Learning with Fitzpatrick Losses

Seta Rakotomandimby, Jean-Philippe Chancelier
Michel De Lara, Mathieu Blondel

When using a **primal-dual link** between a score space and an output space to predict labels are Fenchel-Young losses the only convex **primal-dual losses** that can be used at training time?

Primal-dual link at prediction time

Identity: $\nabla\Omega^*(\theta) = \theta$ for $\Omega(y) = \frac{1}{2}\|y\|^2$

Argmax: $\nabla\Omega^*(\theta) = \arg\max_{y' \in \triangle^k} \langle y', \theta\rangle$ for $\Omega(y) = \iota_{\triangle^k}(y) := \begin{cases} 0, & \text{if } y \in \triangle^k \\ +\infty, & \text{otherwise} \end{cases}$

Sparsemax: $\nabla\Omega^*(\theta) = P_{\triangle^k}(\theta)$ for $\Omega(y) = \frac{1}{2}\|y\|^2 + \iota_{\triangle^k}(y)$

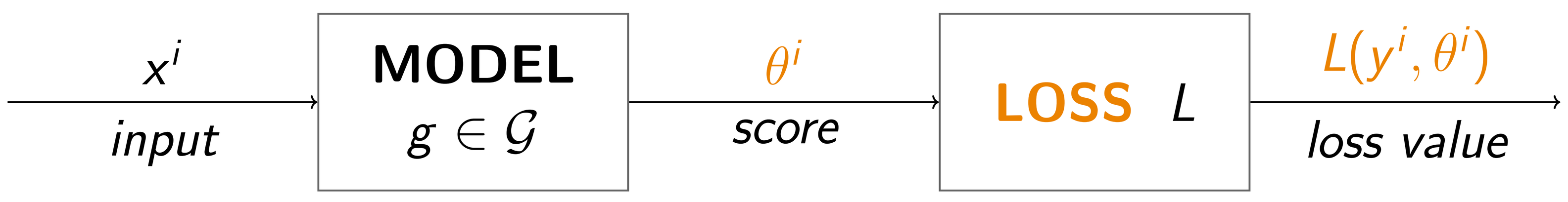
Softmax: $\nabla\Omega^*(\theta) = \frac{\exp(\theta)}{\sum_{i=1}^k \exp(\theta_i)}$ for $\Omega(y) = \langle y, \log y\rangle + \iota_{\triangle^k}(y)$

Simplex projection: $P_{\triangle^k}(y) = \arg\min_{y' \in \triangle^k} \|y' - y\|^2$ **Conjugate function:** $\Omega^*(\theta) := \sup_{y' \in \mathbb{R}^k} \langle y', \theta\rangle - \Omega(y')$

⚠ Composing such link with the squared error usually results a nonconvex loss!

Using primal-dual losses will allow us to have convexity at the last layer at training time!

Primal-dual loss at training time



Minimization problem

- Dataset $(x^i, y^i)_{i=1,\dots,N}$
- Class of models $\mathcal{G}$
- Associated loss $L(y, \theta)$

$$\min_{g \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} L(y^i, \overbrace{g(x^i)}^{=\theta^i})$$

Desired properties

- $L(y, \theta) \geq 0$
- $L(y, \theta)$ convex in θ
- $L(y, \theta)$ diff. in θ

Fenchel-Young (FY) and Fitzpatrick (FP) losses

Which primal-dual losses L are we considering?

Representations[3] of the link

$$L(y, \theta) = 0 \iff \nabla\Omega^*(\theta) = y$$

We **define** for a given proper lower semicontinuous convex function $\Omega : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$

- (Usual choice) **FY loss** [2]
$$L_{\Omega \oplus \Omega^*}(y, \theta) = \Omega(y) + \Omega^*(\theta) - \langle y, \theta\rangle$$
- (**NEW** choice) **FP loss**
$$L_{F[\partial\Omega]}(y, \theta) = \sup_{(y', \theta') \in \partial\Omega} \langle y' - y, \theta - \theta'\rangle$$

Subdifferential: $(y', \theta') \in \partial\Omega \iff \langle y'' - y', \theta\rangle - \Omega(y') \leq \Omega(y''), \ \forall y''$

Case of sparsemax and softmax

Sparsemax

- **FY sparsemax loss**
$$L_{\Omega \oplus \Omega^*}(y, \theta) = \frac{1}{2}\|y - \theta\|^2 - \frac{1}{2}\|P_{\triangle^k}(\theta) - \theta\|^2$$
- **FP sparsemax loss**
$$L_{F[\partial\Omega]} = \|y - \frac{y+\theta}{2}\|^2 - \|P_{\triangle^k}\left(\frac{y+\theta}{2}\right) - \frac{y+\theta}{2}\|^2$$
- The simplex projection is computed using a **sorting algorithm** [6].

Softmax

- **FY logistic loss**
$$L_{\Omega \oplus \Omega^*}(y, \theta) = \log \sum_{i=1}^{k} \exp(\theta_i) + \langle y, \log y\rangle - \langle y, \theta\rangle$$
- **FP logistic loss**
$$L_{F[\partial\Omega]} = \langle y^\star - y, \theta - \log y^\star\rangle$$
- Computation of $y^\star = y^\star(y, \theta)$
$$y_i^\star := \begin{cases} \mathrm{e}^{-\lambda^\star}\mathrm{e}^{\theta_i}, & \text{if } y_i = 0 \\ \frac{y_i}{W(y_i \mathrm{e}^{\lambda^\star - \theta_i})}, & \text{if } y_i > 0 \end{cases}$$
by the **bisection** of $\lambda^\star = \lambda^\star(y, \theta)$ which is the unique solution of
$$\mathrm{e}^{-\lambda^\star} \sum_{i:y_i=0} \mathrm{e}^{\theta_i} + \sum_{i:y_i>0} \frac{y_i}{W(y_i \mathrm{e}^{-(\theta_i - \lambda^\star)})} = 1$$

The nonanalytic **Lambert function** W is defined [5] as the unique nonnegative solution w of $u = w\mathrm{e}^w$, for $u \geq 0$.

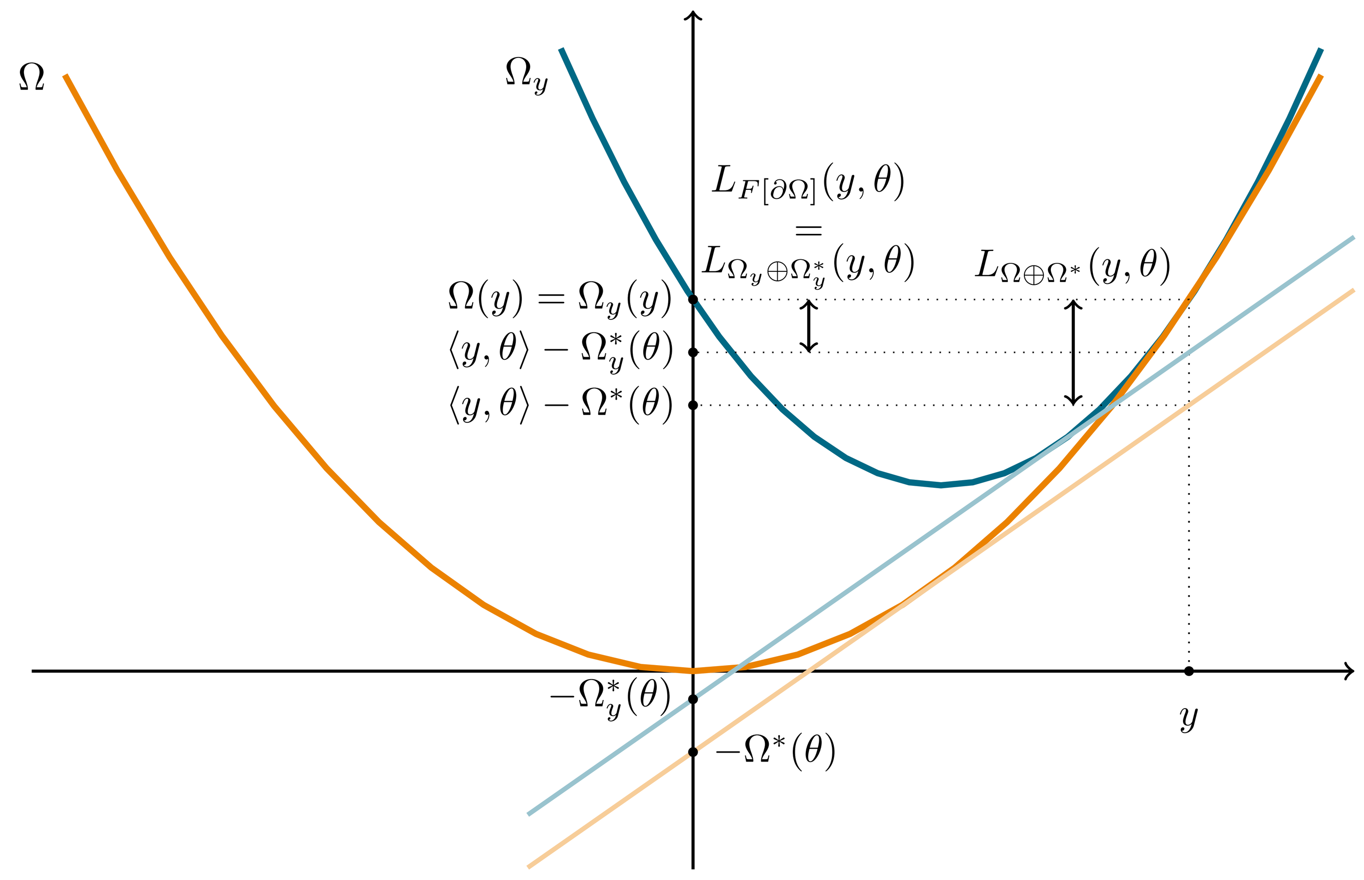
Properties of Fitzpatrick losses

- **FP losses** are tighter than **FY losses**
$$0 \leq L_{F[\partial\Omega]}(y, \theta) \leq L_{\Omega \oplus \Omega^*}(y, \theta)$$
- **Proposition 7**

FP losses as target-dependent FY losses

for some proper lower semicontinuous convex function Ω
$$L_{F[\partial\Omega]}(y, \theta) = L_{\Omega_y \oplus \Omega_y^*}(y, \theta)$$
where the **target-dependent** Ω_y is defined by $\Omega_y(y') = \Omega(y') + D_\Omega(y, y')$

Generalized Bregman divergence: $D_\Omega(y, y') = \Omega(y) - \Omega(y') - \sup_{\theta' \in \partial\Omega(y')} \langle y - y', \theta'\rangle$,



Geometric illustration of Proposition 7 for $\Omega(y) = \frac{1}{2}\|y'\|_2^2$

- **Proposition 8**

Lower bound for **FP losses**
$$\langle y - y^\star, \nabla^2\Omega(y^\star)(y - y^\star)\rangle \leq L_{F[\partial\Omega]}(y, \theta)$$
where $y^\star - y \ (= y^\star(y, \theta) - y) = \nabla_\theta L_{F[\partial\Omega]}(y, \theta)$ and $\nabla^2\Omega$ is the Hessian of Ω

Numerical experiments

Label proportion estimation

Dataset	FY sparsemax	FP sparsemax	FY logistic	FP 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

Test performance measured in mean squared error (the lower the better)

- The **FY sparsemax** and the **FP sparsemax** losses are **comparable on most datasets**.
- The **FY sparsemax** loss significantly wins on only 1 **datasets out of** 11 and the **FP sparsemax** loss significantly wins on 2 **datasets out of** 11.
- The two losses have similar computational cost: **the Fitzpatrick sparsemax loss is a serious contender to the sparsemax loss**.
- The **FY logistic** and the **FP logistic** losses are **comparable on most datasets**.
- The **FY logistic** loss significantly wins on 2 **datasets out of** 11 and the **FP logistic** loss significantly wins on 2 **datasets out of** 11.
- The FP logistic loss is computationally demanding, **the FY logistic loss remains the best choice when we wish to use the softmax**.

Conclusion

- We proposed new nonnegative convex losses **from the maximal monotone operator**theory [4, 3] that share the same primal-dual link as Fenchel-Young losses.
- The **Fitzpatrick sparsemax** loss is a **serious contender** to the **sparsemax loss**.

References

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