# Geometric losses for distributional learning

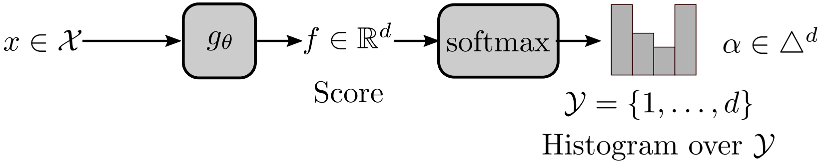

*Arthur Mensch*[(1)], Mathieu Blondel[(2)], Gabriel Peyré[(1)]

[(1)] École Normale Supérieure, DMA
Centre National pour la Recherche Scientifique
Paris, France

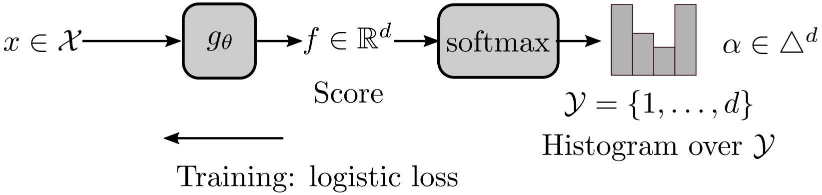
[(2)] NTT Communication Science Laboratories
Kyoto, Japan


June 12, 2019

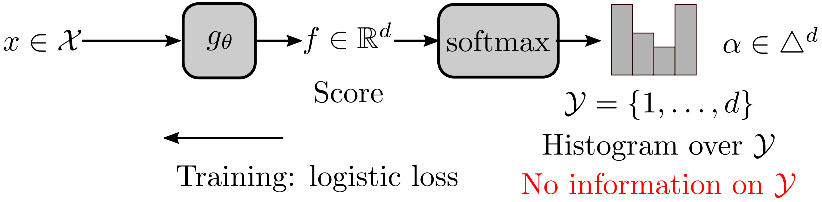
# Introduction: Predicting distributions



# Introduction: Predicting distributions

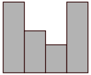


# Introduction: Predicting distributions



# Introduction: Predicting distributions

$x \in \mathcal{X} \longrightarrow \boxed{g_\theta} \longrightarrow f \in \mathbb{R}^d \longrightarrow \boxed{\text{softmax}} \longrightarrow$



$\alpha \in \triangle^d$

Score

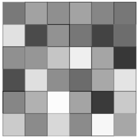
$\mathcal{Y} = \{1, \dots, d\}$

Histogram over $\mathcal{Y}$

No information on $\mathcal{Y}$

$\longleftarrow$

Training: logistic loss

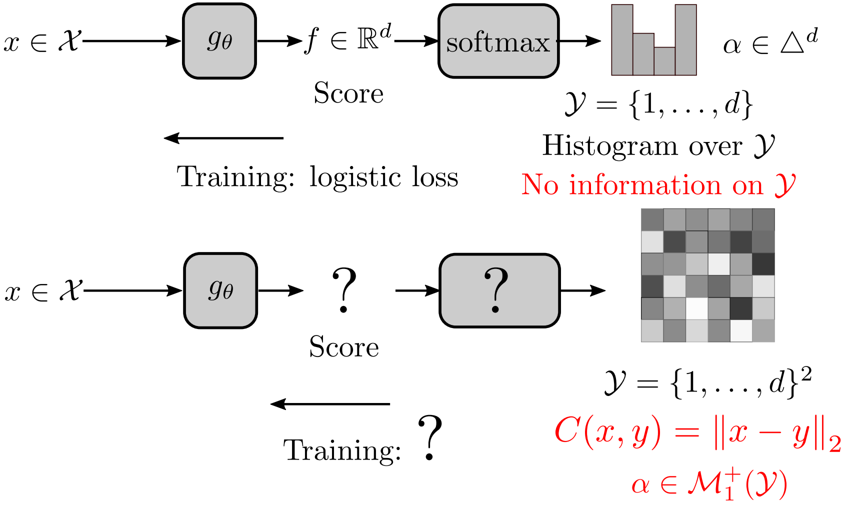


$\mathcal{Y} = \{1, \dots, d\}^2$

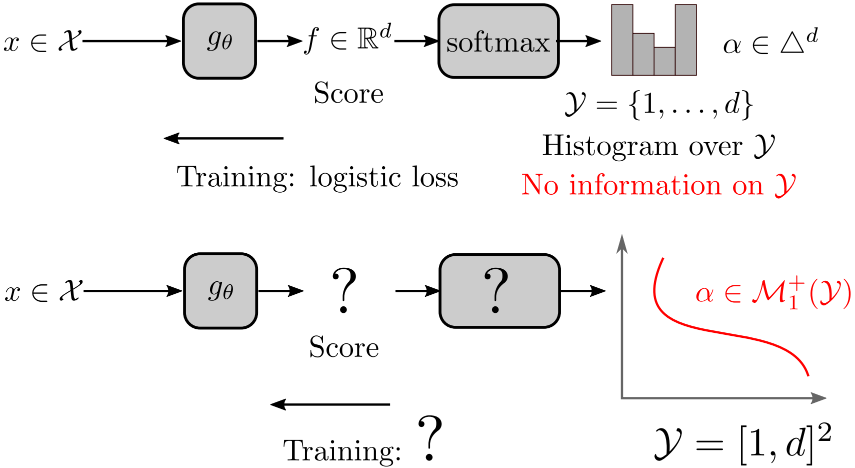
$C(x, y) = \|x - y\|_2$

$\alpha \in \mathcal{M}_1^+(\mathcal{Y})$

# Introduction: Predicting distributions



# Introduction: Predicting distributions



# Contribution: losses and links for continuous metrized output

## Handling output geometry

- Link and loss with cost between classes

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

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

# Contribution: losses and links for continuous metrized output

### Handling output geometry

- Link and loss with cost between classes

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

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

**New geometric losses and associated link functions:**

1 Construction from duality between distributions and scores

# Contribution: losses and links for continuous metrized output

**Handling output geometry**

- Link and loss with cost between classes

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

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

**New geometric losses and associated link functions:**

1. Construction from duality between distributions and scores
2. **Need:** Convex functional on distribution space

- Provided by **regularized optimal transport**

# Background: learning with a cost over outputs $\mathcal{Y}$

**Cost augmentation** of losses[1,2]:

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

⚠ **Undefined link functions**: $\mathbb{R}^d \to \triangle^d$: what to predict at test time ?

[1] Ioannis Tsochantaridis et al. "Large margin methods for structured and interdependent output variables". In: *JMLR* (2005).

[2] Kevin Gimpel and Noah A Smith. "Softmax-margin CRFs: Training log-linear models with cost functions". In: *NAACL*. 2010.

# Background: learning with a cost over outputs $\mathcal{Y}$

**Cost augmentation** of losses[1,2]:

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

⚠ **Undefined link functions**: $\mathbb{R}^d \to \triangle^d$: what to predict at test time ?

Use a Wasserstein **distance between output distributions**[3]:

- Ground metric $C$ defines a distance $W_C$ between distributions
- Prediction with a softmax link

$$\ell(\alpha, f) \triangleq W_C(\mathrm{softmax}(f), \alpha))$$

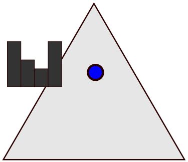
⚠ **Non-convex loss** and costly to compute

[1] Ioannis Tsochantaridis et al. "Large margin methods for structured and interdependent output variables". In: *JMLR* (2005).

[2] Kevin Gimpel and Noah A Smith. "Softmax-margin CRFs: Training log-linear models with cost functions". In: *NAACL*. 2010.

[3] Charlie Frogner et al. "Learning with a Wasserstein loss". In: *NIPS*. 2015.

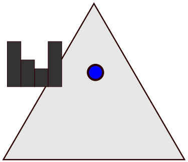
# Predicting distributions from topological duality



Primal histograms

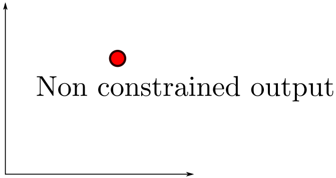
$\triangle^3, \|\cdot\|_1$

# Predicting distributions from topological duality



Primal histograms

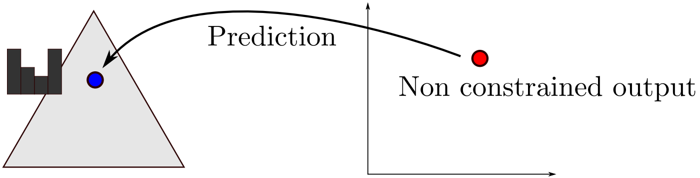
$\triangle^3, \|\cdot\|_1$




Dual scores

$\mathbb{R}^3, \|\cdot\|_\infty$

# Predicting distributions from topological duality



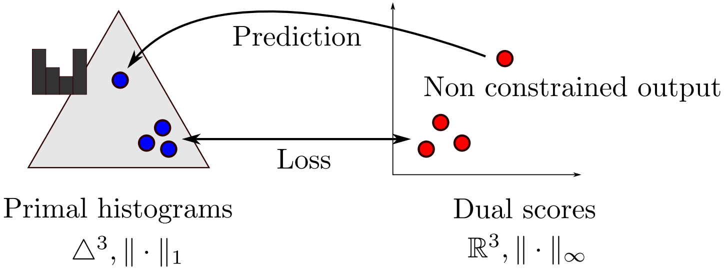

Primal histograms

$\triangle^3, \|\cdot\|_1$

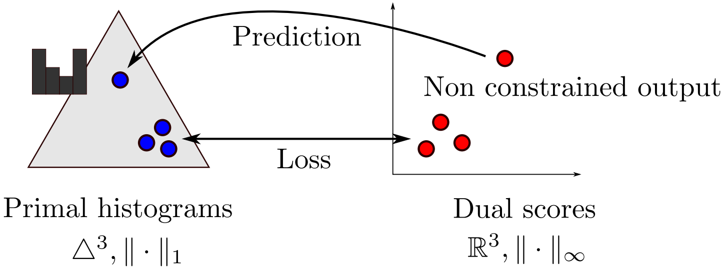
Dual scores

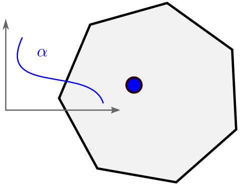
$\mathbb{R}^3, \|\cdot\|_\infty$

# Predicting distributions from topological duality



# Predicting distributions from topological duality

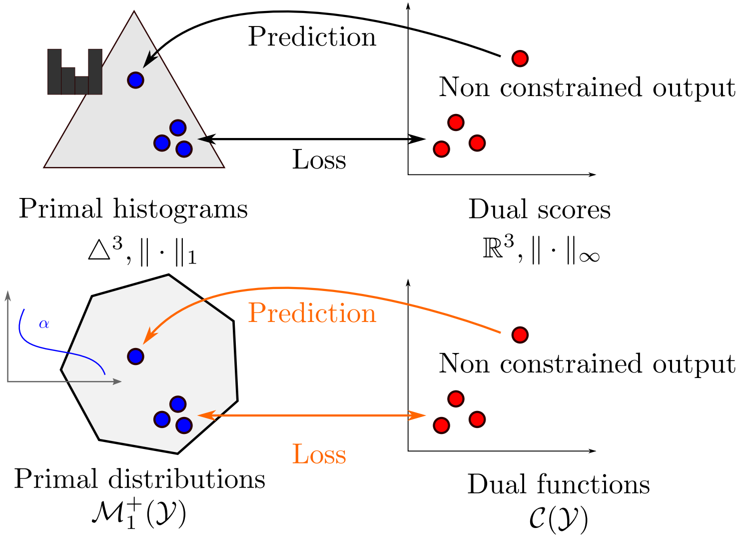


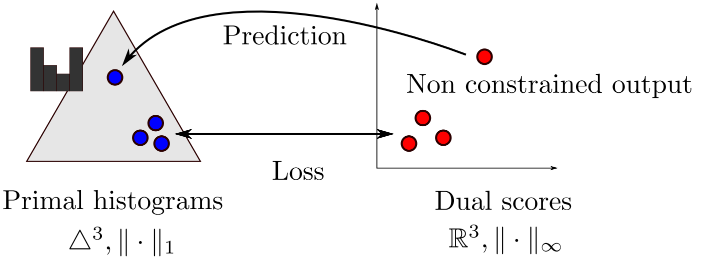

Primal distributions

$\mathcal{M}_1^+(\mathcal{Y})$

# Predicting distributions from topological duality



# All you need is a convex functional

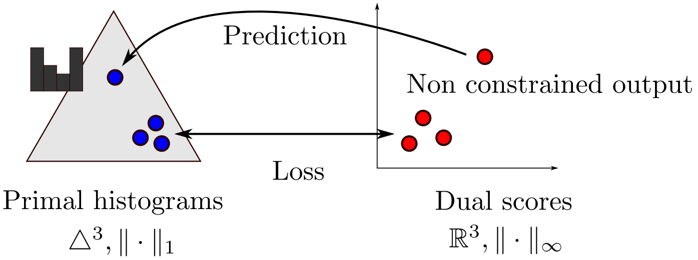



**Fenchel-Young losses**[4 5]: Convex function $\Omega : \triangle^d \to \mathbb{R}$

[4] John C. Duchi et al. "Multiclass Classification, Information, Divergence, and Surrogate Risk". In: *Annals of Statistics* (2018).

[5] Mathieu Blondel et al. "Learning Classifiers with Fenchel-Young Losses: Generalized Entropies, Margins, and Algorithms". In: *AISTATS*. 2019.

# All you need is a convex functional



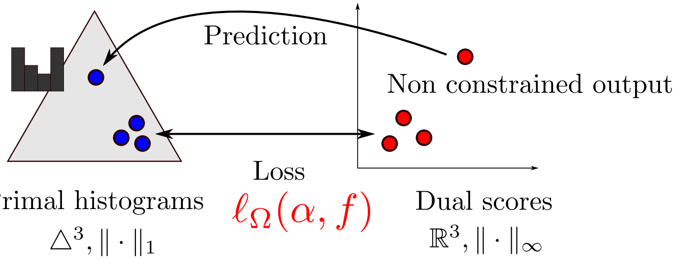

**Fenchel-Young losses**[4 5]: Convex function $\Omega : \triangle^d \to \mathbb{R}$ and conjugate

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

[4] John C. Duchi et al. "Multiclass Classification, Information, Divergence, and Surrogate Risk". In: *Annals of Statistics* (2018).

[5] Mathieu Blondel et al. "Learning Classifiers with Fenchel-Young Losses: Generalized Entropies, Margins, and Algorithms". In: *AISTATS*. 2019.

# All you need is a convex functional



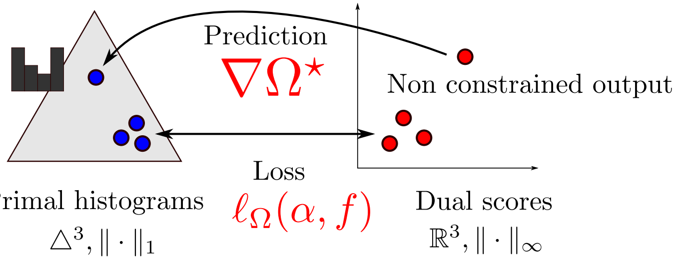

**Fenchel-Young losses**[4 5]: Convex function $\Omega : \triangle^d \to \mathbb{R}$ and conjugate

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

[4] John C. Duchi et al. "Multiclass Classification, Information, Divergence, and Surrogate Risk". In: *Annals of Statistics* (2018).

[5] Mathieu Blondel et al. "Learning Classifiers with Fenchel-Young Losses: Generalized Entropies, Margins, and Algorithms". In: *AISTATS*. 2019.

# All you need is a convex functional




**Fenchel-Young losses**[4 5]: Convex function $\Omega : \triangle^d \to \mathbb{R}$ and 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 between dual and primal

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

[4] John C. Duchi et al. "Multiclass Classification, Information, Divergence, and Surrogate Risk". In: *Annals of Statistics* (2018).

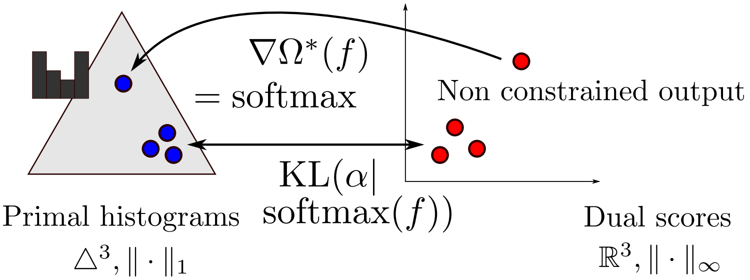
[5] Mathieu Blondel et al. "Learning Classifiers with Fenchel-Young Losses: Generalized Entropies, Margins, and Algorithms". In: *AISTATS*. 2019.

# Discrete canonical example: Shannon entropy

$$\Omega(\alpha) = -H(\alpha) = \sum_{i=1}^{d} \alpha_i \log \alpha_i \qquad \Omega^*(f) = \mathrm{logsumexp}(f)$$

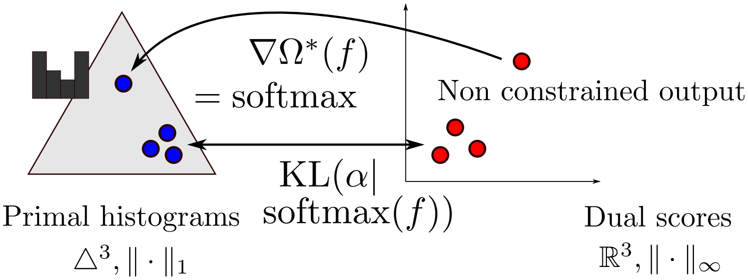
# Discrete canonical example: Shannon entropy

$$\Omega(\alpha) = -H(\alpha) = \sum_{i=1}^{d} \alpha_i \log \alpha_i \qquad \Omega^*(f) = \mathrm{logsumexp}(f)$$



# Discrete canonical example: Shannon entropy

$$\Omega(\alpha) = -H(\alpha) = \sum_{i=1}^{d} \alpha_i \log \alpha_i \qquad \Omega^*(f) = \mathrm{logsumexp}(f)$$

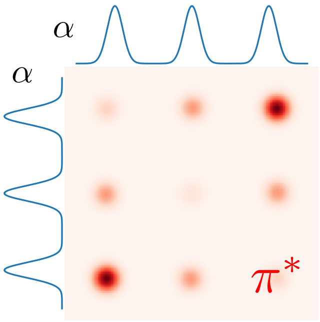



Not defined on continuous distributions, cost-agnostic

# Sinkhorn entropies from regularized optimal transport

**Self regularized optimal transportation distance:**

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

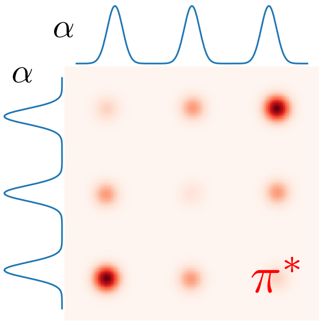


# Sinkhorn entropies from regularized optimal transport

**Self regularized optimal transportation distance:**

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

**Continuous convex**

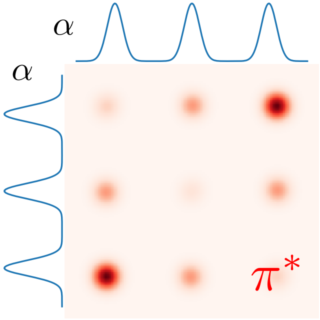


# Sinkhorn entropies from regularized optimal transport

**Self regularized optimal transportation distance:**

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

**Continuous convex**




**Special cases**

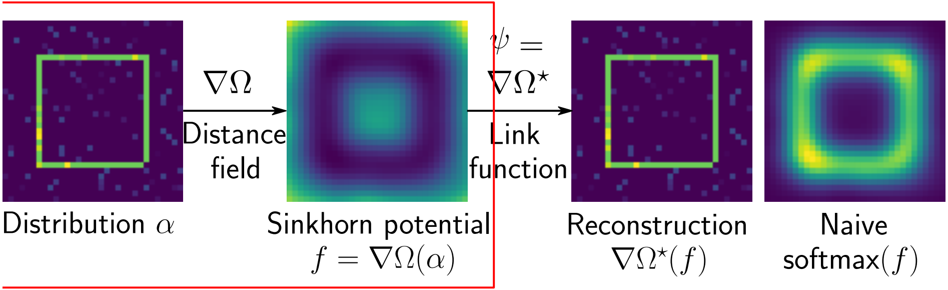
$\varepsilon \to \infty$: MMD autocorrelation

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

Shannon entropy

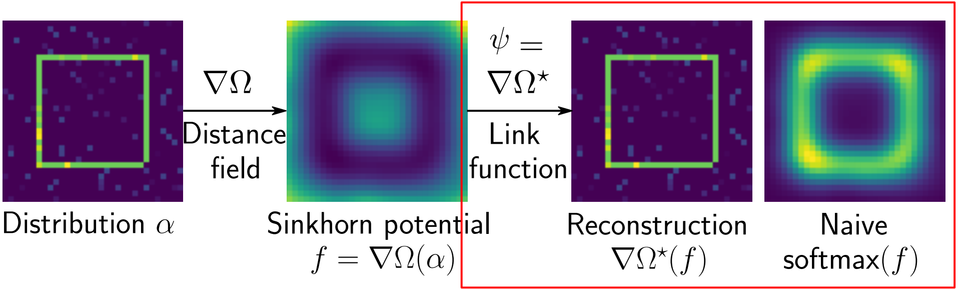
Gini index

# Dual mapping from Sinkhorn negentropy




Sinkhorn entropy: $$\Omega(\alpha) = -\max_{f \in \mathcal{C}(\mathcal{Y})} \langle \alpha, f \rangle - \log \langle \alpha \otimes \alpha, e^{\frac{f \oplus f - C}{2}} \rangle$$

# Returning to primal: geometric softmax




$$\Omega^* = \text{g-logsumexp} : f \mapsto -\log \min_{\alpha \in \mathcal{M}_1^+(\mathcal{Y})} \langle \alpha \otimes \alpha, \exp(-\frac{f \oplus f + {\color{red}C}}{2}) \rangle$$

$\nabla\Omega^* =$ geometric-softmax. Minimizes a **simple quadratic.**

# Geomtric loss construction and computation

**Training with the geometric logistic loss:**

$$\ell_C(\alpha, f) = \text{geometric-LSE}_C(f) + \text{sinkhorn-negentropy}_C(\alpha) - \langle \alpha, f \rangle$$

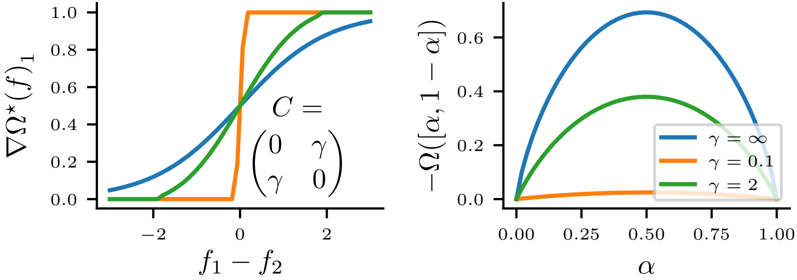
**Tractable discrete $\mathcal{Y}$:** use mirror descent/L-BFGS

- $10\times$ as costly as a softmax
- Backpropagation: $\nabla\Omega^\star = \text{geometric-softmax}$

**Continuous case:**

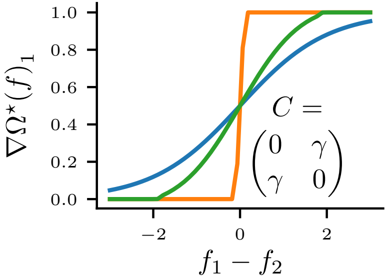
- Frank-Wolfe scheme, adding one Dirac at each iteration

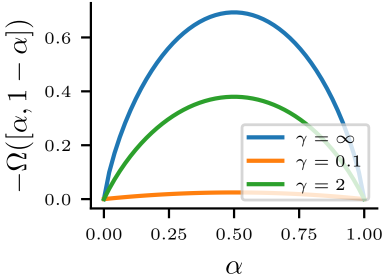
# Properties of the geometric-softmax




- $\nabla\Omega^\star$ **returns from Sinkhorn potentials**: $\nabla\Omega^\star \circ \nabla\Omega = \mathrm{Id}$

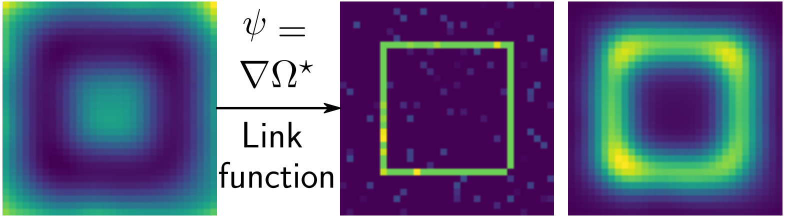
# Properties of the geometric-softmax




- $\nabla\Omega^\star$ **returns from Sinkhorn potentials**: $\nabla\Omega^\star \circ \nabla\Omega = \mathrm{Id}$
- $\nabla\Omega \circ \nabla\Omega^*$ **projects** $f$ onto $\mathcal{F}$ the set of symmetric Sinkhorn potentials
- **Sparse** $\nabla\Omega^*(f) \leftarrow$ minimization on the simplex

# Properties of the geometric-softmax




Sinkhorn potential $f = \nabla\Omega(\alpha)$ | Reconstruction $\nabla\Omega^\star(f)$ | Naive softmax$(f)$

$\varepsilon \to 0$ Mode finding

$\varepsilon \to \infty$ Positive deconvolution

# Consistent learning with the geometric logistic loss

**Bregman divergence** from Sinkhorn negentropy:

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

# Consistent learning with the geometric logistic loss

**Bregman divergence** from Sinkhorn negentropy:

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

Sample distribution $(x_i, \alpha_i)_i \in \mathcal{X} \times \mathcal{M}_1^+(\mathcal{Y})$.

# Consistent learning with the geometric logistic loss

**Bregman divergence** from Sinkhorn negentropy:

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

Sample distribution $(x_i, \alpha_i)_i \in \mathcal{X} \times \mathcal{M}_1^+(\mathcal{Y})$.

**Fisher consistency:**

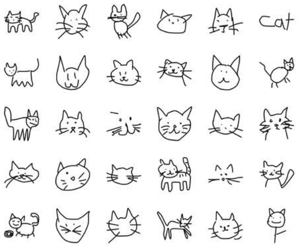
$$\min_{\beta : \mathcal{X} \to \mathcal{M}_1^+(\mathcal{Y})} \mathbb{E}\left[D_g\left(\alpha, \beta\left(x\right)\right)\right] = \min_{g : \mathcal{X} \to \mathcal{C}(\mathcal{Y})} \mathbb{E}\left[\ell_g\left(\alpha, \nabla\Omega^*\left(g\left(x\right)\right)\right)\right]$$

# Applications: variational auto-encoder

**Goal:** Generate nearly 1D distribution in 2D images

- Dataset: Google Quickdraw
- Traditional sigmoid activation layer

$\leftarrow$ replaced by **geometric softmax**

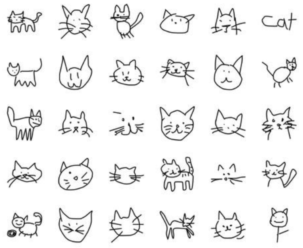


# Applications: variational auto-encoder

**Goal:** Generate nearly 1D distribution in 2D images

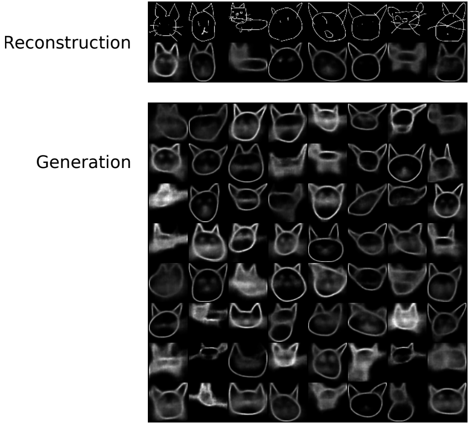
- Dataset: Google Quickdraw
- Traditional sigmoid activation layer

$\leftarrow$ replaced by **geometric softmax**

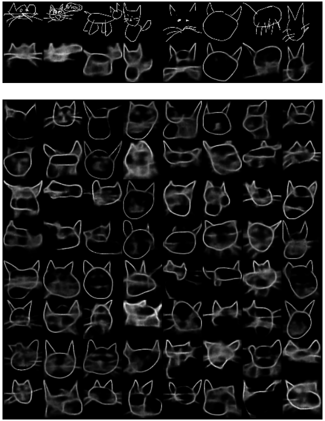



- Deconvolutional effect
- Cost-informed non-linearity

# Applications: variational auto-encoders




softmax



g-softmax

Better defined generated images

# Conclusion

**Geometric softmax:**

- New loss and projector onto output probabilities
- Discrete/continuous, aware of a cost between outputs
- Fenchel duality in Banach spaces + regularized optimal transport
- Application in VAE and ordinal regression

# Conclusion

**Geometric softmax:**

- New loss and projector onto output probabilities
- Discrete/continuous, aware of a cost between outputs
- Fenchel duality in Banach spaces + regularized optimal transport
- Application in VAE and ordinal regression

**Future directions:**

- How to improve computation methods (continuous FW)
- Geometric logisitic loss in super resolution[6]

[6] Nicholas Boyd et al. "DeepLoco: Fast 3D localization microscopy using neural networks". In: *BioRxiv* (2018), p. 267096.

Mathieu Blondel

Gabriel Peyré

Poster # 179