The Adaptive Lasso and its Oracle Properties

Hui Zou, 2006

Beyond Lasso Reading Group
UofA, Edmonton

Barnabás Póczos

April 2, 2009
Brief Summary

- Necessary conditions for Lasso variable selection to be consistent.

- Scenarios where Lasso variable selection is inconsistent.

- Lasso cannot be an oracle procedure.
  - Consistent variable selection
  - Performs as well as if true model were given
    → new version of Lasso, Adaptive Lasso.

- Adaptive Lasso enjoys the oracle properties.
Brief Summary

- Adaptive Lasso can be solved by the same algorithms as Lasso.

- Adaptive Lasso can be extended to Generalized Linear Models, and the oracle properties will still hold.

- Nonnegative Garotte is shown to be consistent for variable selection.
Notations

\[
y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \text{response vector} \quad n : \# \, \text{Patients}
\]

\[
X = \begin{bmatrix} x_1, \ldots, x_p \end{bmatrix}, \text{linearly independent predictors} \quad p : \# \, \text{Features}
\]

\[
x_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}, \text{feature } j
\]

Assume \( \mathbb{E}[y|x] = \beta_1^* x_1 + \ldots + \beta_p^* x_p \)

i.e assume for patient \( i \) that

\[
\mathbb{E}[y_i|X_{i,:}] = \beta_1^* x_{1i} + \ldots + \beta_p^* x_{ip}
\]

\[
\mathbb{E}[y_i|X_{i,:}] = X_{i,:} \beta^*
\]

\( \mathcal{A} \doteq \{ j : \beta_j^* \neq 0 \} \). Assume sparsity: \( |\mathcal{A}| = p_0 < p \)
Definitions

- **Convergence in *Probability***

  \[ X_n \stackrel{p}{\to} X \iff \{ \forall \epsilon > 0 : \lim_{n \to \infty} P(\| X_n - X \| > \epsilon) = 0 \} \]

- **Convergence in *Distribution***

  \[ X_n \stackrel{d}{\to} X \iff \{ F_{X_n} \to w F_X \} \quad \text{The cdf-s are weakly convergent} \]

- **Weak** convergence of cdf-s

  \[ F_n(\cdot) \to w F(\cdot) \iff \{ \lim_{n \to \infty} F_n(x) = F(x), \forall x \in C \}, \]
  
  where \( C \) is the set of continuity points of \( F \).

- **Remark:**

  \[ X_n \stackrel{p}{\to} X \not\implies X_n \stackrel{d}{\to} X \]
Oracle Property

Definition: method $\delta$ is an *oracle procedure* if

- $\lim_{n \to \infty} P(\{j : \hat{\beta}_j(\delta) \neq 0\} = A) = 1$. Identifies the right subset model

- $\sqrt{n}(\hat{\beta}(\delta)_A - \beta^*_A) \to_d \mathcal{N}(0, \Sigma^*)$, where $\Sigma^*$ is the covariance matrix knowing the true subset model. Optimal root-n estimation rate

For an optimal procedure we need continuous shrinkage property, too.
Lasso

Regularization technique for simultaneous estimation and variable selection

- (L₁ penalization is also called basis pursuit)

- Theoretical work
  - **Pro:** L₁ approach is able to discover the right sparse representations (Donoho et. al. 2002, 2004)
  - **Contra:** Variable selection with Lasso can and cannot be consistent (Meinshausen and Bühlmann, 2004)
Asymptotic Properties of Lasso

Setup from Knight and Fu (2000)

- \( y_i = X_i \beta^* + \epsilon_i, \)

  where \( X \in \mathbb{R}^{n \times p}, \ y_i \in \mathbb{R}, \ \beta^* \in \mathbb{R}^p, \ X_{i,:} \in \mathbb{R}^{1 \times p} \)
  
  \( \epsilon_i \in \mathbb{R} \text{ iid, } i = 1, \ldots, n \)
  
  \( \mathbb{E}[\epsilon_i] = 0, \ \text{Var}[\epsilon_i] = \sigma^2 \)

- \( \frac{1}{n} X^T X \rightarrow C, \text{ where } C > 0, \ C \in \mathbb{R}^{p \times p} \text{ pos. def. matrix} \)

- Active components: without loss of generality assume
  
  \[ A = \{1, 2, \ldots, p_0\}, \ p_0 < p \]
  
  \[ C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \text{ where } C_{11} \in \mathbb{R}^{p_0 \times p_0} \]
Asymptotic Properties of Lasso  (Knight & Fu, 2000)

Lasso estimate for \( n \) variables:

\[
\hat{\beta}^{(n)} = \arg \min_{\beta} \| y - \sum_{j=1}^{p} x_j \beta_j \|^2 + \lambda \sum_{j=1}^{p} |\beta_j|
\]

Lasso selected variables:

\[ A_n \doteq \{ j : \hat{\beta}_j^{(n)} \neq 0 \} \]

Definition: Lasso variable selection is consistent

\[
\lim_{n \to \infty} \mathbb{P}(A_n = A) = 1
\]
Lemma 1 (about estimation consistency)

\[ \left\{ \frac{\lambda_n}{n} \to \lambda_0 \geq 0 \right\} \Rightarrow \left\{ \hat{\beta}^{(n)} \to_p u^* = \arg \min_u V_1(u) \right\} \]

where

\[ V_1(u) = (u - \beta^*)^T C (u - \beta^*) + \lambda_0 \sum_{j=1}^{p} |u_j|. \]

Def: \( \left\{ \hat{\beta}^{(n)} \to_p u^* \right\} \Leftrightarrow \{ \forall \epsilon > 0 : \lim_{n \to \infty} \mathbb{P}(\| \hat{\beta}^{(n)} - u^* \| > \epsilon) = 0 \}. \)

Note: \( u^* \) is a deterministic quantity.

- \( C > 0, \Rightarrow \) only \( \lambda_0 = 0 \) guarantees estimation consistency.
- \( \lambda_n = o(n) \) is sufficient for estimation consistency.
Asymptotic Properties of Lasso (Knight & Fu, 2000)

Lemma 2 (about root-n estimation consistency)

\[
\left\{ \frac{\lambda_n}{\sqrt{n}} \to \lambda_0 \geq 0 \right\} \Rightarrow \left\{ \sqrt{n} \left( \hat{\beta}^{(n)} - \beta^* \right) \to_d u^* = \arg \min_u V_2(u) \right\}
\]

where \( W \sim \mathcal{N}(0, \sigma^2 C) \) vector valued random variable, and

\[
V_2(u) = -2 u^T W + u^T Cu + \lambda_0 \sum_{j=1}^{p} \left[ u_j \text{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0) \right]
\]

Def: \( \{ \hat{\beta}^{(n)} \to_d u^* \} \iff \{ F_{\hat{\beta}^{(n)}}(x) \to F_{u^*}(x), \forall x \in C_{F_{u^*}} \} \)

Note: \( u^* \) is a random quantity.

\( \Rightarrow \) The Lasso estimation is root-n consistent if \( \frac{\lambda_n}{\sqrt{n}} \to \lambda_0 \geq 0 \)
Lasso Variable Selection can be Inconsistent

- **Proposition 1** (root-n estimation consistent Lasso cannot be selection consistent…grrrr..):

\[
\left\{ \frac{\lambda_n}{\sqrt{n}} \to \lambda_0 \geq 0 \right\} \Rightarrow \left\{ \limsup_n \mathbb{P}\{A_n = A\} \leq c < 1 \right\}
\]

where \( c \) is a constant depending on the true model, but does not depend on \( n \).
Lemma 3 (about not too slow, not too fast $\lambda_n$)

$$\frac{\lambda_n}{n} \to 0$$
$$\frac{\lambda_n}{\sqrt{n}} \to \infty$$

$$\tag*{\Rightarrow} \left\{ \frac{n}{\lambda_n} (\widehat{\beta}(n) - \beta^*) \to_p u^* = \arg\min_u V_3(u) \right\}$$

$$V_3(u) = u^T C u + \sum_{j=1}^{p} [u_j \text{sgn}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0)]$$

- $u^*$ is a deterministic quantity.
- The estimation is consistent, but the convergence rate $\frac{n}{\lambda_n} < \sqrt{n}$
- The optimal rate is when $\lambda_n = O(\sqrt{n})$ but then we already know the variable selection is inconsistent.
$\lambda_n \sim n^{1/4}$ \quad \Rightarrow \quad \frac{\lambda_n}{\sqrt{n}} \rightarrow \lambda_0 \geq 0 \Rightarrow \quad \text{root-n consistent estimation selection is not consistent}$

$\lambda_n \sim n^{1/2}$ \quad \Rightarrow \quad \left\{ \begin{array}{l}
\frac{\lambda_n}{\sqrt{n}} \rightarrow \infty \\
\frac{\lambda_n}{n} \rightarrow \lambda_0 \geq 0
\end{array} \right\} \Rightarrow \quad \left\{ \begin{array}{l}
\frac{n}{\lambda_n} \quad \text{rate, slower than } \sqrt{n}
\end{array} \right\}$

$\lambda_n \sim n^{3/4}$ \quad \Rightarrow \quad \null \text{model estimation, not consistent}$
The previous lemmas investigated the estimation consistency. We need conditions for selection consistency, too.

**Theorem 1**
(Necessary condition for selection consistency)

\[
\lim_{n \to \infty} P(A_n = A) = 1 \quad \text{(selection is consistent)} \quad \Rightarrow \quad \exists s = (s_1, \ldots, s_{p_0}) \in \{-1, 1\}^{p_0} \text{ sign vector such that } |C_{21}C_{11}^{-1}s| \leq 1 \quad \text{(componentwise)}
\]

When this condition fails \(\Rightarrow\) Lasso selection is not consistent

- Further results on model selection consistency:
  - Zhao & Yu (2006) + Zhao’s JMLR article published today
Example when Lasso Selection is Not Consistent

**Corollary 1**
(An example when Lasso selection is not consistent)

\[
C = \begin{pmatrix}
C_{11} & C_{12} \\
C_{21} & C_{22}
\end{pmatrix}
\]

\[
C_{11} = (1 - \rho_1)I_{p_0} + \rho_1 J_{p_0} \in \mathbb{R}^{p_0 \times p_0}
\]

\[
C_{12} = \rho_2 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{p_0}
\]

\[
C_{22} = 1
\]

where \(I_{p_0} \in \mathbb{R}^{p_0 \times p_0}\) is the identity, \(J_{p_0} \in \mathbb{R}^{p_0 \times p_0}\) is the all 1 matrix.

\[
p_0 = 2m + 1, p = p_0 + 1
\]

\[
\frac{-1}{p_0 - 1} < \rho_1 < \frac{-1}{p_0}
\]

\[
1 + (p_0 - 1)\rho_1 < |\rho_2| < \sqrt{1 + (p_0 - 1)\rho_1/p_0}
\]

\[
\implies \begin{cases} 
\triangleleft \text{Condition of Theorem 1} \\
\bullet \text{inconsistent selection}
\end{cases}
\]
Special Cases

- Orthogonal design guarantees the consistency of Lasso selection.

Two features (p=2)
- Necessary condition is satisfied.
- ∃ closed form Lasso solution.
- If \( \lambda_n \) is appropriately chosen \( \Rightarrow \) selection is consistent.
Meinshausen & Bühlman:
In Lasso there is a conflict between

- Optimal prediction
- Consistent variable selection

*Optimal $\lambda$ for prediction gives inconsistent variable selection results, many noise features will be included in the predictive models.*
Weighted Lasso

- We have shown that Lasso cannot be an oracle procedure. Let us try to fix this.

- **Definition: Weighted Lasso**

  Weighted Lasso estimate for $n$ variables:

  $$
  \text{arg min}_{\beta} \| y - \sum_{j=1}^{p} x_j \beta_j \|^2 + \lambda n \sum_{j=1}^{p} w_j |\beta_j| 
  $$

  $w = (w_1, \ldots, w_p)$ known weight vector. $w_i \geq 0$

- **Proposition:** $w$ weights are data dependent and cleverly chosen $\Rightarrow$ Weighted Lasso can have oracle properties.
Adaptive Lasso (AdaLasso)

Let $\hat{\beta}$ a root-n consistent estimator to $\beta^*$, e.g. $\hat{\beta}(OLS)$. Let $\gamma > 0$, $\hat{w}_j = |\hat{\beta}_j|^{-\gamma} \geq 0$.

The Adaptive Lasso estimation:

$$
\hat{\beta}^*(n) = \arg\min_{\beta} \|y - \sum_{j=1}^{p} x_j \beta_j\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j|
$$

$$
A^*_n = \{j : \hat{\beta}^*_j(n) \neq 0\}. \text{ (Adaptive Lasso estimated model)}
$$

Remark: Adaptive Lasso is still a convex optimization problem, LARS can be used for solving it.
The Oracle Properties of Adaptive Lasso

**Theorem 2**

Width the proper choice of $\lambda_n$ the Adaptive Lasso enjoys the oracle properties.

$\begin{align*}
\lambda_n n^{-1/2} &\to 0 \\
\lambda_n n^{(-1/2+\gamma/2)} &\to \infty
\end{align*}
\begin{align*}
\Rightarrow \\
\lim_{n \to \infty} \mathbb{P}(A_n^* = A) &= 1 \\
\text{(consistent selection)} \\
\sqrt{n} \left( \hat{\beta}_A^*(n) - \beta^*_A \right) &\to_d \mathcal{N}(0, \sigma^2 C_{11}^{-1}) \\
\text{asymptotic normality with root-n rate}
\end{align*}$

e.g. $\lambda_n \sim n^{(1/2-\gamma/4)}$
Remarks

- \( \hat{\beta} \) is not required to be root-n consistent for the Adaptive Lasso. This condition can be weakened.

- \( \hat{w}_j = \hat{\beta}_j^{-\gamma} \) is the key.

\[
\text{When } n \to \infty \Rightarrow \begin{cases} 
\hat{\beta}_i \to 0 \\
\text{or} \\
\hat{\beta}_i \to c 
\end{cases} \Rightarrow \begin{cases} 
\hat{w}_i = \frac{1}{\hat{\beta}_i^\gamma} \to \infty \\
\text{or} \\
\hat{w}_i = \frac{1}{\hat{\beta}_i^\gamma} \to 1/c < \infty 
\end{cases}
\]

- Adaptive Lasso is continuous.

(Discontinuity results in instability in model prediction)
Bridge regression with \( L_q \) penalty is not continuous when \( q < 1 \), but has the oracle properties when \( 0 < q < 1 \).
Thresholding Functions

(a) Hard
(b) Bridge, $L_{0.5}$
(c) Lasso
(d) SCAD
(e) Adaptive Lasso, $\gamma=0.5$
(f) Adaptive Lasso, $\gamma=2$
The Nonnegative Garotte Problem (NGP).

\[
\min_{c_1 \geq 0, \ldots, c_p \geq 0} \| y - \sum_{j=1}^{p} x_j \hat{\beta}_j (OLS) c_j \|^2 + \lambda_n \sum_{j=1}^{p} c_j
\]

NGP can be used for sparse modeling.
(Large \( \lambda_n \) shrinks some \( c_j \) to 0)

Corollary 2: The NGP is consistent for variable selection with proper \( \lambda_n \).

\[
\frac{\lambda_n}{\sqrt{n}} \to 0 \quad \frac{\lambda_n}{\sqrt{n}} \to \infty
\]
\[\Rightarrow\] NGP is consistent for variable selection.
Yuan & Lin also proved the variable selection consistency of NGP.

NGP can be considered as an Adaptive Lasso with $\gamma=1$, and additional sign constraints.

**Proof:**

The Adaptive Lasso estimation with $\gamma = 1$:

$$\hat{\beta}^*(n) = \arg \min_{\beta} \| y - \sum_{j=1}^{p} x_j \beta_j \|^2 + \lambda n \sum_{j=1}^{p} \frac{|\beta_j|}{\hat{\beta}_j(OLS)}$$

Let $\beta_j(NGP) \doteq c_j \hat{\beta}_j(OLS)$

NGP cost $\Rightarrow \begin{cases} 
\arg \min_{\beta} \| y - \sum_{j=1}^{p} x_j \beta_j \|^2 + \lambda n \sum_{j=1}^{p} \frac{|\beta_j|}{\hat{\beta}_j(OLS)} \\
such that \frac{\beta_j}{\hat{\beta}_j(OLS)} \geq 0, \quad \forall j = 1, \ldots, p
\end{cases}$
The Adaptive Lasso problem is a convex problem and can be solved e.g. by LARS.

The Adaptive Lasso estimation:

\[ \hat{\beta}^*(n) = \arg \min_{\beta} \| y - \sum_{j=1}^{p} x_j \beta_j \|^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{\hat{\beta}_j(\text{OLS})} \]

\[ \lambda_n \sum_{j=1}^{p} w_j |\beta_j| = \lambda_n \sum_{j=1}^{p} |\hat{\beta}_j| \]

- Let \( \hat{w}_j = \frac{1}{\hat{\beta}_j(\text{OLS})} \), \( x^{**}_j = x_j / \hat{w}_j \), \( \hat{\beta}^{**}_j = \hat{w}_j \hat{\beta}^*(n)_j \), \( j = 1, \ldots, p \)
- \( \hat{\beta}^{**} = \arg \min_{\tilde{\beta}} \| y - \sum_{j=1}^{p} x^{**}_j \tilde{\beta}_j \|^2 + \lambda_n \sum_{j=1}^{p} |\tilde{\beta}_j| \) (Lasso)
- OUTPUT: \( \hat{\beta}^*_j(n) = \hat{\beta}^{**}_j / \hat{w}_j \), \( j = 1, \ldots, p \)
Adaptive Lasso, Tuning

Tuning:

- we want to find the optimal \((\gamma, \lambda_n)\) for a given \(n\).
  \[ \Rightarrow \text{two-dimensional cross validation.} \]

(For a given \(\gamma\) LARS can calculate the entire path for \(\lambda_n\))

- In principle we can replace \(\beta(\text{OLS})\) to any other consistent estimator (e.g. \(\beta(\text{Ridge})\)).
  \[ \Rightarrow \text{three-dimensional cross validation} \ (\beta, \gamma, \lambda_n) \ldots \]

- In case of collinearity \(\beta(\text{Ridge})\) can be more stable than \(\beta(\text{OLS})\).
Other Oracle methods

- Oracle procedure goal
  - Simultaneously achieving:
    - Consistent variable selection
    - Optimal estimation prediction

- Other oracle methods:
  - Bach: Bolasso
  - Fan & Li: SCAD
  - Bridge ($q<1$)

- Oracle properties do not hold for Lasso usually
Numerical Experiments

Generative model: \( y = x^T \beta^* + N(0, \sigma^2) \)

We compare Adaptive Lasso, Lasso, SCAD, NGP

- Adaptive Lasso: OLS for adaptive weighting + LARS
- Lasso: LARS
- SCAD: LQA algorithm of Fan & Lin (2001)
- NGP: Quadratic Programming
Model 0, Inconsistent Lasso

\[ \beta^* = (5.6, 5.6, 5.6, 0)^T \]

\[ \rho_1 = -0.39 \quad \rho_2 = 0.23 \quad \Rightarrow \text{C does not allow consistent Lasso estimation} \]

The empirical probability of the solution path is containing the true model (100 repetitions)

<table>
<thead>
<tr>
<th></th>
<th>( n = 60, \sigma = 9 )</th>
<th>( n = 120, \sigma = 5 )</th>
<th>( n = 300, \sigma = 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>0.55</td>
<td>0.51</td>
<td>0.53</td>
</tr>
<tr>
<td>adalasso(( \gamma = 0.5 ))</td>
<td>0.59</td>
<td>0.68</td>
<td>0.93</td>
</tr>
<tr>
<td>adalasso(( \gamma = 1 ))</td>
<td>0.67</td>
<td>0.89</td>
<td>1</td>
</tr>
<tr>
<td>adalasso(( \gamma = 2 ))</td>
<td>0.73</td>
<td>0.97</td>
<td>1</td>
</tr>
<tr>
<td>adalasso(( \gamma ) by cv)</td>
<td>0.67</td>
<td>0.91</td>
<td>1</td>
</tr>
</tbody>
</table>

As \( n \) increases, \( \sigma \) decreases the variable selection problem is expected to become easier
Numerical Experiments, Relative Prediction Error

Model 1

\[ \beta^* = (3, 1.5, 0, 0, 2, 0, 0)^T \]

\[ x_i, \ i = 1, \ldots, n \text{ normal, iid} \]

\[ \text{corr}(X_{i,j_1}, X_{i,j_2}) = (1/2)^{|j_1-j_2|} \]

\[ \sigma = \begin{pmatrix} 1 \\ 3 \\ 6 \end{pmatrix} \Rightarrow \text{SNR} = \begin{pmatrix} 21.25 \\ 2.35 \\ 0.59 \end{pmatrix} \]

\[ n = 20, \ 60 \]

Model 2

\[ \beta^* = (.85, .85, 0, 0, .85, 0, 0)^T \]

\[ \sigma = \begin{pmatrix} 1 \\ 3 \\ 6 \end{pmatrix} \Rightarrow \text{SNR} = \begin{pmatrix} 14.46 \\ 1.61 \\ 0.40 \end{pmatrix} \]

\[ n = 40, \ 80 \]
Lasso performs best when SNR is low, ($\sigma$ is high)

Oracle methods are more accurate when SNR is high ($\sigma$ small)

Adaptive Lasso seems to be able to combine the strength of Lasso and SCAD

Medium SNR $\Rightarrow$ Adaptive Lasso outperforms SCAD and Garotte

High SNR $\Rightarrow$ adaptive Lasso outperforms Lasso

Adaptive Lasso is more stable than SCAD

None of the four methods can universally dominate the three others
Performance in Variable Selection

### Table 3. Median Number of Selected Variables for Model 1 With $n = 60$

<table>
<thead>
<tr>
<th></th>
<th>$\sigma = 1$</th>
<th></th>
<th>$\sigma = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C$</td>
<td>$l$</td>
<td>$C$</td>
</tr>
<tr>
<td>Truth</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Lasso</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Adaptive lasso</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>SCAD</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Garotte</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

NOTE: The column labeled “C” gives the number of selected nonzero components, and the column labeled “l” presents the number of zero components incorrectly selected into the final model.

- All of them can identify the true significant variables.
- Lasso tends to select two noise variables into the final model.
We can generalize AdaLasso from linear regression with Gaussian noise to Generalized Linear Models (GLM).

GLM is related to the Exponential Family:

A regular exponential family is a family of probability distributions on $\mathbb{R}^d$:

$$f_{\psi}(y|\theta) = \exp\{y^T \theta - \psi(\theta) - g_{\psi}(y)\}$$

$\psi$ is convex, the so-called cumulant function. $\theta$ is called natural parameter.

In GLM we assume that $\theta = x^T \beta^*$, $y \in \mathbb{R}$.

$$f_{\psi}(y|x, \beta^*) = \exp\{yx^T \beta^* - \psi(x^T \beta^*) - g_{\psi}(y)\}$$

Instead of the generative model with additive noise we had before…
Extension to GLM

The **Maximum Likelihood** estimation of a GLM is defined as:

\[
\hat{\beta}(MLE) = \arg \min_{\beta} \sum_{i=1}^{n} [-y_i(x_i^T \beta) + \psi(x_i^T \beta)]
\]

Let \( \gamma > 0 \), \( \tilde{w}_i = |\hat{\beta}_i(MLE)|^{-\gamma} > 0 \)

The **Adaptive Lasso** estimates of the GLM parameters:

\[
\hat{\beta}^*(n)(GLM) = \arg \min_{\beta} \sum_{i=1}^{n} [-y_i(x_i^T \beta) + \psi(x_i^T \beta)] + \lambda_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j|
\]

The **Logistic Regression** special case:

\[
\hat{\beta}^*(n)(Log) = \arg \min_{\beta} \sum_{i=1}^{n} [-y_i(x_i^T \beta + \log(1 + \exp(x_i^T \beta)))] + \lambda_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j|
\]

The **Poisson log-linear** regression special case:

\[
\hat{\beta}^*(n)(Poisson) = \arg \min_{\beta} \sum_{i=1}^{n} [-y_i(x_i^T \beta + \exp(x_i^T \beta))] + \lambda_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j|
\]
The Fisher Information matrix

Definition: Fisher Information matrix

\[
I(\theta) = \mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f(X|\theta) \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^T \right]
\]

\[A_n^* \doteq \{ j : \hat{\beta}_j^*(n) \neq 0 \}. \text{ (Adaptive Lasso estimated model)}\]

\[A \doteq \{ j : \beta_j^* \neq 0 \} = \{1, 2, \ldots, p_0\}. \text{ (True model)}\]

\[
I(\theta) = \begin{pmatrix}
I_{11} & I_{12} \\
I_{21} & I_{22}
\end{pmatrix}, \quad I_{11} \in \mathbb{R}^{p_0 \times p_0}
\]

Cramer Rao lower bound:
If \( \hat{\theta} \) is an unbiased estimator of \( \theta \) \( \Rightarrow \) \( \text{Var}[\hat{\theta}] \geq I(\theta)^{-1} \).

**MLE Asymptotic Normality:**
If \( \hat{\theta} \) is an MLE estimator of \( \theta \) \( \Rightarrow \) \( \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1}) \).
The Oracle Properties of AdaLasso on GLM

**Theorem 4**

Under some mild regularity conditions the AdaLasso estimates enjoys the oracle properties, if $\lambda_n$ is chosen appropriately.

$$
A^*_n = \{ j : \hat{\beta}^*_j(n) (GLM) \neq 0 \} \\
\frac{\lambda_n}{\sqrt{n}} \to 0 \\
\lambda_n n^{(\gamma-1)/2} \to \infty
$$

some regular conditions

$$
\Rightarrow \\
\sqrt{n}(\hat{\beta}^*_n - \beta^*_A) \to_d \mathcal{N}(0, I_{11}^{-1})
$$

asymptotic normality

$$
\lim_{n \to \infty} \mathbb{P}(A^*_n = A) = 1
$$

(consistency in variable selection)

$$
e.g. \lambda_n \sim n^{(1/2-\gamma/4)}
$$
GLM results, Logistic regression

\[ \beta^* = (3, 0, 0, 1.5, 0, 0, 7, 0, 0)^T \]

- n=200 observations
- 100 repetitions
- \( x_i, i = 1, \ldots, n \) normal, iid
- \( \text{corr}(X_{i,j_1}, X_{i,j_2}) = 0.75 \)

![Graph showing comparison between Lasso, SCAD, and AdaLasso methods.](image)

*Figure 2. Simulation Example: Logistic Regression Model.*
Not Discussed Parts from the Article

- Standard Error Formula
  - Definitions in Tibshirani (1996)

- Near Minimax Optimality
  - Definitions in Donoho and Johnstone (1994)

- Proofs
  - Bit sketchy…
  - Heavily use results from other articles
  - Some of these results can be found in other articles.
Conclusions

- Proposed Adaptive Lasso for simultaneous.
  - Estimation
  - Variable selection
- Lasso variable selection can be inconsistent.
- AdaLasso enjoys the oracle properties using weighted $L_1$ penalty.
- AdaLasso is still convex problem, LARS can be used for solving it.
- Oracle properties does **NOT** automatically result in optimal prediction: Lasso can be advantageous in difficult noisy problems.