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

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; yet their predictions are produced using the same link function. As an example, we introduce the Fitzpatrick logistic loss and the Fitzpatrick sparsemax loss, counterparts of the logistic and the sparsemax losses, two instances of Fenchel-Young losses. This allows us to obtain two new tighter losses associated with the soft argmax and the sparse argmax, two of the most popular 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 probabilistic classification.

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) [12, 11] 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 primal space. They are typically explicitly composed with a link function [17, 19], 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 [5, 6], which include 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. The key difference with proper losses is that the link function, mapping the dual-space model output to a primal-space prediction, is implicit. This crucial difference makes Fenchel-Young losses always convex.

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 [10, 8, 2]. So far, the Fitzpatrick function had been used as a theoretical tool to represent maximal monotone operators [18] and to construct Bregman-like primal-primal divergences [7], 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

Preprint. Last update: May 7, 2024.

same link function, suggesting that we can use Fitzpatrick losses as a tighter replacement for the corresponding Fenchel-Young losses. 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 details the properties of Fitzpatrick losses. In particular, 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 seven datasets.

2 Background

2.1 Convex analysis

We denote the probability simplex by $\triangle^k \coloneqq \{p \in \mathbb{R}_+^k \colon \sum_{i=1}^k p_i = 1\}$. We denote 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 Ω by $\operatorname{dom} \Omega \coloneqq \{y \in \mathbb{R}^k : f(y) < +\infty\}$. We denote the Euclidean projection onto a convex set $\mathcal{C}$ by $P_{\mathcal{C}}(\theta) = \operatorname{argmin}_{y \in \mathcal{C}} \|y - \theta\|_2^2$.

For a 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 $\mathcal{C}$ 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

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

We define the **generalized Bregman divergence** [13] $D_\Omega : \mathbb{R}^k \times \mathbb{R}^k \to \overline{\mathbb{R}}_+$ generated by a convex l.s.c. function $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$ by

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

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. Ω is defined by [6]

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

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, 6], where y belongs to the primal space and θ belongs to the dual space.

Link functions

To map the dual-space variable θ to the primal space y, we can use the canonical link function $\partial\Omega^*$, as it satisfies

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

In particular when Ω is strictly convex, the Fenchel-Young loss is positive definite, meaning that

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

Since Ω^* is convex, its gradient $\nabla\Omega^*$ is monotone. As shown in [6], the monotonicity implies that θ and $y \in \partial\Omega^*(\theta)$ are sorted the same way, i.e., $\theta_i > \theta_j \implies y_i \geq y_j$. Link functions also play an important role in the loss gradient, as we have

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

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

We give a few examples of instance 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**

$$\nabla\Omega^*(\theta) = \theta.$$

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**

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

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** [16]

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

In particular, when the set is $\mathcal{C} = \triangle^k$, we obtain the **sparsemax loss** [15]. The link is the **Euclidean projection** onto $\mathcal{C}$,

$$\nabla\Omega^*(\theta) \coloneqq P_{\mathcal{C}}(\theta).$$

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**

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

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, 18].

A well-known object in monotone operator theory, the **Fitzpatrick function** associated with a monotone operator A [10, 8, 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 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, we also have

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

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.

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 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:*** *$L_{\Omega \oplus \Omega^*}(y, \theta) = L_{F[\partial\Omega]}(y, \theta) = 0 \iff y \in \partial\Omega^*(\theta)$.*
3. ***Convexity:*** *$L_{F[\partial\Omega]}(y, \theta)$ is convex in y and θ separately.*
4. ***(Sub-)Gradient:*** *$\partial_\theta L_{F[\partial\Omega]}(y, \theta) = y^\star_{F[\partial\Omega]}(y, \theta) - y$.*
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 A.1.

Because the Fitzpatrick loss and the Fenchel-Young loss generated by the same Ω have the same link function, they share the same minimizers. However, the fact that the Fitzpatrick loss is a **lower bound** suggests that it can be used to make the link function predict the correct target faster than with the corresponding Fenchel-Young loss. As our notation suggests, $\Omega \oplus \Omega^*$ uses **decoupled** functions Ω and Ω^* w.r.t. y and θ. In contrast, $F[\partial\Omega]$ is a **coupled** function w.r.t. y and θ.

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 Appendix A.2. This expression shows that the Fitzpatrick loss behaves locally like a squared Mahalanobis distance.

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 A.3. 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 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 A.4. 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 [16].

Proposition 5 *Fitzpatrick sparseMAP loss*

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

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

and the loss is smooth w.r.t. θ with gradient

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

where 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 A.5. 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 [15].

In constrast, the gradient of the original sparsemax loss is $\nabla_\theta L_{\Omega+\Omega^*}(y,\theta) = y' - y$, where $y' = \nabla\Omega^*(\theta) = P_{\mathcal{C}}(\theta)$.

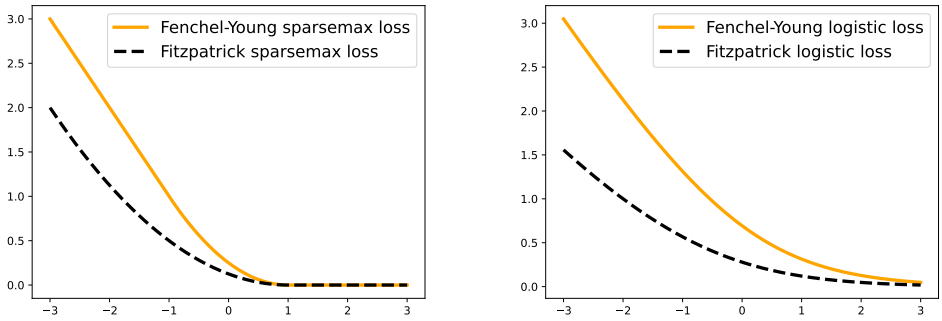

Figure 1: **Left:** Fitzpatrick vs. Fenchel-Young sparsemax losses. **Right:** Fitzpatrick vs. Fenchel-Young logistic losses. In both cases, we plot $L(y, \theta)$, where $y = e_1$ and $\theta = (s, 0)$. This confirms that Fitzpatrick losses are a **lower bound** of Fenchel-Young losses.

Fitzpatrick logistic loss.

Proposition 6 *Formula for Fitzpatrick logistic loss*

When $\Omega(y) = \sum_{i=0}^{k} y_i \ln y_i + \iota_{\triangle^k}(y)$, $L_{F[\partial\Omega]}(y, \theta)$ *is differentiable in* θ *and*

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

where $y^\star_{F[\partial\Omega]}$ *depends on* y *and* θ *and is defined by*

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

where $W(t)$ *is the* Lambert function *[9] defined for* $t > -1/\mathrm{e}$ *as the unique* $W(t) > -1$ *such that* $W(t)\mathrm{e}^{W(t)} = t$ *and where* $\lambda \in \mathbb{R}$ *satisfies*

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

$$\ln \sum_{i=0}^{k} \mathrm{e}^{\theta_i} \leq \lambda - \ln 2 \leq \max \left\{ \ln \sum_{i:y_i=0}^{k} \mathrm{e}^{\theta_i}, \ln K_+ + \max_{i:y_i>0} 2y_i K_+ + \theta_i \right\},$$

where $K_+ = \mathrm{Card}(j : y_j > 0)$.

3.3 Relation with Fenchel-Young losses

On first sight, Fitzpatrick losses and Fenchel-Young losses appear quite different. In the next proposition, we derive a new characterization of the Fitzpatrick function $F[\partial\Omega]$ based on the generalized Bregman divergence D_Ω defined in (1), allowing us to rewrite Fitzpatrick losses $L_{F[\partial\Omega]}$ as Fenchel-Young losses with a modified generating function.

Proposition 7 *Characterization of* $F[\partial\Omega]$ *and* $L_{F[\partial\Omega]}$ *using* D_Ω

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

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

where we defined

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

In particular, when $\Omega = \Psi + \iota_{\mathcal{C}}$, where $\mathcal{C} \subseteq \operatorname{dom} \Psi$, we have for all $y \in \mathcal{C}$

$$\Omega_y(y') = \Psi(y') + D_\Psi(y, y') + \iota(y').$$

A proof is given in Appendix A.8. The Fitzpatrick loss generated by Ω can therefore be seen as a Fenchel-Young loss generated by a modified, target-dependent function $\Omega_y := \Omega + D_\Omega(y, \cdot)$. 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 [6, 4]. It also also provides a mean to compute Fitzpatrick losses and their gradient.

3.4 Lower bound on Fitzpatrick losses

Proposition 8 *Lower bound*

Let $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$ be a twice differentiable μ-strongly convex function. Then, we have for all $(y, \theta) \in \mathbb{R}^k \times \mathbb{R}^k$

$$\frac{\mu}{2} \|y - y^\star\|_2^2 \leq L_{F[\partial\Omega]}(y, \theta).$$

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

A proof is given in Appendix A.9.

4 Experiments

4.1 Label proportion estimation tests

Experimental setup. We follow the same experimental setup as in [6] that we recall here. We consider a dataset of n pairs (x_i, y_i) of features 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. Knowing the dataset and given an unknown input vector $x \in \mathbb{R}^d$, we want to estimate a vector of label proportions $y \in \triangle^k$. We train a model, given by a matrix $W \in \mathbb{R}^{k \times d}$ and a function $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$ which is twice differentiable in the interior of its domain, such that our label proportions prediction is $\widehat{y}_\Omega(Wx) = \nabla\Omega^*(Wx)$. At training time, we estimate the matrix $W \in \mathbb{R}^{k \times d}$ by minimizing the convex objective

$$R_{\Omega,L}(W) := \sum_i^n L(Wx_i, y_i) + \frac{\lambda}{2} \|W\|^2,$$

where $\|W\|^2 = \sum_h^k \sum_j^d W_{hj}^2$, Ω is either the squared 2-norm restricted to the probability simplex or the Shannon negentropy restricted to the probability simplex, and $L \in \left\{L_{\Omega \oplus \Omega^*}, L_{F[\partial\Omega]}\right\}$ is either the Fenchel-Young loss or the Fitzpatrick loss.

We optimize the convex objective using L-BFGS algorithm [14]. We denote $Y^\star, Y, X$ the matrices whose rows gather $y^\star(y_i, Wx_i)$, y_i, x_i, for $i = 1, \ldots, n$. According to Equation (2), $y^\star(y_i, Wx_i) = \nabla\Omega^*(Wx_i)$, for Fenchel-Young losses. Proposition 5 and Proposition 6 give $y^\star$ respectively for Fitzpatrick sparsemax loss and for Fitzpatrick logistic loss. By applying the chain rule to the convex objective $R_{\Omega,L}$, we obtain the gradient expression $\nabla R_{\Omega,L}(W) = (Y^\star - Y)^T X + \lambda W$.

Real data experiments. We ran experiments on 6 standard multi-label benchmark datasets[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 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.

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

Dataset	FYsparsemax	FPsparsemax	FYlogistic	FPlogistic
Birds	0.531	**0.513**	0.519	0.522
Cal500	0.035	0.035	0.034	0.034
Emotions	0.317	0.318	0.327	**0.320**
Mediamill	**0.191**	0.203	**0.207**	0.220
Scene	0.363	**0.355**	**0.344**	0.368
Yeast	0.186	0.187	0.183	0.185

Table 1: Test-set performance comparison for Fenchel-Young and Fitzpatrick versions of sparsemax loss and logistic loss on the task of label proportion estimation. For each datasets errors are measured using the mean squared error. For each double column, we write in bold letters the best error if it is at least 0.05 lower than its counterpart.

4.2 Multiclass classification tests

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, 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.

[5] 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.

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

[7] 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.

[8] 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.

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

[10] 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.

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

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

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

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

[15] 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.

[16] 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.

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

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

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

A Proofs

A.1 Proof of Proposition 1 (Properties of Fitzpatrick losses)

Apart from differentiability, the proofs follow from the study of Fitzpatrick functions found in [10, 2, 18]. 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

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

A.2 Proof of Proposition 2 (twice differentiable Ω)

We recall that

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

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 optimal y' and $\theta' = y'$, we then obtain

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

A.3 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

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

A.4 Proof of Proposition 4 (perceptron loss)

A proof 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 we used Lemma 3 in the last line. Therefore, for all $y \in \mathbb{R}^k$ and $\theta \in \mathbb{R}^k$,

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

A.5 Proof of Proposition 5 (Fitzpatrick sparseMAP loss)

A proof was given in [2, Example 3.13]. We provide an alternative proof.

From Proposition 7, we know that

$$F_{\partial f}(y, \theta) = \Omega_y(y) + \Omega^*_y(\theta) = \Omega(y) + \Omega^*_y(\theta),$$

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,

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

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

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

A.6 Proof of Proposition ?? (Formula for Fitzpatrick logistic loss)

A.7 Lemmas

Lemma 1 *Generalized Bregman divergence for sum of functions*

Lemma 2 *Generalized Bregman divergence for constrained Ω*

Let $\Omega = \Psi + \iota_{\mathcal{C}}$, where $\mathcal{C} \subseteq \operatorname{dom}\Psi$. Then, for all $y, y' \in \operatorname{dom}\Psi$

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

Proof. Since $\Omega(y') = \Psi(y') + \iota_{\mathcal{C}}(y')$, 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 3 *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 $y' \in \mathcal{C}$,

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

Therefore, $D_{\iota_{\mathcal{C}}}(y, y') = \iota_{\mathcal{C}}(y)$.

When $y' \notin \mathcal{C}$, $N_{\mathcal{C}}(y') = \emptyset$.

Lemma 4 *Bregman divergence of Ψ_y Let $\Psi_y(y') \coloneqq \Psi(y') + D_\Psi(y, y')$. Then,*

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

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 $y, y' \in \operatorname{dom} \Psi$,

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

Lemma 5 *Gradient of Ψ_y^**

Let $\Psi_y(y') \coloneqq \Psi(y') + D_\Psi(y, y')$. Then,

$$\nabla \Psi_y^*(\theta) = \tilde{y}$$

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. Using Danskin's theorem,

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

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.

Lemma 7 *Gradient of Ω_y^**

Let $\Omega_y(y') \coloneqq \Psi_y(y') + \iota_{\mathcal{C}}(y') = \Psi(y') + D_\Psi(y, y') + \iota_{\mathcal{C}}(y')$, where $\mathcal{C} \subseteq \operatorname{dom} \Psi$. Then,

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

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

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

or equivalently the solution w.r.t. y of

$$\operatorname*{argmin}_{y \in \mathcal{C}} \ \langle y - \tilde{y}, \nabla^2 \Psi(\tilde{y})(y - \tilde{y}) \rangle$$

where

$$\tilde{y} = \nabla \Psi_y^*(\theta).$$

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

$$\nabla (g + \iota_{\mathcal{C}})^*(v) = \operatorname*{argmin}_{u \in \mathcal{C}} D_g(u, \nabla g^*(v))$$

(see [6, Proposition 3]) with $g = \Psi_y$.

A.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 2 and 3, we have for all $y \in \mathcal{C}$

$$\Omega_y(y') = \Psi(y') + D_\Psi(y, y') + \iota(y').$$

A.9 Proof of Proposition 8 (lower bound)

We recall that if Ω is μ-strongly convex and twice differentiable, then for all $y, y' \in \mathbb{R}^k$

$$\frac{\mu}{2}\|y\|^2 \le \langle y, \nabla^2\Omega(y')\cdot y\rangle.$$

Using Proposition 2, we therefore obtain

$$\begin{aligned}
L_{F[\partial\Omega]}(y,\theta) &= \langle y' - y, \nabla^2\Omega(y')(y' - y)\rangle \\
&\ge \frac{\mu}{2}\|y - y'\|_2^2 .
\end{aligned}$$

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: **[TODO]**

   Justification: **[TODO]**

   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: **[TODO]**

   Justification: **[TODO]**

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: **[TODO]**

   Justification: **[TODO]**

   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: **[TODO]**

   Justification: **[TODO]**

   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: **[TODO]**

   Justification: **[TODO]**

   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: **[TODO]**

   Justification: **[TODO]**

   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: **[TODO]**

   Justification: **[TODO]**

   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: **[TODO]**

   Justification: **[TODO]**

   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: **[TODO]**

   Justification: **[TODO]**

   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: **[TODO]**

    Justification: **[TODO]**

    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: **[TODO]**

    Justification: **[TODO]**

    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: **[TODO]**

    Justification: **[TODO]**

    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: **[TODO]**

    Justification: **[TODO]**

    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: **[TODO]**

    Justification: **[TODO]**

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: **[TODO]**

    Justification: **[TODO]**

    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.