Learning with Fitzpatrick Losses

Seta Rakotomandimby
Ecole des Ponts
seta.rakotomandimby@enpc.fr

Jean-Philippe Chancelier
Ecole des Ponts
jean-philippe.chancelier@enpc.fr

Michel De Lara
Ecole des Ponts
michel.delara@enpc.fr

Mathieu Blondel
Google DeepMind
mblondel@google.com

Abstract

Fenchel-Young losses are a family of convex loss functions, encompassing the squared, logistic and sparsemax losses, among others. Each Fenchel-Young loss is implicitly associated with a link function, for mapping model outputs to predictions. For instance, the logistic loss is associated with the soft argmax link function. Can we build new loss functions associated with the same link function as Fenchel-Young losses? In this paper, we introduce Fitzpatrick losses, a new family of convex loss functions based on the Fitzpatrick function. A well-known theoretical tool in maximal monotone operator theory, the Fitzpatrick function naturally leads to a refined Fenchel-Young inequality, making Fitzpatrick losses tighter than Fenchel-Young losses, while maintaining the same link function for prediction. As an example, we introduce the Fitzpatrick logistic loss and the Fitzpatrick sparsemax loss, counterparts of the logistic and the sparsemax losses. This yields two new tighter losses associated with the soft argmax and the sparse argmax, two of the most ubiquitous output layers used in machine learning. We study in details the properties of Fitzpatrick losses and in particular, we show that they can be seen as Fenchel-Young losses using a modified, target-dependent generating function. We demonstrate the effectiveness of Fitzpatrick losses for label proportion estimation.

1 Introduction

Loss functions are a cornerstone of statistics and machine learning: they measure the difference, or "loss," between a ground-truth target and a model prediction. As such, they have attracted a wealth of research. Proper losses (a.k.a. proper scoring rules) [16, 15] measure the discrepancy between a target distribution and a probability forecast. They are essentially primal-primal Bregman divergences, with both the target and the prediction belonging to the same primal space. They are typically explicitly composed with a link function [24, 27], in order to map the model output to a prediction. A disadvantage of this explicit composition is that it often makes the resulting composite loss function nonconvex. A related family of loss functions are Fenchel-Young losses [6, 7], which encompass many commonly-used loss functions in machine learning including the squared, logistic, sparsemax and perceptron losses. Fenchel-Young losses can be seen as primal-dual Bregman divergences [1], with the target belonging to the primal space and the model output belonging to the dual space. In contrast to proper losses, each Fenchel-Young loss is implicitly associated with a given link function, mapping the dual-space model output to a primal-space prediction (for instance, the soft argmax is the link function associated with the logistic loss). This crucial difference makes Fenchel-Young losses always convex. Can we build new convex losses associated with the same link function as Fenchel-Young losses?

Preprint. Last update: May 22, 2024.

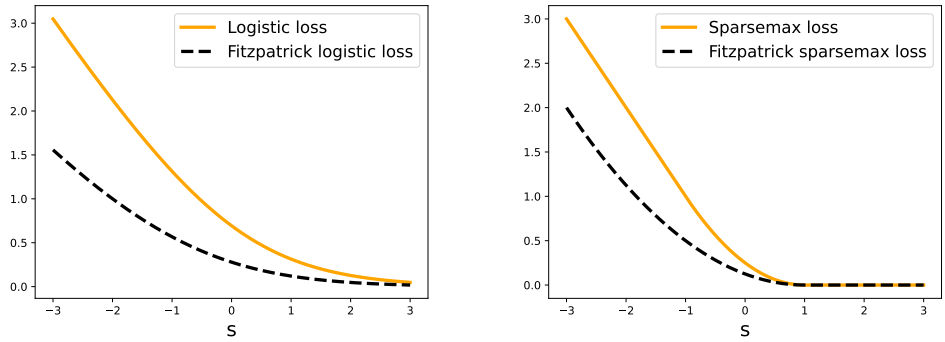

Figure 1: We introduce **Fitzpatrick losses**, a new family of loss functions generated by a convex regularization function Ω, that **lower-bound** Fenchel-Young losses generated by the same Ω, while maintaining the **same** link function $\widehat{y}_\Omega = \nabla\Omega^*$. In particular, we use our framework to instantiate the counterparts of the **logistic** and **sparsemax** losses, two instances of Fenchel-Young losses, associated with the **soft argmax** and the **sparse argmax**. In the figures above, we plot $L(y, \theta)$, where $y = e_1$, $\theta = (s, 0)$ and $L \in \{L_{F[\partial\Omega]}, L_{\Omega\oplus\Omega^*}\}$, confirming the lower-bound property.

In this paper, we introduce Fitzpatrick losses, a new family of primal-dual convex loss functions. Our proposal builds upon the Fitzpatrick function, a well-known theoretical object in maximal monotone operator theory [14, 10, 2]. So far, the Fitzpatrick function had been used as a theoretical tool to represent maximal monotone operators [25] and to construct Bregman-like primal-primal divergences [9], but it had not been used to construct primal-dual loss functions for machine learning, as we do. Crucially, the Fitzpatrick function naturally leads to a refined Fenchel-Young inequality, making Fitzpatrick losses tighter than Fenchel-Young losses. Yet, their predictions are produced using the same link function, suggesting that we can use Fitzpatrick losses as a tighter replacement for the corresponding Fenchel-Young losses (Figure 1). We make the following contributions.

- After reviewing some background, we introduce Fitzpatrick losses. They can be thought as a tighter version of Fenchel-Young losses, that use the same link function.
- We instantiate two new loss functions in this family: the Fitzpatrick logistic loss and the Fitzpatrick sparsemax loss. They are the counterparts of the logistic and sparsemax losses, two instances of Fenchel-Young losses. We therefore obtain two new tighter losses for the soft argmax and the sparse argmax, two of the most popular output layers in machine learning.
- We study in detail the properties of Fitzpatrick losses. We show that Fitzpatrick losses are equivalent to Fenchel-Young losses with a modified, target-dependent generating function.
- We demonstrate the effectiveness of Fitzpatrick losses for probabilistic classification on 11 datasets.

2 Background

2.1 Convex analysis

We define $[k] \coloneqq \{1, \dots, k\}$. We denote the probability simplex by $\triangle^k \coloneqq \{p \in \mathbb{R}_+^k : \sum_{i=1}^k p_i = 1\}$ and the extended reals by $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{+\infty\}$. We denote the indicator function of a set $\mathcal{C}$ by $\iota_\mathcal{C}(y) = 0$ if $y \in \mathcal{C}$, $+\infty$ otherwise. We denote the effective domain of a function $\Omega \colon \mathbb{R}_+^k \to \overline{\mathbb{R}}$ by $\operatorname{dom}\Omega := \{y \in \mathbb{R}^k : \Omega(y) < +\infty\}$. We denote the Euclidean projection onto a closed convex set $\mathcal{C}$ by $P_\mathcal{C}(\theta) = \operatorname{argmin}_{y\in\mathcal{C}} \|y - \theta\|_2^2$.

For a convex function $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$, its **subdifferential** $\partial\Omega$ is defined by

$$(y', \theta') \in \partial\Omega \iff \theta' \in \partial\Omega(y') \iff \Omega(y) \geq \Omega(y') + \langle y - y', \theta' \rangle \; \forall y.$$

When Ω is differentiable, the subdifferential is a singleton and we have $\partial\Omega(y') = \{\nabla\Omega(y')\}$. The **normal cone** to a set $\mathcal{C}$ at y' is defined by

$$\theta' \in N_{\mathcal{C}}(y') \iff \langle y - y', \theta' \rangle \leq 0 \quad \forall y \in \mathcal{C}$$

if $y' \in \mathcal{C}$ and $N_{\mathcal{C}}(y') = \emptyset$ if $y' \notin \mathcal{C}$. The **Fenchel conjugate** $\Omega^* : \mathbb{R}^k \to \overline{\mathbb{R}}$ of a function $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$ is defined by

$$\Omega^*(\theta) \coloneqq \sup_{y' \in \mathbb{R}^k} \langle y', \theta \rangle - \Omega(y').$$

From standard convex analysis, when $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$ is a convex l.s.c. (lower semicontinuous) function,

$$\partial\Omega^*(\theta) = \operatorname*{argmax}_{y' \in \mathbb{R}^k} \langle y', \theta \rangle - \Omega(y').$$

When the argmax is unique, it is equal to $\nabla\Omega^*(\theta)$. We define the **generalized Bregman divergence** [17] $D_\Omega : \mathbb{R}^k \times \mathbb{R}^k \to \overline{\mathbb{R}}_+$ generated by a convex l.s.c. function $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$ by

$$D_\Omega(y, y') \coloneqq \Omega(y) - \Omega(y') - \sup_{\theta' \in \partial\Omega(y')} \langle y - y', \theta' \rangle, \tag{1}$$

with the convention $+\infty + (-\infty) = +\infty$. When Ω is differentiable, it recovers the classical **Bregman divergence**

$$D_\Omega(y, y') \coloneqq \Omega(y) - \Omega(y') - \langle y - y', \nabla\Omega(y') \rangle.$$

Both y and y' belong to the **primal space**.

2.2 Fenchel-Young losses

Definition and properties

The **Fenchel-Young loss** $L_{\Omega \oplus \Omega^*} : \mathbb{R}^k \times \mathbb{R}^k \to \overline{\mathbb{R}}$ generated by a convex l.s.c. function Ω [7] is

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

As its name indicates, it is grounded in the Fenchel-Young inequality

$$\langle y, \theta \rangle \leq \Omega(y) + \Omega^*(\theta) \quad \forall y, \theta \in \mathbb{R}^k.$$

The Fenchel-Young loss enjoys many desirable properties, notably it is **non-negative** and it is **convex** in y and θ separately. The Fenchel-Young loss can be seen as a **primal-dual** Bregman divergence [1, 7], where y belongs to the primal space and θ belongs to the dual space.

Link functions

To map a dual-space θ to a primal-space y, we can use the canonical link function $\partial\Omega^*$, since

$$L_{\Omega \oplus \Omega^*}(y, \theta) = 0 \iff y \in \partial\Omega^*(\theta).$$

In particular when Ω is strictly convex, the Fenchel-Young loss is positive, meaning that it satisfies the identity of indiscernibles

$$L_{\Omega \oplus \Omega^*}(y, \theta) = 0 \iff y = \nabla\Omega^*(\theta).$$

In the remainder of this paper, we will use the notation $\widehat{y}_\Omega(\theta)$ to denote the gradient $\nabla\Omega^*(\theta)$ or any subgradient in $\partial\Omega^*(\theta)$. Since Ω^* is convex, $\widehat{y}_\Omega$ is monotone. As shown in [7], the monotonicity implies that θ and $\widehat{y}_\Omega(\theta)$ are sorted the same way, i.e., $\theta_i > \theta_j \implies \widehat{y}_\Omega(\theta)_i \geq \widehat{y}_\Omega(\theta)_j$. Link functions also play an important role in the loss gradient, as we have

$$\partial_\theta L_{\Omega \oplus \Omega^*}(y, \theta) = \widehat{y}_\Omega(\theta) - y. \tag{2}$$

Examples of Fenchel-Young loss instances and their associated link function

We give a few examples of Fenchel-Young losses. With the squared 2-norm, $\Omega(y') = \frac{1}{2}\|y'\|_2^2$, we obtain the **squared loss**

$$L_{\Omega \oplus \Omega^*}(y, \theta) = L_{\text{squared}}(y, \theta) \coloneqq \frac{1}{2}\|y - \theta\|_2^2$$

and the **identity link**

$$\widehat{y}_\Omega(\theta) = \theta.$$

With the indicator of a convex set $\mathcal{C}$, $\Omega(y') = \iota_\mathcal{C}(y')$, we obtain the **perceptron loss**

$$L_{\Omega\oplus\Omega^*}(y,\theta) = L_{\text{perceptron}}(y,\theta) \coloneqq \max_{y'\in\mathcal{C}} \langle y',\theta\rangle - \langle y,\theta\rangle.$$

and the **argmax link**

$$\widehat{y}_\Omega(\theta) = \operatorname*{argmax}_{y\in\mathcal{C}} \langle y,\theta\rangle.$$

With the squared 2-norm restricted to some convex set $\mathcal{C}$, $\Omega(y') = \frac{1}{2}\|y'\|_2^2 + \iota_\mathcal{C}(y')$, we obtain the **sparseMAP loss** [22]

$$L_{\Omega\oplus\Omega^*}(y,\theta) = L_{\text{sparseMAP}}(y,\theta) \coloneqq \frac{1}{2}\|y-\theta\|_2^2 - \frac{1}{2}\|P_\mathcal{C}(y)-\theta\|_2^2.$$

The link is the **Euclidean projection** onto $\mathcal{C}$,

$$\widehat{y}_\Omega(\theta) = P_\mathcal{C}(\theta).$$

When the set is $\mathcal{C} = \triangle^k$, we obtain the **sparsemax loss** [20] and the **sparsemax link** $\widehat{y}_\Omega(\theta) = P_{\triangle^k}(\theta)$, which is known to produce sparse probability distributions. With the Shannon negentropy restricted to the probability simplex, $\Omega(y) \coloneqq \langle y', \log y'\rangle + \iota_{\triangle^k}(y')$, we obtain the **logistic loss**

$$L_{\Omega\oplus\Omega^*}(y,\theta) = L_{\text{logistic}}(y,\theta) \coloneqq \log\sum_{i=1}^k \exp(\theta_i) + \langle y, \log y\rangle - \langle y,\theta\rangle,$$

and the **soft argmax link** (also know as softmax)

$$\widehat{y}_\Omega(\theta) = \text{softargmax}(\theta) \coloneqq \exp(\theta)/\sum_{i=1}^k \exp(\theta_i).$$

2.3 Maximal monotone operators and the Fitzpatrick function

An operator A is called **monotone** if for all $(y,\theta)\in A$ and all $(y',\theta')\in A$, we have

$$\langle y'-y, \theta'-\theta\rangle \geq 0.$$

We overload the notation to denote $A(y) \coloneqq \{\theta\colon (y,\theta)\in A\}$. A monotone operator A is said to be **maximal** if there does not exist $(y,\theta)\notin A$ such that $A\cup\{(y,\theta)\}$ is still monotone. It is well-known that the subdifferential $\partial\Omega$ of a convex function Ω is maximal monotone. For more details on monotone operators, see [3, 25].

A well-known object in monotone operator theory, the **Fitzpatrick function** associated with a monotone operator A [14, 10, 2], denoted $F[A]:\mathbb{R}^k\times\mathbb{R}^k\to\overline{\mathbb{R}}$, is defined by

$$F[A](y,\theta) \coloneqq \sup_{(y',\theta')\in A} \langle y-y',\theta'\rangle + \langle y',\theta\rangle.$$

In particular, with $A=\partial\Omega$, we have

$$F[\partial\Omega](y,\theta) = \sup_{(y',\theta')\in\partial\Omega} \langle y-y',\theta'\rangle + \langle y',\theta\rangle = \sup_{y'\in\operatorname{dom}\Omega} \langle y',\theta\rangle + \sup_{\theta'\in\partial\Omega(y')} \langle y-y',\theta'\rangle.$$

The Fitzpatrick function was studied in depth in [2]. In particular, it is jointly convex and satisfies

$$\langle y,\theta\rangle \leq F[\partial\Omega](y,\theta) \leq \Omega\oplus\Omega^*(y,\theta) = \Omega(y)+\Omega^*(\theta) \quad \forall y,\theta\in\mathbb{R}^k. \tag{3}$$

From Danskin's theorem, when $\operatorname{dom}\Omega$ is compact, we also have

$$y^\star_{F[\partial\Omega]}(y,\theta) \coloneqq \partial_\theta F[\partial\Omega](y,\theta) = \operatorname*{argmax}_{y'\in\operatorname{dom}\Omega} \langle y',\theta\rangle + \sup_{\theta'\in\partial\Omega(y')} \langle y-y',\theta'\rangle. \tag{4}$$

The Fitzpatrick function $F[\partial\Omega](y,\theta)$ and $\Omega\oplus\Omega^*(y,\theta) = \Omega(y)+\Omega^*(\theta)$ play a similar role but the latter is **separable** in y and θ, while the former is **not**. In particular this makes the subdifferential $\partial_\theta F[\partial\Omega](y,\theta)$ depend on both y and θ, while $\partial_\theta(\Omega\oplus\Omega^*)(y,\theta) = \partial\Omega^*(\theta)$ depends only on θ.

The Fitzpatrick function was used in [9] to theoretically study primal-primal Bregman-like divergences. As discussed in more detail in Section 3.4, using these divergences for machine learning would require us to compose them with an explicit link function, which would typically break convexity. In the next section, we introduce new primal-dual losses based on the Fitzpatrick function.

3 Fitzpatrick losses

3.1 Definition and properties

Inspired by the inequality in (3), which we can view as a refined Fenchel-Young inequality, we introduce Fitzpatrick losses, a new family of loss functions generated by a convex l.s.c. function Ω.

Definition 1 *Fitzpatrick loss generated by a convex l.s.c. function Ω*

When $y \in \operatorname{dom}\Omega$ and $\theta \in \mathbb{R}^k$, we define the Fitzpatrick loss $L_{F[\partial\Omega]} : \mathbb{R}^k \times \mathbb{R}^k \to \overline{\mathbb{R}}$ generated by a proper convex l.s.c. function $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$ by

$$\begin{aligned} L_{F[\partial\Omega]}(y, \theta) &\coloneqq F[\partial\Omega](y, \theta) - \langle y, \theta \rangle \\ &= \sup_{(y', \theta') \in \partial\Omega} \langle y - y', \theta' \rangle + \langle y', \theta \rangle - \langle y, \theta \rangle \\ &= \sup_{(y', \theta') \in \partial\Omega} \langle y' - y, \theta - \theta' \rangle. \end{aligned}$$

When $y \notin \operatorname{dom}\Omega$, $L_{F[\partial\Omega]}(y, \theta) = +\infty$.

Fitzpatrick losses enjoy similar properties as Fenchel-Young losses, but they are **tighter**.

Proposition 1 *Properties of Fitzpatrick losses*

1. ***Non-negativity:*** *for all $(y, \theta) \in \mathbb{R}^k$, $L_{F[\partial\Omega]}(y, \theta) \geq 0$.*
2. ***Same link function:*** *$L_{\Omega \oplus \Omega^*}(y, \theta) = L_{F[\partial\Omega]}(y, \theta) = 0 \iff y = \widehat{y}_\Omega(\theta)$.*
3. ***Convexity:*** *$L_{F[\partial\Omega]}(y, \theta)$ is convex in y and θ separately.*
4. ***(Sub-)Gradient:*** *$\partial_\theta L_{F[\partial\Omega]}(y, \theta) = y^\star_{F[\partial\Omega]}(y, \theta) - y$ where $y^\star_{F[\partial\Omega]}(y, \theta)$ is given by (4).*
5. ***Tighter inequality:*** *for all $(y, \theta) \in \mathbb{R}^k$, $0 \leq L_{F[\partial\Omega]}(y, \theta) \leq L_{\Omega \oplus \Omega^*}(y, \theta)$.*

A proof is given in Appendix B.2. Because the Fitzpatrick loss and the Fenchel-Young loss generated by the same Ω have the same link function $\widehat{y}_\Omega$, they share the same minimizers w.r.t. θ for y fixed. However, the Fitzpatrick loss is always a **lower bound** of the corresponding Fenchel-Young loss. Moreover, they have different gradients w.r.t. θ: $\partial_\theta L_{\Omega \oplus \Omega^*}(y, \theta) = \widehat{y}_\Omega(\theta) - y$ vs. $\partial_\theta L_{F[\partial\Omega]}(y, \theta) = y^\star_{F[\partial\Omega]}(y, \theta) - y$. It is worth noticing that $y^\star_{F[\partial\Omega]}(y, \theta)$ depends on both y and θ, contrary to $\widehat{y}_\Omega(\theta)$.

When Ω is an unconstrained twice differential function on its domain (which is for instance the case of the squared 2-norm or the negentropy), we next show that Fitzpatrick losses enjoy a particularly simple expression and become a squared Mahalanobis-like distance.

Proposition 2 *Expressions of $F[\partial\Omega](y, \theta)$ and $L_{F[\partial\Omega]}(y, \theta)$ when Ω is twice differentiable*

Suppose Ω is twice differentiable. Then,

$$\begin{aligned} F[\partial\Omega](y, \theta) &= \langle y, \nabla\Omega(y^\star) \rangle + \langle y^\star, \theta \rangle - \langle y^\star, \nabla\Omega(y^\star) \rangle \\ L_{F[\partial\Omega]}(y, \theta) &= \langle y^\star - y, \theta - \nabla\Omega(y^\star) \rangle \\ &= \langle y^\star - y, \nabla^2\Omega(y^\star)(y^\star - y) \rangle \end{aligned}$$

where $y^\star = y^\star_{F[\partial\Omega]}(y, \theta)$ is the solution w.r.t. y' of

$$\nabla^2\Omega(y')(y' - y) = \theta - \nabla\Omega(y').$$

A proof is given in B.3. When Ω is constrained (i.e., when it contains an indicator function), we show in Section 3.5 that the above expression becomes a lower bound.

3.2 Examples

We now present the Fitzpatrick loss counterparts of various Fenchel-Young losses.

Squared loss.

Proposition 3 *Squared loss as a Fitzpatrick loss*

When $\Omega(y') = \frac{1}{2}\|y'\|_2^2$*, we have for all* $y \in \mathbb{R}^k$ *and* $\theta \in \mathbb{R}^k$

$$L_{F[\partial\Omega]}(y, \theta) = \frac{1}{4}\|y - \theta\|_2^2 = \frac{1}{2} L_{\text{squared}}(y, \theta).$$

A proof is given in Appendix B.4. Therefore, the Fenchel-Young and Fitzpatrick losses generated by Ω coincide, up to a factor $\frac{1}{2}$.

Perceptron loss.

Proposition 4 *Perceptron loss as a Fitzpatrick loss*

When $\Omega(y') = \iota_{\mathcal{C}}(y')$*, where* $\mathcal{C}$ *is a closed convex set, we have for all* $y \in \mathcal{C}$ *and* $\theta \in \mathbb{R}^k$

$$L_{F[\partial\Omega]}(y, \theta) = L_{\text{perceptron}}(y, \theta) = \max_{y' \in \mathcal{C}} \langle y', \theta \rangle - \langle y, \theta \rangle.$$

A proof is given in Appendix B.5. Therefore, the Fenchel-Young and Fitzpatrick losses generated by Ω exactly coincide in this case.

Fitzpatrick sparseMAP and Fitzpatrick sparsemax losses. As our first example where Fenchel-Young and Fitzpatrick losses substantially differ, we introduce the **Fitzpatrick sparseMAP** loss, which is the Fitzpatrick counterpart of the sparseMAP loss [22].

Proposition 5 *Fitzpatrick sparseMAP loss*

When $\Omega(y') = \frac{1}{2}\|y'\|_2^2 + \iota_{\mathcal{C}}(y')$*, where* $\mathcal{C}$ *is a closed convex set, we have for all* $y \in \mathcal{C}$ *and* $\theta \in \mathbb{R}^k$

$$L_{F[\partial\Omega]}(y, \theta) = 2\Omega^*\left((y + \theta)/2\right) - \langle y, \theta \rangle = \langle y^\star - y, \theta - y^\star \rangle$$

where we used $y^\star$ *as a shorthand for*

$$y^\star_{F[\partial\Omega]}(y, \theta) = \nabla\Omega^*((y + \theta)/2) = P_{\mathcal{C}}((y + \theta)/2).$$

A proof is given in Appendix B.6. As a special case, when $\mathcal{C} = \triangle^k$, we call the obtained loss the **Fitzpatrick sparsemax loss**, as it is the counterpart of the sparsemax loss [20]. Like the sparseMAP and sparsemax losses, these new losses rely on the Euclidean projection as a core building block. The Euclidean projection onto the probability simplex $\triangle^k$ can be computed exactly in $O(k)$ expected time and $O(k \log k)$ worst-case time [8, 21, 13, 11].

Fitzpatrick logistic loss. We now derive the Fitzpatrick couterpart of the logistic loss. Before stating the next proposition, we recall the definition of the Lambert W function [12]. For $z \geq 0$, $W(z)$ is the inverse of the function $f(w) = w \exp(w)$. That is, $W(z) = f^{-1}(z) = w$.

Proposition 6 *Fitzpatrick logistic loss*

When $\Omega(y') = \langle y', \log y' \rangle + \iota_{\triangle^k}(y')$*, we have for all* $y \in \triangle^k$ *and* $\theta \in \mathbb{R}^k$

$$L_{F[\partial\Omega]}(y, \theta) = \langle y^\star - y, \theta - \log y^\star - 1 \rangle$$

where we used $y^\star$ *as a shorthand for* $y^\star_{F[\partial\Omega]}(y, \theta)$ *defined by*

$$y^\star_{F[\partial\Omega]}(y, \theta)_i = \begin{cases} \mathrm{e}^{-\lambda^\star}\mathrm{e}^{\theta_i}, \text{ if } y_i = 0, \\ \frac{y_i}{W(y_i \mathrm{e}^{\lambda^\star - \theta_i})}, \text{ if } y_i > 0. \end{cases}$$

A proof and the value of $\lambda^\star = \lambda^\star(y, \theta) \in \mathbb{R}$ are given in Appendix B.7. To obtain $\lambda^\star(y, \theta)$, we need to solve a one-dimensional root equation, which can be done using for instance a bisection.

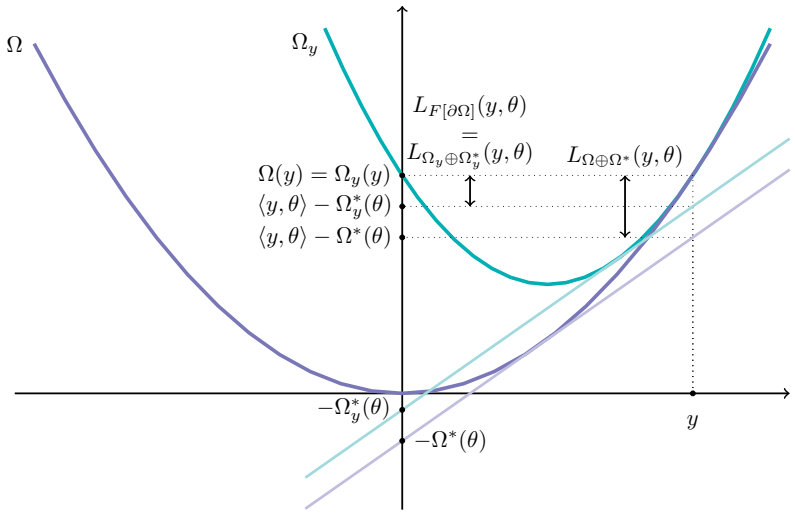

Figure 2: **Geometric interpretation**, with $\Omega(y') = \frac{1}{2}\|y'\|_2^2$. The Fenchel-Young loss $L_{\Omega \oplus \Omega^*}(y, \theta)$ is the gap (depicted with a double-headed arrow) between $\Omega(y)$ and $\langle y, \theta \rangle - \Omega^*(\theta)$, the value at y of the tangent with slope θ and intercept $-\Omega^*(\theta)$. As per Proposition 7, the Fitzpatrick loss $L_{F[\partial\Omega]}(y, \theta)$ is equal to $L_{\Omega_y \oplus \Omega_y^*}(y, \theta)$ and is therefore equal to the gap between $\Omega_y(y) = \Omega(y)$ and $\langle y, \theta \rangle - \Omega_y^*(\theta)$, the value at y of the tangent with slope θ and intercept $-\Omega_y^*(\theta)$. Since $\Omega_y(y') = \Omega(y') + D_\Omega(y, y')$, we have that $\Omega_y(y') \geq \Omega(y')$, with equality when $y = y'$. We therefore have $\Omega_y^*(\theta) \leq \Omega^*(\theta)$, implying that the Fitzpatrick loss is a lower bound of the Fenchel-Young loss.

3.3 Relation with Fenchel-Young losses

On first sight, Fitzpatrick losses and Fenchel-Young losses appear quite different. In the next proposition, we show that the Fitzpatrick loss generated by Ω is in fact equal to the Fenchel-Young loss generated by the modified, target-dependent function

$$\Omega_y(y') \coloneqq \Omega(y') + D_\Omega(y, y'),$$

where D_Ω is the generalized Bregman divergence defined in (1). In particular, Lemma 1 in the appendix shows that if $\Omega = \Psi + \iota_\mathcal{C}$, then $\Omega_y(y') = \Psi_y(y') + \iota_\mathcal{C}(y') = \Psi(y') + D_\Psi(y, y') + \iota_\mathcal{C}(y')$.

Proposition 7 *Characterization of $F[\partial\Omega]$, $L_{F[\partial\Omega]}$ and $y^\star_{F[\partial\Omega]}$ using Ω_y*

Let $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$ be a proper convex l.s.c. function. Then, for all $y \in \operatorname{dom} \Omega$ and all $\theta \in \mathbb{R}^k$,

$$\begin{aligned} F[\partial\Omega](y, \theta) &= \Omega_y(y) + \Omega_y^*(\theta) \\ L_{F[\partial\Omega]}(y, \theta) &= L_{\Omega_y \oplus \Omega_y^*}(y, \theta) \\ y^\star_{F[\partial\Omega]}(y, \theta) &= \widehat{y}_{\Omega_y}(\theta). \end{aligned}$$

This characterization of the Fitzpatrick function $F[\partial\Omega]$ is also new to our knowledge. A proof is given in Appendix B.8. Proposition 7 is very useful, as it means that Fitzpatrick losses inherit from all the known properties of Fenchel-Young losses, analyzed in prior works [7, 5]. In particular, Fenchel-Young losses are smooth (i.e., with Lipschitz gradients) when Ω is strongly convex. We therefore immediately obtain that Fitzpatrick losses are smooth if Ω is strongly convex and D_Ω is convex in its second argument, which is the case when $\Omega(y') = \frac{1}{2}\|y'\|_2^2$ and $\Omega(y') = \langle y', \log y' \rangle$. Therefore, the Fitzpatrick sparsemax and logistic losses are smooth. Proposition 7 also provides a mean to compute Fitzpatrick losses and their gradient. Finally, it suggests a very natural geometric interpretation of Fitzpatrick losses, as presented in Figure 2.

3.4 Relation with generalized Bregman divergences

As we stated before, the generalized Bregman divergence $D_\Omega(y, y')$ in (1) is a primal-primal divergence, as both y and y' belong to the same primal space. In contrast, Fenchel-Young losses

$L_{\Omega\oplus\Omega^*}(y,\theta)$ are primal-dual, since y belongs to the primal space and θ belongs to the dual space. Both can however be related, since

$$\begin{aligned} D_\Omega(y,y') &= \inf_{\theta'\in\partial\Omega(y')} L_{\Omega\oplus\Omega^*}(y,\theta') \\ &= \inf_{\theta'\in\partial\Omega(y')} \Omega(y)+\Omega^*(\theta')-\langle y,\theta'\rangle \\ &= \Omega(y)+\inf_{\theta'\in\partial\Omega(y')} \Omega^*(\theta')-\langle y,\theta'\rangle \\ &= \Omega(y)-\sup_{\theta'\in\partial\Omega(y')} -\Omega^*(\theta')+\langle y,\theta'\rangle \\ &= \Omega(y)-\Omega(y')-\sup_{\theta'\in\partial\Omega(y')} \langle y-y',\theta'\rangle, \end{aligned}$$

where in the last line we used that that $\Omega^*(\theta') = \langle y',\theta'\rangle - \Omega(y')$, as $\theta' \in \partial\Omega(y')$. This identity suggests that we can create Bregman-like primal-primal divergences by replacing $\Omega\oplus\Omega^*$ with $F[\partial\Omega]$,

$$\mathcal{D}_{F[\partial\Omega]}(y,y') \coloneqq \inf_{\theta'\in\partial\Omega(y')} L_{F[\partial\Omega]}(y,\theta') = \inf_{\theta'\in\partial\Omega(y')} F[\partial\Omega](y,\theta')-\langle y,\theta'\rangle.$$

This recovers one of the two Bregman-like divergences proposed in [9], the other one replacing the $\inf$ above by a $\sup$. As stated in [9], $F[\partial\Omega]$ and $\Omega\oplus\Omega^*$ are **representations** of $\partial\Omega$. More generally, Bregman divergences can be defined for any representation of the subdifferential $\partial\Omega$.

In order to use a primal-primal divergence as a loss, we need to explicitly compose it with a link function, such as $\widehat{y}_\Omega(\theta) = \nabla\Omega^*(\theta)$. Unfortunately, $D_\Omega(y,\widehat{y}_\Omega(\theta))$ or $\mathcal{D}_{F[\partial\Omega]}(y,\widehat{y}_\Omega(\theta))$ are typically **nonconvex** functions of θ, while Fenchel-Young and Fitzpatrick losses are always **convex**. In addition, differentiating through $\widehat{y}_\Omega(\theta)$ typically requires implicit differentiation [18, 4], while Fenchel-Young and Fitzpatrick losses enjoy easy-to-compute gradients, thanks to Danskin's theorem.

3.5 Lower bound on Fitzpatrick losses

If $\Omega = \Psi + \iota_\mathcal{C}$, where Ψ is a convex Legendre-type function and $\mathcal{C} \subseteq \operatorname{dom}\Psi$, then it was shown in [7, Proposition 3] that Fenchel-Young losses satisfy the lower bound

$$D_\Psi(y,\widehat{y}) \leq L_{\Omega\oplus\Omega^*}(y,\theta),$$

with equality if $\mathcal{C} = \operatorname{dom}\Psi$, where we used $\widehat{y}$ as a shorthand for $\widehat{y}_\Omega(\theta)$. We now show that a similar result holds for Fitzpatrick losses.

Proposition 8 *Lower bound on Fitzpatrick losses*

Let $\Omega = \Psi + \iota_\mathcal{C}$, where Ψ is a convex Legendre-type function and $\mathcal{C} \subseteq \operatorname{dom}\Psi$. Then,

$$D_{\Psi_y}(y,y^\star) = \langle y-y^\star, \nabla^2\Psi(y^\star)(y-y^\star)\rangle \leq L_{F[\partial\Omega]}(y,\theta),$$

with equality if $\operatorname{dom}\Psi = \mathcal{C}$, where we used $y^\star$ as a shorthand for $y^\star_{F[\partial\Omega]}(y,\theta)$.

A proof is given in Appendix B.9. If Ψ_y is μ-strongly convex, we obtain $\frac{\mu}{2}\|y-y^\star\|_2^2 \leq D_{\Psi_y}(y,y^\star)$.

4 Experiments

Experimental setup. We follow exactly the same experimental setup as in [6, 7]. We consider a dataset of n pairs (x_i,y_i) of feature vector $x_i \in \mathbb{R}^d$ and label proportions $y_i \in \triangle^k$, where d is the number of features and k is the number of classes. At inference time, given an unknown input vector $x \in \mathbb{R}^d$, our goal is to estimate a vector of label proportions $\widehat{y} \in \triangle^k$. A model is specified by a matrix $W \in \mathbb{R}^{k\times d}$ and a convex l.s.c. function $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$. Predictions are then produced by the generalized linear model $x \mapsto \widehat{y}_\Omega(Wx)$. At training time, we estimate the matrix $W \in \mathbb{R}^{k\times d}$ by minimizing the convex objective

$$R_{L,\lambda}(W) \coloneqq \sum_{i=1}^n L(y_i, Wx_i) + \frac{\lambda}{2}\|W\|_2^2, \tag{5}$$

Dataset	Sparsemax	Fitzpatrick-sparsemax	Logistic	Fitzpatrick-logistic
Birds	0.531	**0.513**	0.519	0.522
Cal500	0.035	0.035	0.034	0.034
Delicious	0.051	0.052	0.056	0.055
Ecthr A	0.514	0.514	0.431	**0.423**
Emotions	0.317	0.318	0.327	**0.320**
Flags	0.186	0.188	0.184	0.187
Mediamill	**0.191**	0.203	**0.207**	0.220
Scene	0.363	**0.355**	**0.344**	0.368
Tmc	0.151	0.152	0.161	0.160
Unfair	0.149	0.148	0.157	0.158
Yeast	0.186	0.187	0.183	0.185

Table 1: Test performance comparison between the sparsemax loss, the logistic loss and their Fitzpatrick counterparts on the task of label proportion estimation, with regularization parameter λ against the validation set. For each dataset, label proportion errors are measured using the mean squared error (MSE). We use bold if the error is at least 0.05 lower than its counterpart.

where $L \in \left\{L_{\Omega \oplus \Omega^*}, L_{F[\partial \Omega]}\right\}$. We focus on the (Fitzpatrick) sparsemax and the (Fitzpatrick) logistic losses. We optimize (5) using the L-BFGS algorithm [19]. The gradient of the Fenchel-Young loss is given in (2), while the gradient of the Fitzpatrick loss is given in Proposition 1, item 4. Experiments were conducted on a Intel Xeon E5-2667 clocked at 3.30GHz with 192 GB of RAM running on Linux. Our implementation relies on the SciPy [26] and scikit-learn [23] libraries.

We ran experiments on 11 standard multi-label benchmark datasets[1] (see Table 2 in Appendix A for statistics on the datasets). 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 the hyperparameter $\lambda \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$ against the validation set. We report test set mean squared error in Table 1.

Results. We found that the logistic loss and the Fitzpatrick logistic loss are comparable on most datasets, with the logistic loss significantly winning on 2 datasets and the Fitzpatrick logistic loss significantly winning on 2 datasets, out of 11. Since the Fitzpatrick logistic loss is slightly more computationally demanding, requiring to solve a root equation while the logistic loss does not, we believe that that the logistic loss remains the best choice when we wish to use the softargmax as link function $\widehat{y}_\Omega$.

Similarly, we found that the sparsemax loss and the Fitzpatrick sparsemax loss are comparable on most datasets, with the sparsemax loss significantly winning on only 1 dataset out 11 and the Fitzpatrick loss significantly winning on 2 datasets out of 11. Since the two losses both use the Euclidean projection onto the simplex $P_{\triangle^k}$ as their link function $\widehat{y}_\Omega$, we conclude that the Fitzpatrick sparsemax loss is a serious contender to the sparsemax loss, especially when predicting sparse label proportions is important.

5 Conclusion

We proposed to leverage the Fitzpatrick function, a theoretical tool from monotone operator theory, in order to build a new family of primal-dual convex loss functions for machine learning. We showed that Fitzpatrick losses are lower bounds of Fenchel-Young losses, while maintaining the same link function. Our paper therefore challenges the idea that there can only be one loss function associated with a certain link function. For instance, we created the Fitzpatrick logistic and sparsemax losses, that are associated with the soft argmax and sparse argmax links, traditionally associated with the logistic and sparsemax losses, respectively. We believe that even more loss functions with the same link can be created, which calls for a systematic study of their properties and respective benefits.

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

References

[1] S.-i. Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[2] H. Bauschke, D. McLaren, and H. Sendov. Fitzpatrick functions: Inequalities, examples, and remarks on a problem by S. Fitzpatrick. *Journal of Convex Analysis*, 13, 07 2005.

[3] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer-Verlag, New York, second edition, 2017.

[4] M. Blondel, Q. Berthet, M. Cuturi, R. Frostig, S. Hoyer, F. Llinares-Lopez, F. Pedregosa, and J.-P. Vert. Efficient and modular implicit differentiation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5230–5242. Curran Associates, Inc., 2022.

[5] M. Blondel, F. Llinares-López, R. Dadashi, L. Hussenot, and M. Geist. Learning energy networks with generalized Fenchel-Young losses. *Advances in Neural Information Processing Systems*, 35:12516–12528, 2022.

[6] M. Blondel, A. Martins, and V. Niculae. Learning classifiers with Fenchel-Young losses: Generalized entropies, margins, and algorithms. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 606–615. PMLR, 2019.

[7] M. Blondel, A. F. Martins, and V. Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

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

[9] R. S. Burachik and J. E. Martínez-Legaz. On Bregman-type distances for convex functions and maximally monotone operators. *Set-Valued and Variational Analysis*, 26:369–384, 2018.

[10] R. S. Burachik and B. F. Svaiter. Maximal monotone operators, convex functions and a special family of enlargements. *Set-Valued Analysis*, 10:297–316, 2002.

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

[12] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5(1):329–359, Dec 1996.

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

[14] S. Fitzpatrick. Representing monotone operators by convex functions. In *Workshop/Miniconference on Functional Analysis and Optimization*, volume 20, pages 59–66. Australian National University, Mathematical Sciences Institute, 1988.

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

[16] P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. 2004.

[17] K. C. Kiwiel. Proximal minimization methods with generalized Bregman functions. *SIAM journal on control and optimization*, 35(4):1142–1168, 1997.

[18] S. G. Krantz and H. R. Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002.

[19] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

[20] A. Martins and R. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623. PMLR, 2016.

[21] C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. *Journal of Optimization Theory and Applications*, 50(1):195–200, 1986.

[22] V. Niculae, A. Martins, M. Blondel, and C. Cardie. Sparsemap: Differentiable sparse structured inference. In *International Conference on Machine Learning*, pages 3799–3808. PMLR, 2018.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[24] M. D. Reid and R. C. Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422, 2010.

[25] E. K. Ryu and W. Yin. *Large-scale convex optimization: algorithms & analyses via monotone operators*. Cambridge University Press, 2022.

[26] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[27] R. C. Williamson, E. Vernet, and M. D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17(222):1–52, 2016.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: As stated in the abstract and in the introduction, we have defined Fitzpatrick losses in Definition 1 and studied their properties in Proposition 1. We instantiated the Fitzpatrick sparseMAP and the Fitzpatrick logistic loss in Proposition 5 and Proposition 6; we have studied the properties of Fitzpatrick losses in Proposition 1 and Proposition 7; we have gathered the results of the label proportion estimation in Table 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the computational limitations of the Fitzpatrick logistic loss. As stated in Proposition 6 and in Section 4, the computation of the loss value and gradient involves solving a root equation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All results in the paper are proved in the appendix. We strived to make the proofs self-contained.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Section 4, we describe the training setup, the linear model we use for predictions and give a link to access the datasets we use. We also indicate that we use the L-BFGS algorithm for training. Furthermore, Proposition 5 and Proposition 6 yield formulae for the gradients of the new losses we introduce.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provided a link for the used datasets in the experiments. The code will be opened upon release. For now, we provide instructions to reproduce the main experimental results. These instructions involved basic supervised learning tools such as L-BFGS algorithm, linear prediction model and standard cross-validation for the hyperparameter.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The selection of hyperparameter are discussed in section 4. We use predetermined train-test splits that come with the datasets.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The split between training sets and test sets are fixed in advance so there is no variability when conducting the label proportion estimation tests. Furthermore, as we minimize convex losses, the convergence of the minimization algorithm is independent of the starting point.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The characteristics of the computer we used to conduct the tests. On the largest dataset, our experiment only takes a couple of hours to run.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted does not involve human subjects. We did not create new datasets and used standard open access datasets for multiclassification. After reviewing the "societal impact and potential harmful consequences" from the Code of Ethics, we conclude that the research conduct in the paper does not pose a risk of harmful consequences.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The research conducted in this paper is theoretical and has no societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The datasets used in the tests are publicly available. No new prediction model has been released in this paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: In Section 4, the libraries used for the implementation of the numerical experiments are credited and cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Datasets 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	26
Delicious	Text	9682	3228	3181	500	983	19
Ecthr A	Text	6683	228	847	92401	10	1
Emotions	Music	293	98	202	72	6	2
Flags	Images	96	33	65	19	7	3
Mediamill	Video	22353	7451	12373	120	101	5
Scene	Images	908	303	1196	294	6	1
Tmc	Text	16139	5380	7077	48099	896	6
Unfair	Text	645	215	172	6290	8	1
Yeast	Micro-array	1125	375	917	103	14	4

Table 2: Datasets statistics

B Proofs

B.1 Lemmas

Lemma 1 *Generalized Bregman divergence for constrained Ω*

Let $\Omega = \Psi + \iota_{\mathcal{C}}$, where Ψ is convex differentiable and $\mathcal{C} \subseteq \operatorname{dom}\Psi$ such that $\operatorname{ri}\mathcal{C} \cap \operatorname{ri}\operatorname{dom}\Psi \neq \emptyset$, where $\operatorname{ri}\mathcal{C}$ is the relative interior of $\mathcal{C}$. Then, for all $y, y' \in \operatorname{dom}\Psi$

$$D_\Omega(y, y') = D_\Psi(y, y') + D_{\iota_{\mathcal{C}}}(y, y').$$

Proof. As $\mathcal{C}, \operatorname{dom}\Psi \subset \mathbb{R}^k$ and $\operatorname{ri}\mathcal{C} \cap \operatorname{ri}\operatorname{dom}\Psi \neq \emptyset$, we can apply [3, Proposition 6.19] and [3, Theorem 16.46] to write $\partial\Omega(y') = \partial\Psi(y') + N_{\mathcal{C}}(y')$.

Thus, we have

$$\theta' \in \partial\Omega(y') \iff \theta' - \nabla\Psi(y') \in N_{\mathcal{C}}(y') \iff \delta' \in N_{\mathcal{C}}(y'),$$

where

$$\delta' \coloneqq \theta' - \nabla\Psi(y') \iff \theta' \coloneqq \delta' + \nabla\Psi(y').$$

We then have

$$\begin{aligned}
D_\Omega(y, y') &\coloneqq \Omega(y) - \Omega(y') - \sup_{\theta' \in \partial\Omega(y')} \langle y - y', \theta' \rangle \\
&= \Omega(y) - \Omega(y') - \sup_{\delta' \in N_{\mathcal{C}}(y')} \langle y - y', \delta' + \nabla\Psi(y') \rangle \\
&= \Psi(y) + \iota_{\mathcal{C}}(y) - \Psi(y') - \iota_{\mathcal{C}}(y') - \langle y - y', \nabla\Psi(y') \rangle - \sup_{\delta' \in N_{\mathcal{C}}(y')} \langle y - y', \delta' \rangle \\
&= D_\Psi(y, y') + D_{\iota_{\mathcal{C}}}(y, y').
\end{aligned}$$

Lemma 2 *Generalized Bregman divergence of indicator function*

$$D_{\iota_{\mathcal{C}}}(y, y') = \begin{cases} \iota_{\mathcal{C}}(y) & \text{if } y' \in \mathcal{C} \\ \infty & \text{if } y' \notin \mathcal{C} \end{cases} = \iota_{\mathcal{C}}(y) + \iota_{\mathcal{C}}(y').$$

Proof.

$$D_{\iota_{\mathcal{C}}}(y, y') \coloneqq \iota_{\mathcal{C}}(y) - \iota_{\mathcal{C}}(y') - \sup_{\theta' \in N_{\mathcal{C}}(y')} \langle y - y', \theta' \rangle.$$

When $y' \in \mathcal{C}$ and $y \in \mathcal{C}$,

$$\sup_{\theta' \in N_{\mathcal{C}}(y')} \langle y - y', \theta' \rangle = \sup_{\substack{\theta' \in \mathbb{R}^k \\ \langle z - y', \theta' \rangle \leq 0 \\ \forall z \in \mathcal{C}}} \langle y - y', \theta' \rangle = 0.$$

When $y' \in \mathcal{C}$ and $y \notin \mathcal{C}$, $D_{\iota_\mathcal{C}}(y, y') = +\infty$, as $+\infty + (-\infty) = +\infty$ in the definition of the Bregman divergence. Therefore, when $y' \in \mathcal{C}$ $D_{\iota_\mathcal{C}}(y, y') = \iota_\mathcal{C}(y)$.

When $y' \notin \mathcal{C}$, $N_\mathcal{C}(y') = \emptyset$. Again, in the definition of the Bregman divergence, $+\infty + (-\infty) = +\infty$ and we use the convention $\sup_\emptyset = -\infty$.

Lemma 3 *Bregman divergence of Ψ_y*

Let Ψ be convex and twice differentiable. Let $\Psi_y(y') \coloneqq \Psi(y') + D_\Psi(y, y')$.

Then, for all $y, y', y'' \in \operatorname{dom} \Psi$,

$$\begin{aligned} D_{\Psi_y}(y', y'') &= D_\Psi(y, y') - D_\Psi(y, y'') + D(y', y'') + \langle y' - y'', \nabla^2\Psi(y'')(y - y'')\rangle \\ &= \langle y' - y, \nabla\Psi(y')\rangle - \langle y' - y, \nabla\Psi(y'')\rangle + \langle y' - y'', \nabla^2\Psi(y'')(y - y'')\rangle \end{aligned}$$

and in particular for all $y, y' \in \operatorname{dom} \Psi$

$$D_{\Psi_y}(y, y') = \langle y - y', \nabla^2\Psi(y')(y - y')\rangle.$$

Proof. For all $y, y' \in \operatorname{dom} \Psi$,

$$\begin{aligned} \Psi_y(y') &= \Psi(y') + D_\Psi(y, y') \\ &= \Psi(y') + \Psi(y) - \Psi(y') - \langle y - y', \nabla\Psi(y')\rangle \\ &= \Psi(y) - \langle y - y', \nabla\Psi(y')\rangle. \end{aligned}$$

and therefore

$$\begin{aligned} \nabla\Psi_y(y') &= -\nabla^2\Psi(y')y + \nabla\Psi(y') + \nabla^2\Psi(y')y' \\ &= \nabla^2\Psi(y')(y' - y) + \nabla\Psi(y'). \end{aligned}$$

Therefore, for all $y, y', y'' \in \operatorname{dom} \Psi$,

$$\begin{aligned} D_{\Psi_y}(y', y'') &= \Psi_y(y') - \Psi_y(y'') - \langle y' - y'', \nabla\Psi_y(y'')\rangle \\ &= \Psi(y') + D_\Psi(y, y') - \Psi(y'') - D_\Psi(y, y'') - \langle y' - y'', \nabla^2\Psi(y'')(y'' - y)\rangle - \langle y' - y'', \nabla\Psi(y'')\rangle \\ &= D_\Psi(y, y') - D_\Psi(y, y'') + D(y', y'') + \langle y' - y'', \nabla^2\Psi(y'')(y - y'')\rangle \\ &= \langle y' - y, \nabla\Psi(y')\rangle - \langle y' - y, \nabla\Psi(y'')\rangle + \langle y' - y'', \nabla^2\Psi(y'')(y - y'')\rangle \end{aligned}$$

and in particular, for all $y, y' \in \operatorname{dom} \Psi$,

$$\begin{aligned} D_{\Psi_y}(y, y') &= D_\Psi(y, y) - D_\Psi(y, y') + D_\Psi(y, y') + \langle y - y', \nabla^2\Psi_y(y')(y - y')\rangle \\ &= \langle y - y', \nabla^2\Psi(y')(y - y')\rangle. \end{aligned}$$

Lemma 4 *Generalized Bregman divergence of negentropy*

Let $\alpha \in \mathbb{R}$. Let $\Psi(y') \coloneqq \sum_{i=1}^k y'_i \log y'_i - \alpha \sum_{i=1}^k y'_i$ be defined for $y' \in \mathbb{R}^k_+$. Then, for $y, y' \in \mathbb{R}^k_+$,

$$D_\Psi(y, y') = \sum_{i=1}^k y_i \log \frac{y_i}{y'_i} - \sum_{i=1}^k (y_i - y'_i) + \iota_{\mathbb{R}^k_{++}}(y').$$

Proof. If $y' \in \mathbb{R}^k_{++}$, Ψ is differentiable at y' and $\nabla\Psi(y')_i = \log y_i + 1 - \alpha$. Thus, $\partial\Psi = \{\nabla\Psi\}$ and

$$\begin{aligned} D_\Psi(y, y') &= \Psi(y) - \Psi(y') - \sup_{\theta' \in \partial\Psi(y')} \langle y - y', \theta'\rangle, \\ &= \Psi(y) - \Psi(y') - \langle y - y', \nabla\Psi(y')\rangle, \\ &= \sum_{i=1}^k y_i \log \frac{y_i}{y'_i} - \sum_{i=1}^k (y_i - y'_i). \end{aligned}$$

If we prove that $\partial\Psi(y') = \emptyset$ when there is $y'_i = 0$, we can conclude the proof, as $\sup_\emptyset = -\infty$ by convention. Let us assume that $y'_i = 0$. Suppose that $\theta' \in \partial\Psi(y')$. Then, by definition of subgradients,

$$\langle y'' - y', \theta' \rangle + \Psi(y') \leq \Psi(y''),\ \forall y'' \in \mathbb{R}^k_{++}.$$

We choose $y'' = y' + \varepsilon e_i$, where $\varepsilon > 0$ and e_i is the i-th canonical base vector. Thus, we obtain

$$\begin{aligned}
\varepsilon\theta_i &\leq \Psi(y' + \varepsilon e_i) - \Psi(y'), \\
&= \sum_{j=1}^{k} y'_j \log y'_j + \varepsilon \log \varepsilon - \alpha \sum_{j=1}^{k} y'_j - \alpha\varepsilon - \big(\sum_{j=1}^{k} y'_j \log y'_j - \alpha \sum_{j=1}^{k} y'_j \big), \\
&= \varepsilon \log \varepsilon - \alpha\varepsilon,
\end{aligned}$$

as $y_i = 0$ and $0 \log 0 = 0$ by convention. By noticing that $\lim_{\varepsilon \to 0^+} \big(\varepsilon \log \varepsilon - \alpha\varepsilon\big)/\varepsilon = -\infty$, we get a contradiction, which concludes the proof.

Lemma 5 *Value and gradient of Ψ_y^**

Let $\Psi_y(y') \coloneqq \Psi(y') + D_\Psi(y, y')$, where Ψ is strictly convex and twice differentiable, $D_\Psi(y, y')$ is convex w.r.t. y' and $y \in \operatorname{dom} \Psi$. Then, for all $\theta \in \mathbb{R}^k$,

$$\begin{aligned}
\Psi_y^*(\theta) &= \langle \tilde{y}, \theta \rangle - \Psi(y) + \langle y - \tilde{y}, \nabla\Psi(\tilde{y}) \rangle \\
\nabla\Psi_y^*(\theta) &= \tilde{y}
\end{aligned}$$

where $\tilde{y}$ is the solution w.r.t. y' of

$$\begin{aligned}
&\operatorname*{argmax}_{y' \in \operatorname{dom}\Psi} \langle y', \theta \rangle + \langle y - y', \nabla\Psi(y') \rangle \\
&\iff \nabla^2\Psi(y')(y' - y) = \theta - \nabla\Psi(y').
\end{aligned}$$

Proof. As $\Psi \leq \Psi_y$, we have $\operatorname{dom}\Psi_y \subset \operatorname{dom}\Psi$. Thus, we get

$$\begin{aligned}
\Psi_y^*(\theta) &= \sup_{y' \in \operatorname{dom}\Psi} \langle y', \theta \rangle - \Psi_y(y') \\
&= \sup_{y' \in \operatorname{dom}\Psi} \langle y', \theta \rangle - (\Psi(y') + \Psi(y) - \Psi(y') - \langle y - y', \nabla\Psi(y') \rangle) \\
&= \sup_{y' \in \operatorname{dom}\Psi} \langle y', \theta \rangle - \Psi(y) + \langle y - y', \nabla\Psi(y') \rangle.
\end{aligned}$$

Using Danskin's theorem,

$$\begin{aligned}
\nabla\Psi_y^*(\theta) &= \operatorname*{argmax}_{y' \in \operatorname{dom}\Psi} \langle y', \theta \rangle - \Psi(y) + \langle y - y', \nabla\Psi(y') \rangle \\
&= \operatorname*{argmax}_{y' \in \operatorname{dom}\Psi} \langle y', \theta \rangle + \langle y - y', \nabla\Psi(y') \rangle.
\end{aligned}$$

Setting the gradient of the inner function to zero concludes the proof.

Lemma 6 *Gradient of Ψ_y^*, squared norm case*

Let $\Psi(y') \coloneqq \frac{1}{2}\|y'\|_2^2$. Then,

$$\nabla\Psi_y^*(\theta) = \frac{y + \theta}{2}.$$

Proof. Using Lemma 5 with $\nabla\Psi(y') = y'$ and $\nabla^2\Psi(y') = I$, we obtain that $\nabla\Psi_y^*(\theta)$ is the solution w.r.t. y' of $y' - y = \theta - y'$. Rearranging the terms concludes the proof.

Before stating the next lemma, we recall the definition of the Lambert W function [12]. For $z \geq 0$, $W(z)$ is the inverse of the function $f(w) = w \exp(w)$. That is, $W(z) = f^{-1}(z) = w$.

Lemma 7 *Gradient of Ψ_y^*, negentropy case*

Let $\Psi(y') \coloneqq \sum_{i=1}^k y'_i \log y'_i - \alpha \sum_{i=1}^k y'_i$ be defined for $y' \in \mathbb{R}_+^k$. Then,

$$\nabla \Psi_y^*(\theta)_i = \begin{cases} \mathrm{e}^{\theta_i - 2 + \alpha}, & \text{if } y_i = 0 \\ \frac{y_i}{W(y_i \mathrm{e}^{-(\theta_i - 2 + \alpha)})}, & \text{if } y_i > 0. \end{cases}$$

Proof. Using Lemma 5, we know that $\tilde{y}$ is the solution of $\nabla^2 \Psi(\tilde{y})(\tilde{y} - y) = \theta - \nabla \Psi(\tilde{y})$. Using $\nabla \Psi(\tilde{y}) = \log \tilde{y} + 1 - \alpha$ and $\nabla^2 \Psi(\tilde{y}) = 1/\tilde{y}$ (where logarithm and division are performed element-wise), we obtain for all $i \in [k]$

$$(\tilde{y}_i - y_i)/\tilde{y}_i = \theta_i - \log \tilde{y}_i - 1 + \alpha \iff 1 - y_i/\tilde{y}_i = \theta_i - \log \tilde{y}_i - 1 + \alpha.$$

When $y_i = 0$, we immediatly have $\tilde{y}_i = \exp(\theta_i - 2 + \alpha)$. When $y_i > 0$, after rearranging, we obtain

$$\frac{y_i}{\tilde{y}_i} \exp\left(\frac{y_i}{\tilde{y}_i}\right) = y_i \exp(-(\theta_i - 2 + \alpha)) \iff \frac{y_i}{\tilde{y}_i} = W(y_i \exp(-(\theta_i - 2 + \alpha))),$$

hence the result.

Lemma 8 *Gradient of Ω_y^**

Let Ψ be a strictly convex function such that $D_\Psi(y, y')$ is convex w.r.t. y'. Let $\Omega_y(y') \coloneqq \Psi_y(y') + \iota_{\mathcal{C}}(y') = \Psi(y') + D_\Psi(y, y') + \iota_{\mathcal{C}}(y')$, where $\mathcal{C} \subseteq \operatorname{dom} \Psi$ is closed convex. Then,

$$\nabla \Omega_y^*(\theta) = y^\star$$

where $y^\star$ is the solution w.r.t. y' of

$$\operatorname*{argmax}_{y' \in \mathcal{C}} \ \langle y', \theta \rangle + \langle y - y', \nabla \Psi(y') \rangle.$$

Proof. The result again follows from Danskin's theorem.

Lemma 9 *Dual of simplex-constrained conjugate*

If Ψ is strictly convex with $\mathbb{R}_+^k \subseteq \operatorname{dom} \Psi$, then,

$$(\Psi + \iota_{\triangle^k})^*(\theta) = \min_{\tau \in \mathbb{R}} \tau + (\Psi + \iota_{\mathbb{R}_+^k})^*(\theta - \tau \mathbf{1}).$$

and

$$\nabla (\Psi + \iota_{\triangle^k})^*(\theta) = \nabla (\Psi + \iota_{\mathbb{R}_+^k})^*(\theta - \tau^\star \mathbf{1}),$$

where $\tau^\star$ denotes the optimal dual variable.

Proof.

$$\begin{aligned}
(\Psi + \iota_{\triangle^k})^*(\theta) &= \max_{y' \in \triangle^k} \langle y', \theta \rangle - \Psi(y') \\
&= \max_{y' \in \mathbb{R}_+^k} \min_{\tau \in \mathbb{R}} \langle y', \theta \rangle - \Psi(y') - \tau(\langle y', \mathbf{1} \rangle - 1) \\
&= \min_{\tau \in \mathbb{R}} \tau + \max_{y' \in \mathbb{R}_+^k} \langle y', \theta - \tau \mathbf{1} \rangle - \Psi(y') \\
&= \min_{\tau \in \mathbb{R}} \tau + (\Psi + \iota_{\mathbb{R}_+^k})^*(\theta - \tau \mathbf{1}),
\end{aligned}$$

where we used that, as the constraints of belonging to the simplex are affine, they are qualified and we can invert the $\max$ and the $\min$.

We use the strict convexity of Ψ and Danskin's theorem to show that $(\Psi + \iota_{\triangle^k})^*$ is differentiable. Then, we use the converse of Danskin's theorem to conclude.

Lemma 10 *Gradient of Ω_y^*, negentropy, constrained to the simplex*

Let $\Omega = \Psi + \iota_{\triangle^k}$, where $\Psi(y') = \langle y', \log y' \rangle$. Then,

$$y_i^\star = \nabla \Omega_y^*(\theta)_i = \begin{cases} \mathrm{e}^{-\lambda^\star} \mathrm{e}^{\theta_i}, & \text{if } y_i = 0, \\ \frac{y_i}{W(y_i \mathrm{e}^{\lambda^\star - \theta_i})}, & \text{if } y_i > 0. \end{cases}$$

where $\lambda^\star$ is the solution of

$$\mathrm{e}^{-\lambda^\star} \sum_{i:y_i=0} \mathrm{e}^{\theta_i} + \sum_{i:y_i>0} \frac{y_i}{W(y_i \mathrm{e}^{-(\theta_i - \lambda^\star)})} = 1.$$

Proof. From Lemma 8 and Lemma 9, since $\operatorname{dom} \Psi_y = \mathbb{R}_+^k$, we have

$$y^\star = \nabla \Omega_y^*(\theta) = \nabla \Psi_y^*(\theta - \tau^\star \mathbf{1})$$

where $\tau^\star$ is the solution of

$$\min_{\tau \in \mathbb{R}} \tau + \Psi_y^*(\theta - \tau \mathbf{1}).$$

Setting the gradient of the inner function to zero, we get

$$\langle \nabla \Psi_y^*(\theta - \tau^\star \mathbf{1}), \mathbf{1} \rangle = 1.$$

Using Lemma 7, we obtain that $\tau^\star$ satisfies

$$\mathrm{e}^{-\tau^\star - 2} \sum_{i:y_i=0} \mathrm{e}^{\theta_i} + \sum_{i:y_i>0} \frac{y_i}{W(y_i \mathrm{e}^{-(\theta_i - \tau^\star - 2)})} = 1.$$

Using the change of variable $\tau^\star = \lambda^\star + 2$ concludes the proof.

B.2 Proof of Proposition 1 (Properties of Fitzpatrick losses)

Apart from differentiability, the proofs follow from the study of Fitzpatrick functions found in [14, 2, 25]. We include the proofs for completeness.

Link function and non-negativity. We recall that

$$\begin{aligned} L_{F[\partial\Omega]}(y, \theta) &= \sup_{(y', \theta') \in \partial\Omega} \langle y' - y, \theta - \theta' \rangle \\ &= - \inf_{(y', \theta') \in \partial\Omega} \langle y' - y, \theta' - \theta \rangle. \end{aligned}$$

From the monotonicity of $\partial\Omega$, we have that if $(y, \theta) \in \partial\Omega$ and $(y', \theta') \in \partial\Omega$, then $\langle y' - y, \theta' - \theta \rangle \geq 0$. Therefore, for all $(y, \theta) \in \partial\Omega$,

$$\inf_{(y', \theta') \in \partial\Omega} \langle y' - y, \theta' - \theta \rangle = 0,$$

with the infimum being attained at $(y', \theta') = (y, \theta)$. This proves the link function.

From the maximality of $\partial\Omega$, if $(y, \theta) \notin \partial\Omega$, there exists $(y', \theta') \in \partial\Omega$ such that $\langle y' - y, \theta' - \theta \rangle < 0$. Therefore, for all $(y, \theta) \notin \partial\Omega$,

$$\inf_{(y', \theta') \in \partial\Omega} \langle y' - y, \theta' - \theta \rangle < 0.$$

This proves the non-negativity.

Convexity. We recall that

$$L_{F[\partial\Omega]}(y, \theta) = F[\partial\Omega](y, \theta) - \langle y, \theta \rangle$$

where

$$F[\partial\Omega](y, \theta) = \sup_{(y', \theta') \in \partial\Omega} \langle y - y', \theta' \rangle + \langle y', \theta \rangle = \sup_{(y', \theta') \in \partial\Omega} \langle y', \theta \rangle + \langle y, \theta' \rangle - \langle y', \theta' \rangle.$$

The function $(y, \theta) \mapsto \langle y', \theta \rangle + \langle y, \theta' \rangle - \langle y', \theta' \rangle$ is jointly convex in (y, θ) for all (y', θ'). Since the supremum preserves convexity, $F[\partial\Omega](y, \theta)$ is jointly convex in (y, θ). The function $\langle y, \theta \rangle$ is separately convex / concave in y and θ but not jointly convex / concave in (y, θ). Therefore, $L_{F[\partial\Omega]}(y, \theta)$ is separately convex in y and θ.

Differentiability. Since $\Omega(y')$ is strictly convex and $y' \mapsto D_\Omega(y, y')$ is convex, $\Omega_y(y') = \Omega(y') + D_\Omega(y, y')$ is strictly convex in y'. From the duality between strict convexity and differentiability, $\Omega_y^*(\theta)$ is differentiable in θ.

Tighter inequality. Using

$$\partial\Omega = \{(y', \theta') \colon \Omega(y) \geq \Omega(y') + \langle y - y', \theta' \rangle \; \forall y\}$$

and

$$\Omega^*(\theta) = \sup_{y' \in \mathbb{R}^k} \langle y', \theta \rangle - \Omega(y'),$$

we get for any $(y', \theta') \in \partial\Omega$,

$$\begin{aligned} \langle y - y', \theta' \rangle + \langle y', \theta \rangle &\leq \Omega(y) - \Omega(y') + \langle y', \theta \rangle \\ &\leq \Omega(y) + \Omega^*(\theta). \end{aligned}$$

Therefore

$$F[\partial\Omega](y, \theta) = \sup_{(y', \theta') \in \partial\Omega} \langle y - y', \theta' \rangle + \langle y', \theta \rangle \leq \Omega(y) + \Omega^*(\theta).$$

B.3 Proof of Proposition 2 (Expression of Fitzpatrick loss when Ω is twice differentiable)

We recall that

$$F[\partial\Omega](y, \theta) = \sup_{(y', \theta') \in \partial\Omega} \langle y, \theta' \rangle + \langle y', \theta \rangle - \langle y', \theta' \rangle = \sup_{y' \in \operatorname{dom} \Omega} \langle y', \theta \rangle + \sup_{\theta' \in \partial\Omega(y')} \langle y, \theta' \rangle - \langle y', \theta' \rangle.$$

Since Ω is differentiable, we have $\partial\Omega(y') = \{\nabla\Omega(y')\}$ and therefore $\theta' = \nabla\Omega(y')$, which gives

$$F[\partial\Omega](y, \theta) = \sup_{y' \in \mathbb{R}^k} \langle y, \nabla\Omega(y') \rangle + \langle y', \theta \rangle - \langle y', \nabla\Omega(y') \rangle.$$

Setting the gradient of the inner function w.r.t. y' to zero, we get

$$\nabla^2\Omega(y')y + \theta - \nabla\Omega(y') - \nabla^2\Omega(y')y' = 0.$$

Using the $y' = y^\star$ and $\theta' = \nabla\Omega(y')$ in 1, we then obtain

$$\begin{aligned} L_{F[\partial\Omega]}(y, \theta) &= \langle y' - y, \theta - \theta' \rangle \\ &= \langle y' - y, \theta - \nabla\Omega(y') \rangle \\ &= \langle y' - y, \nabla^2\Omega(y')(y' - y) \rangle. \end{aligned}$$

B.4 Proof of Proposition 3 (squared loss)

Using Proposition 2 with $\nabla\Omega(y') = y'$ and $\nabla^2\Omega(y') = I$, we obtain

$$y + \theta - 2y' = 0 \iff y' = \frac{y + \theta}{2}.$$

We therefore obtain

$$\begin{aligned} L_{F[\partial\Omega]}(y, \theta) &= \left\langle \frac{y + \theta}{2} - y, \theta - \frac{y + \theta}{2} \right\rangle \\ &= \left\langle \frac{\theta - y}{2}, \frac{\theta - y}{2} \right\rangle \\ &= \frac{1}{4} \|y - \theta\|_2^2. \end{aligned}$$

B.5 Proof of Proposition 4 (perceptron loss)

A proof of the Fitzpatrick function for this case was given in [2, Example 3.1]. We include a proof for completeness. Since $\Omega = \iota_{\mathcal{C}}$, we have $\partial\Omega = N_{\mathcal{C}}$ and $\operatorname{dom}\Omega = \mathcal{C}$. Therefore, for all $y \in \mathcal{C}$ and $\theta \in \mathbb{R}^k$,

$$\begin{aligned}
F[\partial\Omega](y,\theta) &= \sup_{y' \in \operatorname{dom}\Omega} \langle y', \theta\rangle + \sup_{\theta' \in \partial\Omega(y')} \langle y - y', \theta'\rangle \\
&= \sup_{y' \in \mathcal{C}} \langle y', \theta\rangle + \sup_{\theta' \in N_{\mathcal{C}}(y')} \langle y - y', \theta'\rangle \\
&= \sup_{y' \in \mathcal{C}} \langle y', \theta\rangle - \left(\iota_{\mathcal{C}}(y) - \iota_{\mathcal{C}}(y') - \sup_{\theta' \in N_{\mathcal{C}}(y')} \langle y - y', \theta'\rangle\right) \\
&= \sup_{y' \in \mathcal{C}} \langle y', \theta\rangle - D_{\iota_{\mathcal{C}}}(y, y') \\
&= \sup_{y' \in \mathcal{C}} \langle y', \theta\rangle,
\end{aligned}$$

where in the third line we used that $\iota_{\mathcal{C}}(y) = \iota_{\mathcal{C}}(y') = 0$ and where in the last line we used Lemma 2. Therefore, for all $y \in \mathbb{R}^k$ and $\theta \in \mathbb{R}^k$,

$$F[\partial\Omega](y,\theta) = \sup_{y' \in \mathcal{C}} \langle y', \theta\rangle + \iota_{\mathcal{C}}(y) = \iota_{\mathcal{C}}(y) + \iota^*_{\mathcal{C}}(\theta).$$

B.6 Proof of Proposition 5 (Fitzpatrick sparseMAP loss)

A proof of the Fitzpatrick function for this case was given in [2, Example 3.13]. We provide an alternative proof.

From Proposition 7, we know that

$$F[\partial\Omega](y,\theta) = \Omega_y(y) + \Omega^*_y(\theta) = \Omega(y) + \Omega^*_y(\theta),$$

where

$$\begin{aligned}
\Omega_y(y') &= \frac{1}{2}\|y'\|_2^2 + \frac{1}{2}\|y - y'\|_2^2 + \iota_{\mathcal{C}}(y') \\
&= \|y'\|_2^2 + \frac{1}{2}\|y\|_2^2 - \langle y, y'\rangle + \iota_{\mathcal{C}}(y') \\
&= 2\Omega(y') + \Omega(y) - \langle y, y'\rangle.
\end{aligned}$$

Using conjugate calculus, we obtain

$$\Omega^*_y(\theta) = 2\Omega^*\left(\frac{y+\theta}{2}\right) - \Omega(y).$$

Therefore,

$$F[\partial\Omega](y,\theta) = 2\Omega^*\left(\frac{y+\theta}{2}\right).$$

From Proposition 7, the supremum w.r.t. y' is achieved at $y^\star = \nabla\Omega^*((y+\theta)/2) = P_{\mathcal{C}}((y+\theta)/2)$. We therefore obtain

$$L_{F[\partial\Omega]}(y,\theta) = \langle y^\star - y, \theta - y^\star\rangle.$$

B.7 Proof of Proposition 6 (Fitzpatrick logistic loss)

Differentiability w.r.t. θ and formula of gradient. According to Proposition 7, we have

$$L_{F[\partial\Omega]}(y,\theta) = \Omega_y(y) + \Omega^*_y(\theta) - \langle y, \theta\rangle.$$

Thus the differentiability w.r.t. θ of $L_{F[\partial\Omega]}(y,\theta)$ follows from the differentiability of $\Omega^*_y(\theta)$. Lemma 10 yields the differentiability of $\Omega^*_y(\theta)$ and a formula for its gradient $y^\star_{F[\partial\Omega]}(y,\theta) \coloneqq \nabla\Omega^*_y(\theta)$.

$$y^\star_{F[\partial\Omega]}(y,\theta)_i = \begin{cases} \mathrm{e}^{-\lambda^\star}\mathrm{e}^{\theta_i}, \text{ if } y_i = 0, \\ \frac{y_i}{W(y_i \mathrm{e}^{\lambda^\star - \theta_i})}, \text{ if } y_i > 0. \end{cases}$$

It follows that $\nabla_\theta L_{F[\partial\Omega]}(y,\theta) = y^\star_{F[\partial\Omega]}(y,\theta) - y$.

Formula of the Fitzpatrick logistic loss. We use $y^\star$ as a shorthand for $y^\star_{F[\partial\Omega]}(y,\theta)$. As we know that $\Omega_y^*(\theta) = \langle y^\star, \theta\rangle - \Omega_y(y^\star)$, we use again Proposition 7 to get

$$\begin{aligned} L_{F[\partial\Omega]}(y,\theta) &= \Omega_y(y) + \langle y^\star, \theta\rangle - \Omega_y(y^\star) - \langle y, \theta\rangle \\ &= \Omega(y) - (\Omega(y^\star) + D_\Omega(y, y^\star)) + \langle y^\star - y, \theta\rangle \end{aligned}$$

as $\Omega_y(y') = \Omega(y) + D_\Omega(y, y')$ and in particular $\Omega_y(y) = \Omega(y)$. Furthermore, as $y^\star \in \triangle^k \cap \mathbb{R}^k_{++}$, Ω is differentiable at $y^\star$ and $D_\Omega(y, y^\star) = \Omega(y) - \Omega(y^\star) - \langle y - y^\star, \nabla\Omega(y^\star)\rangle$, where $\nabla\Omega(y^\star) = \log y^\star + 1$. Thus

$$\begin{aligned} L_{F[\partial\Omega]}(y,\theta) &= \langle y - y^\star, \nabla\Omega(y^\star)\rangle + \langle y^\star - y, \theta\rangle \\ &= \langle y^\star - y, \theta - \log y^\star - 1\rangle. \end{aligned}$$

Bisection formula for $\lambda^\star$ and bounds. We also get from Lemma 10 a bisection formula for $\lambda^\star$, which is a shorthand for $\lambda^\star_{F[\partial\Omega]}(y,\theta)$.

$$\mathrm{e}^{-\lambda^\star} \sum_{i:y_i=0} \mathrm{e}^{\theta_i} + \sum_{i:y_i>0} \frac{y_i}{W(y_i \mathrm{e}^{-(\theta_i - \lambda^\star)})} = 1.$$

We focus here on a lower bound and an upper bound for $\lambda^\star \in \mathbb{R}$. Let us prove that

$$\log \sum_{i=1}^k \mathrm{e}^{\theta_i} \le \lambda^\star \le \log 2 + \max\Big\{ \log \sum_{i:y_i=0}^k \mathrm{e}^{\theta_i}, \log \ell_0(y) + \max_{i:y_i>0} \theta_i + 2\ell_0(y) y_i \Big\},$$

where $\ell_0(y) = \mathrm{Card}(j : y_j \neq 0)$.

For the lower bound, we use the concavity of the Lambert function W, which implies $\frac{1}{W(y_i \mathrm{e}^{\lambda^\star - \theta_i})} \ge \frac{1}{y_i \mathrm{e}^{\lambda^\star - \theta_i}}$. Thus,

$$1 \ge \mathrm{e}^{-\lambda^\star} \sum_{i:y_i=0} \mathrm{e}^{\theta_i} + \sum_{i:y_i>0} \frac{y_i}{y_i \mathrm{e}^{\lambda^\star - \theta_i}},$$

which in turn implies

$$\mathrm{e}^{\lambda^\star} \ge \sum_{i:y_i=0} \mathrm{e}^{\theta_i} + \sum_{i:y_i>0} \mathrm{e}^{\theta_i}$$

and yields the lower bound.

For the upper bound, the function $g(\lambda) = \mathrm{e}^{-\lambda} \sum_{i:y_i=1} \mathrm{e}^{\theta_i} + \sum_{i:y_i>0} \frac{y_i}{W(y_i \mathrm{e}^{\lambda - \theta_i})}$ is continuous and decreasing (as it is a positive combination of decreasing functions) and $g(-\infty) = +\infty$. Thus if we find a λ such that $g(\lambda) < 1$, we know that $\lambda^\star \le \lambda$.

We deal with each term of $g(\lambda)$ separately. If $\lambda \in \mathbb{R}$ satisfies

$$\begin{aligned} \mathrm{e}^{-\lambda} \sum_{i:y_i=0} \mathrm{e}^{\theta_i} &\le \frac{1}{2} \\ \max_{i:y_i>0} \frac{y_i}{W(y_i \mathrm{e}^{\lambda - \theta_i})} &\le \frac{1}{2\ell_0(y)}, \end{aligned}$$

then

$$g(\lambda) = \underbrace{\mathrm{e}^{-\lambda} \sum_{i:y_i=0} \mathrm{e}^{\theta_i}}_{\le 1/2} + \sum_{i:y_i>0} \underbrace{\frac{y_i}{W(y_i \mathrm{e}^{\lambda - \theta_i})}}_{\le 1/\left(2\ell_0(y)\right)} \le 1.$$

Thus, all λ satisfying the following inequalties are upper bounds of $\lambda^\star$

$$\begin{aligned} 2 \sum_{i:y_i=0} \mathrm{e}^{\theta_i} &\le \mathrm{e}^{\lambda} \\ 2\ell_0(y) y_i &\le W(y_i \mathrm{e}^{\lambda - \theta_i}), \forall i : y_i > 0. \end{aligned}$$

As W is monotone and $W^{-1}(t) = te^t$, we get

$$\log 2 + \log \sum_{i:y_i=0} \mathrm{e}^{\theta_i} \leq \lambda$$
$$2\ell_0(y)\mathrm{e}^{2\ell_0(y)y_i} \leq \mathrm{e}^{\lambda^\star - \theta_i}, \forall i : y_i > 0.$$

Thus taking $\lambda = \max\left\{ \log 2 + \log \sum_{i:y_i=0}^{k} \mathrm{e}^{\theta_i}, \max_{i:y_i>0} \log 2 + \log \ell_0(y) + \theta_i + 2\ell_0(y)y_i \right\}$ yields an upper bound of $\lambda^\star$.

B.8 Proof of Proposition 7 (characterization of $F[\partial\Omega]$ using D_Ω)

Let $(y, \theta) \in \operatorname{dom}\Omega \times \mathbb{R}^k$. We have

$$\begin{aligned}
F[\partial\Omega](y,\theta) &= \sup_{(y',\theta')\in\partial\Omega} \langle y - y', \theta'\rangle + \langle y', \theta\rangle \\
&= \sup_{y'\in\operatorname{dom}\Omega} \left\{ \langle y', \theta\rangle + \sup_{\theta'\in\partial\Omega(y')} \langle y - y', \theta'\rangle \right\} \\
&= \sup_{y'\in\operatorname{dom}\Omega} \left\{ \langle y', \theta\rangle - \Omega(y') + \Omega(y') + \sup_{\theta'\in\partial\Omega(y')} \langle y - y', \theta'\rangle \right\} \\
&= \Omega(y) + \sup_{y'\in\operatorname{dom}\Omega} \langle y', \theta\rangle - \left(\Omega(y') + \Omega(y) - \Omega(y') - \sup_{\theta'\in\partial\Omega(y')} \langle y - y', \theta'\rangle \right) \\
&= \Omega(y) + \sup_{y'\in\operatorname{dom}\Omega} \langle y', \theta\rangle - (\Omega(y') + D_\Omega(y, y')) \\
&= \Omega(y) + (\Omega + D_\Omega(y,\cdot))^*(\theta) \\
&= \Omega_y(y) + \Omega_y^*(\theta).
\end{aligned}$$

The supremum above is achieved at $y' \in \partial\Omega_y^*(\theta) = y^\star_{F[\partial\Omega]}(y,\theta)$.

When $\Omega = \Psi + \iota_{\mathcal{C}}$, where $\mathcal{C} \subseteq \operatorname{dom}\Psi$, using Lemma 1 and 2, we have for all $y \in \mathcal{C}$

$$\Omega_y(y') = \Psi(y') + D_\Psi(y, y') + \iota_{\mathcal{C}}(y').$$

B.9 Proof of Proposition 8 (lower bound)

It was shown in [7, Proposition 3] that if $f = g + \iota_{\mathcal{C}}$, where g is Legendre type with $\mathcal{C} \subseteq \operatorname{dom}\Psi$, then for all $y \in \mathcal{C}$ and $\theta \in \mathbb{R}^k$,

$$0 \leq D_g(y, \nabla f^*(\theta)) \leq L_{f\oplus f^*}(y,\theta),$$

with equality if $\mathcal{C} = \operatorname{dom} g$. Using $g = \Psi_y$, $f = \Omega_y = \Psi_y + \iota_{\mathcal{C}}$, $y^\star = \nabla\Omega_y^*(\theta) = y^\star_{F[\partial\Omega]}(y,\theta)$, and Lemma 3, we therefore obtain

$$D_{\Psi_y}(y, y^\star) = \langle y - y^\star, \nabla^2\Psi(y^\star)(y - y^\star)\rangle \leq L_{\Omega_y\oplus\Omega_y^*}(y,\theta) = L_{F[\partial\Omega]}(y,\theta).$$

If Ψ_y is μ-strongly convex and D_Ψ is convex in its second argument, then Ψ_y is μ-strongly convex as well. Therefore, we also have

$$\frac{\mu}{2}\|y - y^\star\|_2^2 \leq D_{\Psi_y}(y, y^\star).$$