A tutorial on the LASSO and the "shooting algorithm"

Gautam V. Pendse*[1]

[1] P.A.I.N Group, Imaging and Analysis Group (IMAG), McLean Hospital, Harvard Medical School

February 8, 2011

*To whom correspondence should be addressed. e-mail: gpendse@mclean.harvard.edu

Contents

List of Figures

1 Abstract

The LASSO is an L_1 penalized regression technique introduced by Tibshirani [1996]. An efficient algorithm called the "shooting algorithm" was proposed by Fu [1998] for solving the LASSO problem in the multiparameter case. In this tutorial, we present a simple and self-contained derivation of the LASSO shooting algorithm.

2 Code distribution for LASSO shooting

MATLAB (`www.mathworks.com`) code for solving a LASSO problem using the "shooting algorithm" and estimating the regularization parameter can be downloaded from:

`http://www.gautampendse.com/software/lasso/webpage/lasso_shooting.html`

3 Notation

- Scalars will be denoted in a non-bold font possibly with subscripts (e.g. λ, β_i). We will use bold face lower case letters possibly with subscripts to denote vectors (e.g. $\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{z_1}$) and bold face upper case letters possibly with subscripts to denote matrices (e.g. $\boldsymbol{X}, \boldsymbol{B_1}$). The ith element of a vector $\boldsymbol{x}$ will be denoted by x_i in non-bold font.
- The transpose of a matrix $\boldsymbol{X}$ will be denoted by $\boldsymbol{X}^T$ and its inverse will be denoted by $\boldsymbol{X}^{-1}$. We will denote the $p \times p$ identity matrix by $\boldsymbol{I}_p$. A vector or matrix of all zeros will be denoted by a bold face zero $\mathbf{0}$ whose size should be clear from context.
- The q-norm of a $p \times 1$ vector $\boldsymbol{\beta}$ will be denoted by $||\boldsymbol{\beta}||_q = \left(\sum_{i=1}^{p} |\beta_i|^q \right)^{\frac{1}{q}}$ where $|\beta_i|$ denotes the absolute value of β_i.

4 Introduction

Given n feature vectors of length p arranged in the rows of a design matrix $\boldsymbol{X}$ we would like to predict the $n \times 1$ observed response vector $\boldsymbol{y}$ via a linear model. LASSO solves the following L_1

regularized optimization problem:

$$
\begin{aligned}
&\min_{\boldsymbol{\beta}} h(\boldsymbol{\beta}) = \tfrac{1}{2}\,||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1, \text{ where } \lambda \geq 0 && (4.1)\\
&\quad\text{where} && (4.2)\\
&\qquad \boldsymbol{\beta} \text{ is a } p \times 1 \text{ vector} && (4.3)\\
&\qquad \boldsymbol{y} \text{ is a } n \times 1 \text{ vector} && (4.4)\\
&\qquad \boldsymbol{X} \text{ is a } n \times p \text{ matrix} && (4.5)
\end{aligned}
$$

We assume that $n > p$. The penalty term in 4.1 is a 1-norm penalty or simply the sum of the absolute values of the components of $\boldsymbol{\beta}$. As we shall see, this penalty term encourages sparsity in the components of the solution vector $\boldsymbol{\beta}$ and thus automatically leads to feature/model selection. In addition, the penalty term regularizes the solution vector $\boldsymbol{\beta}$ and hence prevents overfitting.

5 Preliminaries

In this section, we give some background material that is necessary for a clear understanding of how LASSO works. We will cover some basic relationships between convexity, positive semidefiniteness, local and global minimizers.

Definition 5.1 (Convexity)**.** A set $\mathcal{D}$ is convex if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{D}$ and all $\alpha \in (0,1)$, $\boldsymbol{x} = \alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2 \in \mathcal{D}$. A function $f(\boldsymbol{x})$ is convex if (1) its domain $\mathcal{D}$ is convex and (2) $f(\boldsymbol{x}) = f(\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \leq \alpha f(\boldsymbol{x}_1) + (1-\alpha)f(\boldsymbol{x}_2)$.

Definition 5.2 (PSD)**.** A $p \times p$ matrix $\boldsymbol{H}$ is positive semidefinite (PSD) if for all $p \times 1$ vectors $\boldsymbol{z}$ we have $\boldsymbol{z}^T\boldsymbol{H}\boldsymbol{z} \geq 0$.

Proposition 5.3 (PSD Hessian implies Convexity)**.** *Suppose $\boldsymbol{x}$ is a $p \times 1$ vector and $f(\boldsymbol{x})$ is a scalar function of p variables with continuous second order derivatives defined on a convex domain $\mathcal{D}$. If the Hessian $\nabla^2 f(\boldsymbol{x})$ is positive semidefinite for all $x \in \mathcal{D}$ then f is convex.*

Proof. By Taylor's theorem for all $\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h} \in \mathcal{D}$ we can write:

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T\boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T\nabla^2 f(\boldsymbol{x}+\theta\boldsymbol{h})\boldsymbol{h} \tag{5.1}$$

for some $\theta \in (0,1)$. By assumption, the Hessian $\nabla^2 f(\boldsymbol{x}+\theta\boldsymbol{h})$ is positive semidefinite and hence $\boldsymbol{h}^T\nabla^2 f(\boldsymbol{x}+\theta\boldsymbol{h})\boldsymbol{h} \geq 0$. Hence for all $\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h} \in \mathcal{D}$ we can write:

$$f(\boldsymbol{x}+\boldsymbol{h}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T\boldsymbol{h} \tag{5.2}$$

Letting $\boldsymbol{x} + \boldsymbol{h} = \boldsymbol{y}$ we can also write the above equation as:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \ \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D} \tag{5.3}$$

Now let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be any two points in $\mathcal{D}$ and let $\alpha \in (0,1)$ be a scalar. Then by the convexity of $\mathcal{D}$, $\boldsymbol{x} = \alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2 \in \mathcal{D}$.

By 5.3 we can write:

$$f(\boldsymbol{x}_1) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{x}_1 - \boldsymbol{x}) \tag{5.4}$$

and

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{x}_2 - \boldsymbol{x}) \tag{5.5}$$

Multiplying 5.4 by α and 5.5 by $(1-\alpha)$ and adding we get:

$$\begin{aligned} \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2) &\geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2 - \boldsymbol{x}) \\ &= f(\boldsymbol{x}) \end{aligned} \tag{5.6}$$

Hence $f(\boldsymbol{x})$ is convex.

□

Proposition 5.4. *If $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are convex functions defined on a convex domain $\mathcal{D}$ then $r(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ is also convex on $\mathcal{D}$.*

Proof. Suppose $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{D}$ and let $\boldsymbol{x} = \alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2$ for some $\alpha \in (0,1)$. Since $\mathcal{D}$ is convex we have $\boldsymbol{x} \in \mathcal{D}$. Now

$$\begin{aligned} r(\boldsymbol{x}) &= r(\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) && (5.7) \\ &= f(\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) + g(\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) && (5.8) \\ &\leq \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2) + \alpha g(\boldsymbol{x}_1) + (1-\alpha) g(\boldsymbol{x}_2) \qquad \text{by convexity of } f \text{ and } g && (5.9) \\ &= \alpha r(\boldsymbol{x}_1) + (1-\alpha) r(\boldsymbol{x}_2) && (5.10) \end{aligned}$$

Hence $r(\boldsymbol{x})$ is convex.

□

Proposition 5.5 (LASSO objective is convex)**.** *The LASSO objective function $h(\boldsymbol{\beta})$ in equation 4.1 is convex.*

Proof. We can write the LASSO objective as:

$$h(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + g(\boldsymbol{\beta}) \tag{5.11}$$

where $f(\boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$ and $g(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1$. Note that the domain of both functions f and g is $\mathbf{R}^p$ which is convex.

The Hessian of $f(\boldsymbol{\beta})$ is $\nabla^2 f(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{X}$. For any $p \times 1$ vector $\boldsymbol{z}$: $\boldsymbol{z}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{z} = ||\boldsymbol{X}\boldsymbol{z}||_2^2 \geq 0$. Hence $\nabla^2 f(\boldsymbol{\beta})$ is positive semidefinite. Hence by proposition 5.3 $f(\boldsymbol{\beta})$ is convex.

For any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and any $\alpha \in (0,1)$, let $\boldsymbol{\beta} = \alpha\boldsymbol{\beta}_1 + (1-\alpha)\boldsymbol{\beta}_2$. Then

$$\begin{aligned}
g(\boldsymbol{\beta}) &= \lambda\,||\alpha\boldsymbol{\beta}_1 + (1-\alpha)\boldsymbol{\beta}_2||_1 && (5.12)\\
&\leq \lambda\,||\alpha\boldsymbol{\beta}_1||_1 + \lambda\,||(1-\alpha)\boldsymbol{\beta}_2||_1 \qquad \text{Triangle inequality} && (5.13)\\
&= \lambda\,\alpha\,||\boldsymbol{\beta}_1||_1 + \lambda\,(1-\alpha)\,||\boldsymbol{\beta}_2||_1 && (5.14)\\
&= \alpha\,g(\boldsymbol{\beta}_1) + (1-\alpha)\,g(\boldsymbol{\beta}_2) && (5.15)
\end{aligned}$$

Hence $g(\boldsymbol{\beta})$ is convex. Since $f(\boldsymbol{\beta})$ and $g(\boldsymbol{\beta})$ are both convex, by proposition 5.4 $h(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + g(\boldsymbol{\beta})$ is also convex.

□

Proposition 5.6. *If $f(\boldsymbol{x})$ is a convex function defined for $\boldsymbol{x} \in \mathcal{D}$ with convex $\mathcal{D}$ then any local minimizer of f on $\mathcal{D}$ is a global minimizer of f on $\mathcal{D}$.*

Proof. Suppose $\boldsymbol{x}^*$ is a local minimizer but not a global minimizer. Then there exists a global minimizer $\boldsymbol{x}_g^*$ such that:

$$f(\boldsymbol{x}_g^*) < f(\boldsymbol{x}^*) \tag{5.16}$$

In addition, since $\boldsymbol{x}^*$ is a local minimizer we must have:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*) \text{ for all } \boldsymbol{y} \in \text{nbhd}(\boldsymbol{x}^*) \tag{5.17}$$

Here $\text{nbhd}(\boldsymbol{x}^*)$ is a local neighborhood of $\boldsymbol{x}^*$. By the convexity of f and $\mathcal{D}$, for any $\alpha \in (0,1)$ we can write:

$$\begin{aligned}
f(\alpha\boldsymbol{x}^* + (1-\alpha)\boldsymbol{x}_g^*) &\leq \alpha f(\boldsymbol{x}^*) + (1-\alpha) f(\boldsymbol{x}_g^*) && (5.18)\\
&< \alpha f(\boldsymbol{x}^*) + (1-\alpha) f(\boldsymbol{x}^*) \qquad \text{using 5.16} && (5.19)\\
&= f(\boldsymbol{x}^*) && (5.20)
\end{aligned}$$

For sufficiently small α such that $\boldsymbol{y} = \alpha\boldsymbol{x}^* + (1-\alpha)\boldsymbol{x}_g^* \in \text{nbhd}(\boldsymbol{x}^*)$ we get:

$$f(\boldsymbol{y}) < f(\boldsymbol{x}^*) \text{ with } \boldsymbol{y} \in \text{nbhd}(\boldsymbol{x}^*) \qquad \text{using 5.18} \tag{5.21}$$

Comparing 5.17 and 5.21 we have a contradiction. Hence we must have $f(\boldsymbol{x}_g^*) \geq f(\boldsymbol{x}^*)$. However, since $\boldsymbol{x}_g^*$ is a global minimizer we must also have $f(\boldsymbol{x}_g^*) \leq f(\boldsymbol{x}^*)$. Therefore we must have $f(\boldsymbol{x}_g^*) = f(\boldsymbol{x}^*)$. In other words, the local minimizer $\boldsymbol{x}^*$ is also a global minimizer as claimed. □

Remark 5.7. Note that $\boldsymbol{x}^*$ is not necessarily equal to $\boldsymbol{x}_g^*$ in proposition 5.6. It is quite possible that $\boldsymbol{x}^* \neq \boldsymbol{x}_g^*$ but at the same time the convexity of f and $\mathcal{D}$ will imply that $f(\boldsymbol{x}_g^*) = f(\boldsymbol{x}^*)$.

6 Derivation of the LASSO "shooting algorithm"

In this section, we present a simple derivation of the "shooting algorithm". First, we consider the case of single variable optimization, i.e., when $p = 1$. Next, we show how this simple case can be applied to the multi parameter situation via the "shooting algorithm".

6.1 Single variable case: $p = 1$

The optimization problem 4.1 is non-smooth because of the presence of the L_1 penalty term. We can convert this problem into a smooth one by introducing a new scalar variable t. The next proposition establishes the link between the two optimization problems.

Proposition 6.1. *Suppose $\beta \in \mathbf{R}$ is a scalar and $\boldsymbol{x}$ and $\boldsymbol{y}$ are $n \times 1$ vectors. Consider the 1-D optimization problem*

$$min_\beta \ h(\beta) = \frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta||_2^2 + \lambda|\beta|, \ \textit{where } \lambda \geq 0 \tag{6.1}$$

Suppose β_1^* is the solution to 6.1. Consider another 1-D optimization problem:

$$min_\beta \ \bar{h}(\beta) = \frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta||_2^2 + \lambda t, \ \textit{where } \lambda \geq 0 \tag{6.2}$$

$$t - \beta \geq 0 \tag{6.3}$$

$$t + \beta \geq 0 \tag{6.4}$$

Suppose (β^*, t^*) is the solution to 6.2. Then $\beta^* = \beta_1^*$.

Proof. By proposition 5.5 the objective function in 6.1 is convex. Suppose (t_1, β_1) and (t_2, β_2) satisfy the constraints in 6.2. Then $t_1 - \beta_1 \geq 0$ and $t_1 + \beta_1 \geq 0$. Also $t_2 - \beta_2 \geq 0$ and $t_2 + \beta_2 \geq 0$. Now let $\alpha \in (0, 1)$ and let $t = \alpha t_1 + (1-\alpha)t_2$ and $\beta = \alpha\beta_1 + (1-\alpha)\beta_2$. Then $t - \beta = \alpha(t_1 - \beta_1) + (1-\alpha)(t_2 - \beta_2) \geq 0$. Similarly, $t + \beta = \alpha(t_1 + \beta_1) + (1-\alpha)(t_2 + \beta_2) \geq 0$. Hence (t, β) also satisfy the constraints. This implies that the constraints define a convex set. The Hessian of the objective function in 6.2 is $H(\beta, t) = \begin{pmatrix} \boldsymbol{x}^T\boldsymbol{x} & 0 \\ 0 & 0 \end{pmatrix}$. Clearly, this is positive semidefinite. Hence by proposition 5.3, the optimization problem in 6.2 is convex. Hence by proposition 5.6 any local minimizer of 6.1 or 6.2 is also a global minimizer.

Since (β^*, t^*) is the local (and hence global) solution of 6.2, for all (β, t) such that $t - \beta \geq 0$ and $t + \beta \geq 0$ we can write:

$$\frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta^*||_2^2 + \lambda t^* \leq \frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta||_2^2 + \lambda t \tag{6.5}$$

In particular, $\beta = \beta_1^*$ and $t = |\beta_1^*|$ satisfy the constraints in 6.2 and hence we can write:

$$\frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta^*||_2^2 + \lambda t^* \leq \frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta_1^*||_2^2 + \lambda|\beta_1^*| \tag{6.6}$$

Since β_1^* is a global minimizer of 6.1 we can write:

$$\tfrac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta_1^*||_2^2 + \lambda|\beta_1^*| \leq \tfrac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta^*||_2^2 + \lambda|\beta^*| \tag{6.7}$$

Adding 6.6 and 6.7 and simplifying we get:

$$t^* \leq |\beta^*| \tag{6.8}$$

But t^* satisfies $t^* \geq \beta^*$ and $t^* \geq -\beta^*$ i.e., $t^* \geq |\beta^*|$. From 6.8 we must therefore have:

$$t^* = |\beta^*| \tag{6.9}$$

Substituting 6.9 in 6.6 we get:

$$\tfrac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta^*||_2^2 + \lambda|\beta^*| \leq \tfrac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta_1^*||_2^2 + \lambda|\beta_1^*| \tag{6.10}$$

From 6.10 and 6.7 we must have:

$$\tfrac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta^*||_2^2 + \lambda|\beta^*| = \tfrac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta_1^*||_2^2 + \lambda|\beta_1^*| \tag{6.11}$$

Expanding we get:

$$\tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{y} + \tfrac{1}{2}(\beta^*)^2\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{y}^T\boldsymbol{x}\beta^* + \lambda|\beta^*| = \tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{y} + \tfrac{1}{2}(\beta_1^*)^2\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{y}^T\boldsymbol{x}\beta_1^* + \lambda|\beta_1^*| \tag{6.12}$$

Case 1: $\lambda = 0$: In this case, $\frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta_1^*||_2^2 + \lambda|\beta_1^*| = \frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta_1^*||_2^2$ which is minimized for $\beta_1^* = \boldsymbol{y}^T\boldsymbol{x}/\boldsymbol{x}^T\boldsymbol{x}$. Similarly $\frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta^*||_2^2 + \lambda t^* = \frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta^*||_2^2$ which is minimized for $\beta^* = \boldsymbol{y}^T\boldsymbol{x}/\boldsymbol{x}^T\boldsymbol{x}$. Hence, in this case we have $\beta^* = \beta_1^* = \boldsymbol{y}^T\boldsymbol{x}/\boldsymbol{x}^T\boldsymbol{x}$.

Case 2: $\lambda \neq 0$ and $\boldsymbol{y}^T\boldsymbol{x} = 0$: In this case, $\frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta_1^*||_2^2 + \lambda|\beta_1^*| = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} + \frac{1}{2}(\beta_1^*)^2\boldsymbol{x}^T\boldsymbol{x} + \lambda|\beta_1^*|$ which is minimized for $\beta_1^* = 0$. Similarly, $\frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta^*||_2^2 + \lambda t^* = \frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta^*||_2^2 + \lambda|\beta^*| = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} + \frac{1}{2}(\beta^*)^2\boldsymbol{x}^T\boldsymbol{x} + \lambda|\beta^*|$ which is minimized for $\beta^* = 0$. Hence, in this case we have $\beta^* = \beta_1^* = 0$.

Case 3: $\lambda \neq 0$ and $\boldsymbol{y}^T\boldsymbol{x} \neq 0$: Equation 6.12 holds for all values of λ, $\boldsymbol{x}$ and $\boldsymbol{y}$. Equating the terms containing λ we must have:

$$|\beta^*| = |\beta_1^*| \tag{6.13}$$

Equation 6.13 already ensures that $\frac{1}{2}(\beta^*)^2\boldsymbol{x}^T\boldsymbol{x} = \frac{1}{2}(\beta_1^*)^2\boldsymbol{x}^T\boldsymbol{x}$. Equating the coefficient of $\boldsymbol{y}^T\boldsymbol{x}$ on both sides of 6.12 we get:

$$-\boldsymbol{y}^T\boldsymbol{x}\beta^* = -\boldsymbol{y}^T\boldsymbol{x}\beta_1^* \tag{6.14}$$

Since $\boldsymbol{y}^T\boldsymbol{x} \neq 0$, we must have $\beta^* = \beta_1^*$ which is consistent with 6.13.

Hence in all cases, we have $\beta^* = \beta_1^*$ as claimed.

□

Proposition 6.2. *Consider another 1-D optimization problem:*

$$min_\beta \ \bar{h}(\beta) = \tfrac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta||_2^2 + \lambda t, \ \textit{where } \lambda \geq 0 \tag{6.15}$$

$$t - \beta \geq 0 \tag{6.16}$$

$$t + \beta \geq 0 \tag{6.17}$$

Suppose $\boldsymbol{x} \neq \mathbf{0}$ and suppose (β^*, t^*) is the solution to 6.15. Then β^* is given by:

$$\beta^* = \begin{cases} \frac{(\boldsymbol{y}^T\boldsymbol{x} - \lambda)}{\boldsymbol{x}^T\boldsymbol{x}} & \textit{if } \boldsymbol{y}^T\boldsymbol{x} - \lambda > 0, \\ \\ \frac{(\boldsymbol{y}^T\boldsymbol{x} + \lambda)}{\boldsymbol{x}^T\boldsymbol{x}} & \textit{if } \boldsymbol{y}^T\boldsymbol{x} + \lambda < 0, \\ \\ 0 & \textit{if } -\lambda \leq \boldsymbol{y}^T\boldsymbol{x} \leq \lambda. \end{cases} \tag{6.18}$$

Proof. The Lagrangian for the optimization problem 6.15 is:

$$\mathcal{L}(\beta, t, \lambda_1, \lambda_2) = \tfrac{1}{2}\,||\boldsymbol{y} - \boldsymbol{x}\beta||_2^2 + \lambda t - \lambda_1(t - \beta) - \lambda_2(t + \beta) \tag{6.19}$$

The Karush-Kuhn-Tucker (KKT) necessary conditions of optimality for (β^*, t^*) are:

$$\begin{aligned} &\frac{\partial \mathcal{L}}{\partial \beta} = 0 \implies \beta\,\boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{x} + \lambda_2 - \lambda_1 \\ &\frac{\partial \mathcal{L}}{\partial t} = 0 \implies \lambda_1 + \lambda_2 = \lambda \\ &\left.\begin{aligned} t - \beta \geq 0 \\ t + \beta \geq 0 \end{aligned}\right\} \quad \text{Inequality constraints} \\ &\left.\begin{aligned} \lambda_1 \geq 0 \\ \lambda_2 \geq 0 \end{aligned}\right\} \quad \text{Positivity of } \lambda_1, \lambda_2 \\ &\left.\begin{aligned} \lambda_1(t - \beta) = 0 \\ \lambda_2(t + \beta) = 0 \end{aligned}\right\} \quad \text{Complementarity constraints} \end{aligned} \tag{6.20}$$

If $\boldsymbol{y}^T\boldsymbol{x} = 0$ then as shown in proposition 6.1 Case 1 and Case 2, $\beta^* = 0$. Thus we assume without loss of generality that $\boldsymbol{y}^T\boldsymbol{x} \neq 0$.

Case 1: $\boldsymbol{y}^T\boldsymbol{x} - \lambda > 0$: From 6.20 $\beta\,\boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{x} + \lambda_2 - \lambda_1$ and $\lambda_1 + \lambda_2 = \lambda$. Thus $\beta\,\boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{x} - \lambda + 2\lambda_2$. Since $\lambda_2 \geq 0$ (by 6.20) and $\boldsymbol{y}^T\boldsymbol{x} - \lambda > 0$ (by assumption in Case 1) we have that:

$$\beta\,\boldsymbol{x}^T\boldsymbol{x} = (\boldsymbol{y}^T\boldsymbol{x} - \lambda) + 2\lambda_2 > 0 \tag{6.21}$$

Since $\boldsymbol{x} \neq \mathbf{0}$ we must have $\beta > 0$. Also, adding the inequality constraints in 6.20 we have $t \geq 0$. Hence in Case 1, we must have $(t + \beta) > 0$. Hence the complementarity constraints in 6.20 imply that $\lambda_2 = 0$. Hence from 6.21 we have:

$$\beta = \frac{(\boldsymbol{y}^T\boldsymbol{x} - \lambda)}{\boldsymbol{x}^T\boldsymbol{x}} \tag{6.22}$$

Case 2: $\boldsymbol{y}^T\boldsymbol{x} + \lambda < 0$: From 6.20 $\beta\,\boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{x} + \lambda_2 - \lambda_1$ and $\lambda_1 + \lambda_2 = \lambda$. Thus $\beta\,\boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{x} + \lambda - 2\lambda_1$. Since $\lambda_1 \geq 0$ (by 6.20) and $\boldsymbol{y}^T\boldsymbol{x} + \lambda < 0$ (by assumption in Case 2) we have that:

$$\beta\,\boldsymbol{x}^T\boldsymbol{x} = (\boldsymbol{y}^T\boldsymbol{x} + \lambda) - 2\lambda_1 < 0 \tag{6.23}$$

Since $\boldsymbol{x} \neq \mathbf{0}$ we must have $\beta < 0$. Since $t \geq |\beta| \geq 0$, in Case 2, we must have $(t - \beta) > 0$. Hence the complementarity constraints in 6.20 imply that $\lambda_1 = 0$. Hence from 6.23 we have that:

$$\beta = \frac{(\boldsymbol{y}^T\boldsymbol{x} + \lambda)}{\boldsymbol{x}^T\boldsymbol{x}} \tag{6.24}$$

Case 3: $-\lambda \leq \boldsymbol{y}^T\boldsymbol{x} \leq \lambda$: If $\beta > 0$ then $(t+\beta) > 0$ which implies $\lambda_2 = 0$ (complementarity) and as in 6.22 $\beta = \frac{(\boldsymbol{y}^T\boldsymbol{x} - \lambda)}{\boldsymbol{x}^T\boldsymbol{x}}$. However $\boldsymbol{y}^T\boldsymbol{x} - \lambda \leq 0$ in Case 3 which means $\beta \leq 0$ which is a contradiction.

Similarly, if $\beta < 0$ then $(t - \beta) > 0$ which implies $\lambda_1 = 0$ (complementarity) and as in 6.24 $\beta = \frac{(\boldsymbol{y}^T\boldsymbol{x} + \lambda)}{\boldsymbol{x}^T\boldsymbol{x}}$. By assumption in Case 3 $\boldsymbol{y}^T\boldsymbol{x} + \lambda \geq 0$ which means $\beta \geq 0$ which is a contradiction.

The only way to avoid contradiction is to choose $\beta = 0$ which leads to the following valid selection of lagrange multipliers:

$$\lambda_1 = \frac{\lambda + \boldsymbol{y}^T\boldsymbol{x}}{2} \geq 0 \tag{6.25}$$

$$\lambda_2 = \frac{\lambda - \boldsymbol{y}^T\boldsymbol{x}}{2} \geq 0 \tag{6.26}$$

It can be checked that $\beta = 0$, $t = 0$ and λ_1, λ_2 as given in 6.25 satisfy the all the KKT conditions of optimality in 6.20. Hence in all cases, β^* is given by 6.18 as claimed. $\square$

6.2 Multiple variable case: $p > 1$

In this section we describe the co-ordinate wise optimization approach of Fu [1998] which is also known as the "shooting algorithm" and show that it converges to the global minimum of the LASSO objective function.

The LASSO objective function is a sum of two convex functions one of which is non-differentiable. However, the non-differentiable part is separable in the individual co-ordinate wise components. As shown in Tseng [1988], for optimization problems with this structure, the co-ordinate wise optimization approach converges to a global minimum. This same property also holds in the case of blockwise co-ordinate optimization as shown in Tseng [2001]. As discussed in Friedman et al. [2007], a similar co-ordinate wise approach can also be applied to other methods related to LASSO such as the "elastic net".

Proposition 6.3. *Consider the LASSO optimization problem:*

$$min_{\boldsymbol{\beta}}\ h(\boldsymbol{\beta}) = \frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1,\ \ where\ \lambda \geq 0 \tag{6.27}$$

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p]$, $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^T$, $\boldsymbol{X}^{(-i)} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_p]$ *and* $\boldsymbol{\beta}^{(-i)} = [\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_p]^T$. *Consider the following solution approach:*

- *Initialize* $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ *(using for instance least squares, regularized least squares or randomly)*
- *For* $k = 0, 1, 2, \ldots, m$ *repeat*
  - *Compute* $f_k = h(\boldsymbol{\beta})$.
  - *For* $i = 1, 2, \ldots, p$
    1. *Using the current value of* $\boldsymbol{\beta}^{(-i)}$ *solve the following 1-D optimization problem w.r.t.* β_i
    $$min_{\beta_i}\ h^{'}(\beta_i) = \frac{1}{2}\,||\boldsymbol{y}_i - \boldsymbol{x}_i\beta_i||_2^2 + \lambda|\beta_i| + \lambda||\boldsymbol{\beta}^{(-i)}||_1 \tag{6.28}$$
    where
    $$\boldsymbol{y}_i = \boldsymbol{y} - \boldsymbol{X}^{(-i)}\boldsymbol{\beta}^{(-i)} \tag{6.29}$$
    2. *Suppose* β_i^* *is the solution to 6.28 then update the ith element of* $\boldsymbol{\beta}$ *to be equal to* β_i^* *i.e., set* $\beta_i = \beta_i^*$

Then the sequence of iterates $f_1, f_2, \ldots, f_m$ *converge to the co-ordinate wise minimum of* $h(\boldsymbol{\beta})$ *in 6.27 as* $m \to \infty$.

Proof. It is easy to see that:

$$h(\boldsymbol{\beta}) = \frac{1}{2}\,||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 = \frac{1}{2}\,||\boldsymbol{y}_i - \boldsymbol{x}_i\beta_i||_2^2 + \lambda|\beta_i| + \lambda||\boldsymbol{\beta}^{(-i)}||_1 \tag{6.30}$$

where $\boldsymbol{y}_i$ is defined in 6.29. If β_i^* solves the convex optimization problem 6.28 then we must have:

$$\tfrac{1}{2}\,||\boldsymbol{y}_i - \boldsymbol{x}_i\beta_i^*||_2^2 + \lambda|\beta_i^*| + \lambda||\boldsymbol{\beta}^{(-i)}||_1 \leq \tfrac{1}{2}\,||\boldsymbol{y}_i - \boldsymbol{x}_i\beta_i||_2^2 + \lambda|\beta_i| + \lambda||\boldsymbol{\beta}^{(-i)}||_1 = h(\boldsymbol{\beta}) \tag{6.31}$$

If $\boldsymbol{\beta}_{new}$ is the new vector obtained by updating the ith component of $\boldsymbol{\beta}$ to be equal to β_i^* then we can re-write 6.31 as:

$$h(\boldsymbol{\beta}_{new}) \leq h(\boldsymbol{\beta}) \tag{6.32}$$

Hence we see that every iteration in the inner for loop ($i = 1, 2, \ldots, p$) decreases the objective function. This implies that:

$$f_{k+1} \leq f_k \text{ for all } k \tag{6.33}$$

In addition f_k is bounded below by 0 i.e., $f_k \geq 0$ for all k. Suppose $\hat{f}$ is the greatest lower bound on the sequence $\{f_k\}$. Then $\hat{f} \leq f_k$ for all k. Choose any $\varepsilon > 0$. Then

$$f_k + \varepsilon > \hat{f} \tag{6.34}$$

Also $\hat{f} + \varepsilon$ is not the greatest lower bound. Hence there exists n_0 such that

$$f_{n_0} < \hat{f} + \varepsilon \tag{6.35}$$

Since $k > n_0$ implies $f_k \leq f_{n_0}$ we conclude that:

$$f_k \leq f_{n_0} < \hat{f} + \varepsilon \text{ if } k > n_0 \tag{6.36}$$

Hence for all $k > n_0$ we have:

$$\hat{f} - \varepsilon < f_k < \hat{f} + \varepsilon \tag{6.37}$$

In other words, the sequence $\{f_k\}$ converges to $\hat{f}$. If we cycle through all the co-ordinate directions until convergence then $\hat{f}$ will be the co-ordinate wise minimum of $h(\boldsymbol{\beta})$.

□

Proposition 6.4. *Suppose $\hat{\boldsymbol{\beta}}$ is the co-ordinate wise minimum of $h(\boldsymbol{\beta})$:*

$$h(\hat{\boldsymbol{\beta}} + \delta_i \boldsymbol{e}_i) \geq h(\hat{\boldsymbol{\beta}}) \text{ where } \delta_i \neq 0 \tag{6.38}$$

and $\boldsymbol{e}_i$ is a vector with a 1 at position i and zeros elsewhere. Then for any vector $\boldsymbol{p}$ in some open neighborhood of $\hat{\boldsymbol{\beta}}$:

$$h(\hat{\boldsymbol{\beta}} + \boldsymbol{p}) \geq h(\hat{\boldsymbol{\beta}}) \tag{6.39}$$

i.e., $\hat{\boldsymbol{\beta}}$ is a local minimizer of $h(\boldsymbol{\beta})$. Since $h(\boldsymbol{\beta})$ is convex this implies that $\hat{\boldsymbol{\beta}}$ is also a global minimizer.

Proof. Recall that we can write the LASSO objective as:

$$h(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + g(\boldsymbol{\beta}) \tag{6.40}$$

where

$$f(\boldsymbol{\beta}) = \tfrac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \tag{6.41}$$

$$g(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1 = \lambda \sum_{i=1}^{p} |\beta_i| \tag{6.42}$$

Hence we can write:

$$
\begin{aligned}
h(\hat{\boldsymbol{\beta}} + \boldsymbol{p}) &= f(\hat{\boldsymbol{\beta}} + \boldsymbol{p}) + g(\hat{\boldsymbol{\beta}} + \boldsymbol{p}) && (6.43)\\
&= f(\hat{\boldsymbol{\beta}}) + \boldsymbol{p}^T \nabla f(\hat{\boldsymbol{\beta}}) + \tfrac{1}{2}\boldsymbol{p}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{p} + \lambda \sum_{i=1}^{p} |\hat{\beta}_i + p_i| && (6.44)\\
&= f(\hat{\boldsymbol{\beta}}) + \lambda \sum_{i=1}^{p} |\hat{\beta}_i| + \boldsymbol{p}^T \nabla f(\hat{\boldsymbol{\beta}}) + \tfrac{1}{2}\boldsymbol{p}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{p} + \lambda \sum_{i=1}^{p} |\hat{\beta}_i + p_i| - \lambda \sum_{i=1}^{p} |\hat{\beta}_i| && (6.45)\\
&= f(\hat{\boldsymbol{\beta}}) + g(\hat{\boldsymbol{\beta}}) + \boldsymbol{p}^T \nabla f(\hat{\boldsymbol{\beta}}) + \tfrac{1}{2}\boldsymbol{p}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{p} + \lambda \sum_{i=1}^{p} |\hat{\beta}_i + p_i| - \lambda \sum_{i=1}^{p} |\hat{\beta}_i| && (6.46)\\
&= h(\hat{\boldsymbol{\beta}}) + \boldsymbol{p}^T \nabla f(\hat{\boldsymbol{\beta}}) + \tfrac{1}{2}\boldsymbol{p}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{p} + \lambda \sum_{i=1}^{p} |\hat{\beta}_i + p_i| - \lambda \sum_{i=1}^{p} |\hat{\beta}_i| && (6.47)\\
& && (6.48)
\end{aligned}
$$

Let $\boldsymbol{p} = \delta_i \boldsymbol{e}_i$ in 6.43 with $\delta_i \neq 0$ then we can write:

$$h(\hat{\boldsymbol{\beta}} + \delta_i \boldsymbol{e}_i) = h(\hat{\boldsymbol{\beta}}) + \delta_i \boldsymbol{e}_i^T \nabla f(\hat{\boldsymbol{\beta}}) + \tfrac{1}{2}\delta_i^2 \boldsymbol{e}_i^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{e}_i + \lambda|\hat{\beta}_i + \delta_i| - \lambda|\hat{\beta}_i| \tag{6.49}$$

By assumption $h(\hat{\boldsymbol{\beta}} + \delta_i \boldsymbol{e}_i) \geq h(\hat{\boldsymbol{\beta}})$ and so we must have:

$$\delta_i \boldsymbol{e}_i^T \nabla f(\hat{\boldsymbol{\beta}}) + \tfrac{1}{2}\delta_i^2 \boldsymbol{e}_i^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{e}_i + \lambda|\hat{\beta}_i + \delta_i| - \lambda|\hat{\beta}_i| \geq 0 \tag{6.50}$$

The above relationship holds for all δ_i not matter how small. By choosing $|\delta_i|$ sufficiently small, we can make the term $\frac{1}{2}\delta_i^2 \boldsymbol{e}_i^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{e}_i$ arbitrarily close to 0. Hence there exists $\theta_i > 0$ such that for all $\delta_i \in (-\theta_i, \theta_i)$ the following holds:

$$\delta_i \boldsymbol{e}_i^T \nabla f(\hat{\boldsymbol{\beta}}) + \lambda|\hat{\beta}_i + \delta_i| - \lambda|\hat{\beta}_i| \geq 0 \tag{6.51}$$

Now let

$$\boldsymbol{p} = \sum_{i=1}^{p} \delta_i \boldsymbol{e}_i \tag{6.52}$$

then from 6.43 we get:

$$h(\hat{\boldsymbol{\beta}} + \boldsymbol{p}) = h(\hat{\boldsymbol{\beta}}) + \sum_{i=1}^{p} \delta_i \boldsymbol{e}_i^T \nabla f(\hat{\boldsymbol{\beta}}) + \tfrac{1}{2}\boldsymbol{p}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{p} + \lambda \sum_{i=1}^{p} |\hat{\beta}_i + \delta_i| - \lambda \sum_{i=1}^{p} |\hat{\beta}_i| \tag{6.53}$$

Note that 6.51 implies:

$$\sum_{i=1}^{p} \delta_i \boldsymbol{e}_i^T \nabla f(\hat{\boldsymbol{\beta}}) + \lambda \sum_{i=1}^{p} |\hat{\beta}_i + \delta_i| - \lambda \sum_{i=1}^{p} |\hat{\beta}_i| \geq 0 \tag{6.54}$$

Therefore from 6.53 and 6.54 we must have:

$$\begin{aligned} h(\hat{\boldsymbol{\beta}} + \boldsymbol{p}) &= h(\hat{\boldsymbol{\beta}}) + \sum_{i=1}^{p} \delta_i \boldsymbol{e}_i^T \nabla f(\hat{\boldsymbol{\beta}}) + \tfrac{1}{2}\boldsymbol{p}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{p} + \lambda \sum_{i=1}^{p} |\hat{\beta}_i + \delta_i| - \lambda \sum_{i=1}^{p} |\hat{\beta}_i| && (6.55) \\ &\geq h(\hat{\boldsymbol{\beta}}) + \tfrac{1}{2}\boldsymbol{p}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{p} && (6.56) \\ &\geq h(\hat{\boldsymbol{\beta}}) \qquad \text{by the positive semi-definiteness of } \boldsymbol{X}^T \boldsymbol{X} && (6.57) \end{aligned}$$

In other words, we have found an open neighborhood with $\delta_i \in (-\theta_i, \theta_i), \theta_i > 0$ such that for all $\boldsymbol{p}$ of the form 6.52, $h(\hat{\boldsymbol{\beta}} + \boldsymbol{p}) \geq h(\hat{\boldsymbol{\beta}})$. This implies that the co-ordinate wise minimizer $\hat{\boldsymbol{\beta}}$ is actually a local minimizer (and hence by convexity a global minimizer) of $h(\boldsymbol{\beta})$. □

7 How to choose λ?

The L_1 regularization parameter for LASSO can be chosen using cross validation. In brief, given data $(\boldsymbol{X}, \boldsymbol{y})$, we partition the rows of $\boldsymbol{X}$ and $\boldsymbol{y}$ into K parts giving us K data/response pairs: $(\boldsymbol{X}_1, \boldsymbol{y}_1), (\boldsymbol{X}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{X}_K, \boldsymbol{y}_K)$. Let $\left(\boldsymbol{X}^{(-i)}, \boldsymbol{y}^{(-i)}\right)$ be the data/response pair obtained by deleting the ith part $(\boldsymbol{X}_i, \boldsymbol{y}_i)$ from $(\boldsymbol{X}, \boldsymbol{y})$. Let $\boldsymbol{\beta}_{lasso}^{(-i)}$ be the LASSO solution obtained using $\left(\boldsymbol{X}^{(-i)}, \boldsymbol{y}^{(-i)}\right)$. Let n_i be the number of data points in the ith data/response pair $(\boldsymbol{X}_i, \boldsymbol{y}_i)$. For a given value of λ define the average cross validated mean squared error as:

$$\overline{CV}_{MSE}(\lambda) = \tfrac{1}{K} \sum_{i=1}^{K} \frac{1}{n_i} \left\| \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_{lasso}^{(-i)} \right) \right\|_2^2 \tag{7.1}$$

Given a range of palusible values for λ we choose the optimal λ as the one that minimizes the average cross validated mean squared error:

$$\lambda^* = \arg\min_\lambda \overline{CV}_{MSE}(\lambda) \tag{7.2}$$

Figure 1 shows the process of choosing λ for an example data set using 10-fold cross-validation. Figure 2 shows the optimal LASSO fit using λ^* from Figure 1 as well as the estimated LASSO coefficients.

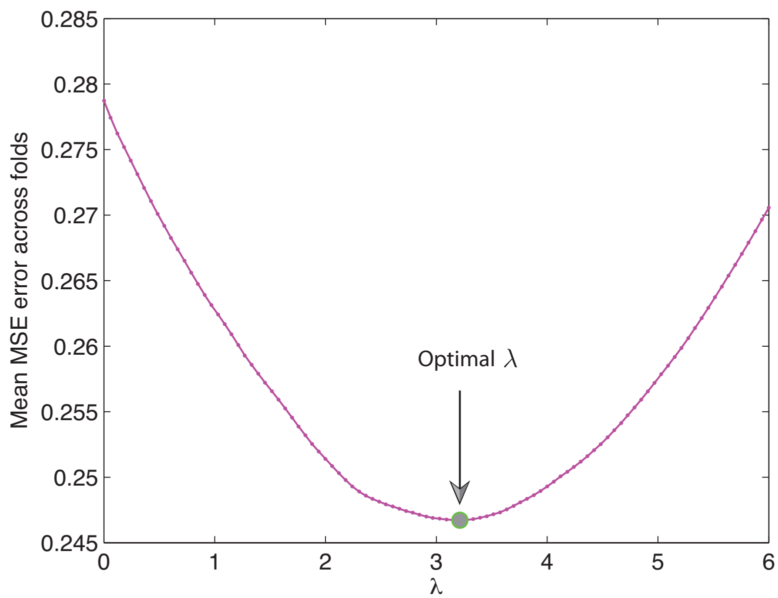
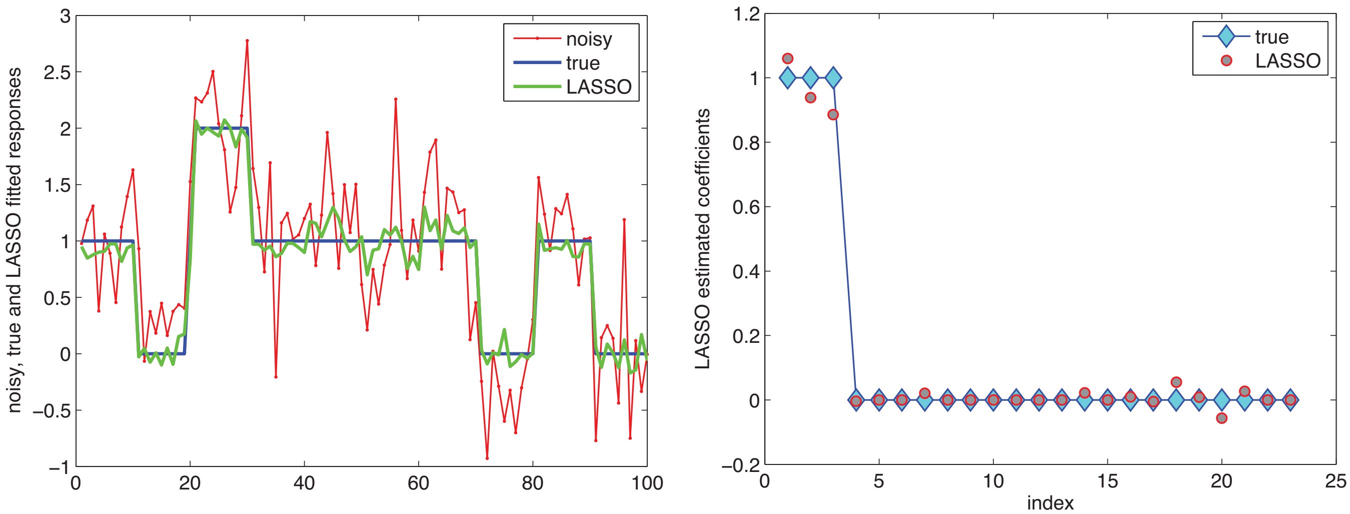


Figure 1: Average mean squared error across cross-validation folds (10-fold cross-validation) versus the regularization parameter λ for an example data set. Arrow shows the location of λ^*.

8 Conclusions

This goal of this tutorial was to provide a simple yet self-contained introduction to the LASSO [Tibshirani, 1996] technique for L_1 regularized linear regression. We discussed an efficient algorithm for optimizing the LASSO objective function - the "shooting algorithm" of Fu [1998]. From a practical point of view, we suggest a cross-validation based approach for choosing the regularization parameter λ. We encourage the reader to learn more about LASSO by visiting Rob Tibshirani's LASSO page: `http://www-stat.stanford.edu/~tibs/lasso.html`.

MATLAB code for estimating a LASSO model along with example data can be downloaded from: `http://www.gautampendse.com/software/lasso/webpage/lasso_shooting.html`.



(a) (b)

Figure 2: (a) Overlay of noisy data, true data and the LASSO fit obtained using λ^* from Figure 1 (b) The true coefficients versus the LASSO estimated coefficients using λ^* from Figure 1.

References

J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

W. J. Fu. Penalized Regressions: The Bridge Versus the Lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

P. Tseng. Coordinate Ascent for Maximizing Nondifferentiable Concave Functions. Technical Report LIDS-P-1840, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 1988.

P. Tseng. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.