Spectral k-Support Norm Regularization

Andrew M. McDonald, Massimiliano Pontil, Dimitris Stamos
Department of Computer Science
University College London
{a.mcdonald,m.pontil,d.stamos}@cs.ucl.ac.uk

Abstract

The k-support norm has successfully been applied to sparse vector prediction problems. We observe that it belongs to a wider class of norms, which we call the box-norms. Within this framework we derive an efficient algorithm to compute the proximity operator of the squared norm, improving upon the original method for the k-support norm. We extend the norms from the vector to the matrix setting and we introduce the spectral k-support norm. We study its properties and show that it is closely related to the multitask learning cluster norm. We apply the norms to real and synthetic matrix completion datasets. Our findings indicate that spectral k-support norm regularization gives state of the art performance, consistently improving over trace norm regularization and the matrix elastic net.

1 Introduction

In recent years there has been a great deal of interest in the problem of learning a low rank matrix from a set of linear measurements. A widely studied and successful instance of this problem arises in the context of matrix completion or collaborative filtering, in which we want to recover a low rank (or approximately low rank) matrix from a small sample of its entries, see e.g. [1, 2]. One prominent method to solve this problem is trace norm regularization: we look for a matrix which closely fits the observed entries and has a small trace norm (sum of singular values) [3, 4, 5]. Besides collaborative filtering, this problem has important applications ranging from multitask learning, to computer vision and natural language processing, to mention but a few.

In this paper, we propose new techniques to learn low rank matrices. These are inspired by the notion of the k-support norm [6], which was recently studied in the context of sparse vector prediction and shown to empirically outperform the Lasso [7] and Elastic Net [8] penalties. We note that this norm can naturally be extended to the matrix setting and its characteristic properties relating to the cardinality operator translate in a natural manner to matrices. Our approach is suggested by the observation that the k-support norm belongs to a broader class of norms, which makes it apparent that they can be extended to spectral matrix norms. Moreover, it provides a link between the spectral k-support norm and the *cluster norm*, a regularizer introduced in the context of multitask learning [9]. This result allows us to interpret the spectral k-support norm as a special case of the cluster norm and furthermore adds a new perspective of the cluster norm as a perturbation of the former.

The main contributions of this paper are threefold. First, we show that the k-support norm can be written as a parametrized infimum of quadratics, which we term the *box-norms*, and which are symmetric gauge functions. This allows us to extend the norms to orthogonally invariant matrix norms using a classical result by von Neumann [10]. Second, we show that the spectral box-norm is essentially equivalent to the cluster norm, which in turn can be interpreted as a perturbation of the spectral k-support norm, in the sense of the Moreau envelope [11]. Third, we use the infimum framework to compute the box-norm and the proximity operator of the squared norm in $\mathcal{O}(d \log d)$ time. Apart from improving on the $\mathcal{O}(d(k + \log d))$ algorithm in [6], this method allows one to use optimal first order optimization algorithms [12] with the cluster norm. Finally, we present numerical

experiments which indicate that the spectral k-support norm shows a significant improvement in performance over regularization with the trace norm and the matrix elastic net, on four popular matrix completion benchmarks.

The paper is organized as follows. In Section 2 we recall the k-support norm, and define the box-norm. In Section 3 we study their properties, we introduce the corresponding spectral norms, and we observe the connection to the cluster norm. In Section 4 we compute the norm and we derive a fast method to compute the proximity operator. Finally, in Section 5 we report on our numerical experiments. The supplementary material contains derivations of the results in the body of the paper.

2 Preliminaries

In this section, we recall the k-support norm and we introduce the box-norm and its dual. The k-support norm $\|\cdot\|_{(k)}$ was introduced in [6] as the norm whose unit ball is the convex hull of the set of vectors of cardinality at most k and ℓ_2-norm no greater than one. The authors show that the k-support norm can be written as the infimal convolution [11]

$$\|w\|_{(k)} = \inf\left\{\sum_{g\in\mathcal{G}_k} \|v_g\|_2 : v_g \in \mathbb{R}^d,\ \mathrm{supp}(v_g) \subseteq g,\ \sum_{g\in\mathcal{G}_k} v_g = w\right\},\quad w\in\mathbb{R}^d, \tag{1}$$

where $\mathcal{G}_k$ is the collection of all subsets of $\{1,\dots,d\}$ containing at most k elements, and for any $v\in\mathbb{R}^d$, the set $\mathrm{supp}(v) = \{i : v_i \neq 0\}$ denotes the support of v. When used as a regularizer, the norm encourages vectors w to be a sum of a limited number of vectors with small support. The k-support norm is a special case of the group lasso with overlap [13], where the cardinality of the support sets is at most k. Despite the complicated form of the primal norm, the dual norm has a simple formulation, namely the ℓ_2-norm of the k largest components of the vector

$$\|u\|_{*,(k)} = \sqrt{\sum_{i=1}^{k}(|u|_i^{\downarrow})^2},\quad u\in\mathbb{R}^d, \tag{2}$$

where $|u|^{\downarrow}$ is the vector obtained from u by reordering its components so that they are non-increasing in absolute value [6]. The k-support norm includes the ℓ_1-norm and ℓ_2-norm as special cases. This is clear from the dual norm since for $k=1$ and $k=d$, it is equal to the ℓ_∞-norm and ℓ_2-norm, respectively. We note that while definition (1) involves a combinatorial number of variables, [6] observed that the norm can be computed in $\mathcal{O}(d\log d)$.

We now define the box-norm, and in the following section we will show that the k-support norm is a special case of this family.

Definition 2.1. Let $0\le a\le b$ and $c\in[ad, bd]$ and let $\Theta = \{\theta\in\mathbb{R}^d : a\le\theta_i\le b, \sum_{i=1}^d\theta_i\le c\}$. The box-norm is defined as

$$\|w\|_\Theta = \sqrt{\inf_{\theta\in\Theta}\sum_{i=1}^{d}\frac{w_i^2}{\theta_i}},\quad w\in\mathbb{R}^d. \tag{3}$$

This formulation will be fundamental in deriving the proximity operator in Section 4.1. Note that we may assume without loss of generality that $b=1$, as by rescaling we obtain an equivalent norm, however we do not explicitly fix b in the sequel.

Proposition 2.2. *The norm* (3) *is well defined and the dual norm is* $\|u\|_{*,\Theta} = \sqrt{\sup_{\theta\in\Theta}\sum_{i=1}^{d}\theta_i u_i^2}$.

The result holds true in the more general case that Θ is a bounded convex subset of the strictly positive orthant (for related results see [14, 15, 16, 17, 18, 19] and references therein). In this paper we limit ourselves to the box constraints above. In particular we note that the constraints are invariant with respect to permutation of the components of Θ, and as we shall see this property is key to extend the norm to matrices.

3 Properties of the Norms

In this section, we study the properties of the vector norms, and we extend the norms to the matrix setting. We begin by deriving the dual box-norm.

Proposition 3.1. *The dual box-norm is given by*

$$\|u\|_{*,\Theta} = \sqrt{a\|u\|_2^2 + (b-a)\|u\|_{*,(k)}^2 + (b-a)(\rho-k)(|u|_{k+1}^{\downarrow})^2}, \tag{4}$$

where $\rho = \frac{c-da}{b-a}$ *and* k *is the largest integer not exceeding* ρ*.*

We see from (4) that the dual norm decomposes into two ℓ_2-norms plus a residual term, which vanishes if $\rho = k$, and for the rest of this paper we assume this holds, which loses little generality.

Note that setting $a = 0, b = 1$, and $c = k \in \{1, \ldots, d\}$, the dual box-norm (4) is the ℓ_2-norm of the largest k components of u, and we recover the dual k-support norm in equation (2). It follows that the k-support norm is a box-norm with parameters $a = 0$, $b = 1$, $c = k$.

The following infimal convolution interpretation of the box-norm provides a link between the box-norm and the k-support norm, and illustrates the effect of the parameters.

Proposition 3.2. *If* $0 < a \leq b$ *and* $c = (b-a)k + da$*, for* $k \in \{1, \ldots, d\}$*, then*

$$\|w\|_\Theta = \inf \left\{ \sum_{g \in \mathcal{G}_k} \sqrt{\sum_{i \in g} \frac{v_{g,i}^2}{b} + \sum_{i \notin g} \frac{v_{g,i}^2}{a}} : v_g \in \mathbb{R}^d, \ \sum_{g \in \mathcal{G}_k} v_g = w \right\}. \tag{5}$$

Notice that if $b = 1$, then as a tends to zero, we obtain the expression of the k-support norm (1), recovering in particular the support constraints. If a is small and positive, the support constraints are not imposed, however effectively most of the weight for each v_g tends to be concentrated on $\text{supp}(g)$. Hence, Proposition 3.2 suggests that the box-norm regularizer will encourage vectors w whose dominant components are a subset of a union of a small number of groups $g \in \mathcal{G}_k$.

The previous results have characterized the k-support norm as a special case of the box-norm. Conversely, the box-norm can be seen as a perturbation of the k-support norm with a quadratic term.

Proposition 3.3. *Let* $\|\cdot\|_\Theta$ *be the box-norm on* $\mathbb{R}^d$ *with parameters* $0 < a < b$ *and* $c = k(b-a)+da$*, for* $k \in \{1, \ldots, d\}$*, then*

$$\|w\|_\Theta^2 = \min_{z \in \mathbb{R}^d} \left\{ \frac{1}{a}\|w - z\|_2^2 + \frac{1}{b-a}\|z\|_{(k)}^2 \right\}. \tag{6}$$

Consider the regularization problem $\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda \|w\|_\Theta^2$, with data X and response y. Using Proposition 3.3 and setting $w = u + z$, we see that this problem is equivalent to

$$\min_{u,z \in \mathbb{R}^d} \left\{ \|X(u+z) - y\|_2^2 + \frac{\lambda}{a}\|u\|_2^2 + \frac{\lambda}{b-a}\|z\|_{(k)}^2 \right\}.$$

Furthermore, if $(\hat{u}, \hat{z})$ solves this problem then $\hat{w} = \hat{u} + \hat{z}$ solves problem (6). The solution $\hat{w}$ can therefore be interpreted as the superposition of a vector which has small ℓ_2 norm, and a vector which has small k-support norm, with the parameter a regulating these two components. Specifically, as a tends to zero, in order to prevent the objective from blowing up, $\hat{u}$ must also tend to zero and we recover k-support norm regularization. Similarly, as a tends to b, $\hat{z}$ vanishes and we have a simple ridge regression problem.

3.1 The Spectral k-Support Norm and the Spectral Box-Norm

We now turn our focus to the matrix norms. For this purpose, we recall that a norm $\|\cdot\|$ on $\mathbb{R}^{d \times m}$ is called orthogonally invariant if $\|W\| = \|UWV\|$, for any orthogonal matrices $U \in \mathbb{R}^{d \times d}$ and $V \in \mathbb{R}^{m \times m}$. A classical result by von Neumann [10] establishes that a norm is orthogonally invariant if and only if it is of the form $\|W\| = g(\sigma(W))$, where $\sigma(W)$ is the vector formed by the singular values of W in nonincreasing order, and g is a symmetric gauge function, that is a norm which is invariant under permutations and sign changes of the vector components.

Lemma 3.4. *If Θ is a convex bounded subset of the strictly positive orthant in $\mathbb{R}^d$ which is invariant under permutations, then $\|\cdot\|_\Theta$ is a symmetric gauge function.*

In particular, this readily applies to both the k-support norm and box-norm. We can therefore extend both norms to orthogonally invariant norms, which we term the spectral k-support norm and the spectral box-norm respectively, and which we write (with some abuse of notation) as $\|W\|_{(k)} = \|\sigma(W)\|_{(k)}$ and $\|W\|_\Theta = \|\sigma(W)\|_\Theta$. We note that since the k-support norm subsumes the ℓ_1 and ℓ_2-norms for $k = 1$ and $k = d$ respectively, the corresponding spectral k-support norms are equal to the trace and Frobenius norms respectively. We first characterize the unit ball of the spectral k-support norm.

Proposition 3.5. *The unit ball of the spectral k-support norm is the convex hull of the set of matrices of rank at most k and Frobenius norm no greater than one.*

Referring to the unit ball characterization of the k-support norm, we note that the restriction on the cardinality of the vectors whose convex hull defines the unit ball naturally extends to a restriction on the rank operator in the matrix setting. Furthermore, as noted in [6], regularization using the k-support norm encourages vectors to be sparse, but less so that the ℓ_1-norm. In matrix problems, as the extreme points of the unit ball have rank k, Proposition 3.5 suggests that the spectral k-support norm for $k > 1$ should encourage matrices to have low rank, but less so than the trace norm.

3.2 Cluster Norm

We end this section by briefly discussing the cluster norm, which was introduced in [9] as a convex relaxation of a multitask clustering problem. The norm is defined, for every $W \in \mathbb{R}^{d\times m}$, as

$$\|W\|_{\mathrm{cl}} = \sqrt{\inf_{S\in\mathcal{S}_m} \operatorname{tr}(S^{-1}W^\top W)} \tag{7}$$

where $\mathcal{S}_m = \{S \in \mathbb{R}^{m\times m}, S \succeq 0 : aI \preceq S \preceq bI,\ \operatorname{tr} S = c\}$, and $0 < a \leq b$. In [9] the authors state that the cluster norm of W equals the box-norm of the vector formed by the singular values of W where $c = (b-a)k + da$. Here we provide a proof of this result. Denote by $\lambda_i(\cdot)$ the eigenvalues of a matrix which we write in nonincreasing order $\lambda_1(\cdot) \geq \lambda_2(\cdot) \geq \cdots \geq \lambda_d(\cdot)$. Note that if θ_i are the eigenvalues of S then $\theta_i = \lambda_{d-i+1}(S^{-1})$. We have that

$$\operatorname{tr}(S^{-1}W^\top W) = \operatorname{tr}(S^{-1}U\Sigma^2U^\top) \geq \sum_{i=1}^m \lambda_{d-i+1}(S^{-1})\lambda_i(W^\top W) = \sum_{i=1}^d \frac{\sigma_i^2(W)}{\theta_i}$$

where we have used the inequality [20, Sec. H.1.h] for S^{-1}, $W^\top W \succeq 0$. Since this inequality is attained whenever $S = U\mathrm{Diag}(\theta)U$, where U are the eigenvectors of $W^\top W$, we see that $\|W\|_{\mathrm{cl}} = \|\sigma(W)\|_\Theta$, that is, the cluster norm coincides with the spectral box-norm. In particular, we see that the spectral k-support norm is a special case of the cluster norm, where we let a tend to zero, $b = 1$ and $c = k$. Moreover, the methods to compute the norm and its proximity operator described in the following section can directly be applied to the cluster norm.

As in the case of the vector norm (Proposition 3.3), the spectral box-norm or cluster norm can be written as a perturbation of spectral k-support norm with a quadratic term.

Proposition 3.6. *Let $\|\cdot\|_\Theta$ be a matrix box-norm with parameters a, b, c and let $k = \frac{c-da}{b-a}$. Then*

$$\|W\|_\Theta^2 = \min_Z \frac{1}{a}\|W - Z\|_F^2 + \frac{1}{b-a}\|Z\|_{(k)}^2.$$

In other words, this result shows that the cluster norm can be seen as the Moreau envelope [11] of a spectral k-support norm.

4 Computing the Norms and their Proximity Operator

In this section, we compute the norm and the proximity operator of the squared norm by explicitly solving the optimization problem in (3). We begin with the vector norm.

Theorem 4.1. *For every $w \in \mathbb{R}^d$ it holds that*

$$\|w\|_\Theta^2 = \frac{1}{b}\|w_Q\|_2^2 + \frac{1}{p}\|w_I\|_1^2 + \frac{1}{a}\|w_L\|_2^2, \tag{8}$$

where $w_Q = (|w|_1^\downarrow, \ldots, |w|_q^\downarrow)$, $w_I = (|w|_{q+1}^\downarrow, \ldots, |w|_{d-\ell}^\downarrow)$, $w_L = (|w|_{d-\ell+1}^\downarrow, \ldots, |w|_d^\downarrow)$, and q and ℓ are the unique integers in $\{0, \ldots, d\}$ that satisfy $q + \ell \leq d$,

$$\frac{|w_q|}{b} \geq \frac{1}{p}\sum_{i=q+1}^{d-\ell} |w_i| > \frac{|w_{q+1}|}{b}, \quad \frac{|w_{d-\ell}|}{a} \geq \frac{1}{p}\sum_{i=q+1}^{d-\ell} |w_i| > \frac{|w_{d-\ell+1}|}{a}, \tag{9}$$

$p = c - qb - \ell a$ and we have defined $|w_0| = \infty$ and $|w_{d+1}| = 0$.

Proof. (Sketch) We need to solve the optimization problem

$$\inf_\theta \left\{ \sum_{i=1}^d \frac{w_i^2}{\theta_i} : a \leq \theta_i \leq b, \sum_{i=1}^d \theta_i \leq c \right\}. \tag{10}$$

We assume without loss of generality that the w_i are ordered nonincreasing in absolute values, and it follows that at the optimum the θ_i are also ordered nonincreasing. We further assume that $w_i \neq 0$ for all i and $c \leq db$, so the sum constraint will be tight at the optimum. The Lagrangian is given by

$$L(\theta, \alpha) = \sum_{i=1}^d \frac{w_i^2}{\theta_i} + \frac{1}{\alpha^2}\left(\sum_{i=1}^d \theta_i - c\right)$$

where $1/\alpha^2$ is a strictly positive multiplier to be chosen such that $S(\alpha) := \sum_{i=1}^d \theta_i(\alpha) = c$. We can then solve the original problem by minimizing the Lagrangian over the constraint $\theta \in [a, b]^d$. Due to the decoupling effect of the multiplier we can solve the simplified problem componentwise, obtaining the solution

$$\theta_i = \theta_i(\alpha) = \min(b, \max(a, \alpha|w_i|)) \tag{11}$$

where $S(\alpha) = c$. The minimizer has the form $\theta = (b, \ldots, b, \theta_{q+1}, \ldots, \theta_{d-\ell}, a, \ldots, a)$, where q, ℓ are determined by the value of α. From $S(\alpha) = c$ we get $\alpha = p/(\sum_{i=q+1}^{d-\ell} |w_i|)$. The value of the norm in (8) follows by substituting θ into the objective. Finally, by construction we have $\theta_q \geq b > \theta_{q+1}$ and $\theta_{d-\ell} > a \geq \theta_{d-\ell+1}$, which give rise to the conditions in (9). ∎

Theorem 4.1 suggests two methods for computing the box-norm. First we find α such that $S(\alpha) = c$; this value uniquely determines θ in (11), and the norm follows by substitution into (10). Alternatively, we identify q and ℓ that jointly satisfy (9) and we compute the norm using (8). Taking advantage of the structure of θ in the former method leads to a computation time that is $\mathcal{O}(d \log d)$.

Theorem 4.2. *The computation of the box-norm can be completed in $\mathcal{O}(d \log d)$ time.*

The k-support norm is a special case of the box-norm, and as a direct corollary of Theorem 4.1 and Theorem 4.2, we recover [6, Proposition 2.1].

4.1 Proximity Operator

Proximal gradient methods can be used to solve optimization problems of the form $\min_w f(w) + \lambda g(w)$, where f is a convex loss function with Lipschitz continuous gradient, $\lambda > 0$ is a regularization parameter, and g is a convex function for which the proximity operator can be computed efficiently, see [12, 21, 22] and references therein. The proximity operator of g with parameter $\rho > 0$ is defined as

$$\text{prox}_{\rho g}(w) = \text{argmin}\left\{ \frac{1}{2}\|x - w\|^2 + \rho g(x) : x \in \mathbb{R}^d \right\}.$$

We now use the infimum formulation of the box-norm to derive the proximity operator of the squared norm.

Algorithm 1 Computation of $x = \operatorname{prox}_{\frac{\lambda}{2}\|\cdot\|^2_\Theta}(w)$.

Require: parameters a, b, c, λ.

1. Sort points $\left\{\alpha^i\right\}_{i=1}^{2d} = \left\{\frac{a+\lambda}{|w_j|}, \frac{b+\lambda}{|w_j|}\right\}_{j=1}^{d}$ such that $\alpha^i \leq \alpha^{i+1}$;
2. Identify points α^i and α^{i+1} such that $S(\alpha^i) \leq c$ and $S(\alpha^{i+1}) \geq c$ by binary search;
3. Find α^* between α^i and α^{i+1} such that $S(\alpha^*) = c$ by linear interpolation;
4. Compute $\theta_i(\alpha^*)$ for $i = 1, \ldots, d$;
5. Return $x_i = \frac{\theta_i w_i}{\theta_i + \lambda}$ for $i = 1, \ldots, d$.

Theorem 4.3. *The proximity operator of the square of the box-norm at point $w \in \mathbb{R}^d$ with parameter $\frac{\lambda}{2}$ is given by $\operatorname{prox}_{\frac{\lambda}{2}\|\cdot\|^2_\Theta}(w) = (\frac{\theta_1 w_1}{\theta_1+\lambda}, \ldots, \frac{\theta_d w_d}{\theta_d+\lambda})$, where*

$$\theta_i = \theta_i(\alpha) = \min(b, \max(a, \alpha|w_i| - \lambda)) \tag{12}$$

and α is chosen such that $S(\alpha) := \sum_{i=1}^d \theta_i(\alpha) = c$. Furthermore, the computation of the proximity operator can be completed in $\mathcal{O}(d \log d)$ time.

The proof follows a similar reasoning to the proof of Theorem 4.1. Algorithm 1 illustrates the computation of the proximity operator for the squared box-norm in $\mathcal{O}(d \log d)$ time. This includes the k-support as a special case, where we let a tend to zero, and set $b = 1$ and $c = k$, which improves upon the complexity of the $\mathcal{O}(d(k + \log d))$ computation provided in [6], and we illustrate the improvement empirically in Table 1.

4.2 Proximity Operator for Orthogonally Invariant Norms

The computational considerations outlined above can be naturally extended to the matrix setting by using von Neumann's trace inequality (see, e.g. [23]). Here we comment on the computation of the proximity operator, which is important for our numerical experiments in the following section. The proximity operator of an orthogonally invariant norm $\|\cdot\| = g(\sigma(\cdot))$ is given by

$$\operatorname{prox}_{\|\cdot\|}(W) = U \operatorname{diag}(\operatorname{prox}_g(\sigma(W)))V^\top, \quad W \in \mathbb{R}^{m \times d},$$

where U and V are the matrices formed by the left and right singular vectors of W (see e.g. [24, Prop 3.1]). Using this result we can employ proximal gradient methods to solve matrix regularization problems using the squared spectral k-support norm and spectral box-norm.

5 Numerical Experiments

In this section, we report on the statistical performance of the spectral regularizers in matrix completion experiments. We also offer an interpretation of the role of the parameters in the box-norm and we empirically verify the improved performance of the proximity operator computation (see Table 1). We compare the trace norm (*tr*) [25], matrix elastic net (*en*) [26], spectral k-support (*ks*) and the spectral box-norm (*box*). The Frobenius norm, which is equal to the spectral k-support norm for $k = d$, performed considerably worse than the trace norm and we omit the results here. We report test error and standard deviation, matrix rank (r) and optimal parameter values for k and a, which were determined by validation, as were the regularization parameters. When comparing performance, we used a *t-test* to determine statistical significance at a level of $p < 0.001$. For the optimization we used an accelerated proximal gradient method (FISTA), see e.g. [12, 21, 22], with the percentage change in objective as convergence criterion, with a tolerance of 10^{-5} for the simulated datasets and 10^{-3} for the real datasets. As is typical with spectral regularizers we found that the spectrum of the learned matrix exhibited a rapid decay to zero. In order to explicitly impose a low rank on the solution we included a final step where we hard-threshold the singular values of the final matrix below a level determined by validation. We report on both sets of results below.

5.1 Simulated Data

Matrix Completion. We applied the norms to matrix completion on noisy observations of low rank matrices. Each $m \times m$ matrix was generated as $W = AB^\top + E$, where $A, B \in \mathbb{R}^{m \times r}$, $r \ll m$, and

Table 1: Comparison of proximity operator algorithms for the k-support norm (time in s), $k = 0.05d$. Algorithm 1 is the method in [6], Algorithm 2 is our method.

d	1,000	2,000	4,000	8,000	16,000	32,000
Alg. 1	0.0443	0.1567	0.5907	2.3065	9.0080	35.6199
Alg. 2	0.0011	0.0016	0.0026	0.0046	0.0101	0.0181

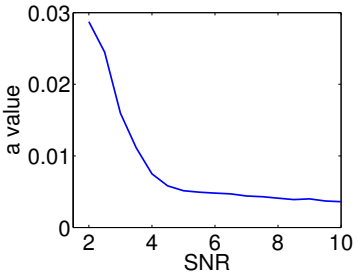

Figure 1: Impact of signal to noise on a.

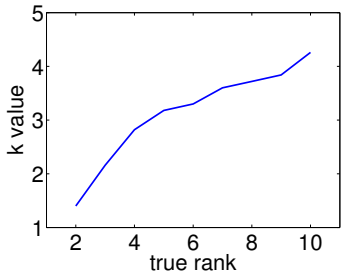

Figure 2: Impact of matrix rank on k.

the entries of A, B and E are i.i.d. standard Gaussian. We set $m = 100$, $r \in \{5, 10\}$ and we sampled uniformly a percentage $\rho \in \{10\%, 20\%, 30\%\}$ of the entries for training, and used a fixed 10% for validation. The error was measured as $\|\text{true} - \text{predicted}\|^2 / \|\text{true}\|^2$ [5] and averaged over 100 trials. The results are summarized in Table 2. In the thresholding case, all methods recovered the rank of the true noiseless matrix. The spectral box-norm generated the lowest test errors in all regimes, with the spectral k-support a close second, in particular in the thresholding case. This suggests that the non zero parameter a in the spectral box-norm counteracted the noise to some extent.

Role of Parameters. In the same setting we investigated the role of the parameters in the box-norm. As previously discussed, parameter b can be set to 1 without loss of generality. Figure 1 shows the optimal value of a chosen by validation for varying signal to noise ratios (SNR), keeping k fixed. We see that for greater noise levels (smaller SNR), the optimal value for a increases. While for $a > 0$, the recovered solutions are not sparse, as we show below this can still lead to improved performance in experiments, in particular in the presence of noise. Figure 2 shows the optimal value of k chosen by validation for matrices with increasing rank, keeping a fixed. We notice that as the rank of the matrix increases, the optimal k value increases, which is expected since it is an upper bound on the sum of the singular values.

Table 2: Matrix completion on simulated data sets, without (left) and with (right) thresholding.

dataset	norm	test error	r	k	a
rank 5	tr	0.8184 (0.03)	20	-	-
ρ=10%	en	0.8164 (0.03)	20	-	-
	ks	0.8036 (0.03)	16	3.6	-
	box	0.7805 (0.03)	87	2.9	1.7e-2
rank 5	tr	0.4085 (0.03)	23	-	-
ρ=20%	en	0.4081 (0.03)	23	-	-
	ks	0.4031 (0.03)	21	3.1	-
	box	0.3898 (0.03)	100	1.3	9e-3
rank 10	tr	0.6356 (0.03)	27	-	-
ρ=20%	en	0.6359 (0.03)	27	-	-
	ks	0.6284 (0.03)	24	4.4	-
	box	0.6243 (0.03)	89	1.8	9e-3
rank 10	tr	0.3642 (0.02)	36	-	-
ρ=30%	en	0.3638 (0.002	36	-	-
	ks	0.3579 (0.02)	33	5.0	-
	box	0.3486 (0.02)	100	2.5	9e-3

dataset	norm	test error	r	k	a
rank 5	tr	0.7799 (0.04)	5	-	-
ρ=10%	en	0.7794 (0.04)	5	-	-
	ks	0.7728 (0.04)	5	4.23	-
	box	0.7649 (0.04)	5	3.63	8.1e-3
rank 5	tr	0.3449 (0.02)	5	-	-
ρ=20%	en	0.3445 (0.02)	5	-	-
	ks	0.3381 (0.02)	5	2.97	-
	box	0.3380 (0.02)	5	3.28	1.9e-3
rank 10	tr	0.6084 (0.03)	10	-	-
ρ=20%	en	0.6074 (0.03)	10	-	-
	ks	0.6000 (0.03)	10	5.02	-
	box	0.6000 (0.03)	10	5.22	1.9e-3
rank 10	tr	0.3086 (0.02)	10	-	-
ρ=30%	en	0.3082 (0.02)	10	-	-
	ks	0.3025 (0.02)	10	5.13	-
	box	0.3025 (0.02)	10	5.16	3e-4

Table 3: Matrix completion on real data sets, without (left) and with (right) thresholding.

dataset	norm	test error	r	k	a
MovieLens 100k $\rho = 50\%$	tr	0.2034	87	-	-
	en	0.2034	87	-	-
	ks	0.2031	102	1.00	-
	box	0.2035	943	1.00	1e-5
MovieLens 1M $\rho = 50\%$	tr	0.1821	325	-	-
	en	0.1821	319	-	-
	ks	0.1820	317	1.00	-
	box	0.1817	3576	1.09	3e-5
Jester 1 20 per line	tr	0.1787	98	-	-
	en	0.1787	98	-	-
	ks	0.1764	84	5.00	-
	box	0.1766	100	4.00	1e-6
Jester 3 8 per line	tr	0.1988	49	-	-
	en	0.1988	49	-	-
	ks	0.1970	46	3.70	-
	box	0.1973	100	5.91	1e-3

dataset	norm	test error	r	k	a
MovieLens 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
MovieLens 1M $\rho = 50\%$	tr	0.1790	17	-	-
	en	0.1789	17	-	-
	ks	0.1782	17	1.80	-
	box	0.1777	19	2.00	1e-6
Jester 1 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
Jester 3 8 per line	tr	0.1959	3	-	-
	en	0.1959	3	-	-
	ks	0.1942	3	2.13	-
	box	0.1940	3	4.00	8e-4

5.2 Real Data

Matrix Completion (MovieLens and Jester). In this section we report on matrix completion on real data sets. We observe a percentage of the (user, rating) entries of a matrix and the task is to predict the unobserved ratings, with the assumption that the true matrix has low rank. The datasets we considered were MovieLens 100k and MovieLens 1M (*http://grouplens.org/datasets/movielens/*), which consist of user ratings of movies, and Jester 1 and Jester 3 (*http://goldberg.berkeley.edu/jester-data/*), which consist of users and ratings of jokes (Jester 2 showed essentially identical performance to Jester 1). Following [4], for MovieLens we uniformly sampled $\rho = 50\%$ of the available entries for each user for training, and for Jester 1 and Jester 3 we sampled 20, respectively 8, ratings per user, and we used 10% for validation. The error was measured as normalized mean absolute error, $\frac{\|\text{true}-\text{predicted}\|^2}{\#\text{observations}/(r_{\max}-r_{\min})}$, where $r_{\min}$ and $r_{\max}$ are lower and upper bounds for the ratings [4]. The results are outlined in Table 3. In the thresholding case, the spectral box and k-support norms had the best performance. In the absence of thresholding, the spectral k-support showed slightly better performance. Comparing to the synthetic data sets, this suggests that in the absence of noise the parameter a did not provide any benefit. We note that in the absence of thresholding our results for the trace norm on MovieLens 100k agreed with those in [3].

6 Conclusion

We showed that the k-support norm belongs to the family of box-norms and noted that these can be naturally extended from the vector to the matrix setting. We also provided a connection between the k-support norm and the cluster norm, which essentially coincides with the spectral box-norm. We further observed that the cluster norm is a perturbation of the spectral k-support norm, and we were able to compute the norm and its proximity operator. Our experiments indicate that the spectral box-norm and k-support norm consistently outperform the trace norm and the matrix elastic net on various matrix completion problems. With a single parameter to validate, compared to two for the spectral box-norm, our results suggest that the spectral k-support norm is a powerful alternative to the trace norm and the elastic net, which has the same number of parameters. In future work, we would like to study the application of the norms to clustering problems in multitask learning [9], in particular the impact of centering. It would also be valuable to derive statistical inequalities and Rademacher complexities for these norms.

Acknowledgements

We would like to thank Andreas Maurer, Charles Micchelli and especially Andreas Argyriou for useful discussions. Part of this work was supported by EPSRC Grant EP/H027203/1.

References

[1] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems 17*, 2005.

[2] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, Vol. 10:803–826, 2009.

[3] M Jaggi and M. Sulovsky. A simple algorithm for nuclear norm regularized problems. *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[4] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *SIAM Journal on Imaging Sciences*, 4:573–596, 2011.

[5] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

[6] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems 25*, pages 1466–1474, 2012.

[7] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Vol. 58:267–288, 1996.

[8] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.

[9] L. Jacob, F. Bach, and J.-P. Vert. Clustered multi-task learning: a convex formulation. *Advances in Neural Information Processing Systems (NIPS 21)*, 2009.

[10] J. Von Neumann. *Some matrix-inequalities and metrization of matric-space*. Tomsk. Univ. Rev. Vol I, 1937.

[11] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[12] Y. Nesterov. Gradient methods for minimizing composite objective function. *Center for Operations Research and Econometrics*, 76, 2007.

[13] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. *Proc of the 26th Int. Conf. on Machine Learning*, 2009.

[14] Y. Grandvalet. Least absolute shrinkage is equivalent to quadratic penalization. In *ICANN 98*, pages 201–206. Springer London, 1998.

[15] C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

[16] M. Szafranski, Y. Grandvalet, and P. Morizet-Mahoudeaux. Hierarchical penalization. In *Advances in Neural Information Processing Systems 21*, 2007.

[17] C. A. Micchelli, J. M. Morales, and M. Pontil. Regularizers for structured sparsity. *Advances in Comp. Mathematics*, 38:455–489, 2013.

[18] A. Maurer and M. Pontil. Structured sparsity and generalization. *The Journal of Machine Learning Research*, 13:671–690, 2012.

[19] G. Obozinski and F. Bach. Convex relaxation for combinatorial penalties. *CoRR*, 2012.

[20] A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and its Applications*. Academic Press, 1979.

[21] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inv Prob*. Springer, 2011.

[22] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

[23] A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2:173–183, 1995.

[24] A. Argyriou, C. A. Micchelli, M. Pontil, L. Shen, and Y. Xu. Efficient first order methods for linear composite regularizers. *CoRR*, abs/1104.1436, 2011.

[25] J.-F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2008.

[26] H. Li, N. Chen, and L. Li. Error analysis for matrix elastic-net regularization algorithms. *IEEE Transactions on Neural Networks and Learning Systems*, 23-5:737–748, 2012.

[27] W. Rudin. *Functional Analysis*. McGraw Hill, 1991.

[28] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[29] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

Supplementary Material

In this appendix, we collect some auxiliary results and we provide proofs of the results stated in the main body of the paper.

A Auxiliary Results

Recall that a subset A of a real vector space X is called *balanced* if $\alpha A \subset A$ whenever $|\alpha| \leq 1$. Furthermore, A is called *absorbing* if for any $x \in X$, $x \in \lambda A$ for some $\lambda(x) > 0$. For a proof of the following lemma see e.g. [27, §1.35].

Lemma A.1. *Let X be a real vector space and let $A \subset X$ be a convex, balanced, and absorbing set. The Minkowski functional μ_A of A, given, for every $x \in X$, by the formula*

$$\mu_A(x) = \inf\{\lambda > 0 : x \in \lambda A\}$$

defines a seminorm on X. In addition, if $\mu_A(x) > 0$ for every $x \neq 0$, then μ_A defines a norm on X.

The next result is due to von Neumann [10], see also [23].

Theorem A.2 (Von Neumann's trace inequality)**.** *For any $d \times m$ matrices X and Y,*

$$\operatorname{tr}(XY^\top) \leq \langle \sigma(X), \sigma(Y) \rangle.$$

Equality holds if and only if X and Y admit a simultaneous singular value decomposition, that is

$$X = U \operatorname{diag}(\sigma(X)) V^\top, \qquad Y = U \operatorname{diag}(\sigma(Y)) V^\top,$$

where $U \in \mathbb{R}^{d \times d}$ and $V \in \mathbb{R}^{m \times m}$ are orthogonal matrices.

The following result, which is presented in [6, Section 2] is key for the proof of Theorem 3.5.

Proposition A.3. *The unit ball of the vector k-support norm is equal to the convex hull of the set $\{w \in \mathbb{R}^d : \operatorname{card}(w) \leq k, \|w\|_2 \leq 1\}$.*

Theorems 4.1 and 4.3 make use of the following result, which follows from [17], Theorem 3.1.

Lemma A.4. *Let $w \in \mathbb{R}, \beta > 0$, and define $g(\theta) = \frac{w^2}{\theta} + \beta^2 \theta (\theta > 0)$. For $0 < a \leq b$, the unique solution to the problem $\min\{g(\theta) : a \leq \theta \leq b\}$ is given by*

$$\theta = \begin{cases} a, & \text{if } \frac{|w|}{\beta} < a, \\ \frac{|w|}{\beta}, & \text{if } a \leq \frac{|w|}{\beta} \leq b, \\ b, & \text{if } \frac{|w|}{\beta} > b. \end{cases}$$

Proof. For fixed w, the objective function is strictly convex on $\mathbb{R}^d_{++}$ and has a unique minimum on $(0, \infty)$ (see Figure 1.b in [17] for a one-dimensional illustration). The derivative of the objective function is zero for $\theta = \theta^* := |w|/\beta$, strictly positive below θ^* and strictly increasing above θ^*. Considering these three cases we recover the expression in statement of the lemma. ∎

B Proofs

Proof of Proposition 2.2. Consider the expression for the dual norm. The function $\|\cdot\|_\Theta$ is a norm since it is a supremum of norms. Recall that the Fenchel conjugate h^* of a function $h : \mathbb{R}^d \to \mathbb{R}$ is defined for every $u \in \mathbb{R}^d$ as $h^*(u) = \sup\left\{\langle u, w \rangle - h(w) : w \in \mathbb{R}^d\right\}$. It is a standard result from convex analysis that for any norm $\|\cdot\|$, the Fenchel conjugate of the function $h := \frac{1}{2}\|\cdot\|^2$ satisfies

$h^* = \frac{1}{2}\|\cdot\|_*^2$, where $\|\cdot\|_*$ is the corresponding dual norm (see, e.g. [23]). By the same result, for any norm the biconjugate is equal to the norm, that is $(\|\cdot\|^*)^* = \|\cdot\|$. Applying this to the dual norm we have, for every $w \in \mathbb{R}^d$,

$$h(w) = \sup_{u\in\mathbb{R}^d} \{\langle w, u\rangle - h^*(u)\} = \sup_{u\in\mathbb{R}^d} \inf_{\theta\in\Theta} \left\{ \sum_{i=1}^d \left(w_i u_i - \frac{1}{2}\theta_i u_i^2 \right) \right\}.$$

This is a minimax problem in the sense of von Neumann [28], and we can exchange the order of the $\inf$ and the $\sup$, and solve the latter (which is in fact a maximum) componentwise. The gradient with respect to u_i is zero for $u_i = \frac{w_i}{\theta_i}$, and substituting this into the objective we get

$$h(w) = \frac{1}{2} \inf_{\theta\in\Theta} \sum_{i=1}^d \frac{w_i^2}{\theta_i}.$$

It follows that the infimum expression in (3) defines a norm, and the two norms are duals of each other as required. ∎

Proof of Proposition 3.1. We make the change of variable $\phi_i = \frac{\theta_i - a}{b-a}$ and observe that the constraints on θ induce the constraint set $\{\phi \in (0,1]^d,\ \sum_{i=1}^d \phi_i \le \rho\}$, where $\rho = \frac{c-da}{b-a}$. Furthermore

$$\sum_{i=1}^d \theta_i u_i^2 = a\|u\|_2^2 + (b-a)\sum_{i=1}^d \phi_i u_i^2.$$

The result then follows by taking the supremum over ϕ. ∎

Proof of Proposition 3.2. Equation 5 defines a norm and we will show that its norm coincides with the dual of the Θ-norm given by equation (4). To simplify the exposition we define the norm

$$\|v\|_g^2 = \sum_{i\in g} \frac{v_i^2}{b} + \sum_{i\notin g} \frac{v_i^2}{a}, \quad v \in \mathbb{R}^d,$$

whose corresponding dual norm is

$$\|u\|_{*,g}^2 = b\sum_{i\in g} u_i^2 + a\sum_{i\notin g} u_i^2, \quad u \in \mathbb{R}^d.$$

Furthermore for every $u \in \mathbb{R}^d$ and $g \subseteq \{1,\dots,d\}$, we define the vectors $u_{|g} = (u_i I_{\{i\in g\}})_{i=1}^d$ and $u_{|g^c} = (u_i I_{\{i\notin g\}})_{i=1}^d$.

We have, for every $u \in \mathbb{R}^d$, $u \neq 0$, that

$$\begin{aligned}
\sup_{w\in\mathbb{R}^d} \frac{\langle w, u\rangle}{\|w\|} &= \sup_{\{v_g\}} \frac{\sum_{g\in\mathcal{G}_k} \langle v_g, u\rangle}{\sum_{g\in\mathcal{G}_k} \|v_g\|_g} \\
&\le \sup_{\{v_g\}} \frac{\sum_{g\in\mathcal{G}_k} \|v_g\|_g \|u\|_{*,g}}{\sum_{g\in\mathcal{G}_k} \|v_g\|_g} \\
&\le \max_{g\in\mathcal{G}_k} \|u\|_{*,g}, \qquad (13)
\end{aligned}$$

where we have used Cauchy-Schwarz and Hölder inequalities. We can make the first inequality tight by setting $v_g = \lambda_g(bu_{|g} + au_{|g^c})$ and the second inequality tight by requiring $\lambda_g = 0$ whenever $g \notin \operatorname{argmax}_{g'\in\mathcal{G}_k} \|u\|_{*,g'}$, see e.g. [29, Sects. 5.4.14. and 5.4.15]. Note that the right hand side in (13) is maximized when $g = \{i_1,\dots,i_k\}$ such that $|u_{i_1}| \ge \cdots |u_{i_k}|$ and the expression coincides with (4) for $\rho = k$. ∎

Proof of Proposition 3.3. Consider the definition of the norm $\|w\|_\Theta$ in (3). We make the change of variables $\phi_i = \frac{\theta_i - a}{b-a}$, and write

$$\|w\|_\Theta^2 = \min_{\theta \in \Theta} \sum_{i=1}^d \frac{w_i^2}{\theta_i} = \frac{\gamma}{a} \min_{\phi \in \Phi} \sum_{i=1}^d \frac{w_i^2}{\phi_i + \gamma}, \tag{14}$$

where we have defined $\gamma = \frac{a}{b-a}$ and $\Phi = \{\phi \in (0,1]^d : \sum_{i=1}^d \phi_i \leq k\}$. We observe that

$$\min_{z \in \mathbb{R}^d} \left\{ \|w - z\|_2^2 + \gamma \|z\|_\Phi^2 \right\} = \min_{z \in \mathbb{R}^d} \min_{\phi \in \Phi} \left\{ \sum_{i=1}^d (w_i - z_i)^2 + \gamma \frac{z_i^2}{\phi_i} \right\} = \gamma \min_{\phi \in \Phi} \sum_{i=1}^d \frac{w_i^2}{\phi_i + \gamma}, \tag{15}$$

where we have interchanged the order of the minimization problems and solved for z_i componentwise. The result follows by combining equations (14) and (15). ∎

Proof of Lemma 3.4. Let $g(w) = \|w\|_\Theta$. We need to show that g is a norm which is invariant under permutations and sign changes. By Proposition 2.2, g is a norm, so it remains to show that $g(w_1, ..., w_d) = g(w_{\pi(1)}, \ldots, w_{\pi(d)})$ for every permutation π, and $g(Jw) = g(w)$ for every diagonal matrix J with entries ± 1. The latter property is immediate. The former property follows since the set Θ-norm is permutation invariant. ∎

Proof of Proposition 3.5. For any $W \in \mathbb{R}^{d \times m}$, define the following sets

$$T_k = \{W \in \mathbb{R}^{d \times m} : \text{rank}(W) \leq k, \|W\|_F \leq 1\}, \quad A_k = \text{co}(T_k),$$

and consider the following functional

$$\lambda(W) = \inf\{\lambda > 0 : W \in \lambda A_k\}, \quad W \in \mathbb{R}^{d \times m}. \tag{16}$$

By Lemma A.1, λ defines a norm on $\mathbb{R}^{d \times m}$ with unit ball equal to A_k. Since the constraints in T_k involve spectral functions, the sets T_k and A_k are invariant to left and right multiplication by orthogonal matrices. It follows that λ is a spectral function, that is $\lambda(W)$ is defined in terms of the singular values of W, and by von Neumann's Theorem [10] the norm it defines is orthogonally invariant and we have

$$\begin{aligned} \lambda(W) &= \inf\{\lambda > 0 : W \in \lambda A_k\} \\ &= \inf\{\lambda > 0 : \sigma(W) \in \lambda C_k\} \\ &= \|\sigma(W)\|_{(k)}, \end{aligned}$$

where we have defined the set $C_k = \text{co}\{w \in \mathbb{R}^d : \|w\|_2 \leq 1, \ \text{card}(w) \leq k\}$ and we have used the fact that the unit ball of the k-support norm is the convex hull of C_k [6, Section 2] in the penultimate step. It follows that the norm defined by (16) is the spectral k-support norm.

∎

Proof of Proposition 3.6. By von Neumann's trace inequality (Theorem A.2) we have

$$\begin{aligned} \frac{1}{a}\|W - Z\|_F^2 + \frac{1}{b-a}\|Z\|_{(k)}^2 &= \frac{1}{a}\left(\|W\|_F^2 + \|Z\|_F^2 - 2\langle W, Z\rangle\right) + \frac{1}{b-a}\|Z\|_{(k)}^2 \\ &\geq \frac{1}{a}\left(\|\sigma(W)\|_2^2 + \|\sigma(Z)\|_2^2 - 2\langle \sigma(W), \sigma(Z)\rangle\right) + \frac{1}{b-a}\|\sigma(Z)\|_{(k)}^2 \\ &= \frac{1}{a}\|\sigma(W) - \sigma(Z)\|_2^2 + \frac{1}{b-a}\|\sigma(Z)\|_{(k)}^2. \end{aligned}$$

Furthermore the inequality is tight if W and Z have the same ordered set of singular vectors. Hence

$$\min_{Z\in\mathbb{R}^{d\times m}}\left\{\frac{1}{a}\|W-Z\|_F^2+\frac{1}{b-a}\|Z\|_{(k)}^2\right\}=\min_{z\in\mathbb{R}^d}\left\{\frac{1}{a}\|\sigma(W)-z\|_2^2+\frac{1}{b-a}\|z\|_{(k)}^2\right\}=\|\sigma(W)\|_{(k)}^2,$$

where the last equality follows by Proposition 3.3 ∎

Proof of Theorem 4.1. We solve the constrained optimization problem

$$\inf\left\{\sum_{i=1}^d\frac{w_i^2}{\theta_i}:a\le\theta_i\le b,\sum_{i=1}^d\theta_i\le c\right\}. \tag{17}$$

To simplify notation we assume without loss of generality that w_i are positive and ordered nonincreasing, and note that the optimal θ_i are ordered nonincreasing. To see this, let $\theta^*=\operatorname{argmin}_{\theta\in\Theta}\sum_{i=1}^d\frac{w_i^2}{\theta_i}$. Now suppose that $\theta_i^*<\theta_j^*$ for some $i<j$ and define $\hat{\theta}$ to be identical to θ^*, except with the i and j elements exchanged. The difference in objective values is

$$\sum_{i=1}^d\frac{w_i^2}{\hat{\theta}_i}-\sum_{i=1}^d\frac{w_i^2}{\theta_i^*}=(w_i^2-w_j^2)\left(\frac{1}{\theta_j^*}-\frac{1}{\theta_i^*}\right),$$

which is negative so θ^* cannot be a minimizer.

We further assume without loss of generality that $w_i\neq 0$ for all i, and $c\le db$ (see Remark B.1 below). The objective is continuous and we take the infimum over a closed bounded set, so a solution exists, the solution is a minimum, and it is unique by strict convexity. Furthermore, since $c\le db$, the sum constraint will be tight at the optimum.

Consider the Lagrangian function

$$L(\theta,\alpha)=\sum_{i=1}^d\frac{w_i^2}{\theta_i}+\frac{1}{\alpha^2}\left(\sum_{i=1}^d\theta_i-c\right), \tag{18}$$

where $1/\alpha^2$ is a strictly positive multiplier, and α is to be chosen to make the sum constraint tight, call this value α^*. Let θ^* be the minimizer of $L(\theta,\alpha^*)$ over θ subject to $a\le\theta_i\le b$.

We claim that θ^* solves equation (17). Indeed, for any $\theta\in[a,b]^d$, $L(\theta^*,\alpha^*)\le L(\theta,\alpha^*)$, which implies that

$$\sum_{i=1}^d\frac{w_i^2}{\theta_i^*}\le\sum_{i=1}^d\frac{w_i^2}{\theta_i}+\frac{1}{(\alpha^*)^2}\left(\sum_{i=1}^d\theta_i-c\right).$$

If in addition we impose the constraint $\sum_{i=1}^d\theta_i\le c$, the second term on the right hand side is at most zero, so we have for all such θ

$$\sum_{i=1}^d\frac{w_i^2}{\theta_i^*}\le\sum_{i=1}^d\frac{w_i^2}{\theta_i},$$

whence it follows that θ^* is the minimizer of (17).

We can therefore solve the original problem by minimizing the Lagrangian (18) over the box constraint. Due to the coupling effect of the multiplier, the problem is separable, and we can solve the simplified problem componentwise using Lemma A.4. It follows that

$$\theta_i=\begin{cases}a, & \text{if}\quad \alpha<\frac{a}{|w_i|},\\ \alpha|w_i|, & \text{if}\quad \frac{a}{|w_i|}\le\alpha\le\frac{b}{|w_i|},\\ b, & \text{if}\quad \alpha>\frac{b}{|w_i|},\end{cases}$$

where $\alpha > 0$ is such that $\sum_{i=1}^{d} \theta_i(\alpha) = c$. Note also that in the main body of the paper we use the equivalent compact notation $\theta_i = \theta_i(\alpha) = \min(b, \max(a, \alpha|w_i|))$.

The minimizer then has the form

$$\theta = (\underbrace{b, \dots, b}_{q}, \theta_{q+1}, \dots, \theta_{d-\ell}, \underbrace{a, \dots, a}_{\ell}),$$

where $q, \ell \in \{0, \dots, d\}$ are determined by the value of α which satisfies

$$S(\alpha) = \sum_{i=1}^{d} \theta_i(\alpha) = qb + \sum_{i=q+1}^{d-\ell} \alpha|w_i| + \ell a = c,$$

i.e. $\alpha = p / \left(\sum_{i=q+1}^{d-\ell} |w_i| \right)$, where $p = c - qb - \ell a$.

The value of the norm follows by substituting θ into the objective and we get

$$\begin{aligned} \|w\|_\Theta^2 &= \sum_{i=1}^{q} \frac{|w_i|^2}{b} + \frac{1}{p}\Big(\sum_{i=q+1}^{d-\ell} |w_i| \Big)^2 + \sum_{i=d-\ell+1}^{d} \frac{|w_i|^2}{a} \\ &= \frac{1}{b}\|w_Q\|_2^2 + \frac{1}{p}\|w_I\|_1^2 + \frac{1}{a}\|w_L\|_2^2, \end{aligned}$$

as required. We can further characterize q and ℓ by considering the form of θ_i. By construction we have $\theta_q \geq b > \theta_{q+1}$ and $\theta_{d-\ell} > a \geq \theta_{d-\ell+1}$, or equivalently

$$\begin{aligned} \frac{|w_q|}{b} \geq \frac{1}{p} \sum_{i=q+1}^{d-\ell} |w_i| > \frac{|w_{q+1}|}{b}, \text{ and} \\ \frac{|w_{d-\ell}|}{a} \geq \frac{1}{p} \sum_{i=q+1}^{d-\ell} |w_i| > \frac{|w_{d-\ell+1}|}{a}, \end{aligned}$$

and we are done. ∎

Remark B.1. The case where some w_i are zero follows from the case that we have considered in the theorem. If $w_i = 0$ for $n < i \leq d$, then clearly we must have $\theta_i = a$ for all such i. We then consider the n-dimensional problem of finding $(\theta_1, \dots, \theta_n)$ that minimizes $\sum_{i=1}^{n} \frac{w_i^2}{\theta_i}$, subject to $a \leq \theta_i \leq b$, and $\sum_{i=1}^{n} \theta_i \leq c'$, where $c' = c - (d-n)a$. As $c \geq da$ by assumption, we also have $c' \geq na$, so a solution exists to the n-dimensional problem. If $c' < bn$, then a solution is trivially $\theta_i = b$ for all $i = 1 \dots n$. In general, $c' \geq bn$, and we proceed as per the proof of the theorem. Finally, a vector that solves the original d-dimensional problem will be given by $(\theta_1, \dots, \theta_n, a, \dots, a)$.

Proof of Theorem 4.2. Following Theorem 4.1, we need to determine α^* to satisfy the coupling constraint $S(\alpha^*) = c$. Each component θ_i is a piecewise linear function in the form of a step function with a constant positive slope between the values $a/|w_i|$ and $b/|w_i|$. Let the set $\{\alpha^i\}_{i=1}^{2d}$ be the set of the $2d$ critical points, where the α^i are ordered nondecreasing. The function $S(\alpha)$ is a nondecreasing piecewise linear function with at most $2d$ critical points. We can find α^* by first sorting the points $\{\alpha^i\}$, finding α^i and α^{i+1} such that

$$S(\alpha^i) \leq c \leq S(\alpha^{i+1})$$

by binary search, and then interpolating α^* between the two points. Sorting takes $\mathcal{O}(d \log d)$. Computing $S(\alpha^i)$ at each step of the binary search is $\mathcal{O}(d)$, so $\mathcal{O}(d \log d)$ overall. Given α^i and α^{i+1}, interpolating α^* is $\mathcal{O}(1)$, so the algorithm overall is $\mathcal{O}(d \log d)$ as claimed. ∎

Proof of Theorem 4.3. Using the infimum formulation of the norm, we solve

$$\min_{x\in\mathbb{R}^d}\inf_{\theta\in\Theta}\left\{\frac{1}{2}\sum_{i=1}^{d}(x_i-w_i)^2+\frac{\lambda}{2}\sum_{i=1}^{d}\frac{x_i^2}{\theta_i}\right\}.$$

We can exchange the order of the optimization and solve for x first. The problem is separable and a direct computation yields that $x_i=\frac{\theta_i w_i}{\theta_i+\lambda}$. Discarding a multiplicative factor of $\lambda/2$, and noting that the infimum is a minimum, the problem in θ becomes

$$\min_{\theta}\left\{\sum_{i=1}^{d}\frac{w_i^2}{\theta_i+\lambda}:a\leq\theta_i\leq b,\sum_{i=1}^{d}\theta_i\leq c\right\}.$$

This is exactly like problem (17) after the change of variable $\theta_i'=\theta_i+\lambda$. The remaining part of the proof then follows in a similar manner to the proof of Theorem 4.1. ■