Multi-Stage Convex Relaxation Methods for Machine Learning

Tong Zhang
Rutgers University

March 2013
Final Report

DISTRIBUTION A: Approved for public release.
Many problems in machine learning can be naturally formulated as non-convex optimization problems. However, such direct nonconvex formulations have been largely replaced by convex relaxation methods (such as support vector machines for classification or L1 regularization for sparse learning) to avoid the usual local optima issues. Many significant theoretical results have been developed in recent years to show that the convex relaxation approach solves the original problem asymptotically. However in practice, the standard simple convex relaxation schemes can be sub-optimal.

In the proposed work, we consider a more general framework of multi-stage convex relaxation methods, which remedies the above gap between theory and practice. 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. We will develop theoretical properties of this method and algorithmic consequences. Related convex and nonconvex machine learning methods will also be investigated.
Multi-Stage Convex Relaxation Methods for Machine Learning

Award FA9550-09-1-0425
Program managers: Jun Zhang and Jay Myung

PI: Tong Zhang
Statistics Department
Rutgers University, NJ 08854

February 8, 2013

1 Introduction

We consider the general regularization framework for machine learning, where a risk (or 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.

In this research project, we consider a more general framework of multi-stage convex relaxation methods, which remedies the above gap between theory and practice. 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. Since each stage is a convex optimization problem, the approach is computationally efficient. Moreover, using mathematical tools from convex analysis, we can analyze the effectiveness of the resulting procedure. This research can significantly improve the widely used convex relaxation methods in machine learning, by extending the standard one-stage convex learning algorithms to more general and sophisticated multi-stage convex learning algorithms that are both computationally efficient and theoretically superior.

2 Scientific Objectives of Research

The combination of regularization and risk minimization is essential in modern machine learning. We shall first motivate this class of learning algorithms from supervised learning as follows. Consider a set of input vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, with corresponding desired output variables $y_1, \ldots, y_n$. The
task of supervised learning is to estimate the functional relationship $y \approx f(x)$ between the input $x$ and the output variable $y$ from the training examples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. The quality of prediction is often measured through a loss function $\phi(f(x), y)$. In this work, we are especially interested in linear prediction model $f(x) = w^T x$. As in boosting or kernel methods, nonlinearity can be easily incorporated in our approach by including nonlinear features in $x$. Hence we shall focus our description on linear models for simplicity. 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 $w$ to obtain a regularized problem. This leads to the following regularized empirical risk minimization method:

$$\hat{w} = \arg \min_{w \in \mathbb{R}^d} \left[ \sum_{i=1}^{n} \phi(w^T x_i, y_i) + \lambda g(w) \right], \quad (1)$$

Supervised learning can be solved using general empirical risk minimization formulation in (1). Both $\phi$ 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 the proposed work, we try to extend this idea, by looking at a more general multi-stage convex relaxation method, which leads to more accurate approximations.

We consider an optimization formulation more general than (1) as follows:

$$\hat{w} = \arg \min_w R(w),$$

$$R(w) = R_0(w) + \sum_{k=1}^{K} R_k(w), \quad (2)$$

where $R(w)$ is the general form of a regularized objective function. Moreover, for convenience, we assume that $R_0(w)$ is convex in $w$, and each $R_k(w)$ is non-convex. In the proposed work, we shall employ convex/concave duality to derive convex relaxations of (2) that can be efficiently solved. More generally, we will study computational procedures and develop statistical theory for nonconvex formulations.

3 **Technical Approach**

We are specifically interested in sparse estimation problems, and try to understand the effectiveness of convex methods versus nonconvex methods. Of special interests, we want to investigated the so-called multi-stage convex relaxation approach described as follows. We consider a single nonconvex component $R_k(w)$ in (2), which we shall rewrite using concave duality. Let $h_k(w) : \mathbb{R}^d \to \Omega_k \subset \mathbb{R}^{d_k}$ be a vector function with range $\Omega_k$. It may not be a one-to-one map. However, we assume that there exists a function $\bar{R}_k$ defined on $\Omega_k$ so that we can express $R_k(w)$ as

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

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

$$R_k(w) = \inf_{v_k \in \mathbb{R}^{d_k}} [v_k^T h_k(w) + \bar{R}_k^*(v_k)] \quad (3)$$

2
using concave duality. In this case, $R^*_k(v_k)$ is the concave dual of $\bar{R}_k(u_k)$ given below

$$R^*_k(v_k) = \inf_{u_k \in \Omega_k} [-v_k^T u_k + \bar{R}_k(u_k)].$$

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

$$\hat{v}_k = \nabla_u \bar{R}_k(u)|_{u = h_k(w)}.$$  

(4)

This is a very general framework.

Using concave duality given in the previous section, we can derive a general convex relaxation based procedure for solving (2).

Let $h_k(w)$ be a convex relaxation of $R_k(w)$ that dominates $R_k(w)$ (for example, it can be the smallest convex upperbound (i.e., the inf over all convex upperbounds) of $R_k(w)$). A simple convex relaxation of (1) becomes

$$\hat{w} = \arg \min_{w \in \mathbb{R}^d} \left[ R_0(w) + \sum_{k=1}^K h_k(w)^T v_k \right].$$

(5)

This simple relaxation yields a solution that is different from the solution of (1). However, it is possible to write $R_k(w)$ using (3). Now, with this new representation, we can rewrite (1) as

$$[\hat{w}, \hat{v}] = \arg \min_{w, \{v_k\}} \left[ R_0(w) + \sum_{k=1}^K (h_k(w)^T v_k + R^*_k(v_k)) \right].$$

(6)

This is clearly equivalent to (1) because of (3). If we can find a good approximation of $\hat{v} = \{\hat{v}_k\}$ that improves upon the initial value of $\hat{v}_k = 1$, then the above formulation can lead to a refined convex problem in $w$ that is a better convex relaxation than (5). Our numerical procedure exploits the above fact, which tries to improve the estimation of $v_k$ over the initial choice of $v_k = 1$ 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 $w$ with $v$ fixed: this is a convex problem in $w$ with appropriately chosen $h(w)$.
- Second we optimize $v$ with $w$ fixed: although non-convex, it has a closed form solution that is given by (4).

The general procedure is presented in Figure 1.

4 Progress Made & Results Obtained

I have made several major progresses during this research project. In particular, I studied the theoretical properties of multi-stage convex relaxation for sparse recovery problems. The analysis resulted in one paper in JMLR and one paper in the Bernoulli journal that analyzed multi-stage convex relaxation for sparse regularization. These papers showed that in comparison to standard convex relaxation with Lasso (L1 regularization), the multi-stage convex relaxation method can
Initialize $\hat{v} = 1$

Repeat the following two steps until convergence:

- Let
  \[
  \hat{w} = \text{arg min}_w \left[ R_0(w) + \sum_{k=1}^{K} h_k(w)^T \hat{v}_k \right]. 
  \] (7)

- Let $\hat{v}_k = \nabla_u R_k(u) \big|_{u = h_k(w)}$ ($k = 1, \ldots, K$)

Figure 1: Multi-stage Convex Relaxation Method

recover sparse target more accurately by solving appropriate nonconvex objective functions with sparse regularization. Moreover, the solutions can be obtained efficiently.

In sparse recovery, we observe a set of input vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, with corresponding desired output variables $y_1, \ldots, y_n$. In general, we may assume that there exists a target $\bar{w} \in \mathbb{R}^d$ such that
\[
y_i = \bar{w}^T x_i + \epsilon_i \quad (i = 1, \ldots, n),
\] (8)
where $\epsilon_i$ are zero-mean independent random noises (but not necessarily identically distributed). Moreover, we assume that the target vector $\bar{w}$ is sparse. That is, there exists $\bar{k} = \|\bar{w}\|_0$ is small.

Let $y$ denote the vector of $[y_i]$ and $X$ be the $n \times d$ matrix with each row a vector $x_i$. We are interested in recovering $\bar{w}$ from noisy observations using the following sparse regression method:
\[
\hat{w} = \text{arg min}_w \left[ \frac{1}{n} \|Xw - y\|_2^2 + \lambda \sum_{j=1}^{d} g(|w_j|) \right], 
\] (9)
where $g(|w_j|)$ 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 $[g(|w_1|), \ldots, g(|w_d|)]$ is concave with respect to $h(w) = [|w_1|^q, \ldots, |w_d|^q]$ for some $q \geq 1$. It follows that (9) can be solved using the multi-stage convex relaxation algorithm. The main difficulty is the nonconvexity of the formulation, which hasn’t been successfully studied previously. We overcome this difficulty by introducing new techniques that allow us to obtain strong theoretical results on the procedure. The results can be summarized as follows: under standard conditions, multi-stage convex relaxation with appropriate nonconvex regularizer $g(w)$ gives a solution that recovers the support of the true target vector $\bar{w}$ after no more than $O(\log(s))$ stages where $s = \|\bar{w}\|_0$ is the sparsity of the true target.

In addition to these results on multi-stage convex relaxation, the PI has also looked at a number of research directions during the past year. These studies have been fruitful, and resulted in many conference/journal publications that are supported by the grant. Specifically, I extended fundamental theoretical investigation of regularization methods and studied new application problems.

- Together with collaborators in USC, we applied multi-stage convex relaxation to the problem of finding co-expressions in multiple biological networks. The work appeared in Plos Computational Biology.

- I studied fundamental properties and limitations of convex L1 regularization problem, and published results in the Annals of statistics.
• Together with student Junzhou Huang, we investigated the structured sparsity problems and published results in the Annals of statistics and in JMLR.

• Together with collaborator John Langford at Yahoo (who has moved to Microsoft), we applied non-convex procedures to time-series prediction problems. The work appeared in ICML.

• I worked with Dr Wan at MSU on another application of nonconvex regularization to influenza prediction, which resulted in several papers appeared in bioinformatics journals such as Plos Computational Biology.

• I applied nonconvex analysis to image recognition problems, jointly with Kai Yu’s group at NEC. We successfully improved state of the art image classification algorithms. This resulted in several conference publications in NIPS, ICML, and ECCV. Moreover, the technique was used in the winning system of ImageNet large scale image classification Challenge in 2010 (http://www.image-net.org/challenges/LSVRC/2010).

• I investigated greedy algorithms for solving nonconvex formulations, resulted in several journal papers in top machine learning, optimization, and engineering journals such as JMLR, SIAM Journal on optimization, and IEEE Transaction on information theory.

• I studied matrix regularization problems for robust matrix reconstruction. This is a relatively new problem with many applications that have drawn significant attention. We studied new algorithms and analysis for this problem, together with postdoc Daniel Hsu and collaborator Sham Kakade at U Penn. This resulted in a journal paper in IEEE Trans. Information Theory. Moreover, we studied some new spectral algorithms for nonconvex formulations such as the hidden Markov model problem, and presented new solutions; the resulting work was published in NIPS and Journal of Computer and System Sciences.

• I worked with professor Cunhui Zhang at Rutgers on extending theoretical investigations of multi-stage convex relaxation, which resulted in one paper on general nonconvex formulation published in Statistical Science journal.

• I worked with a graduate student Dai Dong on model averaging methods that can greatly improve prediction accuracy. This resulted in a paper published in the Annals of statistics.

5 Significance of Results & Impact on Science

My results on multi-stage convex relaxation was the first major result that demonstrated the possibility to work with nonconvex optimization, and design a provably efficient algorithm to find a local optimal solution superior to standard convex relaxation solution. Experiments demonstrated the superiority of the multi-stage procedure as well. This important milestone rigorously shows that the multi-stage convex relaxation is viable choice for nonconvex problems, which allows us to expand into more general problems and applications. The general work supported by this research, described in the previous section with publications listed in Section 6 have made significant impact in the scientific community. To show this, I will list the Google scholar citations of some papers resulted from this research grant:

• [Huang and Zhang, 2010] (combined with arxiv version): 120
6 Publications Resulted from Research


