Pattern Recognition 48 (2015) 3728–3738

Contents lists available at ScienceDirect

Pattern Recognition

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

On group-wise ℓ_p regularization: Theory and efficient algorithms

Duc-Son Pham *

Department of Computing, Curtin University, Perth, Australia

ARTICLE INFO

Article history:
Received 16 April 2014
Received in revised form
10 March 2015
Accepted 11 May 2015
Available online 27 May 2015

Keywords:
ℓ_p Regularization
Convex optimization algorithms
ADMM
FISTA
Algorithmic stability
Lasso
Group Lasso
Bridge regression
Group bridge regression
Splice detection

ABSTRACT

Following advances in compressed sensing and high-dimensional statistics, many pattern recognition methods have been developed with ℓ_1 regularization, which promotes sparse solutions. In this work, we instead advocate the use of ℓ_p $(2 \geq p > 1)$ regularization in a group setting which provides a better trade-off between sparsity and algorithmic stability. We focus on the simplest case with squared loss, which is known as group bridge regression. On the theoretical side, we prove that group bridge regression is uniformly stable and thus generalizes, which is an important property of a learning method. On the computational side, we make group bridge regression more practically attractive by deriving provably convergent and computationally efficient optimization algorithms. We show that there are at least several values of p over (1,2) at which the iterative update is analytical, thus it is even suitable for large-scale settings. We demonstrate the clear advantage of group bridge regression with the proposed algorithms over other competitive alternatives on several datasets. As ℓ_p-regularization allows one to achieve flexibility in sparseness/denseness of the solution, we hope that the algorithms will be useful for future applications of this regularization.

1. Introduction

Regularization is an important issue in pattern recognition for developing learning algorithms with high predictive power. In this work, we consider algorithms for solving a regularization problem of the following form:

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + g(\mathbf{x}). \tag{1}$$

Here, we restrict our attention to the squared loss function and norm-based regularization $g(\mathbf{x})$. Extensions to other convex loss functions, such as logistic, may be obtained similarly.

In learning theory, such a regularization is known to avoid over-fitting and thus it allows the developed algorithm to generalize. Regularization has been an important principle in machine learning and statistics [1], especially when one is faced with increasing challenges of massive data-sets wherein the dimension can be very large [2]. Recently, with the explosive growth of interest in compressed sensing [3,4] and high-dimensional statistics [2], a great deal of literature has been devoted to study the learning problem with ℓ_1 regularization, i.e. $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1 = \lambda\sum_i |x_i|$. The theoretical arguments for such a choice have been put forward in, for example, [3–6]. It is known that ℓ_1 regularization promotes sparsity, which is conceived to be desirable in many learning problems. As such, optimization algorithms have been specifically developed to solved the Lasso-type problem [5] efficiently. The compressed sensing repository[1] contains numerous references on optimization algorithms for solving compressed sensing recovery via ℓ_1 regularization. Consequently, the literature has seen an increasing number of applications of ℓ_1 regularization, such as face recognition [7], graph optimization [8], object categorization [9].

As structure constraints are shown to be beneficial to learning algorithms [10], the statistics literature has also seen an extension of the basic Lasso scheme to situations where grouped variables are available, known as group Lasso [11,12,2]. In this setting, the variable vector $\mathbf{x}$ is naturally divided into G groups

$$\mathbf{x} = [\mathbf{x}_1;\ \mathbf{x}_2;\ \ldots;\ \mathbf{x}_G], \quad \mathbf{A} = [\mathbf{A}_1\mathbf{A}_2\ldots,\mathbf{A}_G] \tag{2}$$

$$\mathbf{A}\mathbf{x} = \sum_{i=1}^{G} \mathbf{A}_i\mathbf{x}_i. \tag{3}$$

* Tel.: +61 8 9266 4453; fax: +61 8 9266 2819.
E-mail address: dspham@ieee.org

[1] http://dsp.rice.edu/cs.

http://dx.doi.org/10.1016/j.patcog.2015.05.009
0031-3203/

Encouraging applications that exploits the group information can be found in a wide range of problems from image tagging [13] to face recognition [6].

However, there are cases where ℓ_1 regularization does not achieve competitive results as other dense regularization [14]. Theoretically, Xu et al. [15] have established that certain sparse algorithms, including Lasso and group Lasso, are not algorithmically stable, an important property of a good learning algorithm. Whilst not being algorithmically stable does not mean that sparsity algorithms do not generalize, it implies that they can potentially have poor predictive performance in the worst case scenarios. More recently, it has been shown in [16] in the context of multiple kernel learning that dense solution via ℓ_p-norm performs better than sparse solutions and achieves state-of-the-art performance over a wide range of problems. Likewise, [17] found that ℓ_p regularization with group settings attains best compromise between prediction and robustness for $p \in [1.5, 2]$. It appears that ℓ_p regularization is an alternative that provides a natural trade-off between sparsity and stability [15]. However, ℓ_p regularization is still of infrequent use in practice, especially in the group setting. This could be of two reasons, both theoretically and computationally. On the theoretical side, though there are some published works in the statistics literature such as [18,19], little is known about the generalization property of group bridge regression. On the computational side, efficient algorithms for ℓ_p regularization in general, especially in large-scale problems, seem to be lacking compared with ℓ_1 regularization. We note that ℓ_p regularization is strictly convex for $p > 1$, and thus gradient techniques can be used. However, they tend to have rather poor convergence property especially when p is close to 1 (which we demonstrate subsequently).

In this work, we further advocate the use of ℓ_p regularization in a group setting. For the squared loss, this is known as group bridge regression [18]. Though it is not new, we revisit this powerful regression method in the large-scale pattern recognition context and make two contributions. Theoretically, we prove that group bridge regression is also algorithmically stable, and thus it generalizes. Computationally, we develop the novel and efficient algorithms under two powerful optimization frameworks: alternative directions method of multipliers (ADMM) [20] and fast iterative shrinkage thresholding (FISTA) [21]. We show that there are values of p distributed over the range [1,2] where group bridge regression have *analytical* solutions for the iterative updates, just like the Lasso. This implies one can achieve varying degrees of sparseness in the solution efficiently with the proposed algorithms. This is particularly useful in cases where compressible data is present [22]. When analytical updates are not available, we propose an algorithm to compute the updates with an efficient warm-start strategy. Whilst the studied examples in this work subsequently show the advantage of ℓ_p regularization over ℓ_1, note that we do not claim it is always better. There will be cases where ℓ_1 might be more suitable. What we try to convey here is an alternative method for pattern recognition, which clearly allows flexibility between achieving sparse or dense solutions with the most desirable property of a learning algorithm.

The paper is organized as follows. Section 2 establishes the algorithmic stability of ℓ_p regularization in group bridge regression settings. In Section 3, we derive efficient ADMM- and FISTA-based algorithms for solving group bridge regression. Section 4 examines the numerical properties of the proposed algorithms and demonstrate the competitive advantage of group bridge regression over other sparse alternatives on a synthetic dataset and a real-world splice detection problem. Finally, Section 5 concludes.

The Matlab implementation of all developed methods is made publicly available at the following website: https://sites.google.com/site/dspham.

2. Algorithmic stability with ℓ_p-norm regularization

Algorithmic stability [23] is one powerful concept for assessing the predictive power of a supervised learning method. We now show that group bridge regression of problem (1) with

$$g(\mathbf{x}) = \lambda \sum_{i=1}^{G} \|\mathbf{x}_i\|_2^p = \lambda \|\mathbf{x}\|_{\ell_2/\ell_p}, \quad p \in (1, 2], \tag{4}$$

is indeed algorithmically stable. Our approach is based on the key result in [24], and we tailor it to the group setting.

First, we briefly revisit the common setting in supervised learning, where a set of data points $\mathbf{z} = \{(\mathbf{a}_1, y_1), \ldots, (\mathbf{a}_n, y_n)\}$, and $\mathbf{a}_i \in \mathbb{R}^d$. The aim is to learn a function f from $\mathbf{z}$ that allow us to predict y given a future $\mathbf{a}$. Here, for the formulation (1) the function to be learnt is linear $f(\mathbf{a}; \mathbf{x}) = \mathbf{a}^T\mathbf{x}$ and the squared loss function $V(y_1, y) = \frac{1}{2}(y_1 - y)^2$. Up to a scaling by a factor of $1/n$, the formulation (1) is known in learning theory as empirical risk minimization where the first term essential represents the empirical risk $R_{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} V(\mathbf{a}_i^T\mathbf{x}, y_i)$. An algorithm is said to be consistent if the empirical risk converges asymptotically to the risk, i.e.

$$\lim_{n\to\infty} R_{\mathbf{z}} = R = E_{\mathbf{a}}(1/2)(y - \mathbf{a}^T\mathbf{x})^2,$$

assuming bounded risk. For the finite sample case, the learning theory is interested in the bound on the deviation of the empirical risk from the risk. In algorithmic stability theory [23], an algorithm is said to be uniformly β-stable if there exists a finite β that upper bounds the maximum deviation in the loss due to replacement of the sample $\mathbf{z}$ with any possible $\mathbf{z}'$ from the same distribution. Under regularity conditions on the loss function, including convexity, boundedness at $\mathbf{0}$, and L-Lipschitz in the first variable (which are met by the actual squared loss considered here), [23] showed that the bound is

$$|R_{\mathbf{z}} - R| \le \beta + (2n\beta + L\sqrt{\kappa}\tau + B)\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{5}$$

with the probability of at least $1-\delta$. Here, κ is an upper bound on the feature functions in the functional space of f, i.e. $\kappa = \sup\|\mathbf{x}\|_2^2$, and B is an upper bound on the loss function when the first variable is zero, i.e. $B = \frac{1}{2}\max y^2$, and L is the Lipschitz constant of the lost function $V(y_1, y)$ in terms of the first variable y_1 subject to the regular conditions, i.e. $L = y_{\max} - y_{\min}$. Detail can be found in [23,24]. Clearly, when $\beta = o(n^{-1/2})$ then stability implies generalization. Though uniform stability appears rather strict, it requires no further assumptions on the data than other weaker notion of stability in the literature [24].

Though algorithmic stability is a powerful tool to characterize a learning algorithm, there was not an easy way to verify uniform stability for a particular method until recently when Wibisono et al. [24] discovered a sufficient condition to do so. Consider the class of norm regularization where $g(\mathbf{x}) = \lambda P(\mathbf{x})$ where λ is the regularization parameter and $P(\mathbf{x})$ is some suitable norm. Denote as $\mathbf{x}_{\mathbf{z}}$ and $\mathbf{x}_{\mathbf{z}^j}$ respectively the solution of the regularized empirical risk minimization on original data $\mathbf{z}$ and when the jth sample is replaced with another from the same distribution. It was established in [24] that

Theorem 2.1. *Suppose that for some constant $C > 0$ and $\xi > 1$, the penalty function satisfies*

$$P(\mathbf{x}_{\mathbf{z}}) + P(\mathbf{x}_{\mathbf{z}^j}) - 2P\left(\frac{\mathbf{x}_{\mathbf{z}} + \mathbf{x}_{\mathbf{z}^j}}{2}\right) \ge C\|\mathbf{x}_{\mathbf{z}} - \mathbf{x}_{\mathbf{z}^j}\|_2^{\xi}$$

then the regularization is uniformly β-stable with $\beta = \left(\frac{L^{\xi}\kappa^{\xi/2}}{n\lambda C}\right)^{1/(\xi-1)}$.

Using this important result and following the strategy in [24], we also establish algorithmic stability for group bridge regression as follows: