence

$$\liminf_{r \downarrow 0} K_p^f(r) = \kappa_p^f.$$

Assume that there is no $r \in (0, 2^{1/p}]$ so that $K_p^f(r) = \kappa_p^f$. Then, there exists $(r_j)_{j \in \mathbb{N}} \subseteq (0, 2^{1/p}]$ converging to 0 such that

$$\lim_{j \to \infty} K_p^f(r_j) = \inf \left\{ K_p^f(r) \;\middle|\; r \in (0, 2^{\frac{1}{p}}] \right\} = \kappa_p^{f, \frac{1}{2}} \leqslant \liminf_{r \downarrow 0} K_p^f(r) \leqslant \lim_{j \to \infty} K_p^f(r_j),$$

where the second equality follows from (18). This with Theorem 13 proves the first assertion.

For $\mathbf{q}, \breve{\mathbf{q}} \in \triangle$, we calculate

$$J(\mathbf{q}, \breve{\mathbf{q}}) \geqslant \omega(\|\mathbf{q} - \breve{\mathbf{q}}\|_p) = \frac{K_p^f(\|\mathbf{q} - \breve{\mathbf{q}}\|_p)}{8} \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2 \geqslant \frac{\kappa_p^f}{8} \|\mathbf{q} - \breve{\mathbf{q}}\|_p^2.$$

This is the second assertion.

For $r \in (0, 2^{1/p}]$, we can pick $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ such that $\omega(r) = J(\mathbf{q}, \breve{\mathbf{q}})$ and $\|\mathbf{q} - \breve{\mathbf{q}}\|_p = r$ from Theorem 9. For $\tau \in (0, 1/2)$, we have

$$\omega((1-2\tau)r) \leqslant \omega(r) - \frac{\kappa_p^f}{2}\tau(1-\tau)r^2$$

from Theorem 21, which yields

$$D^- \omega(r) = \limsup_{\tau \downarrow 0} \frac{\omega(r) - \omega((1-2\tau)r)}{r - (1-2\tau)r} \geqslant \limsup_{\tau \downarrow 0} \frac{\frac{\kappa_p^f}{2}\tau(1-\tau)r^2}{2\tau r} = \frac{\kappa_p^f}{4} r.$$

This completes the proof of the lemma. □

Appendix F. Derivation of modulus for general N

In §6, we mainly consider examples of ω only for $(N, p) = (2, 1)$. Since we have extended the moduli on general $(\triangle^N, \|\cdot\|_p)$ in Theorem 7, it is nice to have an example beyond the binary case. To this end, we calculate ω for the negative Shannon entropy $f(\mathbf{q}) = \langle \mathbf{q}, \ln \mathbf{q} \rangle$ with

general $N \geqslant 2$ and $p = 2$. In what follows, we focus on $f(\mathbf{q}) = \langle \mathbf{q}, \ln \mathbf{q} \rangle$, and the modulus ω and midpoint Jensen gap J is defined based on this particular f throughout this section.

Let

$$\mathcal{U}^{N-1} := \left\{ \mathbf{u} \in (0,1)^{N-1} \;\middle|\; \sum_{n \in [N-1]} u_n < 1 \right\}$$

and define $\psi : \mathcal{U}^{N-1} \times \mathcal{U}^{N-1} \to \mathbb{R}$ and $\phi : \mathcal{U}^{N-1} \to \mathbb{R}$ by

$$\psi(\mathbf{u}, \mathbf{w}) := \frac{1}{2} \left\{ \sum_{n \in [N-1]} (u_n - w_n)^2 + \left[\sum_{n \in [N-1]} (u_n - w_n) \right]^2 \right\} \qquad \text{for } \mathbf{u}, \mathbf{w} \in \mathcal{U}^{N-1},$$

$$\phi(\mathbf{u}) := \sum_{n \in [N-1]} u_n \ln u_n + \left(1 - \sum_{n \in [N-1]} u_n \right) \ln \left(1 - \sum_{n \in [N-1]} u_n \right) \quad \text{for } \mathbf{u} \in \mathcal{U}^{N-1},$$

respectively. Here, ϕ is the negative Shannon entropy of $[\mathbf{u} \;\; 1 - \langle \mathbf{u}, \mathbf{1} \rangle] \in \triangle^N$. Indeed, we have

$$\begin{bmatrix} u_1 \\ \vdots \\ u_{N-1} \\ 1 - \langle \mathbf{u}, \mathbf{1} \rangle \end{bmatrix} \in \triangle_+^N \quad \text{and} \quad \psi(\mathbf{u}, \mathbf{w}) = \frac{1}{2} \left\| \begin{bmatrix} u_1 \\ \vdots \\ u_{N-1} \\ 1 - \langle \mathbf{u}, \mathbf{1} \rangle \end{bmatrix} - \begin{bmatrix} w_1 \\ \vdots \\ w_{N-1} \\ 1 - \langle \mathbf{w}, \mathbf{1} \rangle \end{bmatrix} \right\|_2^2,$$

which yields $\psi(\mathbf{u}, \mathbf{w}) \in [0, 1]$. First, we present a couple of necessary lemmas.

Lemma 27 *For $r \in (0, 2^{1/2})$, there exist $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ such that*

$$\|\mathbf{q} - \breve{\mathbf{q}}\|_2 = r, \quad \operatorname{supp}(\mathbf{q}) \cap \operatorname{supp}(\breve{\mathbf{q}}) \neq \varnothing, \quad J(\mathbf{q}, \breve{\mathbf{q}}) < \ln 2.$$

Proof Fix $r \in (0, 2^{1/2})$ and set $a := 2^{-1/2} r \in (0, 1)$. Define $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ by

$$q_n := \delta_{1n}, \quad \breve{q}_n := (1 - a)\delta_{1n} + a\delta_{2n} \quad \text{for } n \in [N].$$

Then, we have $\|\mathbf{q} - \breve{\mathbf{q}}\|_2 = r$, $\operatorname{supp}(\mathbf{q}) \cap \operatorname{supp}(\breve{\mathbf{q}}) \neq \varnothing$, and

$$J(\mathbf{q}, \breve{\mathbf{q}}) = \frac{1 - a}{2} \ln(1 - a) - \left(1 - \frac{a}{2} \right) \ln \left(1 - \frac{a}{2} \right) + \frac{a}{2} \ln 2 =: \overline{J}(a).$$

Since we have

$$\overline{J}(1) = \ln 2, \quad \overline{J}'(a) = \frac{1}{2} \ln \frac{2 - a}{1 - a} > 0 \quad \text{for } a < 1,$$

we conclude $J(\mathbf{q}, \breve{\mathbf{q}}) < \ln 2$ as desired. □

Lemma 28 *For $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$, if*

$$\operatorname{supp}(\mathbf{q}) \cap \operatorname{supp}(\breve{\mathbf{q}}) \neq \varnothing, \quad \operatorname{supp}(\mathbf{q}) \neq \operatorname{supp}(\breve{\mathbf{q}}),$$

then there exist $\mathbf{q}', \breve{\mathbf{q}}' \in \triangle^N$ such that

$$\|\mathbf{q} - \mathbf{q}\|_2 = \|\mathbf{q}' - \breve{\mathbf{q}}'\|_2, \quad \operatorname{supp}(\mathbf{q}') = \operatorname{supp}(\breve{\mathbf{q}}'), \quad J(\mathbf{q}', \breve{\mathbf{q}}') < J(\mathbf{q}, \breve{\mathbf{q}}).$$

Proof Take $i \in \operatorname{supp}(\mathbf{q}) \cap \operatorname{supp}(\breve{\mathbf{q}})$ and write

$$S_1 \coloneqq [\operatorname{supp}(\mathbf{q}) \cap \operatorname{supp}(\breve{\mathbf{q}})] \setminus \{i\}, \quad S_2 \coloneqq \operatorname{supp}(\mathbf{q}) \setminus \operatorname{supp}(\breve{\mathbf{q}}), \quad S_3 \coloneqq \operatorname{supp}(\breve{\mathbf{q}}) \setminus \operatorname{supp}(\mathbf{q}),$$

and $m \coloneqq |S_2 \cup S_3|$. For sufficiently small $\varepsilon > 0$, define $\mathbf{q}^\varepsilon, \breve{\mathbf{q}}^\varepsilon \in \triangle^N$ by

$$q_n^\varepsilon \coloneqq \begin{cases} q_i - m\varepsilon & \text{for } n = i, \\ q_n & \text{for } n \in S_1, \\ q_n + \varepsilon & \text{for } n \in S_2 \cup S_3, \\ 0 & \text{otherwise,} \end{cases} \qquad \breve{q}_n^\varepsilon \coloneqq \begin{cases} \breve{q}_i - m\varepsilon & \text{for } n = i, \\ \breve{q}_n & \text{for } n \in S_1, \\ \breve{q}_n + \varepsilon & \text{for } n \in S_2 \cup S_3, \\ 0 & \text{otherwise.} \end{cases}$$

We find that $\|\mathbf{q}^\varepsilon - \breve{\mathbf{q}}^\varepsilon\|_2 = \|\mathbf{q} - \breve{\mathbf{q}}\|_2$, $\operatorname{supp}(\mathbf{q}^\varepsilon) = \operatorname{supp}(\breve{\mathbf{q}}^\varepsilon)$, and

$$\lim_{\varepsilon \downarrow 0} J(\mathbf{q}^\varepsilon, \breve{\mathbf{q}}^\varepsilon) = J(\mathbf{q}, \breve{\mathbf{q}}),$$

thanks to the continuity of J. Since we have

$$\frac{\partial}{\partial \varepsilon} J(\mathbf{q}^\varepsilon, \breve{\mathbf{q}}^\varepsilon) = -\frac{m}{2} \ln \frac{(q_i - m\varepsilon)(\breve{q}_i - m\varepsilon)}{\left(\frac{q_i + \breve{q}_i}{2} - m\varepsilon\right)^2} + \frac{1}{2} \sum_{n \in S_2} \ln \frac{(q_n + \varepsilon)\varepsilon}{\left(\frac{q_n}{2} + \varepsilon\right)^2} + \frac{1}{2} \sum_{n \in S_3} \ln \frac{(\breve{q}_n + \varepsilon)\varepsilon}{\left(\frac{\breve{q}_n}{2} + \varepsilon\right)^2},$$

which diverges to $-\infty$ as $\varepsilon \downarrow 0$, we have $J(\mathbf{q}^\varepsilon, \breve{\mathbf{q}}^\varepsilon) < J(\mathbf{q}, \breve{\mathbf{q}})$ for sufficiently small $\varepsilon > 0$. This completes the proof of the lemma. □

Corollary 29 *For $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ with $\|\mathbf{q} - \breve{\mathbf{q}}\|_2 < 2^{1/2}$, if $\omega(\|\mathbf{q} - \breve{\mathbf{q}}\|_2) = J(\mathbf{q}, \breve{\mathbf{q}})$ holds then $\operatorname{supp}(\mathbf{q}) = \operatorname{supp}(\breve{\mathbf{q}})$.*

Proof Let $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ satisfy $\|\mathbf{q} - \breve{\mathbf{q}}\|_2 < 2^{1/2}$. By Theorem 27, we have $\omega(\|\mathbf{q} - \breve{\mathbf{q}}\|_2) < \ln 2$. It is easy to see if $\operatorname{supp}(\mathbf{q}') \cap \operatorname{supp}(\breve{\mathbf{q}}') = \varnothing$, then $J(\mathbf{q}', \breve{\mathbf{q}}') = \ln 2$ holds. This with Theorem 28 leads to $\omega(\|\mathbf{q} - \breve{\mathbf{q}}\|_2) = J(\mathbf{q}, \breve{\mathbf{q}})$ implies $\operatorname{supp}(\mathbf{q}) = \operatorname{supp}(\breve{\mathbf{q}})$. Thus, the proof of the corollary is complete. □

Subsequently, we show Theorems 30 and 31, which are needed to invoke the method of Lagrangian multipliers later.

Lemma 30 *For $\mathbf{u}, \mathbf{w} \in \mathcal{U}^{N-1}$, the rank of the Jacobian of ψ at $(\mathbf{u}, \mathbf{w})$ is zero if and only if $\mathbf{u} = \mathbf{w}$.*

Proof Since we have

$$\frac{\partial \psi}{\partial u_i}(\mathbf{u}, \mathbf{w}) = u_i - w_i + \sum_{n \in [N-1]} (u_n - w_n) = -\frac{\partial \psi}{\partial w_i} \quad \text{for } i \in [N-1],$$

the rank of the Jacobian of ψ at $(\mathbf{u}, \mathbf{w})$ is zero if and only if

$$u_i - w_i + \sum_{n \in [N-1]} (u_n - w_n) = 0 \quad \text{for } i \in [N-1].$$

Summing the above equation up gives

$$N \sum_{n \in [N-1]} (u_n - w_n) = 0,$$

and hence $u_i - w_i = 0$ for all $i \in [N-1]$, which shows $\mathbf{u} = \mathbf{w}$. The converse implication is trivial. □

Lemma 31 *For $r \in (0, 2^{1/2})$ with $r^2 < N/(N-1)$, let $\mathbf{u}, \mathbf{w} \in \mathcal{U}^{N-1}$ satisfy $\psi(\mathbf{u}, \mathbf{w}) = r^2/2$ and set*

$$u_N \coloneqq 1 - \sum_{n \in [N-1]} u_n, \quad w_N \coloneqq 1 - \sum_{n \in [N-1]} w_n.$$

If there exists $\lambda \in \mathbb{R}$ such that

$$\begin{aligned} &\frac{1}{2}\frac{\partial \phi}{\partial u_n}(\mathbf{u}) - \frac{\partial \phi}{\partial u_n}\left(\frac{\mathbf{u}+\mathbf{w}}{2}\right) + \lambda \frac{\partial \psi}{\partial u_n}(\mathbf{u}, \mathbf{w}) = 0 \quad \textit{and} \\ &\frac{1}{2}\frac{\partial \phi}{\partial w_n}(\mathbf{u}) - \frac{\partial \phi}{\partial w_n}\left(\frac{\mathbf{u}+\mathbf{w}}{2}\right) + \lambda \frac{\partial \psi}{\partial w_n}(\mathbf{u}, \mathbf{w}) = 0 \end{aligned} \tag{25}$$

for $n \in [N-1]$, then there exists $\mathcal{I} \subseteq [N]$ with $I \coloneqq |\mathcal{I}|$ such that $0 < I(N-I) < r^{-2}N$ and

$$\begin{aligned} &u_n = \frac{1 + r\sqrt{a_N(I)}}{2I}, \quad w_n = \frac{1 - r\sqrt{a_N(I)}}{2I} \quad \textit{for } n \in \mathcal{I}, \\ &u_n = \frac{1 - r\sqrt{a_N(I)}}{2(N-I)}, \quad w_n = \frac{1 + r\sqrt{a_N(I)}}{2(N-I)} \quad \textit{for } n \in [N] \backslash \mathcal{I}, \quad \textit{where } a_N(I) \coloneqq \frac{I(N-I)}{N}. \end{aligned}$$

Proof For simplicity, set

$$\mathbf{v} \coloneqq \frac{\mathbf{u}+\mathbf{w}}{2}, \quad \text{and} \quad v_N \coloneqq 1 - \sum_{n \in [N-1]} v_n.$$

Assuming Eq. (25), we calculate

$$\begin{aligned} 0 &= \frac{1}{2}\frac{\partial \phi}{\partial u_n}(\mathbf{u}) - \frac{\partial \phi}{\partial u_n}(\mathbf{v}) + \lambda \frac{\partial \psi}{\partial u_n}(\mathbf{u}, \mathbf{w}) = \frac{1}{2}\left(\ln \frac{u_n}{u_N} - \ln \frac{v_n}{v_N}\right) + \lambda[u_n - w_n - (u_N - w_N)], \\ 0 &= \frac{1}{2}\frac{\partial \phi}{\partial w_n}(\mathbf{w}) - \frac{\partial \phi}{\partial w_n}(\mathbf{v}) + \lambda \frac{\partial \psi}{\partial w_n}(\mathbf{u}, \mathbf{w}) = \frac{1}{2}\left(\ln \frac{w_n}{w_N} - \ln \frac{v_n}{v_N}\right) - \lambda[u_n - w_n - (u_N - w_N)], \end{aligned}$$

for $n \in [N-1]$, which yields

$$\ln \frac{u_n}{u_N} - \ln \frac{v_n}{v_N} = -2\lambda[u_n - w_n - (u_N - w_N)] = -\ln \frac{w_n}{w_N} + \ln \frac{v_n}{v_N} \tag{26}$$

Thus, we have

$$\frac{(u_n + w_n)^2}{4u_n w_n} = \frac{v_n^2}{u_n w_n} = \frac{v_N^2}{u_N w_N} =: \Lambda > 1 \quad \text{for } n \in [N-1],$$

where $\Lambda > 1$ holds; otherwise $u_n = w_n$ holds for all $n \in [N]$, which contradicts $\psi(\mathbf{u}, \mathbf{w}) = r^2/2$. By rearranging in w_n, we have

$$w_n^2 - 2u_n(2\Lambda - 1)w_n + u_n^2 = 0 \quad \text{for } n \in [N].$$

Setting

$$\mu \coloneqq (2\Lambda - 1) - 2\sqrt{\Lambda^2 - \Lambda} \in (0, 1),$$

we have either $w_n = \mu u_n$ or $w_n = \mu^{-1} u_n$ for $n \in [N]$, by solving the quadratic equation with respect to w_n. Set

$$\mathcal{I} \coloneqq \{i \in [N] \mid w_i = \mu u_i\}, \quad I \coloneqq |\mathcal{I}|, \qquad \mathcal{J} \coloneqq \{j \in [N] \mid w_j = \mu^{-1} u_j\}, \quad J \coloneqq |\mathcal{J}|.$$

Then $I + J = N$ holds.

Assume $N \in \mathcal{I}$. If $\mathcal{J} = \varnothing$, then

$$\begin{aligned} w_N = 1 - \sum_{i \in [N-1]} w_i &= 1 - \mu \sum_{i \in [N-1]} u_i \\ &\neq \mu - \mu \sum_{i \in [N-1]} u_i = \mu \left(1 - \sum_{i \in [N-1]} u_i \right) = \mu u_N = w_N, \end{aligned}$$

which is a contradiction. Thus $\mathcal{J} \neq \varnothing$ holds. On one hand, for $j \in \mathcal{J}$, we observe from Eq. (26) that

$$\ln \frac{1+\mu}{1+\mu^{-1}} = \ln \frac{u_j}{u_N} - \ln \frac{v_j}{v_N} = -2\lambda[u_j - w_j - (u_N - w_N)] = 2\lambda(1-\mu)(\mu^{-1} u_j + u_N).$$

Since the left-hand side is independent of j and not zero, then so is the right-hand side hence u_j is determined independent of j and $\lambda \neq 0$. On the other hand, for $i \in \mathcal{I}$, it turns out that

$$\ln \frac{u_i}{u_N} - \ln \frac{v_i}{v_N} = 0$$

and consequently $u_i = u_N$ holds by Eq. (26) together with the property $\lambda \neq 0$. This yields $u_i = u_N$ for $i \in \mathcal{I}$. Thus there exist $s, t \in (0, 1)$ such that

$$u_i = s, \ w_i = \mu s \quad \text{for } i \in \mathcal{I}, \quad u_j = t, \ w_j = \mu^{-1} t \quad \text{for } j \in \mathcal{J}. \tag{27}$$

We see that

$$1 = \sum_{n \in [N]} u_N = Is + Jt, \quad 1 = \sum_{n \in [N]} w_N = \mu I s + \mu^{-1} J t,$$

that is, $Jt = 1 - Is = \mu(1 - \mu I s)$. This can be simplified as follows:

$$\mu = \frac{1}{Is} - 1 = \frac{Jt}{Is}.$$

We also find that

$$r^2 = 2\psi(\mathbf{u}, \mathbf{w}) = (1-\mu)^2 \left(I s^2 + \mu^{-2} J t^2 \right) = (1-\mu)^2 \left(I s^2 + \frac{I^2}{J} s^2 \right) = \frac{I}{J} N (1-\mu)^2 s^2,$$

which in turn shows

$$r \sqrt{\frac{J}{NI}} = (1-\mu)s = 2s - \frac{1}{I}.$$

We conclude

$$s = \frac{1}{2I} \left(1 + r \sqrt{\frac{IJ}{N}} \right), \quad t = \frac{1}{2J} \left(1 - r \sqrt{\frac{IJ}{N}} \right)$$

as desired.

The case $N \in \mathcal{J}$ is proved by switching the role of $\mathbf{u}$ and $\mathbf{w}$ in the argument for the case $N \in \mathcal{I}$, and the proof is achieved. □

By combining these lemmas, we have the following claim, which is the minimizers $(\mathbf{q}, \breve{\mathbf{q}})$ for the negative Shannon entropy we show in §6.

Proposition 32 *For $r \in (0, 2^{1/2})$ with $r^2 < N/(N-1)$, $\mathbf{q}, \breve{\mathbf{q}} \in \triangle^N$ satisfy $\|\mathbf{q} - \breve{\mathbf{q}}\|_2 = r$ and $\omega(r) = J(\mathbf{q}, \breve{\mathbf{q}})$ if and only if there exist distinct $i, j \in [N]$ such that*

$$q_n \coloneqq \frac{1 + 2^{-1/2} r}{2} \delta_{ni} + \frac{1 - 2^{-1/2} r}{2} \delta_{nj}, \quad \breve{q}_n \coloneqq \frac{1 - 2^{-1/2} r}{2} \delta_{ni} + \frac{1 + 2^{-1/2} r}{2} \delta_{nj}.$$

In this case,

$$\omega(r) = J(\mathbf{q}, \breve{\mathbf{q}}) = \frac{1}{2} \left[\left(1 + \frac{r}{\sqrt{2}}\right) \ln \left(1 + \frac{r}{\sqrt{2}}\right) + \left(1 - \frac{r}{\sqrt{2}}\right) \ln \left(1 - \frac{r}{\sqrt{2}}\right) \right].$$

Proof There exist $\mathbf{q}', \breve{\mathbf{q}}' \in \triangle^N$ such that

$$\|\mathbf{q}' - \breve{\mathbf{q}}'\|_2 = r, \quad \omega(r) = J(\mathbf{q}, \breve{\mathbf{q}}), \quad \text{and} \quad \operatorname{supp}(\mathbf{q}') = \operatorname{supp}(\breve{\mathbf{q}}')$$

from Theorem 9 and Theorem 29. By relabeling the indices $n \in [N]$ and switching $\mathbf{q}'$ and $\breve{\mathbf{q}}'$ if necessary, we may assume that and there exists $N' \in [N]$ such that $N' \geqslant 2$ with

$$\operatorname{supp}(\mathbf{q}') = \operatorname{supp}(\breve{\mathbf{q}}') = [N'],$$

and $I \in [N' - 1]$ with $I \leqslant N'/2$ such that $0 < I(N' - I) < r^{-2} N'$ with

$$\begin{aligned} q'_n &= \frac{1 + r\sqrt{a_{N'}(I)}}{2I}, \quad \breve{q}'_n = \frac{1 - r\sqrt{a_{N'}(I)}}{2I} \quad \text{for } n \in [I], \\ q'_n &= \frac{1 - r\sqrt{a_{N'}(I)}}{2(N' - I)}, \quad \breve{q}'_n = \frac{1 + r\sqrt{a_{N'}(I)}}{2(N' - I)} \quad \text{for } n \in [N']\backslash[I], \end{aligned} \tag{28}$$

by Theorem 31, where

$$a_{N'}(x) \coloneqq \frac{x(N' - x)}{N'}.$$

We need to identify I and N' hereafter. Note that for any $N' \in [N]$, we have

$$1 \cdot \left(1 - \frac{1}{N'}\right) \leqslant 1 \cdot \left(1 - \frac{1}{N}\right) \leqslant r^{-2}$$

hence

$$I' \coloneqq \max \left\{ i \in [N' - 1] \;\middle|\; i(N' - i) < r^{-2} N', i \leqslant N'/2 \right\}$$

is well-defined.

Setting

$$\overline{J}_{N'}(x) \coloneqq \left[1 + r\sqrt{a_{N'}(x)}\right] \ln \left[1 + r\sqrt{a_{N'}(x)}\right] + \left[1 - r\sqrt{a_{N'}(x)}\right] \ln \left[1 - r\sqrt{a_{N'}(x)}\right]$$

for $1 \leqslant x \leqslant N' - 1$, we see that

$$J(\mathbf{q}', \breve{\mathbf{q}}') = \frac{1}{2} \overline{J}_{N'}(I).$$

From this with the relation

$$\frac{\mathrm{d}}{\mathrm{d}x} \overline{J}_{N'}(x) = \frac{r}{2\sqrt{a_{N'}(x)}} \left(1 - \frac{2x}{N'}\right) \ln \frac{1 + r\sqrt{a_{N'}(x)}}{1 - r\sqrt{a_{N'}(x)}} \geqslant 0 \quad \text{for } x \in [I'],$$

we observe

$$\min_{I \in [I']} \overline{J}_{N'}(I) = \overline{J}_{N'}(1) = \overline{J}(N')$$

where

$$\overline{J}(y) \coloneqq \left(1 + r\sqrt{1 - y^{-1}}\right) \ln \left(1 + r\sqrt{1 - y^{-1}}\right) + \left(1 - r\sqrt{1 - y^{-1}}\right) \ln \left(1 - r\sqrt{1 - y^{-1}}\right)$$

for $y \geqslant 2$. We calculate

$$\frac{\mathrm{d}}{\mathrm{d}y} \overline{J}(y) = \frac{r y^{-2}}{2\sqrt{1 - y^{-1}}} \ln \frac{1 + r\sqrt{1 - y^{-1}}}{1 - r\sqrt{1 - y^{-1}}} > 0 \quad \text{for } y \geqslant 2,$$

which leads to

$$\min_{N' \in [N], N' \geqslant 2} \overline{J}(N') = \overline{J}(2).$$

Thus, in Eq. (28), the correct choice is $(N', I) = (2, 1)$ and this completes the proof of the proposition. □