# On the $k$-Support and Related Norms

*Massimiliano Pontil*

Department of Computer Science
Centre for Computational Statistics and Machine Learning
University College London

(Joint work with **Andrew McDonald** and **Dimitris Stamos**)

# Plan

- Problem
- Spectral regularization
- $k$-support norm
- Box norm
- Link to cluster norm

# Problem

- Learn a matrix from a set of linear measurements:

$$y_i = \langle W^*, X_i \rangle + \text{noise}_i, \quad i = 1, ..., n$$

- Method

$$\min_{W \in \mathbb{R}^{d \times m}} \sum_{i=1}^{n} (y_i - \langle W, X_i \rangle)^2 + \lambda \Omega(W)$$

- Matrix completion: $X_i = e_r e_c^\top$
- Multitask learning: $X_i = e_r x_i^\top$
- Regularizer $\Omega$ encourages matrix structure

# Spectral Regularization

$$\min_{W \in \mathbb{R}^{d \times m}} \sum_{i=1}^{n} (y_i - \langle W, X_i \rangle)^2 + \lambda \Omega(W)$$

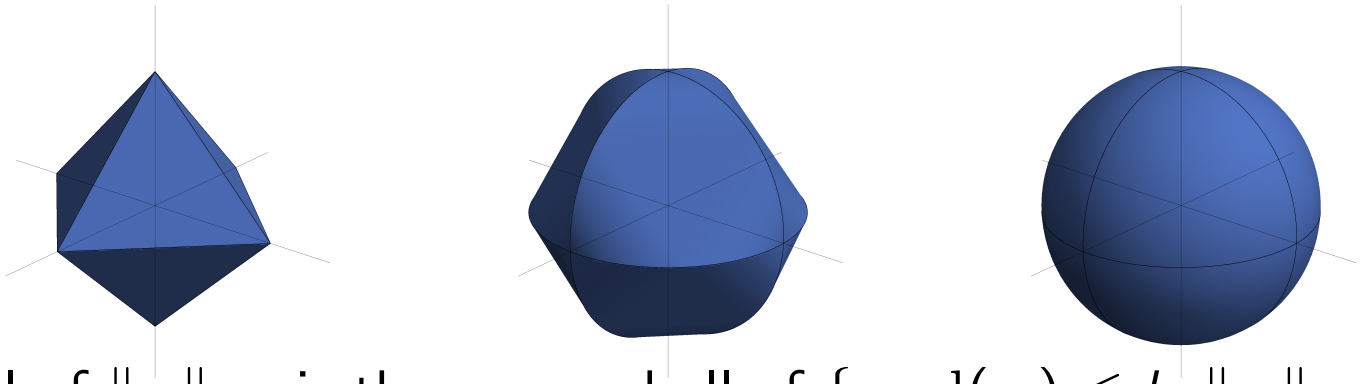
- $\Omega$ favors matrix structure (low rank, low variance, clustering, etc.)
- Choose an **OI-norm**: $\Omega(W) \equiv \|W\| = \|UWV\|, \ \forall U, V$ orthogonal
- von Neumann (1937): $\|W\| = g(\sigma(W))$, with $g$ is an SG-function
- Well studied example is **trace norm**: $g(\cdot) = \|\cdot\|_1$

# $k$-Support Norm [Argyriou et al. 2012]

- Special case of group lasso with overlap [Jacob et al., 2009]

$$\|w\|_{(k)} = \inf \left\{ \sum_{|J| \leq k} \|v_J\|_2 : \sum_{|J| \leq k} v_J = w, \ \mathrm{supp}(v_J) \subset J \right\}$$

- Includes the $\ell_1$-norm ($k = 1$) and $\ell_2$-norm ($k = d$)



- Unit ball of $\|\cdot\|_{(k)}$ is the convex hull of $\{\mathrm{card}(w) \leq k, \|w\|_2 \leq 1\}$

- Dual norm: $\|u\|_{*,(k)} = \sqrt{\sum_{i=1}^{k} (|u|_i^{\downarrow})^2}$

# Spectral $k$-Support Norm

$k$-support norm is an SG-function, inducing the OI-norm

$$\|W\|_{(k)} := \|\sigma(W)\|_{(k)}$$

- **Proposition.** Unit ball of $\|\sigma(\cdot)\|_{(k)}$ is the convex hull of

$$\{\mathrm{rank}(W) \leq k,\ \|W\|_{\mathrm{F}} \leq 1\}$$

- Includes trace norm ($k = 1$) and Frobenius norm ($k = d$)

# Matrix Completion Experiment

| dataset | norm | test error | $r$ | $k$ | $a$ |
|---|---|---|---|---|---|
| ML 100k $\rho = 50\%$ | tr | 0.2017 | 13 | - | - |
| | en | 0.2017 | 13 | - | - |
| | ks | 0.1990 | 9 | 1.87 | - |
| | box | 0.1989 | 10 | 2.00 | 1e-5 |
| ML 1M $\rho = 50\%$ | tr | 0.1790 | 17 | - | - |
| | en | 0.1789 | 15 | - | - |
| | ks | 0.1782 | 17 | 1.80 | - |
| | box | 0.1777 | 19 | 2.00 | 1e-6 |
| Jester1 20 per line | tr | 0.1752 | 11 | - | - |
| | en | 0.1752 | 11 | - | - |
| | ks | 0.1739 | 11 | 6.38 | - |
| | box | 0.1726 | 11 | 6.40 | 2e-5 |

# MTL Experiment

Table: Multitask learning clustering on Lenk dataset, with simple thresholding.

| dataset | norm | test error | $k$ | $a$ |
|---|---|---|---|---|
| Lenk 8 per task | fr | 3.7869 (0.07) | - | - |
| | tr | 1.9058 (0.04) | - | - |
| | en | 1.8974 (0.04) | - | - |
| | ks | 1.8933 (0.04) | 1.02 | - |
| | box | 1.8916 (0.04) | 1.01 | 5.5e-3 |
| | c-fr | 1.8667 (0.08) | - | - |
| | c-tr | 1.7904 (0.03) | - | - |
| | c-en | 1.7896 (0.03) | - | - |
| | c-ks | 1.7775 (0.03) | 1.89 | - |
| | c-box | 1.7754 (0.03) | 1.12 | 9.5e-3 |

# Box Norm

Let $\Theta \subseteq \mathbb{R}^d_{++}$, bounded and convex and consider the norm:

$$\|w\|^2_\Theta = \inf_{\theta\in\Theta} \sum_{i=1}^d \frac{w_i^2}{\theta_i}, \quad \|u\|^2_{*,\Theta} = \sup_{\theta\in\Theta} \sum_{i=1}^d \theta_i u_i^2$$

- Box norm: $\Theta = \left\{ a < \theta_i \leq b, \ \sum_{i=1}^d \theta_i \leq c \right\}$
- **Includes $k$-support norm** for $a = 0, b = 1, c = k$
- Unit ball is the convex hull of

$$\bigcup_{|J|\leq k} \left\{ w \in \mathbb{R}^d : \sum_{i\in J} \frac{w_i^2}{b} + \sum_{i\notin J} \frac{w_i^2}{a} \leq 1 \right\}$$

# Unit Balls

Figure: Unit balls of the box norm in $\mathbb{R}^2$ for $k = 1$, $a \in \{0.01, 0.25, 0.50\}$.

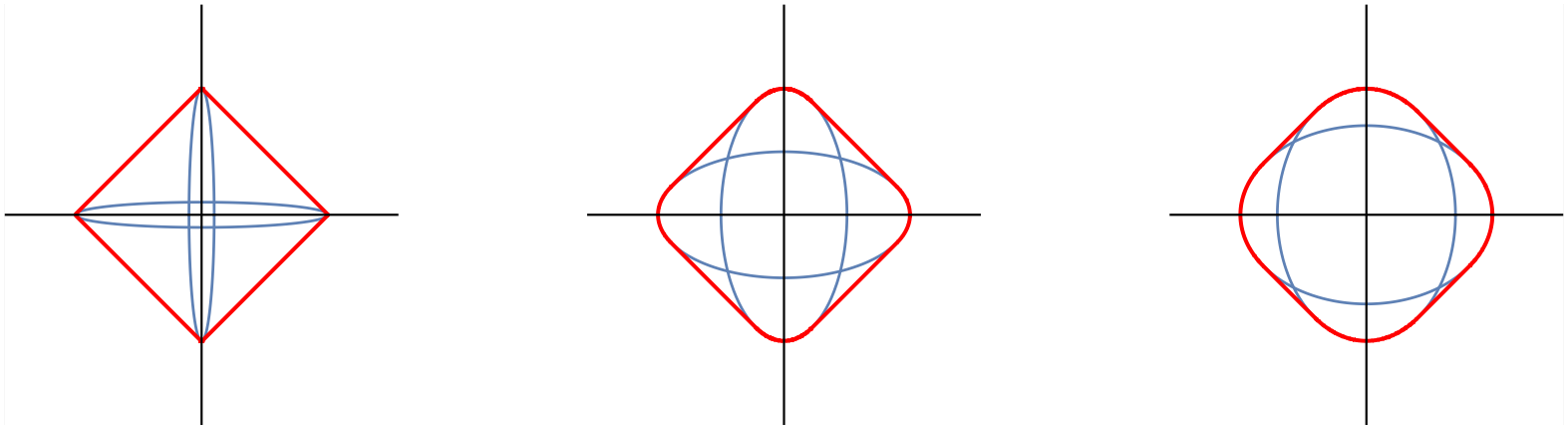
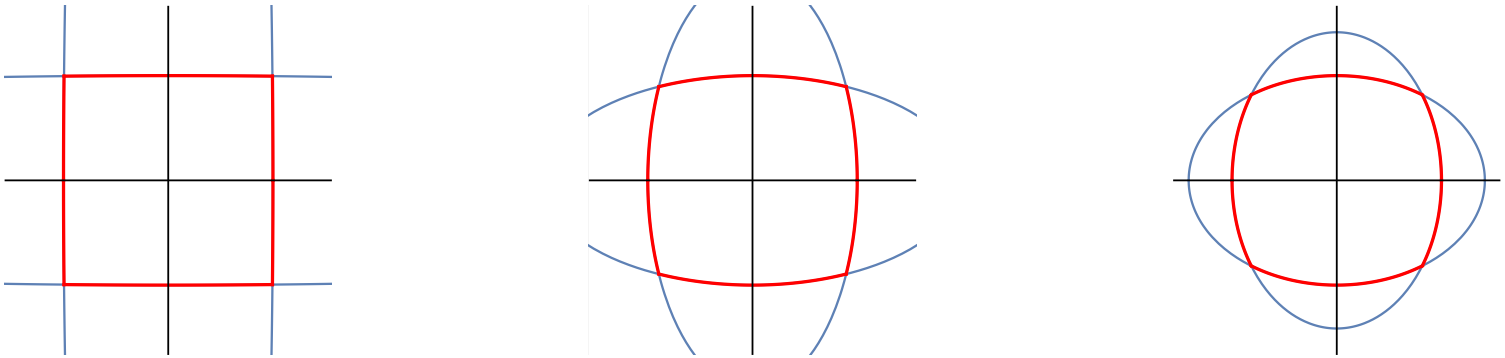


Figure: Unit balls of the dual box norm in $\mathbb{R}^2$ for $k = 1$, $a \in \{0.01, 0.25, 0.50\}$.



# Cluster Norm

- Box norm is an SG-function, inducing the OI-norm

$$\|W\|_\Theta^2 = \|\sigma(W)\|_\Theta^2 = \inf\left\{\sum_{i=1}^d \frac{\sigma_i(W)^2}{\theta_i} : \theta \in (a,b]^d, \sum_{i=1}^d \theta_i \leq c\right\}$$

- Associated OI-norm has been used to favour task clustering [Jacob et al. 2008]. It can be written as

$$\|W\|_\Theta^2 = \inf\left\{\mathrm{tr}(W\Sigma^{-1}W^T) : aI \preceq \Sigma \preceq bI,\ \mathrm{tr}\,\Sigma \leq c\right\}$$

- Includes spectral $k$-support norm for $a = 0$, $b = 1$, $c = k$

# Interpretation of "$a$"

**Proposition.** If $c = da + k(b-a)$, the solution of the regularization problem is given by $\hat{W} = \hat{V} + \hat{Z}$, where

$$(\hat{V}, \hat{Z}) = \arg\min_{V,Z} \sum_{i=1}^{n} (y_i - \langle V + Z, X_i \rangle)^2 + \lambda \left( \frac{1}{a} \|V\|_{\mathrm{F}}^2 + \frac{1}{b-a} \|Z\|_{(k)}^2 \right)$$

- Parameter '$a$' balances the relative importance of the two components

- Cluster norm is the Moureau envelope of spectral $k$-support norm:

$$\|W\|_{\Theta}^2 = \min_{Z \in \mathbb{R}^{d \times m}} \left\{ \frac{1}{a} \|W - Z\|_{\mathrm{F}}^2 + \frac{1}{b-a} \|Z\|_{(k)}^2 \right\}$$

# Computation of the $\Theta$ norm

Assume w.l.o.g. $w \geq 0$ with non increasing components

$$\|w\|_\Theta^2 = \tfrac{1}{b}\|w_{[1:q]}\|_2^2 + \tfrac{1}{c-qb-\ell a}\|w_{[q+1:d-\ell]}\|_1^2 + \tfrac{1}{a}\|w_{[\ell+1:d]}\|_2^2,$$

where $q, \ell \in \{0, \ldots, d\}$ are uniquely determined

In particular: $\|w\|_{(k)} = \|w_{[1:q]}\|_2^2 + \frac{1}{k-q}\|w_{[q+1:d]}\|_1^2$

where $q \in \{0, \ldots, k-1\}$ is determined by $|w|_q^\downarrow \geq \frac{1}{k-q}\sum_{j=q+1}^{d} |w|_j^\downarrow > |w|_{q+1}^\downarrow$

- Computation of norm is $O(d \log(d))$
- For $k$-support improves previous $O(kd)$ method
- Efficient optimization using proximal-gradient methods

# Extensions/Open Problems

- Other sets $\Theta$ allow for exact prox, e.g. $\Theta = \{\theta_1 \geq \ldots \theta_d > 0\}$. Can give a general characterization?
- Online learning / stochastic optimization
- Kernel extensions