Outline

• Model selection
• Lasso estimator
• Risk inflation
• Minimax risk for sparse vectors
Asymptotic notation

- \( f(n) \lesssim g(n) \) or \( f(n) = O(g(n)) \) means
  \[
  \lim_{n \to \infty} \frac{|f(n)|}{|g(n)|} \leq \text{const}
  \]

- \( f(n) \gtrsim g(n) \) or \( f(n) = \Omega(g(n)) \) means
  \[
  \lim_{n \to \infty} \frac{|f(n)|}{|g(n)|} \geq \text{const}
  \]

- \( f(n) \asymp g(n) \) or \( f(n) = \Theta(g(n)) \) means
  \[
  \text{const}_1 \leq \lim_{n \to \infty} \frac{|f(n)|}{|g(n)|} \leq \text{const}_2
  \]

- \( f(n) = o(g(n)) \) means
  \[
  \lim_{n \to \infty} \frac{|f(n)|}{|g(n)|} = 0
  \]
Model selection

All models are wrong but some are useful.

— George Box
Basic linear model

\[ y = X \beta + \eta, \]

- \( y = [y_1, \cdots, y_n]^\top \in \mathbb{R}^n \): observed data / response variables
- \( X = [x_1, \cdots, x_n]^\top \): design matrix / feature matrix (known)
  - assumed to be full rank
- \( \beta = [\beta_1, \cdots, \beta_p]^\top \in \mathbb{R}^p \): unknown signal / regression coefficients
- \( \eta = [\eta_1, \cdots, \eta_n]^\top \in \mathbb{R}^n \): noise

Throughout this lecture, we assume Gaussian noise

\[ \eta \sim \mathcal{N}(0, \sigma^2 I_n) \]
Regression:

- find relationship between response $y_i$ and explanatory variables $x_{i,1}, \ldots, x_{i,p}$

- use the fitted model to make prediction

**Question:** which (sub)-set of variables / features should we include?

- Myth: nothing is lost by including every feature / variable available

- Paradoxically, we can often achieve better predictions by discarding a fraction of variables
Tradeoff

- Model too small $\implies$ large bias (underfitting)
- Model too large $\implies$ large variance and poor prediction (overfitting)

How to achieve a desired tradeoff between predictive accuracy and parsimony (model complexity)?
Underfitting

Recall that the least squares (LS) estimate is \( \hat{\beta} = (X^\top X)^{-1} X^\top y \)

- Divide the design matrix into 2 parts: \( X = [X^{(1)}, X^{(2)}] \)
- \( \tilde{x} = \begin{bmatrix} \tilde{x}^{(1)} \\ \tilde{x}^{(2)} \end{bmatrix} \): new data
- LS estimate based only on \( X^{(1)} \):
  \[ \hat{\beta}^{(1)} := (X^{(1)^\top} X^{(1)})^{-1} X^{(1)^\top} y \]
  with prediction at \( \tilde{x} \) given by
  \[ \hat{y}_{\text{underfit}} = \tilde{x}^{(1)^\top} \hat{\beta}^{(1)} \]
- LS estimate based on true model
  \[ \hat{\beta} := (X^\top X)^{-1} X^\top y \]
  with prediction at \( \tilde{x} \) given by
  \[ \hat{y}_{\text{true}} = [\tilde{x}^{(1)^\top}, \tilde{x}^{(2)^\top}] \hat{\beta} \]
Bias due to underfitting

Suppose the ground truth is $\beta = \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \end{bmatrix}$, then

$$
\mathbb{E} \left[ \hat{\beta}^{(1)} \right] = \left( X^{(1)\top} X^{(1)} \right)^{-1} X^{(1)\top} \left( X^{(1)} \beta^{(1)} + X^{(2)} \beta^{(2)} \right)
$$

$$
= \beta^{(1)} + \underbrace{\left( X^{(1)\top} X^{(1)} \right)^{-1} X^{(1)\top} X^{(2)} \beta^{(2)}}_{\text{bias}}
$$

$\implies \hat{\beta}^{(1)}$ is a biased estimate of $\beta^{(1)}$
Prediction variance due to underfitting

Fact 7.1

\[ \text{Var} [\hat{y}_{\text{true}}] \geq \text{Var} [\hat{y}_{\text{underfit}}] \]

• **Implications:** the “apparent” prediction variance tends to decrease when we adopt small models

• (Exercise): compute the prediction variance under overfitting
Proof of Fact 7.1

Observe that
\[
\text{Cov}[\hat{\beta}] = (X^\top X)^{-1} X^\top \text{Cov}[y] X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}
\]
\[
= \sigma^2 \begin{bmatrix}
(X^{(1)^\top} X^{(1)})^{-1} + LML^\top & -LM \\
-ML^\top & M
\end{bmatrix}
\]
(matrix inversion identity)

where \(L = (X^{(1)^\top} X^{(1)})^{-1} X^{(1)^\top} X^{(2)}\) and
\[
M = \left\{ X^{(2)^\top} \left( I - X^{(1)} \left( X^{(1)^\top} X^{(1)} \right)^{-1} X^{(1)^\top} \right) X^{(2)} \right\}^{-1} \succeq 0.
\]

This gives
\[
\text{Var}[\hat{y}_{\text{true}}] = \begin{bmatrix}
\tilde{x}^{(1)^\top} \\
\tilde{x}^{(2)^\top}
\end{bmatrix} \text{Cov}[\hat{\beta}] \begin{bmatrix}
\tilde{x}^{(1)} \\
\tilde{x}^{(2)}
\end{bmatrix}
\]
\[
= \sigma^2 \tilde{x}^{(1)^\top} \left( X^{(1)^\top} X^{(1)} \right)^{-1} \tilde{x}^{(1)} + \sigma^2 \left( L^\top \tilde{x}^{(1)} - \tilde{x}^{(2)} \right)^\top M \left( L^\top \tilde{x}^{(1)} - \tilde{x}^{(2)} \right)
\]
\[
\geq \sigma^2 \tilde{x}^{(1)^\top} \left( X^{(1)^\top} X^{(1)} \right)^{-1} \tilde{x}^{(1)} = \text{Var}[\hat{y}_{\text{underfit}}]
\]
Choosing a subset of explanatory variables might improve prediction

**Question:** which subset shall we select?

One strategy

(1) pick a criterion that measures how well a model performs
(2) evaluate the criterion for each subset and pick the best

One popular choice: choose a model that predicts well
Prediction error and model error

- training set: \( y, X \)
- \( \hat{\beta} \): an estimate based on training set
- new data: \( \tilde{y} = \tilde{X}\beta + \tilde{\eta} \in \mathbb{R}^m \), where \( \tilde{\eta} \sim \mathcal{N}(0, I_m) \)
- Goal: use \( \hat{\beta} \) to predict \( \tilde{y} \)

One may assess the quality of the estimate based on its *prediction error* on \( \tilde{y} \), i.e.

\[
\text{PE} := \mathbb{E} \left[ \| \tilde{X}\beta - \tilde{y} \|^2 \right] \\
= \mathbb{E} \left[ \| \tilde{X}(\hat{\beta} - \beta) \|^2 \right] + 2\mathbb{E} \left[ (\tilde{X}(\hat{\beta} - \beta))^\top (\tilde{y} - \tilde{X}\beta) \right] + \mathbb{E} \left[ \| \tilde{y} - \tilde{X}\beta \|^2 \right] \\
= \mathbb{E} \left[ \| \tilde{X}(\hat{\beta} - \beta) \|^2 \right] + m\sigma^2 \\
:= \text{ME (model error)} + \text{variability of data}
\]
Residual sum of squares (RSS)

We shall set $\tilde{\mathbf{X}} = \mathbf{X}$ (and hence $m = n$) out of simplicity

- the case where the structures of new and old data are the same

Unfortunately, we do not have access to PE (as we don’t know $\mathbf{\beta}$) ➞ need an operational criterion for estimating PE

- One candidate: estimate PE via residual sum of squares

\[
\text{RSS} := \left\| \mathbf{y} - \mathbf{X}\hat{\mathbf{\beta}} \right\|_2^2
\]

⇒ training error
Training error underestimates prediction error

Suppose \( X\hat{\beta} = \Pi y \) for some given \( \Pi \) with \( \text{Tr}(\Pi) > 0 \) (e.g. LS), then

\[
PE = \mathbb{E}[\text{RSS}] + 2\sigma^2 \text{Tr}(\Pi) > \mathbb{E}[\text{RSS}]
\]  \hspace{1cm} (7.1)

Proof:

\[
PE - \mathbb{E}[\text{RSS}] = \mathbb{E} \left[ \| \tilde{y} - X\hat{\beta} \|^2 \right] - \mathbb{E} \left[ \| y - X\hat{\beta} \|^2 \right]
\]

\[
= \mathbb{E} \left[ \| \tilde{y} \|^2 + \| X\hat{\beta} \|^2 - 2\langle \tilde{y}, X\hat{\beta} \rangle \right]
\]

\[
- \mathbb{E} \left[ \| y \|^2 + \| X\hat{\beta} \|^2 - 2\langle y, X\hat{\beta} \rangle \right]
\]

\[
= 2\mathbb{E} \left[ \langle y - \tilde{y}, X\hat{\beta} \rangle \right] = 2\mathbb{E} \left[ \langle \eta - \tilde{\eta}, \Pi y \rangle \right]
\]

\[
= 2\mathbb{E} \left[ \langle \eta, \Pi \eta \rangle \right] \overset{(a)}{=} 2\text{Tr} \left( \Pi \mathbb{E} \left[ \eta \eta^\top \right] \right)
\]

\[
= 2\sigma^2 \text{Tr}(\Pi),
\]

where (a) follows from the identity \( \text{Tr}(A^\top B) = \text{Tr}(BA^\top) \).
Example: least squares (LS) estimator

The least squares solution is

$$\hat{\beta}_{ls} := \arg \min_{\hat{\beta}} \| y - X \hat{\beta} \|^2_2 = (X^\top X)^{-1} X^\top y$$

The fitted values $\hat{y}^{ls}$ is given by

$$\hat{y}^{ls} = \Pi^{ls} y := X (X^\top X)^{-1} X^\top y.$$ 

$$\implies PE = \mathbb{E}[RSS] + 2\sigma^2 \text{Tr}(\Pi^{ls}) = \mathbb{E}[RSS] + 2\sigma^2 P$$
LS estimator for a given model

Suppose the model (i.e. support of $\beta$) is $S \subseteq \{1, \cdots, p\}$. The least squares solution given $S$ is

$$\hat{\beta}_S := \arg \min_{\hat{\beta}} \{ \| y - X\hat{\beta} \|_2^2 : \hat{\beta}_i = 0 \text{ for all } i \notin S \}$$

The fitted values $\hat{y}$ is then given by

$$\hat{y} = \Pi_S y := X_S (X_S^\top X_S)^{-1} X_S^\top y,$$

where $X_S$ is formed by the columns of $X$ at indices in $S$. 
Mallows’ $C_p$ statistic

In view of (7.1),

$$PE(\hat{\beta}_S) = \mathbb{E}[RSS(\hat{\beta}_S)] + 2\sigma^2 \text{Tr}(\Pi_S) = \mathbb{E}[RSS(\hat{\beta}_S)] + 2|S|\sigma^2,$$

since

$$\text{Tr}(\Pi_S) = \text{Tr}(X_S(X_S^T X_S)^{-1} X_S^T) = \text{Tr}((X_S^T X_S)^{-1} X_S^T X_S) = |S|.$$

**Definition 7.2 ($C_p$ statistic, Mallows ’73)**

$$C_p(S) = \underbrace{\text{RSS}(\hat{\beta}_S)}_{\text{training error}} + \underbrace{2\sigma^2|S|}_{\text{model complexity}}$$

$C_p$ is an unbiased estimate of prediction error
Model selection based on $C_p$ statistic

1. Compute $C_p(S)$ for each model $S$
2. Choose $S^* = \arg \min_S C_p(S)$

This is essentially an $\ell_0$-regularized least-squares problem

$$\minimize_{\hat{\beta}} \quad \| y - X\hat{\beta} \|^2_2 + 2\sigma^2\|\hat{\beta}\|_0$$

penalized by model complexity (7.2)
Example: orthogonal design

Suppose $X = I$, then (7.8) reduces to

$$\minimize_{\hat{\beta}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{\beta}_i)^2 + 2\sigma^2 \mathbf{1}\{\hat{\beta}_i \neq 0\}$$

Solving this problem gives

$$\hat{\beta}_i = \begin{cases} 0, & |y_i| \leq \sqrt{2}\sigma \\ y_i, & |y_i| > \sqrt{2}\sigma \end{cases} \text{ hard thresholding}$$

- Keep large coefficients; discard small coefficients
Example: orthogonal design

\[ \hat{\beta}_i = \psi_{ht}(y_i; \sqrt{2}\sigma) := \begin{cases} 0, & |y_i| \leq \sqrt{2}\sigma \\ y_i, & |y_i| > \sqrt{2}\sigma \end{cases} \]

Hard thresholding preserves data outside threshold zone
Lasso estimator
Lasso (Least absolute shrinkage and selection operator)

\[
\minimize_{\hat{\beta}} \quad \frac{1}{2} \| y - X \hat{\beta} \|_2^2 + \lambda \| \hat{\beta} \|_1
\]  

(7.3)

for some regularization parameter \( \lambda > 0 \)

- It is equivalent to

\[
\minimize_{\hat{\beta}} \quad \| y - X \hat{\beta} \|_2^2
\]

s.t. \( \| \hat{\beta} \| \leq t \)

for some \( t \) that depends on \( \lambda \)
  - a quadratic program (QP) with convex constraints

- \( \lambda \) controls model complexity: larger \( \lambda \) restricts the parameters more; smaller \( \lambda \) frees up more parameters
Lasso vs. MMSE (or ridge regression)

\( \hat{\beta} \): least squares solution

\[
\text{minimize}_{\beta} \quad \| y - X \beta \|_2 \\
\text{s.t.} \quad \| \beta \|_1 \leq t
\]

\[
\text{minimize}_{\beta} \quad \| y - X \beta \|_2 \\
\text{s.t.} \quad \| \beta \|_2 \leq t
\]

Fig. credit: Hastie, Tibshirani, & Wainwright
A Bayesian interpretation

Orthogonal design: \( y = \beta + \eta \) with \( \eta \sim \mathcal{N}(0, \sigma^2 I) \).

Impose an i.i.d. prior on \( \beta_i \) to encourage sparsity (Gaussian is not a good choice):

\[
\mathbb{P}(\beta_i = z) = \frac{\lambda}{2} e^{-\lambda |z|}
\]

(Laplacian prior)
A Bayesian interpretation of Lasso

Posterior of $\beta$:

$$
\mathbb{P}(\beta \mid y) \propto \mathbb{P}(y \mid \beta) \mathbb{P}(\beta) \propto \prod_{i=1}^{n} e^{-\frac{(y_i - \beta_i)^2}{2\sigma^2}} \frac{\lambda}{2} e^{-\lambda |\beta_i|}
$$

$$
\propto \prod_{i=1}^{n} \exp \left\{ -\frac{(y_i - \beta_i)^2}{2\sigma^2} - \lambda |\beta_i| \right\}
$$

$\implies$ maximum a posteriori (MAP) estimator:

$$
\arg \min_{\beta} \sum_{i=1}^{n} \left\{ \frac{(y_i - \beta_i)^2}{2\sigma^2} + \lambda |\beta_i| \right\} \quad \text{(Lasso)}
$$

Implication: Lasso is MAP estimator under Laplacian prior

Model selection and Lasso
Example: orthogonal design

Suppose $X = I$, then Lasso reduces to

$$\min_{\hat{\beta}} \quad \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{\beta}_i)^2 + \lambda |\hat{\beta}_i|$$

The Lasso estimate $\hat{\beta}$ is then given by

$$\hat{\beta}_i = \begin{cases} 
    y_i - \lambda, & y_i \geq \lambda \\
    y_i + \lambda, & y_i \leq -\lambda \\
    0, & \text{else}
\end{cases}$$

soft thresholding
Example: orthogonal design

\[ \hat{\beta}_i = \psi_{st}(y_i; \lambda) = \begin{cases} 
  y_i - \lambda, & y_i \geq \lambda \\
  y_i + \lambda, & y_i \leq -\lambda \\
  0, & \text{else} 
\end{cases} \]

Soft thresholding shrinks data towards 0 outside threshold zone
Optimality condition for convex functions

For any convex function $f(\beta)$, $\beta^*$ is an optimal solution iff $0 \in \partial f(\beta^*)$, where $\partial f(\beta)$ is the set of all subgradients at $\beta$.

- $s$ is a subgradient of $f(\beta) = |\beta|$ if

$$
\begin{align*}
  s &= \text{sign}(\beta), & \text{if } \beta \neq 0 \\
  s &= [-1, 1], & \text{if } \beta = 0
\end{align*}
$$

(7.4)
Optimality condition for convex functions

For any convex function $f(\beta)$, $\beta^*$ is an optimal solution iff $0 \in \partial f(\beta^*)$, where $\partial f(\beta)$ is the set of all subgradients at $\beta$.

- The subgradient of $f(\beta) = \frac{1}{2}(y - \beta)^2 + \lambda|\beta|$ can be written as
  $$g = \beta - y + \lambda s$$
  with $s$ defined in (7.4).

- We see that $\hat{\beta} = \psi_{st}(y; \lambda)$ by checking optimality conditions for two cases:
  - If $|y| \leq \lambda$, taking $\beta = 0$ and $s = y/\lambda$ gives $g = 0$.
  - If $|y| > \lambda$, taking $\beta = y - \text{sign}(y)\lambda$ gives $g = 0$. 

Model selection and Lasso 7-29
Consider the case where there is only a single parameter $\hat{\beta} \in \mathbb{R}$:

$$\minimize_{\hat{\beta} \in \mathbb{R}} \frac{1}{2} \| y - \hat{\beta} z \|_2^2 + \lambda |\hat{\beta}|.$$ 

Then one can verify that (homework)

$$\hat{\beta} = \psi_{st} \left( \frac{z^\top y}{\| z \|_2^2}; \frac{\lambda}{\| z \|_2^2} \right) = \begin{cases} 
\frac{z^\top y}{\| z \|_2^2} - \frac{\lambda}{\| z \|_2^2}, & \text{if } z^\top y > \lambda \\
0, & \text{if } |z^\top y| \leq \lambda \\
\frac{z^\top y}{\| z \|_2^2} + \frac{\lambda}{\| z \|_2^2}, & \text{else}
\end{cases}$$
Algorithm: coordinate descent

Idea: repeatedly cycle through the variables and, in each step, optimize only a single variable

- When updating $\hat{\beta}_j$, we solve

$$
\minimize_{\hat{\beta}_j \in \mathbb{R}} \frac{1}{2}\| y - \sum_{i: i \neq j} X_{i:} \hat{\beta}_i - X_{:,j} \hat{\beta}_j \|_2^2 + \lambda |\hat{\beta}_j| + \lambda \sum_{i: i \neq j} |\hat{\beta}_i|
$$

where $X_{:,j}$ is $j$th column of $X$

- This is exactly the single-parameter setting, and hence

$$
\hat{\beta}_j \leftarrow \psi_{\text{st}} \left( \frac{X_{:,j}^\top (y - \sum_{i: i \neq j} X_{i:} \hat{\beta}_i)}{\| X_{:,j} \|_2^2} ; \frac{\lambda}{\| X_{:,j} \|_2^2} \right)
$$
Algorithm: coordinate descent

Algorithm 7.1 Coordinate descent for Lasso

Repeat until convergence

for $j = 1, \cdots, n$:

\[
\hat{\beta}_j \leftarrow \psi_{st} \left( \frac{X_{:j}^\top (y - \sum_{i:i \neq j} X_{:i} \hat{\beta}_i)}{\|X_{:j}\|_2^2}; \frac{\lambda}{\|X_{:j}\|_2^2} \right)
\]  (7.5)
Risk inflation
Ideal risk: orthogonal design

\[ y_i = \beta_i + \eta_i, \quad i = 1, \ldots, n \]

Let’s first select / fix a model and then estimate: for a fixed model \( S \subseteq \{1, \ldots, n\} \), the LS estimate \( \hat{\beta}_S \) is

\[
(\hat{\beta}_S)_i = \begin{cases} 
  y_i, & \text{if } i \in S \\
  0, & \text{else}
\end{cases}
\]

- Mean square estimation error for a fixed model \( S \):

\[
\text{MSE}(\hat{\beta}_S, \beta) := \mathbb{E}[\|\hat{\beta}_S - \beta\|^2] = \sum_{i \in S} \mathbb{E}[(y_i - \beta_i)^2] + \sum_{i \notin S} \beta_i^2
\]

\[
= |S| \sigma^2 + \sum_{i \notin S} \beta_i^2
\]

\( |S| \sigma^2 \) variance due to noise

\( \sum_{i \notin S} \beta_i^2 \) bias (since we don’t estimate all coefficients)
• **Smallest MSE for a fixed model size:** If we fix the model size $|S| = k$, then the best model achieves

$$\text{MSE}_k(\beta) := \min_{S:|S|=k} \text{MSE}(\hat{\beta}_S, \beta) = k\sigma^2 + \min_{S:|S|=k} \sum_{i \notin S} \beta_i^2$$

$$= k\sigma^2 + \sum_{i=k+1}^{n} |\beta|_{(i)}^2$$

where $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(n)}$ are order statistics of \{|$\beta_i$|\}

**Implication:** good estimation is possible when $\beta$ compresses well
Optimizing the risk over all possible models

- **Ideal risk (smallest MSE over all models):** minimizing over all possible model size $k$ gives

\[
\text{MSE}_{\text{ideal}}(\beta) := \min_k \min_{S:|S|=k} \text{MSE}(\hat{\beta}_S, \beta) = \min_k \left\{ k\sigma^2 + \sum_{i=k+1}^{n} |\beta_{(i)}|^2 \right\} \\
= \sum_{i=1}^{n} \min\{ \sigma^2, \beta_i^2 \}
\]
Oracle lower bound

\[
\text{MSE}^{\text{ideal}}(\beta) = \sum_{i=1}^{n} \min\{\sigma^2, \beta_i^2\}
\]

- $\beta_i$ is worth estimating iff $|\beta_i| > \sigma^2$
- $\text{MSE}^{\text{ideal}}$ is the optimal risk \textit{if an oracle reveals which variables are worth estimating and which can be safely ignored}
- With the oracle information, one can achieve $\text{MSE}^{\text{ideal}}$ via

\[
\hat{\beta}_i^{\text{ideal}} = \begin{cases} 
y_i, & \text{if } |\beta_i| > \sigma, \\
0, & \text{else}
\end{cases} \quad \text{(eliminate irrelevant variables)}
\]
Risk inflation

• **Problem:** unfortunately, we do NOT know which model $S$ is the best and hence cannot attain $\text{MSE}^{\text{ideal}}$ ...

• Instead, we shall treat it as an oracle lower bound, and consider the increase in estimation error *due to selecting* rather than knowing the correct model
Definition 7.3 (risk inflation, Foster & George '94)

The risk inflation of an estimator $\hat{\beta}$ is

$$RI(\hat{\beta}) = \sup_{\beta} \frac{\text{MSE}(\hat{\beta}, \beta)}{\text{MSE}_{\text{ideal}}(\beta)},$$

where $\text{MSE}(\hat{\beta}, \beta) := \mathbb{E}[\|\beta - \hat{\beta}\|_2^2].$

- Idea: calibrate the actual risk against the ideal risk for each $\beta$ to better reflect the potential gains / loss

- Suggestion: *find a procedure that achieves low risk inflation!*
Consider identity design $X = I$, and $\hat{\beta}_i = \psi_{ht}(y_i; \lambda)$ or $\hat{\beta}_i = \psi_{st}(y_i; \lambda)$ with threshold zone $[-\lambda, \lambda]$

- For the extreme case where $\beta = 0$,

\[
\text{MSE}^{\text{ideal}}(\beta) = \sum_{i=1}^{p} \min\{\sigma^2, \beta_i^2\} = 0
\]

- In order to control risk inflation, $\lambda$ needs to be sufficiently large so as to ensure $\hat{\beta}_i \approx 0$ for all $i$. In particular,

\[
\max_{1 \leq i \leq p} |y_i| = \max_{1 \leq i \leq p} |\eta_i| \approx \sigma \sqrt{2 \log p} \quad \text{(exercise)}
\]

\[
\implies \quad \lambda \geq \sigma \sqrt{2 \log p}
\]
Risk inflation by soft / hard thresholding

Theorem 7.4 (Foster & George ’94, Johnstone, Candes)

Let \( \hat{\beta} \) be either a soft or hard thresholding procedure with threshold \( \lambda = \sigma \sqrt{2 \log p} \). Then

\[
\text{MSE}(\hat{\beta}, \beta) \leq (2 \log p + c) \left( \sigma^2 + \text{MSE}^{\text{ideal}}(\beta) \right)
\]

where \( c = 1 \) for soft thresholding and \( c = 1.2 \) for hard thresholding.

For large \( p \), one typically has \( \text{MSE}^{\text{ideal}}(\beta) \gg \sigma^2 \). Then Theorem 7.4 implies

\[
\text{RI}(\hat{\beta}) \approx 2 \log p
\]
Proof of Theorem 7.4 for soft thresholding

WLOG, assume that $\sigma = 1$. The risk of soft thresholding for a single coordinate is

$$r_{st}(\lambda, \beta_i) := \mathbb{E}[(\psi_{st}(y_i; \lambda) - \beta_i)^2]$$

where $y_i \sim \mathcal{N}(\beta_i, 1)$.

1. There are 2 very special points that we shall single out: $\beta_i = 0$ and $\beta_i = \infty$. We start by connecting $r_{st}(\lambda, \beta_i)$ with $r_{st}(\lambda, 0)$ and $r_{st}(\lambda, \infty)$.

**Lemma 7.5**

$$r_{st}(\lambda, \beta) \leq r_{st}(\lambda, 0) + \beta^2 \quad (quadratic \ upper \ bound)$$

$$r_{st}(\lambda, \beta) \leq r_{st}(\lambda, \infty) = 1 + \lambda^2$$

Model selection and Lasso
Proof of Theorem 7.4 for soft thresholding

2. The next step is to control \( r_{st}(\lambda, 0) \)

**Lemma 7.6**

\[
 r_{st}(\lambda, 0) \leq 2\phi(\lambda)/\lambda \leq \frac{\sqrt{2\log p}}{\lambda} \ll 1/p \quad \text{(very small)}
\]

where \( \phi(z) := \frac{1}{\sqrt{2\pi}} \exp(-z^2/2). \)

3. With these lemmas in mind, we are ready to prove Theorem 7.4

\[
\sum_{i=1}^{p} \mathbb{E}[(\beta_i - \hat{\beta}_i)^2] \leq pr_{st}(\lambda, 0) + \sum_{i=1}^{p} \min \left\{ \beta_i^2, \lambda^2 + 1 \right\}
\]
\[
< 1 + \sum_{i=1}^{p} \min \left\{ \beta_i^2, 2\log p + 1 \right\}
\]
\[
\leq (2\log p + 1) \left[ 1 + \sum_{i=1}^{p} \min \left\{ \beta_i^2, 1 \right\} \right]
\]
\[
= (2\log p + 1) \left( 1 + \text{MSE}^{\text{ideal}} \right)
\]
Proof of Lemma 7.5

Figure adapted from Johnstone ’15

WLOG, assume $\beta \geq 0$.

(1) To prove $r_{st}(\lambda, \beta) \leq r_{st}(\lambda, 0) + \beta^2$, it suffices to show $\frac{\partial r_{st}}{\partial \beta} \leq 2\beta$, as

$$r_{st}(\lambda, \beta) - r_{st}(\lambda, 0) = \int_0^\beta \frac{\partial r_{st}(\lambda, \beta)}{\partial \beta} d\beta \leq \int_0^\beta 2\beta d\beta = \beta^2. $$

This follows since (exercise)

$$\frac{\partial r_{st}(\lambda, \beta)}{\partial \beta} = 2\beta \mathbb{P}(Z \in [-\lambda - \beta, \lambda - \beta]) \in [0, 2\beta], \quad (7.6)$$

with $Z \sim \mathcal{N}(0, 1)$
WLOG, assume $\beta \geq 0$.

(2) The identity (7.6) also shows $r_{st}$ is increasing in $\beta > 0$, and hence

$$r_{st}(\lambda, \beta) \leq r_{st}(\lambda, \infty) = \mathbb{E}[((\beta + z - \lambda) - \beta)^2] = 1 + \lambda^2$$

(7.7)
Proof of Lemma 7.6 (Candes)

\[ \frac{r_{st}(\lambda, 0)}{2} = \int_\lambda^\infty (y - \lambda)^2 \phi(y) dy \]
\[ = \int_\lambda^\infty (y - 2\lambda) y \phi(y) dy + \lambda^2 \int_\lambda^\infty \phi(y) dy \]
\[ \equiv -\int_\lambda^\infty (y - 2\lambda) \phi'(y) dy + \lambda^2 \mathbb{P} \{ Z > \lambda \} \]
\[ \leq -\lambda \phi(\lambda) + \left(1 + \lambda^2\right) \mathbb{P} \{ Z > \lambda \} \]
\[ \leq -\lambda \phi(\lambda) + \left(1 + \lambda^2\right) \frac{\phi(\lambda)}{\lambda} = \frac{\phi(\lambda)}{\lambda}, \]

where (a) follows since \( \phi'(y) = -y \phi(y) \), (b) follows from integration by parts, and (c) holds since \( \mathbb{P} \{ Z > \lambda \} \leq \frac{\phi(\lambda)}{\lambda} \).
Theorem 7.7 (Foster & George ’94, Jonestone)

\[
\begin{align*}
\inf_{\hat{\beta}} \sup_{\beta} \frac{\text{MSE}(\hat{\beta}, \beta)}{\sigma^2 + \text{MSE}_{\text{ideal}}(\beta)} \geq (1 + o(1))2 \log p
\end{align*}
\]

- Soft and hard thresholding rules—depending only on available data without access to an oracle—can achieve the ideal risk up to the multiplicative factor \((1 + o(1))2 \log p\)
- This \(2 \log p\) factor is asymptotically optimal for unrestricted \(\beta\)
Comparison with canonical selection procedure

1. Minimax-optimal procedure w.r.t. risk inflation

\[ \hat{\beta}_i = \psi_{ht} \left( y_i; \sigma \sqrt{2 \log p} \right) \]

2. Canonical selection based on \( C_p \) statistics

\[ \hat{\beta}_i = \psi_{ht} \left( y_i; \sqrt{2} \sigma \right) \]

- Optimal procedure employs a much larger threshold zone
- **Reason:** \( \min_S C_p(S) \) underestimates \( \min_S \mathbb{E}[PE(S)] \) since

\[
\mathbb{E} \left[ \min_S C_p(S) \right] \leq \min_S \mathbb{E}[C_p(S)] = \min_S \mathbb{E}[PE(S)]
\]

- sometimes \( \ll \)

- e.g. when \( \beta = 0 \), \( \| \psi_{ht} (y; \sqrt{2} \sigma) - \beta \|^2 \preceq n \gg 0 \) with high prob.
More general models

Let’s turn to a general design matrix $X$:

$$y = X\beta + \eta \quad \text{where } \eta \sim \mathcal{N}(0, I)$$

One can take the ideal risk to be

$$\text{MSE}^{\text{ideal}} := \min_S \text{PE}(S) = \min_S \left\{ \mathbb{E}\left[\|X_S \hat{\beta}_S - X\beta\|^2_2 \right] + |S|\sigma^2 \right\}$$

model error
More general models

Consider the $\ell_0$-penalized selection procedure

$$\minimize_{\hat{\beta}} \| y - X \hat{\beta} \|_2^2 + \lambda^2 \sigma^2 \| \hat{\beta} \|_0$$ (7.8)

for some $\lambda \asymp \sqrt{\log p}$

**Theorem 7.8** (Foster & George ’94, Birge & Massart ’01, Jonestone, Candes)

\[
\begin{align*}
\text{(achievability)} & \quad \text{MSE}(\hat{\beta}, \beta) \lesssim (\log p) \left\{ \sigma^2 + \text{MSE}^{\text{ideal}}(\beta) \right\} \\
\text{(minimax lower bound)} & \quad \inf_{\hat{\beta}} \sup_{\beta} \frac{\text{MSE}(\hat{\beta}, \beta)}{\sigma^2 + \text{MSE}^{\text{ideal}}(\beta)} \gtrsim \log p
\end{align*}
\]

(7.8) is nearly minimax optimal for arbitrary designs!
Lasso is suboptimal for coherent design

\[ X = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ \epsilon \\ \epsilon \\ 1 & 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -1/\epsilon \\ 1/\epsilon \end{bmatrix}, \quad \text{and} \quad y = X\beta = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ 0 \end{bmatrix} \]

When \( \epsilon \to 0 \), solution to Lasso is

\[ \hat{\beta} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \quad \text{far from the truth} \]

- Issue: the last 2 columns of \( X \) are too similar / correlated
Minimax risk for sparse vectors
So far we’ve considered risk without any restriction on $\beta$. Practically, prior knowledge (like sparsity of $\beta$) might be exploited to yield more accurate estimates.

**Theorem 7.9**

Suppose $X = I$. For any $k$-sparse $\beta$ with $k \ll p$, the asymptotic minimax risk is

$$\inf_{\hat{\beta}} \sup_{\beta: \|\beta\|_0 \leq k} \text{MSE}(\hat{\beta}, \beta) = (1 + o(1))2\sigma^2k \log(p/k)$$
Minimaxity of soft thresholding estimator

Consider $\hat{\beta}_i = \psi_{st}(y_i; \lambda)$ with $\lambda = \sigma \sqrt{2 \log p}$ as before

- If $\beta = 0$, one has $\hat{\beta} \approx 0$ as discussed before
- If $\beta_1 \gg \beta_2 \gg \cdots \gg \beta_k \gg \sigma$ and $\beta_{k+1} = \cdots = \beta_p = 0$, then

$$
\hat{\beta}_i \approx \begin{cases}
y_i - \lambda, & \text{if } i \leq k \\
0, & \text{else}
\end{cases}
$$

$$
\Rightarrow \quad \text{MSE}(\hat{\beta}, \beta) \approx \sum_{i=1}^{k} \mathbb{E} \left[ (y_i - \beta_i - \lambda)^2 \right] = k(\sigma^2 + \lambda^2)
$$

$$
= k\sigma^2 (2 \log p + 1) > 2k\sigma^2 \log \left( \frac{p}{k} \right)
$$

○ Need to pick a smaller threshold $\lambda$
Minimaxity of soft thresholding estimator

Theorem 7.10

Suppose $X = I$. For any $k$-sparse $\beta$ with $k \ll p$, the soft thresholding estimator $\hat{\beta}_i = \psi_{st}(y_i; \lambda)$ with $\lambda = \sigma \sqrt{2 \log(p/k)}$ obeys

$$\text{MSE}(\hat{\beta}, \beta) \leq (1 + o(1)) 2\sigma^2 k \log(p/k)$$

- Threshold $\lambda$ determined by sparsity level
Sanity check

If $\beta_1 \gg \cdots \gg \beta_k \gg \sigma$ and $\beta_{k+1} = \cdots = \beta_p = 0$, then

$$\hat{\beta}_i \approx \begin{cases} y_i - \lambda, & \text{if } i \leq k \\ 0, & \text{else} \end{cases}$$

$$\Rightarrow \quad \text{MSE}(\hat{\beta}, \beta) \approx k(\sigma^2 + \lambda^2) \quad \text{(as shown before)}$$

$$= k\sigma^2(2 \log(p/k) + 1)$$

$$\approx 2k\sigma^2 \log(p/k) \quad \text{minimax risk}$$
Proof of Theorem 7.11

WLOG, suppose $\sigma = 1$. Under the sparsity constraint,

$$\text{MSE}(\hat{\beta}, \beta) = \sum_{i=1}^{p} r_{st}(\lambda, \beta_i) = \sum_{i: \beta_i \neq 0} r_{st}(\lambda, \beta_i) + (p - k) r_{st}(\lambda, 0)$$

$$\leq k r_{st}(\lambda, \infty) + (p - k) r_{st}(\lambda, 0) \quad (7.9)$$

$$\leq k \left(1 + \lambda^2\right) + 2p \frac{\phi(\lambda)}{\lambda} \quad (7.10)$$

$$= (1 + o(1)) 2k \log(p/k) + \frac{k}{\sqrt{\pi \log(p/k)}}$$

$$= (1 + o(1)) 2k \log(p/k),$$

where (7.9) follows since $r_{st}(\lambda, \beta)$ is increasing in $\beta$, and (7.10) comes from (7.7) and Lemma 7.6.
Adaptivity to unknown sparsity

- **Problem of “optimal” soft thresholding:** knowing the sparsity level *a priori* is often unrealistic

- **Question:** can we develop an estimator that is adaptive to unknown sparsity?

Adaptivity cannot be achieved via soft thresholding with fixed thresholds, but what if we adopt data-dependent thresholds?
Let $|y|(1) \geq \cdots \geq |y|(p)$ be the order statistics of $|y_1|, \cdots, |y_p|$

**Key idea:** use a different threshold for $y_i$ based on its rank

$$\hat{\beta}_i = \psi_{st}(y_i; \lambda_j) \quad \text{if} \quad |y_i| = |y|(j) \quad (7.11)$$

- originally due to Benjamini & Hochberg '95 for controlling false discovery rate
How to set thresholds? (non-rigorous)

Consider $k$-sparse vectors, and WLOG suppose $\sigma = 1$. Recall that when we use soft thresholding with $\lambda = \sqrt{2 \log(p/k)}$, the least favorable signal is

$$\beta_1 \gg \cdots \gg \beta_k \gg \sigma \quad \text{and} \quad \beta_{k+1} = \cdots = \beta_p = 0 \quad (7.12)$$

If we use data-dependent thresholds and if $\{\lambda_i\}_{i > k}$ are sufficiently large, then

$$\hat{\beta}_i \approx \begin{cases} 
  y_i - \lambda_i, & \text{if } i \leq k \\
  0, & \text{else}
\end{cases}$$

$$\text{MSE}(\hat{\beta}, \beta) \approx \sum_{i=1}^{k} \mathbb{E}[(y_i - \lambda_i - \beta_i)^2] = \sum_{i=1}^{k} \left(1 + \lambda_i^2\right)$$
How to set thresholds? (non-rigorous)

If the estimator is minimax for each $k$ and if the worst-case $\beta$ for each $k$ is given by (7.12), then

$$\text{MSE}(\hat{\beta}, \beta) \approx 2k \log(p/k), \quad k = 1, \cdots, p$$

$$\Rightarrow \sum_{i=1}^{k} \lambda_i^2 \approx 2k \log(p/k), \quad k = 1, \cdots, p$$

This suggests a choice (think of $\lambda_i^2$ as the derivative of $g(x) := 2x \log(p/x)$)

$$\lambda_i^2 \approx 2 \log(p/i) - 2 \approx 2 \log(p/i)$$
How to set thresholds? (non-rigorous)

When $\beta_1 = \cdots = \beta_{50} = 0$, $\sigma = 1$
How to set thresholds? (non-rigorous)

When $\beta_1 = \cdots = \beta_5 = 3$, $\beta_6 = \cdots = \beta_{50} = 0$, $\sigma = 1$
Minimaxity

Theorem 7.11 (Abramovich ’06, Su & Candes ’16)

Suppose $X = I$, and $k \ll p$. The estimator (7.11) with 
$\lambda_i = \sigma \sqrt{2 \log(p/i)}$ is minimax, i.e.

$$\text{MSE}(\hat{\beta}, \beta) = (1 + o(1))2\sigma^2k \log(p/k)$$

- Adaptive to unknown sparsity
SLOPE (Sorted L-One Penalized Estimation): a generalization of LASSO

\[
\text{minimize}_{\hat{\beta} \in \mathbb{R}^p} \quad \frac{1}{2} \| y - X \hat{\beta} \|^2_2 + \lambda_1 |\hat{\beta}|_1 + \lambda_2 |\hat{\beta}|_2 + \cdots + \lambda_p |\hat{\beta}|_p
\]

where \( \lambda_i = \sigma \Phi^{-1}(1 - iq/(2p)) \approx \sigma \sqrt{2 \log(p/i)}, 0 < q < 1 \) is constant, and \( \Phi \) is CDF of \( \mathcal{N}(0, 1) \)

- This is a convex program if \( \lambda_1 \geq \cdots \geq \lambda_p \geq 0 \) (homework)
- This can be computed efficiently via proximal methods
- SLOPE is minimax and adaptive to unknown sparsity under i.i.d. Gaussian design \( X \)
Reference

Reference


