1. Review of Convex Analysis

**Definition 1.1.** Let $C$ be a subset of $\mathbb{R}^n$. We say $C$ is convex if

$$ax + (1 - \alpha)y \in C, \quad \forall x, y \in C, \forall \alpha \in [0, 1].$$

**Definition 1.2.** Let $C$ be a convex subset of $\mathbb{R}^n$. A function $f : C \mapsto \mathbb{R}$ is called convex if

$$f(ax + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y), \quad \forall x, y \in C, \forall \alpha \in [0, 1].$$

The function $f$ is called concave if $-f$ is convex. The function $f$ is called strictly convex if the above inequality is strict for all $x, y \in C$ with $x \neq y$, and all $\alpha \in (0, 1)$.

**Definition 1.3.** Given a convex function $f : \mathbb{R}^n \mapsto \mathbb{R}$, we say that a vector $d \in \mathbb{R}^n$ is a subgradient of $f$ at a point $x \in \mathbb{R}^n$ if

$$f(z) \geq f(x) + (z - x)^T d, \quad \forall z \in \mathbb{R}^n.$$ 

The set of all subgradients of a convex function $f$ at $x \in \mathbb{R}^n$ is called the subdifferential of $f$ at $x$, and is denoted by $\partial f(x)$.

**Proposition 1.1.** Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be convex. For every $x \in \mathbb{R}^n$, the following hold.

(a) A vector $d \in \mathbb{R}^n$ is a subgradient of $f$ at $x$ if and only if

$$\lim_{\alpha \downarrow 0} f(x + \alpha y) - f(x) \geq d^T y, \quad \forall y \in \mathbb{R}^n.$$ 

(b) If $f$ is equal to the sum $f_1 + \cdots + f_m$ of convex functions $f_j : \mathbb{R}^n \mapsto \mathbb{R}$, $j = 1, \ldots, m$, then $\partial f(x)$ is equal to the vector sum $\partial f_1(x) + \cdots + \partial f_m(x)$.

(c) $x$ minimizes $f$ over a convex set $C \subset \mathbb{R}^n$ if and only if there exists a subgradient $d \in \partial f$ such that

$$d^T(z - x) \geq 0 \quad \forall z \in C.$$

2. Duality

Consider the problem

$$\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in X, \ g_j(x) \leq 0, \ j = 1, \ldots, r,
\end{align*}
$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $g_j : \mathbb{R}^n \mapsto \mathbb{R}$ are given functions, and $X$ is a subset of $\mathbb{R}^n$. $f$ is called the cost function and $g_j$ are called constraint functions. We also denote $g(x) = (g_1(x), \ldots, g_r(x))^T$ and write the constraints $g_j(x) \leq 0$ compactly as $g(x) \leq 0$. We refer to this problem as the primal problem and we denote by $f^*$ its optimal value:

$$f^* = \inf_{x \in X} f(x).$$

We shall make the following assumption.

**Assumption 2.1** (Feasibility and Boundedness). There exists at least one feasible solution for the primal problem and the cost is finite, i.e. $-\infty < f^* < \infty$. 

---

*This is a supplementary reading for Stat588. Last updated on September 21, 2011.*
We will give a brief review of Lagrange multiplier theory and the notion of duality. For more intuitions and details we recommend Chapter 5 of Bertsekas (1999). Define the *Lagrangian function* $f : \mathbb{R}^{n+r} \mapsto \mathbb{R}$ as

$$L(x, \mu) = f(x) + \sum_{j=1}^{r} \mu_j g_j(x) = f(x) + \mu^T g(x),$$

where $\mu := (\mu_1, \ldots, \mu_r)^T \in \mathbb{R}^r$. To understand the following definition of *Lagrange multiplier*, you may want to draw intuition from the method of Lagrange multiplier from calculus.

**Definition 2.1.** A vector $\mu^* = (\mu^*_1, \ldots, \mu^*_r)$ is said to be a *Lagrange multiplier* for the primal problem (2.1) if

$$\mu^*_j \geq 0, \quad j = 1, \ldots, r \quad \text{and} \quad f^* = \inf_{x \in X} L(x, \mu^*).$$

**Proposition 2.1.** Let $\mu^*$ be a Lagrange multiplier. Then $x^*$ is a global minimum of the primal problem (2.1) if and only if $x^* \in X$ and $x^* = \arg\min_{x \in X} L(x, \mu^*)$, and $\mu^*_j g_j(x^*) = 0, \quad j = 1, \ldots, r.$

We introduce the *dual function* $q$ defined for $\mu \in \mathbb{R}^r$ by

$$q(\mu) = \inf_{x \in X} L(x, \mu).$$

The *dual problem* is

$$\begin{align*}
\text{maximize} & \quad q(\mu) \\
\text{subject to} & \quad \mu \in D, \quad \mu \geq 0,
\end{align*}$$

where $D$ is the *domain* of $q$, defined as the set of $\mu$ for which $q(\mu)$ is finite:

$$D = \{ \mu : q(\mu) > -\infty \}.$$ 

Note that $D$ might be empty. We denote by $q^*$ the optimal value of the dual problem (2.2). The dual problem has nice convexity properties, as shown by the following proposition.

**Proposition 2.2.** The domain $D$ of the dual function $q$ is convex and $q$ is concave over $D$.

An important property is that the optimal dual value is always an underestimate of the optimal primal value.

**Proposition 2.3** (*Weak Duality Theorem*). We have $q^* < f^*$.

If $q^* = f^*$ we say there is no duality gap, and if $q^* < f^*$ we say there is a duality gap.

**Proposition 2.4.** We have the following two cases.

(a) If there is no duality gap, the set of Lagrange multipliers is equal to the set of optimal dual solutions.

(b) If there is a duality gap, the set of Lagrange multipliers is empty.

If there is no duality gap, the following proposition provides a characterization of primal and dual optimal solution pairs.

**Proposition 2.5.** $(x^*, \mu^*)$ is an optimal solution-Lagrange multiplier pair if and only if

$$\begin{align*}
x^* & \in X, \quad g(x^*) \leq 0, \quad \text{(Primal Feasibility)}, \\
\mu^* & \geq 0, \quad \text{(Dual Feasibility)}, \\
x^* & = \arg\min_{x \in X} L(x, \mu^*), \quad \text{(Lagrangian Optimality)}, \\
\mu^*_j g_j(x^*) & = 0, \quad j = 1, \ldots, r, \quad \text{(Complementary Slackness)}.
\end{align*}$$

We now consider the primal problem (2.1) under convexity assumptions.

**Assumption 2.2** (*Convexity and Interior Point*). The set $X$ is a convex subset of $\mathbb{R}^n$ and the functions $f$ and $g_j$ are convex over $X$. In addition, there exists a vector $\bar{x} \in X$ such that

$$g_j(\bar{x}) < 0, \quad \forall j = 1, \ldots, r.$$  

(2.3)

The condition (2.3) is called the *interior point* condition.
The following theorem will be used to explain the duality of LASSO in Section 3

**Theorem 2.1 (Strong Duality Theorem).** Let Assumption 2.1 and Assumption 2.2 hold for the primal problem (2.1). Then there is no duality gap and there exists at least one Lagrange multiplier.

3. LASSO AND ITS DUAL

We adopt notations from the textbook ESL. We have an generic input vector $X = (X_1, \ldots, X_p)^T$, and want to predict a generic real-valued output $Y$. Consider the linear regression model with the predictor

$$f(X) = \beta_0^* + \sum_{j=1}^{p} X_j \beta_j^*.$$  \hfill (3.1)

Typically we have a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ from which to estimate the parameter $\beta^* := (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)^T$. Each $x_i = (x_{i1}, \ldots, x_{ip})^T$ is a vector of inputs for the $i$th case.

Tibshirani (1996) proposed the “least absolute shrinkage and selection operator” (LASSO), which minimizes the residual sum of squares under a constraint on the $\ell^1$-norm of the coefficient vector. The LASSO solves the optimization problem

$$\text{minimize } \frac{1}{2} \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t,$

for some $t \geq 0$. The minimizer of this problem is denoted by $\hat{\beta}_{\text{lasso}}$. We can also write the lasso problem in the equivalent Lagrangian form

$$\text{minimize } \frac{1}{2} \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

for some $\lambda \geq 0$. Using the duality theory introduced in Section 2, we see that problems (3.2) and (3.3) are equivalent in the sense that, for a given $t \geq 0$, there exists a $\lambda \geq 0$ such that the two problems share the same set of minimizers, and vice versa. For a more direct argument we recommend Osborne et al. (2000).

4. SUPPORT RECOVERY

For notational simplicity we assume the output vector $y = (y_1, \ldots, y_N)^T$ is centered, i.e. $\sum_{i=1}^{N} y_i = 0$, and we assume $x_j$, $1 \leq j \leq p$ are also centered, where $x_j = (x_{1j}, x_{2j}, x_{Nj})^T$ is the $N$-vector consisting of all observations on the $j$th feature $X_j$. Under this assumption the intercept is always estimated as $\hat{\beta}_0 = 0$, so we omit it and consider the optimization problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$  \hfill (4.1)

where $X = (x_1, x_2, \ldots, x_p)$, and $\beta = (\beta_1, \ldots, \beta_p)^T$. We also redefine $\beta^* = (\beta_1^*, \ldots, \beta_p^*)^T$ as the true coefficient vector that we want to estimate.

According to our definition of subgradient in Definition 1.3 and part (b) of the Proposition 1.1, a vector $z \in \mathbb{R}^p$ is a subgradient of the $\ell_1$ norm evaluated at $\beta \in \mathbb{R}^p$, written as $\beta \in \partial \|\beta\|_1$, if its entries satisfy the conditions

$$\begin{cases} 
    z_i = \text{sign}(\beta_i) & \text{if } \beta_i \neq 0; \\
    z_i \in [-1, 1], & \text{otherwise.}
\end{cases}$$

For any subset $A \subset \{1, \ldots, p\}$ with cardinality $|A|$, let $X_A$ be the $n \times |A|$ submatrix formed by concatenating columns $\{x_j, j \in A\}$ indexed by $A$. For any vector $\beta \in \mathbb{R}^p$, define its support set

$$S(\beta) = \{1 \leq j \leq p : \beta_j \neq 0\}.$$  \hfill (Wainwright, 2009)
(a) A vector \( \hat{\beta} \in \mathbb{R}^p \) is optimal if and only if there exists a subgradient vector \( \hat{z} \in \partial \| \hat{\beta} \|_1 \) such that

\[
X^T X (\hat{\beta} - \beta^*) - X^T (y - X \beta^*) + \lambda \hat{z} = 0.
\]

(b) Suppose that the subgradient \( \hat{z} \) satisfies the strict dual feasibility condition \( |\hat{z}_j| < 1 \) for all \( j \notin S(\hat{\beta}) \).

Then any optimal solution \( \hat{\beta} \) to the Lasso satisfies \( \hat{\beta}_j = 0 \) for all \( j \notin S(\hat{\beta}) \).

(c) Under the conditions of part (b), if the matrix \( X^T_{S(\hat{\beta})} X_{S(\hat{\beta})} \) is invertible, then \( \hat{\beta} \) is the unique optimal solution of the Lasso problem (4.1).

Contemporary data often has high dimension, where \( p \) is comparable to or much greater than \( N \). In many practical settings, it is reasonable to assume that the vector \( \beta^* \) is sparse, that is, the cardinality \( k := S(\beta^*) \) of its support satisfies \( k = o(p) \). It is desirable that the optimal solution \( \hat{\beta} \) of (4.1) has the same support as \( \beta^* \) i.e. \( S(\hat{\beta}) = S(\beta^*) \), which we refer to as support recovery. A more ambitious goal is the so called signed support recovery, which requires not only the support recovery, but also that the nonzero coefficients are estimated with correct signs. Specifically, for a real number \( a \in \mathbb{R} \), define

\[
\text{sign}(a) = \begin{cases} 
1, & \text{if } a > 0 \\
0, & \text{if } a = 0 \\
-1, & \text{if } a < 0.
\end{cases}
\]

For any vector \( \beta \in \mathbb{R}^p \), define the corresponding sign vector by

\[
S_\pm(\beta) = [\text{sign}(\beta_1), \text{sign}(\beta_2), \ldots, \text{sign}(\beta_p)]^T.
\]

Note that \( S(\beta) \) is a set, and \( S_\pm(\beta) \) is a vector, but they are closely related in the sense that \( S(\beta) \) consists of indexes of all nonzero entries of \( S_\pm(\beta) \). We say the estimate \( \hat{\beta} \) achieves signed support recovery if \( S_\pm(\hat{\beta}) = S_\pm(\beta^*) \).

The following procedure, called primal-dual witness (PDW) method, is not a practical method for solving the Lasso program (4.1), but it provides heuristics on when Lasso achieves support recovery. It consists of constructing a pair \( (\hat{\beta}, \hat{z}) \) according to the following steps. We use \( S \) as a shorthand for the support \( S(\beta^*) \) of the true coefficient vector \( \beta^* \), and assume the \( k \times k \) matrix \( X_S^T X_S \) is invertible.

1) First, we obtain \( \hat{\beta}_S \in \mathbb{R}^k \) by solving the restricted Lasso problem

\[
\hat{\beta}_S = \arg \min_{\hat{\beta}_S \in \mathbb{R}^k} \left\{ \frac{1}{2} \| y - X_S \hat{\beta}_S \|_2^2 + \lambda \| \hat{\beta}_S \|_1 \right\}.
\]

The solution to this restricted convex program is guaranteed to be unique under the invertibility condition of \( X_S^T X_S \). We set \( \hat{\beta}_{S^C} = 0 \).

2) Second, we choose \( \hat{z}_S \in \mathbb{R}^k \) as an element of the subdifferential of the \( \ell_1 \) norm evaluated at \( \hat{\beta}_S \).

3) Third, we solve for a vector \( \hat{z}_{S^C} \in \mathbb{R}^{p-k} \) satisfying the zero subgradient condition (4.2), and check whether or not the dual feasibility condition \( |\hat{z}_j| \leq 1 \) for all \( j \in S^C \) is satisfied. (For ensuring uniqueness, we check for strict dual feasibility condition \( |\hat{z}_j| < 1 \) for all \( j \in S^C \).)

4) Fourth, we check whether the sign consistency condition \( \hat{z}_S = \text{sign}(\beta_S^*) \) is satisfied.

This procedure succeeds if and only if Lasso has a unique optimal solution with the correct signed support, as summarized by the next proposition.

**Proposition 4.2** (Wainwright, 2009). Assume the \( k \times k \) matrix \( X_S^T X_S \) is invertible.

(a) If Steps 1 through 3 of the PDW method succeed with strict dual feasibility in Step 3, then the Lasso (4.1) has a unique solution \( \hat{\beta} \) with \( S(\hat{\beta}) \subset S(\beta^*) \).

(b) If Steps 1 through 4 succeeds with strict dual feasibility in Step 3, then Lasso (4.1) has a unique solution \( \hat{\beta} \) with the correct signed support, i.e. \( S_\pm(\hat{\beta}) = S_\pm(\beta^*) \).

(c) Conversely, if either Steps 3 or 4 of the PDW method fail, then the Lasso fails to recover the correct signed support.
5. Least Angle Regression

Efron et al. (2004) proposed a procedure called least angle regression, abbreviated LARS (the “S” suggesting “Lasso” or “Stagewise”), which works roughly as follows. We start with all coefficients equal to zero, and find the predictor most correlated with the response, say \(x_{j_1}\). We take the largest step possible in the direction of this predictor until some other predictor, say \(x_{j_2}\), has as much correlation with the current residual. Then LARS proceeds in a direction equiangular between the two predictors until a third variable \(x_{j_3}\) earns its way into the “most correlated” set. LARS then proceeds equiangularly among \(x_{j_1}, x_{j_2}, \) and \(x_{j_3}\), that is, along the “least angle direction”, until a forth variable enters, and so on.

Before providing details, let us make the following important assumption.

**Assumption 5.1.**

(i) The output vector \(y\) is centered, that is, \(\sum_{i=1}^{N} y_i = 0\).
(ii) Each predictor \(x_j, 1 \leq j \leq p\) is normalized, i.e.
\[
\sum_{i=1}^{N} x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{N} x_{ij}^2 = 1, \quad \forall 1 \leq j \leq p.
\]
(iii) The \(p\) input vectors \(x_1, \ldots, x_p\) are linearly independent.
(iv) The output vector \(y\) is not contained in the linear space spanned by the input vectors.

**5.1. The LARS algorithm.** LARS builds up estimates \(\hat{\beta} = X \hat{\beta}\) and \(\hat{\beta}\) in successive steps.

0) We start with \(\hat{\mu}_0 = 0, \hat{\beta}(0) = 0\) and an empty index set \(A_0 = \emptyset\), which is viewed as a subset of \(\{1, \ldots, p\}\).

1) At step \(t\) with \(t \geq 1\), we first compute the current correlations
\[
\hat{c}_t = X^T (y - \hat{\mu}_{t-1}),
\]
where the \(j\)th component of \(\hat{c}_t\), denoted by \(\hat{c}_{tj}\), is the correlation of the current residual and the \(j\)th input vector \(x_j\).

2) The active set \(A_t\) is the set of indices corresponding to inputs with the greatest absolute current correlations, that is
\[
\hat{C}_t = \max_{1 \leq j \leq p} |\hat{c}_{tj}| \quad \text{and} \quad A_t = \left\{ j : |\hat{c}_{tj}| = \hat{C}_t \right\}.
\]

3) Set \(s_j = \text{sign}(\hat{c}_{tj})\) for \(j \in A_t\), and define the matrix
\[
\tilde{X}_{A_t} := (\cdots s_j x_j \cdots)_{j \in A_t}.
\]
Let \(1_{A_t}\) be a \(|A_t|\)-dimensional vector with all entries equal to one, and
\[
B_t = \left[1_{A_t}^T (\tilde{X}_{A_t}^T \tilde{X}_{A_t})^{-1} 1_{A_t}\right]^{1/2}.
\]
LARS updates \(\hat{\mu}\) in the following equiangular direction,
\[
u_t = \tilde{X}_{A_t} w_{A_t}, \quad \text{where} \quad w_{A_t} = (w_{tj})_{j \in A_t} := B_t \left(\tilde{X}_{A_t}^T \tilde{X}_{A_t}\right)^{-1} 1_{A_t}.
\]
Observe that \(\nu_t\) is a unit vector making equal angles, less than \(90^\circ\), with the columns of the matrix \(\tilde{X}_{A_t}\). This explains the name “least angle”, because \(\nu_t\) makes the smallest (and equal) angle with each column of \(\tilde{X}_{A_t}\). Set
\[
b_t = (b_{t1}, \ldots, b_{tp})^T = X^T \nu_t.
\]

4) LARS makes the current estimate \(\hat{\mu}_{t-1}\) proceed in the direction \(\nu_t\) by a length \(\gamma_t\), and thus updates the estimates as
\[
\hat{\mu}_t = \hat{\mu}_{t-1} + \gamma_t \nu_t \quad \text{and} \quad \hat{\beta}(t) = \hat{\beta}(t-1) + \gamma_t d_t,
\]
where \(d_t \in \mathbb{R}^p\) is a vector equaling \(s_j w_{tj}\) for \(j \in A_t\) and zero elsewhere; and
\[
\gamma_t = \min\left\{ \frac{\hat{C}_t - \hat{c}_{tj}}{B_t - b_{tj}}, \frac{\hat{C}_t + \hat{c}_{tj}}{B_t + b_{tj}} : j \in A_t^c \right\},
\]


where \( \min^+ \) indicates that the minimum is taken over only finite positive elements of the candidate set.

5) Iterate with \( t \), until all \( p \) inputs have been entered. We need at most \( p \) steps.

5.2. The LARS-Lasso relationship. It turns out a simple modification of the LARS produces the same solution paths as Lasso. Let \( \hat{\beta} \) be the optimal solution of Lasso (4.1) with \( \hat{\mu} = X\hat{\beta} \). Due to the zero subgradient condition (4.2), we see that the sign of any nonzero coefficient \( \hat{\beta}_j \) must have the same sign of the correlation \( \hat{c}_j = x_j^T(y - X\hat{\beta}) \) between the current residual \( y - X\hat{\beta} \) and the \( j \)th input vector \( x_j \). In the LARS procedure, when a variable \( x_j \) comes into the model, the sign of the correlation between \( x_j \) and the residual does not change as LARS proceeds further until it reaches 0 at the final step. However, it is possible that the corresponding coefficient \( \hat{\beta}_j \) will have a sign change before the LARS terminates. Example 5.1 explains why this could happen. As a result, LARS cannot give a legitimate Lasso solution when there is a sign change, because the correlation never has a sign change. In order to produce Lasso solutions, we must prohibit such a sign change in LARS.

**Lasso modification.** If there is a sign change on \( \hat{\beta}_j \) as LARS proceeds, stop the ongoing LARS at the point the sign change happens i.e. \( \hat{\beta}_j \) reaches zero; and remove \( x_j \) from the calculation of the next equiangular direction.

**Theorem 5.1** (Efron et al. 2004). Under the Lasso modification, and assuming the “one at a time” condition discussed below, the LARS algorithm yields all Lasso solution.

For a proof of Theorem 5.1 we refer the readers to Efron et al. (2004). “One at a time” means that the increases and decreases of the active set never involve more than one index \( j \). This is usually true for quantitative data and can always be realized by adding a small amount of random noise to the output \( y \).

**Example 5.1.** This is an artificial example with \( N = 5 \) objects and \( p = 3 \) input variables. \( y \) is centered, and \( x_1, x_2, x_3 \) are normalized so that the Assumption 5.1 is satisfied. The next display shows the values of these vectors and the correlations between the output \( y \) and each \( x_j \).

```R
> print(round(t(y),digits=2))
[1,] 3.09  0.4 -1.12 -0.3  -2.07
> print(round(t(x1),digits=2))
 [,1] 0.89 -0.22 -0.22 -0.22 -0.22
> print(round(t(x2),digits=2))
 [,1] 0.82 -0.09 -0.55 -0.09 -0.09
> print(round(t(x3),digits=2))
 [,1] -0.22  0.89 -0.22 -0.22 -0.22
> cor(y,cbind(x1,x2,x3))
     x1    x2    x3
[1,] 0.8823568 0.851197 0.1145183
```

Clearly \( x_1 \) enters the model first as it has the largest absolute correlation with the output \( y \). Figure 1 plots how the correlations change as LIRAS moves along the direction \( x_1 \). Because \( x_2 \) reaches as much absolute correlation as \( x_1 \) with the current residual, it is the second variable entering the model. At this point, the correlation between \( x_2 \) and the current residual is positive, so the coefficient of \( x_2 \) will move in the positive direction and become positive as LARS proceeds in the equiangular direction between \( x_1 \) and \( x_2 \). At the end of the second step, \( x_3 \) comes into the model. In the third step, the LARS moves \( \hat{\mu} \) in the equiangular direction among \( x_1, x_2, x_3 \) and finally reaches the projection of \( y \) onto the linear space spanned by \( x_1, x_2, x_3 \), in the meanwhile the coefficients reach the least squares estimate as shown in the next display.

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|-----------|---------|---------|
|          |           |         |         |
Observe that the coefficient of $x_2$ is negative, so at some point in the third step, there is a sign change on $\hat{\beta}_2$ because it becomes positive at the end of the second step. Here is the reason. Although the equiangular direction $u_3$ among $x_1, x_2, x_3$ makes the same positive correlation with each of them; the coefficient of $x_2$, when expressing $u_3$ as a linear combination of $x_1, x_2, x_3$, is negative:

$$u_3 = 1.17x_1 - .64x_2 + 1.42x_3.$$ 

**Figure 1.** The first step of LARS for Example 5.1. It shows how the correlations change as LARS proceeds in the direction $x_1$. We see that $x_2$ is the second variable entering the model.

### 5.3. Degrees-of-freedom formula

We need a more general definition for the effective degrees of freedom an adaptively fitted model. We have a set of training data $(x_i, y_i), 1 \leq i \leq N$. Assume the input vectors are fixed, and given the $x_j$’s, the outputs $y_i, 1 \leq i \leq N$ are uncorrelated, with mean $\mu_i$ and variance $\sigma^2$, denoted in symbol by $y \sim (\mu, \sigma^2 I_N)$, where $\mu = (\mu_1, \ldots, \mu_N)^T$ and $I_N$ is the $N \times N$ identity matrix. Suppose a statistical model yields an estimate $\hat{\mu} = g(y)$ of $\mu$. The mean square error $E\|\hat{\mu} - \mu\|^2$ is a measure of the risk of the estimate $g(y)$. Denote by $\hat{\mu}_i$ the $i$th component of $\hat{\mu}$. Taking expectations in the identity

$$(\hat{\mu}_i - \mu_i)^2 = (y_i - \hat{\mu}_i)^2 - (y_i - \mu_i)^2 + 2(y_i - \mu_i)(\hat{\mu}_i - \mu_i)$$

and summing over $i$ yields

$$E\left(\frac{\|\hat{\mu} - \mu\|^2}{\sigma^2}\right) = E\left(\frac{\|y - \hat{\mu}\|^2}{\sigma^2} - n\right) + 2\sum_{i=1}^{N} \text{Cov}(\hat{\mu}_i, y_i).$$
Intuitively speaking, if we fit a model harder, then $E\|y - \hat{\mu}\|^2$ becomes smaller, but in the meanwhile, the covariances between $\hat{\mu}_i$ and $y_i$ increases, so the last term in the above identity increases. We define the degree of freedom of the estimate $\hat{\mu}$ as

$$df(\hat{\mu}) = \sum_{i=1}^{N} \frac{\text{Cov}(\hat{\mu}_i, y_i)}{\sigma^2}, \quad (5.3)$$

which “roughly” measures the model complexity. For an ordinary least square model with $k$ predictors (including the intercept), this definition coincides with the familiar degree of freedom $k$. For a general fitting procedure $\hat{\mu} = g(y)$, typically there is no close form method for estimating $df(\hat{\mu})$, but we can always use the bootstrapping approach. For LAR and Lasso, something magical happens.

**Theorem 5.2** (Efron et al., 2004). **Under the “positive cone” condition, the degree of freedom of the k-step LARS estimate $\hat{\mu}_k$ is EXACTLY $k$, that is $df(\hat{\mu}_k) = k$.**

The “positive cone” condition basically excludes the situation as described in Example 5.1. For more details, see Efron et al. (2004).

For Lasso, we do not have the same result, because the degrees of freedom of the full model is $p$, but Lasso can take more than $p$ steps to reach the full model. However, at any stage of Lasso, the degree of freedom approximately equals the number of predictors in the model, as shown by the next theorem. Zou et al. (2007) also discussed the consistency of the degree-of-freedom estimate.

**Theorem 5.3** (Zou et al., 2007). Let $\hat{\mu}_\lambda$ and $\hat{\beta}(\lambda)$ be the estimated mean vector and coefficient vector respectively for a given penalty parameter $\lambda$. Denote by $\hat{D}_\lambda$ the number of nonzero components of $\hat{\beta}(\lambda)$. Then for any $\lambda \geq 0$,

$$E(\hat{D}_\lambda) = df(\hat{\mu}).$$

6. **Generalized Lasso**

The generalized Lasso, proposed by Tibshirani and Taylor (2011), uses the $\ell_1$ norm to enforce certain structural constraints–instead of pure sparsity–on the coefficients in a linear regression. The generalized Lasso solves the optimization problem

$$\minimize \left\{ \frac{1}{2} \| y - X\beta \|^2 + \lambda \| D\beta \|_1 \right\}, \quad (6.1)$$

where $D \in \mathbb{R}^{m \times p}$ is a specified penalty matrix. The sparsity of $D\beta$ corresponds to some other desired behavior for $\beta$, typically one that is structural or geometric in nature. Various choices of $D$ give problems that are already well known in the literature: the fussed lasso, trend filtering and wavelet smoothing etc.

6.1. **Examples.**

6.2. **Algorithms.**

6.3. Connection to LARS.

**References**


