Geometric losses for distributional learning

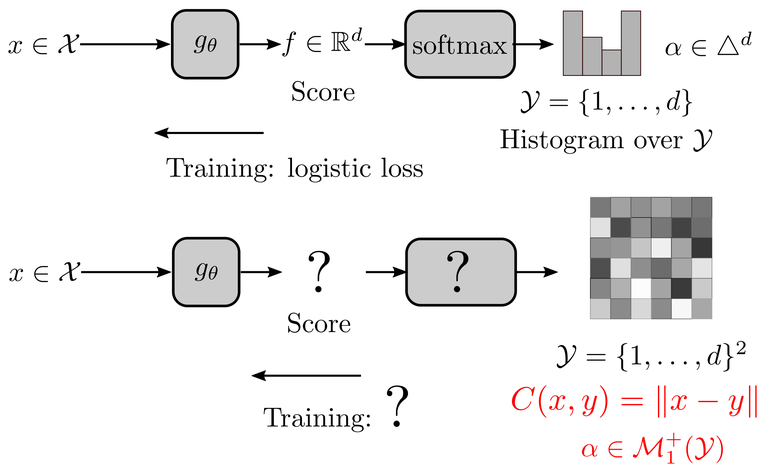
Arthur Mensch[1] Mathieu Blondel[2] Gabriel Peyré[1]

[1]ENS, DMA, CNRS, Pars, France [2]NTT Communication Science Laboratories, Kyoto, Japan

1 - Summary

- Extend the **softmax operator** and logistic loss to handle a **ground metric** in the output space
- The geometric softmax output potentially **continuous distributions**, from a continuous score function ($\sim$ logits)
- Construction based on **Fenchel duality** and a weakly continuous entropy defined on all distributions
- Obtained through self **regularized optimal transport**
- Consistency theorem, geometric study, new g-softmax layer

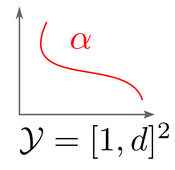
2 - Predicting distribution on metrized output



- Link and loss with cost between classes:

$$C : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

- Output distribution over **continuous space** $\mathcal{Y}$



3 - Prior approaches

Cost augmentation of losses:

- Convex cost-aware loss $L_c : [1, d] \times \mathbb{R}^d \to \mathbb{R}$
- **Undefined link functions**: $\mathbb{R}^d \to \triangle^d$

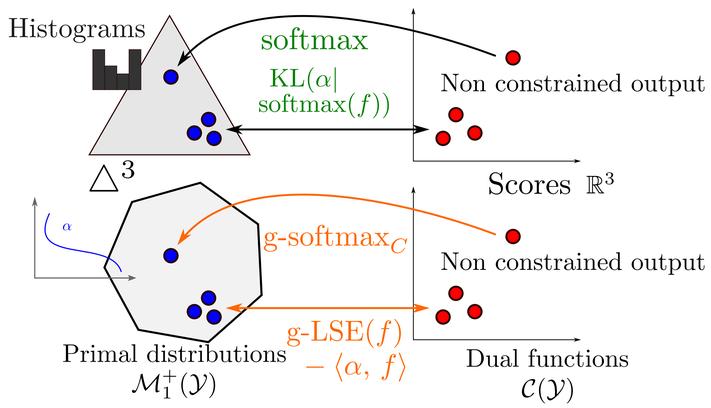
Use a Wasserstein distance between output distributions:

- Prediction with a softmax link:[1] $\ell(\alpha, f) \triangleq W_C(\mathrm{softmax}(f), \alpha))$
- **Non-convex loss** and costly to compute

References

1. Frogner, C. *et al. Learning with a Wasserstein loss.* in *NIPS* (2015).
2. Blondel, M. *et al. Learning Classifiers with Fenchel-Young Losses: Generalized Entropies, Margins, and Algorithms.* in *AISTATS* (2019).
3. Feydy, J. *et al. Interpolating between Optimal Transport and MMD Using Sinkhorn Divergences.* in *AISTATS* (2019).
4. Mensch, A. *et al. Geometric losses for distributional learning.* in *ICML* (2019).

4 - Prediction function and loss from duality



- Predict a continuous function $f = g_\theta(x)$, and go back to primal

4 - Fenchel-Young losses[2] from convex entropies

Convex function $\Omega : \triangle^d \to \mathbb{R}$ and Legendre-Fenchel conjugate

$$\Omega^\star(f) = \min_{\alpha \in \triangle^d} \Omega(\alpha) - \langle \alpha,\ f \rangle \qquad \ell_\Omega(\alpha, f) = \Omega(\alpha) + \Omega^\star(f) - \langle \alpha,\ f \rangle \geqslant 0$$

Define link functions (a.k.a. mirror maps) between dual and primal:

$$\nabla\Omega(\alpha) = \operatorname*{argmin}_{f \in \mathbb{R}^d} \ell_\Omega(\alpha, f) \quad \nabla\Omega^\star(f) = (\nabla\Omega)^{-1}(f) = \operatorname*{argmin}_{\alpha \in \triangle^d} \ell_\Omega(\alpha, f)$$

Shannon ent.: $\Omega(\alpha) = \sum_{i=1}^d \alpha_i \log \alpha_i$, $\nabla\Omega^* = \mathrm{softmax}$, $\Omega^* = \mathrm{LSE}$

5 - Entropies from optimal transport

Self-regularized optimal transportation distance:

$$\Omega_C(\alpha) = -\frac{1}{2}\mathrm{OT}_{C,\varepsilon=2}(\alpha, \alpha) = -\max_{f \in \mathcal{C}(\mathcal{Y})} \langle \alpha, f \rangle - \log\langle \alpha \otimes \alpha, e^{\frac{f \oplus f - C}{2}} \rangle$$

= Sinkhorn negentropy[3]

Weakly continuous and convex

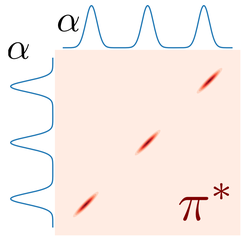
Special cases

$\varepsilon \to \infty$: MMD autocorrelation

$$C = \begin{pmatrix} 0 & \infty & \dots \\ \infty & 0 & \ddots \end{pmatrix}$$

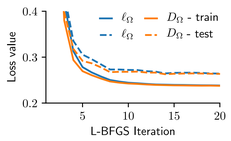
Shannon entropy

Gini index

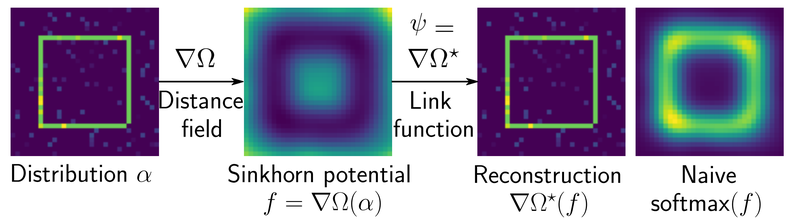


8 - Ordinal regression

	LR	LR(AT)	LR(IT)	**g-logistic**
Haus. div.	.46±.12	.47±.14	.59±.16	**.44±.08**
MAE	.44±.09	**.42±.06**	.44±.08	.45±.09
Acc.	**.66±.07**	.65±.06	.65±.06	.65±.07



6 - Geometric softmax and geometric-LSE



$$\text{g-LSE}(f) \triangleq \Omega^\star(f) = -\log \min_{\alpha \in \mathcal{M}_1^+(\mathcal{Y})} \langle \alpha \otimes \alpha, \exp(-\frac{f \oplus f + C}{2}) \rangle$$

$$\text{g-softmax}(f) \triangleq \nabla\Omega^\star(f) = \operatorname*{argmin}_{\alpha \in \mathcal{M}_1^+(\mathcal{Y})} \langle \alpha \otimes \alpha, \exp(-\frac{f \oplus f + C}{2}) \rangle$$

- Minimization of a quadratic over distribution space / Simplex
- Computable with L-BFGS/mirror descent (discrete distribution)
- Continuous: Frank Wolfe with non-convex linear min. oracle

6 - Learning guarantees

- Sample distribution $(x_i, \alpha_i)_i \in \mathcal{X} \times \mathcal{M}_1^+(\mathcal{Y})$.
- **Bregman divergence** from Sinkhorn negentropy:

$$D_g(\alpha|\beta) = \Omega(\alpha) - \Omega(\beta) - \langle \nabla\Omega(\alpha),\ \alpha - \beta \rangle.$$

- **Fisher consistency:**

$$\min_{\beta : \mathcal{X} \to \mathcal{M}_1^+(\mathcal{Y})} \mathbb{E}\left[D_g\left(\alpha, \beta\left(x\right)\right)\right] = \min_{g : \mathcal{X} \to \mathcal{C}(\mathcal{Y})} \mathbb{E}\left[\ell_g\left(\alpha, \nabla\Omega^*\left(g\left(x\right)\right)\right)\right]$$

7 - Sketch variational autoencoders