Outline

1. Review: Canonical Selection Procedure
2. Risk Inflation of Thresholding Rules
3. Excess Risk for the LASSO

1 Canonical Selection Procedure

Consider the usual linear model

\[ y = X\beta + z \]

and the canonical \( \ell_0 \) selection procedure

\[
\min \|y - X\hat{\beta}\|^2 + \lambda^2 \sigma^2 \|\hat{\beta}\|_0.
\]

We have seen several choices of the parameter \( \lambda \)

- AIC: \( \lambda^2 = 2 \)
- BIC: \( \lambda^2 = \log n \)
- RIC: \( \lambda^2 \approx 2 \log p \)

We define the risk of an variable selection procedure as

\[ R(\mu, \hat{\mu}[S]) = \|\mu - \mu[S]\|^2 + |S|\sigma^2, \]

and the corresponding optimal risk as

\[ R^I(\mu) = \min_S R(\mu, \hat{\mu}[S]). \]

We have seen last lecture that the RIC gives us an oracle inequality. That is, when \( \lambda \propto \sqrt{2\log p} \), we are within a logarithmic factor of the optimal risk,

\[
E\|\mu - \hat{\mu}\|^2 = O(2\log p) \left[ \sigma^2 + R^I(\mu) \right].
\]

The critical problem shared by these methods is that finding the minimum of the empirical risk requires an exhaustive search over all possible models. This search is completely intractable for even moderate values of \( p \). We will get back to this later when we discuss the LASSO. First, we will present a lower bound, arguing that one cannot expect to do much better than a \( \log p \) factor (at least uniformly) of the oracle.
2 Risk Inflation

Foster and George (1994) proved that
\[
\inf_{\hat{\mu}} \sup_{\mu} \frac{R(\mu, \hat{\mu})}{\sigma^2 + R^I(\mu)} \geq 2 \log p (1 + o(1))
\]
for all hard-thresholding rules. Johnstone generalized the result to all estimators.

**Proof sketch:** (For details, see the original 1994 paper by Foster and George available here: [http://diskworld.wharton.upenn.edu/research/risk_inflation.pdf](http://diskworld.wharton.upenn.edu/research/risk_inflation.pdf)) We make the simplifying assumption (as Foster and George did) that \( \hat{\mu} \) is a HT estimator:
\[
\hat{\mu} = \begin{cases} 
  y & |y| \geq \lambda \\
  0 & \text{o.w.}
\end{cases}
\]
in which \( \lambda \) is fixed. Assume that \( \sigma = 1 \). It is sufficient to prove that \( \forall \lambda, \)
\[
RI(\lambda) = \max_k \sup_{||\mu||_0=k} \frac{R(\mu, \hat{\mu})}{1 + k} \geq 2 \log p (1 + o(1))
\]
since for \( k \)-sparse \( \mu \) we certainly have
\[
R^I(\mu) = \sum_i \mu_i^2 \wedge 1 \leq k
\]
The risk of a thresholding rule for \( \mu \in \mathbb{R} \) is
\[
r_H(\mu, \lambda) = \mathbb{E}(Z^2; |Z + \mu| \geq \lambda) + \mu^2 \mathbb{P}(|Z + \mu| \leq \lambda)
\]
The worst-case risk of a HT rule over \( k \)-sparse \( \mu \) is
\[
f(k) = (p - k)r_H(0, \lambda) + k \sup_{\mu} r_H(\mu, \lambda)
\]
Then
\[
RI(\lambda) = \max_k \frac{f(k)}{1 + k}
\]
\[
\geq \max \left( f(0), \frac{f(p)}{1 + p} \right)
\]
\[
= \max \left( pr_H(0, \lambda), \frac{p}{1 + p} \sup_{\mu} r_H(\mu, \lambda) \right)
\]
We know that the risk inflation is proportional to \( p \) if \( \lambda \) is fixed (independent of \( p \)), check the risk inflation at \( \mu = 0 \). We can show
\[
\lambda^2 + o(\lambda) \leq \sup_{\mu} r_H(\mu, \lambda) \leq \lambda^2 + 1
\]
Further,
\[ r_H(0, \lambda) = \mathbb{E}[Z^2; |Z| > \lambda] \approx 2\lambda \phi(\lambda) \]
Combining these with the previous bound, we have (for large \( p \))
\[ \text{RI}(\lambda) \geq \max\left(2p\lambda \phi(\lambda), \lambda^2\right) \]
The \( \lambda \) minimizing this lower bound is the one that equalizes the two arguments to the maximum.
\[ 2p\lambda \phi(\lambda) \approx \lambda^2 \]
i.e. for large \( \lambda \),
\[ \frac{\phi(\lambda)}{\lambda} \approx \frac{1}{2p} \]
or \( \lambda = \sqrt{2 \log p} \). Plugging in this value yields
\[ \text{RI}(\lambda) \geq 2 \log p (1 + o(1)) \]

To obtain our estimate for the worst-case \( r_H(\mu, \lambda) \),
\[
\sup_{\mu} r_H(\mu, \lambda) \geq \mu^2 \mathbb{P}(\mu + z \leq \lambda)
\approx \lambda^2 - \lambda \sqrt{8 \log \lambda} + o\left(\lambda \sqrt{\log \lambda}\right)
\]

**Discussion:** This tells us that try as we might we cannot improve on the logarithmic risk inflation factor in a minimax sense over all \( \mu \). However, it may be possible to improve on it for certain important classes of \( \mu \). We will see more on this later.

### 3 Excess Risk for the LASSO

Consider instead solving the \( \ell_1 \) regularized problem
\[
\min \frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \sigma \|\hat{\beta}\|_1
\]
We can view the \( \ell_1 \) norm as a proxy for the \( \ell_0 \) counting norm. Indeed, \( \ell_1 \) norm is closest convex approximation to \( \ell_0 \) quasi-norm. As we will illustrate, we cannot hope to achieve comparable optimality results for the Lasso.

**Take-away message:** In the following, we will show a case where the Lasso does not work well in terms of risk relative to an oracle, even when the correlation between columns of \( X \) is very low.
**Case study:**  \( X = [I_n F_{2,n}] \), where \( F_{2,n} \) is the discrete cosine transform matrix excluding the first column (the DC component). Specifically,

\[
F_{2,n} = \begin{pmatrix}
\varphi_1(1) & \varphi_2(1) & \cdots & \varphi_{n-1}(1) \\
\varphi_1(2) & \varphi_2(2) & \cdots & \varphi_{n-1}(2) \\
\vdots & \vdots & \ddots & \vdots \\
\varphi_1(n) & \varphi_2(n) & \cdots & \varphi_{n-1}(n)
\end{pmatrix},
\]

where

\[
\varphi_{2k-1}(t) = \sqrt{2/n} \cos\left( \frac{2\pi kt}{n} \right), \quad k = 1, 2, \cdots, n/2 - 1
\]

\[
\varphi_{2k}(t) = \sqrt{2/n} \sin\left( \frac{2\pi kt}{n} \right), \quad k = 1, 2, \cdots, n/2 - 1
\]

\[
\varphi_{n-1}(t) = (-1)^{t}/\sqrt{n}.
\]

In this case, \( p = 2n - 1 \). The maximum correlation between columns of \( X \), \( \mu(X) = \sqrt{2/n} \), which shows that the columns of \( X \) are extremely incoherent. Besides, \( \|X\|^2 \leq 2 \).

To illustrate the issue consider an example where \( n = 256 \) and \( \beta \) (shown in Figure 1) only has 24 nonzeros, chosen such that each entry of \( f = X\beta \) is 1.

![Figure 1: Graphical representation of \( f = X\beta \) and \( \beta \) with \( n = 256 \), \( p = 511 \).](image)

In noisy setting, the observed noisy \( X\beta + z \), the Lasso estimate \( \hat{\beta} \) and \( X\hat{\beta} \) are shown in Figure 2.

In this example, we know \( R^f(\mu) = 24\sigma^2 \). However, the figure demonstrates that the LASSO solution simply soft-thresholds \( y \).

Empirically and provably,

\[
X\hat{\beta} = y - \lambda\sigma 1, \quad \hat{\beta}_i = \begin{cases} y_i - \lambda\sigma, & i \in \{1, \cdots, n\}, \\ 0, & i \in \{n + 1, \cdots, 2n - 1\}. \end{cases}
\]

The Lasso MSE, \( \|\hat{f} - f\|^2 \sim n\sigma^2(1 + \lambda^2) \), is horrible. The \( l_0 \) norm of \( \hat{\beta} \), \( |\text{supp}(\hat{\beta})| = n \), is much higher than that of \( \beta \), \( |\text{supp}(\beta)| = \frac{3}{2}\sqrt{n} \).
Figure 2: Graphical representation of the results with \( n = 256, \ p = 511 \). The lasso estimate \( \hat{\beta} \) has far more non-zero entries than \( \beta \) and results in large MSE.