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

Mathieu Blondel
NTT CS laboratories
Kyoto, Japan

André F.T. Martins
Unbabel, Instituto de Telecomunicações
Lisbon, Portugal

Vlad Niculae
Instituto de Telecomunicações
Lisbon, Portugal

Abstract

This paper studies Fenchel-Young losses, a generic way to *construct* convex loss functions from a regularization function. We analyze their properties in depth, showing that they unify many well-known loss functions and allow to create useful new ones easily. Fenchel-Young losses constructed from a generalized entropy, including the Shannon and Tsallis entropies, induce predictive probability distributions. We formulate conditions for a generalized entropy to yield losses with a separation margin, and probability distributions with sparse support. Finally, we derive efficient algorithms, making Fenchel-Young losses appealing both in theory and practice.

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. 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 (Rosenblatt, 1958) and many other generalization bounds (Vapnik, 1998; Schölkopf and Smola, 2002). For probabilistic classification, the most popular loss is arguably the (multinomial) logistic loss. It is smooth, enabling fast convergence rates, and the softmax operator provides a consistent mapping to probability distributions. However, the logistic loss does not enjoy a margin, and the generated probability distributions have **dense** support, which is undesirable in some applications for interpretability or computational efficiency reasons.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

To address these shortcomings, Martins and Astudillo (2016) proposed a new loss based on the projection onto the simplex. Unlike the logistic loss, this "sparsemax" loss has a natural separation margin and induces a **sparse** probability distribution. However, the sparsemax loss was derived in a relatively ad-hoc manner and it is still relatively poorly understood. Thorough understanding of the core principles underpinning these losses, enabling the creation of new losses combining their strengths, is still lacking.

This paper studies and extends Fenchel-Young (F-Y) losses, recently proposed for structured prediction (Niculae et al., 2018). We show that F-Y losses provide a generic and principled way to **construct** a loss with an associated probability distribution. We uncover a fundamental connection between generalized entropies, margins, and sparse probability distributions. In sum, we make the following contributions.

- We introduce regularized prediction functions to generalize the softmax and sparsemax transformations, possibly beyond the probability simplex (§2).
- We study F-Y losses and their properties, showing that they unify many existing losses, including the hinge, logistic, and sparsemax losses (§3).
- We then show how to seamlessly create entire new families of losses from generalized entropies. We derive efficient algorithms to compute the associated probability distributions, making such losses appealing both in theory and in practice (§4).
- We characterize which entropies yield sparse distributions and losses with a separation margin, notions we prove to be intimately connected (§5).
- Finally, we demonstrate F-Y losses on the task of sparse label proportion estimation (§6).

Notation. We denote the probability simplex by $\triangle^d := \{\boldsymbol{p} \in \mathbb{R}^d_+ \colon \|\boldsymbol{p}\|_1 = 1\}$, the domain of $\Omega \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ by $\mathsf{dom}(\Omega) := \{\boldsymbol{p} \in \mathbb{R}^d \colon \Omega(\boldsymbol{p}) < \infty\}$, the Fenchel conjugate of Ω by $\Omega^*(\boldsymbol{\theta}) := \sup_{\boldsymbol{p} \in \mathsf{dom}(\Omega)} \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle - \Omega(\boldsymbol{p})$, the indicator function of a set $\mathcal{C}$ by $I_{\mathcal{C}}$.

2 Regularized prediction functions

We consider a general predictive setting with input $\boldsymbol{x} \in \mathcal{X}$, and a parametrized model $\boldsymbol{f}_W : \mathcal{X} \to \mathbb{R}^d$, producing a score vector $\boldsymbol{\theta} \coloneqq \boldsymbol{f}_W(\boldsymbol{x})$. To map $\boldsymbol{\theta}$ to predictions, we introduce regularized prediction functions.

Definition 1 *Regularized prediction function*

Let $\Omega \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *be a regularization function, with* $\mathsf{dom}(\Omega) \subseteq \mathbb{R}^d$. *The prediction function regularized by* Ω *is defined by*

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \underset{\boldsymbol{p} \in \mathsf{dom}(\Omega)}{\mathsf{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle - \Omega(\boldsymbol{p}). \tag{1}$$

We emphasize that the regularization is w.r.t. the output and not w.r.t. the model parameters W, as is usually the case in the literature. The optimization problem in (1) balances between two terms: an "affinity" term $\langle \boldsymbol{\theta}, \boldsymbol{p} \rangle$, and a "confidence" term $\Omega(\boldsymbol{p})$ which should be low if $\boldsymbol{p}$ is "uncertain". Two important classes of convex Ω are (squared) norms and, when $\mathsf{dom}(\Omega)$ is the probability simplex, generalized negative entropies. However, our framework does not require Ω to be convex in general. Allowing extended-real Ω further permits general domain constraints in (1) via indicator functions, as we now illustrate.

Examples. When $\Omega = I_{\triangle^d}$, $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is a one-hot representation of the argmax prediction

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \underset{\boldsymbol{p} \in \triangle^d}{\mathsf{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle = \underset{\boldsymbol{y} \in \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_d\}}{\mathsf{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle.$$

We can see that output as a probability distribution that assigns all probability mass on the same class. When $\Omega = -\mathsf{H}^{\mathrm{s}} + I_{\triangle^d}$, where $\mathsf{H}^{\mathrm{s}}(\boldsymbol{p}) \coloneqq -\sum_i p_i \log p_i$ is Shannon's entropy, $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is the well-known softmax

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \mathsf{softmax}(\boldsymbol{\theta}) \coloneqq \frac{\exp(\boldsymbol{\theta})}{\sum_{j=1}^d \exp(\theta_j)}.$$

See Boyd and Vandenberghe (2004, Ex. 3.25) for a derivation. The resulting distribution always has **dense** support. When $\Omega = \frac{1}{2}\|\cdot\|^2 + I_{\triangle^d}$, $\widehat{\boldsymbol{y}}_\Omega$ is the Euclidean projection onto the probability simplex

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \mathsf{sparsemax}(\boldsymbol{\theta}) \coloneqq \underset{\boldsymbol{p} \in \triangle^d}{\mathsf{argmin}} \|\boldsymbol{p} - \boldsymbol{\theta}\|^2,$$

a.k.a. the sparsemax transformation (Martins and Astudillo, 2016). The distribution has **sparse** support (it may assign exactly zero probability to low-scoring classes) and can be computed **exactly** in $O(d)$ time (Brucker, 1984; Duchi et al., 2008; Condat, 2016). This paradigm is not limited to the probability simplex: When $\Omega(\boldsymbol{p}) = -\sum_i \mathsf{H}^{\mathrm{s}}([p_i, 1 - p_i]) + I_{[0,1]^d}(\boldsymbol{p})$, we get

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \mathsf{sigmoid}(\boldsymbol{\theta}) \coloneqq \frac{\mathbf{1}}{\mathbf{1} + \exp(-\boldsymbol{\theta})},$$

i.e., the sigmoid function evaluated coordinate-wise. We can think of its output as a positive measure (unnormalized probability distribution).

Properties. We now discuss simple properties of regularized prediction functions. The first two assume that Ω is a symmetric function, i.e., that it satisfies

$$\Omega(\boldsymbol{p}) = \Omega(\boldsymbol{P}\boldsymbol{p}) \quad \forall \boldsymbol{p} \in \mathsf{dom}(\Omega), \forall \boldsymbol{P} \in \mathcal{P},$$

where $\mathcal{P}$ is the set of $d \times d$ permutation matrices.

Proposition 1 *Properties of* $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$

1. **Effect of a permutation.** *If* Ω *is symmetric, then* $\forall \boldsymbol{P} \in \mathcal{P}$: $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{P}\boldsymbol{\theta}) = \boldsymbol{P}\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$.
2. **Order preservation.** *Let* $\boldsymbol{p} = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$. *If* Ω *is symmetric, then the coordinates of* $\boldsymbol{p}$ *and* $\boldsymbol{\theta}$ *are sorted the same way, i.e.,* $\theta_i > \theta_j \Rightarrow p_i \geq p_j$ *and* $p_i > p_j \Rightarrow \theta_i > \theta_j$.
3. **Gradient mapping.** $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ *is a subgradient of* Ω^* *at* $\boldsymbol{\theta}$, *i.e.,* $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \partial\Omega^*(\boldsymbol{\theta})$. *If* Ω *is strictly convex,* $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ *is the gradient of* Ω^*, *i.e.,* $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta})$.
4. **Temperature scaling.** *For any constant* $t > 0$, $\widehat{\boldsymbol{y}}_{t\Omega}(\boldsymbol{\theta}) \in \partial\Omega^*(\boldsymbol{\theta}/t)$. *If* Ω *is strictly convex,* $\widehat{\boldsymbol{y}}_{t\Omega}(\boldsymbol{\theta}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}/t) = \nabla\Omega^*(\boldsymbol{\theta}/t)$.

The proof is given in §C.1. For classification, the order-preservation property ensures that the highest-scoring class according to $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ agree with each other:

$$\underset{i \in [d]}{\mathsf{argmax}}\, \theta_i = \underset{i \in [d]}{\mathsf{argmax}}\, (\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))_i\,.$$

Temperature scaling is useful to control how close we are to unregularized prediction functions.

3 Fenchel-Young losses

In this section, we introduce Fenchel-Young losses as a natural way to learn models whose output layer is a regularized prediction function.

Definition 2 *Fenchel-Young loss generated by* Ω

Let $\Omega \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *be a regularization function such that the maximum in (1) is achieved for all* $\boldsymbol{\theta} \in \mathbb{R}^d$. *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** $L_\Omega \colon \mathsf{dom}(\Omega^*) \times \mathsf{dom}(\Omega) \to \mathbb{R}_+$ *generated by* Ω *is*

$$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \coloneqq \Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle. \tag{2}$$

Fenchel-Young losses can also be written as $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))$, where $f_{\boldsymbol{\theta}}(\boldsymbol{p}) \coloneqq \Omega(\boldsymbol{p}) - \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle$, highlighting the relation with the regularized prediction

Table 1: **Examples of regularized prediction functions and their associated Fenchel-Young losses.** For multi-class classification, we denote the ground-truth by $\boldsymbol{y} = \boldsymbol{e}_k$, where $\boldsymbol{e}_i$ denotes a standard basis ("one-hot") vector. We denote by $\mathsf{H}^{\mathrm{s}}(\boldsymbol{p}) := -\sum_i p_i \log p_i$ the Shannon entropy of a distribution $\boldsymbol{p} \in \triangle^d$.

Loss	$\mathrm{dom}(\Omega)$	$\Omega(\boldsymbol{p})$	$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$	$L_\Omega(\boldsymbol{\theta};\boldsymbol{y})$
Squared	$\mathbb{R}^d$	$\frac{1}{2}\|\boldsymbol{p}\|^2$	$\boldsymbol{\theta}$	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|^2$
Perceptron (Rosenblatt, 1958)	$\triangle^d$	0	$\mathsf{argmax}(\boldsymbol{\theta})$	$\max_i \ \theta_i - \theta_k$
Hinge (Crammer and Singer, 2001)	$\triangle^d$	$\langle \boldsymbol{p}, \boldsymbol{e}_k - \mathbf{1}\rangle$	$\mathsf{argmax}(\mathbf{1}-\boldsymbol{e}_k+\boldsymbol{\theta})$	$\max_i \ [[i \neq k]] + \theta_i - \theta_k$
Sparsemax (Martins and Astudillo, 2016)	$\triangle^d$	$\frac{1}{2}\|\boldsymbol{p}\|^2$	$\mathsf{sparsemax}(\boldsymbol{\theta})$	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|^2 - \frac{1}{2}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})-\boldsymbol{\theta}\|^2$
Logistic (multinomial)	$\triangle^d$	$-\mathsf{H}^{\mathrm{s}}(\boldsymbol{p})$	$\mathsf{softmax}(\boldsymbol{\theta})$	$\log \sum_i \exp \theta_i - \theta_k$
Logistic (one-vs-all)	$[0,1]^d$	$-\sum_i \mathsf{H}^{\mathrm{s}}([p_i, 1-p_i])$	$\mathsf{sigmoid}(\boldsymbol{\theta})$	$\sum_i \log\big(1+\exp(-(2y_i-1)\theta_i)\big)$

function $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$. Therefore, as long as we can compute $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$, we can evaluate the associated Fenchel-Young loss $L_\Omega(\boldsymbol{\theta};\boldsymbol{y})$. Examples of Fenchel-Young losses are given in Table 1. In addition to the aforementioned multinomial logistic and sparsemax losses, we recover the squared, hinge and one-vs-all logistic losses, for suitable choices of Ω and $\mathrm{dom}(\Omega)$.

Properties. As the name indicates, this family of loss functions is grounded in the Fenchel-Young inequality (Borwein and Lewis, 2010, Proposition 3.3.4)

$$\Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{p}) \geq \langle \boldsymbol{\theta}, \boldsymbol{p}\rangle \ \forall \boldsymbol{\theta} \in \mathrm{dom}(\Omega^*), \boldsymbol{p} \in \mathrm{dom}(\Omega). \tag{3}$$

The inequality, together with well-known properties of convex conjugates, imply the following results.

Proposition 2 *Properties of F-Y losses*

1. **Non-negativity.** *$L_\Omega(\boldsymbol{\theta};\boldsymbol{y}) \geq 0$ for any $\boldsymbol{\theta} \in \mathrm{dom}(\Omega^*) = \mathbb{R}^d$ and $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathrm{dom}(\Omega)$.*
2. **Zero loss.** *If Ω is a lower semi-continuous proper convex function, then $\min_{\boldsymbol{\theta}} L_\Omega(\boldsymbol{\theta};\boldsymbol{y}) = 0$ and $L_\Omega(\boldsymbol{\theta};\boldsymbol{y}) = 0$ iff $\boldsymbol{y} \in \partial\Omega^*(\boldsymbol{\theta})$. If Ω is strictly convex, then $L_\Omega(\boldsymbol{\theta};\boldsymbol{y}) = 0$ iff $\boldsymbol{y} = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta})$.*
3. **Convexity & subgradients.** *L_Ω is convex in $\boldsymbol{\theta}$ and the residual vectors are its subgradients: $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y} \in \partial L_\Omega(\boldsymbol{\theta};\boldsymbol{y})$.*
4. **Differentiability & smoothness.** *If Ω is strictly convex, then L_Ω is differentiable and $\nabla L_\Omega(\boldsymbol{\theta};\boldsymbol{y}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}$. If Ω is strongly convex, then L_Ω is smooth, i.e., $\nabla L_\Omega(\boldsymbol{\theta};\boldsymbol{y})$ is Lipschitz continuous.*
5. **Temperature scaling.** *For any constant $t > 0$, $L_{t\Omega}(\boldsymbol{\theta};\boldsymbol{y}) = tL_\Omega(\boldsymbol{\theta}/t;\boldsymbol{y})$.*

Remarkably, the non-negativity and convexity properties hold even if Ω is not convex. The zero loss property follows from the fact that, if Ω is l.s.c. proper convex, then (3) becomes an equality (i.e., the duality gap is zero) if and only if $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{p})$. It suggests that minimizing a Fenchel-Young loss requires adjusting $\boldsymbol{\theta}$ to produce predictions $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ that are close to the target $\boldsymbol{y}$, reducing the duality gap.

Relation with Bregman divergences. Fenchel-Young losses seamlessly work when ground-truth vectors are label proportions, i.e., $\boldsymbol{y} \in \triangle^d$ instead of $\boldsymbol{y} \in \{\boldsymbol{e}_i\}_{i=1}^d$. For instance, setting Ω to the Shannon negative entropy restricted to $\triangle^d$ yields the cross-entropy loss, $L_\Omega(\boldsymbol{\theta};\boldsymbol{y}) = \mathsf{KL}(\boldsymbol{y}\|\,\mathsf{softmax}(\boldsymbol{\theta}))$, where KL denotes the Kullback-Leibler divergence. From this example, it is tempting to conjecture that a similar result holds for more general Bregman divergences (Bregman, 1967). Recall that the Bregman divergence $B_\Omega\colon \mathrm{dom}(\Omega) \times \mathrm{relint}(\mathrm{dom}(\Omega)) \to \mathbb{R}_+$ generated by a strictly convex and differentiable Ω is

$$B_\Omega(\boldsymbol{y}\|\boldsymbol{p}) := \Omega(\boldsymbol{y}) - \Omega(\boldsymbol{p}) - \langle \nabla\Omega(\boldsymbol{p}), \boldsymbol{y} - \boldsymbol{p}\rangle, \tag{4}$$

the difference at $\boldsymbol{y}$ between Ω and its linearization around $\boldsymbol{p}$. It turns out that $L_\Omega(\boldsymbol{\theta};\boldsymbol{y})$ is not in general equal to $B_\Omega(\boldsymbol{y}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))$. However, when $\Omega = \Psi + I_{\mathcal{C}}$, where Ψ is a Legendre-type function (Rockafellar, 1970; Wainwright and Jordan, 2008), meaning that it is strictly convex, differentiable and its gradient explodes at the boundary of its domain, we have the following proposition, proved in §C.2.

Proposition 3 *Let $\Omega := \Psi + I_{\mathcal{C}}$, where Ψ is of Legendre type and $\mathcal{C} \subseteq \mathrm{dom}(\Psi)$ is a convex set. Then, for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathcal{C}$, we have:*

$$0 \leq \underbrace{B_\Omega(\boldsymbol{y}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))}_{\text{possibly non-convex in } \boldsymbol{\theta}} \leq \underbrace{L_\Omega(\boldsymbol{\theta};\boldsymbol{y})}_{\text{convex in } \boldsymbol{\theta}} \tag{5}$$

with equality when the loss is 0. If $\mathcal{C} = \mathrm{dom}(\Psi)$, i.e., $\Omega = \Psi$, then $L_\Omega(\boldsymbol{\theta};\boldsymbol{y}) = B_\Omega(\boldsymbol{y}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))$.

As an example, applying (5) with $\Psi = \frac{1}{2}\|\cdot\|^2$ and $\mathcal{C} = \triangle^d$, we get that the sparsemax loss is a convex upper-bound for the non-convex $\frac{1}{2}\|\boldsymbol{y} - \mathsf{sparsemax}(\cdot)\|^2$.

This suggests that the sparsemax loss can be useful for sparse label proportion estimation, as we confirm in §6.

The relation between Fenchel-Young losses and Bregman divergences can be further clarified using duality. Letting $\boldsymbol{\theta} = \nabla\Omega(\boldsymbol{p})$ (i.e., $(\boldsymbol{\theta}, \boldsymbol{p})$ is a **dual pair**), we have $\Omega^*(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{p}\rangle - \Omega(\boldsymbol{p})$. Substituting in (4), we get $B_\Omega(\boldsymbol{y}||\boldsymbol{p}) = L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$. In other words, Fenchel-Young losses can be viewed as a "mixed-form Bregman divergence" (Amari, 2016, Theorem 1.1) where the argument $\boldsymbol{p}$ in (4) is **replaced by its dual point** $\boldsymbol{\theta}$. This difference is best seen by comparing the function signatures, $L_\Omega\colon \mathsf{dom}(\Omega^*) \times \mathsf{dom}(\Omega) \to \mathbb{R}_+$ vs. $B_\Omega\colon \mathsf{dom}(\Omega) \times \mathsf{relint}(\mathsf{dom}(\Omega)) \to \mathbb{R}_+$. An important consequence is that Fenchel-Young losses do not impose any restriction on their left argument $\boldsymbol{\theta}$: Our assumption that the maximum in the prediction function (1) is achieved for all $\boldsymbol{\theta} \in \mathbb{R}^d$ implies $\mathsf{dom}(\Omega^*) = \mathbb{R}^d$.

4 New loss functions for sparse probabilistic classification

In the previous section, we presented Fenchel-Young losses in a broad setting. We now restrict to classification over the probability simplex and show how to easily create several entire new families of losses.

Generalized entropies. A natural choice of regularization function Ω over the probability simplex is $\Omega = -\mathsf{H}$, where H is a generalized entropy (DeGroot, 1962; Grünwald and Dawid, 2004): a concave function over $\triangle^d$, used to measure the "uncertainty" in a distribution $\boldsymbol{p} \in \triangle^d$.

Assumptions: We will make the following assumptions about H.

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

A.2. Strict concavity: $\mathsf{H}\big((1-\alpha)\boldsymbol{p} + \alpha\boldsymbol{p}'\big) > (1-\alpha)\mathsf{H}(\boldsymbol{p}) + \alpha\mathsf{H}(\boldsymbol{p}')$, for $\boldsymbol{p} \neq \boldsymbol{p}'$, $\alpha \in (0,1)$.

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

Assumptions A.2 and A.3 imply that H is Schur-concave (Bauschke and Combettes, 2017), a common requirement in generalized entropies. This in turn implies assumption A.1, up to a constant (that constant can easily be subtracted so as to satisfy assumption A.1). As suggested by the next result, proved in §C.3, together, these assumptions imply that H can be used as a sensible uncertainty measure.

Proposition 4 *If H satisfies assumptions A.1-A.3, then it is non-negative and uniquely maximized by the uniform distribution $\boldsymbol{p} = \mathbf{1}/d$.*

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

Induced Fenchel-Young loss. If the ground truth is $\boldsymbol{y} = \boldsymbol{e}_k$ and assumption A.1. holds, (2) becomes

$$L_{-\mathsf{H}}(\boldsymbol{\theta}; \boldsymbol{e}_k) = (-\mathsf{H})^*(\boldsymbol{\theta}) - \theta_k. \tag{6}$$

By using the fact that $\Omega^*(\boldsymbol{\theta} + c\mathbf{1}) = \Omega^*(\boldsymbol{\theta}) + c$ for all $c \in \mathbb{R}$ if $\mathsf{dom}(\Omega) \subseteq \triangle^d$, we can further rewrite it as

$$L_{-\mathsf{H}}(\boldsymbol{\theta}; \boldsymbol{e}_k) = (-\mathsf{H})^*(\boldsymbol{\theta} - \theta_k\mathbf{1}).$$

This expression shows that Fenchel-Young losses over $\triangle^d$ can be written solely in terms of the generalized "cumulant function" $(-\mathsf{H})^*$. Indeed, when H is Shannon's entropy, we recover the cumulant (a.k.a. log-partition) function $(-\mathsf{H}^{\mathrm{S}})^*(\boldsymbol{\theta}) = \log\sum_{i=1}^d \exp(\theta_i)$. When H is strongly concave over $\triangle^d$, we can also see $(-\mathsf{H})^*$ as a smoothed max operator (Niculae and Blondel, 2017; Mensch and Blondel, 2018) and hence $L_{-\mathsf{H}}(\boldsymbol{\theta}; \boldsymbol{e}_k)$ can be seen as a smoothed upper-bound of the perceptron loss $(\boldsymbol{\theta}; \boldsymbol{e}_k) \mapsto \max_{i\in[d]} \theta_i - \theta_k$.

We now give two examples of generalized entropies. The resulting families of prediction and loss functions, new to our knowledge, are illustrated in Figure 1. We provide more examples in §A.

Tsallis α-entropies (Tsallis, 1988). Defined as $\mathsf{H}^{\mathrm{T}}_\alpha(\boldsymbol{p}) \coloneqq k(\alpha-1)^{-1}\big(1 - \|\cdot\|_\alpha^\alpha\big)$, where $\alpha \geq 1$ and k is an arbitrary positive constant, these entropies arise as a generalization of the Shannon-Khinchin axioms to non-extensive systems (Suyari, 2004) and have numerous scientific applications (Gell-Mann and Tsallis, 2004; Martins et al., 2009). 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 uniformly separable form:

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

The limit case $\alpha \to 1$ corresponds to the Shannon entropy. When $\alpha = 2$, we recover the Gini index (Gini, 1912), a popular "impurity measure" for decision trees:

$$\mathsf{H}^{\mathrm{T}}_2(\boldsymbol{p}) = \frac{1}{2}\sum_{j=1}^d p_j(1-p_j) = \frac{1}{2}(1 - \|\boldsymbol{p}\|_2^2) \quad \forall \boldsymbol{p} \in \triangle^d. \tag{7}$$

It is easy to check that $L_{-\mathsf{H}^{\mathrm{T}}_2}$ recovers the sparsemax loss (Martins and Astudillo, 2016) (cf. Table 1).

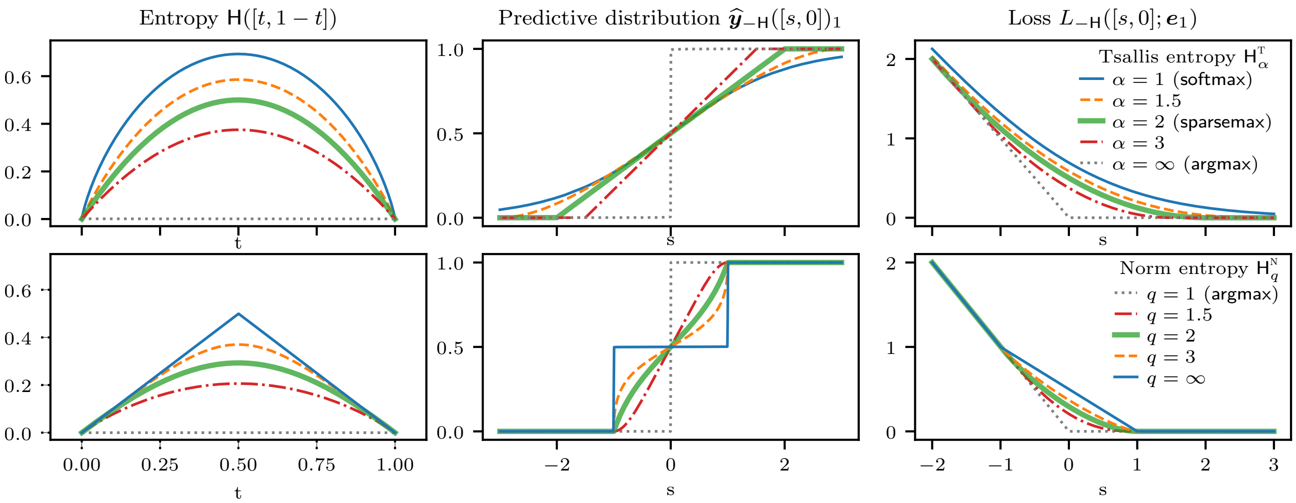

Figure 1: **New families of losses made possible by our framework.** Left: Tsallis and norm entropies. Center: regularized prediction functions. Right: Fenchel-Young loss. Except for softmax, which never exactly reaches 0, all distributions shown in the center can have **sparse support**. As can be checked visually, $\widehat{\boldsymbol{y}}_{-\mathsf{H}}$ is differentiable everywhere when $\alpha, q \in (1, 2)$. Hence, $L_{-\mathsf{H}}$ is **twice differentiable** everywhere for these values.

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 1. The resulting "argmax" distribution puts all probability mass on the top-scoring classes. In summary, $\widehat{\boldsymbol{y}}_{-\mathsf{H}^{\mathrm{T}}_\alpha}$ for $\alpha \in \{1, 2, \infty\}$ is softmax, sparsemax, and argmax, and $L_{-\mathsf{H}^{\mathrm{T}}_\alpha}$ is the logistic, sparsemax and perceptron loss, respectively. Tsallis entropies induce a **continuous parametric family** subsuming these important cases. Since the best surrogate loss often depends on the data (Nock and Nielsen, 2009), tuning α typically improves accuracy, as we confirm in §6.

Norm entropies. 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. From the Minkowski inequality, q-norms with $q > 1$ are strictly convex on the simplex, so $\mathsf{H}^{\mathrm{N}}_q$ satisfies assumptions A.1–A.3 for $q > 1$. The limit case $q \to \infty$ is particularly interesting: in this case, we obtain $\mathsf{H}^{\mathrm{N}}_\infty = 1 - \|\cdot\|_\infty$, recovering the Berger-Parker dominance index (Berger and Parker, 1970), widely used in ecology to measure species diversity. We surprisingly encounter $\mathsf{H}^{\mathrm{N}}_\infty$ again in §5, as a limit case for the existence of separation margins.

Computing $\widehat{\boldsymbol{y}}_{-\mathsf{H}}(\boldsymbol{\theta})$. For non-separable entropies H, the regularized prediction function $\widehat{\boldsymbol{y}}_{-\mathsf{H}}(\boldsymbol{\theta})$ does not generally enjoy a closed-form expression and one must resort to projected gradient methods to compute it. Fortunately, for uniformly separable entropies, which we saw to be the case of Tsallis entropies, we now show that $\widehat{\boldsymbol{y}}_{-\mathsf{H}}(\boldsymbol{\theta})$ can be computed in linear time.

Proposition 5 *Reduction to root finding*

Let $\mathsf{H}(\boldsymbol{p}) = \sum_i h(p_i) + I_{\triangle^d}(\boldsymbol{p})$ *where* $h\colon [0,1] \to \mathbb{R}_+$ *is strictly concave and differentiable. Then,*

$$\widehat{\boldsymbol{y}}_{-\mathsf{H}}(\boldsymbol{\theta}) = \boldsymbol{p}(\tau) := (-h')^{-1}(\max\{\boldsymbol{\theta} - \tau, -h'(0)\})$$

where τ *is a root of* $\phi(t) := \langle \boldsymbol{p}(t), \mathbf{1}\rangle - 1$, *in the tight search interval* $[\tau_{\min}, \tau_{\max}]$, *where* $\tau_{\min} := \max(\boldsymbol{\theta}) + h'(1)$ *and* $\tau_{\max} := \max(\boldsymbol{\theta}) + h'\left({}^{1}\!/_{d}\right)$.

An approximate τ such that $|\phi(\tau)| \leq \epsilon$ can be found in $O({}^{1}\!/_{\log \epsilon})$ time by, e.g., bisection. The related problem of Bregman projection onto the probability simplex was recently studied by Krichene et al. (2015) but our derivation is different and more direct (cf. §C.4).

5 Separation margin of F-Y losses

In this section, we are going to see that the simple assumptions A.1–A.3 about a generalized entropy H are enough to obtain results about the **separation margin** associated with $L_{-\mathsf{H}}$. The notion of margin is well-known in machine learning, lying at the heart of support vector machines and leading to generalization error bounds (Vapnik, 1998; Schölkopf and Smola, 2002; Guermeur, 2007). We provide a definition and will see that many other Fenchel-Young losses also have a "margin," for suitable conditions on H. Then, we take a step further, and connect the existence of a margin with the **sparsity** of the regularized prediction function, providing necessary and sufficient conditions for Fenchel-Young losses to have a margin. Finally, we show how this margin can be computed analytically.

Definition 3 *Separation margin*

Let $L(\boldsymbol{\theta}; \boldsymbol{e}_k)$ be a loss function over $\mathbb{R}^d \times \{\boldsymbol{e}_i\}_{i=1}^d$. We say that L has the separation margin property *if there exists $m > 0$ such that:*

$$\theta_k \geq m + \max_{j \neq k} \theta_j \quad \Rightarrow \quad L(\boldsymbol{\theta}; \boldsymbol{e}_k) = 0. \tag{8}$$

The smallest possible m that satisfies (8) is called the margin *of L, denoted* $\mathrm{margin}(L)$.

Examples. The most famous example of a loss with a separation margin is the **multi-class hinge loss**, $L(\boldsymbol{\theta}; \boldsymbol{e}_k) = \max\{0, \max_{j \neq k} 1 + \theta_j - \theta_k\}$, which we saw in Table 1 to be a Fenchel-Young loss: it is immediate from the definition that its margin is 1. Less trivially, Martins and Astudillo (2016, Prop. 3.5) showed that the **sparsemax loss** also has the separation margin property. On the negative side, the logistic loss does not have a margin, as it is strictly positive. Characterizing which Fenchel-Young losses have a margin is an open question which we address next.

Conditions for existence of margin. To accomplish our goal, we need to characterize the gradient mappings $\partial(-\mathsf{H})$ and $\nabla(-\mathsf{H})^*$ associated with generalized entropies (note that $\partial(-\mathsf{H})$ is never single-valued: if $\boldsymbol{\theta}$ is in $\partial(-\mathsf{H})(\boldsymbol{p})$, then so is $\boldsymbol{\theta} + c\mathbf{1}$, for any constant $c \in \mathbb{R}$). Of particular importance is the subdifferential set $\partial(-\mathsf{H})(\boldsymbol{e}_k)$. The next proposition, whose proof we defer to §C.5, uses this set to provide a necessary and sufficient condition for the existence of a separation margin, along with a formula for computing it.

Proposition 6 *Let H satisfy A.1–A.3. Then:*

1. *The loss $L_{-\mathsf{H}}$ has a separation margin iff there is a $m > 0$ such that $m\boldsymbol{e}_k \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$.*
2. *If the above holds, then the margin of $L_{-\mathsf{H}}$ is given by the smallest such m or, equivalently,*

$$\mathrm{margin}(L_{-\mathsf{H}}) = \sup_{\boldsymbol{p} \in \triangle^d} \frac{\mathsf{H}(\boldsymbol{p})}{1 - \|\boldsymbol{p}\|_\infty}. \tag{9}$$

Reassuringly, the first part confirms that the logistic loss does not have a margin, since $\partial(-\mathsf{H}^{\mathrm{S}})(\boldsymbol{e}_k) = \varnothing$. A second interesting fact is that the denominator of (9) is the generalized entropy $\mathsf{H}^{\mathrm{N}}_\infty(\boldsymbol{p})$ introduced in §4: the ∞**-norm entropy**. As Figure 1 suggests, this entropy provides an upper bound for convex losses with unit margin. This provides some intuition to the formula (9), which seeks a distribution $\boldsymbol{p}$ maximizing the **entropy ratio** between $\mathsf{H}(\boldsymbol{p})$ and $\mathsf{H}^{\mathrm{N}}_\infty(\boldsymbol{p})$.

Equivalence between sparsity and margin. The next result, proved in §C.6, characterizes more precisely the image of $\nabla(-\mathsf{H})^*$. In doing so, it establishes a key result in this paper: **a sufficient condition for the existence of a separation margin in $L_{-\mathsf{H}}$ is the sparsity of the regularized prediction function $\widehat{\boldsymbol{y}}_{-\mathsf{H}} \equiv \nabla(-\mathsf{H})^*$**, i.e., its ability to reach the entire simplex, including the boundary points. If H is uniformly separable, this is also a necessary condition.

Proposition 7 *Equivalence between sparse probability distribution and loss enjoying a margin*

Let H satisfy A.1–A.3 and be uniformly separable, i.e., $\mathsf{H}(\boldsymbol{p}) = \sum_{i=1}^d h(p_i)$. Then the following statements are all equivalent:

1. *$\partial(-\mathsf{H})(\boldsymbol{p}) \neq \varnothing$ for any $\boldsymbol{p} \in \triangle^d$;*
2. *The mapping $\nabla(-\mathsf{H})^*$ covers the full simplex, i.e., $\nabla(-\mathsf{H})^*(\mathbb{R}^d) = \triangle^d$;*
3. *$L_{-\mathsf{H}}$ has the separation margin property.*

For a general H (not necessarily separable) satisfying A.1–A.3, we have (1) $\Leftrightarrow$ (2) $\Rightarrow$ (3).

Let us reflect for a moment on the three conditions stated in Proposition 7. The first two conditions involve the subdifferential and gradient of $-\mathsf{H}$ and its conjugate; the third condition is the margin property of $L_{-\mathsf{H}}$. To provide some intuition, consider the case where H is separable with $\mathsf{H}(\boldsymbol{p}) = \sum_i h(p_i)$ and h is differentiable in $(0,1)$. Then, from the concavity of h, its derivative h' is decreasing, hence the first condition is 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, since $h'(t) = -1 - \log t$ explodes at 0. Functions whose gradient "explodes" in the boundary of their domain (hence failing to meet the first condition in Proposition 7) are called "essentially smooth" (Rockafellar, 1970). For those functions, $\nabla(-\mathsf{H})^*$ maps only to the relative interior of $\triangle^d$, never attaining boundary points (Wainwright and Jordan, 2008); this is expressed in the second condition. This prevents essentially smooth functions from generating a sparse $\boldsymbol{y}_{-\mathsf{H}} \equiv \nabla(-\mathsf{H})^*$ or (if they are separable) a loss $L_{-\mathsf{H}}$ with a margin, as asserted by the third condition. Since Legendre-type functions (§3) are strictly convex *and* essentially smooth, by Proposition 3, loss functions for which the composite form $L_{-\mathsf{H}}(\boldsymbol{\theta}; \boldsymbol{y}) = B_{-\mathsf{H}}(\boldsymbol{y} || \widehat{\boldsymbol{y}}_{-\mathsf{H}}(\boldsymbol{\theta}))$ holds, which is the case of the logistic loss but not of the sparsemax loss, do not enjoy a margin and cannot induce a sparse probability distribution. This is geometrically visible in Figure 1.

Margin computation. For Fenchel-Young losses that have the separation margin property, Proposition 6 provided a formula for determining the margin. While informative, formula (9) is not very practical, as it involves a generally non-convex optimization problem. The next proposition, proved in §C.7, takes a step further and provides a remarkably simple closed-form expression for generalized entropies that are **twice-differentiable**. To simplify notation, we denote by $\nabla_j \mathsf{H}(\boldsymbol{p}) \equiv (\nabla \mathsf{H}(\boldsymbol{p}))_j$ the j^{th} component of $\nabla \mathsf{H}(\boldsymbol{p})$.

Proposition 8 *Assume* H *satisfies the conditions in Proposition 7 and is twice-differentiable on the simplex. Then, for arbitrary* $j \neq k$*:*

$$\operatorname{margin}(L_{-\mathsf{H}}) = \nabla_j \mathsf{H}(\boldsymbol{e}_k) - \nabla_k \mathsf{H}(\boldsymbol{e}_k). \tag{10}$$

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, with* $h(0) = h(1) = 0$*, then*

$$\operatorname{margin}(L_{-\mathsf{H}}) = h'(0) - h'(1) = -\int_0^1 h''(t)dt. \tag{11}$$

The compact formula (10) provides a geometric characterization of separable entropies and their margins: (11) tells us that only the slopes of h at the two extremities of $[0,1]$ are relevant in determining the margin.

Example: case of Tsallis and norm entropies. As seen in §4, Tsallis entropies are separable with $h(t) = (t - t^\alpha)/(\alpha(\alpha-1))$. For $\alpha > 1$, $h'(t) = (1 - \alpha t^{\alpha-1})/(\alpha(\alpha-1))$, hence $h'(0) = 1/(\alpha(\alpha-1))$ and $h'(1) = -1/\alpha$. Proposition 8 then yields

$$\operatorname{margin}(L_{-\mathsf{H}^{\mathrm{T}}_\alpha}) = h'(0) - h'(1) = (\alpha - 1)^{-1}.$$

Norm entropies, while not separable, have gradient $\nabla \mathsf{H}^{\mathrm{N}}_q(\boldsymbol{p}) = -(\boldsymbol{p}/\|\boldsymbol{p}\|_q)^{q-1}$, giving $\nabla \mathsf{H}^{\mathrm{N}}_q(\boldsymbol{e}_k) = -\boldsymbol{e}_k$, so

$$\operatorname{margin}(\mathsf{H}^{\mathrm{N}}_q) = \nabla_j \mathsf{H}^{\mathrm{N}}_q(\boldsymbol{e}_k) - \nabla_k \mathsf{H}^{\mathrm{N}}_q(\boldsymbol{e}_k) = 1,$$

as confirmed visually in Figure 1, in the binary case.

6 Experimental results

As we saw, α-Tsallis entropies generate a family of losses, with the logistic ($\alpha \to 1$) and sparsemax losses ($\alpha = 2$) as important special cases. In addition, they are twice differentiable for $\alpha \in [1, 2)$, produce sparse probability distributions for $\alpha > 1$ and are computationally efficient for any $\alpha \geq 1$, thanks to Proposition 5. In this section, we demonstrate their usefulness on the task of label proportion estimation and compare different solvers for computing $\widehat{\boldsymbol{y}}_{-\mathsf{H}^{\mathrm{T}}_\alpha}$.

Label proportion estimation. Given an input vector $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^p$, where p is the number of features, our goal is to estimate a vector of label proportions $\boldsymbol{y} \in \triangle^d$, where d is the number of classes. If $\boldsymbol{y}$ is sparse, we expect the superiority of Tsallis losses over the conventional logistic loss on this task. At training time, given a set of n $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ pairs, we estimate a matrix $W \in \mathbb{R}^{d \times p}$ by minimizing the convex objective

$$R(W) \coloneqq \sum_{i=1}^n L_\Omega(W\boldsymbol{x}_i; \boldsymbol{y}_i) + \frac{\lambda}{2}\|W\|_F^2.$$

We use L-BFGS (Liu and Nocedal, 1989) for simplicity. From Proposition 2 and using the chain rule, we obtain the gradient expression $\nabla R(W) = (\widehat{Y}_\Omega - Y)^\top X + \lambda W$, where $\widehat{Y}_\Omega$, Y and X are matrices whose rows gather $\widehat{\boldsymbol{y}}_\Omega(W\boldsymbol{x}_i)$, $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$, for $i = 1, \dots, n$. At test time, we predict label proportions by $\boldsymbol{p} = \boldsymbol{y}_{-\mathsf{H}^{\mathrm{T}}_\alpha}(W\boldsymbol{x})$.

We ran experiments on 7 standard multi-label benchmark datasets — see §B for dataset characteristics. For all datasets, we removed samples with no label, normalized samples to have zero mean unit variance, and normalized labels to lie in the probability simplex. We chose $\lambda \in \{10^{-4}, 10^{-3}, \dots, 10^4\}$ and $\alpha \in \{1, 1.1, \dots, 2\}$ against the validation set. We report the test set mean Jensen-Shannon divergence, $\mathrm{JS}(\boldsymbol{p}, \boldsymbol{y}) \coloneqq \frac{1}{2}\mathrm{KL}(\boldsymbol{p}\|\frac{\boldsymbol{p}+\boldsymbol{y}}{2}) + \frac{1}{2}\mathrm{KL}(\boldsymbol{y}\|\frac{\boldsymbol{p}+\boldsymbol{y}}{2})$, and the mean squared error $\frac{1}{2}\|\boldsymbol{p} - \boldsymbol{y}\|^2$ in Table 6. As can be seen, the loss with tuned α achieves the best averaged rank overall. Tuning α allows to choose the best loss in the family in a data-driven fashion. Additional experiments confirm these findings — see §B.

Solver comparison. Next, we compared bisection (binary search) and Brent's method for solving (1) by root finding (Proposition 5). We focus on $\mathsf{H}^{\mathrm{T}}_{1.5}$, i.e. the 1.5-Tsallis entropy, and also compare against using a generic projected gradient algorithm (FISTA) to solve (1) naively. We measure the time needed to reach a solution $\boldsymbol{p}$ with $\|\boldsymbol{p} - \boldsymbol{p}^\star\|_2 < 10^{-5}$, over 200 samples $\boldsymbol{\theta} \in \mathbb{R}^d \sim \mathcal{N}(\boldsymbol{0}, \sigma\boldsymbol{I})$ with $\log\sigma \sim \mathcal{U}(-4, 4)$. Median and 99% CI times reported in Figure 2 reveal that root finding scales better, with Brent's method outperforming FISTA by one to two orders of magnitude.

7 Related work

Proper scoring rules (proper losses) are a well-studied object in statistics (Grünwald and Dawid, 2004; Gneiting and Raftery, 2007) and machine learning (Reid and Williamson, 2010; Williamson et al., 2016), that measures the discrepancy between a ground-truth $\boldsymbol{y} \in \triangle^d$ and a probability forecast $\boldsymbol{p} \in \triangle^d$ in a **Fisher-consistent** manner. From Savage (1971) (see also Gneiting and Raftery (2007)), we can construct a proper scoring rule $S_\Omega \colon \triangle^d \times \triangle^d \to \mathbb{R}_+$ by

$$S_\Omega(\boldsymbol{p}; \boldsymbol{y}) \coloneqq \langle \nabla\Omega(\boldsymbol{p}), \boldsymbol{y} - \boldsymbol{p}\rangle - \Omega(\boldsymbol{p}) = B_\Omega(\boldsymbol{y}\|\boldsymbol{p}) - \Omega(\boldsymbol{y}),$$

	$\alpha = 1$ (logistic)	$\alpha = 1.5$	$\alpha = 2$ (sparsemax)	tuned α
Birds	0.359 / 0.530	0.364 / 0.504	0.364 / 0.504	**0.358** / **0.501**
Cal500	0.454 / **0.034**	0.456 / 0.035	**0.452** / 0.035	0.456 / **0.034**
Emotions	0.226 / 0.327	0.225 / **0.317**	0.225 / **0.317**	**0.224** / 0.321
Mediamill	0.375 / 0.208	0.363 / 0.193	**0.356** / **0.191**	0.361 / 0.193
Scene	**0.175** / **0.344**	0.176 / 0.363	0.176 / 0.363	**0.175** / 0.345
TMC	0.225 / 0.337	0.224 / **0.327**	0.224 / **0.327**	**0.217** / 0.328
Yeast	**0.307** / **0.183**	0.314 / 0.186	0.314 / 0.186	**0.307** / **0.183**
Avg. rank	2.57 / 2.71	2.71 / 2.14	2.14 / 2.00	**1.43** / **1.86**

Table 2: **Test-set performance of Tsallis losses for various α on the task of sparse label proportion estimation:** average Jensen-Shannon divergence (left) and mean squared error (right). Lower is better.

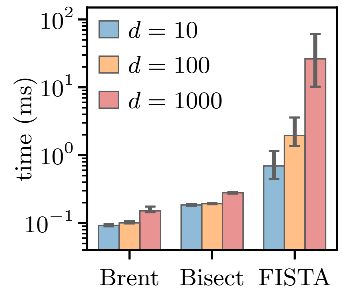

Figure 2: Median time until 10^{-5} accuracy is met for computing $\widehat{\boldsymbol{y}}_{-\mathsf{H}^{\mathrm{T}}_{1.5}}$.

recovering the well-known relation between Bregman divergences and proper scoring rules. For example, using the Gini index $\mathsf{H}(\boldsymbol{p}) = 1 - \|\boldsymbol{p}\|^2$ generates the **Brier score** (Brier, 1950) $S_{-\mathsf{H}}(\boldsymbol{p}; \boldsymbol{e}_k) = \sum_{i=1}^d ([[k = i]] - p_i)^2$, showing that the sparsemax loss and the Brier score share the same generating function. More generally, while a scoring rule S_Ω is related to a **primal-space** Bregman divergence, a Fenchel-Young loss L_Ω can be seen as a **mixed-space** Bregman divergence (§3). This difference has a number of important consequences. First, S_Ω is **not necessarily convex** in $\boldsymbol{p}$ (Williamson et al. (2016, Proposition 17) show that it is in fact quasi-convex). In contrast, L_Ω **is always** convex in $\boldsymbol{\theta}$. Second, the first argument is **constrained** to $\triangle^d$ for S_Ω, while **unconstrained** for L_Ω.

In practice, proper scoring rules (losses) are often composed with an **invertible** link function $\boldsymbol{\psi}^{-1} \colon \mathbb{R}^d \to \triangle^d$. This form of a loss, $S_\Omega(\boldsymbol{\psi}^{-1}(\boldsymbol{\theta}); \boldsymbol{y})$, is sometimes called composite (Buja et al., 2005; Reid and Williamson, 2010; Williamson et al., 2016). Although the decoupling between loss and link has merits (Reid and Williamson, 2010), the composition of $S_\Omega(\cdot; \boldsymbol{y})$ and $\boldsymbol{\psi}^{-1}(\boldsymbol{\theta})$ is not necessarily convex in $\boldsymbol{\theta}$. The **canonical link function** (Buja et al., 2005) of S_Ω is a link function that ensures the convexity of $S_\Omega(\boldsymbol{\psi}^{-1}(\boldsymbol{\theta}); \boldsymbol{y})$ in $\boldsymbol{\theta}$. It also plays a key role in generalized linear models (Nelder and Baker, 1972). Following Proposition 3, when Ω is Legendre type, we obtain

$$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = B_\Omega(\boldsymbol{y} || \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})) = S_\Omega(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}); \boldsymbol{y}) + \Omega(\boldsymbol{y}).$$

Thus, in this case, Fenchel-Young losses and proper composite losses coincide up to the constant term $\Omega(\boldsymbol{y})$ (which vanishes if $\boldsymbol{y} = \boldsymbol{e}_k$ and Ω satisfies assumption A.1), with $\boldsymbol{\psi}^{-1} = \widehat{\boldsymbol{y}}_\Omega$ the canonical inverse link function. Fenchel-Young losses, however, require neither invertible link nor Legendre type assumptions, allowing to express losses (e.g., hinge or sparsemax) that are not expressible in composite form. Moreover, as seen in §5, a Legendre-type Ω precisely precludes sparse probability distributions and losses enjoying a margin.

Other losses. Nock and Nielsen (2009) proposed binary classification losses based on the Legendre transformation but require invertible mappings. Masnadi-Shirazi (2011) studied the Bayes consistency of related binary classification loss functions. Duchi et al. (2018, Proposition 3) derived the multi-class loss (6), a special case of Fenchel-Young loss over the probability simplex, and showed (Proposition 4) that any strictly concave generalized entropy generates a classification-calibrated loss. Amid and Warmuth (2017) proposed a different family of losses based on the Tsallis divergence, to interpolate between convex and non-convex losses, for robustness to label noise.

Smoothing techniques. Fenchel duality also plays a key role in smoothing techniques (Nesterov, 2005; Beck and Teboulle, 2012), which have been used extensively to create smoothed losses (Shalev-Shwartz and Zhang, 2016). However, these techniques were applied on a per-loss basis and were not connected to the induced probability distribution. In contrast, we propose a generic construction, with clear links between smoothing and the distribution induced by $\widehat{\boldsymbol{y}}_\Omega$.

8 Conclusion

We showed that regularization and Fenchel duality provide simple core principles, unifying many existing loss functions, and allowing to create useful new ones easily. In particular, we derived a new family of loss functions based on Tsallis entropies, which includes the logistic, sparsemax and perceptron losses as special cases. With the unique exception of the logistic loss, losses in this family induce sparse probability distributions. We also showed a close and fundamental relationship between generalized entropies, losses enjoying a margin and sparse probability distributions. Remarkably, Fenchel-Young losses can be defined over arbitrary domains, allowing to construct loss functions for a large variety of applications (Blondel et al., 2019).

Acknowledgements

MB thanks Arthur Mensch and Gabriel Peyré for numerous fruitful discussions and Tim Vieira for introducing him to generalized exponential families. AM and VN were 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 and CMU-PERI/TIC/0046/2014 (GoLocal).

References

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

Ehsan Amid and Manfred K Warmuth. Two-temperature logistic regression based on the Tsallis divergence. *arXiv preprint arXiv:1705.07210*, 2017.

Keith Ball, Eric A Carlen, and Elliott H Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones Mathematicae*, 115(1): 463–482, 1994.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6:1705–1749, 2005.

Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2nd edition, 2017.

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

Wolfgang H Berger and Frances L Parker. Diversity of planktonic foraminifera in deep-sea sediments. *Science*, 168(3937):1345–1347, 1970.

Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific Belmont, 1999.

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with fenchel-young losses. *arXiv preprint arXiv:1901.02324*, 2019.

Jonathan Borwein and Adrian S Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer Science & Business Media, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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.

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

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

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, 2005.

Laurent Condat. Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 158(1-2): 575–585, 2016.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

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

Morris H DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of ICML*, 2008.

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

Murray Gell-Mann and Constantino Tsallis. *Nonextensive Entropy: Interdisciplinary Applications*. Oxford University Press, 2004.

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

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

Peter D Grünwald and A Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

Yann Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8:2551–2594, 2007.

Walid Krichene, Syrine Krichene, and Alexandre Bayen. Efficient Bregman projections onto the simplex. In *Proceedings of CDC*, 2015.

Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

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

André FT Martins and Ramón Fernandez Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of ICML*, 2016.

André FT Martins, Mário AT Figueiredo, Pedro MQ Aguiar, Noah A Smith, and Eric P Xing. Nonextensive information theoretic kernels on measures. *Journal of Machine Learning Research*, 10:935–975, 2009.

Hamed Masnadi-Shirazi. *The design of bayes consistent loss functions for classification*. PhD thesis, UC San Diego, 2011.

Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In *Proceedings of ICML*, 2018.

John Ashworth Nelder and R Jacob Baker. *Generalized Linear Models*. Wiley Online Library, 1972.

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

Vlad Niculae and Mathieu Blondel. A regularized framework for sparse and structured neural attention. In *Proceedings of NIPS*, 2017.

Vlad Niculae, André FT Martins, Mathieu Blondel, and Claire Cardie. SparseMAP: Differentiable sparse structured inference. In *Proceedings of ICML*, 2018.

Richard Nock and Frank Nielsen. Bregman divergences and surrogates for learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2048–2059, 2009.

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

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics, and Probability*, volume 1. University of California Press, 1961.

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

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

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

Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, 2002.

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

Hiroki Suyari. Generalization of Shannon-Khinchin axioms to nonextensive systems and the uniqueness theorem for the nonextensive entropy. *IEEE Transactions on Information Theory*, 50(8):1783–1787, 2004.

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

Vladimir N Vapnik. *Statistical Learning Theory*. Wiley, 1998.

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

Robert C Williamson, Elodie Vernet, and Mark D Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17(223):1–52, 2016.