

Facilitating OWL Norm Minimizations

Simon Foucart*— Texas A&M University

Abstract

We present some characterizations of the ordered weighted ℓ_1 norm (aka sorted ℓ_1 norm) and of the vector Ky-Fan norm as solutions to linear programs involving reasonably many variables and constraints. Such linear characterizations can be exploited to recast and effortlessly solve a variety of convex optimization problems involving these norms. Similar linear characterizations are given for the dual norms of the ordered weighted ℓ_1 norm and the Ky-Fan norm.

Key words and phrases: sorted ℓ_1 norm, ordered weighted ℓ_1 norm, Ky-Fan norm, dual norms, structure-promoting minimization, duality in linear programming.

This note is concerned with the sorted ℓ_1 norm occurring in [4], and which is also called OWL norm in [6], as a shorthand for ordered weighted ℓ_1 norm. Given weights $w_1 \geq w_2 \geq \dots \geq w_n \geq 0$ with at least $w_1 > 0$, this norm is defined, for any $x \in \mathbb{R}^n$, by

$$(1) \quad \|x\|_{\text{OWL}} := \sum_{j=1}^n w_j x_j^*,$$

where $x_1^* \geq x_2^* \geq \dots \geq x_n^* \geq 0$ is the nondecreasing rearrangement of $|x_1|, |x_2|, \dots, |x_n|$. The fact that it is a norm is probably most easily seen from the following restatement of a classical rearrangement inequality:

$$(2) \quad \|x\|_{\text{OWL}} = \max_{\sigma \in S_n} \sum_{j=1}^n w_j |x_{\sigma(j)}|,$$

where S_n denotes the set of all permutations of $\{1, \dots, n\}$ (i.e., the symmetric group of degree n). Thus, it is apparent that SLOPE, introduced in [4] and further studied e.g. in [2, 7], which consists in solving

$$(3) \quad \underset{z \in \mathbb{R}^n}{\text{minimize}} \frac{1}{2} \|y - Az\|_2^2 + \|z\|_{\text{OWL}},$$

is a convex optimization problem. Its solution is usually computed by proximal gradient descent. This note aims to showcase an alternative way of solving (3)—in fact, a way of solving many

*S. F. is partially supported by NSF grants DMS-1622134 and DMS-1664803, and also acknowledges the NSF grant CCF-1934904.

convex optimization problems involving the OWL norm, and of doing so without any algorithmic adjustment so that all-purpose solvers can be relied on. For instance, as illustrated in the MATLAB file accompanying this note, the problem

$$(4) \quad \underset{z \in \mathbb{R}^n}{\text{minimize}} \|z\|_{\text{OWL}} \quad \text{subject to } Az = y$$

can be simply solved after recasting it as a linear program, namely as

$$(5) \quad \underset{z, a, b \in \mathbb{R}^n}{\text{minimize}} \sum_{j=1}^n (a_j + b_j) \quad \text{subject to } Az = y, \quad -(a_k + b_\ell) \leq w_k z_\ell \leq a_k + b_\ell \text{ for all } k, \ell.$$

This observation is based on a linear characterization of the OWL norm that strangely seems to have gone unnoticed so far.

Theorem 1. For any $x \in \mathbb{R}^n$,

$$(6) \quad \|x\|_{\text{OWL}} = \max_{S \in \mathbb{R}^{n \times n}} \left\{ \sum_{j=1}^n w_j (S|x|)_j : S \geq 0, \sum_k S_{k,\ell} = 1 \text{ for all } \ell, \sum_\ell S_{k,\ell} = 1 \text{ for all } k \right\}$$

$$(7) \quad = \min_{a, b \in \mathbb{R}^n} \left\{ \sum_{j=1}^n a_j + \sum_{j=1}^n b_j : -(a_k + b_\ell) \leq w_k x_\ell \leq a_k + b_\ell \text{ for all } k, \ell \right\}.$$

Proof. The expression (6) results from (2) and from Birkhoff's theorem [1, page 37] stating that the extreme points of the set of doubly stochastic matrices are the permutation matrices. The expression (7) follows from (6) by invoking duality in linear programming (see e.g. [3, page 225] read from the bottom up). $\square$

Note that the expression (2) would also have provided a linear characterization of the OWL norm by introducing slack variables $c^{(\sigma)} \in \mathbb{R}^n$, $\sigma \in S_n$, such that $|x_{\sigma(j)}| \leq c_j^{(\sigma)}$ for all $j \in \{1, \dots, n\}$, but the resulting number of variables would have been too large for practical purposes. Here, the order n^2 for the number of variables/constraints in the linear characterizations (6)-(7) is manageable. This number can even be reduced in some situations of interest. These include the (vector versions) of the Ky-Fan norms, corresponding to the choice of weights $w_1 = \dots = w_k = 1$ and $w_{k+1} = \dots = w_n = 0$ for some $k \in \{1, \dots, n\}$. Precisely, the norm defined, for any $x \in \mathbb{R}^n$, by

$$(8) \quad \|x\|_{(k)} = \sum_{j=1}^k x_j^* = \max_{j_1 < \dots < j_k} \sum_{i=1}^k |x_{j_i}|$$

admits the following linear characterizations involving a number of variables/constraints of order n .

Theorem 2. For any $x \in \mathbb{R}^n$,

$$(9) \quad \|x\|_{(k)} = \max_{u, v \in \mathbb{R}^n} \left\{ \sum_{j=1}^n v_j x_j : \sum_{j=1}^n u_j \leq k, -u_j \leq v_j \leq u_j, u_j \leq 1 \text{ for all } j \right\}$$

$$(10) \quad = \min_{a, b, \alpha \in \mathbb{R}^n, \beta \in \mathbb{R}} \left\{ \sum_{j=1}^n \alpha_j + k\beta : a + b = x, -\alpha_j \leq a_j \leq \alpha_j, -\beta \leq b_j \leq \beta \text{ for all } j \right\}.$$

Proof. Observe first that the expression (8) can be written as

$$(11) \quad \|x\|_{(k)} = \max_{v \in \mathbb{R}^n} \left\{ \sum_{j=1}^n v_j x_j : \|v\|_\infty \leq 1, \|v\|_0 := \sum_{j=1}^n \mathbb{1}_{\{v_j \neq 0\}} \leq k \right\}.$$

As a consequence of [8, Lemma 5.2 p 465], see also [5, Lemma 1.1], we have

$$(12) \quad \text{conv}\{v \in \mathbb{R}^n : \|v\|_\infty \leq 1, \|v\|_0 \leq k\} = \{v \in \mathbb{R}^n : \|v\|_\infty \leq 1, \|v\|_1 \leq k\},$$

from where we derive that

$$(13) \quad \|x\|_{(k)} = \max_{v \in \mathbb{R}^n} \left\{ \sum_{j=1}^n v_j x_j : \|v\|_\infty \leq 1, \|v\|_1 \leq k \right\}.$$

The expression (9) follows by introducing a vector $u \in \mathbb{R}^n$ of slack variables such that $|v_j| \leq u_j$ for all $j \in \{1, \dots, n\}$. As for the expression (10), by adapting a well-known characterization of the Ky-Fan norm (see [1, Proposition IV.2.3]) from matrices to vectors, we observe that

$$(14) \quad \|x\|_{(k)} = \min_{a, b \in \mathbb{R}^n} \{ \|a\|_1 + k\|b\|_\infty : a + b = x \}.$$

We then conclude by introducing slack variables $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ such that $|a_j| \leq \alpha_j$ and $|b_j| \leq \beta$ for all $j \in \{1, \dots, n\}$. $\square$

As consequences of the linear characterizations for the OWL and Ky-Fan norms, we can now deduce linear characterizations for their dual norms, starting with the dual OWL norm.

Theorem 3. For any $x \in \mathbb{R}^n$,

$$(15) \quad \|x\|_{\text{OWL}}^* = \max_{z, a, b \in \mathbb{R}^n} \left\{ \sum_{j=1}^n x_j z_j : \sum_{j=1}^n (a_j + b_j) \leq 1, -(a_k + b_\ell) \leq w_k z_\ell \leq a_k + b_\ell \text{ for all } k, \ell \right\}$$

$$(16) \quad = \min_{c \in \mathbb{R}, U, V \in \mathbb{R}^{n \times n}} \left\{ c : U, V \geq 0, (U - V)w = x, \sum_{\ell} (U_{i,\ell} + V_{i,\ell}) = c \text{ for all } i \right. \\ \left. \sum_k (U_{k,j} + V_{k,j}) = c \text{ for all } j \right\}.$$

Proof. In view of the definition of the dual OWL norm, i.e., of

$$(17) \quad \|x\|_{\text{OWL}}^* := \max_{z \in \mathbb{R}^n} \{ \langle x, z \rangle : \|z\|_{\text{OWL}} \leq 1 \},$$

the characterization (15) immediately follows from (7). As for the characterization (16), it is deduced from (15) by invoking duality in linear programming (see e.g. [3, page 224] read from the bottom up). $\square$

It has to be noted that a linear characterization—but one involving too many variables— could be derived from the expression of the dual OWL norm obtained in [9, Theorem 1]. In the case of the dual Ky-Fan norm (i.e., taking $w_1 = \dots = w_k = 1$ and $w_{k+1} = \dots = w_n = 0$ for some $k \in \{1, \dots, n\}$), the expression of [9] reduces to

$$(18) \quad \|x\|_{(k)}^* = \max \left\{ \|x\|_{\infty}, \frac{\|x\|_1}{k} \right\}.$$

This identity can easily be explained from (14) and from the well-known fact that, for two arbitrary norms $\|\cdot\|_{(1)}$ and $\|\cdot\|_{(2)}$, the norm defined by $\|x\| = \max\{\|x\|_{(1)}, \|x\|_{(2)}\}$ admits the dual norm given by $\|z\|^* = \inf\{\|a\|_{(1)}^* + \|b\|_{(2)}^* : a + b = z\}$. From here, we conclude this note by presenting some linear characterizations of the dual Ky-Fan norm that involve a number of variables/constraints only of order n , rather than the order nk resulting from an application of Theorem 3 with $w = [1; \dots; 1; 0; \dots; 0]$.

Theorem 4. For $x \in \mathbb{R}^n$,

$$(19) \quad \|x\|_{(k)}^* = \max_{z, a, b, \alpha \in \mathbb{R}^n, \beta \in \mathbb{R}} \left\{ \sum_{j=1}^n x_j z_j : \sum_{j=1}^n \alpha_j + k\beta \leq 1, \ a + b = z, \right. \\ \left. -\alpha_j \leq a_j \leq \alpha_j, \ -\beta \leq b_j \leq \beta \text{ for all } j \right\}$$

$$(20) \quad = \min_{c, \alpha \in \mathbb{R}, \beta \in \mathbb{R}^n} \left\{ c : \alpha \leq c, \ \sum_{j=1}^n \beta_j \leq kc, \ -\alpha \leq x_j \leq \alpha, \ -\beta_j \leq x_j \leq \beta_j \text{ for all } j \right\}.$$

Proof. The characterization (19) follows from the definition $\|x\|_{(k)}^* := \max_{z \in \mathbb{R}^n} \{ \langle x, z \rangle : \|z\|_{(k)} \leq 1 \}$ of the dual Ky-Fan norm and from (10). As for the characterization (20), it is deduced from the abridged expression (18) by writing $\|x\|_{(k)}^* = \min\{c : \|x\|_{\infty} \leq c \text{ and } \|x\|_1/k \leq c\}$ and by introducing slack variables $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^n$ such that $\alpha \leq c$, $|x_j| \leq \alpha$ for all $j \in \{1, \dots, n\}$, $(\sum_j \beta_j)/k \leq c$, and $|x_j| \leq \beta_j$ for all $j \in \{1, \dots, n\}$. $\square$

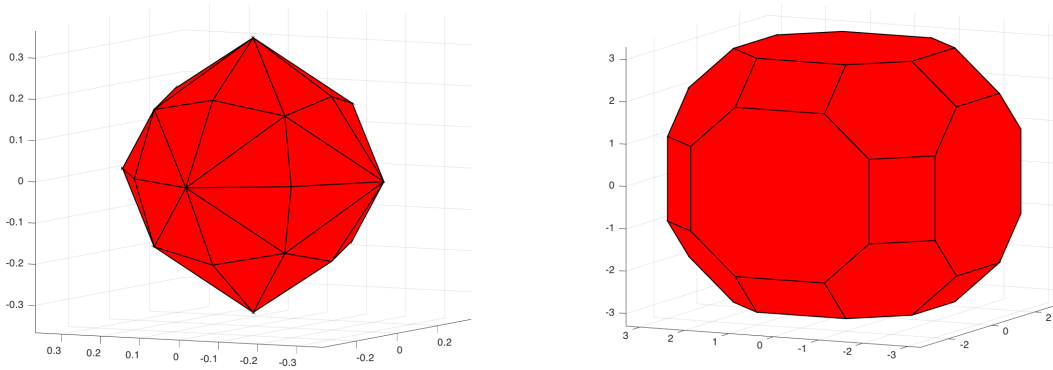


Figure 1: The unit OWL ball and the dual unit OWL ball for the weight $w = [3; 2; 1]$.

References

- [1] R. Bhatia. Matrix Analysis. Vol. 169. Springer Science & Business Media, 2013.
- [2] P. C. Bellec, G. Lécué, and A. B. Tsybakov. *SLOPE meets LASSO: improved oracle bounds and optimality*. The Annals of Statistics 46.6B (2018): 3603–3642.
- [3] S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, 2004.
- [4] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. *SLOPE—adaptive variable selection via convex optimization*. The Annals of Applied Statistics 9.3 (2015): 1103.
- [5] T. Cai and A. Zhang. *Sparse representation of a polytope and recovery of sparse signals and low-rank matrices*. IEEE Transactions on Information Theory 60 (2014): 122–132 .
- [6] M. A. T. Figueiredo and R. D. Nowak. *Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization*. arXiv preprint arXiv:1409.4005 (2014).
- [7] H. Hu and Y. M. Lu. *Asymptotics and optimal designs of SLOPE for sparse linear regression*. arXiv preprint arXiv:1903.11582 (2019).
- [8] G. Lorentz, M. von Golitschek, and Y. Makovoz. Constructive Approximation: Advanced Problems. Springer, Berlin, 1996.
- [9] X. Zeng and M. A. T Figueiredo. *Decreasing weighted sorted ℓ_1 regularization*. IEEE Signal Processing Letters 21.10 (2014): 1240–1244.