# THE GENERALIZED LASSO IS REDUCIBLE TO A SUBSPACE CONSTRAINED LASSO

*Hao Xu, David J. Eis and Peter J. Ramadge*

Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

## ABSTRACT

We investigate connections between the generalized lasso and the standard lasso problem. We show by an efficient direct construction, that the generalized lasso problem is reducible to a subspace constrained lasso. We then derive the dual of the subspace constrained lasso. This dual problem can be projected to the dual of a standard lasso problem with a modified dictionary. Finally, we discuss the application of these ideas to image approximation using the 2D fused lasso.

***Index Terms***— Sparsity, lasso, regularized regression, Lagrange dual

## 1. INTRODUCTION

The lasso problem [1] takes the form

$$\min_{\mathbf{w}\in\mathbb{R}^p} \quad 1/2\|\mathbf{x} - B\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|_1, \tag{1}$$

where $\lambda > 0$ is a regularization parameter. We call $B \in \mathbb{R}^{n\times p}$ the *dictionary*, the columns of $B$ *codewords* and $\mathbf{x}$ the *target vector*. The $\ell_1$ regularization in (1) encourages sparsity in the solution $\tilde{\mathbf{w}}$. The vector $\tilde{\mathbf{w}}$ thus gives a new (feature based) representation of $\mathbf{x}$ as a sparse linear combination of a subset of the codewords. This has proved effective in subsequent stages of processing (e.g., learning, classification). Indeed, such representations have proven effective in applications ranging from image restoration [2, 3], face recognition [4, 5], object recognition [6], speech classification [7], speech recognition [8], music genre classification [9], and topic detection in text documents [10].

Several extensions to the lasso problem have also been proposed. These retain aspects of the sparse representation while encouraging other desired properties in the solution. For example, the group lasso [11] performs sparse variable selection on groups of variables and the elastic net [12] encourages a grouping effect by weighting highly correlated variables similarly.

Recently a new form of the lasso has been introduced and analyzed: the *Generalized Lasso* [13]. This is a least squares problem with modified $\ell_1$ regularization:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \quad 1/2\|\mathbf{x} - B\mathbf{w}\|_2^2 + \lambda\|D\mathbf{w}\|_1. \tag{2}$$

Here $D \in \mathbb{R}^{m\times p}$ is a given matrix of arbitrary form. The idea is to select $D$, according to the application, so as to encourage the solution $\tilde{\mathbf{w}}$ to exhibit desired structural properties. A range of interesting applications of this idea have already emerged including the fused lasso [14], trend filtering [15], wavelet smoothing [16].

Our interest in (2) arises from its potential application to spatially informed analysis of fMRI data. In this application, $D$ would be selected to ensure that the weight vector $\tilde{\mathbf{w}}$ is suitably spatially smooth. This and similar applications of the generalized lasso have been of recent interest. For example, recent work has examined the application of the 3D fused lasso to EEG inverse problems [17] and its application to classification and feature selection of CT images [18]. The 2D fused lasso has also been used in compressed sensing reconstruction to accelerate 3D MRI [19]. Some related applications involve using the generalized elastic net for decoding cognitive states in fMRI [20] and using variants of group lasso and weighted fusion for paradigm free mapping of fMRI [21].

Since fMRI data is high dimensional ($\approx 10^5$ voxels in the whole brain and $\approx 10^3$ voxels in a specific region of interest) and often many subjects must be included in a full analysis, the efficient solution of (2) becomes very important. The first step towards obtaining efficient solution procedures is to fully understand the generalized lasso (2) and its dual. This is the principal objective of the current paper.

Tibshirani and Taylor [13] ask the important question "when is a generalized lasso problem reducible to a standard lasso problem?" They show that such a reduction is possible when $\mathrm{rank}(D) = m$. In this situation, it is also straightforward to obtain the dual of the generalized lasso problem. The authors of [13] point out, however, that many interesting generalized lasso problems have $\mathrm{rank}(D) < m$. They then examine a dual of the generalized lasso in this situation.

We show by direct construction that when $\mathrm{rank}(D) \leq m$ a generalized lasso problem is reducible to a lasso problem with a subspace constraint. We give an efficient procedure for this reduction requiring the one-time computation of two singular value decompositions (SVD). When $\mathrm{rank}(D) = m$, the constraint is always satisfied and the reduced problem is a lasso, in agreement with the result of [13]. A recent study [22] focuses on solving the lasso problem with linear constraints and develops efficient algorithms for this purpose. Hence this reduction of the generalized lasso to a subspace constrained lasso is of particular interest. We then obtain a dual of the

constrained lasso problem and analyze the properties of its solutions. We also show that this dual problem can be projected to the dual of a standard lasso problem with a modified dictionary. This is interesting theoretically, but the worst case complexity of the projection makes practical use unlikely. Finally, we give an example of solving the 2D fused lasso using the proposed reduction.

## 2. REDUCTION TO A CONSTRAINED LASSO

For simplicity of presentation, we will assume that $\operatorname{rank}(B) = n$. The reduction also applies when $\operatorname{rank}(B) < n$. We need the following notation. Let $D \in \mathbb{R}^{m\times p}$ have rank $r \leq m$ and denote its Moore-Penrose inverse by $D^+$, its range by $\mathcal{R}(D)$, and its null space by $\mathcal{N}(D)$. So $\dim \mathcal{R}(D) = r$ and $\dim \mathcal{N}(D) = p - r$.

Let the columns of $V_0 \in \mathbb{R}^{p\times(p-r)}$ form an orthonormal (ON) basis for $\mathcal{N}(D)$ and the columns of $Q \in \mathbb{R}^{m\times(m-r)}$ be basis for $\mathcal{R}(D)^\perp$ (the orthogonal complement of $\mathcal{R}(D)$).

Set $B_0 = BV_0 \in \mathbb{R}^{n\times(p-r)}$ with $\dim \mathcal{R}(B_0) = q \leq n$. Let the columns of $U$ form an ON basis for $\mathcal{R}(B_0)^\perp$ and set $\hat{\mathbf{x}} = U^T\mathbf{x} \in \mathbb{R}^{n-q}$ and $\hat{B} = U^TB \in \mathbb{R}^{(n-q)\times p}$.

The main result of this section is the following theorem.

**Theorem 1.** *The generalized lasso* (2) *is equivalent to the subspace constrained lasso,*

$$\begin{aligned} \min_{\mathbf{u}\in\mathbb{R}^m} \quad & 1/2\|\hat{\mathbf{x}} - \hat{B}D^+\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1 \\ \text{s.t.} \quad & Q^T\mathbf{u} = \mathbf{0}. \end{aligned} \tag{3}$$

*If* $\mathbf{u}^\star$ *is the solution of* (3)*, then*

$$\mathbf{w}^\star = V_0B_0^+\mathbf{x} + (I - V_0B_0^+B)D^+\mathbf{u}^\star \tag{4}$$

*is a solution of* (2)*. Conversely, if* $\mathbf{w}^\star$ *is a solution of* (2)*, then* $\mathbf{u}^\star = D\mathbf{w}^\star$ *is a solution of* (3)*.*

To gain some insights into this result, set $\mathbf{u} = D\mathbf{w}$ and write the problem (2) as:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}^m} \quad & 1/2\|\mathbf{x} - B\mathbf{w}\|_2^2 + \lambda\|\mathbf{u}\|_1 \\ \text{s.t.} \quad & D\mathbf{w} - \mathbf{u} = \mathbf{0}. \end{aligned} \tag{5}$$

For the reduction we must eliminate $\mathbf{w}$ from (5). Now every $\mathbf{w} \in \mathbb{R}^p$ can be uniquely written as $\mathbf{w} = \mathbf{w}_0 + \mathbf{w}_d$ with $\mathbf{w}_0 \in \mathcal{N}(D)$ and $\mathbf{w}_d \in \mathcal{N}(D)^\perp$. Substitution into (5) yields:

$$1/2\|\mathbf{x} - B\mathbf{w}_d - B\mathbf{w}_0\|_2^2 + \lambda\|\mathbf{u}\|_1 \tag{6}$$

and the constraint in (5) becomes $D\mathbf{w}_d - \mathbf{u} = 0$. By selecting $\mathbf{w}_0$ we can reduce the least squares cost without impacting the regularization cost or the constraint. Since $\mathbf{w}_0$ can only achieve points in the subspace $\mathcal{B}_0 = B\mathcal{N}(D)$ of $\mathcal{R}(B)$, an optimal choice of $\mathbf{w}_0$ will cancel some part of the orthogonal projection of $(\mathbf{x} - B\mathbf{w}_d)$ onto $\mathcal{B}_0$. This depends on $\mathbf{w}_d$ because $B\mathbf{w}_d$ can also have a component in the subspace $\mathcal{B}_0$: think of this as crosstalk from $\mathbf{w}_d$ into $\mathcal{B}_0$. To deduce an optimal choice of $\mathbf{w}_0$, let $U_{B_0} \in \mathbb{R}^{n\times q}$ be a matrix with ON columns that form a basis for $\mathcal{B}_0$ and write $B\mathbf{w}_0 = U_{B_0}\mathbf{v}_0$ for some $\mathbf{v}_0 \in \mathbb{R}^q$. Using this representation, the fact that $U_{B_0}$ has orthonormal columns, and some algebra, the objective (6) can be transformed into:

$$\begin{aligned} &1/2\|(I - U_{B_0}U_{B_0}^T)(\mathbf{x} - B\mathbf{w}_d)\|_2^2 \\ &\quad + 1/2\|U_{B_0}^T(\mathbf{x} - B\mathbf{w}_d) - \mathbf{v}_0\|_2^2 + \lambda\|\mathbf{u}\|_1 \end{aligned} \tag{7}$$

From (7) it is clear that $\mathbf{v}_0 = U_{B_0}^T(\mathbf{x} - B\mathbf{w}_d)$ zeroes the second term in (7) and reduces (6) to:

$$1/2\|(I - U_{B_0}U_{B_0}^T)(\mathbf{x} - B\mathbf{w}_d)\|_2^2 + \lambda\|\mathbf{u}\|_1 \tag{8}$$

Let the columns of $U$ form an ON basis for $\mathcal{B}_0^\perp$. Since $I - U_{B_0}U_{B_0}^T$ is orthogonal projection onto $\mathcal{B}_0^\perp$, we can write $(I - U_{B_0}U_{B_0}^T) = UU^T$. Let $\hat{\mathbf{x}} = U^T\mathbf{x}$ and $\hat{B} = U^TB$. With these definitions, we can rewrite the objective (8) as

$$1/2\|\hat{\mathbf{x}} - \hat{B}\mathbf{w}_d\|_2^2 + \lambda\|\mathbf{u}\|_1 \tag{9}$$

Here $\bar{\mathbf{x}} = UU^T\mathbf{x}$ is the projection of $\mathbf{x}$ onto $\mathcal{B}_0^\perp$ and $\hat{\mathbf{x}} = U^T\mathbf{x} \in \mathbb{R}^{n-q}$ is the vector of coordinates of $\bar{\mathbf{x}}$ with respect to the ON basis $U$. Think of $\bar{\mathbf{x}}$ as the component of $\mathbf{x}$ that can't be explained by $\mathbf{w}_0$. Similarly, $\bar{B} = UU^TB$ is the column-by-column projection of the codewords in $B$ onto $\mathcal{B}_0^\perp$ and the columns of $\hat{B} = U^TB \in \mathbb{R}^{(n-q)\times p}$ give the coordinates of the columns of $\bar{B}$ with respect to the basis $U$. The degrees of freedom in $U_{B_0}$ have been exploited by $\mathbf{w}_0$ and hence can be removed from the dictionary when solving for $\mathbf{w}_d$.

The variabale $\mathbf{w}_d$ can be eliminated as follows. $\mathbf{w}_d$ and $\mathbf{u}$ satisfy the constraint $D\mathbf{w}_d = \mathbf{u}$. Now when restricted to $\mathcal{N}(D)^\perp$, $D$ is a one-to-one mapping from $\mathcal{N}(D)^\perp$ onto $\mathcal{R}(D)$ with $D\mathbf{w}_d = \mathbf{u}$. Hence $\mathbf{w}_d = D^+\mathbf{u}$. Substituting this expression into (9) yields the equivalent objective:

$$1/2\|\hat{\mathbf{x}} - \hat{B}D^+\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1 \tag{10}$$

with constraint $\mathbf{u} \in \mathcal{R}(D)$. Since the columns of $Q$ form a basis for $\mathcal{R}(D)^\perp$, (10) and its constraint can be written as (3).

When $\operatorname{rank}(D) = m$, the constraint in (3) is always satisfied. In this case, the generalized lasso reduces to a lasso in agreement with the result in [13].

## 3. DUAL OF SUBSPACE CONSTRAINED LASSO

We now explore the dual of the subspace constrained lasso. To keep the exposition general, let $\mathbf{y} \in \mathbb{R}^k$ denote the target vector and $C \in \mathbb{R}^{k\times l}$ denote the dictionary. Let the columns of $Q \in \mathbb{R}^{l\times(l-r)}$ be a basis for $\mathcal{R}(U)^\perp$ and require that $Q^T\mathbf{u} = 0$. The primal constrained lasso is then:

$$\begin{aligned} \min_{\mathbf{u}\in\mathbb{R}^l} \quad & 1/2\|\mathbf{y} - C\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1 \\ \text{s.t.} \quad & Q^T\mathbf{u} = 0 \end{aligned} \tag{11}$$

If we ignore the constraint in (11), we obtain the associated lasso problem:

$$\min_{\mathbf{u}\in\mathbb{R}^l} \ {}^1\!/_2\|\mathbf{y}-C\mathbf{u}\|_2^2+\lambda\|\mathbf{u}\|_1 \tag{12}$$

To provide an important point of comparison, we quickly review the Lagrangian dual of (12) (see e.g., [23–26]). This can be expressed as [26]:

$$\begin{aligned}\max_{\boldsymbol{\theta}\in\mathbb{R}^k} \ & {}^1\!/_2\|\mathbf{y}\|_2^2-{}^{\lambda^2}\!/_2\|\boldsymbol{\theta}-\mathbf{y}/\lambda\|_2^2\\ \text{s.t.} \ & -\mathbf{1}\le C^T\boldsymbol{\theta}\le\mathbf{1}.\end{aligned} \tag{13}$$

Here $\mathbf{1}\in\mathbb{R}^l$ denotes the vector of all $1's$ and the inequalities in (13) are interpreted component-wise. In addition, the solutions of (12) and (13) are related by

$$C\mathbf{u}^\star=\mathbf{y}-\lambda\boldsymbol{\theta}^\star,\ \mathbf{c}_i^T\boldsymbol{\theta}^\star\in\begin{cases}\{\operatorname{sign} w_i^\star\} & \text{if } w_i^\star\neq 0,\\ [-1,1] & \text{if } w_i^\star=0.\end{cases} \tag{14}$$

The inequality constraints in (13) specify a set $\mathcal{F}$ of dual feasible points: $\boldsymbol{\theta}\in\mathcal{F}\Leftrightarrow\mathbf{c}_i^T\boldsymbol{\theta}\le 1,\ i=1,\ldots,l$. $\mathcal{F}$ is a nonempty ($0\in\mathcal{F}$), closed, convex polyhedron. The objective of (13) seeks a $\boldsymbol{\theta}\in\mathcal{F}$ that is closest to $\mathbf{y}/\lambda$. For every $\mathbf{y}\in\mathbb{R}^k$ and $\lambda>0$, (13) has a unique solution $\boldsymbol{\theta}^\star(\mathbf{y},\lambda)$ - this is a property of projection onto a closed convex set [27].

Now consider the Lagrangian dual of problem (11). Set $\mathbf{z}=\mathbf{y}-C\mathbf{u}$ and rewrite (11) as:

$$\begin{aligned}\min_{\mathbf{z}\in\mathbb{R}^k,\mathbf{u}\in\mathbb{R}^l} \ & {}^1\!/_2\|\mathbf{z}\|_2^2+\lambda\|\mathbf{u}\|_1\\ \text{s.t.} \ & Q^T\mathbf{u}=0\\ & \mathbf{y}-C\mathbf{u}-\mathbf{z}=0\end{aligned} \tag{15}$$

The corresponding Lagrangian for this problem is

$$\begin{aligned}&L(\mathbf{z},\mathbf{u},\mathbf{v},\boldsymbol{\sigma},\boldsymbol{\tau})=\\ &{}^1\!/_2\,\mathbf{z}^T\mathbf{z}+\lambda\|\mathbf{u}\|_1+\boldsymbol{\sigma}^T(\mathbf{y}-C\mathbf{u}-\mathbf{z})+\boldsymbol{\tau}^T(Q^T\mathbf{u}).\end{aligned} \tag{16}$$

Minimizing the Lagrangian with respect to $\mathbf{z}$ gives $\mathbf{z}=\boldsymbol{\sigma}$, and with respect to $\mathbf{u}$ gives:

$$\mathbf{c}_i^T\boldsymbol{\sigma}-\mathbf{q}_i^T\boldsymbol{\tau}\in\begin{cases}\{\operatorname{sign}(u_i)\lambda\} & \text{if } u_i\neq 0\\ [-\lambda,\lambda] & \text{if } u_i=0\end{cases} \tag{17}$$

where $\mathbf{c}_i$ (resp. $\mathbf{q}_i^T$) is $i$-th column (resp. row) of $C$ (resp. $Q$). This in turn implies $\lambda\|\mathbf{u}\|_1-(\boldsymbol{\sigma}^TC-\boldsymbol{\tau}^TQ^T)\mathbf{u}=0$ and the set of constraints $-\lambda\mathbf{1}\le C^T\boldsymbol{\sigma}-Q\boldsymbol{\tau}\le\lambda\mathbf{1}$. Let $\boldsymbol{\theta}=\boldsymbol{\sigma}/\lambda$ and $\boldsymbol{\nu}=-\boldsymbol{\tau}/\lambda$. Then the Langragian simplifies to ${}^1\!/_2\|\mathbf{y}\|_2^2-{}^{\lambda^2}\!/_2\|\boldsymbol{\theta}-\mathbf{y}/\lambda\|_2^2$ and the constraints to: $-\mathbf{1}\le C^T\boldsymbol{\theta}+Q\boldsymbol{\nu}\le\mathbf{1}$.

Putting all of this together yields a dual of (11):

$$\begin{aligned}\max_{\boldsymbol{\theta}\in\mathbb{R}^k,\boldsymbol{\nu}\in\mathbb{R}^{l-r}} \ & {}^1\!/_2\|\mathbf{y}\|_2^2-{}^{\lambda^2}\!/_2\|\boldsymbol{\theta}-\mathbf{y}/\lambda\|_2^2\\ \text{s.t.} \ & -\mathbf{1}\le C^T\boldsymbol{\theta}+Q\boldsymbol{\nu}\le\mathbf{1}\end{aligned} \tag{18}$$

with the solutions of problems (11) and (18) related by

$$\begin{aligned}&C\mathbf{u}^\star=\mathbf{y}-\lambda\boldsymbol{\theta}^\star\\ &\mathbf{c}_i^T\boldsymbol{\theta}^\star+\mathbf{q}_i^T\boldsymbol{\nu}^\star\in\begin{cases}\{\operatorname{sign} u_i^\star\} & \text{if } u_i^\star\neq 0,\\ [-1,1] & \text{if } u_i^\star=0.\end{cases}\end{aligned} \tag{19}$$

The dual variable in (18) is the composite vector $(\boldsymbol{\theta},\boldsymbol{\nu})$ and the constraint can be written in terms of this variable as

$$-\mathbf{1}\le\begin{bmatrix}C\\ Q^T\end{bmatrix}^T\begin{bmatrix}\boldsymbol{\theta}\\ \boldsymbol{\nu}\end{bmatrix}\le\mathbf{1} \tag{20}$$

This augments the dictionary by adjoining $Q^T$ to the rows of $C$. Modulo this modification, the objective and constraint in (18) are similar to those in (13). The key difference is that in (18) the objective only depends on a subset of the dual coordinates ($\boldsymbol{\theta}$).

Let $\mathcal{F}$ denote the dual feasible points of the unconstrained lasso (12). This is a polyhedron in $\mathbb{R}^k$. Similarly, let $\mathcal{F}_c$ denote the dual feasible points of (11). This is a polyhedron in $\mathbb{R}^{k+l-r}$. $\mathcal{F}$ can be embedded in $\mathbb{R}^{k+l-r}$ by padding $\boldsymbol{\theta}\in\mathcal{F}$ with $l-r$ zeros, i.e., $\boldsymbol{\theta}\mapsto(\boldsymbol{\theta},\mathbf{0})$. Denote this embeded set by $\mathcal{F}_e$. Under this embedding, $\mathcal{F}_e\subseteq\mathcal{F}_c$ with equality when $r=l$. So adding a constraint to the lasso enlarges the set of dual feasible points. In general, the constrained lasso can't achieve the same minimal objective value as the associated unconstrained lasso. Hence the dual of the constrained problem compensates by enlarging the dual feasible set. This enables it to achieve the same optimum value as the (constrained) primal problem. We see this in (18) where the extra coordinates $\boldsymbol{\nu}$ in the dual variable give the opportunity to expand the set of dual feasible points beyond that of the unconstrained lasso. Clearly, the $\boldsymbol{\nu}$ of a dual solution need not be unique (see also [13]), but $\theta$ is unique.

**Lemma 1.** *For each $\mathbf{y}\in R^k$ and $\lambda>0$, the dual problem (18) has a solution. In addition, the $\boldsymbol{\theta}$ component is unique, i.e., if $(\tilde{\boldsymbol{\theta}}_j,\tilde{\boldsymbol{\nu}}_j)$ are solutions, $j=1,2$, then $\tilde{\boldsymbol{\theta}}_1=\tilde{\boldsymbol{\theta}}_2$.*

*Proof.* $\mathcal{F}_c$ is nonempty, convex and closed. The linear mapping $(\boldsymbol{\theta},\boldsymbol{\nu})\mapsto\boldsymbol{\theta}$ projects $\mathcal{F}_c$ to a nonempty closed convex set $\mathcal{F}_p$ in $\mathbb{R}^k$. We seek the point in $\mathcal{F}_p$ that is closest to $\mathbf{y}/\lambda$. By standard results, [27, §3.1], there is a unique point $\tilde{\boldsymbol{\theta}}$ in the projected set that is closest to $\mathbf{y}/\lambda$. Since $\tilde{\boldsymbol{\theta}}$ is in the projected set, there exists $\tilde{\boldsymbol{\nu}}$ such that $(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\nu}})\in\mathcal{F}_c$ and this point is a solution of the dual. □

## 4. PROJECTION OF THE DUAL

By (18), the polyhedron $\mathcal{F}_c$ of dual feasible points depends only on the augmented dictionary through the inequality constraints (20). It does not depend on $\mathbf{y}$ or $\lambda$. In addition, only the $\boldsymbol{\theta}$ component of the dual variable plays a role in the objective function. By variable elimination in linear inequalities (e.g., via Fourier-Motzkin elimination), the $\boldsymbol{\nu}$ component can be eliminated and only the $\boldsymbol{\theta}$ component retained.

**Lemma 2.** *Assume $r < m$. Then there exists an integer $d \geq 1$ and a matrix $P \in \mathbb{R}^{d \times l}$ such that $\mathcal{F}_p$ is the set of points $\boldsymbol{\theta} \in \mathbb{R}^k$ satisfying $-\mathbf{1} \leq PC^T\boldsymbol{\theta} \leq \mathbf{1}$.*

*Proof.* Use Fourier-Motzkin elimination [28]. □

By projecting $\mathcal{F}_c$ to $\mathcal{F}_p$, we project (18) to:

$$\begin{aligned} \max_{\boldsymbol{\theta} \in \mathbb{R}^k} \quad & 1/2\|\mathbf{y}\|_2^2 - \lambda^2/2\|\boldsymbol{\theta} - \mathbf{y}/\lambda\|_2^2 \\ \text{s.t.} \quad & -\mathbf{1} \leq PC^T\boldsymbol{\theta} \leq \mathbf{1} \end{aligned} \tag{21}$$

We call (21) the *projected dual*. Let $A = CP^T \in \mathbb{R}^{k \times d}$ denote the dictionary of the projected dual. The projected dual is the dual of the standard lasso:

$$\min_{\mathbf{t} \in \mathbb{R}^d} \quad 1/2\|\mathbf{y} - A\mathbf{t}\|_2^2 + \lambda\|\mathbf{t}\|_1 \tag{22}$$

This interesting result says that a subspace constrained lasso can be transformed into a standard lasso problem by "filtering" the dictionary $C$ via $P^T$. Unfortunately, in general this will remain purely of theoretical interest since in the worst case, variable elimination in linear inequalities has exponential complexity. So in the worst case, during the process of variable elimination the size of the dictionary grows exponentially.

## 5. EXAMPLE: 2D FUSED LASSO

The Fused Lasso is a special case of (2). In the 1D fused lasso [14], $D$ forms the differences in adjacent entries of $\mathbf{w}$ and $\text{rank}(D) = m$. So the 1D fused lasso reduces to a lasso. In the 2D fused lasso, $\mathbf{x}$ represents an image rearranged into a vector and $D$ forms the horizontal and vertical differences at each pixel. If the image is $n \times n$, then $D$ is $m = 2(n^2 - n)$ by $p = n^2$ and won't have full row rank for $n > 2$. Consider the special case when $B = I$ ("signal approximation"). Let $D$ have rank $r \leq m$. To compute the reduction to a subspace constrained lasso we first find a full SVD

$$D = \begin{bmatrix} U_D & X_D \end{bmatrix} \begin{bmatrix} \Sigma_D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_D & Y_D \end{bmatrix}^T.$$

$Y_D$ is $p \times (p - r)$ and forms an ON basis for $\mathcal{N}(D)$. $X_D$ is $m \times (m - r)$ and forms an ON basis for $\mathcal{R}(D)^\perp$. The Moore-Penrose inverse is $D^+ = V_D\Sigma_D^{-1}U_D^T$. We can take $V_0 = Y_D$. Then $\mathcal{R}(BV_0) = \mathcal{R}(Y_D)$ and we can take $U = V_D$. So the new dictionary is $\hat{B}D^+ = U^TBD^+ = \Sigma_D^{-1}U_D^T$ and $\hat{\mathbf{x}} = V_D^T\mathbf{x}$. We need $Q$ to be basis for $\mathcal{R}(D)^\perp$. So set $Q = X_D$. This yields the 2D fused subspace constrained lasso:

$$\begin{aligned} \min_{\mathbf{u} \in \mathbb{R}^m} \quad & 1/2\|\hat{\mathbf{x}} - \Sigma_D^{-1}U_D^T\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1 \\ \text{s.t.} \quad & X_D^T\mathbf{u} = \mathbf{0}. \end{aligned} \tag{23}$$

Fig. 1 displays an example solved using this reduction. The original binary image (top left) is $32 \times 32$. Normal $N(0, 0.1^2)$ noise is added to the image (top right). We applied the reduction procedure, then solved the subspace constrained lasso (we used CVX [29]). The solution ($\lambda = 1$) is shown in Fig. 1 (bottom).

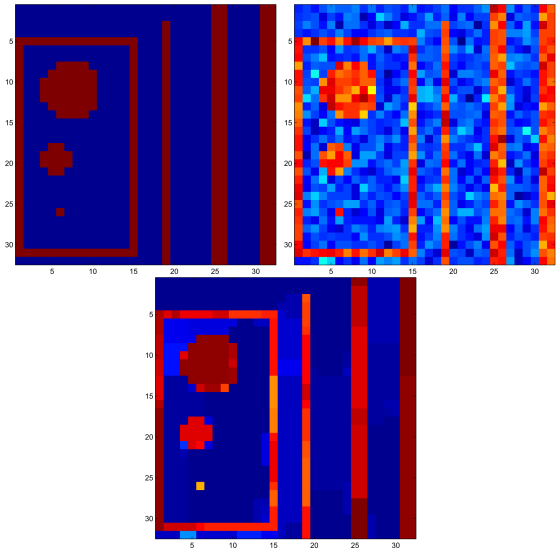

**Fig. 1**. Example: 2D fused lasso solved using a subspace constrained lasso representation. Axes are image pixel indexes. Further details are in the text.

## 6. CONCLUSION

Our reduction results on the generalized lasso add additional understanding to this new and potentially useful sparse regression problem. We have shown that the generalized lasso is easily reducible to a subspace constrained lasso. The main computations required for this reduction are two singular value decompositions. Moreover, recent work [22] has already put in place algorithms for the solution of constrained lasso problems. Current algorithms to solve the constrained lasso include the path algorithm in [13] and the iterative non-linear conjugate gradient algorithm in [30] and general convex program solvers. At the very least the reduction reported here provides an alternative approach for solving the generalized lasso problem. In addition, the dual of the subspace constrained lasso can be projected to the dual problem of a standard lasso with a modified dictionary. Thus by modifying the dictionary, a subspace constrained lasso can be equivalently formulated as a lasso problem. By extension, the generalized lasso can also be equivalently formulated as a lasso problem. This is an interesting theoretical result. Unfortunately, the worst case complexity of the required construction is exponential. So for now, this second result is principally of theoretical interest. Finally we note that while finalizing this accepted paper we came across the very recent manuscript [31] which also considers equality constrained lasso problems.

# References


[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, vol. 58, no. 1, pp. 267–288, 1996.

[2] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, vol. 17, no. 1, pp. 53–69, 2008.

[3] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2272–2279.

[4] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[5] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Towards a practical face recognition system: robust alignment and illumination by sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[6] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in Neural Information Processing Systems*, 2009, vol. 3.

[7] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *ICASSP*, 2010, pp. 4370–4373.

[8] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representation features for speech recognition," in *Interspeech*, 2010.

[9] K. Chang, J. Jang, and C. S. Iliopoulos, "Music genre classification via compressive sampling," in *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, 2010, pp. 387–392.

[10] S. Prasad, P. Melville, A. Banerjee, and V. Sindhwani, "Emerging topic detection using dictionary learning," in *ACM Conference on Information and Knowledge Management*, 2011.

[11] Ming Yuan and Yi Lin, "Model selection and estimation in regression with grouped variables," *J. R. Statist. Soc.: Series B*, vol. 68, pp. 4967, 2006.

[12] Hui Zou and Trevor Hastie, "Regularization and variable selection via the elastic net," *J. R. Statist. Soc.: Series B*, pp. 301–320, 2005.

[13] R. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, 2011.

[14] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistics Society: Series B*, vol. 67, no. 3, pp. 91–108, 2005.

[15] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, "l1 trend filtering," *SIAM Review*, vol. 51, no. 2, pp. 339–360, 2009.

[16] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[17] M. Vega-Hernandez, E. Martinez-Montes, J. M. Sanchez-Bornot, A. Lage-Castellanos, and P. A. Valdes-Sosa, "Penalized least squares methods for solving the eeg inverse problem," *Statistica Sinica*, vol. 18, pp. 1535–1551, 2008.

[18] S. Yang, L. G. Shapiro, M. L. Cunningham, M. Speltz, and S. I. Lee, "Classification and feature selection for craniosynostosis," *ACM Conf. Bioinformatics, Computational Biology and Biomedicine*, 2011.

[19] Yoon-Chul Kim, Shrikanth S. Narayanan, and Krishna S. Nayak, "Accelerated 3d mri of vocal tract shaping using compressed sensing and parallel imaging," *ICASSP*, pp. 389–392, 2009.

[20] B. Ng, A. Vahdat, R. Abugharbieh, , and G Hamarneh, "Generalized sparse classifiers for decoding cognitive states in fmri," *MICCAI MLMI*, 2010.

[21] C. Caballero Gaudes, F. I. Karahanoglu, F. Lazeyras, and D. Van de Ville, "Structured sparse deconvolution for paradigm free mapping of functional mri data," *ISBI*, 2012.

[22] Tianhong He, *LASSO and General $\ell_1$ Regularized Regression under Linear Equality and Inequality Constraints*, Ph.D. thesis, Purdue University, 2011.

[23] M. R. Osborne, B. Presnell, and B. A. Turlach, "On the lasso and its dual," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 319–337, June 2000.

[24] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large scale $\ell_1$-regularized least squares," *IEEE Selected Topics in Signal Processing*, vol. 1, pp. 606–617, 2007.

[25] L. El Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination in sparse supervised learning," Tech. Rep. UCB/EECS-2010-126, EECS Department, University of California, Berkeley, Sep 2010.

[26] Z. J. Xiang, H. Xu, and P. J. Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries," in *Advances in Neural Information Processing Systems*, 2011.

[27] Hiriart-Urruty, Jean-Baptiste, and Claude Lemarechal, *Fundamentals of Convex Analysis*, Springer, 2001.

[28] H. P. Williams, "Fourier's method of linear programming and its dual," *The American Mathematical Monthly*, vol. 93, no. 9, pp. 681–695, 1986.

[29] Inc. CVX Research, "CVX: Matlab software for disciplined convex programming, version 2.0 beta," `http://cvxr.com/cvx`, Sept. 2012.

[30] Michael Lustig, David Donoho, and John M. Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging," *Magnetic Resonance in Medicine*, vol. 58, pp. 1182–1195, 2007.

[31] Gareth M. James, Courtney Paulson, and Paat Rusmevichientong, "The constrained lasso," `http://www-bcf.usc.edu/~gareth/research/CLassoFinal.pdf`, 2013.