Fitting Spectral Decay with the k-Support Norm

Andrew M. McDonald[1] Massimiliano Pontil[1,2]
Dimitris Stamos[2]

(1) Department of Computer Science
University College London
email: $\{$*a.mcdonald,d.stamos.12*$\}$*@ucl.ac.uk*
Gower Street, London WC1E 6BT, UK

(2) Istituto Italiano di Tecnologia
Via Morego 30, 16163 Genova, Italy

January 5, 2016

Abstract

The spectral k-support norm enjoys good estimation properties in low rank matrix learning problems, empirically outperforming the trace norm. Its unit ball is the convex hull of rank k matrices with unit Frobenius norm. In this paper we generalize the norm to the spectral (k,p)-support norm, whose additional parameter p can be used to tailor the norm to the decay of the spectrum of the underlying model. We characterize the unit ball and we explicitly compute the norm. We further provide a conditional gradient method to solve regularization problems with the norm, and we derive an efficient algorithm to compute the Euclidean projection on the unit ball in the case $p = \infty$. In numerical experiments, we show that allowing p to vary significantly improves performance over the spectral k-support norm on various matrix completion benchmarks, and better captures the spectral decay of the underlying model.

Keywords. k-support norm, orthogonally invariant norms, matrix completion, multitask learning, proximal point algorithms.

1 Introduction

The problem of learning a sparse vector or a low rank matrix has generated much interest in recent years. A popular approach is to use convex regularizers which encourage sparsity, and a number of these have been studied with applications including image denoising, collaborative filtering and multitask learning, see for example, [Buehlmann and van der Geer 2011, Wainwright 2014] and references therein.

Recently, the *k-support norm* was proposed by [Argyriou et al. 2012], motivated as a tight relaxation of the set of k-sparse vectors of unit Euclidean norm. The authors argue that as a regularizer for sparse vector estimation, the norm empirically outperforms the Lasso [Tibshirani 1996] and Elastic Net [Zou and Hastie 2005] penalties. Statistical bounds on the Gaussian width of the k-support norm have been provided by [Chatterjee et al. 2014]. The k-support norm has also been extended to the matrix setting. By applying the norm to the vector of singular values of a matrix, [McDonald et al. 2014] obtain the orthogonally invariant *spectral k-support norm*, reporting state of the art performance on matrix completion benchmarks.

Motivated by the performance of the k-support norm in sparse vector and matrix learning problems, in this paper we study a natural generalization by considering the ℓ_p-norms (for $p \in [1, \infty]$) in place of the Euclidean norm. These allow a further degree of freedom when fitting a model to the underlying data. We denote the ensuing norm the *(k, p)-support norm*. As we demonstrate in numerical experiments, using $p = 2$ is not necessarily the best choice in all instances. By tuning the value of p the model can incorporate prior information regarding the singular values. When prior knowledge is lacking, the parameter can be chosen by validation, hence the model can adapt to a variety of decay patterns of the singular values. An interesting property of the norm is that it interpolates between the ℓ_1 norm (for $k = 1$) and the ℓ_p-norm (for $k = d$). It follows that varying both k and p the norm allows one to learn sparse vectors which exhibit different patterns of decay in the non-zero elements. In particular, when $p = \infty$ the norm prefers vectors which are constant.

A main goal of the paper is to study the proposed norm in matrix learning problems. The (k, p)-support norm is a symmetric gauge function hence it induces the orthogonally invariant *spectral (k, p)-support norm*. This interpolates between the trace norm (for $k = 1$) and the Schatten p-norms (for $k = d$) and its unit ball has a simple geometric interpretation as the convex hull of matrices of rank no greater than k and Schatten p-norm no greater than one. This suggests that the new norm favors low rank structure and the effect of varying p allows different patterns of decay in the spectrum. In the special case of $p = \infty$, the (k, p)-support norm is the dual of the Ky-Fan k-norm [Bhatia 1997] and it encourages a flat spectrum when used as a regularizer.

The main contributions of the paper are: i) we propose the (k, p)-support norm as an extension of the k-support norm and we characterize in particular the unit ball of the induced orthogonally invariant matrix norm (Section 3); ii) we show that the norm can be computed efficiently and we discuss the role of the parameter p (Section 4); iii) we outline a conditional gradient method to solve the associated regularization problem for both vector and matrix problems (Section 5);

and in the special case $p = \infty$ we provide an $\mathcal{O}(d \log d)$ computation of the projection operator (Section 5.1); finally, iv) we present numerical experiments on matrix completion benchmarks which demonstrate that the proposed norm offers significant improvement over previous methods, and we discuss the effect of the parameter p (Section 6). The appendix contains derivations of results which are sketched in or are omitted from the main body of the paper.

Notation. We use $\mathbb{N}_n$ for the set of integers from 1 up to and including n. We let $\mathbb{R}^d$ be the d-dimensional real vector space, whose elements are denoted by lower case letters. For any vector $w \in \mathbb{R}^d$, its *support* is defined as $\operatorname{supp}(w) = \{i \in \mathbb{N}_d : w_i \neq 0\}$, and its *cardinality* is defined as $\operatorname{card}(w) = |\operatorname{supp}(w)|$. We let $\mathbb{R}^{d \times m}$ be the space of $d \times m$ real matrices. We denote the rank of a matrix as $\operatorname{rank}(W)$. We let $\sigma(W) \in \mathbb{R}^r$ be the vector formed by the singular values of W, where $r = \min(d, m)$, and where we assume that the singular values are ordered nonincreasing, that is $\sigma_1(W) \geqslant \cdots \geqslant \sigma_r(W) \geqslant 0$. For $p \in [1, \infty)$ the ℓ_p-norm of a vector $w \in \mathbb{R}^d$ is defined as $\|w\|_p = (\sum_{i=1}^d |w_i|^p)^{1/p}$ and $\|w\|_\infty = \max_{i=1}^d |w_i|$. Given a norm $\|\cdot\|$ on $\mathbb{R}^d$ or $\mathbb{R}^{d \times m}$, $\|\cdot\|_*$ denotes the corresponding dual norm, defined by $\|u\|_* = \sup\{\langle u, w\rangle : \|w\| \leqslant 1\}$. The convex hull of a subset S of a vector space is denoted $\operatorname{co}(S)$.

2 Background and Previous Work

For every $k \in \mathbb{N}_d$, the k-support norm $\|\cdot\|_{(k)}$ is defined as the norm whose unit ball is given by

$$\operatorname{co}\left\{w \in \mathbb{R}^d : \text{card}(w) \leqslant k, \|w\|_2 \leqslant 1\right\}, \tag{2.1}$$

that is, the convex hull of the set of vectors of cardinality at most k and ℓ_2-norm no greater than one [Argyriou et al. 2012]. We readily see that for $k = 1$ and $k = d$ we recover the unit ball of the ℓ_1 and ℓ_2-norms respectively.

The k-support norm of a vector $w \in \mathbb{R}^d$ can be expressed as an infimal convolution [Rockafellar 1970, p. 34],

$$\|w\|_{(k)} = \inf_{(v_g)} \left\{ \sum_{g \in \mathcal{G}_k} \|v_g\|_2 : \sum_{g \in \mathcal{G}_k} v_g = w \right\}, \tag{2.2}$$

where $\mathcal{G}_k$ is the collection of all subsets of $\mathbb{N}_d$ containing at most k elements and the infimum is over all vectors $v_g \in \mathbb{R}^d$ such that $\operatorname{supp}(v_g) \subseteq g$, for $g \in \mathcal{G}_k$. Equation (2.2) highlights that the k-support norm is a special case of the group lasso with overlap [Jacob et al. 2009], where the cardinality of the support sets is at most k. This expression suggests that when used as a regularizer, the norm encourages vectors w to be a sum of a limited number of vectors with small support. Due to the variational form of (2.2) computing the norm is not straightforward, however [Argyriou et al. 2012] note that the dual norm has a simple form, namely it is the ℓ_2-norm of the k

largest components,

$$\|u\|_{(k),*} = \sqrt{\sum_{i=1}^{k}(|u|_i^{\downarrow})^2}, \quad u \in \mathbb{R}^d, \tag{2.3}$$

where $|u|^{\downarrow}$ is the vector obtained from u by reordering its components so that they are nonincreasing in absolute value. Note also from equation (2.3) that for $k = 1$ and $k = d$, the dual norm is equal to the ℓ_∞-norm and ℓ_2-norm, respectively, which agrees with our earlier observation regarding the primal norm.

A related problem which has been studied in recent years is learning a matrix from a set of linear measurements, in which the underlying matrix is assumed to have sparse spectrum (low rank). The trace norm, the ℓ_1-norm of the singular values of a matrix, has been shown to perform well in this setting, see e.g. [Argyriou et al. 2008, Jaggi and Sulovsky 2010]. 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 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 [Von Neumann 1937]. In other words, g is a norm which is invariant under permutations and sign changes of the vector components, that is $g(w) = g(Pw) = g(Jw)$, where P is any permutation matrix and J is diagonal with entries equal to ± 1 [Horn and Johnson 1991, p. 438].

Examples of symmetric gauge functions are the ℓ_p norms for $p \in [1, \infty]$ and the corresponding orthogonally invariant norms are called the Schatten p-norms [Horn and Johnson 1991, p. 441]. In particular, those include the trace norm and Frobenius norm for $p = 1$ and $p = 2$ respectively. Regularization with Schatten p-norms has been previously studied by [Argyriou et al. 2007] and a statistical analysis has been performed by [Rohde and Tsybakov 2011]. As the set $\mathcal{G}_k$ includes all subsets of size k, expression (2.2) for the k-support norm reveals that is a symmetric gauge function. [McDonald et al. 2014] use this fact to introduce the spectral k-support norm for matrices, by defining $\|W\|_{(k)} = \|\sigma(W)\|_{(k)}$, for $W \in \mathbb{R}^{d\times m}$ and report state of the art performance on matrix completion benchmarks.

3 The (k, p)-Support Norm

In this section we introduce the (k, p)-support norm as a natural extension of the k-support norm. This follows by applying the ℓ_p-norm, rather than the Euclidean norm, in the infimum convolution definition of the norm.

Definition 1. *Let $k \in \mathbb{N}_d$ and $p \in [1, \infty]$. The (k, p)-support norm of a vector $w \in \mathbb{R}^d$ is defined* *as*

$$\|w\|_{(k,p)} = \inf_{(v_g)} \left\{ \sum_{g \in \mathcal{G}_k} \|v_g\|_p : \sum_{g \in \mathcal{G}_k} v_g = w \right\}. \tag{3.1}$$

where the infimum is over all vectors $v_g \in \mathbb{R}^d$ *such that* $\operatorname{supp}(v_g) \subseteq g$, *for* $g \in \mathcal{G}_k$.

Let us note that the norm is well defined. Indeed, positivity, homogeneity and non degeneracy are immediate. To prove the triangle inequality, let $w, w' \in \mathbb{R}^d$. For any $\epsilon > 0$ there exist $\{v_g\}$ and $\{v'_g\}$ such that $w = \sum_g v_g$, $w' = \sum_g v'_g$, $\sum_g \|v_g\|_p \leqslant \|w\|_{(k,p)} + \epsilon/2$, and $\sum_g \|v'_g\|_p \leqslant \|w'\|_{(k,p)} + \epsilon/2$. As $\sum_g v_g + \sum_g v'_g = w + w'$, we have

$$\begin{aligned} \|w + w'\|_{(k,p)} &\leqslant \sum_g \|v_g\|_p + \sum_g \|v'_g\|_p \\ &\leqslant \|w\|_{(k,p)} + \|w'\|_{(k,p)} + \epsilon, \end{aligned}$$

and the result follows by letting ϵ tend to zero.

Note that, since a convex set is equivalent to the convex hull of its extreme points, Definition 1 implies that the unit ball of the (k,p)-support norm, denoted by C_k^p, is given by the convex hull of the set of vectors with cardinality no greater than k and ℓ_p-norm no greater than 1, that is

$$C_k^p = \text{co}\left\{ w \in \mathbb{R}^d : \text{card}(w) \leqslant k, \|w\|_p \leqslant 1 \right\}. \tag{3.2}$$

Definition 1 gives the norm as the solution of a variational problem. Its explicit computation is not straightforward in the general case, however for $p = 1$ the unit ball (3.2) does not depend on k and is always equal to the ℓ_1 unit ball. Thus, the $(k,1)$-support norm is always equal to the ℓ_1-norm, and we do not consider further this case in this section. Similarly, for $k = 1$ we recover the ℓ_1-norm for all values of p. For $p = \infty$, from the definition of the dual norm it is not difficult to show that $\|\cdot\|_{(k,p)} = \max\{\|\cdot\|_\infty, \|\cdot\|_1/k\}$. We return to this in Section 4 when we describe how to compute the norm for all values of p.

Note further that in Equation (3.1), as p tends to ∞, the ℓ_p-norm of each v_g is increasingly dominated by the largest component of v_g. As the variational formulation tries to identify vectors v_g with small aggregate ℓ_p-norm, this suggests that higher values of p encourage each v_g to tend to a vector whose k entries are equal. In this manner varying p allows us adjust the degree to which the components of vector w can be clustered into (possibly overlapping) groups of size k.

As in the case of the k-support norm, the dual (k,p)-support norm has a simple expression. Recall that the dual norm of a vector $u \in \mathbb{R}^d$ is defined by the optimization problem

$$\|u\|_{(k,p),*} = \max\left\{ \langle u, w \rangle : \|w\|_{(k,p)} = 1 \right\}. \tag{3.3}$$

Proposition 2. *If $p \in (1, \infty]$ then the dual (k,p)-support norm is given by*

$$\|u\|_{(k,p),*} = \left(\sum_{i \in I_k} |u_i|^q \right)^{\frac{1}{q}}, \quad u \in \mathbb{R}^d,$$

where $q = p/(p-1)$ and $I_k \subset \mathbb{N}_d$ is the set of indices of the k largest components of u in absolute value. Furthermore, if $p \in (1, \infty)$ and $u \in \mathbb{R}^d \backslash \{0\}$ then the maximum in (3.3) *is attained for*

$$w_i = \begin{cases} \operatorname{sign}(u_i) \left(\frac{|u_i|}{\|u\|_{(k,p),*}} \right)^{\frac{1}{p-1}} & \textit{if } i \in I_k, \\ 0 & \textit{otherwise}. \end{cases} \tag{3.4}$$

If $p = \infty$ the maximum is attained for

$$w_i = \begin{cases} \operatorname{sign}(u_i) & \textit{if } i \in I_k, u_i \neq 0, \\ \lambda_i \in [-1, 1] & \textit{if } i \in I_k, u_i = 0, \\ 0 & \textit{otherwise}. \end{cases}$$

Note that for $p = 2$ we recover the dual of the k-support norm in (2.3).

3.1 The Spectral (k,p)-Support Norm

From Definition 1 it is clear that the (k,p)-support norm is a symmetric gauge function. This follows since $\mathcal{G}_k$ contains all groups of cardinality k and the ℓ_p-norms only involve absolute values of the components. Hence we can define the spectral (k,p)-support norm as

$$\|W\|_{(k,p)} = \|\sigma(W)\|_{(k,p)}, \quad W \in \mathbb{R}^{d \times m}.$$

Since the dual of any orthogonally invariant norm is given by $\| \cdot \|_* = \|\sigma(\cdot)\|_*$, see e.g. [Lewis 1995], we conclude that the dual spectral (k,p)-support norm is given by

$$\|Z\|_{(k,p),*} = \|\sigma(Z)\|_{(k,p),*}, \quad Z \in \mathbb{R}^{d \times m}.$$

The next result characterizes the unit ball of the spectral (k,p)-support norm. Due to the relationship between an orthogonally invariant norm and its corresponding symmetric gauge function, we see that the cardinality constraint for vectors generalizes in a natural manner to the rank operator for matrices.

Proposition 3. *The unit ball of the spectral (k,p)-support norm is the convex hull of the set of matrices of rank at most k and Schatten p-norm no greater than one.*

In particular, if $p = \infty$, the dual vector norm is given by $u \in \mathbb{R}^d$, by $\|u\|_{(k,\infty),*} = \sum_{i=1}^k |u|_i^\downarrow$. Hence, for any $Z \in \mathbb{R}^{d\times m}$, the dual spectral norm is given by $\|Z\|_{(k,\infty),*} = \sum_{i=1}^k \sigma_i(Z)$, that is the sum of the k largest singular values, which is also known as the Ky-Fan k-norm, see e.g. [Bhatia 1997].

4 Computing the Norm

In this section we compute the norm, illustrating how it interpolates between the ℓ_1 and ℓ_p-norms.

Theorem 4. *Let $p \in (1,\infty)$. For every $w \in \mathbb{R}^d$, and $k \leqslant d$, it holds that*

$$\|w\|_{(k,p)} = \left[\sum_{i=1}^{\ell}(|w|_i^\downarrow)^p + \left(\frac{\sum_{i=\ell+1}^d |w|_i^\downarrow}{\sqrt[q]{k-\ell}}\right)^p\right]^{\frac{1}{p}} \tag{4.1}$$

where $\frac{1}{p}+\frac{1}{q}=1$, and for $k = d$, we set $\ell = d$, otherwise ℓ is the largest integer in $\{0,\dots,k-1\}$ satisfying

$$(k-\ell)|w|_\ell^\downarrow \geqslant \sum_{i=\ell+1}^d |w|_i^\downarrow. \tag{4.2}$$

Furthermore, the norm can be computed in $\mathcal{O}(d\log d)$ time.

Proof. Note first that in (4.1) when $\ell = 0$ we understand the first term in the right hand side to be zero, and when $\ell = d$ we understand the second term to be zero.

We need to compute

$$\|w\|_{(k,p)} = \max\left\{\sum_{i=1}^d u_i w_i : \|u\|_{(k,p),*} \leqslant 1\right\}$$

where the dual norm $\|\cdot\|_{(k,p),*}$ is described in Proposition 2. Let $z_i = |w|_i^\downarrow$. The problem is then equivalent to

$$\max\left\{\sum_{i=1}^d z_i u_i : \sum_{i=1}^k u_i^q \leqslant 1, u_1 \geqslant \cdots \geqslant u_d\right\}. \tag{4.3}$$

This further simplifies to the k-dimensional problem

$$\max\left\{\sum_{i=1}^{k-1} u_i z_i + u_k \sum_{i=k}^d z_i : \sum_{i=1}^k u_i^q \leqslant 1, u_1 \geqslant \cdots \geqslant u_k\right\}.$$

Note that when $k = d$, the solution is given by the dual of the ℓ_q-norm, that is the ℓ_p-norm. For the remainder of the proof we assume that $k < d$. We can now attempt to use Holder's inequality, which states that for all vectors x such that $\|x\|_q = 1$, $\langle x, y\rangle \leqslant \|y\|_p$, and the inequality is tight if and only if

$$x_i = \left(\frac{|y_i|}{\|y\|_p}\right)^{p-1} \mathrm{sign}(y_i).$$

We use it for the vector $y = (z_1, \ldots, z_{k-1}, \sum_{i=k}^d z_i)$. The components of the maximizer u satisfy $u_i = \left(\frac{z_i}{M_{k-1}}\right)^{p-1}$ if $i \leqslant k-1$, and

$$u_k = \left(\frac{\sum_{i=\ell+1}^d z_i}{M_{k-1}}\right)^{p-1}.$$

where for every $\ell \in \{0, \ldots, k-1\}$, M_ℓ denotes the r.h.s. in equation (4.1). We then need to verify that the ordering constraints are satisfied. This requires that

$$(z_{k-1})^{p-1} \geqslant \left(\sum_{i=k}^d z_i\right)^{p-1}$$

which is equivalent to inequality (4.2) for $\ell = k-1$. If this inequality is true we are done, otherwise we set $u_k = u_{k-1}$ and solve the smaller problem

$$\max\left\{\sum_{i=1}^{k-2} u_i z_i + u_{k-1}\sum_{i=k-1}^d z_i \ : \\ \sum_{i=1}^{k-2} u_i^q + 2u_{k-1}^q \leqslant 1, \quad u_1 \geqslant \cdots \geqslant u_{k-1}\right\}.$$

We use again Hölder's inequality and keep the result if the ordering constraints are fulfilled. Continuing in this way, the generic problem we need to solve is

$$\max\left\{\sum_{i=1}^{\ell} u_i z_i + u_{\ell+1}\sum_{i=\ell+1}^d z_i \ : \\ \sum_{i=1}^{\ell} u_i^q + (k-\ell)u_{\ell+1}^q \leqslant 1, \quad u_1 \geqslant \cdots \geqslant u_{\ell+1}\right\}$$

where $\ell \in \{0, \ldots, k-1\}$. Without the ordering constraints the maximum, M_ℓ, is obtained by the change of variable $u_{\ell+1} \mapsto (k-\ell)^{\frac{1}{q}} u_\ell$ followed by applying Hölder's inequality. A direct computation provides that the maximizer is $u_i = \left(\frac{z_i}{M_\ell}\right)^{p-1}$ if $i \leqslant \ell$, and

$$(k-\ell)^{\frac{1}{q}} u_{\ell+1} = \left(\frac{\sum_{i=\ell+1}^{d} z_i}{(k-\ell)^{\frac{1}{q}} M_\ell^p}\right)^{p-1} .$$

Using the relationship $\frac{1}{p} + \frac{1}{q} = 1$, we can rewrite this as

$$u_{\ell+1} = \left(\frac{\sum_{i=\ell+1}^{d} z_i}{(k-\ell) M_\ell^p}\right)^{p-1} .$$

Hence, the ordering constraints are satisfied if

$$z_\ell^{p-1} \geqslant \left(\frac{\sum_{i=\ell+1}^{d} z_i}{(k-\ell)}\right)^{p-1} ,$$

which is equivalent to (4.2). Finally note that M_ℓ is a nondecreasing function of ℓ. This is because the problem with a smaller value of ℓ is more constrained, namely, it solves (4.3) with the additional constraints $u_{\ell+1} = \cdots = u_d$. Moreover, if the constraint (4.2) holds for some value $\ell \in \{0, \ldots, k-1\}$ then it also holds for a smaller value of ℓ, hence we maximize the objective by choosing the largest ℓ.

The computational complexity stems from using the monotonicity of M_ℓ with respect to ℓ, which allows us to identify the critical value of ℓ using binary search. □

Note that for $k = d$ we recover the ℓ_p-norm and for $p = 2$ we recover the result in [Argyriou et al. 2012, McDonald et al. 2014], however our proof technique is different from theirs.

Remark 5 (Computation of the norm for $p \in \{1, \infty\}$)**.** *Since the norm $\|\cdot\|_{(k,p)}$ computed above for $p \in (1, \infty)$ is continuous in p, the special cases $p = 1$ and $p = \infty$ can be derived by a limiting argument. We readily see that for $p = 1$ the norm does not depend on k and it is always equal to the ℓ_1-norm, in agreement with our observation in the previous section. For $p = \infty$ we obtain that $\|w\|_{(k,\infty)} = \max(\|w\|_\infty, \|w\|_1/k)$.*

5 Optimization

In this section, we describe how to solve regularization problems using the vector and matrix (k,p)-support norms. We consider the constrained optimization problem

$$\min\left\{f(w) : \|w\|_{(k,p)} \leqslant \alpha\right\} , \tag{5.1}$$

Algorithm 1 Frank-Wolfe.

```
Choose $w^{(0)}$ such that $\|w^{(0)}\|_{(k,p)} \leqslant \alpha$
for $t = 0, \ldots, T$ do
    Compute $g := \nabla f(w^{(t)})$
    Compute $s :=$ argmin $\left\{\langle s, g)\rangle : \|s\|_{(k,p)} \leqslant \alpha\right\}$
    Update $w^{(t+1)} := (1-\gamma)w^{(t)} + \gamma s$, for $\gamma := \frac{2}{t+2}$
end for
```

where w is in $\mathbb{R}^d$ or $\mathbb{R}^{d\times m}$, $\alpha > 0$ is a regularization parameter and the error function f is assumed to be convex and continuously differentiable. For example, in linear regression a valid choice is the square error, $f(w) = \|Xw - y\|_2^2$, where X is matrix of observations and y a vector of response variables. Constrained problems of form (5.1) are also referred to as Ivanov regularization in the inverse problems literature [Ivanov et al. 1978].

A convenient tool to solve problem (5.1) is provided by the *Frank-Wolfe* method [Frank and Wolfe 1956], see also [Jaggi 2013] for a recent account. The method is outlined in Algorithm 1, and it has worst case convergence rate $\mathcal{O}(1/T)$. The key step of the algorithm is to solve the subproblem

$$\text{argmin } \left\{\langle s, g\rangle : \|s\|_{(k,p)} \leqslant \alpha\right\}, \tag{5.2}$$

where $g = \nabla f(w^{(t)})$, that is the gradient of the objective function at the t-th iteration. This problem involves computing a subgradient of the dual norm at g. It can be solved exactly and efficiently as a consequence of Proposition 2. We discuss here the vector case and postpone the discussion of the matrix case to Section 5.2. By symmetry of the ℓ_p-norm, problem (5.2) can be solved in the same manner as the maximum in Proposition 2, and the solution is given by $s_i = -\alpha w_i$, where w_i is given by (3.4). Specifically, letting $I_k \subset \mathbb{N}_d$ be the set of indices of the k largest components of g in absolute value, for $p \in (1, \infty)$ we have

$$s_i = \begin{cases} -\alpha \operatorname{sign}(g_i) \left(\frac{g_i}{\|g\|_{(k,p),*}}\right)^{\frac{1}{p-1}}, & \text{if } i \in I_k \\ 0, & \text{if } i \notin I_k \end{cases} \tag{5.3}$$

and, for $p = \infty$ we choose the subgradient

$$s_i = \begin{cases} -\alpha \operatorname{sign}(g_i) & \text{if } i \in I_k,\ g_i \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

5.1 Projection Operator

An alternative method to solve (5.1) in the vector case is to consider the equivalent problem

$$\min\left\{f(w) + \delta_{\{\|\cdot\|_{(k,p)} \leqslant \alpha\}}(w) : w \in \mathbb{R}^d\right\}, \tag{5.5}$$

where $\delta_C(\cdot)$ is the indicator function of convex set C. Proximal gradient methods can be used to solve optimization problems of the form $\min\left\{f(w)+\lambda g(w) : w \in \mathbb{R}^d\right\}$, 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 [Beck and Teboulle 2009, Nesterov 2007] and references therein. The proximity operator of g with parameter $\rho > 0$ is defined as $\operatorname{prox}_{\rho g}(w) = \operatorname{argmin}\{\frac{1}{2}\|x-w\|^2 + \rho g(x) : x \in \mathbb{R}^d\}$. The proximity operator for the squared k-support norm was computed by [Argyriou et al. 2012] and [McDonald et al. 2014], and for the k-support norm by [Chatterjee et al. 2014].

In the special case that $g(w) = \delta_C(w)$, where C is a convex set, the proximity operator reduces to the projection operator onto C. For the (k,p)-support norm, for the case $p = \infty$ we can compute the projection onto its unit ball using the following result.

Proposition 6. *For every $w \in \mathbb{R}^d$, the projection x of w onto the unit ball of the (k,∞)-norm is given by*

$$x_i = \begin{cases} \operatorname{sign}(w_i)(|w_i| - \beta), & \textit{if } ||w_i| - \beta| \leqslant 1, \\ \operatorname{sign}(w_i), & \textit{if } ||w_i| - \beta| > 1, \end{cases} \tag{5.6}$$

where $\beta = 0$ if $\|w\|_1 \leqslant k$, otherwise $\beta \in (0,\infty)$ is chosen such that $\sum_{i=1}^d |x_i| = k$. Furthermore, the projection can be computed in $\mathcal{O}(d \log d)$ time.

Proof. (Sketch) We solve the optimization problem

$$\min_{x \in \mathbb{R}^d} \left\{ \sum_{i=1}^d (x_i - w_i)^2 : |x_i| \leqslant 1, \sum_{i=1}^d |x_i| \leqslant k \right\}. \tag{5.7}$$

We consider two cases. If $\sum_{i=1}^d |w_i| \leqslant k$, then the problem decouples and we solve it componentwise. If $\sum_{i=1}^d |w_i| > k$, we solve problem (5.7) by minimizing the Lagrangian function $\mathcal{L}(x,\beta) = \sum_{i=1}^d (x_i - w_i)^2 + 2\beta(\sum_{i=1}^d |x_i| - k)$ with nonnegative multiplier β. This can be done componentwise, and at the optimum the constraint $\sum_{i=1}^d |x_i| \leqslant k$ will be tight. Finally, both cases can be combined into the form of (5.6). The complexity follows by taking advantage of the monotonicity of $x_i(\beta)$. □ We can use Proposition 6 to project onto the unit ball of radius $\alpha > 0$ by a rescaling argument (see the appendix for details).

5.2 Matrix Problem

Given data matrix $X \in \mathbb{R}^{d \times m}$ for which we observe a subset of entries, we consider the constrained optimization problem

$$\min_{W \in \mathbb{R}^{d \times m}} \left\{ \|\Omega(X) - \Omega(W)\|_{\mathrm{F}} : \|W\|_{(k,p)} \leqslant \alpha \right\} \tag{5.8}$$

Table 1: Matrix completion on rank 5 matrix with flat spectrum. The improvement of the (k, p)-support norm over the k-support and trace norms is considerable (statistically significant at a level < 0.001).

dataset	norm	test error	k	p
rank 5	trace	0.8415 (0.03)	-	-
ρ=10%	k-supp	0.8343 (0.03)	3.3	-
	kp-supp	0.8108 (0.05)	5.0	∞
rank 5	trace	0.6161 (0.03)	-	-
ρ=15%	k-supp	0.6129 (0.03)	3.3	-
	kp-supp	0.4262 (0.04)	5.0	∞
rank 5	trace	0.4453 (0.03)	-	-
ρ=20%	k-supp	0.4436 (0.02)	3.5	-
	kp-supp	0.2425 (0.02)	5.0	∞
rank 5	trace	0.1968 (0.01)	-	-
ρ=30%	k-supp	0.1838 (0.01)	5.0	-
	kp-supp	0.0856 (0.01)	5.0	∞

where the operator Ω applied to a matrix sets unobserved values to zero. As in the vector case, the Frank-Wolfe method can be applied to the matrix problems. Algorithm 1 is particularly convenient in this case as we only need to compute the largest k singular values, which can result in a computationally efficient algorithm. The result is a direct consequence of Proposition 2 and von Neumann's trace inequality, see e.g. [Marshall and Olkin 1979, Ch. 9 Sec. H.1.h]. We obtain that the solution of the inner minimization step is $U_k \mathrm{diag}(s) V_k^\top$ where U_k and V_k are the top k left and right singular vectors of the gradient G of the objective function in (5.8) evaluated at the current solution, whose singular values we denote by g, and s is obtained from g as per equations (5.3) and (5.4), for $p \in (1, \infty)$ and $p = \infty$, respectively.

Note also that the proximity operator of the norm and the Euclidean projection on the associated unit ball both require the full singular value decomposition to be performed. Indeed, the proximity operator of an orthogonally invariant norm $\|\cdot\| = g(\sigma(\cdot))$ at $W \in \mathbb{R}^{d\times m}$ is given by $\mathrm{prox}_{\|\cdot\|}(W) = U \mathrm{diag}(\mathrm{prox}_g(\sigma(W)))V^\top$, where U and V are the matrices formed by the left and right singular vectors of W, see e.g. [Argyriou et al. 2011, Prop. 3.1], and this requires the full decomposition.

6 Numerical Experiments

In this section we apply the spectral (k, p)-support norm to matrix completion (collaborative filtering), in which we want to recover a low rank, or approximately low rank, matrix from a small sam-

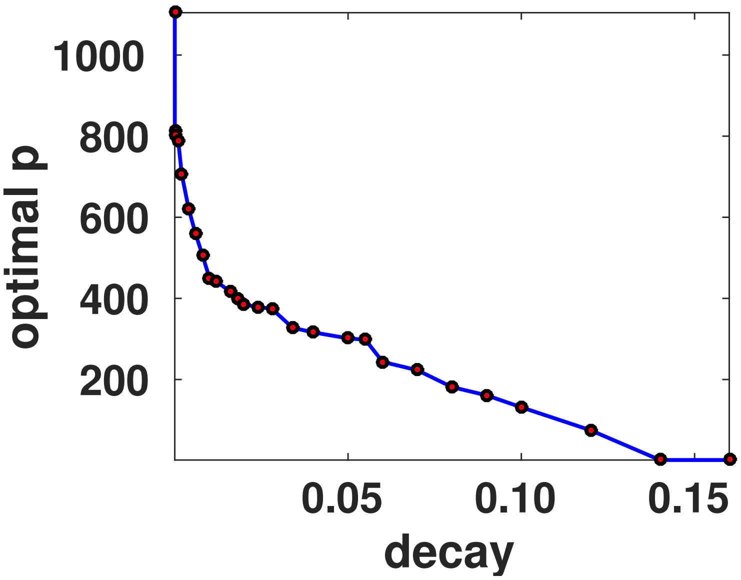

Figure 1: Optimal p vs. decay a.

ple of its entries, see e.g. [Jaggi and Sulovsky 2010]. One prominent method of solving 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) [Jaggi and Sulovsky 2010, Mazumder et al. 2010, Toh and Yun 2011]. We apply the (k,p)-support norm to this framework and we investigate the impact of varying p. Next we compare the spectral (k,p)-support norm to the trace norm and the spectral k-support norm ($p=2$) in both synthetic and real datasets. In each case we solve the optimization problem (5.8) using the Frank-Wolfe method as outlined in Section 5. We determine the values of k and $p \geqslant 1$ by validation, averaged over a number of trials. Specifically, we choose the optimal p, k, as well as the regularization parameter α by validation over a grid. We let alpha vary in 10^0 to 10^5 with step $10^{0.25}$, we let p vary over 20 values from 1 to $50,000$, plus $p=\infty$, and vary k from 1 to 20. Our code is available from *http://www0.cs.ucl.ac.uk/staff/M.Pontil/software.html*.

Impact of p. A key motivation for the additional parameter p is that it allows us to tune the norm to the decay of the singular values of the underlying matrix. In particular the variational formulation of (3.1) suggests that as the spectrum of the true low rank matrix flattens out, larger values of p should be preferred.

We ran the method on a set of 100×100 matrices of rank 12, with decay of the non zero singular values σ_ℓ proportional to $\exp(-\ell a)$, for 26 values of a between 10^{-6} and 0.18, and we determined the corresponding optimal value of p. Figure 1 illustrates the optimum value of p as a function of a. We clearly observe the negative slope, that is the steeper the slope the smaller the

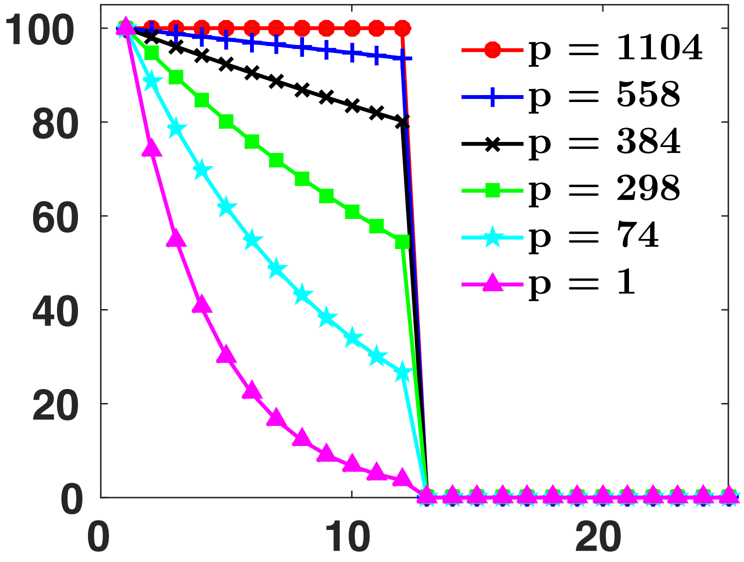

Figure 2: Optimal p fitted to Matrix spectra with various decays.

optimal value of p. Figure 2 shows the spectrum and the optimal p for several decay values.

Note that k is never equal to 1, which is a special case in which the norm is independent of p, and is equal to the trace norm. In each case the improvement of the spectral (k, p)-support norm over the k-support and trace norms is statistically significant at a level < 0.001.

Figure 3 illustrates the impact of the curvature p on the test error on synthetic and real datasets. We observe that the error levels off as p tends to infinity, so for these specific datasets the major gain is to be had for small values of p. The optimum value of p for both the real and synthetic datasets is statistically different from $p = 2$ (k-support norm), and $p = 1$ (trace norm).

Simulated Data. Next we compared the performance of the (k, p)-support norm to that of the k-support norm and the trace norm for a matrix with flat spectrum. As outlined above, as the spectrum of the true low rank matrix flattens out, larger values of p should be preferred. Each 100×100 matrix is generated as $W = ASB^\top + E$, where U and V are the singular vectors of the matrix $UV^\top$, where $U, V \in \mathbb{R}^{100\times 5}$, the entries of U, V and E are i.i.d. standard Gaussian, and S is diagonal with 5 non zero constant entries. Table 1 illustrates the performance of the norms on a synthetic dataset of rank 5, with identical singular values, that is a flat spectrum. In each regime the case $p = \infty$ outperforms the other norms by a substantial margin, with statistical significance at a level < 0.001. We followed the framework of [McDonald et al. 2014] and use ρ to denote the percentage of the data to use in the training set. We further replicated the setting of [McDonald et al. 2014] for synthetic matrix completion, and found that the (k, p)-support norm

Table 2: Matrix completion on real datasets. The improvement of the (k, p)-support norm over the k-support and trace norms is statistically significant at a level < 0.001.

dataset	norm	test error	k	p
MovieLens 100k	trace	0.2017	-	-
	k-supp	0.1990	1.9	-
	kp-supp	0.1921	2.0	∞
Jester 1	trace	0.1752	-	-
	k-supp	0.1739	6.4	-
	kp-supp	0.1744	2.0	∞
	kp-supp	0.1731	2.0	6.5
Jester 3	trace	0.1959	-	-
	k-supp	0.1942	2.1	-
	kp-supp	0.1841	2.0	∞

outperformed the standard k-support norm, as well as the trace norm, at a statistically significant level (see Table 3 in the appendix for details).

We note that although Frank-Wolfe method for the (k, p)-support norm does not generally converge as quickly as proximal methods (which are available in the case of k-support norm [McDonald et al. 2016, McDonald et al. 2014, Chatterjee et al. 2014]), the computational cost can be mitigated using the continuation method. Specifically given an ordered sequence of parameter values for p we can proceed sequentially, initializing its value based on the previously computed value. Empirically we tried this approach for a series of 30 values of p and found that the total computation time increased only moderately.

Real Data. Finally, we applied the norms to real collaborative filtering datasets. We observe a subset of the (user, rating) entries of a matrix and predict the unobserved ratings, with the assumption that the true matrix is likely to have low rank. We report on the MovieLens 100k dataset (*http://grouplens.org/datasets/movielens/*), which consists of ratings of movies, and the Jester 1 and 3 datasets (*http://goldberg.berkeley.edu/jester-data/*), which consist of ratings of jokes. We followed the experimental protocol in [McDonald et al. 2014, Toh and Yun 2011], using normalized mean absolute error [Toh and Yun 2011], and we implemented a final thresholding step as in [McDonald et al. 2014] (see the appendix for further details). The results are outlined in Table 2. The spectral (k, p)-support outperformed the trace norm and the spectral k-support norm, and the improvement is statistically significant at a level < 0.001 (the standard deviations, not shown here, are of the order of 10^{-5}). In summary the experiments suggest that the additional flexibility of the p parameter does allow the model to better fit both the sparsity and the decay of the true spectrum.

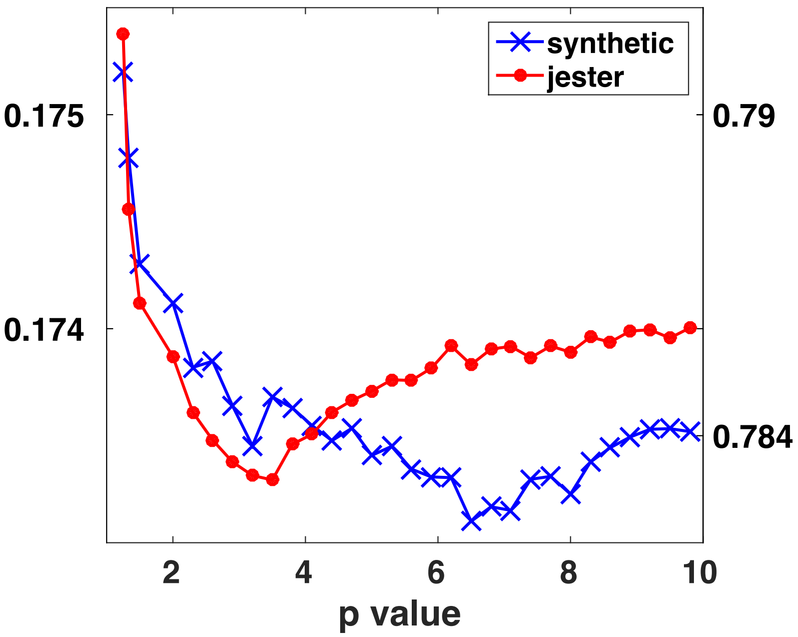

Figure 3: Test error vs curvature (p). Left axis: synthetic data (blue crosses); right axis: Jester 1 dataset (red circles).

7 Conclusion

We presented a generalization of the k-support norm, the (k, p)-support norm, where the additional parameter p is used to better fit the decay of the components of the underlying model. We determined the dual norm, characterized the unit ball and computed an explicit expression for the norm. As the norm is a symmetric gauge function, we further described the induced spectral (k, p)-support norm. We adapted the Frank-Wolfe method to solve regularization problems with the norm, and in the particular case $p = \infty$ we provided a fast computation for the projection operator. In numerical experiments we considered synthetic and real matrix completion problems and we showed that varying p leads to significant performance improvements. Future work could include deriving statistical bounds for the performance of the norms, and situating the norms in the framework of other structured sparsity norms which have recently been studied.

References

[Argyriou et al. 2007] A. Argyriou, C. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. *NIPS*, 2007.

[Argyriou et al. 2008] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[Argyriou et al. 2011] 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.

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

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

[Bhatia 1997] R. Bhatia. *Matrix Analysis*. Springer, 1997.

[Buehlmann and van der Geer 2011] P. Buehlmann and S. A. van der Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

[Chatterjee et al. 2014] S. Chatterjee, S. Chen, and A. Banerjee. Generalized Dantzig selector: application to the k-support norm. In *Advances in Neural Information Processing Systems 28*, 2014.

[Frank and Wolfe 1956] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3 (1-2):95–110, 1956.

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

[Ivanov et al. 1978] V.K. Ivanov, V. V. Vasin, and V.P. Tanana. *Theory of Linear Ill-Posed Problems and its Applications*. De Gruyter, 1978.

[Jacob et al. 2009] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. *Proceedings of the 26th International Conference on Machine Learning*, 2009.

[Jaggi 2013] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

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

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

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

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

[McDonald et al. 2016] A. M. McDonald, M. Pontil, and D. Stamos. New perspectives on k-support and cluster norms. *Journal of Machine Learning Research*, 2016a.

[McDonald et al. 2014] A. M. McDonald, M. Pontil, and D. Stamos. Spectral k-support regularization. In *Advances in Neural Information Processing Systems 28*, 2014b.

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

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

[Rohde and Tsybakov 2011] A. Rohde and A.B. Tsybakov. Estimation of high-dimensional low rank matrices. *Annals of Statistics*, 39:887–930, 2011.

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

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

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

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

[Wainwright 2014] M. Wainwright. Structured regularizers for high-dimensional problems. *Annual Review of Statistics and Its Application*, 1:233–253, 2014.

[Zou and Hastie 2005] 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.

A Appendix

In this appendix, we provide proofs of the results stated in the main body of the paper, and we include experimental details and results that were not included in the paper for space reasons.

A.1 Proof of Proposition 2

For every $u \in \mathbb{R}^d$ we have

$$\begin{aligned}
\|u\|_{(k,p),*} &= \max\left\{\sum_{i=1}^{d} u_i w_i : w \in C_k^p\right\} \\
&= \max\left\{\sum_{i=1}^{d} u_i w_i : \text{card}(w) \leqslant k, \|w\|_p \leqslant 1\right\} \\
&= \max\left\{\sum_{i\in I_k} u_i w_i : \sum_{i\in I_k} |w_i|^p \leqslant 1\right\} \\
&= \left(\sum_{i\in I_k} |u_i|^q\right)^{\frac{1}{q}},
\end{aligned}$$

where the first equality uses the definition of the unit ball (3.2) and the second equality is true because the maximum of a linear functional on a compact set is attained at an extreme point of the set. The third equality follows by using the cardinality constraint, that is we set $w_i = 0$ if $i \notin I_k$. Finally, the last equality follows by Hölder's inequality in $\mathbb{R}^k$ [Marshall and Olkin 1979, Ch. 16 Sec. D.1].

The second claim is a direct consequence of the cardinality constraint and Hölder's inequality in $\mathbb{R}^k$. □

To prove Proposition 3 we require the following auxiliary result. Let X be a finite dimensional vector space. Recall that a subset C of X is called *balanced* if $\alpha C \subseteq C$ whenever $|\alpha| \leqslant 1$. Furthermore, C is called *absorbing* if for any $x \in X$, $x \in \lambda C$ for some $\lambda > 0$.

Lemma 7. *Let $C \subseteq X$ be a bounded, convex, balanced, and absorbing set. The Minkowski functional μ_C of C, defined, for every $w \in X$, as*

$$\mu_C(w) = \inf\left\{\lambda : \lambda > 0, \; \frac{1}{\lambda} w \in C\right\}$$

is a norm on X.

Proof. We give a direct proof that μ_C satisfies the properties of a norm. See also e.g. [Rudin 1991, §1.35] for further details. Clearly $\mu_C(w) \geqslant 0$ for all w, and $\mu_C(0) = 0$. Moreover, as C is bounded, $\mu_C(w) > 0$ whenever $w \neq 0$.

Next we show that μ_C is one-homogeneous. For every $\alpha \in \mathbb{R}$, $\alpha \neq 0$, let $\sigma = \text{sign}(\alpha)$ and note that

$$
\begin{aligned}
\mu_C(\alpha w) &= \inf\left\{\lambda > 0 : \frac{1}{\lambda}\alpha w \in C\right\} \\
&= \inf\left\{\lambda > 0 : \frac{|\alpha|}{\lambda}\sigma w \in C\right\} \\
&= |\alpha| \inf\left\{\lambda > 0 : \frac{1}{\lambda} w \in \sigma C\right\} \\
&= |\alpha| \inf\left\{\lambda > 0 : \frac{1}{\lambda} w \in C\right\} \\
&= |\alpha| \mu_C(w),
\end{aligned}
$$

where we have made a change of variable and used the fact that $\sigma C = C$.

Finally, we prove the triangle inequality. For every $v, w \in X$, if $v/\lambda \in C$ and $w/\mu \in C$ then setting $\gamma = \lambda/(\lambda + \mu)$, we have

$$
\frac{v+w}{\lambda+\mu} = \gamma\frac{v}{\lambda} + (1-\gamma)\frac{w}{\mu}
$$

and since C is convex, then $\frac{v+w}{\lambda+\mu} \in C$. We conclude that $\mu_C(v+w) \leqslant \mu_C(v) + \mu_C(w)$. The proof is completed. □ Note that for such set C, the unit ball of the induced norm μ_C is C. Furthermore, if $\|\cdot\|$ is a norm then its unit ball satisfies the hypotheses of Lemma 7.

A.2 Proof of Proposition 3

Define the set

$$
T_k^p = \{W \in \mathbb{R}^{d\times m} : \text{rank}(W) \leqslant k, \|\sigma(W)\|_p \leqslant 1\},
$$

and its convex hull $A_k^p = \text{co}(T_k^p)$, and consider the Minkowski functional

$$
\lambda(W) = \inf\{\lambda > 0 : W \in \lambda A_k^p\}, \quad W \in \mathbb{R}^{d\times m}. \tag{A.1}
$$

We show that A_k^p is absorbing, bounded, convex and symmetric, and it follows by Lemma 7 that λ defines a norm on $\mathbb{R}^{d\times m}$ with unit ball equal to A_k^p. The set A_k^p is clearly bounded, convex and symmetric. To see that it is absorbing, let W in $\mathbb{R}^{d\times m}$ have singular value decomposition $U\Sigma V^\top$, and let $r = \min(d, m)$. If W is zero then clearly $W \in A_k^p$, so assume it is non zero.

For $i \in \mathbb{N}_r$ let $S_i \in \mathbb{R}^{d\times m}$ have entry (i, i) equal to 1, and all remaining entries zero. We then have

$$\begin{aligned} W &= U\Sigma V^\top \\ &= U\left(\sum_{i=1}^{r}\sigma_i S_i\right)V^\top \\ &= \left(\sum_{i=1}^{d}\sigma_i\right)\sum_{i=1}^{r}\frac{\sigma_i}{\sum_{j=1}^{r}\sigma_j}(US_iV^\top) \\ &=: \lambda\sum_{i=1}^{r}\lambda_i Z_i. \end{aligned}$$

Now for each i, $\|\sigma(Z_i)\|_p = \|\sigma(S_i)\|_p = 1$, and $\mathrm{rank}(Z_i) = \mathrm{rank}(S_i) = 1$, so $Z_i \in T_k^p$ for any $k \geqslant 1$. Furthermore $\lambda_i \in [0,1]$ and $\sum_{i=1}^{r}\lambda_i = 1$, that is $(\lambda_1, \ldots, \lambda_r)$ lies in the unit simplex in $\mathbb{R}^d$, so $\frac{1}{\lambda}W$ is a convex combination of elements of Z_i, in other words $W \in \lambda A_k^p$, and we have shown that A_k^p is absorbing. It follows that A_k^p satisfies the hypotheses of Lemma 7, and λ defines a norm on $\mathbb{R}^{d\times m}$ with unit ball equal to A_k^p.

Since the constraints in T_k^p involve spectral functions, the sets T_k^p and A_k^p 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. By von Neumann's Theorem [Von Neumann 1937] the norm it defines is orthogonally invariant and we have

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

where we have used Equation (3.2), which states that C_k^p is the unit ball of the (k,p)-support norm. It follows that the norm defined by (A.1) is the spectral (k,p)-support norm with unit ball given by A_k^p.

□

A.3 Proof of Proposition 6

We solve the optimization problem

$$\operatorname{argmin}_{x\in\mathbb{R}^d}\left\{\sum_{i=1}^{d}(x_i - w_i)^2 : |x_i| \leqslant 1, \sum_{i=1}^{d}|x_i| \leqslant k\right\}. \tag{A.2}$$

We consider two cases. If $\sum_{i=1}^{d}|w_i| \leqslant k$, then the problem decouples and we solve it componentwise. Specifically we minimize $(x_i - w_i)^2$ subject to $|x_i| \leqslant 1$, and the solution is immediately given by

$$x_i = \begin{cases} -1, & \text{if } w_i < -1, \\ w_i, & \text{if } -1 \leqslant w_i \leqslant 1, \\ 1, & \text{if } w_i > 1. \end{cases} \tag{A.3}$$

We now assume that $\sum_{i=1}^d |w_i| > k$. Consider the Lagrangian function $\mathcal{L}(x, \beta) = \sum_{i=1}^d (x_i - w_i)^2 + 2\beta \left(\sum_{i=1}^d |x_i| - k \right)$ with nonnegative multiplier β. We solve problem (A.2) by minimizing the Lagrangian with respect to x, which can be done componentwise due to the coupling effect of the Lagrangian. Furthermore, at the optimum the constraint $\sum_{i=1}^d |x_i| \leqslant k$ will be tight. The derivative with respect to x_i is zero when $x_i = w_i - \beta \operatorname{sign}(x_i)$. Incorporating the constraint $|x_i| \leqslant 1$ we get the following solution

$$x_i = \begin{cases} -1, & \text{if } w_i + \beta < -1, \\ w_i + \beta, & \text{if } -1 \leqslant w_i + \beta \leqslant 0, \\ w_i - \beta, & \text{if } 0 \leqslant w_i - \beta \leqslant 1, \\ 1, & \text{if } w_i - \beta > 1, \end{cases} \tag{A.4}$$

where $\beta \geqslant 0$ is chosen such that $\sum_{i=1}^d |x_i(\beta)| = k$. Note that for $\beta = 0$, which corresponds to $\|w\|_1 \leqslant k$, (A.4) reduces to (A.3), hence we obtain the compact notation (5.6). Finally, note that the expression $\sum_{i=1}^d |x_i(\beta)|$ decreases monotonically as β increases. In the case that $\|w\|_1 > k$, $\beta \in (0, |w_j| - 1)$, where $|w_j| = \operatorname{argmin} |w_i|$, hence we can determine β by binary search in $\mathcal{O}(d \log d)$ time.

□

In order to project onto the unit ball of radius $\alpha > 0$, we solve the optimization problem $\min\{\sum_{i=1}^d (x_i - w_i)^2 \ : \ x \in \mathbb{R}^d, \ |x_i| \leqslant \alpha, \ \sum_{i=1}^d |x_i| \leqslant \alpha k\}$. To do so, we make the change of variables $x_i' = x_i/\alpha$ and note that the problem reduces to computing the projection x' of w' onto the unit ball of the norm, where $w_i' = w_i/\alpha$, which is the problem that was solved in (5.7). Once this is done, our solution is given by $x_i = \alpha x_i'(\beta)$, where $x'(\beta)$ is determined in accordance with Proposition 6.

A.4 Numerical Experiments

In this section we report further experimental details and results not included in the main body of the paper for space reasons.

Simulated Datasets. We replicated the setting of [McDonald et al. 2014] in order to verify that the additional parameter can improve performance. Each 100×100 matrix is generated as $W = UV^\top + E$, where $U, V \in \mathbb{R}^{100 \times r}$, $r \ll 100$, and the entries of U, V and E are i.i.d. standard

Table 3: Matrix completion on synthetic datasets generated with decaying spectrum. The improvement of the (k,p)-support norm over the k-support and trace norms is statistically significant at a level < 0.001.

dataset	norm	test error	k	p
rank 5	trace	0.8184 (0.03)	-	-
ρ=10%	k-supp	0.8036 (0.03)	3.6	-
	kp-supp	0.7831 (0.03)	1.8	7.3
rank 5	trace	0.4085 (0.03)	-	-
ρ=20%	k-supp	0.4031 (0.03)	3.1	-
	kp-supp	0.3996 (0.03)	2.0	4.7
rank 10	trace	0.6356 (0.03)	-	-
ρ=20%	k-supp	0.6284 (0.03)	4.4	-
	kp-supp	0.6270 (0.03)	2.0	4.4

Gaussian. Table 3 illustrates the results. The error is measured as $\|\text{true} - \text{predicted}\|^2/\|\text{true}\|^2$, standard deviations are shown in brackets and the mean values of k and p are selected by validation. We note that the spectral (k,p)-support norm outperforms the standard spectral k-support norm in each regime, and the improvement is statistically significant at a level < 0.01.

Real Datasets. The MovieLens 100k dataset (*http://grouplens.org/datasets/movielens/*) consists of 943 user ratings of 1,682 movies, the ratings are integers from 1 to 5, and all users have rated a minimum number of 20 films. The Jester 1 dataset (*http://goldberg.berkeley.edu/jester-data/*) consists of ratings of 24,983 users of 100 jokes, and the Jester 3 dataset consists of ratings of 34,938 users of 100 jokes, and the ratings are real values from -10 to 10.

Following [McDonald et al. 2014, Toh and Yun 2011], for MovieLens for each user we uniformly sampled $\rho = 50\%$ of available entries for training, and for Jester 1 and Jester 3 we sampled 20, respectively 8 ratings per user, using 10% for validation. We used normalized mean absolute error,

$$\text{NMAE} = \frac{\|\text{true} - \text{predicted}\|^2}{\#\text{obs.}/(r_{\max} - r_{\min})},$$

where $r_{\min}$ and $r_{\max}$ are lower and upper bounds for the ratings [Toh and Yun 2011], and we implemented a final thresholding step as in [McDonald et al. 2014].