Sparse Prediction with the k-Support Norm

Andreas Argyriou
École Centrale Paris
argyrioua@ecp.fr

Rina Foygel
Department of Statistics, Stanford University
rinafb@stanford.edu

Nathan Srebro
Toyota Technological Institute at Chicago
nati@ttic.edu

Abstract

We derive a novel norm that corresponds to the tightest convex relaxation of sparsity combined with an ℓ_2 penalty. We show that this new *k-support norm* provides a tighter relaxation than the elastic net and can thus be advantageous in in sparse prediction problems. We also bound the looseness of the elastic net, thus shedding new light on it and providing justification for its use.

1 Introduction

Regularizing with the ℓ_1 norm, when we expect a sparse solution to a regression problem, is often justified by $\|w\|_1$ being the "convex envelope" of $\|w\|_0$ (the number of non-zero coordinates of a vector $w \in \mathbb{R}^d$). That is, $\|w\|_1$ is the tightest convex lower bound on $\|w\|_0$. But we must be careful with this statement—for sparse vectors with large entries, $\|w\|_0$ can be small while $\|w\|_1$ is large. In order to discuss convex lower bounds on $\|w\|_0$, we must impose some scale constraint. A more accurate statement is that $\|w\|_1 \leq \|w\|_\infty \|w\|_0$, and so, when the magnitudes of entries in w are bounded by 1, then $\|w\|_1 \leq \|w\|_0$, and indeed it is the largest such convex lower bound. Viewed as a convex outer relaxation,

$$S_k^{(\infty)} := \left\{ w \,\middle|\, \|w\|_0 \leq k, \|w\|_\infty \leq 1 \right\} \subseteq \left\{ w \,\middle|\, \|w\|_1 \leq k \right\} .$$

Intersecting the right-hand-side with the ℓ_∞ unit ball, we get the tightest convex outer bound (convex hull) of $S_k^{(\infty)}$:

$$\left\{ w \,\middle|\, \|w\|_1 \leq k, \|w\|_\infty \leq 1 \right\} = \operatorname{conv}(S_k^{(\infty)}) .$$

However, in our view, this relationship between $\|w\|_1$ and $\|w\|_0$ yields disappointing learning guarantees, and does not appropriately capture the success of the ℓ_1 norm as a surrogate for sparsity. In particular, the sample complexity[1] of learning a linear predictor with k non-zero entries by empirical risk minimization inside this class (an NP-hard optimization problem) scales as $O(k \log d)$, but relaxing to the constraint $\|w\|_1 \leq k$ yields a sample complexity which scales as $O(k^2 \log d)$, because the sample complexity of ℓ_1-regularized learning scales quadratically with the ℓ_1 norm [11, 20].

Perhaps a better reason for the ℓ_1 norm being a good surrogate for sparsity is that, not only do we expect the magnitude of each entry of w to be bounded, but we further expect $\|w\|_2$ to be small. In a regression setting, with a vector of features x, this can be justified when $E[(x^\top w)^2]$ is bounded (a reasonable assumption) and the features are not too correlated—see, e.g. [15]. More broadly,

---

[1] We define this as the number of observations needed in order to ensure expected prediction error no more than ϵ worse than that of the best k-sparse predictor, for an arbitrary constant ϵ (that is, we suppress the dependence on ϵ and focus on the dependence on the sparsity k and dimensionality d).

1

especially in the presence of correlations, we might require this as a modeling assumption to aid in robustness and generalization. In any case, we have $\|w\|_1 \leq \|w\|_2 \sqrt{\|w\|_0}$, and so if we are interested in predictors with bounded ℓ_2 norm, we can motivate the ℓ_1 norm through the following relaxation of sparsity, where the scale is now set by the ℓ_2 norm:

$$\left\{ w \mid \|w\|_0 \leq k, \|w\|_2 \leq B \right\} \subseteq \left\{ w \mid \|w\|_1 \leq B\sqrt{k} \right\} .$$

The sample complexity when using the relaxation now scales as[2] $O(k \log d)$.

Sparse + ℓ_2 constraint.   Our starting point is then that of combining sparsity and ℓ_2 regularization, and learning a sparse predictor with small ℓ_2 norm. We are thus interested in classes of the form

$$S_k^{(2)} := \left\{ w \mid \|w\|_0 \leq k, \|w\|_2 \leq 1 \right\} .$$

As discussed above, the class $\{\|w\|_1 \leq \sqrt{k}\}$ (corresponding to the standard Lasso) provides a convex relaxation of $S_k^{(2)}$. But clearly we can get a tighter relaxation by keeping the ℓ_2 constraint:

$$\operatorname{conv}(S_k^{(2)}) \subseteq \left\{ w \mid \|w\|_1 \leq \sqrt{k}, \|w\|_2 \leq 1 \right\} \subsetneq \left\{ w \mid \|w\|_1 \leq \sqrt{k} \right\} . \tag{1}$$

Constraining (or equivalently, penalizing) both the ℓ_1 and ℓ_2 norms, as in (1), is known as the "elastic net" [5, 21] and has indeed been advocated as a better alternative to the Lasso. In this paper, we ask whether the elastic net is the *tightest* convex relaxation to sparsity plus ℓ_2 (that is, to $S_k^{(2)}$) or whether a tighter, and better, convex relaxation is possible.

A new norm.   We consider the convex hull (tightest convex outer bound) of $S_k^{(2)}$,

$$C_k := \operatorname{conv}(S_k^{(2)}) = \operatorname{conv} \left\{ w \mid \|w\|_0 \leq k, \|w\|_2 \leq 1 \right\} . \tag{2}$$

We study the gauge function associated with this convex set, that is, the norm whose unit ball is given by (2), which we call the *k-support norm*. We show that, for $k > 1$, this is indeed a tighter convex relaxation than the elastic net (that is, both inequalities in (1) are in fact strict inequalities), and is therefore a better convex constraint than the elastic net when seeking a sparse, low ℓ_2-norm linear predictor. We thus advocate using it as a replacement for the elastic net.

However, we also show that the gap between the elastic net and the k-support norm is at most a factor of $\sqrt{2}$, corresponding to a factor of two difference in the sample complexity. Thus, our work can also be interpreted as justifying the use of the elastic net, viewing it as a fairly good approximation to the tightest possible convex relaxation of sparsity intersected with an ℓ_2 constraint. Still, even a factor of two should not necessarily be ignored and, as we show in our experiments, using the tighter k-support norm can indeed be beneficial.

To better understand the k-support norm, we show in Section 2 that it can also be described as the group lasso with overlaps norm [10] corresponding to all $\binom{d}{k}$ subsets of k features. Despite the exponential number of groups in this description, we show that the k-support norm can be calculated efficiently in time $O(d \log d)$ and that its dual is given simply by the ℓ_2 norm of the k largest entries. We also provide efficient first-order optimization algorithms for learning with the k-support norm.

Related Work   In many learning problems of interest, Lasso has been observed to shrink too many of the variables of w to zero. In particular, in many applications, when a group of variables is highly correlated, the Lasso may prefer a sparse solution, but we might gain more predictive accuracy by including all the correlated variables in our model. These drawbacks have recently motivated the use of various other regularization methods, such as the *elastic net* [21], which penalizes the regression coefficients w with a combination of ℓ_1 and ℓ_2 norms:

$$\min \left\{ \frac{1}{2} \|Xw - y\|^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2 : w \in \mathbb{R}^d \right\} , \tag{3}$$

---

[2]More precisely, the sample complexity is $O(B^2 k \log d)$, where the dependence on B^2 is to be expected. Note that if feature vectors are ℓ_∞-bounded (i.e. individual features are bounded), the sample complexity when using only $\|w\|_2 \leq B$ (without a sparsity or ℓ_1 constraint) scales as $O(B^2 d)$. That is, even after identifying the correct support, we still need a sample complexity that scales with B^2.

2

where for a sample of size n, $y \in \mathbb{R}^n$ is the vector of response values, and $X \in \mathbb{R}^{n \times d}$ is a matrix with column j containing the values of feature j.

The elastic net can be viewed as a trade-off between ℓ_1 regularization (the Lasso) and ℓ_2 regularization (Ridge regression [9]), depending on the relative values of λ_1 and λ_2. In particular, when $\lambda_2 = 0$, (3) is equivalent to the Lasso. This method, and the other methods discussed below, have been observed to significantly outperform Lasso in many real applications.

The pairwise elastic net (PEN) [13] is a penalty function that accounts for similarity among features:

$$\|w\|_R^{PEN} = \|w\|_2^2 + \|w\|_1^2 - |w|^\top R |w| \ ,$$

where $R \in [0,1]^{p \times p}$ is a matrix with R_{jk} measuring similarity between features X_j and X_k. The trace Lasso [6] is a second method proposed to handle correlations within X, defined by

$$\|w\|_X^{trace} = \|X \mathrm{diag}(w)\|_* \ ,$$

where $\| \cdot \|_*$ denotes the matrix trace-norm (the sum of the singular values) and promotes a low-rank solution. If the features are orthogonal, then both the PEN and the Trace Lasso are equivalent to the Lasso. If the features are all identical, then both penalties are equivalent to Ridge regression (penalizing $\|w\|_2$). Another existing penalty is OSCAR [3], given by

$$\|w\|_c^{OSCAR} = \|w\|_1 + c \sum_{j<k} \max\{|w_j|, |w_k|\} \ .$$

Like the elastic net, each one of these three methods also "prefers" averaging similar features over selecting a single feature.

2 The k-Support Norm

One argument for the elastic net has been the flexibility of tuning the cardinality k of the regression vector w. Thus, when groups of correlated variables are present, a larger k may be learned, which corresponds to a higher λ_2 in (3). A more natural way to obtain such an effect of tuning the cardinality is to consider the convex hull of cardinality k vectors,

$$C_k = \mathrm{conv}(S_k^{(2)}) = \mathrm{conv}\{w \in \mathbb{R}^d \,|\, \|w\|_0 \leq k, \|w\|_2 \leq 1\}.$$

Clearly the sets C_k are nested, and C_1 and C_d are the unit balls for the ℓ_1 and ℓ_2 norms, respectively. Consequently we define the k-*support norm* as the norm whose unit ball equals C_k (the gauge function associated with the C_k ball).[3] An equivalent definition is the following variational formula:

Definition 2.1. *Let $k \in \{1, \ldots, d\}$. The k-support norm $\| \cdot \|_k^{sp}$ is defined, for every $w \in \mathbb{R}^d$, as*

$$\|w\|_k^{sp} := \min\left\{ \sum_{I \in \mathcal{G}_k} \|v_I\|_2 : \mathrm{supp}(v_I) \subseteq I, \sum_{I \in \mathcal{G}_k} v_I = w \right\},$$

where $\mathcal{G}_k$ denotes the set of all subsets of $\{1, \ldots, d\}$ of cardinality at most k.

The equivalence is immediate by rewriting $v_I = \mu_I z_I$ in the above definition, where $\mu_I \geq 0, z_I \in C_k, \forall I \in \mathcal{G}_k, \sum_{I \in \mathcal{G}_k} \mu_I = 1$. In addition, this immediately implies that $\| \cdot \|_k^{sp}$ is indeed a norm. In fact, the k-support norm is equivalent to the norm used by the group lasso with overlaps [10], when the set of overlapping groups is chosen to be $\mathcal{G}_k$ (however, the group lasso has traditionally been used for applications with some specific known group structure, unlike the case considered here).

Although the variational definition 2.1 is not amenable to computation because of the exponential growth of the set of groups $\mathcal{G}_k$, the k-support norm is computationally very tractable, with an $O(d \log d)$ algorithm described in Section 2.2.

As already mentioned, $\| \cdot \|_1^{sp} = \| \cdot \|_1$ and $\| \cdot \|_d^{sp} = \| \cdot \|_2$. The unit ball of this new norm in $\mathbb{R}^3$ for $k = 2$ is depicted in Figure 1. We immediately notice several differences between this unit ball and the elastic net unit ball. For example, at points with cardinality k and ℓ_2 norm equal to 1, the k-support norm is not differentiable, but unlike the ℓ_1 or elastic-net norm, it *is* differentiable at points with cardinality less than k. Thus, the k-support norm is less "biased" towards sparse vectors than the elastic net and the ℓ_1 norm.

---

[3]The gauge function $\gamma_{C_k} : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is defined as $\gamma_{C_k}(x) = \inf\{\lambda \in \mathbb{R}_+ : x \in \lambda C_k\}$.

3

Figure 1: Unit ball of the 2-support norm (left) and of the elastic net (right) on $\mathbb{R}^3$.

2.1 The Dual Norm

It is interesting and useful to compute the dual of the k-support norm. For $w \in \mathbb{R}^d$, denote $|w|$ for the vector of absolute values, and $w_i^{\downarrow}$ for the i-th largest element of w [2]. We have

$$\|u\|_k^{sp^*} = \max\left\{\langle w, u \rangle : \|w\|_k^{sp} \le 1\right\} = \max\left\{\left(\sum_{i \in I} u_i^2\right)^{\frac{1}{2}} : I \in \mathcal{G}_k\right\} = \left(\sum_{i=1}^{k}(|u|_i^{\downarrow})^2\right)^{\frac{1}{2}} =: \|u\|_{(k)}^{(2)} .$$

This is the ℓ_2-norm of the largest k entries in u, and is known as the 2-k *symmetric gauge norm* [2].

Not surprisingly, this dual norm interpolates between the ℓ_2 norm (when $k = d$ and all entries are taken) and the ℓ_∞ norm (when $k = 1$ and only the largest entry is taken). This parallels the interpolation of the k-support norm between the ℓ_1 and ℓ_2 norms.

2.2 Computation of the Norm

In this section, we derive an alternative formula for the k-support norm, which leads to computation of the value of the norm in $O(d \log d)$ steps.

Proposition 2.1. *For every* $w \in \mathbb{R}^d$, $\|w\|_k^{sp} = \left(\sum_{i=1}^{k-r-1}(|w|_i^{\downarrow})^2 + \frac{1}{r+1}\left(\sum_{i=k-r}^{d}|w|_i^{\downarrow}\right)^2\right)^{\frac{1}{2}}$,

where, letting $|w|_0^{\downarrow}$ *denote* $+\infty$, r *is the unique integer in* $\{0, \ldots, k-1\}$ *satisfying*

$$|w|_{k-r-1}^{\downarrow} > \frac{1}{r+1}\sum_{i=k-r}^{d}|w|_i^{\downarrow} \ge |w|_{k-r}^{\downarrow} . \tag{4}$$

This result shows that $\|\cdot\|_k^{sp}$ trades off between the ℓ_1 and ℓ_2 norms in a way that favors sparse vectors but allows for cardinality larger than k. It combines the uniform shrinkage of an ℓ_2 penalty for the largest components, with the sparse shrinkage of an ℓ_1 penalty for the smallest components.

Proof of Proposition 2.1. We will use the inequality $\langle w, u \rangle \le \langle w^{\downarrow}, u^{\downarrow} \rangle$ [7]. We have

$$\frac{1}{2}(\|w\|_k^{sp})^2 = \max\left\{\langle u, w \rangle - \frac{1}{2}(\|u\|_{(k)}^{(2)})^2 : u \in \mathbb{R}^d\right\} = \max\left\{\sum_{i=1}^{d}\alpha_i|w|_i^{\downarrow} - \frac{1}{2}\sum_{i=1}^{k}\alpha_i^2 :\right.$$

$$\left.\alpha_1 \ge \cdots \ge \alpha_d \ge 0\right\} = \max\left\{\sum_{i=1}^{k-1}\alpha_i|w|_i^{\downarrow} + \alpha_k\sum_{i=k}^{d}|w|_i^{\downarrow} - \frac{1}{2}\sum_{i=1}^{k}\alpha_i^2 : \alpha_1 \ge \cdots \ge \alpha_k \ge 0\right\} .$$

Let $A_r := \sum_{i=k-r}^{d}|w|_i^{\downarrow}$ for $r \in \{0, \ldots, k-1\}$. If $A_0 < |w|_{k-1}^{\downarrow}$ then the solution α is given by $\alpha_i = |w|_i^{\downarrow}$ for $i = 1, \ldots, (k-1)$, $\alpha_i = A_0$ for $i = k, \ldots, d$. If $A_0 \ge |w|_{k-1}^{\downarrow}$ then the optimal α_k, α_{k-1} lie between $|w|_{k-1}^{\downarrow}$ and A_0, and have to be equal. So, the maximization becomes

$$\max\left\{\sum_{i=1}^{k-2}\alpha_i|w|_i^{\downarrow} - \frac{1}{2}\sum_{i=1}^{k-2}\alpha_i^2 + A_1\alpha_{k-1} - \alpha_{k-1}^2 : \alpha_1 \ge \cdots \ge \alpha_{k-1} \ge 0\right\} .$$

4

If $A_0 \geq |w|_{k-1}^{\downarrow}$ and $|w|_{k-2}^{\downarrow} > \frac{A_1}{2}$ then the solution is $\alpha_i = |w|_i^{\downarrow}$ for $i = 1, \ldots, (k-2)$, $\alpha_i = \frac{A_1}{2}$ for $i = (k-1), \ldots, d$. Otherwise we proceed as before and continue this process. At stage r the process terminates if $A_0 \geq |w|_{k-1}^{\downarrow}, \ldots, \frac{A_{r-1}}{r} \geq |w|_{k-r}^{\downarrow}, \frac{A_r}{r+1} < |w|_{k-r-1}^{\downarrow}$ and all but the last two inequalities are redundant. Hence the condition can be rewritten as (4). One optimal solution is $\alpha_i = |w|_i^{\downarrow}$ for $i = 1, \ldots, k-r-1$, $\alpha_i = \frac{A_r}{r+1}$ for $i = k-r, \ldots, d$. This proves the claim. ∎

2.3   Learning with the k-support norm

We thus propose using learning rules with k-support norm regularization. These are appropriate when we would like to learn a sparse predictor that also has low ℓ_2 norm, and are especially relevant when features might be correlated (that is, in almost all learning tasks) but the correlation structure is not known in advance. E.g., for squared error regression problems we have:

$$\min \left\{ \frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \left(\|w\|_k^{sp} \right)^2 : w \in \mathbb{R}^d \right\} \tag{5}$$

with $\lambda > 0$ a regularization parameter and $k \in \{1, \ldots, d\}$ also a parameter to be tuned. As typical in regularization-based methods, both λ and k can be selected by cross validation [8]. Despite the relationship to $S_k^{(2)}$, the parameter k does not necessarily correspond to the sparsity of the actual minimizer of (5), and should be chosen via cross-validation rather than set to the desired sparsity.

3   Relation to the Elastic Net

Recall that the elastic net with penalty parameters λ_1 and λ_2 selects a vector of coefficients given by

$$\arg \min \left\{ \frac{1}{2} \|Xw - y\|^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2 \right\}. \tag{6}$$

For ease of comparison with the k-support norm, we first show that the set of optimal solutions for the elastic net, when the parameters are varied, is the same as for the norm

$$\|w\|_k^{el} := \max \left\{ \|w\|_2, \|w\|_1/\sqrt{k} \right\},$$

when $k \in [1, d]$, corresponding to the unit ball in (1) (note that k is not necessarily an integer). To see this, let $\hat{w}$ be a solution to (6), and let $k := (\|\hat{w}\|_1/\|\hat{w}\|_2)^2 \in [1, d]$. Now for any $w \neq \hat{w}$, if $\|w\|_k^{el} \leq \|\hat{w}\|_k^{el}$, then $\|w\|_p \leq \|\hat{w}\|_p$ for $p = 1, 2$. Since $\hat{w}$ is a solution to (6), therefore, $\|Xw - y\|_2^2 \geq \|X\hat{w} - y\|_2^2$. This proves that, for some constraint parameter B,

$$\hat{w} = \arg \min \left\{ \frac{1}{n} \|Xw - y\|_2^2 : \|w\|_k^{el} \leq B \right\}.$$

Like the k-support norm, the elastic net interpolates between the ℓ_1 and ℓ_2 norms. In fact, when k is an integer, any k-sparse unit vector $w \in \mathbb{R}^d$ must lie in the unit ball of $\|\cdot\|_k^{el}$. Since the k-support norm gives the convex hull of all k-sparse unit vectors, this immediately implies that

$$\|w\|_k^{el} \leq \|w\|_k^{sp} \qquad \forall\, w \in \mathbb{R}^d.$$

The two norms are not equal, however. The difference between the two is illustrated in Figure 1, where we see that the k-support norm is more "rounded".

To see an example where the two norms are not equal, we set $d = 1 + k^2$ for some large k, and let $w = (k^{1.5}, 1, 1, \ldots, 1)^\top \in \mathbb{R}^d$. Then

$$\|w\|_k^{el} = \max \left\{ \sqrt{k^3 + k^2}, \frac{k^{1.5} + k^2}{\sqrt{k}} \right\} = k^{1.5} \left(1 + \frac{1}{\sqrt{k}} \right).$$

Taking $u = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2k}}, \frac{1}{\sqrt{2k}}, \ldots, \frac{1}{\sqrt{2k}})^\top$, we have $\|u\|_{(k)}^{(2)} < 1$, and recalling this norm is dual to the k-support norm:

$$\|w\|_k^{sp} > \langle w, u \rangle = \frac{k^{1.5}}{\sqrt{2}} + k^2 \cdot \frac{1}{\sqrt{2k}} = \sqrt{2} \cdot k^{1.5}.$$

In this example, we see that the two norms can differ by as much as a factor of $\sqrt{2}$. We now show that this is actually the most by which they can differ.

5

Proposition 3.1. $\| \cdot \|_k^{el} \leq \| \cdot \|_k^{sp} < \sqrt{2}\, \| \cdot \|_k^{el}$.

Proof. We show that these bounds hold in the duals of the two norms. First, since $\| \cdot \|_k^{el}$ is a maximum over the ℓ_1 and ℓ_2 norms, its dual is given by

$$\|u\|_k^{(el)^*} := \inf_{a \in \mathbb{R}^d} \left\{ \|a\|_2 + \sqrt{k} \cdot \|u - a\|_\infty \right\}$$

Now take any $u \in \mathbb{R}^d$. First we show $\|u\|_{(k)}^{(2)} \leq \|u\|_k^{(el)^*}$. Without loss of generality, we take $u_1 \geq \cdots \geq u_d \geq 0$. For any $a \in \mathbb{R}^d$,

$$\|u\|_{(k)}^{(2)} = \|u_{1:k}\|_2 \leq \|a_{1:k}\|_2 + \|u_{1:k} - a_{1:k}\|_2 \leq \|a\|_2 + \sqrt{k}\|u - a\|_\infty\,.$$

Finally, we show that $\|u\|_k^{(el)^*} < \sqrt{2}\,\|u\|_{(k)}^{(2)}$. Let $a = (u_1 - u_{k+1}, \ldots, u_k - u_{k+1}, 0, \ldots, 0)^\top$. Then

$$\|u\|_k^{(el)^*} \leq \|a\|_2 + \sqrt{k} \cdot \|u - a\|_\infty = \sqrt{\sum_{i=1}^{k}(u_i - u_{k+1})^2} + \sqrt{k} \cdot |u_{k+1}|$$

$$\leq \sqrt{\sum_{i=1}^{k}(u_i^2 - u_{k+1}^2)} + \sqrt{k\,u_{k+1}^2} \leq \sqrt{2} \cdot \sqrt{\sum_{i=1}^{k}(u_i^2 - u_{k+1}^2) + k\,u_{k+1}^2} = \sqrt{2}\,\|u\|_{(k)}^{(2)}\,.$$

Furthermore, this yields a strict inequality, because if $u_1 > u_{k+1}$, the next-to-last inequality is strict, while if $u_1 = \cdots = u_{k+1}$, then the last inequality is strict. ∎

4 Optimization

Solving the optimization problem (5) efficiently can be done with a first-order proximal algorithm. Proximal methods – see [1, 4, 14, 18, 19] and references therein – are used to solve composite problems of the form $\min\{f(x) + \omega(x) : x \in \mathbb{R}^d\}$, where the loss function $f(x)$ and the regularizer $\omega(x)$ are convex functions, and f is smooth with an L-Lipschitz gradient. These methods require fast computation of the gradient ∇f and the *proximity operator*

$$\mathrm{prox}_\omega(x) := \mathrm{argmin}\left\{ \frac{1}{2}\|u - x\|^2 + \omega(u) : u \in \mathbb{R}^d \right\}\,.$$

To obtain a proximal method for k-support regularization, it suffices to compute the proximity map of $g = \frac{1}{2\beta}(\| \cdot \|_k^{sp})^2$, for any $\beta > 0$ (in particular, for problem (5) β corresponds to $\frac{L}{\lambda}$). This computation can be done in $O(d(k + \log d))$ steps with Algorithm 1.

Algorithm 1 Computation of the proximity operator.

Input $v \in \mathbb{R}^d$
Output $q = \mathrm{prox}_{\frac{1}{2\beta}(\|\cdot\|_k^{sp})^2}(v)$
Find $r \in \{0, \ldots, k-1\}$, $\ell \in \{k, \ldots, d\}$ such that

$$\frac{1}{\beta+1}z_{k-r-1} > \frac{T_{r,\ell}}{\ell - k + (\beta+1)r + \beta + 1} \geq \frac{1}{\beta+1}z_{k-r} \tag{7}$$

$$z_\ell > \frac{T_{r,\ell}}{\ell - k + (\beta+1)r + \beta + 1} \geq z_{\ell+1} \tag{8}$$

where $z := |v|^\downarrow$, $z_0 := +\infty$, $z_{d+1} := -\infty$, $T_{r,\ell} := \sum\limits_{i=k-r}^{\ell} z_i$

$$q_i \leftarrow \begin{cases} \frac{\beta}{\beta+1}z_i & \text{if } i = 1, \ldots, k - r - 1 \\ z_i - \frac{T_{r,\ell}}{\ell - k + (\beta+1)r + \beta + 1} & \text{if } i = k - r, \ldots, \ell \\ 0 & \text{if } i = \ell + 1, \ldots, d \end{cases}$$

Reorder and change signs of q to conform with v

6
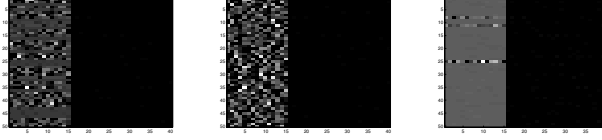
Figure 2: Solutions learned for the synthetic data. Left to right: k-support, Lasso and elastic net.

Proof of Correctness of Algorithm 1. Since the support-norm is sign and permutation invariant, $\mathrm{prox}_g(v)$ has the same ordering and signs as v. Hence, without loss of generality, we may assume that $v_1 \geq \cdots \geq v_d \geq 0$ and require that $q_1 \geq \cdots \geq q_d \geq 0$, which follows from inequality (7) and the fact that z is ordered.

Now, $q = \mathrm{prox}_g(v)$ is equivalent to $\beta z - \beta q = \beta v - \beta q \in \partial \frac{1}{2}(\| \cdot \|_k^{sp})^2(q)$. It suffices to show that, for $w = q$, $\beta z - \beta q$ is an optimal α in the proof of Proposition 2.1. Indeed, A_r corresponds to

$$\sum_{i=k-r}^{d} q_i = \sum_{i=k-r}^{\ell} \left(z_i - \frac{T_{r,\ell}}{\ell-k+(\beta+1)r+\beta+1} \right) = T_{r,\ell} - \frac{(\ell-k+r+1)T_{r,\ell}}{\ell-k+(\beta+1)r+\beta+1} = (r+1)\frac{\beta\, T_{r,\ell}}{\ell-k+(\beta+1)r+\beta+1}$$

and (4) is equivalent to condition (7). For $i \leq k-r-1$, we have $\beta z_i - \beta q_i = q_i$. For $k-r \leq i \leq \ell$, we have $\beta z_i - \beta q_i = \frac{1}{r+1}A_r$. For $i \geq \ell+1$, since $q_i = 0$, we only need $\beta z_i - \beta q_i \leq \frac{1}{r+1}A_r$, which is true by (8). $\blacksquare$

We can now apply a standard accelerated proximal method, such as FISTA [1], to (5), at each iteration using the gradient of the loss and performing a prox step using Algorithm 1. The FISTA guarantee ensures us that, with appropriate step sizes, after T such iterations, we have:

$$\frac{1}{2}\|Xw_T - y\|^2 + \frac{\lambda}{2}\,(\|w_T\|_k^{sp})^2 \leq \left(\frac{1}{2}\|Xw^* - y\|^2 + \frac{\lambda}{2}\,(\|w^*\|_k^{sp})^2 \right) + \frac{2L\|w^* - w_1\|^2}{(T+1)^2} \ .$$

5   Empirical Comparisons

Our theoretical analysis indicates that the k-support norm and the elastic net differ by at most a factor of $\sqrt{2}$, corresponding to at most a factor of two difference in their sample complexities and generalization guarantees. We thus do not expect huge differences between their actual performances, but would still like to see whether the tighter relaxation of the k-support norm does yield some gains.

Synthetic Data   For the first simulation we follow [21, Sec. 5, example 4]. In this experimental protocol, the target (oracle) vector equals $w^* = (\underbrace{3,\ldots,3}_{15}, \underbrace{0\ldots,0}_{25})$, with $y = (w^*)^\top x + \mathcal{N}(0,1)$.

The input data X were generated from a normal distribution such that components $1,\ldots,5$ have the same random mean $Z_1 \sim \mathcal{N}(0,1)$, components $6,\ldots,10$ have mean $Z_2 \sim \mathcal{N}(0,1)$ and components $11,\ldots,15$ have mean $Z_3 \sim \mathcal{N}(0,1)$. A total of 50 data sets were created in this way, each containing 50 training points, 50 validation points and 350 test points. The goal is to achieve good prediction performance on the test data.

We compared the k-support norm with Lasso and the elastic net. We considered the ranges $k = \{1,\ldots,d\}$ for k-support norm regularization, $\lambda = 10^i$, $i = \{-15,\ldots,5\}$, for the regularization parameter of Lasso and k-support regularization and the same range for the λ_1, λ_2 of the elastic net. For each method, the optimal set of parameters was selected based on mean squared error on the validation set. The error reported in Table 5 is the mean squared error with respect to the oracle w^*, namely $MSE = (\hat{w} - w^*)^\top V (\hat{w} - w^*)$, where V is the population covariance matrix of X_{test}.

To further illustrate the effect of the k-support norm, in Figure 5 we show the coefficients learned by each method, in absolute value. For each image, one row corresponds to the w learned for one of the 50 data sets. Whereas all three methods distinguish the 15 relevant variables, the elastic net result varies less within these variables.

South African Heart Data   This is a classification task which has been used in [8]. There are 9 variables and 462 examples, and the response is presence/absence of coronary heart disease. We

7

Table 1: Mean squared errors and classification accuracy for the synthetic data (median over 50 repetition), SA heart data (median over 50 replications) and for the "20 newsgroups" data set. (SE = standard error)

	Synthetic	Heart		Newsgroups	
Method	MSE (SE)	MSE (SE)	Accuracy (SE)	MSE	Accuracy
Lasso	0.2685 (0.02)	0.18 (0.005)	66.41 (0.53)	0.70	73.02
Elastic net	0.2274 (0.02)	0.18 (0.005)	66.41 (0.53)	0.70	73.02
k-support	**0.2143** (0.02)	0.18 (0.005)	66.41 (0.53)	**0.69**	**73.40**

normalized the data so that each predictor variable has zero mean and unit variance. We then split the data 50 times randomly into training, validation, and test sets of sizes 400, 30, and 32 respectively. For each method, parameters were selected using the validation data. In Tables 5, we report the MSE and accuracy of each method on the test data. We observe that all three methods have identical performance.

20 Newsgroups This is a binary classification version of 20 newsgroups created in [12] which can be found in the LIBSVM data repository.[4] The positive class consists of the 10 groups with names of form sci.*, comp.*, or misc.forsale and the negative class consists of the other 10 groups. To reduce the number of features, we removed the words which appear in less than 3 documents. We randomly split the data into a training, a validation and a test set of sizes 14000,1000 and 4996, respectively. We report MSE and accuracy on the test data in Table 5. We found that k-support regularization gave improved prediction accuracy over both other methods.[5]

6   Summary

We introduced the k-support norm as the tightest convex relaxation of sparsity plus ℓ_2 regularization, and showed that it is tighter than the elastic net by exactly a factor of $\sqrt{2}$. In our view, this sheds light on the elastic net as a close approximation to this tightest possible convex relaxation, and motivates using the k-support norm when a tighter relaxation is sought. This is also demonstrated in our empirical results.

We note that the k-support norm has better prediction properties, but not necessarily better sparsity-inducing properties, as evident from its more rounded unit ball. It is well understood that there is often a tradeoff between sparsity and good prediction, and that even if the population optimal predictor is sparse, a denser predictor often yields better predictive performance [3, 10, 21]. For example, in the presence of correlated features, it is often beneficial to include several highly correlated features rather than a single representative feature. This is exactly the behavior encouraged by ℓ_2 norm regularization, and the elastic net is already known to yield less sparse (but more predictive) solutions. The k-support norm goes a step further in this direction, often yielding solutions that are even less sparse (but more predictive) compared to the elastic net.

Nevertheless, it is interesting to consider whether compressed sensing results, where ℓ_1 regularization is of course central, can be refined by using the k-support norm, which might be able to handle more correlation structure within the set of features.

Acknowledgements The construction showing that the gap between the elastic net and the k-overlap norm can be as large as $\sqrt{2}$ is due to joint work with Ohad Shamir. Rina Foygel was supported by NSF grant DMS-1203762.

References

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

[2] R. Bhatia. *Matrix Analysis*. Graduate Texts in Mathematics. Springer, 1997.

---

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[5] Regarding other sparse prediction methods, we did not manage to compare with OSCAR, due to memory limitations, or to PEN or trace Lasso, which do not have code available online.

8

[3] H.D. Bondell and B.J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123, 2008.

[4] P.L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2006.

[5] C. De Mol, E. De Vito, and L. Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–230, 2009.

[6] E. Grave, G. R. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, 2011.

[7] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 1934.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag Series in Statistics, 2001.

[9] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, pages 55–67, 1970.

[10] L. Jacob, G. Obozinski, and J.P. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440. ACM, 2009.

[11] S.M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, volume 22, 2008.

[12] S. S. Keerthi and D. DeCoste. A modified finite Newton method for fast solution of large scale linear SVMs. *Journal of Machine Learning Research*, 6:341–361, 2005.

[13] A. Lorbert, D. Eis, V. Kostina, D.M. Blei, and P.J. Ramadge. Exploiting covariate similarity in sparse regression via the pairwise elastic net. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

[14] Y. Nesterov. *Gradient methods for minimizing composite objective function*. CORE, 2007.

[15] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low-noise and fast rates. In *Advances in Neural Information Processing Systems 23*, 2010.

[16] T. Suzuki and R. Tomioka. SpicyMKL: a fast algorithm for multiple kernel learning with thousands of kernels. *Machine learning*, pages 1–32, 2011.

[17] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 58(1):267–288, 1996.

[18] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Preprint, 2008.

[19] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.

[20] T. Zhang. Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research*, 2:527–550, 2002.

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

9