Neurocomputing 152 (2015) 151–163

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Damping proximal coordinate descent algorithm for non-convex regularization

Zheng Pan, Ming Lin, Guangdong Hou, Changshui Zhang*

Department of Automation, Tsinghua University, State Key Lab of Intelligent Technologies and Systems,
Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, PR China

ARTICLE INFO

Article history:
Received 9 June 2014
Received in revised form
2 September 2014
Accepted 2 November 2014
Communicated by Huaping Liu
Available online 10 November 2014

Keywords:
Non-convex regularization
Non-convex optimization
Coordinate descent
Sparsity regularization

ABSTRACT

Non-convex regularization has attracted much attention in the fields of machine learning, since it is unbiased and improves the performance on many applications compared with the convex counterparts. The optimization is important but difficult for non-convex regularization. In this paper, we propose the Damping Proximal Coordinate Descent (DPCD) algorithms that address the optimization issues of a general family of non-convex regularized problems. DPCD is guaranteed to be globally convergent. The computational complexity of obtaining an approximately stationary solution with a desired precision is only linear to the data size. Our experiments on many machine learning benchmark datasets also show that DPCD has a fast convergence rate and it reduces the time of training models without significant loss of prediction accuracy.

1. Introduction

Regularization is an important way of introducing prior knowledge in machine learning, e.g., the priors of max margin in SVM and the sparsity in LASSO. From the view of optimization, regularization can be separated into convex and non-convex ones.

Beyond convex regularizers, non-convex regularization is unbiased for statistical inference [9,10]. Tremendous works have revealed the superiority of *non-convex regularizers* in many applications, e.g., natural image prior with ℓ_q-regularization ($0 < q < 1$) [12,24,17], model selection with SCAD [9], MCP [27] and Capped-ℓ_1 [29,22] regularization, image inpainting and denoising with MCP regularization [23], compressed sensing with LSP [5] and ℓ_q-regularization [13], robust anisotropic diffusion with Lorentzian error norm [3].

The optimization of non-convex regularization problems is usually expensive or even intractable due to their non-convexity. It cannot be guaranteed to obtain global optimal solutions for general non-convex regularization problems, but it has been proved that the local optima are also the exact or well approximated global optima under proper conditions [27,28,20]. It is useful for non-convex regularization to propose algorithms that have the following three good properties:

1. They are suitable for a general family of regularizers. The regularizers may be highly non-convex and even non-smooth or non-continuous, e.g., ℓ_0-norm.
2. They have global convergence. For arbitrary initial solutions, the algorithms always converge and the limit points are stationary points or well approximated stationary points. Global convergence is not trivial for non-convex optimization, especially for non-smooth or non-continuous cases, e.g., ℓ_0-regularization. From the view of machine learning, the models trained by globally convergent algorithms are stable in the sense that the models do not change a lot after enough iterations. Stable models cause stable predictions, which is crucial since we do not hope that the models and the prediction results are significantly changed just because of one more iteration.
3. They have low computational complexity, e.g., linear time complexity. Given the desired precision, linear time complexity means that the computational time for a required solution is linear to the data size. Linear time complexity, or even lower ones, makes the algorithms applicable to large-scale datasets.

In this paper, we aim to design the algorithms satisfying the above three good properties. First, we formalize the regularization

* Corresponding author.
E-mail addresses: zhengpan.gm@gmail.com (Z. Pan), linming04@gmail.com (M. Lin), hougd05@gmail.com (G. Hou), zcs@mail.tsinghua.edu.cn (C. Zhang).

http://dx.doi.org/10.1016/j.neucom.2014.11.009
0925-2312/© 2014 Elsevier B.V. All rights reserved.

problems concerned in this paper. We consider the following regularized linear learning model, which is actually an optimization problem:

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{j=1}^{n} l(\boldsymbol{x}_j^T\boldsymbol{\theta}, y_j)+\mathcal{R}(\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_n \in \mathbb{R}^p$ are n samples with their labels $y_1,\dots,y_n$ in the label domain,[1] $\boldsymbol{\theta}\in\mathbb{R}^p$ is the parameter that we want to learn from the samples. $l(\cdot,\cdot)$ is called *loss function*. We also call the averaged loss function

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n} l(\boldsymbol{x}_j^T\boldsymbol{\theta}, y_j). \tag{2}$$

as loss function. In Eq. (1), $\mathcal{R}(\boldsymbol{\theta})$ is called a *regularizer*. In this paper, we assume that the regularizers are *decomposable* such that

$$\mathcal{R}(\boldsymbol{\theta}) = \sum_{i=1}^{p} r(|\theta_i|), \tag{3}$$

where $r(\cdot)$ is the *basis function* of the regularizer. Table 1 lists some examples of non-convex regularizers and their basis functions. The basis functions in Table 1 are concave and non-decreasing, which is called sparsity-inducing regularizers since they usually drive the solutions of Eq. (1) to be sparse. In this paper, we mainly concern the sparsity-inducing non-convex regularizers, but the proposed algorithms, as well as the analysis on the global convergence and computational complexity, fit for not only sparsity-inducing non-convex regularizers, but also all the decomposable regularizers.

For simplicity, we denote the objective function of Eq. (1) as

$$\mathcal{F}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta})+\mathcal{R}(\boldsymbol{\theta}).$$

A lot of efficient algorithms have been proposed for convex regularization problems, e.g., SMO [21] and Coordinate Descent [6] for SVM, SpaRSA [25] and FISTA [2] for ℓ_1-regularization. However, their analysis on the convergence does not fit for non-convex regularizers.

Iterative Re-weighted ℓ_1 (IRL1) methods, Iterative Shrinkage-Thresholding Algorithms (ISTA) and Coordinate Descent (CD) methods are the three main methods for general non-convex regularization problems.

IRL1 methods use the convex relaxations of the regularizers at the current points to approximate the original problems [5,29,26]. IRL1 methods iteratively obtain local optimums of the approximated problems with the convex relaxations of the regularizers [7]. However, IRL1 methods are usually time-consuming since they have to solve a sequence of ℓ_1-regularized problems, e.g., LASSO for linear regression. IRL1 methods can also be regarded as DC programming [14] and Majorization-Minimization [16].

ISTA can be employed to non-convex regularizers in the same manner as convex regularizers [25,15,19]. However, ISTA is not always globally convergent. The number of iterations and computational complexity to obtain a local or approximately local optimums are still unknown for general non-convex regularizers.

For CD methods, Mazumder et al. [18] and Breheny and Huang [4] gave a convergence analysis for non-convex regularization, but their regularizers have a restriction on the "degree of non-convexity".[2] For example, their analysis cannot be applied to ℓ_q-norm $(0<q<1)$ or LSP with $\gamma<1$. For general non-convex regularizers, it still lacks guarantees for global convergence. In fact, the original CD may get stuck at a non-stationary point for non-smooth objective functions. The Proximal Coordinate Descent (PCD) [20] overcomes the stuck problem for sparse linear regression problems, but PCD needs $O(p^3)$ iterations for an approximately stationary solutions, which is not suitable for large-scale problems. Like CD, PCD is also not globally convergent without the restriction on the "degree of non-convexity".

If not restricted to special regularizers, IRL1, ISTA and CD have the drawbacks of heavy computational complexity and not being globally convergent. In this paper, we propose the Damping Proximal Coordinate Descent (DPCD) to give an algorithm that is applicable for a wide range of regularizers with global convergence and low computational complexity.

DPCD algorithms only update one dimension of the parameter in the manner of coordinate descent. Instead of directly optimizing the original objective functions in Eq. (1), DPCD algorithms replace the loss function with a quadratic approximation at the current point. More importantly, we further propose the *damping penalization* to guarantee the global convergence, as well as a fast convergence rate.

The convergence of DPCD does not need the convexity of the loss functions or the regularizers. DPCD algorithms are suitable for many existing loss functions and regularizers. The loss functions include the loss of L2-SVM (squared hinge loss) [11], linear regressions, logistic regressions and neurons (sigmoid functions). The regularizers can be set as all the regularizers in Table 1, including the highly non-convex regularizer ℓ_0-norm.

Theorem 1. *For any initial solution, DPCD algorithms are always convergent, i.e., global convergence. For any $\epsilon>0$, DPCD algorithms only need $O(\#\mathrm{nz}/\epsilon^2)$ time to give a solution $\hat{\boldsymbol{\theta}}$ with $\|\nabla\mathcal{F}(\hat{\boldsymbol{\theta}})\|_\infty<\epsilon$, where* #nz *is the number of the non-zero components of all the samples.*

Theorem 1 assumes the differentiability of the loss function and the regularizer. Section 3 will give more general analysis for the non-differentiable cases, where the global convergence still holds and the time complexity remains the same.

The time complexity of DPCD is linear to the data size given the approximately stationary precision ϵ. In our implementation, DPCD only accesses the non-zero elements of the samples. Thus, the time complexity can be tightened to $O(\#\mathrm{nz}/\epsilon^2)$, where #nz is the number of non-zero components of all the samples.

We also test DPCD algorithms on many real-world datasets in Section 4. Compared to the related algorithms in the experiments, DPCD algorithms have faster convergence rates. It takes less time for DPCD algorithms to train a model with good prediction results than the related algorithms.

2. Algorithm

Following the idea of coordinate descent, we update only one component (or dimension) of the parameters each time. However, the updating is not performed by minimizing the original function in Eq. (1). Directly minimizing the original function has the stuck problem of CD algorithms and such minimizing may also be a difficult problem to solve. Instead, we minimize an approximation of the function at the current points.

Let $\boldsymbol{\theta}^{(k)}=(\theta_1^{(k)},\dots,\theta_p^{(k)})$ be the solution of DPCD algorithms *after* the k-th iteration. During the k-th iteration, DPCD algorithms update the component $\theta_i^{(k-1)}$ to $\theta_i^{(k)}$ from $i=1$ to $i=p$. During the k-th iteration, the current point before updating $\theta_i^{(k-1)}$ is denoted as

$$\boldsymbol{z}^{(k,i)} = (\theta_1^{(k)},\dots,\theta_{i-1}^{(k)},\theta_i^{(k-1)},\dots,\theta_p^{(k-1)})^T. \tag{4}$$

[1] Different problems have different label domains, e.g., the label domain of linear regression is $\mathbb{R}$ and the label domain of binary classification is $\{1,-1\}$.

[2] Mazumder et al. [18] and Breheny and Huang [4] focused on sparsity-inducing regularizers which are regarded as approximations to ℓ_0-norm. The degree of non-convexity is actually the degree of approximation to ℓ_0-norm. The degree of non-convexity is controlled by the parameters of regularizers, e.g., the γ of LSP. Also, ℓ_0-norm is treated as the most non-convex case.