Journal of Machine Learning Research 11 (2010) 1081-1107 Submitted 5/09; Revised 1/10; Published 3/10

Analysis of Multi-stage Convex Relaxation for Sparse Regularization

Tong Zhang * TZHANG@STAT.RUTGERS.EDU
Statistics Department
110 Frelinghuysen Road
Rutgers University
Piscataway, NJ 08854

Editor: Francis Bach

Abstract

We consider learning formulations with non-convex objective functions that often occur in practical applications. There are two approaches to this problem:

- Heuristic methods such as gradient descent that only find a local minimum. A drawback of this approach is the lack of theoretical guarantee showing that the local minimum gives a good solution.
- Convex relaxation such as L_1-regularization that solves the problem under some conditions. However it often leads to a sub-optimal solution in reality.

This paper tries to remedy the above gap between theory and practice. In particular, we present a multi-stage convex relaxation scheme for solving problems with non-convex objective functions. For learning formulations with sparse regularization, we analyze the behavior of a specific multi-stage relaxation scheme. Under appropriate conditions, we show that the local solution obtained by this procedure is superior to the global solution of the standard L_1 convex relaxation for learning sparse targets.

Keywords: sparsity, non-convex optimization, convex relaxation, multi-stage convex relaxation

1. Introduction

We consider the general regularization framework for machine learning, where a loss function is minimized, subject to a regularization condition on the model parameter. For many natural machine learning problems, either the loss function or the regularization condition can be non-convex. For example, the loss function is non-convex for classification problems, and the regularization condition is non-convex in problems with sparse parameters.

A major difficulty with nonconvex formulations is that the global optimal solution cannot be efficiently computed, and the behavior of a local solution is hard to analyze. In practice, convex relaxation (such as support vector machine for classification or L_1 regularization for sparse learning) has been adopted to remedy the problem. The choice of convex formulation makes the solution unique and efficient to compute. Moreover, the solution is easy to analyze theoretically. That is, it can be shown that the solution of the convex formulation approximately solves the original problem under appropriate assumptions. However, for many practical problems, such simple convex relaxation schemes can be sub-optimal.

∗. Supported by the following grants: NSF DMS-0706805, NSA-081024, and AFOSR-10097389.

Because of the above gap between practice and theory, it is important to study direct solutions of non-convex optimization problems beyond the standard convex relaxation. Our goal is to design a numerical procedure that leads to a *reproducible solution* which is better than the standard convex relaxation solution. In order to achieve this, we present a general framework of multi-stage convex relaxation, which iteratively refine the convex relaxation formulation to give better solutions. The method is derived from concave duality, and involves solving a sequence of convex relaxation problems, leading to better and better approximations to the original nonconvex formulation. It provides a unified framework that includes some previous approaches (such as LQA Jianqing Fan, 2001, LLA Zou and Li, 2008, CCCP Yuille and Rangarajan, 2003) as special cases. The procedure itself may be regarded as a special case of alternating optimization, which automatically ensures its convergence. Since each stage of multi-stage convex relaxation is a convex optimization problem, the approach is also computationally efficient. Although the method only leads to a local optimal solution for the original nonconvex problem, this local solution is a refinement of the global solution for the initial convex relaxation. Therefore intuitively one expects that the local solution is better than the standard one-stage convex relaxation. In order to prove this observation more rigorously, we consider least squares regression with nonconvex sparse regularization terms, for which we can analyze the effectiveness of the multi-stage convex relaxation. It is shown that under appropriate assumptions, the (local) solution computed by the multi-stage convex relaxation method using nonconvex regularization achieves better parameter estimation performance than the standard convex relaxation with L_1 regularization.

The main contribution of this paper is the analysis of sparse regularized least squares regression presented in Section 3, where we derive theoretical results showing that under appropriate conditions, it is beneficial to use multi-stage convex relaxation with nonconvex regularization as opposed to the standard convex L_1 regularization. This demonstrates the effectiveness of multi-stage convex relaxation for a specific but important problem. Although without theoretical analysis, we shall also present the general idea of multi-stage convex relaxation in Section 2, because it can be applied to other potential application examples as illustrated in Appendix C. The gist of our analysis can be applied to those examples (e.g., the multi-task learning problem in the setting of matrix completion, which has drawn significant attention recently) as well. However, the detailed derivation will be specific to each application and the analysis will not be trivial. Therefore while we shall present a rather general form of multi-stage convex relaxation formulation in order to unify various previous approaches, and put this work in a broader context, the detailed theoretical analysis (and empirical studies) for other important applications will be left to future work.

2. Multi-stage Convex Relaxation

This section presents the general idea of multi-stage convex relaxation which can be applied to various optimization problems. It integrates a number of existing ideas into a unified framework.

2.1 Regularized Learning Formulation

The multi-stage convex relaxation approach considered in the paper can be applied to the following optimization problem, which can be motivated from supervised learning formulation. As back-

ground information, its connection to regularized learning formula (15) is given in Appendix B.

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} R(\mathbf{w}),$$
$$R(\mathbf{w}) = R_0(\mathbf{w}) + \sum_{k=1}^{K} R_k(\mathbf{w}), \tag{1}$$

where $\mathbf{w} = [\mathbf{w}_1, \ldots, \mathbf{w}_d] \in \mathbb{R}^d$ is a d-dimensional parameter vector, and $R(\mathbf{w})$ is the general form of a regularized objective function. Moreover, for convenience, we assume that $R_0(\mathbf{w})$ is convex in $\mathbf{w}$, and each $R_k(\mathbf{w})$ is non-convex. In the proposed work, we shall employ convex/concave duality to derive convex relaxations of (1) that can be efficiently solved.

Related to (1), one may also consider the constrained formulation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} R_0(\mathbf{w}) \quad \text{subject to } \sum_{k=1}^{K} R_k(\mathbf{w}) \le A, \tag{2}$$

where A is a constant. One may also mix (1) and (2).

2.2 Concave Duality

In the following discussion, we consider a single nonconvex component $R_k(\mathbf{w})$ in (1), which we shall rewrite using concave duality. Let $\mathbf{h}_k(\mathbf{w}) : R^d \to \Omega_k \subset R^{d_k}$ be a vector function with Ω_k being the convex hull of its range. It may not be a one-to-one map. However, we assume that there exists a function $\bar{R}_k$ defined on Ω_k so that we can express $R_k(\mathbf{w})$ as

$$R_k(\mathbf{w}) = \bar{R}_k(\mathbf{h}_k(\mathbf{w})).$$

Assume that we can find $\mathbf{h}_k$ so that the function $\bar{R}_k(\mathbf{u}_k)$ is concave on $\mathbf{u}_k \in \Omega_k$. Under this assumption, we can rewrite the regularization function $R_k(\mathbf{w})$ as:

$$R_k(\mathbf{w}) = \inf_{\mathbf{v}_k \in R^{d_k}} \left[\mathbf{v}_k^\top \mathbf{h}_k(\mathbf{w}) - R_k^*(\mathbf{v}_k) \right] \tag{3}$$

using concave duality (Rockafellar, 1970). In this case, $R_k^*(\mathbf{v}_k)$ is the concave dual of $\bar{R}_k(\mathbf{u}_k)$ given below

$$R_k^*(\mathbf{v}_k) = \inf_{\mathbf{u}_k \in \Omega_k} \left[\mathbf{v}_k^\top \mathbf{u}_k - \bar{R}_k(\mathbf{u}_k) \right].$$

Note that using the convention in convex analysis (Rockafellar, 1970), we may assume that $R_k^*(\mathbf{v}_k)$ is defined on R^{d_k} but may take $-\infty$ value. Equivalently, we may consider the subset $\{\mathbf{v}_k : R_k^*(\mathbf{v}_k) > -\infty\}$ as the feasible region of the optimization problem (3), and assume that $R_k^*(\mathbf{v}_k)$ is only defined on this feasible region.

It is well-known that the minimum of the right hand side of (3) is achieved at

$$\hat{\mathbf{v}}_k = \nabla_{\mathbf{u}} \bar{R}_k(\mathbf{u})|_{\mathbf{u}=\mathbf{h}_k(\mathbf{w})}. \tag{4}$$

This is the general framework we suggest in the paper. For illustration, some example non-convex problems encountered in machine learning are included in Appendix C.

2.3 Penalized Formulation

Let $\mathbf{h}_k(\mathbf{w})$ be a vector function with convex components, so that (3) holds. Given an appropriate vector $\mathbf{v}_k \in R^{d_k}$, a simple convex relaxation of (1) becomes

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in R^d} \left[R_0(\mathbf{w}) + \sum_{k=1}^{K} \mathbf{h}_k(\mathbf{w})^\top \mathbf{v}_k \right]. \tag{5}$$

This simple relaxation yields a solution that is different from the solution of (1). However, if each $\mathbf{h}_k$ satisfies the condition of Section 2.2, then it is possible to write $R_k(\mathbf{w})$ using (3). Now, with this new representation, we can rewrite (1) as

$$[\hat{\mathbf{w}}, \hat{\mathbf{v}}] = \arg\min_{\mathbf{w},\{\mathbf{v}_k\}} \left[R_0(\mathbf{w}) + \sum_{k=1}^{K} (\mathbf{h}_k(\mathbf{w})^\top \mathbf{v}_k - R_k^*(\mathbf{v}_k)) \right]. \tag{6}$$

This is clearly equivalent to (1) because of (3). If we can find a good approximation of $\hat{\mathbf{v}} = \{\hat{\mathbf{v}}_k\}$ that improves upon the initial value of $\hat{\mathbf{v}}_k = \mathbf{1}$, then the above formulation can lead to a refined convex problem in $\mathbf{w}$ that is a better convex relaxation than (5).

Our numerical procedure exploits the above fact, which tries to improve the estimation of $\mathbf{v}_k$ over the initial choice of $\mathbf{v}_k = \mathbf{1}$ in (5) using an iterative algorithm. This can be done using an alternating optimization procedure, which repeatedly applies the following two steps:

- First we optimize $\mathbf{w}$ with $\mathbf{v}$ fixed: this is a convex problem in $\mathbf{w}$ with appropriately chosen $\mathbf{h}(\mathbf{w})$.
- Second we optimize $\mathbf{v}$ with $\mathbf{w}$ fixed: although non-convex, it has a closed form solution that is given by (4).

Initialize $\hat{\mathbf{v}} = \mathbf{1}$
Repeat the following two steps until convergence:

- Let
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[R_0(\mathbf{w}) + \sum_{k=1}^{K} \mathbf{h}_k(\mathbf{w})^\top \hat{\mathbf{v}}_k \right]. \tag{7}$$
- Let $\hat{\mathbf{v}}_k = \nabla_{\mathbf{u}} \bar{R}_k(\mathbf{u})|_{\mathbf{u}=\mathbf{h}_k(\hat{\mathbf{w}})}$ $(k = 1, \ldots, K)$

Figure 1: Multi-stage Convex Relaxation Method

The general procedure for solving (6) is presented in Figure 1. It can be regarded as a generalization of CCCP (concave-convex programming) (Yuille and Rangarajan, 2003), which takes $\mathbf{h}(\mathbf{w}) = \mathbf{w}$. It is also more general than LQA (local quadratic approximation) (Jianqing Fan, 2001) or LLA (local linear approximation) (Zou and Li, 2008). Specifically LQA takes $\mathbf{h}_j(\hat{\mathbf{w}}) = \mathbf{w}_j^2$ and LLA takes $\mathbf{h}_j(\hat{\mathbf{w}}) = |\mathbf{w}_j|$. The justification of those procedures rely on the so-called MM (majorization-minimization) principle, where an upper bound of the objective function is minimized at each step (see Zou and Li, 2008 and references therein). However, in order to apply MM, for each particular choice of $\mathbf{h}$, one has to demonstrate that the convex relaxation is indeed an upper bound, which is

necessary to show convergence. In the concave relaxation formulation adopted in this work, the justification of convergence is automatically embedded in (6), which becomes a joint optimization problem. Figure 1 is simply an alternating optimization procedure for solving (6), which is equivalent to (1). Since convex duality of many interesting objective functions (including matrix functions) are familiar to many machine learning researchers, the concave duality derivation presented here can be automatically applied to various applications without the need to worry about convergence justification. This will be especially useful for complex formulations such as structured or matrix regularization, where the more traditional MM idea cannot be easily applied. One may also regard our framework as a principled method to design a class of algorithms that may be interpreted as MM procedures. Some examples illustrate its applications are presented in Appendix C.

Note that by repeatedly refining the parameter $\mathbf{v}$, we can potentially obtain better and better convex relaxation in Figure 1, leading to a solution superior to that of the initial convex relaxation. Since at each step the procedure decreases the objective function in (6), its convergence to a local minimum is easy to show. In fact, in order to achieve convergence, one only needs to approximately minimize (7) and reasonably decrease the objective value at each step. We skip the detailed analysis here, because in the general case, a local solution is not necessarily a good solution, and there are other approaches (such as gradient descent) that can compute a local solution. In order to demonstrate the effectiveness of multi-stage convex relaxation, we shall include a more careful analysis for the special case of sparse regularization in Section 3.1. Our theory shows that the local solution of multi-stage relaxation with a nonconvex sparse regularizer is superior to the convex L_1 regularization solution (under appropriate conditions).

2.4 Constrained Formulation

The multi-stage convex relaxation idea can also be used to solve the constrained formulation (2). The one-stage convex relaxation of (2), given fixed relaxation parameter $\mathbf{v}_k$, becomes

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in R^d} R_0(\mathbf{w}) \quad \text{subject to } \sum_{k=1}^{K} \mathbf{h}_k(\mathbf{w})^\top \mathbf{v}_k \le A + \sum_{k=1}^{K} R_k^*(\mathbf{v}_k).$$

Because of (3), the above formulation is equivalent to (2) if we optimize over $\mathbf{v}$. This means that by optimizing $\mathbf{v}$ in addition to $\mathbf{w}$, we obtain the following algorithm:

- Initialize $\hat{\mathbf{v}} = \mathbf{1}$
- Repeat the following two steps until convergence:
  - Let
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} R_0(\mathbf{w}) \quad \text{subject to } \sum_{k=1}^{K} \mathbf{h}_k(\mathbf{w})^\top \hat{\mathbf{v}}_k \le A + \sum_{k=1}^{K} R_k^*(\hat{\mathbf{v}}_k).$$
  - Let $\hat{\mathbf{v}}_k = \nabla_{\mathbf{u}} \bar{R}_k(\mathbf{u})|_{\mathbf{u}=\mathbf{h}_k(\hat{\mathbf{w}})}$ $(k = 1,\ldots,K)$

If an optimization problem includes both nonconvex penalization and nonconvex constrains, then one may use the above algorithm with Figure 1.

3. Multi-stage Convex Relaxation for Sparse Regularization

The multi-stage convex relaxation method described in the previous section tries to obtain better approximations of the original nonconvex problem by refining the convex relaxation formulation. Since the local solution found by the algorithm is the global solution of a refined convex relaxation formulation, it should be closer to the desired solution than that of the standard one-stage convex relaxation method. Although this high level intuition is appealing, it is still necessarily to present a more rigorous theoretical result which can precisely demonstrate the advantage of the multi-stage approach over the standard single stage method. Unless we can develop a theory to show the effectiveness of the multi-stage procedure in Figure 1, our proposal is yet another local minimum finding scheme that may potentially get stuck at a bad local solution.

In order to obtain some strong theoretical results that can demonstrate the advantage of the multi-stage approach, we consider the special case of sparse learning. This is because this problem has been well-studied in recent years, and the behavior of convex relaxation (L_1 regularization) is well-understood.

3.1 Theory of Sparse Regularization

For a non-convex but smooth regularization condition such as capped-L_1 or smoothed-L_p with $p \in (0,1)$, standard numerical techniques such as gradient descent lead to a local minimum solution. Unfortunately, it is difficult to find the global optimum, and it is also difficult to analyze the quality of the local minimum. Although in practice, such a local minimum solution may outperform the Lasso solution, the lack of theoretical (and practical) performance guarantee prevents the more wide-spread applications of such algorithms. As a matter of fact, results with non-convex regularization are difficult to reproduce because different numerical optimization procedures can lead to different local minima. Therefore the quality of the solution heavily depend on the numerical procedure used.

The situation is very different for a convex relaxation formulation such as L_1-regularization (Lasso). The global optimum can be easily computed using standard convex programming techniques. It is known that in practice, 1-norm regularization often leads to sparse solutions (although often suboptimal). Moreover, its performance has been theoretically analyzed recently. For example, it is known from the compressed sensing literature that under certain conditions, the solution of L_1 relaxation may be equivalent to L_0 regularization asymptotically (e.g., Candes and Tao, 2005). If the target is truly sparse, then it was shown in Zhao and Yu (2006) that under some restrictive conditions referred to as *irrepresentable conditions*, 1-norm regularization solves the feature selection problem. The prediction performance of this method has been considered in Koltchinskii (2008), Zhang (2009a), Bickel et al. (2009) and Bunea et al. (2007).

In spite of its success, L_1-regularization often leads to suboptimal solutions because it is not a good approximation to L_0 regularization. Statistically, this means that even though it converges to the true sparse target when $n \to \infty$ (consistency), the rate of convergence can be suboptimal. The only way to fix this problem is to employ a non-convex regularization condition that is closer to L_0 regularization. In the following, we formally prove a result for multi-stage convex relaxation with non-convex sparse regularization that is superior to the Lasso result. In essence, we establish a performance guarantee for non-convex formulations when they are solved by using the multi-stage convex relaxation approach which is more sophisticated than the standard one-stage convex relaxation.

In supervised learning, we observe a set of input vectors $\mathbf{x}_1,\ldots,\mathbf{x}_n \in R^d$, with corresponding desired output variables $y_1,\ldots,y_n$. In general, we may assume that there exists a target $\bar{\mathbf{w}} \in R^d$ such that

$$y_i = \bar{\mathbf{w}}^\top \mathbf{x}_i + \varepsilon_i \qquad (i = 1,\ldots,n), \tag{8}$$

where ε_i are zero-mean independent random noises (but not necessarily identically distributed). Moreover, we assume that the target vector $\bar{\mathbf{w}}$ is sparse. That is, there exists $\bar{k} = \|\bar{\mathbf{w}}\|_0$ is small. This is the standard statistical model for sparse learning.

Let $\mathbf{y}$ denote the vector of $[y_i]$ and X be the $n \times d$ matrix with each row a vector $\mathbf{x}_i$. We are interested in recovering $\bar{\mathbf{w}}$ from noisy observations using the following sparse regression method:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[\frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \sum_{j=1}^{d} g(|\mathbf{w}_j|) \right], \tag{9}$$

where $g(|\mathbf{w}_j|)$ is a regularization function. Here we require that $g'(u)$ is non-negative which means we penalize larger u more significantly. Moreover, we assume $u^{1-q}g'(u)$ is a non-increasing function when $u > 0$, which means that $[g(|\mathbf{w}_1|),\ldots,g(|\mathbf{w}_d|)]$ is concave with respect to $\mathbf{h}(\mathbf{w}) = [|\mathbf{w}_1|^q,\ldots,|\mathbf{w}_d|^q]$ for some $q \geq 1$. It follows that (9) can be solved using the multi-stage convex relaxation algorithm in Figure 2, which we will analyze. Although this algorithm was mentioned in Zou and Li (2008) as LLA when $q = 1$, they only presented a one-step low-dimensional asymptotic analysis. We present a true multi-stage analysis in high dimension. Our analysis also focuses on $q = 1$ (LLA) for convenience because the Lasso analysis in Zhang (2009a) can be directly adapted; however in principle, one can also analyze the more general case of $q > 1$.

Initialize $\lambda_j^{(0)} = \lambda$ for $j = 1,\ldots,d$
For $\ell = 1,2,\ldots$

- Let
$$\hat{\mathbf{w}}^{(\ell)} = \arg\min_{\mathbf{w}\in R^d} \left[\frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|_2^2 + \sum_{j=1}^{d} \lambda_j^{(\ell-1)} |\mathbf{w}_j|^q \right]. \tag{10}$$
- Let $\lambda_j^{(\ell)} = \lambda q^{-1} |\hat{\mathbf{w}}_j|^{1-q} g'(|\hat{\mathbf{w}}_j^{(\ell)}|)$ $(j = 1,\ldots,d)$

Figure 2: Multi-stage Convex Relaxation for Sparse Regularization

For convenience, we consider fixed design only, where X is fixed and the randomness is with respect to $\mathbf{y}$ only. We require some technical conditions for our analysis. First we assume sub-Gaussian noise as follows.

Assumption 3.1 *Assume that $\{\varepsilon_i\}_{i=1,\ldots,n}$ in (8) are independent (but not necessarily identically distributed) sub-Gaussians: there exists $\sigma \geq 0$ such that $\forall i$ and $\forall t \in R$,*

$$\mathbf{E}_{\varepsilon_i} e^{t\varepsilon_i} \leq e^{\sigma^2 t^2/2}.$$

Both Gaussian and bounded random variables are sub-Gaussian using the above definition. For example, if a random variable $\xi \in [a,b]$, then $\mathbf{E}_\xi e^{t(\xi - \mathbf{E}\xi)} \leq e^{(b-a)^2 t^2/8}$. If a random variable is Gaussian: $\xi \sim N(0,\sigma^2)$, then $\mathbf{E}_\xi e^{t\xi} \leq e^{\sigma^2 t^2/2}$.

We also introduce the concept of sparse eigenvalue, which is standard in the analysis of L_1 regularization.

Definition 1 *Given k, define*

$$\rho_+(k) = \sup\left\{\frac{1}{n}\|X\mathbf{w}\|_2^2/\|\mathbf{w}\|_2^2 : \|\mathbf{w}\|_0 \le k\right\},$$
$$\rho_-(k) = \inf\left\{\frac{1}{n}\|X\mathbf{w}\|_2^2/\|\mathbf{w}\|_2^2 : \|\mathbf{w}\|_0 \le k\right\}.$$

Our main result is stated as follows. The proof is in the appendix.

Theorem 2 *Let Assumption 3.1 hold. Assume also that the target $\bar{\mathbf{w}}$ is sparse, with $\mathbf{E}y_i = \bar{\mathbf{w}}^\top \mathbf{x}_i$, and $\bar{k} = \|\bar{\mathbf{w}}\|_0$. Choose λ such that*

$$\lambda \ge 20\sigma\sqrt{2\rho_+(1)\ln(2d/\eta)/n}.$$

Assume that $g'(z) \ge 0$ is a non-increasing function such that $g'(z) = 1$ when $z \le 0$. Moreover, we require that $g'(\theta) \ge 0.9$ with $\theta = 9\lambda/\rho_-(2\bar{k}+s)$. Assume that $\rho_+(s)/\rho_-(2\bar{k}+2s) \le 1+0.5s/\bar{k}$ for some $s \ge 2\bar{k}$, then with probability larger than $1-\eta$:

$$\|\hat{\mathbf{w}}^{(\ell)} - \bar{\mathbf{w}}\|_2 \le \frac{17}{\rho_-(2\bar{k}+s)}\left[2\sigma\sqrt{\rho_+(\bar{k})}\left(\sqrt{7.4\bar{k}/n} + \sqrt{2.7\ln(2/\eta)/n}\right)\right.$$
$$\left. +\lambda\left(\sum_{j:\bar{\mathbf{w}}_j \ne 0} g'(|\bar{\mathbf{w}}_j| - \theta)^2\right)^{1/2}\right] + 0.7^\ell \frac{10}{\rho_-(2\bar{k}+s)}\sqrt{\bar{k}}\lambda,$$

where $\hat{\mathbf{w}}^{(\ell)}$ is the solution of (10) with $q = 1$.

Note that the theorem allows the situation $d \gg n$, which is what we are interested in. This is the first general analysis of multi-stage convex relaxation for high dimensional sparse learning, although some simpler asymptotic results for low dimensional two-stage procedures were obtained in Zou (2006) and Zou and Li (2008), they are not comparable to ours.

Results most comparable to what we have obtained here are that of the FoBa procedure in Zhang (2009b) and that of the MC+ procedure in Zhang (2010). The former is a forward backward greedy algorithm, which does not optimize (9), while the latter is a path-following algorithm for solving (9). Although results in Zhang (2010) are comparable to ours, we should note that efficient path-following computation in MC+ requires specialized regularizers $g(\cdot)$. Moreover, unlike our procedure, which is efficient because of convex optimization, there is no proof showing that the path-following strategy in Zhang (2010) is always efficient (in the sense that there may be exponentially many switching points). However, empirical experience in Zhang (2010) does indicate its efficiency for a class of regularizers that can be relatively easily handled by path-following. Therefore we are not claiming here that our approach will always be superior to Zhang (2010) in practice. Nevertheless, our result suggests that different local solution procedures can be used to solve the same nonconvex formulation with valid theoretical guarantees. This opens the door for additional theoretical studies of other numerical procedures.

The condition $\rho_+(s)/\rho_-(2\bar{k}+2s) \le 1+0.5s/\bar{k}$ requires the eigenvalue ratio $\rho_+(s)/\rho_-(s)$ to grow sub-linearly in s. Such a condition, referred to as *sparse eigenvalue condition*, is also needed in

the standard analysis of L_1 regularization (Zhang and Huang, 2008; Zhang, 2009a). It is related but weaker than the *restricted isometry property* (RIP) in compressive sensing (Candes and Tao, 2005). Note that in the traditional low-dimensional statistical analysis, one assumes that $\rho_+(s)/\rho_-(2\bar{k}+2s) < \infty$ as $s \to \infty$, which is significantly stronger than the condition we use here. Although in practice it is often difficult to verify the sparse eigenvalue condition for real problems, Theorem 2 nevertheless provides important theoretical insights for multi-stage convex relaxation.

Since in standard Lasso, $g'(|\mathbf{w}_j|) \equiv 1$, we obtain the following bound from Theorem 2

$$\|\hat{\mathbf{w}}_{L_1} - \bar{\mathbf{w}}\|_2 = O(\sqrt{\bar{k}}\lambda),$$

where $\hat{\mathbf{w}}_{L_1}$ is the solution of the standard L_1 regularization. This bound is tight for Lasso, in the sense that the right hand side cannot be improved except for the constant—this can be easily verified with an orthogonal design matrix. It is known that in order for Lasso to be effective, one has to pick λ no smaller than the order $\sigma\sqrt{\ln d/n}$. Therefore, the parameter estimation error of the standard Lasso is of the order $\sigma\sqrt{\bar{k}\ln d/n}$, which cannot be improved.

In comparison, if we consider an appropriate regularization condition $g(|\mathbf{w}_j|)$ that is concave in $|\mathbf{w}_j|$. Since $g'(|\mathbf{w}_j|) \approx 0$ when $|\mathbf{w}_j|$ is large, the bound in Theorem 2 can be significantly better when most non-zero coefficients of $\bar{\mathbf{w}}$ are relatively large in magnitude. For example, consider the capped-L_1 regularizer $g(|\mathbf{w}_j|) = \min(\alpha, |\mathbf{w}_j|)$ with $\alpha \geq \theta$; in the extreme case where $\min_j |\mathbf{w}_j| > \alpha + \theta$ (which can be achieved when all nonzero components of $\bar{\mathbf{w}}$ are larger than the order $\sigma\sqrt{\ln d/n}$), we obtain the better bound

$$\|\hat{\mathbf{w}}^{(\ell)} - \bar{\mathbf{w}}\|_2 = O(\sqrt{\bar{k}/n} + \sqrt{\ln(1/\eta)/n})$$

for the multi-stage procedure for a sufficiently large ℓ at the order of $\ln\ln d$. This bound is superior to the standard one-stage L_1 regularization bound $\|\hat{\mathbf{w}}_{L_1} - \bar{\mathbf{w}}\|_2 = O(\sqrt{\bar{k}\ln(d/\eta)/n})$, which is tight for Lasso. The difference can be significant when $\ln d$ is large.

Generally speaking, with a regularization condition $g(|\mathbf{w}_j|)$ that is concave in $|\mathbf{w}_j|$, the dependency on λ is through $g'(|\bar{\mathbf{w}}_j|)$ which decreases as $|\bar{\mathbf{w}}_j|$ increases. This removes the bias of the Lasso and leads to improved performance. Specifically, if $\bar{\mathbf{w}}_j$ is large, then $g'(|\bar{\mathbf{w}}_j|) \approx 0$. In comparison, the Lasso bias is due to the fact that $g'(|\bar{\mathbf{w}}_j|) \equiv 1$. For illustration, the derivative $g'(\cdot)$ of some sparse regularizers are plotted in Figure 3.

Note that our theorem only applies to regularizers with finite derivative at zero. That is, $g'(0) < \infty$. The result doesn't apply to L_p regularization with $p < 1$ because $g'(0) = \infty$. Although a weaker result can be obtained for such regularizers, we do not include it here. We only include an intuitive example below to illustrate why the condition $g'(0) < \infty$ is necessary for stronger results presented in the paper. Observe that the multi-stage convex relaxation method only computes a local minimum, and the regularization update rule is given by $\lambda_j^{(\ell-1)} = g'(\hat{\mathbf{w}}_j^{(\ell-1)})$. If $g'(0) = \infty$, then $\lambda_j^{(\ell-1)} = \infty$ when $\hat{\mathbf{w}}_j^{(\ell-1)} = 0$. This means that if a feature accidentally becomes zero in some stage, it will always remain zero. This is why only weaker results can be obtained for L_p regularizers ($p < 1$): we need to further assume that $\hat{\mathbf{w}}_j^{(\ell)}$ never becomes close to zero when $\bar{\mathbf{w}}_j \neq 0$. A toy example is presented in Table 1 to demonstrate this point. The example is a simulated regression problem with $d = 500$ variables and $n = 100$ training data. The first five variables of the target $\bar{\mathbf{w}}$ are non-zeros, and the remaining variables are zeros. For both capped-L_1 and L_p regularizers, the first stage is the standard L_1 regularization, which misses the correct feature #2 and wrongly selects some incorrect ones. For capped-L_1 regularization, in the second stage, because most correct features are identified,

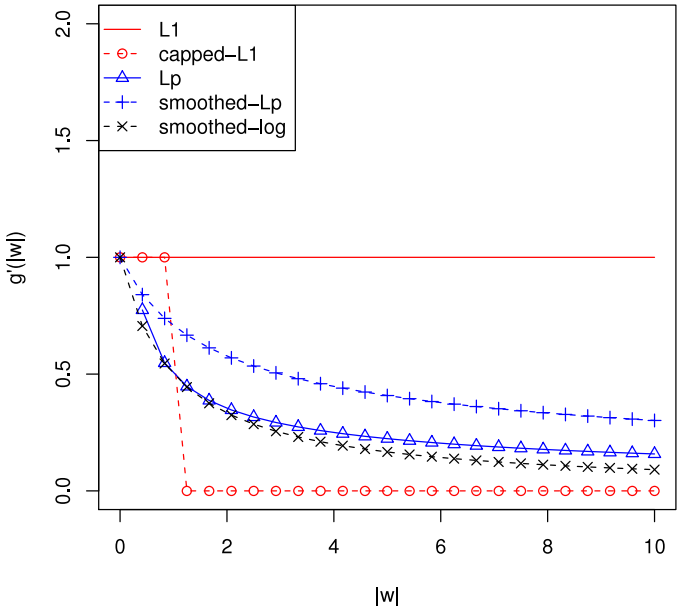

Figure 3: Derivative $g'(|\mathbf{w}_j|)$ of some sparse regularizers

Stage ℓ	coefficients	$\|\hat{\mathbf{w}}^{(\ell)} - \bar{\mathbf{w}}\|_2$
multi-stage capped-L_1		
1	$[6.0, 0.0, 4.7, 4.8, 3.9, 0.6, 0.7, 1.2, 0.0, \ldots]$	4.4
2	$[7.7, 0.4, 5.7, 6.3, 5.7, 0.0, 0.0, 0.2, 0.0, \ldots]$	1.6
3	$[7.8, 1.2, 5.7, 6.6, 5.7, 0.0, 0.0, 0.0, 0.0, \ldots]$	0.98
4	$[7.8, 1.2, 5.7, 6.6, 5.7, 0.0, 0.0, 0.0, 0.0, \ldots]$	0.98
multi-stage $L_{0.5}$		
1	$[6.0, 0.0, 4.7, 4.8, 3.9, 0.6, 0.7, 1.2, 0.0, \ldots]$	4.4
2	$[7.3, 0.0, 5.4, 5.9, 5.3, 0.0, 0.3, 0.3, 0.0, 0.0, \ldots]$	2.4
3	$[7.5, 0.0, 5.6, 6.1, 5.7, 0.0, 0.1, 0.0, 0.0, 0.0, \ldots]$	2.2
4	$[7.5, 0.0, 5.6, 6.2, 5.7, 0.0, 0.1, 0.0, 0.0, 0.0, \ldots]$	2.1
target $\bar{\mathbf{w}}$	$[8.2, 1.7, 5.4, 6.9, 5.7, 0.0, 0.0, 0.0, 0.0, \ldots]$	

Table 1: An Illustrative Example for Multi-stage Sparse Regularization

the corresponding "bias" is reduced by not penalizing the corresponding variables. This leads to improved performance. Since the correct feature #2 shows up in stage 2, we are able to identify it and further improve the convex relaxation in stage 3. After stage 3, the procedure stabilizes because it computes exactly the same relaxation. For L_p regularization, since feature #2 becomes zero in stage 1, it will remain zero thereafter because $\lambda_2^{(\ell)} = \infty$ when $\ell \geq 1$. In order to remedy this problem, one has to use a regularizer with $g'(0) < \infty$ such as the smoothed L_p regularizer.

3.2 Empirical Study

Although this paper focuses on the development of the general multi-stage convex relaxation framework as well as its theoretical understanding (in particular the major result given in Theorem 2), we include two simple numerical examples to verify our theory. More comprehensive empirical comparisons can be found in other related work such as Candes et al. (2008), Zou (2006) and Zou and Li (2008).

In order to avoid cluttering, we only present results with capped-L_1 and L_p ($p = 0.5$) regularization methods. Note that based on Theorem 2, we may tune α in capped-L_1 by using a formula $\alpha = \alpha_0\lambda$ where λ is the regularization parameter. We choose $\alpha_0 = 10$ and $\alpha_0 = 100$.

In the first experiment, we generate an $n \times d$ random matrix with its column j corresponding to $[\mathbf{x}_{1,j}, \ldots, \mathbf{x}_{n,j}]$, and each element of the matrix is an independent standard Gaussian $N(0,1)$. We then normalize its columns so that $\sum_{i=1}^{n} \mathbf{x}_{i,j}^2 = n$. A truly sparse target $\bar{\beta}$, is generated with k nonzero elements that are uniformly distributed from $[-10, 10]$. The observation $\mathbf{y}_i = \bar{\beta}^\top \mathbf{x}_i + \varepsilon_i$, where each $\varepsilon_i \sim N(0, \sigma^2)$. In this experiment, we take $n = 50, d = 200, k = 5, \sigma = 1$, and repeat the experiment 100 times. The average training error and 2-norm parameter estimation error are reported in Figure 4. We compare the performance of multi-stage methods with different regularization parameter λ. As expected, the training error for the multi-stage algorithms are smaller than that of L_1, due to the smaller bias. Moreover, substantially smaller parameter estimation error is achieved by the multi-stage procedures, which is consistent with Theorem 2. This can be regarded as an empirical verification of the theoretical result.

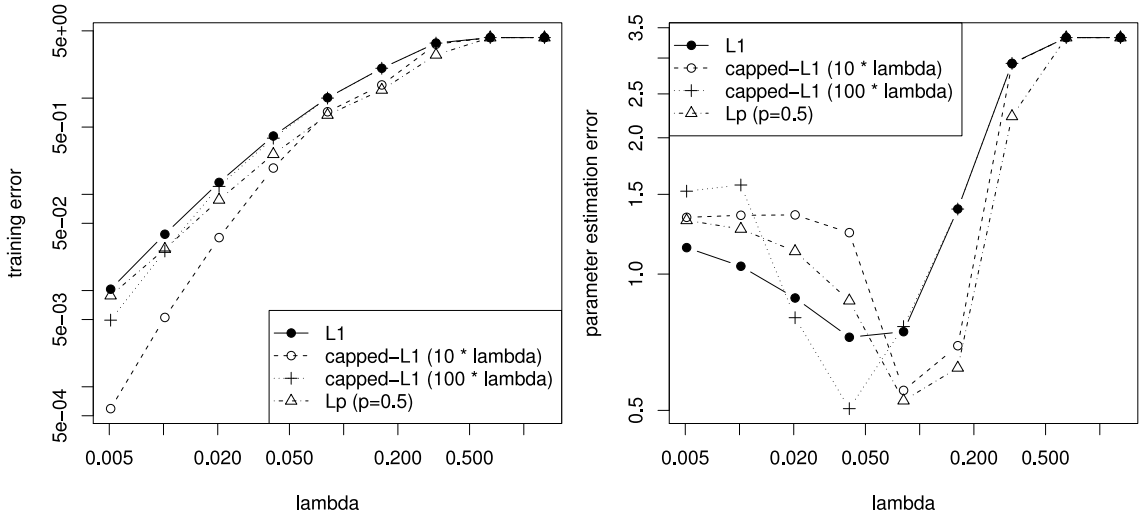

Figure 4: Performance of multi-stage convex relaxation on simulation data. Left: average training squared error versus λ; Right: parameter estimation error versus λ.

In the second experiment, we use the *Boston Housing* data to illustrate the effectiveness of multi-stage convex relaxation. This data set contains 506 census tracts of Boston from the 1970 census, available from the *UCI Machine Learning Database Repository*: `http://archive.ics.`

uci.edu/ml/. Each census tract is a data-point, with 13 features (we add a constant offset on e as the 14th feature), and the desired output is the housing price. In this example, we randomly partition the data into 20 training plus 486 test points. We perform the experiments 100 times, and report training and test squared error versus the regularization parameter λ for different q. The results are plotted in Figure 5. In this case, $L_{0.5}$ is not effective, while capped-L_1 regularization with $\alpha = 100\lambda$ is slightly better than Lasso. Note that this data set contains only a small number ($d = 14$) features, which is not the case where we can expect significant benefit from the multi-stage approach (most of other UCI data similarly contain only small number of features). In order to illustrate the advantage of the multi-stage method more clearly, we also report results on a modified Boston Housing data, where we append 20 random features (similar to the simulation experiments) to the original Boston Housing data, and rerun the experiments. The results are shown in Figure 6. As expected from Theorem 2 and the discussion thereafter, since d becomes large, the multi-stage convex relaxation approach with capped-L_1 regularization and $L_{0.5}$ regularization perform significantly better than the standard Lasso.

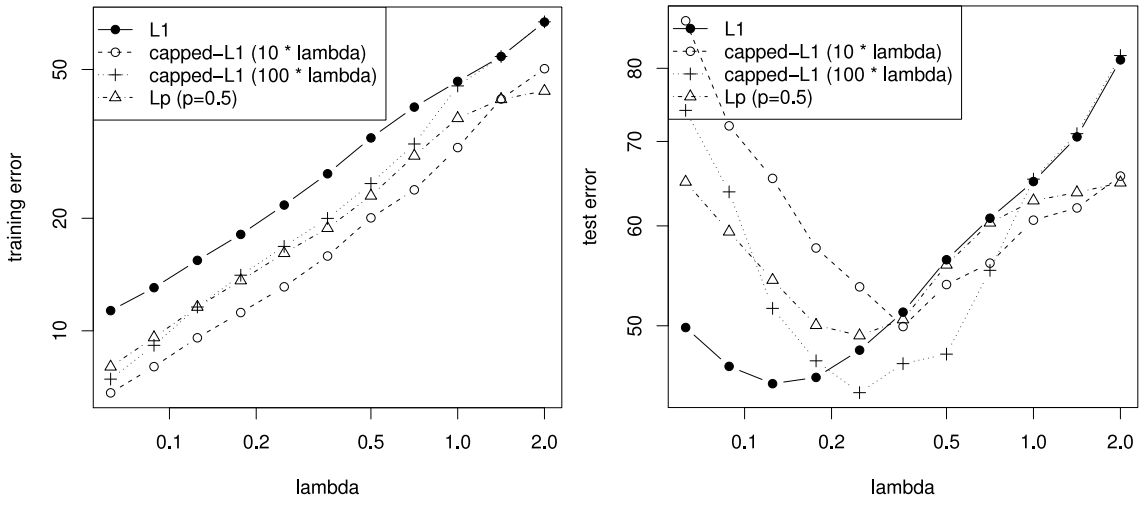

Figure 5: Performance of multi-stage convex relaxation on the original Boston Housing data. Left: average training squared error versus λ; Right: test squared error versus λ.

4. Discussion

Many machine learning applications require solving nonconvex optimization problems. There are two approaches to this problem:

- Heuristic methods such as gradient descent that only find a local minimum. A drawback of this approach is the lack of theoretical guarantee showing that the local minimum gives a good solution.

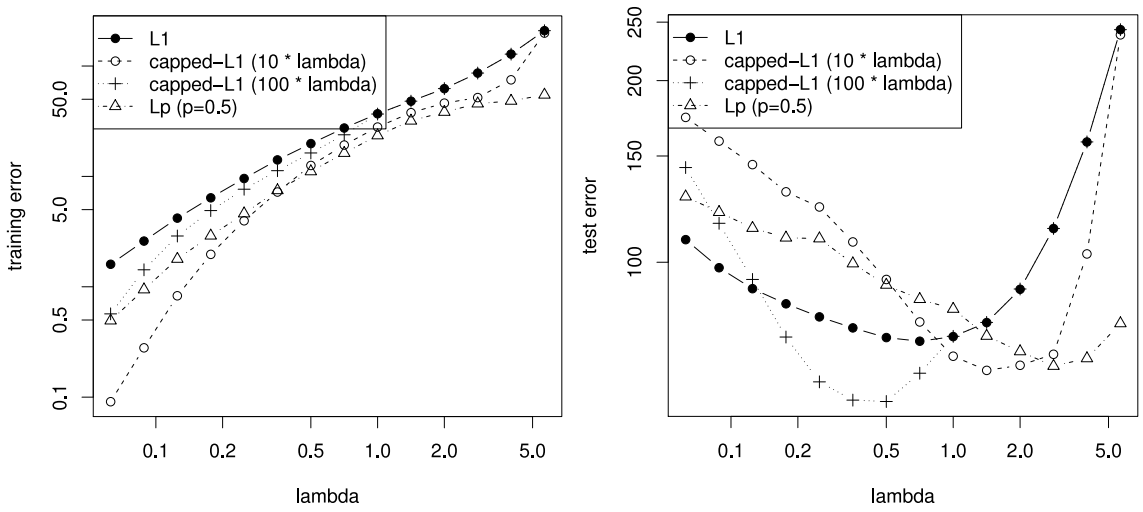

Figure 6: Performance of multi-stage convex relaxation on the modified Boston Housing data. Left: average training squared error versus λ; Right: test squared error versus λ.

- Convex relaxation such as L_1-regularization that solves the problem under some conditions. However it often leads to a sub-optimal solution in reality.

The goal of this paper is to remedy the above gap between theory and practice. In particular, we investigated a multi-stage convex relaxation scheme for solving problems with non-convex objective functions. The general algorithmic technique is presented first, which can be applied to a wide range of problems. It unifies a number of earlier approaches. The intuition is to refine convex relaxation iteratively by using solutions obtained from earlier stages. This leads to better and better convex relaxation formulations, and thus better and better solutions.

Although the scheme only finds a local minimum, the above argument indicates that the local minimum it finds should be closer to the original nonconvex problem than the standard convex relaxation solution. In order to prove the effectiveness of this approach theoretically, we considered the sparse learning problem where the behavior of convex relaxation (Lasso) has been well studied in recent years. We showed that under appropriate conditions, the local solution from the multi-stage convex relaxation algorithm is superior to the global solution of the standard L_1 convex relaxation for learning sparse targets. Experiments confirmed the effectiveness of this method.

We shall mention that our theory only shows that nonconvex regularization behaves better than Lasso under appropriate sparse eigenvalue conditions. When such conditions hold, multi-stage convex relaxation is superior. On the other hand, when such conditions fail, neither Lasso nor (the local solution of) multi-stage convex relaxation can be shown to work well. However, in such case, some features will become highly correlated, and local solutions of non-convex formulations may become unstable. In order to improve stability, it may be helpful to employ ensemble methods such as bagging. Our empirical experience suggests that when features are highly correlated, convex formulations may perform better than (non-bagged) nonconvex formulations due to the added sta-

bility. However, since our analysis doesn't yield any insights in this scenario, further investigation is necessary to theoretically compare convex formulations to bagged nonconvex formulations.

Finally, multi-stage convex relaxation is not the only numerical method that can solve nonconvex formulations with strong theoretical guarantee. For example, the MC+ procedure in Zhang (2010) offers a different method with similar guarantee. This opens the possibility of investigating other local solution methods for nonconvex optimization such as modified gradient descent algorithms that may be potentially more efficient.

Appendix A. Proof of Theorem 2

The analysis is an adaptation of Zhang (2009a). We first introduce some definitions. Consider the positive semi-definite matrix $A = n^{-1}X^\top X \in \mathbb{R}^{d\times d}$. Given $s,k \geq 1$ such that $s+k \leq d$. Let I,J be disjoint subsets of $\{1,\ldots,d\}$ with k and s elements respectively. Let $A_{I,I} \in R^{k\times k}$ be the restriction of A to indices I, $A_{I,J} \in R^{k\times s}$ be the restriction of A to indices I on the left and J on the right. Similarly we define restriction $\mathbf{w}_I$ of a vector $\mathbf{w} \in R^d$ on I; and for convenience, we allow either $\mathbf{w}_I \in R^k$ or $\mathbf{w}_I \in R^d$ (where components not in I are zeros) depending on the context.

We also need the following quantity in our analysis:

$$\pi(k,s) = \sup_{\mathbf{v}\in R^k, \mathbf{u}\in R^s, I, J} \frac{\mathbf{v}^\top A_{I,J}\mathbf{u}\|\mathbf{v}\|_2}{\mathbf{v}^\top A_{I,I}\mathbf{v}\|\mathbf{u}\|_\infty}.$$

The following two lemmas are taken from Zhang (2009a). We skip the proof.

Lemma 3 *The following inequality holds:*

$$\pi(k,s) \leq \frac{s^{1/2}}{2}\sqrt{\rho_+(s)/\rho_-(k+s)-1},$$

Lemma 4 *Consider $k,s>0$ and $G \subset \{1,\ldots,d\}$ such that $|G^c| = k$. Given any $\mathbf{w} \in R^d$. Let J be the indices of the s largest components of $\mathbf{w}_G$ (in absolute values), and $I = G^c \cup J$. Then*

$$\max(0, \mathbf{w}_I^\top A\mathbf{w}) \geq \rho_-(k+s)(\|\mathbf{w}_I\|_2 - \pi(k+s,s)\|\mathbf{w}_G\|_1/s)\|\mathbf{w}_I\|_2.$$

The following lemma gives bounds for sub-Gaussian noise needed in our analysis.

Lemma 5 *Define $\hat{\varepsilon} = \frac{1}{n}\sum_{i=1}^n(\bar{\mathbf{w}}^\top\mathbf{x}_i - \mathbf{y}_i)\mathbf{x}_i$. Under the conditions of Assumption 3.1, with probability larger than $1-\eta$:*

$$\|\hat{\varepsilon}\|_\infty^2 \leq 2\sigma^2\rho_+(1)\ln(2d/\eta)/n. \tag{11}$$

Moreover, for any fixed F, with probability larger than $1-\eta$:

$$\|\hat{\varepsilon}_F\|_2^2 \leq \rho_+(|F|)\sigma^2[7.4|F| + 2.7\ln(2/\eta)]/n. \tag{12}$$

Proof The proof relies on two propositions. The first proposition is a simple application of large deviation bound for sub-Gaussian random variables.

Proposition 6 *Consider a fixed vector $\mathbf{u} = [\mathbf{u}_1,\ldots,\mathbf{u}_n] \in \mathbb{R}^n$, and a random vector $\mathbf{y} = [\mathbf{y}_1,\ldots,\mathbf{y}_n] \in \mathbb{R}^n$ with independent sub-Gaussian components: $\mathbf{E}e^{t(\mathbf{y}_i - \mathbf{E}\mathbf{y}_i)} \leq e^{\sigma^2t^2/2}$ for all t and i, then $\forall\varepsilon>0$:*

$$\Pr\left(\left|\mathbf{u}^\top(\mathbf{y}-\mathbf{E}\mathbf{y})\right| \geq \varepsilon\right) \leq 2e^{-\varepsilon^2/(2\sigma^2\|\mathbf{u}\|_2^2)}.$$

Proof (of Proposition 6). Let $s_n = \sum_{i=1}^n \mathbf{u}_i(\mathbf{y}_i - \mathbf{E}\mathbf{y}_i)$; then by assumption, $\mathbf{E}(e^{ts_n} + e^{-ts_n}) \leq 2e^{\sum_i \mathbf{u}_i^2\sigma^2t^2/2}$, which implies that $\Pr(|s_n| \geq \varepsilon)e^{t\varepsilon} \leq 2e^{\sum_i \mathbf{u}_i^2\sigma^2t^2/2}$. Now let $t = \varepsilon/(\sum_i \mathbf{u}_i^2\sigma^2)$, we obtain the desired bound. ■

The second proposition is taken from Pisier (1989).

Proposition 7 *Consider the unit sphere $S^{k-1} = \{\mathbf{u} : \|\mathbf{u}\|_2 = 1\}$ in $\mathbb{R}^k$ ($k \geq 1$). Given any $\varepsilon > 0$, there exists an ε-cover $Q \subset S^{k-1}$ such that $\min_{q\in Q} \|\mathbf{u} - q\|_2 \leq \varepsilon$ for all $\|\mathbf{u}\|_2 = 1$, with $|Q| \leq (1+2/\varepsilon)^k$.*

Now are are ready to prove (11). Let $\mathbf{x}_{i,j}$ be the j-th component of $\mathbf{x}_i$, then by definition, we have $\sum_{i=1}^n \mathbf{x}_{i,j}^2 \leq n\rho_+(1)$ for all $j = 1,\ldots,d$. It follows from Proposition 6 that for all $\varepsilon > 0$ and j: $\Pr(|\hat{\varepsilon}_j| \geq \varepsilon) \leq 2e^{-n\varepsilon^2/(2\sigma^2\rho_+(1))}$. Taking union bound for $j = 1,\ldots,d$, we obtain $\Pr(\|\hat{\varepsilon}\|_\infty \geq \varepsilon) \leq 2de^{-n\varepsilon^2/(2\sigma^2\rho_+(1))}$, which is equivalent to (11).

Next we are ready to prove (12). Let P be the projection matrix to the column spanned by X_F, and let k be the dimension of P, then $k \leq |F|$.

According to Proposition 7, given $\varepsilon_1 > 0$, there exists a finite set $Q = \{q_i\}$ with $|Q| \leq (1+2/\varepsilon_1)^k$ such that $\|Pq_i\|_2 = 1$ for all i, and $\min_i \|P\mathbf{z} - Pq_i\|_2 \leq \varepsilon_1$ for all $\|P\mathbf{z}\|_2 = 1$. To see the existence of Q, we consider a rotation of the coordinate system (which does not change 2-norm) so that $P\mathbf{z}$ is the projection of $\mathbf{z} \in \mathbb{R}^n$ to its first k coordinates in the new coordinate system. Proposition 7 can now be directly applied to the first k coordinates in the new system, implying that we can pick q_i such that $Pq_i = q_i$.

For each i, Proposition 6 implies that $\forall \varepsilon_2 > 0$:

$$\Pr\left(\left|q_i^\top P(\mathbf{y} - \mathbf{E}\mathbf{y})\right| \geq \varepsilon_2\right) \leq 2e^{-\varepsilon_2^2/(2\sigma^2)}.$$

Taking union bound for all $q_i \in Q$, we obtain with probability exceeding $1 - 2(1+2/\varepsilon_1)^k e^{-\varepsilon_2^2/2\sigma^2}$:

$$\left|q_i^\top P(\mathbf{y} - \mathbf{E}\mathbf{y})\right| \leq \varepsilon_2$$

for all i.

Let $\mathbf{z} = P(\mathbf{y} - \mathbf{E}\mathbf{y})/\|P(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2$, then there exists i such that $\|P\mathbf{z} - Pq_i\|_2 \leq \varepsilon_1$. We have

$$\begin{aligned}\|P(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2 =& \mathbf{z}^\top P(\mathbf{y} - \mathbf{E}\mathbf{y}) \\ \leq& \|P\mathbf{z} - Pq_i\|_2\|P(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2 + |q_i^\top P(\mathbf{y} - \mathbf{E}\mathbf{y})| \\ \leq& \varepsilon_1\|P(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2 + \varepsilon_2.\end{aligned}$$

Therefore

$$\|P(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2 \leq \varepsilon_2/(1 - \varepsilon_1).$$

Let $\varepsilon_1 = 2/15$, and $\eta = 2(1+2/\varepsilon_1)^k e^{-\varepsilon_2^2/2\sigma^2}$, we have

$$\varepsilon_2^2 = 2\sigma^2[(4k+1)\ln 2 - \ln\eta],$$

and thus

$$\begin{aligned}\rho_+(|F|)^{-1/2}\|\hat{\varepsilon}_F\|_2 =& \rho_+(|F|)^{-1/2}\|X_F^\top(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2 \\ \leq& \|P(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2 \leq \frac{15}{13}\sigma\sqrt{2(4k+1)\ln 2 - 2\ln\eta}.\end{aligned}$$

This simplifies to the desired bound.

■

Lemma 8 *Consider $\bar{\mathbf{w}}$ such that $\{j : \bar{\mathbf{w}}_j \neq 0\} \subset F$ and $F \cap G = \emptyset$. Let $\hat{\mathbf{w}} = \hat{\mathbf{w}}^{(\ell)}$ be the solution of (10) with $q = 1$, and let $\Delta\hat{\mathbf{w}} = \hat{\mathbf{w}} - \bar{\mathbf{w}}$. Let $\lambda_G = \min_{j\in G}\lambda_j^{(\ell-1)}$ and $\lambda_0 = \max_j \lambda_j^{(\ell-1)}$. Then*

$$\sum_{j\in G}|\hat{\mathbf{w}}_j| \leq \frac{2\|\hat{\varepsilon}\|_\infty}{\lambda_G - 2\|\hat{\varepsilon}\|_\infty}\sum_{j\notin F\cup G}|\hat{\mathbf{w}}_j| + \frac{2\|\hat{\varepsilon}\|_\infty + \lambda_0}{\lambda_G - 2\|\hat{\varepsilon}\|_\infty}\sum_{j\in F}|\Delta\hat{\mathbf{w}}_j|.$$

Proof For simplicity, let $\lambda_j = \lambda_j^{(\ell-1)}$. The first order equation implies that

$$\frac{1}{n}\sum_{i=1}^n 2(\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_{i,j} + \lambda_j \mathrm{sgn}(\mathbf{w}_j) = 0,$$

where $\mathrm{sgn}(\mathbf{w}_j) = 1$ when $\mathbf{w}_j > 0$, $\mathrm{sgn}(\mathbf{w}_j) = -1$ when $\mathbf{w}_j < 0$, and $\mathrm{sgn}(\mathbf{w}_j) \in [-1,1]$ when $\mathbf{w}_j = 0$. This implies that for all $\mathbf{v} \in \mathbb{R}^d$, we have

$$2\mathbf{v}^\top A\Delta\hat{\mathbf{w}} \leq -2\mathbf{v}^\top\hat{\varepsilon} - \sum_{j=1}^d \lambda_j \mathbf{v}_j \mathrm{sgn}(\hat{\mathbf{w}}_j). \tag{13}$$

Now, let $\mathbf{v} = \Delta\hat{\mathbf{w}}$ in (13), we obtain

$$\begin{aligned}
0 \leq & 2\Delta\hat{\mathbf{w}}^\top A\Delta\hat{\mathbf{w}} \leq 2|\Delta\hat{\mathbf{w}}^\top\hat{\varepsilon}| - \sum_{j=1}^d \lambda_j\Delta\hat{\mathbf{w}}_j\mathrm{sgn}(\hat{\mathbf{w}}_j)\\
\leq & 2\|\Delta\hat{\mathbf{w}}\|_1\|\hat{\varepsilon}\|_\infty - \sum_{j\in F}\lambda_j\Delta\hat{\mathbf{w}}_j\mathrm{sgn}(\hat{\mathbf{w}}_j) - \sum_{j\notin F}\lambda_j\Delta\hat{\mathbf{w}}_j\mathrm{sgn}(\hat{\mathbf{w}}_j)\\
\leq & 2\|\Delta\hat{\mathbf{w}}\|_1\|\hat{\varepsilon}\|_\infty + \sum_{j\in F}\lambda_j|\Delta\hat{\mathbf{w}}_j| - \sum_{j\notin F}\lambda_j|\hat{\mathbf{w}}_j|\\
\leq & \sum_{j\in G}(2\|\hat{\varepsilon}\|_\infty - \lambda_G)|\hat{\mathbf{w}}_j| + \sum_{j\notin G\cup F} 2\|\hat{\varepsilon}\|_\infty|\hat{\mathbf{w}}_j| + \sum_{j\in F}(2\|\hat{\varepsilon}\|_\infty + \lambda_0)|\Delta\hat{\mathbf{w}}_j|.
\end{aligned}$$

By rearranging the above inequality, we obtain the desired bound. ■

Lemma 9 *Using the notations of Lemma 8, and let J be the indices of the largest s coefficients (in absolute value) of $\hat{\mathbf{w}}_G$. Let $I = G^c \cup J$ and $k = |G^c|$. If $(\lambda_0 + 2\|\hat{\varepsilon}\|_\infty)/(\lambda_G - 2\|\hat{\varepsilon}\|_\infty) \leq 3$, then*

$$\|\Delta\hat{\mathbf{w}}\|_2 \leq (1 + (3k/s)^{0.5})\|\Delta\hat{\mathbf{w}}_I\|_2.$$

Proof Using $(\lambda_0 + 2\|\hat{\varepsilon}\|_\infty)/(\lambda_G - 2\|\hat{\varepsilon}\|_\infty) \leq 3$, we obtain from Lemma 8

$$\|\hat{\mathbf{w}}_G\|_1 \leq 3\|\Delta\hat{\mathbf{w}} - \hat{\mathbf{w}}_G\|_1.$$

Therefore $\|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_\infty \leq \|\Delta\hat{\mathbf{w}}_G\|_1/s \leq 3\|\Delta\hat{\mathbf{w}} - \hat{\mathbf{w}}_G\|_1/s$, which implies that

$$\begin{aligned}
\|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_2 \leq & (\|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_1\|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_\infty)^{1/2}\\
\leq & 3^{1/2}\|\Delta\hat{\mathbf{w}} - \hat{\mathbf{w}}_G\|_1 s^{-1/2} \leq (3k/s)^{1/2}\|\Delta\hat{\mathbf{w}}_I\|_2.
\end{aligned}$$

By rearranging this inequality, we obtain the desired bound. ■

Lemma 10 *Let the conditions of Lemma 8 hold, and let $k = |G^c|$. If $t = 1 - \pi(k+s,s)k^{1/2}s^{-1} > 0$, and $(\lambda_0 + 2\|\hat{\varepsilon}\|_\infty)/(\lambda_G - 2\|\hat{\varepsilon}\|_\infty) \leq (4-t)/(4-3t)$, then*

$$\|\Delta\hat{\mathbf{w}}\|_2 \leq \frac{1+(3k/s)^{0.5}}{t\rho_-(k+s)}\left[2\|\hat{\varepsilon}_{G^c}\|_2 + \left(\sum_{j\in F}(\lambda_j^{(\ell-1)})^2\right)^{1/2}\right].$$

Proof Let J be the indices of the largest s coefficients (in absolute value) of $\hat{\mathbf{w}}_G$, and $I = G^c \cup J$. The conditions of the lemma imply that

$$\begin{aligned}\max(0, \Delta\hat{\mathbf{w}}_I^\top A\Delta\hat{\mathbf{w}}) \geq& \rho_-(k+s)[\|\Delta\hat{\mathbf{w}}_I\|_2 - \pi(k+s,s)\|\hat{\mathbf{w}}_G\|_1/s]\|\Delta\hat{\mathbf{w}}_I\|_2 \\ \geq& \rho_-(k+s)[1-(1-t)(4-t)(4-3t)^{-1}]\|\Delta\hat{\mathbf{w}}_I\|_2^2 \\ \geq& 0.5t\rho_-(k+s)\|\Delta\hat{\mathbf{w}}_I\|_2^2.\end{aligned}$$

In the above derivation, the first inequality is due to Lemma 4; the second inequality is due to the conditions of this lemma plus Lemma 8, which implies that

$$\|\hat{\mathbf{w}}_G\|_1 \leq 2\frac{\|\hat{\varepsilon}\|_\infty + \lambda_0}{\lambda_G - 2\|\hat{\varepsilon}\|_\infty}\|\hat{\mathbf{w}}_{G^c}\|_1 \leq \frac{\|\hat{\varepsilon}\|_\infty + \lambda_0}{\lambda_G - 2\|\hat{\varepsilon}\|_\infty}\sqrt{k}\|\hat{\mathbf{w}}_I\|_2;$$

and the last inequality follows from $1-(1-t)(4-t)(4-3t)^{-1} \geq 0.5t$.

If $\Delta\hat{\mathbf{w}}_I^\top \hat{A}\Delta\hat{\mathbf{w}} \leq 0$, then the above inequality, together with Lemma 9, imply the lemma. Therefore in the following, we can assume that

$$\Delta\hat{\mathbf{w}}_I^\top A\Delta\hat{\mathbf{w}} \geq 0.5t\rho_-(k+s)\|\Delta\hat{\mathbf{w}}_I\|_2^2.$$

Moreover, let $\lambda_j = \lambda_j^{(\ell-1)}$. We obtain from (13) with $\mathbf{v} = \Delta\hat{\mathbf{w}}_I$ the following:

$$\begin{aligned}2\Delta\hat{\mathbf{w}}_I^\top\hat{A}\Delta\hat{\mathbf{w}} \leq& -2\Delta\hat{\mathbf{w}}_I^\top\hat{\varepsilon} - \sum_{j\in I}\lambda_j\Delta\hat{\mathbf{w}}_j\mathrm{sgn}(\hat{\mathbf{w}}_j) \\ \leq& 2\|\Delta\hat{\mathbf{w}}_I\|_2\|\hat{\varepsilon}_{G^c}\|_2 + 2\|\hat{\varepsilon}_G\|_\infty\sum_{j\in G}|\Delta\hat{\mathbf{w}}_j| + \sum_{j\in F}\lambda_j|\Delta\hat{\mathbf{w}}_j| - \sum_{j\in G}\lambda_j|\Delta\hat{\mathbf{w}}_j| \\ \leq& 2\|\Delta\hat{\mathbf{w}}_I\|_2\|\hat{\varepsilon}_{G^c}\|_2 + (\sum_{j\in F}\lambda_j^2)^{1/2}\|\Delta\hat{\mathbf{w}}_I\|_2,\end{aligned}$$

where $\lambda_j \geq \lambda_G \geq 2\|\hat{\varepsilon}_G\|_\infty$ is used to derive the last inequality. Now by combining the above two estimates, we obtain

$$\|\Delta\hat{\mathbf{w}}_I\|_2 \leq \frac{1}{t\rho_-(k+s)}\left[2\|\hat{\varepsilon}_{G^c}\|_2 + (\sum_{j\in F}\lambda_j^2)^{1/2}\right].$$

The desired bound now follows from Lemma 9. ■

Lemma 11 *Consider $g(\cdot)$ that satisfies the conditions of Theorem 2. Let $\lambda_j = \lambda g'(|\tilde{\mathbf{w}}_j|)$ for some $\tilde{\mathbf{w}} \in R^d$, then*

$$\left(\sum_{j\in F}\lambda_j^2\right)^{1/2} \le \lambda\left(\sum_{j\in F} g'(|\bar{\mathbf{w}}_j|-\theta)^2\right)^{1/2} + \lambda\theta^{-1}\left(\sum_{j\in F}|\bar{\mathbf{w}}_j-\tilde{\mathbf{w}}_j|^2\right)^{1/2}.$$

Proof By assumption, if $|\bar{\mathbf{w}}_j - \tilde{\mathbf{w}}_j| \ge \theta$, then

$$g'(|\tilde{\mathbf{w}}_j|) \le 1 \le \theta^{-1}|\bar{\mathbf{w}}_j - \tilde{\mathbf{w}}_j|;$$

otherwise, $g'(|\tilde{\mathbf{w}}_j|) \le g'(|\bar{\mathbf{w}}_j|-\theta)$. It follows that the following inequality always holds:

$$g'(|\tilde{\mathbf{w}}_j|) \le g'(|\bar{\mathbf{w}}_j|-\theta) + \theta^{-1}|\bar{\mathbf{w}}_j - \tilde{\mathbf{w}}_j|.$$

The desired bound is a direct consequence of the above result and the 2-norm triangle inequality $(\sum_j (x_j+\Delta x_j)^2)^{1/2} \le (\sum_j x_j^2)^{1/2} + (\sum_j \Delta x_j^2)^{1/2}$. ■

Lemma 12 *Under the conditions of Theorem 2, we have for all $s \ge 2\bar{k}$:*

$$\|\hat{\mathbf{w}}^{(\ell)} - \bar{\mathbf{w}}\|_2 \le \frac{7}{\rho_-(2\bar{k}+s)}\sqrt{|F|}\lambda.$$

Proof Let $t = 0.5$, then using Lemma 5, the condition of the theorem implies that

$$\frac{\lambda + 2\|\hat{\varepsilon}\|_\infty}{\lambda g'(\theta) - 2\|\hat{\varepsilon}\|_\infty} \le \frac{4-t}{4-3t}.$$

Moreover, Lemma 3 implies that the condition

$$t = 0.5 \le 1 - \pi(2\bar{k}+s, s)(2\bar{k})^{0.5}/s$$

is also satisfied.

Now, if we assume that at some $\ell \ge 1$ that

$$|G_\ell^c| \le 2\bar{k}, \quad \text{where } G_\ell = \{j \notin F : \lambda_j^{(\ell-1)} \ge \lambda g'(\theta)\}, \tag{14}$$

then we can obtain from Lemma 10 that

$$\|\hat{\mathbf{w}}^{(\ell)} - \bar{\mathbf{w}}\|_2 \le \frac{1+\sqrt{3}}{t\rho_-(2\bar{k}+s)}\left[2\sqrt{|G_\ell^c|}\|\hat{\varepsilon}\|_\infty + \sqrt{|F|}\lambda\right] \le \frac{3.2}{t\rho_-(2\bar{k}+s)}\sqrt{|F|}\lambda,$$

where we have used the fact that $|G_\ell^c| \le 2\bar{k} \le 2|F|$ and $\lambda \ge 20\|\hat{\varepsilon}\|_\infty$ in the derivation of the second inequality. This shows that (14) implies the lemma.

Therefore next we only need to prove by induction on ℓ that (14) holds for all $\ell = 1, 2, \ldots$. When $\ell = 1$, we have $G_1 = F^c$, which implies that (14) holds.

Now assume that (14) holds at $\ell-1$ for some $\ell > 1$. Since $j \in G_\ell^c - F$ implies that $j \notin F$ and $\lambda g'(|\hat{\mathbf{w}}_j^{(\ell-1)}|) = \lambda_j^{(\ell)} < \lambda g'(\theta)$ by definition, and since $g'(z)$ is non-increasing when $z \geq 0$ (theorem assumption), we know that $|\hat{\mathbf{w}}_j^{(\ell-1)}| \geq \theta$. Therefore by induction hypothesis we obtain

$$\begin{aligned}\sqrt{|G_\ell^c - F|} \leq & \sqrt{\sum_{j \in G_\ell^c - F} |\hat{\mathbf{w}}_j^{(\ell-1)}|^2/\theta^2} \leq \frac{\|\hat{\mathbf{w}}^{(\ell-1)} - \bar{\mathbf{w}}\|_2}{\theta} \\ \leq & \frac{7\lambda}{\rho_-(2k+s)\theta}\sqrt{|F|} \leq \sqrt{|F|},\end{aligned}$$

where the second to the last inequality is due to the fact that (14) implies the lemma at $\ell-1$. The last inequality uses the definition of θ in the theorem. This inequality implies that $|G_\ell^c| \leq 2|F| \leq 2\bar{k}$, which completes the induction step. ■

A.1 Proof of Theorem 2

As in the proof of Lemma 12, if we let $t = 0.5$, then using Lemma 5, the condition of the theorem implies that

$$\frac{\lambda + 2\|\hat{\varepsilon}\|_\infty}{\lambda g'(\theta) - 2\|\hat{\varepsilon}\|_\infty} \leq \frac{4-t}{4-3t}.$$

Moreover, Lemma 3 implies that the condition

$$t = 0.5 \leq 1 - \pi(2\bar{k}+s, s)(2\bar{k})^{0.5}/s$$

is also satisfied.

We prove by induction: for $\ell = 1$, the result follows from Lemma 12. For $\ell > 1$, we let $G^c = F \cup \{j : |\hat{\mathbf{w}}_j^{(\ell-1)}| \geq \theta\}$. From the proof of Lemma 12, we know that

$$k = |G^c| \leq 2\bar{k}.$$

Let $u = \sqrt{\rho_+(\bar{k})}\sigma[\sqrt{7.4\bar{k}/n} + \sqrt{2.7\ln(2/\eta)/n}]$. We know from Lemma 5, and $\lambda \geq 20\|\hat{\varepsilon}\|_\infty$ that with probability $1-2\eta$,

$$\begin{aligned}\|\hat{\varepsilon}_{G^c}\|_2 \leq & \|\hat{\varepsilon}_F\|_2 + \sqrt{|G^c - F|}\|\hat{\varepsilon}\|_\infty \\ \leq & u + \sqrt{|G^c - F|}\lambda/20 \\ \leq & u + (\lambda/20)\sqrt{\sum_{j \in G^c - F} |\hat{\mathbf{w}}_j^{(\ell-1)}|^2/\theta^2} \\ \leq & u + \lambda(20\theta)^{-1}\|\hat{\mathbf{w}}^{(\ell-1)} - \bar{\mathbf{w}}\|_2.\end{aligned}$$

Now, using Lemma 10 and Lemma 11, we obtain

$$\begin{aligned}
&\|\Delta\hat{\mathbf{w}}^{(\ell)}\|_2 \\
\leq&\frac{1+\sqrt{3}}{t\rho_-(k+s)}\left[2\|\hat{\varepsilon}_{G^c}\|_2+(\sum_{j\in F}(\lambda_j^{(\ell-1)})^2)^{1/2}\right] \\
\leq&\frac{1+\sqrt{3}}{t\rho_-(k+s)}\left[2\|\hat{\varepsilon}_{G^c}\|_2+\lambda\left(\sum_{j\in F}g'(|\bar{\mathbf{w}}_j|-\theta)^2\right)^{1/2}+\lambda\theta^{-1}\left(\sum_{j\in F}|\bar{\mathbf{w}}_j-\hat{\mathbf{w}}_j^{(\ell-1)}|^2\right)^{1/2}\right] \\
\leq&\frac{1+\sqrt{3}}{t\rho_-(k+s)}\left[2u+\lambda\left(\sum_{j\in F}g'(|\bar{\mathbf{w}}_j|-\theta)^2\right)^{1/2}+1.1\lambda\theta^{-1}\|\bar{\mathbf{w}}-\hat{\mathbf{w}}^{(\ell-1)}\|_2\right] \\
\leq&\frac{1+\sqrt{3}}{t\rho_-(k+s)}\left[2u+\lambda\left(\sum_{j\in F}g'(|\bar{\mathbf{w}}_j|-\theta)^2\right)^{1/2}\right]+0.67\|\bar{\mathbf{w}}-\hat{\mathbf{w}}^{(\ell-1)}\|_2.
\end{aligned}$$

The desired bound can now be obtained by solving this recursion with respect to

$$\|\Delta\hat{\mathbf{w}}^{(\ell)}\|_2=\|\bar{\mathbf{w}}-\hat{\mathbf{w}}^{(\ell)}\|_2$$

for $\ell=2,3,\ldots$, where $\|\Delta\hat{\mathbf{w}}^{(1)}\|_2$ is given by Lemma 12.

Appendix B. Some Non-convex Formulations in Machine Learning

Consider a set of input vectors $\mathbf{x}_1,\ldots,\mathbf{x}_n\in R^d$, with corresponding desired output variables $y_1,\ldots,y_n$. The task of supervised learning is to estimate the functional relationship $y\approx f(\mathbf{x})$ between the input $\mathbf{x}$ and the output variable y from the training examples $\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}$. The quality of prediction is often measured through a loss function $\phi(f(\mathbf{x}),y)$.

Now, consider linear prediction model $f(\mathbf{x})=\mathbf{w}^\top\mathbf{x}$. As in boosting or kernel methods, nonlinearity can be introduced by including nonlinear features in $\mathbf{x}$. For linear models, we are mainly interested in the scenario that $d\gg n$. That is, there are many more features than the number of samples. In this case, an unconstrained empirical risk minimization is inadequate because the solution overfits the data. The standard remedy for this problem is to impose a constraint on $\mathbf{w}$ to obtain a *regularized* problem. This leads to the following regularized empirical risk minimization method:

$$\hat{\mathbf{w}}=\arg\min_{\mathbf{w}\in R^d}\left[\sum_{i=1}^n\phi(\mathbf{w}^\top\mathbf{x}_i,y_i)+\lambda g(\mathbf{w})\right], \tag{15}$$

where $\lambda>0$ is an appropriately chosen regularization condition. This is the motivation for the general problem formulation (1) in Section 2.

B.1 Loss Function

Examples of loss function $\phi(\mathbf{w}^\top\mathbf{x},y)$ in (15) include least squares for regression: $\phi(\mathbf{w}^\top\mathbf{x},y)=(\mathbf{w}^\top\mathbf{x}-y)^2$, and 0-1 binary classification error: $\phi(\mathbf{w}^\top\mathbf{x},y)=I(\mathbf{w}^\top\mathbf{x}y\leq 0)$, where $y\in\{\pm1\}$ are the class labels, and $I(\cdot)$ is the set indicator function. The latter is nonconvex. In practice, for computational reasons, a convex relaxation such as the SVM loss $\phi(\mathbf{w}^\top\mathbf{x},y)=\max(0,1-\mathbf{w}^\top\mathbf{x}y)$

is often used to substitute the classification error loss. Such a convex loss is often referred to as a surrogate loss function, and the resulting method becomes a convex relaxation method for solving binary classification. This class of methods have been theoretically analyzed in Bartlett et al. (2006) and Zhang (2004). While asymptotically, convex surrogate methods are consistent (that is, they can be used to obtain Bayes optimal classifiers when the sample size approaches infinity), for finite data, these methods can be more sensitive to outliers. In order to alleviate the effect of outliers, one may consider the smoothed classification error loss function $\phi(\mathbf{w}^\top \mathbf{x}, y) = \min(\alpha, \max(0, 1 - \mathbf{w}^\top \mathbf{x} y))$ ($\alpha \geq 1$). This loss function is bounded, and thus more robust to outliers than SVMs under finite sample size; moreover, it is piece-wise differentiable, and thus easier to handle than the discontinuous classification error loss. For comparison purpose, the three loss functions are plotted in Figure 7.

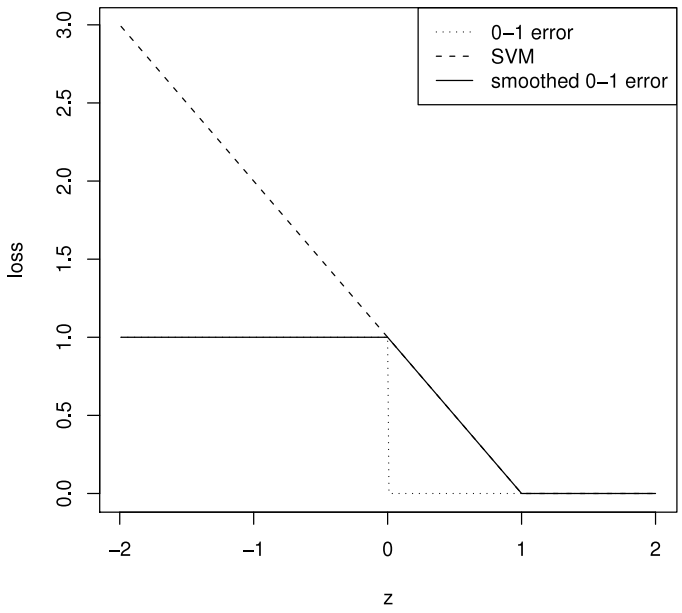

Figure 7: Loss Functions: classification error versus smoothed classification error ($\alpha = 1$) and SVM

B.2 Regularization Condition

Some examples of regularization conditions in (15) include squared regularization $g(\mathbf{w}) = \mathbf{w}^\top \mathbf{w}$, and 1-norm regularization $g(\mathbf{w}) = \|\mathbf{w}\|_1$. The former can be generalized to kernel methods, and the latter leads to sparse solutions. Sparsity is an important regularization condition, which corresponds to the (non-convex) L_0 regularization, defined as $\|\mathbf{w}\|_0 = |\{j : \mathbf{w}_j \neq 0\}| = k$. That is, the measure of complexity is the number of none zero coefficients. If we know the sparsity parameter k for the target vector, then a good learning method is L_0 regularization:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in R^d} \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{w}^\top \mathbf{x}_i, y_i) \quad \text{subject to } \|\mathbf{w}\|_0 \leq k, \tag{16}$$

which applies the standard empirical risk minimization formulation to learning L_0 constrained sparse targets.

If k is not known, then one may regard k as a tuning parameter, which can be selected through cross-validation. This method is often referred to as *subset selection* in the literature. Sparse learning is an essential topic in machine learning, which has attracted considerable interests recently. It can be shown that the solution of the L_0 regularization problem in (16) achieves good prediction accuracy if the target function can be approximated by a sparse $\bar{\mathbf{w}}$. However, a fundamental difficulty with this method is the computational cost, because the number of subsets of $\{1,\ldots,d\}$ of cardinality k (corresponding to the nonzero components of $\mathbf{w}$) is exponential in k.

Due to the computational difficult, in practice, it is necessary to replace (16) by some easier to solve formulations in (15). Specifically, L_0 regularization is equivalent to (15) by choosing the regularization function as $g(\mathbf{w}) = \|\mathbf{w}\|_0$. However, this function is discontinuous. For computational reasons, it is helpful to consider a continuous approximation with $g(\mathbf{w}) = \|\mathbf{w}\|_p^p$, where $p > 0$. If $p \geq 1$, the resulting formulation is convex. In particular, by choosing the closest approximation with $p = 1$, one obtain *Lasso*, which is the standard convex relaxation formulation for sparse learning. With $p \in (0,1)$, the L_p regularizer $\|\mathbf{w}\|_p^p$ is non-convex but continuous.

Supervised learning can be solved using general empirical risk minimization formulation in (15). Both ϕ and g can be non-convex in application problems. The traditional approach is to use convex relaxation to approximate it, leading to a single stage convex formulation. In this paper, we try to extend the idea by looking at a more general multi-stage convex relaxation method, which leads to more accurate approximations.

For illustration, we consider the following examples which will be used in our later discussion.

- Smoothed classification error loss: formulation (15) with convex regularization $g(\mathbf{w})$ and nonconvex loss function (with $\alpha \geq 1$)

$$\phi(\mathbf{w}^\top \mathbf{x}, y) = \min(\alpha, \max(0, 1 - \mathbf{w}^\top \mathbf{x} y)).$$

  This corresponds to $R_0(\mathbf{w}) = \lambda g(\mathbf{w})$, and $R_k(\mathbf{w}) = \phi(\hat{\mathbf{w}}^\top \mathbf{x}_k, y_k)$ for $k = 1,\ldots,n$ in (1).

- L_p regularization ($0 \leq p \leq 1$): formulation (15) with nonconvex regularization $g(\mathbf{w}) = \|\mathbf{w}\|_p^p$ and a loss function $\phi(\cdot,\cdot)$ that is convex in $\mathbf{w}$. This corresponds to $R_0(\mathbf{w}) = n^{-1}\sum_{i=1}^n \phi(\mathbf{w}^\top \mathbf{x}_i, y_i)$, and $R_k(\mathbf{w}) = \lambda |\mathbf{w}_k|^p$ for $k = 1,\ldots,d$ in (1).

- Smoothed L_p regularization (with parameters $\alpha > 0$ and $0 \leq p \leq 1$): formulation (15) with nonconvex regularization $g(\mathbf{w}) = \sum_k [(\alpha + |\mathbf{w}_k|)^p - \alpha^p]/(p\alpha^{p-1})$, and a loss function $\phi(\cdot,\cdot)$ that is convex in $\mathbf{w}$. This corresponds to $R_0(\mathbf{w}) = n^{-1}\sum_{i=1}^n \phi(\mathbf{w}^\top \mathbf{x}_i, y_i)$, and $R_k(\mathbf{w}) = \lambda[(\alpha + |\mathbf{w}_k|)^p - \alpha^p]/(p\alpha^{p-1})$ for $k = 1,\ldots,d$ in (1). The main difference between standard L_p and smoothed L_p is at $|\mathbf{w}_k| = 0$, where the smoothed L_p regularization is differentiable, with derivative 1. This difference is theoretically important as discussed in Section 3.1.

- Smoothed log regularization (with parameter $\alpha > 0$): formulation (15) with nonconvex regularization $g(\mathbf{w}) = \sum_k \alpha \ln(\alpha + |\mathbf{w}_k|)$, and a loss function $\phi(\cdot,\cdot)$ that is convex in $\mathbf{w}$. This corresponds to $R_0(\mathbf{w}) = n^{-1}\sum_{i=1}^n \phi(\mathbf{w}^\top \mathbf{x}_i, y_i)$, and $R_k(\mathbf{w}) = \lambda\alpha \ln(\alpha + |\mathbf{w}_k|)$ for $k = 1,\ldots,d$ in (1). Similar to the smoothed L_p regularization, the smoothed log-loss has derivative 1 at $|\mathbf{w}_k| = 0$.

- Capped-L_1 regularization (with parameter $\alpha > 0$): formulation (15) with nonconvex regularization $g(\mathbf{w}) = \sum_{j=1}^{d} \min(|\mathbf{w}_j|, \alpha)$, and a loss function $\phi(\cdot,\cdot)$ that is convex in $\mathbf{w}$. This corresponds to $R_0(\mathbf{w}) = n^{-1}\sum_{i=1}^{n}\phi(\mathbf{w}^\top \mathbf{x}_i, y_i)$, and $R_k(\mathbf{w}) = \lambda \min(|\mathbf{w}_k|, \alpha)$ for $k = 1,\ldots,d$ in (1). The capped-L_1 regularization is a good approximation to L_0 because as $\alpha \to 0$, $\sum_k \min(|\mathbf{w}_k|, \alpha)/\alpha \to \|\mathbf{w}\|_0$. Therefore when $\alpha \to 0$, this regularization condition is equivalent to the sparse L_0 regularization up to a rescaling of λ. Capped-L_1 regularization is a simpler but less smooth version of the SCAD regularization (Jianqing Fan, 2001). SCAD is more complicated, but its advantage cannot be shown through our analysis.

Appendix C. Some Examples of Multi-stage Convex Relaxation Methods

The multi-stage convex relaxation method can be used with examples in Section 2.2 to obtain concrete algorithms for various formulations. We describe some examples here.

C.1 Smoothed Classification Loss

We consider a loss term of the form $R_k(\mathbf{w}) = \min(\alpha, \max(0, 1 - \mathbf{w}^\top \mathbf{x}_k y_k))$ for $k = 1,\ldots,n$ (with $\alpha \geq 1$), and relax it to the SVM loss $\mathbf{h}_k(\mathbf{w}) = \max(0, 1 - \mathbf{w}^\top \mathbf{x}_k y_k)$.

The optimization problem is

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[\sum_{i=1}^{n} \min(\alpha, \max(0, 1 - \mathbf{w}^\top \mathbf{x}_i y_i)) + \lambda g(\mathbf{w}) \right],$$

where we assume that $g(\mathbf{w})$ is a convex regularization condition such as $g(\mathbf{w}) = \lambda \|\mathbf{w}\|_2^2$.

Consider concave duality in Section 2.2. Each $\mathbf{u}_k$ is a scalar in the range $\Omega_k = [0,\infty)$, and $\bar{R}_k(\mathbf{u}_k) = \min(\alpha, \mathbf{u}_k)$. We have $R_k^*(\mathbf{v}_k) = \alpha(\mathbf{v}_k - 1) I(\mathbf{v}_k \in [0,1])$, defined on the domain $\mathbf{v}_k \geq 0$. The solution in (4) is given by $\hat{\mathbf{v}}_k = I(\mathbf{w}^\top \mathbf{x}_k y_k \geq 1 - \alpha)$ for $k = 1,\ldots,n$. Therefore Section 2.2 implies that the multi-stage convex relaxation solves the weighted SVM formulation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[\sum_{i=1}^{n} \hat{\mathbf{v}}_i \max(0, 1 - \mathbf{w}^\top \mathbf{x}_i y_i) + \lambda g(\mathbf{w}) \right],$$

where the relaxation parameter $\mathbf{v}$ is updated as

$$\hat{\mathbf{v}}_i = I(\hat{\mathbf{w}}^\top \mathbf{x}_i y_i \geq 1 - \alpha) \quad (i = 1,\ldots,n).$$

Intuitively, the mis-classified points $\hat{\mathbf{w}}^\top \mathbf{x}_i y_i < 1 - \alpha$ are considered as outliers, and ignored.

C.2 L_p and Smoothed L_p Regularization

In sparse regularization, we may consider a regularization term $R_k(\mathbf{w}) = \lambda |\mathbf{w}_k|^p / p$ ($k = 1,\ldots,d$) for some $p \in (0,1)$. Given any $q > p$, (3) holds with $\mathbf{u}_k = \mathbf{h}_k(\mathbf{w}) = |\mathbf{w}_k|^q \in [0,\infty)$, and $\bar{R}_k(\mathbf{u}_k) = \lambda |\mathbf{u}_k|^{p/q}/p$, where $\mathbf{u}_k \in \Omega_k = [0,\infty)$. The dual is $R_k^*(\mathbf{v}_k) = -\lambda c(p,q)(\mathbf{v}_k/\lambda)^{p/(p-q)}$, defined on the domain $\mathbf{v}_k \geq 0$, where $c(p,q) = (q-p)p^{-1}q^{q/(p-q)}$. The solution in (4) is given by $\hat{\mathbf{v}}_k = (\lambda/q)|\mathbf{w}_k|^{p-q}$.

An extension is to consider a regularization term $R_k(\mathbf{w}) = \lambda[(\alpha + |\mathbf{w}_k|)^p - \alpha^p]/(p\alpha^{p-1})$ ($k = 1,\ldots,d$) for some $p \in (0,1)$ and $\alpha > 0$. Given any $q > p$, (3) holds with $\mathbf{u}_k = \mathbf{h}_k(\mathbf{w}) = (\alpha +$

$|\mathbf{w}_k|)^q \in [\alpha^q, \infty)$, and $\bar{R}_k(\mathbf{u}_k) = \lambda[\mathbf{u}_k^{p/q} - \alpha^p]/(p\alpha^{p-1})$, where $\mathbf{u}_k \in \Omega_k = [0,\infty)$. The dual is $R_k^*(\mathbf{v}_k) = -\lambda c(p,q)\alpha^{p/(p-q)}(\mathbf{v}_k/\lambda)^{p/(p-q)} + \lambda\alpha/p$, defined on the domain $\mathbf{v}_k \geq 0$, where $c(p,q) = (q-p)p^{p/(q-p)}q^{q/(p-q)}$. The solution in (4) is given by $\hat{\mathbf{v}}_k = \lambda/(q\alpha^{p-1})(\alpha + |\mathbf{w}_k|)^{p-q}$.

An alternative is to relax smoothed L_p regularization ($p \in (0,1)$) directly to L_q regularization for $q \geq 1$ (one usually takes either $q = 1$ or $q = 2$). In this case, $\mathbf{u}_k = \mathbf{h}_k(\mathbf{w}) = |\mathbf{w}_k|^q \in [0,\infty)$, and $\bar{R}_k(\mathbf{u}_k) = \lambda[(\alpha + \mathbf{u}_k^{1/q})^p - \alpha^p]/(p\alpha^{p-1})$. Although it is not difficult to verify that $\bar{R}_k(\mathbf{u}_k)$ is concave, we do not have a simple closed form for $R_k^*(\mathbf{v}_k)$. However, it is easy to check that the solution in (4) is given by $\hat{\mathbf{v}}_k = \lambda/(q\alpha^{p-1})(\alpha + |\mathbf{w}_k|)^{p-1}|\mathbf{w}_k|^{1-q}$.

In summary, for L_p and smoothed L_p, we consider the following optimization formulation for some $\alpha \geq 0$ and $p \in (0,1]$:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[R_0(\mathbf{w}) + \lambda \sum_{j=1}^{d} (\alpha + |\mathbf{w}_j|)^p \right],$$

where we assume that $R_0(\mathbf{w})$ is a convex function of $\mathbf{w}$.

From previous discussion, the multi-stage convex relaxation method in Section 2.2 becomes a weighted L_q regularization formulation for $q \geq 1$:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[R_0(\mathbf{w}) + \sum_{j=1}^{d} \hat{\mathbf{v}}_j |\mathbf{w}_j|^q \right],$$

where the relaxation parameter $\mathbf{v}$ is updated as

$$\hat{\mathbf{v}}_j = \lambda(p/q)(\alpha + |\hat{\mathbf{w}}_j|)^{p-1}|\hat{\mathbf{w}}_j|^{1-q} \quad (j = 1,\ldots,d).$$

The typical choices of q are $q = 1$ or $q = 2$. That is, we relax L_p regularization to L_1 or L_2 regularization.

Finally, we note that the two stage version of L_p regularization, relaxed to L_q with $q = 1$, is referred to Adaptive-Lasso (Zou, 2006).

C.3 Smoothed log Regularization

This is a different sparse regularization condition, where we consider a regularization term $R_k(\mathbf{w}) = \lambda\alpha\ln(\alpha + |\mathbf{w}_k|)$ for some $\alpha > 0$. Given any $q > 0$, (3) holds with $\mathbf{u}_k = \mathbf{h}_k(\mathbf{w}) = (\alpha + |\mathbf{w}_k|)^q \in [\alpha^q, \infty)$, and $\bar{R}_k(\mathbf{u}_k) = \lambda(\alpha/q)\ln(\mathbf{u}_k)$, where $\mathbf{u}_k \in \Omega_k = [0,\infty)$. The dual is $R_k^*(\mathbf{v}_k) = \lambda(\alpha/q)[\ln\mathbf{v}_k + 1 - \ln(\lambda\alpha/q)]$, defined on the domain $\mathbf{v}_k \geq 0$. The solution in (4) is given by $\hat{\mathbf{v}}_k = \lambda(\alpha/q)(\alpha + |\mathbf{w}_k|)^{-q}$.

Similar to smoothed L_p, we may relax directly to L_q, with $\mathbf{u}_k = \mathbf{h}_k(\mathbf{w}) = |\mathbf{w}_k|^q \in [0,\infty)$. $\bar{R}_k(\mathbf{u}_k) = \lambda\alpha\ln(\alpha + \mathbf{u}_k^{1/q})$, where The solution in (4) is given by $\hat{\mathbf{v}}_k = \lambda(\alpha/q)(\alpha + |\mathbf{w}_k|)^{-1}|\mathbf{w}_k|^{1-q}$.

Similar to smoothed log regularization, the multi-stage convex relaxation method in Section 2.2 becomes a weighted L_q regularization formulation for $q \geq 1$:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[R_0(\mathbf{w}) + \sum_{j=1}^{d} \hat{\mathbf{v}}_j |\mathbf{w}_j|^q \right],$$

where the relaxation parameter $\mathbf{v}$ is updated as

$$\hat{\mathbf{v}}_j = \lambda(\alpha/q)(\alpha + |\mathbf{w}_j|)^{-1}|\mathbf{w}_j|^{1-q} \quad (j = 1,\ldots,d).$$

This resulting procedure is the same as the one empirically studied in Candes et al. (2008).

C.3.1 CAPPED L_1 REGULARIZATION

We consider another sparse regularization term with $R_k(\mathbf{w}) = \lambda \min(|\mathbf{w}_k|, \alpha)$ ($k = 1, \ldots, d$) for some $\alpha > 0$. In this case, (3) holds with $\mathbf{u}_k = \mathbf{h}_k(\mathbf{w}) = |\mathbf{w}_k| \in [0, \infty)$, and $\bar{R}_k(\mathbf{u}_k) = \lambda \min(\mathbf{u}_k, \alpha)$, where $\mathbf{u}_k \in \Omega_k = [0, \infty)$. The dual is $R_k^*(\mathbf{v}_k) = \lambda\alpha(-1 + \mathbf{v}_k/\lambda) I(\mathbf{v}_k \in [0, \lambda])$ defined $[0, \infty)$, where $I(\cdot)$ is the set indicator function. The solution in (4) is given by $\hat{\mathbf{v}}_k = \lambda I(|\mathbf{w}_k| \leq \alpha)$.

In capped L_1 regularization, we consider the optimization problem

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[R_0(\mathbf{w}) + \lambda \sum_{j=1}^{d} \min(\alpha, |\mathbf{w}_j|) \right],$$

where we assume that $R_0(\mathbf{w})$ is a convex function of $\mathbf{w}$.

From Section 2.2, the multi-stage convex relaxation becomes a weighted L_1 regularization formulation:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[R_0(\mathbf{w}) + \sum_{j=1}^{d} \hat{\mathbf{v}}_j |\mathbf{w}_j| \right],$$

where the relaxation parameter $\mathbf{v}$ is updated as

$$\hat{\mathbf{v}}_j = \lambda I(|\hat{\mathbf{w}}_j| \leq \alpha) \quad (j = 1, \ldots, d).$$

This method has an intuitive interpretation: in order to achieve sparsity, the standard L_1 regularization not only shrinks small coefficients to zero, but also shrinks large coefficients. This causes a bias. The capped-L_1 formulation removes the bias by adaptively adjusting the relaxation parameter $\hat{\mathbf{v}}_j$ so that if $|\hat{\mathbf{w}}_j|$ is large, then we do not penalize the corresponding variable j.

C.3.2 SPARSE EIGENVALUE PROBLEM

We use a simple example to illustrate that the multi-stage convex relaxation idea does not only apply to formulations with convex risks. Consider the sparse eigenvalue problem, where we are interested in finding the largest eigenvalue of a positive semi-definite matrix A. One formulation is

$$\hat{\mathbf{w}} = \arg\max_{\|\mathbf{w}\|_2 \leq 1} \left[\mathbf{w}^\top A \mathbf{w} - \lambda \sum_{j=1}^{d} (\alpha + |\mathbf{w}_j|)^p \right],$$

with parameter $p \in (0, 1)$ and a small parameter $\alpha > 0$ to encourage sparsity. If $\lambda = 0$, then it is the standard eigenvalue problem without sparsity constraints. Although the standard eigenvalue problem is not convex in $\mathbf{w}$, it has a convex relaxation to a semi-definite programming problem, and thus can be efficiently solved. For convenience, we think of the standard eigenvalue problem as "convex" for the purpose of this paper. The multi-stage convex relaxation becomes:

$$\hat{\mathbf{w}} = \arg\max_{\|\mathbf{w}\|_2 \leq 1} \left[\mathbf{w}^\top A \mathbf{w} - \sum_{j=1}^{d} \mathbf{v}_j \mathbf{w}_j^2 \right],$$

which is a standard eigenvalue problem. The relaxation parameter is updated as

$$\hat{\mathbf{v}}_j = \lambda(p/2)(\alpha + |\hat{\mathbf{w}}_j|)^{p-1} |\hat{\mathbf{w}}_j|^{-1} \quad (j = 1, \ldots, d).$$

C.3.3 MATRIX REGULARIZATION

Our final example is multi-task learning with matrix regularization, also considered in Argyriou et al. (2008). In this case, $\mathbf{w}$ is not a vector, but a matrix, with columns (tasks) $\mathbf{w}^\ell$. We solve a problem of the following form:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \left[\sum_{\ell=1}^{m} R^\ell(\mathbf{w}^\ell) + \lambda \mathrm{tr}((\alpha I + \mathbf{w}\mathbf{w}^\top)^{p/2}) \right].$$

In the above formulation, R^ℓ is the risk function for task ℓ. The matrix regularization used here is the counterpart of L_p regularization for vectors. It encourages low-rank if $p < 2$. In particular, the case of $p = 1$ is often called trace norm (or nuclear norm). It is convex and frequently used in the literature. The parameter $\alpha > 0$ gives some smoothness, similar to the vector case.

The case of $p < 1$ gives better low-rank approximation, similar to the vector regularization case. Again, this problem can be solved with multi-stage convex relaxation method. In this case, the relaxation parameter $\mathbf{v}$ is a positive semi-definite matrix, and we relax the regularization term to $\mathbf{h}(\mathbf{w}) = (\alpha I + \mathbf{w}\mathbf{w}^\top)$ as a matrix. Thus the relaxed regularization term becomes $\mathrm{tr}(\mathbf{v}(\alpha I + \mathbf{w}\mathbf{w}^\top))$. This regularization decouples the problems as follows, which allows us to solve each task ℓ separately:

$$\hat{\mathbf{w}}^\ell = \arg\min_{\mathbf{w}^\ell} \left[R^\ell(\mathbf{w}^\ell) + (\mathbf{w}^\ell)^\top \hat{\mathbf{v}} \mathbf{w}^\ell \right] \qquad (\ell = 1, 2, \ldots, m).$$

This is a key advantage of the method. Similar to the vector case, we have the following update formula for the relaxation parameter:

$$\hat{\mathbf{v}} = \lambda(p/2)(\alpha I + \hat{\mathbf{w}}\hat{\mathbf{w}}^\top)^{(p-2)/2}.$$

References

Andreas Argyriou, Charles A. Micchelli, Massimiliano Pontil, and Yiming Ying. A spectral regularization framework for multi-task structure learning. In *NIPS'07*, 2008.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Peter Bickel, Yaacov Ritov, and Alexandre Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

Florentina Bunea, Alexandre Tsybakov, and Marten H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, 2005.

Emmanuel J. Candes, Michael B. Wakin, and Stephen P. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

Runze Li Jianqing Fan. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré*, 2008.

Gilles Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge University Press, 1989.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

Alan L. Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, 15: 915–936, 2003.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 2010. to appear.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004. with discussion.

Tong Zhang. Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009a. ISSN 0090-5364. doi: 10.1214/08-AOS659.

Tong Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *NIPS'08*, 2009b.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.