Dual Augmented Lagrangian Method for Efficient Sparse Reconstruction

Ryota Tomioka* and Masashi Sugiyama†

Abstract

We propose an efficient algorithm for sparse signal reconstruction problems. The proposed algorithm is an augmented Lagrangian method based on the dual problem. It is efficient when the number of unknown variables is much larger than the number of observations because of the dual formulation. Moreover, the primal variable is explicitly updated and the sparsity in the solution is exploited. Numerical comparison with the state-of-the-art algorithms shows that the proposed algorithm is favorable when the design matrix is poorly conditioned or dense and very large.

EDICS category: SAS-STAT, SAS-MALN

I. Introduction

Sparse signal reconstruction has recently gained considerable interests in signal/image processing and machine learning. Sparsity is often a natural assumption in inverse problems, such as MEG/EEG source localization and image/signal deconvolution; sparsity enables us to identify a small number of active components even when the dimension is much larger than the number of observations. In addition, a

RT was supported by the Global COE program (Computationism as a Foundation for the Sciences). MS is supported by MEXT Grant-in-Aid for Young Scientists (A) 20680007, SCAT, and AOARD. This work was done while RT was at Tokyo Institute of Technology.

*(Corresponding author) Department of Mathematical Informatics, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan. TEL: +81-3-5841-6898, FAX: +81-3-5841-6897, E-mail: tomioka@mist.i.u-tokyo.ac.jp.

† Department of Computer Science, Tokyo Institute of Technology, 2-12-1-W8-74, O-okayama, Meguro-ku, Tokyo 152-8552, Japan.

sparse model is also valuable in predictive tasks because it can explain why it is able to make a prediction in contrast to black-box models such as neural networks and support vector machines.

In this paper we consider the following particular problem that typically arises in sparse reconstruction:

$$\text{(P)} \qquad \underset{\boldsymbol{w}\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{A}\boldsymbol{w}-\boldsymbol{b}\|^2+\lambda\|\boldsymbol{w}\|_1 =: f(\boldsymbol{w}), \qquad (1)$$

where $\boldsymbol{w}\in\mathbb{R}^n$ is the coefficient vector to be estimated, $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ is the design matrix, and $\boldsymbol{b}\in\mathbb{R}^m$ is the vector of observations; $f(\boldsymbol{w})$ denotes the objetive function. It is well known that the ℓ_1-norm penalty enforces $\boldsymbol{w}$ to have many zero elements. It is called lasso [1] in the statistics, basis pursuit denoising [2] in the signal processing, and FOCUSS [3] in the brain imaging communities.

Various methods have been proposed to efficiently solve the optimization problem (1) (or its generalized versions). Iteratively reweighted shrinkage (IRS) is a popular approach for solving the problem (1) (see [3]-[7]). The main idea of the IRS approach is to replace a non-differentiable (or non-convex) optimization problem by a series of differentiable convex ones; typically the regularizer (e.g., $\|\cdot\|_1$ in Eq. (1)) is upper bounded by a weighted quadratic regularizer. Then one can use various existing algorithms to minimize the upper bound. The upper bound is tightened after every minimization so that the solution eventually converges to the solution of the original problem (1). The challenge in the IRS framework is the *singularity* [7] around the coordinate axis. For example, in the ℓ_1 problem in Eq. (1), any zero component $w_j = 0$ in the initial vector $\boldsymbol{w}$ will remain zero after any number of iterations. Moreover, it is possible to create a situation that the convergence becomes arbitrarily slow for finite $|w_j|$ because the convergence in the ℓ_1 case is only linear [3]. Another recent work is the split Bregman iteration (SBI) method [8], which is related to the Bregman iteration algorithm [9]. The Bregman iteration algorithm can be considered as an augmented Lagrangian (AL) method (see [9]-[11]). By introducing an auxiliary variable $\tilde{\boldsymbol{w}}$, the SBI approach decouples the minimization of the first and the second term in Eq. (1), which can then be handled independently. The two variables $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ are gradually enforced to coincide with each other. Both IRS and SBI require solving a linear system of the size of the number of unknown variables (n) repeatedly, which may become challenging when $n \gg m$.

Kim *et al.* [12] developed an efficient interior-point (IP) method called l1_ls. They proposed a truncated Newton method for solving the inner minimization that scales well when the design matrix $\boldsymbol{A}$ is sparse.

The iterative shrinkage/thresholding (IST) (see [9], [13]-[15]) is a classic method but it is still an area of active research [16], [17]. It alternately computes the steepest descent direction on the loss term in Eq. (1) and the *soft thresholding* related to the regularization term. The IST method has the advantage that every iteration is extremely light (only computes gradient) and every intermediate solution is sparse.

However the naive version of IST is sensitive to the selection of step-size. Recently several authors have proposed intelligent step-size selection criteria [16], [17].

In this paper we propose an efficient method that scales well when $n \gg m$, which we call the dual augmented Lagrangian (DAL). It is an AL method similarly to the SBI method but it is applied to the dual problem; thus the inner minimization is efficient when $n \gg m$. In addition, in contrast to the "divide and conquer" approach of SBI, the inner minimization can be performed jointly over all the variables; it converges *super linearly* because the inner minimization is solved at sufficient precision (see [10], [11]). Moreover, although the proposed method is based on the dual problem, the primal variable is explicitly updated in the computation as the Lagrangian multiplier. DAL computes soft thresholding after every iteration similarly to the IST approach but with an *improved direction* as well as an automatic step-size selection mechanism; typically the number of outer iterations is less than 10. The proposed approach can be applied to large scale problems with *dense* design matrices because it exploits the sparsity in the coefficient vector $\boldsymbol{w}$ in contrast to the IP methods [12], which exploits the sparsity in the design matrix.

This paper is organized as follows. In Sec. II, the DAL algorithm is presented; two approaches for the inner minimization problem are discussed. In Sec. III, we experimentally compare DAL to the state-of-the-art SpaRSA [17] and l1_ls [12] algorithms. We give a brief summary and future directions in Sec. IV.

II. Method

A. Dual augmented Lagrangian method for sparse reconstruction

The challenge in minimizing Eq. (1) arises from its non-differentiability. The proposed approach is based on the minimization of a differentiable surrogate function $f_\eta(\boldsymbol{w})$, which we derive from the augmented Lagrangian function L_η of the dual problem of Eq. (1).

Using the Fenchel duality (see [18, Sec. 5.4]) and a splitting similar to SBI (in the dual), we obtain the following dual problem of problem (1) (see also [15]):

$$\text{(D)} \qquad \underset{\boldsymbol{v}\in\mathbb{R}^n,\boldsymbol{\alpha}\in\mathbb{R}^m}{\text{maximize}} \qquad \underbrace{-\frac{1}{2}\|\boldsymbol{\alpha}-\boldsymbol{b}\|_2^2+\frac{1}{2}\|\boldsymbol{b}\|_2^2-\delta_\lambda^\infty(\boldsymbol{v}),}_{=:\,d(\boldsymbol{\alpha},\boldsymbol{v})} \tag{2}$$

$$\text{subject to} \qquad \boldsymbol{v}=\boldsymbol{A}^\top\boldsymbol{\alpha}, \tag{3}$$

where $\delta_\lambda^\infty(\boldsymbol{v})$ is the indicator function [15] of the ℓ_∞ ball of radius λ, i.e., $\delta_\lambda^\infty(\boldsymbol{v})=0$ if $\|\boldsymbol{v}\|_\infty\leq\lambda$, and $+\infty$ otherwise. It can be shown that the strong duality holds, i.e., the maximum of Eq. (2) $d(\boldsymbol{\alpha}^*,\boldsymbol{v}^*)$ coincides with the minimum of Eq. (1) $f(\boldsymbol{w}^*)$, where $\boldsymbol{w}^*$ and $(\boldsymbol{\alpha}^*,\boldsymbol{v}^*)$ are the minimizer and the maximizer of the primal and dual problems, respectively.

The *augmented Lagrangian (AL) function* of the dual problem (Eqs. (2) and (3)) is defined as follows:

$$L_\eta(\boldsymbol{\alpha}, \boldsymbol{v}; \boldsymbol{w}) = d(\boldsymbol{\alpha}, \boldsymbol{v}) - \boldsymbol{w}^\top \left(\boldsymbol{A}^\top \boldsymbol{\alpha} - \boldsymbol{v}\right) - \frac{\eta}{2}\|\boldsymbol{A}^\top\boldsymbol{\alpha} - \boldsymbol{v}\|_2^2, \tag{4}$$

where $\boldsymbol{w}$ is the Lagrangian multiplier associated with the equality constraint (Eq. (3)) and corresponds to the coefficient vector in the primal problem. The last term in Eq. (4) is called the barrier term and $\eta \geq 0$ is called the barrier parameter. When $\eta = 0$, the AL function is reduced to the ordinary Lagrangian function. See [10], [11] for the details of the AL method. See also [19] for the ordinary Lagrangian duality. Now we define the surrogate function $f_\eta(\boldsymbol{w})$ as follows:

$$f_\eta(\boldsymbol{w}) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^m, \boldsymbol{v}\in\mathbb{R}^n} L_\eta(\boldsymbol{\alpha}, \boldsymbol{v}; \boldsymbol{w}). \tag{5}$$

Note that from the strong duality $f_0(\boldsymbol{w}) = \max_{\boldsymbol{\alpha},\boldsymbol{v}} L_0(\boldsymbol{\alpha}, \boldsymbol{v}; \boldsymbol{w}) = f(\boldsymbol{w})$. In addition, since $L_0(\boldsymbol{\alpha}, \boldsymbol{v}; \boldsymbol{w}) \geq L_\eta(\boldsymbol{\alpha}, \boldsymbol{v}; \boldsymbol{w})$, the inequality $f(\boldsymbol{w}) \geq f_\eta(\boldsymbol{w})$ holds. Moreover, since $f_\eta(\boldsymbol{w}) \geq L_\eta(\boldsymbol{\alpha}^*, \boldsymbol{v}^*; \boldsymbol{w}) = d(\boldsymbol{\alpha}^*, \boldsymbol{v}^*) = f(\boldsymbol{w}^*)$ (we used $\boldsymbol{A}^\top\boldsymbol{\alpha}^* = \boldsymbol{v}^*$ to obtain the first equality), we have $\min_{\boldsymbol{w}\in\mathbb{R}^n} f_\eta(\boldsymbol{w}) = f(\boldsymbol{w}^*)$ for any nonnegative η. Furthermore, $f_\eta(\boldsymbol{w})$ is differentiable if $\eta > 0$.

The maximization with respect to $\boldsymbol{v}$ in Eq. (5) can be carried out analytically and $\boldsymbol{v}$ can be eliminated from Eq. (4) as follows:

$$\begin{aligned}
&\max_{\boldsymbol{v}\in\mathbb{R}^n} L_\eta(\boldsymbol{\alpha}, \boldsymbol{v}; \boldsymbol{w}) - c(\boldsymbol{w}, \eta) \\
&= -\frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{b}\|_2^2 - \min_{\boldsymbol{v}\in\mathbb{R}^n}\left(\delta_\lambda^\infty(\boldsymbol{v}) + \frac{\eta}{2}\left\|\boldsymbol{v} - \boldsymbol{A}^\top\boldsymbol{\alpha} - \boldsymbol{w}/\eta\right\|_2^2\right) \\
&= -\frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{b}\|_2^2 - \frac{\eta}{2}\|\boldsymbol{A}^\top\boldsymbol{\alpha} + \boldsymbol{w}/\eta - P_\lambda^\infty(\boldsymbol{A}^\top\boldsymbol{\alpha} + \boldsymbol{w}/\eta)\|_2^2 \\
&= -\frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{b}\|_2^2 - \frac{\eta}{2}\|\mathsf{ST}_\lambda(\boldsymbol{A}^\top\boldsymbol{\alpha} + \boldsymbol{w}/\eta)\|_2^2 =: L_\eta(\boldsymbol{\alpha}; \boldsymbol{w}),
\end{aligned} \tag{6}$$

where $c(\boldsymbol{w}, \eta)$ is a constant that only depends on $\boldsymbol{w}$ and η, and P_λ^∞ is a projection on the ℓ_∞ ball of radius λ; note that $\eta P_\lambda^\infty(\boldsymbol{w}) = P_{\eta\lambda}^\infty(\eta\boldsymbol{w})$; in addition, we define the well known *soft thresholding function* ST_λ (see [13]-[15],[9]) as follows:

$$\mathsf{ST}_\lambda(\boldsymbol{w}) = \boldsymbol{w} - P_\lambda^\infty(\boldsymbol{w}) = \left(\max(|w_j| - \lambda, 0)\frac{w_j}{|w_j|}\right)_j \qquad (j = 1, \ldots, n). \tag{7}$$

The coefficient vector $\boldsymbol{w}$ is updated using the gradient of $f_\eta(\boldsymbol{w})$ as follows:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \eta_k(\boldsymbol{A}^\top\boldsymbol{\alpha}_k - \boldsymbol{v}_k), \tag{8}$$

because $\nabla_{\boldsymbol{w}} f_\eta(\boldsymbol{w}_k) = \nabla_{\boldsymbol{w}} L_\eta(\boldsymbol{\alpha}_k, \boldsymbol{v}_k; \boldsymbol{w}_k) = -(\boldsymbol{A}^\top\boldsymbol{\alpha}_k - \boldsymbol{v}_k)$, where $(\boldsymbol{\alpha}_k, \boldsymbol{v}_k)$ is the maximizer of Eq. (5) at the current $\boldsymbol{w}_k$. Moreover, we can show that $(\boldsymbol{\alpha}_k, \boldsymbol{v}_k)$ also maximizes $L_0(\boldsymbol{\alpha}, \boldsymbol{v}; \boldsymbol{w}_{k+1})$ [10,

```
Choose sequences $\eta_1 \leq \eta_2 \leq \cdots$ and $\epsilon_1 \geq \epsilon_2 \geq \cdots$. Let $\boldsymbol{w}_1$ be the initial primal vector. Let $k = 1$.
while Stopping criterion is not satisfied do
    Let $\boldsymbol{\alpha}_k$ be an (approximate) minimizer of $\varphi_k(\boldsymbol{\alpha}) := -L_{\eta_k}(\boldsymbol{\alpha}; \boldsymbol{w}_k)$ (see Eq. (6)) with tolerance $\epsilon_k$ as follows:
        $\boldsymbol{\alpha}_k \simeq \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha} \in \mathbb{R}^m} \left( \frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{b}\|_2^2 + \frac{\eta_k}{2} \left\| \mathsf{ST}_\lambda \left( \boldsymbol{A}^\top \boldsymbol{\alpha} + \boldsymbol{w}_k/\eta_k \right) \right\|_2^2 \right),$   (9)
    where $\|\nabla_{\boldsymbol{\alpha}} \varphi_k(\boldsymbol{\alpha}_k)\|_2 \leq \epsilon_k$. See Eq. (11) for the expression of the gradient $\nabla_{\boldsymbol{\alpha}} \varphi_k(\boldsymbol{\alpha})$.
    Update the primal coefficient vector $\boldsymbol{w}_k$ as:
        $\boldsymbol{w}_{k+1} = \mathsf{ST}_{\lambda\eta_k} \left( \boldsymbol{w}_k + \eta_k \boldsymbol{A}^\top \boldsymbol{\alpha}_k \right).$   (10)
    $k \leftarrow k + 1$.
end while
```

Fig. 1. Dual augmented Lagrangian (DAL) algorithm.

Chap.5]; thus $f(\boldsymbol{w}_{k+1}) = L_0(\boldsymbol{\alpha}_k, \boldsymbol{v}_k; \boldsymbol{w}_{k+1}) = d(\boldsymbol{\alpha}_k, \boldsymbol{v}_k) - \boldsymbol{w}_k{}^\top(\boldsymbol{A}^\top\boldsymbol{\alpha}_k - \boldsymbol{v}_k) - \eta\|\boldsymbol{A}^\top\boldsymbol{\alpha}_k - \boldsymbol{v}_k\|^2 \leq L_\eta(\boldsymbol{\alpha}_k, \boldsymbol{v}_k, \boldsymbol{w}_k) = f_\eta(\boldsymbol{w}_k) \leq f(\boldsymbol{w}_k)$, where the first inequality is strict whenever $\|\boldsymbol{A}^\top\boldsymbol{\alpha}_k - \boldsymbol{v}_k\| > 0$. The barrier parameter η is increased as $\eta_1 \leq \eta_2 \leq \cdots$; this guarantees super linear convergence of the method (see [10]). Accordingly the dual augmented Lagrangian method can be described as in Fig. 1. Note that Eq. (10) is obtained by substituting $\boldsymbol{v}_k = P_\lambda^\infty(\boldsymbol{A}^\top\boldsymbol{\alpha}_k + \boldsymbol{w}_k/\eta_k)$ into Eq. (8).

B. Inner minimization

The inner minimization of $\varphi_k(\boldsymbol{\alpha})$ in Eq. (9) can be efficiently performed through the Newton method because of the special structures of the gradient and the Hessian matrix of $\varphi_k(\boldsymbol{\alpha})$. $\varphi_k(\boldsymbol{\alpha})$ is once differentiable everywhere and also twice differentiable except the points on which the above soft thresholding function switches. The gradient and the Hessian of the objective function $\varphi_k(\boldsymbol{\alpha})$ can be written as follows:

$$\nabla_{\boldsymbol{\alpha}} \varphi_k(\boldsymbol{\alpha}) = \boldsymbol{\alpha} - \boldsymbol{b} + \boldsymbol{A}\mathsf{ST}_{\lambda\eta_k}(\boldsymbol{q}), \tag{11}$$

$$\nabla_{\boldsymbol{\alpha}}^2 \varphi_k(\boldsymbol{\alpha}) = \boldsymbol{I}_m + \eta_k \boldsymbol{A}_+ {\boldsymbol{A}_+}^\top, \tag{12}$$

where $\boldsymbol{q} = \boldsymbol{w}_k + \eta_k \boldsymbol{A}^\top \boldsymbol{\alpha}$, $\boldsymbol{I}_m$ is the identity matrix of size m, and $\boldsymbol{A}_+$ is the submatrix of $\boldsymbol{A}$ that consists of "active" columns with indices $\mathcal{J}_+ = \{j \in \{1, 2, \ldots, n\} : |q_j| > \lambda\eta_k\}$. Note that in both the computation of the gradient and the Hessian, computational complexity is only proportional to the number of active components of $\boldsymbol{q}$. Thus the sparser the (intermediate) solution becomes the faster the computation of a Newton step becomes. The discontinuity of the second derivative is in general not a problem. In fact, we can see from the complementary slackness condition that for finite η the optimal solution-multiplier pair $(\boldsymbol{w}^*, \boldsymbol{\alpha}^*)$ is on a regular point; thus the convergence around the minimum is quadratic.

We propose two approaches for solving the Newton system $\nabla^2_{\boldsymbol{\alpha}}\varphi_k(\boldsymbol{\alpha})\boldsymbol{y} = -\nabla_{\boldsymbol{\alpha}}\varphi_k(\boldsymbol{\alpha})$. The first approach (DALchol) uses the Cholesky factorization of the Hessian matrix $\nabla^2_{\boldsymbol{\alpha}}\varphi_k(\boldsymbol{\alpha})$. The second approach (DALcg) uses a preconditioned conjugate gradient method (the truncated Newton method in [12]) with a preconditioner that only consists of the diagonal elements of the Hessian matrix. Finally the standard backtracking line-search with initial step-size 1 is applied to guarantee decrease in the objective $\varphi_k(\boldsymbol{\alpha})$.

III. Empirical comparisons

We test the computational efficiency of the proposed DAL algorithm on the ℓ_2-ℓ_1 problem (Eq. (1)) under various conditions. The DAL algorithm is compared to two state-of-the-art algorithms, namely l1_ls (interior-point algorithm, [12]) and SpaRSA (step-size improved IST, [17]).

A. Experimental settings

In the first experiment (Fig. 2(a)), the elements of the design matrix $\boldsymbol{A}$ are sampled from the independent zero-mean Gaussian distribution with variance $1/(2n)$. This choice of variance makes the largest singularvalue of $\boldsymbol{A}$ approximately one [17]. The true coefficient vector $\boldsymbol{w}_0$ is generated by randomly filling 4% of its elements by $+1$ or -1 which is also randomly chosen. The remaining elements are zero. The target vector $\boldsymbol{b}$ is generated as $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{w}_0 + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is sampled from the zero-mean Gaussian distribution with variance 10^{-4}. The number of observations (m) is increased from $m = 128$ to $m = 8,196$ while the number of variables (n) is increased proportionally as $n = 4m$. The regularization constant λ is kept constant at 0.025, which is found to approximately correspond to the choice $\lambda = 0.1\|\boldsymbol{A}^\top\boldsymbol{b}\|_\infty$ in [17]. In the second experiment (Fig. 2(b)), the setting is almost the same except that the singular values of $\boldsymbol{A}$ is replaced by a series decreasing as $1/s$ for the s-th singular value. Thus the condition number (the ratio between the smallest and the largest singular values) of $\boldsymbol{A}$ is m. Additionally we set the variance of $\boldsymbol{\xi}$ to zero (no noise) and keep λ constant at 0.0003, which is also found to approximately correspond to the setting in [17]. In the last experiment (Fig. 3), the number of observations (m) is kept at $m = 1,024$ and the number of samples (n) is increased from $n = 4,096$ to $n = 1,048,576$. The design matrix $\boldsymbol{A}$ and the target vector $\boldsymbol{b}$ are generated as in the first experiment. In addition, the regularization constant λ is decreased as $\lambda = 1.6/n^{1/2}$, which equals 0.025 at $n = 4,096$ and is again chosen to approximately match the setting in [17]. In each figure, we show the computation time, the number of iterations, and the sparsity of the solution (the proportion of non-zero elements in the final solution) from top to bottom. All the results are averaged over 10 random initial coefficient vectors $\boldsymbol{w}$. All the experiments are run on MATLAB 7.7 (R2008b) on a workstation with two 3.0GHz quad-core Xeon processors and 16GB of memory.

B. Practical issues

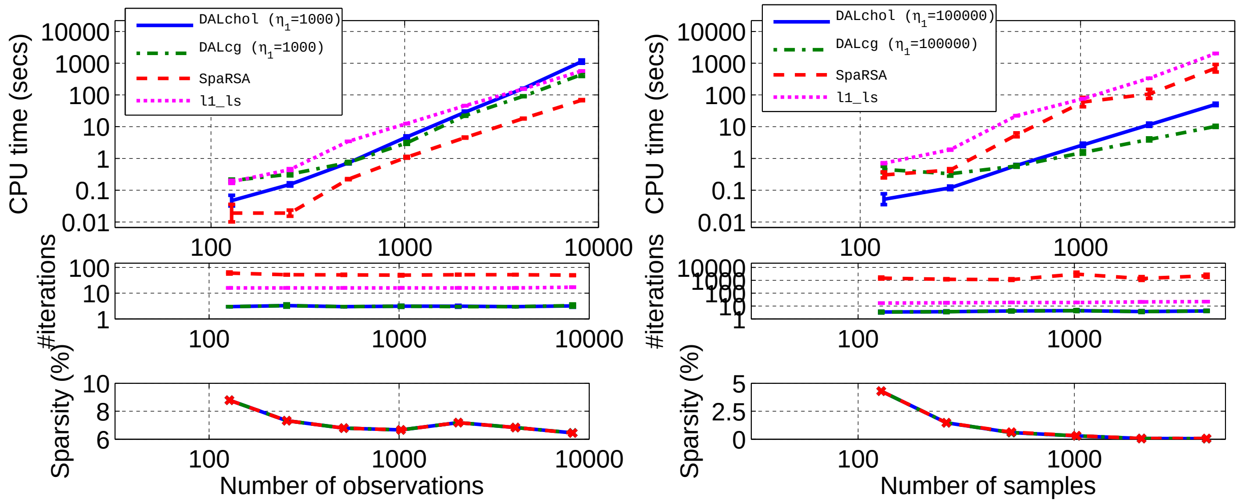
1) Stopping criterion: We use the "duality" stopping criterion proposed in [17] for all the results presented here. More precisely, we generate a dual variable $\hat{\boldsymbol{\alpha}}$ as $\hat{\boldsymbol{\alpha}} = \lambda\tilde{\boldsymbol{\alpha}}/\|\boldsymbol{A}^\top\tilde{\boldsymbol{\alpha}}\|_\infty$, where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}$ is the gradient of the primal loss term in Eq. (1). The above defined $\hat{\boldsymbol{\alpha}}$ is a feasible point of the dual problem (Eq. (2)) by definition, i.e., $\|\boldsymbol{A}^\top\hat{\boldsymbol{\alpha}}\|_\infty \leq \lambda$. Thus we use the primal-dual pair $(\boldsymbol{w}, \hat{\boldsymbol{\alpha}})$ to measure the relative duality gap $(f(\boldsymbol{w}) - d(\hat{\boldsymbol{\alpha}}, \boldsymbol{A}^\top\hat{\boldsymbol{\alpha}}))/f(\boldsymbol{w})$, where f and d are the objective functions in the primal problem (Eq. (1)) and the dual problem (Eq. (2)), respectively. The tolerance 10^{-3} is used.

2) Hyperparameters: The tolerance parameter ϵ_k for the inner minimization is chosen as follows. We use $\epsilon_1 = 10^{-4} \cdot m^{1/2}$ and decrease ϵ_k as $\epsilon_k = \epsilon_{k-1}/2$. Using larger ϵ_k results in cheaper inner minimization but often requires a larger number of outer iterations. The barrier parameter η_k also affects the behavior of the algorithm. Typically larger η_k gives larger reduction in the duality gap at every iteration but makes the inner minimization more difficult. Additionally the best value of η_k depends on the size of the problem, regularization constant λ, and the spectrum of $\boldsymbol{A}$. Here we manually choose η_1 for each problem and increase η_k as $\eta_k = 2\eta_{k-1}$, which guarantees the super-linear convergence [10].

C. Results

When the data is well conditioned (Fig. 2(a)), SpaRSA performs clearly the best within the three algorithms. The proposed DAL algorithm with the conjugate gradient (DALcg) performs comparably to l1_ls. The proposed DAL with the Cholesky factorization (DALchol) is less efficient than DALcg when m is large because the complexity grows as $O(m^3)$; note however that the cost for building the Hessian matrix is only $O(m^2 n_+)$, where n_+ is the number of active components (see Eq. (12)).

In contrast, when the data is poorly conditioned (Fig. 2(b)), the proposed DALcg runs almost 100 times faster than SpaRSA at most. This can be clearly seen in the number of iterations (the middle row). Although the numbers of iterations DAL and l1_ls require are almost constant from Fig. 2(a) to Fig. 2(b), that of SpaRSA is increased at least by the factor 10. Note that the sparsity of the solution is decreasing as the number of samples increases. This may explain why the proposed DAL algorithm is more robust to poor conditioning than l1_ls because l1_ls does not exploit the sparsity in the solution.



(a) Normal conditioning

(b) Poor conditioning

Fig. 2. Comparison of running time and number of iterations of DAL, SpaRSA and l1_ls for problems of various sizes with (a) design matrix $\boldsymbol{A}$ generated from independent normal random variables and (b) the same matrix with singular values replaced by a power-law distribution. The horizontal axis denotes the number of observations (m). The number of variables is $n = 4m$. The regularization constant λ is fixed at $\lambda = 0.025$ in (a) and $\lambda = 0.0003$ in (b). Note that in the second row the number of iterations spent by DALchol and DALcg are the same and in the third row all methods found solutions with the same sparsity except l1_ls, whose sparsity is always 100% (no sparsity).

Finally we compare the three algorithms for very large problems in Fig. 3. Clearly the proposed DAL has milder scaling to the dimensionality than both SpaRSA and l1_ls. This is because the proposed DAL algorithm is based on the dual problem (Eq. (2)). The computational efficiency of DALchol and DALcg is comparable because m is kept constant in this experiment.

IV. Conclusion

In this paper we have proposed a new optimization framework for sparse signal reconstruction, which converges super-linearly. It is based on the dual sparse reconstruction problem. The sparsity of the coefficient vector $\boldsymbol{w}$ is explicitly used in the algorithm. Numerical comparisons have shown that the proposed DAL algorithm is favorable against a state-of-the-art algorithm SpaRSA when the design matrix $\boldsymbol{A}$ is poorly conditioned or $m \ll n$. In fact, it has solved problems with millions of variables in less than 20 minutes even when the design matrix $\boldsymbol{A}$ is dense. In addition, for dense $\boldsymbol{A}$, DAL has shown improved efficiency to l1_ls in most cases. Future work includes generalization of DAL to other loss functions and sparsity measures, continuation strategy, and approximate minimization of the inner problem.

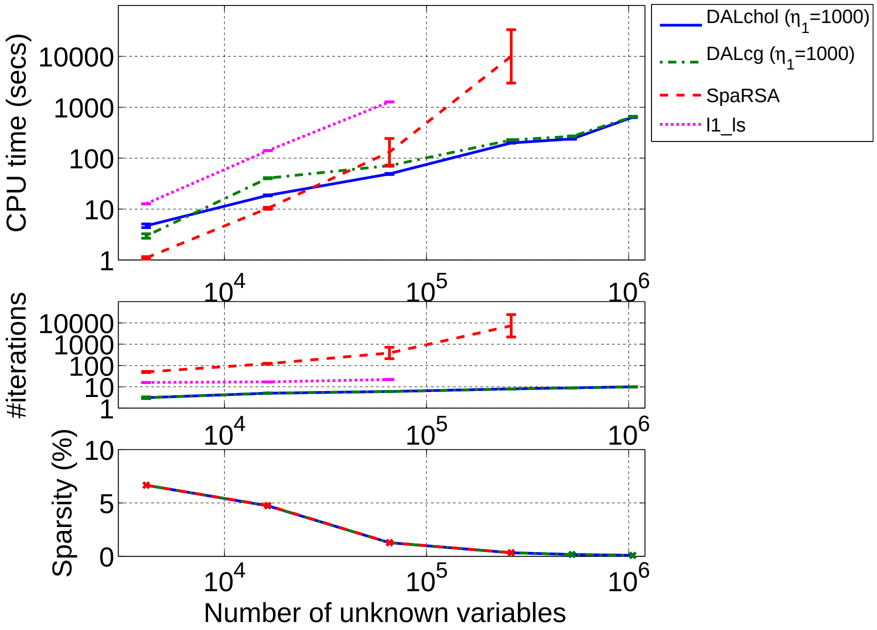


Fig. 3. Comparison of the algorithms for large scale problems when the number of variable (n) is much larger than the number of observations (m). m is kept constant at $m = 1024$. λ is decreased as $\lambda = 1.6/n^{1/2}$. Note that in the third row all methods found solutions with the same sparsity except l1_ls, whose sparsity is always 100% (no sparsity).

ACKNOWLEDGMENT

The authors would like to thank Masakazu Kojima, David Wipf, Srikantan Nagarajan, and Hagai Attias for helpful discussions.

REFERENCES

[1] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Stat. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[2] S. Chen, D. Donoho, and M. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[3] I. F. Gorodnitsky and B. D. Rao, “Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm,” *IEEE Trans. Signal Process.*, vol. 45, no. 3, 1997.

[4] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, “Sparse Solutions to Linear Inverse Problems With Multiple Measurement Vectors,” *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, 2005.

[5] J. Bioucas-Dias, “Bayesian wavelet-based image deconvolution: A GEM algorithm exploiting a class of heavy-tailed priors,” *IEEE Trans. Image Process.*, vol. 15, pp. 937–951, 2006.

[6] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, “Variational em algorithms for non-gaussian latent variable models,” in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 1059–1066.

[7] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-Minimization Algorithm for Wavelet-Based Image Restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, 2007.

[8] T. Goldstein and S. Osher, "Split Bregman Method for L1 Regularized Problems," UCLA Department of Mathematics, Tech. Rep. 08-29, 2008.

[9] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman Iterative Algorithms for L1-Minimization with Applications to Compressed Sensing," *SIAM J. Imaging Sciences*, vol. 1, no. 1, pp. 143–168, 2008.

[10] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.

[11] J. Nocedal and S. Wright, *Numerical Optimization*. Springer, 1999.

[12] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinvesky, "An Interior-Point Method for Large-Scale l-Regularized Least Squares," *IEEE journal of selected topics in signal processing*, vol. 1, pp. 606–617, 2007.

[13] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, pp. 906–916, 2003.

[14] I. Daubechies, M. Defrise, and C. D. Mol, "An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint," *Communications on Pure and Applied Mathematics*, vol. LVII, pp. 1413–1457, 2004.

[15] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[16] Y. Nesterov, "Gradient methods for minimizing composite objective function," Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Tech. Rep. 2007/76, 2007.

[17] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse Reconstruction by Separable Approximation," *IEEE Trans. Signal Process.*, 2009.

[18] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999, 2nd edition.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.