The LASSO
CSci 8980: ML at Large Scale and High Dimensions

Instructor: Arindam Banerjee

January 29, 2014
Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
Regression with OLS

- Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
- Ordinary least squares (OLS)

\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2
\]

Issues/challenges with OLS:
- Accuracy: low bias, high variance
- Interpretation: All coefficients are non-zero
- Cannot determine small subsets with strong effects

Shrinking coefficients:
- Increases bias, lowers variance, improves accuracy

Alternatives:
- Subset selection: Unstable, sensitive to small changes
- Ridge regression: Shrinks coefficients, but not to 0

Instructor: Arindam Banerjee
Regression with OLS

- Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
- Ordinary least squares (OLS)

\[
\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2
\]

- Issues/challenges with OLS
Regression with OLS

- Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
- Ordinary least squares (OLS)
  \[
  \hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2
  \]
- Issues/challenges with OLS
  - Accuracy: low bias, high variance
Regression with OLS

- Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
- Ordinary least squares (OLS)

\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2
\]

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
Regression with OLS

- Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
- Ordinary least squares (OLS)

\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2
\]

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects
Regression with OLS

- Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
- Ordinary least squares (OLS)

\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2
\]

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects
- Shrinking coefficients

Instructor: Arindam Banerjee
Regression with OLS

- Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
- Ordinary least squares (OLS)
  \[
  \hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2
  \]

Issues/challenges with OLS
- Accuracy: low bias, high variance
- Interpretation: All coefficients are non-zero
- Cannot determine small subsets with strong effects

Shrinking coefficients
- Increases bias, lowers variance, improves accuracy

Instructor: Arindam Banerjee
Regression with OLS

- Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
- Ordinary least squares (OLS)

\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2
\]

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects
- Shrinking coefficients
  - Increases bias, lowers variance, improves accuracy
- Alternatives

Instructor: Arindam Banerjee

The LASSO
Regression with OLS

- Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
- Ordinary least squares (OLS)

\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2
\]

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects

- Shrinking coefficients
  - Increases bias, lowers variance, improves accuracy

- Alternatives
  - Subset selection: Unstable, sensitive to small changes
Regression with OLS

- Given training data \((y_i, x_i), i = 1, \ldots, n, x_i \in \mathbb{R}^p\)
- Ordinary least squares (OLS)
  \[
  \hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2
  \]

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects

- Shrinking coefficients
  - Increases bias, lowers variance, improves accuracy

- Alternatives
  - Subset selection: Unstable, sensitive to small changes
  - Ridge regression: Shrinks coefficients, but not to 0
Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0|$
The LASSO

- Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0_j|$
- The non-negative garotte estimator (Breiman, 1996)

$$(\hat{\alpha}, \hat{c}) = \arg\min_{(\alpha, c)} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j} c_j \hat{\beta}^0_j x_{ij})^2 \text{ s.t. } c_j \geq 0, \sum_{j} c_j \leq t$$
Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0_j|$

The non-negative garotte estimator (Breiman, 1996)

$$(\hat{\alpha}, \hat{c}) = \arg\min_{(\alpha, c)} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j} c_j \hat{\beta}^0_j x_{ij})^2 \text{ s.t. } c_j \geq 0, \sum_j c_j \leq t$$

Relies on OLS $\hat{\beta}^0$: may be problematic in certain settings

Parameter $t < t_0$ will cause shrinkage

Some coefficients will become 0
The LASSO

- Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0_j|$
- The non-negative garotte estimator (Breiman, 1996)

$$(\hat{\alpha}, \hat{c}) = \arg\min_{(\alpha, c)} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j} c_j \hat{\beta}_j^0 x_{ij})^2 \text{ s.t. } c_j \geq 0, \sum_{j} c_j \leq t$$

- Relies on OLS $\hat{\beta}^0$: may be problematic in certain settings
- Least absolute shrinkage and selection operator (LASSO)

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha, \beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j} \beta_j x_{ij})^2 \text{ s.t. } \sum_{j} |\beta_j| \leq t$$
The LASSO

- Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0_j|$
- The non-negative garotte estimator (Breiman, 1996)
  
  $$(\hat{\alpha}, \hat{c}) = \arg\min_{(\alpha,c)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j c_j \hat{\beta}^0_j x_{ij})^2 \text{ s.t. } c_j \geq 0, \sum j c_j \leq t$$
  
  Relies on OLS $\hat{\beta}^0$: may be problematic in certain settings
- Least absolute shrinkage and selection operator (LASSO)
  
  $$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha,\beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \text{ s.t. } \sum_j |\beta_j| \leq t$$
  
  Parameter $t < t_0$ will cause shrinkage
The LASSO

- Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j^0|$
- The non-negative garrotte estimator (Breiman, 1996)

$$(\hat{\alpha}, \hat{c}) = \arg\min_{(\alpha, c)} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j} c_j \hat{\beta}_j^0 x_{ij})^2 \text{ s.t. } c_j \geq 0, \sum_{j} c_j \leq t$$

- Relies on OLS $\hat{\beta}^0$: may be problematic in certain settings
- Least absolute shrinkage and selection operator (LASSO)

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha, \beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j} \beta_j x_{ij})^2 \text{ s.t. } \sum_{j} |\beta_j| \leq t$$

- Parameter $t < t_0$ will cause shrinkage
  - Some coefficients will become 0
Design matrix \( X \in \mathbb{R}^{n \times p} \), assume \( X^T X = I \in \mathbb{R}^{p \times p} \)
Orthonormal Design Case

- Design matrix $X \in \mathbb{R}^{n \times p}$, assume $X^T X = I \in \mathbb{R}^{p \times p}$
- Best subset selection picks $k$ largest coefficients
Orthonormal Design Case

- Design matrix $X \in \mathbb{R}^{n \times p}$, assume $X^T X = I \in \mathbb{R}^{p \times p}$
- Best subset selection picks $k$ largest coefficients
- For a suitable constant $\gamma$, the LASSO solution is
  \[
  \hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+
  \]
**Orthonormal Design Case**

- Design matrix $X \in \mathbb{R}^{n \times p}$, assume $X^T X = I \in \mathbb{R}^{p \times p}$
- Best subset selection picks $k$ largest coefficients
- For a suitable constant $\gamma$, the LASSO solution is
  \[ \hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+ \]
- Ridge regression shrinks the coefficients
  \[ \hat{\beta}_j^{\text{ridge}} = \frac{1}{1 + \gamma \hat{\beta}_j^0} \]
Orthonormal Design Case

- Design matrix $X \in \mathbb{R}^{n \times p}$, assume $X^T X = I \in \mathbb{R}^{p \times p}$
- Best subset selection picks $k$ largest coefficients
- For a suitable constant $\gamma$, the LASSO solution is
  \[ \hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+ \]
- Ridge regression shrinks the coefficients
  \[ \hat{\beta}_{\text{ridge}}^j = \frac{1}{1 + \gamma \hat{\beta}_j^0} \]
- Garotte estimates
  \[ \hat{\beta}_{\text{garotte}}^j = \left(1 - \frac{\gamma}{(\hat{\beta}_j^0)^2}\right)_+ \hat{\beta}_j^0 \]
Orthonormal Design Case

Shrinkage due to (a) subset selection, (b) ridge regression, (c) the lasso, and (b) the garotte
Elliptical contour of the objective

\[(\beta - \hat{\beta}^0)^T X^T X (\beta - \hat{\beta}^0)\]
Elliptical contour of the objective

$$(\beta - \hat{\beta}^0)^T X^T X (\beta - \hat{\beta}^0)$$

Level sets of the contour intersects with $L_q$ norm ball
Geometry of LASSO

Elliptical contour of the objective

\[(\beta - \hat{\beta}^0)^T X^T X (\beta - \hat{\beta}^0)\]

Level sets of the contour intersects with \(L_q\) norm ball
- \(q = 2\): Ridge regression, shrinkage but no sparsity
Geometry of LASSO

- Elliptical contour of the objective
  \[(\beta - \hat{\beta}^0)^T X^T X (\beta - \hat{\beta}^0)\]

- Level sets of the contour intersects with $L_q$ norm ball
  - $q = 2$: Ridge regression, shrinkage but no sparsity
  - $q = 1$: Lasso, shrinkage and sparsity
Elliptical contour of the objective

\[(\beta - \hat{\beta}^0)^T X^T X (\beta - \hat{\beta}^0)\]

Level sets of the contour intersects with $L_q$ norm ball
- $q = 2$: Ridge regression, shrinkage but no sparsity
- $q = 1$: Lasso, shrinkage and sparsity

Ridge vs Lasso: Can the sign change from OLS estimate?
Geometry of LASSO: $p = 2$

Estimation in (a) the lasso, and (b) ridge regression
Geometry of LASSO: $p > 2$

Sign change in LASSO vs OLS is possible for $p > 2$
Example: Regularization Path

Shrinkage of parameters over $s = \frac{t}{\sum_j \beta_j^0}$
Estimating “t”

- The ‘regularized’ version of Lasso

\( (\hat{\alpha}, \hat{\beta}) = \text{argmin}_{(\alpha, \beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j| \)
Estimating “t”

- The ‘regularized’ version of Lasso

\[(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha, \beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|\]

- Cross-validation over \(\lambda\) (or \(t\))
Estimating “t”

- The ‘regularized’ version of Lasso

\[
(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha, \beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|
\]

- Cross-validation over $\lambda$ (or $t$)
  - Pick the value that leads to smallest error

Instructor: Arindam Banerjee
The ‘regularized’ version of Lasso

\[
(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha, \beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|
\]

- Cross-validation over \( \lambda \) (or \( t \))
  - Pick the value that leads to smallest error
- Resampling based estimates, e.g., stability selection
Generalized Regression Models

- General regression problem formulation

Examples: logistic regression, generalized linear models, etc.

We will consider efficient algorithms for such general problems.

Instructor: Arindam Banerjee
Generalized Regression Models

- General regression problem formulation
  - Constrained version
    \[
    \hat{\beta} = \arg\min_{\beta} L(y, X, \beta) \text{ s.t. } \|\beta\|_1 \leq t
    \]

Examples: logistic regression, generalized linear models, etc.

We will consider efficient algorithms for such general problems.
Generalized Regression Models

- General regression problem formulation
  - Constrained version
    \[ \hat{\beta} = \arg\min_{\beta} L(y, X, \beta) \text{ s.t. } \|\beta\|_1 \leq t \]
  - Regularized version
    \[ \hat{\beta} = \arg\min_{\beta} L(y, X, \beta) + \lambda \|\beta\|_1 \]

Examples: logistic regression, generalized linear models, etc.

Instructor: Arindam Banerjee

The LASSO
Generalized Regression Models

- General regression problem formulation
  - Constrained version
    \[ \hat{\beta} = \arg\min_{\beta} L(y, X, \beta) \text{ s.t. } \|\beta\|_1 \leq t \]
  - Regularized version
    \[ \hat{\beta} = \arg\min_{\beta} L(y, X, \beta) + \lambda \|\beta\|_1 \]
  - The other constrained version
    \[ \hat{\beta} = \arg\min_{\beta} \|\beta\|_1 \text{ s.t. } L(y, X, \beta) \leq a \]

Examples: logistic regression, generalized linear models, etc.

We will consider efficient algorithms for such general problems.

Instructor: Arindam Banerjee
Generalized Regression Models

- General regression problem formulation
  - Constrained version
    \[ \hat{\beta} = \arg\min_{\beta} L(y, X, \beta) \text{ s.t. } \|\beta\|_1 \leq t \]
  - Regularized version
    \[ \hat{\beta} = \arg\min_{\beta} L(y, X, \beta) + \lambda \|\beta\|_1 \]
  - The other constrained version
    \[ \hat{\beta} = \arg\min_{\beta} \|\beta\|_1 \text{ s.t. } L(y, X, \beta) \leq a \]
- Examples: logistic regression, generalized linear models, etc.
Generalized Regression Models

- General regression problem formulation
  - Constrained version
    \[
    \hat{\beta} = \arg\min_{\beta} L(y, X, \beta) \text{ s.t. } \|\beta\|_1 \leq t
    \]
  - Regularized version
    \[
    \hat{\beta} = \arg\min_{\beta} L(y, X, \beta) + \lambda \|\beta\|_1
    \]
  - The other constrained version
    \[
    \hat{\beta} = \arg\min_{\beta} \|\beta\|_1 \text{ s.t. } L(y, X, \beta) \leq a
    \]
- Examples: logistic regression, generalized linear models, etc.
- We will consider efficient algorithms for such general problems
Consider orthonormal design $X^T X = I$, so Lasso estimate is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$
Consider orthonormal design $X^TX = I$, so Lasso estimate is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

Let $\beta$ be the ‘true’ parameter:

$$y = \beta^T x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$
Consider orthonormal design $X^TX = I$, so Lasso estimate is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

Let $\beta$ be the ‘true’ parameter:

$$y = \beta^T x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Consider risk

$$R(\hat{\beta}, \beta) = E \| \hat{\beta} - \beta \|^2$$
Consider orthonormal design $X^TX = I$, so Lasso estimate is
\[
\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma) + 
\]

Let $\beta$ be the ‘true’ parameter:
\[
y = \beta^T x + \epsilon, \quad \epsilon \sim N(0, \sigma^2) 
\]

Consider risk
\[
R(\hat{\beta}, \beta) = E \| \hat{\beta} - \beta \|^2 
\]

Let $R_{DP}$ be the loss of the ‘optimal’ predictor
\[
T_{DP}(\hat{\beta}^0, \delta) = (\delta_j\hat{\beta}_j^0), \quad \delta_j = I(|\beta_j| > \sigma) \in \{0, 1\} 
\]
Consider orthonormal design $X^T X = I$, so Lasso estimate is
\[ \hat{\beta}_j = \text{sign}(\hat{\beta}^0_j)(|\hat{\beta}^0_j| - \gamma)_+ \]

Let $\beta$ be the ‘true’ parameter:
\[ y = \beta^T x + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \]

Consider risk
\[ R(\hat{\beta}, \beta) = E\|\hat{\beta} - \beta\|^2 \]

Let $R_{DP}$ be the loss of the ‘optimal’ predictor
\[ T_{DP}(\hat{\beta}^0, \delta) = (\delta_j \hat{\beta}_j^0), \quad \delta_j = I(|\beta_j| > \sigma) \in \{0, 1\} \]

$T_{DP}$ needs knowledge of $\beta$, not practical
Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma)$
Bounds on the Risk: Donoho et al.

- Hard threshold estimator \( \tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma) \)

  - Has risk

\[
R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})
\]
Bounds on the Risk: Donoho et al.

- Hard threshold estimator \( \hat{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma) \)
  - Has risk \( R(\hat{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP}) \)
  - Threshold \( \gamma = \sigma(2 \log n)^{1/2} \) to get smallest asymptotic risk
Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma)$
- Has risk $R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})$
- Threshold $\gamma = \sigma(2 \log n)^{1/2}$ to get smallest asymptotic risk
Soft threshold estimator $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$
Bounds on the Risk: Donoho et al.

- Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma)$
  - Has risk $R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})$
  - Threshold $\gamma = \sigma (2 \log n)^{1/2}$ to get smallest asymptotic risk

- Soft threshold estimator $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$
  - With $\gamma = \sigma (2 \log n)^{1/2}$, has same behavior
Bounds on the Risk: Donoho et al.

- Hard threshold estimator \( \tilde{\beta}_j = \tilde{\beta}_j^0 I(|\tilde{\beta}_j^0| > \gamma) \)
  - Has risk \( R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP}) \)
  - Threshold \( \gamma = \sigma (2 \log n)^{1/2} \) to get smallest asymptotic risk

- Soft threshold estimator \( \hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+ \)
  - With \( \gamma = \sigma (2 \log n)^{1/2} \), has same behavior

- General design matrices
Hard threshold estimator $\hat{\beta}_j = \hat{\beta}_j^0 \mathbb{I}(|\hat{\beta}_j^0| > \gamma)$

- Has risk $R(\hat{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})$

- Threshold $\gamma = \sigma (2 \log n)^{1/2}$ to get smallest asymptotic risk

Soft threshold estimator $\tilde{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$

- With $\gamma = \sigma (2 \log n)^{1/2}$, has same behavior

General design matrices
- Lasso estimator continues to have good properties
- **Hard threshold estimator** \( \tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma) \)
  - Has risk \( R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP}) \)
  - Threshold \( \gamma = \sigma (2 \log n)^{1/2} \) to get smallest asymptotic risk
- **Soft threshold estimator** \( \hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+ \)
  - With \( \gamma = \sigma (2 \log n)^{1/2} \), has same behavior
- **General design matrices**
  - Lasso estimator continues to have good properties
  - Generalized to other sparsity inducing norms
$L_q$ norm level sets: (a) $q = 4$, (b) $q = 2$, (c) $q = 1$, (d) $q = 0.5$, (e) $q = 0.1$