1 The Lasso

Assume we want to find the best $C_p$ model $\beta$ for the linear model.

$$y = X\beta + z$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ and $z_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. This is equivalent to solving the following optimisation problem

$$\arg \min_{\beta} \frac{1}{2} \|y - X\beta\|^2 + \lambda \sigma^2 \|\beta\|_0,$$

where $\|\cdot\|_0$ is used to represent the $\ell_0$ norm of a vector (which is the number of non-zero entries in the vector). For the $C_p$ model, $\lambda^2 = 2$ but we may be interested in general values of $\lambda$. In addition to being overly optimistic in predicting the risk of the selected model (when $\lambda^2 = 2$) as discussed in the last lecture, this is also an NP-hard problem, and thus there no efficient algorithm is known to solve this.

The Lasso is a relaxation of the optimization problem of Eq (1) to

$$\arg \min_{\beta} \frac{1}{2} \|y - X\beta\|^2 + \lambda \sigma \|\beta\|_1,$$

where $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$ is the $\ell_1$ norm of $\beta$.

We note that the Lasso optimization is a convex optimization problem—in fact it is a quadratic program—which can be solved efficiently.

As an aside, we note that efficient algorithms are not known to solve all convex optimization problems. For example, no efficient algorithm is known to solve the following feasibility problem for a general matrix $A$:

$$\text{decide whether } A \in \text{CoPos}$$

where CoPos is the convex cone of co-positive matrices $M$ obeying $x \geq 0 \implies x^T M x \geq 0$. 

2 \(\ell_0 - \ell_1\) equivalence

We note that the optimization problem

\[
\begin{align*}
\min & \|\beta\|_{\ell_0}, \\
\text{s.t.} & \quad X\beta = y
\end{align*}
\]

(3)

is an NP-hard problem.

However the problem

\[
\begin{align*}
\min & \|\beta\|_{\ell_1}, \\
\text{s.t.} & \quad X\beta = y
\end{align*}
\]

(4)

is just a linear program and thus can be efficiently solved.

Further we note that the \(\ell_1\) norm is the closest convex approximation of the \(\ell_0\) quasi-norm. That is, the \(\ell_1\) ball is the smallest convex ball that contains \(\pm e_i = (0, \cdots (i \text{ zeros}) \cdots, \pm 1, \cdots, 0)\).

Thus relaxing the optimization of Eq. (3) to the optimization of Eq. (4) is reasonable. In fact [1, 2] and later extensions show that under broad conditions, the minimizers of the two optimization problems are equal. These are studied widely in the research area of \textit{Compressed Sensing}.

3 \textbf{The Lasso dual}

To compute a of the lasso, we begin by rewriting Eq. (2) (with \(\sigma = 1\)) as

\[
\begin{align*}
\min_b & \quad \frac{1}{2}\|y - Xb\|^2 + \lambda \|b\|_1 = \min_b \frac{1}{2}\|y - \mu\|^2 + \lambda \|b\|_1, \\
\text{s.t.} & \quad \mu = Xb.
\end{align*}
\]

Writing the Lagrangian \(\mathcal{L}(\mu, b, r)\) for the RHS, we get that

\[
\begin{align*}
\mathcal{L}(\mu, b, r) &= \frac{1}{2}\|y - \mu\|^2 + \lambda \|b\|_1 + r^T(\mu - Xb), \\
&= \frac{1}{2}\|\mu - (y - r)\|^2 - \frac{1}{2}\|r\|^2 + y^T r - \langle X^T r, b \rangle + \lambda \|b\|_1.
\end{align*}
\]

We further note that,

\[
\begin{align*}
\min_b & \quad \frac{1}{2}\|y - Xb\|^2 + \lambda \|b\|_1 = \min_{b, \mu} \max_r \mathcal{L}(\mu, b, r) \\
&\overset{(a)}{=} \max_r \min_{b, \mu} \mathcal{L}(\mu, b, r),
\end{align*}
\]

where (a) follows from strong convex duality.

Further,

\[
\min_\mu \mathcal{L}(\mu, b, r) = -\frac{1}{2}\|r\|^2 + y^T r - \langle X^T r, b \rangle + \lambda \|b\|_1,
\]
which is achieved when $\mu = y - r$.

Next, we observe that,
\[
\min_b \mathcal{L}(y - r, b, r) = \min_b -\frac{1}{2}||r||_2^2 + y^T r - \langle X^T r, b \rangle + \lambda ||b||_1,
\]
\[
= -\frac{1}{2}||r||_2^2 + y^T r + \min_b \langle \lambda \text{sgn}(b) - X^T r, |b| \rangle,
\]
\[
= \begin{cases} 
-\infty & \text{if } ||X^T r||_\infty > \lambda \text{ for some co-ordinate of } b \uparrow \infty, \\
-\frac{1}{2}||r||_2^2 + y^T r & \text{if } ||X^T r||_\infty \leq \lambda \text{ for } b = 0.
\end{cases}
\]

where we have that $\text{sgn}(b)_i = \text{sgn}(b_i)_i$, $|b|_i = |b_i|$ and $||.||_\infty$ is the $\ell_\infty$-norm of a vector (that is the largest entry in absolute value in the vector).

Thus we have that,
\[
\min_b \frac{1}{2}||y - X b||^2 + \lambda ||b||_1 = \max_r \mathcal{L}(y - r, 0, r),
\]
\[
\text{s.t. } ||X^T r||_\infty \leq \lambda
\]
\[
= \max_r -\frac{1}{2}||r||_2^2 + y^T r,
\]
\[
\text{s.t. } ||X^T r||_\infty \leq \lambda
\]
\[
= \min_r \frac{1}{2}||r||_2^2 - y^T r,
\]
\[
\text{s.t. } ||X^T r||_\infty \leq \lambda
\]
\[
= \min_r \frac{1}{2}||y - r||_2^2.
\]
\[
\text{s.t. } ||X^T r||_\infty \leq \lambda
\]

where the last equality follows by adding the constant $||y||_2^2$.

Define
\[
C = \{ z \in \mathbb{R}^n : ||X^T z||_\infty \leq \lambda \}
\]

We note that $C \subseteq \mathbb{R}^n$ is a convex polytope as shown in Figure 1. Let $\Pi_C : \mathbb{R}^n \to C$ be the projection operator onto $C$.

Then we have that the estimate of $\mu$, $\hat{\mu}$ and that of $b$, $\hat{b}$ are given by,
\[
\hat{\mu} = y - \Pi_C(y),
\]
\[
\hat{b} = X^\dagger \hat{\mu},
\]

where $X^\dagger$ is the pseudo-inverse of $X$.

4 SURE formula for Lasso prediction error

An unbiased estimate for prediction error is given by the SURE formula (previous lecture)
\[
\text{SURE} = \text{RSS} + 2\sigma^2(n - \text{div}(\Pi_C(y))),
\]
Figure 1: A geometric interpretation of the dual Lasso problem. We assume $n = 2$, and $X_1$ and $X_2$ are the two rows of $X$, $\|X^T r\|_{\infty} \leq \lambda$ gives us a polytope in $\mathbb{R}^2$ and want to find the projection of $y$ onto the polytope to give us the residual $r$. The estimate is the signal remaining after removing the residual signal.

where RSS is the residual sum of squares.

To reason about the $\text{div}(\Pi_C(y))$, a simple example with $p = 3$ and $n = 2$ is discussed in Figure 2. More formally, we have that

$$\text{div}(\Pi_C(y)) = \text{Dimension of the affine space projected onto.}$$

Thus define

$$S = \{j \in [n] | |X_j^T r| = \lambda\},$$

to be the set of constraints that are active when doing the projection. We note that as $\hat{\mu} = y - r$, the co-ordinates in $S$ are precisely the non-zero co-ordinates in $\hat{b}$.

Further note that the affine space being projected into is $T = \{z | X_j^T z = \lambda \epsilon_j, j \in S\}$, where $\epsilon_j = 1$ if $X_j^T r = \lambda$ and $\epsilon_j = -1$ if $X_j^T r = -\lambda$ for some $j \in S$. Clearly the dimension of the affine-space $T$ is $n - |S|$.

This gives us that

$$\text{SURE}(\lambda) = \text{RSS}(\lambda) + 2\sigma^2 |S(\lambda)|,$$

$$= \text{RSS}(\lambda) + 2\sigma^2 |\{j : \hat{b}_j(\lambda) \neq 0\}|.$$
Figure 2: A simple example to reason about $\text{div}(\Pi_C(y))$. We consider a simple example with $p = 3$ and $n = 2$. Let the polytope in white denote the constraints imposed by $\|X^T r\|_\infty \leq \lambda$. Hence the set $C$ is the polytope and its interior (and thus the region in white). Hence, if $y$ was in the white region, we would have had $\Pi_C(y) = y$. Thus $\text{div}(\Pi_C(y)) = n = 2$. If $y$ is in one of the red regions, the projection is to one of the faces of the polytope. We note that any perturbation of $y$ in the direction orthogonal to the face does not change its projection. Thus $\text{div}(\Pi_C(y)) = n - 1 = 1$. Similarly, if $y$ lies in a blue region, then it gets projected to a vertex of the polytope, and thus any local perturbation does not really change the projection. Thus in this case $\text{div}(\Pi_C(y)) = 0$.

References
