Structured Prediction with Projection Oracles

Mathieu Blondel, NTT Communication Science laboratories, Kyoto, Japan

Structured prediction

Target loss

- Goal: Learn a mapping $f\colon \mathcal{X} \to \mathcal{Y}$
- Target loss $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$
- Minimize target expected risk

$$\mathcal{L}(f) \coloneqq \mathbb{E}_{(X,Y)\sim\rho}\, L(f(X), Y)$$

Surrogate loss

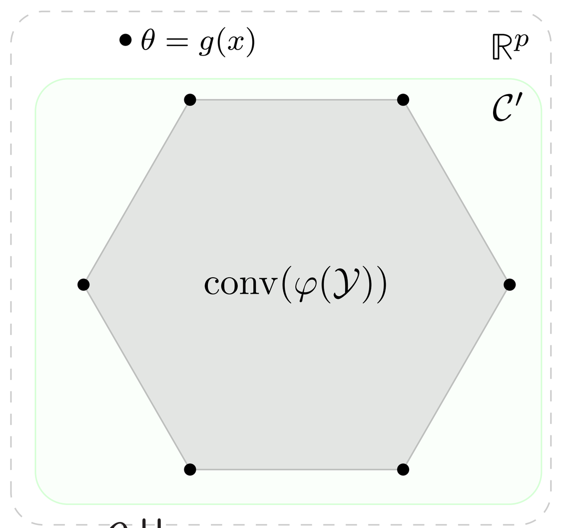
- Vector space $\Theta \subseteq \mathbb{R}^p$
- Label encoding (embedding) $\varphi\colon \mathcal{Y} \to \Theta$
- Model $g\colon \mathcal{X} \to \Theta$, $\theta = g(x)$
- Surrogate loss $S\colon \Theta \times \Theta \to \mathbb{R}$
- Minimize surrogate expected risk

$$\mathcal{S}(g) \coloneqq \mathbb{E}_{(X,Y)\sim\rho}\, S(g(X), \varphi(Y))$$

- Pull-back: decoder $d\colon \Theta \to \mathcal{Y}$
- Example: Maximum A-Posteriori decoder

$$\mathrm{MAP}(\theta) \coloneqq \operatorname*{argmax}_{y \in \mathcal{Y}} \langle \theta, \varphi(y) \rangle$$

Why projections?



- Euclidean projection onto a closed convex set $\mathcal{C}$

$$P_{\mathcal{C}}(\theta) \coloneqq \operatorname*{argmin}_{u \in \mathcal{C}} \|u - \theta\|$$

- As long as $\varphi(y) \in \mathcal{C}$, we have

$$\|\varphi(y) - P_{\mathcal{C}}(\theta)\|^2 \leq \|\varphi(y) - \theta\|^2$$

$\to$ projection achieves smaller squared loss!
- Projections as an output layer

$$x \in \mathcal{X} \xrightarrow[\text{model}]{g} \theta \in \Theta \xrightarrow[\text{projection}]{P_{\mathcal{C}}} u \in \mathcal{C}$$

Loss functions

Composite loss

- $\ell_{\mathcal{C}}(\theta, y) \coloneqq \frac{1}{2}\|P_{\mathcal{C}}(\theta) - \varphi(y)\|^2$ Non-convex!

Fenchel-Young losses

- Conjugate $\Omega^*(\theta) \coloneqq \max_{\mu \in \mathrm{dom}(\Omega)} \langle \mu, \theta \rangle - \Omega(\mu)$
- Fenchel-Young loss generated by Ω

$$S_\Omega(\theta, y) \coloneqq \Omega^*(\theta) + \Omega(\varphi(y)) - \langle \theta, \varphi(y) \rangle$$

- Convex, smooth if Ω strongly convex, non-negative

Projection loss

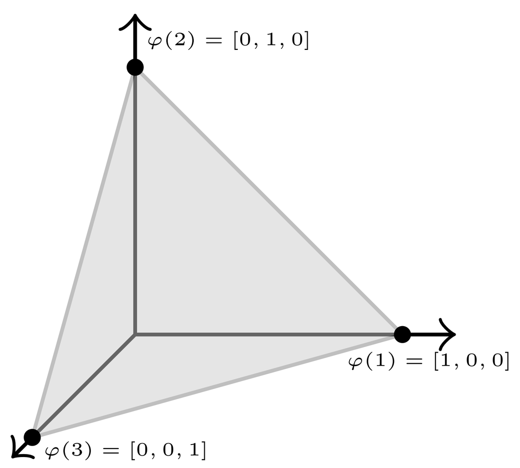
- Choose $\Omega = \frac{1}{2}\|\cdot\|^2 + I_{\mathcal{C}}$ and define shorthand

$$S_{\mathcal{C}}(\theta, y) \coloneqq S_{\frac{1}{2}\|\cdot\|^2 + I_{\mathcal{C}}}(\theta, y)$$

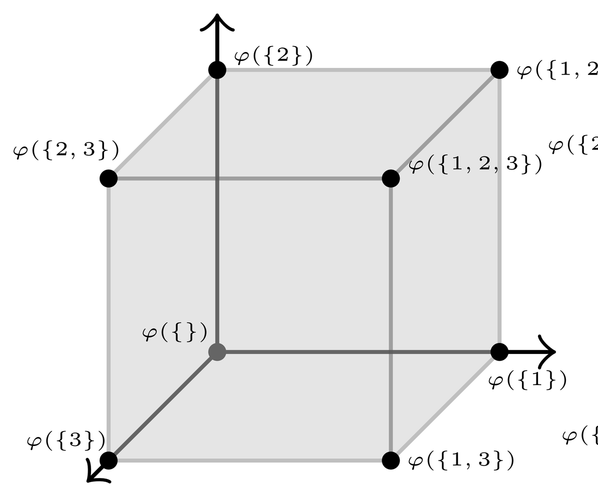
- Zero loss: $S_{\mathcal{C}}(\theta, y) = 0 \Leftrightarrow P_{\mathcal{C}}(\theta) = \varphi(y)$
- Convex upper-bound: $\ell_{\mathcal{C}}(\theta, y) \leq S_{\mathcal{C}}(\theta, y)$
- Smaller sets enjoy smaller loss

$$S_{\mathcal{C}}(\theta, y) \leq S_{\mathcal{C}'}(\theta, y) \quad \forall \mathcal{C} \subseteq \mathcal{C}'$$

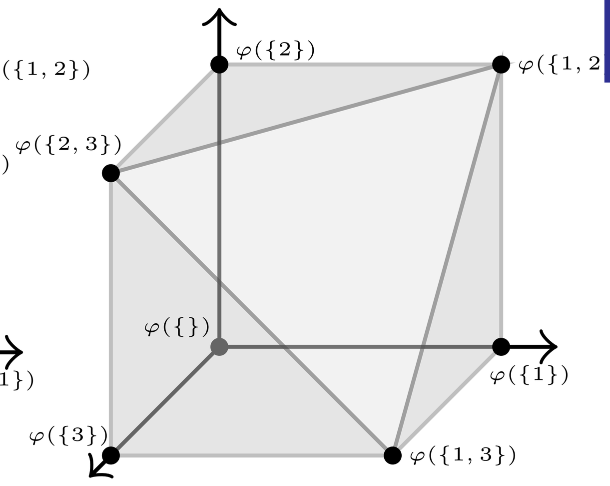
- Smallest convex set: the convex hull of $\varphi(\mathcal{Y})$



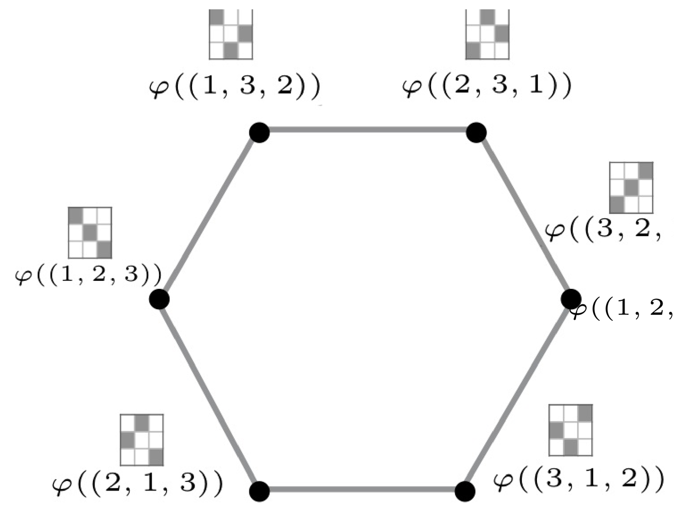
Probability simplex



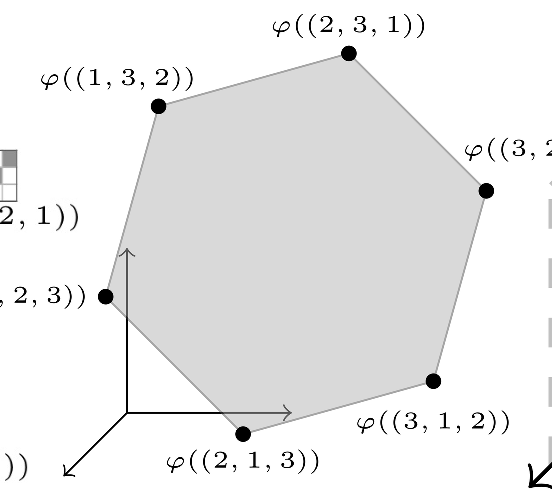
Unit cube



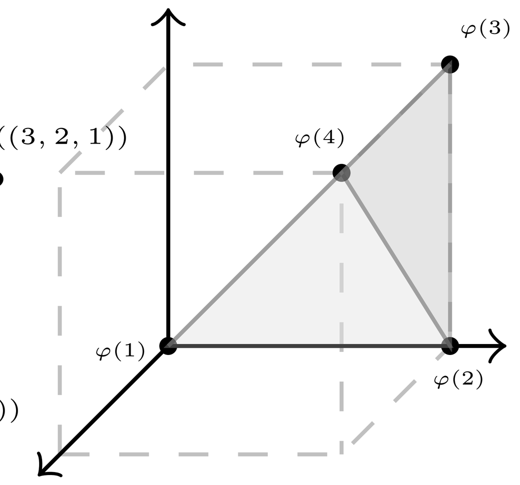
Knapsack polytope



Birkhoff polytope



Permutahedron



Order simplex

Calibration analysis

- Affine decomposition

$$L(\widehat{y}, y) = \langle \varphi(\widehat{y}), V\varphi(y) + b \rangle + c(y)$$

- Decoding calibrated for loss L

$$\widehat{y}_L(u) \coloneqq \operatorname*{argmin}_{y' \in \mathcal{Y}} \langle \varphi(y'), Vu + b \rangle = \mathrm{MAP}(-Vu - b)$$

- Analyzed prediction pipeline

$$x \in \mathcal{X} \xrightarrow[\text{model}]{g} \theta \in \Theta \xrightarrow[\text{projection}]{P_{\mathcal{C}}} u \in \mathcal{C} \xrightarrow[\text{decoding}]{\widehat{y}_L} \widehat{y} \in \mathcal{Y}$$

- Excess of risks

$$\delta\mathcal{L}(f) \coloneqq \mathcal{L}(f) - \mathcal{L}(f^\star), \delta\mathcal{S}_{\mathcal{C}}(g) \coloneqq \mathcal{S}_{\mathcal{C}}(g) - \mathcal{S}_{\mathcal{C}}(g^\star)$$

- Theorem

$$\forall g\colon \mathcal{X} \to \Theta : \quad \frac{\delta\mathcal{L}(\widehat{y}_L \circ P_{\mathcal{C}} \circ g)^2}{8\beta\sigma^2} \leq \delta\mathcal{S}_{\mathcal{C}}(g)$$

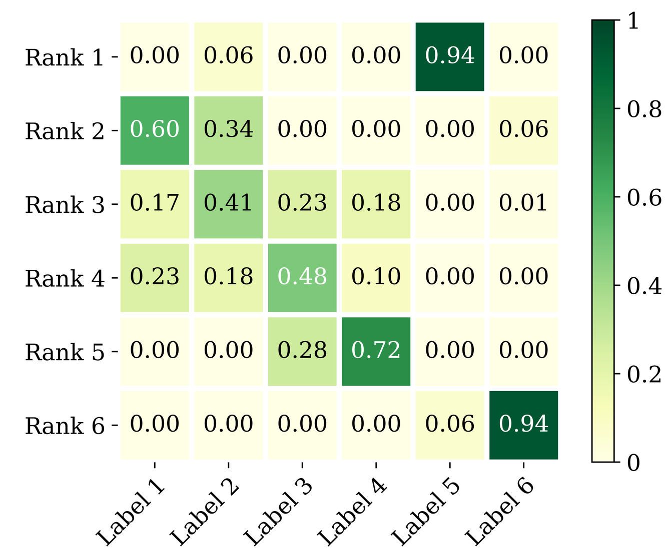
$1/\beta$: strong-convexity constant of Ω w.r.t. $\|\cdot\|$

$\sigma \coloneqq \sup_{\widehat{y} \in \mathcal{Y}} \|V^\top \varphi(\widehat{y})\|_*$

Experiment: label ranking

Projection	$[0,1]^{k\times k}$	$\triangle^{k\times k}$	$\mathbb{R}^{k\times k}$	$[0,1]^{k\times k}$	$\triangle^{k\times k}$	$\mathcal{M}$
Decoding	$[0,1]^{k\times k}$	$\triangle^{k\times k}$	$\mathcal{M}$	$\mathcal{M}$	$\mathcal{M}$	$\mathcal{M}$
Authorship	12.83	5.62	5.70	5.18	5.70	**5.10**
Glass	24.35	5.43	7.11	5.68	5.04	**4.65**
Iris	27.78	10.37	19.26	4.44	**1.48**	2.96
Vehicle	26.36	7.43	9.04	7.57	6.99	**5.88**
Vowel	43.71	9.65	10.57	9.56	9.18	**8.76**
Wine	10.19	1.85	**1.23**	1.85	1.85	1.85

$\mathcal{M}$: Birkhoff polytope



Multi-label classification and ordinal regression experiments in the paper!

github.com/mblondel/projection-losses