Fast and Accurate $\ell_1$ Penalized Estimators

Andrew Barron and Cong Huang

Yale University, Department of Statistics

Presentation at Rutgers University, December 12, 2007
Outline

• Penalization for Least Squares and Log Likelihood Criteria
• The $\ell_1$ Penalized Greedy Pursuit (LPGP) algorithm
  Description of the algorithm
  Analysis of its performance
• Advantages and Disadvantages of LPGP
• Key Ideas of the Proof
• Risk Characterization
  What forms of penalty permit desirable risk bounds?
• Conclusion
ℓ₁ Penalized Least Squares

• Suppose the data are \((X_i, Y_i)_{i=1}^n\) and a library \(\mathcal{H} = \{h\}\) is given. Find a function in the linear span of \(\mathcal{H}\) to minimize the following objective function.

\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - \sum_{h} \beta_h h(X_i))^2 + \lambda \sum_{h} |\beta_h|
\]

• This optimization is also called the Lasso (Tibshirani 1996) and Basis Pursuit (Chen and Donoho 1996).
\[ \ell_1 \text{ Penalized Log Likelihood} \]

- For an exponential family with statistics taken from a given library \( \mathcal{H} \) of functions of the data \( X \).

- Find the parameters \( \beta \) with which these statistics are to be linearly combined to optimize the objective function

\[ -\log \text{likelihood}(\beta) + \lambda \sum_{h} |\beta_h| \]

$\ell_1$ Penalized Greedy Pursuit (LPGP)

First suppose the library $\mathcal{H}$ is normalized in that $\|h\| = 1$ for all $h \in \mathcal{H}$.

- Algorithm

  Initialize $\hat{f}_0 = 0$.

  Then for $m = 1, 2, \ldots$, iteratively, given $\hat{f}_{m-1} = \sum_{j=1}^{m-1} \beta_{j,m-1} h_j$, we seek

  \[
  \hat{f}_m(x) = \alpha \hat{f}_{m-1}(x) + \beta h(x)
  \]

  to minimize the objective function over choices of $h, \alpha, \beta$,

  \[
  \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha f_{m-1}(X_i) - \beta h(X_i))^2 + \lambda (|\beta| + \alpha \sum_{j=1}^{m-1} |\beta_{j,m-1}|)
  \]

  yielding $h_m, \alpha_m, \beta_{m,m}$ and $\beta_{j,m} = \alpha_m \beta_{j,m-1}$ for $j = 1, 2, \ldots, m - 1$. 

**ℓ₁ Penalized Greedy Pursuit (LPGP)**

First suppose the library $\mathcal{H}$ is normalized in that $\|h\| = 1$ for all $h \in \mathcal{H}$.

- **Algorithm**

  Initialize $\hat{f}_0 = 0$.

  Then for $m = 1, 2, \ldots$, iteratively, given $\hat{f}_{m-1} = \sum_{j=1}^{m-1} \beta_{j,m-1} h_j$, we seek

  $$\hat{f}_m(x) = \alpha \hat{f}_{m-1}(x) + \beta h(x)$$

  to minimize the objective function over choices of $h, \alpha, \beta$.

  \[
  \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha f_{m-1}(X_i) - \beta h(X_i))^2 + \lambda (|\beta| + \alpha \sum_{j=1}^{m-1} |\beta_{j,m-1}|) \leq \inf_\beta \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2 + \lambda \sum_h |\beta_{f,h}| + \frac{4(\sum_h |\beta_{f,h}|)^2}{m + 1} \right\},
  \]

  where $f_\beta = \sum_h \beta_{f,h} h$. 

- **Key Conclusion**

  \[
  h_m, \alpha_m, \beta_{m,m} \text{ and } \beta_{j,m} = \alpha_{m} \beta_{j,m-1} \text{ for } j = 1, 2, \ldots, m - 1.
  \]
Advantages and Disadvantages of \textit{LPGP}

- Let $p = \text{Card}(\mathcal{H})$, typically much larger than the data size $n$.

- As we shall see, the number of steps $m$ for statistical accurate fit is typically much less than $n$.

- Advantages
  - Computation of $\ell_1$ penalized solution with explicit guarantee of accuracy
  - Time cost $pnm$ v.s. $pn^2$ for an alternative strategies (LARS)

- Disadvantages
  - Not an exact solution
  - Algorithm basically is for the case of fixed $\lambda$. 
Key Ideas of the Proof

- Assume $\mathcal{H}$ is closed under sign-change (otherwise replaced by $\mathcal{H} \cup -\mathcal{H}$), so the coefficients of linear combination are kept non-negative.

- Denote $v_m = \sum_{j=1}^{m} \beta_{j,m}$ and $v = \sum_{h} \beta_{f,h}$ for a particular $f_{\beta} = \sum_{h} \beta_{f,h} h_f$. Let

$$e_m^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_m(X_i))^2 - \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_{\beta}(X_i))^2 + \lambda v_m.$$
Key Ideas of the Proof

- Assume \( H \) is closed under sign-change (otherwise replaced by \( H \cup -H \)), so the coefficients of linear combination are kept non-negative.

- Denote \( v_m = \sum_{j=1}^{m} \beta_{j,m} \) and \( v = \sum_{h} \beta_{f,h} \) for a particular \( f \). Let

\[
[w + \frac{1}{\lambda}]X + \frac{u}{\lambda} - \frac{u}{\lambda} + \frac{1}{\lambda} \geq \frac{\lambda}{\lambda}
\]

where

\[
\alpha - 1 = \alpha \quad \text{and} \quad \frac{1}{\lambda} = \lambda
\]

By the choice of \( \alpha \) and \( \beta \), we have

\[
\sum_{i=1}^{n} (Y_i - \alpha \hat{f}_m (X_i))^2 - \sum_{i=1}^{n} (Y_i - f \beta (X_i))^2 + \lambda v_m
\]

- Denote \( \eta \eta' \eta' \eta' \) for a particular \( \eta \), let \( \eta' \eta' \eta' \) for a particular \( \eta' \) and \( \eta' \eta' \eta' \) for a particular \( \eta' \). Hence, we have

\[
\sum_{i=1}^{n} (Y_i - \hat{f}_m (X_i))^2 - \sum_{i=1}^{n} (Y_i - \hat{f} (X_i))^2 = \lambda v_m
\]

Assume \( H \) is closed under sign-change (otherwise replaced by \( H \cup H \)), so \( H - H \cup H \) by
Key Ideas of the Proof

• Assume $\mathcal{H}$ is closed under sign-change (otherwise replaced by $\mathcal{H} \cup -\mathcal{H}$), so the coefficients of linear combination are kept non-negative.

• Denote $v_m = \sum_{j=1}^{m} \beta_{j,m}$ and $v = \sum_{h} \beta_{f,h}$ for a particular $f_{\beta} = \sum_{h} \beta_{f,h} h_f$. Let

$$e_m^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_m(X_i))^2 - \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_{\beta}(X_i))^2 + \lambda v_m.$$ 

• By the choice of $\alpha_m$, $\beta_{m,m}$ and $h_m$, the value is at least as good as if we use $\alpha = 1 - \frac{2}{m+1}$ and $\beta = \bar{\alpha} v$, we have

$$e_m^2 \leq \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha \hat{f}_{m-1} - \bar{\alpha} v h(X_i))^2 - \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_{\beta}(X_i))^2 + \lambda [\alpha v_{m-1} + \bar{\alpha} v],$$

where $\bar{\alpha} = 1 - \alpha$. We may rearrange it as

$$e_m^2 \leq \alpha e_{m-1}^2 + \bar{\alpha}^2 b_h + \bar{\alpha} \lambda v$$

$$- \frac{2\alpha \bar{\alpha}}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_{m-1}(X_i))(v h(X_i) - f_{\beta}(X_i))$$

$$- \frac{\alpha \bar{\alpha}}{n} \sum_{i=1}^{n} (\hat{f}_{m-1}(X_i) - f_{\beta}(X_i))^2,$$

where $b_h = \frac{1}{n} \sum_{i=1}^{n} (Y_i - v h(X_i))^2 - \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_{\beta}(X_i))^2.$
Key Ideas of the Proof

• Since the inequality holds for all \( h \), this \( e_m^2 \) is less than the average of the right side for any convenient distribution on the choices of \( h \). We consider the distribution that \( h \) is chosen to be \( h_f \) with probability \( \frac{\beta_{f,h}}{v} \) so that the expectation of \( vh(x) \) is \( f_\beta(x) \).

• Then \( (Y_i - \hat{f}_{m-1}(X_i))(vh(X_i) - f_\beta(X_i)) \) has expectation 0 and \( b_h \) has expectation not more than \( v^2 \). Thus
  \[
e_m^2 \leq \alpha e_{m-1}^2 + \bar{\alpha}^2 v^2 + \lambda \bar{\alpha} v,
\]
  where \( \bar{\alpha} = \frac{2}{m+1} \).
Key Ideas of the Proof

• Since the inequality holds for all $h$, $e_m^2$ is less than the average of the right side for any convenient distribution on the choices of $h$. We consider the distribution that $h$ is chosen to be $h_f$ with probability $\frac{\beta_{f,h}}{v}$ so that the expectation of $vh(x)$ is $f_\beta(x)$.

• Then $(Y_i - \hat{f}_{m-1}(X_i))(vh(X_i) - f_\beta(X_i))$ has expectation 0 and $b_h$ has expectation not more than $v^2$. Thus

$$e_m^2 \leq \alpha e_{m-1}^2 + \bar{\alpha}^2v^2 + \lambda\bar{\alpha}v,$$

where $\bar{\alpha} = \frac{2}{m+1}$.

• Initially $e_0^2 \leq v^2 + \lambda v$. By induction assuming that $e_{m-1}^2 \leq \frac{4v^2}{m} + \lambda v$, we establish that

$$e_m^2 \leq \frac{4v^2}{m+1} + \lambda v.$$

• Thus

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_m(X_i))^2 + \lambda v_m \leq \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2 + \lambda v + \frac{4v^2}{m+1}.$$
Results Re-expressed for Un-normalized $\mathcal{H}$

Drop the normalization condition.

- **Algorithm**
  
  Initialize $\hat{f}_0 = 0$.
  
  Then for $m = 1, 2, \ldots$, iteratively, given the terms of $\hat{f}_{m-1} = \sum_{j=1}^{m-1} \beta_{j,m-1} h_j$, we seek $\hat{f}_m = \alpha \hat{f}_{m-1} + \beta h$ to minimize the objective function over choices of $h, \alpha, \beta$.

  
  \[
  \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \alpha \hat{f}_{m-1}(X_i) - \beta h(X_i) \right)^2 + \lambda \left( |\beta| \|h\| + \alpha \sum_{j=1}^{m-1} |\beta_{j,m-1}| \|h_j\| \right)
  \]

  yielding $h_m, \alpha_m, \beta_{m,m}$ and $\beta_{j,m} = \alpha_m \beta_{j,m-1}$ for $j = 1, 2, \ldots, m - 1$.

- **Key Conclusion**
  
  Let $V(f) = \|f\|_{1,\mathcal{H}} = \inf \{ \sum_h |\beta_{f,h}| \|h\| : f = \sum_h \beta_{f,h} h \}$. Thus,

  \[
  \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_m(X_i))^2 + \lambda \sum_{j=1}^{m} |\beta_{j,m}| \|h_j\| \leq \inf \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda V(f) + \frac{4V^2(f)}{m+1} \right\}.
  \]
Risk Bounds of the Estimators obtained from \textit{LPGP}

• Suppose \((X_i, Y_i)_{i=1}^n\) are independently drawn from the distribution of \((X, Y)\). The target regression function \(f^*(x) = E[Y|X = x]\) is unknown and is to be estimated. The error \(\epsilon = Y - f^*(X)\) is assumed to have a conditional distribution given \(X\) which satisfies certain moment conditions.

• We work with the set \(\mathcal{F}\), the linear span of library \(\mathcal{H}\).

• Suppose \(\{\hat{f}_m, m = 1, 2, \ldots\}\) is the sequence of estimators formulated from the \textit{LPGP} algorithm.

• Measure of loss is the generalization error for \(\mu = P_x\),

\[
\|f - f^*\|^2 = \int (f(x) - f^*(x))^2 \mu(dx).
\]
Risk Bounds of the Estimators obtained from LPGP

• Risk bounds for $\ell_1$ penalization

If $\lambda_n > B\sqrt{\frac{\log p}{n}}$, we may run LPGP for many steps to reach an approximation of the Lasso solution $\hat{f}$. It has the following risk bound.

$$E\|\hat{f} - f^*\|^2 \leq (1 + \delta) \inf_{f \in F} \{ \|f - f^*\|^2 + \lambda_n V(f) \} + \frac{C_\delta}{n}.$$
Risk Bounds of the Estimators obtained from \textit{LPGP}

\begin{itemize}
\item **Risk bounds for $\ell_1$ penalization**

  If $\lambda_n > B \sqrt{\frac{\log p}{n}}$, we may run \textit{LPGP} for many steps to reach an approximation of the Lasso solution $\hat{f}$. It has the following risk bound.

  \[
  \mathbb{E} \| \hat{f} - f^* \|^2 \\
  \leq (1 + \delta) \inf_{f \in \mathcal{F}} \left\{ \| f - f^* \|^2 + \lambda_n V(f) \right\} + \frac{C_\delta}{n}.
  \]

\item **Risk bounds for model selection**

  If $\lambda_n$ is chosen much smaller (e.g. of the order of $1/n$). We choose $\hat{m}$ to minimize the penalized least squares

  \[
  \frac{1}{n} \sum_{i=1}^{m} (Y_i - \hat{f}_m(X_i))^2 + \lambda_n \left( \sum_{j=1}^{n} |\beta_{j,m}| \| h_j \|_n \right) + \frac{\gamma m \log p}{n},
  \]

  where $\gamma$ is a constant. Then the risk of the estimator $\hat{f}_\hat{m}$ is bounded by

  \[
  \mathbb{E} \| \hat{f}_\hat{m} - f^* \|^2 \\
  \leq (1 + \delta) \inf_{\hat{m}} \inf_{f \in \mathcal{F}} \left\{ \| f - f^* \|^2 + \lambda_n V(f) + \frac{4V^2(f)}{m} + \frac{\gamma m \log p}{n} \right\} \leq \frac{C_\delta}{n} \\
  \leq (1 + \delta) \inf_{f \in \mathcal{F}} \left\{ \| f - f^* \|^2 + \lambda'_n V(f) \right\} \leq \frac{C_\delta}{n},
  \]

  where $\lambda'_n = \lambda_n + B_1 \sqrt{\frac{\log p}{n}}$.
\end{itemize}
Conclusion

- Subset selection procedures may be used in $\ell_1$-penalized least squares optimization.

- An $m$-term chosen by relaxed greedy pursuit or by $\ell_1$-penalized greedy pursuit provides accuracy within order $V^2(f)/m$ of the minimal objective function.

- Ultimate penalty is
  \[ \min \left\{ \lambda_n V(f), \frac{m \log p}{n} \right\} \]

- Risk of the estimate is captured by the ideal tradeoff between $\|f - f^*\|^2$ and the penalty.