Journal of Machine Learning Research 21 (2020) 1-69 Submitted 1/19; Revised 10/19; Published 1/20

Learning with Fenchel-Young Losses

Mathieu Blondel mathieu@mblondel.org
NTT Communication Science Laboratories
Kyoto, Japan

André F. T. Martins andre.martins@unbabel.com
Unbabel & Instituto de Telecomunicações
Lisbon, Portugal

Vlad Niculae vlad@vene.ro
Instituto de Telecomunicações
Lisbon, Portugal

Editor: Sathiya Keerthi

Abstract

Over the past decades, numerous loss functions have been been proposed for a variety of supervised learning tasks, including regression, classification, ranking, and more generally structured prediction. Understanding the core principles and theoretical properties underpinning these losses is key to choose the right loss for the right problem, as well as to create new losses which combine their strengths. In this paper, we introduce Fenchel-Young losses, a generic way to *construct* a convex loss function for a regularized prediction function. We provide an in-depth study of their properties in a very broad setting, covering all the aforementioned supervised learning tasks, and revealing new connections between sparsity, generalized entropies, and separation margins. We show that Fenchel-Young losses unify many well-known loss functions and allow to create useful new ones easily. Finally, we derive efficient predictive and training algorithms, making Fenchel-Young losses appealing both in theory and practice.

Keywords: loss functions, output regularization, convex duality, structured prediction

1. Introduction

Loss functions are a cornerstone of statistics and machine learning: They measure the difference, or "loss," between a ground truth and a prediction. As such, much work has been devoted to designing loss functions for a variety of supervised learning tasks, including regression (Huber, 1964), classification (Crammer and Singer, 2001), ranking (Joachims, 2002) and structured prediction (Lafferty et al., 2001; Collins, 2002; Tsochantaridis et al., 2005), to name only a few well-known directions.

For the case of probabilistic classification, proper composite loss functions (Reid and Williamson, 2010; Williamson et al., 2016) offer a principled framework unifying various existing loss functions. A proper composite loss is the composition of a proper loss between two probability distributions, with an invertible mapping from real vectors to probability distributions. The theoretical properties of proper loss functions, also known as proper scoring rules (Grünwald and Dawid (2004); Gneiting and Raftery (2007); and references therein), such as their Fisher consistency (classification calibration) and correspondence with Bregman divergences, are now well-understood. However, not all existing losses are proper composite loss functions; a notable example is the hinge loss used in support vector machines. In fact, we shall see that any loss function enjoying a 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), cannot be written in composite proper loss form.

At the same time, loss functions are often intimately related to an underlying statistical model and prediction function. For instance, the logistic loss corresponds to the multinomial distribution and the softmax operator, while the conditional random field (CRF) loss (Lafferty et al., 2001) for structured prediction is tied with marginal inference (Wainwright and Jordan, 2008). Both are instances of generalized linear models (Nelder and Baker, 1972; McCullagh and Nelder, 1989), associated with exponential family distributions. More recently, Martins and Astudillo (2016) proposed a new classification loss based on the projection onto the simplex. Unlike the logistic loss, this "sparsemax" loss induces probability distributions with **sparse** support, which is desirable in some applications for interpretability or computational efficiency reasons. However, the sparsemax loss was derived in a relatively ad-hoc manner and it is still relatively poorly understood. Is it one of a kind or can we generalize it in a principled manner? Thorough understanding of the core principles underpinning existing losses and their associated predictive model, potentially enabling the creation of useful new losses, is one of the main quests of this paper.

This paper. The starting point of this paper are the notions of **output** regularization and regularized prediction functions, which we use to provide a variational perspective on many existing prediction functions, including the aforementioned softmax, sparsemax and marginal inference. Based on simple convex duality arguments, we then introduce Fenchel-Young losses, a new way to automatically construct a loss function associated with any regularized prediction function. As we shall see, our proposal recovers many existing loss functions, which is in a sense surprising since many of these losses were originally proposed by independent efforts. Our framework goes beyond the simple probabilistic classification setting: We show how to create loss functions over various structured domains, including convex polytopes and convex cones. Our framework encourages the loss designer to think **geometrically** about the outputs desired for the task at hand. Once a (regularized) prediction function has been designed, our framework generates a corresponding loss function automatically. We will demonstrate the ease of creating loss functions, including useful new ones, using abundant examples throughout this paper.

Previous papers. This paper builds upon two previously published shorter conference papers. The first (Niculae et al., 2018) introduced Fenchel-Young losses in the structured prediction setting but only provided a limited analysis of their properties. The second (Blondel et al., 2019) provided a more in-depth analysis but focused on unstructured probabilistic classification. This paper provides a comprehensive study of Fenchel-Young losses across various domains. Besides a much more thorough treatment of previously covered topics, this paper contributes entirely new sections, including §6 on losses for positive measures, §8 on primal and dual training algorithms, and §A.2 on loss "Fenchel-Youngization". We provide in §7 a new unifying view between structured predictions losses, and discuss at length various convex polytopes, promoting a geometric approach to structured prediction loss design; we also provide novel results in this section regarding structured separation margins (Proposition 8), proving the unit margin of the SparseMAP loss. We demonstrate how to use our framework to create useful new losses, including ranking losses, not covered in the previous two papers.

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

$$I_{\mathcal{C}}(\boldsymbol{\mu}) := \begin{cases} 0 & \text{if } \boldsymbol{\mu} \in \mathcal{C} \\ \infty & \text{otherwise} \end{cases} \tag{1}$$

and its support function by $\sigma_{\mathcal{C}}(\boldsymbol{\theta}) := \sup_{\boldsymbol{\mu} \in \mathcal{C}} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle$. We define the proximity operator (a.k.a. proximal operator) of Ω by

$$\mathsf{prox}_{\Omega}(\boldsymbol{\eta}) := \operatorname*{\mathsf{argmin}}_{\boldsymbol{\mu} \in \mathsf{dom}(\Omega)} \frac{1}{2}\|\boldsymbol{\mu} - \boldsymbol{\eta}\|^2 + \Omega(\boldsymbol{\mu}). \tag{2}$$

We denote the interior and relative interior of $\mathcal{C}$ by $\mathsf{int}(\mathcal{C})$ and $\mathsf{relint}(\mathcal{C})$, respectively. We denote $[\boldsymbol{x}]_+ := \max(\boldsymbol{x}, \mathbf{0})$, evaluated element-wise.

Table of contents.

2. Regularized prediction functions

In this section, we introduce the concept of regularized prediction function (§2.1), which is central to this paper. We then give simple and well-known examples of such functions (§2.2) and discuss their properties in a general setting (§2.3,§2.4).

2.1. Definition

We consider a general predictive setting with input variables $\boldsymbol{x} \in \mathcal{X}$, and a parametrized model $\boldsymbol{f}_W : \mathcal{X} \to \mathbb{R}^d$ (which could be a linear model or a neural network), producing a score vector $\boldsymbol{\theta} \coloneqq \boldsymbol{f}_W(\boldsymbol{x}) \in \mathbb{R}^d$. In a simple multi-class classification setting, the score vector is typically used to pick the highest-scoring class among d possible ones

$$\widehat{y}(\boldsymbol{\theta}) \in \operatorname*{argmax}_{j \in [d]} \theta_j. \tag{3}$$

This can be generalized to an arbitrary output space $\mathcal{Y} \subseteq \mathbb{R}^d$ by using instead

$$\widehat{\boldsymbol{y}}(\boldsymbol{\theta}) \in \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle, \tag{4}$$

where intuitively $\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$ captures the affinity between $\boldsymbol{x}$ (since $\boldsymbol{\theta}$ is produced by $f_W(\boldsymbol{x})$) and $\boldsymbol{y}$. Therefore, (4) seeks the output $\boldsymbol{y}$ with greatest affinity with $\boldsymbol{x}$. The support function $\sigma_{\mathcal{Y}}(\boldsymbol{\theta}) = \max_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle = \langle \boldsymbol{\theta}, \widehat{\boldsymbol{y}}(\boldsymbol{\theta}) \rangle$ can be interpreted as the largest projection of any element of $\mathcal{Y}$ onto the line generated by $\boldsymbol{\theta}$.

Clearly, (4) recovers (3) with $\mathcal{Y} = \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_d\}$, where $\boldsymbol{e}_j$ is a standard basis vector, $\boldsymbol{e}_j \coloneqq [0, \dots 0, \underbrace{1}_{j}, 0, \dots 0]$. In this case, the cardinality $|\mathcal{Y}|$ and the dimensionality d coincide, but this need not be the case in general. Eq. (4) is often called a linear maximization oracle or maximum a-posteriori (MAP) oracle (Wainwright and Jordan, 2008). The latter name comes from the fact that (4) coincides with the mode of the Gibbs distribution defined by

$$p(\boldsymbol{y}; \boldsymbol{\theta}) \propto \exp \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle.$$

Prediction over convex hulls. We now extend the prediction function (4) by replacing $\mathcal{Y}$ with its convex hull $\mathsf{conv}(\mathcal{Y}) \coloneqq \{\mathbb{E}_{\boldsymbol{p}}[Y] \colon \boldsymbol{p} \in \triangle^{|\mathcal{Y}|}\}$ and introducing a regularization function Ω into the optimization problem:

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \operatorname*{argmax}_{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \Omega(\boldsymbol{\mu}).$$

We emphasize that the regularization is w.r.t. predictions (outputs) and not w.r.t. model parameters (denoted by W in this paper), as is usually the case in the literature. We illustrate the regularized prediction function pipeline in Figure 1.

Unsurprisingly, the choice $\Omega = 0$ recovers the unregularized prediction function (4). This follows from the fundamental theorem of linear programming (Dantzig et al., 1955, Theorem

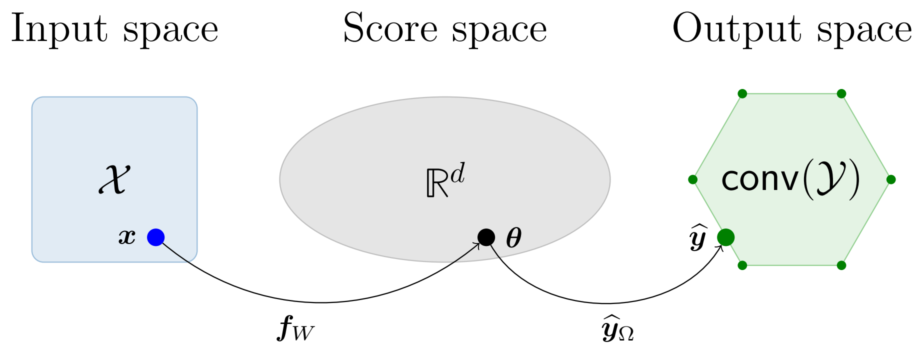

Figure 1: **Illustration of the proposed regularized prediction framework over a convex hull** $\mathsf{conv}(\mathcal{Y})$. A parametrized model $\boldsymbol{f}_W : \mathcal{X} \to \mathbb{R}^d$ (linear model, neural network, etc.) produces a score vector $\boldsymbol{\theta} \in \mathbb{R}^d$. The regularized prediction function $\widehat{\boldsymbol{y}}_\Omega$ produces a prediction $\widehat{\boldsymbol{y}} \in \mathsf{conv}(\mathcal{Y})$. Regularized prediction functions are not limited to convex hulls and can be defined over arbitrary domains (Def. 1).

6), which states that the maximum of a linear form over a convex polytope is always achieved at one of its vertices:

$$\widehat{\boldsymbol{y}}_0(\boldsymbol{\theta}) \in \operatorname*{argmax}_{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle.$$

Why regularize outputs? The regularized prediction function (5) casts computing a prediction as a variational problem. It involves an optimization problem that balances between two terms: an "affinity" term $\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle$, and a "confidence" term $\Omega(\boldsymbol{\mu})$ which should be low if $\boldsymbol{\mu}$ 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.

Introducing $\Omega(\boldsymbol{\mu})$ in (5) tends to move the prediction away from the vertices of $\mathsf{conv}(\mathcal{Y})$. Unless the regularization term $\Omega(\boldsymbol{\mu})$ is negligible compared to the affinity term $\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle$, a prediction becomes a convex combination of several vertices. As we shall see in §7, which is dedicated to structured prediction over convex hulls, we can interpret this prediction as the mean under some underlying distribution. This contrasts with (4), which always outputs the most likely vertex, i.e., the mode.

Prediction over arbitrary domains. Regularized prediction functions are in fact not limited to convex hulls. We now state their precise definition in complete generality.

Definition 1 *Prediction function regularized by* Ω

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*

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \operatorname*{argmax}_{\boldsymbol{\mu} \in \mathsf{dom}(\Omega)} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \Omega(\boldsymbol{\mu}). \tag{5}$$

Allowing extended-real Ω permits general domain constraints in (5) via indicator functions. For instance, choosing $\Omega = I_{\mathsf{conv}(\mathcal{Y})}$, where $I_\mathcal{C}$ is the indicator function defined in (1), recovers the MAP oracle (4). Importantly, the choice of domain $\mathsf{dom}(\Omega)$ is not limited to convex hulls. For instance, we will also consider conic hulls, $\mathsf{cone}(\mathcal{Y})$, later in this paper.

Choosing $\boldsymbol{\Omega}$. Regularized prediction functions $\widehat{\boldsymbol{y}}_\Omega$ involve two main design choices: the domain $\mathsf{dom}(\Omega)$ over which Ω is defined and Ω itself. The choice of $\mathsf{dom}(\Omega)$ is mainly dictated by the type of output we want from $\widehat{\boldsymbol{y}}_\Omega$, such as $\mathsf{dom}(\Omega) = \mathsf{conv}(\mathcal{Y})$ for convex combinations of elements of $\mathcal{Y}$, and $\mathsf{dom}(\Omega) = \mathsf{cone}(\mathcal{Y})$ for conic combinations. The choice of the regularization Ω itself further governs certain properties of $\widehat{\boldsymbol{y}}_\Omega$, including, as we shall see in the sequel, its sparsity or its use of prior knowledge regarding the importance or misclassification cost of certain outputs. The choices of $\mathsf{dom}(\Omega)$ and Ω may also be constrained by computational considerations. Indeed, while computing $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ involves a potentially challenging constrained maximization problem in general, we will see that certain choices of Ω lead to closed-form expressions. The power of our framework is that the user can focus solely on designing and computing $\widehat{\boldsymbol{y}}_\Omega$: We will see in §3 how to automatically construct a loss function associated with $\widehat{\boldsymbol{y}}_\Omega$.

2.2. Examples

To illustrate regularized prediction functions, we give several concrete examples enjoying a **closed-form expression**.

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

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

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 $\mathsf{H}^{\mathrm{s}}(\boldsymbol{p}) := -\sum_i p_i \log p_i$ is Shannon's entropy, $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is the well-known softmax

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \mathsf{softmax}(\boldsymbol{\theta}) := \frac{\exp(\boldsymbol{\theta})}{\sum_{j=1}^d \exp(\theta_j)}. \tag{6}$$

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

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

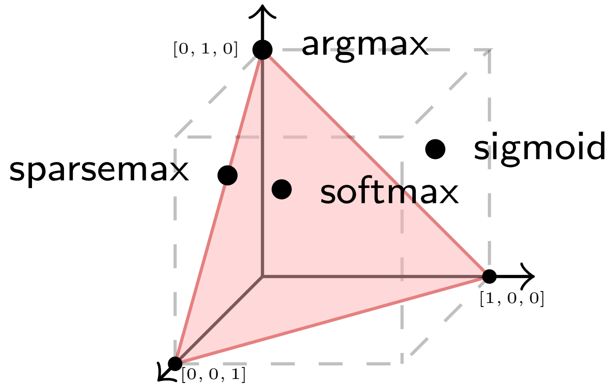

Figure 2: **Examples of regularized prediction functions**. The unregularized argmax prediction always hits a vertex of the probability simplex, leading to a probability distribution that puts all probability mass on the same class. Unlike, softmax which always occurs in the relative interior of the simplex and thus leads to a dense distribution, sparsemax (Euclidean projection onto the probability simplex) may hit the boundary, leading to a **sparse** probability distribution. We also display the sigmoid operator which lies in the unit cube and is thus not guaranteed to output a valid probability distribution.

a.k.a. the sparsemax transformation (Martins and Astudillo, 2016). It is well-known that

$$\operatorname*{argmin}_{\boldsymbol{p} \in \triangle^d} \|\boldsymbol{p} - \boldsymbol{\theta}\|^2 = [\boldsymbol{\theta} - \tau \mathbf{1}]_+, \tag{7}$$

for some threshold $\tau \in \mathbb{R}$. Hence, the predicted distribution can have **sparse** support (it may assign exactly zero probability to low-scoring classes). The threshold τ can be computed **exactly** in $O(d)$ time (Brucker, 1984; Duchi et al., 2008; Condat, 2016).

The regularized prediction function paradigm is, however, 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

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

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

We will see in §4 that the first three examples (argmax, softmax and sparsemax) are particular instances of a broader family of prediction functions, using the notion of **generalized entropy**. The last example is a special case of regularized prediction function over positive measures, developed in §6. Regularized prediction functions also encompass more complex convex polytopes for structured prediction, as we shall see in §7.

2.3. Gradient mapping and dual objective

From Danskin's theorem (Danskin, 1966) (see also Bertsekas (1999, Proposition B.25)) $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is a subgradient of Ω^* at $\boldsymbol{\theta}$, i.e., $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \partial\Omega^*(\boldsymbol{\theta})$. If, furthermore, Ω is strictly convex, then $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is the gradient of Ω^* at $\boldsymbol{\theta}$, i.e., $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta})$. This interpretation of $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ as a (sub)gradient mapping will play a crucial role in the next section for deriving a loss function associated with $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$.

Viewing $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ as the (sub)gradient of $\Omega^*(\boldsymbol{\theta})$ is also useful to derive the dual of (5). Let $\Omega \coloneqq \Psi + \Phi$. It is well-known (Borwein and Lewis, 2010; Beck and Teboulle, 2012) that

$$\Omega^*(\boldsymbol{\theta}) = (\Psi + \Phi)^*(\boldsymbol{\theta}) = \inf_{\boldsymbol{u} \in \mathbb{R}^d} \Phi^*(\boldsymbol{u}) + \Psi^*(\boldsymbol{\theta} - \boldsymbol{u}) =: (\Phi^* \Box \Psi^*)(\boldsymbol{\theta}), \tag{8}$$

where $f \Box g$ denotes the infimal convolution of f with g. Furthermore, from Danskin's theorem, $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \nabla\Psi^*(\boldsymbol{\theta} - \boldsymbol{u}^\star)$, where $\boldsymbol{u}^\star$ denotes an optimal solution of the infimum in (8). We can think of that infimum as the dual of the optimization problem in (5). When $\Psi = \Psi^* = \frac{1}{2}\|\cdot\|^2$, $\Omega^*(\boldsymbol{\theta})$ is known as the Moreau envelope of Φ^* (Moreau, 1965) and using Moreau's decomposition, we obtain $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \boldsymbol{\theta} - \mathsf{prox}_{\Phi^*}(\boldsymbol{\theta}) = \mathsf{prox}_{\Phi}(\boldsymbol{\theta})$. As another example, when $\Omega = \Psi + I_\mathcal{C}$, we obtain

$$\Omega^*(\boldsymbol{\theta}) = \inf_{\boldsymbol{u} \in \mathbb{R}^d} \sigma_\mathcal{C}(\boldsymbol{u}) + \Psi^*(\boldsymbol{\theta} - \boldsymbol{u}),$$

where we used $I_\mathcal{C}^* = \sigma_\mathcal{C}$, the support function of $\mathcal{C}$. In particular, when $\mathcal{C} = \mathsf{conv}(\mathcal{Y})$, we have $\sigma_\mathcal{C}(\boldsymbol{u}) = \max_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{u}, \boldsymbol{y} \rangle$. This dual view is informative insofar as it suggests that regularized prediction functions $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ with $\Omega = \Psi + I_\mathcal{C}$ minimize a trade-off between maximizing the value achieved by the unregularized prediction function $\sigma_\mathcal{C}(\boldsymbol{u})$, and a proximity term $\Psi^*(\boldsymbol{\theta} - \boldsymbol{u})$.

2.4. Properties

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

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

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

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

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 $\boldsymbol{\mu} = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$. If Ω is symmetric, then the coordinates of $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ are sorted the same way, i.e., $\theta_i > \theta_j \Rightarrow \mu_i \geq \mu_j$ and $\mu_i > \mu_j \Rightarrow \theta_i > \theta_j$.*
3. **Approximation error.** *Assume $\mathcal{Y} \subseteq \mathsf{dom}(\Omega)$. If Ω is γ-strongly convex and bounded with $L \leq \Omega(\boldsymbol{\mu}) \leq U$ for all $\boldsymbol{\mu} \in \mathsf{dom}(\Omega)$, then $\frac{1}{2}\|\widehat{\boldsymbol{y}}(\boldsymbol{\theta}) - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\|^2 \leq \frac{U-L}{\gamma}$.*
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, $\widehat{\boldsymbol{y}}_{t\Omega}(\boldsymbol{\theta}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}/t) = \nabla\Omega^*(\boldsymbol{\theta}/t)$.*

5. **Constant invariance.** *For any constant* $c \in \mathbb{R}$, $\widehat{\boldsymbol{y}}_{\Omega+c}(\boldsymbol{\theta}) = \widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta}; \boldsymbol{y})$.

The proof is given in Appendix B.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:

$$\operatorname*{argmax}_{i \in [d]} \theta_i = \operatorname*{argmax}_{i \in [d]} \left(\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta})\right)_i.$$

Temperature scaling is useful to control how close we are to unregularized prediction functions. Clearly, $\widehat{\boldsymbol{y}}_{t\Omega}(\boldsymbol{\theta}) \to \widehat{\boldsymbol{y}}(\boldsymbol{\theta})$ as $t \to 0$, where $\widehat{\boldsymbol{y}}(\boldsymbol{\theta})$ is defined in (4).

3. Fenchel-Young losses

In the previous section, we introduced regularized prediction functions over arbitrary domains, as a generalization of classical (unregularized) decision functions. In this section, we introduce Fenchel-Young losses for learning models whose output layer is a regularized prediction function. We first give their definitions and state their properties (§3.1). We then discuss their relationship with Bregman divergences (§3.2) and their Bayes risk (§3.3). Finally, we show how to construct a cost-sensitive loss from any Fenchel-Young loss (§3.4).

3.1. Definition and properties

Given a regularized prediction function $\widehat{\boldsymbol{y}}_{\Omega}$, we define its associated loss as follows.

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* (5) *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*

$$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \coloneqq \Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle. \tag{9}$$

It is easy to see that Fenchel-Young losses can be rewritten as

$$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta})),$$

where $f_{\boldsymbol{\theta}}(\boldsymbol{\mu}) \coloneqq \Omega(\boldsymbol{\mu}) - \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle$, highlighting the relation with regularized prediction functions. Therefore, as long as we can compute a regularized prediction function $\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta})$, we can automatically obtain an associated Fenchel-Young loss $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$. Conversely, we also have that $\widehat{\boldsymbol{y}}_{\Omega}$ outputs the prediction minimizing the loss:

$$\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta}) \in \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathsf{dom}(\Omega)} L_\Omega(\boldsymbol{\theta}; \boldsymbol{\mu}).$$

Examples of existing losses that fall into the Fenchel-Young loss family are given in Table 1. Some of these examples will be discussed in more details in the sequel of this paper. Note

Table 1: **Examples of regularized prediction functions and their corresponding 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 (Wainwright and Jordan, 2008). We denote by $\mathsf{H}^{\mathrm{s}}(\boldsymbol{p}) \coloneqq -\sum_i p_i \log p_i$ the Shannon entropy of a distribution $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$.

Loss	$\mathsf{dom}(\Omega)$	$\Omega(\boldsymbol{\mu})$	$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$	$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$
Squared	$\mathbb{R}^d$	$\frac{1}{2}\|\boldsymbol{\mu}\|^2$	$\boldsymbol{\theta}$	$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|^2$
Perceptron	$\triangle^{|\mathcal{Y}|}$	0	$\mathsf{argmax}(\boldsymbol{\theta})$	$\max_i \theta_i - \theta_k$
Logistic	$\triangle^{|\mathcal{Y}|}$	$-\mathsf{H}^{\mathrm{s}}(\boldsymbol{\mu})$	$\mathsf{softmax}(\boldsymbol{\theta})$	$\log \sum_i \exp \theta_i - \theta_k$
Hinge	$\triangle^{|\mathcal{Y}|}$	$\langle \boldsymbol{\mu}, \boldsymbol{e}_k - \mathbf{1} \rangle$	$\mathsf{argmax}(\mathbf{1} - \boldsymbol{e}_k + \boldsymbol{\theta})$	$\max_i \ [[i \neq k]] + \theta_i - \theta_k$
Sparsemax	$\triangle^{|\mathcal{Y}|}$	$\frac{1}{2}\|\boldsymbol{\mu}\|^2$	$\mathsf{sparsemax}(\boldsymbol{\theta})$	$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|^2 - \frac{1}{2}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{\theta}\|^2$
Logistic (one-vs-all)	$[0,1]^{|\mathcal{Y}|}$	$-\sum_i \mathsf{H}^{\mathrm{s}}([\mu_i, 1 - \mu_i])$	$\mathsf{sigmoid}(\boldsymbol{\theta})$	$\sum_i \log(1 + \exp(-(2y_i - 1)\theta_i))$
Structured perceptron	$\mathsf{conv}(\mathcal{Y})$	0	$\mathsf{MAP}(\boldsymbol{\theta})$	$\max_{\boldsymbol{y}'} \langle \boldsymbol{\theta}, \boldsymbol{y}' \rangle - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$
Structured hinge	$\mathsf{conv}(\mathcal{Y})$	$-\langle \boldsymbol{\mu}, \boldsymbol{c}_{\boldsymbol{y}} \rangle$	$\mathsf{MAP}(\boldsymbol{\theta} + \boldsymbol{c}_{\boldsymbol{y}})$	$\max_{\boldsymbol{y}'} \langle \boldsymbol{c}_{\boldsymbol{y}}, \boldsymbol{y}' \rangle + \langle \boldsymbol{\theta}, \boldsymbol{y}' \rangle - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$
CRF	$\mathsf{conv}(\mathcal{Y})$	$\min\limits_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|} \colon \mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}} -\mathsf{H}^{\mathrm{s}}(\boldsymbol{p})$	$\mathsf{marginals}(\boldsymbol{\theta})$	$\log \sum_{\boldsymbol{y}'} \exp \langle \boldsymbol{\theta}, \boldsymbol{y}' \rangle - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$
SparseMAP	$\mathsf{conv}(\mathcal{Y})$	$\frac{1}{2}\|\boldsymbol{\mu}\|^2$	$\mathsf{SparseMAP}(\boldsymbol{\theta})$	$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|^2 - \frac{1}{2}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{\theta}\|^2$

that we will allow Ω in some cases to depend on the ground-truth $\boldsymbol{y}$. Since we do not know $\boldsymbol{y}$ at test time, this requires us to use another prediction function as a replacement for the regularized prediction function. As we explain in §3.4, this discrepancy is well motivated and allows us to express popular cost-sensitive losses, such as the structured hinge loss.

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)

$$\Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{\mu}) \geq \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle \quad \forall \boldsymbol{\theta} \in \mathsf{dom}(\Omega^*), \boldsymbol{\mu} \in \mathsf{dom}(\Omega). \tag{10}$$

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

Proposition 2 *Properties of Fenchel-Young losses*

1. **Non-negativity.** *$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \geq 0$ for any $\boldsymbol{\theta} \in \mathsf{dom}(\Omega^*) = \mathbb{R}^d$ and $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathsf{dom}(\Omega)$.*
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 $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{y} \in \partial\Omega^*(\boldsymbol{\theta})$. If Ω is strictly convex, then $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{y} = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta}) = \mathsf{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$.*
3. **Convexity & subgradients.** *L_Ω is convex in $\boldsymbol{\theta}$ and the residual vectors are its subgradients: $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y} \in \partial L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$.*

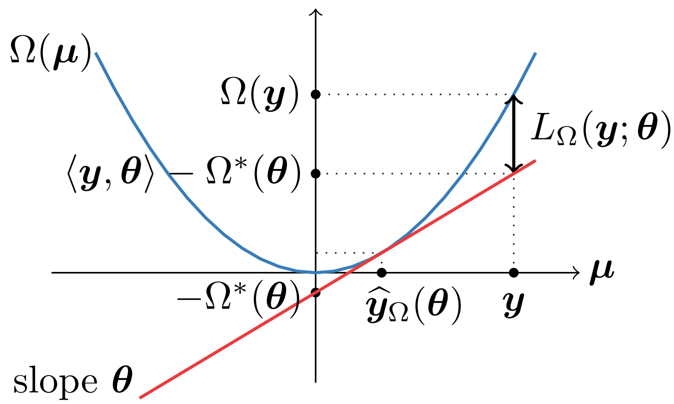

Figure 3: **Illustration of the Fenchel-Young loss** $L_\Omega(\boldsymbol{\theta};\boldsymbol{y}) = \Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y}\rangle$, here with $\Omega(\boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{\mu}\|^2$ and $\mathsf{dom}(\Omega) = \mathbb{R}^d$. Minimizing $L_\Omega(\boldsymbol{\theta};\boldsymbol{y})$ w.r.t. $\boldsymbol{\theta}$ can be seen as minimizing the duality gap, the difference between $\Omega(\boldsymbol{\mu})$ and the tangent $\boldsymbol{\mu} \mapsto \langle \boldsymbol{\theta}, \boldsymbol{\mu}\rangle - \Omega^*(\boldsymbol{\theta})$, at $\boldsymbol{\mu} = \boldsymbol{y}$ (the ground truth). The regularized prediction $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is the value of $\boldsymbol{\mu}$ at which the tangent touches $\Omega(\boldsymbol{\mu})$. When Ω is of Legendre type, $L_\Omega(\boldsymbol{\theta};\boldsymbol{y})$ is equal to the Bregman divergence generated by Ω between $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ (cf. §3.2). However, we do not require that assumption in this paper.

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 $t > 0$, $L_{t\Omega}(\boldsymbol{\theta};\boldsymbol{y}) = tL_\Omega(\boldsymbol{\theta}/t;\boldsymbol{y})$.*
6. **Constant invariance.** *For any constant $c \in \mathbb{R}$, $L_{\Omega+c}(\boldsymbol{\theta};\boldsymbol{y}) = L_\Omega(\boldsymbol{\theta};\boldsymbol{y})$.*

Remarkably, the non-negativity, convexity and constant invariance 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 (10) becomes an equality (i.e., the duality gap is zero) if and only if $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{\mu})$. It suggests that the minimization of Fenchel-Young losses attempts 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. This is illustrated with $\Omega = \frac{1}{2}\|\boldsymbol{\mu}\|^2$ and $\mathsf{dom}(\Omega) = \mathbb{R}^d$ (leading to the squared loss) in Figure 3.

Domain of Ω^*. Our assumption that the maximum in the regularized prediction function (5) is achieved for all $\boldsymbol{\theta} \in \mathbb{R}^d$ implies that $\mathsf{dom}(\Omega^*) = \mathbb{R}^d$. This assumption is quite mild and does not require $\mathsf{dom}(\Omega)$ to be bounded. Minimizing $L_\Omega(\boldsymbol{\theta};\boldsymbol{y})$ w.r.t. $\boldsymbol{\theta}$ is therefore an **unconstrained** convex optimization problem. This contrasts with proper loss functions, which are defined over the probability simplex, as discussed in §10.

3.2. Relation with Bregman divergences

Fenchel-Young losses seamlessly work when $\mathcal{Y} = \mathsf{dom}(\Omega)$ instead of $\mathcal{Y} \subset \mathsf{dom}(\Omega)$. For example, in the case of the logistic loss, where $-\Omega$ is the Shannon entropy restricted to

$\triangle^d$, allowing $\boldsymbol{y} \in \triangle^d$ instead of $\boldsymbol{y} \in \{\boldsymbol{e}_i\}_{i=1}^d$ yields the cross-entropy loss, $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \mathsf{KL}(\boldsymbol{y} \| \mathsf{softmax}(\boldsymbol{\theta}))$, where KL denotes the (generalized) Kullback-Leibler divergence

$$\mathsf{KL}(\boldsymbol{y}\|\boldsymbol{\mu}) \coloneqq \sum_i y_i \log \frac{y_i}{\mu_i} - \sum_i y_i + \sum_i \mu_i.$$

This can be useful in a multi-label setting with supervision in the form of *label proportions*.

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 $B_\Omega \colon \mathsf{dom}(\Omega) \times \mathsf{relint}(\mathsf{dom}(\Omega)) \to \mathbb{R}_+$ generated by a strictly convex and differentiable Ω is

$$B_\Omega(\boldsymbol{y}\|\boldsymbol{\mu}) \coloneqq \Omega(\boldsymbol{y}) - \Omega(\boldsymbol{\mu}) - \langle \nabla\Omega(\boldsymbol{\mu}), \boldsymbol{y} - \boldsymbol{\mu} \rangle. \tag{11}$$

In other words, this is the difference at $\boldsymbol{y}$ between Ω and its linearization around $\boldsymbol{\mu}$. It turns out that $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ is not in general equal to $B_\Omega(\boldsymbol{y}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))$. In fact the latter is **not necessarily convex** in $\boldsymbol{\theta}$ while the former always is. However, there is a duality relationship between Fenchel-Young losses and Bregman divergences, as we now discuss.

A "mixed-space" Bregman divergence. 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 (11), 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" (Amari, 2016, Theorem 1.1) where the argument $\boldsymbol{\mu}$ in (11) 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 (5) is achieved for all $\boldsymbol{\theta} \in \mathbb{R}^d$ implies $\mathsf{dom}(\Omega^*) = \mathbb{R}^d$. In contrast, a Bregman divergence would typically need to be composed with a mapping from $\mathbb{R}^d$ to $\mathsf{dom}(\Omega)$, such as $\widehat{\boldsymbol{y}}_\Omega$, resulting in a possibly non-convex function.

Case of Legendre-type functions. We can make the relationship with Bregman divergences further precise when $\Omega = \Psi + I_\mathcal{C}$, where Ψ is restricted to the class of so-called Legendre-type functions (Rockafellar, 1970; Wainwright and Jordan, 2008). We first recall the definition of this class of functions and then state our results.

Definition 3 *Essentially smooth and Legendre type functions*

A function Ψ is essentially smooth if

- $\mathsf{dom}(\Psi)$ *is non-empty,*
- Ψ *is differentiable throughout* $\mathsf{int}(\mathsf{dom}(\Psi))$*,*
- *and* $\lim_{i\to\infty} \nabla\Psi(\boldsymbol{\mu}^i) = +\infty$ *for any sequence* $\{\boldsymbol{\mu}^i\}$ *contained in* $\mathsf{dom}(\Psi)$*, and converging to a boundary point of* $\mathsf{dom}(\Psi)$*.*

A function Ψ is of Legendre type if

- *it is strictly convex on* $\mathsf{int}(\mathsf{dom}(\Psi))$

- *and essentially smooth.*

For instance, $\Psi(\boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{\mu}\|^2$ is Legendre-type with $\mathsf{dom}(\Psi) = \mathbb{R}^d$, and $\Psi(\boldsymbol{\mu}) = \sum_i \mu_i \log \mu_i$ is Legendre-type with $\mathsf{dom}(\Psi) = \mathbb{R}^d_+$. However, $\Omega(\boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{\mu}\|^2 + I_{\mathbb{R}^d_+}(\boldsymbol{\mu})$ is not Legendre-type, since the gradient of Ω does not explode everywhere on the boundary of $\mathbb{R}^d_+$. The Legendre-type assumption crucially implies that

$$\nabla\Psi(\nabla\Psi^*(\boldsymbol{\theta})) = \boldsymbol{\theta} \quad \text{for all} \quad \boldsymbol{\theta} \in \mathsf{dom}(\Psi^*).$$

We can use this fact to derive the following results, proved in Appendix B.2.

Proposition 3 *Relation with Bregman divergences*

Let Ψ be of Legendre type with $\mathsf{dom}(\Psi^*) = \mathbb{R}^d$ and let $\mathcal{C} \subseteq \mathsf{dom}(\Psi)$ be a convex set. Let Ω be the restriction of Ψ to $\mathcal{C} \subseteq \mathsf{dom}(\Psi)$, i.e., $\Omega := \Psi + I_{\mathcal{C}}$.

1. **Bregman projection.** *The prediction function regularized by Ω, $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$, reduces to the Bregman projection of $\widehat{\boldsymbol{y}}_\Psi(\boldsymbol{\theta})$ onto $\mathcal{C}$:*

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\mu}\in\mathcal{C}} \langle \boldsymbol{\theta}, \boldsymbol{\mu}\rangle - \Psi(\boldsymbol{\mu}) = \operatorname*{argmin}_{\boldsymbol{\mu}\in\mathcal{C}} B_\Psi(\boldsymbol{\mu}\|\widehat{\boldsymbol{y}}_\Psi(\boldsymbol{\theta})).$$

2. **Difference of divergences.** *For all $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathcal{C}$:*

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

3. **Bound.** *For all $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathcal{C}$:*

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

with equality when the loss is minimized

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \boldsymbol{y} \Leftrightarrow L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = 0 \Leftrightarrow B_\Psi(\boldsymbol{y}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})) = 0.$$

4. **Composite form.** *When $\mathcal{C} = \mathsf{dom}(\Psi)$, i.e., $\Omega = \Psi$, we have equality for all $\boldsymbol{\theta} \in \mathbb{R}^d$*

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

We illustrate these properties using $\Psi = \frac{1}{2}\|\cdot\|^2$ as a running example. From the first property, since $\widehat{\boldsymbol{y}}_\Psi(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $B_\Psi(\boldsymbol{y}\|\boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\mu}\|^2$, we get

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\mu}\in\mathcal{C}} B_\Psi(\boldsymbol{\mu}\|\widehat{\boldsymbol{y}}_\Psi(\boldsymbol{\theta})) = \operatorname*{argmin}_{\boldsymbol{\mu}\in\mathcal{C}} B_\Psi(\boldsymbol{\mu}\|\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\mu}\in\mathcal{C}} \|\boldsymbol{\mu} - \boldsymbol{\theta}\|^2,$$

recovering the Euclidean projection onto $\mathcal{C}$. The reduction of regularized prediction functions to Bregman projections (when Ψ is of Legendre type) is useful because there exist efficient algorithms for computing the Bregman projection onto various convex sets (Yasutake et al., 2011; Suehiro et al., 2012; Krichene et al., 2015; Lim and Wright, 2016). Therefore, we can use these algorithms to compute $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ provided that $\widehat{\boldsymbol{y}}_\Psi(\boldsymbol{\theta})$ is available.

From the second property, we obtain for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathcal{C}$

$$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|^2 - \frac{1}{2}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{\theta}\|^2.$$

This recovers the expression of the sparsemax loss given in Table 1 with $\mathcal{C} = \triangle^d$.

From the third claim, we obtain for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathcal{C}$

$$\frac{1}{2}\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\|^2 \leq L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}).$$

This shows that $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ provides a convex upper-bound for the possibly non-convex composite function $B_\Psi(\boldsymbol{y}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))$. In particular, when $\mathcal{C} = \triangle^d$, we get $\frac{1}{2}\|\boldsymbol{y} - \mathsf{sparsemax}(\boldsymbol{\theta})\|^2 \leq L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$. This suggests that the sparsemax loss is useful for sparse label proportion estimation, as confirmed in our experiments (§9).

Finally, from the last property, if $\Omega = \Psi = \frac{1}{2}\|\cdot\|^2$, we obtain $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|^2$, which is indeed the squared loss given in Table 1.

3.3. Expected loss, Bayes risk and Bregman information

In this section, we discuss the relation between the pointwise Bayes risk (minimal achievable loss) of a Fenchel-Young loss and Bregman information (Banerjee et al., 2005).

Expected loss. Let Y be a random variable taking values in $\mathcal{Y}$ following the distribution $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$. The expected loss (a.k.a. expected risk) is then

$$\begin{aligned} \mathbb{E}_{\boldsymbol{p}}[L_\Omega(\boldsymbol{\theta}; Y)] &= \sum_{\boldsymbol{y} \in \mathcal{Y}} p(\boldsymbol{y}) L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \\ &= \sum_{\boldsymbol{y} \in \mathcal{Y}} p(\boldsymbol{y})(\Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y}\rangle) \\ &= \mathbb{E}_{\boldsymbol{p}}[\Omega(Y)] + \Omega^*(\boldsymbol{\theta}) - \langle \boldsymbol{\theta}, \mathbb{E}_{\boldsymbol{p}}[Y]\rangle \\ &= L_\Omega(\boldsymbol{\theta}; \mathbb{E}_{\boldsymbol{p}}[Y]) + \mathbb{I}_\Omega(Y; \boldsymbol{p}). \end{aligned} \tag{12}$$

Here, we defined the **Bregman information** of Y by

$$\mathbb{I}_\Omega(Y; \boldsymbol{p}) \coloneqq \min_{\boldsymbol{\mu} \in \mathsf{dom}(\Omega)} \mathbb{E}_{\boldsymbol{p}}[B_\Omega(Y\|\boldsymbol{\mu})] = \mathbb{E}_{\boldsymbol{p}}\left[B_\Omega(Y\|\mathbb{E}_{\boldsymbol{p}}[Y])\right] = \mathbb{E}_{\boldsymbol{p}}[\Omega(Y)] - \Omega(\mathbb{E}_{\boldsymbol{p}}[Y]).$$

We refer the reader to Banerjee et al. (2005) for a detailed discussion as to why the last two equalities hold. The r.h.s. is exactly equal to the difference between the two sides of Jensen's inequality $\mathbb{E}[\Omega(Y)] \geq \Omega(\mathbb{E}[Y])$ and is therefore non-negative. For this reason, it is sometimes also called Jensen gap (Reid and Williamson, 2011).

Bayes risk. From Proposition 2, we know that $\min_{\boldsymbol{\theta}} L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = 0$ for all $\boldsymbol{y} \in \mathsf{dom}(\Omega)$. Therefore, the pointwise Bayes risk coincides precisely with the Bregman information of Y,

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{E}_{\boldsymbol{p}}[L_\Omega(\boldsymbol{\theta}; Y)] = \min_{\boldsymbol{\theta} \in \mathbb{R}^d} L_\Omega(\boldsymbol{\theta}; \mathbb{E}_{\boldsymbol{p}}[Y]) + \mathbb{I}_\Omega(Y; \boldsymbol{p}) = \mathbb{I}_\Omega(Y; \boldsymbol{p}), \tag{13}$$

provided that $\mathbb{E}_{\boldsymbol{p}}[Y] \in \mathsf{dom}(\Omega)$. A similar relation between Bayes risk and Bregman information exists for proper losses (Reid and Williamson, 2011). We can think of (13) as a measure of the "difficulty" of the task. Combining (12) and (13), we obtain

$$\mathbb{E}_{\boldsymbol{p}}[L_\Omega(\boldsymbol{\theta}; Y)] - \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{E}_{\boldsymbol{p}}[L_\Omega(\boldsymbol{\theta}; Y)] = L_\Omega(\boldsymbol{\theta}; \mathbb{E}_{\boldsymbol{p}}[Y]),$$

the pointwise "regret" of $\boldsymbol{\theta} \in \mathbb{R}^d$ w.r.t. $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$. If $\mathcal{Y} = \{\boldsymbol{e}_i\}_{i=1}^d$, $L_\Omega(\boldsymbol{\theta}; \mathbb{E}_{\boldsymbol{p}}[Y]) = L_\Omega(\boldsymbol{\theta}; \boldsymbol{p})$.

3.4. 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 may depend on the ground truth $\boldsymbol{y}$. For example, $\boldsymbol{c}_{\boldsymbol{y}} = \mathbf{1} - \boldsymbol{y}$ corresponds to the 0/1 cost and can be used to impose a margin. 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 (Crammer and Singer (2001); $\Omega = 0$), the softmax-margin loss (Gimpel and Smith (2010); $\Omega = -\mathsf{H}^{\mathrm{s}}$), and the cost-augmented sparsemax (Shalev-Shwartz and Zhang (2016, Eq. (13)), Niculae et al. (2018); $\Omega = \frac{1}{2}\|\cdot\|^2$). It is easy to see that the associated regularized prediction function is

$$\widehat{\boldsymbol{y}}_\Psi(\boldsymbol{\theta}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta} + \boldsymbol{c}_{\boldsymbol{y}}).$$

For the 0/1 cost $\boldsymbol{c}_{\boldsymbol{y}} = \mathbf{1} - \boldsymbol{y}$ and $\boldsymbol{y} = \boldsymbol{e}_k$, we have

$$\underset{i \in [d]}{\mathsf{argmax}} \ (\widehat{\boldsymbol{y}}_\Psi(\boldsymbol{\theta}))_i = k \quad \Longrightarrow \quad \underset{i \in [d]}{\mathsf{argmax}} \ (\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))_i = k,$$

justifying the use of $\widehat{\boldsymbol{y}}_\Omega$ at prediction time.

4. Probabilistic prediction with Fenchel-Young losses

In the previous section, we presented Fenchel-Young losses in a broad setting. We now restrict to classification over the probability simplex. More precisely, we restrict to the case $\mathcal{Y} = \{\boldsymbol{e}_i\}_{i=1}^d$, (i.e., unstructured multi-class classification), and assume that $\mathsf{dom}(\Omega) \subseteq \mathsf{conv}(\mathcal{Y}) = \triangle^d$. In this case, the regularized prediction function (5) becomes

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \underset{\boldsymbol{p} \in \triangle^d}{\mathsf{argmax}} \, \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle - \Omega(\boldsymbol{p}),$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a vector of (possibly negative) prediction scores produced by a model $\boldsymbol{f}_W(\boldsymbol{x})$ and $\boldsymbol{p} \in \triangle^d$ is a discrete probability distribution. It is a generalized exponential family distribution (Grünwald and Dawid, 2004; Frongillo and Reid, 2014) with natural parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ and regularization Ω. Of particular interest is the case where $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is **sparse**, meaning that there are scores $\boldsymbol{\theta}$ for which the resulting $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ assigns zero probability to some classes. As seen in §2.2, this happens for example with the sparsemax transformation, but not with the softmax. Later, in §5, we will establish conditions for the regularized prediction function to be sparse and will connect it to the notion of separation margin.

We first discuss generalized entropies and the properties of the Fenchel-Young losses they induce (§4.1). We then discuss their expected loss, Bayes risk and Fisher consistency (§4.2). We then give examples of generalized entropies and corresponding loss functions, several of them new to our knowledge (§4.3). Finally, we discuss the binary classification setting, recovering several examples of commonly-used loss functions (§4.4).

4.1. Fenchel-Young loss generated by a generalized entropy

Generalized entropies. A natural choice of regularization function Ω over the probability simplex is $\Omega = -\mathsf{H}$, where H is a generalized entropy (Grünwald and Dawid, 2004), also called uncertainty function by DeGroot (1962): 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 §B.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 $\boldsymbol{p} = \mathbf{1}/d$.*

That is, assumptions A.1-A.3 ensure that the uniform distribution is the maximum entropy distribution. 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 $h : [0,1] \to \mathbb{R}_+$ 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, the Fenchel-Young loss definition (9) becomes

$$L_{-\mathsf{H}}(\boldsymbol{\theta}; \boldsymbol{e}_k) = (-\mathsf{H})^*(\boldsymbol{\theta}) - \theta_k. \tag{14}$$

This form was also recently proposed by Duchi et al. (2018, Proposition 3). 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

$$L_{-\mathsf{H}}(\boldsymbol{\theta}; \boldsymbol{e}_k) = (-\mathsf{H})^*(\boldsymbol{\theta} - \theta_k \mathbf{1}). \tag{15}$$

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 (Nesterov, 2005; 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 $(\boldsymbol{\theta}; \boldsymbol{e}_k) \mapsto \max_{i \in [d]} \theta_i - \theta_k$.

It is well-known that minimizing the logistic loss $(\boldsymbol{\theta}; \boldsymbol{e}_k) \mapsto \log \sum_i \exp \theta_i - \theta_k$ is equivalent to minimizing the KL divergence between $\boldsymbol{e}_k$ and $\mathsf{softmax}(\boldsymbol{\theta})$, which in turn is equivalent to maximizing the likelihood of the ground-truth label, $\mathsf{softmax}(\boldsymbol{\theta})_k$. Importantly, this equivalence does not carry over for generalized entropies H: minimizing $L_{-\mathsf{H}}(\boldsymbol{\theta}; \boldsymbol{e}_k)$ is not in general equivalent to minimizing $B_{-\mathsf{H}}(\boldsymbol{e}_k, \widehat{\boldsymbol{y}}_{-\mathsf{H}}(\boldsymbol{\theta}))$ or maximizing $\widehat{\boldsymbol{y}}_{-\mathsf{H}}(\boldsymbol{\theta})_k$. In fact, maximizing the likelihood is generally a non-concave problem. Fenchel-Young losses can be seen as a principled way to construct a convex loss regardless of H.

4.2. Expected loss, Bayes risk and Fisher consistency

Let Y be a random variable taking values in $\mathcal{Y} = \{\boldsymbol{e}_i\}_{i=1}^d$. If H satisfies assumption A.1, we can use (14) to obtain simpler expressions of the expected loss and pointwise Bayes risk than the ones derived in §3.3.

Expected loss. Indeed, the expected loss (risk) for all $\boldsymbol{p} \in \triangle^d$ simplifies to:

$$\mathbb{E}_{\boldsymbol{p}}[L_{-\mathsf{H}}(\boldsymbol{\theta}; Y)] = \sum_{i=1}^d p_i L_{-\mathsf{H}}(\boldsymbol{\theta}; \boldsymbol{e}_i) = \sum_{i=1}^d p_i((-\mathsf{H})^*(\boldsymbol{\theta}) - \theta_j) = (-\mathsf{H})^*(\boldsymbol{\theta}) - \langle \boldsymbol{p}, \boldsymbol{\theta} \rangle. \tag{16}$$

Bayes risk and entropy. The pointwise (conditional) Bayes risk thus becomes

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{E}_{\boldsymbol{p}}[L_{-\mathsf{H}}(\boldsymbol{\theta}; Y)] = \min_{\boldsymbol{\theta} \in \mathbb{R}^d} (-\mathsf{H})^*(\boldsymbol{\theta}) - \langle \boldsymbol{p}, \boldsymbol{\theta} \rangle = \mathsf{H}(\boldsymbol{p}). \tag{17}$$

Therefore, the pointwise Bayes risk is equal to the generalized entropy H generating $L_{-\mathsf{H}}$, evaluated at $\boldsymbol{p}$. This is consistent with (13), which states that the pointwise Bayes risk is equal to the Bregman information of Y under $\boldsymbol{p}$, because

$$\mathbb{I}_{-\mathsf{H}}(Y; \boldsymbol{p}) = -\mathbb{I}_{\mathsf{H}}(Y; \boldsymbol{p}) = \mathsf{H}(\mathbb{E}_{\boldsymbol{p}}[Y]) - \mathbb{E}_{\boldsymbol{p}}[\mathsf{H}(Y)] = \mathsf{H}(\boldsymbol{p}),$$

where we used $\mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{p}$ and assumption A.1.

A similar relation between generalized entropies and pointwise (conditional) Bayes risk is well-known in the proper loss (scoring rule) literature (Grünwald and Dawid, 2004; Gneiting and Raftery, 2007; Reid and Williamson, 2010; Williamson et al., 2016). The main difference is that the minimization above is over $\mathbb{R}^d$, while it is over $\triangle^d$ in that literature (§10.1). As noted by Reid and Williamson (2011, §4.6), Bregman information can also be connected the notion of statistical information developed by DeGroot (1962), the reduction between prior and posterior uncertainty H, of which mutual information is a special case.

Fisher consistency. From (16), $\mathbb{E}_{\boldsymbol{p}}[L_{-\mathsf{H}}(\boldsymbol{\theta}; Y)] = L_{-\mathsf{H}}(\boldsymbol{\theta}; \boldsymbol{p}) + \Omega(\boldsymbol{p})$. Combined with Proposition 2, we have that the pointwise Bayes risk (17) is achieved if and only if $\widehat{\boldsymbol{y}}_{-\mathsf{H}}(\boldsymbol{\theta}) = \boldsymbol{p}$. Such losses are Fisher consistent estimators of probabilities (Williamson et al., 2016).

4.3. Examples

We now give examples of generalized entropies over the simplex $\triangle^d$ (we omit the indicator function $I_{\triangle^d}$ from the definitions since there is no ambiguity). We illustrate them together with the regularized prediction function and loss they produce in Figure 4. Several of the resulting loss functions are new to our knowledge.

Shannon entropy (Shannon and Weaver, 1949). This is the foundation of information theory, defined as

$$\mathsf{H}^{\mathrm{S}}(\boldsymbol{p}) := -\sum_{j=1}^{d} p_j \log p_j.$$

As seen in Table 1, the resulting Fenchel-Young loss $L_{-\mathsf{H}^{\mathrm{S}}}$ corresponds to the logistic loss. The associated distribution is the classical softmax, Eq. (6).

Tsallis α-entropies (Tsallis, 1988). These entropies are defined as

$$\mathsf{H}^{\mathrm{T}}_{\alpha}(\boldsymbol{p}) := k(\alpha - 1)^{-1} \left(1 - \sum_{j=1}^{d} p_j^{\alpha} \right),$$

where $\alpha \geq 0$ and k is an arbitrary positive constant. They 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., 2009a). 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:

$$\mathsf{H}^{\mathrm{T}}_{\alpha}(\boldsymbol{p}) := \sum_{j=1}^{d} h_{\alpha}(p_j) \quad \text{with} \quad h_{\alpha}(t) := \frac{t - t^{\alpha}}{\alpha(\alpha - 1)}. \tag{18}$$

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:

$$\mathsf{H}^{\mathrm{T}}_{2}(\boldsymbol{p}) = \frac{1}{2} \sum_{j=1}^{d} p_j (1 - p_j) = \frac{1}{2}(1 - \|\boldsymbol{p}\|_2^2) \quad \forall \boldsymbol{p} \in \triangle^d. \tag{19}$$

Using the constant invariance property in Proposition 2, it can be checked that $L_{-\mathsf{H}_2^{\mathrm{T}}}$ recovers the sparsemax loss (Martins and Astudillo, 2016) (cf. Table 1).

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

Norm entropies. An interesting class of non-separable entropies are entropies generated by a q-norm, defined as

$$\mathsf{H}_q^{\mathrm{N}}(\boldsymbol{p}) \coloneqq 1 - \|\boldsymbol{p}\|_q. \tag{20}$$

We call them **norm entropies**. By the Minkowski inequality, q-norms with $q > 1$ are strictly convex on the simplex, so $\mathsf{H}_q^{\mathrm{N}}$ satisfies assumptions A.1–A.3 for $q > 1$. The resulting norm entropies differ from Tsallis entropies in that the norm is not raised to the power of q: a subtle but important difference. The limit case $q \to \infty$ is particularly interesting: in this case, we obtain $\mathsf{H}_\infty^{\mathrm{N}} = 1 - \|\cdot\|_\infty$, recovering the Berger-Parker dominance index (Berger and Parker, 1970), widely used in ecology to measure species diversity. We surprisingly encounter $\mathsf{H}_\infty^{\mathrm{N}}$ again in Section 5, as a limit case for the existence of separation margins.

Squared norm entropies. Inspired by Niculae and Blondel (2017), as a simple extension of the Gini index (19), we consider the generalized entropy based on squared q-norms:

$$\mathsf{H}_q^{\mathrm{SQ}}(\boldsymbol{p}) \coloneqq \frac{1}{2}(1 - \|\boldsymbol{p}\|_q^2) = \frac{1}{2} - \frac{1}{2}\left(\sum_{j=1}^d p_j^q\right)^{\frac{2}{q}}.$$

The constant term $\frac{1}{2}$, omitted by Niculae and Blondel (2017), ensures satisfaction of A.1. For $q \in (1, 2]$, it is known that the squared q-norm is strongly convex w.r.t. $\|\cdot\|_q$ (Ball et al., 1994), implying that $(-\mathsf{H}_q^{\mathrm{SQ}})^*$, and therefore $L_{-\mathsf{H}_q^{\mathrm{SQ}}}$, is smooth. Although $\widehat{\boldsymbol{y}}_{-\mathsf{H}_q^{\mathrm{SQ}}}(\boldsymbol{\theta})$ cannot to our knowledge be solved in closed form for $q \in (1, 2)$, efficient iterative algorithms such as projected gradient are available.

Rényi $\boldsymbol{\beta}$-entropies. Rényi entropies (Rényi, 1961) are defined for any $\beta \geq 0$ as:

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

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 $\mathsf{H}_\beta^{\mathrm{R}}(\boldsymbol{p}) = -\log \|\boldsymbol{p}\|_\infty$. 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 (Mangasarian, 1965), meaning that, for all $\boldsymbol{p}, \boldsymbol{q} \in \triangle^d$, $\langle \nabla \mathsf{H}_\beta^{\mathrm{R}}(\boldsymbol{p}), \boldsymbol{q} - \boldsymbol{p} \rangle \leq 0$ implies $\mathsf{H}_\beta^{\mathrm{R}}(\boldsymbol{q}) \leq \mathsf{H}_\beta^{\mathrm{R}}(\boldsymbol{p})$. This implies, among other things, that points $\boldsymbol{p} \in \triangle^d$ with zero gradient are maximizers of $\langle \boldsymbol{p}, \boldsymbol{\theta} \rangle + \mathsf{H}_\beta^{\mathrm{R}}(\boldsymbol{p})$, which allows us to compute the predictive distribution $\widehat{\boldsymbol{y}}_{-\mathsf{H}_\beta^{\mathrm{R}}}$ with gradient-based methods.

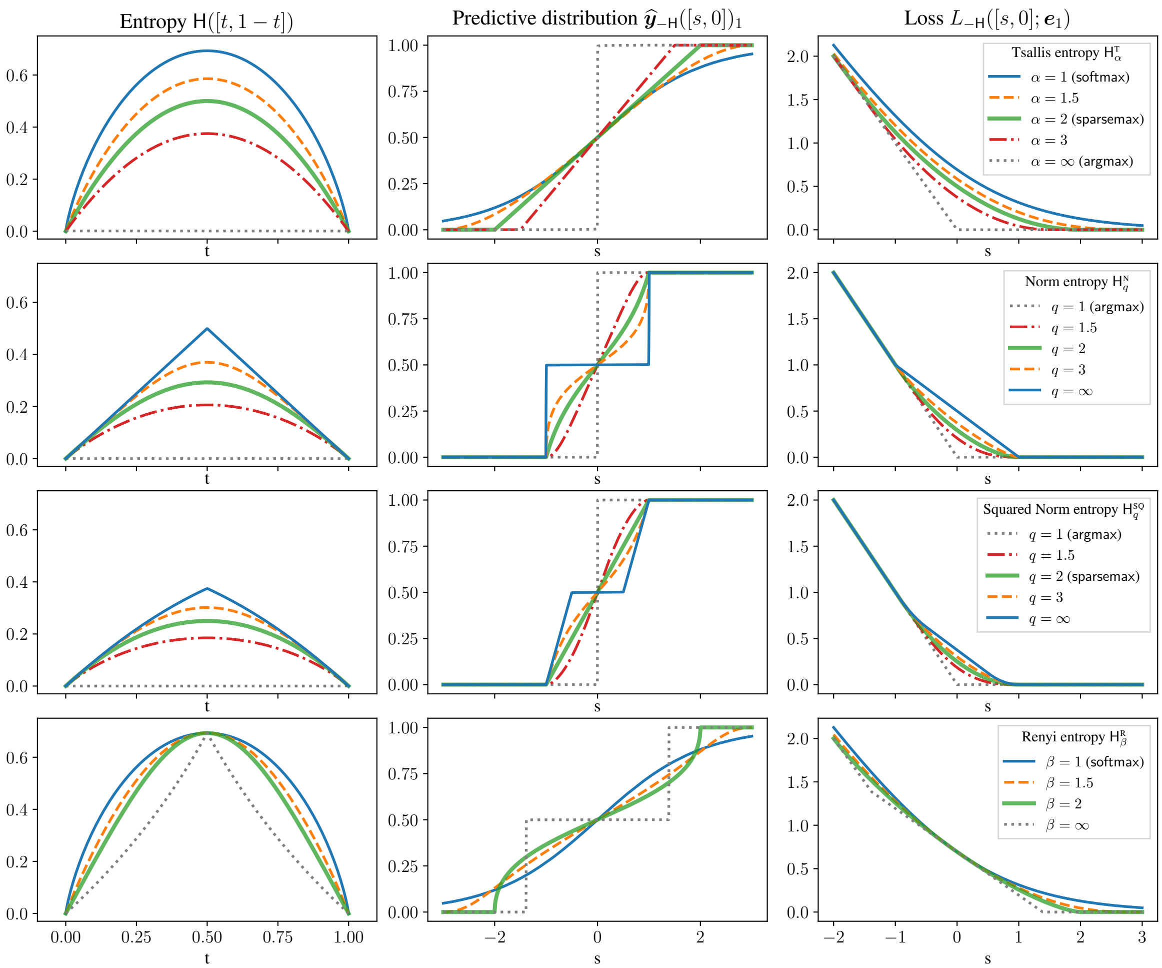

Figure 4: Examples of **generalized entropies** (left) along with their **prediction distribution** (middle) and **Fenchel-Young losses** (right) for the binary case, where $\boldsymbol{p} = [t, 1-t] \in \triangle^2$ and $\boldsymbol{\theta} = [s, 0] \in \mathbb{R}^2$. Except for softmax, which never exactly reaches 0, all distributions shown on the center can have **sparse support**.

4.4. Binary classification case

When $d = 2$, the convex conjugate expression simplifies and reduces to a univariate maximization problem over $[0, 1]$:

$$\Omega^*(\boldsymbol{\theta}) = \max_{\boldsymbol{p} \in \triangle^2} \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle - \Omega(\boldsymbol{p}) = \max_{p \in [0,1]} \theta_1 p + \theta_2 (1-p) - \Omega([p, 1-p]) = \phi^*(\theta_1 - \theta_2) + \theta_2,$$

where we defined $\phi(p) \coloneqq \Omega([p, 1-p])$. Likewise, it is easy to verify that we also have $\Omega^*(\boldsymbol{\theta}) = \phi^*(\theta_2 - \theta_1) + \theta_1$. Let us choose $\boldsymbol{\theta} = [s, -s] \in \mathbb{R}^2$ for some $s \in \mathbb{R}$. Since the ground truth is $\boldsymbol{y} \in \{\boldsymbol{e}_1, \boldsymbol{e}_2\}$, where $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ represent positive and negative classes, from (15), we can write

$$L_\Omega(\boldsymbol{\theta}; \boldsymbol{e}_j) = \Omega^*(\boldsymbol{\theta} - \theta_j \mathbf{1}) = \begin{cases} \Omega^*([0, -s]) = \phi^*(-s) & \text{if } j = 1 \\ \Omega^*([s, 0]) = \phi^*(s) & \text{if } j = 2 \end{cases}.$$

This can be written concisely as $\phi^*(-ys)$ where $y \in \{+1, -1\}$. Losses that can be written in this form are sometimes called margin losses (Reid and Williamson, 2010). These are losses that treat the positive and negative classes symmetrically. We have thus made a connection between margin losses and the regularization function $\phi(p) \coloneqq \Omega([p, 1-p])$. Note, however, that this is different from the notion of margin we develop in §5.

Examples. Choosing $\phi(p) = -\mathsf{H}^{\mathrm{S}}([p, 1-p])$ leads to the binary logistic loss

$$\phi^*(-ys) = \log(1 + \exp(-ys)).$$

Choosing $\phi(p) = -\mathsf{H}_2^{\mathrm{T}}([p, 1-p]) = p^2 - p$ leads to

$$\phi^*(u) = \begin{cases} 0 & \text{if } u \leq -1 \\ u & \text{if } u \geq 1 \\ \frac{1}{4}(u+1)^2 & \text{o.w.} \end{cases}.$$

The resulting loss, $\phi(-ys)$, is known as the modified Huber loss in the literature (Zhang, 2004). Hence, the sparsemax loss is the multiclass extension of the modified Huber loss, as as already noted in (Martins and Astudillo, 2016).

Finally, we note that the modified Huber loss is closely related to the smoothed hinge loss (Shalev-Shwartz and Zhang, 2016). Indeed, choosing $\phi(p) = \frac{1}{2}p^2 - p$ (notice the $\frac{1}{2}$ factor), we obtain

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

It can be verified that $\phi^*(-ys)$ is indeed the smoothed hinge loss.

5. Separation margin of Fenchel-Young 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 4 *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:*

$$\theta_k \geq m + \max_{j \neq k} \theta_j \quad \Rightarrow \quad L(\boldsymbol{\theta}; \boldsymbol{e}_k) = 0. \tag{21}$$

The smallest possible m that satisfies (21) is called the margin *of L, denoted* $\operatorname{margin}(L)$.

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 important 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 §B.4, 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 5 *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,*

$$\operatorname{margin}(L_{-\mathsf{H}}) = \sup_{\boldsymbol{p} \in \triangle^d} \frac{\mathsf{H}(\boldsymbol{p})}{1 - \|\boldsymbol{p}\|_\infty}. \tag{22}$$

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 (22) is the generalized entropy $\mathsf{H}^{\mathrm{N}}_\infty(\boldsymbol{p})$ introduced in §4: the **∞-norm entropy**. As Figure 4 suggests, this entropy provides an upper bound for convex losses with unit margin. This provides some intuition to the formula (22), which seeks a distribution $\boldsymbol{p}$ maximizing the **entropy ratio** between $\mathsf{H}(\boldsymbol{p})$ and $\mathsf{H}^{\mathrm{N}}_\infty(\boldsymbol{p})$.

Relationship between sparsity and margins. The next result, proved in §B.5, 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 6 *Relationship between margin losses and sparse predictive probabilities*

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 6. 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. As stated in Definition 3, functions whose gradient "explodes" in the boundary of their domain (hence failing to meet the first condition in Proposition 6) 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 (Definition 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 4.

Margin computation. For Fenchel-Young losses that have the separation margin property, Proposition 5 provided a formula for determining the margin. While informative, formula (22) is not very practical, as it involves a generally non-convex optimization problem. The next proposition, proved in §B.6, 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 7 *Assume* H *satisfies the conditions in Proposition 6 and is twice-differentiable on the simplex. Then, for arbitrary* $j \neq k$*:*

$$\operatorname{margin}(L_{-\mathsf{H}}) = \nabla_j \mathsf{H}(\boldsymbol{e}_k) - \nabla_k \mathsf{H}(\boldsymbol{e}_k). \tag{23}$$

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*

$$\operatorname{margin}(L_{-\mathsf{H}}) = h'(0) - h'(1) = -\int_0^1 h''(t) dt. \tag{24}$$

The compact formula (23) provides a geometric characterization of separable entropies and their margins: (24) 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 7 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 4, in the binary case.

6. Positive measure prediction with Fenchel-Young losses

In this section, we again restrict to classification and $\mathcal{Y} = \{\boldsymbol{e}_i\}_{i=1}^d$ but now assume that $\mathsf{dom}(\Omega) \subseteq \mathsf{cone}(\mathcal{Y}) = \mathbb{R}^d_+$, where $\mathsf{cone}(\mathcal{Y})$ is the conic hull of $\mathcal{Y}$. In this case, the regularized prediction function (5) becomes

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \operatorname*{argmax}_{\boldsymbol{m} \in \mathbb{R}^d_+} \langle \boldsymbol{\theta}, \boldsymbol{m} \rangle - \Omega(\boldsymbol{m}),$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is again a vector of prediction scores and $\boldsymbol{m} \in \mathbb{R}^d_+$ can be interpreted as a discrete positive measure (unnormalized probability distribution).

We first demonstrate how Fenchel-Young losses over positive measures allow to recover **one-vs-all reductions** (§6.1), theoretically justifying this popular scheme. We then give examples of loss instantiations (§6.2).

6.1. Uniformly separable regularizers and one-vs-all loss functions

A particularly simple case is that of uniformly separable Ω, i.e., $\Omega(\boldsymbol{m}) = \sum_{j=1}^d \phi(m_j) + I_{\mathbb{R}_+^d}(\boldsymbol{m})$, for some $\phi\colon \mathbb{R}_+ \to \mathbb{R}$. In that case the regularized prediction function can be computed in a coordinate-wise fashion:

$$(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))_j = \operatorname*{argmax}_{m \in \mathbb{R}_+} \; m\theta_j - \phi(m).$$

As we shall later see, this simplified optimization problem often enjoys a closed-form solution. Intuitively, $(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))_j$ can be interpreted as the "unnormalized probability" of class j. Likewise, the corresponding Fenchel-Young loss is separable over classes:

$$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle = \sum_{j=1}^d \phi^*(\theta_j) + \phi(y_j) - \theta_j y_j = \sum_{j=1}^d L_\phi(\theta_j; y_j),$$

where $y_j \in \{1, 0\}$ indicates membership to class j. The separability allows to train the model producing each θ_j in an embarrassingly parallel fashion.

Let us now consider the case $\phi(p) = -\mathsf{H}([p, 1-p])$, where H is a generalized entropy and ϕ is restricted to $[0,1]$. From §4.4, we know that $\phi^*(s) = \phi^*(-s) + s$ for all $s \in \mathbb{R}$. If H further satisfies assumption A.1, meaning that $\phi(1) = \phi(0) = 0$, then we obtain

$$L_\phi(s; 1) = \phi^*(s) + \phi(1) - s = \phi^*(-s) \quad \text{and} \quad L_\phi(s; 0) = \phi^*(s) + \phi(0) = \phi^*(s).$$

Combining the two, we obtain

$$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{j=1}^d \phi^*(-(2y_i - 1)\theta_i),$$

where we used that $(2y_i - 1) \in \{+1, -1\}$. Fenchel-Young losses thus recover classical **one-vs-all** loss functions. Since Fenchel-Young losses satisfy $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = 0 \Leftrightarrow \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \boldsymbol{y}$, our framework provides a theoretical justification for one-vs-all, a scheme that works well in practice (Rifkin and Klautau, 2004). Further, our framework justifies using $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ as a measure of class membership, as commonly implemented (with post-normalization) in software packages (Pedregosa et al., 2011; Buitinck et al., 2013).

Finally, we point out that in the binary case, choosing $\boldsymbol{\theta} = [s, -s]$, we have the relationship

$$L_\Omega(\boldsymbol{\theta}; \boldsymbol{e}_1) = 2\phi^*(-s) \quad \text{and} \quad L_\Omega(\boldsymbol{\theta}; \boldsymbol{e}_2) = 2\phi^*(s).$$

Thus, we recover the same loss as in §4.4, up to a constant 2 factor (i.e., learning over the simplex $\triangle^d$ or over the unit cube $[0,1]^d$ is equivalent for binary classification).

6.2. Examples

Recall that $\Omega(\boldsymbol{m}) = \sum_{j=1}^d \phi(m_j) + I_{\mathbb{R}_+^d}(\boldsymbol{m})$. Choosing $\phi(p) = \frac{1}{2}p^2$ gives

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = [\boldsymbol{\theta}]_+ \quad \text{and} \quad L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^d \frac{1}{2}[\theta_i]_+^2 + \frac{1}{2}y_i^2 - \theta_i y_i.$$

Choosing $\phi(p) = -\mathsf{H}^{\text{S}}([p, 1-p]) + I_{[0,1]}(p)$ leads to

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \mathsf{sigmoid}(\boldsymbol{\theta}) \coloneqq \frac{\mathbf{1}}{\mathbf{1} + \exp(-\boldsymbol{\theta})}, \quad \text{and} \quad L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{d} \log(1 + \exp(-(2y_i - 1)\theta_i))$$

the classical sigmoid function and the one-vs-all logistic function. Note that $-\Omega(\boldsymbol{p}) = \sum_j \mathsf{H}^{\text{S}}([p_j, 1-p_j]) = -\sum_j p_j \log p_j + (1-p_j)\log(1-p_j)$ is sometimes known as the Fermi-Dirac entropy (Borwein and Lewis, 2010, Commentary of Section 3.3).

Choosing $\phi(p) = -\mathsf{H}_2^{\text{T}}([p, 1-p]) + I_{[0,1]}(p) = p^2 - p + I_{[0,1]}(p)$ leads to

$$(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))_j = (\phi^*)'(s_j) = \begin{cases} 0 & \text{if } u \leq -1 \\ 1 & \text{if } u \geq 1 \\ \frac{1}{2}(u+1) & \text{o.w.} \end{cases} \quad \forall j \in [d].$$

We can think of this function as a "sparse sigmoid".

7. Structured prediction with Fenchel-Young losses

In this section, we now turn to prediction over the convex hull of a set $\mathcal{Y}$ of structured objects (sequences, trees, assignments, etc.), represented as d-dimensional vectors, i.e. $\mathcal{Y} \subseteq \mathbb{R}^d$. In this case, the regularized prediction function (5) becomes

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})} \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - \Omega(\boldsymbol{\mu}).$$

Typically, $|\mathcal{Y}|$ will be exponential in d, i.e., $d \ll |\mathcal{Y}|$. Because $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} p(\boldsymbol{y})\boldsymbol{y} = \mathbb{E}_{\boldsymbol{p}}[Y]$ for some $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$, $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ can be interpreted as the expectation under some (not necessarily unique) underlying distribution.

We first discuss the concepts of probability distribution regularization (§7.1) and mean regularization (§7.2), and the implications in terms of computational tractability and identifiability (note that in the unstructured setting, probability distribution and mean regularizations coincide). We then give several examples of polytopes and show how a Fenchel-Young loss can seamlessly be constructed over them given access to regularized prediction functions (§7.3). Finally, we extend the notion of separation margin to the structured setting (§7.4).

7.1. Distribution regularization, marginal inference and structured sparsemax

In this section, we discuss the concept of **probability distribution regularization**. Our treatment follows closely the variational formulations of exponential families (Barndorff-Nielsen, 1978; Wainwright and Jordan, 2008) and generalized exponential families (Grünwald and Dawid, 2004; Frongillo and Reid, 2014) but adopts the novel viewpoint of regularized prediction functions. We discuss two instances of that framework: the structured counterpart of the softmax, marginal inference, and a new structured counterpart of the sparsemax, structured sparsemax.

Let us start with a probability-space regularized prediction function

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

where H is a generalized entropy and $\boldsymbol{s}_{\boldsymbol{\theta}} \coloneqq (\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle)_{\boldsymbol{y} \in \mathcal{Y}} \in \mathbb{R}^{|\mathcal{Y}|}$ is a vector that contains the scores of all possible structures. Note that in the unstructured setting, where $\mathcal{Y} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$, we have $\boldsymbol{s}_{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and hence the distinction between $\boldsymbol{\theta}$ and $\boldsymbol{s}_{\boldsymbol{\theta}}$ is not necessary. In the structured prediction setting, however, the above optimization problem is defined over a space of size $|\mathcal{Y}| \gg d$. The regularized prediction function outputs a probability distribution over structures and from Danskin's theorem, we have

$$\widehat{\boldsymbol{y}}_{-\mathsf{H}}(\boldsymbol{s}_{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{s}_{\boldsymbol{\theta}}} (-\mathsf{H})^*(\boldsymbol{s}_{\boldsymbol{\theta}}) = \boldsymbol{p}^\star \in \triangle^{|\mathcal{Y}|}.$$

Using the chain rule, differentiating w.r.t. $\boldsymbol{\theta}$ instead of $\boldsymbol{s}_{\boldsymbol{\theta}}$ gives

$$\nabla_{\boldsymbol{\theta}} (-\mathsf{H})^*(\boldsymbol{s}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{p}^\star}[Y] \in \mathsf{conv}(\mathcal{Y}).$$

Therefore, the gradient w.r.t. $\boldsymbol{\theta}$ corresponds to the mean under $\boldsymbol{p}^\star$. This can be equivalently expressed under our framework by defining a regularization function directly in mean space. Let us define the regularization function Ω over $\mathsf{conv}(\mathcal{Y})$ as

$$-\Omega(\boldsymbol{\mu}) \coloneqq \max_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \mathsf{H}(\boldsymbol{p}) \quad \text{s.t.} \quad \mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}. \tag{25}$$

That is, among all distributions satisfying the first-moment matching condition $\mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}$, we seek the distribution $\boldsymbol{p}^\star$ with maximum (generalized) entropy. This allows to make the underlying distribution unique and identifiable. With this choice of Ω, simple calculations show that the corresponding regularized prediction function is precisely the mean under that distribution:

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

Similar results hold for higher order moments: If Ω^* is twice-differentiable, the Hessian corresponds to the second moment under $\boldsymbol{p}^\star$:

$$\operatorname{cov}_{\boldsymbol{p}^\star}[Y] = \nabla^2 \Omega^*(\boldsymbol{\theta}).$$

The relation between these various maps is summarized in Figure 5.

Marginal inference. A particular case of (25) is

$$-\Omega(\boldsymbol{\mu}) = \max_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \mathsf{H}^{\mathrm{S}}(\boldsymbol{p}) \quad \text{s.t.} \quad \mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}, \tag{26}$$

where H^{S} is the Shannon entropy. The distribution achieving maximum Shannon entropy, which is unique, is the Gibbs distribution $p(\boldsymbol{y}; \boldsymbol{\theta}) \propto \exp(\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle)$.

As shown in Table 1, the resulting loss L_Ω is the CRF loss (Lafferty et al., 2001) and the resulting regularized prediction function $\widehat{\boldsymbol{y}}_\Omega$ is known as marginal inference in the literature (Wainwright and Jordan, 2008):

$$\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta}) = \mathsf{marginals}(\boldsymbol{\theta}) \coloneqq \sum_{\boldsymbol{y} \in \mathcal{Y}} \exp(\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle) \boldsymbol{y} / Z(\boldsymbol{\theta}),$$

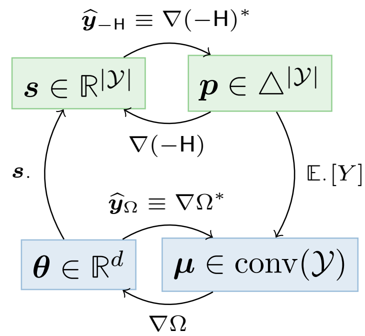

Figure 5: **Summary of maps between spaces.** We define Ω over $\mathsf{conv}(\mathcal{Y})$ using a generalized maximum entropy principle: $-\Omega(\boldsymbol{\mu}) \coloneqq \max_{\boldsymbol{p}\in\triangle^{|\mathcal{Y}|}} \mathsf{H}(\boldsymbol{p})$ s.t. $\mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}$ (cf. §7.1). We also define $\boldsymbol{s_\theta} \coloneqq (\langle \boldsymbol{\theta}, \boldsymbol{y}\rangle)_{\boldsymbol{y}\in\mathcal{Y}} \in \mathbb{R}^{|\mathcal{Y}|}$, the vector that contains the scores of all possible structures, a linear map from $\mathbb{R}^d$ to $\mathbb{R}^{|\mathcal{Y}|}$. Note that the diagram is not necessarily commutative.

where $Z(\boldsymbol{\theta}) \coloneqq \sum_{\boldsymbol{y}\in\mathcal{Y}} \exp(\langle\boldsymbol{\theta}, \boldsymbol{y}\rangle)$ is the partition function (normalization constant) of the Gibbs distribution. Although marginal inference is intractable in general, we will see in §7.3 that it can be computed exactly and efficiently for specific polytopes. The conjugate of $\Omega(\boldsymbol{\mu})$, $\Omega^*(\boldsymbol{\theta})$, corresponds to the log partition function: $\Omega^*(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$.

Structured sparsemax. From the above perspective, it is tempting to define a sparsemax counterpart of the CRF loss and of marginal inference by replacing Shannon's entropy H^{S} with the Gini index $\mathsf{H}_2^{\mathrm{T}}$ (see (19) for a definition) in (26):

$$-\Omega(\boldsymbol{\mu}) = \max_{\boldsymbol{p}\in\triangle^{|\mathcal{Y}|}} \mathsf{H}_2^{\mathrm{T}}(\boldsymbol{p}) \quad \text{s.t.} \quad \mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}.$$

This is equivalent to seeking a distribution $\boldsymbol{p}^\star$ with minimum squared norm

$$\Omega(\boldsymbol{\mu}) + \frac{1}{2} = \min_{\boldsymbol{p}\in\triangle^{|\mathcal{Y}|}} \frac{1}{2}\|\boldsymbol{p}\|^2 \quad \text{s.t.} \quad \mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}.$$

The corresponding $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is then $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{p}^\star}[Y]$. Recalling that $\boldsymbol{p}^\star = \widehat{\boldsymbol{y}}_{-\mathsf{H}_2^{\mathrm{T}}}(\boldsymbol{s_\theta})$, a naive approach would be to first compute the score vector $\boldsymbol{s_\theta} = (\langle\boldsymbol{\theta}, \boldsymbol{y}\rangle)_{\boldsymbol{y}\in\mathcal{Y}} \in \mathbb{R}^{|\mathcal{Y}|}$ and then to project that vector on the simplex to obtain $\boldsymbol{p}^\star$. Unfortunately, $\boldsymbol{s_\theta}$ could be exponentially large and $\boldsymbol{p}^\star$ could be arbitrarily dense in the worst case, making that approach intractable in general. As also noted by Pillutla et al. (2018), we can approximate the problem by further imposing an upper-bound on the sparsity of the distribution, $\|\boldsymbol{p}\|_0 \leq k$. Since the optimal k-sparse distribution solely depends on the top-k elements of $\boldsymbol{s_\theta}$ (Kyrillidis et al., 2013), we can first use a k-best oracle to retrieve the top-k elements of $\boldsymbol{s_\theta}$ and project them onto the simplex. A disadvantage of that approach is that k-best oracles are usually more complex than MAP oracles and k could be arbitrarily large in order to guarantee an exact solution to the original problem. For shortest path problems over a directed acyclic graph, an alternative approximation is to smooth the maximum operator directly in the dynamic programming recursion (Mensch and Blondel, 2018).

7.2. Mean regularization and SparseMAP

Marginal inference is computationally tractable only for certain polytopes and structured sparsemax requires approximations, even for polytopes for which exact marginal inference is available. In this section, we discuss an alternative approach which only requires access to a MAP oracle, broadening the set of applicable polytopes. The key idea is **mean regularization**: We directly regularize the mean $\boldsymbol{\mu}$ rather than the distribution $\boldsymbol{p}$.

The mean regularization counterpart of marginal inference is

$$\widehat{\boldsymbol{y}}_{-\mathsf{H}^{\mathrm{s}}}(\boldsymbol{\theta}) = \underset{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})}{\mathsf{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle + \mathsf{H}^{\mathrm{s}}(\boldsymbol{\mu}) = \underset{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})}{\mathsf{argmin}} \ \mathsf{KL}(\boldsymbol{\mu} \| e^{\boldsymbol{\theta} - \mathbf{1}}).$$

It has been used for specific convex polytopes, most importantly in the optimal transport literature (Cuturi, 2013; Peyré and Cuturi, 2017) but also for learning to predict permutation matrices (Helmbold and Warmuth, 2009) or permutations (Yasutake et al., 2011; Ailon et al., 2016). The mean regularization counterpart of sparsemax is known as SparseMAP (Niculae et al., 2018):

$$\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta}) = \mathsf{SparseMAP}(\boldsymbol{\theta}) \coloneqq \underset{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})}{\mathsf{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \frac{1}{2} \|\boldsymbol{\mu}\|^2 = \underset{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})}{\mathsf{argmin}} \ \|\boldsymbol{\mu} - \boldsymbol{\theta}\|^2.$$

The main advantage of mean regularization is computational. Indeed, computing $\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta})$ now simply involves a d-dimensional optimization problem instead of a $|\mathcal{Y}|$-dimensional one and can be cast as a Bregman projection onto $\mathsf{conv}(\mathcal{Y})$(§3.2). For specific polytopes, that projection can often be computed directly (§7.3). More generally, it can always be computed to arbitrary precision given access to a MAP oracle

$$\underset{\boldsymbol{y} \in \mathsf{conv}(\mathcal{Y})}{\mathsf{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle = \underset{\boldsymbol{y} \in \mathcal{Y}}{\mathsf{argmax}} \, \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle \eqqcolon \mathsf{MAP}(\boldsymbol{\theta}),$$

thanks to conditional gradient algorithms (§8.3). Since MAP inference is a cornerstone of structured prediction, efficient algorithms have been developed for many kinds of structures (§7.3). In addition, as conditional gradient algorithms maintain a convex combination of vertices, they can also return a (not necessarily unique) distribution $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$ such that $\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{p}}[Y]$. From Carathéodory's theorem, the support of $\boldsymbol{p}$ contains at most $d \ll |\mathcal{Y}|$ elements. This ensures that $\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta})$ can be written as a "small" number of elementary structures. The price to pay for this computational tractability is that the underlying distribution $\boldsymbol{p}$ is not necessarily unique.

7.3. Examples

Deriving loss functions for structured outputs can be challenging. In this section, we give several examples of polytopes $\mathsf{conv}(\mathcal{Y})$ for which efficient computational oracles (MAP, marginal inference, projection) are available, thus allowing to obtain a Fenchel-Young loss for learning over these polytopes. A summary is given in Table 7.3.

Table 2: **Examples of convex polytopes and computational cost of the three main computational oracles:** MAP, marginal inference and projection (in the Euclidean distance sense or more generally in the Bregman divergence sense). For polytopes for which direct marginal inference or projection algorithms are not available, we can always use conditional gradient algorithms to compute an optimal solution to arbitrary precision — see §8.3.

Polytope	Vertices	Dim.	MAP	Marginal	Projection
Probability simplex	Basis vectors	d	$O(d)$	$O(d)$	$O(d)$
	Sequences	nm^2	$O(nm^2)$	$O(nm^2)$	N/A
Arborescence	Spanning trees	$n(n-1)$	$O(n^2)$	$O(n^3)$	N/A
	Alignments	nm	$O(nm)$	$O(nm)$	N/A
Permutahedron	Permutations	d	$O(d \log d)$	N/A	$O(d \log d)$
Birkhoff	Permutation matrices	n^2	$O(n^3)$	#P-complete	$O(n^2/\epsilon)$

Sequences. We wish to tag a sequence $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ of vectors in $\mathbb{R}^p$ (*e.g.,* word representations) with the most likely output sequence (*e.g.,* entity tags) $\boldsymbol{s} = (s_1, \dots, s_n) \in [m]^n$. It is convenient to represent each sequence $\boldsymbol{s}$ as a $n \times m \times m$ binary tensor $\boldsymbol{y} \in \mathcal{Y}$, such that $y_{t,i,j} = 1$ if $\boldsymbol{y}$ transitions from node j to node i on time t, and 0 otherwise. The potentials $\boldsymbol{\theta}$ can similarly be organized as a $n \times m \times m$ real tensor, such that $\theta_{t,i,j} = \phi_t(\boldsymbol{x}_t, i, j)$, where ϕ_t is a potential function. Using the above binary tensor representation, the Frobenius inner product $\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$ is equal to $\sum_{t=1}^{n} \phi_t(\boldsymbol{x}_t, s_t, s_{t-1})$, the cumulated score of $\boldsymbol{s}$.

MAP inference, $\mathsf{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{y}, \boldsymbol{\theta} \rangle$, seeks the highest-scoring sequence and can be computed using Viterbi's algorithm (Viterbi, 1967) in $O(nm^2)$ time. Marginal inference can be computed in the same cost using the forward-backward algorithm (Baum and Petrie, 1966) or equivalently using backpropagation (Eisner, 2016; Mensch and Blondel, 2018).

When Ω is defined over $\mathsf{conv}(\mathcal{Y})$ as in (26), the Fenchel-Young loss corresponds to a linear-chain conditional random field loss (Lafferty et al., 2001). In recent years, this loss has been used to train various end-to-end natural language pipelines based on neural networks (Collobert et al., 2011; Lample et al., 2016).

Alignments. Let $\boldsymbol{A} \in \mathbb{R}^{m \times p}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times p}$ be two time-series of lengths m and n, respectively. We denote by $\boldsymbol{a}_i \in \mathbb{R}^p$ and $\boldsymbol{b}_j \in \mathbb{R}^p$ their i^{th} and j^{th} observations. Our goal is to find an alignment between $\boldsymbol{A}$ and $\boldsymbol{B}$, matching their observations. We define $\boldsymbol{\theta}$ as a $m \times n$ matrix, such that $\theta_{i,j}$ is the similarity between observations $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$. Likewise, we represent an alignment $\boldsymbol{y}$ as a $m \times n$ binary matrix, such that $y_{i,j} = 1$ if $\boldsymbol{a}_i$ is aligned with $\boldsymbol{b}_j$, and 0 otherwise. We write $\mathcal{Y}$ the set of all monotonic alignment matrices, such that the path that connects the upper-left $(1, 1)$ matrix entry to the lower-right (m, n) one uses only $\downarrow, \rightarrow, \searrow$ moves.

In this setting, MAP inference, $\mathsf{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{y}, \boldsymbol{\theta} \rangle$, corresponds to seeking the maximal similarity (or equivalently, minimal cost) alignment between the two time-series. It can be computed in $O(mn)$ time using dynamic time warping, DTW (Sakoe and Chiba, 1978).

Marginal inference can be computed in the same cost by backpropagation, if we replace the hard minimum with a soft one in the DTW recursion. This algorithm is known as soft-DTW (Cuturi and Blondel, 2017; Mensch and Blondel, 2018).

Structured SVMs were combined with DTW to learn to predict music-to-score alignments (Garreau et al., 2014). Our framework easily enables extensions of this work, such as replacing DTW with soft-DTW, which amounts to introducing entropic regularization w.r.t. the probability distribution over alignments.

Spanning trees. When $\mathcal{Y}$ is the set of possible directed spanning trees (arborescences) of a complete graph $\mathcal{G}$ with n vertices, the convex hull $\mathsf{conv}(\mathcal{Y}) \subset \mathbb{R}^{n(n-1)}$ is known as the arborescence polytope (Martins et al., 2009b) (each $\boldsymbol{y} \in \mathcal{Y}$ is a binary vector which indicates which arcs belong to the arborescence). MAP inference may be performed by maximal arborescence algorithms (Chu and Liu, 1965; Edmonds, 1967) in $O(n^2)$ time (Tarjan, 1977), and the Matrix-Tree theorem (Kirchhoff, 1847) provides an $O(n^3)$ marginal inference algorithm (Koo et al., 2007; Smith and Smith, 2007). Spanning tree structures are often used in natural language processing for (non-projective) *dependency parsing*, with graph edges corresponding to dependencies between the words in a sentence. (McDonald et al., 2005; Kiperwasser and Goldberg, 2016).

Permutations. We view ranking as a structured prediction problem. Let $\mathcal{Y}$ be the set of d-permutations of a prescribed vector $\boldsymbol{w} \in \mathbb{R}^d$, i.e., $\mathcal{Y} = \{[w_{\pi_1}, \dots, w_{\pi_d}] \in \mathbb{R}^d \colon \boldsymbol{\pi} \in \Pi\}$, where Π denotes the permutations of $(1, \dots, d)$. We assume without loss of generality that $\boldsymbol{w}$ is sorted in descending order, i.e., $w_1 \geq \dots \geq w_d$. MAP inference seeks the permutation of $\boldsymbol{w}$ whose inner product with $\boldsymbol{\theta}$ is maximized:

$$\max_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle = \max_{\boldsymbol{y} \in \mathcal{Y}} \sum_{i=1}^{d} \theta_i y_i = \max_{\boldsymbol{\pi} \in \Pi} \sum_{i=1}^{d} \theta_i w_{\pi_i} = \max_{\boldsymbol{\pi} \in \Pi} \sum_{i=1}^{d} \theta_{\pi_i} w_i.$$

Since $\boldsymbol{w}$ is assumed sorted, an optimal solution $\boldsymbol{\pi}^\star$ of the last optimization problem is a permutation sorting $\boldsymbol{\theta}$ in descending order. The function $\boldsymbol{\theta} \mapsto \sum_{i=1}^d \theta_{\pi_i^\star} w_i$ is called ordered weighted averaging (OWA) operator (Yager, 1988) and includes the mean and max operator as special cases. MAP inference over $\mathcal{Y}$ can be seen as the variational formulation of the OWA operator. An optimal solution $\boldsymbol{y}^\star$ is simply $\boldsymbol{w}$ sorted using the inverse permutation of $\boldsymbol{\pi}^\star$. The overall computational cost is therefore $O(d \log d)$, for sorting $\boldsymbol{\theta}$.

The convex hull of $\mathcal{Y}$, $\mathsf{conv}(\mathcal{Y})$, is known as the *permutahedron* when $\boldsymbol{w} = [d, \dots, 1]$. For arbitrary $\boldsymbol{w}$, we follow Lim and Wright (2016) and call $\mathsf{conv}(\mathcal{Y})$ the permutahedron induced by $\boldsymbol{w}$. Its vertices correspond to permutations of $\boldsymbol{w}$.

We can define mean-regularized ranking prediction functions $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ if we choose $\boldsymbol{w}$ such that each w_i represents the "preference" of being in position i. Intuitively, the score vector $\boldsymbol{\theta} \in \mathbb{R}^d$ should be such that each θ_i is the score of instance i (e.g., a document or a label). We give several examples below.

- Choosing $\boldsymbol{w} = [1, 0, \dots, 0]$, $\mathsf{conv}(\mathcal{Y})$ is equal to $\triangle^d$, the probability simplex. We thus recover probabilistic classification as a natural special case.

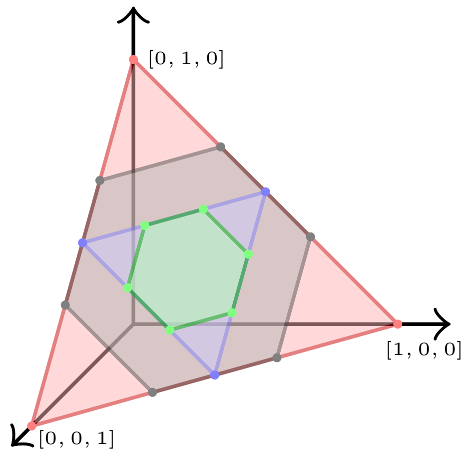

Figure 6: **Examples of instances of permutahedron induced by $\boldsymbol{w}$.** Round circles indicate vertices of the permutahedron, permutations of $\boldsymbol{w}$. Choosing $\boldsymbol{w} = [1, 0, 0]$ recovers the probability simplex (red) while choosing $\boldsymbol{w} = [\frac{1}{2}, \frac{1}{2}, 0]$ recovers the capped probability simplex (blue). The other two instances are obtained by $\boldsymbol{w} = [\frac{2}{3}, \frac{1}{3}, 0]$ (gray) and $\boldsymbol{w} = [\frac{1}{2}, \frac{1}{3}, \frac{1}{6}]$ (green). Euclidean projection onto these polytopes can be cast as isotonic regression. More generally, Bregman projection reduces to isotonic *optimization*.

- Choosing $\boldsymbol{w} = \frac{1}{k}[\underbrace{1, \dots, 1}_{k \text{ times}}, 0, \dots, 0]$, $\mathsf{conv}(\mathcal{Y})$ is equal to $\{\boldsymbol{p} \in \triangle^d \colon \|\boldsymbol{p}\|_\infty \leq \frac{1}{k}\}$, sometimes referred to as the capped probability simplex (Warmuth and Kuzmin, 2008; Lim and Wright, 2016). This setting corresponds to predicting k-subsets.
- Choosing $\boldsymbol{w} = \frac{2}{d(d+1)}[d, d-1, \dots, 1]$ corresponds to predicting full rankings.
- Choosing $\boldsymbol{w} = \frac{2}{k(k+1)}[k, k-1, \dots, 1, \underbrace{0, \dots, 0}_{d-k-1 \text{ times}}]$ corresponds to predicting partial rankings.

The corresponding polytopes are illustrated in Figure 6. In all the examples above, $\boldsymbol{w} \in \triangle^d$, implying $\mathsf{conv}(\mathcal{Y}) \subseteq \triangle^d$. Therefore, $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ outputs a probability distribution.

As discussed in §3.2, computing the regularized prediction function $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is equivalent to a Bregman projection when $\Omega = \Psi + I_{\mathcal{C}}$, where Ψ is Legendre type. The Euclidean projection onto $\mathcal{C} = \mathsf{conv}(\mathcal{Y})$ reduces to isotonic regression (Zeng and Figueiredo, 2015; Negrinho and Martins, 2014). The computational cost is $O(d \log d)$. More generally, Bregman projections reduce to isotonic *optimization* (Lim and Wright, 2016). This provides a unified way to compute $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ efficiently, regardless of $\boldsymbol{w}$.

The generated Fenchel-Young loss $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$, is illustrated in Figure 7 for various choices of Ω. When $\Omega = 0$, as expected, the loss is zero as long as the predicted ranking is correct. Note that in order to define a meaningful loss, it is necessary that $\boldsymbol{y} \in \mathcal{Y}$ or more generally $\boldsymbol{y} \in \mathsf{conv}(\mathcal{Y})$. That is, $\boldsymbol{y}$ should belong to the convex hull of the permutations of $\boldsymbol{w}$.

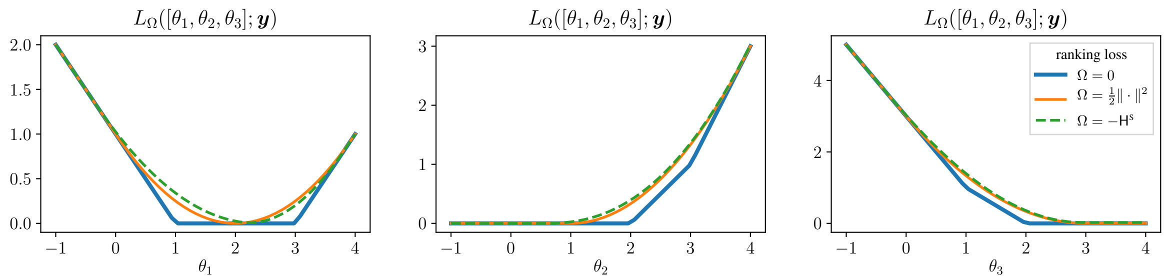

Figure 7: **Ranking losses** generated by restricting $\mathsf{dom}(\Omega)$ to the permutahedron induced by $\boldsymbol{w} = [3, 2, 1]$. We set the ground-truth $\boldsymbol{y} = [2, 1, 3]$, i.e., a permutation of $\boldsymbol{w}$. We set $\boldsymbol{\theta} = \boldsymbol{y}$ and inspect how the loss changes when varying each θ_i. When $\Omega = 0$, the loss is zero as expected when $\theta_1 \in [1, 3]$, $\theta_2 \in (-\infty, 2]$ and $\theta_3 \in [2, \infty)$, since the ground-truth ranking is still correctly predicted in these intervals. When $\Omega = \frac{1}{2}\|\cdot\|^2$ and $\Omega = -\mathsf{H}^{\mathrm{s}}$, we obtain a smooth approximation. Although not done here for clarity, if $\boldsymbol{w}$ is normalized such that $\boldsymbol{w} \in \triangle^d$, then $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \triangle^d$ as well.

Permutahedra have been used to derive online learning to rank algorithms (Yasutake et al., 2011; Ailon et al., 2016) but it is not obvious how to extract a loss from these works. Ordered weighted averaging (OWA) operators have been used to define related top-k multiclass losses (Usunier et al., 2009; Lapin et al., 2015) but without identifying the connection with permutahedra. Our construction follows directly from the general Fenchel-Young loss framework and provides a novel geometric perspective.

Permutation matrices. Let $\mathcal{Y} \subset \{0, 1\}^{n \times n}$ be the set of $n \times n$ permutation matrices. MAP inference is the solution of the linear assignment problem

$$\max_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle = \max_{\boldsymbol{\pi} \in \Pi} \sum_{i=1}^{n} \theta_{i, \pi_i},$$

where Π denotes the permutations of $(1, \ldots, n)$. The problem can be solved exactly in $O(n^3)$ time using the Hungarian algorithm (Kuhn, 1955) or the Jonker-Volgenant algorithm (Jonker and Volgenant, 1987). Noticeably, marginal inference is known to be #P-complete (Valiant, 1979; Taskar, 2004, Section 3.5). This makes it an open problem how to solve marginal inference for this polytope.

In contrast, projections on the convex hull of $\mathcal{Y}$, the set of doubly stochastic matrices known as Birkhoff polytope (Birkhoff, 1946), can be computed efficiently. Since the Birkhoff polytope is a special case of transportation polytope, we can leverage algorithms from the optimal transport literature to compute the mean-regularized prediction function. When $\Omega(\boldsymbol{\mu}) = \langle \boldsymbol{\mu}, \log \boldsymbol{\mu} \rangle$, $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ can be computed using the Sinkhorn algorithm (Sinkhorn and Knopp, 1967; Cuturi, 2013). For other regularizers, we can use Dykstra's algorithm (Dessein et al., 2016) or dual approaches (Blondel et al., 2018). The cost of obtaining an ϵ-approximate solution is typically $O(n^2/\epsilon)$.

The Birkhoff polytope has been used to define continuous relaxations of non-convex ranking losses (Adams and Zemel, 2011). In contrast, the Fenchel-Young loss over the Birkhoff polytope (new to our knowledge) is convex by construction. Although working with permutation matrices is more computationally expensive than working with permutations, it brings different modeling power, since it allows to take into account similarity between instances (e.g., text documents) through the similarity matrix $\boldsymbol{\theta}$. It also enables other applications, such as learning to match.

7.4. Structured separation margins

We end this section by extending some of the results in §5 for the structured prediction case, showing that **there is also a relation between structured entropies and margins.** In the sequel, we assume that structured outputs are contained in a sphere of radius r, i.e., $\|\boldsymbol{y}\| = r$ for any $\boldsymbol{y} \in \mathcal{Y}$. This holds for all the examples above (sequences, alignments, spanning trees, permutations of a given vector, and permutation matrices). In particular, it holds whenever outputs are represented as binary vectors with a constant number of entries set to 1; this includes overcomplete parametrizations of discrete graphical models (Wainwright and Jordan, 2008).

Definition 5 *Structured separation margin*

Let $L(\boldsymbol{\theta}; \boldsymbol{y})$ be a loss function over $\mathbb{R}^d \times \mathcal{Y}$. We say that L has the structured separation margin property *if there exists $m > 0$ such that, for any $\boldsymbol{y} \in \mathcal{Y}$:*

$$\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle \geq \max_{\boldsymbol{y}' \in \mathcal{Y}} \left(\langle \boldsymbol{\theta}, \boldsymbol{y}' \rangle + \frac{m}{2} \|\boldsymbol{y} - \boldsymbol{y}'\|^2 \right) \quad \Rightarrow \quad L(\boldsymbol{\theta}; \boldsymbol{y}) = 0. \tag{27}$$

The smallest possible m that satisfies (27) is called the margin *of L, denoted* $\mathrm{margin}(L)$.

Note that this definition generalizes the unstructured case (Definition 4), which is recovered when $\mathcal{Y} = \{\boldsymbol{e}_i\}_{i=1}^d$. Note also that, when outputs are represented as binary vectors, the term $\|\boldsymbol{y} - \boldsymbol{y}'\|^2$ is a **Hamming distance**, which counts how many bits need to be flipped to transform $\boldsymbol{y}'$ into $\boldsymbol{y}$. The most famous example of a loss with a structured separation margin in the **structured hinge loss** used in structured support vector machines (Taskar, 2004; Tsochantaridis et al., 2005).

The next proposition extends Proposition 5. We defer its proof to §B.7.

Proposition 8 *Assume Ω is convex and $\mathcal{Y}$ is contained in a sphere of radius r. Then:*

1. *The loss L_Ω has a structured separation margin iff there is a $m > 0$ such that, for any $\boldsymbol{y} \in \mathcal{Y}$, $m\boldsymbol{y} \in \partial\Omega(\boldsymbol{y})$.*
2. *If the above holds, then the margin of L_Ω is given by the smallest such m or, equivalently,*

$$\mathrm{margin}(L_\Omega) = \sup_{\boldsymbol{\mu} \in \mathrm{conv}(\mathcal{Y}),\, \boldsymbol{y} \in \mathcal{Y}} \frac{\Omega(\boldsymbol{y}) - \Omega(\boldsymbol{\mu})}{r^2 - \langle \boldsymbol{y}, \boldsymbol{\mu} \rangle}. \tag{28}$$

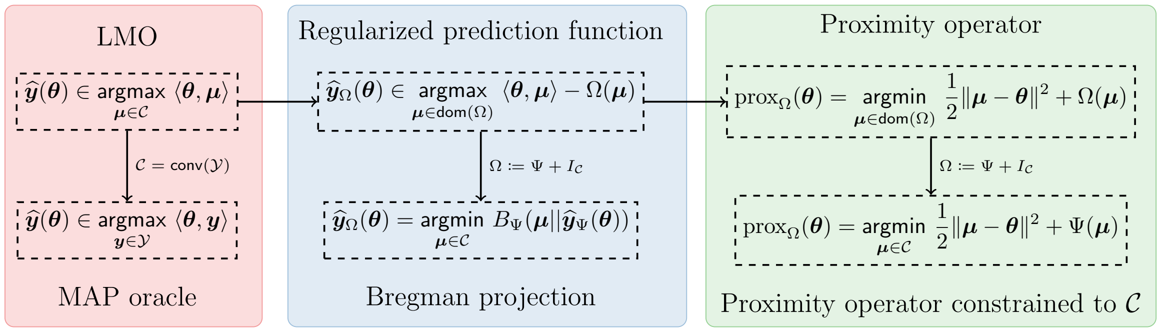

Figure 8: **Summary of computational oracles:** $\widehat{\boldsymbol{y}}(\boldsymbol{\theta})$ is used to compute unregularized predictions or as a linear maximization oracle in conditional gradient algorithms; $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is used to compute regularized predictions, loss values $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))$, where $f_{\boldsymbol{\theta}}(\boldsymbol{\mu}) \coloneqq \Omega(\boldsymbol{\mu}) - \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle$, and loss gradients $\nabla L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}$; finally, $\mathsf{prox}_\Omega(\boldsymbol{\theta})$ is used by training algorithms, in particular block coordinate ascent algorithms. In the above, we assume Ψ is Legendre-type and $\mathcal{C} \subseteq \mathsf{dom}(\Psi)$.

Unit margin of the SparseMAP loss. We can invoke Proposition 8 to show that the SparseMAP loss of Niculae et al. (2018) has a structured margin of 1, a novel result that follows directly from our construction. Indeed, from (28), we have that, for $\Omega(\boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{\mu}\|^2$:

$$\operatorname{margin}(L_\Omega) = \sup_{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y}),\, \boldsymbol{y} \in \mathcal{Y}} \frac{\frac{1}{2}r^2 - \frac{1}{2}\|\boldsymbol{\mu}\|^2}{r^2 - \langle \boldsymbol{y}, \boldsymbol{\mu} \rangle} = 1 - \inf_{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y}),\, \boldsymbol{y} \in \mathcal{Y}} \frac{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\mu}\|^2}{\langle \boldsymbol{y}, \boldsymbol{y} - \boldsymbol{\mu} \rangle} \leq 1,$$

where the last inequality follows from the fact that both the numerator and denominator in the second term are non-negative, the latter due to the Cauchy-Schwartz inequality. We now show that, for any $\boldsymbol{y} \in \mathcal{Y}$, we have $\inf_{\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})} \frac{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\mu}\|^2}{\langle \boldsymbol{y}, \boldsymbol{y} - \boldsymbol{\mu} \rangle} = 0$. Choosing $\boldsymbol{\mu} = t\boldsymbol{y}' + (1-t)\boldsymbol{y}$, for an arbitrary $\boldsymbol{y}' \in \mathcal{Y} \setminus \{\boldsymbol{y}\}$, and letting $t \to 0^+$, we obtain $\frac{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\mu}\|^2}{\langle \boldsymbol{y}, \boldsymbol{y} - \boldsymbol{\mu} \rangle} = \frac{\frac{1}{2}t\|\boldsymbol{y} - \boldsymbol{y}'\|^2}{\langle \boldsymbol{y}, \boldsymbol{y} - \boldsymbol{y}' \rangle} \to 0$.

8. Algorithms for learning with Fenchel-Young losses

In this section, we present generic primal (§8.1) and dual (§8.2) algorithms for training predictive models with a Fenchel-Young loss L_Ω for arbitrary Ω. In doing so, we obtain unified algorithms for training models with a wealth of existing and new loss functions. We then discuss algorithms for computing regularized prediction functions (§8.3) and proximity operators (§8.4). We summarize these "computational oracles" in Figure 8.

8.1. Primal training

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

$$\min_W F(W) + G(W) \coloneqq \sum_{i=1}^n L_\Omega(\boldsymbol{f}_W(\boldsymbol{x}_i); \boldsymbol{y}_i) + G(W), \tag{29}$$

where G is a regularization function w.r.t. parameters W. Typical examples are Tikhonov regularization, $G(W) = \frac{\lambda}{2}\|W\|_F^2$, and elastic net regularization, $G(W) = \frac{\lambda}{2}\|W\|_F^2 + \lambda\rho\|W\|^2$, for some $\lambda > 0$ and $\rho \geq 0$.

The objective in (29) is very broad and allows to learn, e.g., linear models or neural networks using Fenchel-Young losses. Since L_Ω is convex, if $\boldsymbol{f}_W$ is a linear model and G is convex, then (29) is convex as well. Assuming further that Ω is strongly convex, then L_Ω is smooth and (29) can be solved globally using proximal gradient algorithms (Wright et al., 2009; Beck and Teboulle, 2009; Defazio et al., 2014). From Proposition 2, the gradient of $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ w.r.t. $\boldsymbol{\theta}$ is the residual vector

$$\nabla L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y} \in \mathbb{R}^d.$$

Using the chain rule, the gradient of F w.r.t. W is

$$\nabla F(W) = \sum_{i=1}^n D_{\boldsymbol{f}_W(\boldsymbol{x}_i)}^\top \nabla L_\Omega(\boldsymbol{f}_W(\boldsymbol{x}_i); \boldsymbol{y}_i),$$

where $D_{\boldsymbol{f}_W(\boldsymbol{x}_i)}$ is the Jacobian of $\boldsymbol{f}_W(\boldsymbol{x}_i)$ w.r.t. W, a linear map from the space of W to $\mathbb{R}^d$. For linear models, we set $\boldsymbol{\theta}_i = \boldsymbol{f}_W(\boldsymbol{x}_i) = W\boldsymbol{x}_i$, where $W \in \mathbb{R}^{d\times p}$ and $\boldsymbol{x}_i \in \mathbb{R}^p$, and thus get

$$\nabla F(W) = (\widehat{Y}_\Omega - Y)^\top X,$$

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 $i = 1, \ldots, n$.

In summary, the two main computational ingredients to solve (29) are $\widehat{\boldsymbol{y}}_\Omega$ and the proximity operator of G (when G is non-differentiable). This separation of concerns results in very modular implementations.

Remark on temperature scaling. When $\boldsymbol{f}_W$ is a linear model and G is a homogeneous function, it is easy to check that the temperature scaling parameter $t > 0$ from Proposition 2 and the regularization strength $\lambda > 0$ above are redundant. Hence, we can set $t = 1$ at training time, without loss of generality. However, at test time, using $\widehat{\boldsymbol{y}}_{t\Omega}(\boldsymbol{\theta})$ and tuning t can potentially improve accuracy. For sparse prediction functions, t also gives control over sparsity (the smaller, the sparser).

8.2. Dual training

We now derive dual training of Fenchel-Young losses. Although our treatment follows closely Shalev-Shwartz and Zhang (2016), it is different in that we put the output regularization Ω at the center of all computations, leading to a new perspective.

Dual objective. For linear models $\boldsymbol{f}_W(\boldsymbol{x}) = W\boldsymbol{x}$, where $W \in \mathbb{R}^{d\times p}$, it can sometimes be more computationally efficient to solve the corresponding dual problem, especially in the $n \ll p$ setting. From Fenchel's duality theorem (see for instance Borwein and Lewis (2010, Theorem 3.3.5)), we find that the dual problem of (29) is

$$\max_{\alpha \in \mathbb{R}^{n\times d}} -\sum_{i=1}^n L_\Omega^*(-\boldsymbol{\alpha}_i; \boldsymbol{y}_i) - G^*\left(\alpha^\top X\right),$$

and where L_Ω^* is the convex conjugate of L_Ω in its first argument.

We now rewrite the dual problem using the specific form of L_Ω. Using $L_\Omega^*(-\boldsymbol{\alpha}; \boldsymbol{y}) = \Omega(\boldsymbol{y} - \boldsymbol{\alpha}) - \Omega(\boldsymbol{y})$ and using the change of variable $\boldsymbol{\mu}_i \coloneqq \boldsymbol{y}_i - \boldsymbol{\alpha}_i$, we get

$$\max_{\mu \in \mathbb{R}^{n \times d}} -D(\mu) \text{ s.t. } \boldsymbol{\mu}_i \in \mathsf{dom}(\Omega) \ \forall i \in [n], \tag{30}$$

where we defined

$$D(\mu) \coloneqq \sum_{i=1}^n \Omega(\boldsymbol{\mu}_i) - \Omega(\boldsymbol{y}_i) + G^*(V(\mu)) \quad \text{and} \quad V(\mu) \coloneqq (Y - \mu)^\top X.$$

This expression is informative as we can interpret each $\boldsymbol{\mu}_i$ as regularized predictions, i.e., $\boldsymbol{\mu}_i$ should belong to $\mathsf{dom}(\Omega)$, the same domain as $\widehat{\boldsymbol{y}}_\Omega$. The fact that $G^*(V(\mu))$ is a function of the predictions μ is similar to the value regularization framework of Rifkin and Lippert (2007). A key difference with the regularization Ω, however, is that $G^*(V(\mu))$ depends on the training data X through $V(\mu)$. When $G(W) = \frac{\lambda}{2}\|W\|_F^2 \Leftrightarrow G^*(V) = \frac{1}{2\lambda}\|V\|_F^2$, we obtain $G^*(V(\mu)) = \frac{1}{2\lambda}\text{Trace}((Y - \mu)^\top K (Y - \mu))$, where $K \coloneqq XX^\top$ is the Gram matrix.

Primal-dual relationship. Assuming G is a λ-strongly convex regularizer, given an optimal dual solution $\mu^\star$ to (30), we may retrieve the optimal primal solution $W^\star$ by

$$W^\star = \nabla G^*(V(\mu^\star)).$$

Coordinate ascent. We can solve (30) using block coordinate ascent algorithms. At every iteration, we pick $i \in [n]$ and consider the following sub-problem associated with i:

$$\operatorname*{argmin}_{\boldsymbol{\mu}_i \in \mathsf{dom}(\Omega)} \Omega(\boldsymbol{\mu}_i) + G^*(V(\mu)). \tag{31}$$

Since this problem could be hard to solve, we follow Shalev-Shwartz and Zhang (2016, Option I) and consider instead a quadratic upper-bound. If G is λ-strongly convex, G^* is $\frac{1}{\lambda}$-smooth w.r.t. the dual norm $\|\cdot\|$ and it holds that

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

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

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

where $\sigma_i \coloneqq \frac{\|\boldsymbol{x}_i\|^2}{\lambda}$. Substituting $G^*(V(\mu))$ by the above upper bound into (31) and ignoring constant terms, we get the following approximate sub-problem:

$$\operatorname*{argmin}_{\boldsymbol{\mu}_i \in \mathsf{dom}(\Omega)} \Omega(\boldsymbol{\mu}_i) - \boldsymbol{\mu}_i^\top \boldsymbol{v}_i + \frac{\sigma_i}{2}\|\boldsymbol{\mu}_i\|^2 = \mathsf{prox}_{\frac{1}{\sigma_i}\Omega}\left(\frac{\boldsymbol{v}_i}{\sigma_i}\right), \tag{32}$$

where $\boldsymbol{v}_i \coloneqq \nabla G^*(V(\bar{\mu}))\boldsymbol{x}_i + \sigma_i \bar{\boldsymbol{\mu}}_i$. Note that when G^* is a quadratic function, (32) is an optimal solution of (31).

Examples of parameter regularization G. We now give examples of $G^*(W)$ and $\nabla G^*(W)$ for two commonly used regularization: squared ℓ_2 and elastic-net regularization.

When $G(W) = \frac{\lambda}{2}\|W\|_F^2$, we obtain

$$G^*(V) = \frac{1}{2\lambda}\|V\|_F^2 \quad \text{and} \quad \nabla G^*(V) = \frac{1}{\lambda}V.$$

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

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

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

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

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

where all operations above are performed element-wise.

Summary. To summarize, on each iteration, we pick $i \in [n]$ and perform the update

$$\boldsymbol{\mu}_i \leftarrow \mathsf{prox}_{\frac{1}{\sigma_i}\Omega}(\boldsymbol{v}_i/\sigma_i),$$

where we defined $\boldsymbol{v}_i \coloneqq \nabla G^*(V(\mu))\boldsymbol{x}_i + \sigma_i\boldsymbol{\mu}_i$ and $\sigma_i \coloneqq \frac{\|\boldsymbol{x}_i\|^2}{\lambda}$. This update does not require choosing any learning rate. Interestingly, if $\mathsf{prox}_{\tau\Omega}$ is sparse, then **so are the dual variables $\{\boldsymbol{\mu}_i\}_{i=1}^n$**. Sparsity in the dual variables is particularly useful when kernelizing models, as it makes computing predictions more efficient.

When i is picked uniformly at random, this block coordinate ascent algorithm is known to converge to an optimal solution $W^\star$, at a linear rate if L_Ω is smooth (i.e., if Ω is strongly convex) (Shalev-Shwartz and Zhang, 2016). When $\mathsf{dom}(\Omega)$ is a compact set, another option to solve (30) that does not involve proximity operators is the block Frank-Wolfe algorithm (Lacoste-Julien et al., 2012).

8.3. Regularized prediction functions

The regularized prediction function $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ does not generally enjoy a closed-form expression and one must resort to generic algorithms to compute it. In this section, we first discuss two such algorithms: projected gradient and conditional gradient methods. Then, we present a new more efficient algorithm when $\mathsf{dom}(\Omega) \subseteq \triangle^d$ and Ω is uniformly separable.

Generic algorithms. In their greater generality, regularized prediction functions involve the following optimization problem

$$\min_{\boldsymbol{\mu} \in \mathsf{dom}(\Omega)} f_{\boldsymbol{\theta}}(\boldsymbol{\mu}) \coloneqq \Omega(\boldsymbol{\mu}) - \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle. \tag{33}$$

Assuming Ω is convex and smooth (differentiable and with Lipschitz-continuous gradient), we can solve this problem to arbitrary precision using projected gradient methods. Unfortunately, the projection onto $\mathsf{dom}(\Omega)$, which is needed by these algorithms, is often as challenging to solve as (33) itself. This is especially the case when $\mathsf{dom}(\Omega)$ is $\mathsf{conv}(\mathcal{Y})$, the convex hull of a combinatorial set of structured objects.

Assuming further that $\mathsf{dom}(\Omega)$ is a compact set, an appealing alternative which sidesteps these issues is provided by conditional gradient (a.k.a. Frank-Wolfe) algorithms (Frank and Wolfe, 1956; Dunn and Harshbarger, 1978; Jaggi, 2013). Their main advantage stems from the fact that they access $\mathsf{dom}(\Omega)$ only through the linear maximization oracle $\mathsf{argmax}_{\boldsymbol{y} \in \mathsf{dom}(\Omega)} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$. CG algorithms maintain the current solution as a sparse convex combination of vertices of $\mathsf{dom}(\Omega)$. At each iteration, the linear maximization oracle is used to pick a new vertex to add to the combination. Despite its simplicity, the procedure converges to an optimal solution, albeit at a sub-linear rate (Jaggi, 2013). Linear convergence rates can be obtained using away-step, pairwise and full-corrective variants (Lacoste-Julien and Jaggi, 2015). When Ω is a quadratic function (as in the case of SparseMAP), an approximate correction step can achieve finite convergence efficiently. The resulting algorithm is known as the active set method (Nocedal and Wright, 1999, chapters 16.4 & 16.5), and is a generalization of Wolfe's min-norm point algorithm (Wolfe, 1976). See also Vinyes and Obozinski (2017) for a more detailed discussion on these algorithms and Niculae et al. (2018) for an instantiation, in the specific case of SparseMAP.

Although not explored in this paper, similar algorithms have been developed to minimize a function over conic hulls (Locatello et al., 2017). Such algorithms allow to compute a regularized prediction function that outputs a conic combination of elementary structures, instead of a convex one. It can be seen as the structured counterpart of the regularized prediction function over positive measures (§6).

Reduction to root finding. When Ω is a strictly convex regularization function over $\triangle^d$ and is uniformly separable, i.e., $\Omega(\boldsymbol{p}) = \sum_{i=1}^d g(p_i)$ for some strictly convex function g, we now show that $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ can be computed in linear time.

Proposition 9 *Reduction to root finding*

Let $g \colon [0,1] \to \mathbb{R}_+$ be a strictly convex and differentiable function. Then,

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \mathsf{argmax}_{\boldsymbol{p} \in \triangle^d} \langle \boldsymbol{p}, \boldsymbol{\theta} \rangle - \sum_{j=1}^d g(p_j) = \boldsymbol{p}(\tau^\star)$$

where

$$\boldsymbol{p}(\tau) \coloneqq (g')^{-1}(\max\{\boldsymbol{\theta} - \tau, g'(0)\})$$

and where $\tau^\star$ is a root of $\phi(\tau) \coloneqq \langle \boldsymbol{p}(\tau), \mathbf{1} \rangle - 1$.

Moreover, $\tau^\star$ belongs to the tight search interval $[\tau_{\min}, \tau_{\max}]$, where

$$\tau_{\min} \coloneqq \max(\boldsymbol{\theta}) - g'(1) \quad \textit{and} \quad \tau_{\max} \coloneqq \max(\boldsymbol{\theta}) - g'\left(1/d\right).$$

The equation of $\boldsymbol{p}(\tau)$ can be seen as a generalization of (7). 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. §B.8).

For example, when Ω is the negative α-Tsallis entropy $-\mathsf{H}^{\mathrm{T}}_\alpha$, which we saw can be written in separable form in (18), we obtain

$$g(t) = \frac{t^\alpha - t}{\alpha(\alpha - 1)}, \quad g'(t) = \frac{t^{\alpha-1} - 1}{\alpha - 1} \quad \text{and} \quad (g')^{-1}(s) = \big(1 + (\alpha - 1)s\big)^{\frac{1}{\alpha - 1}},$$

yielding

$$\boldsymbol{p}(\tau) = \big(1 + (\alpha - 1)\max(\boldsymbol{\theta} - \tau, {-1}/{\alpha - 1})\big)^{\frac{1}{\alpha-1}}.$$

From the root $\tau^\star$ of $\phi(\tau) = \langle \boldsymbol{p}(\tau), \mathbf{1} \rangle - 1$, we obtain $\widehat{\boldsymbol{y}}_{-\mathsf{H}^{\mathrm{T}}_\alpha}(\boldsymbol{\theta}) = \boldsymbol{p}(\tau^\star)$.

8.4. Proximity operators

Computing the proximity operator $\mathsf{prox}_{\tau\Omega}(\boldsymbol{\eta})$, defined in (2), usually involves a more challenging optimization problem than $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$. For instance, when Ω is Shannon's negative entropy over $\triangle^d$, $\widehat{\boldsymbol{y}}_\Omega$ enjoys a closed-form solution (the softmax) but not prox_Ω. However, we can always compute $\mathsf{prox}_{\tau\Omega}(\boldsymbol{\eta})$ by first-order gradient methods given access to $\widehat{\boldsymbol{y}}_\Omega$. Indeed, using the Moreau decomposition, we have

$$\mathsf{prox}_{\tau\Omega}(\boldsymbol{\eta}) = \boldsymbol{\eta} - \mathsf{prox}_{\frac{\Omega^*}{\tau}}(\boldsymbol{\eta}/\tau).$$

Since $\mathsf{dom}(\Omega^*) = \mathbb{R}^d$, computing $\mathsf{prox}_{\frac{\Omega^*}{\tau}}(\boldsymbol{\eta}/\tau)$ only involves an unconstrained optimization problem, $\mathsf{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\eta}/\tau\|^2 + \frac{1}{\tau}\Omega^*(\boldsymbol{\theta})$. Since $\nabla\Omega^* = \widehat{\boldsymbol{y}}_\Omega$, that optimization problem can easily be solved by any first-order gradient method given access to $\widehat{\boldsymbol{y}}_\Omega$. For specific choices of Ω, $\mathsf{prox}_{\tau\Omega}(\boldsymbol{\eta})$ can be computed directly more efficiently, as we now discuss.

Closed-form expressions. We now give examples of commonly-used loss functions for which $\mathsf{prox}_{\tau\Omega}$ enjoys a closed-form expression.

For the squared loss, we choose $\Omega(\boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{\mu}\|^2$. Hence:

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

For the perceptron loss (Rosenblatt, 1958; Collins, 2002), we choose $\Omega = I_{\triangle^d}$. Hence:

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

For the sparsemax loss (Martins and Astudillo, 2016), we choose $\Omega = \frac{1}{2}\|\cdot\|^2 + I_{\triangle^d}$. Hence:

$$\mathsf{prox}_{\tau\Omega}(\boldsymbol{\eta}) = \underset{\boldsymbol{p}\in\triangle^d}{\mathsf{argmin}} \, \frac{1}{2}\|\boldsymbol{p}-\boldsymbol{\eta}\|^2 + \frac{\tau}{2}\|\boldsymbol{p}\|^2 = \underset{\boldsymbol{p}\in\triangle^d}{\mathsf{argmin}} \, \frac{1}{2}\left\|\boldsymbol{p} - \frac{\boldsymbol{\eta}}{\tau+1}\right\|^2.$$

For the cost-sensitive multiclass hinge loss, we choose $\Omega = I_{\triangle^d} - \langle\cdot, \boldsymbol{c_y}\rangle$. Hence:

$$\mathsf{prox}_{\tau\Omega}(\boldsymbol{\eta}) = \underset{\boldsymbol{p}\in\triangle^d}{\mathsf{argmin}} \, \frac{1}{2}\|\boldsymbol{p}-\boldsymbol{\eta}\|^2 - \tau\langle\boldsymbol{p}, \boldsymbol{c_y}\rangle = \underset{\boldsymbol{p}\in\triangle^d}{\mathsf{argmin}} \, \frac{1}{2}\|\boldsymbol{p} - (\boldsymbol{\eta} + \tau\boldsymbol{c_y})\|^2.$$

Choosing $\boldsymbol{c_y} = \mathbf{1} - \boldsymbol{y}$, where $\boldsymbol{y} \in \{\boldsymbol{e}_i\}_{i=1}^d$ is the ground-truth label, gives the proximity operator for the usual multiclass hinge loss (Crammer and Singer, 2001).

For the Shannon entropy and 1.5-Tsallis entropy, we show that $\mathsf{prox}_{-\mathsf{H}}$ reduces to root finding in Appendix A.1.

9. Experiments

In this section, we demonstrate one of the key features of Fenchel-Young losses: their ability to induce **sparse** probability distributions. We focus on two tasks: label proportion estimation (§9.1) and dependency parsing (§9.2).

9.1. Label proportion estimation experiments

As we saw in §4, 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 9. In this section, we demonstrate their effectiveness on the task of label proportion estimation and compare different solvers for computing the regularized prediction function $\widehat{\boldsymbol{y}}_{-\mathsf{H}^{\mathrm{T}}_\alpha}$.

Experimental setup. 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 pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, 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 optimize the loss using L-BFGS (Liu and Nocedal, 1989). 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 $i = 1, \ldots, n$. At test time, we predict label proportions by $\boldsymbol{p} = \widehat{\boldsymbol{y}}_{-\mathsf{H}^{\mathrm{T}}_\alpha}(W\boldsymbol{x})$.

Table 3: Dataset statistics

Dataset	Type	Train	Dev	Test	Features	Classes	Avg. labels
Birds	Audio	134	45	172	260	19	2
Cal500	Music	376	126	101	68	174	25
Emotions	Music	293	98	202	72	6	2
Mediamill	Video	22,353	7,451	12,373	120	101	5
Scene	Images	908	303	1,196	294	6	1
SIAM TMC	Text	16,139	5,380	7,077	30,438	22	2
Yeast	Micro-array	1,125	375	917	103	14	4

Real data experiments. We ran experiments on 7 standard multi-label benchmark datasets — see Table 3 for dataset characteristics[1]. 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, $\mathrm{JS}(\boldsymbol{p}, \boldsymbol{y}) := {}^1\!/_2 \, \mathrm{KL}(\boldsymbol{p} \| \frac{\boldsymbol{p}+\boldsymbol{y}}{2}) + {}^1\!/_2 \, \mathrm{KL}(\boldsymbol{y} \| \frac{\boldsymbol{p}+\boldsymbol{y}}{2})$, and the mean squared error ${}^1\!/_2 \, \|\boldsymbol{p} - \boldsymbol{y}\|^2$ in Table 4. 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.

Synthetic data experiments. We follow Martins and Astudillo (2016) and generate a document $\boldsymbol{x} \in \mathbb{R}^p$ from a mixture of multinomials and label proportions $\boldsymbol{y} \in \triangle^d$ from a multinomial. The number of words in $\boldsymbol{x}$ and labels in $\boldsymbol{y}$ is sampled from a Poisson distribution — see Martins and Astudillo (2016) for a precise description of the generative process. We use 1200 samples as training set, 200 samples as validation set and 1000 samples as test set. We tune $\lambda \in \{10^{-6}, 10^{-5}, \dots, 10^0\}$ and $\alpha \in \{1.0, 1.1, \dots, 2.0\}$ against the validation set. We report the Jensen-Shannon divergence in Figure 10. Results using the mean squared error (MSE) were entirely similar. When the number of classes is 10, Tsallis and sparsemax losses perform almost exactly the same, both outperforming softmax. When the number of classes is 50, Tsallis losses outperform both sparsemax and softmax.

Solver comparison. We also compared bisection (binary search) and Brent's method for solving (5) by root finding (Proposition 9). 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 (5) naively. We measure the time needed to reach a solution $\boldsymbol{p}$ with $\|\boldsymbol{p} - \boldsymbol{p}^\star\|_2 < 10^{-5}$, over 200 samples $\boldsymbol{\theta} \in \mathbb{R}^d \sim \mathcal{N}(\mathbf{0}, \sigma \boldsymbol{I})$ with $\log \sigma \sim \mathcal{U}(-4, 4)$. Median and 99% CI times reported in Figure 9 reveal that root finding scales better, with Brent's method outperforming FISTA by one to two orders of magnitude.

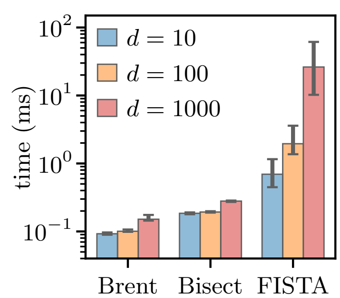

Figure 9: Solver comparison.

1. The datasets can be downloaded from `http://mulan.sourceforge.net/datasets-mlc.html` and `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Table 4: **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.

	$\alpha = 1$ (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**

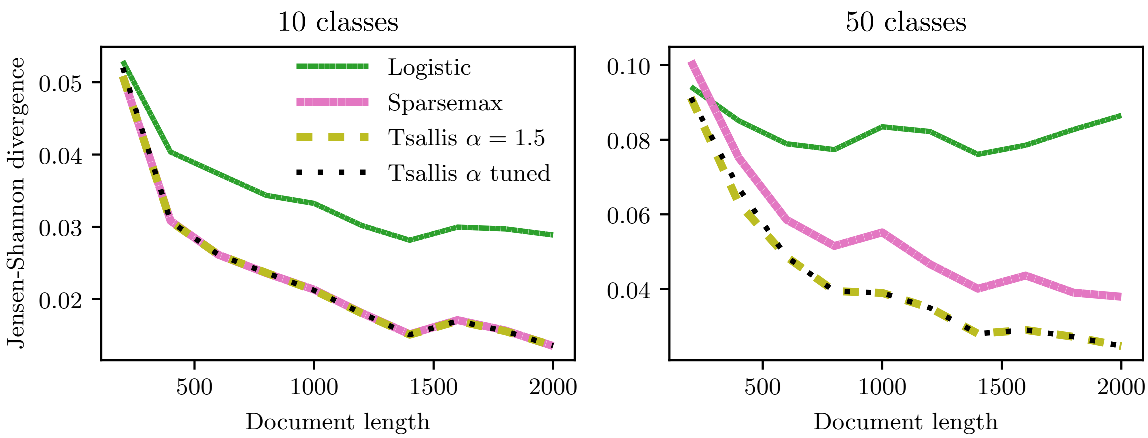

Figure 10: Jensen-Shannon divergence between predicted and true label proportions, when varying document length, of various losses generated by a Tsallis entropy.

9.2. Dependency parsing experiments

We next compare Fenchel-Young losses for structured output prediction, as formulated in §7, for non-projective *dependency parsing*. The application consists of predicting the directed tree of grammatical dependencies between words in a sentence (Jurafsky and Martin, 2018, Chapter 14). We tackle it by learning structured models over the arborescence polytope, using the structured SVM, CRF, and SparseMAP losses (Niculae et al., 2018), which we saw to all be instances of Fenchel-Young loss. Training data, as in the CoNLL 2017 shared task (Zeman et al., 2017), comes from the Universal Dependency project (Nivre et al., 2016). To isolate the effect of the loss, we use the provided gold tokenization and part-of-speech tags. We follow closely the bidirectional LSTM arc-factored parser of Kiperwasser and Goldberg (2016), using the same model configuration; the only exception is not using externally pretrained embeddings. Parameters are trained using Adam (Kingma and Ba,

Table 5: Unlabeled attachment accuracy scores for dependency parsing, using a bi-LSTM model (Kiperwasser and Goldberg, 2016). For context, we include the scores of the CoNLL 2017 UDPipe baseline (Straka and Straková, 2017).

Loss	en	zh	vi	ro	ja
Structured SVM	87.02	81.94	69.42	**87.58**	**96.24**
CRF	86.74	83.18	69.10	87.13	96.09
SparseMAP	86.90	**84.03**	**69.71**	87.35	96.04
UDPipe baseline	**87.68**	82.14	69.63	87.36	95.94

2015), tuning the learning rate on the grid $\{.5, 1, 2, 4, 8\} \times 10^{-3}$, expanded by a factor of 2 if the best model is at either end.

We experiment with 5 languages, diverse both in terms of family and in terms of the amount of training data (ranging from 1,400 sentences for Vietnamese to 12,525 for English). Test set results (Table 5) indicate that SparseMAP outperforms the CRF loss and is competitive with the structured SVM loss, outperforming it substantially on Chinese. The Fenchel-Young perspective sheds further light onto the empirical results, as the SparseMAP loss and the structured SVM (unlike the CRF) both enjoy the margin property with margin 1, while both SparseMAP and the CRF loss (unlike structured SVM) are differentiable.

10. Related work

10.1. Loss functions

Proper losses. Proper losses, a.k.a. proper scoring rules, are a well-studied object in statistics (Grünwald and Dawid (2004); Gneiting and Raftery (2007); references therein) and more recently in machine learning (Reid and Williamson, 2010; Williamson et al., 2016). A loss $\ell(\boldsymbol{p}; \boldsymbol{y})$, between a ground truth $\boldsymbol{y} \in \triangle^d$ and a probability forecast $\boldsymbol{p} \in \triangle^d$, where d is the number of classes, is said to be proper if it is minimized when estimating the true $\boldsymbol{y}$. Formally, a proper loss ℓ satisfies

$$\ell(\boldsymbol{y}; \boldsymbol{y}) \le \ell(\boldsymbol{p}; \boldsymbol{y}) \quad \forall \boldsymbol{p}, \boldsymbol{y} \in \triangle^d.$$

It is strictly proper if the inequality is strict when $\boldsymbol{p} \neq \boldsymbol{y}$, implying that it is uniquely minimized by predicting the correct probability. Strictly proper losses induce **Fisher consistent** estimators of probabilities (Williamson et al., 2016). A key result, which dates back to Savage (1971) (see also Gneiting and Raftery (2007)), is that given a regularization function $\Omega \colon \triangle^d \to \mathbb{R}$, the function $\ell_\Omega \colon \triangle^d \times \triangle^d \to \mathbb{R}_+$ defined by

$$\ell_\Omega(\boldsymbol{p}; \boldsymbol{y}) \coloneqq \langle \nabla \Omega(\boldsymbol{p}), \boldsymbol{y} - \boldsymbol{p} \rangle - \Omega(\boldsymbol{p})$$

is proper. It is easy to see that $\ell_\Omega(\boldsymbol{p}; \boldsymbol{y}) = B_\Omega(\boldsymbol{y} \| \boldsymbol{p}) - \Omega(\boldsymbol{y})$, which recovers the well-known Bregman divergence representation of proper losses. For example, using the Gini index

$\mathsf{H}(\boldsymbol{p}) = 1 - \|\boldsymbol{p}\|^2$ generates the **Brier score** (Brier, 1950)

$$\ell_{-\mathsf{H}}(\boldsymbol{p}; \boldsymbol{e}_k) = \sum_{i=1}^{d} ([[k = i]] - p_i)^2,$$

showing that the sparsemax loss and the Brier score share the same generating function.

More generally, while a proper loss ℓ_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, ℓ_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 ℓ_Ω, while **unconstrained** for L_Ω.

In practice, proper losses are often composed with an **invertible** link function $\boldsymbol{\psi}^{-1} \colon \mathbb{R}^d \to \triangle^d$. This form of a loss, $\ell_\Omega(\boldsymbol{\psi}^{-1}(\boldsymbol{\theta}); \boldsymbol{y})$, is sometimes called composite (Buja et al., 2005; Reid and Williamson, 2010; Williamson et al., 2016). However, the composition of $\ell_\Omega(\cdot; \boldsymbol{y})$ and $\boldsymbol{\psi}^{-1}(\boldsymbol{\theta})$ is not necessarily convex in $\boldsymbol{\theta}$. The **canonical link function** (Buja et al., 2005) of ℓ_Ω is a link function that ensures the convexity of $\ell_\Omega(\boldsymbol{\psi}^{-1}(\boldsymbol{\theta}); \boldsymbol{y})$ in $\boldsymbol{\theta}$. It also plays a key role in generalized linear models (Nelder and Baker, 1972; McCullagh and Nelder, 1989). Following Proposition 3, when Ω is Legendre type, we obtain

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

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. However, the decoupling between loss and link as promoted by proper composite losses also has some merits (Reid and Williamson, 2010). For instance, using a non-canonical link is useful to express the exponential loss of Adaboost (Nowak-Vila et al., 2019).

Other related losses. Nock and Nielsen (2009) proposed a binary classification loss construction. Technically, their loss is based on the Legendre transformation (a subset of Fenchel conjugate functions), precluding non-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 (14), 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. Finally, several works have explored connections between another divergence, the f-divergence (Csiszár, 1975), and surrogate loss functions (Nguyen et al., 2009; Reid and Williamson, 2011; Garcia-Garcia and Williamson, 2012; Duchi et al., 2018).

Consistency. In this paper, we have not addressed the question of the consistency of Fenchel-Young losses when used as a surrogate for a (possibly non-convex) loss. Although consistency has been widely studied in the multiclass setting (Zhang, 2004; Bartlett et al., 2006; Tewari and Bartlett, 2007; Mroueh et al., 2012) and in other specific settings (Duchi et al., 2010; Ravikumar et al., 2011), it is only recently that it was studied for general losses in a fully general structured prediction setting. Since the publication of this paper, sufficient conditions for consistency have been established for composite losses (Nowak-Vila et al., 2019) and for projection-based losses (Blondel, 2019), both a subset of Fenchel-Young losses. As shown in these works, the strong convexity of Ω plays a crucial role.

10.2. Fenchel duality in machine learning

In this paper, we make extensive use of Fenchel duality to derive loss functions. Fenchel duality has also played a key role in several machine learning studies before. It was used to provide a unifying perspective on convex empirical risk minimization and representer theorems (Rifkin and Lippert, 2007). It was also used for deriving regret bounds (Shalev-Shwartz and Singer, 2007; Shalev-Shwartz and Kakade, 2009), risk bounds (Kakade et al., 2009), and unified analyses of boosting (Shen and Li, 2010; Shalev-Shwartz and Singer, 2010; Telgarsky, 2012). However, none of these works propose loss functions, as we do. More closely related to our work are smoothing techniques (Nesterov, 2005; Beck and Teboulle, 2012), which have been used extensively to create smoothed versions of existing losses (Song et al., 2014; Shalev-Shwartz and Zhang, 2016). However, these techniques were applied on a per-loss basis and were not connected to an induced probability distribution. In contrast, we propose a generic loss construction, with clear links between smoothing / regularization and the probability distribution produced by regularized prediction functions.

10.3. Approximate inference with conditional gradient algorithms

In this paper, we suggest conditional gradient (CG) algorithms as a powerful tool for computing regularized prediction functions and Fenchel-Young losses over potentially complex output domains. CG algorithms have been used before to compute an approximate solution to intractable marginal inference problems (Belanger et al., 2013; Krishnan et al., 2015) or to sample from an intractable distribution so as to approximate its mean (Bach et al., 2012; Lacoste-Julien et al., 2015). When combined with reproducing kernel Hilbert spaces (RKHS), these ideas are closely related maximum mean discrepancy (Gretton et al., 2012).

11. Conclusion

We showed that the notion of output regularization and Fenchel duality provide simple core principles, unifying many existing loss functions, and allowing to create useful new ones easily, on a large spectrum of tasks. We characterized a tight connection between sparse distributions and losses with a separation margin, and showed that these losses are precisely the ones that cannot be written in proper composite loss form. We established the computational tools to efficiently learn with Fenchel-Young losses, whether in unstructured or structured settings. We expect that this groundwork will enable the creation of many more novel losses in the future, by exploring other convex polytopes and regularizers.

Acknowledgments

MB thanks Arthur Mensch, Gabriel Peyré and Marco Cuturi for numerous fruitful discussions and Tim Vieira for introducing him to generalized exponential families. This work was 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 CMUPERI/TIC/0046/2014 (GoLocal).

Appendix A. Additional materials

A.1. Computing proximity operators by root finding

In this section, we first show how the proximity operator of uniformly separable generalized entropies reduces to unidimensional root finding. We then illustrate the case of two important entropies. The proximity operator is defined as

$$\begin{aligned}
\mathsf{prox}_{-\frac{1}{\sigma}\mathsf{H}}(\boldsymbol{x}) &= \operatorname*{argmin}_{\boldsymbol{p}\in\triangle^d} \frac{\sigma}{2}\|\boldsymbol{p}-\boldsymbol{x}\|^2 - \mathsf{H}(\boldsymbol{p}) \\
&= \operatorname*{argmax}_{\boldsymbol{p}\in\triangle^d} \langle \boldsymbol{p}, \boldsymbol{x}\rangle + \frac{1}{\sigma}\mathsf{H}(\boldsymbol{p}) - \frac{1}{2}\|\boldsymbol{p}\|^2 \\
&= \operatorname*{argmax}_{\boldsymbol{p}\in\triangle^d} \langle \boldsymbol{p}, \boldsymbol{x}\rangle - \sum_{i=1}^d g(p_i),
\end{aligned}$$

where

$$g(t) \coloneqq \frac{1}{2}t^2 - \frac{1}{\sigma}h(t).$$

Note that $g(t)$ is strictly convex even when h is only concave. We may thus apply Proposition 9 to compute $\mathsf{prox}_{-\frac{1}{\sigma}\mathsf{H}}$. The ingredients necessary for using Algorithm 1 are $g'(t) = t - \frac{1}{\sigma}h'(t)$, and its inverse $(g')^{-1}(x)$. We next derive closed-form expressions for the inverse for two important entropies.

Shannon entropy. We have

$$g(t) \coloneqq \frac{t^2}{2} + \frac{t\log t}{\sigma} \quad \text{and} \quad g'(t) = t + \frac{1+\log t}{\sigma}.$$

To compute the inverse, $(g')^{-1}(x) = t$, we solve for t satisfying

$$\frac{1+\log t}{\sigma} + t = x.$$

Making the change of variable $s \coloneqq \sigma t$ gives

$$s + \log s = \sigma x - 1 + \log \sigma \iff s = \omega(\sigma x - 1 + \log \sigma).$$

Therefore, $(g')^{-1}(x) = \frac{1}{\sigma}\omega(\sigma x - 1 + \log \sigma)$, where ω denotes the Wright omega function.

Tsallis entropy with $\alpha = 1.5$. Up to a constant term, we have

$$g(t) = \frac{t^2}{2} + \frac{4}{3\sigma} t^{3/2} \quad \text{and} \quad g'(t) = t + \frac{2}{\sigma}\sqrt{t}.$$

To obtain the inverse, $(g')^{-1}(x) = t$ we seek t satisfying

$$\frac{2}{\sigma}\sqrt{t} + t = x \iff \sqrt{t} = \frac{\sigma}{2}(x - t).$$

If $t > x$ there are no solutions, otherwise we may square both sides, yielding

$$\frac{\sigma^2 t^2}{4} - t\left(\frac{\sigma^2 x}{2} + 1\right) + \frac{\sigma^2 x^2}{4} = 0.$$

The discriminant is

$$\Delta = 1 + \sigma^2 x > 0,$$

resulting in two solutions

$$t_{\pm} = x + \frac{2}{\sigma^2}\left(1 \pm \sqrt{1 + \sigma^2 x}\right).$$

However, $t_+ > x$, therefore

$$(g')^{-1}(x) = x + \frac{2}{\sigma^2}\left(1 - \sqrt{1 + \sigma^2 x}\right).$$

A.2. Loss "Fenchel-Youngization"

Not all loss functions proposed in the literature can be written in Fenchel-Young loss form. In this section, we present a natural method to approximate (in a sense we will clarify) any loss $\ell \colon \mathbb{R}^d \times \{\boldsymbol{e}_i\}_{i=1}^d \to \mathbb{R}_+$ with a Fenchel-Young loss. This has two main advantages. First, the resulting loss is convex even when ℓ is not. Second, the associated regularized prediction function can be used for probabilistic prediction even if ℓ is not probabilistic.

From loss to entropy. Equations (16) and (17), which relate entropy and conditional Bayes risk, suggest a reverse construction from entropy to loss function. This direction has been explored in previous works (Grünwald and Dawid, 2004; Duchi et al., 2018). The entropy H_ℓ generated from ℓ is defined as follows:

$$\mathsf{H}_\ell(\boldsymbol{p}) \coloneqq \inf_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{E}_{\boldsymbol{p}}[\ell(\boldsymbol{\theta}; Y)] = \inf_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^d p_i \ell(\boldsymbol{\theta}; \boldsymbol{e}_i). \tag{34}$$

This is the infimum of a linear and thus concave function of $\boldsymbol{p}$. Hence, by Danskin's theorem (Danskin, 1966; Bertsekas, 1999, Proposition B.25), $H_\ell(\boldsymbol{p})$ is concave (note that this is true

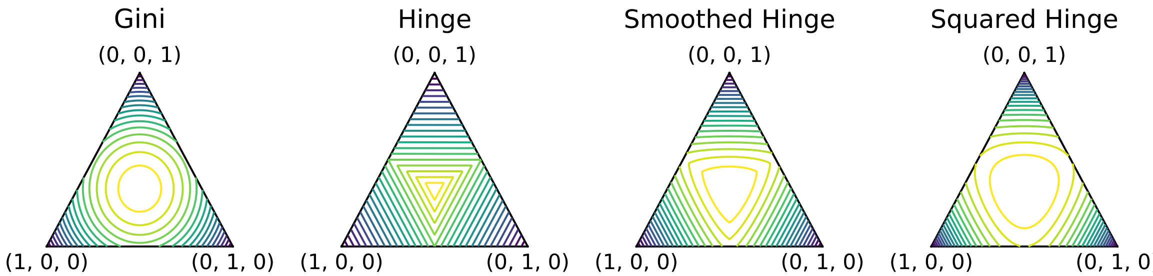

Figure 11: Contour comparison of Gini index $\mathsf{H}_2^{\mathrm{T}}$ and pairwise hinge entropies H_ℓ.

even if ℓ is not convex). As an example, Duchi et al. (2018, Example 1) show that the entropy associated with the zero-one loss is $\mathsf{H}_\ell(\boldsymbol{p}) = 1 - \max_j p_j$. As we discussed in §4.3, this entropy is known as the Berger-Parker dominance index (Berger and Parker, 1970), and is a special case of norm entropy (20).

The following new proposition shows how to compute the entropy of pairwise hinge losses.

Proposition 10 *Entropy generated by pairwise hinge losses*

Let $\ell(\boldsymbol{\theta}; \boldsymbol{e}_k) \coloneqq \sum_{j \neq k} \phi(\theta_j - \theta_k)$ *and* $\tau(\boldsymbol{p}) \coloneqq \min_j 1 - p_j$. *Then, for all* $\boldsymbol{p} \in \triangle^d$:

1. *If* $\phi(t)$ *is the hinge function* $[1+t]_+$ *then* $\mathsf{H}_\ell(\boldsymbol{p}) = \tau(\boldsymbol{p}) d$.
2. *If* $\phi(t)$ *is the smoothed hinge function* (48), *then* $\mathsf{H}_\ell(\boldsymbol{p}) = \frac{-\tau(\boldsymbol{p})^2}{2} \sum_{j=1}^d \frac{1}{1-p_j} + \tau(\boldsymbol{p}) d$.
3. *If* $\phi(t)$ *is the squared hinge function* $\frac{1}{2}[1+t]_+^2$, *then* $\mathsf{H}_\ell(\boldsymbol{p}) = \frac{\frac{1}{2} d^2}{\sum_{j=1}^d 1/(1-p_j)}$.

See Appendix B.9 for a proof. The first one recovers (Duchi et al., 2018, Example 5) with a simpler proof. The last two are new. These entropies are illustrated in Figure 11. For the last two choices, H_ℓ is strictly concave over $\triangle^d$ and the associated prediction function $\widehat{\boldsymbol{y}}_{-\mathsf{H}_\ell}(\boldsymbol{\theta})$ is typically sparse, although we are not aware of a closed-form expression.

And back to Fenchel-Young loss. A natural idea is then to construct a Fenchel-Young from H_ℓ. The resulting loss is convex by construction. Let us assume for simplicity that ℓ achieves its minimum at 0, implying $\mathsf{H}_\ell(\boldsymbol{e}_k) = 0$. Combining (14) and (34), we obtain

$$\begin{aligned}
L_{-\mathsf{H}_\ell}(\boldsymbol{\theta}; \boldsymbol{e}_k) &= (-\mathsf{H}_\ell)^*(\boldsymbol{\theta}) - \theta_k \\
&= \max_{\boldsymbol{p} \in \triangle^d} \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle + \mathsf{H}_\ell(\boldsymbol{p}) - \theta_k \\
&= \max_{\boldsymbol{p} \in \triangle^d} \min_{\boldsymbol{\theta}' \in \mathbb{R}^d} \sum_{i=1}^d p_i (\ell(\boldsymbol{\theta}'; \boldsymbol{e}_i) + \theta_i) - \theta_k.
\end{aligned}$$

By weak duality, we have for every (possibly non-convex) ℓ

$$
\begin{aligned}
L_{-\mathsf{H}_\ell}(\boldsymbol{\theta}; \boldsymbol{e}_k) &\le \min_{\boldsymbol{\theta}' \in \mathbb{R}^d} \max_{\boldsymbol{p} \in \triangle^d} \sum_{i=1}^d p_i(\ell(\boldsymbol{\theta}'; \boldsymbol{e}_i) + \theta_i) - \theta_k \\
&= \min_{\boldsymbol{\theta}' \in \mathbb{R}^d} \max_{i \in [d]} \ell(\boldsymbol{\theta}'; \boldsymbol{e}_i) + \theta_i - \theta_k.
\end{aligned}
$$

We can see $(\boldsymbol{\theta}; \boldsymbol{e}_k) \mapsto \max_{i \in [d]} \ell(\boldsymbol{\theta}'; \boldsymbol{e}_i) + \theta_i - \theta_k$ as a family of cost-augmented hinge losses (Tsochantaridis et al., 2005) parametrized by $\boldsymbol{\theta}'$. Hence, $L_{-\mathsf{H}_\ell}(\boldsymbol{\theta}; \boldsymbol{e}_k)$ is upper-bounded by the tightest loss in this family.

If ℓ is convex, strong duality holds and we obtain equality

$$
L_{-\mathsf{H}_\ell}(\boldsymbol{\theta}; \boldsymbol{e}_k) = \min_{\boldsymbol{\theta}' \in \mathbb{R}^d} \max_{i \in [d]} \ell(\boldsymbol{\theta}'; \boldsymbol{e}_i) + \theta_i - \theta_k.
$$

Margin. We can easily upper-bound the margin of $L_{-\mathsf{H}_\ell}$. Indeed, assuming $\ell(\boldsymbol{\theta}; \boldsymbol{e}_i)$ is upper-bounded by m for all $\boldsymbol{\theta}$ and $i \in [d]$, we get

$$
0 \le L_{-\mathsf{H}_\ell}(\boldsymbol{\theta}; \boldsymbol{e}_k) \le \max_{i \in [d]} \ell(\boldsymbol{\theta}; \boldsymbol{e}_i) + \theta_i - \theta_k \le \max_{i \in [d]} m + \theta_i - \theta_k.
$$

Hence, from Definition 4, we have $\mathrm{margin}(L_{-\mathsf{H}_\ell}) \le m$. Note that if $\mathrm{margin}(L_{-\mathsf{H}_\ell}) \le \mathrm{margin}(\ell)$, then $L_{-\mathsf{H}_\ell}(\boldsymbol{\theta}; \boldsymbol{e}_k) = 0 \Rightarrow \ell(\boldsymbol{\theta}; \boldsymbol{e}_k) = 0$.

Appendix B. Proofs

B.1. Proof of Proposition 1

Effect of a permutation. Let $\Omega(\boldsymbol{\mu})$ be symmetric. We first prove that Ω^* is symmetric as well. Indeed, we have

$$
\Omega^*(\boldsymbol{P\theta}) = \sup_{\boldsymbol{\mu} \in \mathsf{dom}(\Omega)} (\boldsymbol{P\theta})^\top \boldsymbol{\mu} - \Omega(\boldsymbol{\mu}) = \sup_{\boldsymbol{\mu} \in \mathsf{dom}(\Omega)} \boldsymbol{\theta}^\top \boldsymbol{P}^\top \boldsymbol{\mu} - \Omega(\boldsymbol{P}^\top \boldsymbol{\mu}) = \Omega^*(\boldsymbol{\theta}).
$$

The last equality was obtained by a change of variable $\boldsymbol{\mu}' = \boldsymbol{P}^\top \boldsymbol{\mu}$, from which $\boldsymbol{\mu}$ is recovered as $\boldsymbol{\mu} = \boldsymbol{P\mu}'$, which proves $\nabla\Omega^*(\boldsymbol{P\mu}) = \boldsymbol{P}\nabla\Omega^*(\boldsymbol{\mu})$.

Order preservation. Since Ω^* is convex, the gradient operator $\nabla\Omega^*$ is monotone, i.e.,

$$
(\boldsymbol{\theta}' - \boldsymbol{\theta})^\top (\boldsymbol{\mu}' - \boldsymbol{\mu}) \ge 0
$$

for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, $\boldsymbol{\mu} = \nabla\Omega^*(\boldsymbol{\theta})$ and $\boldsymbol{\mu}' = \nabla\Omega^*(\boldsymbol{\theta}')$. Let $\boldsymbol{\theta}'$ be obtained from $\boldsymbol{\theta}$ by swapping two coordinates, i.e., $\theta'_j = \theta_i$, $\theta'_i = \theta_j$, and $\theta'_k = \theta_k$ for any $k \notin \{i, j\}$. Then, since Ω is symmetric, we obtain:

$$
2(\theta_j - \theta_i)(\mu_j - \mu_i) \ge 0,
$$

which implies $\theta_i > \theta_j \Rightarrow \mu_i \ge \mu_j$ and $\mu_i > \mu_j \Rightarrow \theta_i \ge \theta_j$. To fully prove the claim, we need to show that the last inequality is strict: to do this, we simply invoke $\nabla\Omega^*(\boldsymbol{P\mu}) = \boldsymbol{P}\nabla\Omega^*(\boldsymbol{\mu})$ with a matrix $\boldsymbol{P}$ that permutes i and j, from which we must have $\theta_i = \theta_j \Rightarrow \mu_i = \mu_j$.

Approximation error. Let $\boldsymbol{y}^\star := \mathsf{argmax}_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{\theta}, \boldsymbol{y}\rangle$ and $\boldsymbol{\mu}^\star := \mathsf{argmax}_{\boldsymbol{\mu}\in\mathsf{dom}(\Omega)} \langle \boldsymbol{\theta}, \boldsymbol{\mu}\rangle - \Omega(\boldsymbol{\mu})$. If f is a differentiable γ-strongly convex function with unique minimizer $\boldsymbol{\mu}^\star$, then

$$\frac{\gamma}{2}\|\boldsymbol{y} - \boldsymbol{\mu}^\star\|_2^2 \leq f(\boldsymbol{y}) - f(\boldsymbol{\mu}^\star) \quad \forall \boldsymbol{y} \in \mathsf{dom}(f).$$

We assume that $\mathcal{Y} \subseteq \mathsf{dom}(\Omega)$ so it suffices to upper-bound $f(\boldsymbol{y}^\star) - f(\boldsymbol{\mu}^\star)$, where $f(\boldsymbol{\mu}) := \Omega(\boldsymbol{\mu}) - \langle \boldsymbol{\theta}, \boldsymbol{\mu}\rangle$. Since $L \leq \Omega(\boldsymbol{\mu}) \leq U$ for all $\boldsymbol{\mu} \in \mathsf{dom}(\Omega)$, we have $\langle \boldsymbol{\theta}, \boldsymbol{\mu}^\star\rangle \geq -f(\boldsymbol{\mu}^\star) + L$ and $-f(\boldsymbol{y}^\star) + U \geq \langle \boldsymbol{\theta}, \boldsymbol{y}^\star\rangle$. Together with $\langle \boldsymbol{\theta}, \boldsymbol{y}^\star\rangle \geq \langle \boldsymbol{\theta}, \boldsymbol{\mu}^\star\rangle$, this implies $f(\boldsymbol{y}^\star) - f(\boldsymbol{\mu}^\star) \leq U - L$. Therefore, $\frac{1}{2}\|\boldsymbol{y}^\star - \boldsymbol{\mu}^\star\|^2 \leq \frac{U-L}{\gamma}$.

Temperature scaling & constant invariance. These immediately follow from properties of the argmax operator.

B.2. Proof of Proposition 3

We set $\Omega := \Psi + I_{\mathcal{C}}$.

Bregman projections. If Ψ is Legendre type, then $\nabla\Psi(\nabla\Psi^*(\boldsymbol{\theta})) = \boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \mathsf{int}(\mathsf{dom}(\Psi^*))$. Using this and our assumption that $\mathsf{dom}(\Psi^*) = \mathbb{R}^d$, we get for all $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$B_\Psi(\boldsymbol{p}\|\nabla\Psi^*(\boldsymbol{\theta})) = \Psi(\boldsymbol{p}) - \langle \boldsymbol{\theta}, \boldsymbol{p}\rangle + \langle \boldsymbol{\theta}, \nabla\Psi^*(\boldsymbol{\theta})\rangle - \Psi(\nabla\Psi^*(\boldsymbol{\theta})). \tag{36}$$

The last two terms are independent of $\boldsymbol{p}$ and therefore

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \underset{\boldsymbol{p}\in\mathcal{C}}{\mathsf{argmax}}\, \langle \boldsymbol{\theta}, \boldsymbol{p}\rangle - \Psi(\boldsymbol{p}) = \underset{\boldsymbol{p}\in\mathcal{C}}{\mathsf{argmin}}\, B_\Psi(\boldsymbol{p}\|\nabla\Psi^*(\boldsymbol{\theta})),$$

where $\mathcal{C} \subseteq \mathsf{dom}(\Psi)$. The r.h.s. is the Bregman projection of $\nabla\Psi^*(\boldsymbol{\theta})$ onto $\mathcal{C}$.

Difference of Bregman divergences. Let $\boldsymbol{p} = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$. Using (36), we obtain

$$\begin{aligned} B_\Psi(\boldsymbol{y}\|\nabla\Psi^*(\boldsymbol{\theta})) - B_\Psi(\boldsymbol{p}\|\nabla\Psi^*(\boldsymbol{\theta})) &= \Psi(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y}\rangle + \langle \boldsymbol{\theta}, \boldsymbol{p}\rangle - \Psi(\boldsymbol{p}) \\ &= \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y}\rangle + \Omega^*(\boldsymbol{\theta}) \\ &= L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}), \end{aligned} \tag{37}$$

where we assumed $\boldsymbol{y} \in \mathcal{C}$ and $\mathcal{C} \subseteq \mathsf{dom}(\Psi)$, implying $\Psi(\boldsymbol{y}) = \Omega(\boldsymbol{y})$.

If $\mathcal{C} = \mathsf{dom}(\Psi)$ (i.e., $\Omega = \Psi$), then $\boldsymbol{p} = \nabla\Psi^*(\boldsymbol{\theta})$ and $B_\Psi(\boldsymbol{p}\|\nabla\Psi^*(\boldsymbol{\theta})) = 0$. We thus get the **composite form** of Fenchel-Young losses

$$B_\Omega(\boldsymbol{y}\|\nabla\Omega^*(\boldsymbol{\theta})) = B_\Omega(\boldsymbol{y}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})) = L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}).$$

Bound. Let $\boldsymbol{p} = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$. Since $\boldsymbol{p}$ is the Bregman projection of $\nabla\Psi^*(\boldsymbol{\theta})$ onto $\mathcal{C}$, we can use the well-known Pythagorean theorem for Bregman divergences (see, e.g., Banerjee et al. (2005, Appendix A)) to obtain for all $\boldsymbol{y} \in \mathcal{C} \subseteq \mathsf{dom}(\Psi)$:

$$B_\Psi(\boldsymbol{y}\|\boldsymbol{p}) + B_\Psi(\boldsymbol{p}\|\nabla\Psi^*(\boldsymbol{\theta})) \leq B_\Psi(\boldsymbol{y}\|\nabla\Psi^*(\boldsymbol{\theta})).$$

Using (37), we obtain for all $\boldsymbol{y} \in \mathcal{C}$:

$$0 \leq B_\Psi(\boldsymbol{y} \| \boldsymbol{p}) \leq L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}).$$

Since Ω is a l.s.c. proper convex function, from Proposition 2, we immediately get

$$\boldsymbol{p} = \boldsymbol{y} \Leftrightarrow L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = 0 \Leftrightarrow B_\Psi(\boldsymbol{y} \| \boldsymbol{p}) = 0.$$

B.3. Proof of Proposition 4

The two facts stated in Proposition 4 (H is always non-negative and maximized by the uniform distribution) follow directly from Jensen's inequality. Indeed, for all $\boldsymbol{p} \in \triangle^d$:

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

where $\mathcal{P}$ is the set of $d \times d$ permutation matrices. Strict concavity ensures that $\boldsymbol{p} = \mathbf{1}/d$ is the unique maximizer.

B.4. Proof of Proposition 5

We start by proving the following lemma.

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

1. *We have $\boldsymbol{\theta} \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$ iff $\theta_k = (-\mathsf{H})^*(\boldsymbol{\theta})$. That is:*

$$\partial(-\mathsf{H})(\boldsymbol{e}_k) = \{\boldsymbol{\theta} \in \mathbb{R}^d \colon \theta_k \geq \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle + \mathsf{H}(\boldsymbol{p}),\ \forall \boldsymbol{p} \in \triangle^d\}.$$

2. *If $\boldsymbol{\theta} \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$, then, we also have $\boldsymbol{\theta}' \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$ for any $\boldsymbol{\theta}'$ such that $\theta'_k = \theta_k$ and $\theta'_i \leq \theta_i$, for all $i \neq k$.*

Proof of the lemma: Let $\Omega = -\mathsf{H}$. From Proposition 1 (order preservation), we can consider $\partial\Omega(\boldsymbol{e}_1)$ without loss of generality, in which case any $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{e}_1)$ satisfies $\theta_1 = \max_j \theta_j$. We have $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{e}_1)$ iff $\Omega(\boldsymbol{e}_1) = \langle \boldsymbol{\theta}, \boldsymbol{e}_1 \rangle - \Omega^*(\boldsymbol{\theta}) = \theta_1 - \Omega^*(\boldsymbol{\theta})$. Since $\Omega(\boldsymbol{e}_1) = 0$, we must have $\theta_1 = \Omega^*(\boldsymbol{\theta}) \geq \sup_{\boldsymbol{p} \in \triangle^d} \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle - \Omega(\boldsymbol{p})$, which proves part 1. To see 2, note that we have $\theta'_k = \theta_k \geq \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle - \Omega(\boldsymbol{p}) \geq \langle \boldsymbol{\theta}', \boldsymbol{p} \rangle - \Omega(\boldsymbol{p})$, for all $\boldsymbol{p} \in \triangle^d$, from which the result follows. ■

We now proceed to the proof of Proposition 5. Let $\Omega = -\mathsf{H}$, and suppose that L_Ω has the separation margin property. Then, $\boldsymbol{\theta} = m\boldsymbol{e}_1$ satisfies the margin condition $\theta_1 \geq m + \max_{j \neq 1} \theta_j$, hence $L_\Omega(m\boldsymbol{e}_1, \boldsymbol{e}_1) = 0$. From the first part of Proposition 2, 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 Lemma 1, this implies that $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{e}_1)$ for any $\boldsymbol{\theta}$ such that $\theta_1 = m$ and $\theta_i \leq 0$ for all $i \geq 2$; and more generally we have $\boldsymbol{\theta} + c\mathbf{1} \in \partial\Omega(\boldsymbol{e}_1)$. That is, any $\boldsymbol{\theta}$ with $\theta_1 \geq m + \max_{i \neq 1} \theta_i$ satisfies $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{e}_1)$. From Proposition 2, this is equivalent to $L_\Omega(\boldsymbol{\theta}; \boldsymbol{e}_1) = 0$.

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 Lemma 1, this is equivalent to $m \geq mp_1 - \Omega(\boldsymbol{p})$ for any $\boldsymbol{p} \in \triangle^d$, i.e., $\frac{-\Omega(\boldsymbol{p})}{1-p_1} \leq m$. Note that by Proposition 1 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 (22).

B.5. Proof of Proposition 6

Let us start by showing that conditions 1 and 2 are equivalent. To show that $2 \Rightarrow 1$, take an arbitrary $\boldsymbol{p} \in \triangle^d$. From Fenchel-Young duality and the Danskin's theorem, we have that $\nabla(-\mathsf{H})^*(\boldsymbol{\theta}) = \boldsymbol{p} \;\Rightarrow\; \boldsymbol{\theta} \in \partial(-\mathsf{H})(\boldsymbol{p})$, which implies the subdifferential set is non-empty everywhere in the simplex. Let us now prove that $1 \Rightarrow 2$. Let $\Omega = -\mathsf{H}$, and assume that Ω has non-empty subdifferential everywhere in $\triangle^d$. We need to show that for any $\boldsymbol{p} \in \triangle^d$, there is some $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $\boldsymbol{p} \in \mathsf{argmin}_{\boldsymbol{p}' \in \triangle^d} \Omega(\boldsymbol{p}') - \langle \boldsymbol{\theta}, \boldsymbol{p}' \rangle$. The Lagrangian associated with this minimization problem is:

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

The KKT conditions are:

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

For a given $\boldsymbol{p} \in \triangle^d$, we seek $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{p})$. From the assumption that Ω has non-empty subdifferential in all the simplex, we have that for any $\boldsymbol{p} \in \triangle^d$ we can find a $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $(\boldsymbol{p}, \boldsymbol{\theta})$ are a dual pair, i.e., $\boldsymbol{p} = \nabla\Omega^*(\boldsymbol{\theta})$, which proves that $\nabla\Omega^*(\mathbb{R}^d) = \triangle^d$.

Next, we show that condition $1 \Rightarrow 3$. Since $\partial(-\mathsf{H})(\boldsymbol{p}) \neq \varnothing$ everywhere in the simplex, we can take an arbitrary $\boldsymbol{\theta} \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$. From Lemma 1, item 2, we have that $\boldsymbol{\theta}' \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$ for $\theta'_k = \theta_k$ and $\theta'_j = \min_\ell \theta_\ell$; since $(-\mathsf{H})^*$ is shift invariant, we can without loss of generality have $\theta' = m\boldsymbol{e}_k$ for some $m > 0$, which implies from Proposition 5 that L_Ω has a margin.

Let us show that, if $-\mathsf{H}$ is separable, then $3 \Rightarrow 1$, which establishes equivalence between all conditions 1, 2, and 3. From Proposition 5, the existing of a separation margin implies that there is some m such that $m\boldsymbol{e}_k \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$. Let $\mathsf{H}(\boldsymbol{p}) = \sum_{i=1}^d h(p_i)$, with $h : [0,1] \to \mathbb{R}_+$ concave. Due to assumption A.1, h must satisfy $h(0) = h(1) = 0$. Without loss of generality, suppose $\boldsymbol{p} = [\tilde{\boldsymbol{p}}; \mathbf{0}_k]$, where $\tilde{\boldsymbol{p}} \in \mathsf{relint}(\triangle^{d-k})$ and $\mathbf{0}_k$ is a vector with k zeros. We will see that there is a vector $\boldsymbol{g} \in \mathbb{R}^d$ such that $\boldsymbol{g} \in \partial(-\mathsf{H})(\boldsymbol{p})$, i.e., satisfying

$$-\mathsf{H}(\boldsymbol{p}') \geq -\mathsf{H}(\boldsymbol{p}) + \langle \boldsymbol{g}, \boldsymbol{p}' - \boldsymbol{p} \rangle, \quad \forall \boldsymbol{p}' \in \triangle^d. \tag{39}$$

Since $\tilde{\boldsymbol{p}} \in \mathsf{relint}(\triangle^{d-k})$, we have $\tilde{p}_i \in]0,1[$ for $i \in \{1,\ldots,d-k\}$, hence $\partial(-h)(\tilde{p}_i)$ must be nonempty, since $-h$ is convex and $]0,1[$ is an open set. We show that the following $\boldsymbol{g} = (g_1,\ldots,g_d) \in \mathbb{R}^d$ is a subgradient of $-\mathsf{H}$ at $\boldsymbol{p}$:

$$g_i = \begin{cases} \partial(-h)(\tilde{p}_i), & i = 1,\ldots,d-k \\ m, & i = d-k+1,\ldots,d. \end{cases}$$

By definition of subgradient, we have

$$-\psi(p'_i) \geq -\psi(\tilde{p}_i) + \partial(-h)(\tilde{p}_i)(p'_i - \tilde{p}_i), \quad \text{for } i = 1,\ldots,d-k. \tag{40}$$

Furthermore, since m upper bounds the separation margin of H, we have from Proposition 5 that $m \geq \frac{\mathsf{H}([1-p'_i,p'_i,0,\ldots,0])}{1-\max\{1-p'_i,p'_i\}} = \frac{h(1-p'_i)+h(p'_i)}{\min\{p'_i,1-p'_i\}} \geq \frac{h(p'_i)}{p'_i}$ for any $p'_i \in]0,1]$. Hence, we have

$$-\psi(p'_i) \geq -\psi(0) - m(p'_i - 0), \quad \text{for } i = d-k+1,\ldots,d. \tag{41}$$

Summing all inequalities in Eqs. (40)–(41), we obtain the expression in Eq. (39), which finishes the proof.

B.6. Proof of Proposition 7

Define $\Omega = -\mathsf{H}$. Let us start by writing the margin expression (22) as a unidimensional optimization problem. This is done by noticing that the max-generalized entropy problem constrained to $\max(\boldsymbol{p}) = 1-t$ 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 4. 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 6. 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 $\nabla_j\Omega(\boldsymbol{p}) = -h'(p_j)$, in particular $\nabla_1\Omega(\boldsymbol{e}_1) = -h'(1)$ and $\nabla_2\Omega(\boldsymbol{e}_1) = -h'(0)$, yielding $\text{margin}(L_\Omega) = h'(0) - h'(1)$. Since h is twice differentiable, this equals $-\int_0^1 h''(t)dt$, completing the proof.

B.7. Proof of Proposition 8

We start by proving the following lemma, which generalizes Lemma 1.

Lemma 2 *Let Ω be convex. Then:*

1. *We have $\partial\Omega(\boldsymbol{y}) = \{\boldsymbol{\theta} \in \mathbb{R}^d \colon \langle\boldsymbol{\theta}, \boldsymbol{y}\rangle - \Omega(\boldsymbol{y}) \geq \langle\boldsymbol{\theta}, \boldsymbol{\mu}\rangle - \Omega(\boldsymbol{\mu}),\ \forall\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})\}$.*
2. *If $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{y})$, then, we also have $\boldsymbol{\theta}' \in \partial\Omega(\boldsymbol{y})$ for any $\boldsymbol{\theta}'$ such that $\langle\boldsymbol{\theta}' - \boldsymbol{\theta}, \boldsymbol{y}'\rangle \leq \langle\boldsymbol{\theta}' - \boldsymbol{\theta}, \boldsymbol{y}\rangle$, for all $\boldsymbol{y}' \in \mathcal{Y}$.*

Proof of the lemma: The first part comes directly from the definition of Fenchel conjugate. To see the second part, note that, if $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{y})$, then $\langle\boldsymbol{\theta}', \boldsymbol{y}\rangle - \Omega(\boldsymbol{y}) = \langle\boldsymbol{\theta}' - \boldsymbol{\theta}, \boldsymbol{y}\rangle + \langle\boldsymbol{\theta}, \boldsymbol{y}\rangle - \Omega(\boldsymbol{y}) \geq \langle\boldsymbol{\theta}' - \boldsymbol{\theta}, \boldsymbol{y}\rangle + \langle\boldsymbol{\theta}, \boldsymbol{\mu}\rangle - \Omega(\boldsymbol{\mu})$ for every $\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})$. Let $\boldsymbol{\mu} = \sum_{\boldsymbol{y}'\in\mathcal{Y}} p(\boldsymbol{y}')\boldsymbol{y}'$ for some distribution $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$. Then, we have $\langle\boldsymbol{\theta}' - \boldsymbol{\theta}, \boldsymbol{y}\rangle + \langle\boldsymbol{\theta}, \boldsymbol{\mu}\rangle - \Omega(\boldsymbol{\mu}) = \langle\boldsymbol{\theta}' - \boldsymbol{\theta}, \boldsymbol{y}\rangle + \sum_{\boldsymbol{y}'\in\mathcal{Y}} p(\boldsymbol{y}')\langle\boldsymbol{\theta}, \boldsymbol{y}'\rangle - \Omega(\boldsymbol{\mu}) \geq \langle\boldsymbol{\theta}' - \boldsymbol{\theta}, \boldsymbol{y}\rangle + \sum_{\boldsymbol{y}'\in\mathcal{Y}} p(\boldsymbol{y}')(\langle\boldsymbol{\theta}', \boldsymbol{y}'\rangle - \langle\boldsymbol{\theta}' - \boldsymbol{\theta}, \boldsymbol{y}\rangle) - \Omega(\boldsymbol{\mu}) = \langle\boldsymbol{\theta}', \boldsymbol{\mu}\rangle - \Omega(\boldsymbol{\mu})$.

We now proceed to the proof of Proposition 8. Suppose that L_Ω has the separation margin property, and let $\boldsymbol{\theta} = m\boldsymbol{y}$. Then, for any $\boldsymbol{y}' \in \mathcal{Y}$, we have $\langle\boldsymbol{\theta}, \boldsymbol{y}'\rangle + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{y}'\|^2 = m\langle\boldsymbol{y}, \boldsymbol{y}'\rangle + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{y}'\|^2 = \frac{m}{2}\|\boldsymbol{y}\| + \frac{m}{2}\|y'\|^2 = mr^2 = \langle m\boldsymbol{y}, \boldsymbol{y}\rangle = \langle\boldsymbol{\theta}, \boldsymbol{y}\rangle$, which implies that $L_\Omega(\boldsymbol{\theta}, \boldsymbol{y}) = 0$. From the first part of Proposition 2, this implies $m\boldsymbol{y} \in \partial\Omega(\boldsymbol{y})$.

Conversely, let us assume that $m\boldsymbol{y} \in \partial\Omega(\boldsymbol{y})$. From Lemma 2, this implies that $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{y})$ for any $\boldsymbol{\theta}$ such that $\langle\boldsymbol{\theta} - m\boldsymbol{y}, \boldsymbol{y}'\rangle \leq \langle\boldsymbol{\theta} - m\boldsymbol{y}, \boldsymbol{y}\rangle$ for all $\boldsymbol{y}' \in \mathcal{Y}$. Therefore, we have $\langle\boldsymbol{\theta}, \boldsymbol{y}\rangle - mr^2 \geq \langle\boldsymbol{\theta}, \boldsymbol{y}'\rangle - m\langle\boldsymbol{y}, \boldsymbol{y}'\rangle$, that is, $\langle\boldsymbol{\theta}, \boldsymbol{y}\rangle = \langle\boldsymbol{\theta}, \boldsymbol{y}'\rangle + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{y}'\|^2$. That is, any such $\boldsymbol{\theta}$ satisfies $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{y})$. From Proposition 2, this is equivalent to $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = 0$.

Let us now determine the margin of L_Ω, i.e., the smallest m such that $m\boldsymbol{y} \in \partial\Omega(\boldsymbol{y})$. From Lemma 2, this is equivalent to $mr^2 - \Omega(\boldsymbol{y}) \geq \langle m\boldsymbol{y}, \boldsymbol{\mu}\rangle - \Omega(\boldsymbol{\mu})$ for any $\boldsymbol{\mu} \in \mathsf{conv}(\mathcal{Y})$, i.e., $\frac{-\Omega(\boldsymbol{\mu})+\Omega(\boldsymbol{y})}{r^2 - \langle \boldsymbol{y}, \boldsymbol{\mu}\rangle} \leq m$, which leads to the expression (28).

B.8. Proof of Proposition 9

Let $\Omega(\boldsymbol{p}) = \sum_{j=1}^d g(p_j) + I_{\triangle^d}(\boldsymbol{p})$, where $g\colon [0,1] \to \mathbb{R}_+$ is a non-negative, strictly convex, differentiable function. Therefore, g' is strictly monotonic on $[0,1]$, thus invertible. We show how computing $\nabla(\Omega)^*$ reduces to finding the root of a monotonic scalar function, for which efficient algorithms are available.

From strict convexity and the definition of the convex conjugate,

$$\nabla(\Omega)^*(\boldsymbol{\theta}) = \operatorname*{\mathsf{argmax}}_{\boldsymbol{p} \in \triangle^d} \langle \boldsymbol{p}, \boldsymbol{\theta}\rangle - \sum_j g(p_j).$$

The constrained optimization problem above has Lagrangian

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\nu}, \tau) \coloneqq \sum_{j=1}^d g(p_j) - \langle \boldsymbol{\theta} + \boldsymbol{\nu}, \boldsymbol{p}\rangle + \tau(\mathbf{1}^\top \boldsymbol{p} - 1).$$

A solution $(\boldsymbol{p}^\star, \boldsymbol{\nu}^\star, \tau^\star)$ must satisfy the KKT conditions

$$\begin{cases} g'(p_j) - \theta_j - \nu_j + \tau = 0 & \forall j \in [d] \\ \langle \boldsymbol{p}, \boldsymbol{\nu}\rangle = 0 & \\ \boldsymbol{p} \in \triangle^d,\ \boldsymbol{\nu} \geq 0. & \end{cases} \tag{44}$$

Let us define

$$\tau_{\min} \coloneqq \max(\boldsymbol{\theta}) - g'(1) \quad \text{and} \quad \tau_{\max} \coloneqq \max(\boldsymbol{\theta}) - g'\left(\frac{1}{d}\right).$$

Since g is strictly convex, g' is increasing and so $\tau_{\min} < \tau_{\max}$. For any $\tau \in [\tau_{\min}, \tau_{\max}]$, we construct $\boldsymbol{\nu}$ as

$$\nu_j \coloneqq \begin{cases} 0, & \theta_j - \tau \geq g'(0) \\ g'(0) - \theta_j + \tau, & \theta_j - \tau < g'(0) \end{cases}$$

By construction, $\nu_j \geq 0$, satisfying dual feasibility. Injecting ν into (44) and combining the two cases, we obtain

$$g'(p_j) = \max\{\theta_j - \tau,\ g'(0)\}. \tag{45}$$

We show that i) the stationarity conditions have a unique solution given τ, and ii) $[\tau_{\min}, \tau_{\max}]$ forms a sign-changing bracketing interval, and thus contains $\tau^\star$, which can then be found by one-dimensional search. The solution verifies all KKT conditions, thus is globally optimal.

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

$$\begin{aligned} g'(0) \leq g'(p_j) = \theta_j - \tau &\leq \max(\boldsymbol{\theta}) - \tau_{\min} \\ &= \max(\boldsymbol{\theta}) - \max(\boldsymbol{\theta}) + g'(1) \\ &= g'(1). \end{aligned}$$

Otherwise, $g'(p_j) = g'(0)$. This verifies that the r.h.s. of (45) is always within the domain of $(g')^{-1}$. We can thus apply the inverse to both sides to solve for p_j, obtaining

$$p_j(\tau) = (g')^{-1}(\max\{\theta_j - \tau,\ g'(0)\}). \tag{46}$$

Strict convexity implies the optimal $\boldsymbol{p}^\star$ is unique; it can be seen that $\tau^\star$ is also unique. Indeed, assume optimal $\tau_1^\star, \tau_2^\star$. Then, $\boldsymbol{p}(\tau_1^\star) = \boldsymbol{p}(\tau_2^\star)$, so $\max(\boldsymbol{\theta} - \tau_1^\star, g'(0)) = \max(\boldsymbol{\theta} - \tau_2^\star, g'(0))$. This implies either $\tau_1^\star = \tau_2^\star$, or $\boldsymbol{\theta} - \tau_{\{1,2\}}^\star \leq g'(0)$, in which case $\boldsymbol{p} = \mathbf{0} \notin \triangle^d$, which is a contradiction.

Validating the bracketing interval. Consider the primal infeasibility function $\phi(\tau) \coloneqq \langle \boldsymbol{p}(\tau), \mathbf{1} \rangle - 1$; $\boldsymbol{p}(\tau)$ is primal feasible iff $\phi(\tau) = 0$. We show that ϕ is decreasing on $[\tau_{\min}, \tau_{\max}]$, and that it has opposite signs at the two extremities. From the intermediate value theorem, the unique root $\tau^\star$ must satisfy $\tau^\star \in [\tau_{\min}, \tau_{\max}]$.

Since g' is increasing, so is $(g')^{-1}$. Therefore, for all j, $p_j(\tau)$ is decreasing, and so is the sum $\phi(\tau) = \sum_j p_j(\tau) - 1$. It remains to check the signs at the boundaries.

$$\begin{aligned} \sum_i p_i(\tau_{\max}) &= \sum_i (g')^{-1}(\max\{\theta_i - \max(\boldsymbol{\theta}) + g'\left({}^1\!/_d\right),\ g'(0)\}) \\ &\leq d\ (g')^{-1}(\max\{g'\left({}^1\!/_d\right),\ g'(0)\}) \\ &= d\ (g')^{-1}\left(g'\left({}^1\!/_d\right)\right) = 1, \end{aligned}$$

where we upper-bounded each term of the sum by the largest one. At the other end,

$$\begin{aligned} \sum_i p_i(\tau_{\min}) &= \sum_i (g')^{-1}(\max\{\theta_i - \max(\boldsymbol{\theta}) + g'(1),\ g'(0)\}) \\ &\geq (g')^{-1}(\max\{g'(1),\ g'(0)\}) \\ &= (g')^{-1}(g'(1)) = 1, \end{aligned}$$

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

Algorithm 1 is an example of a bisection algorithm for finding an approximate solution; more advanced root finding methods can also be used. We note that the resulting algorithm resembles the method provided in Krichene et al. (2015), with a non-trivial difference being the order of the thresholding and $(-g)^{-1}$ in Eq. (46).

Algorithm 1 Bisection for $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta})$

Input: $\boldsymbol{\theta} \in \mathbb{R}^d$, $\Omega(\boldsymbol{p}) = I_{\triangle^d} + \sum_i g(p_i)$
$\boldsymbol{p}(\tau) \coloneqq (g')^{-1}(\max\{\boldsymbol{\theta} - \tau, g'(0)\})$
$\phi(\tau) \coloneqq \langle \boldsymbol{p}(\tau), \mathbf{1} \rangle - 1$
$\tau_{\min} \leftarrow \max(\boldsymbol{\theta}) - g'(1)$;
$\tau_{\max} \leftarrow \max(\boldsymbol{\theta}) - g'\left(1/d\right)$
$\tau \leftarrow (\tau_{\min} + \tau_{\max})/2$
while $|\phi(\tau)| > \epsilon$
  if $\phi(\tau) < 0$ $\tau_{\max} \leftarrow \tau$
  else $\tau_{\min} \leftarrow \tau$
  $\tau \leftarrow (\tau_{\min} + \tau_{\max})/2$
Output: $\nabla\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \approx \boldsymbol{p}(\tau)$

B.9. Proof of Proposition 10

We first need the following lemma.

Lemma 3 *Let $\ell(\boldsymbol{\theta}; \boldsymbol{e}_k)$ be defined as*

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

where $\phi \colon \mathbb{R} \to \mathbb{R}_+$ is convex. Then, H_ℓ defined in (34) equals

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

Proof We want to solve

$$\mathsf{H}_\ell(\boldsymbol{p}) = \min_{\boldsymbol{\theta}^\top \mathbf{1} = 0} \sum_j p_j \ell(\boldsymbol{\theta}; \boldsymbol{e}_j) = \min_{\boldsymbol{\theta}^\top \mathbf{1} = 0} \sum_j p_j \sum_i c_{j,i}\phi(\theta_i) = \min_{\boldsymbol{\theta}^\top \mathbf{1} = 0} \sum_i (\boldsymbol{p}^\top \boldsymbol{c}_i)\phi(\theta_i),$$

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

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

Strong duality holds and we can swap the order of the min and max. After routine calculations, we obtain

$$-\mathsf{H}_\ell(\boldsymbol{p}) = \min_{\tau \in \mathbb{R}} \sum_i (\boldsymbol{p}^\top \boldsymbol{c}_i)\phi^*\left(\frac{-\tau}{\boldsymbol{p}^\top \boldsymbol{c}_i}\right). \tag{47}$$

■

We now prove Proposition 10. First, we rewrite $\ell(\boldsymbol{\theta}; \boldsymbol{e}_k) = \sum_{j \neq k} \phi(\theta_j - \theta_k)$ as $\ell(\boldsymbol{\theta}; \boldsymbol{e}_k) =$

$\sum_j c_{k,j}\phi(\theta_j - \theta_k)$, where we choose $c_{k,j} = 0$ if $k = j$, 1 otherwise, leading to $\boldsymbol{p}^\top \boldsymbol{c}_j = 1 - p_j$. Because $\phi(\theta_j - \theta_k)$ is shift invariant w.r.t. $\boldsymbol{\theta}$, without loss of generality, we can further rewrite the loss as $\ell(\boldsymbol{\theta}; \boldsymbol{e}_k) = \sum_j c_{k,j}\phi(\theta_j)$ with $\mathsf{dom}(\ell) = \{\boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{Y}|} \colon \boldsymbol{\theta}^\top \mathbf{1} = 0\}$. Hence, Lemma 3 applies. We now derive closed form for specific choices of ϕ.

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

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

This recovers Duchi et al. (2018, 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 (Shalev-Shwartz and Zhang, 2016):

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

Going back to ϕ, we obtain:

$$\phi(t) = \begin{cases} 0 & \text{if } t \le -1 \\ t + \frac{1}{2} & \text{if } t \ge 0 \\ \frac{1}{2}(1+t)^2 & \text{o.w.} \end{cases}. \tag{48}$$

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 (47), we obtain

$$-\mathsf{H}_\ell(\boldsymbol{p}) = \min_{\tau\in\mathcal{C}} \frac{\tau^2}{2} \sum_{j=1}^d \frac{1}{\boldsymbol{p}^\top \boldsymbol{c}_j} + d\,\tau. \tag{49}$$

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

$$\tau = -d \Big/ \Big(\sum_{j=1}^d 1/(\boldsymbol{p}^\top \boldsymbol{c}_j)\Big). \tag{50}$$

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 (49) gives the claimed expression of H_ℓ.

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

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

The constraint is now $\tau \leq 0$. Hence, the optimal solution of the problem w.r.t. τ in (47) is now (50) for all $\boldsymbol{p} \in \triangle^d, \boldsymbol{c}_j \geq \boldsymbol{0}$. Simplifying, we get

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

References

Ryan P Adams and Richard S Zemel. Ranking via sinkhorn propagation. *arXiv e-prints*, 2011.

Nir Ailon, Kohei Hatano, and Eiji Takimoto. Bandit online optimization over the permutahedron. *Theoretical Computer Science*, 650:92–108, 2016.

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.

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. *arXiv preprint arXiv:1203.4523*, 2012.

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. *JMLR*, 6:1705–1749, 2005.

Ole Barndorff-Nielsen. *Information and Exponential Families: In Statistical Theory*. John Wiley & Sons, 1978.

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

Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

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

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

David Belanger, Dan Sheldon, and Andrew McCallum. Marginal inference in MRFs using Frank-Wolfe. In *NeurIPS Workshop on Greedy Opt., FW and Friends*, 2013.

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.

Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucumán Rev. Ser. A*, 5:147–151, 1946.

Mathieu Blondel. Structured prediction with projection oracles. In *Proc. of NeurIPS*, 2019.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *Proc. of AISTATS*, 2018.

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning classifiers with Fenchel-Young losses: Generalized entropies, margins, and algorithms. In *Proc. of AISTATS*, 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.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013.

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.

Yoeng-Jin Chu and Tseng-Hong Liu. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400, 1965.

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

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *JMLR*, 12:2493–2537, 2011.

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. *JMLR*, 2:265–292, 2001.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. In *Proc. of NeurIPS*, 2013.

Marco Cuturi and Mathieu Blondel. Soft-DTW: A differentiable loss function for time-series. In *Proc. of ICML*, 2017.

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

George B Dantzig, Alex Orden, and Philip Wolfe. The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific Journal of Mathematics*, 5(2):183–195, 1955.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Proc. of NeurIPS*, 2014.

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

Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the Rot Mover's Distance. *arXiv preprint arXiv:1610.06447*, 2016.

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

John C Duchi, Lester W Mackey, and Michael I Jordan. On the consistency of ranking algorithms. In *Proc. of ICML*, pages 327–334, 2010.

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

Joseph C Dunn and S Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

Jack Edmonds. Optimum branchings. *J. Res. Nat. Bur. Stand.*, 71B:233–240, 1967.

Jason Eisner. Inside-outside and forward-backward algorithms are just backprop (tutorial paper). In *Proc. of the Workshop on Structured Prediction for NLP*, 2016.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

Rafael Frongillo and Mark D Reid. Convex foundations for generalized MaxEnt models. In *Proc. of AIP*, 2014.

Dario Garcia-Garcia and Robert C Williamson. Divergences and risks for multiclass experiments. In *Proc. of COLT*, 2012.

Damien Garreau, Rémi Lajugie, Sylvain Arlot, and Francis Bach. Metric learning for temporal sequence alignment. In *Proc. of NeurIPS*, 2014.

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

Kevin Gimpel and Noah A Smith. Softmax-margin CRFs: Training log-linear models with cost functions. In *Proc. of NAACL*, 2010.

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.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(Mar):723–773, 2012.

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

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

David P Helmbold and Manfred K Warmuth. Learning permutations with exponential weights. *JMLR*, 10(Jul):1705–1736, 2009.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proc. of ICML*, 2013.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proc. of ACM SIGKDD*. ACM, 2002.

Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

Dan Jurafsky and James H Martin. *Speech and Language Processing (3rd ed.)*. draft, 2018.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proc. of NeurIPS*, 2009.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Eliyahu Kiperwasser and Yoav Goldberg. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *TACL*, 4:313–327, 2016.

Gustav Kirchhoff. Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. *Annalen der Physik*, 148(12): 497–508, 1847.

Terry Koo, Amir Globerson, Xavier Carreras Pérez, and Michael Collins. Structured prediction models via the matrix-tree theorem. In *EMNLP*, 2007.

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

Rahul G Krishnan, Simon Lacoste-Julien, and David Sontag. Barrier Frank-Wolfe for marginal inference. In *Proc. of NeurIPS*, 2015.

Harold W Kuhn. The Hungarian method for the assignment problem. *Nav. Res. Log.*, 2 (1-2):83–97, 1955.

Anastasios Kyrillidis, Stephen Becker, Volkan Cevher, and Christoph Koch. Sparse projections onto the simplex. In *Proc. of ICML*, 2013.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Proc. of NeurIPS*, 2015.

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *Proc. of ICML*, 2012.

Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank-Wolfe optimization for particle filtering. In *Proc. of AISTATS*, 2015.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, 2001.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proc. of NAACL*, 2016.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Top-k multiclass SVM. In *Proc. of NeurIPS*, 2015.

Cong Han Lim and Stephen J Wright. Efficient bregman projections onto the permutahedron and related polytopes. In *Artificial Intelligence and Statistics*, pages 1205–1213, 2016.

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

Francesco Locatello, Michael Tschannen, Gunnar Rätsch, and Martin Jaggi. Greedy algorithms for cone constrained optimization with convergence guarantees. In *Proc. of NeurIPS*, 2017.

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 *Proc. 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. *JMLR*, 10:935–975, 2009a.

André FT Martins, Noah A Smith, and Eric P Xing. Concise integer linear programming formulations for dependency parsing. In *ACL-IJCNLP*, 2009b.

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

Peter McCullagh and John A Nelder. *Generalized Linear Models*, volume 37. CRC press, 1989.

Ryan T McDonald, Fernando CN Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proc. of HLT-EMNLP*, 2005.

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

Jean-Jacques Moreau. Proximit et dualit dans un espace hilbertien. *Bullet de la Société Mathémathique de France*, 93(2):273–299, 1965.

Youssef Mroueh, Tomaso Poggio, Lorenzo Rosasco, and Jean-Jeacques Slotine. Multiclass learning with simplex coding. In *Proc. of NeurIPS*, pages 2789–2797, 2012.

Renato Negrinho and Andre Martins. Orbit regularization. In *Proc. of NeurIPS*, 2014.

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.

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

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

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

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. Universal Dependencies v1: A multilingual treebank collection. In *LREC*, 2016.

Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer New York, 1999.

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

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. A general theory for structured prediction with smooth convex surrogates. *arXiv preprint arXiv:1902.01958*, 2019.

Fabian Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *JMLR*, 12:2825–2830, 2011.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 2017.

Venkata Krishna Pillutla, Vincent Roulet, Sham M Kakade, and Zaid Harchaoui. A Smoother Way to Train Structured Prediction Models. In *Proc. of NeurIPS*, 2018.

Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On ndcg consistency of listwise ranking methods. In *Proc. of AISTATS*, pages 618–626, 2011.

Mark D Reid and Robert C Williamson. Composite binary losses. *JMLR*, 11:2387–2422, 2010.

Mark D Reid and Robert C Williamson. Information, divergence and risk for binary experiments. *JMLR*, 12:731–817, 2011.

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

Ryan M Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *JMLR*, 5(Jan): 101–141, 2004.

Ryan M Rifkin and Ross A Lippert. Value regularization and Fenchel duality. *JMLR*, 8 (Mar):441–479, 2007.

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.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26: 43–49, 1978.

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 Sham M Kakade. Mind the duality gap: Logarithmic regret algorithms for online optimization. In *Proc. of NeurIPS*, 2009.

Shai Shalev-Shwartz and Yoram Singer. Convex repeated games and Fenchel duality. In *Proc. of NeurIPS*, 2007.

Shai Shalev-Shwartz and Yoram Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. *Machine Learning*, 80 (2-3):141–163, 2010.

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

Claude E Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois, 1949.

Chunhua Shen and Hanxi Li. On the dual formulation of boosting algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2216–2231, 2010.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

David A Smith and Noah A Smith. Probabilistic models of nonprojective dependency trees. In *EMNLP*, 2007.

Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *Proc. of ICML*, 2014.

Milan Straka and Jana Straková. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *CoNLL Shared Task*, 2017.

Daiki Suehiro, Kohei Hatano, Shuji Kijima, Eiji Takimoto, and Kiyohito Nagano. Online prediction under submodular constraints. In *International Conference on Algorithmic Learning Theory*. Springer, 2012.

Hiroki 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.

Robert E Tarjan. Finding optimum branchings. *Networks*, 7(1):25–35, 1977.

Ben Taskar. *Learning Structured Prediction Models: A Large Margin Approach*. PhD thesis, Stanford University, 2004.

Matus Telgarsky. A primal-dual convergence analysis of boosting. *JMLR*, 13(Mar):561–606, 2012.

Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *JMLR*, 8(May):1007–1025, 2007.

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

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.

Nicolas Usunier, David Buffoni, and Patrick Gallinari. Ranking with ordered weighted pairwise classification. In *Proc. of ICML*, 2009.

Leslie G Valiant. The complexity of computing the permanent. *Theor. Comput. Sci.*, 8(2): 189–201, 1979.

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

Marina Vinyes and Guillaume Obozinski. Fast column generation for atomic norm regularization. In *Proc. of AISTATS*, 2017.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

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.

Manfred K Warmuth and Dima Kuzmin. Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *JMLR*, 9:2287–2320, 2008.

Robert C Williamson, Elodie Vernet, and Mark D Reid. Composite multiclass losses. *JMLR*, 2016.

Philip Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1): 128–149, 1976.

Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

Ronald R Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):183–190, 1988.

Shota Yasutake, Kohei Hatano, Shuji Kijima, Eiji Takimoto, and Masayuki Takeda. Online linear optimization over permutations. In *International Symposium on Algorithms and Computation*, pages 534–543. Springer, 2011.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. *CoNLL*, 2017.

Xiangrong Zeng and Mário AT Figueiredo. The ordered weighted ℓ_1 norm: Atomic formulation and conditional gradient algorithm. In *Proc. of SPARS*, 2015.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.