Learning with Fitzpatrick Losses

Seta Rakotomandimby Jean-Philippe Chancelier

Michel De Lara Mathieu Blondel

ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES

INSTITUT POLYTECHNIQUE DE PARIS

Google DeepMind

NeurIPS in Paris — December 5, 2024

Paris

Outline of the presentation

From primal-dual links to primal-dual losses

Primal-dual losses from regularized arg max

Theoretical and experimental results on Fitzpatrick losses

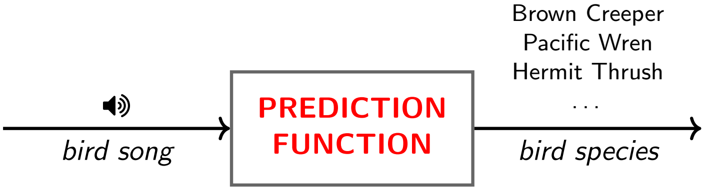
Outline of the presentation

From primal-dual links to primal-dual losses

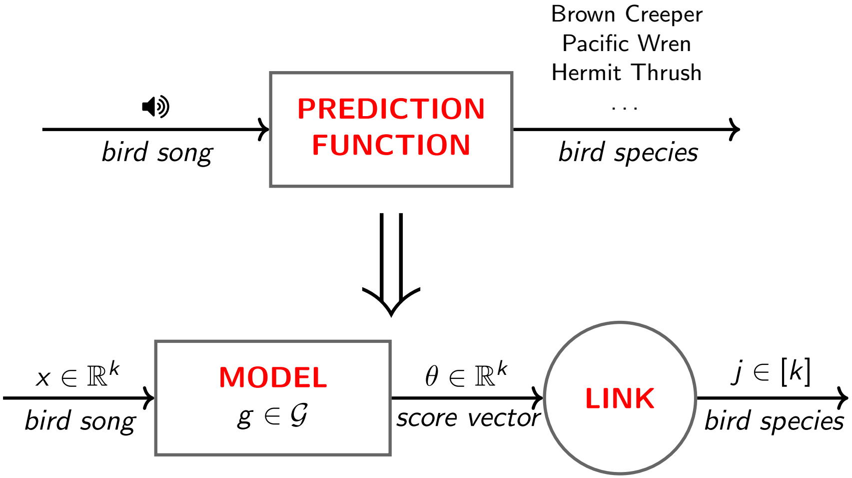
Primal-dual losses from regularized arg max

Theoretical and experimental results on Fitzpatrick losses

Multiclass classification

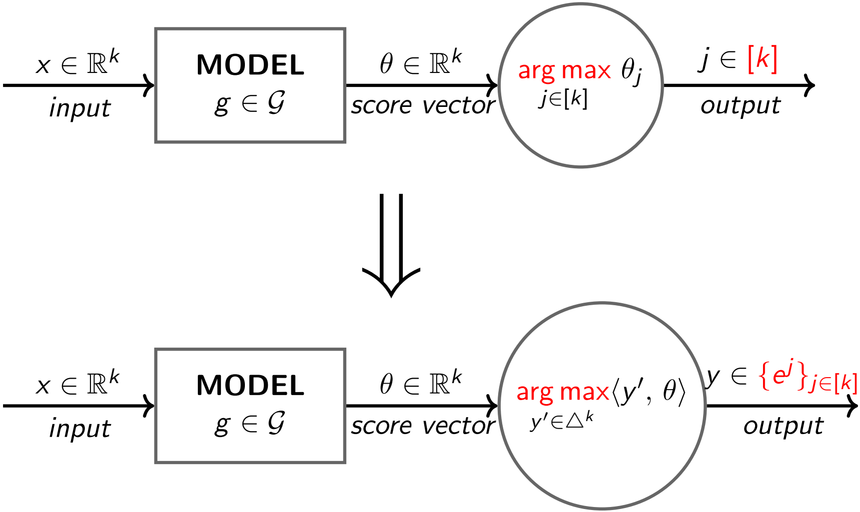


Splitting the prediction function in two



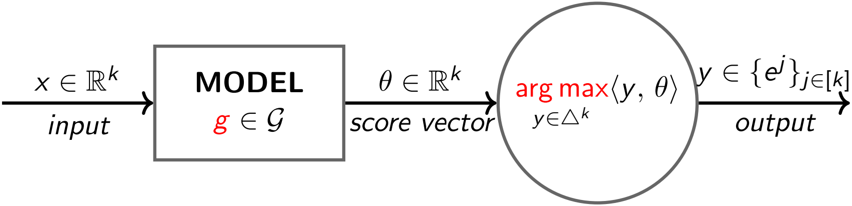
$$[k] = \{1, \ldots, k\}$$

Unregularized link: arg max



$[k] = \{1, \ldots, k\}$, e^j: j-th canonical base vector

Unregularized link: arg max



$[k] = \{1, \ldots, k\}$, e^j: j-th canonical base vector

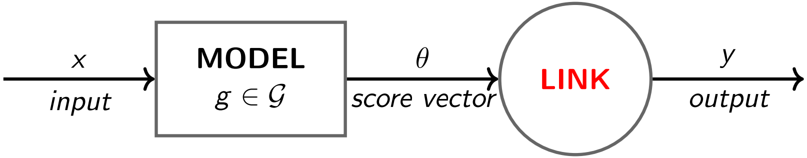
- How to train this?
- Which loss?
- What about the squared loss?

$$\mathbb{R}^k \ni \theta \mapsto \left\| \arg\max_{y' \in \triangle^k} \langle y', \theta \rangle - y_t \right\|^2$$

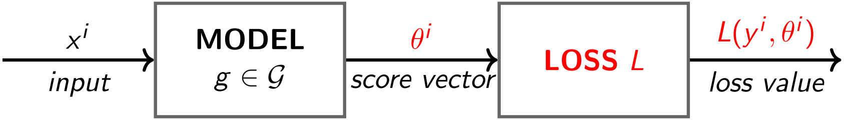
What if we want a convex differentiable loss?

Idea primal-dual loss instead of primal-primal loss

PREDICTION with primal-dual link



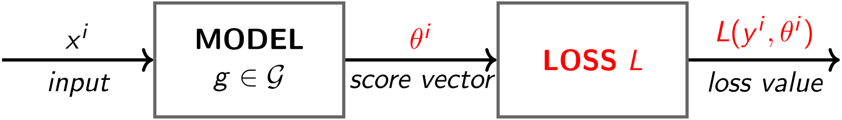
TRAINING with primal-dual loss



Properties of the primal dual loss L

- $L(y, \theta) \geq 0$
- $L(y, \theta)$ convex in θ
- $L(y, \theta)$ differentiable in θ
- $$L(y, \theta) = 0 \iff \textbf{LINK}(\theta) = y$$

The training optimization problem



- **What is given:** dataset $(x^i, y^i)_{i \in [N]}$
- **What is chosen before training:**
  - class of models $\mathcal{G}$
  - link and associated loss L
- **What is trained:** model g

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

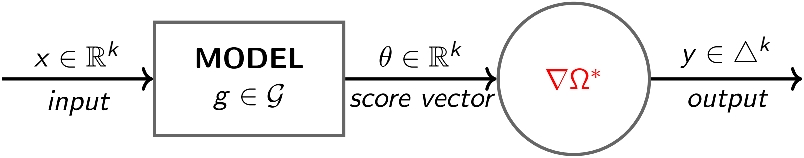
Outline of the presentation

From primal-dual links to primal-dual losses

Primal-dual losses from regularized arg max

Theoretical and experimental results on Fitzpatrick losses

Regularized arg max link



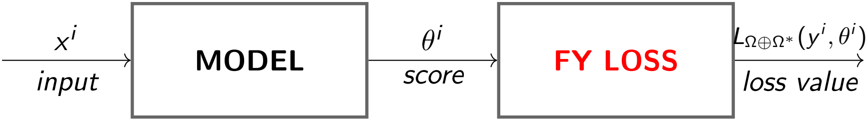
- Regularized arg max:

$$\nabla\Omega^*(\theta) = \underset{y' \in \triangle^k}{\arg\max} \langle y', \theta \rangle - \Omega(y')$$

  where Ω is a strongly convex function
- **Examples**: sparsemax, softmax

Fenchel-Young loss: a natural choice for the loss

[Blondel, Martins, and Niculae, 2020]



- Fenchel-Young (FY) loss associated with $\nabla\Omega^*(\theta)$

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

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

- $L_{\Omega\oplus\Omega^*} \geq 0$, convex in θ
- $L_{\Omega\oplus\Omega^*}(y, \theta) = 0 \iff y = \nabla\Omega^*(\theta)$

Maximal monotone operator representation

[Burachik and Svaiter, 2002, Burachik and Martínez-Legaz, 2018]

- $\nabla\Omega^*$ is a (maximal) monotone operator

$$\langle \nabla\Omega^*(\theta) - \Omega^*(\theta'), \, \theta - \theta' \rangle \geq 0 \,, \quad \forall \theta, \theta' \in \mathbb{R}^k$$

- Smallest convex representation of $\nabla\Omega^*$:
  the Fitzpatrick function $F[\partial\Omega]$ satisfying

$$F[\partial\Omega](y,\theta) \geq \langle y, \, \theta \rangle$$
$$F[\partial\Omega](y,\theta) = \langle y, \, \theta \rangle \iff \nabla\Omega^*(\theta) = y$$
$$F[\partial\Omega] \text{ is convex}$$

- We define the Fitzpatrick loss

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

New loss: the Fitzpatrick loss

[Fitzpatrick, 1988, Bauschke, McLaren, and Sendov, 2005]

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

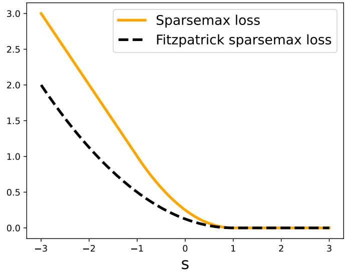
Subdifferential: $(y', \theta') \in \partial\Omega \iff \langle y'', \theta' \rangle - \Omega(y'') \leq \langle y', \theta' \rangle - \Omega(y')$

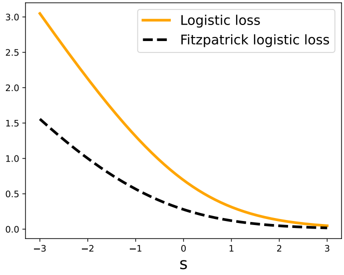
New loss: the Fitzpatrick loss

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

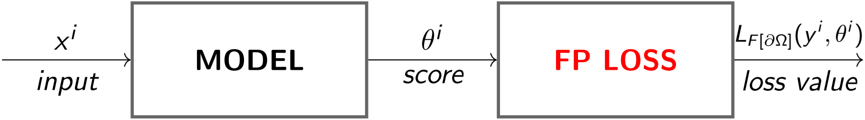
Fitzpatrick losses are tighter than Fenchel-Young losses

$$0 \leq \underbrace{L_{F[\partial\Omega]}}_{\text{Fitzpatrick}} \leq \underbrace{L_{\Omega\oplus\Omega^*}}_{\text{Fenchel-Young}}$$





New loss: the Fitzpatrick loss



- Fitzpatrick (FP) loss associated with $\nabla\Omega^*(\theta)$

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

- $L_{F[\partial\Omega]} \geq 0$, convex in θ
- $L_{F[\partial\Omega]}(y, \theta) = 0 \iff y = \nabla\Omega^*(\theta)$

Appendices

Appendix: usual Ω

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

Argmax: $\nabla\Omega^*(\theta) = \underset{y' \in \triangle^k}{\arg\max}\, \langle y', \theta \rangle$ **for** $\Omega(y) = \iota_{\triangle^k}(y)$

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** $\Omega(y) = \langle y, \log y \rangle + \iota_{\triangle^k}(y)$

$$\iota_{\triangle^k}(y) := \begin{cases} 0\ , & \text{if } y \in \triangle^k \\ +\infty\ , & \text{otherwise} \end{cases}$$

Appendix: Sparsemax

- **FY sparsemax loss**

$$L_{\Omega \oplus \Omega^*}(y, \theta) = \frac{1}{2}\|y - \theta\|^2 - \frac{1}{2}\|P_{\triangle^k}(y) - \theta\|^2$$

- **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** [Martins and Astudillo, 2016].

Appendix: 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

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

Outline of the presentation

From primal-dual links to primal-dual losses

Primal-dual losses from regularized arg max

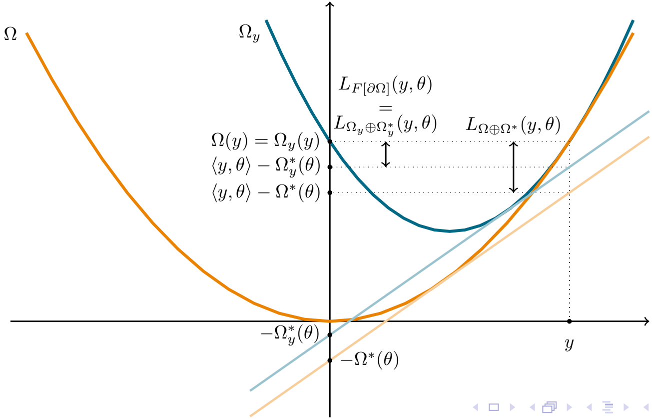
Theoretical and experimental results on Fitzpatrick losses

Theoretical contribution: target dependent FY loss

Theorem

Fitzpatrick losses are FY losses with target dependent Ω_y

$$\underbrace{L_{F[\partial\Omega]}(y,\theta)}_{\text{Fitzpatrick}} = \underbrace{L_{\Omega_y \oplus \Omega_y^*}(y,\theta)}_{\text{“Fenchel-Young”}}$$



Theoretical contribution: target dependent FY loss

Theorem

Fitzpatrick losses are FY losses with target dependent Ω_y

$$\underbrace{L_{F[\partial\Omega]}(y,\theta)}_{\text{Fitzpatrick}} = \underbrace{L_{\Omega_y \oplus \Omega_y^*}(y,\theta)}_{\text{“Fenchel-Young”}}$$

- Target dependent $\Omega_y(y') = \Omega(y') + D_\Omega(y, y')$

  Generalized Bregman divergence:

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

- Ω_y not necessarily strongly convex
- Surprising news:
  for sparsemax and softmax Ω_y is strongly convex

Test performance comparison FY vs FP

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 comparison for **label proportion estimation** with LBFGS algorithm

- FY and FP sparsemax: same computational cost
- FP logistic higher computational cost than FY logistic

Theoretical contribution: lowerbound on FP loss

Theorem

Lowerbound on Fitzpatrick losses

$$\left\langle y^\star - y, \nabla^2\Omega(y^\star)(y^\star - y)\right\rangle \leq L_{F[\partial\Omega]}(y, \theta)$$

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

Thank you for your attention!

Thank you for your attention!

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

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

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

Regina Sandra Burachik and Benar F Svaiter. Maximal monotone operators, convex functions and a special family of enlargements. *Set-Valued Analysis*, 10:297–316, 2002.

Simon 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.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: