An Efficient Augmented Lagrangian Based Method for Constrained Lasso

Zengde Deng 1  Anthony Man-Cho So 1

Abstract

Variable selection is one of the most important tasks in statistics and machine learning. To incorporate more prior information about the regression coefficients, the constrained Lasso model has been proposed in the literature. In this paper, we present an inexact augmented Lagrangian method to solve the Lasso problem with linear equality constraints. By fully exploiting second-order sparsity of the problem, we are able to greatly reduce the computational cost and obtain highly efficient implementations. Furthermore, numerical results on both synthetic data and real data show that our algorithm is superior to existing first-order methods in terms of both running time and solution accuracy.

1. Introduction

With the advent of big data era, variable selection has received great attention in statistics and machine learning since contemporary applications often involve a large number of variables. There exist a host of methods to address this problem, such as Lasso (Tibshirani, 1996), SCAD (Fan & Li, 2001), elastic net (Zou & Hastie, 2005), adaptive Lasso (Zou, 2006), relaxed Lasso (Meinshausen, 2007) and so on. Benefiting from the simple formulation and the powerful modeling concerning the variable selection task, Lasso has been extensively applied in various instances (Burnham & Anderson, 2003; Candès & Wakin, 2008; Chen et al., 2001).

In spite of the overwhelming success of Lasso, it still suffers from the limited information induced by $l_1$ norm. To circumvent these issues, researchers have proposed the constrained Lasso model (Gaines et al., 2018; James et al., 2013) to incorporate more prior information. In the view of above discussions, our goal in this paper is to propose an efficient algorithm to tackle the following constrained Lasso problem

$$\min_{x} \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1,$$

s.t. $Bx = d,$

(1)

where $b \in \mathbb{R}^m$ is the response vector, $A \in \mathbb{R}^{m \times n}$ is the design matrix of covariates, $x \in \mathbb{R}^n$ is the regression estimator, and $B \in \mathbb{R}^{s \times n}, d \in \mathbb{R}^s$ are given constraints.

An important example which falls into the constrained Lasso problem is Lasso with sum-to-zero constraints, i.e. $e^T x = 0$. This type of constraint has been adopted in microbiome data regression (Shi et al., 2016) and variable selection (Lin et al., 2014) where the covariates come from compositional data.

Another example which is widely used in statistics is the generalized Lasso problem

$$\min_{x} \frac{1}{2} \|Ax - b\|^2 + \|Dx\|_1,$$

(2)

where $D \in \mathbb{R}^{p \times n}$. When rank$(D) = p$ and $p \leq n$, Tibshirani (2011) has derived that (2) can be transformed into a Lasso problem. In fact, (2) is a special case of constrained Lasso with $d = 0$ (Gaines et al., 2018; James et al., 2013) when $p \geq n$ and $D$ has full column rank $n$ and we elaborate on this in section 6.

Our contributions In this paper, we propose a semismooth Newton augmented Lagrangian method to solve the constrained Lasso problem. To fully exploit the sparsity structure, we mainly focus on the dual formulation of our problem and an inexact augmented Lagrangian method is proposed. The main challenge lies in how to solve the subproblem of augmented Lagrangian method efficiently. To overcome this difficulty, we apply the semismooth Newton method to resolve the inner subproblem. As we will show in numerical experiments, we only need about tens or dozens of outer iterations and about ten iterations to solve each subproblem to the desired accuracy and the total running time is small. The key insights behind this impressive performance are three-fold: (a) Regarding the outer loop, we have superlinear convergence to achieve highly accurate solution; (b) Besides, we also attain superlinear convergence in the inner subproblem solver and hence the total iteration number of our algorithm is still small; (c) For each iteration...

1Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong. Correspondence to: Zengde Deng <zddeng@se.cuhk.edu.hk>.
in solving the inner subproblem, we extensively exploit the second-order sparsity of the problem to greatly reduce the computational cost. In summary, not only can we prove the theoretical effectiveness of our algorithm, but also provide highly efficient implementations by exploring the hidden structure of our problem.

2. Related work

The constrained Lasso problem can be seen as a special case of the composite optimization problem with linear equality constraints. Without linear constraints, the proximal gradient method can achieve $O(1/k)$ convergence rate and accelerated proximal gradient (Beck & Teboulle, 2009; Nesterov, 2013) obtain $O(1/k^2)$ rate which is optimal for first-order methods to solve composite convex optimization problems. Although the methods proposed in Beck et al. (2009), Nesterov (2013) can also handle constrained problems, they require projections onto the constraint set, which can be computational prohibitive. Fortunately, there are other choices to tackle structured linear constrained problems. One of the most popular methods is the augmented Lagrangian method (ALM) (Bertsekas, 2014). Due to the difficulty of solving the subproblem, linearization technique is utilized in Yang et al. (Yang & Yuan, 2013), which results in linearized ALM (LALM). The recent work Xu (2017) proposes an accelerated linearized ALM to solve linearly constrained composite convex problem. Additionally, alternating direction method of multipliers (ADMM) (Boyd et al., 2011) is applied to solve the linear constrained problem and Goldstein et al. (2014) propose a faster version of ADMM (A-ADMM). Recently, first-order primal-dual methods (Chambolle & Pock, 2011) are also utilized to handle this kind of problems. Unfortunately, the acceleration rates of above algorithms all need the objective function to be strongly convex which is not satisfied by our problem.

More recently, semismooth Newton augmented Lagrangian method is attracting more and more attention due to its fast convergence and good experimental performance. This kind of method has been used to tackle problems such as SDP (Zhao et al., 2010), quadratic SDP (Li et al., 2015), Lasso (Li et al., 2018b) and fused Lasso (Li et al., 2018a), convex clustering (Yuan et al., 2018) and so on. Inspired by this, we propose a new ALM based framework to cope with constrained Lasso problem in this paper.

3. Problem formulation and an augmented Lagrangian method

In this section, we propose an augmented Lagrangian method to solve problem (1). Before that, we introduce some properties of our problem.

3.1. Dual problem and optimality conditions

In this subsection we first consider the general case of our problem (P) and derive the dual problem (D). Moreover, we give the optimality conditions associated with (D) which will play an important role in the convergence analysis in section 5.

Our constrained Lasso problem (2) can be seen as a special case of following problem:

$$\min_{x} \{ f(x) = h(Ax) + p(x) \}$$

subject to $Bx = d,$

where $A \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n, B \in \mathbb{R}^{s \times n}, d \in \mathbb{R}^s, h : \mathcal{U} \rightarrow (-\infty, +\infty]$ is continuous differentiable on dom($h$) and $p : \mathcal{X} \rightarrow (-\infty, +\infty]$ is a closed proper convex function. The dual problem of (P) can be written as

$$\min_{u,v,w} \{ g(z) := h^*(u) - \langle v, d \rangle + p^*(w) \}$$

subject to $A^Tu - B^Tv + w = 0,$

where $h^*$ and $p^*$ are conjugate functions of $h$ and $g$ respectively and we denote $z := (u, v, w) \in \mathcal{Z} := \mathcal{U} \times \mathcal{V} \times \mathcal{W}$ with $u \in \mathbb{R}^m, v \in \mathbb{R}^s$ and $w \in \mathbb{R}^n$. In our constrained Lasso model, we have $h(x) = \frac{1}{2} \| x - \bar{b} \|^2$ and $p(x) = \lambda \| x \|_1$. Therefore $h^*(u) = \frac{1}{2} \| u \|^2 + b^*u$ and $p^*(w) = \mathbb{I}_{\| w \|_\infty \leq \lambda}$ where $\mathbb{I}$ is the indicator function.

Denote the Lagrangian function associated with (D) by $l$ for $\forall (z, x) \in \mathcal{Z} \times \mathcal{X}$, i.e.,

$$l(z; x) \equiv l(u, v; w; x) := h^*(u) - \langle v, d \rangle + p^*(w) - \langle x, A^Tu - B^Tv + w \rangle.$$  \hspace{1cm} (3)

For the convex-concave function $l$, we further define

$$\psi(z) := \sup_{x \in \mathcal{X}} l(z; x), \quad \forall z \in \mathcal{Z},$$

$$\phi(x) := \inf_{z \in \mathcal{Z}} l(z; x), \quad \forall x \in \mathcal{X}.$$  \hspace{1cm} (4)

Moreover, we define the following mappings corresponding to $l, \psi$ and $\phi$:

$$T_l(z, x) := \{ (\hat{z}, \hat{x}) \in \mathcal{Z} \times \mathcal{X} | (\hat{z}, -\hat{x}) \in \partial l(z, x) \},$$

$$T_\psi(z) := \partial \psi(z), \quad \forall z \in \mathcal{Z},$$

$$T_\phi(x) := \partial \phi(x), \quad \forall x \in \mathcal{X}.$$  \hspace{1cm} (5)

The KKT systems associated with (D) is given by

$$\begin{cases}
0 \in \partial h^*(u) - Ax,
0 \in \partial p^*(w) - x,
0 = Bx - d,
0 = A^Tu - B^Tv + w,
\end{cases} \quad (u, v, w; x) \in \mathcal{Z} \times \mathcal{X}. \hspace{1cm} (5)$$

Suppose that the KKT system (5) has at least one solution. Let $\tilde{z} := (\tilde{u}, \tilde{v}, \tilde{w})$ be an optimal solution to (D). Denote $\tilde{x} \in \mathcal{X}$
\( \mathcal{M}_\varphi(\bar{z}) \), where \( \mathcal{M}_\varphi(\bar{z}) \) is the set of Lagrangian multipliers associated with \( \bar{z} \), then \( (\bar{z}, \bar{x}) \) is a solution of KKT system (5). By the property of conjugate function, we see that \((\bar{x}, \bar{v})\) solves the following KKT system:

\[
\begin{align*}
0 & \in A^T \nabla h(Ax) - B^T v + \partial p(x), \\
0 & = Bx - d,
\end{align*}
\]

(6)

Conversely, if we set \( \bar{u} = \nabla h(Ax) \) and \( \bar{w} = -A^T \bar{u} + B^T \bar{v} \), then \((\bar{u}, \bar{v}, \bar{w}, \bar{x})\) solves the KKT system (5). From Rockafellar (2015), we know that \((\bar{z}, \bar{x})\) solves KKT system (5) if and only if \( \bar{x} \) solves (P) and \( \bar{z} \) solves (D).

### 3.2. An augmented Lagrangian method

In this section, we introduce the inexact augmented Lagrangian method to solve (D). The augmented Lagrangian function associated with (D) is given by

\[
\mathcal{L}_\sigma(u, v, w; x) = l(u, v, w; x) + \frac{\sigma}{2} \| A^T u - B^T v + w \|^2.
\]

We propose the following inexact augmented Lagrangian method for (D).

**Algorithm 1** Inexact ALM for (D)

1. **Input:** \( u^0, v^0, w^0, x^0 \).
2. **for** \( k = 0, 1, \ldots \) **do**
3. **Get an approximate solution**
   \[
   z^{k+1} = (u^{k+1}, v^{k+1}, w^{k+1}) = \text{argmin}_{z=(u,v,w)} \left\{ \Theta_k(z) := \mathcal{L}_\sigma(u, v, w; x^k) \right\}
   \]
   \quad \approx \text{argmin}_{z=(u,v,w)} \left\{ \Theta_k(z) := \mathcal{L}_\sigma(u, v, w; x^k) \right\}
\]
4. **Update** \( x \) by \( x^{k+1} = x^k - \sigma_k (A^T u^{k+1} - B^T v^{k+1} + w^{k+1}) \) and update \( \sigma_{k+1} \uparrow \sigma_\infty \leq \infty \).
5. **end for**

To obtain an efficient implementation of Algorithm 1, we need to solve the subproblem (7) inexactly. From Rockafellar’s work (1976), we use one of the following stopping criteria:

\[
\begin{align*}
\Theta_k(z^{k+1}) - \inf \Theta_k & \leq \frac{\varepsilon_k^2}{2\sigma_k}, \\
\Theta_k(z^{k+1}) - \inf \Theta_k & \leq \zeta_k \| x^{k+1} - x^k \|^2 / 2\sigma_k,
\end{align*}
\]

(8)

where \( \sum_{k=0}^{\infty} \varepsilon_k < \infty \) and \( \sum_{k=0}^{\infty} \zeta_k < \infty \).

### 4. Inexact semismooth Newton method to solve ALM subproblem

In section 3.2, we design an inexact ALM to solve (D), but the main challenge lies in how to solve the subproblem (7) efficiently. In this section we propose a semismooth Newton method to solve ALM subproblem (7) and provide highly efficient implementations by exploiting second-order sparsity in the subproblem.

Here, we define the proximal mapping \( \text{Prox}_{p} (\cdot) \) associated with \( p \) as

\[
\text{Prox}_{p}(x) := \text{argmin}_{u} \left\{ p(u) + \frac{1}{2}\| u - x \|^2 \right\}.
\]

Moreover, by the Moreau decomposition, we have \( x = \text{Prox}_{p}(x) + \text{Prox}_{-1,p}(x/t) \) for \( t > 0 \).

Note that the augmented Lagrangian can be expressed as

\[
\mathcal{L}_\sigma(u, v, w; x) = h^*(u) - \langle v, d \rangle + p^*(w) - \langle x, A^T u - B^T v + w \rangle + \sigma\| A^T u - B^T v + w \|^2.
\]

For a fixed \( x \) and given \( \sigma \), let \( y = (u, v) \in \mathbb{R}^{m+n} \). we consider

\[
\min_{y,w} \Theta(y, w) := \mathcal{L}_\sigma(y, w; x),
\]

(9)

For convenience we set \( \hat{h}^*(y) = h^*(u) - \langle v, d \rangle \). Then we denote \( \theta(y) \) by

\[
\theta(y) = \inf_w \mathcal{L}_\sigma(y, w; x) = \hat{h}^*(y) + p^*(\text{Prox}_{\sigma^{-1}}(x/\sigma - A^T y)) + \frac{1}{2\sigma} \| \text{Prox}_{\sigma^{-1}}(x - \sigma(A^T y)) \|^2 - \frac{1}{2\sigma} \| x \|^2,
\]

where \( \hat{A} = \begin{bmatrix} A \\ -B \end{bmatrix} \), and we can also write \( \theta(y) \) as \( \theta(u, v) \).

Hence, if we let \( (\hat{y}, \hat{w}) = \text{argmin} \Theta(y, w) \), then \( (\hat{y}, \hat{w}) \) can be computed in the following manner:

\[
\begin{align*}
\hat{y} & = \text{argmin} \Theta(y), \\
\hat{w} & = \text{Prox}_{\sigma^{-1}}(x/\sigma - A^T \hat{y}).
\end{align*}
\]

(10)

Note that \( \theta(\cdot) \) is a continuous differentiable function with

\[
\nabla \theta(y) := \begin{bmatrix} \nabla^2 h^*(u) + \sigma \hat{A}^T \text{Prox}_{\sigma^{-1}}(x - \sigma(A^T y)) \hat{A}^T \\ -d + B \text{Prox}_{\sigma^{-1}}(x - \sigma(A^T y)) \hat{A}^T \end{bmatrix}.
\]

Moreover, (10) is equivalent to the following:

\[
\nabla \theta(y) = 0, \quad \forall y \in \text{dom}(y).
\]

For any \( y \in \text{dom}(y) \), we define

\[
\hat{\partial} \theta(y) := \begin{bmatrix} \nabla^2 h^*(u) + \sigma \hat{A}^T \text{Prox}_{\sigma^{-1}}(x - \sigma(A^T y)) \hat{A}^T \\ -d + B \text{Prox}_{\sigma^{-1}}(x - \sigma(A^T y)) \hat{A}^T \end{bmatrix},
\]

(11)

where \( \partial \text{Prox}_{\sigma^{-1}}(x - \sigma(A^T y)) \) is the Clarke subdifferential (Clarke, 1990) of \( \text{Prox}_{\sigma^{-1}}(\cdot) \) at \( x - \sigma(A^T y) \). From Hiriart-Urruty et al., (1984), we know that

\[
\hat{\partial} \theta(y)(d_u, d_v) = \partial^2 \theta(y)(d_u, d_v), \quad \forall (d_u, d_v) \in \text{dom}(y),
\]
where $\partial^2 \theta(y)$ denotes the generalized Hessian of $\theta(\cdot)$ at $y$. Define

$$V := H + \sigma \bar{A}Q\bar{A}^T$$  \hspace{1cm} (12)$$

with $Q \in \partial\text{Prox}_{\sigma}(x - \sigma(\bar{A}^Ty))$, then we have $V \in \partial^2 \theta(y)$. Since both $H$ and $\bar{A}Q\bar{A}^T$ are positive semidefinite, $V$ is positive semidefinite. Now, let us introduce the notion of semismoothness (Mifflin, 1977; Sun & Sun, 2002).

**Definition 1.** (Semismoothness) Suppose that $F : \mathcal{X} \to \mathcal{Y}$ is a locally Lipschitz continuous function. $F$ is said to be semismooth at $x \in \mathcal{X}$ if $F$ is directionally differentiable at $x$ and for any $V \in \partial F(x + \Delta x)$ and $\Delta x \to 0$,

$$F(x + \Delta x) - F(x) - V(\Delta x) = o(\|\Delta x\|).$$

$F$ is said to be strongly semismooth at $x \in \mathcal{X}$ if

$$F(x + \Delta x) - F(x) - V(\Delta x) = O(\|\Delta x\|^2).$$

$F$ is said to be a semismooth (strongly semismooth) function on $\mathcal{X}$ if it is semismooth (strongly semismooth) everywhere in $\mathcal{X}$.

Note that all twice continuous differentiable functions and piecewise linear functions are strongly semismooth everywhere (Li et al., 2018b). In our constrained Lasso problem, $h^*(\cdot)$ is twice continuous differentiable and $\text{Prox}_{\|\cdot\|_1}(\cdot)$ is piecewise linear. Hence, they are all strongly semismooth.

Now, we are ready to give an efficient inexact semismooth Newton (SSN) method to solve equation (9) in Algorithm 2.

### 4.1. Efficient implementation of SSN

As mentioned before, the key step to determine the efficiency of the whole algorithmic framework is how to solve (7) quickly. We use the semismooth Newton method to tackle (7) and the main computational cost lies in (13), which computes the inexact Newton direction. Thus we will give efficient implementations to compute (13).

Recall the definition of $V$ in (12), we write (13) as

$$(H_e + \sigma \bar{A}Q\bar{A}^T)d_y = -\nabla \theta(y)$$  \hspace{1cm} (14)$$

with $H_e$ is defined by $H_e := \begin{bmatrix} I_m & eI_n \end{bmatrix}$, where $I_m$ denote the $m \times m$ identity matrix. Since $H_e$ is positive definite, our linear system is well defined.

Before solving (14), we do Cholesky decomposition on $H_e$ via $H_e = LL^T$ with $L$ being a lower triangular matrix. Consequently, we rewrite (14) as

$$(I_{m+s} + \sigma(L^{-1}\bar{A})Q(L^{-1}\bar{A})^T)(Ld_y) = -L^{-1}\nabla \theta(y).$$  \hspace{1cm} (15)$$

Fortunately, for our constrained Lasso problem we have

$h^*(u) = \frac{1}{2}\|u\|^2 + b^T u$. Then $\nabla h^*(u) = u + b$, $\nabla^2 h^*(u) = I_m + \sigma \bar{A}Q\bar{A}^T$.  

### Algorithm 2 Semismooth Newton (SSN) for subproblem

1: **Input:** Given $\mu \in (0, 1/2)$, $\eta \in (0, 1)$, $\tau \in (0, 1], \gamma \in (0, 1)$, and $\delta \in (0, 1)$. Choose $y^0 = (u^0, v^0)$.

2: **for** $j = 0, 1, \ldots$ **do**

3: \hspace{1cm} Choose $Q^j \in \partial\text{Prox}_{\tau}(x - \sigma(\bar{A}^Ty^j))$. Let $V_j$ be given as in (12) and $\epsilon_j = \tau_j \min\{\tau_2, \|\nabla \theta(y^j)\|\}$.

4: \hspace{1cm} Solve the following linear system

$$V_j(d_u, d_v) + \epsilon_j(0, d_v) = -\nabla \theta(y^j)$$  \hspace{1cm} (13)$$

exactly or by CG algorithm to find an approximate solution such that

$$\|V_j(d_u, d_v) + \epsilon_j(0, d_v) + \nabla \theta(y^j)\| \leq \min(\bar{\eta}, \|\nabla \theta(y^j)\|^{1+\gamma}).$$

5: \hspace{1cm} Set $u^{j+1} = u^j + \alpha_j d_u$ and $v^{j+1} = v^j + \alpha_j d_v$.

6: **end for**

$I_m$ and it is easy to check that $L$ can be directly computed as $L = L^T = \begin{bmatrix} I_m & \sqrt{\bar{I}_n} \end{bmatrix}$ and $L^{-1} = \begin{bmatrix} I_m & \sqrt{1/\bar{I}_n} \end{bmatrix}$. Thus, the cost of computing Cholesky decomposition of $H_e$ is negligible. For notation convenience, we can simplify (15) as

$$(I_{m+s} + \sigma \bar{A}Q\bar{A}^T)d_y = -\nabla \theta(y),$$  \hspace{1cm} (16)$$

where $\bar{A} = L^{-1}\bar{A} = [A; -\sqrt{1/\bar{I}_n}], \bar{d}_y = L^{-1}d_y$ and $\nabla \theta(y) = L^{-1}\nabla \theta(y)$. Note that the cost of computing both $\bar{A}Q\bar{A}$ and $\bar{A}Q\bar{A}^T$ are $O((m + s)^2 n)$. Therefore, the matrix multiplication can be computational prohibitive when $n$ is large. Fortunately, we can overcome this difficulty by exploiting the sparsity structure of our problem in the following manners.

For the subdifferential of proximal mapping, from Li et al., (2018b) we can always choose $Q \in \partial\text{Prox}_{\tau}(x) \in (0, 1)$ to be $Q = \text{diag}(q)$, a diagonal matrix whose $i$-th element is given by

$$q_i = \begin{cases} 1, & \text{if } |x_i| > \sigma \lambda, \\ 0, & \text{otherwise}. \end{cases}$$  \hspace{1cm} (17)$$

Set $J = \{j : |x_j| > \sigma \lambda\}$ with cardinality $|J| = r$. By utilizing the diagonal structure of $Q$, we can write

$$\bar{A}Q\bar{A}^T = \bar{A}_J\bar{A}_J^T, \hspace{1cm} \bar{A}Q\bar{A}^T = \bar{A}_J\bar{A}_J^T,$$  \hspace{1cm} (18)$$
where $\tilde{A}_J \in \mathbb{R}^{(m+s) \times r}$ is the submatrix of $\tilde{A}$ with those columns contained in $J$ preserved and the same for $\tilde{A}_J$.

Now we analyze the reduction of computational cost by exploring the second-order sparsity of the problem. By utilizing (17), we can reduce the cost of computing $\tilde{A}Q\tilde{A}^T$ and $\tilde{A}Q\tilde{A}^T$ from $O((m+s)^2n)$ to $O((m+s)^2r)$. Due to the sparsity-inducing property of $p(x) = \lambda\|x\|_1$, $r$ is usually much smaller than $n$, greatly reducing the computational cost. Consequently, the total computational cost of solving (14) reduces from $O((m+s)^2(m+s+n))$ to $O((m+s)^2(m+s+r))$, meaning that the computational cost has no relationship with $n$. Thus, even for large dimension $n$, we can tackle the linear system (14) by Cholesky factorization.

As a result, the total computational cost to solve the Newton linear system can be reduced from $O((m+s)^2(m+s+r))$ to $O(r^2(m+s+r))$. Although this drastic computational reduction seems exciting, in practice we have to determine when to solve the linear system by Cholesky factorization and when to compute the inverse directly in the aforementioned way. We balance these two choices by judging whether $r \leq \frac{1}{2}(m+s)$ holds. Note that whichever way we choose to solve (14), the computational cost only depends on $m+s$. Thus, when $m+s$ is not too large (smaller than 10$^4$), we can always solve (14) exactly by Cholesky factorization or by computing the inverse. Otherwise, we can choose CG to solve the Newton linear system inexactly.

5. Convergence analysis

In this section, we establish the convergence of both the augmented Lagrangian method and the semismooth method under some mild assumptions.

5.1. Convergence of the augmented Lagrangian method

We first introduce some definitions that are key to our analysis.

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite-dimensional Euclidean spaces. We say that $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is a multi-valued mapping if $F$ maps each element $x \in \mathcal{X}$ to a subset $F(x) \subseteq \mathcal{Y}$. Then, the graph of $F$ is defined as

$$\text{gph}(F) := \{(x, y) \in \mathcal{X} \times \mathcal{Y} | y \in F(x)\},$$

and the inverse mapping of $F$ is denoted by $F^{-1}$. The following definition is from Chapter 3 in (Dontchev & Rockafellar, 2009).

**Definition 2.** A multi-valued mapping $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is said to be metrically subregular at $\bar{x} \in \mathcal{X}$ for $\bar{y} \in \mathcal{Y}$ with modulus $\kappa \geq 0$, where $(\bar{x}, \bar{y}) \in \text{gph}(F)$, if there exist neighborhoods $\mathcal{E}_1$ of $\bar{x}$ and $\mathcal{E}_2$ of $\bar{y}$ such that

$$\text{dist}(x, F^{-1}(\bar{y})) \leq \kappa\text{dist}(\bar{y}, F(x) \cap \mathcal{E}_2), \quad \forall x \in \mathcal{E}_1.$$

The following result shows that for the constrained Lasso problem, the map $T_\phi$ is metrically subregular at $\bar{x}$ for $0$. The proof can be found in the supplementary material.

**Theorem 1.** Assume that $T_\phi^{-1}(0)$ is non-empty and there exists $(\bar{u}, \bar{v}, \bar{w}) \in T_\psi^{-1}(0)$. For $h(\cdot)$ and $p(\cdot)$ chosen as in the constrained Lasso problem, the map $T_\phi$ is metrically subregular at $\bar{x}$ for $0$.

Recall the stopping criteria (A) and (B). Following Rockafellar (1976), we now establish the global convergence and local linear convergence of our inexact ALM algorithm under the metric subregularity of $T_\phi$.

**Theorem 2.** Assume that the solution set $T_\phi^{-1}(0)$ to (P) is non-empty. Let $\{z^k, x^k\}$ be the sequence generated by Algorithm 1 with stopping criterion (A). Then the sequence $\{x^k\}$ is bounded and converges to some $x^\infty \in T_\phi^{-1}(0)$, and the sequence $\{z^k\}$ satisfies the following for all $k \geq 0, z^k \in \mathbb{Z}$:

$$\|A^Tw^{k+1} - B^Tv^{k+1} + w^{k+1}\| = \|x^{k+1} - x^k\|/\sigma_k \to 0,$$

$$g(z^{k+1}) - \inf g \leq \Theta_k(z^{k+1}) - \inf \Theta_k + (\|x^{k+1}\| - \|x^{k+1}\|)/2\sigma_k,$$

where $g(z)$ is the dual objective value defined in (D) and $\Theta_k$ is defined in (7). Moreover, if (D) has a non-empty and bounded solution set, then the sequence $\{z^k\}$ is bounded and converges to an optimal solution to (D).

**Theorem 3.** Suppose that $T_\phi$ is metrically subregular at $x^\infty$ for 0 with modulus $\kappa_\phi$. Suppose further that the sequence $\{z^k, x^k\}$ is generated under the criterion (B). Then, for all sufficiently large $k$, we have

$$\text{dist}(x^{k+1}, T_\phi^{-1}(0)) \leq \rho_k\text{dist}(x^k, T_\phi^{-1}(0),$$

$$\|A^Tw^{k+1} - B^Tv^{k+1} + w^{k+1}\| \leq \zeta_k\text{dist}(x^k, T_\phi^{-1}(0),$$

$$g(z^{k+1}) - \inf g \leq \zeta'_k\text{dist}(x^k, T_\phi^{-1}(0),$$

where

$$1 > \rho_k = (\kappa_\phi/\sqrt{\kappa_\phi^2 + \sigma_k^2} + 2\zeta_k)(1 - \zeta_k^{-1})$$

$$\to \rho_\infty = \kappa_\phi/\sqrt{\kappa_\phi^2 + \sigma_\infty^2} = \sigma_\infty = 0 \text{ if } \sigma_\infty = \infty, \quad \zeta_k = \sigma_k^2(1 - \zeta_k)^{-1} \to \zeta_\infty = 1/\sigma_\infty = \sigma_\infty = \infty, \quad \zeta'_k = \sigma_k(\zeta_k^2\|x^{k+1} - x^k\| + \|x^{k+1}\| + \|x^k\|)/2$$

$$\to \zeta'_\infty = \|x^\infty\|/\sigma_\infty = \sigma_\infty = \infty, \quad \zeta_\infty = \|x^\infty\|/\sigma_\infty = \sigma_\infty = \infty.$$
and $\zeta_k$ comes from stopping criterion (B) and $g(z)$ is the dual objective value defined in (D).

**Remark 1.** Under the assumptions in Theorem 2, we establish the global convergence of the sequence $\{(x^k, z^k)\}$ corresponding to the primal and dual problems. Theorem 6 shows that $\mathcal{T}_0$ is metrically subregular for the constrained Lasso problem. Thus, it satisfies the assumptions in Theorem 3. Moreover, Theorem 3 shows that the KKT residuals corresponding to (P) and (D) converge superlinearly.

5.2. Convergence of the semismooth Newton method

We first establish the convergence of Algorithm 2 under the following mild assumption:

**Assumption 1.** The matrix $B$ has full row rank: $\text{rank}(B) = s$.

Using Assumption 1 and following the proof of Theorem 3.4 of (Zhao et al., 2010), we obtain the following result.

**Theorem 4.** Suppose the Assumption 1 holds. Then Algorithm 2 is well defined and any accumulation point $\hat{(u, v)}$ is an optimal solution to problem (10).

Next, we introduce the notion of constraint nondegeneracy.

**Definition 3.** Let $z \in \mathbb{R}^n$ and $\hat{Q} = \text{diag}(\hat{q}) \in \partial \text{Prox}_{\sigma\lambda, ||x||_1}(z)$ with $\hat{q}$ defined in (17). We say that the constraint nondegeneracy condition holds at $z$ if

$$\text{lin}(B\hat{Q}) = \mathbb{R}^s$$

holds at $z$ where $\text{lin}(B\hat{Q})$ denotes the linear space of $B\hat{Q}$.

Now we can establish the superlinear convergence of Algorithm 2. The proof of the following result can be found in the supplementary material.

**Theorem 5.** Let $\{(\hat{u}, \hat{v})\}$ be an accumulation point of the sequence $\{(u^j, v^j)\}$ generated by Algorithm 2. Suppose that the constraint nondegeneracy condition holds at $\hat{z} = x - \sigma(A^T\hat{u} - B^T\hat{v})$. Then, the sequence $\{(u^j, v^j)\}$ converges to $\{(\hat{u}, \hat{v})\}$ and

$$\|(u^{j+1}, v^{j+1}) - (\hat{u}, \hat{v})\| = O(\|(u^j, v^j) - (\hat{u}, \hat{v})\|^{1+\eta}).$$

6. Numerical experiments

In this section, we evaluate the performance of our algorithm to solve the constrained Lasso problem on both synthetic and real datasets. We compare with four state-of-the-art methods: primal-dual method (Chambolle & Pock, 2011), linearized augmented Lagrangian method (Yang & Yuan, 2013), ADMM (Boyd et al., 2011), accelerated ADMM (A-ADMM) (Goldstein et al., 2014). For both the ADMM and A-ADMM method, we set the step size to be 1.618.

We set the penalty parameter $\lambda$ in the constrained Lasso problem as $\lambda = \lambda_i \|A^Tb\|_\infty$, where $0 < \lambda_i < 1$. In the numerical experiments, we measure the accuracy of solution $\{x, u, v, w\}$ generated by our algorithm by the following residual:

$$\eta_{\text{Klasso}} = \max\{\eta_P, \eta_D\},$$

where

$$\eta_P = \frac{\|Bx - d\|}{1 + \|d\|}, \quad \eta_D = \|A^Tu - B^Tv + w\|$$

are the primal and dual feasibility. Moreover, we compute the relative primal dual gap defined by

$$\eta_{\text{relgap}} = \frac{\text{obj}_P - \text{obj}_D}{1 + |\text{obj}_P| + |\text{obj}_D|},$$

where $\text{obj}_P = \frac{1}{2}||Ax - b||^2 + \lambda ||x||_1$ and $\text{obj}_D = -\frac{1}{2}||u||^2 + \langle b, u \rangle - \langle d, v \rangle$.

We stop our algorithm when $\eta_{\text{Klasso}} < \varepsilon$ for a given tolerance $\varepsilon$ and stop other compared algorithms when both the primal and dual residuals are smaller than $\varepsilon$. Before the comparison, we run our algorithm with high accuracy $\varepsilon = 10^{-10}$ and set this optimal value as the baseline. The optimality gap is measured by $\eta_{\text{gap}} = f(x) - f(x^*)$. For our numerical experiments, we set $\varepsilon = 10^{-6}$. All the algorithms will be stopped when they reach the maximum iteration number, which is set at 100 for our algorithm and at 10000 for other algorithms. The codes were written in MATLAB and run on a server with 10 cores with 20 threads Intel Xeon E7-4870 CPU at 2.4 GHz with 1000 GB memory.

Our numerical tests have the following scenarios on both synthetic and real datasets:

- sum to zero constraints,
- $B$ and $d$ are randomly generated,
- transform generalized Lasso problem to an equivalent constrained Lasso problem.

Due to limited space, we only present the first and last ones in this paper, and full details of the experiments for all three scenarios are deferred to the supplementary material.

6.1. Synthetic data

In this subsection we consider the sum to zero constraints, i.e., $e^Tx = 0$. Recently this
Table 1. Performance of our SSNAL method(a), primal dual method(b), linearized ALM(c), ADMM(d) and A-ADMM(e) with sum to zero constraints on synthetic data sets. ‘nnz’ denotes the number of nonzeros in the solution obtained by our algorithm.

| size m; n   | λₖ | nnz | running time (seconds) |
|------------|-----|-----|------------------------|
| 200;2000   | 10⁻² | 189 | 4.5  | 5.0  | 6.2  | 8.3  |
|            | 10⁻³ | 193 | 1.0  | 6.5  | 4.6  | 5.9  | 7.9  |
|            | 10⁻⁴ | 195 | 1.3  | 6.3  | 6.7  | 7.3  | 6.4  |
| 300;3000   | 10⁻² | 286 | 1.2  | 19.8 | 18.2 | 22.5 |
|            | 10⁻³ | 289 | 2.3  | 18.7 | 17.3 | 22.2 |
|            | 10⁻⁴ | 291 | 3.2  | 18.5 | 25.4 | 24.2 |
| 500;5000   | 10⁻² | 471 | 2.7  | 76.8 | 80.1 | 155.162 |
|            | 10⁻³ | 481 | 3.8  | 70.9 | 95.1 | 158.16 |
|            | 10⁻⁴ | 484 | 5.3  | 78.8 | 81.1 | 159. |
| 800;8000   | 10⁻² | 746 | 6.2  | 151.149 | 355.359 |
|            | 10⁻³ | 760 | 10   | 153 | 159 | 357 | 361 |
|            | 10⁻⁴ | 764 | 15   | 157.169 | 355.363 |
| 1000;10000 | 10⁻² | 935 | 9.0  | 225.249 | 551.553 |
|            | 10⁻³ | 958 | 18   | 252.263 | 546 | 556 |
|            | 10⁻⁴ | 964 | 23   | 236.231 | 550 | 556 |

Table 5 reports the performance of all methods we test on the generalized Lasso problem to an equivalent constrained Lasso problem using techniques from (Gaines et al., 2018).

Recall the generalized Lasso model in (2). If \( \text{rank} (D) = n \) and set \( p = n + s \), a singular value decomposition of \( D \) is given as follows:

\[
D = U \Sigma V^T = [U_1, U_2] \begin{bmatrix} \Sigma_1 & 0 \\ \Sigma_2 & \end{bmatrix} V_1^T,
\]

where \( U_1 \in \mathbb{R}^{p \times n} \), \( U_2 \in \mathbb{R}^{p \times s} \), \( \Sigma_1 \in \mathbb{R}^{n \times n} \), \( V_1 \in \mathbb{R}^{n \times n} \) and \( \theta \in \mathbb{R}^{s \times n} \). Then (2) is equivalent to the constrained Lasso problem:

\[
\min \begin{bmatrix} \frac{1}{2} \|AD^Tz - b\|^2 + \lambda \|z\|_1 \end{bmatrix} \quad \text{s.t.} \quad U_2^Tz = 0,
\]

where \( AD^T \) is the optimal solution of (21). In our experiments, we construct \( D = [D_1 \ D_2] \), where \( D_1 \) is an \( n \times n \) identity matrix and \( D_2 \) is an \( s \times n \) random matrix.

In Table 7 we summarize the numerical results on Problem (21). Observe that our algorithm is 5-10 times faster than other algorithms for all choices of \( \lambda_t \).

Table 2. Performance of our SSNAL method(a), primal dual method(b), linearized ALM(c), ADMM(d) and A-ADMM(e) with generalized Lasso problem on synthetic datasets.

| size m; n | \( \lambda_t \) | nnz | running time (seconds) |
|------------|-----------------|-----|------------------------|
| 200;2000   | 10⁻² | 221 | 3.7  | 8.7  | 18  | 25  | 27  |
| 10⁻³ | 220 | 1.9  | 8.5  | 17  | 23  | 25  |
| 10⁻⁴ | 219 | 2.2  | 10  | 17  | 24  |
| 300;3000   | 10⁻² | 311 | 4.6  | 14 | 28  | 43  | 43  |
| 10⁻³ | 317 | 1.3  | 10  | 17  | 23  | 25  |
| 10⁻⁴ | 318 | 2.6  | 13 | 23  | 41  | 44  |
| 500;5000   | 10⁻² | 495 | 12  | 74 | 111 | 176 | 179 |
| 10⁻³ | 505 | 10  | 77  | 113 | 180 | 179 |
| 10⁻⁴ | 505 | 7.5  | 85  | 111 | 177 | 179 |
| 800;8000   | 10⁻² | 779 | 27  | 152 | 200 | 383 | 387 |
| 10⁻³ | 789 | 24  | 158 | 216 | 383 | 386 |
| 10⁻⁴ | 795 | 17  | 152 | 210 | 390 | 392 |
| 1000;10000 | 10⁻² | 975 | 51  | 254 | 311 | 589 | 589 |
| 10⁻³ | 993 | 41  | 234 | 325 | 597 | 595 |
| 10⁻⁴ | 995 | 27  | 250 | 322 | 589 | 587 |

We plot the optimality gap with running time in Figure 2 for the case \( m = 800, n = 8000 \) and \( s = 30 \). The figures demonstrate that our algorithm is superior to other methods both in terms of solution accuracy and running time.

6.2. Real Data

In this section we test all algorithms on LIBSVM regression datasets (Chang & Lin, 2011). We preprocess the dataset...
abalone, bodyfat, housing, mpg and space.ga to expand the original features based on polynomial basis functions as stated in (Huang et al., 2010). For example, abalone7 means that we expand the feature of abalone by an order 7 polynomial basis function. As mentioned before, we run numerical experiments with the sum to zero constraints and generalized Lasso problem and set $\lambda_l = 10^{-3}$ and $10^{-4}$.

**Sum to zero constraints** We consider the sum to zero constraints and summarize our experimental results on UCI regression datasets in Table 8. Moreover, we display the details of running time with optimality gap on bodyfat5 dataset in Figure 3, which highlight that our algorithm attains a more accurate solution much faster than other algorithms.

Table 4. Performance of our SSNAL method(a), primal dual method(b), linearized ALM(c), ADMM(d) and A-ADMM(e) with generalized Lasso problem on UCI regression datasets.

| problem name | $\lambda_l$ | nnz | running time (seconds) |
|--------------|-------------|-----|------------------------|
|              | $m; n$     | a   | b   | c   | d   | e   |
| abalone7     | $10^{-3}$  | 23  | 21  | 105 | 115 | 250 | 247 |
| 4177:6435    | $10^{-3}$  | 63  | 30  | 104 | 118 | 165 | 169 |
| bodyfat5     | $10^{-3}$  | 39  | 3.6 | 302 | 296 | 534 | 544 |
| 252:11628    | $10^{-3}$  | 79  | 4.3 | 296 | 298 | 235 | 229 |
| housing5     | $10^{-3}$  | 113 | 5.0 | 188 | 185 | 371 | 365 |
| 506:8568     | $10^{-3}$  | 216 | 6.7 | 179 | 189 | 121 | 96  |
| mpg7         | $10^{-3}$  | 43  | 1.3 | 35  | 36  | 19  | 18  |
| 392:3432     | $10^{-3}$  | 132 | 2.2 | 35  | 40  | 9.1 | 7.5 |
| space.ga9    | $10^{-3}$  | 14  | 6.5 | 78  | 89  | 161 | 167 |
| 3107:5005    | $10^{-3}$  | 36  | 11  | 80  | 97  | 50  | 40  |

7. Conclusion

In this paper we propose a semismooth Newton augmented Lagrangian method to solve the dual problem of the constrained Lasso problem. Under some mild assumptions, we establish the superlinear convergence for both outer loop and inner subproblem solver. By exploiting the sparse structure of the problem, we provide efficient implementations to solve the subproblem and greatly reduce the computational cost. We present extensive numerical experiments to show the efficiency and robustness of our algorithm on both synthetic and real datasets. As a future work, we plan to extend our algorithmic framework to handle more general constraints. Moreover, stochastic and distributed versions of this algorithmic framework are also interesting.
References

Altenbuchinger, M., Rehberg, T., Zacharias, H., Stämmler, F., Dettmer, K., Weber, D., Hiergeist, A., Gessner, A., Holler, E., Oefner, P. J., et al. Reference point insensitive molecular data analysis. *Bioinformatics*, 33(2):219–226, 2016.

Aragón Artacho, F. J. and Geoffroy, M. H. Characterization of metric regularity of subdifferentials. *Journal of Convex Analysis*, 15(2):365–380, 2008.

Bauschke, H. H. and Borwein, J. M. On projection algorithms for solving convex feasibility problems. *SIAM review*, 38(3):367–426, 1996.

Bauschke, H. H., Borwein, J. M., and Li, W. Strong conical hull intersection property, bounded linear regularity, jamesons property (g), and error bounds in convex optimization. *Mathematical Programming*, 86(1):135–160, 1999.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Bertsekas, D. P. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Burnham, K. P. and Anderson, D. R. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.

Candès, E. J. and Wakin, M. B. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

Chambolle, A. and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Clarke, F. H. *Optimization and nonsmooth analysis*, volume 5. Siam, 1990.

Dontchev, A. L. and Rockafellar, R. T. Implicit functions and solution mappings. *Springer Monogr. Math.*, 2009.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Gaines, B. R., Kim, J., and Zhou, H. Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, (just-accepted), 2018.

Goldstein, T., O’Donoghue, B., Setzer, S., and Baraniuk, R. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

Golub, G. H. and Van Loan, C. F. *Matrix computations*, volume 3. JHU Press, 2012.

Hiriart-Urruty, J.-B., Strodiot, J.-J., and Nguyen, V. H. Generalized hessian matrix and second-order optimality conditions for problems with 1, 1 data. *Applied mathematics and optimization*, 11(1):43–56, 1984.

Huang, L., Jia, J., Yu, B., Chun, B.-G., Maniatis, P., and Naik, M. Predicting execution time of computer programs using sparse polynomial regression. In *Advances in neural information processing systems*, pp. 883–891, 2010.

James, G. M., Paulson, C., and Rusmevichientong, P. Penalized and constrained regression. *Unpublished Manuscript, available at http://www-bcf.usc.edu/~gareth/research/Research.html*, 2013.

Li, X., Sun, D., and Toh, K.-C. Qsdpnal: A two-phase proximal augmented lagrangian method for convex quadratic semidefinite programming. *arXiv preprint arXiv:1512.08872*, pp. 1–35, 2015.

Li, X., Sun, D., and Toh, K.-C. On efficiently solving the subproblems of a level-set method for fused lasso problems. *SIAM Journal on Optimization*, 28(2):1842–1866, 2018a.

Li, X., Sun, D., and Toh, K.-C. A highly efficient semismooth newton augmented lagrangian method for solving lasso problems. *SIAM Journal on Optimization*, 28(1):433–458, 2018b.

Lin, W., Shi, P., Feng, R., and Li, H. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014.

Luo, Z.-Q. and Tseng, P. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425, 1992.

Meinshausen, N. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.
Efficient Algorithm for Solving Constrained Lasso

Mifflin, R. Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, 15(6):959–972, 1977.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Rockafellar, R. T. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1(2):97–116, 1976.

Rockafellar, R. T. *Convex analysis*. Princeton university press, 2015.

Shi, P., Zhang, A., Li, H., et al. Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040, 2016.

Sun, D. and Sun, J. Semismooth matrix-valued functions. *Mathematics of Operations Research*, 27(1):150–169, 2002.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Tibshirani, R. J. *The solution path of the generalized lasso*. Stanford University, 2011.

Xu, Y. Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. *SIAM Journal on Optimization*, 27(3):1459–1484, 2017.

Yang, J. and Yuan, X. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation*, 82(281):301–329, 2013.

Yuan, Y., Sun, D., and Toh, K.-C. An efficient semismooth Newton based algorithm for convex clustering. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5718–5726, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Zhao, X.-Y., Sun, D., and Toh, K.-C. A newton-cg augmented lagrangian method for semidefinite programming. *SIAM Journal on Optimization*, 20(4):1737–1765, 2010.

Zhou, Z. and So, A. M.-C. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.
A. Proofs

A.1. Proofs for convergence of augmented Lagrangian method

Recall the definition of metric subregularity.

**Definition 4.** A multi-valued mapping $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is said to be metrically subregular at $\bar{x} \in \mathcal{X}$ for $\bar{y} \in \mathcal{Y}$ with modulus $\kappa \geq 0$ where $(\bar{x}, \bar{y}) \in \text{gph}(F)$, if there exist neighborhoods $E_1$ of $\bar{x}$ and $E_2$ of $\bar{y}$ such that

$$\text{dist}(x, F^{-1}(\bar{y})) \leq \kappa \text{dist}(\bar{y}, F(x) \cap E_2), \quad \forall x \in E_1.$$  

We give the following proposition from (Aragón Artacho & Geoffroy, 2008) to provide a relatively easy way to check the metrically subregularity of closed proper convex function.

**Proposition 1.** Denote $\mathcal{H}$ as the Hilbert space. Let $f : \mathcal{H} \to (-\infty, +\infty]$ be a proper lower semicontinuous convex function and $(\bar{x}, \bar{s}) \in \text{gph}(\partial f)$. Then $\partial f$ is metrically subregular at $\bar{x}$ for $\bar{s}$ if and only if the following holds for $\forall x \in \mathcal{E}$:

$$f(x) \geq f(\bar{x}) + \langle \bar{s}, x - \bar{x} \rangle + \kappa \text{dist}^2(x, (\partial f)^{-1}(\bar{s})), \quad (22)$$

where constant $\kappa > 0$ and $\mathcal{E}$ is a neighborhood of $\bar{x}$.

Then we give the definition of boundedly linearly regularity for a collection of $m$ closed convex sets $\{C_1, \ldots, C_m\}$ from (Bauschke & Borwein, 1996).

**Definition 5.** For $m$ closed convex sets $\{C_1, \ldots, C_m\}$ which belong to $\mathcal{X}$. Suppose the intersection $C := C_1 \cap C_2 \cap \cdots \cap C_m$ is non-empty. We say that collection $\{C_1, \ldots, C_m\}$ is boundedly linearly regular if for every bounded set $S \subseteq \mathcal{X}$, there exist a constant $\kappa > 0$ such that for $\forall x \in S$,

$$\text{dist}(x, C) \leq \kappa \max\{\text{dist}(x, C_1), \ldots, \text{dist}(x, C_m)\}.$$  

From corollary 3 in (Bauschke et al., 1999), we have the following sufficient condition to guarantee the boundedly linearly regularity.

**Proposition 2.** Let $C_1, \ldots, C_m$ be closed convex sets in $\mathcal{X}$. Suppose $C_{r+1}, \ldots, C_m$ are polyhedral for some $r \in \{0, \ldots, m-1\}$. We say that the collection $\{C_1, \ldots, C_m\}$ is boundedly linearly regular if

$$\bigcap_{i=1}^{r} r_i(C_i) \cap \bigcap_{i=r+1}^{m} C_i \neq \emptyset.$$  

Here we give following lemma to show the invariant property of $Ax$ over $\mathcal{T}_{\phi}^{-1}(0)$, the proof follows from (Luo & Tseng, 1992) and we omit here.

**Lemma 1.** $Ax$ is invariant over $x \in \mathcal{T}_{\phi}^{-1}(0)$, which means that if $x, x' \in \mathcal{T}_{\phi}^{-1}(0)$, then $Ax = Ax'$.

Combine above lemma, recall the primal problem (P) and KKT system (23) as follows,

$$\begin{cases} 0 \in \partial h^*(u) - Ax, \\
0 \in \partial p^*(w) - x, \\
0 = Bx - d, \\
0 = A^T u - B^T v + w, \\
(u, v, w; x) \in \mathcal{Z} \times \mathcal{X},
\end{cases} \quad (23)$$

it is easy to check that the following observation holds.

**Proposition 3.** Let $(\bar{u}, \bar{v}, \bar{w}, \bar{x})$ be a solution to KKT system (23), then we can represent the optimal solution set $\mathcal{T}_{\phi}^{-1}(0)$ of (P) as

$$\mathcal{T}_{\phi}^{-1}(0) = \{x \in \mathcal{X} | Ax = \bar{\xi}, 0 = Bx - d, 0 \in \bar{\eta} + \partial p(x) - B^T \bar{v}\} = D_1 \cap D_2 \cap D_3,$$

where $\bar{\xi} := A\bar{x}$, $\bar{\eta} = A^T \nabla h(\bar{\xi})$, $D_1 = \{x \in \mathcal{X} | Ax = \bar{\xi}\}$, $D_2 = \{x \in \mathcal{X} | Bx = d\}$ and $D_3 = \{x \in \mathcal{X} | 0 \in \bar{\eta} + \partial p(x) - B^T \bar{v}\}$. 

In the next step, for the constrained Lasso model, we have that \( h(x) = \frac{1}{2}\|x - b\|^2 \) and \( p(x) = \lambda \|x\|_1 \). Then we give key properties of \( h(\cdot) \) and \( p(\cdot) \) to help establish the metric subregularity of \( T_\delta \).

**Proposition 4.** The following properties hold:
(a) For any \( r \in \text{dom}(h) \), there exists a constant \( \kappa_1 > 0 \) and neighborhood \( E_3 \) of \( r \) such that for \( \forall r' \in E_3 \)
\[
    h(r') \geq h(r) + \langle \nabla h(r), r' - r \rangle + \kappa_1 \|r - r'\|^2,
\]
(b) \( \partial p \) is metric subregular with constant \( \kappa_2 > 0 \), i.e., for any \( (x, s) \in \text{gph}(\partial p) \), there exists a constant \( \kappa_2 > 0 \) and a neighborhood \( E_4 \) of \( x \) such that for \( \forall x' \in E_4 \)
\[
    p(x') \geq p(x) + \langle s, x' - x \rangle + \kappa_2 \text{dist}^2(x', (\partial p)^{-1}(s)).
\]

**Proof.**
(a) Since \( h(r) \) is 1-strongly convex, (24) holds with \( \kappa_1 = 1 \).
(b) \( \partial(||x||_1) \) is metric subregular by Proposition 11 in (Zhou & So, 2017) since \( l_1 \) norm of vector is a special case of nuclear norm \( \cdot ||_n \) of matrix.

For problem \( \text{(P)} \), assume there exists at least one optimal solution \( \bar{x} \in T_\phi^{-1}(0) \). We say that the second order growth condition for \( \text{(P)} \) holds at \( \bar{x} \) w.r.t set \( T_\phi^{-1}(0) \) if the following holds for \( \forall x \in E \cap \{x \in X|Bx = d\} \):
\[
    f(x) \geq f(\bar{x}) + \kappa \text{dist}^2(x, T_\phi^{-1}(0)),
\]
where \( \kappa > 0 \) and \( E \) is a neighborhood of \( \bar{x} \). From Proposition 1 we get that the metrically subregularity of \( T_\phi \) holds at \( \bar{x} \) if and only if the second order growth condition holds at \( \bar{x} \) w.r.t set \( T_\phi^{-1}(0) \). Hence, we can just show the second order growth conditions holds to state that \( T_\phi \) is metrically subregular. The following lemma is a key step to prove the second order growth condition.

**Lemma 2.** For \( \bar{x} \in T_\phi^{-1}(0) \), the following holds
\[
    \text{dist}(x, D_2) \leq \kappa \|d - Bx\|_2,
\]
where \( \kappa > 0 \) and \( E \) is a neighborhood of \( \bar{x} \).

**Proof.** Define \( C_1 = \{(x, r) \in X \times V|Bx - d = r\} \) and \( C_2 = \{(x, r) \in X \times V|r = 0\} \). From Proposition 2, we know that \( C_1 \) and \( C_2 \) are boundedly linearly regular since both \( C_1 \) and \( C_2 \) are polyhedral and the intersection \( D_2 = C_1 \cap C_2 \) has an element \( (\bar{x}, d - B\bar{x}) \) which is non-empty. Hence, we have that there exists a constant \( \kappa > 0 \) and a neighborhood of \( E \) of \( \bar{x} \) such that
\[
    \text{dist}(x, d - Bx), D_2) \leq \kappa(\text{dist}(x, d - Bx), C_1) + \text{dist}(x, d - Bx), C_2) = \kappa \|d - Bx\|_2
\]
where the equality holds since \( \text{dist}(x, d - Bx), C_1 = 0 \). Moreover, it’s easy to see that there exists \( (x', v') \in D_2 \) such that
\[
    \text{dist}(x, d - Bx), D_2) = \sqrt{\|x - x'\|^2 + \|d - Bx - v'\|^2} \geq \|x - x'\|_2 = \text{dist}(x, D_2).
\]
Hence we complete the proof.

Based on above results, now we claim that the metrically subregularity of \( T_\phi \) holds at \( \bar{x} \) for our problem in Theorem 6 where \( \bar{x} \) is the optimal solution to \( \text{(P)} \).

**Theorem 6.** Assume that \( T_\phi^{-1}(0) \) is non-empty and there exists \( (\bar{u}, \bar{v}, \bar{w}) \in T_\psi^{-1}(0) \). For \( h(\cdot) \) and \( p(\cdot) \) chosen as in constrained Lasso model. Then metrically subregularity of \( T_\delta \) holds at \( \bar{x} \).

**Proof.** For \( \text{(P)} \), we can rewrite it as
\[
    \min_x \delta_0(d - Bx) + h(Ax) + p(x),
\]
We prove the local superlinear convergence rate of our semismooth Newton method based on the following propositions.

\[ \delta_0(d - Bx) \geq \delta_0(d - B\bar{x}) + \langle \bar{v}, d - Bx - (d - B\bar{x}) + \kappa_3 \text{dist}^2(d - Bx, 0), \quad \forall x \in \mathcal{E}, \tag{27} \]

where \( \kappa_3 > 0 \) and \( \mathcal{E} \) is a neighborhood of \( \bar{x} \). Since there exists \((\bar{u}, \bar{v}, \bar{w}) \in \mathcal{T}_\phi^{-1}(0)\), we get that \( \{D_1, D_2, D_3\} \) are boundedly linearly regular and \( \mathcal{T}_\phi^{-1}(0) = D_1 \cap D_2 \cap D_3 \), by definition we get that

\[ \text{dist}^2(x, \mathcal{T}_\phi^{-1}(0)) = \text{dist}^2(x, D_1 \cap D_2 \cap D_3) \leq \kappa_4(\text{dist}^2(x, D_1) + \text{dist}^2(x, D_2) + \text{dist}^2(x, D_3)) \leq \kappa_5(\|Ax - \xi\|^2 + \|d - Bx\|^2 + \text{dist}^2(x, (\partial p)^{-1}(B^Tv - \eta))), \tag{28} \]

where \( \kappa_4, \kappa_5 > 0 \) and \( x \in \mathcal{E} \), the first term of last inequality comes from Hoffman’s error bound and second term comes from Lemma 2. Since we choose \( h(x) = \frac{1}{2}\|x - b\|^2 \) and \( p(x) = \lambda\|x\|_1 \), combine Proposition 4 and above results, we can get that for any \( x \in \mathcal{E} \cap \{x \in X | Bx = d\} \),

\[
\begin{align*}
    f(x) &= f(x) + \delta_0(d - Bx) \\
    &= h(Ax + p(x) + \delta_0(d - Bx) \\
    &\geq h(\bar{x}) + \langle \nabla h(\bar{x}), Ax - \bar{x} \rangle + \kappa_1\|Ax - \bar{x}\|^2 + p(\bar{x}) + \langle B^T\bar{v} - \bar{\eta}, x - \bar{x} \rangle \\
    &\quad + \kappa_2\text{dist}^2(x, (\partial p)^{-1}(B^T\bar{v} - \bar{\eta})) + \delta_0(d - B\bar{x}) + \langle \bar{v}, d - Bx - (d - B\bar{x}) + \kappa_3\|d - Bx\|^2 \\
    &= f(\bar{x}) + \langle A^T\nabla h(\bar{x}) - \bar{\eta}, x - \bar{x} \rangle + \langle \kappa_1\|Ax - \bar{x}\|^2 + \kappa_3\text{dist}^2(x, (\partial p)^{-1}(B^T\bar{v} - \bar{\eta})) + \kappa_3\|d - Bx\|^2 \\
    &\geq f(\bar{x}) + \min(\kappa_1, \kappa_2, \kappa_3)(\|Ax - \bar{x}\|^2 + \text{dist}^2(x, (\partial p)^{-1}(B^T\bar{v} - \bar{\eta})) + \|d - Bx\|^2) \\
    &\geq f(\bar{x}) + \frac{\min(\kappa_1, \kappa_2, \kappa_3)}{\kappa_5}\text{dist}^2(x, \mathcal{T}_\phi^{-1}(0)).
\end{align*}
\]

Therefore, we prove the second order growth condition holds for \( \mathcal{P} \) at \( \bar{x} \) w.r.t \( \mathcal{T}_\phi^{-1}(0) \). As a consequence, the metrically subregularity of \( \mathcal{T}_\phi \) holds at \( \bar{x} \).

\[ \Box \]

**A.2. Proofs for convergence of semismooth Newton method**

Recall that we use semismooth Newton method to solve the following Newton linear system:

\[
\nabla \theta(y) = 0, \quad \forall y \in \text{dom}(y), \tag{29}
\]

where \( y = (u, v) \).

We prove the local superlinear convergence rate of our semismooth Newton method based on the following propositions.

**Proposition 5.** Let \( Q \in \mathbb{S}^n \) be a positive semidefinite matrix and \( \sigma > 0 \). Then \( BQB^T \) is positive definite if and only if

\[
\left< \begin{bmatrix} u \\ v \end{bmatrix}, \begin{bmatrix} I_m \\ \theta \end{bmatrix} + \sigma \begin{bmatrix} A \\ -B \end{bmatrix} Q \begin{bmatrix} A^T \\ -B^T \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \right> > 0, \tag{30}
\]

for all \( (u, v) \in \text{dom}(y) \}\setminus\{(0, 0)\} \).

**Proof.** For \"⇐\" case, since (30) holds for all \((u, v) \in \text{dom}(y) \}\setminus\{(0, 0)\} \), we can always choose \((u, v) = (0, v') \) where arbitrary \( v' \in \mathbb{R}^n \setminus \{0\} \), hence \((v')^T(BQB^T)v' > 0 \) for all \( v' \in \mathbb{R}^n \setminus \{0\} \). Thus \( BQB^T \) is positive definite.

For \"⇒\" case, suppose \( BQB^T \) is positive definite, we proof it by contradictory. We assume there exists \((u', v') \neq (0, 0) \), such that for (30), we have

\[
(u')^T u' + \sigma \left< \begin{bmatrix} u' \\ v' \end{bmatrix}, \begin{bmatrix} A \\ -B \end{bmatrix} Q \begin{bmatrix} A^T \\ -B^T \end{bmatrix} \begin{bmatrix} u' \\ v' \end{bmatrix} \right> = 0. \tag{31}
\]
Since $Q$ is positive semidefinite, we know that \[
\begin{bmatrix} A & -B \\ -B & \end{bmatrix} Q \begin{bmatrix} A^T \\ -B^T \end{bmatrix}
\] is also positive semidefinite. So both first term and second term in (31) are nonnegative, we must have both term to be zero, which means $u' = 0$ and moreover $(v')^T (BQB^T) v' = 0$. Since $BQB^T$ is positive definite, we have $v' = 0$. Hence $(u', v') = (0, 0)$ which makes the contradiction. This completes the proof of proposition.

Recall that $Q \in \partial \text{Prox}_{\sigma \lambda \|x\|_1}(x)$ is $Q = \text{diag}(q)$, a diagonal matrix with $i$-th element chosen as

\[
q_i = \begin{cases} 1, & \text{if } |x_i| > \sigma \lambda, \\ 0, & \text{otherwise}, \end{cases}
\] (32)

and set $J = \{j : |x_j| > \sigma \lambda\}$ definition of $Q$, $J$ and $r$, we can also ensure the positive definiteness of Newton linear system (29) under the constraint nondegeneracy condition in the following proposition.

**Proposition 6.** Let $(\hat{u}, \hat{v})$ be the solution of Newton linear system (29). Let $\tilde{z} = x - \sigma (A^T \hat{u} - B^T \hat{v})$ and $\check{Q} = \text{diag}(\check{q}) \in \partial \text{Prox}_{\sigma \lambda \|x\|_1}(\tilde{z})$ with $\check{q}$ is defined as in (32). If the constraint nondegeneracy condition, i.e.,

\[
\text{lin}(B\check{Q}) = \mathbb{R}^s
\] (33)

holds at $\tilde{z}$ with $\text{lin}(B\check{Q})$ means the linear space of $B\check{Q}$. Then

\[
\begin{bmatrix} I_d \\ 0 \end{bmatrix} + \sigma \begin{bmatrix} A & -B \\ -B & \end{bmatrix} \check{Q} \begin{bmatrix} A^T \\ -B^T \end{bmatrix}
\]

is positive definite on $\text{dom}(y)$.

**Proof.** By the definition of $\check{Q}$ we know that it is a idempotent matrix which means that $\check{Q}^2 = \check{Q}$. The constraint nondegeneracy conditions implies that $BQB^T = (B\check{Q})(B\check{Q})^T$ is positive definite. Then by Proposition 5, we get the desired results.

Now we can establish the superlinear convergence of Algorithm 2.

**Theorem 7.** Let $\{\hat{u}, \hat{v}\}$ be the accumulation point of the sequence $\{(w^j, v^j)\}$ generated by Algorithm 2. Assume the constraint nondegeneracy condition holds at $\tilde{z} = x - \sigma (A^T \hat{u} - B^T \hat{v})$. Then the sequence $\{(w^j, v^j)\}$ converges to $\{(\hat{u}, \hat{v})\}$ and

\[
\|(w^{j+1}, v^{j+1}) - (\hat{u}, \hat{v})\| = O((\|(w^j, v^j) - (\hat{u}, \hat{v})\|)^{1+r}).
\]

**Proof.** Since constraint nondegenerate condition holds, by Proposition 6, $V^* \in \partial^2 \theta(y^*)$ is positive definite. Hence we can obtain the superlinear convergence result from Theorem 3.5 in (Zhao et al., 2010).

For a special scenario that sum to zero constraint which means $B = e^T$, we have the following corollary.

**Corollary 1.** For $B = e^T$, which is the sum to zero constraint. Assume the optimal solution to Newton linear system is non-zero. Let $\{\hat{u}, \hat{v}\}$ be the accumulation point of the sequence $\{(w, v)\}$ generated by Algorithm 2. Then the sequence $\{(w^j, v^j)\}$ converges to $\{(\hat{u}, \hat{v})\}$ with superlinear convergence.

**Proof.** Since $B = e^T$, which means $B$ is full row rank with rank$(B) = 1$. Moreover, when optimal solution to Newton linear system is non-zero, we get that $r \geq 1 = s$ and constraint nondegeneracy condition holds. Therefore we know our result holds from Theorem 7.
B. Numerical Experiments on Synthetic Data

For the iteration numbers in our tables, \((p)/q\) denote the outer loop \(p\) and total inner iterations \(q\). \((-)\) means that the algorithm achieves maximum iteration number 10000. Moreover, we denote our algorithm as SSNAL. Running time counts in seconds.

B.1. Sum to zero constraints

In this scenario, we set \(B = e^T\) and \(d = 0\).

Table 5. Performance of SSNAL, primal dual method, linearized ALM, ADMM and A-ADMM with sum to zero constraints on synthetic data sets. \(d\) is the sample size and \(n\) is the dimension of each sample. \(\lambda\) controls the penalty parameter in (P). ‘nnz’ denotes the number of nonzeros of the solution obtained by our algorithm. ‘opt’ is the optimal function value of (P). \(\tau_{\text{opt}}\) is the optimal gap. ‘a’ = our algorithm, ‘b’ = primal dual method, ‘c’ = linearized alm, ‘d’ = ADMM, ‘e’ = A-ADMM. Running time counts in seconds.

| size \(m; n\) | \(\lambda\) | nnz | opt | \(a\) | \(b\) | \(c\) | \(d\) | \(e\) | \(a\) | \(b\) | \(c\) | \(d\) | \(e\) |
|-------------|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 200;2000    | \(10^{-2}\) | 189 | 1.3689 + 4 | 3.2 - 8 | 6.1 + 0 | 2.0 + 1 | 2.0 - 2 | 1.5 - 2 | \(0.4(10)^{100}\) | 5.3(-) | 5.0(-) | 6.2(-) | 8.3(-) |
|             | \(10^{-3}\) | 193 | 1.3843 + 3 | 3.4 - 8 | 5.5 + 1 | 8.8 + 1 | 1.6 - 3 | 1.8 - 3 | \(1.0(16)^{180}\) | 6.5(-) | 4.6(-) | 5.9(-) | 7.9(-) |
|             | \(10^{-4}\) | 195 | 1.3859 + 2 | 1.9 - 9 | 3.5 + 1 | 4.5 + 2 | 1.4 - 4 | 1.6 - 4 | \(1.3(17)^{100}\) | 6.3(-) | 6.7(-) | 7.3(-) | 6.4(-) |
| 300;3000    | \(10^{-2}\) | 286 | 3.8649 + 4 | 6.6 - 8 | 1.5 + 1 | 3.7 + 1 | 4.7 - 1 | 4.7 - 1 | \(1.2(11)^{110}\) | 19(-)  | 18(-)  | 22(-)  | 25(-)  |
|             | \(10^{-3}\) | 289 | 3.9137 + 3 | 1.2 - 9 | 1.0 + 2 | 1.7 + 2 | 1.1 - 2 | 1.3 - 2 | \(2.3(13)^{150}\) | 18(-)  | 17(-)  | 23(-)  | 22(-)  |
|             | \(10^{-4}\) | 291 | 3.9186 + 2 | 9.8 - 9 | 8.5 + 1 | 1.1 + 2 | 1.5 - 3 | 1.5 - 3 | \(3.2(16)^{193}\) | 18(-)  | 25(-)  | 24(-)  | 24(-)  |
| 500;5000    | \(10^{-2}\) | 471 | 1.2391 + 5 | 2.6 - 7 | 2.1 + 1 | 7.5 + 1 | 7.0 + 0 | 7.0 + 0 | \(2.7(11)^{106}\) | 76(-)  | 80(-)  | 155(-) | 162(-) |
|             | \(10^{-3}\) | 481 | 1.2555 + 4 | 4.7 - 8 | 3.3 + 2 | 5.6 + 2 | 2.6 - 1 | 2.6 - 1 | \(3.8(13)^{153}\) | 70(-)  | 95(-)  | 158(-) | 158(-) |
|             | \(10^{-4}\) | 484 | 1.2572 + 3 | 4.5 - 9 | 2.7 + 2 | 3.5 + 2 | 4.5 - 3 | 4.9 - 3 | \(5.3(14)^{175}\) | 78(-)  | 81(-)  | 155(-) | 159(-) |
| 800;8000    | \(10^{-2}\) | 746 | 3.1162 + 5 | 1.5 - 5 | 8.4 + 1 | 3.0 + 2 | 7.4 + 1 | 7.4 + 1 | \(6.2(10)^{110}\) | 151(-) | 149(-) | 355(-) | 359(-) |
|             | \(10^{-3}\) | 760 | 3.1602 + 4 | 1.2 - 6 | 7.7 + 2 | 1.3 + 3 | 6.4 - 1 | 6.2 - 1 | \(10(13)^{175}\) | 153(-) | 159(-) | 357(-) | 361(-) |
|             | \(10^{-4}\) | 764 | 3.1647 + 3 | 5.4 - 9 | 6.4 + 2 | 8.4 + 2 | 3.0 - 2 | 3.4 - 2 | \(15(19)^{231}\) | 157(-) | 169(-) | 355(-) | 363(-) |
| 1000;10000  | \(10^{-2}\) | 935 | 5.4143 + 5 | 6.4 - 6 | 1.5 + 2 | 4.8 + 2 | 3.3 + 2 | 3.3 + 2 | \(9.0(10)^{101}\) | 225(-) | 249(-) | 551(-) | 553(-) |
|             | \(10^{-3}\) | 958 | 5.4901 + 4 | 3.3 - 7 | 1.2 + 3 | 2.2 + 3 | 1.5 + 0 | 1.5 + 0 | \(18(15)^{200}\) | 252(-) | 263(-) | 546(-) | 556(-) |
|             | \(10^{-4}\) | 964 | 5.4978 + 3 | 3.3 - 8 | 1.1 + 3 | 1.5 + 3 | 4.5 - 2 | 4.5 - 2 | \(23(17)^{232}\) | 236(-) | 231(-) | 550(-) | 556(-) |

Figure 5. Constrained Lasso with sum to zero constraints, \(\lambda = 10^{-2}, 10^{-3}, 10^{-4}\). Top row is \(m = 500, n = 5000\), bottom row is \(m = 1000, n = 10000\).
B.2. Randomized $B$

In this case, we generate $B \in \mathbb{R}^{s \times n}$ and $d \in \mathbb{R}^s$ randomly and set $s = 30$.

Table 6. Performance of SSNAL, primal dual method, linearized ALM, ADMM and A-ADMM with randomized generated constraints on synthetic data sets. $d$ is the sample size and $n$ is the dimension of each sample. $\lambda_l$ controls the penalty parameter in (P). ‘nnz’ denotes the number of nonzeros of the solution obtained by our algorithm. ‘opt’ is the optimal function value of (P). $\eta_{gap}$ is the optimal gap. ‘a’ = our algorithm, ‘b’ = primal dual method, ‘c’ = linearized alm, ‘d’ = ADMM, ‘e’ = A-ADMM. Running time counts in seconds.

| size $m; n$ | $\lambda_l$ | nnz | opt | optimal gap. $\eta_{gap}$ | running time (iteration number) |
|------------|-------------|-----|-----|-----------------------------|--------------------------------|
|            |             |     |     |                             | a | b | c | d | e | a | b | c | d | e |
| 200;2000   | $10^{-2}$   | 220 | 1.4046+4 | 2.1-8 | 3.6+0 | 1.9+2 | 5.8-2 | 5.3-2 | 1.4(12|114) | 7.9(-) | 15(-) | 23(-) | 25(-) |
|           | $10^{-3}$   | 222 | 1.4203+3 | 6.7-9 | 4.2+1 | 2.3+2 | 1.3-3 | 1.4-3 | 1.9(17|157) | 8.1(-) | 17(-) | 24(-) | 25(-) |
|           | $10^{-4}$   | 224 | 1.4218+2 | 9.9-9 | 3.4+1 | 6.8+1 | 3.0-4 | 3.0-4 | 2.8(19|292) | 8.7(-) | 15(-) | 16(-) | 18(-) |
| 300;3000   | $10^{-2}$   | 308 | 3.9596+4 | 1.8-6 | 1.1+1 | 3.9+2 | 4.0-1 | 4.0-1 | 1.4(11|114) | 7.9(-) | 12(-) | 26(-) | 43(-) |
|           | $10^{-3}$   | 320 | 4.0082+3 | 6.1-9 | 9.7+1 | 5.4+2 | 2.5-2 | 2.4-2 | 1.7(13|136) | 9.7(-) | 35(-) | 43(-) | 44(-) |
|           | $10^{-4}$   | 318 | 4.0131+2 | 1.1-8 | 8.2+1 | 1.8+2 | 1.1-3 | 1.0-3 | 2.5(15|181) | 14(-) | 31(-) | 43(-) | 44(-) |
| 500;5000   | $10^{-2}$   | 501 | 1.2547+5 | 1.37-6 | 2.9+1 | 1.1+3 | 1.1+3 | 1.0+3 | 3.2(12|116) | 7.6(-) | 113(-) | 180(-) | 180(-) |
|           | $10^{-3}$   | 512 | 1.2713+4 | 6.4-9 | 2.8+2 | 1.6+3 | 2.4-1 | 2.4-1 | 4.4(18|227) | 79(-) | 117(-) | 179(-) | 182(-) |
|           | $10^{-4}$   | 511 | 1.2729+3 | 2.0-8 | 2.5+2 | 5.6+2 | 4.9-3 | 4.6-3 | 7.4(18|227) | 79(-) | 117(-) | 180(-) | 182(-) |
| 800;8000   | $10^{-2}$   | 776 | 3.1414+5 | 1.6-4 | 8.8+1 | 2.9+3 | 7.1+1 | 7.1+1 | 6.5(10|110) | 172(-) | 205(-) | 380(-) | 385(-) |
|           | $10^{-3}$   | 780 | 3.1859+4 | 5.0-8 | 7.2+2 | 4.0+3 | 7.3-1 | 7.3-1 | 11(14|175) | 163(-) | 217(-) | 388(-) | 394(-) |
|           | $10^{-4}$   | 799 | 3.1903+3 | 8.0-8 | 6.3+2 | 1.4+3 | 3.2-2 | 3.1-2 | 16(18|251) | 152(-) | 227(-) | 385(-) | 387(-) |
| 1000;10000 | $10^{-2}$   | 974 | 5.4457+5 | 1.0-5 | 1.3+2 | 4.7+3 | 2.9+2 | 2.9+2 | 11(10|117) | 246(-) | 316(-) | 602(-) | 601(-) |
|           | $10^{-3}$   | 986 | 5.5229+4 | 1.2-6 | 1.2+3 | 6.8+3 | 1.3+0 | 1.2+0 | 17(15|185) | 252(-) | 321(-) | 610(-) | 606(-) |
|           | $10^{-4}$   | 990 | 5.5307+3 | 5.2-9 | 1.1+3 | 2.5+3 | 6.8-2 | 6.8-2 | 22(16|214) | 258(-) | 318(-) | 596(-) | 598(-) |

Figure 6. Constrained Lasso with randomized generated constraints, $\lambda_l = 10^{-2}, 10^{-3}, 10^{-4}$. Top row is $m = 200, n = 2000, s = 30$ and bottom row is $m = 500, n = 5000, s = 30$. 
B.3. Generalized Lasso

In this scenario, we construct \( D = \begin{bmatrix} D_1 & D_2 \end{bmatrix} \), where \( D_1 \) is an \( n \times n \) identity matrix and \( D_2 \) is an \( s \times n \) random matrix. Moreover, we set \( s = 30 \).

Table 7. Performance of SSNAL, primal dual method, linearized ALM, ADMM and A-ADMM with generalized Lasso problem on synthetic data sets. \( d \) is the sample size and \( n \) is the dimension of each sample. \( \lambda \) controls the penalty parameter in (P). 'nnz' denotes the number of nonzeros of the solution obtained by our algorithm. 'opt' is the optimal function value of (P). \( \eta_{\text{gap}} \) is the optimal gap. 'a' = our algorithm, 'b' = primal dual method, 'c' = linearized alm, 'd' = ADMM, 'e' = A-ADMM. Running time counts in seconds.

| size \( m; n \) | \( \lambda \) | nnz | opt | optimal gap, \( \eta_{\text{gap}} \) | running time (iteration number) |
|----------------|-------------|-----|-----|-------------------------------|--------------------------------|
| 200;2000       | \( 10^{-2} \) | 221 | 1.4371+4 | 1.3-6 | 5.8-2 | 6.6-2 | 3.7(13)367 | 8.7(-) 18(-) 25(-) 27(-) |
|                | \( 10^{-3} \) | 220 | 1.4530+3 | 4.3-8 | 4.3+1 | 1.5-3 | 1.9(14)245 | 8.5(-) 17(-) 23(-) 25(-) |
|                | \( 10^{-4} \) | 219 | 1.4546+2 | 2.2-8 | 3.1+1 | 3.1+1 | 2.2(17)207 | 10(-) 17(-) 24(-) 24(-) |
| 300;3000       | \( 10^{-2} \) | 311 | 3.8919+4 | 2.6-6 | 1.2+1 | 1.1+1 | 4.6(12)470 | 14(-) 28(-) 43(-) 43(-) |
|                | \( 10^{-3} \) | 317 | 3.9406+3 | 4.0-8 | 8.3+1 | 8.3+1 | 3.0(14)271 | 13(-) 30(-) 43(-) 44(-) |
|                | \( 10^{-4} \) | 318 | 3.9455+2 | 3.9-9 | 8.0+1 | 8.1+1 | 2.6(17)208 | 13(-) 32(-) 41(-) 44(-) |
| 500;5000       | \( 10^{-2} \) | 495 | 1.2179+5 | 6.8-6 | 5.1+1 | 4.7+1 | 12(10)515 | 74(-) 111(-) 176(-) 179(-) |
|                | \( 10^{-3} \) | 505 | 1.2339+4 | 6.4-8 | 3.2+2 | 3.2+2 | 10(14)397 | 77(-) 113(-) 180(-) 179(-) |
|                | \( 10^{-4} \) | 505 | 1.2355+3 | 4.2-8 | 2.5+2 | 2.5+2 | 7.5(15)249 | 85(-) 111(-) 177(-) 179(-) |
| 800;8000       | \( 10^{-2} \) | 779 | 3.1099+5 | 6.9-6 | 1.2+2 | 1.1+2 | 27(10)525 | 152(-) 200(-) 383(-) 387(-) |
|                | \( 10^{-3} \) | 789 | 3.1355+4 | 5.3-7 | 7.2+1 | 7.2+1 | 24(14)415 | 158(-) 216(-) 383(-) 386(-) |
|                | \( 10^{-4} \) | 795 | 3.1578+3 | 8.5-8 | 6.2+2 | 6.2+2 | 17(15)254 | 152(-) 210(-) 390(-) 392(-) |
| 1000;10000     | \( 10^{-2} \) | 975 | 5.2804+5 | 1.9-5 | 1.5+2 | 1.5+2 | 51(11)661 | 254(-) 311(-) 589(-) 589(-) |
|                | \( 10^{-3} \) | 993 | 5.3519+4 | 1.8-8 | 1.2+3 | 1.2+3 | 41(14)479 | 234(-) 325(-) 597(-) 595(-) |
|                | \( 10^{-4} \) | 995 | 5.3591+3 | 7.0-8 | 1.1+3 | 1.1+3 | 27(16)282 | 250(-) 322(-) 589(-) 587(-) |

Figure 7. Constrained Lasso with generalized lasso problem, \( \lambda = 10^{-2}, 10^{-3}, 10^{-4} \). Top row is \( m = 300, n = 3000, s = 30 \) and bottom row is \( m = 800, n = 8000, s = 30 \).
C. Numerical Results on Real Data

C.1. Sum to zero constraint

In this scenario, we set \( B = e^T \) and \( d = 0 \).

Table 8. Performance of SSNAL, primal dual method, linearized ALM, ADMM and A-ADMM with sum to zero constraints on UCI regression data sets. \( d \) is the sample size and \( n \) is the dimension of each sample. \( \lambda \) controls the penalty parameter in (P). 'nnz' denotes the number of nonzeros of the solution obtained by our algorithm. 'opt' is the optimal function value of (P). \( \eta_{gap} \) is the optimal gap. 'a' = our algorithm, 'b' = primal dual method, 'c' = linearized alm, 'd' = ADMM, 'e' = A-ADMM. Running time counts in seconds.

| problem name | \( \lambda \) | nnz | opt | optimal gap, \( \eta_{gap} \) | running time (iteration number) |
|--------------|------------|-----|-----|------------------|-------------------------------|
| \( 10^{-3} \) | \( 10^{-4} \) |
| abalone7     | 10^{-3}    | 23  | 1.1440+4 | 1.9-5 9.2+2 9.3+2 6.8+1 6.8+1 | 21(12|96) 105( ) 115( ) 250( ) 247( ) |
| 10^{-4}      | 63         | 9.2897+3 | 1.3-7 8.6+2 8.7+2 1.8-4 1.8-4 | 30(13|130) 104( ) 118( ) 165(6572) 169(6334) |
| bodyfat5     | 10^{-3}    | 39  | 5.3609-1 | 5.0-8 7.4-2 1.4-1 4.5-6 4.5-6 | 3.6(18|941) 302( ) 296( ) 534(6892) 544(6837) |
| 10^{-4}      | 79         | 5.5219-2 | 1.8-8 7.2-2 1.1-1 2.0-6 2.0-6 | 4.3(19|110) 296( ) 296( ) 235(2389) 229(2160) |
| housing5     | 10^{-3}    | 113 | 2.8392+3 | 6.1-9 1.4+2 1.4+2 5.8-4 5.8-4 | 5.8(12|106) 188( ) 185( ) 371(8558) 365(8368) |
| 10^{-4}      | 216        | 1.0340+3 | 7.2-9 2.6+2 2.7+2 7.8-5 7.8-5 | 6.7(15|139) 179( ) 189( ) 121(2179) 96(1554) |
| mpg7         | 10^{-3}    | 43  | 1.6769+3 | 1.1-7 1.2+1 1.9+1 2.1-4 2.1-4 | 1.3(10|781) 35( ) 36( ) 19(3755) 18(3484) |
| 10^{-4}      | 132        | 8.9060+2 | 8.3-8 4.5+1 5.2+1 4.2-5 4.1-5 | 2.2(14|120) 35( ) 40( ) 9.1(1583) 7.5(1014) |
| space_ga9    | 10^{-3}    | 14  | 3.1931+1 | 2.7-11 1.1-2 1.3-2 1.9-2 1.9-2 | 6.5(8|48) 78( ) 89( ) 161( ) 167( ) |
| 10^{-4}      | 36         | 1.9924+1 | 1.2-8 4.9-1 1.8+0 6.4-6 6.4-6 | 11(11|82) 80( ) 97( ) 50(2597) 40(1949) |

Figure 8. Constrained Lasso with sum to zero constraints, \( \lambda = 10^{-3}, 10^{-4} \). Top two figures are abalone7 dataset; bottom two figures are bodyfat5 dataset.
C.2. Randomized $B$

In this case, we generate $B \in \mathbb{R}^{s \times n}$ and $d \in \mathbb{R}^s$ randomly and set $s = 30$.

Table 9. Performance of SSNAL, primal dual method, linearized ALM, ADMM and A-ADMM with randomized generated constraints on UCI regression data sets. $d$ is the sample size and $n$ is the dimension of each sample. $\lambda_l$ controls the penalty parameter in (P). 'nnz' denotes the number of nonzeros of the solution obtained by our algorithm. 'opt' is the optimal function value of (P). $\eta_{\text{gap}}$ is the optimal gap. 'a' = our algorithm, 'b' = primal dual method, 'c' = linearized alm, 'd' = ADMM, 'e' = A-ADMM. Running time counts in seconds.

| problem name | $\lambda_l$ | nnz | opt | a | b | c | d | e | running time (iteration number) |
|--------------|-------------|-----|-----|---|---|---|---|---|--------------------------------|
| abalone7     | 10$^{-3}$   | 39  | 1.2054+4 | 3.2-7 | 8.46+2 | 9.06+2 | 1.8+1 | 1.8+1 | 19(8) | 112(1) | 163(1) | 273(1) | 276(1) |
|              | 10$^{-4}$   | 70  | 9.3489+3 | 1.8-7 | 8.5+2  | 9.0+2  | 2.6-4 | 2.6-4 | 24(11) | 172(1) | 149(5298) | 142(4957) |
| bodyfat5     | 10$^{-3}$   | 49  | 9.9783-1 | 4.2-7 | 9.4-2  | 5.6-1  | 2.3-5 | 2.3-5 | 3.7(12) | 362(1) | 394(1) | 263(3146) | 237(2789) |
|              | 10$^{-4}$   | 64  | 1.0369-1 | 5.2-9 | 7.4-2  | 3.3-1  | 2.3-6 | 2.2-6 | 4.5(13) | 325(1) | 395(1) | 123(1294) | 98(928) |
| housing5     | 10$^{-3}$   | 104 | 2.8589+3 | 1.7-8 | 1.4+2  | 1.9+2  | 6.5-4 | 6.5-4 | 5.1(12) | 200(1) | 239(1) | 423(9295) | 420(9096) |
|              | 10$^{-4}$   | 218 | 1.0403+3 | 2.0-7 | 2.6+2  | 3.1+2  | 8.0-5 | 8.0-5 | 6.6(17) | 195(1) | 217(4670) | 214(4473) |
| mpg7         | 10$^{-3}$   | 66  | 1.8380+3 | 9.2-7 | 1.4+1  | 6.3+1  | 4.1-4 | 4.1-4 | 2.3(11) | 38(1)   | 62(1)   | 183(3007) | 14(1800) |
|              | 10$^{-4}$   | 121 | 8.9682+2 | 2.8-9 | 4.5+1  | 9.4+1  | 4.2-5 | 4.1-5 | 2.7(14) | 43(1)   | 59(1)   | 13(1745) | 8.3(1075) |
| space_ga9    | 10$^{-3}$   | 49  | 3.6325+1 | 3.6-8 | 2.9-3  | 1.8-1  | 1.1-4 | 1.1-4 | 8.1(10) | 85(1)   | 123(1)  | 165(8885) | 166(8760) |
|              | 10$^{-4}$   | 74  | 2.1154+1 | 5.6-8 | 4.4-1  | 4.8+0  | 9.7-6 | 8.6-6 | 15(13) | 100(1)  | 120(1)  | 24(1074) | 21(885) |

Figure 9. Constrained Lasso with randomized generated constraints, $\lambda_l = 10^{-3}, 10^{-4}$. Top two figures are housing5 dataset; bottom two figures are mpg7 dataset.
C.3. Generalized Lasso

In this scenario, we construct $D = \begin{bmatrix} D_1 & D_2 \end{bmatrix}$, where $D_1$ is an $n \times n$ identity matrix and $D_2$ is an $s \times n$ random matrix. Moreover, we set $s = 30$.

Table 10. Performance of SSNAL, primal dual method, linearized ALM, ADMM and A-ADMM with generalized Lasso problem on UCI regression data sets. $d$ is the sample size and $n$ is the dimension of each sample. $\lambda_l$ controls the penalty parameter in (P). 'nnz' denotes the number of nonzeros of the solution obtained by our algorithm. 'opt' is the optimal function value of (P). $\eta_{gap}$ is the optimal gap. 'a' = our algorithm, 'b' = primal dual method, 'c' = linearized alm, 'd' = ADMM, 'e' = A-ADMM. Running time counts in seconds.

| problem name | $\lambda_l$ | nnz | opt | optimal gap, $\eta_{gap}$ | running time (iteration number) |
|---------------|--------------|-----|-----|---------------------------|---------------------------------|
| abalone7      | $10^{-3}$    | 46  | 1.2013+4 | 2.4+1 | 2.4+1 | 51(22/2759) | 127(-) | 165(-) | 267(-) | 270(-) |
|               | $10^{-4}$    | 78  | 9.355+3   | 2.5+4 | 2.5+4 | 56(12/2888) | 114(-) | 176(-) | 98(3386) | 84(2879) |
| bodyfat5      | $10^{-3}$    | 44  | 8.2521-1  | 4.5-5 | 4.1-5 | 3.7(10/911) | 321(-) | 395(-) | 772(-) | 783(-) |
| 252;11628     | $10^{-4}$    | 68  | 8.6951-2  | 2.3-6 | 1.7-6 | 4.3(13/1050) | 309(-) | 393(3386) | 121(1244) | 93(907) |
| housing5      | $10^{-3}$    | 105 | 2.8593+3  | 6.0-4 | 7.1-4 | 11(12/254) | 198(-) | 246(-) | 458(-) | 463(9971) |
| 506;8568      | $10^{-4}$    | 219 | 1.0453+3  | 8.0-5 | 8.5-5 | 7.7(15/178) | 257(-) | 153(3165) | 146(2923) |
| mpg7          | $10^{-3}$    | 61  | 1.8728+3  | 2.8-10 | 2.5+2 | 1.1(12/254) | 194(-) | 257(-) | 458(-) | 463(9971) |
| 392;3432      | $10^{-4}$    | 127 | 8.9665+2  | 3.8-10 | 2.5+2 | 7.7(15/178) | 257(-) | 153(3165) | 146(2923) |
| space_ga9     | $10^{-4}$    | 47  | 3.5451+1  | 4.0-4 | 4.0-4 | 5.6(10/300) | 46(-)  | 47(-)  | 58(9378) | 59(9198) |
| 3107;5005     | $10^{-4}$    | 78  | 2.1266+1  | 4.0-4 | 4.0-4 | 5.6(10/300) | 46(-)  | 47(-)  | 58(9378) | 59(9198) |

Figure 10. Constrained Lasso with randomized generated constraints, $\lambda_l = 10^{-3}, 10^{-4}$. Top two figures are **bodyfat5** dataset; bottom two figures are **mpg7** dataset.