Non-Convex Feature Learning via $\ell_{p,\infty}$ Operator

Deguang Kong and Chris Ding
Department of Computer Science & Engineering,
University of Texas, Arlington,
500 UTA Blvd, Arlington, TX 76010
doogkong@gmail.com; chqding@uta.edu

Abstract

We present a feature selection method for solving sparse regularization problem, which has a composite regularization of ℓ_p norm and ℓ_∞ norm. We use proximal gradient method to solve this $\ell_{p,\infty}$ operator problem, where a simple but efficient algorithm is designed to minimize a relatively simple objective function, which contains a vector of ℓ_2 norm and ℓ_∞ norm. Proposed method brings some insight for solving sparsity-favoring norm, and extensive experiments are conducted to characterize the effect of varying p and to compare with other approaches on real world multi-class and multi-label datasets.

Introduction

Feature selection is a process of selecting a subset of relevant features from all the original features for robust classification, clustering and other learning tasks. Feature selection plays an important role in machine learning. A large number of feature selection approaches have been developed in literature. In general, they can be divided into two categories. (C1) Filter methods (Langley 1994), (C2) Wrapper methods(Kohavi and John 1997), *etc*.

In recent years, sparsity regularization has been widely investigated and applied into multi-task learning and feature selection studies, where $\ell_{1,\infty}$ variable selection/projection have been proposed and well investigated in (Masaeli, Fung, and Dy 2010), (Turlach, Venablesy, and Wright 2005),(Tropp et al. 2006), (Quattoni, Collins, and Darrell 2008), (Schmidt et al. 2008) and (Liu, Palatucci, and Zhang 2009). One general conclusion (Masaeli, Fung, and Dy 2010) is that $\ell_{1,\infty}$ regularization usually performs significant better for classification tasks than both independent ℓ_1 and independent ℓ_2 regularizations. In other words, $\ell_{1,\infty}$ related optimization is a well behaved algorithm.

We note regularization with non-convex ℓ_p penalty has gained increasing interest (e.g., smoothly clipped absolute deviation method (Fan and Li 2001), bridge regularization (Hastie, Tibshirani, and Friedman 2001)). As considering ℓ_p-norm with p smaller than 1, the penalty is not differentiable as soon as its argument vanishes. To deal with this issue, (Huang, Horowitz, and Ma 2008) considered a parameterized differentiable approximation of the ℓ_p penalty and use gradient descent to solve it; (Chartrand and Staneva 2008), (Kong and Ding 2013) use an iteratively re-weighted least square algorithm.

Above observations motivate us to combine ℓ_∞ with ℓ_p penalty. In this paper, we consider more general $\ell_{p,\infty}$ type operator where $p < 1$. This model is a generalization of the $\ell_{1,\infty}$ group lasso with enforced sparsity of the solution. The attractive property of $\ell_{p,\infty}$ operator is that, at different p (although it is not convex at $p < 1$), it gives interesting property to approach the real number of non-zero features/variables, which is the desired goal in feature/variable selection tasks.

The algorithms designed for ℓ_p penalty (e.g., (Chartrand and Staneva 2008), (Kong, Zhang, and Ding 2013)) cannot be simply modified to deal with ℓ_∞ problem because ℓ_∞ is discontinuous, which is harder to deal with than ℓ_1-norm which is continuous (although its derivative is non-continuous). The challenge of our problem is to solve the projection for $\ell_{p,\infty}$ operator. Another contribution of this paper is that, a very simple and efficient algorithm is derived to solve $\ell_{p,\infty}$ operator with rigorous analysis. We give the structure of the optimal solution for $\ell_{p,\infty}$ projection, to our knowledge, which has not been emphasized before. Our algorithm also has very clear differences with the other methods used for $\ell_{1,\infty}$ computation, like blockwise coordinate descent method in (Liu, Palatucci, and Zhang 2009), double coordinate descent method in (Zhang 2011) and the interior-point method in (Berwin A. Turlach 2005).

We note (Vogt and Roth 2012) studied convex group lasso with $p \geq 1$ for coupling multiple regression tasks. They studied loss function $f(\mathbf{B})$ w.r.t coefficient $\mathbf{B} = (\beta_1, \cdot \cdot \cdot, \beta_J) = \sum_{j=1}^{J} ||\beta_j||_p$ for J tasks. The p in their model is very different from our model. It also seems their paper is not on feature selection: they set the parameter k so that sparsity of $\mathbf{B}$ is a fixed value (this is *flat* sparsity, different from the *structured* sparsity) and give the prediction error. In our feature selection of Eq.(2), we set λ such that entire rows of $\mathbf{W}$ to be zero (*structured* sparsity).

To summarize, the main contributions of our paper are in three-fold. **(1)** We propose to use $\ell_{p,\infty}$ operator (i.e., a composite regularization of ℓ_p-norm and ℓ_∞ norm) for feature selection. **(2)** An efficient algorithm is presented to solve $\ell_{p,\infty}$ proximal operator (i.e., a simple function containing a vector of ℓ_2 norm and ℓ_∞ norm) with rigorous analysis. **(3)** The experiment results suggest that our algorithm is well-

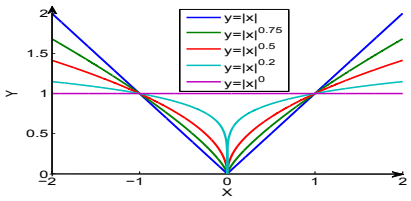


Figure 1: The behaviors of $y = x^p$ when $p = \{1, 0.75, 0.5, 0.2, 0\}$.

behaved for smaller p values on real-world multi-class and multi-label classification tasks.

Feature Selection via $\ell_{p,\infty}$ operator

Notations Let $\mathbf{X} \in \Re^{d\times n} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ denote the matrix of input data of d-dimension over n samples, and $\mathbf{Y} \in \Re^{c\times n} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n)$ denote the matrix of output data for c outputs. In a general multi-class learning task, we use a linear model for the k-th class: $\mathbf{y}_k = \mathbf{W}^T\mathbf{x}_k + \epsilon_k$, $1 \le k \le c$, where $\mathbf{W} \in \Re^{d\times c}$ is the regression coefficient for all the c classes, and ϵ_k is the Gaussian noise.

Proposed feature selection using $\ell_{p,\infty}$

In this paper, we consider the following multi-class sparse regression problem by using $\ell_{p,\infty}$ operator penalization,

$$\min_{\mathbf{W}\in\Re^{d\times c}} J(\mathbf{W}) = f(\mathbf{W}) + \phi_\lambda(\mathbf{W}), \tag{1}$$

where $f(\mathbf{W})$ is a smooth/non-smooth convex loss function, e.g., least square loss, and

$$\phi_\lambda(\mathbf{W}) = \lambda(||\mathbf{W}||_{p,\infty})^p = \lambda\sum_{j=1}^{d}(\max_{1\leqslant k\leqslant c}|\mathbf{W}_{jk}|)^p, \tag{2}$$

is $\ell_{p,\infty}$ penalty, and $\lambda \ge 0$ is regularization parameters. Note $\ell_{p,\infty}$-norm of $\mathbf{W}$ is defined as,

$$||\mathbf{W}||_{p,\infty} = \Big(\sum_{j=1}^{d}(\max_{1\leqslant k\leqslant c}|\mathbf{W}_{jk}|)^p\Big)^{\frac{1}{p}}.$$

In $(||\mathbf{W}||_{p,\infty})^p$ of Eq.(2), p-th root is dropped and it is not a norm any more. The penalty of Eq.(2) is a more general case of $\ell_{1,\infty}$ penalty (Boyd and Vandenberghe 2004). When $p = 1$, it is $\ell_{1,\infty}$-norm penalty. $\ell_{p,\infty}$ operator is a convex relaxation of a pseudo-norm which counts the number of non-zero row in $\mathbf{W}$. When $p = 0$, it counts the number of non-zero rows if we assume $0^0 = 0$. Usually p is set to $0 \le p \le 1$.

Motivation of our method

Let $f(\mathbf{W}^j) = \max_{1\leqslant k\leqslant c}|\mathbf{W}_{jk}|$. Thus the behaviors of $\sum_j f(\mathbf{W}^j)^p$ is determined by the behaviors of $|x|^p$, where $|x| = f(\mathbf{W}^j)$ is a scalar. One can say the different behaviors of function x^p when p is set to different values in Fig.(1).

$$p \to 0, \quad |x|^p \to 1.$$

Thus, when p is small, the behaviors of $\sum_j f(\mathbf{W}^j)^p \to \sum_j \delta(f(\mathbf{W}^j))$, where δ is a 0-1 function, $\delta(a) = 1$ if $a \neq 0$, else $\delta(a) = 0$. It is the desired goal for feature/variable selection because the number of non-zero rows can be more accurately estimated by $\sum_j f(\mathbf{W}^j)^p$ when p is small. Note $\mathbf{W} = (\mathbf{w}^1, \mathbf{w}^2, \cdots, \mathbf{w}^d)^T$, where $\mathbf{w}^d \in \Re^{c\times 1}$ is regression coefficient for feature d across all classes. The mixed ℓ_p/ℓ_∞ penalty in Eq.(2) is used to set all the coefficient in each dimension (i.e., $\mathbf{w}^d$) to zeros or non-zero values, for variable selection purpose.

Overview of computational algorithm

Proximal gradient method (a.k.a FISTA method)(Nesterov 2004; Beck and Teboulle 2009b) is widely used to solve $\min_{\mathbf{W}} f(\mathbf{W})+\phi_\lambda(\mathbf{W})$ problem. Here we adopt this method to solve Eq.(1) due to its fast convergence. The key idea of proximal gradient method is to solve the following objective at each iteration t,

$$\begin{aligned}\mathbf{W}^{t+1} &= \arg\min_{\mathbf{U}} \quad f(\mathbf{W}^t) + \nabla f(\mathbf{W}^t)^T(\mathbf{U} - \mathbf{W}^t) \\ &\qquad + \frac{L_t}{2}\|\mathbf{U} - \mathbf{W}^t\|_F^2 + \phi_\lambda(\mathbf{U}) \qquad (3)\\ &= \arg\min_{\mathbf{U}} \quad \frac{1}{2}\|\mathbf{U} - \mathbf{A}\|_F^2 + \phi_\rho(\mathbf{U}), \qquad (4)\end{aligned}$$

where $\mathbf{A} = \mathbf{W}^t - \frac{1}{L_t}\nabla f(\mathbf{W}^t)$, and $\rho = \frac{\lambda}{L_t}$, L_t is a parameter chosen at each iteration using some search strategy. Thus the problem is transformed to minimization problem of Eq.(3).

Their major computation efforts in each iteration is to compute the gradient of $\nabla f(\mathbf{W}^t)$, which costs $\mathcal{O}(dnK)$ for a generic dense matrix $\mathbf{W}$, where d is feature dimension, n is number of data points, K is number of classes. With appropriate choice of step size L_t, the proximal gradient method (Beck and Teboulle 2009a) will achieve ϵ-optimal solution in $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ iterations. Then the key step is to solve the associated proximal operator:

$$\min_{\mathbf{U}} \quad \frac{1}{2}\|\mathbf{U} - \mathbf{A}\|_F^2 + \phi_\rho(\mathbf{U}). \tag{5}$$

One main contribution of our paper is for the computation of proximal operator of Eq.(5).

Main contribution: An effective algorithm for associated proximal operator computation

In this section, we present an efficient algorithm to solve the proximal operator in Eq.(5), Lots of previous works (Yuan, Liu, and Ye 2011; Beck and Teboulle 2009b) have shown the efficient computation of proximal operator is the key to many sparse learning problems.

First, we note

$$\begin{aligned}&\frac{1}{2}||\mathbf{W} - \mathbf{A}||_F^2 + \rho||\mathbf{W}||_{p,\infty} \\ &= \sum_{i=1}^{d}\Big(\frac{1}{2}||\mathbf{w}^i - \mathbf{a}^i||^2 + \rho(||\mathbf{w}^i||_\infty)^p\Big), \qquad (6)\end{aligned}$$

where $\mathbf{w}^i$, $\mathbf{a}^i$ is the i-th row of matrix $\mathbf{W}, \mathbf{A}$; $||\mathbf{w}^i||_\infty = \max_j|\mathbf{W}_{ij}|$ is infinity norm for vector $\mathbf{w}^i$.

The problem of Eq.(5) is therefore decomposed into d *independent* sub-problems, and each one optimizing $\mathbf{w}^i$. Each sub-problem has the following general formulation,

$$\mathbf{u} = \arg\min_{\mathbf{u}} J_1(\mathbf{u}) = \arg\min_{\mathbf{u}} \left(\frac{1}{2}||\mathbf{u}-\mathbf{a}||^2 + \rho(||\mathbf{u}||_\infty)^p\right)$$
$$= \arg\min_{\mathbf{u}} \left(\frac{1}{2}||\mathbf{u}-\mathbf{a}||^2 + \rho \max_{1\leqslant i\leqslant d}|u_i|^p\right), \quad (7)$$

where $\mathbf{u} = [u_1, u_2, ..., u_d]$ is a row vector to be optimized. We call this optimization problem as $\ell_{p,\infty}$ **proximal projection** problem. Fortunately, we have very efficient algorithm to solve Eq.(7).

Simplify the problem

First, we present Lemmas 1 and 2 to simplify this optimization problem.

Lemma 1. *(A) The optimal solution* $\mathbf{u}^*$ *for Eq.(7) satisfies* $sign(u_i^*) = sign(a_i)$. *(B) Let* $\mathbf{v}^*$ *be the optimal solution for objective function* J_2,

$$J_2 = \min_{v_i\geqslant 0} \frac{1}{2}||\mathbf{v}-\mathbf{b}||^2 + \rho(\max_{1\leqslant i\leqslant d} v_i)^p$$

where $b_i = |a_i|$. *The optimal solution* $\mathbf{u}^*$ *in Eq.(7) is given by*

$$u_i^* = sign(a_i)v_i^*.$$

Proof (A) We prove it by contradiction. Suppose in the optimal solution $\mathbf{u}^*$, there exists i_0 where $sign(u_{i_0}^*) \neq sign(a_{i_0})$, i.e., $sign(u_{i_0}^*) = -sign(a_{i_0})$. Then we can construct another solution $\mathbf{u}^{**}$, where

$$u_i^{**} = \begin{cases} -u_i^* & \textbf{if } i = i_0 \\ u_i^* & \textbf{if } i \neq i_0 \end{cases}.$$

Note that $\max_{1\leqslant i\leqslant d}|u_i^*| = \max_{1\leqslant i\leqslant d}|u_i^{**}|$, we have

$$J_1(\mathbf{u}^{**}) - J_1(\mathbf{u}^*) = \frac{1}{2}\sum_i (u_i^{**}-a_i)^2 - \frac{1}{2}\sum_i (u_i^*-a_i)^2$$
$$= 2u_{i_0}^* a_{i_0} \leqslant 0.$$

Thus, $J_1(\mathbf{u}^{**}) \leqslant J_1(\mathbf{u}^*)$, which contradicts the assumption that $\mathbf{u}^*$ is the optimal solution.

(B) For the optimal solution $\mathbf{u}^*$, we have

$$\frac{1}{2}\sum_i (u_i^*-a_i)^2 + \rho(\max_{1\leqslant i\leqslant d}|u_i^*|)^p$$
$$= \frac{1}{2}\sum_i (sign(u_i^*)|u_i^*| - sign(a_i)|a_i|)^2$$
$$+\rho(\max_{1\leqslant i\leqslant d}|u_i^*|)^p \quad (8)$$
$$= \frac{1}{2}\sum_i (|u_i^*|-|a_i|)^2 + \rho(\max_{1\leqslant i\leqslant d}|u_i^*|)^p. \quad (9)$$

The deduction from Eq.(8) to Eq.(9) holds because $sign(u_i^*) = sign(a_i)$, which is proved in Lemma 1(A). Let $b_i = |a_i|$, $v_i = |u_i^*|$, Eq.(9) recovers the objective function in J_2. □.

Remark From Eq.(7) to objective function of J_2, the absolute value operation $|.|$ is removed. Clearly, to optimize J_2 is much easier than to optimize Eq.(7).

Lemma 2. *If elements of* $\mathbf{b}$ *is sorted in descending order, i.e.,* $b_1 \geqslant b_2 \geqslant ... \geqslant b_d$, *the optimal solution* $\mathbf{v}^*$ *for* J_2 *must satisfy* $v_1^* \geqslant v_2^* \geqslant ... \geqslant v_d^*$.

It is straightforward to verify Lemma 2. For the convenience of computing $max(v_i)$, first we sort $\mathbf{b}$ in descending order. Next we show how to solve J_2.

Optimal solution at $p = 0$ Note when $p = 0$, for objective function J_2, the minimal function value is given by $\min(\rho, \frac{1}{2}||\mathbf{b}||^2)$, where optimal $\mathbf{v} = \mathbf{b}$ or $\mathbf{0}$. [1] Next we discuss about the optimal solution when $0 < p \leq 1$.

The structure of optimal solution when $0 < p \leq 1$

The key observation is that the optimal solution for objective function of J_2, there may be more than one element achieving the maximum value at the same time even if the elements of $\mathbf{b}$ are distinct. For example, if we set

$$\mathbf{b} = (5, 4, 3, 2, 1), \rho = 1.5, p = 1,$$

the optimal solution for $\mathbf{v}^* = (3.75, 3.75, 3, 2, 1)$ instead of $\mathbf{v} = (3.5, 4, 3, 2, 1)$. Thus, we introduce the *maximum set* S defined as,

$$S = \{i | v_i^* = v^m\},$$

where v^m is the maximum value in the optimal solution v^* because more than one element may reach to the same *maximum* value.

The next problem is to decide the set S. It is natural to split the elements of v^* into two parts based on whether they fall into set S or not. To determine the values of optimal solution $\mathbf{v}$ after separation, we have **Theorem 3** to outline the structure of an optimal solution. To determine the elements falling into set S, we have Lemma 4 to characterize the property of set S and give the cardinality of set S.

Theorem 3. *Assume* maximum *set* S *is known, the optimal solution* v_i^* *for objective function* J_2 *is*

$$v_i^* = \begin{cases} b_i & \textbf{if } i \notin S \\ v^m & \textbf{if } i \in S \end{cases}$$

where v^m *is a constant and can be obtained through Newton's method.*

Proof According to the definition of set S, object function J_2 can be written as J_3,

$$J_3 = \frac{1}{2}\sum_{i\in S}(v^m - b_i)^2 + \frac{1}{2}\sum_{i\notin S}(v_i - b_i)^2 + \rho(v^m)^p \quad (10)$$

Clearly, the optimal solution v_i^* can be split into two parts based on $i \in S$ or $i \notin S$, and these two parts are independent in the optimal function J_3.

First, consider $v_i^*(i \notin S)$. The solution for minimization of $\frac{1}{2}\sum_{i\notin S}(v_i - b_i)^2$ of Eq.(10) is given by $v_i^* = b_i$.

Next, consider $v_i^*(i \in S)$. Note

$$\arg\min_{v^m}(J_3) = \arg\min_{v^m} \frac{1}{2}\sum_{i\in S}(v^m - b_i)^2 + \rho(v^m)^p,$$

[1] To our knowledge, there is no well-defined value for 0^0. Here we assume $0^0 = 0$.

thus

$$J_3'(v_m) = \sum_{i \in S}(v^m - b_i) + \rho p (v^m)^{p-1},$$

$$J_3''(v_m) = |S| + \rho p(p-1)(v^m)^{p-2}.$$

Note at $p = 1$, v_m has closed form solution, and is given by

$$v_0^m = \frac{1}{|S|}(\sum_{j \in S} b_j - \rho).$$

When $0 < p < 1$, using standard Newton's method, we can iteratively compute v^m using

$$v_{t+1}^m = v_t^m - \frac{J_3'(v_t^m)}{J_3''(v_t^m)}.$$

Starting from an initial guess for v_0^m, e.g.,

$$v_0^m = \frac{1}{|S|}(\sum_{j \in S} b_j - \rho),$$

in our experiment, we obtain the optimal v^m after 20 iterations to machine precision. □

Lemma 4. *The optimal solution* $\mathbf{v}^*$ *of Eq.(10) is characterized by* $\begin{cases} v^m > b_i & \textbf{if } i \notin S \\ v^m \leqslant b_i & \textbf{if } i \in S \end{cases}$.

Proof. We prove it by contradiction.

(A) We first prove the case $i \notin S$. Suppose for those elements of v_i^*, $v^m \leqslant b_i (i \notin S)$. From Theorem 3, we have $v_i^* = b_i (i \notin S)$, and therefore, $v^m \leqslant b_i = v_i^* (i \notin S)$. This contradicts the definition of the set S where $v^m = \max_{1 \leqslant i \leqslant d} v_i^*$. Thus the assumption $v^m \leqslant b_i (i \notin S)$ does not hold.

(B) Now we prove the case $i \in S$. Suppose for those elements of v_i^*, $v^m > b_i \ \ (i \in S)$.

It is easy to see there exists $\varepsilon = v^m - \max_{i \in S}\{b_i\} > 0$. Thus, for $\forall i \in S$,

$$v^m - b_i - \varepsilon = (v^m - b_i) - (v^m - \max_{i \in S}\{b_i\}) = (\max_{i \in S}\{b_i\} - b_i) \geqslant 0.$$

Then $(2v^m - 2b_i - \varepsilon) = (v^m - b_i) + (v^m - b_i - \varepsilon) > 0$. We can construct another solution $\mathbf{v}^{**}$, such that

$$\begin{cases} v_i^{**} = v_i^* & \textbf{if } i \notin S \\ v_i^{**} = v^m - \varepsilon & \textbf{if } i \in S \end{cases},$$

and we have

$$\begin{aligned} J_2(\mathbf{v}^{**}) - J_2(\mathbf{v}^*) &= [\frac{1}{2}\sum_{i \in S}(v^m - \varepsilon - b_i)^2 \\ &- \frac{1}{2}\sum_{i \in S}(v^m - b_i)^2] + \rho[(\max_{1 \leqslant i \leqslant d} v_i^{**})^p - (\max_{1 \leqslant i \leqslant d} v_i^*)^p] \\ &= \frac{1}{2}\sum_{i \in S} -\varepsilon(2v^m - 2b_i - \varepsilon) + \rho[(v^m - \varepsilon)^p - (v^m)^p)] \\ &< 0. \end{aligned} \tag{11}$$

The second term in above Equation is less than 0 because $v^m \geq 0, v^m - \varepsilon \geq 0$, and function $f(x) = x^p$ is monotonically increasing at $x > 0$ (see Fig.1). Thus

$$J_2(\mathbf{v}^{**}) - J_2(\mathbf{v}^*) < 0.$$

This implies $\mathbf{v}^{**}$ is the optimal solution, which contradicts that $\mathbf{v}^*$ is the optimal solution. Thus the assumption does not hold, $v^m \leq b_i \ \ (i \in S)$. □

Algorithm 1 Proximal operator solution for Eq.(7) when $0 < p \leq 1$

Input: $\mathbf{v}, \rho$
Output: $\mathbf{v}$

1: $\mathbf{b} \leftarrow sort(\mathbf{b})$ to ensure $b_1 \geqslant b_2 \geqslant ... \geqslant b_d$, record the *mapping order* before and after sorting
2: $k \leftarrow d$, let $S = \{1, 2, \cdots k\}$, solve for v^m according to Theorem 1.
3: **while** $v^m > b_k$ **and** $k > 1$ **do**
4: $\quad k \leftarrow (k-1)$
5: $\quad$ let $S = \{1, 2, \cdots, k\}$, solve for v^m according to Theorem 1.
6: **end while**
7: $\mathbf{v}' \leftarrow \mathbf{b}, v_i' \leftarrow v^m (1 \leq i \leq k)$
8: map $\mathbf{v}'$ back into $\mathbf{v}$ according to *mapping order*

Complete algorithm

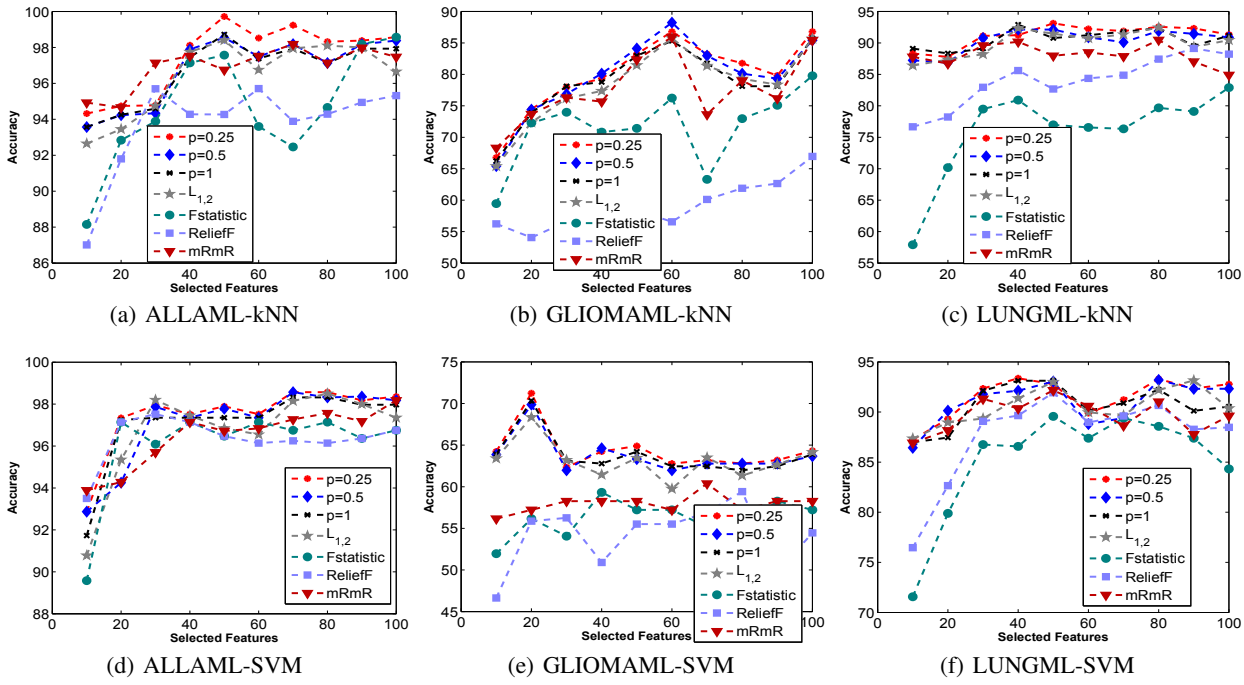
Above we have discussed the structure of optimal solution when $0 < p \leq 1$. Next we complete the whole algorithm and give the optimal solution $\mathbf{v}$ for Eq.(7). Based on Lemma 4, we can use a linear search algorithm to determine the *maximum* set S by making comparisons in the boundary conditions (e.g., $v^m > b_i$). The time cost for linear search algorithm is $O(d)$, which is proportional to the dimension of $\mathbf{b}$.

In summary, we present the detailed algorithm in Algorithm 1. Actually, in order to further improve the efficiency, we can use a binary search to determine set S with time cost $O(\log d)$. Since we can first arrange $\mathbf{b}$ in descending order, i.e, $b_1 \geqslant b_2 \geqslant ... \geqslant b_d$. Another interesting observation is that $v^m \leqslant b_1$ according to Lemma 4. These properties help to obtain a better understanding of the structure of optimal solution.

Experiments

We perform extensive experiment on both single-label multi-class and multi-label datasets to validate the effectiveness of proposed $\ell_{p,\infty}$ algorithm. For multi-class datasets, we use four widely used biology datasets: ALLAML[1], GLIOMAML[2], LUNGML[3] and CARML[4], which all have high-dimensional features (more than 3000) and very few samples (less than 250). We use three widely used multi-label datasets, Barcelona[5], MSRCv2[6], TRECVID2005[7]. We extract 384-dimensional color moment feature on datasets Barcelona and MSRCv2, and 512-dimensional GIST (Oliva and Torralba 2001) features on TRECVID (*e.g.*, (Chang et al. 2014)). A summary of both multi-class and multi-label datasets is shown in Table.1.

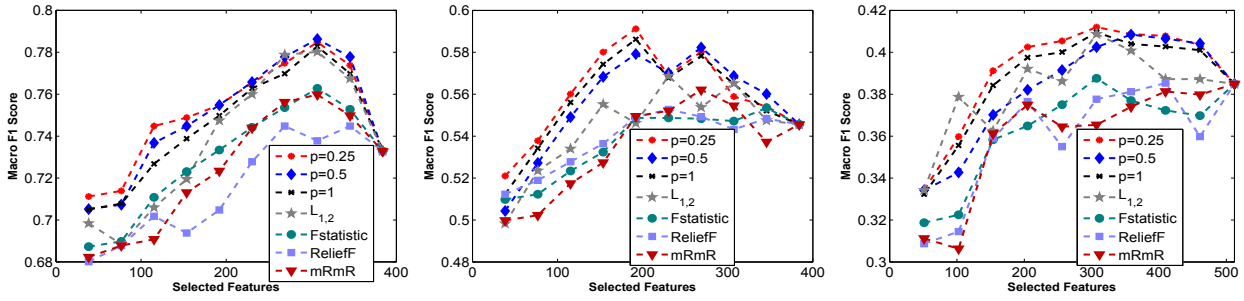
[1] http://www.sciencemag.org/content/277/5324/393.full
[2] http://cancerres.aacrjournals.org/content/63/7/1602.long
[3] http://www.ncbi.nlm.nih.gov/pubmed/11707567
[4] http://www.ncbi.nlm.nih.gov/pubmed/11606367
[5] http://mlg.ucd.ie/content/view/61
[6] http://research.microsoft.com/en-us/projects/objectclassrecognition/
[7] http://www-nlpir.nist.gov/projects/tv2005/



(a) ALLAML-kNN (b) GLIOMAML-kNN (c) LUNGML-kNN

(d) ALLAML-SVM (e) GLIOMAML-SVM (f) LUNGML-SVM

Figure 2: Classification accuracy using selected features on kNN and SVM classifier on three datasets. $p = 0.25, p = 0.5, p = 1$ are our methods at different p. Four other methods: $\ell_{2,1}$ feature selection of Eq.(12), F-statistic, reliefF and mRmR.



(a) Barcelona (b) MSRC (c) Trevid

Figure 3: Multi-label feature selection results using SVM classifier on three datasets. $p = 0.25, p = 0.5, p = 1$ are our methods at different p. Four other methods: $\ell_{2,1}$ feature selection of Eq.(12), F-statistic, reliefF and mRmR.

Numerical experiments with two initializations

It is worth noting that for $p < 1$, it is not a convex optimization problem to solve Eq.(2). We are interested to see the influences of different initializations. We used two initializations: (a) initialize $\mathbf{W}$ using ridge regression, i.e., replace the $\ell_{p,\infty}$ in Eq.(2) with simple Frobenius norm, which gives closed form solution; (b) experiment with another scheme, i.e., start with $p = 1$ global solution, then use this solution as initialization for $p = 0.75, 0.5$, and so on. The results shown for $p < 1$ are the best of these two initializations.

The results obtained from $\ell_{p,\infty}$ model can be used for feature selection. We adjust the parameter λ in Eq.(1), and select the non-zeros rows in the results of $\mathbf{W}$ as the selected features. After selecting top $r < d$ features, we obtain the new data $\tilde{\mathbf{X}} \in \Re^{r \times n}$, which is composed of only top r features across all data samples. We use standard least square loss for the error function $f(\mathbf{W})$ of Eq.(1). The same least square function is also used for multi-label experiments. More formally, we define the residue $\mathbf{R} = ||\mathbf{Y} - \mathbf{B}^T\tilde{\mathbf{X}}||_F$, where $\mathbf{Y} \in \Re^{k \times n}$ is class labels, and $\mathbf{B}$ is obtained by minimizing: $\min_{\mathbf{B}} ||\mathbf{Y} - \mathbf{B}^T\tilde{\mathbf{X}}||_F^2$ on the selected features $\tilde{\mathbf{X}}$. The smaller the residue $\mathbf{R}$, the better the selected features obtained from $\mathbf{X}$. we show the residues $\mathbf{R}$, obtained using top 20 features with different initializations, i.e., (a) *ridge regression*; (b) $p = 1$ *result* on 4 datasets.

We make several important observations from the results shown in Table 1. **(1)** Generally, smaller p values give much smaller residues, which indicates better feature selection results; but it is not always consistent. We believe that if we could find the true global solution at $p < 1$, the residues will

Table 1: Multi-class dataset and multi-label dataset descriptions.

Single-label multi-class datasets				Multi-label multi-class datasets			
dataset	#data	#dimension	# class	dataset	#data	#dimension	# class
ALLAML	72	7129	2	Trevid	3721	512	39
GLIOMAML	50	4434	4	MSRC	591	384	23
LUNGML	203	3312	5	Barcelona	139	384	4
CARML	174	9182	11				

Table 2: Residue $\mathbf{R}$ on selected top 20 features on 4 datasets

Dataset	initialization	$p = 0.1$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1$	$\ell_{2,1}$
ALLAML	$p = 1$	6.4217	**6.3647**	6.4314	6.3892	6.4487	6.4257
	ridge regression	6.4141	6.3841	6.4054	**6.3672**		
GLIOMAML	$p = 1$	**4.5208**	4.7877	4.7935	4.7832	4.9421	4.8324
	ridge regression	**4.5145**	4.7182	4.7877	4.7968		
LUNGML	$p = 1$	7.6889	**7.5701**	7.7916	8.3251	8.3572	8.3612
	ridge regression	7.6762	**7.5534**	7.8216	8.2851		
CARML	$p = 1$	8.3947	8.3837	**8.3687**	8.8135	8.9578	8.9321
	ridge regression	8.3747	8.3921	**8.3723**	8.7335		

be more consistent. However, because at $p < 1$, the problem is *non-convex*, we cannot guarantee the global minima, and the solution is *not unique* and depends on initialization. **(2)** Ridge regression initialization gives slightly better results as compared to $p = 1$ initialization. **(3)** At $p < 1$, the selected features are slightly different (usually several different features depending on how many features to select) for different initialization because we cannot get the global optimization now. **(4)** We also compare the proposed $\ell_{p,\infty}(p = 1)$ against $\ell_{2,1}$ (Liu, Ji, and Ye 2009) feature selection method. To be exact, we follow (Liu, Ji, and Ye 2009), (Kong, Ding, and Huang 2011), use $\ell_{2,1}$ regularization term for feature selection, i.e.,

$$\min_{\mathbf{W}\in\Re^{d\times c}} J(\mathbf{W}) = f(\mathbf{W}) + \lambda||\mathbf{W}||_{2,1}, \tag{12}$$

where $f(\mathbf{W})$ is a smooth/non-smooth convex loss function,e.g., least square loss, and $||\mathbf{W}||_{2,1} = \sum_{j=1}^{d}\sqrt{\sum_{k=1}^{c}\mathbf{W}_{jk}^2}$, and $\lambda \geq 0$ is regularization parameters. Residues obtained from $\ell_{p,\infty}$ feature selection results are generally better than those from $\ell_{2,1}$ results of Eq.(12).

Application for multi-class feature selection tasks

To further validate our feature selection results, we use another two widely used classifiers: SVM, kNN to compute the classification accuracy after using $\ell_{p,\infty}$ feature selection methods. We did 5-fold cross-validation on both classifiers on 3 datasets: ALLAML, GLIOMAML and LUNGML. We compare the proposed feature selection method $\ell_{p,\infty}$ at differen p against several popularly used feature selection methods, such as F-statistic (Liu and Motoda 1998), reliefF (Kononenko 1994), (Kong et al. 2012) and mRmR (Ding and Peng 2003).

Fig. 2 shows the classification accuracy comparisons of different feature selection methods on three data sets. In our methods, p is set to $p = 0.25, p = 0.5, p = 1$. The shown results are the average accuracy over 5 runs. From the results shown in Fig.2, we observe that, **(1)** $\ell_{p,\infty}$ feature selection produces better results as compared to F-statistics, ReliefF and mRmR in terms of classification accuracy; **(2)** the classification accuracy is slightly better when p is small (say $p = 0.25, 0.5$) on both kNN and SVM classifiers; **(3)** the classification accuracy obtained from $\ell_{p,\infty}$ feature selection at smaller p values are also generally better than $\ell_{2,1}$ method, which again verifies the effectiveness of our method.

Application for multi-label feature selection tasks

$\ell_{p,\infty}$ feature selection is naturally extended for multi-label feature selection tasks. In multi-label classification problems, a data point can be attributed to multiple classes simultaneously. For the other multi-label feature selection algorithms, we extend the general F-statistic (Liu and Motoda 1998), reliefF (Kononenko 1994) and mRmR (Ding and Peng 2003) for multi-label classification using binary relevance (Tsoumakas, Katakis, and Vlahavas 2010). We adopt the macro-F1 score defined in (Yang 1999) to evaluate the multi-label classification performance. The higher the macro-F1 score, the better classification accuracy.

We use standard SVM classifier (linear kernel, $C = 1$) to validate the feature selection results. We report the macro-F1 score by using 5 round 5-fold cross validation in Fig.3. Fig.3 indicate $\ell_{p,\infty}$ method performs much better than the other three feature selection methods (e.g., ReliefF, mRmR, etc). Moreover, when p is small, the feature selection results are generally better than $p = 1$.

Conclusion

In this paper, we propose to use $\ell_{p,\infty}$ operator for feature selection. An efficient algorithm is presented to solve $\ell_{p,\infty}$ regularization problem with rigorous analysis. Extensive experiments on multi-class and multi-label datasets indicate the well behaviors of our algorithm at smaller p values.

Acknowledgement. This research is partially supported by NSF-CCF-0917274 and NSF-DMS-0915228 grants.

References

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

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

Berwin A. Turlach, William N. Venables, S. J. W. 2005. Simultaneous variable selection. In *Technometrics*, 349–363.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press.

Chang, X.; Shen, H.; Wang, S.; Liu, J.; and Li, X. 2014. Semi-supervised feature analysis for multimedia annotation by mining label correlation. In *Advances in Knowledge Discovery and Data Mining*, 74–85.

Chartrand, R., and Staneva, V. 2008. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems* 24(035020):1–14.

Ding, C., and Peng, H. 2003. Minimum redundancy feature selection from microarray gene expression data. In *J Bioinform Comput Biol*, 523–529.

Fan, J., and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. 96(456):1348–1359.

Hastie, T.; Tibshirani, R.; and Friedman, J. H. 2001. *Elements of Statistical Learning*. Springer.

Huang, J.; Horowitz, J.; and Ma, S. 2008. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics* 36(2):587–613.

Kohavi, R., and John, G. H. 1997. Wrappers for feature subset selection. *ARTIFICIAL INTELLIGENCE* 97(1):273–324.

Kong, D., and Ding, C. H. Q. 2013. Efficient algorithms for selecting features with arbitrary group constraints via group lasso. In *ICDM*, 379–388.

Kong, D.; Ding, C. H. Q.; Huang, H.; and Zhao, H. 2012. Multi-label relieff and f-statistic feature selections for image annotation. In *CVPR*, 2352–2359.

Kong, D.; Ding, C. H. Q.; and Huang, H. 2011. Robust nonnegative matrix factorization using l21-norm. In *CIKM*, 673–682.

Kong, D.; Zhang, M.; and Ding, C. H. Q. 2013. Minimal shrinkage for noisy data recovery using schatten-p norm objective. In *ECML/PKDD (2)*, 177–193.

Kononenko, I. 1994. Estimating attributes: Analysis and extensions of relief. 171–182. Springer Verlag.

Langley, P. 1994. Selection of relevant features in machine learning. In *In Proceedings of the AAAI Fall symposium on relevance*, 140–144. AAAI Press.

Liu, H., and Motoda, H. 1998. *Feature Selection for Knowledge Discovery and Data Mining*. Springer.

Liu, J.; Ji, S.; and Ye, J. 2009. Multi-task feature learning via efficient l2, 1-norm minimization. In *UAI*, 339–348.

Liu, H.; Palatucci, M.; and Zhang, J. 2009. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *ICML*.

Masaeli, M.; Fung, G.; and Dy, J. G. 2010. From transformation-based dimensionality reduction to feature selection. In *ICML*.

Nesterov, Y. 2004. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42(3):145–175.

Quattoni, A.; Collins, M.; and Darrell, T. 2008. Transfer learning for image classification with sparse prototype representations. In *CVPR*.

Schmidt, M.; Murphy, K.; Fung, G.; and Rosales, R. 2008. Structure learning in random fields for heart motion abnormality detection. In *In CVPR*.

Tropp, J. A.; Gilbert, A. C.; Martin; Strauss, J.; Tropp, J. A.; Gilbert, A. C.; and Strauss, M. J. 2006. Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. *Signal Processing* 589–602.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2010. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook*.

Turlach, B. A.; Venablesy, W. N.; and Wright, S. J. 2005. Simultaneous variable selection. In *Technometrics*, 349–363.

Vogt, J. E., and Roth, V. 2012. A complete analysis of the $l_{1,p}$ group-lasso.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval* 1:67–88.

Yuan, L.; Liu, J.; and Ye, J. 2011. Efficient methods for overlapping group lasso. In *NIPS*, 352–360.

Zhang, Y. 2011. A probabilistic framework for learning task relationships in multi-task learning. In *Ph.D Thesis, The Hong Kong University of Science and Technology*.