Structured Prediction with Projection Oracles

Mathieu Blondel

NTT, Kyoto, Japan

June 19th, 2019

Outline

1. Background

2. Proposed framework

3. Experiments

Outline

1. Background

2. Proposed framework

3. Experiments

Structured prediction

Goal

Learn a mapping from input space to output space

$$f\colon \mathcal{X} \to \mathcal{Y}$$

Structured prediction

Goal

Learn a mapping from input space to output space

$$f\colon \mathcal{X} \to \mathcal{Y}$$ ← exponentially large

Structured prediction

Goal

Learn a mapping from input space to output space

$$f\colon \mathcal{X} \to \mathcal{Y}$$ ← exponentially large

Decomposition

Typically assume $f = dec \circ g$

$$x \in \mathcal{X} \xrightarrow[\text{model}]{g} \theta \in \Theta = \mathbb{R}^d \xrightarrow[\text{decoding}]{dec} \hat{y} \in \mathcal{Y}$$

Structured prediction

Goal

Learn a mapping from input space to output space

$$f\colon \mathcal{X} \to \mathcal{Y}$$ ← exponentially large

Decomposition

Typically assume $f = dec \circ g$

$$x \in \mathcal{X} \xrightarrow[\text{model}]{g} \theta \in \Theta = \mathbb{R}^d \xrightarrow[\text{decoding}]{dec} \hat{y} \in \mathcal{Y}$$

Target loss risk

$$\mathcal{L}(f) \triangleq \mathbb{E}_{(X,Y)\sim p}\, L(f(X), Y) \qquad L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$$

Structured prediction

Goal

Learn a mapping from input space to output space

$$f\colon \mathcal{X} \to \mathcal{Y}$$

← exponentially large

Decomposition

Typically assume $f = dec \circ g$

$$x \in \mathcal{X} \xrightarrow[\text{model}]{g} \theta \in \Theta = \mathbb{R}^d \xrightarrow[\text{decoding}]{dec} \hat{y} \in \mathcal{Y}$$

Non-convex, discontinuous!

Target loss risk

$$\mathcal{L}(f) \triangleq \mathbb{E}_{(X,Y)\sim p}\, L(f(X), Y) \qquad L\colon \mathcal{Y} \times \mathcal{Y}$$

Surrogate losses

Surrogate loss risk

$$\mathcal{S}(g) \triangleq \mathbb{E}_{(X,Y)\sim p}\, S(g(X), Y) \qquad S : \Theta \times \mathcal{Y} \to \mathbb{R}_+$$

Surrogate losses

Surrogate loss risk

$$\mathcal{S}(g) \triangleq \mathbb{E}_{(X,Y)\sim p} \, S(g(X), Y) \qquad S : \Theta \times \mathcal{Y} \to \mathbb{R}_+$$

Fisher consistency

$$\mathcal{S}(g_n) \to \inf_{g \in \mathcal{G}} \mathcal{S}(g) \longrightarrow \mathcal{L}(dec \circ g_n) \to \inf_{g \in \mathcal{G}} \mathcal{L}(dec \circ g)$$

Surrogate losses

Surrogate loss risk

$$\mathcal{S}(g) \triangleq \mathbb{E}_{(X,Y)\sim p}\, S(g(X), Y) \qquad S : \Theta \times \mathcal{Y} \to \mathbb{R}_+$$

Fisher consistency

$$\mathcal{S}(g_n) \to \inf_{g \in \mathcal{G}} \mathcal{S}(g) \longrightarrow \mathcal{L}(dec \circ g_n) \to \inf_{g \in \mathcal{G}} \mathcal{L}(dec \circ g)$$

Extensively studied in the multiclass setting [Zhang 2004, Bartlett et al. 2006]

Only recently studied in the structured setting [Ciliberto et al 2016, Osokin et al. 2017, Nowak-Vila et al. 2019]

Structured perceptron loss

[Collins, 2002]

Loss

$$S(\theta, y) \triangleq \max_{y' \in \mathcal{Y}} \langle \theta, \varphi(y') \rangle - \langle \theta, \varphi(y) \rangle$$

$$\varphi : \mathcal{Y} \to \mathbb{R}^d$$

Structured perceptron loss

[Collins, 2002]

Loss

$$S(\theta, y) \triangleq \max_{y' \in \mathcal{Y}} \langle \theta, \varphi(y') \rangle - \langle \theta, \varphi(y) \rangle$$

$$\varphi : \mathcal{Y} \to \mathbb{R}^d$$

Training oracle

$$MAP(\theta) \triangleq \arg\max_{y \in \mathcal{Y}} \langle \theta, \varphi(y) \rangle$$

Structured perceptron loss

[Collins, 2002]

Loss

$$S(\theta, y) \triangleq \max_{y' \in \mathcal{Y}} \langle \theta, \varphi(y') \rangle - \langle \theta, \varphi(y) \rangle$$

$$\varphi : \mathcal{Y} \to \mathbb{R}^d$$

Training oracle

$$MAP(\theta) \triangleq \arg\max_{y \in \mathcal{Y}} \langle \theta, \varphi(y) \rangle$$

Decoding oracle

$$dec = MAP$$

Structured perceptron loss

[Collins, 2002]

Loss

$$S(\theta, y) \triangleq \max_{y' \in \mathcal{Y}} \langle \theta, \varphi(y') \rangle - \langle \theta, \varphi(y) \rangle$$

$$\varphi : \mathcal{Y} \to \mathbb{R}^d$$

Training oracle

$$MAP(\theta) \triangleq \arg\max_{y \in \mathcal{Y}} \langle \theta, \varphi(y) \rangle$$

Decoding oracle

$$dec = MAP$$

× **Not smooth**

× **Not consistent**

Structured hinge loss

[Tsochantaridis et al., 2005]

Loss

$$S(\theta, y) \triangleq \max_{y' \in \mathcal{Y}} \; L(y', y) + \langle \theta, \varphi(y') \rangle - \langle \theta, \varphi(y) \rangle$$

Structured hinge loss

[Tsochantaridis et al., 2005]

Loss

$$S(\theta, y) \triangleq \max_{y' \in \mathcal{Y}} \; L(y', y) + \langle \theta, \varphi(y') \rangle - \langle \theta, \varphi(y) \rangle$$

Training oracle

$$\arg\max_{y' \in \mathcal{Y}} \; L(y, y') + \langle \theta, \varphi(y') \rangle$$

Structured hinge loss

[Tsochantaridis et al., 2005]

Loss

$$S(\theta, y) \triangleq \max_{y' \in \mathcal{Y}} \; L(y', y) + \langle \theta, \varphi(y') \rangle - \langle \theta, \varphi(y) \rangle$$

Training oracle

$$\arg\max_{y' \in \mathcal{Y}} \; L(y, y') + \langle \theta, \varphi(y') \rangle$$

Decoding oracle

$$dec = MAP$$

Structured hinge loss

[Tsochantaridis et al., 2005]

Loss

$$S(\theta, y) \triangleq \max_{y' \in \mathcal{Y}} \; L(y', y) + \langle \theta, \varphi(y') \rangle - \langle \theta, \varphi(y) \rangle$$

Training oracle

$$\arg\max_{y' \in \mathcal{Y}} \; L(y, y') + \langle \theta, \varphi(y') \rangle$$

Decoding oracle

$$dec = MAP$$

× Not smooth

× Not consistent

Conditional Random Field (CRF) loss

[Lafferty et al., 2001]

Loss

$$S(\theta, y) \triangleq \log \sum_{y' \in \mathcal{Y}} e^{\langle \theta, \varphi(y') \rangle} - \langle \theta, \varphi(y) \rangle$$

Conditional Random Field (CRF) loss

[Lafferty et al., 2001]

Loss

$$S(\theta, y) \triangleq \log \sum_{y' \in \mathcal{Y}} e^{\langle \theta, \varphi(y') \rangle} - \langle \theta, \varphi(y) \rangle$$

Training oracle

$$marginal(\theta) \triangleq \mathbb{E}_{Y \sim p(\cdot;\theta)}[\varphi(Y)]$$

$$p(y; \theta) \propto \exp\langle \varphi(y), \theta \rangle$$

Conditional Random Field (CRF) loss

[Lafferty et al., 2001]

Loss

$$S(\theta, y) \triangleq \log \sum_{y' \in \mathcal{Y}} e^{\langle \theta, \varphi(y') \rangle} - \langle \theta, \varphi(y) \rangle$$

Training oracle

$$marginal(\theta) \triangleq \mathbb{E}_{Y \sim p(\cdot;\theta)}[\varphi(Y)]$$

$$p(y; \theta) \propto \exp\langle \varphi(y), \theta \rangle$$

Decoding oracle

$$MAP$$

Calibrated decoding

Conditional Random Field (CRF) loss

[Lafferty et al., 2001]

Loss

$$S(\theta, y) \triangleq \log \sum_{y' \in \mathcal{Y}} e^{\langle \theta, \varphi(y') \rangle} - \langle \theta, \varphi(y) \rangle$$

Training oracle

$$marginal(\theta) \triangleq \mathbb{E}_{Y \sim p(\cdot;\theta)}[\varphi(Y)]$$

$$p(y;\theta) \propto \exp\langle \varphi(y), \theta \rangle$$

Decoding oracle

$$MAP$$

Calibrated decoding

✓ Smooth

✓ Consistent (w/ calibrated decoding) [Nowak-Vila et al., 2019]

Conditional Random Field (CRF) loss

[Lafferty et al., 2001]

Loss

$$S(\theta, y) \triangleq \log \sum_{y' \in \mathcal{Y}} e^{\langle \theta, \varphi(y') \rangle} - \langle \theta, \varphi(y) \rangle$$

Training oracle

$$marginal(\theta) \triangleq \mathbb{E}_{Y \sim p(\cdot;\theta)}[\varphi(Y)]$$

$$p(y;\theta) \propto \exp\langle \varphi(y), \theta \rangle$$

Decoding oracle

$$MAP$$

Calibrated decoding

✓ Smooth

✓ Consistent (w/ calibrated decoding) [Nowak-Vila et al., 2019]

× Marginal inference is intractable for some tasks

Squared loss

[Ciliberto et al., 2016]

Loss

$$S(\theta, y) \triangleq \frac{1}{2} \|\varphi(y) - \theta\|^2$$

Squared loss

[Ciliberto et al., 2016]

Loss

$$S(\theta, y) \triangleq \frac{1}{2}\|\varphi(y) - \theta\|^2$$

Training oracle

None!

Squared loss

[Ciliberto et al., 2016]

Loss

$$S(\theta, y) \triangleq \frac{1}{2}\|\varphi(y) - \theta\|^2$$

Training oracle

None!

Decoding oracle

Calibrated decoding

Squared loss

[Ciliberto et al., 2016]

Loss

$$S(\theta, y) \triangleq \frac{1}{2}\|\varphi(y) - \theta\|^2$$

Training oracle

None!

Decoding oracle

Calibrated decoding

✓ **Smooth**

✓ **Consistent (when using calibrated decoding)**

Squared loss

[Ciliberto et al., 2016]

Loss

$$S(\theta, y) \triangleq \frac{1}{2}\|\varphi(y) - \theta\|^2$$

Training oracle

None!

Decoding oracle

Calibrated decoding

✓ Smooth

✓ Consistent (when using calibrated decoding)

✗ Ignores structural information at training time

Summary

Loss	Training oracle	Decoding	Smooth	Consistent
Perceptron	MAP	MAP	No	No
SVM	Loss-augmented MAP	MAP	No	No
CRF	Marginal	MAP Calibrated decoding	Yes	No Yes
Squared	None	Calibrated decoding	Yes	Yes

Summary

Loss	Training oracle	Decoding	Smooth	Consistent
Perceptron	MAP	MAP	No	No
SVM	Loss-augmented MAP	MAP	No	No
CRF	Marginal	MAP Calibrated decoding	Yes	No Yes
Squared	None	Calibrated decoding	Yes	Yes
Proposed	**Projection**	**Calibrated decoding**	**Yes**	**Yes**

Outline

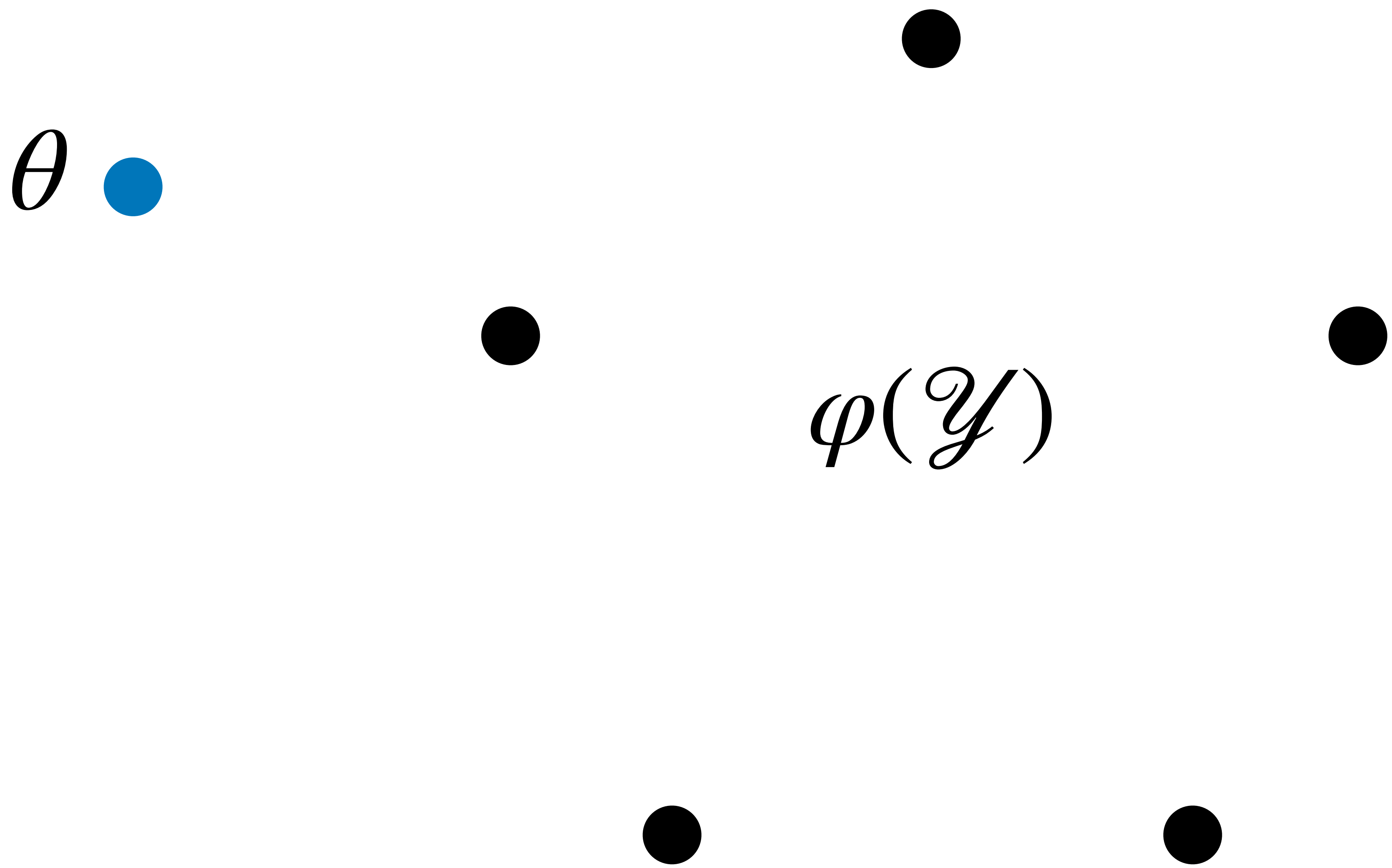
1. Background

2. Proposed framework

3. Experiments

Proposed inference pipeline

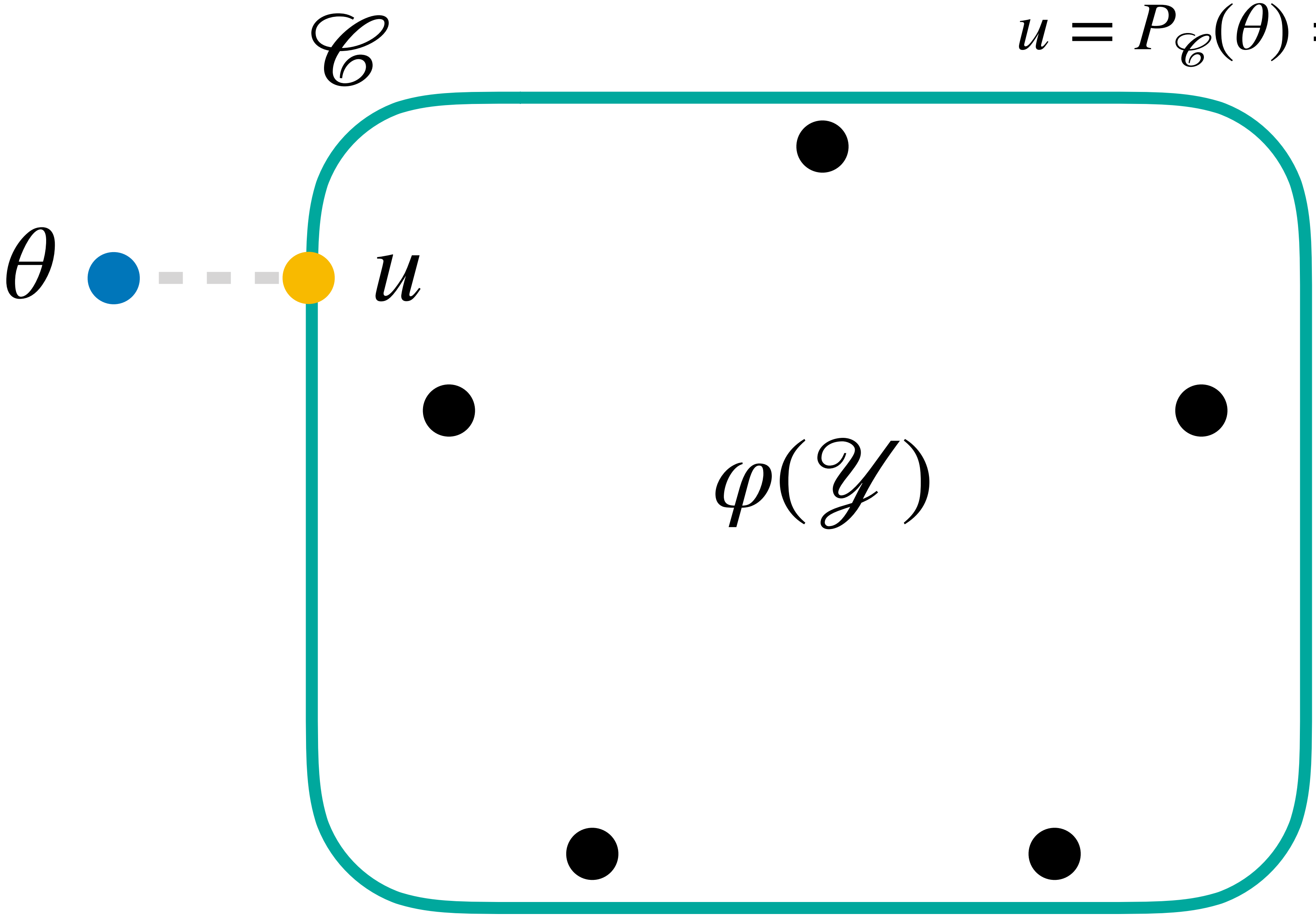
$$x \in \mathscr{X} \xrightarrow[\text{model}]{g} \theta$$



Proposed inference pipeline

$$x \in \mathscr{X} \xrightarrow[\text{model}]{g} \theta \xrightarrow[\text{Projection}]{P_{\mathscr{C}}} u \in \mathscr{C}$$

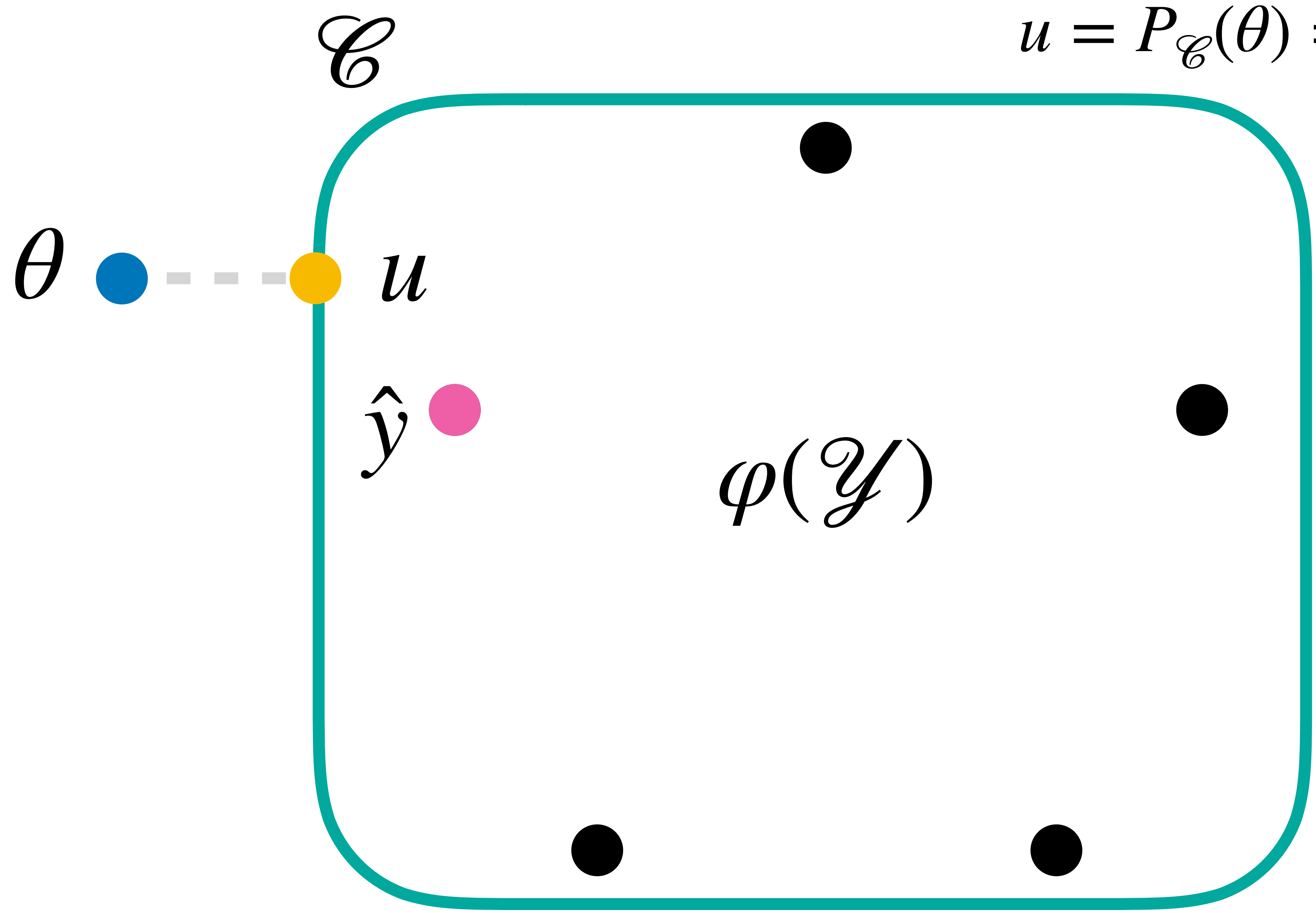
$$u = P_{\mathscr{C}}(\theta) \triangleq \arg\min_{\mu \in \mathscr{C}} \|\mu - \theta\|^2$$



Proposed inference pipeline

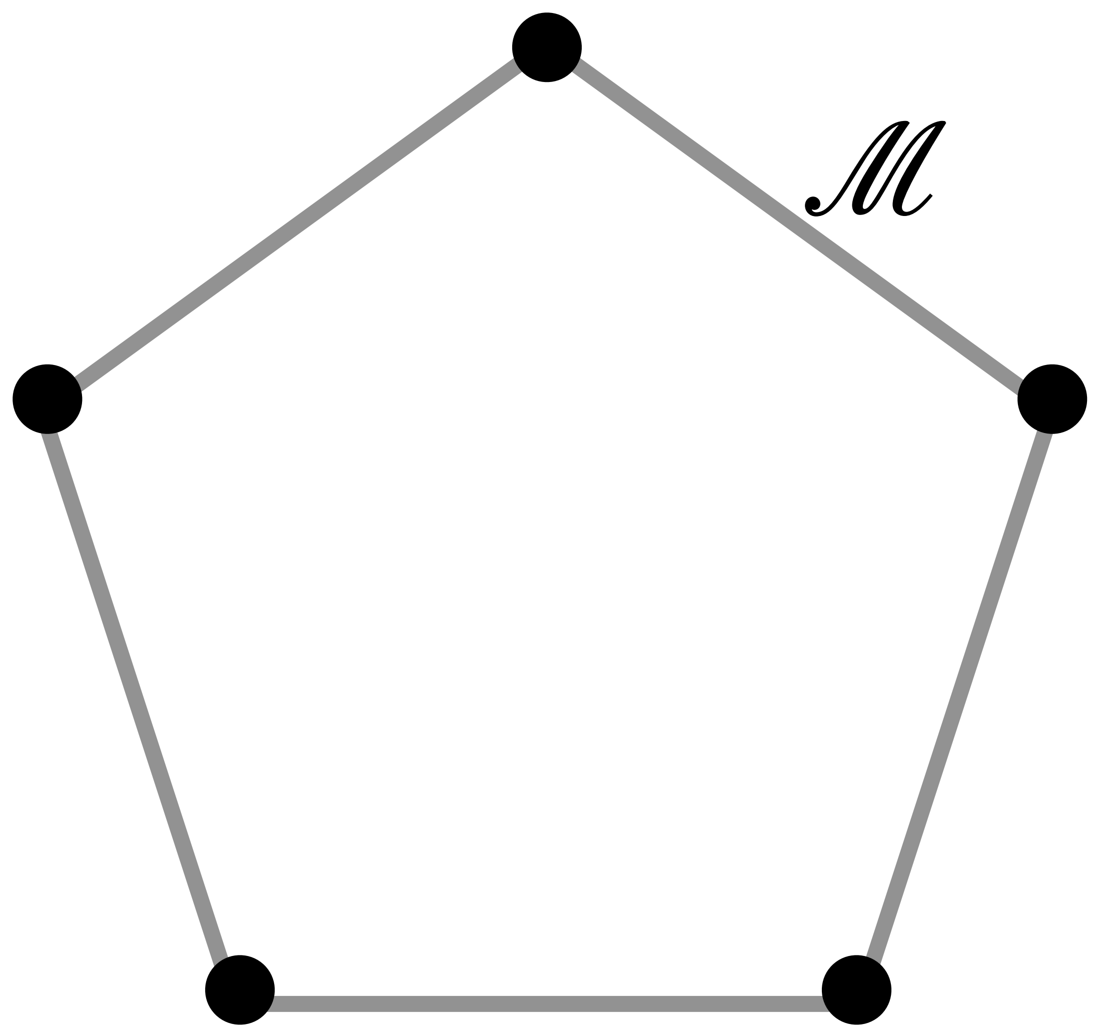
$$x \in \mathscr{X} \xrightarrow[\text{model}]{g} \theta \xrightarrow[\text{Projection}]{P_{\mathscr{C}}} u \in \mathscr{C} \xrightarrow[\text{Calibrated decoding}]{\hat{y}_L} \hat{y} \in \mathscr{Y}$$

$$u = P_{\mathscr{C}}(\theta) \triangleq \underset{\mu \in \mathscr{C}}{\arg\min} \|\mu - \theta\|^2$$



Choice of the convex set

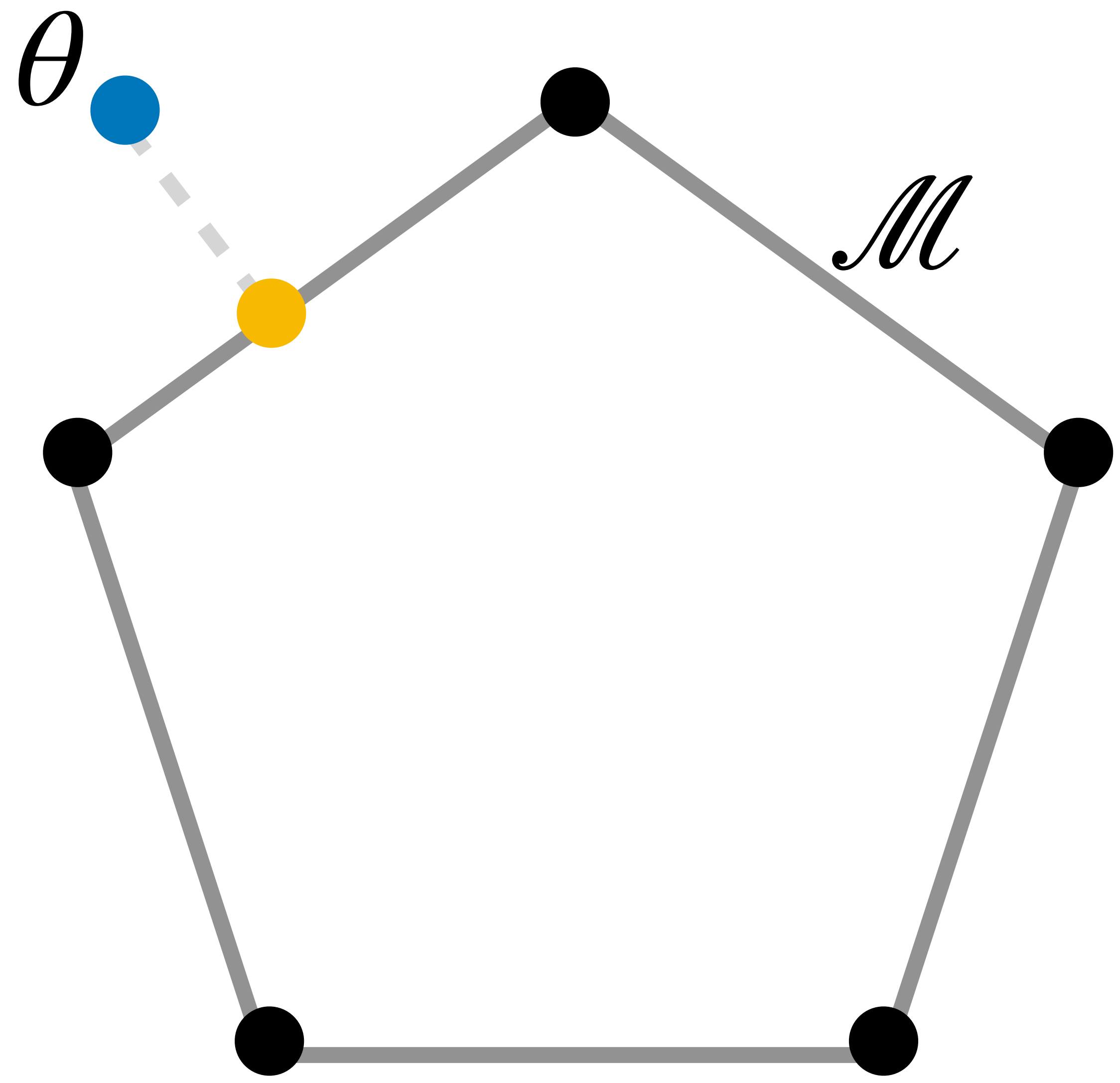
Smallest convex set = **convex hull** (a.k.a. marginal polytope)



$$\mathscr{M} \triangleq conv(\varphi(\mathscr{Y}))$$

Choice of the convex set

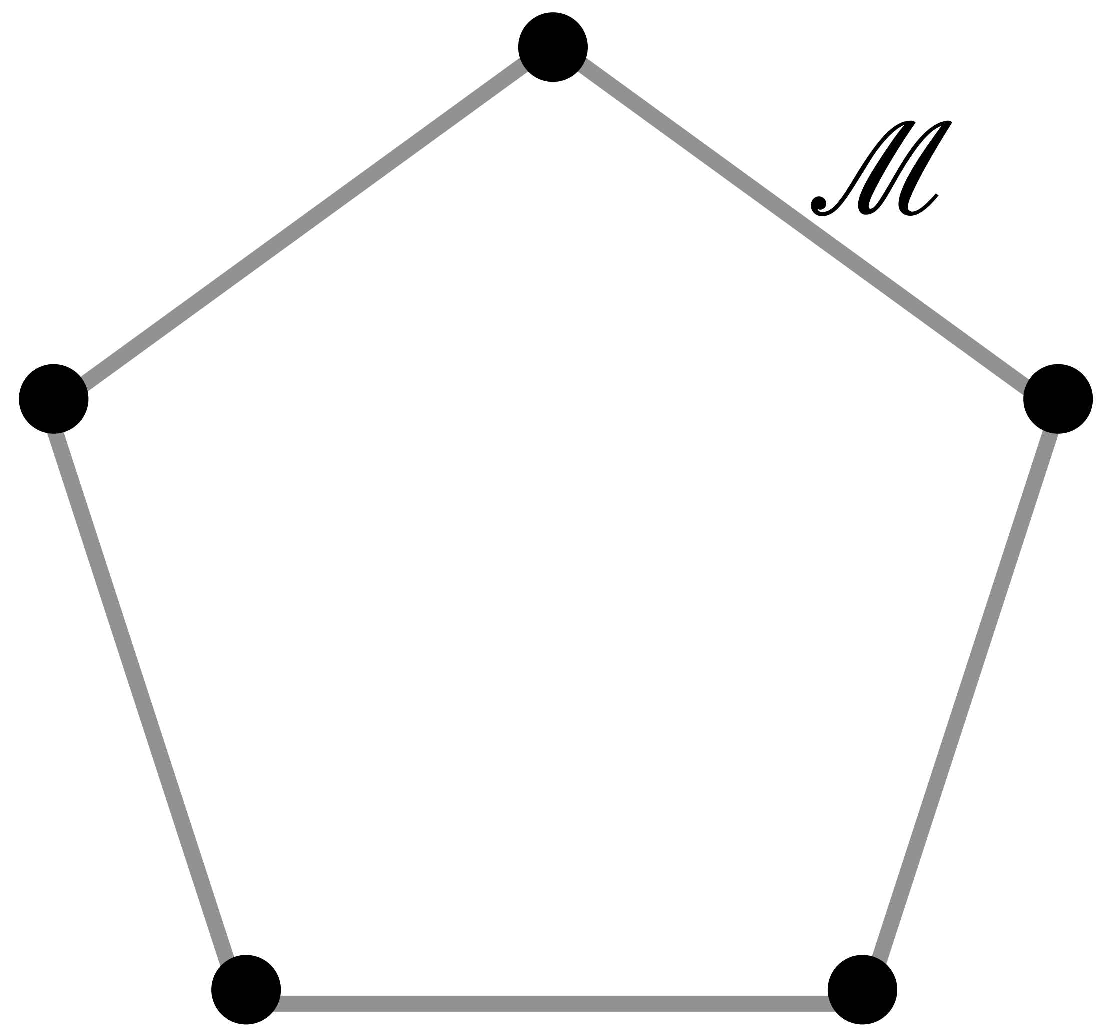
Smallest convex set = **convex hull** (a.k.a. marginal polytope)



$$\mathscr{M} \triangleq conv(\varphi(\mathscr{Y}))$$

Choice of the convex set

Smallest convex set = **convex hull** (a.k.a. marginal polytope)

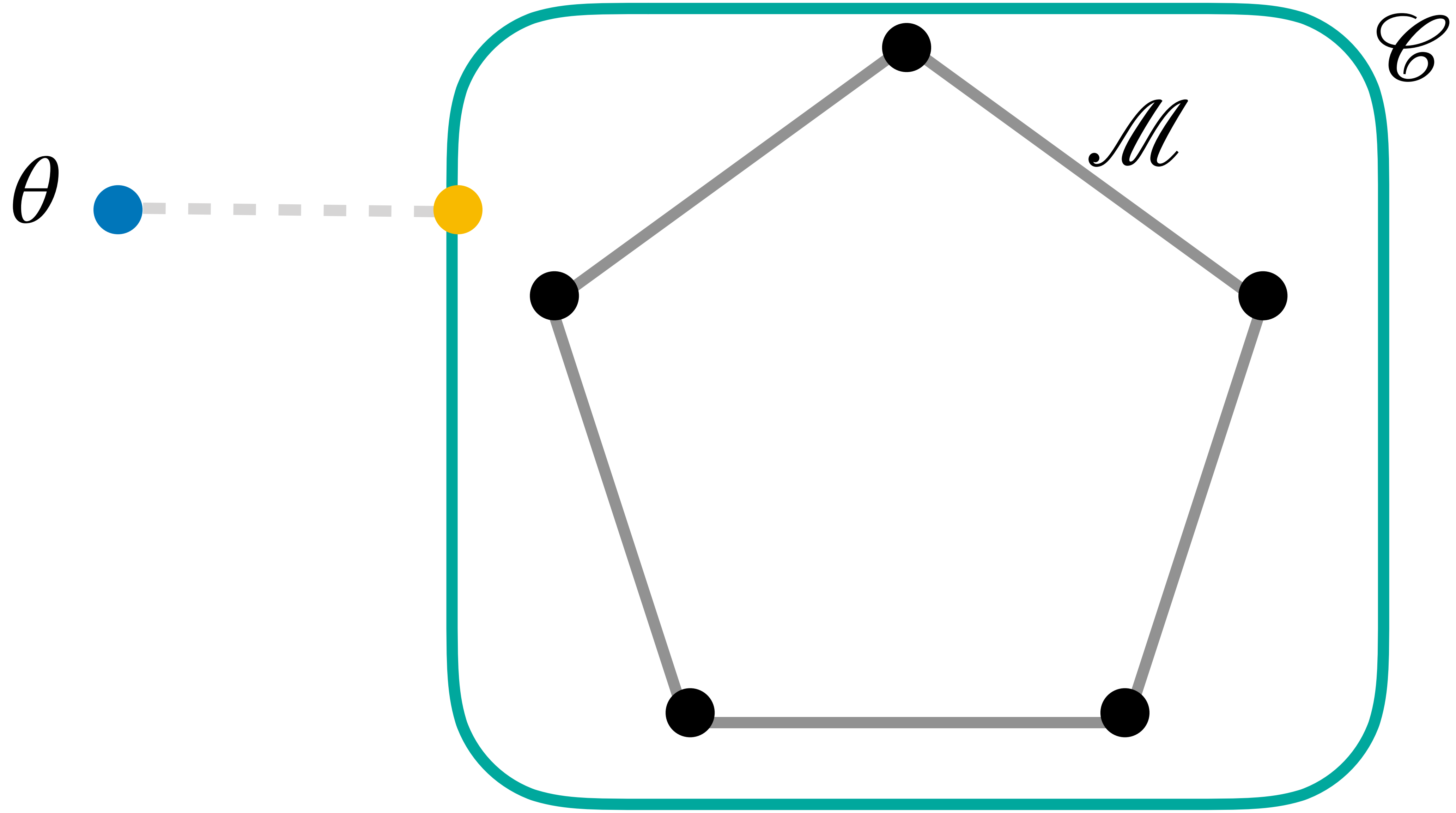


$$\mathscr{M} \triangleq conv(\varphi(\mathscr{Y}))$$

Choice of the convex set

Smallest convex set = **convex hull** (a.k.a. marginal polytope)

Can use any **superset** with cheaper to compute projection



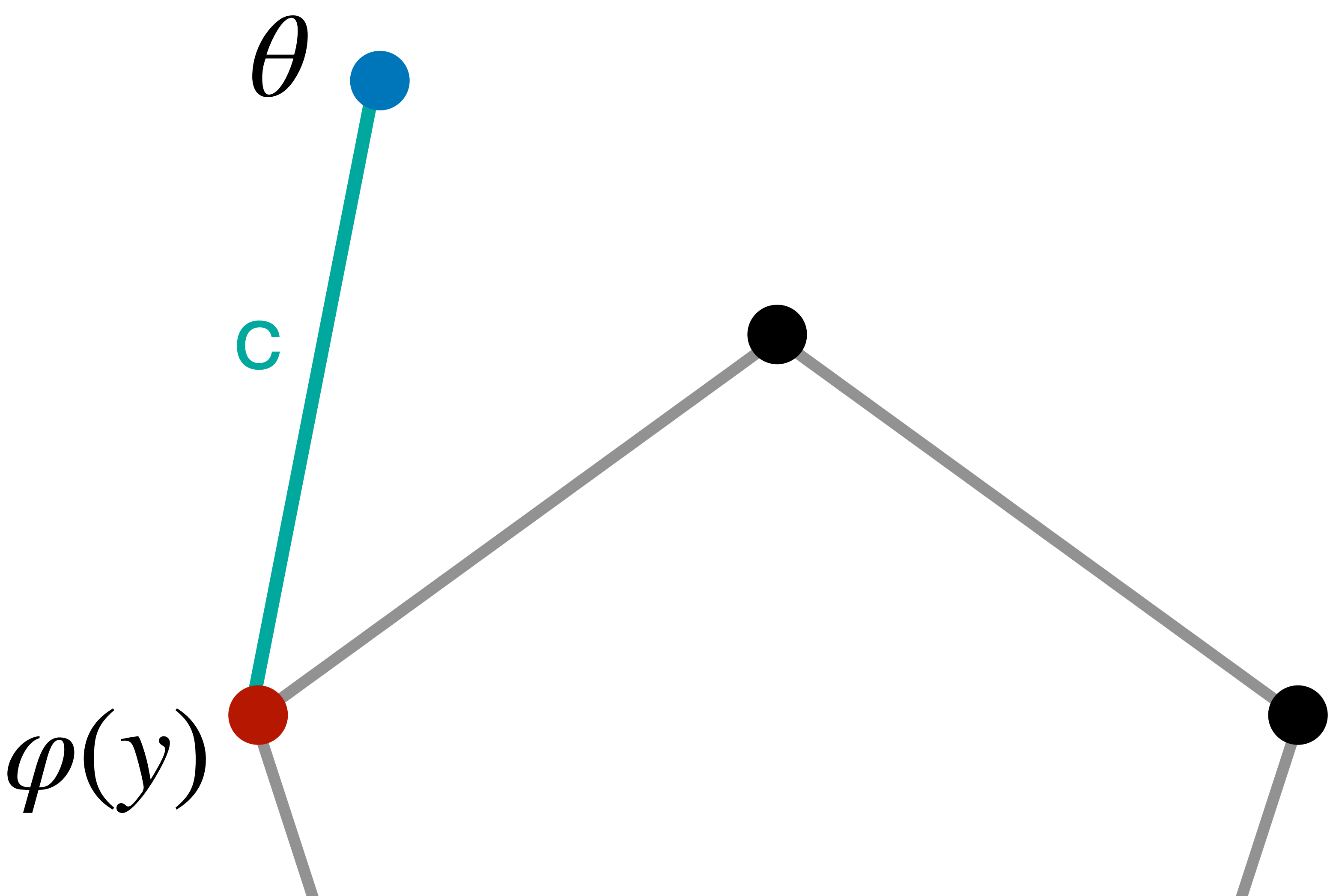
$$\mathscr{M} \triangleq conv(\varphi(\mathscr{Y}))$$

Associated loss function

Associated loss function

Squared loss

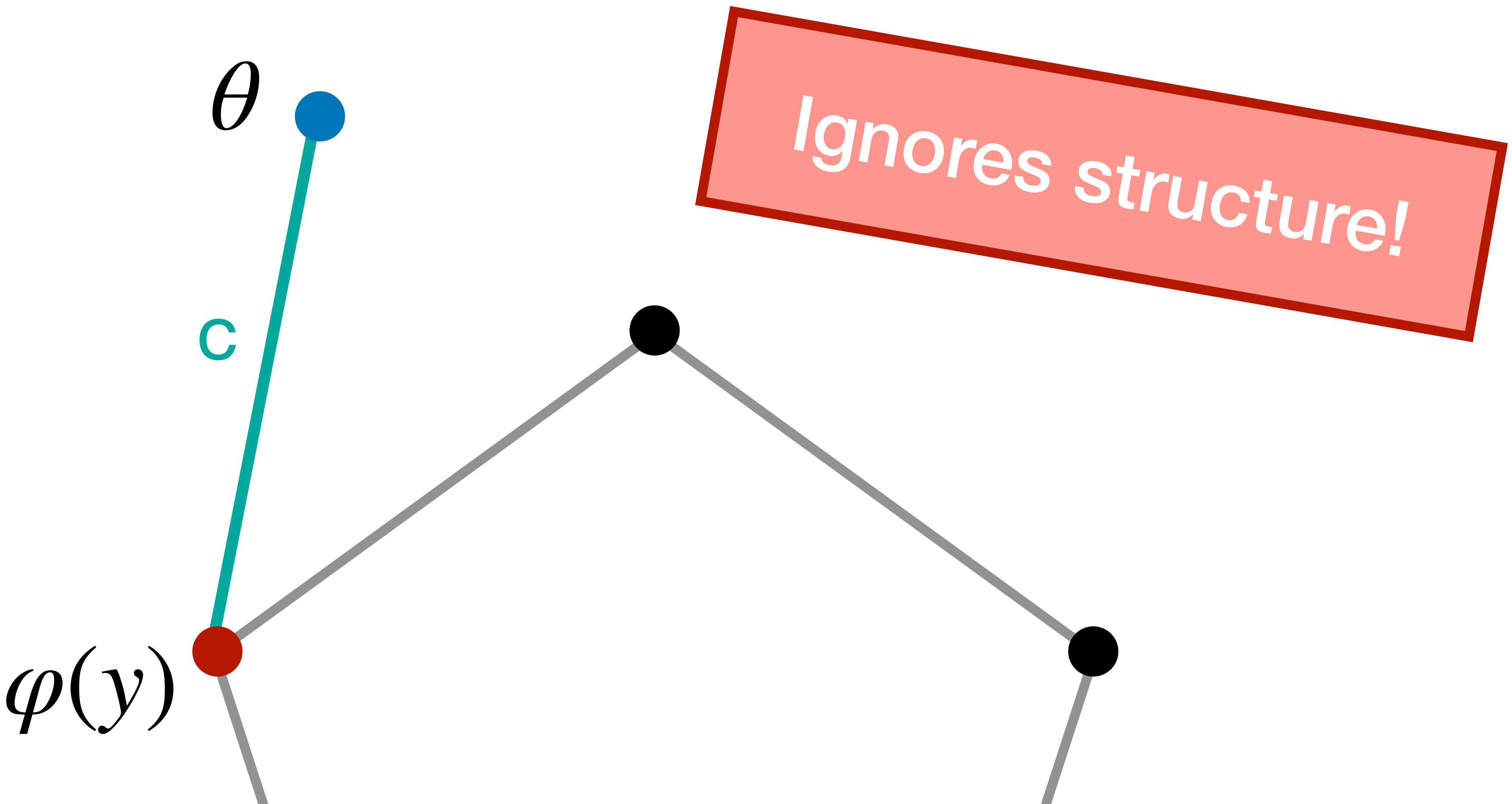
$$SQ(\theta, y) \triangleq \frac{1}{2}\|\varphi(y) - \theta\|^2 = c$$



Associated loss function

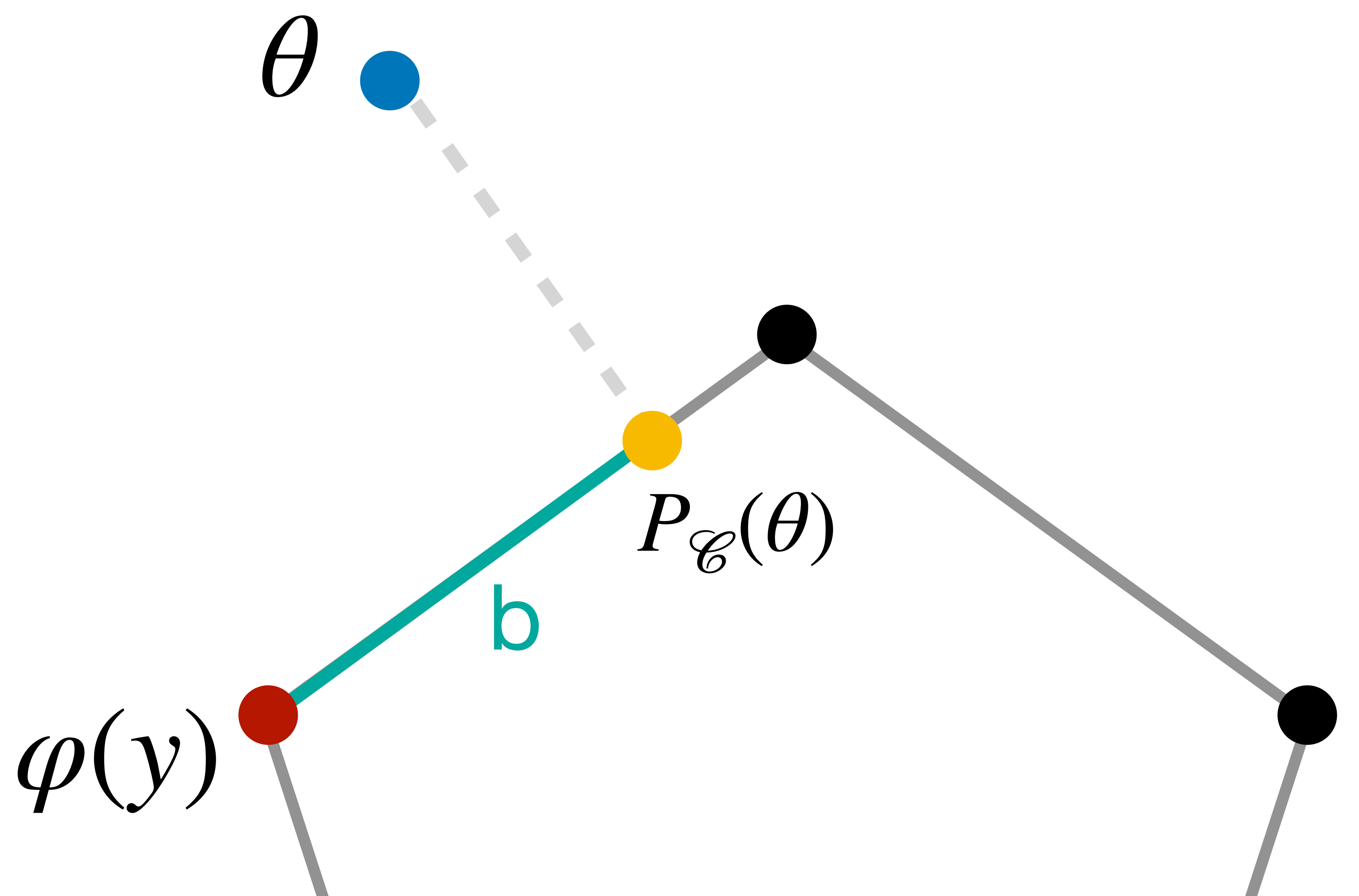
Squared loss

$$SQ(\theta, y) \triangleq \frac{1}{2}\|\varphi(y) - \theta\|^2 = c$$



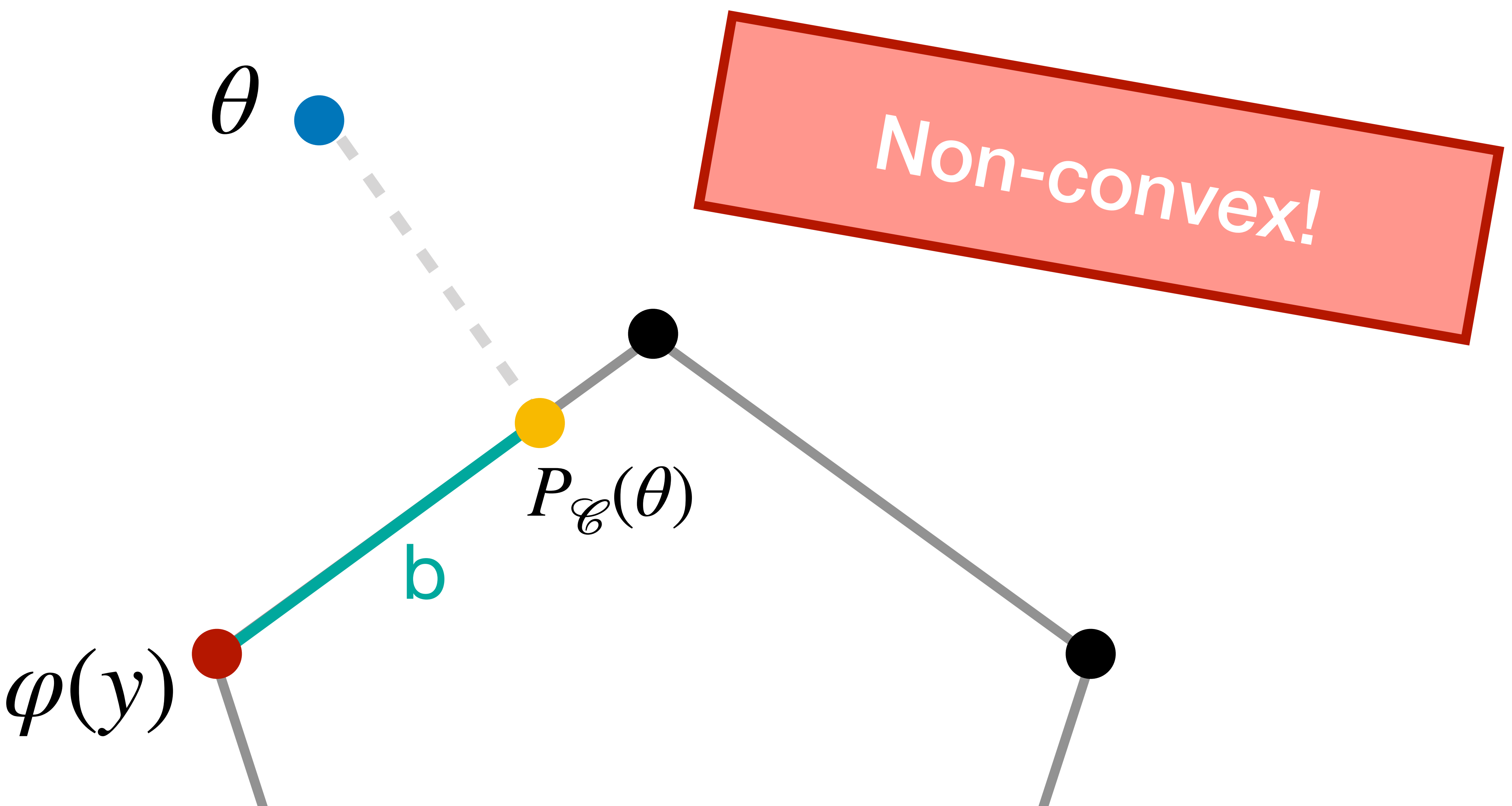
Associated loss function

$$NC_{\mathscr{C}}(\theta, y) \triangleq \frac{1}{2}\|\varphi(y) - P_{\mathscr{C}}(\theta)\|^2 = b$$



Associated loss function

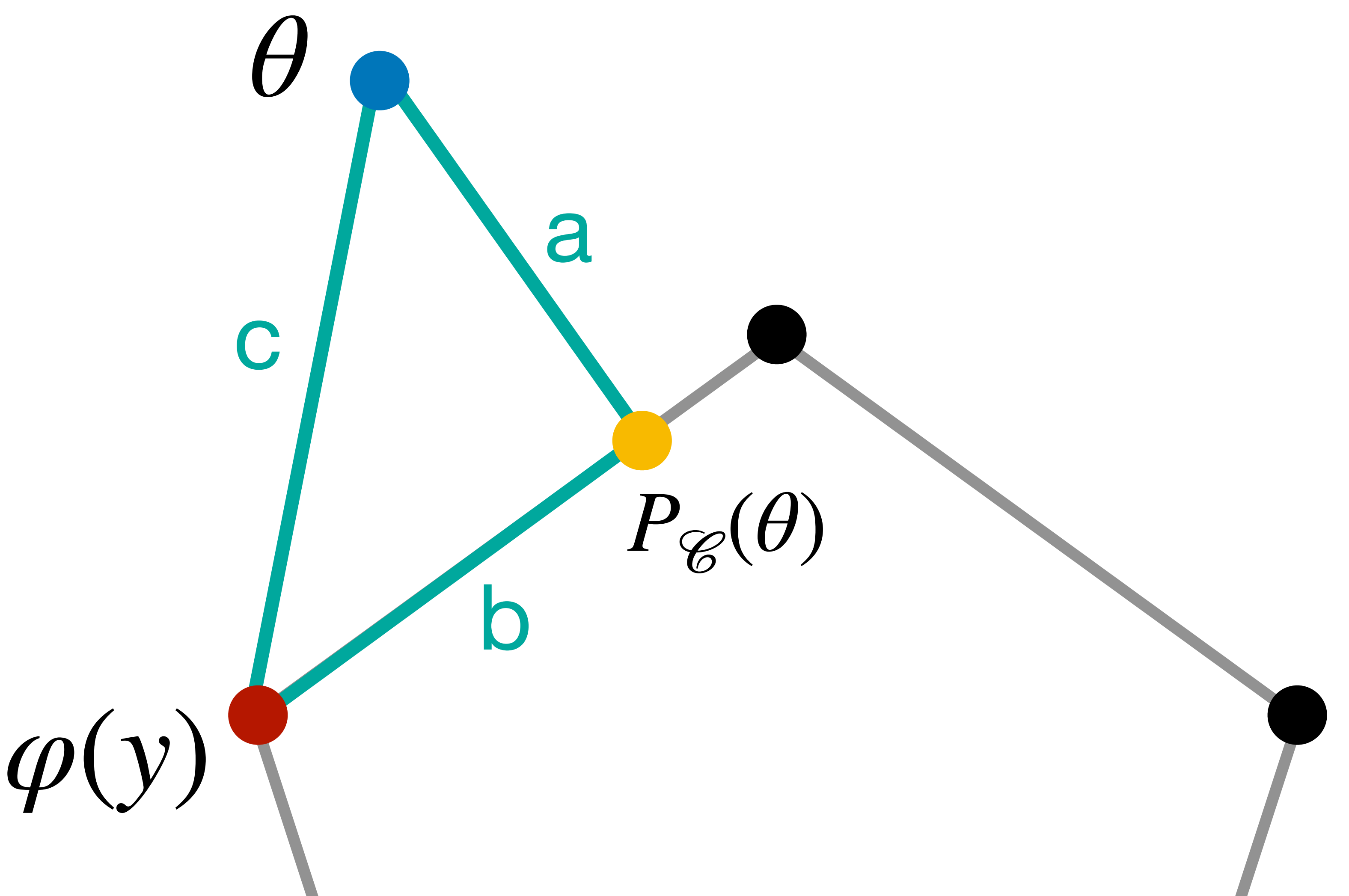
$$NC_{\mathscr{C}}(\theta, y) \triangleq \frac{1}{2}\|\varphi(y) - P_{\mathscr{C}}(\theta)\|^2 = b$$



Associated loss function

Proposed loss

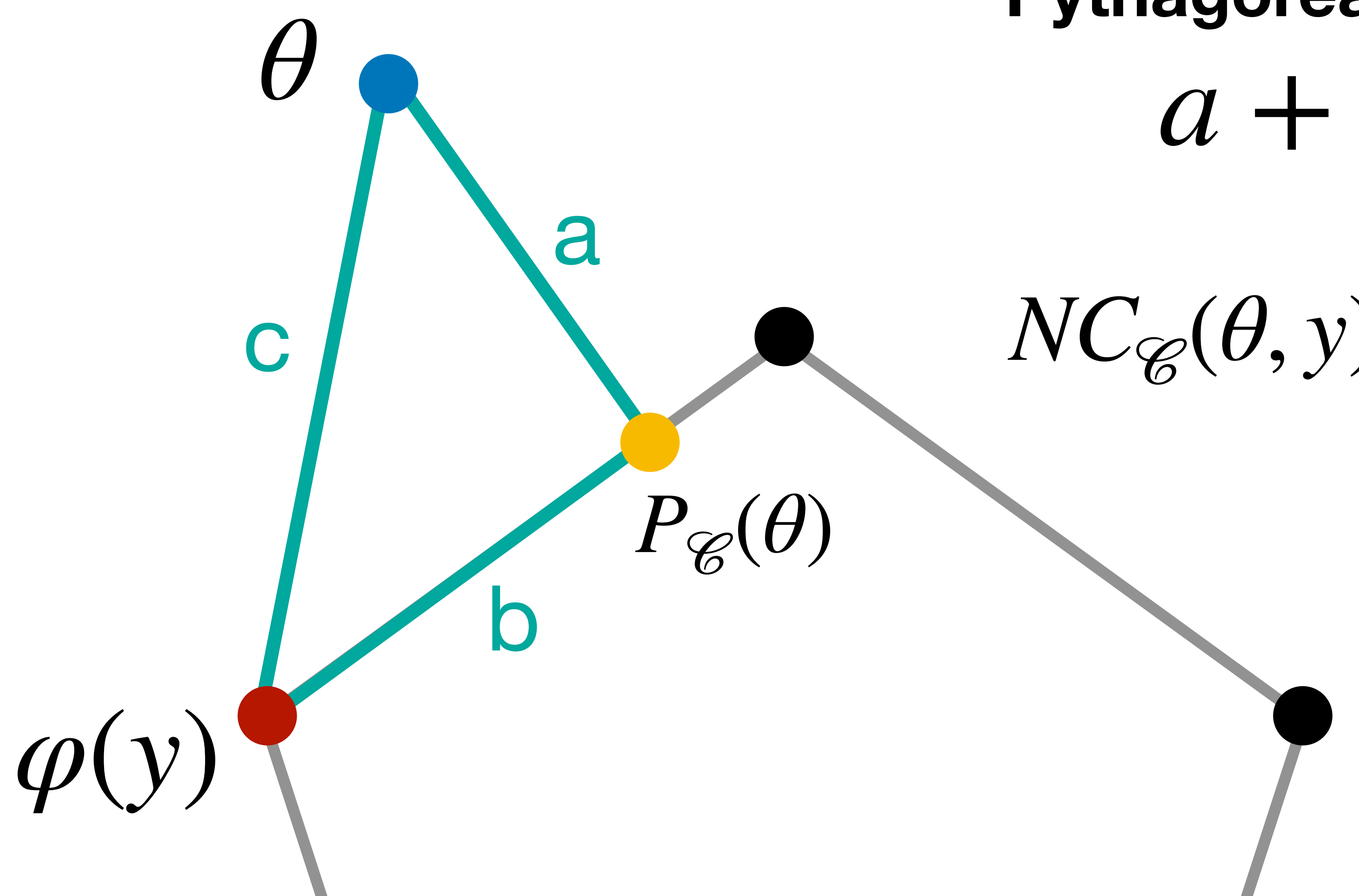
$$S_{\mathscr{C}}(\theta, y) \triangleq SQ(\theta, y) - \frac{1}{2}\|\theta - P_{\mathscr{C}}(\theta)\|^2 = c - a$$



Associated loss function

Proposed loss

$$S_{\mathscr{C}}(\theta, y) \triangleq SQ(\theta, y) - \frac{1}{2}\|\theta - P_{\mathscr{C}}(\theta)\|^2 = c - a$$



Generalized Pythagorean theorem

$$a + b \leq c$$

$$\downarrow$$

$$NC_{\mathscr{C}}(\theta, y) \leq S_{\mathscr{C}}(\theta, y)$$

Properties

1. $S_{\mathscr{C}}(\theta, y)$ is convex w.r.t. θ

2. $S_{\mathscr{C}}(\theta, y)$ is smooth w.r.t. θ (gradient is Lipschitz cont.)

3. $S_{\mathscr{C}}(\theta, y) \geq 0$

4. $S_{\mathscr{C}}(\theta, y) = 0 \Leftrightarrow P_{\mathscr{C}}(\theta) = \varphi(y)$

Upper bounds

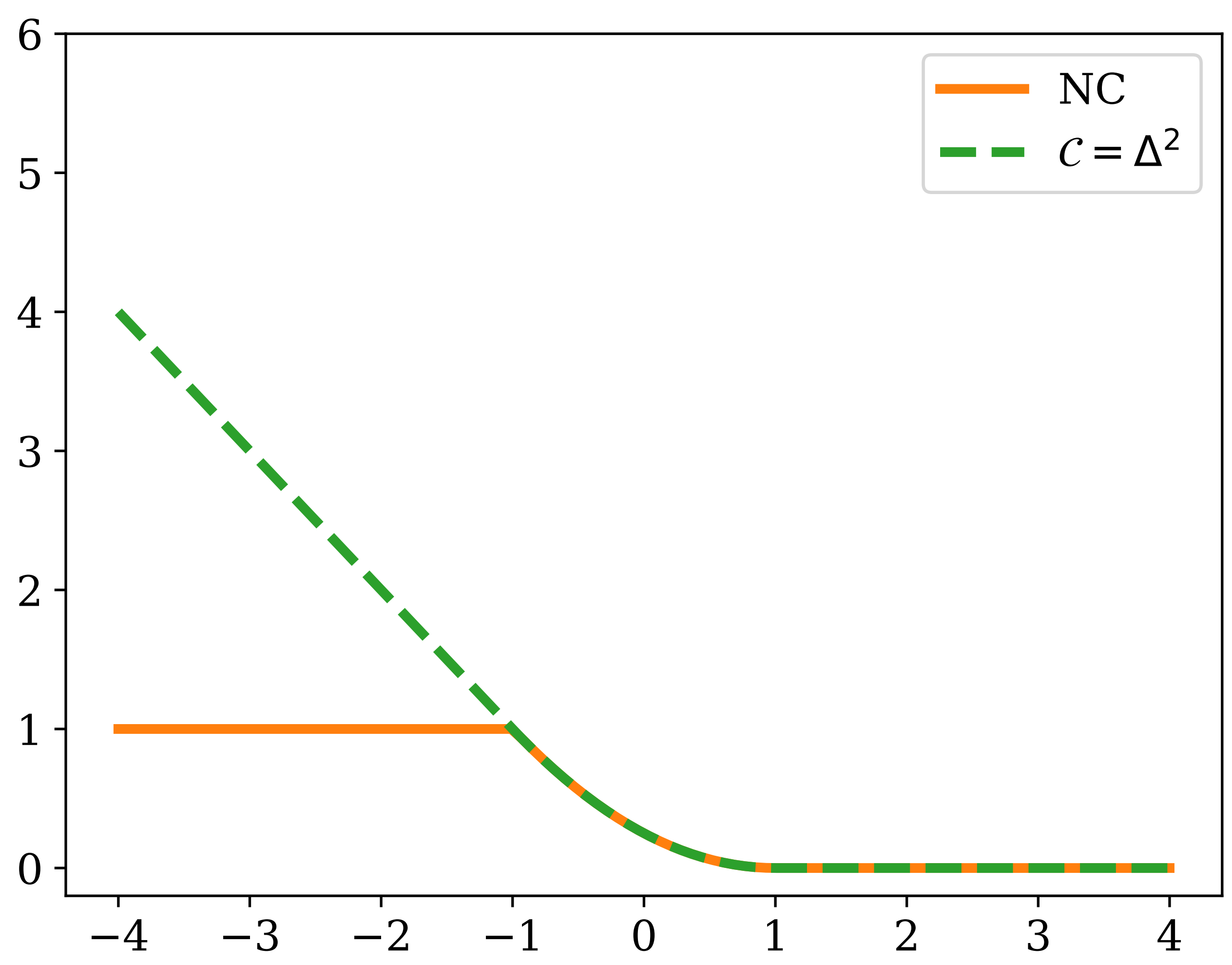
Convex upper bound

$$NC_{\mathscr{C}}(\theta, y) \leq S_{\mathscr{C}}(\theta, y) \quad \forall \theta, \varphi(y) \in \mathscr{C}$$

Upper bounds

Convex upper bound

$$NC_{\mathscr{C}}(\theta, y) \leq S_{\mathscr{C}}(\theta, y) \quad \forall \theta, \varphi(y) \in \mathscr{C}$$



Upper bounds

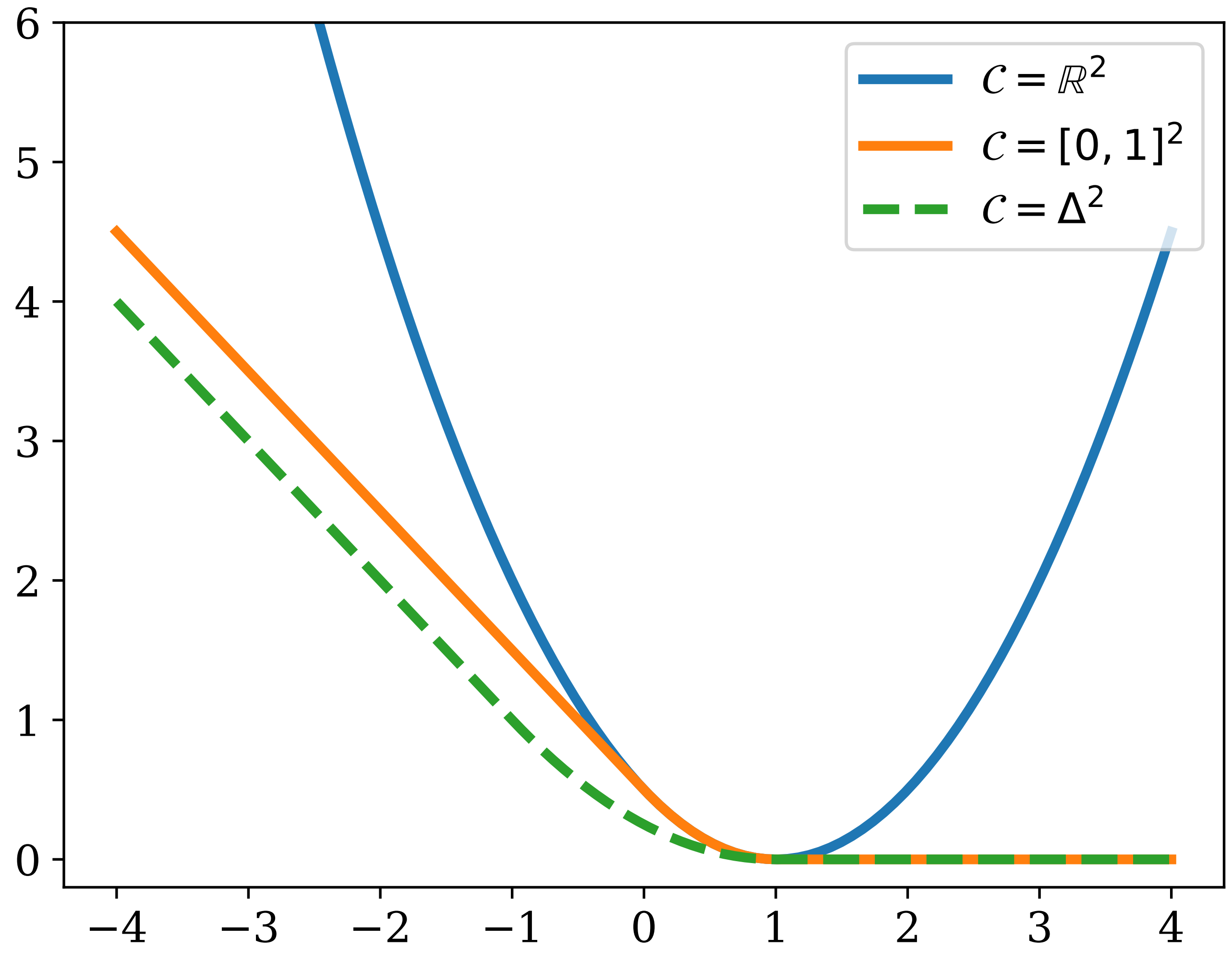
Superset upper bound

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

Upper bounds

Superset upper bound

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



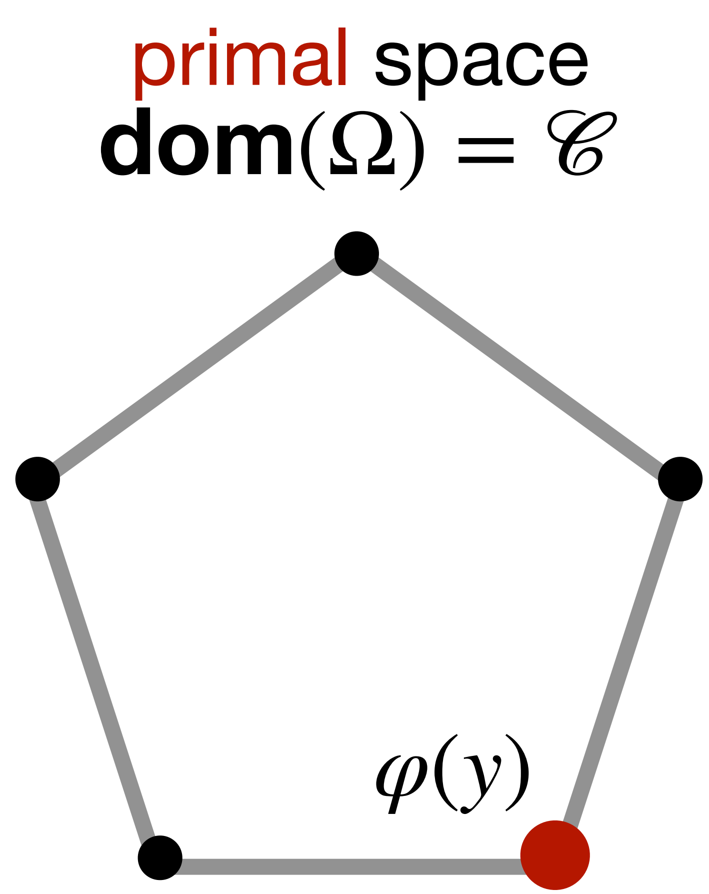
Link with Fenchel duality

Let $\Omega(u) \triangleq \frac{1}{2}\|u\|^2$ **if** $u \in \mathscr{C}$, ∞ **otherwise**

"Fenchel-Young losses", Blondel, Martins, Niculae, 2019

Link with Fenchel duality

Let $\Omega(u) \triangleq \frac{1}{2}\|u\|^2$ **if** $u \in \mathscr{C}$, ∞ **otherwise**

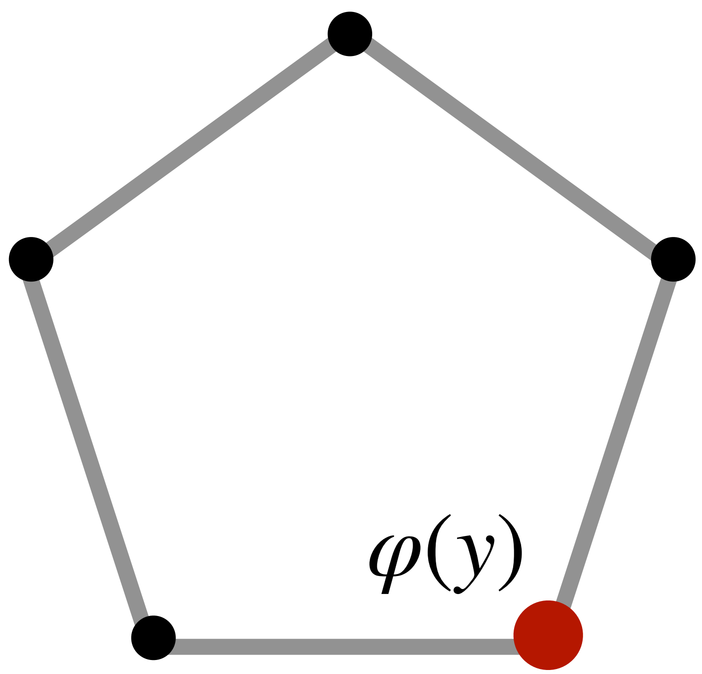


"Fenchel-Young losses", Blondel, Martins, Niculae, 2019

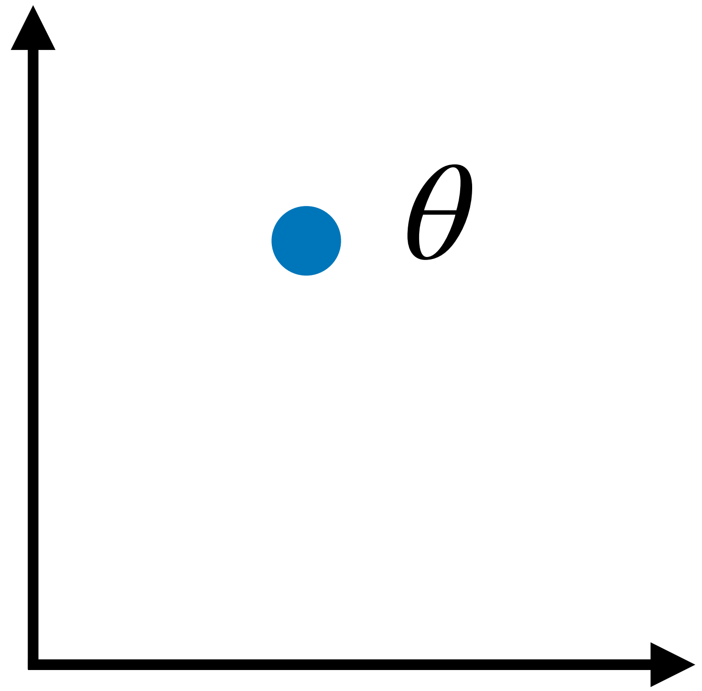
Link with Fenchel duality

Let $\Omega(u) \triangleq \frac{1}{2}\|u\|^2$ **if** $u \in \mathscr{C}$, ∞ **otherwise**

primal space
dom$(\Omega) = \mathscr{C}$



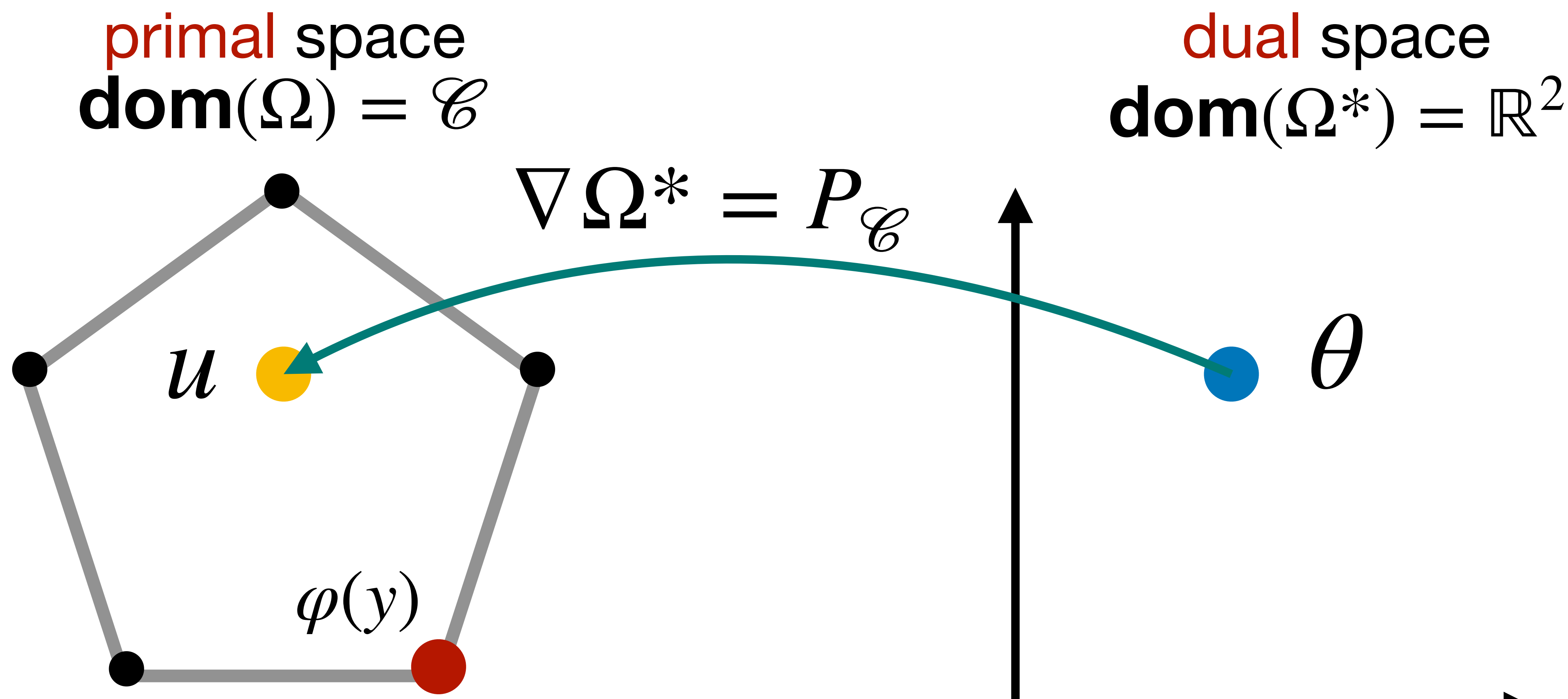
dual space
dom$(\Omega^*) = \mathbb{R}^2$



"Fenchel-Young losses", Blondel, Martins, Niculae, 2019

Link with Fenchel duality

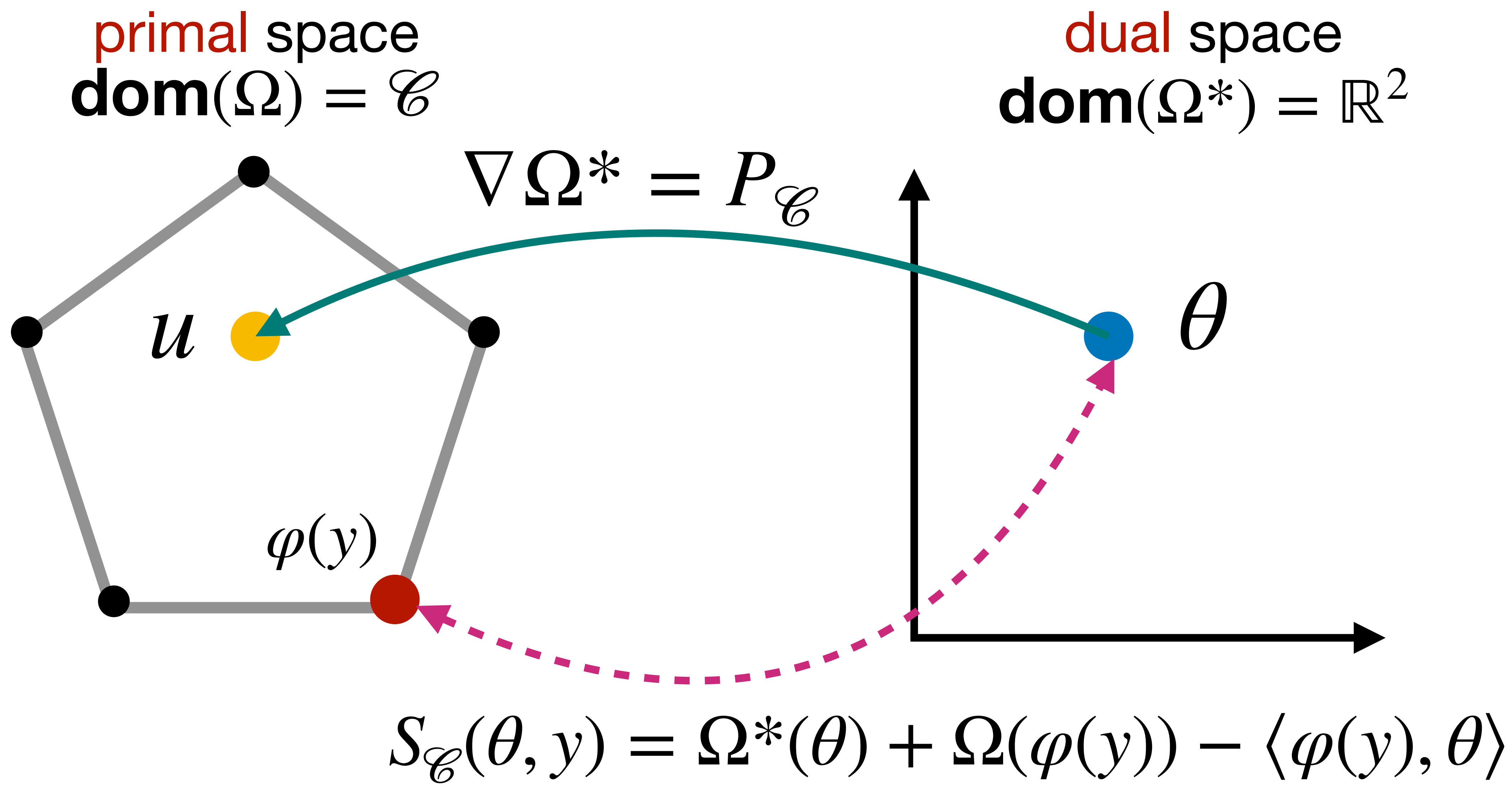
Let $\Omega(u) \triangleq \frac{1}{2}\|u\|^2$ **if** $u \in \mathscr{C}$, ∞ **otherwise**



"Fenchel-Young losses", Blondel, Martins, Niculae, 2019

Link with Fenchel duality

Let $\Omega(u) \triangleq \frac{1}{2}\|u\|^2$ **if** $u \in \mathscr{C}$, ∞ **otherwise**



"Fenchel-Young losses", Blondel, Martins, Niculae, 2019

Kullback Leibler geometry

$\Omega(u) \triangleq \langle u, \log u \rangle$ **if** $u \in \mathscr{C}$, ∞ **otherwise**

Kullback Leibler geometry

$\Omega(u) \triangleq \langle u, \log u \rangle$ **if** $u \in \mathscr{C}$, ∞ **otherwise**

$$\nabla \Omega^*(\theta) = \arg\min_{u \in \mathscr{C}} KL(u, e^{\theta - 1})$$

Kullback Leibler geometry

$$\Omega(u) \triangleq \langle u, \log u \rangle \text{ **if** } u \in \mathscr{C}, \ \infty \text{ **otherwise**}$$

$$\nabla \Omega^*(\theta) = \arg\min_{u \in \mathscr{C}} KL(u, e^{\theta - 1})$$

$$S_{\mathscr{C}}(\theta, y) = \Omega^*(\theta) + \Omega(\varphi(y)) - \langle \varphi(y), \theta \rangle$$

Kullback Leibler geometry

$\Omega(u) \triangleq \langle u, \log u \rangle$ **if** $u \in \mathscr{C}$, ∞ **otherwise**

$$\nabla \Omega^*(\theta) = \arg\min_{u \in \mathscr{C}} KL(u, e^{\theta - 1})$$

$$S_{\mathscr{C}}(\theta, y) = \Omega^*(\theta) + \Omega(\varphi(y)) - \langle \varphi(y), \theta \rangle$$

Proposition

Let $\beta = \max_{u \in \mathscr{C}} \|u\|_1$. Then,

$S_{\mathscr{C}}(\theta, y)$ is β-smooth with respect to $\|\cdot\|_\infty$.

Kullback Leibler geometry

$\Omega(u) \triangleq \langle u, \log u \rangle$ **if** $u \in \mathscr{C}$, ∞ **otherwise**

$$\nabla \Omega^*(\theta) = \arg\min_{u \in \mathscr{C}} KL(u, e^{\theta - 1})$$

$$S_{\mathscr{C}}(\theta, y) = \Omega^*(\theta) + \Omega(\varphi(y)) - \langle \varphi(y), \theta \rangle$$

Smaller set → smoother loss!

Proposition

Let $\beta = \max_{u \in \mathscr{C}} \|u\|_1$. Then,

$S_{\mathscr{C}}(\theta, y)$ is β-smooth with respect to $\|\cdot\|_\infty$.

Calibrated decoding

Calibrated decoding

Affine decomposition of the target loss

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

Calibrated decoding

Affine decomposition of the target loss

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

Decoding calibrated for loss L

$$\hat{y}_L(u) \triangleq \arg\min_{y' \in \mathcal{Y}} \langle \varphi(y'), Vu + b \rangle$$

Calibrated decoding

Affine decomposition of the target loss

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

Decoding calibrated for loss L

$$\hat{y}_L(u) \triangleq \arg\min_{y' \in \mathcal{Y}} \langle \varphi(y'), Vu + b \rangle$$

$$= MAP(-Vu - b)$$

Calibrated decoding

Affine decomposition of the target loss

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

Decoding calibrated for loss L

$$\hat{y}_L(u) \triangleq \arg\min_{y' \in \mathcal{Y}} \langle \varphi(y'), Vu + b \rangle$$

$$= MAP(-Vu - b)$$

Decomposition important both for computational tractability and theoretical analysis

Consistency

Excess risks

$$\delta \mathscr{L}(f) \triangleq \mathscr{L}(f) - \inf_{f' : \mathscr{X} \to \mathscr{Y}} \mathscr{L}(f')$$

$$\delta \mathcal{S}_{\mathscr{C}}(g) \triangleq \mathcal{S}_{\mathscr{C}}(g) - \inf_{g' : \mathscr{X} \to \Theta} \mathcal{S}_{\mathscr{C}}(g')$$

Consistency

Excess risks

$$\delta\mathscr{L}(f) \triangleq \mathscr{L}(f) - \inf_{f' : \mathscr{X} \to \mathscr{Y}} \mathscr{L}(f')$$

$$\delta\mathcal{S}_{\mathscr{C}}(g) \triangleq \mathcal{S}_{\mathscr{C}}(g) - \inf_{g' : \mathscr{X} \to \Theta} \mathcal{S}_{\mathscr{C}}(g')$$

Calibration between excess risks

$$\forall g : \mathscr{X} \to \Theta : \quad \frac{\delta\mathscr{L}(dec \circ g)^2}{8\beta\sigma^2} \leq \delta\mathcal{S}_{\mathscr{C}}(g)$$

$dec \triangleq \hat{y}_L \circ P_{\mathscr{C}}$

$\beta \triangleq$ Lipschitz constant of $P_{\mathscr{C}}$ w.r.t. $\|\cdot\|$

$\sigma \triangleq \sup_{y \in \mathscr{Y}} \|V^\top \varphi(y)\|$

Probability simplex

Output set

$$\mathcal{Y} = [k] \triangleq \{1, \ldots, k\}$$

Probability simplex

Output set

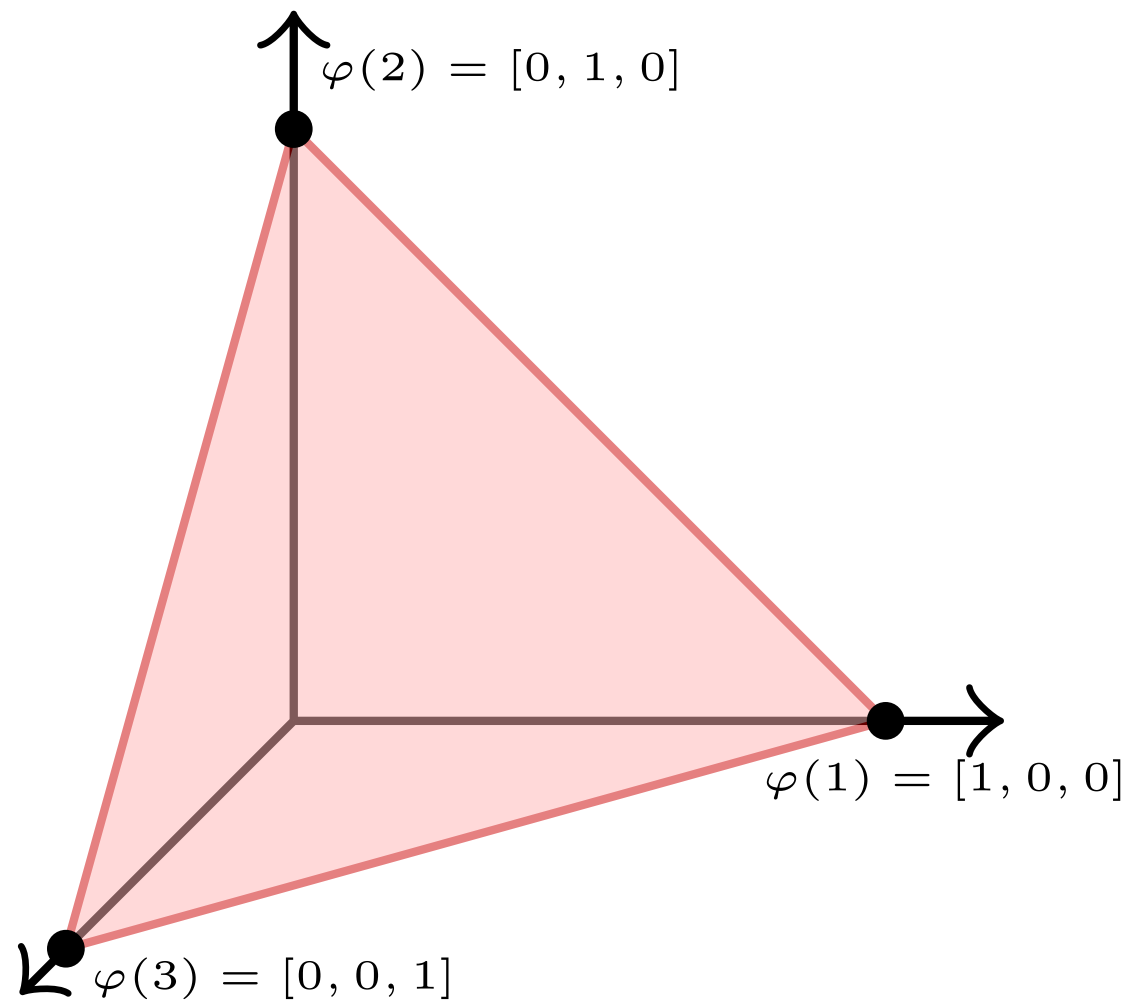
$$\mathscr{Y} = [k] \triangleq \{1, \ldots, k\}$$

Encoding

$$\varphi(y) = e_y$$

Marginal polytope

$$\mathscr{M} = conv(\varphi(\mathscr{Y})) = \triangle^k$$



Probability simplex

Output set

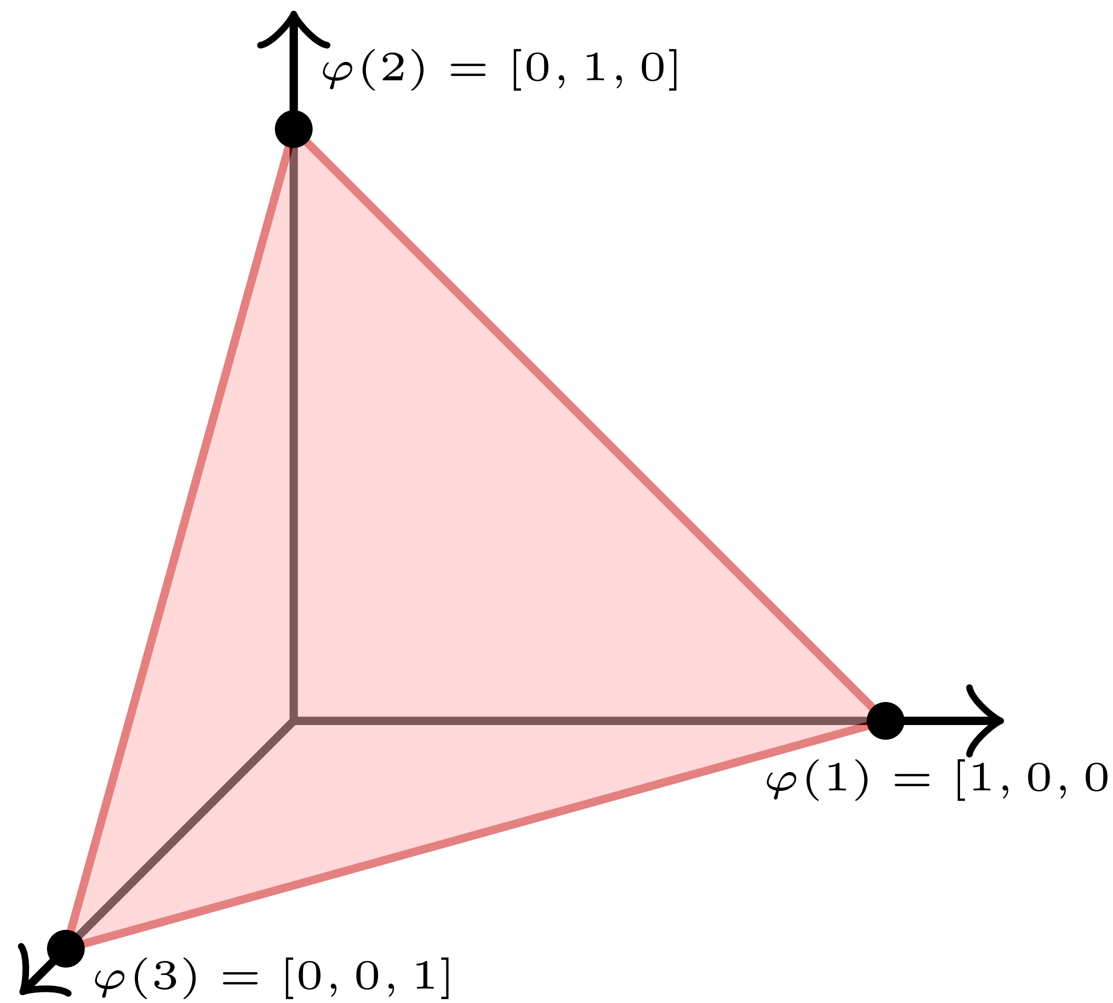
$$\mathcal{Y} = [k] \triangleq \{1, \ldots, k\}$$

Encoding

$$\varphi(y) = e_y$$

Marginal polytope

$$\mathcal{M} = conv(\varphi(\mathcal{Y})) = \triangle^k$$



Oracles

MAP: O(k)

Euclidean: sparsemax, O(k) or O(k log k)

KL: softmax, O(k)

Unit cube

Output set

$$\mathscr{Y} = 2^{[k]}$$

Unit cube

Output set

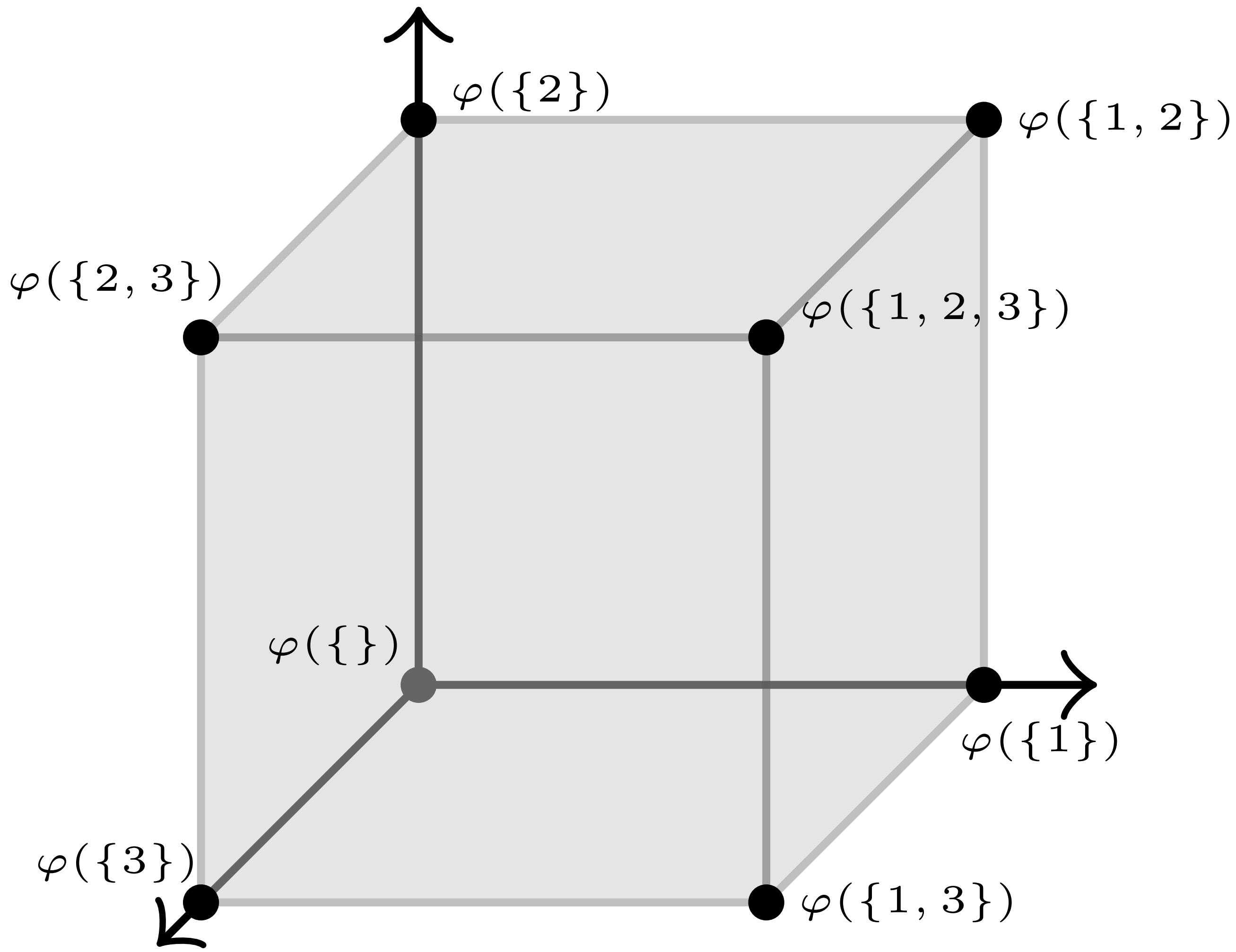
$$\mathscr{Y} = 2^{[k]}$$

Encoding

$$\varphi(y) = \sum_{i=1}^{|y|} e_{y_i}$$

Marginal polytope

$$\mathscr{M} = [0,1]^k$$



Unit cube

Output set

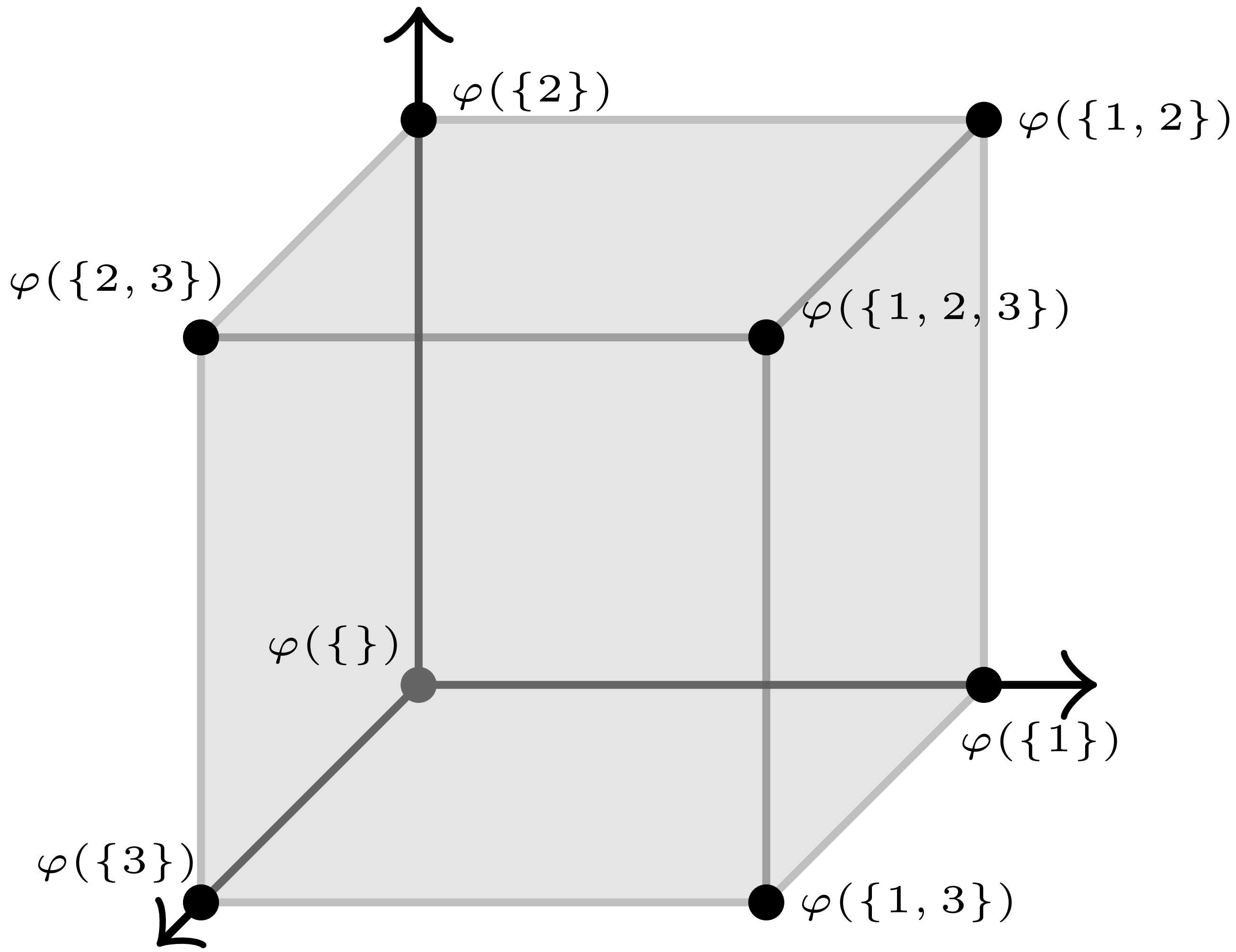
$$\mathcal{Y} = 2^{[k]}$$

Encoding

$$\varphi(y) = \sum_{i=1}^{|y|} e_{y_i}$$

Marginal polytope

$$\mathcal{M} = [0,1]^k$$



Oracles

MAP: O(k)

Euclidean: clipping to [0,1], O(k)

KL: O(k)

Budget polytope

Output set

$$\mathcal{Y} = \{y \in 2^{[k]} \colon l \le |y| \le u\}$$

Budget polytope

Output set

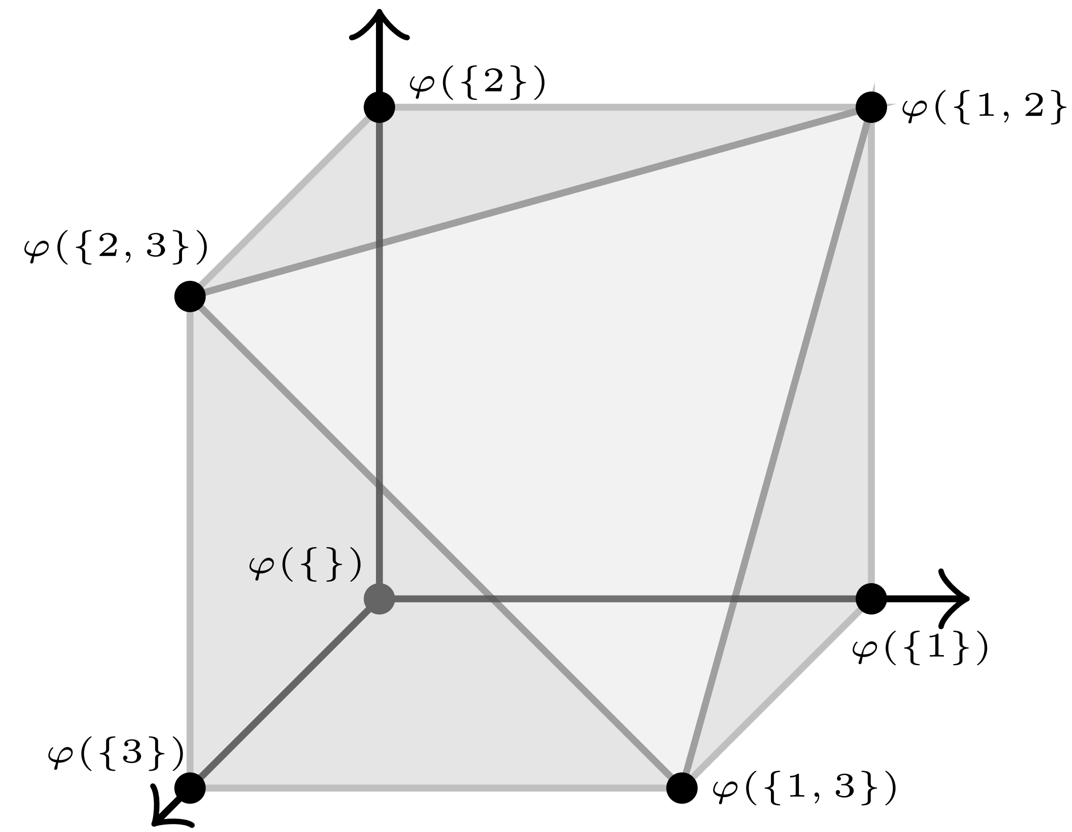
$$\mathcal{Y} = \{y \in 2^{[k]} \colon l \leq |y| \leq u\}$$

Encoding

$$\varphi(y) = \sum_{i=1}^{|y|} e_{y_i}$$

Marginal polytope

$$\mathcal{M} = \{y \in [0,1]^k \colon l \leq y^\top \mathbf{1} \leq m\}$$



Budget polytope

Output set

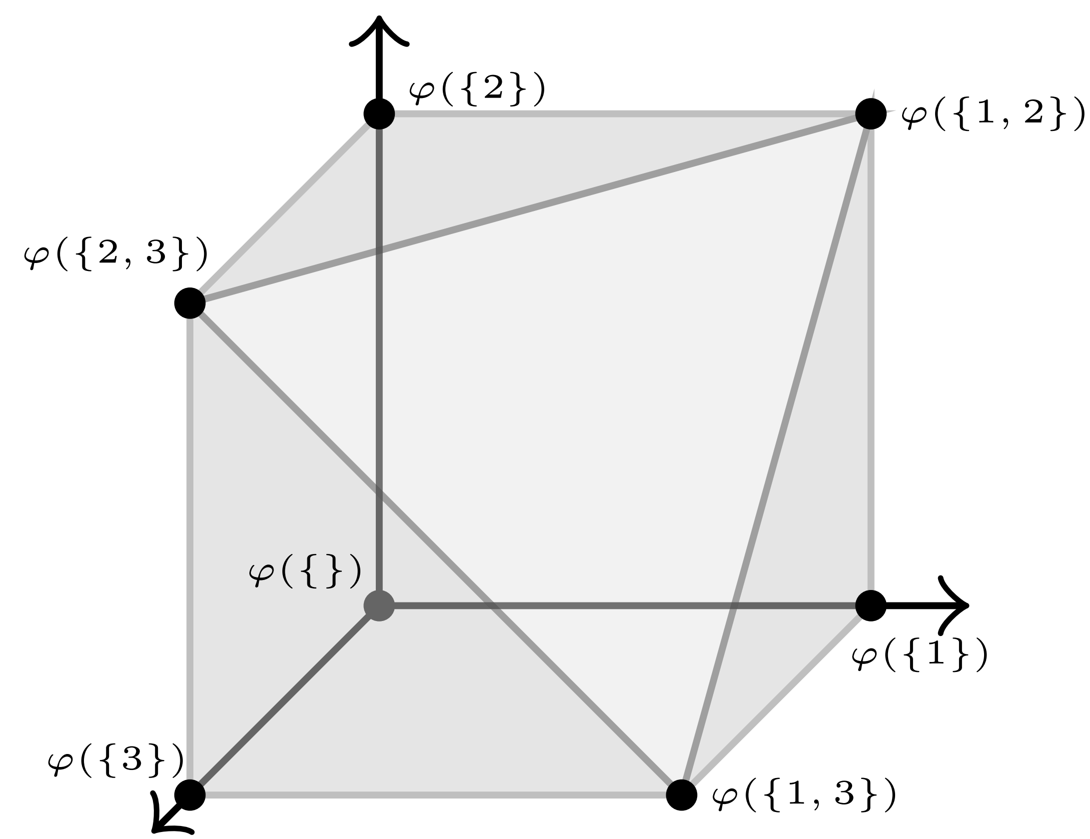
$$\mathcal{Y} = \{y \in 2^{[k]} \colon l \leq |y| \leq u\}$$

Encoding

$$\varphi(y) = \sum_{i=1}^{|y|} e_{y_i}$$

Marginal polytope

$$\mathcal{M} = \{y \in [0,1]^k \colon l \leq y^\top \mathbf{1} \leq m\}$$



Oracles

MAP: O(k log k)

Euclidean: O(k)

KL: O(k log k)

Order simplex

Output set

$\mathcal{Y} = [k] \quad 1 \prec \ldots \prec k$

Order simplex

Output set

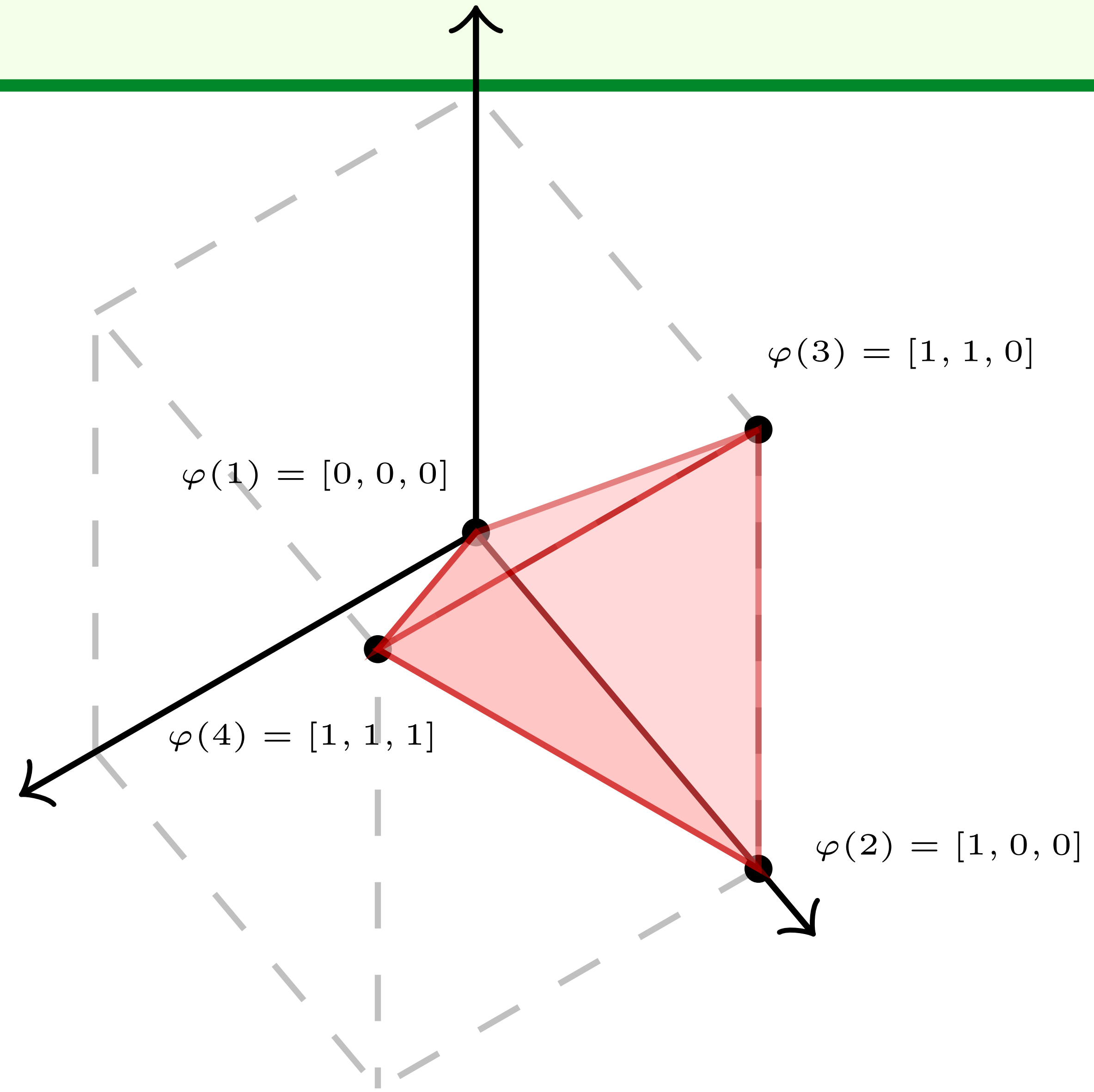
$$\mathcal{Y} = [k] \qquad 1 \prec \ldots \prec k$$

Encoding

$$\varphi(y) = \sum_{1 \leq i < y \leq k} e_i \in \mathbb{R}^{k-1}$$

Marginal polytope

$$\mathcal{M} = \{\mu \in \mathbb{R}^{k-1} \colon 1 \geq \mu_1 \geq \mu_2 \geq \ldots \geq \mu_{k-1} \geq 0\}$$



Order simplex

Output set

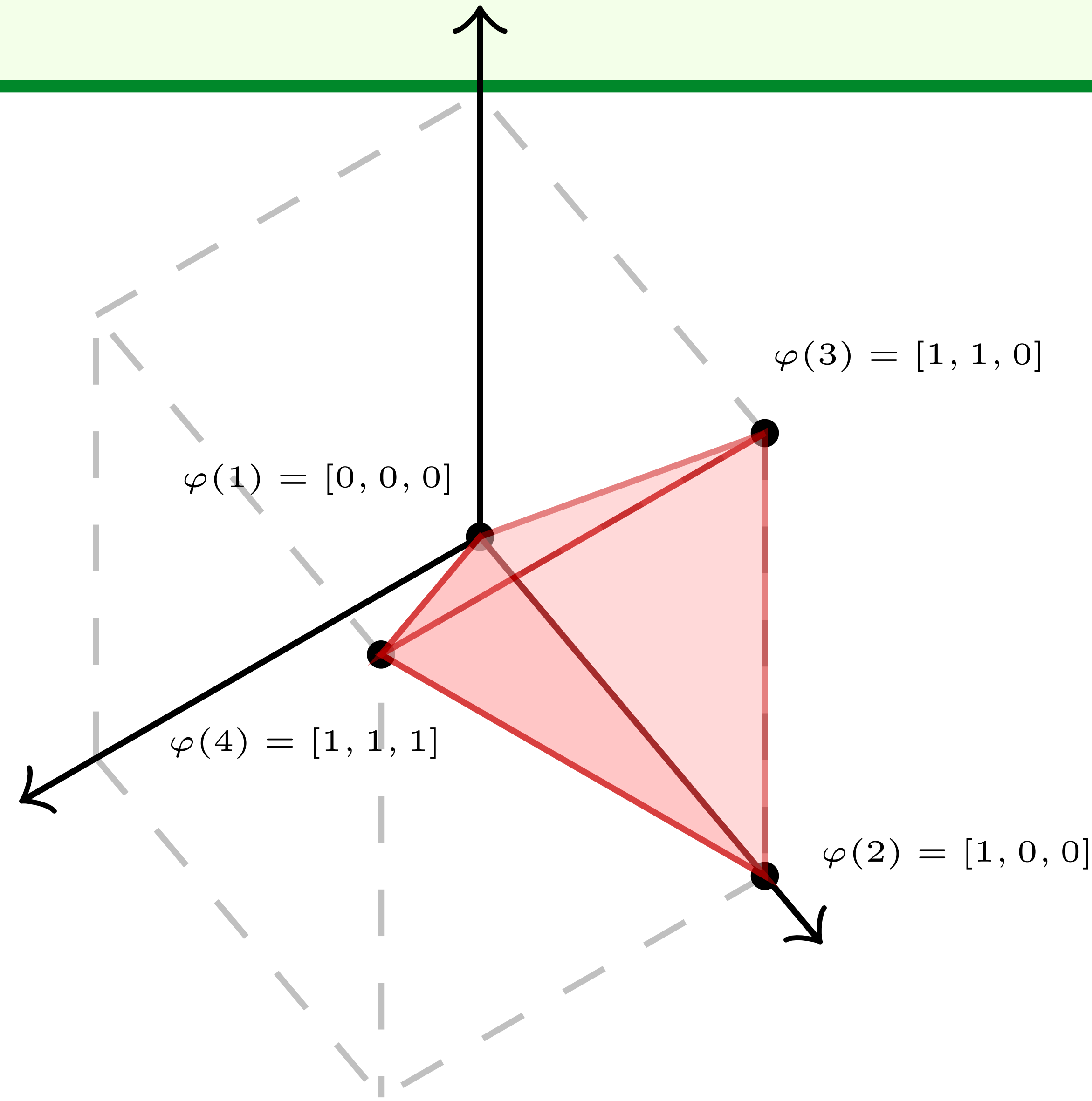
$$\mathcal{Y} = [k] \qquad 1 \prec \ldots \prec k$$

Encoding

$$\varphi(y) = \sum_{1 \le i < y \le k} e_i \in \mathbb{R}^{k-1}$$

Marginal polytope

$$\mathcal{M} = \{\mu \in \mathbb{R}^{k-1} \colon 1 \geq \mu_1 \geq \mu_2 \geq \ldots \geq \mu_{k-1} \geq 0\}$$



Oracles

MAP: O(k)

Eucl: isotonic reg, O(k)

KL: isotonic optimization

Birkhoff polytope

Output set

$\mathscr{Y} =$ **Permutations**$([m])$

Birkhoff polytope

Output set

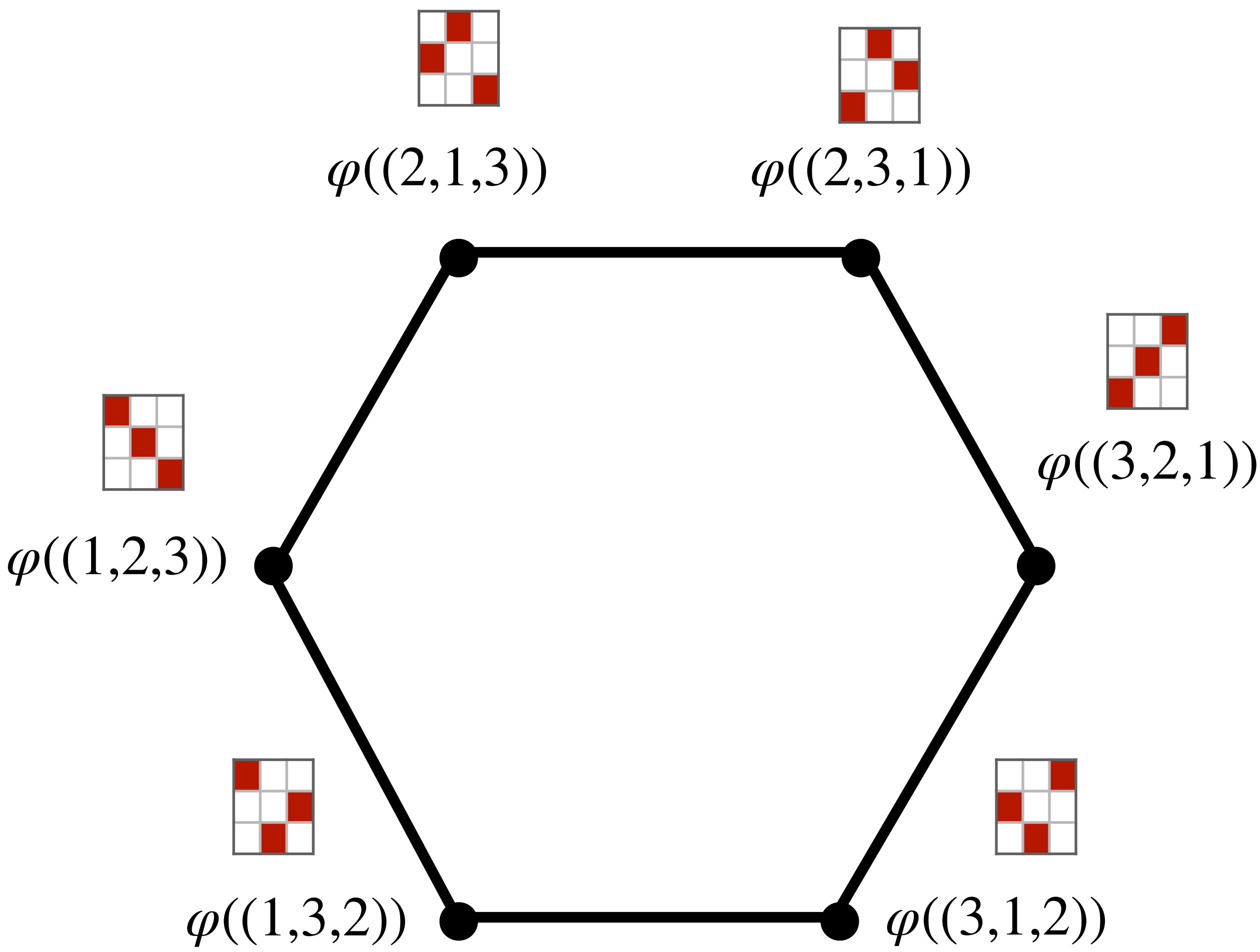
$\mathcal{Y} = \textbf{Permutations}([m])$

Encoding

$\varphi(y) =$ permutation matrix associated with y

Marginal polytope

$$\mathcal{M} = \{P \in \mathbb{R}^{m \times m} \colon P^\top \mathbf{1}_m = 1, P\mathbf{1}_m = 1, 0 \leq P \leq 1\}$$



Birkhoff polytope

Output set

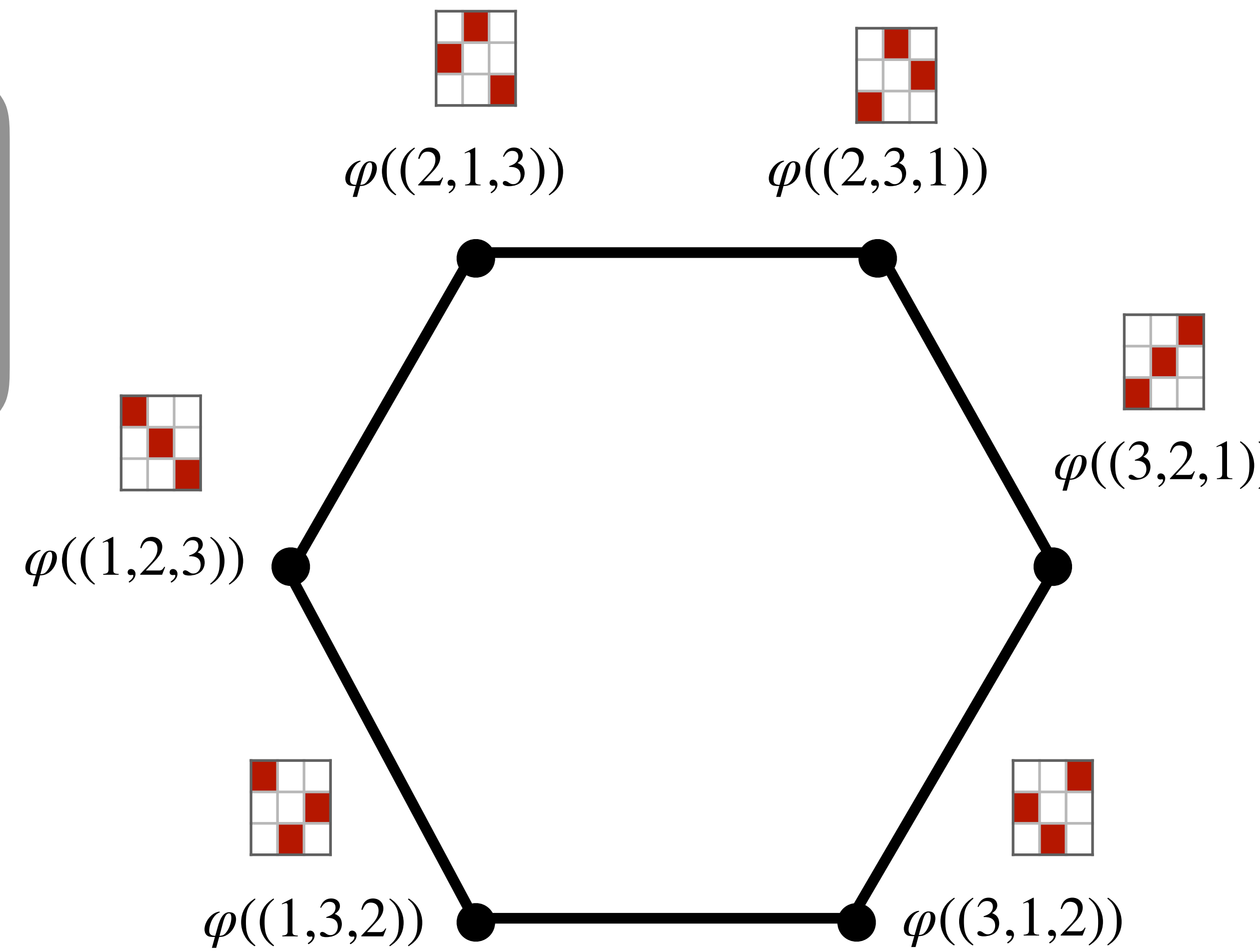
$$\mathcal{Y} = \textbf{Permutations}([m])$$

Encoding

$\varphi(y) =$ permutation matrix associated with y

Marginal polytope

$$\mathcal{M} = \{P \in \mathbb{R}^{m \times m} : P^\top \mathbf{1}_m = 1, P\mathbf{1}_m = 1, 0 \leq P \leq 1\}$$



Oracles

MAP: Hungarian, $O(m^3)$

Eucl: LBFGS dual, $O(m^2/\varepsilon)$

KL: Sinkhorn, $O(m^2/\varepsilon)$

Marginal: **intractable**

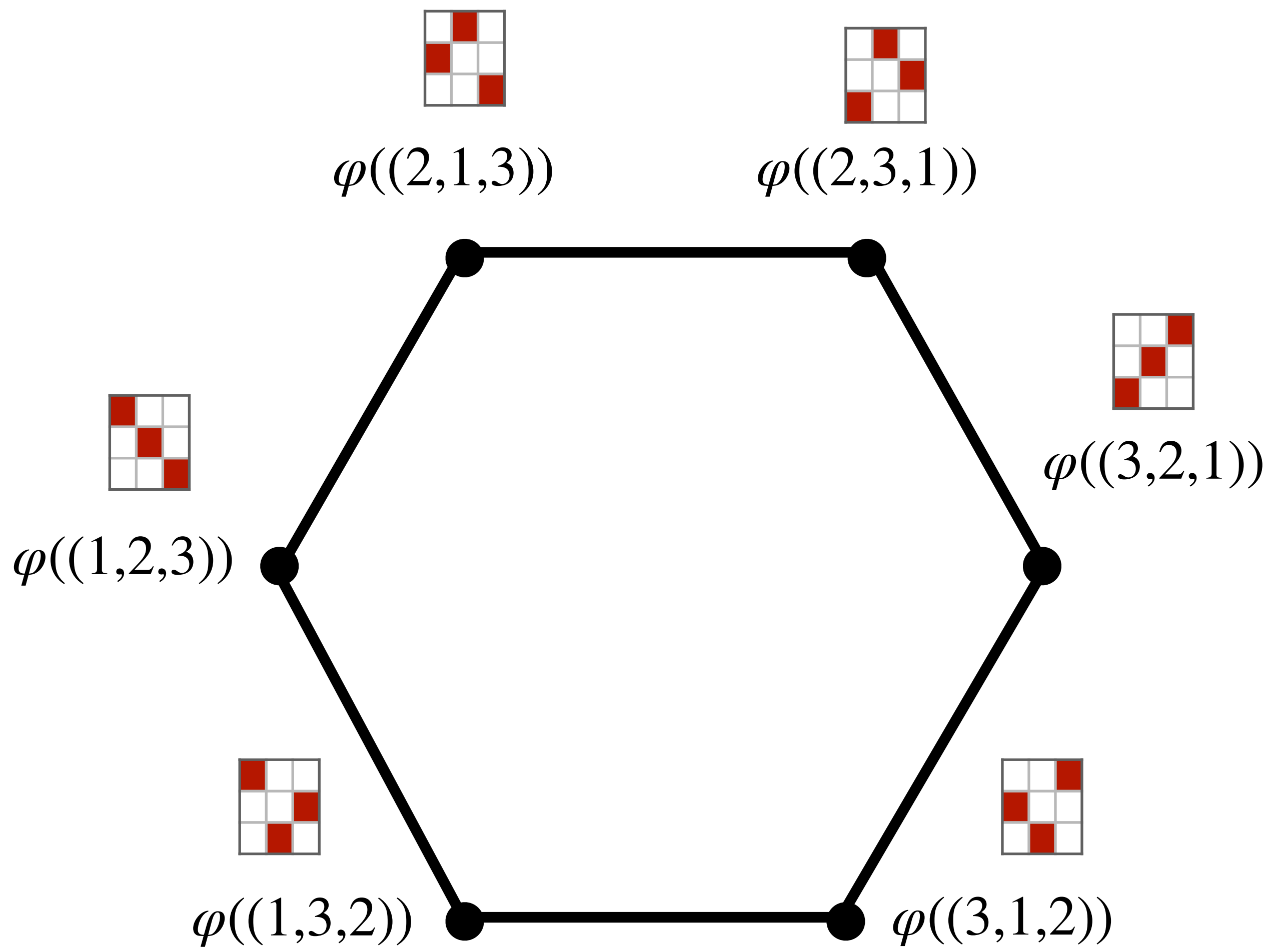
Birkhoff polytope

Output set

$\mathcal{Y} = \textbf{Permutations}([m])$

Encoding

$\varphi(y) =$ permutation matrix associated with y



Marginal polytope

$\mathcal{M} = \{P \in \mathbb{R}^{m \times m} \colon P^\top \mathbf{1}_m = 1, P\mathbf{1}_m = 1, 0 \leq P \leq 1\}$

$\triangle^{m \times m} \triangleq \{P \in \mathbb{R}^{m \times m} \colon P^\top \mathbf{1}_m = 1, 0 \leq P \leq 1\} \supset \mathcal{M}$

Row-stochastic matrices

Oracles

MAP: Hungarian, O(m³)

Eucl: LBFGS dual, O(m²/ε)

KL: Sinkhorn, O(m²/ε)

Marginal: **intractable**

Permutahedron

Output set

$\mathscr{Y} = \textbf{Permutations}([m])$

Permutahedron

Output set

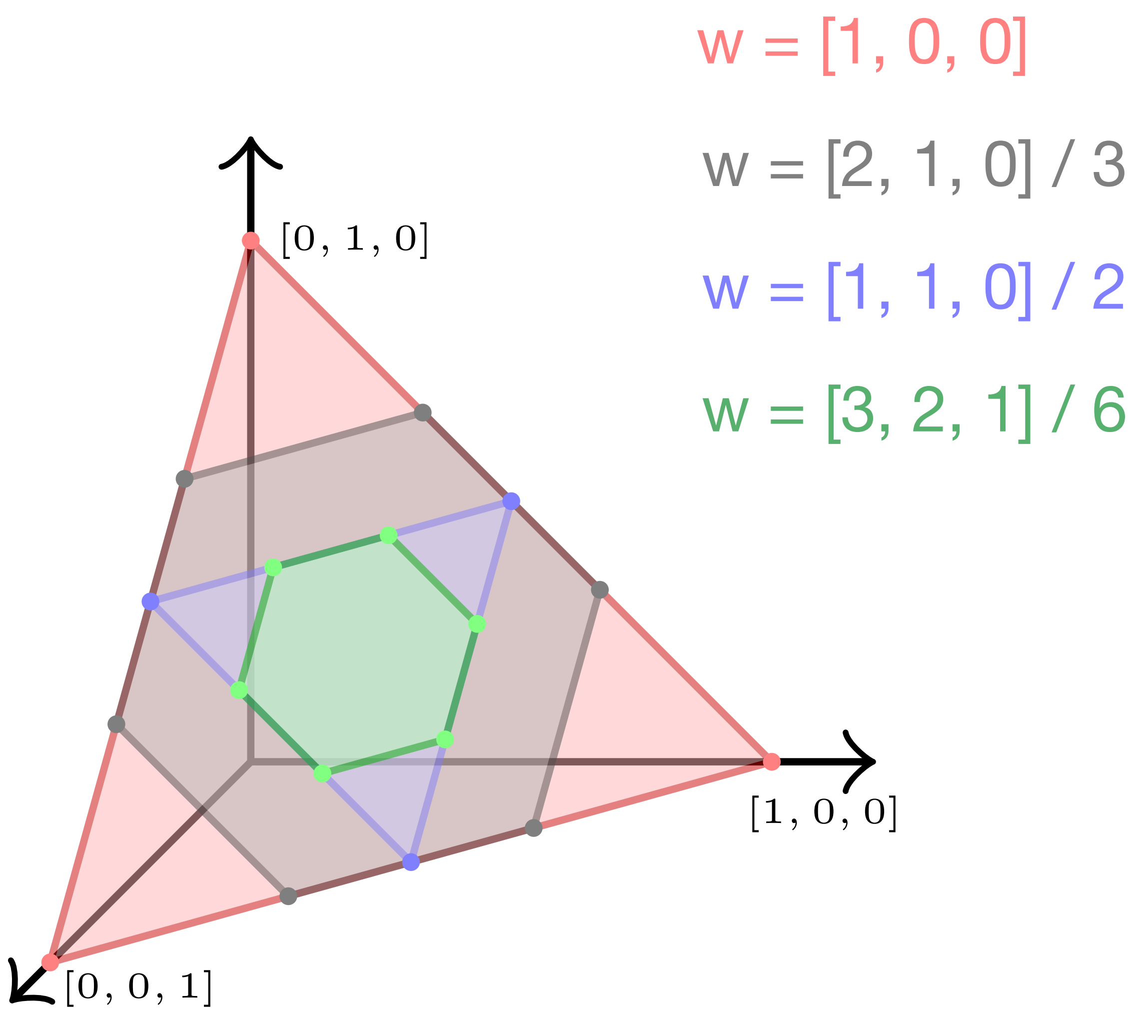
$$\mathcal{Y} = \textbf{Permutations}([m])$$

Encoding

$\varphi(y) =$ permutation of a vector w according to y

Marginal polytope

$$\mathcal{M} = \{\mu \in \mathbb{R}^m : \sum_{i \in S} \mu_i \leq \sum_{i=1}^{|S|} w_i \forall S \subset [m], \sum_{i=1}^{m} \mu_i = \sum_{i=1}^{m} w_i\}$$



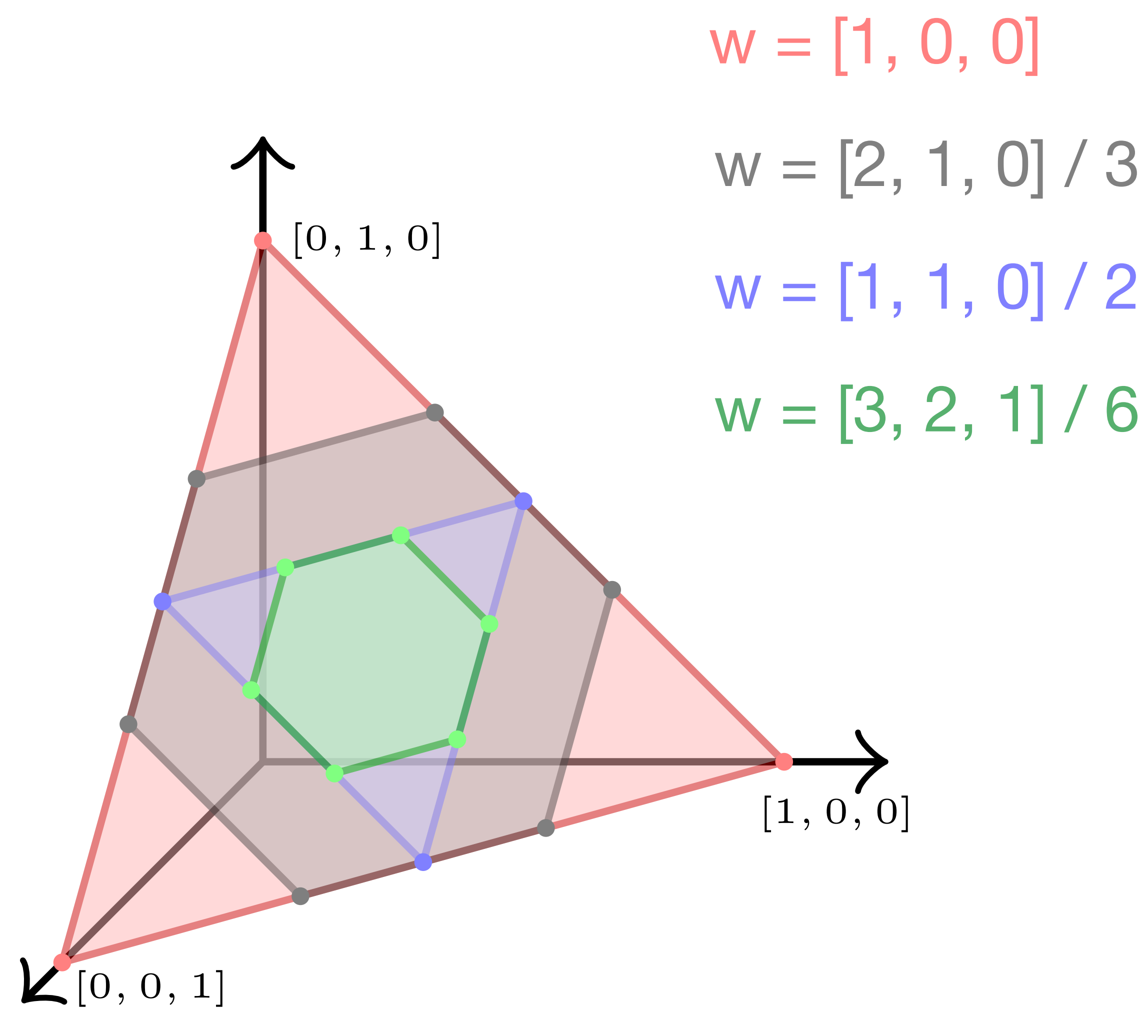
Permutahedron

Output set

$$\mathcal{Y} = \textbf{Permutations}([m])$$

Encoding

$\varphi(y) =$ permutation of a vector w according to y



Marginal polytope

$$\mathcal{M} = \{\mu \in \mathbb{R}^m \colon \sum_{i \in S} \mu_i \leq \sum_{i=1}^{|S|} w_i \forall S \subset [m], \sum_{i=1}^{m} \mu_i = \sum_{i=1}^{m} w_i\}$$

Oracles

MAP: O(m log m)

Eucl: isotonic reg, O(m log m)

KL: isotonic optimization

Outline

1. Background

2. Proposed framework

3. Experiments

Experiments

$$\frac{1}{n}\sum_{i=1}^{n} S_{\mathscr{C}}(Wx_i, y_i) + \lambda \|W\|_F^2$$

- Label ranking
- Ordinal regression
- Multilabel classification

Label ranking

Full-ranking supervision setting (no relevance scores)

e.g. $2 \succ 1 \succ 3 \succ 4$

Label ranking

Full-ranking supervision setting (no relevance scores)

e.g. $2 \succ 1 \succ 3 \succ 4$

Projection / Decoding	$\mathbb{R}^{m \times m}$ / $\mathcal{B}$
Authorship	5.70
Glass	7.11
Iris	19.26
Vehicle	9.04
Vowel	10.57
Wine	**1.23**

= squared loss

L = Hamming loss

Using Euclidean projections

Label ranking

Full-ranking supervision setting (no relevance scores)

e.g. $2 \succ 1 \succ 3 \succ 4$

Projection Decoding	$\mathbb{R}^{m \times m}$ $\mathcal{B}$	$[0, 1]^{m \times m}$ $\mathcal{B}$
Authorship	5.70	5.18
Glass	7.11	5.68
Iris	19.26	4.44
Vehicle	9.04	7.57
Vowel	10.57	9.56
Wine	**1.23**	1.85

= squared loss

L = Hamming loss

Using Euclidean projections

Label ranking

Full-ranking supervision setting (no relevance scores)

e.g. $2 \succ 1 \succ 3 \succ 4$

Projection Decoding	$\mathbb{R}^{m \times m}$ $\mathcal{B}$	$[0, 1]^{m \times m}$ $\mathcal{B}$	$\triangle^{m \times m}$ $\mathcal{B}$
Authorship	5.70	5.18	5.70
Glass	7.11	5.68	5.04
Iris	19.26	4.44	**1.48**
Vehicle	9.04	7.57	6.99
Vowel	10.57	9.56	9.18
Wine	**1.23**	1.85	1.85

= squared loss

L = Hamming loss

Using Euclidean projections

Label ranking

Full-ranking supervision setting (no relevance scores)

e.g. $2 \succ 1 \succ 3 \succ 4$

Projection	$\mathbb{R}^{m \times m}$	$[0,1]^{m \times m}$	$\triangle^{m \times m}$	$\mathcal{B}$
Decoding	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$
Authorship	5.70	5.18	5.70	**5.10**
Glass	7.11	5.68	5.04	**4.65**
Iris	19.26	4.44	**1.48**	2.96
Vehicle	9.04	7.57	6.99	**5.88**
Vowel	10.57	9.56	9.18	**8.76**
Wine	**1.23**	1.85	1.85	1.85

= squared loss

L = Hamming loss

Using Euclidean projections

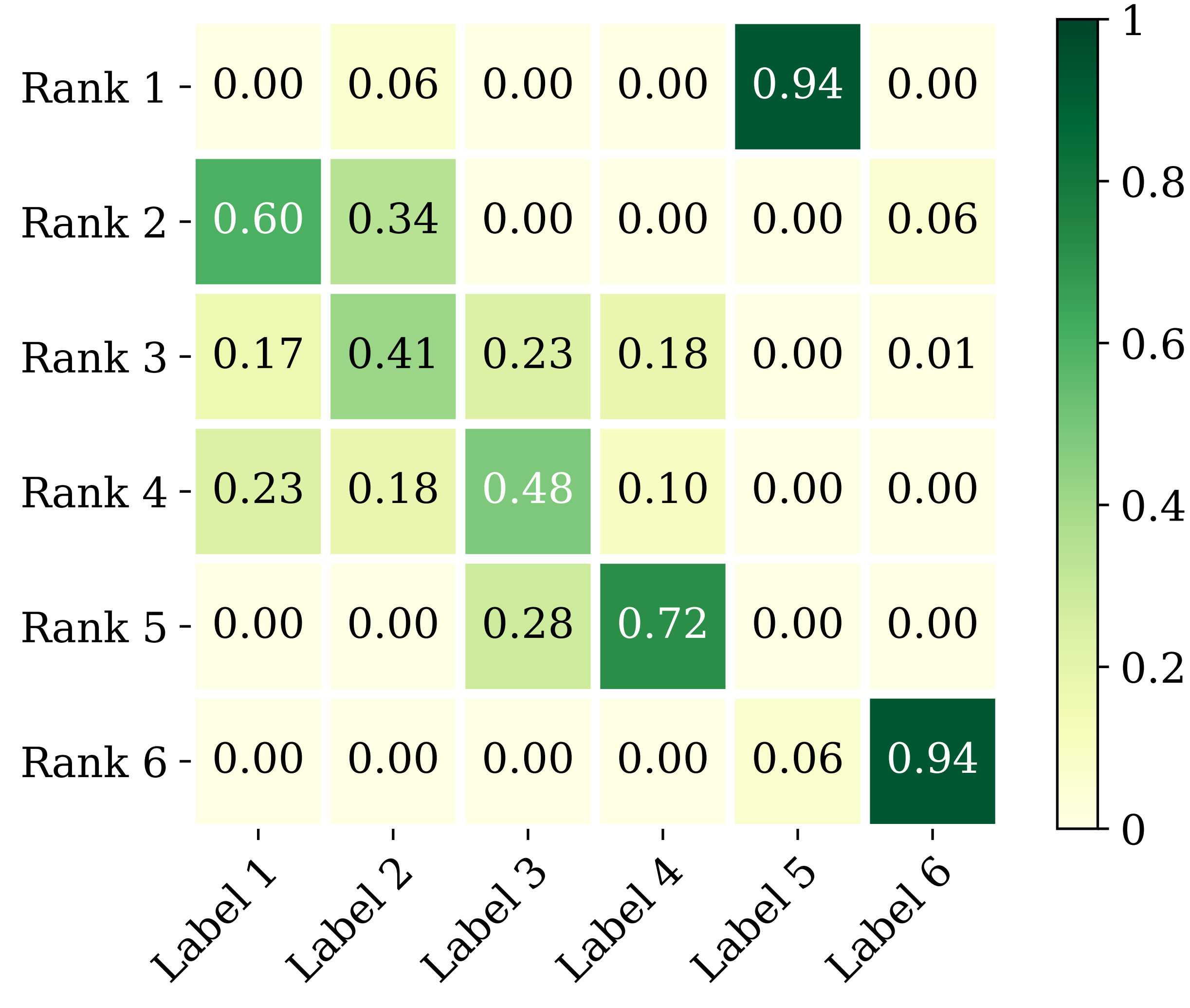
Label ranking

	Euclidean	KL
Projection	$\mathcal{B}$	$\mathcal{B}$
Decoding	$\mathcal{B}$	$\mathcal{B}$
Authorship	**5.10**	**5.10**
Glass	**4.65**	**4.65**
Iris	2.96	2.96
Vehicle	**5.88**	**6.25**
Vowel	**8.76**	**9.17**
Wine	1.85	**1.85**

Label ranking

	Euclidean	vs. KL
Projection	$\mathcal{B}$	$\mathcal{B}$
Decoding	$\mathcal{B}$	$\mathcal{B}$
Authorship	**5.10**	**5.10**
Glass	**4.65**	**4.65**
Iris	2.96	2.96
Vehicle	**5.88**	**6.25**
Vowel	**8.76**	**9.17**
Wine	1.85	**1.85**

$$P_{\mathscr{B}}(Wx)$$



“soft permutation matrix”

Label ranking

Birkhoff vs. permutahedron

		Linear	Poly 2	Poly 3
Projection	$\mathcal{B}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$
Decoding	$\mathcal{B}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$
Authorship	**5.10**	10.06	10.50	8.59
Glass	**4.65**	7.49	7.10	8.14
Iris	2.96	27.41	20.00	5.93
Vehicle	**5.88**	11.62	8.30	9.26
Vowel	**8.76**	14.35	11.74	10.21
Wine	1.85	8.02	3.08	6.79
	$W \in \mathbb{R}^{p \times m^2}$	$W \in \mathbb{R}^{p \times m}$	$W \in \mathbb{R}^{n \times m}$	$W \in \mathbb{R}^{n \times m}$

Using Euclidean projections

Ordinal regression

$$\mathcal{Y} = [k] \qquad 1 \prec \ldots \prec k$$

Ordinal regression

$$\mathscr{Y} = [k] \qquad 1 \prec \ldots \prec k$$

	Projection Decoding	Baseline
Average MAE		0.78
Average rank		4.75

Averaged over 16 datasets

L = MAE = Mean Absolute Error

Ordinal regression

$$\mathscr{Y} = [k] \qquad 1 \prec \ldots \prec k$$

Projection Decoding	Baseline	$\mathbb{R}$ Round
Average MAE	0.78	0.72
Average rank	4.75	2.9

Averaged over 16 datasets

L = MAE = Mean Absolute Error

Ordinal regression

$$\mathcal{Y} = [k] \qquad 1 \prec \ldots \prec k$$

Projection Decoding	Baseline	$\mathbb{R}$ Round
Average MAE	0.78	0.72
Average rank	4.75	2.9

Averaged over 16 datasets

L = MAE = Mean Absolute Error   OS = Order Simplex

Ordinal regression

$$\mathcal{Y} = [k] \qquad 1 \prec \ldots \prec k$$

Projection Decoding	Baseline	$\mathbb{R}$ Round	$\mathbb{R}^{k-1}$ $\mathcal{OS}$
Average MAE	0.78	0.72	0.47
Average rank	4.75	2.9	2.1

Averaged over 16 datasets

L = MAE = Mean Absolute Error   OS = Order Simplex

Ordinal regression

$$\mathcal{Y} = [k] \qquad 1 \prec \ldots \prec k$$

Projection Decoding	Baseline	$\mathbb{R}$ Round	$\mathbb{R}^{k-1}$ $\mathcal{OS}$	$[0,1]^{k-1}$ $\mathcal{OS}$
Average MAE	0.78	0.72	0.47	0.45
Average rank	4.75	2.9	2.1	1.6

Averaged over 16 datasets

L = MAE = Mean Absolute Error   OS = Order Simplex

Ordinal regression

$$\mathcal{Y} = [k] \qquad 1 \prec \ldots \prec k$$

Projection Decoding	Baseline	$\mathbb{R}$ Round	$\mathbb{R}^{k-1}$ $\mathcal{OS}$	$[0,1]^{k-1}$ $\mathcal{OS}$	$\mathcal{OS}$ $\mathcal{OS}$
Average MAE	0.78	0.72	0.47	0.45	0.43
Average rank	4.75	2.9	2.1	1.6	**1.5**

Averaged over 16 datasets

L = MAE = Mean Absolute Error OS = Order Simplex

Multilabel classification

lower bound = 0

upper bound = $\left\lceil \mathbb{E}[\,|Y|\,] + \sqrt{\mathbb{V}[\,|Y|\,]} \right\rceil$

F_1 score

Multilabel classification

lower bound = 0 upper bound = $\lceil \mathbb{E}[\,|Y|\,] + \sqrt{\mathbb{V}[\,|Y|\,]}\rceil$

Projection	$[0, 1]^k$
Decoding	$[0, 1]^k$
Birds	38.87
Emotions	56.60
Scene	61.06

$\mathcal{K}$: budget polytope

F_1 score

Multilabel classification

lower bound = 0 upper bound = $\lceil \mathbb{E}[\,|Y|\,] + \sqrt{\mathbb{V}[\,|Y|\,]}\rceil$

Projection	$[0,1]^k$	$\mathbb{R}^k$
Decoding	$[0,1]^k$	$\mathcal{K}$
Birds	38.87	37.75
Emotions	56.60	51.73
Scene	61.06	50.33

$\mathcal{K}$: budget polytope

F_1 score

Multilabel classification

lower bound = 0 upper bound = $\lceil \mathbb{E}[\,|Y|\,] + \sqrt{\mathbb{V}[\,|Y|\,]} \rceil$

Projection	$[0,1]^k$	$\mathbb{R}^k$	$[0,1]^k$
Decoding	$[0,1]^k$	$\mathcal{K}$	$\mathcal{K}$
Birds	38.87	37.75	39.21
Emotions	56.60	51.73	53.98
Scene	61.06	50.33	58.95

$\mathcal{K}$: budget polytope

F_1 score

Multilabel classification

lower bound = 0 upper bound = $\lceil \mathbb{E}[\,|Y|\,] + \sqrt{\mathbb{V}[\,|Y|\,]}\rceil$

Projection	$[0,1]^k$	$\mathbb{R}^k$	$[0,1]^k$	$\mathcal{K}$
Decoding	$[0,1]^k$	$\mathcal{K}$	$\mathcal{K}$	$\mathcal{K}$
Birds	38.87	37.75	39.21	**39.43**
Emotions	56.60	51.73	53.98	**62.57**
Scene	61.06	50.33	58.95	**69.01**

$\mathcal{K}$: budget polytope

F_1 score

Multilabel classification

lower bound = 0 upper bound = $\lceil \mathbb{E}[\,|Y|\,] + \sqrt{\mathbb{V}[\,|Y|\,]} \rceil$

Projection	$[0,1]^k$	$\mathbb{R}^k$	$[0,1]^k$	$\mathcal{K}$
Decoding	$[0,1]^k$	$\mathcal{K}$	$\mathcal{K}$	$\mathcal{K}$
Birds	38.87	37.75	39.21	**39.43**
Cal500	34.62	**35.86**	34.63	34.61
Emotions	56.60	51.73	53.98	**62.57**
Mediamill	**56.22**	55.35	**56.22**	54.53
Scene	61.06	50.33	58.95	**69.01**
TMC	**60.45**	58.61	60.37	60.25
Yeast	**60.24**	60.20	60.23	60.06

$\mathcal{K}$: budget polytope

F_1 score

Conclusion

Conclusion

- We proposed a generic framework for deriving a **loss** from the **projection** onto a convex set

Conclusion

- We proposed a generic framework for deriving a **loss** from the **projection** onto a convex set
- If its projection is affordable,
  the **marginal polytope** is the best convex set

Conclusion

- We proposed a generic framework for deriving a **loss** from the **projection** onto a convex set
- If its projection is affordable, the **marginal polytope** is the best convex set
- If not, any convex **superset** with cheaper projection can be used (e.g., unit cube)