Projection onto an ℓ_1-norm Ball with Application to Identification of Sparse Autoregressive Models

Jitkomut Songsiri
Department of Electrical Engineering
Faculty of Engineering, Chulalongkorn University
254 Phayathai Rd., Pathumwan, Bangkok, 10330 Thailand
Email: jitkomut.s@chula.ac.th

***Abstract*—We consider a problem of finding the Euclidean projection of a vector in $\mathbf{R}^n$ onto an ℓ_1-norm ball. This can be casted as a simple convex optimization problem. We present an efficient method for the projection, obtained via the dual problem which reduces to an optimization with a scalar variable. The method involves sorting elements of the vector and performing a linear interpolation. We demonstrate the effectiveness of the method and compare the results with solving the primal problem by an interior-point method. Our approach is useful for sparse system identification problems that can be represented as a minimization of a loss function subject to ℓ_1-norm constraints. Sparse autoregressive model estimation is included as an example of this problem type, where zeros in autoregressive coefficients indicates a causality structure of variables. Numerical examples with synthetic data sets are included. With the proposed method we are able to solve problems of dimensions in the order of several thousand efficiently.**

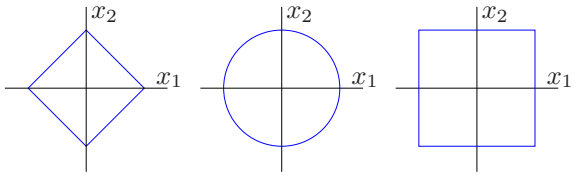
***Index Terms*—ℓ_1-norm, sparse estimation, autoregressive models, convex optimization**

I. Introduction

The problem of computing the Euclidean projection of a vector $a \in \mathbf{R}^n$ onto the unit ℓ_p-norm ball can be casted as

$$\begin{array}{ll} \text{minimize} & \|y-a\|_2^2 \\ \text{subject to} & \|y\|_p \le 1 \end{array} \tag{1}$$

where $p = 1, 2$ or ∞. Examples of the three unit norm balls in $\mathbf{R}^2$ can be shown in Figure 1.



(a) ℓ_1-ball (b) ℓ_2-ball (c) ℓ_∞-ball

Fig. 1. The unit norm balls

The problem (1) is a convex optimization which can be efficiently solved in polynomial time. However, a projection onto the ℓ_2-norm ball has a closed-form expression and therefore turns to be an easy task to perform. From figure 1, intuition suggests (and it can be shown) that the projection is simply given by the normalization of a.

$$y^* = \begin{cases} a, & \text{if } \|a\|_2 \le 1 \\ a/\|a\|_2, & \text{otherwise.} \end{cases}$$

Likewise, finding a projection onto the unit ℓ_∞-norm ball is an inexpensive task and the solution is given by

$$y_k^* = \begin{cases} a_k, & \text{if } |a_k| \le 1 \\ 1, & \text{otherwise} \end{cases}$$

for $k = 1, 2, \ldots, n$. This is a thresholding operator applied to the entries of a that the modulus exceed 1. The solution can be written more compactly in a vector format as

$$y^* = \min(1, \max(-1, a)).$$

The ℓ_∞-norm constraint in (1) is sometimes referred to as a *box constraint*.

Finding a projection onto an ℓ_1-norm ball is more involved. The main goal of this paper is to present an efficient procedure to perform this task as fast as computing the projection onto the unit ℓ_2- or ℓ_∞- norm ball. Section II will first present recent applications of the projection onto an ℓ_1-norm ball which include a sparse system identification problem. Section III, describes the main result of this paper. We derive the dual problem of (1) and show that it reduces to an optimization problem with a scalar variable. As a result, we develop an efficient procedure to compute the projection from the dual problem. Section IV applies the result from section III to consider a problem of finding a projection onto the sum of ℓ_2-norm ball. An example of this problem is the estimation of sparse autoregressive models, which will be demonstrated in section V.

II. Applications

There has been a substantial interest of parameter estimation problems with sparsity-promoting regularization. This type of problem can be expressed by

$$\text{minimize } f(x) \text{ subject to } \|x\|_1 \le \rho \tag{2}$$

where f is a loss function (possibly convex) and ρ is a given positive parameter. The optimization variable is $x \in \mathbf{R}^n$. It is a well-known property that the ℓ_1-norm constraint encourages sparsity in x for a sufficiently small ρ. This is an appealing property for model selection problems. For example, f could represent a likelihood function and entries in x are model parameters to be estimated. Having many zeros in x allows us to pick a model with less number of parameters. However,

at the same time, the number of parameters should be large enough so that the model can adequately explain the data. This concept is known as *principle of parsimony* which is a trade off between model complexity and goodness of fit [1].

The application of this paper is when solving the convex optimization problem (2) using the projected gradient method [2] or its variants. This method minimizes a convex function $f(x)$ subject to the constraint $x \in \mathcal{C}$ for $\mathcal{C}$ convex (here $\mathcal{C}$ is an ℓ_1 ball.) The method is based on the update

$$x^{(k+1)} = P_{\mathcal{C}}(x^{(k)} - t^{(k)} \nabla f(x^{(k)}))$$

where $t^{(k)}$ is a step size, ∇f is the gradient of f, and $P_{\mathcal{C}}$ is a Euclidean projection onto $\mathcal{C}$ and is defined by

$$P_{\mathcal{C}}(y) = \text{argmin}_x \|x - y\| \text{ subject to } x \in \mathcal{C}.$$

The projected gradient method is obviously suitable when a projection step can be done cheaply. The main focus of this paper is therefore to show that the projection on the ℓ_1-norm ball can be efficiently computed.

The rest of this section will show some specific examples of the problem (2).

Example I: LASSO

The first example is the problem of finding sparse approximate solutions to the least-square problem:

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 \\ \text{subject to} & \|x\|_1 \leq \rho \end{array}$$

where $A \in \mathbf{R}^{m \times n}, b \in \mathbf{R}^m, x \in \mathbf{R}^n$. By making ρ small enough, some coefficients of x become zero because of the nature of ℓ_1 norm, so this framework is used as a heuristic for regression selection to find a sparse solution (in opposed to a least-squares solution which is typically dense). This problem is frequently referred to as the *least absolute shrinkage and selection operator (LASSO)* [3] and very closely related to the ℓ_1-regularized least-squares formulation:

$$\text{minimize } \|Ax - b\|_2^2 + \gamma \|x\|_1.$$

This problem is now widely used in many applications; including signal processing [4], wavelet-based image reconstruction [5], and compressed sensing [6], [7], [8]. This formulation can be also interpreted as a maximum a posterior probability (MAP) estimation where the $\|\cdot\|_1$ term corresponds to the log prior of the Laplace distribution for each x_i.

Example II: Group Sparsity

Consider a problem with a constraint on the sum of ℓ_2-norm:

$$\text{minimize } f(x) \text{ subject to } \|x\|_{1,2} = \sum_{k=1}^{m} \|x_k\|_2 \leq \rho,$$

with variables $x_k \in \mathbf{R}^n$, for $k = 1, 2, \ldots, m$. In this problem, the sum of $\|\cdot\|_2$ play as a role of the ℓ_1 norm applied to the ℓ_2-norm of m-subset variables. For a sufficiently small ρ, the constraint on the group norm $\|x\|_{1,2}$ leads to sparsity in terms of groups. It is a popular technique for achieving sparsity of groups of variables and is known as *group lasso* [9], [10] in machine learning community.

In section IV, we show that, the projection onto the sum of $\|\cdot\|_2$ can be performed efficiently using a similar idea from the projection onto the ℓ_1-norm ball.

Example III: Sparse Autoregressive (AR) Models

This example is a direct application of the group sparsity problem in the previous example. Consider a multivariate autoregressive process of order p given by

$$y(t) = \sum_{k=1}^{p} A_k y(t-k) + \nu(t)$$

where $y(t) \in \mathbf{R}^n, A_k \in \mathbf{R}^{n \times n}$, $k = 1, 2, \ldots, p$ and $\nu(t)$ is noise. The least-squares method is a common approach used for fitting an AR model to the measurements $y(1), y(2), \ldots, y(N)$. The model parameters A_k's are chosen such that

$$\sum_{t=p+1}^{N} \|y(t) - \sum_{k=1}^{p} A_k y(t-k)\|^2 \tag{3}$$

is minimized. In many applications, it is of interest to obtain A_k's that satisfy

$$\text{for some } (i,j) \quad (A_k)_{ij} = 0, \quad \text{for all } k \tag{4}$$

(where $(A_k)_{ij}$ denotes the (i,j) entry of A_k.) This is the characterization of *Granger causality* [11] of autoregressive models. The condition (4) explains that y_i is not *Granger-caused* by y_j, or knowing y_j does not help to improve the prediction of y_i. This idea was originally established in economics by Granger but recently has been adopted in many applications in neuroscience and system biology; see examples in [12], [13], [14], [15], [16], [17], [18], just to name a few. In these references, a sparse AR model is fitted to time series data where a zero pattern in AR coefficients reveals causal relationships among the variables. The formulation of this problem can be explained as follows.

If we define $A = \begin{bmatrix} A_1 & A_2 & \cdots & A_p \end{bmatrix}$ and

$$Y = \begin{bmatrix} y(p+1) & y(p+2) & \cdots & y(N) \end{bmatrix}$$

$$H = \begin{bmatrix} y(p) & y(p+1) & \cdots & y(N-1) \\ y(p-1) & y(p) & \cdots & y(N-2) \\ \vdots & \vdots & & \vdots \\ y(1) & y(2) & \cdots & y(N-p) \end{bmatrix}$$

then the quadratic loss in (3) can be rewritten more compactly as $\|Y - AH\|_F^2$ where $\|\cdot\|_F$ denotes the Frobenious norm. To promote a sparse solution A_k, we introduce an additional constraint in the least-squares problem as

$$\begin{array}{ll} \text{minimize} & \|Y - AH\|_F^2 \\ \text{subject to} & \sum\limits_{i \neq j} \| \begin{bmatrix} (A_1)_{ij} & (A_2)_{ij} & \cdots & (A_p)_{ij} \end{bmatrix} \|_2 \leq \rho \end{array} \tag{5}$$

with variables $A_k \in \mathbf{R}^{n \times n}$ for $k = 1, 2, \ldots, p$. The summation over (i,j) with $i \neq j$ plays a role of ℓ_1-type norm, which causes some (i,j) entries of A_k to zero for a sufficiently small

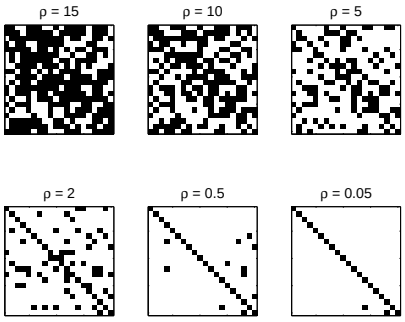


Fig. 2. Common zero patterns of a solution A_k, $k = 1, 2, \ldots, p$, of the problem (5) with $n = 20, p = 3$. As ρ decreases, A_k's contains more zeros.

value of ρ. Furthermore, using the ℓ_2 norm (or any norm) of p-tuple of $(A_k)_{ij}$ will force all p matrices A_k's to have the same sparsity pattern. Figure 2 shows an example of the zero patterns in solutions A_k to (5) as ρ varies.

The problem (5) is a convex optimization problem, which can be solved efficiently by a second-order algorithm such as an interior-point method [19, §11]. However, many applications involve a large number of variables (n can be in order of thousand) and hence solving by a second-order algorithm will suffer from an expensive computational cost. For this reason, there has been much attention to applying a gradient-based method to a high-dimensional problem. The result in this work will serve for this purpose. One can solve (5) using the projected gradient method. In each iteration, it is required to compute a projection onto the constraint set, which can be shown to be an inexpensive task in the following sections.

III. Projection onto the ℓ_1-Ball

The problem of finding the Euclidean projection of a vector $a \in \mathbf{R}^n$ onto the unit ℓ_1-norm ball can be described by the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|y - a\|_2^2 \\ \text{subject to} & \|y\|_1 \leq 1 \end{array} \tag{6}$$

with variable $y \in \mathbf{R}^n$. When $\|a\|_1 \leq 1$, obviously the optimal solution is $y^* = a$. Therefore, in what follows, we only consider the nontrivial case. We will show that an efficient algorithm to compute a projection can be developed from the dual problem of (6).

The derivation of the dual problem starts with the Lagrangian defined as the cost function plus a weighted sum of the constraints. We define λ as a weight of the ℓ_1-norm inequality constraint. The Lagrangian is therefore

$$\begin{aligned} L(y, \lambda) &= \|y - a\|_2^2 + 2\lambda(\|y\|_1 - 1) \\ &= \sum_{k=1}^{n} \left((y_k - a_k)^2 + 2\lambda|y_k|\right) - 2\lambda. \end{aligned} \tag{7}$$

To find the Lagrange dual function [19, §5] of problem (6), define

$$g_k(\lambda) = \inf_{y_k} \ (y_k - a_k)^2 + 2\lambda|y_k|.$$

The solution y_k^* that minimizes g_k is given by

$$y_k^* = \begin{cases} a_k + \lambda, & a_k \leq -\lambda \\ 0, & |a_k| < \lambda \\ a_k - \lambda, & a_k \geq \lambda, \end{cases} \tag{8}$$

and g_k can be expressed as

$$g_k(\lambda) = \begin{cases} -(\lambda - |a_k|)^2 + a_k^2, & \lambda < |a_k| \\ a_k^2, & \lambda \geq |a_k|. \end{cases}$$

The dual function is the minimum value of the Lagrangian over y. Hence, by the above notation, the dual function is

$$g(\lambda) = \inf_y \ L(y, \lambda) = \sum_k g_k(\lambda) - 2\lambda.$$

The Lagrange dual problem is to maximize $g(\lambda)$ over $\lambda \geq 0$. Hence, the dual problem of (6) is then

$$\begin{array}{ll} \text{maximize} & g(\lambda) := \sum_k g_k(\lambda) - 2\lambda \\ \text{subject to} & \lambda \geq 0, \end{array} \tag{9}$$

with variable $\lambda \in \mathbf{R}$. In the primal problem (6), if $\|a\|_1 > 1$, then there exists a strictly feasible point y such that $\|y\|_1 < 1$. Hence, Slater's condition holds, which implies strong duality [19, §5.2.3]. In other words, the optimal values of (6) and (9) are equal. Therefore, it is more efficient to solve the dual problem where the optimization variable is simply a *scalar*.

The dual objective in (9) is a piecewise quadratic function and by definition, a concave function. To find an optimal solution λ^*, we set the gradient of $g(\lambda)$ to zero (find the stationary point), and if the stationary point is positive, then it is the optimal solution. Otherwise, the optimal solution is 0. In other words,

$$\lambda^* = \max\{0, \nu\}$$

where ν is the root of $g'(\nu) = 0$.

To find the stationary point of $g(\lambda)$, note that

$$g_k'(\lambda) = \begin{cases} 2(|a_k| - \lambda), & \lambda < |a_k| \\ 0, & \lambda \geq |a_k|. \end{cases}$$

From (9), if $\|a\|_1 > 1$, then the dual optimal point λ^* is given by the root of

$$g'(\lambda) = \sum_{k=1}^{n} \max(|a_k| - \lambda, 0) - 1. \tag{10}$$

In what follows, we will describe how to find the root of (10) numerically. Without loss of generality, we can sort a_k in ascending order, *i.e.*,

$$|a_1| \leq |a_2| \leq \ldots \leq |a_n|.$$

Therefore, $g'(\lambda)$ is a piecewise linear function in λ where the slope changes at points $|a_1|, |a_2|, \ldots, |a_n|$. Initially at $\lambda = 0$, the slope of $g'(\lambda)$ is $-2n$ and increases by 2 when $\lambda = |a_1|$. The slope keeps increasing as λ reaches the points $|a_k|$ and eventually the slope is -2 when $\lambda \geq |a_n|$, as shown in Figure 3. This means we can simply make a plot of $g'(\lambda)$

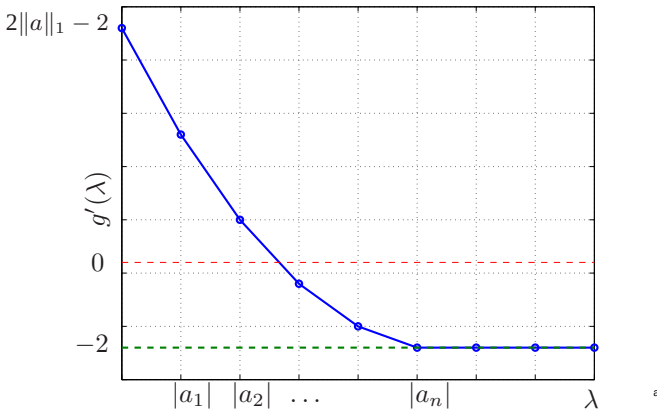


Fig. 3. The gradient of the dual objective, $g'(\lambda)$

and checks for which interval $g'(\lambda)$ changes its sign from a negative to a positive value. Once we can locate the interval, *i.e.*, says between $|a_i|$ and $|a_{i+1}|$, we can do an interpolation to find λ such that $g'(\lambda) = 0$, because for any interval $g'(\lambda)$ is a linear function in λ.

When $\lambda = 0$, $g'(0) = 2\|a\|_1 - 2$. Therefore, if $\|a\| < 1$, the plot of $g'(\lambda)$ is initially negative and never touches zero. This simply means λ^* must be zero since the original vector is already inside the unit ℓ_1-ball.

Using this scheme, we can develop a procedure to find λ^* efficiently as follows.

Algorithm I:

1) If $\|a\|_1 \le 1$, then $\lambda^* = 0$.
2) Otherwise, define $a_0 = 0$ and sort $|a_k|$ in ascending order. Compute the values of $g'(\lambda)$ at points $\lambda = |a_0|, |a_1|, |a_2|, \ldots, |a_n|$ as shown in the following table.

λ	$g'(\lambda)/2$
$\|a_0\| = 0$	$\|a\|_1 - 1$
$\|a_1\|$	$(1-n)\|a_1\| + \sum_{k=2}^{n} \|a_k\| - 1$
$\|a_2\|$	$(2-n)\|a_2\| + \sum_{k=3}^{n} \|a_k\| - 1$
$\vdots$	$\vdots$
$\|a_{n-1}\|$	$-\|a_{n-1}\| + \|a_n\| - 1$
$\|a_n\|$	-1

3) Locate the interval where $g'(\lambda)$ changes its sign, *i.e.*, Find k such that $g'(|a_k|) \ge 0$ and $g'(|a_{k+1}|) \le 0$, where k can take values from 0 to $n-1$.
4) In this interval, the graph of $g'(\lambda)/2$ is a linear function described by

$$g'(\lambda)/2 = -(n-k)\lambda + \sum_{j=k+1}^{n} |a_j| - 1, \quad |a_k| \le \lambda \le |a_{k+1}|.$$

Hence, the point where $g'(\lambda) = 0$ is

$$\lambda^* = \frac{\left(\sum_{j=k+1}^{n} |a_j|\right) - 1}{(n-k)}.$$

5) Using λ^* to compute the projection y^* from (8).

This scheme involves sorting n elements of a vector and thus requires $\mathcal{O}(n \log(n))$ time. We note that this approach was also developed independently by [20] and [21]. By replacing the procedure of sorting elements with finding the median to compute the partial sum more efficiently, the authors provided algorithms for computing the projection in linear time.

IV. Projection onto the sum of ℓ_2-norm ball

First we investigate a solution of the problem:

$$\text{minimize} \quad f(x) := \|x - a\|^2 + 2\lambda\|x\|$$

with variable $x \in \mathbf{R}^n$ and a positive scalar λ. Throughout this section, $\|\cdot\|$ denotes the ℓ_2 norm. The function $f(x)$ is a convex function but not differentiable at zero. The zero gradient condition is

$$x - a + \lambda s = 0 \tag{11}$$

where s is a subgradient of $\|x\|$ and given by

$$s = \begin{cases} \frac{x}{\|x\|}, & x \neq 0 \\ \text{any vector } s \text{ such that } \|s\| < 1, & x = 0. \end{cases}$$

The solution is $x = 0$ when $\|a\| < \lambda$. When $x \neq 0$, $s = x/\|x\|$ and (11) gives $x = a\|x\|/(\|x\| + \lambda)$, which further implies that $\|x\| = \|a\| - \lambda$. As a result, we can write the solution x^* as

$$x^* = \frac{a}{\|a\|} t \tag{12}$$

where $t = \|x^*\| = \max(\|a\| - \lambda, 0) := (\|a\| - \lambda)^+$. The operator $+$ denotes a projection onto the nonnegative orthant $\mathbf{R}_+$. The minimized cost objective becomes

$$\begin{aligned} f^* &= \|x^*\|^2 - 2\langle x^*, a\rangle + \|a\|^2 + 2\lambda\|x^*\| \\ &= t^2 - 2t(\|a\| - \lambda) + \|a\|^2 \\ &= (\|a\| - \lambda)^{+2} - 2(\|a\| - \lambda)^+(\|a\| - \lambda) + \|a\|^2 \\ &= -(\|a\| - \lambda)^{+2} + \|a\|^2 \end{aligned} \tag{13}$$

In what follows, we use the results (12) and 13) in the projection problem. Given m vectors in $\mathbf{R}^n$, a_k's, for $k = 1, \ldots, m$, the problem is to find the projection x_k of each a_k under a constraint on the sum of norms of x_k's. This problem can be casted as

$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^{m} \|x_k - a_k\|^2 \\ \text{subject to} & \sum_{k=1}^{m} \|x_k\| \le 1. \end{array} \tag{14}$$

Similar to the section III, we will find the optimal solution via the dual problem. The Lagrangian of (14) is

$$L(x_1, x_2, \ldots, x_m, \lambda) = \sum_{k=1}^{m} \left(\|x_k - a_k\|^2 + 2\lambda\|x_k\|\right) - 2\lambda$$

where λ is the Lagrange multiplier associated with the inequality constraint in (14). Using (12) and (13), each term in the summation in L can be independently minimized over x_k, when $x_k = a_k t_k / \|a_k\|$ with $t_k = (\|a_k\| - \lambda)^+$. Therefore, the

dual function of (14) which is the infimum of L over x_k's is given by

$$g(\lambda) := \inf_{x_k} L = -\sum_{k=1}^{m} \left((\|a_k\| - \lambda)^{+2} + \|a_k\|^2 \right) - 2\lambda.$$

The dual problem of (14) is then

$$\begin{array}{ll} \text{maximize} & -\sum_{k=1}^{m} \left((\|a_k\| - \lambda)^{+2} + \|a_k\|^2 \right) - 2\lambda \\ \text{subject to} & \lambda \geq 0, \end{array} \tag{15}$$

with variable $\lambda \in \mathbf{R}$. The solution is given by the root of

$$g'(\lambda)/2 = \sum_{k=1}^{m} (\|a_k\| - \lambda)^{+} - 1 = 0. \tag{16}$$

This equation resembles (10) and shows that λ is obtained by projecting the m-dimensional vector $(\|a_1\|, \|a_2\|, \ldots, \|a_m\|)$ onto the unit ℓ_1-norm ball in $\mathbf{R}^m$.

Therefore, a procedure to solve (14) is as follows.

Algorithm II:

1) If $\sum_{k=1}^{m} \|a_k\| \leq 1$, then $\lambda^* = 0$ and $x_k^* = a_k$ for all k.
2) Otherwise, define $\mathbf{a} = \begin{bmatrix} \|a_1\| & \|a_2\| & \cdots & \|a_m\| \end{bmatrix}^T$ and project $\mathbf{a}$ onto the unit ℓ_1-norm ball using the procedure in section III. In another word, solving (16) to obtain λ^*.
3) Use λ^* to calculate the projection via

$$t_k = (\|a_k\| - \lambda^*)^{+}, \quad x_k^* = \frac{a_k}{\|a_k\|} t_k.$$

V. Numerical Examples

In this section, we demonstrate the effectiveness of the proposed projection algorithms. We randomly generate 10 samples of n-dimensional vectors a where n ranges from 800 to 80000. We solve the dual problem (9) using Algorithm I and compare the result with solving the primal problem (6) using an interior-point method implemented in a software CVX [22]. The CPU time used to solve the problems are averaged over the 10 samples.

Figure 4 compares the average CPU times used to perform a projection task. The solver option in CVX is SDPT3 as we find that it is more reliable when the problem size increases. The algorithm implemented in SDPT3 is a primal-dual interior-point algorithm [19, §11.7]. The plot illustrates that our algorithm requires much less time than using SDPT3 (almost 10^4 times less). The results indicate that in high-dimensional problems, the convergence of the interior-point method is limited due to dependency of the problem size.

In the second experiment, we generate 10 samples of matrices $A_1, A_2, \ldots, A_p$ where $A_k \in \mathbf{R}^n$ where n ranges from 40 to 200 and $p = 3$. This results in $n^2 p$ variables in total which ranges from 4800 to 120000. We compute a projection of A_k onto the set

$$\sum_{i \neq j} \| \begin{bmatrix} (A_1)_{ij} & (A_2)_{ij} & \cdots & (A_p)_{ij} \end{bmatrix} \|_2 \leq \rho$$

with $\rho = 5$ by solving (15) and used Algorithm II in section IV. Figure 5 compares the average CPU times used to perform such projection. The plot demonstrates that the proposed projection algorithm is generally much faster than the interior-point method implemented in SDPT3.

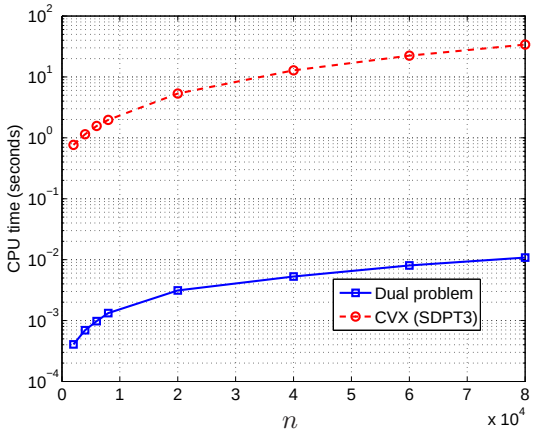


Fig. 4. The CPU time used to solve a projection onto an ℓ_1 ball in high dimensional setting

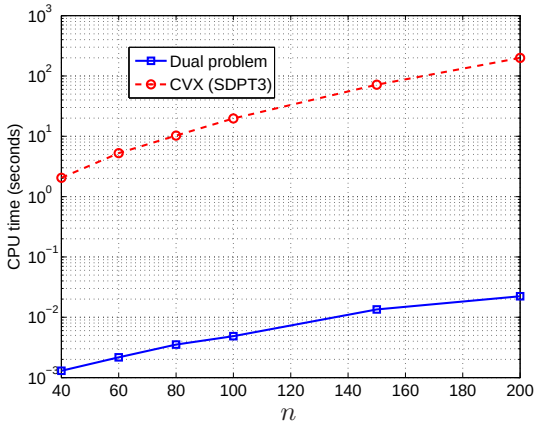


Fig. 5. The CPU time used to solve the projection onto a sum of ℓ_2-norm ball in high dimension setting

Our last experiment is to estimate a sparse AR model from a synthetic data set. We generate 500 time points from a sparse (and stable) AR process of size $n = 50$ and order $p = 3$. The total number of variables is 7500. The process is corrupted by noise with variance 1. The true AR coefficients have a common sparsity pattern shown in Figure 6 (left). We solve the problem (5) using the projected gradient method [2] and the projection step is computed using the procedure in section IV. An initial start in the algorithm is given by the least-squares estimate. Figure 7 shows the speed of convergence of the projected gradient method in combination with the proposed algorithm to compute a projection step.

Using $\rho = 300$ we obtain a sparse model with a zero pattern shown in Figure 6 (right), while the middle plot shows the zero pattern obtained from the least-squares estimate. The figure illustrates that with a limited number of data and the presence of noise, it is almost unlikely to reconstruct the underlying zero pattern of the AR coefficients using a simple approach as the least-squares method.

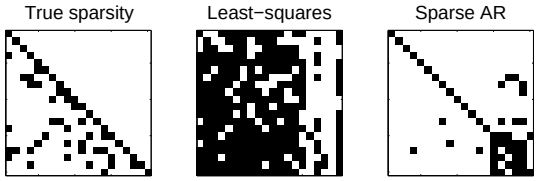


Fig. 6. Comparison of estimated sparsity patterns with the true sparsity. *Middle.* Thresholding the small entries of the least-squares estimate of AR model using a tolerance value of 10^{-2}. *Right.* Sparse AR model obtained by solving (5) using $\rho = 20$.

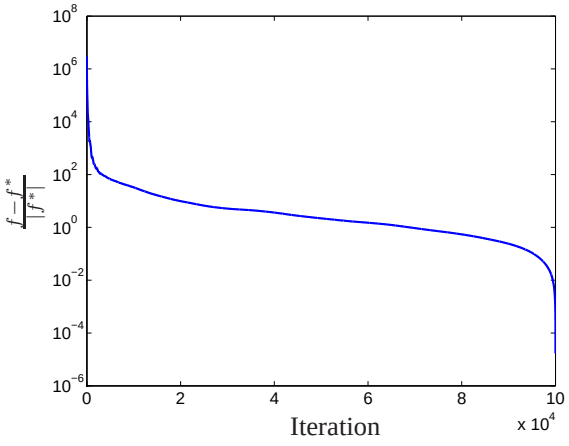


Fig. 7. The relative error of the cost objective versus the number of iterations. We solve the problem (5) using $n = 50, p = 3$ (7500 variables) and $\rho = 300$. The total computational time is approximately 25 minutes.

VI. Conclusion

An efficient method for computing the projection onto an ℓ_1-norm ball was presented. The problem can be casted as a convex optimization problem in n-dimensional space. We applied the duality theory and showed that the optimal values of the dual and primal (original) problems are the same. Therefore, we resort to solve the dual problem instead as it has only a scalar variable. The proposed procedure involves sorting elements of a vector that needs to be projected and thus requires $\mathcal{O}(n \log n)$ time in opposed to solving the primal problem by an interior-point method which requires $\mathcal{O}(n^3)$ time. The method finds many applications in statistical learning problems where discovering sparsity structures in model parameters is a main goal. As an application on sparse system identification, we considered a problem of estimating autoregressive models where many zeros in autoregressive coefficients are favored. This has a statistical interpretation as Granger causality between two variables in AR processes. Therefore, the problem formulation that includes ℓ_1-norm constraints has an advantage over a typical formulation such as the least-squares method. Numerical results showed that given that the true model is sparse, we can obtained a sparse estimate of AR model using such formulation while the least-squares cannot. Furthermore, the efficient ℓ_1 projection algorithm allows us to be able to solve the estimation problem in large scale.

References

[1] K. Burnham and D. Anderson, *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. New York: Springer, 2002.

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

[3] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[4] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[5] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 906–916, 2003.

[6] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[7] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[8] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.

[9] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006.

[10] Y. Kim, J. Kim, and Y. Kim, "Blockwise sparse regression," *Statistica Sinica*, vol. 16, no. 2, pp. 375–390, 2006.

[11] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

[12] R. Goebel, A. Roebroeck, D. Kim, and E. Formisano, "Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping," *Magnetic Resonance Imaging*, vol. 21, no. 10, pp. 1251–1261, 2003.

[13] R. Salvador, J. Suckling, C. Schwarzbauer, and E. Bullmore, "Undirected graphs of frequency-dependent functional connectivity in whole brain networks," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1457, pp. 937–946, 2005.

[14] P. Valdés-Sosa, J. M. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez, "Estimating brain functional connectivity with sparse multivariate autoregression," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1457, pp. 969–981, 2005.

[15] A. Fujita, J. Sato, H. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. Sogayar, and C. Ferreira, "Modeling gene expression regulatory networks with the sparse vector autoregressive model," *BMC Systems Biology*, vol. 1, no. 1, p. 39, 2007.

[16] A. Shojaie and G. Michailidis, "Discovering graphical granger causality using the truncating lasso penalty," *Bioinformatics*, vol. 26, no. 18, p. i517, 2010.

[17] A. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical granger modeling for gene expression regulatory networks discovery," *Bioinformatics*, vol. 25, no. 12, p. i110, 2009.

[18] A. Fujita, P. Severino, J. Sato, and S. Miyano, "Granger causality in systems biology: modeling gene networks in time series microarray data using vector autoregressive models," *Advances in Bioinformatics and Computational Biology*, pp. 13–24, 2010.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004, `www.stanford.edu/~boyd/cvxbook`.

[20] E. V. D. Berg, M. Schmidt, M. Friedlander, and K. Murphy, "Group sparsity via linear-time projection," Dept. of Computer Science, Univ. of British Columbia, Vancouver, BC, Canada, Tech. Rep., 2008.

[21] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l 1-ball for learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 272–279.

[22] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming (web page and software)," `http://stanford.edu/~boyd/cvx`, August 2008.