Proof of variable selection consistency

(Theorem 7.21 and Corollary 7.22)

Kaveh S. Nobari

Lectures in High-Dimensional Statistics

Department of Mathematics and Statistics
Lancaster University

Contents

Theorem (Part I)

Consider an S-sparse linear regression model, with a design matrix that satisfies the lower eigenvalue and mutual incoherence assumptions. Then, for any regularization parameter

$$\lambda_n \geq \frac{2}{1-\alpha}\left\| X'_{S^c}\Pi_{S^\perp}(X)\frac{\varepsilon}{n}\right\|_\infty \tag{1}$$

the Lagrangian Lasso has these properties:

(a) *Uniqueness: There exists a unique optimal solution $\hat{\theta}$.*

(b) *No erroneous inclusion: The optimal solution has its support set $\hat{S}$ contained within the support set S, or in other words*

$$supp(\hat{S}) \subseteq supp(S)$$

Theorem (Part II)

(c) *ℓ_∞ bounds: The error $\hat{\theta} - \theta^*$ satisfies*

$$\|\hat{\theta}_S - \theta^*_S\|_\infty \leq \underbrace{\left\| \left(\frac{X_S' X_S}{n}\right)^{-1} X_S' \frac{\varepsilon}{n} \right\|_\infty + \left|\left|\left| \left(\frac{X_S' X_S}{n}\right)^{-1} \right|\right|\right|_\infty \lambda_n}_{B(\lambda_n, X)}, \quad (2)$$

where $|||A|||_\infty = \max_{i \in [S]} \sum |A_{ij}|$ is the matrix ℓ_∞-norm.

(d) *No erroneous exclusion: The Lasso includes all indices $i \in S$, such that $|\theta_i^*| > B(\lambda_n; X)$, and hence it is sparsistent if $\min_{i \in S} |\theta_i^*| > B(\lambda_n; X)$.*

Subdifferentials

- To prove Theorem 7.21 of Wainwright (2019), we first develop the necessary and sufficient conditions for optimality in Lasso.
- The complication arises as the ℓ_1-norm is not differentiable, despite being convex, as it has a kink at the origin.
- To show this, we will go through an example, and that is the subdifferential of $f(x) = |x|$ at 0.

We know that

$$|x| = \begin{cases} x, & x \geq 0 \\ -x, & x < 0 \end{cases}$$

.

A subdifferential of a convex function $f : I \to \mathbb{R}$ at a point x_0 is $c \in \mathbb{R}$, such that

$$f(x) - f(x_0) \geq c(x - x_0), \quad \forall x \in [a, b]. \tag{3}$$

where I is an open interval.

We may find many subderivatives that satisfy inequality (3). What subdifferentials satisfy (3)? We say that c is in a closed interval and bounded by - i.e., $c \in [a, b]$. Furthermore, it is rather straightforward to find a and b, where

$$a = \lim_{x \to x_0^-} \frac{f(x) - f(x_0)}{x - x_0}$$

$$b = \lim_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0}$$

An example

Let us consider $f(x) = |x|$, where we are interested at the subdifferential of $f(x)$ at 0. We can derive the terms a and b as follows:

$$a = \lim_{x \to 0^-} \frac{|x| - 0}{x - 0} = \lim_{x \to 0^-} \frac{-x}{x} = -1$$
$$b = \lim_{x \to 0^+} \frac{|x| - 0}{x - 0} = \lim_{x \to 0^+} \frac{+x}{x} = +1$$

thus, $c \in [-1, +1]$ and

$$\partial |x| = \begin{cases} -1, & x < 0 \\ [-1, +1], & x = 0 \\ +1, & x > 0 \end{cases}$$

To generalize these results to a d-dimensional vector space, such as $\|\theta\|_1$, we say that for a convex function $f : \mathbb{R}^d \to \mathbb{R}$, $z \in \mathbb{R}^d$ is a subgradient of f at θ, denoted by $z \in \partial f(\theta)$, if

$$f(\theta + \Delta) - f(\theta) \geq \langle z, \Delta \rangle, \quad \forall \Delta \in \mathbb{R}^d.$$

When $f(\theta) = \|\theta\|_1$, $z \in \partial \|\theta\|_1$ iff as with the scalar example earlier

$$z_j = sgn(\theta_j), \quad \text{and} \quad sgn(0) = [-1, +1].$$

Primal Dual Witness

For the Lagrangian Lasso program

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}. \tag{4}$$

It is said that a pair $(\hat{\theta}, \hat{z})$ is primal-dual optimal, if $\hat{\theta}$ is a minimizer and $\hat{z} \in \partial \|\hat{\theta}\|_1$.
Note that (4) can be expressed as follows,

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} (y - X\theta)'(y - X\theta) + \lambda_n \|\theta\|_1 \right\} \\ &= \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \left[y'y - \theta' X' y - y' X\theta + \theta' X' X\theta + \lambda_n \|\theta\|_1 \right] \right\} \end{aligned}$$

Primal Dual Witness

which may alternatively be expressed as

$$\begin{aligned} \frac{1}{2n}\left[\frac{\partial}{\partial\hat{\theta}}\left\{y'y-\hat{\theta}'X'y-y'X\hat{\theta}+\hat{\theta}'X'X\hat{\theta}+\lambda_n\|\hat{\theta}\|_1\right\}\right]&=0 \\ \frac{1}{2n}\left[2X'X\hat{\theta}-2X'y+\lambda_n\hat{z}\right]&=0 \\ \frac{1}{n}X'\left(X\hat{\theta}-y\right)+\lambda_n\hat{z}&=0 \end{aligned} \tag{5}$$

Thus, any pair $(\hat{\theta},\hat{z})$, must satisfy the last line of equation (5).

The proof of Theorem 2.1 is based on a constructive approach, known as a primal-dual witness, which is as follows:

1. Construct a pair $(\hat{\theta}, \hat{z})$ that satisfies the zero-subgradient condition (5) and such that $\hat{\theta}$ has the correct signed-support.
2. When this procedure is successful, the constructed pair is primal-dual optimal.
3. The constructed pair now serves as a witness for the fact that the Lasso has a unique optimal solution with correct signed-support.

Formally, the procedure has been outlined as in the next frame in Wainwright (2019).

Theorem (Primal-dual witness (PDW) construction)

1. *Set* $\hat{\theta}_{S^c} = 0$
2. *Determine* $(\hat{\theta}_S, \hat{z}_S) \in \mathbb{R}^s \times \mathbb{R}^s$ *by solving the oracle subproblem*

$$\hat{\theta}_S \in \arg\min_{\theta_S \in \mathbb{R}^s} \left\{ \underbrace{\frac{1}{2n}\|y - X_S\theta_S\|_2^2}_{=:f(\theta_S)} + \lambda_n \|\theta_S\|_1 \right\}, \tag{6}$$

and then choosing $\hat{z}_S \in \partial\|\hat{\theta}_S\|_1$, *such that* $\nabla f(\theta_S)\Big|_{\theta_S = \hat{\theta}_S} + \lambda_n \hat{z}_S = 0.$

3. *Solve for* $\hat{z}_{S^c} \in \mathbb{R}^{d-s}$ *via the zero-subgradient equation (5), and check whether or not the strict dual feasibility condition* $\|\hat{z}_{S^c}\|_\infty < 1$ *holds.*

PDW intuition

- The vector $\hat{\theta}_{S^c} \in \mathbb{R}^{d-s}$ is determined in the first step.
- The remaining sub-vectors $\hat{\theta}_S, \hat{z}_S$ and $\hat{z}_{S^c}$ are determined in the second and third steps of the method.
- By construction the latter three sub-vectors satisfy the zero-subgradient condition (5).
- Using the fact that $\hat{\theta}_{S^c} = \theta_{S^c} = 0$, and writing out the zero-subgradient condition in a block matrix form, we obtain

$$\frac{1}{n}\begin{bmatrix} X_S' X_S & X_S' X_{S^c} \\ X_{S^c}' X_S & X_{S^c}' X_{S^c} \end{bmatrix}\begin{bmatrix} \hat{\theta}_S - \theta_S^* \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} X_S' \varepsilon \\ X_{S^c}' \varepsilon \end{bmatrix} + \lambda_n \begin{bmatrix} \hat{z}_S \\ \hat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (7)$$

- It is said that the PDW construction succeeds if $\hat{z}_{S^c}$ satisfies the strict dual feasibility condition.

PDW intuition

Note that (7) is the consequence of the following manipulation of (5):

$$\frac{1}{n}X'\left(X\hat{\theta}-y\right)+\lambda_n\hat{z}=0$$

$$\frac{1}{n}X'\left(X\hat{\theta}-\underbrace{X\theta^*+\varepsilon}_{=y}\right)+\lambda_n\hat{z}=0$$

$$\frac{1}{n}X'X\left(\hat{\theta}-\theta^*\right)+\frac{1}{n}X'\varepsilon+\lambda_n\hat{z}=0$$

The following Lemma copied from Wainwright (2019) shows that the success of PDW acts as a witness for the Lasso.

Lemma

If the lower eigenvalue condition

$$\gamma_{\min}\left(\frac{X_S' X_S}{n}\right) \geq c_{\min} > 0$$

holds, then the success of the PDW construction implies that the vector $(\hat{\theta}_S, 0) \in \mathbb{R}^d$ *is the unique optimal solution of the Lasso.*

Proof

- When PDW succeeds, $\hat{\theta} = (\hat{\theta}_S, 0)$ is optimal solution with associated with subgradient vector $\hat{z} \in \mathbb{R}^d$, satisfying $\|\hat{z}_{S^c}\|_\infty < 1$, and $\langle \hat{z}, \hat{\theta} \rangle = \|\hat{\theta}\|_1$.
- Now suppose there exists another optimal solution $\tilde{\theta}$.
- Introduce the shorthand notation $F(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.
- We are then guaranteed that

$$F(\hat{\theta}) + \lambda_n \langle \hat{z}, \hat{\theta} \rangle = F(\tilde{\theta}) + \lambda_n \|\tilde{\theta}\|_1$$

- subtracting $\lambda_n \langle \hat{z}, \tilde{\theta} \rangle$ from both sides, we then obtain

$$F(\hat{\theta}) - \lambda_n \langle \hat{z}, \tilde{\theta} - \hat{\theta} \rangle = F(\tilde{\theta}) + \lambda_n \left(\|\tilde{\theta}\|_1 - \langle \hat{z}, \tilde{\theta} \rangle \right).$$

- But by the zero-subgradient condition (5), we know that $\nabla F(\hat{\theta}) = -\lambda_n \hat{z}$, implying

$$F(\hat{\theta}) + \langle \nabla F(\hat{\theta}), \tilde{\theta} - \hat{\theta} \rangle - F(\tilde{\theta}) = \lambda_n \left(\|\tilde{\theta}\|_1 - \langle \hat{z}, \tilde{\theta} \rangle \right).$$

Proof

- By convexity of F, the left hand side is negative, implying that $\|\tilde{\theta}\|_1 \leq \langle \hat{z}, \tilde{\theta} \rangle$.
- From Holder's inequality, we know that $\langle \hat{z}, \tilde{\theta} \rangle \leq \|\hat{z}\|_\infty \|\tilde{\theta}\|_1$, thus we must have $\|\tilde{\theta}\|_1 = \langle \hat{z}, \tilde{\theta} \rangle$.
- However, since $\|\hat{z}_{S^c}\|_\infty < 1$, this equality can only occur if $\tilde{\theta}_j = 0$, for all $j \in S^c$.
- Hence, all optimal solutions are supported only on S, and can be obtained by solving the oracle sub-problem (6).
- Given the lower eigenvalue condition, this sub-problem is strictly convex, and so has a unique minimizer.
- Therefore, to prove Theroem 7.21 (a) and (b), it is sufficient to show that $\hat{z}_{S^c} \in \mathbb{R}^{d-s}$ in the third step satisfies the strict dual feasibility condition.

Proof

- The latter can be solved using the zero-subgradient conditions (7), - i.e.,

$$
\begin{aligned}
\frac{1}{n}X'_{S^c}X_S(\hat{\theta}_S - \theta^*_S) - \frac{1}{n}X'_{S^c}\varepsilon + \lambda_n \hat{z}_{S^c} &= 0 \\
\lambda_n \hat{z}_{S^c} &= \frac{1}{n}X'_{S^c}\varepsilon - \frac{1}{n}X'_{S^c}X_S(\hat{\theta}_S - \theta^*_S) \\
\hat{z}_{S^c} &= X'_{S^c}\left(\frac{\varepsilon}{\lambda_n n}\right) - \frac{1}{\lambda_n n}X'_{S^c}X_S(\hat{\theta}_S - \theta^*_S)
\end{aligned}
$$

- Similarly, using the invertibility of $X'_S X_S$, we solve for $\hat{\theta}_S - \theta^*_S$ as follows

$$
\begin{aligned}
\frac{1}{n}X'_S X_S(\hat{\theta}_S - \theta^*_S) - \frac{1}{n}X'_S\varepsilon + \lambda_n \hat{z}_S &= 0 \\
\frac{1}{n}X'_S X_S(\hat{\theta}_S - \theta^*_S) &= \frac{1}{n}X'_S\varepsilon - \lambda_n \hat{z}_S \\
\hat{\theta}_S - \theta^*_S &= (X'_S X_S)^{-1}X'_S\varepsilon - n\lambda_n (X'_S X_S)^{-1}\hat{z}_S
\end{aligned}
$$

Proof

- Combining these two, we obtain

$$\hat{z}_{S^c} = X'_{S^c}\left(\frac{\varepsilon}{\lambda_n n}\right) - \frac{1}{\lambda_n n}X'_{S^c}X_S(\hat{\theta}_S - \theta^*_S)$$

$$\hat{z}_{S^c} = X'_{S^c}\left(\frac{\varepsilon}{\lambda_n n}\right) - \frac{1}{\lambda_n n}X'_{S^c}X_S(X'_SX_S)^{-1}X'_S\varepsilon + X'_{S^c}X_S(X'_SX_S)^{-1}\hat{z}_S$$

$$\hat{z}_{S^c} = \underbrace{X_{S^c}\left[I - X_S(X'_SX_S)^{-1}X'_S\right]\left(\frac{\varepsilon}{n\lambda_n}\right)}_{V_{S^c}} + \underbrace{X'_{S^c}X_S(X'_SX_S)^{-1}\hat{z}_S}_{\mu}$$

- By triangle inequality, we have

$$\|\hat{z}_{S^c}\| \leq \|V_{S^c}\|_\infty + \|\mu\|_\infty.$$

Proof

- By the mutual incoherence condition, - i.e.,

$$\max_{j \in S^c} \|(X_S' X_S)^{-1} X_S' X_j\|_1 \leq \alpha, \quad \alpha \in [0,1),$$

we have $\|\mu\|_\infty \leq \alpha$. Furthermore, by the choice of the regularization parameter, - i.e.,

$$\lambda_n \geq \frac{2}{1-\alpha} \left\| X_{S^c}' \Pi_{S^\perp}(X) \frac{\varepsilon}{n} \right\|_\infty$$

where

$$\Pi_{S^\perp}(X) = I_n - X_S (X_S' X_S)^{-1} X_S'$$

is an orthogonal projection matrix, we have $\|V_{S^c}\|_\infty \leq \frac{1}{2}(1-\alpha)$. Putting together the pieces, it can be concluded that $\|\hat{z}_{S^c}\|_\infty \leq \frac{1}{2}(1+\alpha) < 1$, which establishes the strict dual feasibility condition.

Proof

- Finally, it remains to establish a bound on the ℓ_∞-norm of the error $\hat{\theta}_S - \theta_S^*$. Using triangle inequality, we have

$$\|\hat{\theta}_S - \theta_S^*\|_\infty \leq \left\| \left(\frac{X_S' X_s}{n} \right) X_S' \frac{\varepsilon}{n} \right\|_\infty + \left|\left|\left| \left(\frac{X_S' X_S}{n} \right)^{-1} \right|\right|\right|_\infty \lambda_n$$

hence, completing the proof.

Corollary

Consider the S-sparse linear model, with noise vector ε with zero-mean i.i.d. entries that is sub-Gaussian with the sub-Gaussianity parameter σ. Furthermore, suppose that the deterministic design matrix X satisfies the lower eigenvalue and mutual incoherence assumptions, as well as the C-column normalization condition ($\max_{j=1,\cdots,d}\|X_j\|_2/\sqrt{n} \leq C$). Suppose, we solve the Lagrangian Lasso with regularization parameter

$$\lambda_n = \frac{2C\sigma}{1-\alpha}\left\{\sqrt{\frac{2\log(d-s)}{n}} + \delta\right\} \tag{8}$$

for $\delta > 0$. Then the optimal solution is unique, with supp$(\hat{\theta}) \subseteq$ supp(θ^*), and satisfied the ℓ_∞ bound

$$P\left[\|\hat{\theta}_S - \theta^*_S\|_\infty \leq \frac{\sigma}{\sqrt{c_{\min}}}\left\{\sqrt{\frac{2\log s}{n}} + \delta\right\} + \left|\!\left|\!\left|\left(\frac{X_S' X_S}{n}\right)^{-1}\right|\!\right|\!\right|_\infty \lambda_n\right] \geq 1 - 4\exp\left(-\frac{n\delta^2}{2}\right).$$

Proof

- First we must show that the choice of the regularization parameter λ_n (8) satisfies the bound (1) with high probability.
- This can be accomplished by bounding the maximum absolute value of the random variables,

$$Z_j := X_j' \Pi_{S^\perp}(X) \left(\frac{\varepsilon}{n}\right), \quad \text{for } j \in S^c.$$

- Since $\Pi_{S^\perp}(X)$ is an orthogonal projection matrix, we have

$$\|\Pi_{S^\perp}(X) X_j\|_2 \leq \|X_j\|_2 \overset{(i)}{\leq} C\sqrt{n}$$

  where (i) follows the column normalization assumption.
- Hence, each variable Z_j is sub-Gaussian with parameter at most $C^2\sigma^2/n$.

Proof

- From the sub-Gaussian tail bounds, we have

$$P\left[\max_{j \in S^c} |Z_j| \geq t\right] \leq 2(d-s)\exp\left(-\frac{-nt^2}{2C^2\sigma^2}\right)$$

  from which it is evident that the choice λ_n in (8) ensures that (1) holds with claimed probability.
- It now remains to bound the ℓ_∞-bound (2).
- As the second term is deterministic, the problem consists of bounding the first term.

Proof

- For $i = 1, \cdots, s$ consider the random variable

$$\tilde{Z}_j := e_i' \left(\frac{1}{n} X_S' X_S \right)^{-1} X_s \varepsilon / n.$$

, Since the elements of ε are i.i.d. σ-sub-Gaussian, Z_i is also zero-mean and sub-Gaussian with parameter at most

$$\frac{\sigma^2}{n} \left|\left|\left| \left(\frac{1}{n} X_S' X_S \right)^{-1} \right|\right|\right|_2 \leq \frac{\sigma^2}{c_{\min} n},$$

where the lower eigenvalue condition has been used. Consequently, for any $\delta > 0$, we have

$$P \left[\max_{i=1,\cdots,s} |\tilde{Z}_i| > \frac{\sigma}{\sqrt{c_{\min}}} \left\{ \sqrt{\frac{2 \log s}{n}} + \delta \right\} \right] \leq 2 \exp \left(-\frac{n\delta^2}{2} \right).$$

References

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.