Lasso oracle inequality and prediction error bounds

Kaveh S. Nobari

Lectures in High-Dimensional Statistics

Department of Mathematics and Statistics
Lancaster University

Contents

Motivation

- In the presence of a restricted eigenvalue condition on random designs, it is possible to obtain a more general result on the Lasso error, which is known as oracle inequality.
- Said results hold without any assumptions on the vector of parameters $\theta^* \in \mathbb{R}^d$, and it provides a family of upper bounds, with a tunable parameter to be optimized.
- The flexibility in tuning this parameter is akin to that of an oracle, which would have access to the ordered coefficients of θ^*.
- Finally, we divert our attention from the problem of parameter recovery (i.e. $\|\hat{\theta} - \theta^*\|_2$) to finding a good predictor $\hat{\theta} \in \mathbb{R}^d$, such that

$$\frac{\|X(\hat{\theta} - \theta^*)\|_2^2}{n} = \frac{1}{n} \sum_{i=1}^{n} \left(\langle x_i, \hat{\theta} - \theta^* \rangle \right)^2$$

is small.

Lasso oracle inequality

Theorem (Lasso oracle inequality)

Let $X \in \mathbb{R}^{n \times d}$ be a random matrix, with rows $x_i \in \mathbb{R}^d \sim N(0, \Sigma)$. Then there are universal positive constants $c_1 < 1 < c_2$, such that for all $\theta \in \mathbb{R}^d$

$$P\left[\frac{\|X\theta\|_2^2}{n} \geq c_1 \|\sqrt{\Sigma}\theta\|_2^2 - c_2 \rho^2(\Sigma) \frac{\log d}{n} \|\theta\|_1^2\right] \geq 1 - \frac{\exp(-n/32)}{1 - \exp(-n/32)}$$

Given this condition and considering the Lagrangian Lasso with $\lambda_n \geq 2\|X'w/n\|_\infty$ for any $\theta^* \in \mathbb{R}^d$, any optimal solution $\hat{\theta}$ satisfies the bound

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \underbrace{\frac{144}{c_1^2} \frac{\lambda_n^2}{\tilde{\kappa}^2} |S|}_{\text{estimation error}} + \underbrace{\frac{16}{c_1} \frac{\lambda_n}{\tilde{\kappa}} \|\theta^*_{S^c}\|_1 + \frac{32 c_2}{c_1} \frac{\rho^2(\Sigma)}{\tilde{\kappa}} \frac{\log d}{n} \|\theta^*_{S^c}\|_1^2}_{\text{approximation error}}$$

valid for any subset S with cardinaltiy $S \leq \frac{c_1}{64 c_2} \frac{\tilde{\kappa}}{\rho^2(\Sigma)} \frac{n}{\log d}$.

- In the above inequalities $\tilde{\kappa} := \gamma_{\min}(\Sigma)$ and $\rho^2(\Sigma)$ is the maximum diagonal entry of the covariance matrix Σ.
- It is evident that for each choice of S the Lasso oracle inequality provides a family of upper bounds.
- Obtaining the optimal choice S is based on trading off between the approximation error and estimation error.
- Estimation error grows linearly with $|S|$ and corresponds to the error associated with estimating $|S|$ unknown coefficients.
- Approximation error depends on the unknown regression regression vector through tail sum $\|\theta^*_{S^c}\|_1 = \sum_{j \notin S} |\theta^*_j|$.
- Optimal bound requires nominating a choice of S to balance between said two errors.

Proof

Following Wainwright (2019), let us denote $\rho^2(\Sigma)$ using ρ^2. Returning to the derivation of the the error bounds for Lagrangian Lasso using the restricted eigenvalue condition for fixed designs, the initial arguments are identical: we wish to show that under the condition $\lambda_n \geq 2\|\frac{X'w}{n}\|_\infty$, the error vector $\hat{\Delta} := \hat{\theta} - \theta^* \in \mathbb{C}_3(S)$. Defining the Lagrangian Lasso by

$$L(\theta; \lambda_n) = \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n\|\theta\|_1,$$

since $\hat{\theta}$ is optimal, we have

$$L(\hat{\theta}; \lambda_n) \leq L(\theta^*; \lambda_n) = \frac{1}{2n}\|w\|_2^2 + \lambda_n\|\theta^*\|_1.$$

which by expanding and rearranging both sides of the inequality, we obtain the Lagrangian basic inequality

$$0 \leq \frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \frac{w'X\hat{\Delta}}{n} + \lambda_n\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\}.$$

Proof

Knowing that θ^* is S-sparse, we can write

$$\begin{aligned}\|\theta^*\|_1 - \|\hat{\theta}\|_1 &= \|\theta^*_S\|_1 + \|\theta^*_{S^c}\|_1 - \|\theta^* + \hat{\Delta}\|_1 \\ &= \|\theta^*_S\|_1 - \|\theta^*_S + \hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1.\end{aligned}$$

Thus, substituting the above in the Lagrangian basic inequality, we obtain

$$0 \leq \frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \frac{w'X\hat{\Delta}}{n} + \lambda_n\{\|\theta^*_S\|_1 - \|\theta^*_S + \hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\}.$$

Applying Holder's inequality, the lower bound of the triangle inequality, and multiplying both sides by 2, we will obtain

$$\frac{1}{n}\|X\hat{\Delta}\|_2^2 \leq 2\left\|\frac{X'w}{n}\right\|_\infty \|\hat{\Delta}\|_1 + 2\lambda_n\{\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\},$$

Proof

which using $\lambda_n \geq 2\|X'w/n\|_\infty$, yields

$$\frac{1}{n}\|X\hat{\Delta}\|_2^2 \leq \lambda_n\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\}.$$

A similar argument can be used in this case. However, in this scenario the vanishing terms $\|\theta^*_{S^c}\|_1$ must be tracked. If said values are not eliminated, we would instead obtain

$$0 \leq \frac{1}{n}\|X\hat{\Delta}\|_2^2 \leq \lambda_n\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 + 2\|\theta^*_{S^c}\|_1\}.$$

Notice that $\lambda_n \geq 0$ and hence

$$0 \leq 3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 + 2\|\theta^*_{S^c}\|_1.$$

Adding and subtracting $\|\hat{\Delta}_S\|_1$ to the right-hand side of the above inequality yields

$$0 \leq 3\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_S\|_1 - \underbrace{(\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1)}_{\|\hat{\Delta}\|_1} + 2\|\theta^*_{S^c}\|_1.$$

Proof

Rearranging and squaring yields

$$\|\hat{\Delta}\|_1^2 \leq (4\|\hat{\Delta}_S\|_1 + 2\|\theta^*_{S^c}\|_1)^2.$$

Recall (using Cauchy-Schwarz inequality), the relationship between ℓ_1 and ℓ_2 norms - i.e. for $\nu \in \mathbb{R}^s$, $\|\nu\|_1 \leq \sqrt{s}\|\nu\|_2$. Therefore,

$$\|\hat{\Delta}\|_1^2 \leq (4\|\hat{\Delta}_S\|_1 + 2\|\theta^*_{S^c}\|_1)^2 \leq (4\sqrt{s}\|\hat{\Delta}_S\|_2 + 2\|\theta^*_{S^c}\|_1)^2.$$

Hence, it can be concluded that

$$\|\hat{\Delta}\|_1^2 \leq (4\|\hat{\Delta}_S\|_1 + 2\|\theta^*_{S^c}\|_1)^2 \leq 32|S|\|\hat{\Delta}\|_2^2 + 8\|\theta_{S^c}\|_1^2.$$

Returning to the Theorem concerning the restricted eigenvalue condition of random design matrices, we can substitute $\hat{\Delta}$ for θ, to obtain

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \geq c_1\|\sqrt{\Sigma}\hat{\Delta}\|_2^2 - c_2\rho^2(\Sigma)\frac{\log d}{n}\|\hat{\Delta}\|_1^2, \quad \forall\hat{\Delta} \in \mathbb{R}^d.$$

Proof

Combining this with the earlier inequality - i.e.

$$\|\hat{\Delta}\|_1^2 \leq 32|S|\|\hat{\Delta}\|_2^2 + 8\|\theta_{S^c}\|_1^2.$$

yields the following results

$$\begin{aligned}\frac{\|X\hat{\Delta}\|_2^2}{n} &\geq c_1\|\sqrt{\Sigma}\hat{\Delta}\|_2^2 - c_2\rho^2\frac{\log d}{n}\left\{32|S|\|\hat{\Delta}\|_2^2 + 8\|\theta_{S^c}\|_1^2\right\}. \\ &\geq c_1\|\sqrt{\Sigma}\hat{\Delta}\|_2^2 - c_2 32|S|\rho^2\frac{\log d}{n}\|\hat{\Delta}\|_2^2 - c_2 8\rho^2\frac{\log d}{n}\|\theta_{S^c}\|_1^2.\end{aligned}$$

Moreover, we know that

$$\begin{aligned}\|\sqrt{\Sigma}\hat{\Delta}\|_2 &\geq \left|\left|\left|\sqrt{\Sigma}\right|\right|\right|_2\|\hat{\Delta}\|_2 \\ &\geq \sqrt{\tilde{\kappa}}\|\hat{\Delta}\|_2.\end{aligned}$$

Proof

Therefore,

$$\begin{aligned}\frac{\|X\hat{\Delta}\|_2^2}{n} &\geq c_1\|\sqrt{\Sigma}\hat{\Delta}\|_2^2 - c_2 32|S|\rho^2\frac{\log d}{n}\|\hat{\Delta}\|_2^2 - c_2 8\rho^2\frac{\log d}{n}\|\theta_{S^c}\|_1^2 \\ &\geq c_1\tilde{\kappa}\|\hat{\Delta}\|_2^2 - c_2 32|S|\rho^2\frac{\log d}{n}\|\hat{\Delta}\|_2^2 - c_2 8\rho^2\frac{\log d}{n}\|\theta_{S^c}\|_1^2 \\ &\geq \left\{c_1\tilde{\kappa} - c_2 32|S|\rho^2\frac{\log d}{n}\right\}\|\hat{\Delta}\|_2^2 - c_2 8\rho^2\frac{\log d}{n}\|\theta_{S^c}\|_1^2\end{aligned}$$

Now recall the cardinality condition from the Theorem - i.e.

$$\begin{aligned}|S| &\leq \frac{c_1}{64c_2}\frac{\tilde{\kappa}}{\rho^2}\frac{n}{\log d} \\ 32|S| &\leq \frac{c_1}{2c_2}\frac{\tilde{\kappa}}{\rho^2}\frac{n}{\log d} \\ 32c_2\rho^2|S|\frac{\log d}{n} &\leq c_1\frac{\tilde{\kappa}}{2}.\end{aligned}$$

Proof

Thus, we may express the earlier derivation as follows

$$\begin{aligned}\frac{\|X\hat{\Delta}\|_2^2}{n} &\geq \left\{c_1\tilde{\kappa} - c_2 32|S|\rho^2\frac{\log d}{n}\right\}\|\hat{\Delta}\|_2^2 - c_2 8\rho^2\frac{\log d}{n}\|\theta_{S^c}\|_1^2 \\ &\geq \left\{c_1\tilde{\kappa} - c_1\frac{\tilde{\kappa}}{2}\right\}\|\hat{\Delta}\|_2^2 - c_2 8\rho^2\frac{\log d}{n}\|\theta_{S^c}\|_1^2 \\ &\geq c_1\frac{\tilde{\kappa}}{2}\|\hat{\Delta}\|_2^2 - c_2 8\rho^2\frac{\log d}{n}\|\theta_{S^c}\|_1^2.\end{aligned}$$

The analysis is now split into two cases:

Case 1:

First, suppose that $c_1\frac{\tilde{\kappa}}{4}\|\hat{\Delta}\|_2^2 \geq 8c_2\rho^2\frac{\log d}{n}\|\theta^*_{S^c}\|_1^2$. Combining the bounds

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \geq c_1\frac{\tilde{\kappa}}{2}\|\hat{\Delta}\|_2^2 - c_2 8\rho^2\frac{\log d}{n}\|\theta_{S^c}\|_1^2.$$

Proof

and

$$0 \leq \frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \frac{\lambda_n}{2}\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 + 2\|\theta^*_{S^c}\|_1\}.$$

we would obtain

$$\begin{aligned}
0 &\leq c_1\frac{\tilde{\kappa}}{4}\|\hat{\Delta}\|_2^2 - c_2 4\rho^2\frac{\log d}{n}\|\theta_{S^c}\|_1^2 \leq \frac{\lambda_n}{2}\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 + 2\|\theta^*_{S^c}\|_1\} \\
&\leq c_1\frac{\tilde{\kappa}}{4}\|\hat{\Delta}\|_2^2 - \frac{1}{2}c_2 8\rho^2\frac{\log d}{n}\|\theta_{S^c}\|_1^2 \leq \frac{\lambda_n}{2}\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 + 2\|\theta^*_{S^c}\|_1\} \\
&\leq c_1\frac{\tilde{\kappa}}{4}\|\hat{\Delta}\|_2^2 - \frac{1}{2}c_1\frac{\tilde{\kappa}}{4}\|\hat{\Delta}\|_2^2 \leq \frac{\lambda_n}{2}\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 + 2\|\theta^*_{S^c}\|_1\} \\
&\leq c_1\frac{\tilde{\kappa}}{8}\|\hat{\Delta}\|_2^2 \leq \frac{\lambda_n}{2}\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 + 2\|\theta^*_{S^c}\|_1\}.
\end{aligned}$$

Moreover, recall from earlier that

$$\|\hat{\Delta}\|_1 = \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \leq 3\|\hat{\Delta}_S\|_1 \leq 3\sqrt{|S|}\|\hat{\Delta}\|_2$$

Proof

Thus,

$$0 \leq c_1 \frac{\tilde{\kappa}}{8} \|\hat{\Delta}\|_2^2 \leq \frac{\lambda_n}{2} \{3\sqrt{|S|}\|\hat{\Delta}\|_2 + 2\|\theta^*_{S^c}\|_1\}.$$

Notice that the bounds involve a quadratic form in $\|\hat{\Delta}\|_2$ - i.e.

$$c_1 \frac{\tilde{\kappa}}{8} \|\hat{\Delta}\|_2^2 - \frac{3\sqrt{|S|}\lambda_n}{2} \|\hat{\Delta}\|_2 - \lambda_n \|\theta^*_{S^c}\|_1 \leq 0$$
$$c_1 \tilde{\kappa} \|\hat{\Delta}\|_2^2 - 12\sqrt{|S|}\lambda_n \|\hat{\Delta}\|_2 - 8\lambda_n \|\theta^*_{S^c}\|_1 \leq 0$$

where calculating the zeros of this quadratic form, we obtain

$$\|\hat{\Delta}\|_2^2 \leq \frac{144\lambda_n^2}{c_1^2 \tilde{\kappa}^2} |S| + \frac{16\lambda_n \|\theta^*_{S^c}\|_1}{c_1 \tilde{\kappa}}$$

Case 2: Otherwise, we must have

$$c_1 \frac{\tilde{\kappa}}{4} \|\hat{\Delta}\|_2^2 < 8c_2 \rho^2 \frac{\log d}{n} \|\theta_{S^c}\|_1^2$$

Proof

Taking into account both cases, we combine this bound with earlier inequality to derive the claim of the Theorem.

Bounds on prediction error

- In the earlier Section, we focused mainly on the problem of parameter recovery in either noisy or noiseless settings.
- In many situations the true value of θ^* may not be of interest and our interest may lie in finding a good predictor $\hat{\theta} \in \mathbb{R}^d$, such that the mean-squared prediction error

$$\frac{\|X(\hat{\theta} - \theta^*)\|_2^2}{n} = \frac{1}{n} \sum_{i=1}^{n} \left(\langle x_i, \hat{\theta} - \theta^* \rangle \right)^2$$

  is small.
- To understand why the above quantity is a measure of prediction error, we provide the following example.

Bounds on prediction error

Suppose $\hat{\theta}$ is estimated using the response vector

$$y = X\theta^* + w.$$

Now let us assume that we obtain a fresh vector of responses -i.e.

$$\tilde{y} = X\theta^* + \tilde{w}, \quad \tilde{w} \sim i.i.d(0, \sigma^2)$$

The quality of our estimated vector $\hat{\theta}$ can then be assessed by measuring its success in predicting vector $\tilde{y}$ in terms of squared error. Using some algebra, it can be shown

$$\frac{1}{n}\mathbb{E}[\|\tilde{y} - X\hat{\theta}\|_2^2] = \frac{1}{n}\|X(\hat{\theta} - \theta^*)\|_2^2 + \sigma^2.$$

Bounds on prediction error

Proof:

$$\begin{aligned}
\frac{1}{n}\mathbb{E}[\|\tilde{y} - X\hat{\theta}\|_2^2] &= \frac{1}{n}\mathbb{E}[\|\tilde{y} - X\theta^* + X\theta^* - X\hat{\theta}\|_2^2] \\
&= \frac{1}{n}\mathbb{E}[\|\tilde{w} + X(\theta^* - \hat{\theta})\|_2^2] \\
&= \frac{1}{n}\mathbb{E}[\|\tilde{w}\|_2^2 + 2\langle X(\theta^* - \hat{\theta}), \tilde{w}\rangle + \|X(\theta^* - \hat{\theta})\|_2^2] \\
&= \frac{1}{n}\mathbb{E}[\|\tilde{w}\|_2^2] + \frac{1}{n}\underbrace{\mathbb{E}[2\langle X(\theta^* - \hat{\theta}), \tilde{w}\rangle]}_{=0} + \frac{1}{n}\mathbb{E}[\|X(\theta^* - \hat{\theta})\|_2^2] \\
&= \sigma^2 + \frac{1}{n}\|X(\theta^* - \hat{\theta})\|_2^2
\end{aligned}$$

Hence, apart from additive factor σ^2, the mean-squared prediction error measures how well we can predict.

Bounds on prediction error

In general the problem of finding a good predictor must be easier than estimating θ^* in the ℓ_2-norm. Prediction does not require that θ^* is identifiable, and unlike the parameter recovery setting, we may solve the problem if two columns of the designs matrix X are identical.

Bounds on prediction error

Theorem (Prediction error bounds)

Once again consider the Lagrangian Lasso with a strictly positive $\lambda_n \geq 2\|X'w/n\|_\infty$.

(a) *Any optimal solution* $\hat{\theta}$ *satisfies*

$$\frac{\|X(\hat{\theta} - \theta^*)\|_2^2}{n} \leq 12\|\theta^*\|_1\lambda_n$$

(b) *If* θ^* *is supported on subset* S*, such that* $|S| = s$*, and the design matrix satisfies the* $(\kappa, 3)$*-RE condition over* S*, then any optimal solution satisfies the bound*

$$\frac{\|X(\hat{\theta} - \theta^*)\|_2^2}{n} \leq \frac{9}{\kappa}s\lambda_n^2$$

References

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