AFRL-OSR-VA-TR-2013-0111

Multi-Stage Convex Relaxation Methods for Machine Learning

Tong Zhang

Rutgers University

March 2013
Final Report

DISTRIBUTION A: Approved for public release.

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
08-02-2013

2. REPORT TYPE
FINAL

3. DATES COVERED *(From - To)*
May 2009 to June 2012

4. TITLE AND SUBTITLE
Multi-Stage Convex Relaxation Methods for Machine Learning

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-09-1-0425

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Tong Zhang

5d. PROJECT NUMBER
09NL083

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Rutgers University
Piscataway, NJ 08854

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 North Randolph Street
Suite 325, Room 3112
Arlington, Virginia 22203-1768

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2013-0111

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

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.

15. SUBJECT TERMS
regularization, convex relaxation, machine learning, nonconvex optimization

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE
U	U	U

17. LIMITATION OF ABSTRACT
U

18. NUMBER OF PAGES

19a. NAME OF RESPONSIBLE PERSON
Dr. Jay Myung

19b. TELEPHONE NUMBER *(Include area code)*
703-696-8478

Reset

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 $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^d$, 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 $\phi(f(\mathbf{x}), y)$. In this work, we are especially interested in linear prediction model $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. As in boosting or kernel methods, nonlinearity can be easily incorporated in our approach by including nonlinear features in $\mathbf{x}$. 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 $\mathbf{w}$ to obtain a *regularized* problem. This leads to the following regularized empirical risk minimization method:

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

Supervised learning can be solved using general empirical risk minimization formulation in (1). 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 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:

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

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

where $R(\mathbf{w})$ 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 (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 $R_k(\mathbf{w})$ in (2), 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 range Ω_k. 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:

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

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

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

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

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

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

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

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).

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 $\hat{\mathbf{v}} = \mathbf{1}$

Repeat the following two steps until convergence:

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

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 $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^d$, 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

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

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.

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. 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 $g(\mathbf{w})$ gives a solution that recovers the support of the true target vector $\bar{\mathbf{w}}$ after no more than $O(\log(s))$ stages where $s = \|\bar{\mathbf{w}}\|_0$ 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 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

- [Yu, Zhang, and Gong, 2009]: 114
- [Zhang, 2009a]: 114
- [Huang, Zhang, and Metaxas, 2011] (combined with conference version): 113
- [Langford, Li, and Zhang, 2009a]: 106
- [Zhang, 2011a] (combined with conference version): 105
- [Zhou, Yu, Zhang, and Huang, 2010]: 75
- [Hsu, Kakade, and Zhang, 2012] (combined with conference version): 74
- [Zhang, 2009b]: 64
- [Zhang, 2010] (combined with early conference version): 44

6 Publications Resulted from Research

Animashree Anandkumar, Kamalika Chaudhuri, Daniel Hsu, Sham M. Kakade, Le Song, and Tong Zhang. Spectral methods for learning multivariate latent tree structure. In *NIPS' 11*, 2011.

Zhipeng Cai, Tong Zhang, and Xiu-Feng Wan. A computational framework for influenza antigenic cartography. *PLoS Comput Biol*, 6(10):e1000949, 10 2010. doi: 10.1371/journal.pcbi.1000949. URL `http://dx.doi.org/10.1371%2Fjournal.pcbi.1000949`.

Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy Q-aggregation. *Annals of Statistics*, 40:1878–1905, 2012.

Daniel Hsu, Sham M. Kakade, John Langford, and Tong Zhang. Multi-label prediction via compressed sensing. In *NIPS' 09*, 2009.

Daniel Hsu, Sham Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Trans. Info. Th.*, 57:7221–7234, 2011.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Junzhou Huang and Tong Zhang. The benefit of group sparsity. *Annals of Statistics*, 38:1978–2004, 2010.

Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12:3371–3412, 2011.

John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009a.

John Langford, Ruslan Salakhutdinov, and Tong Zhang. Learning nonlinear dynamic models. In *ICML' 09*, 2009b.

Wenyuan Li, Chun-Chi Liu, Tong Zhang, Haifeng Li, Michael S. Waterman, and Xianghong Jasmine Zhou. Integrative analysis of many weighted co-expression networks using tensor computation. *PLoS Comput Biol*, 7(6):e1001106, 06 2011. doi: 10.1371/journal.pcbi.1001106. URL `http://dx.doi.org/10.1371%2Fjournal.pcbi.1001106`.

Yuanqing Lin, Tong Zhang, Shenghuo Zhu, and Kai Yu. Deep coding network. In *NIPS' 10*, 2010.

Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *Siam Journal on Optimization*, 20:2807–2832, 2010.

Lin Xiao and Tong Zhang. A proximal-gradient homotopy method for the L1-regularized least-squares problem. In *ICML'12*, 2012.

Kai Yu and Tong Zhang. Improved local coordinate coding using local tangents. In *ICML' 10*, 2010.

Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. In *NIPS' 09*, 2009.

Cunhui Zhang and Tong Zhang. A general theory of concave regularization for high dimensional sparse estimation problems. *Statistical Science*, 27:576–593, 2012.

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. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009b.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1087–1107, 2010.

Tong Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57:4689–4708, 2011a.

Tong Zhang. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Transactions on Information Theory*, 57:6215 – 6221, 2011b.

Tong Zhang. Multistage convex relaxation for feature selection. *Bernoulli*, 2012.

Xi Zhou, Kai Yu, Tong Zhang, and Thomas Huang. Image classification using super-vector coding of local image descriptors. In *ECCV'10*, 2010.