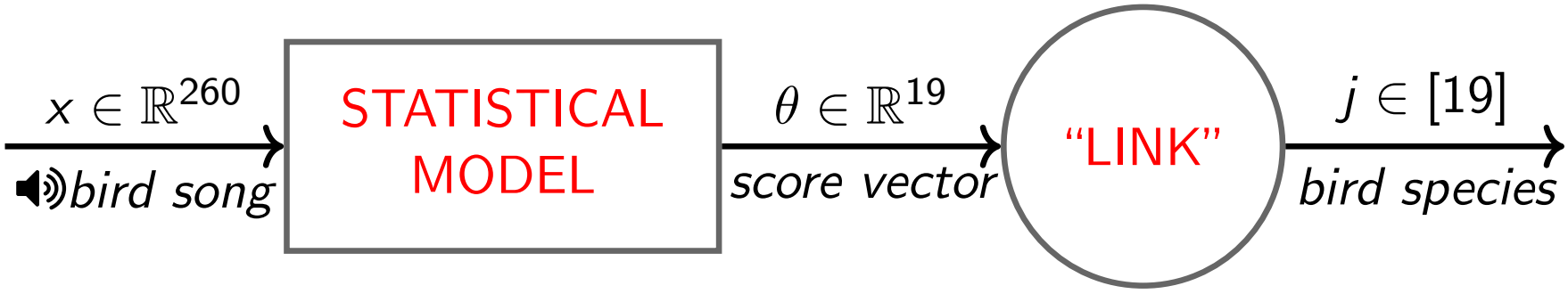
Learning with Representation Losses

Michel De Lara and
Seta Rakotomandimby

Eighth conference on
Discrete Optimization and Machine Learning

Institute of Science Tokyo, Tokyo, Japan
July 13-16 2026

Formal prediction pipeline with a link



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

Input in $\mathbb{R}^{260}$: features

- shape features
- texture features
- noise-robust profile
- ...

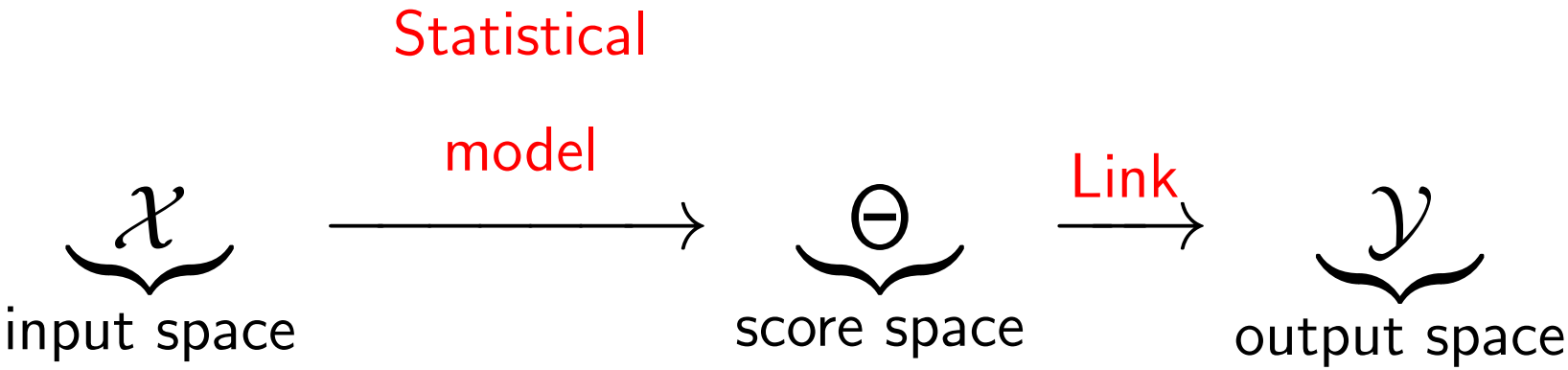
Output in [19]: bird species

- Brown Creeper
- Pacific Wren
- Pacific-slope Flycatcher
- ...

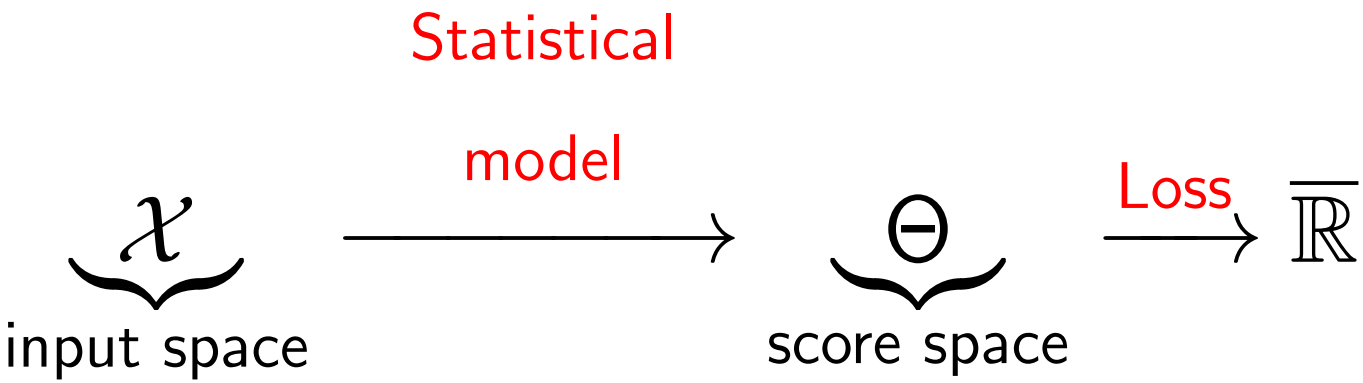
Dataset `birds` from `https://mulan.sourceforge.net/datasets-mlc.html`

From "predicting with a link" to "learning a link"

Predicting:



Training:



Talk outline

Learning links with losses

Learning with Fitzpatrick losses

Learning with representation losses

Outline of the presentation

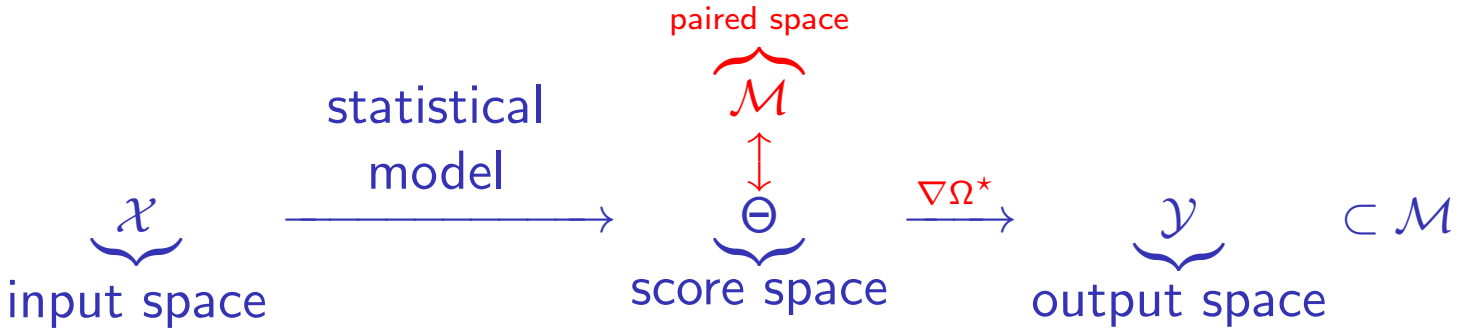
Learning links with losses

Learning with Fitzpatrick losses

Learning with representation losses

Outline of the presentation

Gradient link for predicting [Blondel, Martins, and Niculae, 2020]



- Bilinear pairing $\langle \, , \rangle \colon \mathcal{M} \times \Theta \to \mathbb{R}$
- Regularization function $\Omega \colon \mathcal{M} \to \overline{\mathbb{R}}$

$$\Omega^\star \colon \Theta \to \overline{\mathbb{R}} \, , \quad \Omega^\star(\theta) = \sup_{\mu \in \mathcal{M}} \big(\langle \mu, \, \theta \rangle - \Omega(\mu) \big)$$

Classic scalar links are gradient links

- $\mathrm{relu}(\theta) = \nabla\Omega^\star(\theta) = \max\{0, \theta\}$

$$\text{where } \Omega(\mu) = \frac{1}{2}\mu^2 + \iota_{\mathbb{R}_+}(\mu)$$

$$\text{and } \Omega^\star(\theta) = \frac{1}{2}\underbrace{\max\{0,\theta\}^2}_{\text{differentiable at 0}}$$

- $\mathrm{sigmoid}(\theta) = \nabla\Omega^\star(\theta) = 1/(1+\mathrm{e}^{-\theta})$

$$\text{where } \Omega(\mu) = \mu\ln(\mu) + (1-\mu)\ln(1-\mu) + \iota_{[0,1]}(\mu)$$

$$\text{and } \Omega^\star(\theta) = \ln(1+\mathrm{e}^{\theta})$$

ι indicator function

Classic vectorial links are gradient links

- $\mathrm{sparsemax}(\theta) = \nabla\Omega^\star(\theta) = P_{\triangle^k}(\theta)$ (projection)

$$\text{where } \Omega(\mu) = \frac{1}{2}\|\mu\|_2^2 + \iota_{\triangle^k}(\mu)$$

$$\text{and } \Omega^\star(\theta) = \frac{1}{2}\|\theta\|_2^2 - \frac{1}{2}\underbrace{\left\|\theta - P_{\triangle^k}(\theta)\right\|_2^2}_{=d_{\triangle^k}(\theta)^2\text{, hence is diff.}}$$

- $\mathrm{softmax}(\theta) = \nabla\Omega^\star(\theta) = (\mathrm{e}^{\theta_1}, \ldots, \mathrm{e}^{\theta^k}) / \sum_{i=1}^k \mathrm{e}^{\theta_i}$

$$\text{where } \Omega(\mu) = \sum_{i=1}^k \mu_i \ln \mu_i + \iota_{\triangle^k}(\mu)$$

$$\text{and } \Omega^\star(\theta) = \ln\Big(\sum_{i=1}^k \mathrm{e}^{\theta_i}\Big)$$

Simplex $\triangle^k = \left\{\mu \in \mathbb{R}_+^k \;\middle|\; \sum_i \mu_i = 1\right\}$

Outline of the presentation

Parameterizing the statistical model

- Parameterizing the statistical model $\varphi_w : \mathcal{X} \to \Theta$

$$\underbrace{\mathcal{X}}_{\text{input space}} \xrightarrow{\varphi_w} \underbrace{\Theta}_{\text{score space}} \xrightarrow{\nabla\Omega^\star} \underbrace{\mathcal{Y}}_{\text{output space}} \subset \mathcal{M}$$

- Parameter w over which the optimization will, ultimately, be done
- Typical case: w are weights in a neural network
- Special case: w matrix W, linear model $\varphi_W(x) = Wx$

How to learn the link with a loss?

- Being given an input-output dataset
  $(x^1, \mu^1), \ldots, (x^i, \mu^i), \ldots, (x^N, \mu^N) \in \mathcal{X} \times \mathcal{Y}$
- Measuring error with a loss function $L : \mathcal{M} \times \Theta \to \overline{\mathbb{R}}$

$$x^i \in \mathcal{X} \xrightarrow{\varphi_w} \Theta \xrightarrow{L(\mu^i, \cdot)} \overline{\mathbb{R}}$$

- Minimizing the empirical mean loss

$$\min_w \frac{1}{N} \sum_{i=1}^{N} L\big(\mu^i, \underbrace{\varphi_w(x^i)}_{\theta^i}\big)$$

How to choose a loss to learn a link?

$$\mathcal{M} \times \Theta \ni (\mu, \theta) \mapsto L(\mu, \theta) \in \overline{\mathbb{R}}$$

- nonnegative $L(\mu, \theta) \geq 0$ and vanishing on the link:

$$L(\mu, \theta) = 0 \iff \mu = \nabla \Omega^\star(\theta)$$

- convex in θ, so that, when $\overbrace{\varphi_W(x) = Wx}^{\text{linear stat. model}}$, the training

$$\min_{W \text{ matrix}} \frac{1}{N} \sum_{i=1}^{N} L\big(\mu^i, \overbrace{Wx^i}^{\theta^i}\big) \text{ is a convex minimization problem}$$

- differentiable in θ for the chain rule in gradient descent

Fenchel-Young losses

[Blondel, Martins, and Niculae, 2020]

For $\Omega^\star$ differentiable,

$$L_{FY[\Omega]}(\mu, \theta) = \overbrace{\Omega(\mu) \dot{+} \Omega^\star(\theta)}^{FY[\Omega](\mu,\theta)} - \langle \mu,\ \theta \rangle$$

- nonnegative: $L_{FY[\Omega]}(\mu, \theta) \geq 0$
- vanishing on the link: $L_{FY[\Omega]}(\mu, \theta) = 0 \iff \mu = \nabla\Omega^\star(\theta)$
- convex and differentiable in θ

Other examples of convex losses vanishing on the link
are the so-called Fitzpatrick losses
introduced in
[Rakotomandimby, Chancelier, De Lara, and Blondel, 2024]

Outline of the presentation

Outline of the presentation

New losses: Fitzpatrick losses

- Fitzpatrick function (of a subdifferential)
  [Fitzpatrick, 1988],[Bauschke, McLaren, and Sendov, 2005]

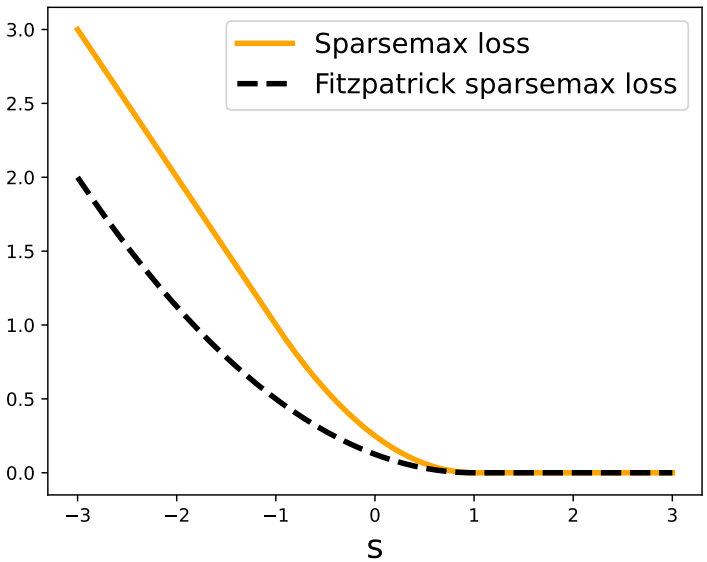
$$FP[\Omega](\mu,\theta) = \sup_{(\mu',\theta')\in\partial\Omega} \left(\langle \mu', \theta\rangle + \langle \mu, \theta'\rangle - \langle \mu', \theta'\rangle\right)$$

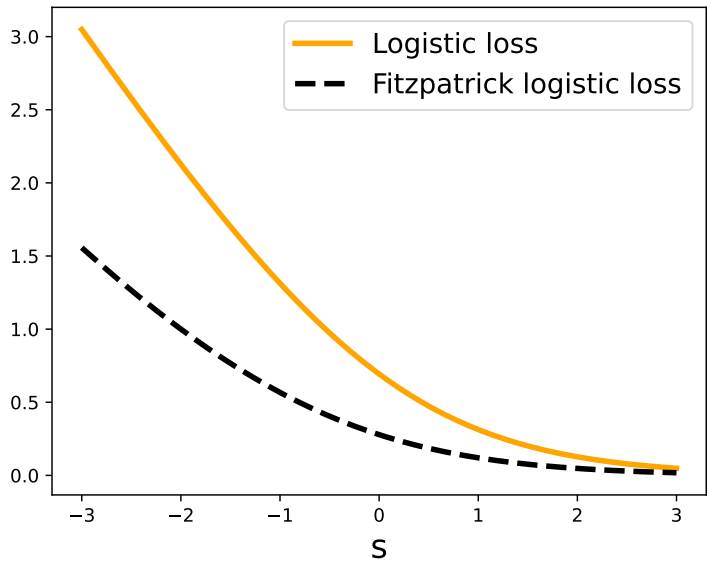
- Fitzpatrick loss
  [Rakotomandimby, Chancelier, De Lara, and Blondel, 2024]

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

Fitzpatrick losses are tighter than Fenchel-Young losses

$$0 \leq \underbrace{L_{FP[\Omega]}}_{\text{Fitzpatrick}} \leq \underbrace{L_{FY[\Omega]}}_{\text{Fenchel-Young}}$$





Fitzpatrick losses display desirable properties

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

- nonnegative: $L_{FP[\Omega]}(\mu, \theta) \geq 0$
- vanishing on the link: $L_{FP[\Omega]}(\mu, \theta) = 0 \iff \mu = \nabla\Omega^\star(\theta)$ when $\Omega^\star$ is differentiable
- convex and ~~differentiable~~ in θ

Outline of the presentation

Unexpected differentiability

In the following cases

- sparsemax: $\nabla\Omega^\star(\theta) = P_{\triangle^k}(\theta)$,
  for $\Omega(\mu) = \frac{1}{2}\|\mu\|_2^2 + \iota_{\triangle^k}(\mu)$

- softmax: $\nabla\Omega^\star(\theta) = (\mathrm{e}^{\theta_1}, \ldots, \mathrm{e}^{\theta^k}) / \sum_{i=1}^k \mathrm{e}^{\theta_i}$,
  for $\Omega(\mu) = \langle \mu, \ln \mu \rangle + \iota_{\triangle^k}(\mu)$

the Fitzpatrick loss $(\mu, \theta) \mapsto L_{FP[\Omega]}(\mu, \theta)$ is differentiable in θ

$\triangle^k = \{\mu \in \mathbb{R}_+^k \mid \sum_i \mu_i = 1\}$ $\quad$ ι indicator function

FY sparsemax *versus* FP sparsemax

From calculations in [Bauschke, McLaren, and Sendov, 2005]

- FY sparsemax loss

$$L_{FY[\Omega]}(\mu,\theta) = \frac{1}{2}\left\|\mu-\theta\right\|_2^2 - \frac{1}{2}\underbrace{\left\|\theta - P_{\triangle^k}(\theta)\right\|_2^2}_{\text{squared distance to } \triangle^k}$$

- FP sparsemax loss

$$L_{FP[\Omega]}(\mu,\theta) = \left\|\mu - \frac{\mu+\theta}{2}\right\|_2^2 - \underbrace{\left\|\frac{\mu+\theta}{2} - P_{\triangle^k}\left(\frac{\mu+\theta}{2}\right)\right\|_2^2}_{\text{squared distance to } \triangle^k}$$

- FY sparsemax loss $\theta \rightarrow \frac{\mu+\theta}{2}$ in FP sparsemax loss

Test performance comparison for label proportion estimation with LBFGS algorithm

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

Outline of the presentation

Theoretical contribution: target dependent FY loss

[Rakotomandimby, Chancelier, De Lara, and Blondel, 2024]

Theorem

Fitzpatrick losses are FY losses with target dependent Ω_μ

$$\underbrace{L_{FP[\Omega]}(\mu, \theta)}_{\text{Fitzpatrick}} = \underbrace{L_{FY[\Omega_\mu]}(\mu, \theta)}_{\text{“Fenchel-Young”}}, \quad \forall \mu \in \mathrm{dom}\Omega, \forall \theta \in \Theta$$

- Target dependent $\Omega_\mu(\mu') = \Omega(\mu') + D^\flat_\Omega(\mu, \mu')$

  Generalized Bregman distance [Kiwiel, 1997]:

  $$D^\flat_\Omega(\mu, \mu') = \Omega(\mu) \dot{+} \big(- \Omega(\mu') \big) \dot{+} \inf_{\theta' \in \partial\Omega(\mu')} \langle \mu' - \mu, \theta' \rangle$$

- This is an example of what [Andrade, Peyré, and Poon, 2025] have later coined sharpened FY losses

Why FP sparsemax and FP softmax are differentiable

- $\Omega_\mu(\mu') = \Omega(\mu') + D^\flat_\Omega(\mu, \mu')$

Ω_μ strongly convex

$\Longrightarrow (\Omega_\mu)^\star$ differentiable

$\Longrightarrow \theta \mapsto L_{FY[\Omega_\mu]}(\mu, \theta) = \Omega_\mu(\mu) \dot{+} (\Omega_\mu)^\star(\theta) - \langle \mu, \theta \rangle$ differentiable

- Sparsemax case:

$$\Omega_\mu(\mu') = \frac{1}{2} \left\| \mu - \mu' \right\|_2^2 + \iota_{\triangle^k}(\mu) + \iota_{\triangle^k}(\mu')$$

- Softmax case:

$$\Omega_\mu(\mu') = \underbrace{\sum_{i=1}^{k} \mu_i (\ln \mu_i - \ln \mu'_i)}_{\text{Kullback-Leibler}} + \iota_{\triangle^k}(\mu) + \iota_{\text{int}\,\triangle^k}(\mu')$$

We will now show that
Fenchel-Young losses and Fitzpatrick losses
are part
of a broader class of convex losses vanishing on the link:
the representation losses
[Rakotomandimby and De Lara, 2026]

Outline of the presentation

References

- Representing monotone operators by convex functions [Fitzpatrick, 1988]
- Maximal monotone operators, convex functions and a special family of enlargements [Burachik and Svaiter, 2002]
- Monotone operators representable by l.s.c. convex functions [Martínez-Legaz and Svaiter, 2005]
- Enlargements: a bridge between maximal monotonicity and convexity [Burachik, 2022]
- Set-valued mappings and enlargements of monotone operators [Burachik and Iusem, 2007]

Outline of the presentation

Maximal monotone operators

Bilinear pairing $\langle\, , \rangle \colon \mathcal{M} \times \Theta \to \mathbb{R}$

Definition

An operator $T \colon \Theta \rightrightarrows \mathcal{M}$

- is called monotone, if

$$\mu \in T(\theta), \mu' \in T(\theta') \implies \langle \mu - \mu', \theta - \theta' \rangle \geq 0 \,, \quad \forall \mu, \mu', \theta, \theta'$$

- is called maximal, if for any monotone $S \colon \mathcal{M} \rightrightarrows \Theta$

$$\mathcal{G}(T) \subset \mathcal{G}(S) \implies T = S$$

Graph $\mathcal{G}(T) = \left\{ (\mu, \theta) \in \mathcal{M} \times \Theta \,\middle|\, \mu \in T(\theta) \right\}$

Representation functions

[Burachik and Svaiter, 2002, Martínez-Legaz and Svaiter, 2005]

Definition

Given $T : \Theta \rightrightarrows \mathcal{M}$ a maximal monotone operator
the bivariate function $H : \mathcal{M} \times \Theta \to \overline{\mathbb{R}}$ is
a representation function of T, if

- $H(\mu, \theta) \geq \langle \mu, \theta \rangle$
- $H(\mu, \theta) = \langle \mu, \theta \rangle \iff \mu \in T(\theta)$
- H is l.s.c. convex in (μ, θ)

Smallest and largest representation

[Burachik and Svaiter, 2002, Corollary 4.1]

Let $T : \Theta \rightrightarrows \mathcal{M}$ be a maximal monotone operator

Proposition

For any representation $H : \mathcal{M} \times \Theta \to \overline{\mathbb{R}}$ of T

$$\overbrace{FP[T]}^{\text{Fitzpatrick}} \leq H \leq \overbrace{FP[T]^\star}^{\substack{\text{conjugate} \\ \text{Fitzpatrick}}}$$

$$FP[T](\mu, \theta) = \sup_{(\mu', \theta') \in T} \left(\langle \mu, \theta' \rangle + \langle \mu', \theta \rangle - \langle \mu', \theta' \rangle \right)$$

$$FP[T]^\star(\mu, \theta) = \sup_{(\mu', \theta')} \left(\langle \mu, \theta' \rangle + \langle \mu', \theta \rangle - FP[T](\mu', \theta') \right)$$

New representation losses for training

[Rakotomandimby and De Lara, 2026]

Given $T : \Theta \rightrightarrows \mathcal{M}$ a maximal monotone operator and $H : \mathcal{M} \times \Theta \to \overline{\mathbb{R}}$ a representation of T, we define the representation loss by

$$L_H(\mu, \theta) = H(\mu, \theta) - \langle \mu, \theta \rangle$$

- nonnegative: $L_H(\mu, \theta) \geq 0$
- vanishing on the link: $L_H(\mu, \theta) = 0 \iff \mu \in T(\theta)$
- convex and ~~differentiable~~ in θ

Outline of the presentation

What distinguishes one representation loss from another?

ε-subdifferential with FY losses

- We have

$$L_{FY[\Omega]}(\mu, \theta) = 0 \iff \mu \in \partial\Omega^\star(\theta)$$

- But also

$$0 \leq \underbrace{L_{FY[\Omega]}(\mu, \theta) \leq \varepsilon}_{\text{robustness}} \iff \underbrace{\mu \in \partial_\varepsilon \Omega^\star(\theta)}_{\text{FY enlargement}}$$

ε-subdifferential

$$\partial_\varepsilon \Omega(\mu) = \left\{ \theta \in \Theta \,\middle|\, \Omega(\mu') \geq \langle \mu' - \mu, \theta \rangle + \Omega(\mu) - \varepsilon \,, \;\; \forall \mu' \in \mathcal{M} \right\}$$

Bijection between representations and enlargements

[Burachik, 2022, Theorem 30]

Theorem

The closed enlargements $\left(\mathbb{E}_c(T), \subset\right)$
and the representations functions $\left(\mathcal{H}(T), \leq\right)$
are in a nondecreasing bijection

- Fenchel-Young enlargement of $\partial\Omega^\star$:

$$\partial_\varepsilon\Omega^\star(\theta) = \left\{\mu \in \mathcal{M} \,\middle|\, \Omega^\star(\theta') \geq \langle\mu,\, \theta' - \theta\rangle + \Omega^\star(\theta) - \varepsilon\ ,\ \ \forall\theta' \in \Theta\right\}$$

- Fitzpatrick enlargement of $\partial\Omega^\star$: $\forall\theta \in \mathrm{dom}\Omega^\star$

$$\left(\partial\Omega^\star\right)^e(\varepsilon,\theta) = \left\{\mu \in \mathcal{M} \,\middle|\, \langle\mu' - \mu,\, \theta' - \theta\rangle \geq -\varepsilon, \forall(\mu',\theta') \in \partial\Omega^\star\right\}$$

- Inclusions, as $L_{FP[\Omega]} \leq L_{FY[\Omega]}$

$$\partial_\varepsilon\Omega^\star(\theta) \subset \left(\partial\Omega^\star\right)^e(\varepsilon,\theta)\ ,\ \ \forall\theta \in \Theta, \varepsilon > 0$$

Family of enlargements

- Let $T\colon \mathcal{M} \rightrightarrows \Theta$ be a maximal monotone operator

[Burachik and Svaiter, 2002, Definition 2.3] [Burachik, 2022, Definition 7]

Definition

We say that $E\colon \mathbb{R}_+ \times \mathcal{M} \rightrightarrows \Theta$ is an enlargement of T if

(i) $T(\mu) \subset E(\varepsilon, \mu)$ e.g. $\partial\Omega(\mu) \subset \partial_\varepsilon\Omega(\mu)$

(ii) If $0 \leq \varepsilon_1 \leq \varepsilon_2$, then $E(\varepsilon_1, \mu) \subset E(\varepsilon_2, \mu)$ e.g. $\partial_{\varepsilon_1}\Omega(\mu) \subset \partial_{\varepsilon_2}\Omega(\mu)$

(iii) The transportation formula is satisfied, if

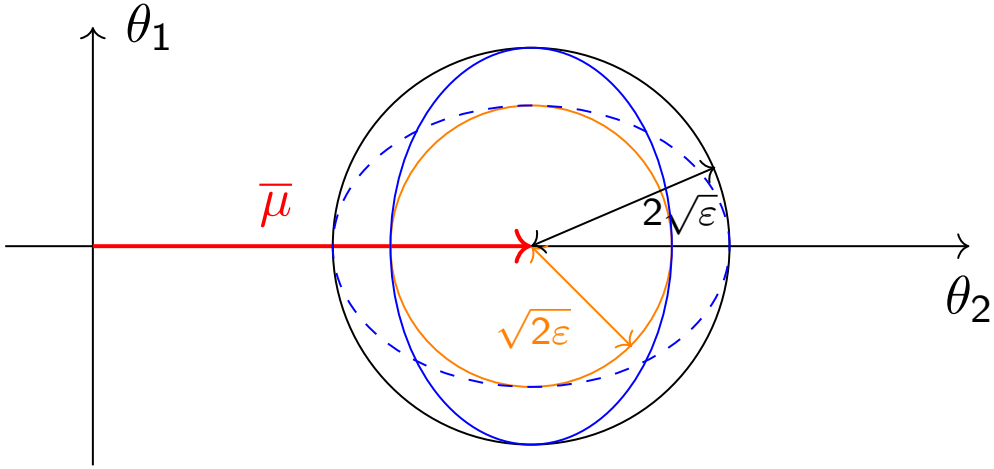
- $\theta_1 \in E(\varepsilon_1, \mu_1)$, $\theta_2 \in E(\varepsilon_2, \mu_2)$
- $\alpha \in [0, 1]$
- $\varepsilon = \alpha\varepsilon_1 + (1-\alpha)\varepsilon_2 + \alpha(1-\alpha)\langle \mu_1 - \mu_2, \theta_1 - \theta_2 \rangle$

then $\varepsilon \geq 0$ and $\alpha\theta_1 + (1-\alpha)\theta_2 \in E(\varepsilon, \alpha\mu_1 + (1-\alpha)\mu_2)$

If $\mathcal{G}(E)$ is (strongly) closed, we say that E is a closed enlargement

Four examples of enlargements of the identity link

$$I = \nabla\Omega^\star \colon \Theta \to \mathcal{M} \ , \quad \text{where } \Omega^\star(\theta) = \frac{1}{2}\|\theta\|_2^2$$



$L(\overline{\mu}, \theta) \leq \varepsilon$

- Black circle: border of $\left\{\theta \,\middle|\, L_{FP[\Omega]}(\overline{\mu}, \theta) \leq \varepsilon\right\}$
- Orange circle: border of $\left\{\theta \,\middle|\, L_{FY[\Omega]}(\overline{\mu}, \theta) \leq \varepsilon\right\}$
- Blue ellipse in full line: border of $\left\{\theta \,\middle|\, L_{H_1}(\overline{\mu}, \theta) \leq \varepsilon\right\}$
- Blue ellipse in dashed line: border of $\left\{\theta \,\middle|\, L_{H_2}(\overline{\mu}, \theta) \leq \varepsilon\right\}$

Outline of the presentation

References

- Learning with Fenchel-Young Losses
  [Blondel, Martins, and Niculae, 2020]
- Learning with Fitzpatrick Losses
  [Rakotomandimby, Chancelier, De Lara, and Blondel, 2024]
- Learning from Samples: Inverse Problems over measures via Sharpened Fenchel-Young Losses
  [Andrade, Peyré, and Poon, 2025]
- Representation theory and sharpened Fenchel-Young losses
  [Rakotomandimby and De Lara, 2026]

Sharpened Fenchel-Young losses

[Andrade, Peyré, and Poon, 2025]

Definition

$H\colon \mathcal{M} \times \Theta \to \overline{\mathbb{R}}$ is a sharpened Fenchel-Young function if

$$H(\mu, \theta) = FY[\Omega(\cdot) \dot{+} \Delta(\mu, \cdot)](\mu, \theta)$$

where

- $\Omega\colon \mathcal{M} \to \overline{\mathbb{R}}$ is a regularizer
- $\Delta\colon \mathcal{M} \times \mathcal{M} \to \overline{\mathbb{R}}$ is a sharpener

Associated sharpened Fenchel-Young loss

$$L_{\Omega,\Delta}(\mu, \theta) = FY[\Omega(\cdot) \dot{+} \Delta(\mu, \cdot)](\mu, \theta) - \langle \mu, \theta \rangle$$

We give a sufficient condition
for a sharpened FY loss
to be a representation loss of $\partial\Omega^\star$

From representations to sharpeners

[Rakotomandimby and De Lara, 2026]

Theorem

Let $\Omega \colon \mathcal{M} \to \overline{\mathbb{R}}$ be a proper l.s.c. convex function
Let $H \colon \mathcal{M} \times \Theta \to \overline{\mathbb{R}}$ be a representation of $\partial\Omega^\star$

$$\text{If} \qquad FP[\Omega] \leq H \leq FY[\Omega]$$

then there exists a nonnegative sharpener
$\Delta \colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+ \cup \{+\infty\}$ such that

$$\Delta(\mu, \mu) = 0 \ , \ \ \forall \mu \in \mathrm{dom}\partial\Omega \ ,$$

$$H(\mu, \theta) = \underbrace{FY[\Omega(\cdot) \dot{+} \Delta(\mu, \cdot)](\mu, \theta)}_{\text{sharpened Fenchel-Young function}} \ , \ \ \forall \mu \in \mathrm{dom}\Omega$$

We give a sufficient condition
for a sharpened FY loss
to be a representation loss of $\partial\Omega^\star$

From sharpeners to representations

[Rakotomandimby and De Lara, 2026]

Theorem

Let $\Omega \colon \mathcal{M} \to \overline{\mathbb{R}}$ be a proper l.s.c. convex function
Let $\Delta \colon \mathcal{M} \times \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ be a sharpener such that

$$0 \leq \Delta(\mu, \mu') \leq \overbrace{D_\Omega^\flat(\mu, \mu')}^{\text{Bregman distance}} \quad , \ \forall \mu' \in \mathcal{M} \ ,$$
$$\Delta(\mu, \mu) = 0 \ , \ \forall \mu \in \mathrm{dom}\Omega \ ,$$
$$\Omega(\cdot) \dot{+} \Delta(\mu, \cdot) \text{ is proper l.s.c. convex}$$

Then, the sharpened function $FY[\Omega(\cdot) \dot{+} \Delta(\mu, \cdot)](\mu, \theta)$ is a representation of $\partial\Omega^\star$

Recovering previous result for the Fitzpatrick function

[Rakotomandimby, Chancelier, De Lara, and Blondel, 2024, Proposition 7]

Choosing the Bregman divergence as a sharpener allows to write the Fitzpatrick function as a sharpened Fenchel-Young function

$$\Delta(\mu, \mu') = D_\Omega^\flat(\mu, \mu') \ , \ \ \forall \mu, \mu'$$

$$\Longrightarrow$$

$$\underbrace{FP[\Omega](\mu, \theta)}_{\text{Fitzpatrick}} = \underbrace{FY[\Omega(\cdot) \dot{+} \Delta(\mu, \cdot)](\mu, \theta)}_{\text{sharpened Fenchel-Young}}$$

$$\forall \mu \in \mathrm{dom}\Omega, \forall \theta \in \Theta$$

And differentiability in θ?

Producing differentiable representation losses

Proposition

If $\Omega(\cdot) + \Delta(\overline{\mu}, \cdot)$ is strongly convex, then

$$\theta \mapsto L_{\Omega,\Delta}(\overline{\mu}, \theta) \text{ is differentiable}$$

Conclusion

Conclusion

- For a given link, how can we design loss functions?
- We have proposed new convex losses
  - Fitzpatrick losses
  - representation losses
- We have exhibited a class of representation functions of subdifferential of l.s.c. convex functions (links) that display Fenchel-Young sharpness
- Through sharpening, we can obtain differentiable losses
- Representation theory of monotone operators, including enlargments, could be a guide to design suitable loss functions

Thank you for your attention!

Francisco Andrade, Gabriel Peyré, and Clarice Poon. Learning from Samples: Inverse Problems over measures via Sharpened Fenchel-Young Losses. *arXiv preprint arXiv:2505.07124*, 2025.

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

Mathieu Blondel, André F.T. Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020. URL `http://jmlr.org/papers/v21/19-021.html`.

Regina S Burachik. Enlargements: a bridge between maximal monotonicity and convexity. In *Proc. Int. Cong. Math*, volume 7, pages 5212–5233, 2022.

Regina S. Burachik and Alfredo N. Iusem. *Set-Valued Mappings and Enlargements of Monotone Operators*. Springer New York, NY, 2007. ISBN 978-0-387-69755-0. doi: https://doi.org/10.1007/978-0-387-69757-4.

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.

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

J-E Martínez-Legaz and Benar Fux Svaiter. Monotone operators representable by lsc convex functions. *Set-Valued Analysis*, 13:21–46, 2005.

Seta Rakotomandimby and Michel De Lara. Learning through Monotone Links with Representation Losses: the Case of Sharpened Fenchel-Young Losses. working paper or preprint, March 2026. URL `https://hal.science/hal-05551266`.

Seta Rakotomandimby, Jean-Philippe Chancelier, Michel De Lara, and Mathieu Blondel. Learning with Fitzpatrick losses. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=7Dep87TMJs`.