Learning Classifiers with Fenchel-Young Losses: Generalized Entropies, Margins, and Algorithms

Mathieu Blondel[1], André F. T. Martins[2], and Vlad Niculae[3]

[1] *NTT Communication Science Laboratories, Kyoto, Japan*
mathieu@mblondel.org

[2] *Unbabel & Instituto de Telecomunicações, Lisbon, Portugal*
andre.martins@unbabel.com

[3] *Cornell University, Ithaca, NY*
vlad@vene.ro

Abstract

We study in this paper Fenchel-Young losses, a generic way to *construct* convex loss functions from a convex regularizer. We provide an in-depth study of their properties in a broad setting and show that they unify many well-known loss functions. When constructed from a generalized entropy, which includes well-known entropies such as Shannon and Tsallis entropies, we show that Fenchel-Young losses induce a predictive probability distribution and develop an efficient algorithm to compute that distribution for separable entropies. We derive conditions for generalized entropies to yield a distribution with sparse support and losses with a separation margin. Finally, we present both primal and dual algorithms to learn predictive models with generic Fenchel-Young losses.

1 Introduction

Loss functions are a cornerstone of statistics and machine learning: They measure the difference, or "loss," between a ground-truth label and a prediction. For this reason, much work has been devoted to designing loss functions [17, 26, 37, 23] or to studying their theoretical properties [53, 30, 57]. Some loss functions, such as the hinge loss of support vector machines, are intimately connected to the notion of separation margin—a prevalent concept in statistical learning theory, which has been used to prove the famous perceptron mistake bound [47] and many other generalization bounds [55, 49]. For probabilistic classification, proper scoring rules [29, 28] provide a generic way to construct loss functions; however, they expect predictions to lie in the probability simplex and hence typically need to be composed with an invertible link function [57], which can render them non-convex [14]. Loss functions are often tightly connected to the statistical model they learn. For instance, the framework of generalized linear models [40] provides a wealth of loss functions for learning conditional probabilistic models in the exponential family [3]. A well-known instance of this framework is the logistic loss, and its associated "softmax distribution." The logistic loss has many appealing properties, notably, its well-known relationship with the principle of maximum entropy. Yet, the logistic loss lacks a separation margin. Martins & Astudillo [37] were the first to explicitly seek a **sparse distribution** and to derive an associated loss. This "sparsemax" loss is differentiable and, simultaneously, has a natural separation margin. Despite much progress, a general understanding of how to construct differentiable convex losses (potentially with a margin) is still lacking.

Towards this goal, this paper studies and extends Fenchel-Young losses, recently proposed for structured prediction [43]. We show that Fenchel-Young losses provide a generic and principled way to **construct** a loss function with an associated predictive probability distribution. We further show that there is a tight and fundamental relation between generalized entropies, margins, and sparse probability distributions. In summary, we make the following contributions.

- We provide an in-depth study of Fenchel-Young losses, deriving their properties and showing that they unify many well-known loss functions, including the logistic and sparsemax losses (§2).
- When based on a generalized entropy, which includes well-known instances such as the Shannon, Tsallis, and Rényi entropies, we show that Fenchel-Young losses induce a predictive probability distribution. We extend this view to a mean parametrization, useful for structured prediction (§3).

- We derive conditions for generalized entropies to yield distributions with sparse support and losses with a separation margin. For separable entropies, we provide a simple formula that relates the margin with analytic properties of the entropy function (§4).
- Finally, we show that, despite their generality, Fenchel-Young losses can be used to learn (generalized) linear models very simply, using both primal and dual algorithms. For separable entropies, we show that computing their associated probability distribution reduces to unidimensional root finding, thus allowing efficient computation of a large class of probability distributions (§5).

Notation. We denote the $(d-1)$-dimensional probability simplex by $\triangle^d \coloneqq \{\boldsymbol{p} \in \mathbb{R}^d_+ \colon \|\boldsymbol{p}\|_1 = 1\}$. We denote the convex hull of a set $\mathcal{Y} \subseteq \mathbb{R}^d$ by $\mathsf{conv}(\mathcal{Y})$. We denote the domain of a convex function $\Omega \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ by $\mathsf{dom}(\Omega) \coloneqq \{\boldsymbol{\mu} \in \mathbb{R}^d \colon \Omega(\boldsymbol{\mu}) < \infty\}$. The (Fenchel) conjugate of Ω is defined by $\Omega^*(\boldsymbol{\theta}) \coloneqq \sup_{\boldsymbol{\mu} \in \mathsf{dom}(\Omega)} \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - \Omega(\boldsymbol{\mu})$.

2 Fenchel-Young loss functions

The Fenchel-Young family of loss functions was recently proposed for structured prediction [43]. We extend it to a broader setting: a general learning problem with input variables $\boldsymbol{x} \in \mathcal{X}$, ground-truth $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^d$, and parametrized model $\boldsymbol{f}_W : \mathcal{X} \to \mathbb{R}^d$ which produces a score vector $\boldsymbol{\theta} \coloneqq \boldsymbol{f}_W(\boldsymbol{x})$. Let Ω be a lower-semicontinuous (l.s.c.) proper convex regularizer, with $\mathsf{dom}(\Omega) \subseteq \mathbb{R}^d$ a closed set. We consider **regularized prediction functions** [42, 39] of the kind

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \partial\Omega^*(\boldsymbol{\theta}) = \operatorname*{\mathsf{argmax}}_{\boldsymbol{\mu} \in \mathsf{dom}(\Omega)} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \Omega(\boldsymbol{\mu}). \tag{1}$$

A typical choice is $\mathsf{dom}(\Omega) = \mathsf{conv}(\mathcal{Y})$, implying that $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is the mean prediction under some underlying distribution, as described in §3. For this choice of $\mathsf{dom}(\Omega)$, $\widehat{\boldsymbol{y}}_0(\boldsymbol{\theta}) \in \mathsf{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$, recovering the class with highest score. In this broad scenario, Fenchel-Young losses arise as a natural choice to learn the model parameters W (cf. §5 for primal and dual formulations).

Definition 1 *Let Ω be a l.s.c. proper convex function, with $\mathsf{dom}(\Omega) \subseteq \mathbb{R}^d$ a closed set. Let $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathsf{dom}(\Omega)$ be a ground-truth label and $\boldsymbol{\theta} \in \mathsf{dom}(\Omega^*) = \mathbb{R}^d$ be a vector of prediction scores. The* **Fenchel-Young loss** *$L_\Omega \colon \mathsf{dom}(\Omega^*) \times \mathcal{Y} \to \mathbb{R}_+$ generated by Ω is*

$$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \coloneqq \Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle.$$

As their name indicates, these losses are grounded in the **Fenchel-Young inequality** [9, Proposition 3.3.4]

$$\Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{\mu}) \geq \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle \quad \forall \boldsymbol{\theta} \in \mathsf{dom}(\Omega^*), \boldsymbol{\mu} \in \mathsf{dom}(\Omega). \tag{2}$$

Under the assumption on Ω in Definition 1, Danskin's theorem [18] implies that (2) becomes an equality (i.e., the duality gap is zero) if and only if $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{\mu})$. This implies the following properties.

Proposition 1 *Properties of Fenchel-Young losses*

1. **Non-negativity.** *$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \geq 0$ for any $\boldsymbol{\theta} \in \mathsf{dom}(\Omega^*)$ and $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathsf{dom}(\Omega)$.*
2. **Zero loss.** *The loss is zero iff $\boldsymbol{y} \in \partial\Omega^*(\boldsymbol{\theta})$, i.e., iff $\boldsymbol{y} \in \mathsf{argmax}_{\boldsymbol{\mu} \in \mathsf{dom}(\Omega)} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \Omega(\boldsymbol{\mu})$.*
3. **Convexity & gradient.** *L_Ω is convex and the residual vectors are its subgradients: $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y} \in \partial L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$. If Ω is strictly convex, then L_Ω is differentiable and $\nabla L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}$.*
4. **Constant invariance.** *For any constant $c \in \mathbb{R}$, $L_{\Omega + c}(\boldsymbol{\theta}; \boldsymbol{y}) = L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$.*
5. **Temperature scaling.** *For any $t > 0$, $L_{t\Omega}(\boldsymbol{\theta}; \boldsymbol{y}) = t L_\Omega(\frac{\boldsymbol{\theta}}{t}; \boldsymbol{y})$ and $\widehat{\boldsymbol{y}}_{t\Omega}(\boldsymbol{\theta}) \in \partial\Omega^*(\frac{\boldsymbol{\theta}}{t})$.*

The properties above suggest that the minimization of Fenchel-Young losses attempt to adjust the model to produce predictions $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ that are close to the target $\boldsymbol{y}$, reducing the duality gap.

Table 2 shows examples of well-known loss functions that fall into this category: simple choices of Ω yield the squared loss, perceptron loss, logistic loss, sparsemax loss, as well as their extensions to structured classification; see §3.2 for details. More losses can also be recovered, as discussed next.

Table 1: Examples of Fenchel-Young losses. For multi-class classification, we assume $\mathcal{Y} = \{\boldsymbol{e}_i\}_{i=1}^d$ and the ground-truth is $\boldsymbol{y} = \boldsymbol{e}_k$, where $\boldsymbol{e}_i$ denotes a standard basis ("one-hot") vector. For structured classification, we assume that elements of $\mathcal{Y}$ are d-dimensional binary vectors with $d \ll |\mathcal{Y}|$, and we denote by $\mathsf{conv}(\mathcal{Y}) = \{\mathbb{E}_{\boldsymbol{p}}[Y] \colon \boldsymbol{p} \in \triangle^{|\mathcal{Y}|}\}$ the corresponding marginal polytope [56]. We denote by $\mathsf{H}^{\mathrm{S}}(\boldsymbol{p}) \coloneqq -\sum_i p_i \log p_i$ the Shannon entropy of a distribution $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$.

	$\mathsf{dom}(\Omega)$	$\Omega(\boldsymbol{\mu})$	$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$	$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$
Squared loss	$\mathbb{R}^{\lvert\mathcal{Y}\rvert}$	$\frac{1}{2}\lVert\boldsymbol{\mu}\rVert^2$	$\boldsymbol{\theta}$	$\frac{1}{2}\lVert\boldsymbol{y} - \boldsymbol{\theta}\rVert^2$
Perceptron loss [47]	$\triangle^{\lvert\mathcal{Y}\rvert}$	0	$\mathsf{argmax}(\boldsymbol{\theta})$	$\mathsf{max}_i\, \theta_i - \theta_k$
Logistic loss	$\triangle^{\lvert\mathcal{Y}\rvert}$	$-\mathsf{H}^{\mathrm{S}}(\boldsymbol{\mu})$	$\mathsf{softmax}(\boldsymbol{\theta})$	$\log \sum_i \exp \theta_i - \theta_k$
Sparsemax loss [37]	$\triangle^{\lvert\mathcal{Y}\rvert}$	$\frac{1}{2}\lVert\boldsymbol{\mu}\rVert^2$	$\mathsf{sparsemax}(\boldsymbol{\theta})$	$\frac{1}{2}\lVert\boldsymbol{y} - \boldsymbol{\theta}\rVert^2 - \frac{1}{2}\lVert\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{\theta}\rVert^2$
Structured perceptron [15]	$\mathsf{conv}(\mathcal{Y})$	0	$\mathsf{MAP}(\boldsymbol{\theta})$	$\mathsf{max}_{\boldsymbol{y}'}\, \langle \boldsymbol{\theta}, \boldsymbol{y}' \rangle - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$
CRF [35]	$\mathsf{conv}(\mathcal{Y})$	$\min_{\mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}} -\mathsf{H}^{\mathrm{S}}(\boldsymbol{p})$	$\mathsf{marginals}(\boldsymbol{\theta})$	$\log \sum_{\boldsymbol{y}'} \exp \langle \boldsymbol{\theta}, \boldsymbol{y}' \rangle - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$
SparseMAP [43]	$\mathsf{conv}(\mathcal{Y})$	$\frac{1}{2}\lVert\boldsymbol{\mu}\rVert^2$	$\mathsf{sparseMAP}(\boldsymbol{\theta})$	$\frac{1}{2}\lVert\boldsymbol{y} - \boldsymbol{\theta}\rVert^2 - \frac{1}{2}\lVert\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{\theta}\rVert^2$

Cost-sensitive losses. Fenchel-Young losses also include the hinge loss of support vector machines. Indeed, from any classification loss L_Ω, we can construct a cost-sensitive version of it as follows. Define $\Psi(\boldsymbol{\mu}; \boldsymbol{y}) \coloneqq \Omega(\boldsymbol{\mu}) - \langle \boldsymbol{c}_{\boldsymbol{y}}, \boldsymbol{\mu} \rangle$, where $\boldsymbol{c}_{\boldsymbol{y}} \in \mathbb{R}_+^{|\mathcal{Y}|}$ is a fixed cost vector that depends on the groundtruth $\boldsymbol{y}$; for example $\boldsymbol{c}_{\boldsymbol{y}} = \boldsymbol{1} - \boldsymbol{y}$ corresponds to the 0/1 cost. Then, L_Ψ is a cost-sensitive version of L_Ω, which can be written as $L_{\Psi(\cdot;\boldsymbol{y})}(\boldsymbol{\theta}; \boldsymbol{y}) = L_\Omega(\boldsymbol{\theta} + \boldsymbol{c}_{\boldsymbol{y}}; \boldsymbol{y}) = \Omega^*(\boldsymbol{\theta} + \boldsymbol{c}_{\boldsymbol{y}}) + \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta} + \boldsymbol{c}_{\boldsymbol{y}}, \boldsymbol{y} \rangle$. This construction recovers the multi-class hinge loss ([17]; $\Omega = 0$), the softmax-margin loss ([26]; $\Omega = -\mathsf{H}^{\mathrm{S}}$), and the cost-augmented sparsemax ([50, Eq. (13)], [43]; $\Omega = \frac{1}{2}\lVert \cdot \rVert^2$).

Relation to Bregman divergences. All the losses in Table 2 can be extended to work over $\mathcal{Y} = \mathsf{dom}(\Omega)$. For example, in the case of the logistic loss, where $\Omega(\boldsymbol{y}) = -\mathsf{H}^{\mathrm{S}}(\boldsymbol{y})$, allowing $\boldsymbol{y} \in \triangle^{|\mathcal{Y}|}$ instead of $\boldsymbol{y} \in \{\boldsymbol{e}_i\}_{i=1}^d$ yields the cross-entropy loss, $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \mathsf{KL}(\boldsymbol{y} \| \mathsf{softmax}(\boldsymbol{\theta}))$. More generally, there is a relation between Fenchel-Young losses and Bregman divergences. Recall that the Bregman divergence [11] generated by a strictly convex and differentiable Ω is

$$B_\Omega(\boldsymbol{y} || \boldsymbol{\mu}) \coloneqq \Omega(\boldsymbol{y}) - \Omega(\boldsymbol{\mu}) - \langle \nabla\Omega(\boldsymbol{\mu}), \boldsymbol{y} - \boldsymbol{\mu} \rangle. \tag{3}$$

In other words, this is the difference at $\boldsymbol{y}$ between Ω and its linearization around $\boldsymbol{\mu}$. Letting $\boldsymbol{\theta} = \nabla\Omega(\boldsymbol{\mu})$ (i.e., $(\boldsymbol{\theta}, \boldsymbol{\mu})$ is a **dual pair**), we have $\Omega^*(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \Omega(\boldsymbol{\mu})$. Substituting in (3), we get $B_\Omega(\boldsymbol{y} || \boldsymbol{\mu}) = L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$. In other words, Fenchel-Young losses can be viewed as a "mixed-form Bregman divergence" [1, Theorem 1.1] where the argument $\boldsymbol{\mu}$ in (3) is **replaced by its dual point** $\boldsymbol{\theta}$.

3 Categorical distribution induced by Fenchel-Young losses

In the previous section, we presented Fenchel-Young losses in a broad setting. We now restrict to classification and show that Fenchel-Young losses induce a probability distribution over classes.

3.1 Generalized entropies and probability spaces

We first restrict to the case $\mathcal{Y} = \{\boldsymbol{e}_i\}_{i=1}^d$. In this case, we have $(\boldsymbol{\theta}, \boldsymbol{\mu}) = (\boldsymbol{s}, \boldsymbol{p})$, where $\boldsymbol{s} \in \mathbb{R}^{|\mathcal{Y}|}$ is a vector of prediction scores and $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$ a vector of probabilities (this is not the case in structured prediction, as will become clear in §3.2). We further assume that $\Omega(\boldsymbol{p}) = -\mathsf{H}(\boldsymbol{p})$, where H is a **generalized entropy** [29]. We say that H is a generalized entropy if $\mathsf{dom}(\mathsf{H}) = \triangle^{|\mathcal{Y}|}$ and H is concave. Using a general concave function $\mathsf{H}(\boldsymbol{p})$ to measure the "uncertainty" in a distribution $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$ dates back to at least [21]. The corresponding Fenchel-Young loss is then $L_{-\mathsf{H}}(\boldsymbol{s}; \boldsymbol{y})$.

Assumptions: We will occasionally make the following assumptions about H.

A.1. Zero entropy: $\mathsf{H}(\boldsymbol{p}) = 0$ if $\boldsymbol{p}$ is a delta distribution, i.e., $\boldsymbol{p} \in \mathcal{Y}$.

A.2. Strict concavity: this implies that $\nabla(-\mathsf{H})^*(\boldsymbol{s})$ exists.

A.3. Symmetry: $\mathsf{H}(\boldsymbol{p}) = \mathsf{H}(\boldsymbol{P}\boldsymbol{p})$ for any permutation matrix $\boldsymbol{P}$.

Assumptions A.2 and A.3 imply that H is Schur-concave [4], a common requirement in generalized entropies. This in turn implies assumption A.1, up to a constant. Together these assumptions imply that H can be used as a sensible uncertainty measure, as shown next (cf. Appendix B.1 for a proof).

Proposition 2 *Assumptions A.1–A.3 imply the following:*

1. H *is non-negative and uniquely maximized by the uniform distribution* $\boldsymbol{p} = \mathbf{1}/|\mathcal{Y}|$*;*
2. *The conjugate* $(-\mathsf{H})^*$ *is convex, symmetric, and differentiable;*
3. $\nabla(-\mathsf{H})^*(\boldsymbol{P}\boldsymbol{s}) = \boldsymbol{P}\nabla(-\mathsf{H})^*(\boldsymbol{s})$ *for any permutation matrix* $\boldsymbol{P}$*;*
4. $\nabla(-\mathsf{H})^*(\boldsymbol{s})$ *is "order-preserving": If* $\boldsymbol{p} = \nabla(-\mathsf{H})^*(\boldsymbol{s})$*, then the coordinates of* $\boldsymbol{p}$ *and* $\boldsymbol{s}$ *are sorted the same way, i.e.,* $s_i > s_j \Rightarrow p_i \geq p_j$ *and* $p_i > p_j \Rightarrow s_i > s_j$*.*

A particular case of generalized entropies satisfying assumptions A.1–A.3 are separable functions of the form $\mathsf{H}(\boldsymbol{p}) = \sum_{j=1}^{|\mathcal{Y}|} h(p_j)$, where $h : [0,1] \to \mathbb{R}_+$ is a non-negative strictly concave function such that $h(0) = h(1) = 0$. However, our framework does not require to restrict to this form.

The prediction function $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{s}) = \nabla(-\mathsf{H})^*(\boldsymbol{s})$ produces a distribution over classes, which can be seen as a special case of generalized exponential family [29, 24]. Because, in the unstructured setting, $\mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{p}$, the prediction function $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{s})$ can also be seen as the mean under a distribution $\boldsymbol{p}$. We now give examples of generalized entropies, some of them enjoying a closed form for $\nabla(-\mathsf{H})^*(\boldsymbol{s})$.

Shannon entropy [51]. This is the foundation of information theory, defined as $\mathsf{H}^{\mathrm{S}}(\boldsymbol{p}) \coloneqq -\sum_{j=1}^{|\mathcal{Y}|} p_j \log p_j$. As seen in Table 2, the resulting Fenchel-Young loss $L_{-\mathsf{H}^{\mathrm{S}}}$ corresponds to the logistic loss. The associated distribution is the classical softmax (cf. [10, Ex. 3.25] for a derivation):

$$\nabla(-\mathsf{H}^{\mathrm{S}})^*(\boldsymbol{s}) = \mathsf{softmax}(\boldsymbol{s}) \coloneqq \frac{\exp(\boldsymbol{s})}{\sum_{j=1}^{d} \exp(s_j)} > \mathbf{0}.$$

Tsallis α-entropies [54]. These are defined as $\mathsf{H}^{\mathrm{T}}_\alpha(\boldsymbol{p}) \coloneqq k(\alpha-1)^{-1}\left(1 - \sum_{j=1}^{|\mathcal{Y}|} p_j^\alpha\right)$, where $\alpha \geq 0$ and k is an arbitrary positive constant. These entropies arise as a generalization of the Shannon-Khinchin axioms to non-extensive systems [52] and have numerous scientific applications [25, 38]. For convenience, we set $k = \alpha^{-1}$ for the rest of this paper. Tsallis entropies satisfy assumptions A.1–A.3 and can also be written in separable form:

$$\mathsf{H}^{\mathrm{T}}_\alpha(\boldsymbol{p}) \coloneqq \sum_{j=1}^{|\mathcal{Y}|} h_\alpha(p_j) \quad \text{with} \quad h_\alpha(t) \coloneqq \frac{t - t^\alpha}{\alpha(\alpha-1)}.$$

The limit case $\alpha \to 1$ corresponds to the Shannon entropy. When $\alpha = 2$, we recover the Gini index [27], a popular "impurity measure" for decision trees:

$$\mathsf{H}^{\mathrm{T}}_2(\boldsymbol{p}) = \frac{1}{2}\sum_{j=1}^{|\mathcal{Y}|} p_j(1-p_j) = \frac{1}{2}(1 - \|\boldsymbol{p}\|_2^2). \tag{4}$$

It can be checked that $L_{-\mathsf{H}^{\mathrm{T}}_2}$ equals the sparsemax loss [37] (cf. Table 2). The gradient mapping is

$$\nabla(-\mathsf{H}^{\mathrm{T}}_2)^*(\boldsymbol{s}) = \mathsf{sparsemax}(\boldsymbol{s}) \coloneqq \operatorname*{\mathsf{argmax}}_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \langle \boldsymbol{p}, \boldsymbol{s}\rangle + \mathsf{H}^{\mathrm{T}}_2(\boldsymbol{p}) = \operatorname*{\mathsf{argmin}}_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \|\boldsymbol{p} - \boldsymbol{s}\|^2,$$

which is the Euclidean projection onto the probability simplex. This recovers the sparsemax transformation introduced in [37]. The resulting distribution has **sparse** support: it may assign exactly zero probability to some classes. The projection can be computed **exactly** in $O(|\mathcal{Y}|)$ time [31, 13, 22, 16].

Another interesting case is $\alpha \to +\infty$, which gives $\mathsf{H}^{\mathrm{T}}_\infty(\boldsymbol{p}) = 0$, hence $L_{-\mathsf{H}^{\mathrm{T}}_\infty}$ is the perceptron loss in Table 2. The resulting "argmax" distribution puts all probability mass on the top-scoring classes. In sum, the prediction functions for $\alpha = 1, 2, \infty$ are respectively softmax, sparsemax, and argmax. Tsallis entropies can be seen as a **continuous parametric family** subsuming these important cases.

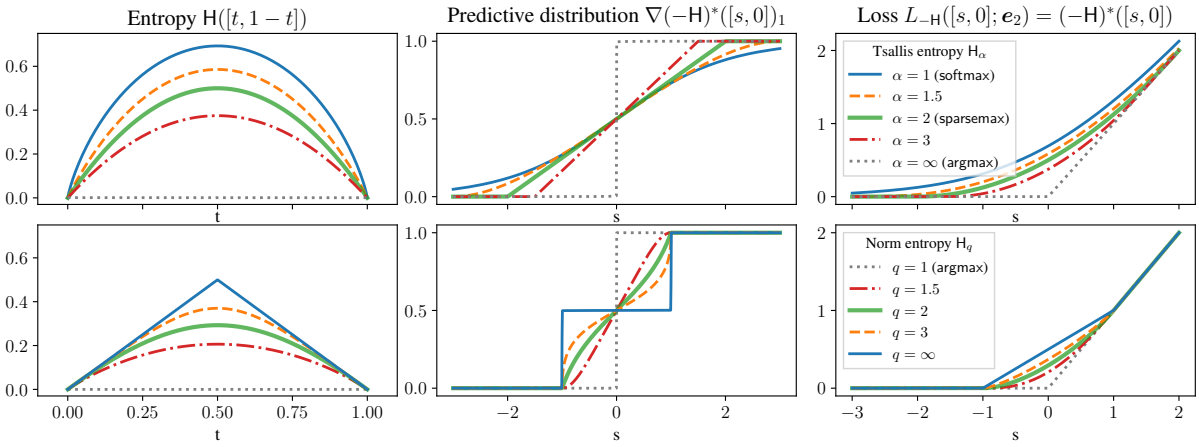


Figure 1: Tsallis and norm entropies (left) along with their prediction functions (middle) and Fenchel-Young losses (right) for the binary case, where $\boldsymbol{p} = (t, 1-t) \in \triangle^2$ and $\boldsymbol{s} = (s, 0) \in \mathbb{R}^2$. Except for softmax, which never exactly reaches 0, all mappings shown on the center lead to **sparse outputs**.

Other entropy families. An interesting class of non-separable entropies are entropies generated by a q-norm, defined as $\mathsf{H}^{\mathrm{N}}_q(\boldsymbol{p}) := 1 - \|\boldsymbol{p}\|_q$; we call them **norm entropies**. Since norms are convex, these entropies satisfy assumptions A.1–A.3 for $q \geq 1$. They differ from Tsallis entropies in that the norm is not raised to q: a subtle but important difference. We illustrate H, $\nabla(-\mathsf{H})^*$, and $L_{-\mathsf{H}}$ for Tsallis and norm entropies in Figure 1. The limit case $q \to \infty$ is particularly interesting: in this case, we obtain $\mathsf{H}^{\mathrm{N}}_\infty = 1 - \|\boldsymbol{p}\|_\infty$, recovering the Berger-Parker dominance index [7], widely used in ecology to measure species diversity. Other interesting entropies include the **squared-norm entropies** [42] and **Rényi entropies** [45]; cf. Appendix A for more details.

3.2 Structured prediction and mean spaces

We now extend this probabilistic perspective to the structured prediction setting, where $\mathcal{Y}$ is a set of structured objects. In this setting, probability distributions and means no longer coincide, i.e., $\mathbb{E}_{\boldsymbol{p}}[Y] \neq \boldsymbol{p}$, and they live in different spaces. We assume that Ω can be written in the form

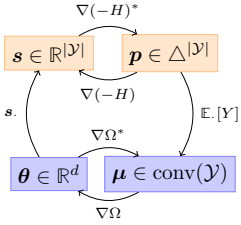
$$\mathsf{H}(\boldsymbol{p}) = -\Omega(\mathbb{E}_{\boldsymbol{p}}[Y]) \quad \text{for all } \boldsymbol{p} \in \triangle^{|\mathcal{Y}|}. \tag{5}$$

That is, $\mathsf{dom}(\Omega) = \mathsf{conv}(\mathcal{Y}) \subseteq \mathbb{R}^d$. In particular, this is the case when using a generalized maximum entropy principle [29, 24]: for all $\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})$ we define $-\Omega(\boldsymbol{\mu}) = \sup_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \mathsf{H}(\boldsymbol{p})$ s.t. $\mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}$.

As with exponential families [3], we can alternatively characterize a distribution induced by L_Ω with its mean. Let $\boldsymbol{s_\theta} \in \mathbb{R}^{|\mathcal{Y}|}$ be a vector of predictions scores, with elements defined by $s_{\boldsymbol{\theta}}(\boldsymbol{y}) := \langle \boldsymbol{y}, \boldsymbol{\theta} \rangle \; \forall \boldsymbol{y} \in \mathcal{Y}$. Assuming that Ω is of the form (5), we have the simple but far-reaching identity

$$(-\mathsf{H})^*(\boldsymbol{s_\theta}) = \sup_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \langle \boldsymbol{p}, \boldsymbol{s_\theta} \rangle + \mathsf{H}(\boldsymbol{p}) = \sup_{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})} \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - \Omega(\boldsymbol{\mu}) = \Omega^*(\boldsymbol{\theta}),$$

where we used $\langle \boldsymbol{p}, \boldsymbol{s_\theta} \rangle = \langle \mathbb{E}_{\boldsymbol{p}}[Y], \boldsymbol{\theta} \rangle$. This identity connects the conjugates $(-\mathsf{H})^*$ in probability space $\triangle^{|\mathcal{Y}|}$ and Ω^* in mean space $\mathsf{conv}(\mathcal{Y}) \subseteq \mathbb{R}^d$. This is useful in structured prediction, since $\Omega^*(\boldsymbol{\theta})$ just involves a d-dimensional optimization problem instead of a $|\mathcal{Y}|$-dimensional one for $(-H)^*$. The optimal distribution $\boldsymbol{p}^\star$ is related to $\boldsymbol{\mu}^\star$ by $\boldsymbol{\mu}^\star = \sum_{\boldsymbol{y} \in \mathcal{Y}} p^\star(\boldsymbol{y}) \boldsymbol{y}$ and, from Carathéodory's theorem, the support of $\boldsymbol{p}^\star$ contains at most $d \ll |\mathcal{Y}|$ elements. If Ω^* is twice-differentiable, its gradient and Hessian equal the first and second moments under $\boldsymbol{p}^\star$:



$$\mathbb{E}_{\boldsymbol{p}^\star}[Y] = \nabla\Omega^*(\boldsymbol{\theta}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \mathsf{conv}(\mathcal{Y}) \quad \text{and} \quad \mathrm{cov}_{\boldsymbol{p}^\star}[Y] = \nabla^2\Omega^*(\boldsymbol{\theta}).$$

As indicated in Table 2, $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ recovers MAP, marginals and sparseMAP for suitable choices of Ω. In the sparseMAP case [43], the mean $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ along with the distribution $\boldsymbol{p}^\star$ can be computed using conditional gradient or active set algorithms even when $\mathsf{conv}(\mathcal{Y})$ does not enjoy an exact expression. Each iteration of these algorithms only requires a call to a maximum a-posteriori (MAP) oracle, a.k.a. linear minimization oracle (LMO) in the conditional gradient literature [32].

4 Separation margin of Fenchel-Young losses

In this section, we are going to see that the simple assumptions A.1–A.3 about the generalized entropy H are enough to obtain results about the separation margin associated with $L_{-\mathsf{H}}$. The notion of separation margin is well-known in machine learning, lying at the heart of support vector machines and leading to generalization error bounds [55, 49, 30]. We provide a definition and will see next that many other Fenchel-Young losses also have a "margin," for suitable conditions on H.

Definition 2 *Let $L(\boldsymbol{s}; \boldsymbol{y})$ be a loss function over $\mathbb{R}^{|\mathcal{Y}|} \times \{\boldsymbol{e}_i\}_{i=1}^{|\mathcal{Y}|}$, where $\boldsymbol{y} = \boldsymbol{e}_k$ is a ground truth label. We say that L has a **separation margin property** if there exists $m > 0$ such that:*

$$s_k \geq m + \max_{j \neq k} s_j \quad \Rightarrow \quad L(\boldsymbol{s}; \boldsymbol{y}) = 0. \tag{6}$$

The smallest possible m that satisfies (6) *is called the* margin *of L, denoted* $\operatorname{margin}(L)$.

The multi-class hinge loss, $L(\boldsymbol{s}; \boldsymbol{e}_k) = \max\{0, \max_{j \neq k} 1 + s_j - s_k\}$, which we saw in §2 to be an instance of a Fenchel-Young loss, is a famous example with margin 1.

To see that other entropies also induce a margin, we need first to characterize the gradient mappings $\partial(-\mathsf{H})$ and $\nabla(-\mathsf{H})^*$ associated with these entropies. (Note that the former mapping is never single-valued: if $\boldsymbol{s}$ is in $\partial(-\mathsf{H})(\boldsymbol{p})$, then so is $\boldsymbol{s} + c\mathbf{1}$, for any constant $c \in \mathbb{R}$.) We show in Appendix B.3 (Lemma 1) that $\boldsymbol{s} \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$ iff $s_k = (-\mathsf{H})^*(\boldsymbol{s})$. This confirms that the Shannon entropy H^{S} has $\partial(-\mathsf{H}^{\mathrm{S}})(\boldsymbol{e}_k) = \varnothing$: since $(-\mathsf{H}^{\mathrm{S}})^*$ is the log-partition function, which upper bounds the maximal score, there is no vector $\boldsymbol{s}$ satisfying the required condition and **hence $L_{-\mathsf{H}^{\mathrm{S}}}$ does not have a margin**.

The next result, proved in Appendix B.2, characterizes more precisely the images of $\nabla(-\mathsf{H})^*$.

Proposition 3 *Let H satisfy assumptions A.1–A.3. If $\partial(-\mathsf{H})(\boldsymbol{p}) \neq \varnothing$ for any $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$, then:*

1. *The mapping $\nabla(-\mathsf{H})^*$ covers the full simplex, i.e., $\nabla(-\mathsf{H})^*(\mathbb{R}^{|\mathcal{Y}|}) = \triangle^{|\mathcal{Y}|}$;*
2. *Suppose $\nabla(-\mathsf{H})^*(\boldsymbol{s}) = \boldsymbol{p}$ where $\boldsymbol{p}$ is a relative boundary point, with support $\mathcal{I} \coloneqq \{1 \leq i \leq |\mathcal{Y}| \colon p_i > 0\}$. Then, we have $\nabla(-\mathsf{H})^*(\boldsymbol{s}') = \boldsymbol{p}$ for any $\boldsymbol{s}'$ such that $s'_i = s_i$ for $i \in \mathcal{I}$ and $s'_i \leq s_i$ otherwise.*

If H is separable with $\mathsf{H}(\boldsymbol{p}) = \sum_i h(p_i)$, the conditions required by Proposition 3 are met if $\lim_{t=0^+} h'(t) < \infty$ and $\lim_{t=1^-} h'(t) > -\infty$. This is the case with Tsallis entropies for $\alpha > 1$, but not Shannon entropy, for which $h'(t) = -1 - \log t$. Functions whose gradient "explode" in the boundary of their domain (hence failing to meet the condition in Proposition 3) are called "essentially smooth" [46]. For those functions, $\nabla(-\mathsf{H})^*$ maps only to the relative interior of $\triangle^{|\mathcal{Y}|}$, never attaining boundary points [56]. This prevents these functions from generating a sparse $\nabla(-\mathsf{H})^*$ or a loss $L_{-\mathsf{H}}$ with a margin. Hence, there is a **close link between sparse distributions and losses with a margin**.

We are now ready to state the main result of this section, proved in Appendix B.3: an explicit condition for a Fenchel-Young loss to have a margin, along with an expression to compute it.

Proposition 4 *The loss $L_{-\mathsf{H}}$ has the separation margin property iff there is a $m > 0$ such that $m\boldsymbol{e}_k \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$. The margin of $L_{-\mathsf{H}}$ is given by the smallest such m, or equivalently by:*

$$\operatorname{margin}(L_{-\mathsf{H}}) = \sup_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \frac{\mathsf{H}(\boldsymbol{p})}{1 - \|\boldsymbol{p}\|_\infty}. \tag{7}$$

Interestingly, the denominator of (7) is the generalized entropy $\mathsf{H}^{\mathrm{N}}_\infty(\boldsymbol{p})$ introduced in §3. As Figure 1 suggests, this entropy provides an upper bound for convex losses with unit margin. The next proposition, which we prove in full detail in Appendix B.4, takes a step even further and provides a remarkably simple formula for twice-differentiable generalized entropies.

Proposition 5 *Assume* H *satisfies the conditions in Proposition 3 and is twice-differentiable on the simplex. Then, for arbitrary* $j \neq k$*:*

$$\operatorname{margin}(L_{-\mathsf{H}}) = \nabla_j \mathsf{H}(\boldsymbol{e}_k) - \nabla_k \mathsf{H}(\boldsymbol{e}_k).$$

In particular, if H *is separable, i.e.,* $\mathsf{H}(\boldsymbol{p}) = \sum_{i=1}^{|\mathcal{Y}|} h(p_i)$*, where* $h : [0,1] \to \mathbb{R}_+$ *is concave, twice differentiable, and satisfies* $h(0) = h(1) = 0$*, we have*

$$\operatorname{margin}(L_{-\mathsf{H}}) = h'(0) - h'(1) = -\int_0^1 h''(t)dt. \tag{8}$$

This provides a geometric characterization of separable entropies and their margins: (8) tells us that only the slopes of h at the two extremities of $[0,1]$ are relevant to determine its margin. It is also a **constructive result**, since it allows to design an entropy that leads to a loss with a prescribed margin.

Margins of Tsallis and norm entropies. We now apply the results above to obtain margin values of Tsallis and norm entropies (which can be confirmed visually in Figure 1).

Proposition 6 *1. Let* $\mathsf{H}^{\mathrm{T}}_\alpha$ *be the Tsallis* α*-entropy. Then, for* $\alpha \in [0,1]$*,* $L_{-\mathsf{H}^{\mathrm{T}}_\alpha}$ *does not have the separation margin property. For* $\alpha > 1$*, we have* $\operatorname{margin}(L_{-\mathsf{H}^{\mathrm{T}}_\alpha}) = 1/(\alpha - 1)$*.*

2. Let $\mathsf{H}^{\mathrm{N}}_q$ *be the* q*-norm entropy. Then, for any* $q \geq 1$*, we have* $\operatorname{margin}(L_{-\mathsf{H}^{\mathrm{N}}_q}) = 1$*.*

Proof: As seen in §3, Tsallis entropies are separable with $h(t) = (t - t^\alpha)/(\alpha(\alpha-1))$. For $\alpha > 1$, the derivative is $h'(t) = (1 - \alpha t^{\alpha-1})/(\alpha(\alpha-1))$, so we have $h'(0) = 1/(\alpha(\alpha-1))$ and $h'(1) = -1/\alpha$. Proposition 5 then yields $\operatorname{margin}(L_{-\mathsf{H}^{\mathrm{T}}}) = h'(0) - h'(1) = 1/(\alpha-1)$. Norm entropies are not separable, but their gradient can be expressed as $\nabla \mathsf{H}^{\mathrm{N}}_q(\boldsymbol{p}) = -(\boldsymbol{p}/\|\boldsymbol{p}\|_q)^{q-1}$. We thus have $\nabla \mathsf{H}^{\mathrm{N}}_q(\boldsymbol{e}_k) = -\boldsymbol{e}_k$, so $\operatorname{margin}(\mathsf{H}^{\mathrm{N}}_q) = \nabla_j \mathsf{H}^{\mathrm{N}}_q(\boldsymbol{e}_k) - \nabla_k \mathsf{H}^{\mathrm{N}}_q(\boldsymbol{e}_k) = 1$. ■

5 Learning with Fenchel-Young losses

We now turn to learning predictive models with a Fenchel-Young loss L_Ω, for arbitrary $\mathsf{dom}(\Omega)$.

Primal gradient-based algorithms. Let $\boldsymbol{f}_W : \mathcal{X} \to \mathbb{R}^d$ be a model parametrized by W. To learn W from training data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{X} \times \mathcal{Y}$, we minimize the regularized empirical risk

$$\min_W \sum_{i=1}^n L_\Omega(\boldsymbol{f}_W(\boldsymbol{x}_i); \boldsymbol{y}_i) + G(W), \tag{9}$$

where G is a regularizer. This objective is very broad and allows to learn, e.g., neural networks using Fenchel-Young losses. If $\boldsymbol{f}_W$ is a linear model and G is convex, then (9) is convex. Assuming further that Ω is strongly convex, then L_Ω is smooth and (9) can be solved globally using proximal gradient algorithms [58, 5, 20]. From Proposition 1 (item 3), the gradient of the loss term with respect to $\boldsymbol{f}_W(\boldsymbol{x})$ involves the regularized prediction function (1), $\widehat{\boldsymbol{y}}_\Omega = \nabla \Omega^*$. The two key computational ingredients to solve (9) are therefore $\nabla \Omega^*$ and the proximity operator of G, if G is non-differentiable.

Computing $\nabla \Omega^*(\boldsymbol{\theta}) = \mathsf{argmax}_{\boldsymbol{y} \in \mathsf{dom}(\Omega)} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle - \Omega(\boldsymbol{y})$ is a concave optimization problem, which enjoys a closed form when $\Omega = -\mathsf{H}$ is the negative Shannon entropy or Gini index (§3). More generally, for separable entropies defined over the simplex, we next show that computing $\nabla(-\mathsf{H})^*$ **reduces to unidimensional root finding** (cf. §B.5 for a proof).

Proposition 7 *Let* $\mathsf{H}(\boldsymbol{p}) = \sum_{j=1}^{|\mathcal{Y}|} h(p_j)$*, where* h *is strictly concave, with derivative* h'*. Then we have* $\nabla(-\mathsf{H})^*(\boldsymbol{s}) = \widehat{\boldsymbol{p}}(\tau^\star)$*, where* $\tau^\star$ *solves the equation* $f(\tau^*) = 0$*, and where we defined*

$$\widehat{\boldsymbol{p}}(\tau) \coloneqq (-h')^{-1}(\mathsf{max}\{\boldsymbol{s} - \tau, -h'(0)\}), \quad f(\tau) \coloneqq \widehat{\boldsymbol{p}}(\tau)^\top \mathbf{1} - 1.$$

The search interval is $[\tau_{\mathsf{min}}, \tau_{\mathsf{max}}]$*, where* $\tau_{\mathsf{min}} \coloneqq \mathsf{max}(\boldsymbol{s}) + h'(1)$*,* $\tau_{\mathsf{max}} \coloneqq \mathsf{max}(\boldsymbol{s}) + h'\left(1/|\mathcal{Y}|\right)$*.*

An approximate τ such that $|f(\tau)| \leq \epsilon$ can be found in $O(1/\log \epsilon)$ time by, e.g., bisection. Root finding has been employed for related projections [33] but our result applies more broadly (cf. Appendix B.5).

Dual algorithms. When $\boldsymbol{f}_W(\boldsymbol{x}) = W\boldsymbol{x}$, we may instead solve the corresponding dual problem

$$\max_{\beta} -D(\beta) \text{ s.t. } \boldsymbol{\beta}_i \in \mathsf{dom}(\Omega)\ \forall i \in [n], \quad \text{where} \quad D(\beta) \coloneqq \sum_i \Omega(\boldsymbol{\beta}_i) - \Omega(\boldsymbol{y}_i) + G^*(V(\beta)) \tag{10}$$

and where we defined $V(\beta) \coloneqq \sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\beta}_i)\boldsymbol{x}_i^\top$. Assuming G is a λ-strongly convex regularizer, given an optimal dual solution $\beta^\star$, we retrieve the optimal primal solution by $W^\star = \nabla G^*(V(\beta^\star))$. We can solve (10) using stochastic block coordinate ascent. At every iteration, we pick $i \in [n]$ and follow [50, Option I] to solve an approximate sub-problem w.r.t. $\boldsymbol{\beta}_i$. By using the specific form of L_Ω losses, we are able to formulate the update w.r.t. $\boldsymbol{\beta}_i$ as a proximity operator. Namely, we perform $\boldsymbol{\beta}_i \leftarrow \mathrm{prox}_{\frac{1}{\sigma_i}\Omega}(\boldsymbol{v}_i/\sigma_i)$, where $\boldsymbol{v}_i \coloneqq \nabla G^*(V(\beta))\boldsymbol{x}_i + \sigma_i\boldsymbol{\beta}_i$, $\sigma_i \coloneqq \frac{\|\boldsymbol{x}_i\|^2}{\lambda}$ and where we defined

$$\mathrm{prox}_{\tau\Omega}(\boldsymbol{x}) \coloneqq \operatorname*{argmin}_{\boldsymbol{y} \in \mathsf{dom}(\Omega)} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + \tau\Omega(\boldsymbol{y}). \tag{11}$$

Note that if $\mathrm{prox}_{\tau\Omega}$ is sparse, **then so are the dual variables** $\{\boldsymbol{\beta}_i\}_{i=1}^n$. Sparsity in the dual variables is useful when kernelizing models, as it makes computing predictions more efficient. In many cases, (11) can be solved in closed form; see §B.7 for examples and for a detailed derivation. However, computing $\mathrm{prox}_{\tau\Omega}$ can sometimes be more challenging than $\nabla\Omega^*$. For instance, when Ω is the Shannon negative entropy $-\mathsf{H}^{\mathrm{S}}$, $\nabla\Omega^*$ enjoys a closed form but not $\mathrm{prox}_{\tau\Omega}$. For such cases, assuming $\mathsf{dom}(\Omega)$ is a compact set, another option to solve (10) is the block Frank-Wolfe algorithm [34].

6 Related work

Proper scoring rules (a.k.a. proper losses) [29, 28, 57] are a key tool in the context of probability forecasting [19] to express the goodness-of-fit (lower is better) between a ground truth $\boldsymbol{y} \in \mathcal{Y}$ and a vector of predictions $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$ (note that $\boldsymbol{p}$ must be in the simplex). The fact that any scoring rule induces a generalized entropy is well-known [21, 29]. The reverse mapping has also been studied. Let H be a generalized entropy. Then we can construct a proper scoring rule $S_{-\mathsf{H}} \colon \triangle^{|\mathcal{Y}|} \times \mathcal{Y} \to \mathbb{R}$ as follows [48, 28, 44]

$$S_\Omega(\boldsymbol{p}; \boldsymbol{y}) \coloneqq \langle \nabla\Omega(\boldsymbol{p}), \boldsymbol{p}\rangle - \Omega(\boldsymbol{p}) - \nabla\Omega(\boldsymbol{p})(\boldsymbol{y}) = \Omega^*(\nabla\Omega(\boldsymbol{p})) - \nabla\Omega(\boldsymbol{p})(\boldsymbol{y}).$$

As an example, choosing the Gini index $\mathsf{H}(\boldsymbol{p}) = 1 - \|\boldsymbol{p}\|^2$ generates the **Brier score** [12] $S_{-\mathsf{H}}(\boldsymbol{p}; \boldsymbol{y}) = \sum_{\boldsymbol{y}' \in \mathcal{Y}} ([\![\boldsymbol{y} = \boldsymbol{y}']\!] - p(\boldsymbol{y}))^2$. Thus sparsemax and the Brier score share the same generating function. A crucial difference between S_Ω and L_Ω, however, is that S_Ω is **not necessarily convex** in its first argument ([57, Proposition 17] shows that it is in fact quasi-convex) while L_Ω **always is**. In addition, the first argument is **constrained** to $\triangle^{|\mathcal{Y}|}$ for S_Ω, while it is **unconstrained** for L_Ω. For this reason, S_Ω is usually composed with an **invertible** link function [14, 57], while this is not necessary with L_Ω. Finally, a classification loss construction was recently proposed in [23, Proposition 3]. Fenchel-Young losses are broader, supporting e.g. regression, and in fact include [23, Proposition 3] as a special case.

Smoothing techniques. Another way to view our work is through the prism of smoothing techniques [41, 6, 42, 39]. Indeed, if $\mathsf{dom}(\Omega) = \triangle^{|\mathcal{Y}|}$, Ω^* can be seen as a smoothed max operator (see Figure 1, right), with $(t\Omega)^*(\boldsymbol{s}) \to \mathsf{max}(\boldsymbol{s})$ when $t \to 0$. Smoothing techniques were used extensively in [50] to create smoothed losses. However, these techniques were applied on a per-loss basis, while we propose a generic loss construction, with clear links between smoothing / regularization and the probability distribution over classes induced by $\nabla\Omega^*$.

7 Conclusion

Fenchel-Young losses not only unify many existing loss functions, they offer a principled way to construct new ones from any convex regularizer Ω. When the regularizer is a generalized entropy, we showed that Fenchel-Young losses induce a probability distribution over classes and under precise conditions we derived, enjoy a separation margin: a useful property leading to generalization error bounds [55, 49], and as we showed, closely related to sparse gradient mappings. Importantly, Fenchel-Young losses support regularizers Ω defined over arbitrary domains $\mathsf{dom}(\Omega)$. Leveraging this fact allows to construct loss functions for a large variety of applications.

Acknowledgements

We are grateful to Tim Vieira for pointing us to [24], which kicked off this project. We thank Mário Figueiredo and Nicolas Keriven for reading a draft of this paper and for their insightful comments. AM was partially supported by the European Research Council (ERC StG DeepSPIN 758969) and by the Fundação para a Ciência e Tecnologia through contracts UID/EEA/50008/2013, PTDC/EEI-SII/7092/2014 (LearnBig), and CMUPERI/TIC/0046/2014 (GoLocal).

References

[1] S. Amari. *Information Geometry and Its Applications*. Springer, 2016.

[2] K. Ball, E. A. Carlen, and E. H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones Mathematicae*, 115(1):463–482, 1994.

[3] O. Barndorff-Nielsen. *Information and Exponential Families: In Statistical Theory*. John Wiley & Sons, 1978.

[4] H. H. Bauschke, P. L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 2011. Springer, 2017.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[6] A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

[7] W. H. Berger and F. L. Parker. Diversity of planktonic foraminifera in deep-sea sediments. *Science*, 168(3937):1345–1347, 1970.

[8] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific Belmont, 1999.

[9] J. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

[10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[11] L. 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.

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

[13] P. Brucker. An $O(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 3(3):163–166, 1984.

[14] A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, 2005.

[15] M. Collins. Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proc. of EMNLP*, 2002.

[16] L. Condat. Fast projection onto the simplex and the $\ell_1$ball. *Mathematical Programming*, 158(1-2):575–585, 2016.

[17] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[18] J. M. Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

[19] A. P. Dawid. Probability forecasting. *Encyclopedia of Statistical Sciences*, 1986.

[20] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Proc. of NIPS*, 2014.

[21] M. H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, pages 404–419, 1962.

[22] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proc. of ICML*, 2008.

[23] J. C. Duchi, K. Khosravi, and F. Ruan. Multiclass classification, information, divergence, and surrogate risk. *Annals of Statistics*, 2018.

[24] R. Frongillo and M. D. Reid. Convex foundations for generalized MaxEnt models. In *Proc. of AIP*, 2014.

[25] M. Gell-Mann and C. Tsallis. *Nonextensive Entropy: Interdisciplinary Applications*. Oxford University Press, 2004.

[26] K. Gimpel and N. A. Smith. Softmax-margin CRFs: Training log-linear models with cost functions. In *Proc. of NAACL*, 2010.

[27] C. Gini. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*, 1912.

[28] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[29] P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics*, pages 1367–1433, 2004.

[30] Y. Guermeur. Vc theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8:2551–2594, 2007.

[31] M. Held, P. Wolfe, and H. P. Crowder. Validation of subgradient optimization. *Mathematical Programming*, 6(1):62–88, 1974.

[32] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proc. of ICML*, 2013.

[33] W. Krichene, S. Krichene, and A. Bayen. Efficient Bregman projections onto the simplex. In *Proc. of CDC*. IEEE, 2015.

[34] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *Proc. of ICML*, 2012.

[35] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, 2001.

[36] O. Mangasarian. Pseudo-convex functions. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 3(2):281–290, 1965.

[37] A. F. Martins and R. F. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proc. of ICML*, 2016.

[38] A. F. T. Martins, M. A. T. Figueiredo, P. M. Q. Aguiar, N. A. Smith, and E. P. Xing. Nonextensive information theoretic kernels on measures. *Journal of Machine Learning Research*, 10:935–975, 2009.

[39] A. Mensch and M. Blondel. Differentiable Dynamic Programming for Structured Prediction and Attention. *arXiv preprint arXiv:1802.03676*, 2018.

[40] J. A. Nelder and R. J. Baker. *Generalized Linear Models*. Wiley Online Library, 1972.

[41] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[42] V. Niculae and M. Blondel. A regularized framework for sparse and structured neural attention. In *Proc. of NIPS*, 2017.

[43] V. Niculae, A. F. Martins, M. Blondel, and C. Cardie. SparseMAP: Differentiable sparse structured inference. In *Proc. of ICML*, 2018.

[44] M. D. Reid, R. M. Frongillo, R. C. Williamson, and N. Mehta. Generalized mixability via entropic duality. In *Proc. of COLT*, 2015.

[45] A. Rényi. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics, and Probability*, volume 1, pages 547–561, Berkeley, 1961. University of California Press.

[46] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[47] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

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

[49] B. Schölkopf and A. J. Smola. *Learning With Kernels*. The MIT Press, Cambridge, MA, 2002.

[50] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1):105–145, 2016.

[51] C. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Ill., 1949.

[52] H. Suyari. Generalization of Shannon-Khinchin axioms to nonextensive systems and the uniqueness theorem for the nonextensive entropy. *IEEE Trans. Information Theory*, 50(8):1783–1787, 2004.

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

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

[55] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[56] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[57] R. C. Williamson, E. Vernet, and M. D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 2016.

[58] S. Wright, R. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

[59] H. Zou, J. Zhu, and T. Hastie. New multicategory boosting algorithms based on multicategory fisher-consistent losses. *The Annals of Applied Statistics*, 2(4):1290, 2008.

Appendix

A More examples of generalized entropies and illustrations

We now discuss further examples of generalized entropy H. In all these examples, $\nabla(-\mathsf{H})^*$ can typically not be computed in closed form and we must resort to projected gradient solvers such as FISTA [5] to solve it. Figure 2 shows plots of these entropies, as well as their corresponding prediction functions and loss functions (we also repeat the Tsallis and norm entropies defined in §3).

A.1 Rényi β-entropies.

Rényi entropies [45] are defined for any $\beta \geq 0$ as:

$$\mathsf{H}^{\mathrm{R}}_\beta(\boldsymbol{p}) \coloneqq \frac{1}{1-\beta} \log \sum_{j=1}^{|\mathcal{Y}|} p_j^\beta.$$

Unlike Shannon and Tsallis entropies, Rényi entropies are not separable, with the exception of $\beta \to 1$, which also recovers Shannon entropy as a limit case. The case $\beta \to +\infty$ gives $\mathsf{H}^{\mathrm{R}}_\beta(\boldsymbol{p}) = -\log \|\boldsymbol{p}\|_\infty$. For $\beta \in [0, 1]$, Rényi entropies satisfy assumptions A.1–A.3; for $\beta > 1$, Rényi entropies fail to be concave. They are however pseudo-concave [36], meaning that, for all $\boldsymbol{p}, \boldsymbol{q} \in \triangle^{|\mathcal{Y}|}$, $\langle \nabla \mathsf{H}^{\mathrm{R}}_\beta(\boldsymbol{p}), \boldsymbol{q} - \boldsymbol{p} \rangle \leq 0$ implies $\mathsf{H}^{\mathrm{R}}_\beta(\boldsymbol{q}) \leq \mathsf{H}^{\mathrm{R}}_\beta(\boldsymbol{p})$. This implies, among other things, that points $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$ with zero gradient are maximizers of $\langle \boldsymbol{p}, \boldsymbol{\theta} \rangle + \mathsf{H}^{\mathrm{R}}_\beta(\boldsymbol{p})$, which allows to compute the predictive distribution $\nabla(-\mathsf{H}^{\mathrm{R}}_\beta)^*$ with gradient-based methods. This predictive distribution and the corresponding loss function are shown in Figure 2.

A.2 Squared q-norm entropies

Inspired by [42], as a simple extension of the Gini index (4), we consider the following generalized entropy based on squared q-norms:

$$\mathsf{H}^{\mathrm{SQ}}_q(\boldsymbol{p}) \coloneqq \frac{1}{2}(1 - \|\boldsymbol{p}\|_q^2) = \frac{1}{2} - \frac{1}{2}\left(\sum_{j=1}^{|\mathcal{Y}|} p_j^q\right)^{\frac{2}{q}} \geq 0 \quad \forall \boldsymbol{p} \in \triangle^{|\mathcal{Y}|}.$$

Contrary to [42], we include the constant term $\frac{1}{2}$ to ensure positivity.

For $q \in (1, 2]$, it is known that the squared q-norm is strongly convex w.r.t. $\|\cdot\|_q$ [2], implying that $(-\mathsf{H}^{\mathrm{SQ}}_q)^*$ is smooth. Although $\nabla(-\mathsf{H}^{\mathrm{SQ}}_q)^*(\boldsymbol{s}) = \mathsf{argmax}_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \ \langle \boldsymbol{p}, \boldsymbol{s} \rangle + \mathsf{H}^{\mathrm{SQ}}_q(\boldsymbol{p})$ can no longer be solved in closed form for $q \in (1, 2)$, it can be solved efficiently using projected gradient descent.

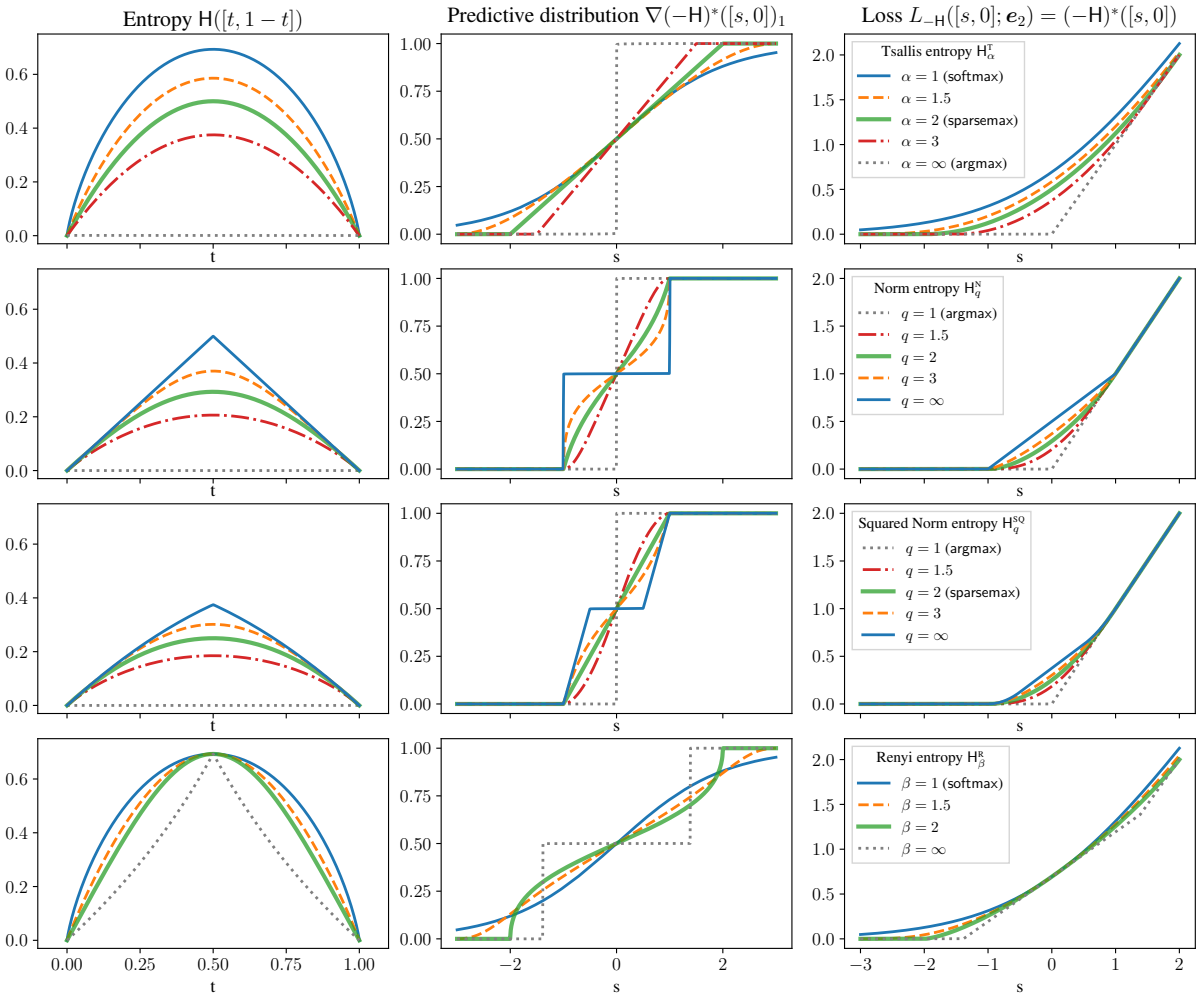


Figure 2: More examples of entropies (left) along with their prediction functions (middle) and Fenchel-Young losses (right) for the binary case, where $\boldsymbol{p} = (t, 1-t) \in \triangle^2$ and $\boldsymbol{s} = (s, 0) \in \mathbb{R}^2$.

A.3 Pairwise hinge entropies

We now turn to the following optimization problem, dating back to [21] (see also [29, 23]), which provides a *reverse* construction from a loss $L \colon \mathbb{R}^{|\mathcal{Y}|} \times \mathcal{Y} \to \mathbb{R}_+$ to an entropy $\mathsf{H}_L \colon \triangle^{|\mathcal{Y}|} \to \mathbb{R}_+$:

$$\mathsf{H}_L(\boldsymbol{p}) \coloneqq \inf_{\boldsymbol{s} \in \mathbb{R}^{|\mathcal{Y}|}} \sum_{\boldsymbol{y} \in \mathcal{Y}} p(\boldsymbol{y}) L(\boldsymbol{s}; \boldsymbol{y}). \tag{12}$$

This is the infimum of a linear and thus concave function of $\boldsymbol{p}$. Hence, by Danskin's theorem [18] [8, Proposition B.25], $H_L(\boldsymbol{p})$ is concave (note that this is true even if L is not convex).

The following proposition shows how to compute the entropy generated by pairwise hinge losses.

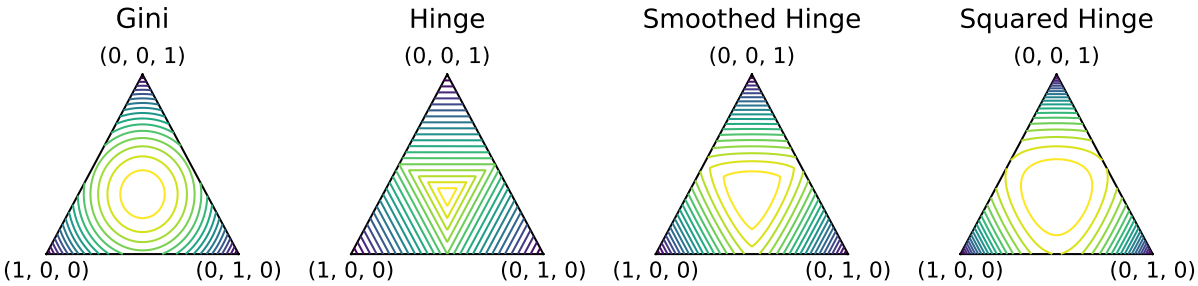


Figure 3: Contour comparison of Gini index $\mathsf{H}_2^{\mathrm{T}}$ and pairwise hinge entropies H_L; cf. Proposition 8.

Proposition 8 *Entropy generated by pairwise hinge losses*

Let $L(\boldsymbol{s}; \boldsymbol{e}_k) \coloneqq \sum_{j \neq k} \phi(s_j - s_k)$ *and* $\tau(\boldsymbol{p}) \coloneqq \min_j 1 - p_j$. *Then, for all* $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$:

1. *If* $\phi(t)$ *is the hinge function* $[1+t]_+$ *then* $\mathsf{H}_L(\boldsymbol{p}) = \tau(\boldsymbol{p})|\mathcal{Y}|$.
2. *If* $\phi(t)$ *is the smoothed hinge function* (20)*, then* $\mathsf{H}_L(\boldsymbol{p}) = \frac{-\tau(\boldsymbol{p})^2}{2} \sum_{j=1}^{|\mathcal{Y}|} \frac{1}{1-p_j} + \tau(\boldsymbol{p})|\mathcal{Y}|$.
3. *If* $\phi(t)$ *is the squared hinge function* $\frac{1}{2}[1+t]_+^2$*, then* $\mathsf{H}_L(\boldsymbol{p}) = \frac{\frac{1}{2}m^2}{\sum_{j=1}^{|\mathcal{Y}|} 1/(1-p_j)}$.

See Appendix B.6 for a proof. The first one recovers [23, Example 5] with a simpler proof. The last two are new. These entropies are illustrated in Figure 3. For the last two choices, H_L is strictly concave over $\triangle^{|\mathcal{Y}|}$ and the associated gradient mapping $\nabla(-\mathsf{H}_L)^*$ typically sparse, although it cannot be computed in closed form.

Our proof also readily supports loss functions of the form $L(\boldsymbol{s}; \boldsymbol{e}_k) = \phi(s_k)$ with $\mathsf{dom}(L) = \{\boldsymbol{s} \in \mathbb{R}^{|\mathcal{Y}|} \colon \boldsymbol{s}^\top \mathbf{1} = 0\}$ [53, 59]. It can be checked that we then simply need to replace all occurences of $1 - p_j$ with p_j in Proposition 8.

B Proofs and derivations

B.1 Proof of Proposition 2

The two facts stated in (i) (H is always non-negative and maximized by the uniform distribution) follow directly from Jensen's inequality. Indeed, for any $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$, we have:

- $\mathsf{H}(\boldsymbol{p}) \geq \sum_{j=1}^{|\mathcal{Y}|} p_j \mathsf{H}(\boldsymbol{e}_j) = 0$;
- $\mathsf{H}(\mathbf{1}/|\mathcal{Y}|) = \mathsf{H}\left(\sum_{\boldsymbol{P} \in \mathcal{P}} \frac{1}{|\mathcal{Y}|!} \boldsymbol{P}\boldsymbol{p}\right) \geq \sum_{\boldsymbol{P} \in \mathcal{P}} \frac{1}{|\mathcal{Y}|!} \mathsf{H}(\boldsymbol{P}\boldsymbol{p}) = \mathsf{H}(\boldsymbol{p})$,

where $\mathcal{P}$ is the set of permutation matrices. In the second case, the maximizer is unique due to strict concavity.

Let $\Omega(\boldsymbol{p}) \coloneqq -\mathsf{H}(\boldsymbol{p})$. To see (ii), note that the convexity and differentiability of Ω^* follow from basic Fenchel duality and the fact that Ω is *strictly* convex. It remains to prove that Ω^* is symmetric. In fact, we have $\Omega^*(\boldsymbol{P}\boldsymbol{s}) = \sup_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} (\boldsymbol{P}\boldsymbol{s})^\top \boldsymbol{p} - \Omega(\boldsymbol{p}) = \sup_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \boldsymbol{s}^\top \boldsymbol{P}^\top \boldsymbol{p} - \Omega(\boldsymbol{P}^\top \boldsymbol{p}) = \Omega^*(\boldsymbol{s})$. The last equality was obtained by a change of variable $\boldsymbol{p}' = \boldsymbol{P}^\top \boldsymbol{p}$, from which $\boldsymbol{p}$ is recovered as $\boldsymbol{p} = \boldsymbol{P}\boldsymbol{p}'$, which proves (iii). Let us turn to (iv). Since Ω is convex, so is Ω^*, hence the gradient operator $\nabla\Omega^*$ is monotone, i.e., we have $(\boldsymbol{s}' - \boldsymbol{s})^\top (\boldsymbol{p}' - \boldsymbol{p}) \geq 0$ for any $\boldsymbol{s}, \boldsymbol{s}' \in \mathbb{R}^{|\mathcal{Y}|}$, $\boldsymbol{p} = \nabla\Omega^*(\boldsymbol{s})$ and $\boldsymbol{p}' = \nabla\Omega^*(\boldsymbol{s}')$. Let $\boldsymbol{s}'$ be obtained from $\boldsymbol{s}$ by swapping two coordinates, i.e., $s'_j = s_i$, $s'_i = s_j$, and $s'_k = s_k$ for any $k \notin \{i, j\}$. Then, since Ω is symmetric, we obtain:

$$2(s_j - s_i)(p_j - p_i) \geq 0,$$

which implies $s_i > s_j \Rightarrow p_i \geq p_j$ and $p_i > p_j \Rightarrow s_i \geq s_j$. To fully prove (iv), we need to show that the last inequality is strict: to do this, we simply invoke (iii) with a permutation matrix that permutes i and j, from which we must have $s_i = s_j \Rightarrow p_i = p_j$.

B.2 Proof of Proposition 3

Let $\Omega = -\mathsf{H}$, and assume that Ω has non-empty subdifferential everywhere in $\triangle^{|\mathcal{Y}|}$. We need to show that for any $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$, there is some $\boldsymbol{s} \in \mathbb{R}^{|\mathcal{Y}|}$ such that $\boldsymbol{p} \in \mathsf{argmin}_{\boldsymbol{p}' \in \triangle^{|\mathcal{Y}|}} \Omega(\boldsymbol{p}') - \langle \boldsymbol{s}, \boldsymbol{p}' \rangle$. The Lagrangian associated with this minimization problem is:

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\mu}, \lambda) = \Omega(\boldsymbol{p}) - \langle \boldsymbol{s} + \boldsymbol{\mu}, \boldsymbol{p} \rangle + \lambda(\mathbf{1}^\top \boldsymbol{p} - 1).$$

The KKT conditions are:

$$\begin{cases} 0 \in \partial_p \mathcal{L}(\boldsymbol{p}, \boldsymbol{\mu}, \lambda) = \partial\Omega(\boldsymbol{p}) - \boldsymbol{s} - \boldsymbol{\mu} + \lambda \mathbf{1} \\ \langle \boldsymbol{p}, \boldsymbol{\mu} \rangle = 0 \\ \boldsymbol{p} \in \triangle^{|\mathcal{Y}|}, \ \boldsymbol{\mu} \geq 0. \end{cases}$$

For a given $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$, we seek $\boldsymbol{s}$ such that $(\boldsymbol{p}, \boldsymbol{\mu}, \lambda)$ are a solution to the KKT conditions for some $\boldsymbol{\mu} \geq 0$ and $\lambda \in \mathbb{R}$.

We will show that such $\boldsymbol{s}$ exists by simply choosing $\boldsymbol{\mu} = \mathbf{0}$ and $\lambda = 0$. Those choices are dual feasible and guarantee that the slackness complementary condition is satisfied. In this case, we have from the first condition that $\boldsymbol{s} \in \partial\Omega(\boldsymbol{p})$. From the assumption that Ω has non-empty subdifferential in all the simplex, we have that for any $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$ we can find a $\boldsymbol{s} \in \mathbb{R}^{|\mathcal{Y}|}$ such that $(\boldsymbol{p}, \boldsymbol{s})$ are a dual pair, i.e., $\boldsymbol{p} = \nabla\Omega^*(\boldsymbol{s})$, which proves that $\nabla\Omega^*(\mathbb{R}^{|\mathcal{Y}|}) = \triangle^{|\mathcal{Y}|}$.

To prove item 2, let us now fix $\boldsymbol{p}$ and pick $\boldsymbol{s} \in \partial\Omega(\boldsymbol{p})$. Naturally, $\boldsymbol{s}' = \boldsymbol{s} + \lambda\mathbf{1}$ also maps to $\boldsymbol{p}$. Moreover, let $\mathcal{I} = \{1 \leq i \leq |\mathcal{Y}| \colon p_i = 0\}$. Then, any $\mu_i \geq 0$ for $i \in \mathcal{I}$ also leads to a solution, which shows any $\boldsymbol{s}'$ with $s'_i = s_i$ for $i \notin \mathcal{I}$ and $s'_i \leq s_i$ for $i \in \mathcal{I}$ also maps to $\boldsymbol{p}$.

B.3 Proof of Proposition 4

We start by proving the following lemma.

Lemma 1 *Let* H *satisfy assumptions A.1–A.3. Then:*

1. *We have* $\boldsymbol{s} \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$ *iff* $s_k = (-\mathsf{H})^*(\boldsymbol{s})$. *That is:*
$$\partial(-\mathsf{H})(\boldsymbol{e}_k) = \{\boldsymbol{s} \in \mathbb{R}^{|\mathcal{Y}|} \colon s_k \geq \langle \boldsymbol{s}, \boldsymbol{p} \rangle + \mathsf{H}(\boldsymbol{p}), \ \forall \boldsymbol{p} \in \triangle^{|\mathcal{Y}|}\}.$$
2. *If* $\boldsymbol{s} \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$, *then, we also have* $\boldsymbol{s}' \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$ *for any* $\boldsymbol{s}'$ *such that* $s'_k = s_k$ *and* $s'_i \leq s_i$, *for all* $i \neq k$.

Proof: (of the lemma.) Let $\Omega = -\mathsf{H}$. From Proposition 2, we can consider $\partial\Omega(\boldsymbol{e}_1)$ without loss of generality, in which case any $\boldsymbol{s} \in \partial\Omega(\boldsymbol{e}_1)$ satisfies $s_1 = \mathsf{max}_j \, s_j$. We have $\boldsymbol{s} \in \partial\Omega(\boldsymbol{e}_1)$ iff $\Omega(\boldsymbol{e}_1) = \langle \boldsymbol{s}, \boldsymbol{e}_1 \rangle - \Omega^*(\boldsymbol{s}) = s_1 - \Omega^*(\boldsymbol{s})$. Since $\Omega(\boldsymbol{e}_1) = 0$, we must have $s_1 = \Omega^*(\boldsymbol{s}) \geq \sup_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \langle \boldsymbol{s}, \boldsymbol{p} \rangle - \Omega(\boldsymbol{p})$, which proves part 1. To see 2, note that we have $s'_k = s_k \geq \langle \boldsymbol{s}, \boldsymbol{p} \rangle - \Omega(\boldsymbol{p}) \geq \langle \boldsymbol{s}', \boldsymbol{p} \rangle - \Omega(\boldsymbol{p})$, for all $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$, from which the result follows. ∎

We now proceed to the proof of Proposition 4. Let $\Omega = -\mathsf{H}$, and suppose that L_Ω has the separation margin property. Then, $\boldsymbol{s} = m\boldsymbol{e}_1$ satisfies the margin condition $s_1 \geq m + \mathsf{max}_{j \neq 1} \, s_j$, hence $L_\Omega(m\boldsymbol{e}_1, \boldsymbol{e}_1) = 0$. From the first part of Proposition 1, this implies $m\boldsymbol{e}_1 \in \partial\Omega(\boldsymbol{e}_1)$.

Conversely, let us assume that $m\boldsymbol{e}_1 \in \partial\Omega(\boldsymbol{e}_1)$. From the second part of Proposition 1, this implies that $\boldsymbol{s} \in \partial\Omega(\boldsymbol{e}_1)$ for any $\boldsymbol{s}$ such that $s_1 = m$ and $s_i \leq 0$ for all $i \geq 2$; and more generally we have $\boldsymbol{s} + c\mathbf{1} \in \partial\Omega(\boldsymbol{e}_1)$. That is, any $\boldsymbol{s}$ with $s_1 \geq m + \mathsf{max}_{i \neq 1} \, s_i$ satisfies $\boldsymbol{s} \in \partial\Omega(\boldsymbol{e}_1)$. From Proposition 1, this is equivalent to $L_\Omega(\boldsymbol{s}; \boldsymbol{e}_1) = 0$.

Let us now determine the margin of L_Ω, i.e., the smallest m such that $m\boldsymbol{e}_1 \in \partial\Omega(\boldsymbol{e}_1)$. From Proposition 1, this is equivalent to $m \geq mp_1 - \Omega(\boldsymbol{p})$ for any $\boldsymbol{p} \in \Delta^d$, i.e., $-\Omega(\boldsymbol{p})(1 - p_1) \leq m$. Note that by Proposition 2 the "most competitive" $\boldsymbol{p}$'s are sorted as $\boldsymbol{e}_1$, so we may write $p_1 = \|\boldsymbol{p}\|_\infty$ without loss of generality. The margin of L_Ω is the smallest possible such margin, given by (7).

B.4 Proof of Proposition 5

Define $\Omega = -\mathsf{H}$. Let us start by writing the margin expression (7) as a unidimensional optimization problem. This is done by noticing that the max-generalized entropy problem constrained to $\mathsf{max}(\boldsymbol{p}) = 1 - t$ gives

$\boldsymbol{p} = \left[1-t, \frac{t}{d-1}, \ldots, \frac{t}{d-1}\right]$, for $t \in \left[0, 1-\frac{1}{d}\right]$ by a similar argument as the one used in Proposition 2, item 1. We obtain:

$$\mathrm{margin}(L_\Omega) = \sup_{t \in \left[0, 1-\frac{1}{d}\right]} \frac{-\Omega\left(\left[1-t, \frac{t}{d-1}, \ldots, \frac{t}{d-1}\right]\right)}{t}.$$

We write the argument above as $A(t) = \frac{-\Omega(\boldsymbol{e}_1 + t\boldsymbol{v})}{t}$, where $\boldsymbol{v} := [-1, \frac{1}{d-1}, \ldots, \frac{1}{d-1}]$. We will first prove that A is decreasing in $[0, 1-\frac{1}{d}]$, which implies that the supremum (and the margin) equals $A(0)$. Note that we have the following expression for the derivative of any function $f(\boldsymbol{e}_1 + t\boldsymbol{v})$:

$$(f(\boldsymbol{e}_1 + t\boldsymbol{v}))' = \boldsymbol{v}^\top \nabla f(\boldsymbol{e}_1 + t\boldsymbol{v}).$$

Using this fact, we can write the derivative $A'(t)$ as:

$$A'(t) = \frac{-t\boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}) + \Omega(\boldsymbol{e}_1 + t\boldsymbol{v})}{t^2} := \frac{B(t)}{t^2}.$$

In turn, the derivative $B'(t)$ is:

$$\begin{aligned} B'(t) &= -\boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}) - t(\boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}))' + \boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}) \\ &= -t(\boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}))' \\ &= -t\boldsymbol{v}^\top \nabla\nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v})\boldsymbol{v} \\ &\leq 0, \end{aligned}$$

where we denote by $\nabla\nabla\Omega$ the Hessian of Ω, and used the fact that it is positive semi-definite, due to the convexity of Ω. This implies that B is decreasing, hence for any $t \in [0,1]$, $B(t) \leq B(0) = \Omega(\boldsymbol{e}_1) = 0$, where we used the fact $\|\nabla\Omega(\boldsymbol{e}_1)\| < \infty$, assumed as a condition of Proposition 3. Therefore, we must also have $A'(t) = \frac{B(t)}{t^2} \leq 0$ for any $t \in [0,1]$, hence A is decreasing, and $\sup_{t\in[0,1-1/d]} A(t) = \lim_{t\to 0+} A(t)$. By L'Hôpital's rule:

$$\begin{aligned} \lim_{t\to 0+} A(t) &= \lim_{t\to 0+} (-\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}))' \\ &= -\boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1) \\ &= \nabla_1\Omega(\boldsymbol{e}_1) - \frac{1}{d-1}\sum_{j\geq 2} \nabla_j\Omega(\boldsymbol{e}_1) \\ &= \nabla_1\Omega(\boldsymbol{e}_1) - \nabla_2\Omega(\boldsymbol{e}_1), \end{aligned}$$

which proves the first part.

If Ω is separable, then $\nabla\Omega(\boldsymbol{p}) = -\sum_{i=1}^d h(p_i)$, from which we obtain $\nabla_1\Omega(\boldsymbol{e}_1) = -h'(1)$ and $\nabla_2\Omega(\boldsymbol{e}_1) = -h'(0)$, yielding $\mathrm{margin}(L_\Omega) = h'(0) - h'(1)$. Since h is twice differentiable, this equals $-\int_0^1 h''(t)dt$, completing the proof.

B.5 Proof of Proposition 7

Let H be a separable entropy on the simplex, i.e., $\mathsf{H}(\boldsymbol{p}) = \sum_{j=1}^{|\mathcal{Y}|} h(p_j)$, where $h : [0,1] \to \mathbb{R}_+$ is a non-negative, strictly concave, differentiable function such that $h(0) = h(1) = 0$, and $\mathsf{dom}(\mathsf{H}) = \triangle^{|\mathcal{Y}|}$; impying, importantly, that h' is invertible. As discussed in §3, such a H satisfies assumptions A.1–A.3. We show how computing $\nabla(-\mathsf{H})^*$ reduces to finding the root of a monotonic scalar function, for which efficient algorithms exist.

Recall the optimization problem we are trying to solve

$$\nabla(-\mathsf{H})^* := \operatorname*{\mathsf{argmax}}_{\boldsymbol{p}\in\triangle^{|\mathcal{Y}|}} \langle \boldsymbol{p}, \boldsymbol{s}\rangle + \mathsf{H}(\boldsymbol{p}).$$

The Lagrangian is

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\mu}, \tau) := -\mathsf{H}(\boldsymbol{p}) - \langle \boldsymbol{s} + \boldsymbol{\mu}, \boldsymbol{p}\rangle + \tau(\mathbf{1}^\top \boldsymbol{p} - 1).$$

The unique solution $(\boldsymbol{p}^\star, \boldsymbol{\mu}^\star, \tau^\star)$ must satisfy the KKT conditions, which, in the separable case, are

$$\begin{aligned} h'(p_i) + s_i - \tau + \mu_i = 0 \qquad & \forall 1 \leq i \leq |\mathcal{Y}| \\ \langle \boldsymbol{p}, \boldsymbol{\mu}\rangle = 0 & \\ \boldsymbol{p} \in \triangle^{|\mathcal{Y}|},\ \boldsymbol{\mu} \geq 0. & \end{aligned}$$

Let us define

$$\tau_{\mathsf{min}} \coloneqq \mathsf{max}(\boldsymbol{s}) + h'(1) \quad \text{and} \quad \tau_{\mathsf{max}} \coloneqq \mathsf{max}(\boldsymbol{s}) + h'\left(\frac{1}{|\mathcal{Y}|}\right).$$

Since h is strictly concave, h' is decreasing and so $\tau_{\mathsf{min}} < \tau_{\mathsf{max}}$. For any $\tau \in [\tau_{\mathsf{min}}, \tau_{\mathsf{max}}]$, we construct $\boldsymbol{\mu}$ as

$$\mu_i \coloneqq \begin{cases} 0, & s_i - \tau \geq -h'(0) \\ -h'(0) - s_i + \tau, & s_i - \tau < -h'(0) \end{cases}$$

By construction, $\mu_i \geq 0$, satisfying dual feasability. Injecting μ_i in $-h'(p_i) = s_i + \mu_i - \tau$ and combining the two cases, we obtain

$$-h'(p_i) = \mathsf{max}\{s_i - \tau, \ -h'(0)\}. \tag{17}$$

We show that i) the stationarity conditions have a unique solution given τ, and ii) $[\tau_{\mathsf{min}}, \tau_{\mathsf{max}}]$ forms a bracketing interval containing the primal-feasible solution, which can then be found by bisection. Since any such point verifies the KKT conditions, it must be the global solution.

Solving the stationarity conditions. Since $-h$ is strictly convex, its derivative $-h'$ is continuous and strictly increasing, and is thus a one-to-one mapping between $[0, 1]$ and $[-h'(0), -h'(1)]$. Denote by $(-h')^{-1} : [-h'(0), -h'(1)] \to [0, 1]$ its inverse. If $s_i - \tau \geq -h'(0)$, we have that

$$\begin{aligned} -h'(0) \leq -h'(p_i) = s_i - \tau &\leq \mathsf{max}(\boldsymbol{s}) - \tau_{\mathsf{min}} \\ &= \mathsf{max}(\boldsymbol{s}) - \mathsf{max}(\boldsymbol{s}) - h'(1) \\ &= -h'(1). \end{aligned}$$

Otherwise, $-h'(p_i) = -h'(0)$. This verifies that the r.h.s. of Eq. (17) is always within the domain of $(-h')^{-1}$. We can thus apply the inverse to both sides to solve for p_i, obtaining

$$p_i(\tau) = (-h')^{-1}(\mathsf{max}\{s_i - \tau, \ -h'(0)\}). \tag{18}$$

Validating the bracketing interval. Consider the primal feasability function $f(\tau) = \mathbf{1}^\top \boldsymbol{p}(\tau) - 1$; $\boldsymbol{p}(\tau)$ is primal feasible iff τ is a root of f. We show that f is decreasing on $[\tau_{\mathsf{min}}, \tau_{\mathsf{max}}]$, and that it has opposite signs at the two extremities. From the intermediate value theorem, the unique root $\tau^\star$ must satisfy $\tau^\star \in [\tau_{\mathsf{min}}, \tau_{\mathsf{max}}]$.

As the inverse of an increasing function, $(-h')^{-1}$ is increasing. Therefore, for all i, $p_i(\tau)$ is decreasing, and so is the sum $f(\tau) = \sum_i p_i(\tau) - 1$. It remains to check the signs at the boundaries.

$$\begin{aligned} \sum_i p_i(\tau_{\mathsf{max}}) &= \sum_i (-h')^{-1}(\mathsf{max}\{s_i - \mathsf{max}(\boldsymbol{s}) - h'\left(1/|\mathcal{Y}|\right), \ -h'(0)\}) \\ &\leq |\mathcal{Y}|(-h')^{-1}(\mathsf{max}\{-h'\left(1/|\mathcal{Y}|\right), \ -h'(0)\}) \\ &= |\mathcal{Y}|(-h')^{-1}\left(-h'\left(1/|\mathcal{Y}|\right)\right) = 1 \end{aligned}$$

upper-bounding each term of the sum by the largest one. At the other end,

$$\begin{aligned} \sum_i p_i(\tau_{\mathsf{min}}) &= \sum_i (-h')^{-1}(\mathsf{max}\{s_i - \mathsf{max}(\boldsymbol{s}) - h'(1), \ -h'(0)\}) \\ &\geq (-h')^{-1}(\mathsf{max}\{-h'(1), \ -h'(0)\}) \\ &= (-h')^{-1}(-h'(1)) = 1 \end{aligned}$$

using that a sum of non-negative terms is no less than its largest term. Therefore, $f(\tau_{\mathsf{min}}) \geq 0$ and $f(\tau_{\mathsf{max}}) \leq 0$. This implies that there must exist $\tau^\star$ in $[\tau_{\mathsf{min}}, \tau_{\mathsf{max}}]$ satisfying $f(\tau^\star) = 0$. A corresponding triplet $(\boldsymbol{p}(\tau^\star), \boldsymbol{\mu}(\tau^\star), \tau^\star)$ thus satisfies all of the KKT conditions, confirming that it is the globally optimal solution.

Algorithm 1 is an example of a bisection algorithm for finding the optimal solution; more advanced root finding methods can also be used. We note that the resulting algorithm resembles the method provided in [33], with a non-trivial difference being the order of the thresholding and $(-h)^{-1}$ in Eq. (18). Our result thus applies for important cases like the Tsallis entropy with $\alpha = 1.5$, where [33] is not applicable.

Algorithm 1 Compute $\nabla(-\mathsf{H})^*(\boldsymbol{s})$

Input: $\boldsymbol{s} \in \mathbb{R}^{|\mathcal{Y}|}$, $\mathsf{H}(\boldsymbol{p}) = \sum_i h(p_i)$
$\widehat{\boldsymbol{p}}(\tau) \coloneqq (-h')^{-1}(\max\{\boldsymbol{s} - \tau, -h'(0)\})$
$f(\tau) \coloneqq \widehat{\boldsymbol{p}}(\tau)^\top \mathbf{1} - 1$
$\tau_{\min} \leftarrow \max(\boldsymbol{s}) + h'(1)$;
$\tau_{\max} \leftarrow \max(\boldsymbol{s}) + h'\left(1/|\mathcal{Y}|\right)$
$\tau \leftarrow (\tau_{\min} + \tau_{\max})/2$
while $|f(\tau)| > \epsilon$
  $\tau \leftarrow (\tau_{\min} + \tau_{\max})/2$
  if $f(\tau) < 0$ $\tau_{\max} \leftarrow \tau$
  else $\tau_{\min} \leftarrow \tau$
Output: $\nabla(-\mathsf{H})^*(\boldsymbol{s}) \approx \widehat{\boldsymbol{p}}(\tau)$

B.6 Proof of Proposition 8

We first need the following lemma.

Lemma 2 *Let $L(\boldsymbol{s}; \boldsymbol{e}_k)$ be defined as*

$$L(\boldsymbol{s}; \boldsymbol{e}_k) \coloneqq \begin{cases} \sum_j c_{k,j}\phi(s_j) & \text{if } \boldsymbol{s}^\top \mathbf{1} = 0 \\ \infty & \text{o.w.} \end{cases},$$

where $\phi\colon \mathbb{R} \to \mathbb{R}_+$ is convex. Then, H_L defined in (12) *equals*

$$-\mathsf{H}_L(\boldsymbol{p}) = \min_{\tau \in \mathbb{R}} \sum_j (\boldsymbol{p}^\top \boldsymbol{c}_j)\phi^*\left(\frac{-\tau}{\boldsymbol{p}^\top \boldsymbol{c}_j}\right).$$

Proof: We want to solve

$$\mathsf{H}_L(\boldsymbol{p}) = \min_{\boldsymbol{s}^\top \mathbf{1} = 0} \sum_j p_j L(\boldsymbol{s}; \boldsymbol{e}_j) = \min_{\boldsymbol{s}^\top \mathbf{1} = 0} \sum_j p_j \sum_i c_{j,i}\phi(s_i) = \min_{\boldsymbol{s}^\top \mathbf{1} = 0} \sum_i (\boldsymbol{p}^\top \boldsymbol{c}_i)\phi(s_i),$$

where $\boldsymbol{c}_i$ is a vector gathering $c_{j,i}$ for all j. Introducing a Lagrange multiplier we get

$$\mathsf{H}_L(\boldsymbol{p}) = \min_{\boldsymbol{s} \in \mathbb{R}^{|\mathcal{Y}|}} \max_{\tau \in \mathbb{R}} \sum_i (\boldsymbol{p}^\top \boldsymbol{c}_i)\phi(s_i) + \tau \sum_i s_i.$$

Strong duality holds and we can swap the order of the min and max. Routine calculations then give

$$-\mathsf{H}_L(\boldsymbol{p}) = \min_{\tau \in \mathbb{R}} \sum_i (\boldsymbol{p}^\top \boldsymbol{c}_i)\phi^*\left(\frac{-\tau}{\boldsymbol{p}^\top \boldsymbol{c}_i}\right). \tag{19}$$

■

We now prove Proposition 8. First, we rewrite $L(\boldsymbol{s}; \boldsymbol{e}_k) = \sum_{j \neq k} \phi(s_j - s_k)$ as $L(\boldsymbol{s}; \boldsymbol{e}_k) = \sum_j c_{k,j}\phi(s_j - s_k)$, where we choose $c_{k,j} = 0$ if $k = j$, 1 otherwise, leading to $\boldsymbol{p}^\top \boldsymbol{c}_j = 1 - p_j$. Because $\phi(s_j - s_k)$ is shift invariant w.r.t. $\boldsymbol{s}$, without loss of generality, we can further rewrite the loss as $L(\boldsymbol{s}; \boldsymbol{e}_k) = \sum_j c_{k,j}\phi(s_j)$ with $\mathsf{dom}(L) = \{\boldsymbol{s} \in \mathbb{R}^{|\mathcal{Y}|}\colon \boldsymbol{s}^\top \mathbf{1} = 0\}$. Hence, Lemma 2 applies. We now derive closed form for specific choices of ϕ (we let $d \coloneqq |\mathcal{Y}|$ for convenience).

Hinge loss. When $\phi(t) = [1 + t]_+$, the conjugate is

$$\phi^*(u) = \begin{cases} -u & \text{if } u \in [0, 1] \\ \infty & \text{o.w.} \end{cases}.$$

The constraint set for τ is therefore $\mathcal{C} \coloneqq \bigcap_{j\in[d]} \left[-\boldsymbol{p}^\top \boldsymbol{c}_j, 0\right] = \left[-\min_{j\in[d]} \boldsymbol{p}^\top \boldsymbol{c}_j, 0\right]$. Hence

$$-\mathsf{H}_L(\boldsymbol{p}) = \min_{\tau\in\mathcal{C}} d\tau = -d \min_{j\in[d]} \boldsymbol{p}^\top \boldsymbol{c}_j.$$

This recovers [23, Example 5, §A.6] with a simpler proof. We next turn to the following new results.

Smoothed hinge loss. We add quadratic regularization to the conjugate [50]:

$$\phi^*(u) = \begin{cases} -u + \frac{1}{2}u^2 & \text{if } u \in [0,1] \\ \infty & \text{o.w.} \end{cases}.$$

Going back to ϕ, we obtain:

$$\phi(t) = \begin{cases} 0 & \text{if } t \le -1 \\ 1 + t - \frac{\gamma}{2} & \text{if } t \ge \gamma - 1 \\ \frac{1}{2\gamma}(1+t)^2 & \text{o.w.} \end{cases}. \tag{20}$$

The constraint set for τ is the same as before, $\mathcal{C} = \left[-\min_{j\in[d]} \boldsymbol{p}^\top \boldsymbol{c}_j, 0\right]$.

Plugging ϕ^* into (19), we obtain

$$-\mathsf{H}_L(\boldsymbol{p}) = \min_{\tau\in\mathcal{C}} \frac{\tau^2}{2} \sum_{j=1}^{d} \frac{1}{\boldsymbol{p}^\top \boldsymbol{c}_j} + d\,\tau. \tag{21}$$

Since the problem is unidimensional, let us solve for τ unconstrained first:

$$\tau = -d \Big/ (\sum_{j=1}^{d} 1/(\boldsymbol{p}^\top \boldsymbol{c}_j)). \tag{22}$$

We notice that $\tau \le -\min_j \boldsymbol{p}^\top \boldsymbol{c}_j$ for $\boldsymbol{c}_j \ge \mathbf{0}$ since $\sum_j \frac{\min_i \boldsymbol{p}^\top \boldsymbol{c}_i}{\boldsymbol{p}^\top \boldsymbol{c}_j} \in [0, d]$. This expression of τ is not feasible. Hence the optimal solution is at the boundary and $\tau^\star = -\min_j \boldsymbol{p}^\top \boldsymbol{c}_j$. Plugging that expression back into (21) gives the claimed expression of H_L.

Squared hinge loss. When $\phi(t) = \frac{1}{2}[1+t]_+^2$, the conjugate is

$$\phi^*(u) = \begin{cases} -u + \frac{1}{2}u^2 & \text{if } u \ge 0 \\ \infty & \text{o.w.} \end{cases}.$$

The constraint is now $\tau \le 0$. Hence, the optimal solution of the problem w.r.t. τ in (19) is now (22) for all $\boldsymbol{p} \in \triangle^d, \boldsymbol{c}_j \ge \mathbf{0}$. Simplifying, we get

$$\mathsf{H}_L(\boldsymbol{p}) = \frac{\frac{1}{2}d^2}{\sum_{j=1}^{d} 1/(\boldsymbol{p}^\top \boldsymbol{c}_j)}.$$

B.7 Derivation of the dual coordinate ascent algorithm

We rewrite the primal objective (9) as

$$\min_{W, \{\boldsymbol{\theta}_i\}} \sum_{i=1}^{n} L_\Omega(\boldsymbol{\theta}_i; \boldsymbol{y}_i) + G(W) \text{ s.t. } \boldsymbol{\theta}_i = W\boldsymbol{x}_i \; \forall i \in [n].$$

After routine calculations, we find that its dual is

$$-\min_{\alpha} \sum_{i=1}^{n} L_\Omega^*(-\boldsymbol{\alpha}_i; \boldsymbol{y}_i) + G^*\left(\sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{x}_i^\top\right).$$

Using $L^*_\Omega(-\boldsymbol{\alpha}; \boldsymbol{y}) = \Omega(\boldsymbol{y} - \boldsymbol{\alpha}) - \Omega(\boldsymbol{y})$ and using the change of variable $\boldsymbol{\beta}_i \coloneqq \boldsymbol{y}_i - \boldsymbol{\alpha}_i$, we get

$$-\min_{\beta} \sum_{i=1}^{n} \Omega(\boldsymbol{\beta}_i) - \Omega(\boldsymbol{y}_i) + G^*(V(\beta)) \text{ s.t. } \boldsymbol{\beta}_i \in \mathsf{dom}(\Omega)\ \forall i \in [n],$$

where $V(\beta) \coloneqq \sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\beta}_i)\boldsymbol{x}_i^\top$. The sub-problem associated with $i \in [n]$ is

$$\operatorname*{argmin}_{\boldsymbol{\beta}_i \in \mathsf{dom}(\Omega)} \Omega(\boldsymbol{\beta}_i) + G^*(V(\beta)). \tag{23}$$

Since this problem could be hard to solve, we follow [50, Option I] and consider instead an upper-bound. Since G^* is $\frac{1}{\lambda}$-smooth w.r.t. the dual norm $\|\cdot\|$, it holds that

$$G^*(V(\beta)) \leq G^*(V(\bar{\beta})) + \langle \nabla G^*(V(\bar{\beta})), V(\beta) - V(\bar{\beta}) \rangle + \frac{1}{2\lambda}\|V(\beta) - V(\bar{\beta})\|^2,$$

where $\bar{\beta}$ denotes the current iterate of β. Using $V(\beta) - V(\bar{\beta}) = \sum_{i=1}^n (\bar{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)\boldsymbol{x}_i^\top$, we get

$$G^*(V(\beta)) \leq G^*(V(\bar{\beta})) + \sum_{i=1}^{n} \langle \nabla G^*(V(\bar{\beta}))\boldsymbol{x}_i, \bar{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i \rangle + \frac{\sigma_i}{2}\|\bar{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|^2,$$

where $\sigma_i \coloneqq \frac{\|\boldsymbol{x}_i\|^2}{\lambda}$. Substituting $G^*(V(\beta))$ by the above upper bound into (23) and ignoring constant terms, we get the following approximate sub-problem:

$$\operatorname*{argmin}_{\boldsymbol{\beta}_i \in \mathsf{dom}(\Omega)} \Omega(\boldsymbol{\beta}_i) - \boldsymbol{\beta}_i^\top \boldsymbol{v}_i + \frac{\sigma_i}{2}\|\boldsymbol{\beta}_i\|^2 = \mathrm{prox}_{\frac{1}{\sigma_i}\Omega}\left(\frac{\boldsymbol{v}_i}{\sigma_i}\right), \tag{24}$$

where $\boldsymbol{v}_i \coloneqq \nabla G^*(V(\bar{\beta}))\boldsymbol{x}_i + \sigma_i \bar{\boldsymbol{\beta}}_i$ and where we defined the proximity operator

$$\mathrm{prox}_{\tau\Omega}(\boldsymbol{x}) \coloneqq \operatorname*{argmin}_{\boldsymbol{y} \in \mathsf{dom}(\Omega)} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + \tau\Omega(\boldsymbol{y}).$$

Note that when G^* is a quadratic function, (24) is an optimal solution of (23).

Summary. A block-wise update is $\boldsymbol{\beta}_i \leftarrow \mathrm{prox}_{\frac{1}{\sigma_i}\Omega}(\boldsymbol{v}_i/\sigma_i)$, where $\boldsymbol{v}_i \coloneqq \nabla G^*(V(\beta))\boldsymbol{x}_i + \sigma_i\boldsymbol{\beta}_i$ and $\sigma_i \coloneqq \frac{\|\boldsymbol{x}_i\|^2}{\lambda}$. A block coordinate ascent algorithm picks $i \in [n]$, performs this update and repeats, until converging to an optimal dual solution $\beta^\star$. Since G is a λ-strongly convex regularizer, given $\beta^\star$, we retrieve the optimal primal solution by $W^\star = \nabla G^*(V(\beta^\star))$.

Examples of regularization. When $G(W) = \frac{\lambda}{2}\|W\|_F^2$, $G^*(V) = \frac{1}{2\lambda}\|V\|_F^2$ and $\nabla G^*(V) = \frac{1}{\lambda}V$.

When $G(W) = \frac{\lambda}{2}\|W\|_F^2 + \lambda\rho R(W)$, for some regularizer $R(W)$, we have

$$\begin{aligned}
\nabla G^*(V) &= \operatorname*{argmax}_W \langle W, V \rangle - \frac{\lambda}{2}\|W\|_F^2 - \lambda\rho R(W) \\
&= \operatorname*{argmin}_W \frac{1}{2}\left\|W - \frac{V}{\lambda}\right\|_F^2 + \rho R(W) \\
&= \mathrm{prox}_{\rho R}(V/\lambda).
\end{aligned}$$

The conjugate is equal to $G^*(V) = \langle \nabla G^*(V), V \rangle - G(\nabla G^*(V))$.

For instance, if we choose $R(W) = \|W\|_1 \coloneqq \sum_{i,j} |w_{i,j}|$, then $G(W)$ is the **elastic-net** regularization and $\mathrm{prox}_{\rho R}$ is the well-known **soft-thresholding** operator

$$\mathrm{prox}_{\rho R}(V) = \mathrm{sign}(V)[|V| - \rho]_+,$$

where all operations above are performed element-wise.

Examples of losses. For the squared loss, $\mathsf{dom}(\Omega) = \mathbb{R}^d$ and $\Omega(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|^2$. Hence:

$$\mathrm{prox}_{\tau\Omega}(\boldsymbol{x}) = \operatorname*{\mathsf{argmin}}_{\boldsymbol{y}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 + \frac{\tau}{2}\|\boldsymbol{y}\|^2 = \frac{\boldsymbol{x}}{\tau+1}.$$

For the perceptron loss [47, 15], $\mathsf{dom}(\Omega) = \triangle^{|\mathcal{Y}|}$ and $\Omega = 0$. Hence:

$$\mathrm{prox}_{\tau\Omega}(\boldsymbol{x}) = \operatorname*{\mathsf{argmin}}_{\boldsymbol{p}\in\triangle^{|\mathcal{Y}|}} \frac{1}{2}\|\boldsymbol{p}-\boldsymbol{x}\|^2.$$

For the sparsemax loss [37], $\mathsf{dom}(\Omega) = \triangle^{|\mathcal{Y}|}$ and $\Omega(\boldsymbol{p}) = \frac{1}{2}(\|\boldsymbol{p}\|^2 - 1)$. Hence:

$$\mathrm{prox}_{\tau\Omega}(\boldsymbol{x}) = \operatorname*{\mathsf{argmin}}_{\boldsymbol{p}\in\triangle^{|\mathcal{Y}|}} \frac{1}{2}\|\boldsymbol{p}-\boldsymbol{x}\|^2 + \frac{\tau}{2}\|\boldsymbol{p}\|^2 = \operatorname*{\mathsf{argmin}}_{\boldsymbol{p}\in\triangle^{|\mathcal{Y}|}} \frac{1}{2}\left\|\boldsymbol{p} - \frac{\boldsymbol{x}}{\tau+1}\right\|^2.$$

For the cost-sensitive multiclass hinge loss, $\mathsf{dom}(\Omega) = \triangle^{|\mathcal{Y}|}$ and since $\Omega^*(\boldsymbol{\theta} + \boldsymbol{c}_{\boldsymbol{y}}) = \max_{\boldsymbol{y}'\in\mathcal{Y}} \theta(\boldsymbol{y}') + c_{\boldsymbol{y}}(\boldsymbol{y}')$, we have $\Omega(\boldsymbol{p}) = -\langle \boldsymbol{p}, \boldsymbol{c}_{\boldsymbol{y}}\rangle$. Hence:

$$\mathrm{prox}_{\tau\Omega}(\boldsymbol{x}) = \operatorname*{\mathsf{argmin}}_{\boldsymbol{p}\in\triangle^{|\mathcal{Y}|}} \frac{1}{2}\|\boldsymbol{p}-\boldsymbol{x}\|^2 - \tau\langle \boldsymbol{p}, \boldsymbol{c}_{\boldsymbol{y}}\rangle = \operatorname*{\mathsf{argmin}}_{\boldsymbol{p}\in\triangle^{|\mathcal{Y}|}} \frac{1}{2}\|\boldsymbol{p}-(\boldsymbol{x}+\tau\boldsymbol{c}_{\boldsymbol{y}})\|^2.$$

Choosing $\boldsymbol{c}_{\boldsymbol{y}} = \mathbf{1} - \boldsymbol{y}$, where $\boldsymbol{y} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{|\mathcal{Y}|}\}$, recovers the usual multiclass hinge loss [17].

For the multiclass logistic loss, $\mathsf{dom}(\Omega) = \triangle^{|\mathcal{Y}|}$ and Ω is the negative Shannon entropy. Unlike $\nabla\Omega^*$, $\mathrm{prox}_{\tau\Omega}(\boldsymbol{x})$ cannot be computed in closed form but we can compute it approximately using a few iterations of proximal gradient descent.