Orbit Regularization

Renato Negrinho [1,2] André Martins [1,2]

[1]Priberam Labs, Lisbon, Portugal [2]Instituto de Telecomunicações, Lisbon, Portugal

Regularized Empirical Risk Minimization

$$\underset{w}{\text{minimize}} \quad \underbrace{L(w;\mathcal{D})}_{\text{Loss function}} + \underbrace{\lambda R(w)}_{\text{Regularizer}}$$

- Tradeoff between **data fitting** and **model complexity**
- Model w in vector space V (typically $\mathbb{R}^d$ or $\mathbb{R}^{m\times n}$)
- $R(w)$ can be designed to induce **structure** in the model (sparsity, low rank, variable grouping, etc.)

Our Approach: Majorization Theory [1]

$$\underset{w}{\text{minimize}} \quad L(w;\mathcal{D}) \qquad \text{subject to } w \preceq_G v$$

- **Key ingredients:** a group G and a prototype v
- Complexity is defined **relative to** v (via relation $\preceq_G$)
- Group G captures desired **complexity invariances**

Groups and Group Actions

A **group** is a tuple $(G,\cdot)$ satisfying *closure*, *associativity*, *existence of identity* and *existence of inverses*

- $\mathcal{P}$, **permutation matrices** under multiplication
- $\mathcal{P}_\pm$, **signed permutation matrices** under multiplic.
- $O(d)$, **orthogonal matrices** under multiplication

Examples of group actions $\phi: G\times V \to V$:

- $\mathcal{P}$ acting on $\mathbb{R}^d$ by permuting the coordinates
- $O(d)$ acting on $\mathbb{R}^d$ by left matrix multiplication

Orbits and Orbitopes

Orbit of $v\in V$ under the action of G:

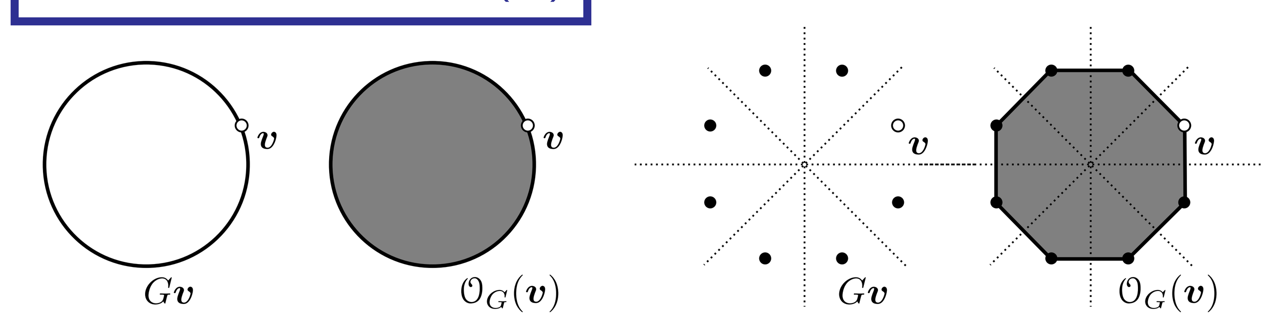
$$Gv := \{gv \mid g\in G\}, \qquad \text{where } gv \equiv \phi(g,v)$$

The convex hull of the orbit is called the **orbitope**:

$$\mathcal{O}_G(v) := \mathbf{conv}(Gv).$$

Orbitopes induce a **pre-order**:

$\boxed{w \preceq_G v \Leftrightarrow w \in \mathcal{O}_G(v)}$



Examples: Vector Case

- **ℓ_2 ball:**

$$\begin{aligned}\mathcal{O}_{O(d)}(v) &= \mathbf{conv}\{Uv \mid U\in O(d)\}\\ &= \mathbf{conv}\{w\in\mathbb{R}^d \mid \|w\|_2 = \|v\|_2\}\end{aligned}$$

- **Permutahedron:**

$$\begin{aligned}\mathcal{O}_{\mathcal{P}}(v) &= \mathbf{conv}\{Pv \mid P\in\mathcal{P}\}\\ &= \{Mv \mid M\mathbf{1}=\mathbf{1}, M^T\mathbf{1}=\mathbf{1}, M\geq 0\}\end{aligned}$$

- **Signed permutahedron:**

$$\mathcal{O}_{\mathcal{P}_\pm}(v) = \mathbf{conv}\{\mathbf{Diag}(s)Pv \mid P\in\mathcal{P}, s\in\{\pm1\}^d\}$$

Particular cases: **ℓ_1-ball** ($v=\gamma e_1$); **ℓ_∞-ball** ($v=\gamma\mathbf{1}$)

Examples: Matrix Case

- **Symmetric matrices with majorized eigenvalues:**

$$\begin{aligned}\mathcal{O}_{O(d)}(A) &= \mathbf{conv}\{UAU^T \mid U\in O(d)\}\\ &= \{B\in\mathbb{S}^d \mid \lambda(B)\preceq_{\mathcal{P}}\lambda(A)\}\end{aligned}$$

- **Squared matrices with majorized singular values:**

$$\begin{aligned}\mathcal{O}_{O(d)\times O(d)}(A) &= \mathbf{conv}\{UAV^T \mid U\in O(d), V\in O(d)\}\\ &= \{B\in\mathbb{R}^{d\times d} \mid \sigma(B)\preceq_{\mathcal{P}}\sigma(A)\}\end{aligned}$$

Particular cases: **spectral norm ball** ($A=\gamma I_d$); **nuclear norm ball** ($A=\gamma\,\mathbf{Diag}(e_1)$)

Example: Atomic Norms [2]

Proposition:
$-v\in\mathcal{O}_G(v) \;\Rightarrow\; \mathcal{O}_G(v)$ is an **atomic norm ball**

Corollaries:

- $\mathcal{O}_{\mathcal{P}_\pm}(v)$ is an atomic norm ball for any v
- $\mathcal{O}_{\mathcal{P}}(v)$ is an atomic norm ball for v of the form (v_+, v_-)

Duality: Permutahedra & Sorted ℓ_1 [3, 4]

Sorted ℓ_1 norm:

$$\|w\|_{\text{slope},v} := \textstyle\sum_{j=1}^d v_j|w|_{(j)}, \quad (v_1\geq\ldots\geq v_d\geq 0)$$

Proposition: the two norms are dual!

$$\|\cdot\|^\star_{\mathcal{P}_\pm v} = \|\cdot\|_{\text{slope},v}$$

Corollary: evaluating the prox of $\|\cdot\|_{\text{SLOPE},v}$ is all we need to project onto $\mathcal{O}_{\mathcal{P}_\pm}(v)$ (Moreau decomposition).
In the paper: similar result for unsigned permutahedron

Two Key Concepts

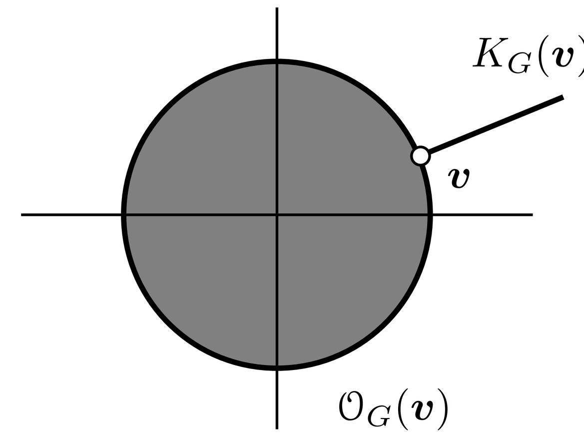
All we need for optimizing with orbitopes (arbitrary G):

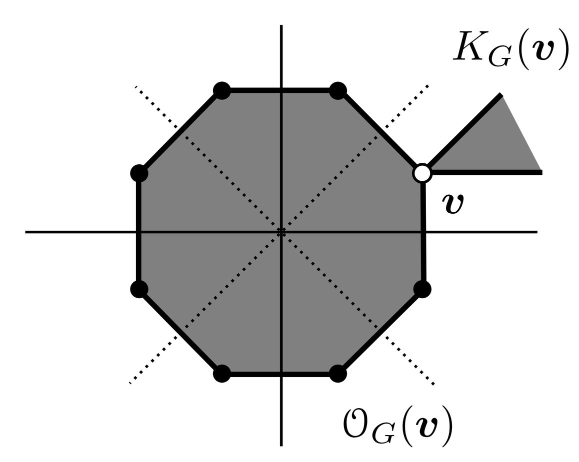
- **Matching function:**

$$m_G(u,v) = \sup\{\langle u,w\rangle \mid w\in Gv\}$$

- **Region cone:**

$$K_G(v) = \{u\in V \mid m_G(u,v) = \langle u,v\rangle\}$$





Frank Wolfe and Projected Gradient

The two ingredients above are all we need from G to train with Frank Wolfe or projected gradient:

	evaluate **matching function**	project onto **region cone**
Frank Wolfe	✓	
Projected Grad.	✓	✓

- **Both steps are easy and efficient for the permutahedron and signed permutahedron cases**

Continuation Algorithm

Given G, how to choose a good prototype v?
Answer: search in the space of all prototypes!

Continuation algorithm: gradually increase the ball (as homotopy methods), but also **shapes** it along the way

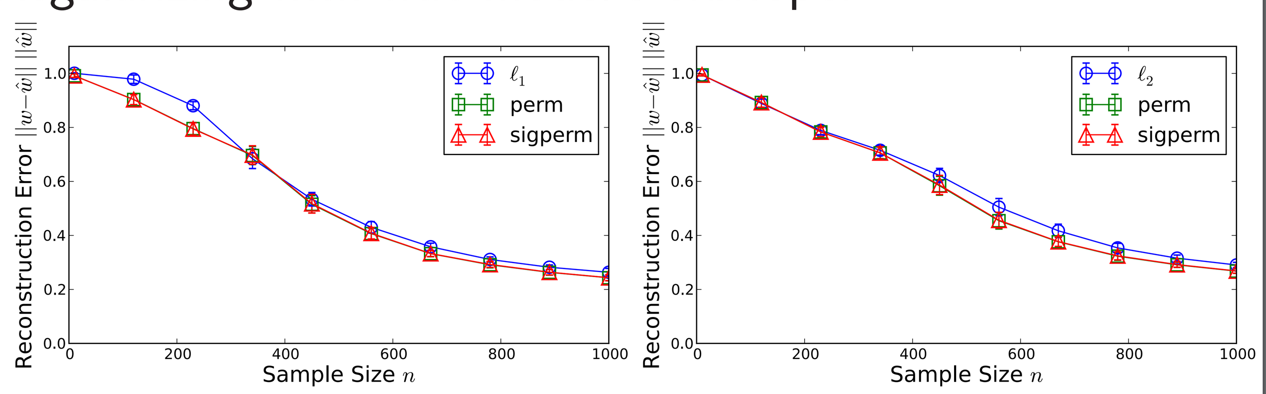
Require: Factor $\epsilon>0$, interpolation parameter $\alpha\in[0,1]$

1: Initialize prototype v_0 randomly and set $\|v_0\|=\epsilon$
2: **repeat** {for $t=0,1,\ldots$}
3: Solve $w_t = \arg\min_{w\preceq_G v_t} L(w;\mathcal{D})$
4: Pick $v_t'\in Gv_t\cap K_G(w_t)$
5: Set next prototype $v_{t+1} = (1+\epsilon)(\alpha v_t' + (1-\alpha)w_t)$
6: **until** $\|w_t\|_{Gv_t} < 1$.
7: Choose the best $\widehat{w}\in\{w_1,w_2,\ldots\}$ with C/V

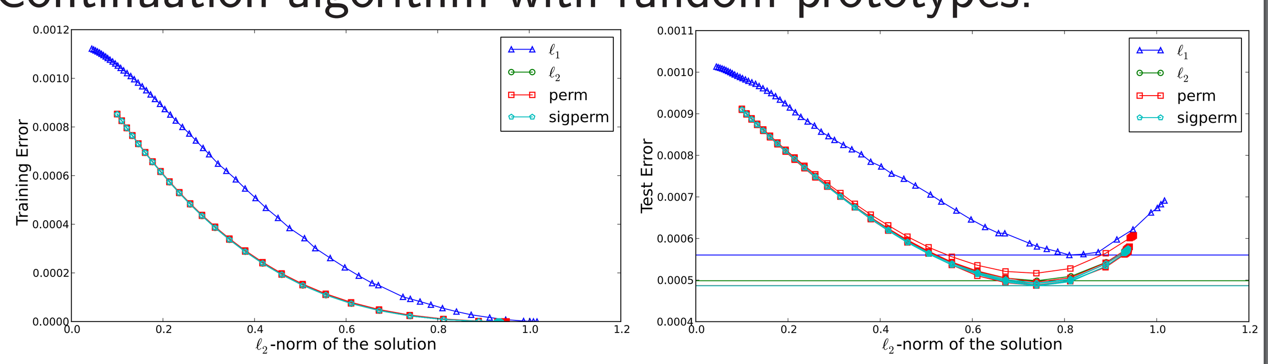
- Depends on prototype initialization
- Guaranteed to stop after a finite number of iterations

Simulation Results

Regularizing with true model's orbitope:



Continuation algorithm with random prototypes:



Conclusions and Future Work

Conclusions:

- New group-based regularization scheme via orbitopes
- Relation with atomic norms and sorted ℓ_1-norms
- Continuation algorithm for exploring regularization paths

Future work:

- Analysis for reflection groups
- Theoretical analysis of the continuation algorithm

Acknowledgements

This work was supported by FCT grants PTDC/EEI-SII/2312/2012 and PEst-OE/EEI/LA0008/2011, and by the EU/FEDER programme, QREN/POR Lisboa (Portugal), under the Intelligo project (contract 2012/24803).

References

[1] Marshall, Albert W., Ingram Olkin, and Barry C. Arnold. "Inequalities: Theory of Majorization and Its Applications." Springer, 2010.

[2] Chandrasekaran, Venkat, et al. "The convex geometry of linear inverse problems." Foundations of Computational Mathematics 12.6 (2012): 805-849.

[3] Bogdan, Malgorzata, et al. "Statistical estimation and testing via the ordered ℓ_1 norm." arXiv preprint arXiv:1310.1969 (2013).

[4] Zeng, Xiangrong, and Mário AT Figueiredo. "Decreasing Weighted Sorted ℓ_1 Regularization." arXiv preprint arXiv:1404.3184 (2014).