Speeding Up Graph Regularized Sparse Coding by Dual Gradient Ascent

Rui Jiang, Hong Qiao, *Senior Member, IEEE*, and Bo Zhang, *Member, IEEE*

***Abstract*—Graph regularized Sparse Coding (GSC) considers data relationships during Sparse Coding (SC) and thus has better performance in certain image analysis tasks. However, it is very time consuming. This letter aims at speeding up GSC. The alternating optimization framework for GSC involves repeatedly solving a variant of ℓ_1 minimization referred to as GSRsub in this letter. Traditional ways to deal with GSRsub are to generalize optimization strategies for ℓ_1 minimization to solve its primal problem that is strongly convex but non-differentiable, thus converging slowly. We propose that GSC can be accelerated by solving a new dual problem of GSRsub called D-GSRsub. Compared with the primal form and the existing dual form of GSRsub, D-GSRsub has a strongly convex and smooth objective function with less variables. Based on these properties, four dual gradient ascent strategies with lower computational complexities are developed. Experimental results on real-world datasets demonstrate that these strategies can dramatically and stably speed up GSC without affecting its performance in the corresponding image analysis tasks.**

***Index Terms*—Graph regularized sparse coding, image classification, image clustering.**

I. Introduction

SPARSE CODING (SC) has been a popular coding technique for generating representations in image analysis. Recent studies found that SC does not consider data relationships during the coding process, thus leading to information loss. To address this issue, Graph regularized SC (GSC) based methods have been proposed. ScSPM [1] is a representation learning method for image classification and uses spatial-pyramid max pooling of SIFT sparse codes. Gao *et al.* [2], [3] pointed out that SC can not encode similar SIFT features as similar sparse codes. To reduce the loss of locality information, Laplacian Sparse Coding (LSc) was proposed to consider the similarity between SIFT features in ScSPM. As a holistic representation learning method for image classification and clustering, SC fails to consider the geometrical structure of images satisfying the manifold assumption. GraphSC [4] was presented for preserving the geometrical information during SC. GSC indeed improves the performance of SC. However, the great execution time of GSC has been a bottleneck that restricts its practical use [5]. This letter focuses on speeding up GSC.

To solve the non-convex GSC problem, an alternating optimization framework is often employed. This framework treats GSC as two problems, Graph regularized Sparse Representation (GSR) problem and Dictionary Update (DU) problem, and then alternatively solves these two problems until convergence. GSR can be separated into several subproblems with a common form (GSRsub) that is a variant of ℓ_1 minimization. And these subproblems are sequentially solved one after another. Traditional ways of solving GSRsub focus on its primal with a strongly convex but non-differentiable objective. Typical methods for ℓ_1 minimization, such as Feature-sign search (Feature-sign) algorithm [6], Iterative Shrinkage-Threholding Algorithm (ISTA) [7] and its improved version FISTA [8], can be directly extended to the primal problem of GSRsub. Note that Feature-sign was first proposed in [6] for SC and then modified in [2]–[4] for GSRsub. GSRsub can also be treated as an ℓ_1 regularized QP (ℓ_1-QP) problem (see, e.g.,[9], [10]) and solved by the Cyclical Coordinate Descent (CCD) method. Further, a dual form of ℓ_1-QP (D-ℓ_1-QP) was also developed in [10]. In this letter, we propose another dual problem of GSRsub (D-GSRsub) which has two advantages over the primal problem and D-ℓ_1-QP: (1) the number of variables of D-GSRsub is less than that of both D-ℓ_1-QP and the primal problem, and (2) the objective function of D-GSRsub is smooth and strongly convex. These advantages make it possible for fast gradient ascent strategies to be applied to D-GSRsub with lower computational complexities. In this letter, four such strategies are applied to D-GSRsub, and their performances are verified in image classification and clustering experiments on real-world datasets.

II. Graph regularized Sparse Coding

A. Model

Data and their relationships can be modeled by a weighted graph whose nodes are data points $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N)$ and edge weight matrix $\boldsymbol{W} = (w_{ij})_{N\times N}$ represents pairwise relationships between data. Generally, $\boldsymbol{W}$ is symmetric and nonnegative. The idea of GSC is to incorporate the objective of SC model with a graph regularization term: $(1/2)\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$, where $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^N$ are sparse codes of $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ respectively, and w_{ij} describes their pairwise relationship. Thus the GSC model is

$$\begin{aligned}\min_{\boldsymbol{X},\boldsymbol{D}} & \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X} \quad \|_{\mathrm{F}}^2 + \frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \\ & \qquad\qquad + \nu\sum_{i=1}^{N}\|\boldsymbol{x}_i\|_1 \\ \text{s.t} \quad & \|\boldsymbol{d}_l\|_2 \le 1 \ (l = 1, \cdots, K) \end{aligned} \tag{1}$$

Manuscript received July 09, 2014; revised September 11, 2014; accepted September 14, 2014. Date of publication September 17, 2014; date of current version September 29, 2014. This work was supported in part by NNSFC Grants 61210009, 61033011, 61379093 and 11131006. The associate editor coordinating the review for this manuscript and approving it for publication was Prof. Waheed U. Bajwa

R. Jiang and H. Qiao are with State Key Lab of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jiangrui627@163.com; hong.qiao@ia.ac.cn).

B. Zhang is with LSEC & Institute of Applied Mathematics, AMSS, Chinese Academy of Sciences, Beijing 100190, China (e-mail: b.zhang@amt.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2358853

where $\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_K) \in \mathbb{R}^{n\times K}$ is an overcomplete dictionary with $n < K$, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N) \in \mathbb{R}^{n\times N}$ are sparse codes, λ is the graph regularization parameter and ν is the sparsity regularization parameter.

B. Optimization

The GSC model (1) is non-convex, so finding a global solution is unrealistic. However, it is feasible to attain a local minimum by alternatively solving two convex problems:

(i) the Graph regularized Sparse Representation (GSR) problem: Fixing $\boldsymbol{D} = \boldsymbol{D}^{(k)}$, we have

$$\min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{D}^{(k)}\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 + \nu\sum_{i=1}^{N}\|\boldsymbol{x}_i\|_1 \quad (2)$$

(ii) the Dictionary Update (DU) problem: Once $\boldsymbol{X}$ is updated, fix $\boldsymbol{X} = \boldsymbol{X}^{(k+1)}$ and update $\boldsymbol{D}$ by solving

$$\min_{\boldsymbol{D}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}^{(k+1)}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \|\boldsymbol{d}_l\|_2 \le 1 (l = 1, \cdots, K). \quad (3)$$

This letter mainly considers the optimization of GSR. The GSR problem (2) can be split into N subproblems:

$$\min_{\boldsymbol{x}_i} \|\boldsymbol{y}_i - \boldsymbol{D}^{(k)}\boldsymbol{x}_i\|_2^2 + \alpha_i\|\boldsymbol{x}_i\|_2^2 - \boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{x}_i + \nu\|\boldsymbol{x}_i\|_1$$

where $\alpha_i = \lambda\sum_{j\neq i} w_{ij}$ and $\boldsymbol{\beta}_i = 2\lambda(\sum_{j\neq i} w_{ij}\boldsymbol{x}_j)$, $i = 1, \cdots, N$, that is, updating each sparse code $\boldsymbol{x}_i$ individually while holding other sparse codes fixed. Dropping the superscripts and subscripts, we have the GSR subproblem (GSRsub):

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 + \alpha\|\boldsymbol{x}\|_2^2 - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} + \nu\|\boldsymbol{x}\|_1. \quad (4)$$

Traditional ways of dealing with GSRsub are to generalize optimization strategies for ℓ_1 minimization to solve the primal problem (4) which has a strongly convex but non-differentiable objective function. Typical strategies are the Feature-sign search (Feature-sign) algorithm [2]–[4], the ISTA algorithm [7] and its improvement FISTA [8]. Feature-sign solves GSRsub by first iteratively searching and refining the signs of the elements of $\boldsymbol{x}$ and then solving a least square problem based on the search result. ISTA and FISTA consider the minimization of a composite objective function:

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x}) \quad (5)$$

where f and g satisfy the following three assumptions [8]: (i) $g : \mathbb{R}^m \to \mathbb{R}$ is a continuous convex function; (ii) $f : \mathbb{R}^m \to \mathbb{R}$ is convex with Lipschitz continuous first-order derivatives, i.e., $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \le L(\nabla f)\|\boldsymbol{x} - \boldsymbol{y}\|_2$, where $L(\nabla f)$ is the Lipchitz constant of ∇f; (iii) Problem (5) is solvable. Let $f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 + \alpha\|\boldsymbol{x}\|_2^2 - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}$, $g(\boldsymbol{x}) = \nu\|\boldsymbol{x}\|_1$. Then it is obvious that GSRsub is a special case of problem (5) with $m = K$ and $L(\nabla f) = 2(\|\boldsymbol{D}\|_2^2 + \alpha)$, where $\|\boldsymbol{D}\|_2$ is the spectral norm of $\boldsymbol{D}$. ISTA is a first-order method with a sublinear global convergence rate $\mathcal{O}((\|\boldsymbol{D}\|_2^2 + \alpha)/k)$ [8]. And FISTA[1]is an improved ISTA with a better global convergence rate $\mathcal{O}((\|\boldsymbol{D}\|_2^2+\alpha)/(k+1)^2)$ [8]. Here, k is the iteration counter. Except for the two strategies, inspired by other lines of work on sparse representation, such as Graphical LASSO [9], [10], we can also treat the primal problem of GSRsub as a ℓ_1 regularized QP (ℓ_1-QP) problem:

$$\min_{\boldsymbol{x}} (1/2)\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x} + \nu\|\boldsymbol{x}\|_1$$

with $\boldsymbol{A} = 2(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D} + \alpha\boldsymbol{I})$ and $\boldsymbol{\eta} = -2\boldsymbol{D}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{\beta}$, which can be solved by the CCD algorithm [10] with convergence rate $\mathcal{O}(1/k)$ [11]. Further, a dual problem of GSRsub can be derived as a dual problem of ℓ_1-QP (D-ℓ_1-QP):

$$\min_{\boldsymbol{\xi}} (1/2)(\boldsymbol{\xi} + \boldsymbol{b})^{\mathrm{T}}\tilde{\boldsymbol{A}}(\boldsymbol{\xi} + \boldsymbol{b}) + \delta_\nu^\infty(\boldsymbol{\xi})$$

(see [10]) with $\tilde{\boldsymbol{A}} = (1/2)(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D} + \alpha\boldsymbol{I})^{-1}$ and $\boldsymbol{b} = \boldsymbol{\eta}$. Here, $\delta_\nu^\infty(\boldsymbol{\xi})$ is the indicator function of the ℓ_∞ ball of radius ν. Note that the number of variables in D-ℓ_1-QP is the same as in the primal problem of GSRsub and that the objective function of D-ℓ_1-QP is still non-differentiable. In next section, we propose a novel dual form of GSRsub whose objective function is smooth and strongly convex with less variables.

III. Dual Gradient Ascent Strategies for GSRsub

A. Novel Dual Problem of GSRsub (D-GSRsub)

Let $R(\boldsymbol{z}) = \|\boldsymbol{y} - \boldsymbol{z}\|_2^2$ and $r(\boldsymbol{x}) = \alpha\|\boldsymbol{x}\|_2^2 - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} + \nu\|\boldsymbol{x}\|_1$. We recast (4) as an equality constrained problem

$$\min_{\boldsymbol{x}} \quad F_{\mathrm{P}}(\boldsymbol{x}) = R(\boldsymbol{z}) + r(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{z} = \boldsymbol{D}\boldsymbol{x}$$

Its associated Lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\mu}) = R(\boldsymbol{z}) + r(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{x})$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ is the dual variables. Note that both R and r are strictly convex, so their conjugate functions R^* and r^* are well defined. Thus the dual objective function, denoted by $-F_{\mathrm{D}}(\boldsymbol{\mu})$, can be written as follows

$$-F_{\mathrm{D}}(\boldsymbol{\mu}) = \inf_{\boldsymbol{x},\boldsymbol{z}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\mu}) = -R^*(-\boldsymbol{\mu}) - r^*(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}).$$

This leads to a dual problem of (4): $\max_{\boldsymbol{\mu}} -F_{\mathrm{D}}(\boldsymbol{\mu})$. We refer to this problem as D-GSRsub and recast it as

$$\min_{\boldsymbol{\mu}} F_{\mathrm{D}}(\boldsymbol{\mu}) = R^*(-\boldsymbol{\mu}) + r^*(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}) \quad (6)$$

where $R^*(-\boldsymbol{\mu}) = (1/4)\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\mu}$ and $r^*(\boldsymbol{u}) = [1/(4\alpha)]\sum_{j=1}^{K}[\max(|u_j + \beta_j| - \nu, 0)]^2$. Detailed derivation of $R^*(-\boldsymbol{\mu})$ and $r^*(\boldsymbol{u})$ is given in Appendix A.

Before proceeding further, we recall the following properties of strongly convex functions, the proof of which can be found in [12], [13].

Proposition 3.1: f is strongly convex with parameter C if and only if there is a C such that $g(\boldsymbol{x}) = f(\boldsymbol{x}) - (C/2)\|\boldsymbol{x}\|_2^2$ is convex.

Proposition 3.2: Suppose f is closed and strongly convex with parameter C. Then its conjugate f^* is differentiable, and $\nabla f^*(\boldsymbol{u}) = \hat{\boldsymbol{x}}(\boldsymbol{u}) = \arg\max_{\boldsymbol{x}}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x} - f(\boldsymbol{x}))$ is Lipschitz continuous with $L(\nabla f^*) = 1/C$, i.e., f^* is smooth.

Thanks to the graph regularization term, we have $\alpha > 0$. By this fact and the convexity of $\nu\|\boldsymbol{x}\| - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}$, it follows from Proposition 3.1 that r is strongly convex. This, together with Proposition 3.2, implies that r^* is differentiable and $\nabla r^*(\boldsymbol{u}) = \arg\max_{\boldsymbol{x}}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x} - r(\boldsymbol{x}))$ is Lipschitz continuous with $L(\nabla r^*) = 1/(2\alpha)$, i.e., r^* is smooth. Note that

[1]To prevent confusions, we call the strategy using FISTA on the primal problem of GSRsub as *Primal FISTA* (PFISTA) hereafter.

$\nabla r^*(\boldsymbol{u}) = (1/(2\alpha))\mathrm{Soft}(\boldsymbol{u}+\boldsymbol{\beta}, \nu\mathbf{1})$, where $\mathrm{Soft}(\boldsymbol{u}+\boldsymbol{\beta}, \nu\mathbf{1}) = (\mathrm{soft}(u_1+\beta_1, \nu), \mathrm{soft}(u_2+\beta_2, \nu), \cdots, \mathrm{soft}(u_K+\beta_K, \nu))^{\mathrm{T}}$ and $\mathrm{soft}(a, b) = \mathrm{sign}(a)\max(|a|-b, 0)$ is the soft-thresholding function [14]. The derivation of $\nabla r^*(\boldsymbol{u})$ is given in Appendix A.

We conclude this subsection by summarizing the better properties of D-GSRsub: (i) The number of variables in D-GSRsub is n which is less than K, the number of the primal variables and the variables in D-ℓ_1-QP in the overcomplete dictionary setting; (ii) F_{D} is smooth with $\nabla F_{\mathrm{D}}(\boldsymbol{\mu}) = 0.5\boldsymbol{\mu} - \boldsymbol{y} + \boldsymbol{D}\nabla r^*(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu})$ and $L(\nabla F_{\mathrm{D}}) = 0.5 + L(\nabla r^*)\|\boldsymbol{D}\|_2^2$; Moreover, by Proposition 3.1, F_{D} is strongly convex with $C = 0.5$.

B. Dual Gradient Ascent (DGA) Strategies for GSRsub

The better properties of F_{D} lead to efficient gradient ascent strategies for D-GSRsub, i.e., gradient descent strategies for problem (6). In this letter, we only consider four such strategies.

1) *DGA with Fixed Step Size*:Due to the strongly convexity and smoothness of F_{D}, a fixed step size $s = 2/(1 + L(\nabla r^*)\|\boldsymbol{D}\|_2^2)$ can achieve a linear convergence rate $\mathcal{O}(\exp(-2k/(1 + L(\nabla r^*)\|\boldsymbol{D}\|_2^2)))$ [15].
2) *DGA with Variable Step Size*: Inspired by a kind of accelerated ISTA called SpaRSA [16], we employ a progressive quadratic approximation

$$\begin{aligned} Q_{\mathrm{I}}(\boldsymbol{\mu}, \boldsymbol{\mu}^{(k)}) = {} & r^*(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}^{(k)}) \\ & + (\boldsymbol{D}\nabla r^*(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}^{(k)}))^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^{(k)}) \\ & + (\gamma^{(k)}/2)\|\boldsymbol{\mu} - \boldsymbol{\mu}^{(k)}\|_2^2 + R^*(-\boldsymbol{\mu}), \end{aligned}$$

and in each iteration use Barzilai-Borwein (BB) method [17] shown at the bottom of the page, with a subsequent update procedure to choose $\gamma^{(k)}$. Hence, we have $s_{\mathrm{BB}}^{(k)} = 1/(0.5 + \gamma^{(k)})$. As shown in [18], for strongly convex objectives the convergence rate of SpaRSA is R-linear.
3) *Dual FISTA (DFISTA)*: Recall the assumptions for ISTA and FISTA mentioned in Section II-B. Let $f = r^*(\boldsymbol{u})$ and $g = R^*(-\boldsymbol{\mu})$. Problem (6) is also a special case of problem (5) with $m = n$ and $L(\nabla f) = L(\nabla r^*)$. We use a quadratic upper bound approximation of F_D at $\boldsymbol{\mu}^{(k)}$, i.e.,

$$\begin{aligned} Q_{\mathrm{II}}(\boldsymbol{\mu}, \boldsymbol{\mu}^{(k)}) = {} & r^*(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}^{(k)}) \\ & + (\boldsymbol{D}\nabla r^*(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}^{(k)}))^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^{(k)}) \\ & + (L(\nabla r^*)/2)\|\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu} - \boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}^{(k)}\|_2^2 \\ & + R^*(-\boldsymbol{\mu}), \end{aligned}$$

at $(k+1)$ th iteration. Following the rationale of FISTA [8], we can obtain an accelerated dual ascent strategy referred to as *Dual FISTA* (DFISTA). Let $t^{(0)} = 1$ and $\boldsymbol{\zeta}^{(1)} = \boldsymbol{\mu}^{(0)}$. Then the updates are as follows:

$$\begin{aligned} t^{(k)} &= (1 + \sqrt{1 + 4{t^{(k-1)}}^2})/2 \\ \boldsymbol{\zeta}^{(k)} &= \boldsymbol{\mu}^{(k-1)} + \left(\frac{t^{(k-1)} - 1}{t^{(k)}}\right)(\boldsymbol{\mu}^{(k-1)} - \boldsymbol{\mu}^{(k-2)}) \\ \boldsymbol{\mu}^{(k)} &= \arg\min_{\boldsymbol{\mu}} Q_{\mathrm{II}}(\boldsymbol{\mu}, \boldsymbol{\zeta}^{(k)}) = \boldsymbol{\zeta}^{(k)} - \boldsymbol{M}^{-1}\nabla F_{\mathrm{D}}(\boldsymbol{\zeta}^{(k)}), \end{aligned}$$

where $\boldsymbol{M} = 0.5\boldsymbol{I} + L(\nabla r^*)\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}$. The convergence rate of DFISTA is $\mathcal{O}(L(\nabla r^*)\|\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}^*\|_2^2/(k+1)^2)$ [8].
4) *Nesterov's Accelerated DGA (NADGA)*: For problem (6), we can employ Nesterov's Accelerated Gradient Descent [15] for strongly convex and smooth unconstrained cases. This strategy starts at $\boldsymbol{\zeta}^{(0)} = \boldsymbol{\mu}^{(0)}$ and iterates as follows:

$$\begin{aligned} \boldsymbol{\zeta}^{(k)} &= \boldsymbol{\mu}^{(k-1)} \\ &\quad - (1/(0.5 + L(\nabla r^*)\|\boldsymbol{D}\|_2^2))\nabla F_{\mathrm{D}}(\boldsymbol{\mu}^{(k-1)}), \\ \boldsymbol{\mu}^{(k)} &= \left(1 + \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1}\right)\boldsymbol{\zeta}^{(k)} - \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1}\boldsymbol{\zeta}^{(k-1)} \end{aligned}$$

where $Q = 1 + 2L(\nabla r^*)\|\boldsymbol{D}\|_2^2$. The convergence rate of NADGA is $\mathcal{O}(\exp(-(k-1)/\sqrt{Q}))$ [15].

In summary, the discussed optimization strategies for GSRsub, except for Feature-sign with unknown convergence rate, arranged in the increasing order of convergence rate are CCD, PFISTA, DFISTA, DGA $+ s_{\mathrm{BB}}^{(k)}$, DGA $+ s$ and NADGA. Let $\mathcal{C}_{\mathrm{SN}}$ be the complexity to compute $\|\boldsymbol{D}\|_2^2$ and let k be the iteration counter. Then the complexity of CCD and PFISTA is respectively $\mathcal{O}(K^2n) + k \times \mathcal{O}(K^2)$ and $\mathcal{O}(K^2n) + \mathcal{C}_{\mathrm{SN}} + k \times \mathcal{O}(K^2)$, the complexity of DFISTA is $\mathcal{O}(Kn^2) + \mathcal{O}(n^3) + k \times \mathcal{O}(Kn)$, and the complexity of DGA$+s_{\mathrm{BB}}^{(k)}$, DGA$+s$ and NADGA is $\mathcal{C}_{\mathrm{SN}} + k \times \mathcal{O}(Kn)$. Since $K > n$, the complexity of DFISTA, DGA$+s_{\mathrm{BB}}^{(k)}$, DGA $+ s$ and NADGA is lower than that of CCD and PFISTA. This indicates that solving D-GSRsub by DFISTA, DGA $+ s_{\mathrm{BB}}^{(k)}$, DGA $+ s$ and NADGA is faster than solving its primal problem by CCD and PFISTA.

IV. Experiments

We present two experiments on real-world datasets to compare the performance of seven optimization strategies for GSRsub in GSC based methods: Feature-sign [2]–[4], CCD [10], PFISTA [8], DFISTA [8], DGA $+ s_{\mathrm{BB}}^{(k)}$[2][16], DGA $+ s$ [15] and NADGA [15]. For fair comparison, we only employ the Lagrange dual method (LagDual) [6] for solving the DU problem (3). All experiments are performed in Matlab R2009a on a Lenovo Windows 7 PC with Intel Core i3-2120 CPU (3.30 GHz) and 3.5 GB RAM. We set the maximum number of iterations as 5000 for Feature-sign and 10000 for both CCD and PFISTA. We use the relative change of the primal objective, i.e., $|F_{\mathrm{P}}(\boldsymbol{x}^{(k)}) - F_{\mathrm{P}}(\boldsymbol{x}^{(k-1)})|/|F_{\mathrm{P}}(\boldsymbol{x}^{(k-1)})|$, as stopping criteria for both CCD and PFISTA. For the other four, we set the maximum number of iterations as 100000, and choose the step size of the dual variables, i.e., $\|\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(k-1)}\|_2$, as the stopping criterion for DGA$+s_{\mathrm{BB}}^{(k)}$ and the dual gap, i.e., $F_{\mathrm{P}}(\hat{\boldsymbol{x}}^{(k)}) - F_{\mathrm{D}}(\boldsymbol{\mu}^{(k)})$, as stopping criteria for the other three.

The first experiment is learning representations for image clustering using GraphSC [4]. We use CMU PIE database with 68 objects and 42 images of size 32×32 for each object and set $k = 5$ for the heat kernel weighted k-nearest-neighbor graph construction. PCA is performed on the images first to reduce the

[2]The key parameters are set as $M = 0, \sigma = 0.0001$ and $\eta = 2$ in both two experiments. See [16] for more details.

$$\gamma^{(k)} = \frac{(\boldsymbol{\mu}^{(k-1)} - \boldsymbol{\mu}^{(k-2)})^{\mathrm{T}}(\boldsymbol{D}\nabla r^*(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}^{(k-1)}) - \boldsymbol{D}\nabla r^*(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{\mu}^{(k-2)}))}{(\boldsymbol{\mu}^{(k-1)} - \boldsymbol{\mu}^{(k-2)})^{\mathrm{T}}(\boldsymbol{\mu}^{(k-1)} - \boldsymbol{\mu}^{(k-2)})}$$

TABLE I
CLUSTERING RESULTS ON CMU PIE DATASET (THE CLUSTERS NUMBER IS 60 AND THE NUMBER OF ITERATIONS IS 20. THE BEST RESULTS ARE MARKED IN BOLDFACE.)

Strategies for GSC	Accuracy (%)	NMI (%)	Objective Value of GSC Model	CPU Time (s)
Feature-sign + LagDual	58.0 ± 3.8	78.8 ± 1.1	323.1	565.6
CCD + LagDual	56.4 ± 3.7	78.5 ± 1.4	323.6	2098.5
PFISTA + LagDual	58.0 ± 3.8	78.8 ± 1.1	323.1	397.6
DFISTA + LagDual	67.1 ± 4.1	82.7 ± 1.6	331.5	259.9
(DGA+$s_{\mathrm{BB}}^{(k)}$) + LagDual	68.5 ± 3.3	84.5 ± 1.1	323.1	230.0
(DGA+s) + LagDual	69.9 ± 4.0	85.7 ± 1.3	327.0	**227.2**
NADGA + LagDual	$\mathbf{71.3 \pm 2.7}$	$\mathbf{86.2 \pm 0.9}$	328.2	227.4

TABLE II
CLASSIFICATION RESULTS ON 20 CLASSES OF CALTECH 256 (THE NUMBER OF ITERATIONS IS 50. THE LAST COLUMN SHOWS AVERAGE CPU TIME FOR INFERRING A REPRESENTATION FOR ONE IMAGE. THE BEST RESULTS ARE MARKED IN BOLDFACE.)

Strategies for GSC	Accuracy (%)	CPU Time for Training $\boldsymbol{D}$ (s)	Objective Value of GSC Model	CPU Time for Inference (s)
Feature-sign + LagDual	66.8 ± 1.4	39046	3661.0	178.6
CCD + LagDual	66.6 ± 1.1	366060	3661.2	1692.8
PFISTA + LagDual	66.6 ± 1.1	69431	3661.2	543.6
DFISTA + LagDual	67.0 ± 1.4	7099	3661.2	63.9
(DGA+$s_{\mathrm{BB}}^{(k)}$) + LagDual	$\mathbf{67.3 \pm 1.7}$	**3534**	3726.0	48.9
(DGA+s) + LagDual	66.3 ± 1.0	21795	3660.4	197.4
NADGA + LagDual	66.3 ± 0.7	6206	3661.1	**39.5**

dimensionality to 100. The dictionary size is set as 200. By conducting the cross validation described in [4], $\lambda = 1$ and $\nu = 0.1$ are adopted. The threshold values for stopping criteria of CCD, PFISTA and $\mathrm{DGA} + s_{\mathrm{BB}}^{(k)}$ are 10^{-10}, 10^{-10} and 10^{-2}, respectively and the threshold value for dual gap is 10^{-2}.

Clustering is carried out with the clusters number 60, and 10 tests are conducted. The average performances after 20 iterations, evaluated by the accuracy, normalized mutual information (NMI) [4] and CPU time, are reported in Table I. It is observed that the results support our analysis in Section III-B. Compared with solving the primal problem by Feature-sign, CCD and PFISTA, solving D-GSRsub by the given four strategies dramatically reduces the execution time of GSC. Precisely, the CPU time used by $(\mathrm{DGA} + s_{\mathrm{BB}}^{(k)}) + \mathrm{LagDual}$, $(\mathrm{DGA} + s) + \mathrm{LagDual}$ and $\mathrm{NADGA} + \mathrm{LagDual}$ is almost the same. $\mathrm{DFISTA} + \mathrm{LagDual}$ takes a little more CPU time than the former three. The accuracy shows that the given four dual gradient ascent strategies even improve the clustering performance. The objective values in the third column suggest that solving D-GSRsub in our parameters setting may lead to a different local minimum of GSC.

The second experiment is learning representations for image classification by using LScSPM [2], [3]. We select the first to the 20th classes with total 2548 images from Caltech 256 dataset and follow the parameter setting in [2] to extract SIFT features. It should be noted that, for the feasibility of comparison, we randomly sample 10000 SIFT features for training a 128×1024 dictionary by GSC, and employ histogram intersection [2] weighted k-nearest-neighbor graph to model the similarity between SIFT features. $k = 5$ is adopted for graph construction. By following the setting in [2], $\lambda = 0.1$ and $\nu = 0.3$ are used. The threshold values for stopping criteria of CCD, PFISTA and $\mathrm{DGA} + s_{\mathrm{BB}}^{(k)}$ are 10^{-10}, 10^{-10} and 10^{-6}, respectively. The threshold value for dual gap is 10^{-6}. In inference, we use the learned dictionary to encode each SIFT feature by solving a GSRsub (4). Finally, we generate a representation for each image by spatial-pyramid max pooling the corresponding SIFT sparse codes. In classification, we randomly select 60 images from each class and use their representations to train a linear SVM classifier. We then carry out classification on the others. 10 tests are conducted, and the average performances after 50 iterations are reported in Table II. The performances of the seven strategies are evaluated by the accuracy, the CPU time for training the dictionary by GSC and the average CPU time for inferring a representation for one image. The objective values of GSC model are also reported. Table II shows that the accuracy of the seven strategies is almost the same, but $(\mathrm{DGA} + s) + \mathrm{LagDual}$, $(\mathrm{DGA} + s_{\mathrm{BB}}^{(k)}) + \mathrm{LagDual}$, $\mathrm{DFISTA} + \mathrm{LagDual}$ and $\mathrm{NADGA} + \mathrm{LagDual}$ which solve D-GSRsub consume less time in training and inference. Further, $(\mathrm{DGA} + s_{\mathrm{BB}}^{(k)}) + \mathrm{LagDual}$ is the fastest one in training, and NADGA is the fastest one in inference.

V. CONCLUSION

In this letter, we proposed to speed up GSC by solving a novel dual problem of GSRsub (D-GSRsub) which is smooth and strongly convex with less variables, compared with its primal problem and the existing dual form (D-ℓ_1-QP). Based on this dual form, four efficient gradient ascent strategies for D-GSRsub have been introduced to speed up GSC. Image classification and clustering experiments on two real-world datasets are conducted to illustrate the fast and stable performance of the four strategies.

APPENDIX A
DERIVATION OF $R^*(-\boldsymbol{\mu})$, $\nabla r^*(\boldsymbol{u})$ AND $r^*(\boldsymbol{u})$

It is easy to verify that the conjugate of R is

$$R^*(-\boldsymbol{\mu}) = \sup_{\boldsymbol{z}}((-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{z} - \|\boldsymbol{y}-\boldsymbol{z}\|_2^2) = (1/4)\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\mu}.$$

By Proposition 3.2 and the strong convexity of r, it follows that

$$\begin{aligned}\nabla r^*(\boldsymbol{u}) = \hat{\boldsymbol{x}} &= \operatorname*{argmin}_{\boldsymbol{x}}((\boldsymbol{u}+\boldsymbol{\beta}^{\mathrm{T}})\boldsymbol{x} - \alpha\|\boldsymbol{x}\|_2^2 - \nu\|\boldsymbol{x}\|_1) \\ &= \operatorname*{argmin}_{\boldsymbol{x}}(\alpha\|\boldsymbol{x} - (1/(2\alpha))(\boldsymbol{u}+\boldsymbol{\beta})\|_2^2 \\ &\quad + \nu\|\boldsymbol{x}\|_1)\end{aligned}$$

The necessary and sufficient condition for $\hat{\boldsymbol{x}}$ to be the optimal point of the above optimization problem is that the subgradient of its objective function equals to 0, that is,

$$u_i + \beta_i = \begin{cases} 2\alpha\hat{x}_i + \nu & \hat{x}_i > 0 \\ c & \hat{x}_i = 0 \\ 2\alpha\hat{x}_i - \nu & \hat{x}_i < 0 \end{cases} \tag{7}$$

where $c \in [-\nu, \nu]$. Thus, for all i, either $\hat{x}_i = 0$ or $\hat{x}_i = (1/(2\alpha))(u_i + \beta_i - \mathrm{sign}(\hat{x}_i)\nu)$. If $\hat{x}_i \neq 0$ then $|u_i + \beta_i| > \nu$ and $\mathrm{sign}(u_i + \beta_i) = \mathrm{sign}(\hat{x}_i)$, so $\hat{x}_i = (1/(2\alpha))(u_i + \beta_i - \mathrm{sign}(u_i + \beta_i)\nu)$. If $\hat{x}_i = 0$ then $|u_i + \beta_i| \leq \nu$. Conversely, if we set $\hat{x}_i = 0$ when $|u_i + \beta_i| \leq \nu$ and $\hat{x}_i = (1/(2\alpha))(u_i + \beta_i - \mathrm{sign}(u_i + \beta_i)\nu)$ when $|u_i + \beta_i| > \nu$, then (7) holds. This implies that $\nabla_i r^*(\boldsymbol{u}) = \hat{x}_i = (1/(2\alpha))\mathrm{sign}(u_i+\beta_i)\max\{|u_i+\beta_i| - \nu, 0\} = (1/(2\alpha))\mathrm{soft}(u_i + \beta_i, \nu)$ Substituting $\hat{\boldsymbol{x}}$ into $r^*(\boldsymbol{u}) = \boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{x}} - r(\hat{\boldsymbol{x}})$ gives the desired expression of $r^*(\boldsymbol{u})$.)

ACKNOWLEDGMENT

The authors thank the referees for their comments.

REFERENCES

[1] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1794–1801.

[2] S. Gao, I. W. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely–laplacian sparse coding for image classification," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 3555–3561.

[3] S. Gao, I. W. Tsang, and L.-T. Chia, "Laplacian sparse coding, hypergraph laplacian sparse coding, and applications," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 35, no. 1, pp. 92–104, 2013.

[4] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, 2011.

[5] A. Shaban, H. Rabiee, M. Farajtabar, and M. Ghazvininejad, "From local similarity to global coding: An application to image classification," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 2794–2801.

[6] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA, USA: MIT Press, 2007, pp. 801–808.

[7] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[8] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[9] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[10] R. Mazumder and T. Hastie, "The graphical lasso: New insights and alternatives," *Electron. J. Statist.*, vol. 6, pp. 2125–2149, 2012.

[11] A. Saha and A. Tewari, "On the non-asymptotic convergence of cyclic coordinate descent methods," *SIAM J. Optim.*, vol. 23, no. 1, pp. 576–601, 2013.

[12] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.

[13] S. Shalev-Shwartz, "Online learning: Theory, algorithms, and applications," Ph. D. dissertation, The Hebrew University of Jerusalem, 2007 [Online]. Available: http://www.cs.huji.ac.il/~shais

[14] D. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[15] Y. Nesterov*, Introductory Lectures on Convex Optimization: A Basic Course*. Norwell, MA, USA: Kluwer, 2004.

[16] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, 2009.

[17] J. Barzilai and J. Borwein, "Two point step size gradient methods," *IMA J. Numer. Anal.*, vol. 8, no. 1, pp. 141–148, 1988.

[18] W. Hager, D. Phan, and H. Zhang, "Gradient-based methods for sparse recovery," *SIAM J. Imag. Sci.*, vol. 4, no. 1, pp. 146–165, 2011.