On the Sub-Optimality of Proximal Gradient Descent for ℓ^0 Sparse Approximation

Yingzhen Yang YYANG58@ILLINOIS.EDU
Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Jianchao Yang JIANCHAO.YANG@SNAPCHAT.COM
Snapchat, Venice, CA 90291

Wei Han WEIHAN3@ILLINOIS.EDU
Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Thomas. S. Huang T-HUANG1@ILLINOIS.EDU
Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Abstract

We investigate the ℓ^0 sparse approximation in this paper, and propose a proximal gradient descent method which obtains a sub-optimal solution to the nonconvex optimization of ℓ^0 sparse approximation in an iterative shrinkage manner. Our analysis gives the gap between the sub-optimal solution and the globally optimal solution for ℓ^0 sparse approximation, as well as the conditions in which the sub-optimal solution is globally optimal. We also show the application of our algorithm to data clustering with superior results.

1. Introduction

In this paper, we consider the ℓ^0 sparse approximation problem, or the ℓ^0 penalized Least Square Estimation (LSE) problem below:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} L(\boldsymbol{\alpha}) = \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_0 \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d$ is a signal in d-dimensional Euclidean space, $\mathbf{D}$ is the design matrix of dimension $d \times n$ which is also called a dictionary with n atoms in the sparse coding literature. The goal of problem (1) is to approximately represent signal $\mathbf{x}$ by the atoms of the dictionary $\mathbf{D}$ while requiring the representation to be sparse. Due to the nonconvexity imposed by the ℓ^0 norm, previous research works resort to solve its ℓ^1 relaxation

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{2}$$

(2) is convex and also known as Basis Pursuit Denoising which can be solved efficiently by linear programming or iterative shrinkage algorithms (Daubechies et al., 2004; Elad, 2006; Bredies & Lorenz, 2008).

Albeit the convexity of (1), sparse representation methods such as (Mancera & Portilla, 2006; Bao et al., 2014) that directly optimize objective function involving ℓ^0-norm demonstrate compelling performance compared to its ℓ^1 norm counterpart. We use Proximal Gradient Descent (PGD) to obtain a sub-optimal solution to (1) in an iterative shrinkage manner with theoretical guarantee. Although a similar Iterative Hard-Thresholding (IHT) algorithm is proposed by Blumensath et al. (Blumensath & Davies, 2008), we prove the bound for gap between the sub-optimal solution and the globally optimal solution to (1). Moreover, if the solution to proper ℓ^1 sparse approximation problem serves as the initialization for PGD, we present the conditions in which the gap vanishes, i.e. the sub-optimal solution equals to the globally optimal solution. The assumptions made for our theoretical analysis are mostly in terms of the sparse eigenvalues of the dictionary. To the best of our knowledge, there are quite few results in this direction. Our results establish the theoretical soundness of PGD for ℓ^0 sparse approximation, and suggest the merit of initialization by the solution to the ℓ^1 relaxation. In addition, we apply our optimization algorithm to data clustering, and the proposed ℓ^0 sparse graph method achieves better results than other competing methods.

Throughout this paper, we use bold letters for matrices and

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

vectors, regular lower letter for scalars. The bold letter with subscript indicates the corresponding element of a matrix or vector, and $\|\cdot\|_p$ denote the ℓ^p-norm of a vector.

2. Proximal Gradient Descent for ℓ^0 Sparse Approximation

Solving the ℓ^0 sparse approximation problem (1) exactly is NP-hard, and it is impractical to seek for its globally optimal solution. The literature extensively resorts to approximate algorithms, such as Orthogonal Matching Pursuit (Tropp, 2004), or that use surrogate functions (Hyder & Mahata, 2009), for ℓ^0 problems. In this section we present an algorithm that employs PGD to optimize (1) in an iterative shrinkage manner, and obtains a sub-optimal solution with theoretical guarantee.

2.1. Algorithm

The dictionary $\mathbf{D}$ is normalized such that each column has unit ℓ^2-norm. In t-th iteration of PGD for $t \geq 1$, gradient descent is performed on the squared loss term of $L(\boldsymbol{\alpha})$, i.e. $Q(\boldsymbol{\alpha}) = \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}\|_2^2$, to obtain

$$\tilde{\boldsymbol{\alpha}}^{(t)} = \boldsymbol{\alpha}^{(t-1)} - \frac{2}{\tau s}(\mathbf{D}^\top \mathbf{D}\boldsymbol{\alpha}^{(t-1)} - \mathbf{D}^\top \mathbf{D}) \tag{3}$$

where τ is any constant that is greater than 1. s is the Lipschitz constant for the gradient of function $Q(\cdot)$, namely

$$\|\nabla Q(\mathbf{y}) - \nabla Q(\mathbf{z})\|_2 \leq s\|\mathbf{y} - \mathbf{z}\|_2, \ \forall\, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n \tag{4}$$

$\boldsymbol{\alpha}^{(t)}$ is then the solution to the following ℓ^0 regularized problem which is also the proximal mapping:

$$\boldsymbol{\alpha}^{(t)} = \arg\min_{\mathbf{v} \in \mathbb{R}^n} \frac{\tau s}{2}\|\mathbf{v} - \tilde{\boldsymbol{\alpha}}^{(t)}\|_2^2 + \lambda\|\mathbf{v}\|_0 \tag{5}$$

It can be verified that (5) has closed-form solution:

$$\boldsymbol{\alpha}^{(t)} = h_{\sqrt{\frac{2\lambda}{\tau s}}}(\tilde{\boldsymbol{\alpha}}^{(t)}) \tag{6}$$

where h_θ is an element-wise hard thresholding operator:

$$[h_\theta(\mathbf{u})]_j = \begin{cases} 0 & : \ |\mathbf{u}_j| < \theta \\ \mathbf{u}_j & : \ \text{otherwise} \end{cases}, \quad 1 \leq j \leq n$$

The iterations start from $t = 1$ and continue until the sequence $\{L(\boldsymbol{\alpha}^{(t)})\}_t$ or $\{\boldsymbol{\alpha}^{(t)}\}_t$ converges or maximum iteration number is achieved, then a sub-optimal solution is obtained. Our optimization algorithm by PGD is described in Algorithm 1. The time complexity of our iterative proximal method is $\mathcal{O}(Mn^2)$ where M is the number of iterations (or maximum number of iterations) for the iterative proximal method.

2.2. Theoretical Analysis

In this section we present the bound for the gap between the sub-optimal solution by PGD in Algorithm 1 and the globally optimal solution for the ℓ^0 sparse approximation problem (1). Under certain assumption on the sparse eigenvalues of the data $\mathbf{D}$, we show that the sub-optimal solution by PGD is actually a critical point of $L(\boldsymbol{\alpha})$ in Lemma 1, namely the sequence $\{\boldsymbol{\alpha}^{(t)}\}_t$ converges to a critical point of the objective (1). We then show that both this sub-optimal solution and the globally optimal solution to (1) are local solutions of a carefully designed capped-ℓ^1 regularized problem in Lemma 2. Based on (Zhang & Zhang, 2012) which shows the distance between different local solutions to various sparse estimation problems including the capped-ℓ^1 problem, the bound for ℓ^2-distance between the sub-optimal solution and the globally optimal solution is presented in Theorem 1, again under the assumption on the sparse eigenvalues of $\mathbf{D}$. We further show when the bound vanishes in Theorem 2.

Algorithm 1 Proximal Gradient Descent for ℓ^0 Sparse Approximation (1)

Input:
The given signal $\mathbf{x} \in R^d$, the dictionary $\mathbf{D}$, the parameter λ for the weight of the ℓ^0 norm, maximum iteration number M, stopping threshold ε, the initialization $\boldsymbol{\alpha}^{(0)}$.

1: Obtain the sub-optimal solution $\hat{\boldsymbol{\alpha}}$ by the Proximal Gradient Descent (PGD) method with (3) and (6) starting from $t = 1$. The iteration terminates either $\{\boldsymbol{\alpha}^{(t)}\}_t$ or $\{L(\boldsymbol{\alpha}^{(t)})\}_t$ converges under the threshold ε or maximum iteration number is achieved.

Output: Obtain the sub-optimal solution $\hat{\boldsymbol{\alpha}}$.

In the following analysis, we let $\boldsymbol{\beta}_{\mathbf{I}}$ denote the vector formed by the elements of $\boldsymbol{\beta}$ with indices in $\mathbf{I}$ when $\boldsymbol{\beta}$ is a vector, or matrix formed by columns of $\boldsymbol{\beta}$ with indices in $\mathbf{I}$ when $\boldsymbol{\beta}$ is a matrix. Also, we let $\mathbf{S} = \text{supp}(\boldsymbol{\alpha}^{(0)})$ and $|\mathbf{S}| = A$. The definition of sparse eigenvalues and critical points are defined below which is important for our analysis.

Definition 1. *(Sparse eigenvalues) The lower and upper sparse eigenvalues of a matrix $\mathbf{A}$ are defined as*

$$\kappa_-(m) := \min_{\|\mathbf{u}\|_0 \leq m; \|\mathbf{u}\|_2 = 1} \|\mathbf{A}\mathbf{u}\|_2^2 \quad \kappa_+(m) := \max_{\|\mathbf{u}\|_0 \leq m, \|\mathbf{u}\|_2 = 1} \|\mathbf{A}\mathbf{u}\|_2^2$$

It is worthwhile mentioning that the sparse eigenvalues are closely related to the Restricted Isometry Property (RIP) (Candes & Tao, 2005) used frequently in the compressive sensing literature. Typical RIP requires bounds such as $\delta_\tau + \delta_{2\tau} + \delta_{3\tau} < 1$ or $\delta_{2\tau} < \sqrt{2} - 1$ (Cands, 2008) for stably recovering the signal from measurements and τ is the sparsity of the signal, where $\delta_\tau = \max\{\kappa_+(\tau) - 1, 1 - \kappa_-(\tau)\}$. Similar to (Zhang & Zhang, 2012), we use conditions on the sparse eigenvalues in this paper which are more general than RIP in the sense of not requiring bounds in terms of

δ to obtain theoretical results. In the following text, sparse eigenvalues κ_- and κ_+ are for the dictionary $\mathbf{D}$.

Definition 2. *(Critical points) Given the non-convex function $f\colon \mathbb{R}^n \to R \cup \{+\infty\}$ which is a proper and lower semi-continuous function.*

- *for a given $\mathbf{x} \in \mathrm{dom} f$, its Frechet subdifferential of f at $\mathbf{x}$, denoted by $\tilde{\partial} f(x)$, is the set of all vectors $\mathbf{u} \in \mathbb{R}^n$ which satisfy*

$$\limsup_{\mathbf{y}\neq\mathbf{x},\mathbf{y}\to\mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{u}, \mathbf{y}-\mathbf{x}\rangle}{\|\mathbf{y}-\mathbf{x}\|} \geq 0$$

- *The limiting-subdifferential of f at $\mathbf{x} \in \mathbb{R}^n$, denoted by written $\partial f(x)$, is defined by*

$$\partial f(x) = \{\mathbf{u} \in \mathbb{R}^n \colon \exists \mathbf{x}^k \to \mathbf{x}, f(\mathbf{x}^k) \to f(\mathbf{x}), \\ \tilde{\mathbf{u}}^k \in \tilde{\partial} f(\mathbf{x}_k) \to \mathbf{u}\}$$

The point $\mathbf{x}$ is a critical point of f if $0 \in \partial f(x)$.

If the dictionary $\mathbf{D}$ has certain positive lower sparse eigenvalue, Lemma 1 shows that the sequences $\{\boldsymbol{\alpha}^{(t)}\}_t$ produced by PGD converges to a critical point of $L(\boldsymbol{\alpha})$, the objective of the ℓ^0 sparse approximation problem (1).

Lemma 1. *Suppose $\kappa_-(A) > 0$, then the sequence $\{\boldsymbol{\alpha}^{(t)}\}_t$ generated by PDG with (3) and (6) converges to a critical point of $L(\boldsymbol{\alpha})$.*

Denote the critical of $L(\boldsymbol{\alpha})$ by $\hat{\boldsymbol{\alpha}}$ that the sequence $\{\boldsymbol{\alpha}^{(t)}\}_t$ converges to when the assumption of Lemma 1 holds, and denote by $\boldsymbol{\alpha}^*$ the globally optimal solution to the ℓ^0-SSC problem (1). Also, we consider the following capped-ℓ^1 regularized problem, which replaces the noncontinuous ℓ^0-norm with the continuous capped-ℓ^1 regularization term R:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} L_{\mathrm{capped}-\ell^1}(\boldsymbol{\beta}) = \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\beta}\|_2^2 + \mathbf{R}(\boldsymbol{\beta}; b) \tag{7}$$

where $\mathbf{R}(\boldsymbol{\beta}; b) = \sum_{j=1}^{n} R(\boldsymbol{\beta}_j; b)$, $R(t; b) = \lambda \frac{\min\{|t|, b\}}{b}$ for some $b > 0$. It can be seen that $R(t; b)$ approaches the ℓ^0-norm when $b \to 0+$. Our following theoretical analysis aims to obtain the gap between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$. For the sake of this purpose, the definition of local solution and degree of nonconvexity of a regularizer are necessary and presented below.

Definition 3. *(Local solution) A vector $\tilde{\boldsymbol{\beta}}$ is a local solution to the problem (7) if*

$$\|2\mathbf{D}^\top(\mathbf{D}\tilde{\boldsymbol{\beta}} - \mathbf{x}_i) + \dot{\mathbf{R}}(\tilde{\boldsymbol{\beta}}; b)\|_2 = 0 \tag{8}$$

where $\dot{\mathbf{R}}(\tilde{\boldsymbol{\beta}}; b) = [\dot{R}(\tilde{\boldsymbol{\beta}}_1; b), \dot{R}(\tilde{\boldsymbol{\beta}}_2; b), \ldots, \dot{R}(\tilde{\boldsymbol{\beta}}_n; b)]^\top$.

Note that in the above definition and the following text, $\dot{R}(t; b)$ can be chosen as any value between the right differential $\frac{\partial R}{\partial t}(t+; b)$ (or $\dot{R}(t+; b)$) and left differential $\frac{\partial R}{\partial t}(t-; b)$ (or $\dot{R}(t-; b)$).

Definition 4. *(Degree of Nonconvexity of a Regularizer) For $\kappa \geq 0$ and $t \in \mathbb{R}$, define*

$$\theta(t, \kappa) := \sup_s \{-\mathrm{sgn}(s-t)(\dot{P}(s; b) - \dot{P}(t; b)) - \kappa|s-t|\}$$

as the degree of nonconvexity for function P. If $\mathbf{u} = (u_1, \ldots, u_n)^\top \in \mathbb{R}^n$, $\theta(\mathbf{u}, \kappa) = [\theta(u_1, \kappa), \ldots, \theta(u_p, \kappa)]$.

Note that $\theta(t, \kappa) = 0$ for convex function P.

Let $\hat{\mathbf{S}} = \mathrm{supp}(\hat{\boldsymbol{\alpha}})$, $\mathbf{S}^* = \mathrm{supp}(\boldsymbol{\alpha}^*)$, the following lemma shows that both $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ are local solutions to the capped-ℓ^1 regularized problem (7).

Lemma 2. *If*

$$0 < b < \min\{\min_{j\in\hat{\mathbf{S}}} |\hat{\boldsymbol{\alpha}}_j|, \frac{\lambda}{\max_{j\notin\hat{\mathbf{S}}} |\frac{\partial Q}{\partial \boldsymbol{\alpha}_j}|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}|}, \\ \min_{j\in\mathbf{S}^*} |{\boldsymbol{\alpha}_j}^*|, \frac{\lambda}{\max_{j\notin\mathbf{S}^*} |\frac{\partial Q}{\partial \boldsymbol{\alpha}_j}|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}|}\} \tag{9}$$

(if the denominator is 0, $\frac{\lambda}{0}$ is defined to be $+\infty$ in the above inequality), then both $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ are local solutions to the capped-ℓ^1 regularized problem (7).

Theorem 5 in (Zhang & Zhang, 2012) gives the estimation on the distance between two local solutions of the capped-ℓ^1 regularized problem. Based on this result, Theorem 1 shows that under assumptions on the sparse eigenvalues of $\mathbf{D}$, the sub-optimal solution $\hat{\boldsymbol{\alpha}}$ obtained by PGD has bounded ℓ^2-distance to $\boldsymbol{\alpha}^*$ which constitutes one of our main results in this paper.

Theorem 1. *(Sub-optimal solution is close to the globally optimal solution) Suppose $\kappa_-(A) > 0$ and $\kappa_-(|\hat{\mathbf{S}} \cup \mathbf{S}^*|) > \kappa > 0$, and b is chosen according to (9) as in Lemma 2. Then*

$$\|\mathbf{D}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_2^2 \leq \frac{2\kappa_-(|\hat{\mathbf{S}} \cup \mathbf{S}^*|)}{(\kappa_-(|\hat{\mathbf{S}} \cup \mathbf{S}^*|) - \kappa)^2} \tag{10} \\ \big(\sum_{j\in\hat{\mathbf{S}}} (\max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j - b|\})^2 + |\mathbf{S}^* \setminus \hat{\mathbf{S}}|(\max\{0, \frac{\lambda}{b} - \kappa b\})^2\big)$$

In addition,

$$\|(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_2^2 \leq \frac{2}{(\kappa_-(|\hat{\mathbf{S}} \cup \mathbf{S}^*|) - \kappa)^2} \tag{11} \\ \big(\sum_{j\in\hat{\mathbf{S}}} (\max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j - b|\})^2 + |\mathbf{S}^* \setminus \hat{\mathbf{S}}|(\max\{0, \frac{\lambda}{b} - \kappa b\})^2\big)$$

Proof. According to Lemma 2, both $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ are local solutions to problem (7). By Theorem 5 in (Zhang & Zhang, 2012), we have

$$\|\mathbf{D}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_2^2 \leq \frac{2\kappa_-(|\hat{\mathbf{S}} \cup \mathbf{S}^*|)}{(\kappa_-(|\hat{\mathbf{S}} \cup \mathbf{S}^*|) - \kappa)^2} \big(\|\theta(|\hat{\boldsymbol{\alpha}}_{\hat{\mathbf{S}}}, \kappa)\|_2^2 \tag{12} \\ + |\mathbf{S}^* \setminus \hat{\mathbf{S}}|\theta^2(0+, \kappa)\big)$$

By the definition of θ,

$$\theta(t,\kappa) = \sup_s\{-\mathrm{sgn}(s-t)(\dot{R}(s;b) - \dot{R}(t;b)) - \kappa|s-t|\}$$

Since $t > b$, it can be verified that $\theta(t,\kappa) = \max\{0, \frac{\lambda}{b} - \kappa|t-b|\}$. Therefore,

$$\|\theta(\hat{\boldsymbol{\alpha}}_{\hat{\mathbf{S}}},\kappa)\|_2^2 = \sum_{j\in\hat{\mathbf{S}}}\left(\theta(\hat{\boldsymbol{\alpha}}_j,\kappa)\right)^2 = \sum_{j\in\hat{\mathbf{S}}}(\max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j - b|\})^2 \tag{13}$$

It can also be verified that

$$\theta(0+,\kappa) = \max\{0, \frac{\lambda}{b} - \kappa b\} \tag{14}$$

So that (10) is proved. Let $\mathbf{S}' = \hat{\mathbf{S}} \cup \mathbf{S}^*$, since $\sigma_{\min}(\mathbf{D}_{\mathbf{S}'}^\top \mathbf{D}_{\mathbf{S}'}) \geq \kappa_-(|\hat{\mathbf{S}} \cup \mathbf{S}^*|)$, so that $\|\mathbf{D}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_2^2 \geq \kappa_-(|\hat{\mathbf{S}} \cup \mathbf{S}^*|)\|(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_2^2$. It follows that (11) holds. □

Remark 1. *If $\hat{\boldsymbol{\alpha}}$ is sparse, we can expect that $|\hat{\mathbf{S}} \cup \mathbf{S}^*|$ is reasonably small, and a small $|\hat{\mathbf{S}} \cup \mathbf{S}^*|$ often increases the chance of a larger $\kappa_-(|\hat{\mathbf{S}} \cup \mathbf{S}^*|)$. Also note that the bound for distance between the sub-optimal solution and the globally optimal solution presented in Theorem 1 does not require typical RIP conditions. Also, it is always too restrictive to assume the rows of $\mathbf{D}$ are i.i.d. randomly generated, and in many practical cases the rows of $\mathbf{D}$ are correlated (i.e. correlated features), wherein the probabilistic RIPless theory is not applicable. Moreover, when $\frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j - b|$ for nonzero $\hat{\boldsymbol{\alpha}}_j$ and $\frac{\lambda}{b} - \kappa b$ are no greater than 0, or they are small positive numbers, the sub-optimal solution $\hat{\boldsymbol{\alpha}}$ is equal to or very close to the globally optimal solution.*

Remark 1 illustrates the conditions in which the gap between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ is small or even vanishing, where the sparsity of $\hat{\boldsymbol{\alpha}}$ is preferred. Based on this remark, we observe that if all the nonzero elements of $\hat{\boldsymbol{\alpha}}$ are positive and $\frac{\lambda}{b} - \kappa b < 0$, then the bound for the ℓ^2-distance between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ vanishes. In order to obtain a sparse sub-optimal solution $\hat{\boldsymbol{\alpha}}$ with all nonzero elements being positive, we propose to use a positive sparse initialization $\boldsymbol{\alpha}^{(0)}$. The following theorem shows that such $\boldsymbol{\alpha}^{(0)}$ can be obtained as the solution to a proper ℓ^1 sparse approximation problem. Also, under the assumption of Theorem 1, the sub-optimal solution obtained by the proposed PGD is in fact globally optimal, i.e. $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^*$, when λ is chosen from a certain range.

Before stating the theorem, let

$$\hat{\boldsymbol{\alpha}}^{\ell^1} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

Denote the indices of positive elements of $\hat{\boldsymbol{\alpha}}_{\ell^1}$ by $\mathbf{I}^+ = \{j\colon \hat{\boldsymbol{\alpha}}_j^{\ell^1} > 0\}$, and we can obtain a new dictionary $\mathbf{D}^+$ by flipping the sign of columns of $\mathbf{D}$ indexed by $\mathbf{I}_+$, i.e. $\mathbf{D}^+_{\mathbf{I}^+} = \mathbf{D}_{\mathbf{I}^+}$, $\mathbf{D}^-_{\mathbf{I}^{+\complement}} = -\mathbf{D}_{\mathbf{I}^{+\complement}}$.

Theorem 2. *Let*

$$\boldsymbol{\alpha}^+ = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \|\mathbf{x} - \mathbf{D}^+\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{15}$$

then all the nonzero elements of $\boldsymbol{\alpha}^+$ are positive. Suppose the assumption in Theorem 1 holds, and $\boldsymbol{\alpha}^+$ is used as the initialization of PGD, i.e. $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}^+$. If $s > \max\{2A, \frac{2(1+\lambda A)}{\lambda\tau}\}$, then the sub-optimal solution $\hat{\boldsymbol{\alpha}}$ generated by PGD with (3) and (6) satisfies $\hat{\mathbf{S}} \subseteq \mathbf{S}$ where $\mathbf{S} = \mathrm{supp}(\boldsymbol{\alpha}^{(0)})$, indicating that $\hat{\boldsymbol{\alpha}}$ is sparse.

Moreover, $\hat{\boldsymbol{\alpha}}_j > 0$ for any $j \in \hat{\mathbf{S}}$, i.e. all the nonzero elements of $\hat{\boldsymbol{\alpha}}$ are positive. It follows that when $\lambda < \kappa b^2$, $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^*$, namely the sub-optimal solution is also the globally optimal solution.

The detailed proofs of the theorems and lemmas in this paper are included in the supplementary document upon request.

Table 1. Clustering Results on the Extended Yale Face Database B. ℓ^0-SG (ℓ^0 sparse graph) is compared to SSC (Elhamifar & Vidal, 2013), SMCE (Elhamifar & Vidal, 2011) and SSC-OMP (Dyer et al., 2013) which are important subspace and manifold based clustering method using sparse approximation, as well as KM (K-means) and SC (spectral clustering).

	Measure	KM	SC	SSC	SMCE	SSC-OMP	ℓ^0-SG
	AC	0.0954	0.1077	0.7850	0.3293	0.6529	**0.8480**
	NMI	0.1258	0.1485	0.7760	0.3812	0.7024	**0.8612**

3. Application to Data Clustering

In this section, we show the application of ℓ^0 sparse approximation by the proposed PGD for ℓ^0 sparse graph based clustering. Given N data points $\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d\times N}$, ℓ^0 sparse graph based clustering method solves the following ℓ^0 sparse approximation problem wherein the data $\boldsymbol{X}$ serves as the dictionary for each $1 \leq i \leq n$: $\min_{\boldsymbol{\alpha}^i\in\mathbb{R}^n, \boldsymbol{\alpha}_i^i = 0} \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^i\|_2^2 + \lambda\|\boldsymbol{\alpha}^i\|_0$. $\boldsymbol{\alpha}^i$ is the sparse code for $\mathbf{x}_i$, and the constraint $\boldsymbol{\alpha}_i^i = 0$ is to avoid the trivial solution. Then a sparse similarity graph with the weighted adjacency matrix W set by $\mathbf{W}_{ij} = \frac{|\boldsymbol{\alpha}_i^j| + |\boldsymbol{\alpha}_j^i|}{2}$, and spectral clustering is performed on $\mathbf{W}$ to obtain the data clustering result. We show the superiority of our method compared to other methods including those using different kinds of sparse approximation in Table 1.

4. Conclusions

We propose to use proximal gradient descent to obtain a sub-optimal solution to the ℓ^0 sparse approximation problem. Our theoretical analysis renders the bound for the ℓ^2-distance between the sub-optimal solution and the globally

optimal solution, and establishes the conditions in which the sub-optimal solution is also the globally optimal solution to the original ℓ^0 sparse approximation problem. Moreover, we apply our algorithm to data clustering and demonstrate the compelling result of the proposed ℓ^0 sparse graph based clustering.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 1318971. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Bao, Chenglong, Ji, Hui, Quan, Yuhui, and Shen, Zuowei. L0 norm based dictionary learning by proximal methods with global convergence. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3858–3865, 2014.

Blumensath, Thomas and Davies, Mike E. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5):629–654, 2008. ISSN 1531-5851. doi: 10.1007/s00041-008-9035-z. URL http://dx.doi.org/10.1007/s00041-008-9035-z.

Bredies, Kristian and Lorenz, Dirk A. Iterated hard shrinkage for minimization problems with sparsity constraints. *SIAM Journal on Scientific Computing*, 30(2):657–683, 2008. doi: 10.1137/060663556.

Candes, E.J. and Tao, T. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

Cands, Emmanuel J. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(910):589 – 592, 2008. ISSN 1631-073X.

Daubechies, I., Defrise, M., and De Mol, C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004. ISSN 1097-0312. doi: 10.1002/cpa.20042. URL http://dx.doi.org/10.1002/cpa.20042.

Dyer, Eva L., Sankaranarayanan, Aswin C., and Baraniuk, Richard G. Greedy feature selection for subspace clustering. *Journal of Machine Learning Research*, 14:2487–2517, 2013.

Elad, M. Why simple shrinkage is still relevant for redundant representations? *IEEE Transactions on Information Theory*, 52(12):5559–5569, Dec 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.885522.

Elhamifar, Ehsan and Vidal, René. Sparse manifold clustering and embedding. In *NIPS*, pp. 55–63, 2011.

Elhamifar, Ehsan and Vidal, René. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

Hyder, M. and Mahata, K. An approximate l0 norm minimization algorithm for compressed sensing. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 3365–3368, April 2009.

Mancera, L. and Portilla, J. L0-norm-based sparse representation through alternate projections. In *Image Processing, 2006 IEEE International Conference on*, pp. 2089–2092, Oct 2006.

Tropp, Joel A. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50 (10):2231–2242, 2004.

Zhang, Cun-Hui and Zhang, Tong. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.*, 27(4):576–593, 11 2012.