STAT 535: Statistical Machine Learning **Autumn 2019**

Lecture 4: Regression: Linear Model

Instructor: Yen-Chi Chen

Reference: Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations.* Chapman and Hall/CRC, 2015.

4.1 Introduction

Suppose that we observe data

$$(X_1, Y_1), \cdots, (X_n, Y_n)$$

that are IID from an unknown distribution $F_{X,Y}$ such that $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$.

Let $\widehat{m}$ be a regression estimator (estimator of the regression function). We often use the squared error as our measure of accuracy. Under the squared error, the prediction risk is

$$R(\widehat{m}) = \mathbb{E}((Y - \widehat{m}(X))^2),$$

where (X, Y) is a new pair of observation from the same population. Note that the expectation is taken over both new observation (X, Y) and the estimator $\widehat{m}$.

Let m be the true regression function, i.e. $\mathbb{E}(Y|X = x) = m(x)$. The prediction risk can be decomposed into

$$R(\widehat{m}) = \sigma^2 + \underbrace{\mathbb{E}(b_n^2(X))}_{\text{bias}} + \underbrace{\mathbb{E}(V_n(X))}_{\text{variance}},$$

where

$$\sigma^2 = \mathbb{E}((Y - m(X))^2), \quad b_n(x) = \mathbb{E}(\widehat{m}(x)) - m(x), \quad V_n(x) = \mathsf{Var}(\widehat{m}(x)).$$

When using the linear regression, we do not (and should not) assume that the linear model is correct. The linear regression can be viewed as the *best linear predictor* that minimizes $\mathbb{E}((Y - \beta^T X)^2)$. Namely, the optimal coefficients

$$\beta^* = \mathsf{argmin}_\beta \mathbb{E}((Y - \beta^T X)^2)$$

and you can easily see that a sample analogue to β^* is

$$\widehat{\beta}_n = \mathsf{argmin}_\beta \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^T X_i)^2,$$

which is the least squared estimator (LSE).

When $\Sigma = \mathbb{E}(XX^T)$ is non-singular, the minimizer β^* has the following closed-form

$$\beta^* = \Sigma^{-1}\alpha,$$

where $\alpha = \mathbb{E}(XY)$. Similarly, the LSE also has the following closed-form

$$\widehat{\beta}_n = \widehat{\Sigma}_n^{-1}\widehat{\alpha}_n,$$

where $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ is the sample covariance matrix and $\widehat{\alpha}_n = \frac{1}{n}\sum_{i=1}^n X_i Y_i$.

Now we study the *excess risk* of $\widehat{\beta}_n$, i.e.,

$$\mathcal{E}(\widehat{\beta}_n) = R(\widehat{\beta}_n) - R(\beta^*).$$

The excess risk tells us that the expected loss when we are using the LSE compare to using the optimal predictor.

Theorem 4.1 *Assume the distribution F_{XY} is supported on a compact set and Σ is non-singular. Then there exists $c_1, c_2 > 0$ such that*

$$P(R(\widehat{\beta}_n) > R(\beta^*) + 2\epsilon) \leq c_1 e^{-nc_2\epsilon^2}.$$

The above bound is also called the *concentration bound*. It is another way to express how good an estimator is.

Proof: Let $Z = (Y, X)$ and let $\underline{\beta} = (-1, \beta)$. With this notation, $(Y - \beta^T X) = -\underline{\beta}^T Z$. So the prediction risk can be written as

$$R(\beta) = \mathbb{E}((Y - \beta^T X)^2) = \mathbb{E}(\underline{\beta}^T Z Z^T \underline{\beta}) = \underline{\beta}^T \mathbb{E}(ZZ^T)\underline{\beta} = \underline{\beta}^T \Gamma \underline{\beta}.$$

Similarly, the sample version of the prediction risk (called *empirical risk*) is

$$\widehat{R}_n(\beta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^T X_i)^2 = \underline{\beta}^T \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T \underline{\beta} = \underline{\beta}^T \widehat{\Gamma}_n \underline{\beta}.$$

Thus, the difference between the empirical risk and prediction risk is

$$|\widehat{R}_n(\beta) - R(\beta)| = |\underline{\beta}^T \widehat{\Gamma}_n \underline{\beta} - \underline{\beta}^T \Gamma \underline{\beta}| = |\underline{\beta}^T (\widehat{\Gamma}_n - \Gamma)\underline{\beta}| \leq \|\underline{\beta}\|_1^2 \|\widehat{\Gamma}_n - \Gamma\|_{\max}.$$

Note that $\|A\|_{\max} = \max_{j,k} |A_{jk}|$ is the matrix maximum norm. Using the Hoeffding's inequality to each entry with the fact that F_{XY} has a compact support, we conclude that

$$P(\|\widehat{\Gamma}_n - \Gamma\|_{\max} > \epsilon) < (d+1)^2 2e^{-nc_3\epsilon^2},$$

where c_3 is a constant depending on the size of the support. Note that when Σ is non-singular and F_{XY} has a compact support, there exists $\bar{B}$ such that $\|\widehat{\beta}_n\| \leq B$ a.s. so we will assume that $\widehat{\beta}_n$ is bounded. Thus, the above concentration inequality implies that

$$P\left(\sup_{\beta:\|\beta\|_1^2 \leq \bar{B}} |\widehat{R}_n(\beta) - R(\beta)| > \epsilon\right) < (d+1)^2 2e^{-\frac{nc_3}{4\bar{B}^2}\epsilon^2}.$$

Finally, because $\widehat{\beta}_n$ is the minimizer of the empirical risk, i.e., $\widehat{R}_n(\widehat{\beta}_n) < \widehat{R}_n(\beta)$ for all β, on the event that $\sup_{\beta:\|\beta\|_1^2 \leq \bar{B}} |\widehat{R}_n(\beta) - R(\beta)| \leq \epsilon$, we have

$$R(\beta^*) \leq R(\widehat{\beta}_n) \leq \widehat{R}(\widehat{\beta}_n) + \epsilon \leq \widehat{R}(\beta^*) + \epsilon \leq R(\beta^*) + 2\epsilon.$$

Thus, we obtain the desired concentration bound.

■

A refined bound can be obtained in Theorem 11.3 of

> Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media,

which states the following (note that the result is stated in terms of estimation error).

Theorem 4.2 *Assume that* $\sup_x \mathsf{Var}(Y|X=x) < \infty$ *and* F_{XY} *are bounded and* Σ *is non-singular. Then*

$$\mathbb{E}\left(|\widehat{\beta}_n^T X - m(X)|^2\right) \leq 8 \inf_{\beta} \mathbb{E}(|\beta^T X - m(X)|^2) + \frac{Cd(\log n + 1)}{n},$$

where C *is some positive constant.*

As is known that $\widehat{\beta}_n$ has asymptotic normality around β. Under simple regularity conditions,

$$\sqrt{n}(\widehat{\beta}_n - \beta^*) \xrightarrow{d} N(0, \Omega),$$

where

$$\Omega = \Sigma^{-1}\mathbb{E}(\epsilon_*^2 X X^T)\Sigma^{-1} = \Sigma^{-1} M \Sigma^{-1},$$

where $\epsilon_* = Y - \beta^{*T} X$. A consistent estimator of Ω is

$$\widehat{\Omega}_n = \widehat{\Sigma}_n^{-1}\widehat{M}_n\widehat{\Sigma}_n^{-1}, \quad \widehat{M}_n = \frac{1}{n}\sum_{i=1}^n e_i^2 X_i X_i^T,$$

where $e_i = Y_i - \widehat{\beta}_n^T X_i$ is the residual. $\widehat{\Omega}_n$ is also called the *sandwich* estimator.

The above results do not assume that a linear model is correct–it is for the best linear predictor. We can use the sandwich estimator to construct a confidence interval of β. Note that we can also use the *bootstrap* method in this case.

Here is one caveat. In many standard textbooks, there is a common formula for computing the standard errors of the regression coefficients:

$$\widetilde{\Omega}_n = \widehat{\Sigma}_n^{-1}\widehat{\sigma}^2, \quad \widehat{\sigma}^2 = \frac{1}{n-d-1}\sum_{i=1}^n e_i^2.$$

The estimator $\widetilde{\Omega}_n$ is not the sandwich estimator; $\widetilde{\Omega}_n$ works only if 1. the linear model is correct, and 2. the error is homogenous. It is a consistent estimator if the linear model is correct. So you have to be very careful about the conclusion when using this formula. On the other hand, if you are using the sandwich estimator or the bootstrap approach, you can always interpret the confidence interval as covering the best linear predictor. More details are in

> Buja, Andreas, et al. "Models as approximations-a conspiracy of random regressors and model deviations against classical inference in regression." *Statistical Science* (2015): 1.

4.2 High Dimensional Linear Regression

In many cases, we may have many covariates so d is large. However, we believe that some of these covariates are useless covariates – the slope of these covariates are 0. Only a few covariates that have the actual linear relation with the response. Even we know this is true, if we naively apply the least square approach to find

β, we often have all fitted coefficients being non-zero and some of them could even be quiet significant just due to randomness of the data. Note that the least square estimator finds the fitted parameter as

$$\widehat{\beta}_{\mathsf{LSE}} = \underset{\beta}{\mathsf{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta^T X_i)^2.$$

There is something called the L_0 penalty. For a vector β, its L_0 norm is

$$\|\beta\|_0 = \text{ number of non-zero elements.}$$

We can also use the L_0 penalty in regression:

$$\widehat{\beta}_{\mathsf{Best}} = \underset{\beta}{\mathsf{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta^T X_i)^2 + \lambda \|\beta\|_0.$$

The resulting coefficients are related to the so-called *best subset estimators*.

However, a problem of the L_0 penalty is that finding the minimum of $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^T X_i)^2 + \lambda\|\beta\|_0$ is difficult. It is a non-convex problem and is an NP-hard problem (you can just view these two statements as 'computationally very very very difficult'). Thus, in many situations we will replace the L_0 penalty by an L_1 penalty because solving an L_1 penalty problem is still a convex problem, so computationally it is not very challenging. The process of replacing L_0 penalty (or other non-convex problem) by L_1 penalty (or other convex problem) is called *convex relaxation*. A common trick in machine learning and optimization.

The idea of penalization/regularization can help in this case. There are two comment penalized parametric regression model: (i) the ridge regression model, and (ii) LASSO (least absolute shrinkage and selection operator).

4.3 Ridge regression

The ridge regression added a penalty called the L_2 penalty in the minimization criterion. Namely, the ridge regression finds the fitted parameter as

$$\widehat{\beta}_{\mathsf{Ridge}} = \underset{\beta}{\mathsf{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta^T X_i)^2 + \lambda \|\beta\|_2^2,$$

where $\|\beta\|_2^2 = \sum_{j=1}^{d} \beta_j^2$ is the square 2-norm of the vector β. The penalty $\lambda\|\beta\|_2^2$ is called the L_2 penalty because it is based on the L_2 norm of the parameter.

It turns out that the ridge regression has a closed-form solution that is similar to the least square estimator and the spline:

$$\widehat{\beta}_{\mathsf{Ridge}} = \left(\mathbb{X}^T\mathbb{X} + n\lambda\mathbb{I}_d\right)^{-1} \mathbb{X}^T\mathbb{Y},$$

where $\mathbb{X}$ is the $n \times d$ data matrix and $\mathbb{I}_d$ is the $d \times d$ identity matrix.

Let $\widehat{\beta}_{\mathsf{LS}} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}$ be the ordinary least square estimator (no penalty, the classical approach). The ridge regression has a very similar coefficients as the least square estimator but just the coefficients are moved toward 0 because in the matrix inverse, there is an extra $n\lambda\mathbb{I}_d$ term. We will say that the ridge regression shrinks the estimator $\widehat{\beta}_{\mathsf{Ridge}}$ toward 0. As you would expect, the penalty λ trades off between the bias and variance. Large λ leads to a large bias but less variance.

When $\lambda \to 0$ properly, we may establish the consistency of ridge regression.

Theorem 4.3 (Hsu, Kakade, Zhang (2014)) *Assume that $\|X\| \le \bar{B}$ almost surely and Σ is non-singular. If the linear model is correct, i.e., the bias $b(x) = \beta^{*T}x - m(x) = 0$, then*

$$R(\widehat{\beta}_{\mathsf{Ridge}}) - R(\beta^*) = \left(1 + O\left(\frac{1+\bar{B}^2/\lambda}{n}\right)\right) \cdot \frac{\lambda\|\beta^*\|^2}{n} + \frac{\sigma^2}{n} \cdot \frac{\mathsf{tr}(\Sigma)}{2\lambda}.$$

This result can be found in Remark 15 of

> Hsu, Daniel, Sham M. Kakade, and Tong Zhang. "Random design analysis of ridge regression." *Conference on learning theory.* 2012.

Actually, they also derived the convergence rate when the linear model is incorrect–the consistency is with respect to the best linear predictor. If you are interested in ridge regression, you may check the references in the above paper.

The ridge regression can be viewed as a Bayesian estimator (posterior mean). To see this, we assume that the model $Y = \beta^T X + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ and place a prior over the parameter $\beta \sim N(0, \tau^2)$. Then you can show that the posterior mean is the ridge regression estimator with $\lambda = \frac{\sigma^2}{\tau^2}$.

Note that ridge regression is sometimes used in low-dimensional problem as well. One scenario that people would use ridge regression is that when the covariance matrix is singular or nearly singular. The ridge regression stabilizes the estimate.

4.4 LASSO

Recommended reference: Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations.* Chapman and Hall/CRC, 2015.

LASSO (least absolute shrinkage and selection operator) is one of the most famous penalized parametric regression model. It has revolutionized the modern statistical research because of its attractive properties. LASSO finds the regression parameters/coefficients using

$$\widehat{\beta}_{\mathsf{LASSO}} = \underset{\beta}{\mathsf{argmin}} \ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^T X_i)^2 + \lambda\|\beta\|_1 = \underset{\beta}{\mathsf{argmin}} \ \widehat{R}_n(\beta) + \lambda\|\beta\|_1, \tag{4.1}$$

where $\|\beta\|_1 = \sum_{j=1}^{d}|\beta_j|$ is the 1-norm of the vector β. The penalty $\lambda\|\beta\|_1$ is called the L_1 penalty. This is often known as the *Lagrangian/reguarlized LASSO*.

There is a dual form of the LASSO problem:

$$\underset{\beta}{\mathsf{minimize}} \ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^T X_i)^2, \qquad \text{subject to } \|\beta\|_1 \le t. \tag{4.2}$$

When t is chosen to be the value of $\widehat{\beta}_{\mathsf{LASSO}}$ under the λ in the original problem, we obtain the same result. This is often known as the *constrained LASSO*.

If we normalized the covariates so that $\mathbb{X}^T\mathbb{X} = \mathbb{I}_d$, the LASSO estimates can be written as

$$\widehat{\beta}_{\mathsf{LASSO,j}} = \widehat{\beta}_{\mathsf{LS,j}} \times \max\left\{0, 1 - \frac{n\lambda}{|\widehat{\beta}_{\mathsf{LS,j}}|}\right\}$$

for $j = 1, \cdots, d$. Namely, the coefficients from LASSO are those coefficients from the least square method shrinking toward 0 and for those parameters whose value are below $n\lambda$, they will be shrink to 0.

When λ is large or the signal is small, many coefficients will be 0. This is called **sparsity** in statistics (only a few non-zero coefficients). Thus, we will say that the LASSO outputs a **sparse** estimate. Those $\widehat{\beta}_j$ will be 0 if it does not provide much improvement on predicting Y. So it naturally leads to an estimator with an automatic **variable selection** property. The value of λ will affect the estimates $\widehat{\beta}$. Larger λ encourages a sparser $\widehat{\beta}$ (namely, more coefficients are 0) whereas smaller λ leads to a less sparse $\widehat{\beta}$.

Although ridge regression also shrinks the coefficients toward 0, it does not yield a sparse estimator. The coefficients are just smaller but generally non-zero. On the other hand, LASSO not only shrinks the values of coefficients but also set them to be 0 if the effect is very weak. Actually, this is a property of the L_1 penalty – it tends to yield a sparse estimator – an estimator with many 0's.

4.4.1 When linear model is correct

When the linear model is correct, the LASSO is consistent under good conditions.

For a linear regression model, we say that the model is s-sparse if there are at most $s < d$ coefficients that are non-zero. Namely, $\|\beta^*\|_0 = \sum_{j=1}^{d} I(\beta_j \neq 0) \leq s$. For an s-sparse model, without lost of generality, we reorder the coefficients such that

$$\beta^* = (\beta_1^*, \beta_2^*, \cdots, \beta_s^*, 0, 0, \cdots, 0),$$

i.e., only the first s coefficients are possibly non-zero and all these first s values are non-zero. Let $S = \{1, 2, \cdots, s\}$ be the support set (the indices of coefficients that are non-zero). In the high dimensional mode, we write $s = s_n$ to allow the sparsity to change with respect to the sample size and $d = d_n$ to allow the number of parameters to increase as well.

Here we display a convergence rate of LASSO from the following book:

> Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations.* Chapman and Hall/CRC, 2015.

In particular, chapter 11 discusses a couple of other results on the LASSO theory.

This result is based on the *restrictive eigenvalue condition*. Recall that $S = \{1, 2, 3, \cdots, s\}$ is the collection of parameters with non-zero coefficients. Define the set

$$\mathcal{C}(S, \alpha) = \{\beta : \|\beta_{S^C}\|_1 \leq \alpha \|\beta_S\|_1\},$$

where $\beta_S = (\beta_1, \cdots, \beta_s)$ and $\beta_{S^c} = \{\beta_{s+1}, \cdots, \beta_d\}$. The design matrix $\widehat{\Sigma}_n$ is called to have *restrictive eigenvalue* with parameter γ over class $\mathcal{C}(S, \alpha)$ if

$$\min_{\nu \in \mathcal{C}(S,\alpha)} \frac{\nu^T \widehat{\Gamma}_n \nu}{\nu^T \nu} \geq \gamma.$$

With this condition, the LASSO has the following asymptotics

Theorem 4.4 (Theorem 11.1 in Hastie, Tibshirani, and Wainwright (2015)) *Assume the followings:*

1. *The linear model is correct and $s-$sparse.*
2. *The design matrix satisfies the restrictive eigenvalue condition with γ over class $\mathcal{C}(S, 3)$.*

Then 1. for the Lagrangian LASSO in equation (4.1) with parameter $\lambda \geq 2\|\sum_{i=1}^n X_i\epsilon_i\|_\infty/n > 0$,

$$\|\widehat{\beta}_{\mathsf{LASSO}} - \beta^*\| \leq \frac{3}{\gamma}\sqrt{s}\lambda.$$

2. for the constrained LASSO in equation (4.2) with $\|\widehat{\beta}_{\mathsf{LASSO}}\|_1 \leq \|\beta^*\|_1$,

$$\|\widehat{\beta}_{\mathsf{LASSO}} - \beta^*\| \leq \frac{4}{\gamma}\sqrt{\frac{s}{n}}\left\|\frac{\sum_{i=1}^n X_i\epsilon_i}{\sqrt{n}}\right\|_\infty.$$

The proof of the constrained form is simple and inspiring so here we display the proof.

Proof: Consider the empirical risk

$$\widehat{R}_n(\beta) = \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^T\beta)^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i^T(\beta - \beta^*) + \epsilon_i\right)^2.$$

This implies

$$\begin{aligned}
\widehat{R}_n(\widehat{\beta}_{\mathsf{LASSO}}) &= \frac{1}{n}\sum_{i=1}^n \left(X_i^T \underbrace{(\widehat{\beta}_{\mathsf{LASSO}} - \beta^*)}_{=-\delta_\beta} + \epsilon_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left(X_i^T\delta_\beta - \epsilon_i\right)^2 \\
&\leq \widehat{R}_n(\beta^*) = \frac{1}{n}\sum_{i=1}^n \epsilon_i^2,
\end{aligned}$$

where $\delta_\beta = \beta^* - \widehat{\beta}_{\mathsf{LASSO}}$.

Thus, after rearrangements,

$$\delta_\beta^T\widehat{\Gamma}_n\delta_\beta = \frac{1}{n}\sum_{i=1}^n (X_i^T\delta_\beta)^2 \leq \frac{2\delta_\beta^T}{n}\sum_{i=1}^n X_i\epsilon_i. \tag{4.3}$$

For the right-hand side, the Holder's inequality implies that

$$\left|\frac{2\delta_\beta^T}{n}\sum_{i=1}^n X_i\epsilon_i\right| \leq 2\|\delta_\beta\|_1\left\|\frac{1}{n}\sum_{i=1}^n X_i\epsilon_i\right\|_\infty. \tag{4.4}$$

Note that one can show that the constraint $\|\widehat{\beta}_{\mathsf{LASSO}}\|_1 \leq \|\beta^*\|_1$ implies that $\delta_\beta \in \mathcal{C}(S,1)$, which further implies

$$\|\delta_\beta\|_1 = \|\delta_{\beta,S}\|_1 + \|\delta_{\beta,S^c}\|_1 \leq 2\|\delta_{\beta,S}\|_1 \leq 2\sqrt{s}\|\delta_{\beta,S}\|_2 \leq 2\sqrt{s}\|\delta_\beta\|_2,$$

where the last inequality is due to Cauchy-Schwarz inequality. Thus, we can rewrite equation (4.4) by

$$\left|\frac{2\delta_\beta^T}{n}\sum_{i=1}^n X_i\epsilon_i\right| \leq 4\sqrt{s}\|\delta_\beta\|_2\left\|\frac{1}{n}\sum_{i=1}^n X_i\epsilon_i\right\|_\infty.$$

Thus, after rearrangements, equation (4.3) becomes

$$\begin{aligned} \delta_\beta^T \widehat{\Gamma}_n \delta_\beta &\le 4\sqrt{s}\|\delta_\beta\|_2 \left\| \frac{1}{n}\sum_{i=1}^n X_i\epsilon_i \right\|_\infty \\ \Rightarrow \|\delta_\beta\|_2 \cdot \underbrace{\frac{\delta_\beta^T \widehat{\Gamma}_n \delta_\beta}{\|\delta_\beta\|_2^2}}_{\ge \gamma} &\le 4\sqrt{s} \left\| \frac{1}{n}\sum_{i=1}^n X_i\epsilon_i \right\|_\infty \\ \Rightarrow \|\delta_\beta\|_2 &\le \frac{4}{\gamma}\sqrt{s} \left\| \frac{1}{n}\sum_{i=1}^n X_i\epsilon_i \right\|_\infty, \end{aligned}$$

which completes the proof.

■

Note that under the quantity $\|\sum_{i=1}^n X_i\epsilon_i\|_\infty$ can be bounded using the concentration inequality. When ϵ is sub-Gaussian, i.e., $\log \mathbb{E}(e^{t\epsilon}) \le \sigma^2 e^2$ for some finite number $\sigma^2 > 0$ and any $t > 0$. we have

$$\left\| \sum_{i=1}^n X_i\epsilon_i \right\|_\infty \le O_P(\sqrt{n \log d}).$$

Using this fact, Theorem 4.4 implies that

$$\|\widehat{\beta}_{\mathsf{LASSO}} - \beta^*\| = O_P\left(\sqrt{\frac{s \log d}{n}}\right).$$

There are many other theoretical work on the convergence of LASSO. Here is another example. For a matrix C, we define its m-sparse minimum and maximum eigenvalues as

$$\phi_{\min}(m; C) = \min_{\beta:\|\beta\|_0 \le \lceil m \rceil} \frac{\beta^T C \beta}{\beta^T \beta}, \quad \phi_{\max}(m; C) = \max_{\beta:\|\beta\|_0 \le \lceil m \rceil} \frac{\beta^T C \beta}{\beta^T \beta}.$$

These quantity are related to the restricted isometry property (RIP)[1].

Theorem 4.5 (Meinshausen and Yu (2006)) *Assume the followings:*

1. *The linear model is correct.*
2. *The covariates are bounded and the design matrix is standardized (i.e, the diagonal of $\widehat{\Sigma}_n$ consists of 1's.)*
3. *The noise ϵ_i is sub-Exponential, i.e, $\mathbb{E}(e^{|\epsilon_i|}) < \infty$, and has variance $\mathsf{Var}(\epsilon_i) = \sigma^2 < \infty$.*
4. *There exists $0 < \kappa_{\min} \le \kappa_{\max} < \infty$ such that*

$$\liminf_n \phi_{\min}(s_n \log n; \widehat{\Sigma}_n) \ge \kappa_{\min}, \quad \limsup_n \phi_{\max}(s_n + \min\{n, d_n\}; \widehat{\Sigma}_n) \le \kappa_{\max}.$$

5. $\lambda \propto \sigma\sqrt{n \log d_n}$.

[1] https://en.wikipedia.org/wiki/Restricted_isometry_property

Then there exists M such that with a probability tending to 1

$$\|\widehat{\beta}_{\mathsf{LASSO}} - \beta^*\|^2 \leq M\sigma^2 \frac{s_n \log p_n}{n}.$$

Sometimes, you will see that people write $\|\widehat{\beta}_{\mathsf{LASSO}} - \beta^*\| = O_P\left(\sqrt{\frac{s_n \log p_n}{n}}\right)$. This is the common rate for the LASSO estimator. The above theorem is from

> Meinshausen, Nicolai, and Bin Yu. "Lasso-type recovery of sparse representations for high-dimensional data." *The annals of statistics* 37.1 (2009): 246-270.

Note that a design matrix $\widehat{\Sigma}_n$ is called an *incoherent* design if there exists a sequence e_n (also known as *sparsity multiplier sequence*) such that

$$\liminf_{n\to\infty} \frac{\phi_{\min}(e_n s_n^2; \widehat{\Sigma}_n)}{\phi_{\max}(s_n + \min\{m, d_n\}; \widehat{\Sigma}_n)} \geq 18.$$

A more general result can be obtained using the incoherent design.

There is one condition that is particularly restrictive in Theorem 4.5: the condition on the eigenvalues (4th condition). A similar condition is the restrictive eigenvalue condition in Theorem 4.4. Essentially, we need the design matrix to behave almost like an orthonormal matrix. For problems like compressive sensing, this is possible since we can manipulate the design matrix but for many other problems such as genetic studies, the design matrix refers to the gene-gene interaction matrix, which is known to fail this condition.

4.4.2 When linear model is not correct

There is less literature about the behavior of LASSO when the model is incorrect. Here we present a theorem about the convergence of predictive risk of LASSO when the model is incorrect. Note that the convergence here refers to the convergence to a 'population LASSO'. We use the dual form of LASSO to simplify the problem.

Theorem 4.6 *Assume that $|Y| \leq B$ and $\|X\|_{\max} \leq B$. Define the population LASSO*

$$\beta^*_{\mathsf{LASSO}} = \mathsf{argmin}_{\beta:\|\beta\|_1 \leq L} \mathbb{E}(Y_i - \beta^T X_i)^2 = \mathsf{argmin}_{\beta:\|\beta\|_1 \leq L} R(\beta)$$

and the LASSO estimator

$$\widehat{\beta}_{\mathsf{LASSO}} = \mathsf{argmin}_{\beta:\|\beta\|_1 \leq L} \widehat{R}_n(\beta).$$

With a probability of at least $1-\delta$, we have

$$R(\widehat{\beta}_{\mathsf{LASSO}}) \leq R(\beta^*_{\mathsf{LASSO}}) + \sqrt{\frac{8(L+1)^4 B^2}{n} \log\left(\frac{2d^2}{\delta}\right)}.$$

Proof: Define $Z = (Y, X)$ and $Z_i = (Y_i, X_i)$ and $\underline{\beta} = (-1, \beta)$. The prediction risk can be written as

$$R(\beta) = \underline{\beta}^T \Gamma \underline{\beta},$$

where $\Gamma = \mathbb{E}(ZZ^T)$.

Similarly, the empirical prediction risk is

$$\widehat{R}_n(\beta) = \underline{\beta}^T \widehat{\Gamma}_n \underline{\beta},$$

where $\widehat{\Gamma}_n = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T$.

For any parameter β, the difference can be written as

$$\begin{aligned}
\widehat{R}_n(\beta) - R(\beta) &= \underline{\beta}^T(\widehat{\Gamma}_n - \Gamma)\underline{\beta} \\
&\leq \sum_{j,k} |\underline{\beta}_j||\underline{\beta}_k|[\widehat{\Gamma}_n - \Gamma]_{j,k} \\
&\leq \|\underline{\beta}\|_1^2 \|\widehat{\Gamma}_n - \Gamma\|_{\max} \\
&\leq (L+1)^2 \|\widehat{\Gamma}_n - \Gamma\|_{\max}.
\end{aligned}$$

By setting $\eta = (L+1)^2\|\widehat{\Gamma}_n - \Gamma\|_{\max}$, we have

$$R(\widehat{\beta}_{\mathsf{LASSO}}) \leq \widehat{R}_n(\widehat{\beta}_{\mathsf{LASSO}}) + \eta \leq \widehat{R}(\beta^*_{\mathsf{LASSO}}) + \eta \leq R(\beta^*_{\mathsf{LASSO}}) + 2\eta.$$

Using the Hoeffding's inequality,

$$P(\|\widehat{\Gamma}_n - \Gamma\|_{\max} > \epsilon) < d^2 2e^{-\frac{n\epsilon^2}{2B^2}}.$$

Thus, by setting $d^2 2e^{-\frac{n\epsilon^2}{2B^2}} = \delta$, we obtain

$$\epsilon = \sqrt{\frac{2B^2}{n}\log\left(\frac{2d^2}{\delta}\right)}$$

Plugging this into η, we conclude that

$$R(\widehat{\beta}_{\mathsf{LASSO}}) \leq R(\beta^*_{\mathsf{LASSO}}) + \sqrt{\frac{8(L+1)^4B^2}{n}\log\left(\frac{2d^2}{\delta}\right)}.$$

■

Note that a more general version appears in the following paper:

> Greenshtein, Eitan, and Ya'Acov Ritov. "Persistence in high-dimensional linear predictor selection and the virtue of overparametrization." *Bernoulli* 10.6 (2004): 971-988.

Remark (sparsistency). Another way to derive the convergence of LASSO is via the concept of *sparsistency*. An estimator $\widehat{\beta}$ is sparsisteny if its non-zero element is the same as the non-zero element of β^* with a high probability, i.e.,

$$P(\mathsf{supp}(\widehat{\beta}) = \mathsf{supp}(\beta^*)) \to 1,$$

where $\mathsf{supp}(\beta) = \{\beta_j : \beta_j \neq 0\}$. Under good assumptions, the LASSO estimator has sparsistency; see, e.g.,

> Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research*, 7(Nov), 2541-2563.

Remark (WARNING on LASSO). Although we have beautiful theorems on LASSO under sparse and high-dimensional settings, these theorems may not be applicable to the real data. In particular, the restrictive eigenvalue condition is often a too strong condition. It basically requires the covariates to be almost uncorrelated (or even independent). When analyzing genetic data or images from fMRI, it is well known that the covariates (genes or voxel values) are highly correlated with each other. So the theorem is not applicable in this case and we have no idea how will the LASSO behaves (although LASSO is still commonly used in these secnarios). One situation that the restrictive eigenvalue condition works is *compressed sensing*–we can design the covariate so that the restrictive eigenvalue conditions can be obtained by design[2].

4.5 Inference in high dimensional regression

Inference in high-dimensional case is very challenging. The major reason is that the convergence rate we obtain is often done by empirical risk minimization approach. This is different from the usual analysis that we perform a Taylor expansion over the objective function. Despite the challenges, there are still some advancements in this direction. In general, there are two common directions for high-dimensional inference.

Sequential testing and post-selection inference. The first approach considers a sequential procedure of including one and one variable. The challenge is that this procedure runs in to the post-selection inference problem that at each stage, our hypothesis testing depends on all the previously selected parameters. Some famous references are:

- Lockhart, Richard, et al. "A significance test for the lasso." *Annals of statistics* 42.2 (2014): 413.
- Tibshirani, Ryan J., et al. "Exact post-selection inference for sequential regression procedures." *Journal of the American Statistical Association* 111.514 (2016): 600-620.
- Lee, Jason D., et al. "Exact post-selection inference, with application to the lasso." *The Annals of Statistics* 44.3 (2016): 907-927.

Debiased/Desparsified approach. The debiased/desparsified LASSO is another common approach for high-dimensional inference. The main idea is: although the LASSO estimator does not have asymptotic normality when d_n increases much faster than n, the debiased version of the LASSO estimator still have (LASSO estimator minus an estimate of the bias). An interesting fact about the debiased LASSO estimator is no longer a sparse estimate–most of its parameter estimates are non-zero. So people also called it a desparsified LASSO. Here are some famous papers about this idea:

- Zhang, Cun-Hui, and Stephanie S. Zhang. "Confidence intervals for low dimensional parameters in high dimensional linear models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014): 217-242.
- Van de Geer, Sara, et al. "On asymptotically optimal confidence regions and tests for high-dimensional models." *The Annals of Statistics* 42.3 (2014): 1166-1202.
- Javanmard, Adel, and Andrea Montanari. "Confidence intervals and hypothesis testing for high-dimensional regression." *The Journal of Machine Learning Research* 15.1 (2014): 2869-2909.

[2] see https://normaldeviate.wordpress.com/2012/08/07/rip-rip-restricted-isometry-property-rest-in-peace/ for more discussion.

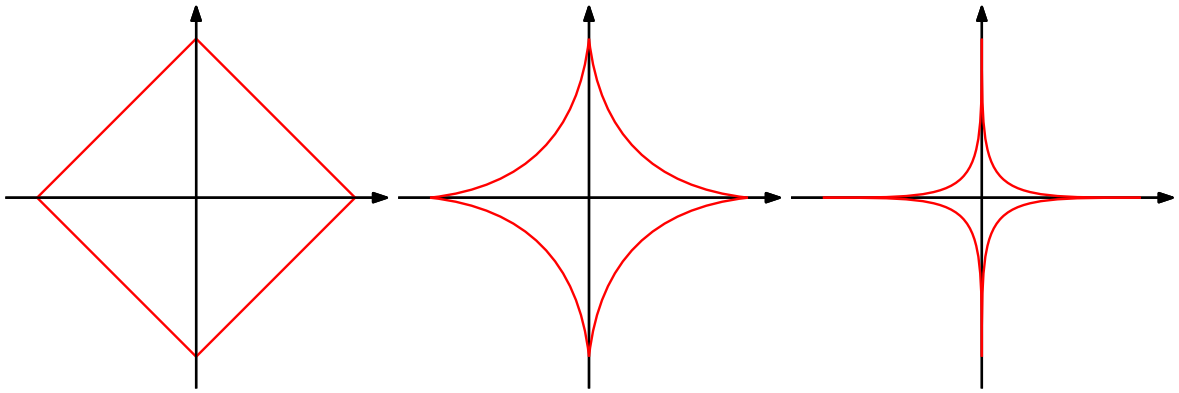

Figure 4.1: How the L_1 norm looks like under different dimensions. The left panel displays the L_1 norm $\|x\|_1$ at $d = 2$. The middle to right panel show the L_1 norm under higher dimensions.

4.6 High-dimenisonal geometry

Why L_1 penalty leads to a sparse estimator? One simple way to explain this is via the high-dimensional geometry. In fact, the geometry in high dimension could be very different from low dimension. To start with, we examine how the L_1 norm behaves when the dimension is high.

4.6.1 L_1 norm in high-dimensions

The first thing that the high-dimensional geometry is very different from the low dimensional geometry is the *shape* of L_1 norm level set. Consider the set

$$B = \{\beta \in \mathbb{R}^d : \|\beta\|_1 \leq 1\}.$$

What will this set looks like relative to the set $[-1, 1]^d$?

In $d = 1$ case, it covers the entire region. In $d = 2$ case, it covers half of the region. In $d = 3$ case, you can show that it covers actually $1/4$ of the region $[-1, 1]^3$.

Then what would happen when d is large? It turns out that this L_1 level set covers $\frac{1}{2^{d-1}}$ volume of the region $[-1, 1]^d$, which means that the regions cover by B will only cover a tiny fraction of the region $[-1, 1]^d$ when d is large and the set B will be the regions around the coordinate axes. Figure 4.1 provides a graphical illustration on this.

The illustration in Figure 4.1 implies that the L_1 norm behaves like a spiky structure under high dimensions. The shape of a squared loss is an ellipse (contour of the squared loss). Thus, when an ellipse hits a spiky structure, it is very like that the hitting point is on the spike, i.e., some parameters are 0. This is why L_1 regularization often leads to a sparse estimator.

In fact, any L_q norm regularization with $q \leq 1$ leads to a sparse estimator. Another interesting fact: the minimization problem of L_q regularization is NP-hard if $q < 1$; or informally, you can say that L_q regularization is 'computable' if $q \geq 1$. We are very fortunate that the intersection of a sparse estimator (requiring $q \leq 1$) and a computable estimator (requiring $q \geq 1$) has an intersection at $q = 1$. Thus, L_1 regularization is a blessing zone that we can enjoy a sparse and computable estimator.

4.6.2 High-dimensional Gaussian

Another bizarre phenomenon of high dimensional geometry occurs when we are working with high-dimensional multivariate Gaussian. To simplify the problem, we consider a d-dimensional Gaussian with unit variance. Let

$$X \sim N(0, \mathbf{I}_d),$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix. The PDF will be

$$p(x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{d} x_j^2\right).$$

This density is symmetric at 0 and decrease with respect to the distance from the origin $r = \sqrt{\sum_{j=1}^{d} x_j^2} = \|x\|_2$. Now we consider the following question: if we are thinking about the density as a function of distance to the origin, which radius will most of the probability mass concentrate?

To study this, we convert the PDF of coordinate x into a PDF with respect to the radius r. Using the polar coordinate transform and the fact that $p(x)$ is isotropic,$dx = r^{d-1} S_{d-1} dr$, where S_{d-1} is the $d-1$ dimensional surface volume of the unit ball $\{x : \|x\|_2 = 1\}$. Thus, the PDF will be

$$p(r) = (2\pi)^{-d/2} S_{d-1} r^{d-1} \exp\left(-\frac{1}{2}r^2\right) \propto r^{d-1} e^{-\frac{1}{2}r^2}.$$

What will the mean and variance be and what will the mode be? Let R be the random variable with a PDF $p(r)$.

A simple approach to compute the mean and variance is to use the fact that by setting $R^2 = S$, we obtain

$$p(s) \propto s^{\frac{d-2}{2}} e^{-\frac{1}{2}s} \sim \mathsf{Gamma}\left(\alpha = \frac{d}{2}, \beta = \frac{1}{2}\right).$$

Using the properties of Gamma distribution, we conclude that

$$\begin{aligned} \mathbb{E}(S) &= d \\ \mathsf{Var}(S) &= 2d \\ \mathsf{Mode}(S) &= d - 2. \end{aligned}$$

What does this tell us about random variable S when d is large? A crucial implication is that the mean and the variance are of the same order, meaning that the standard deviation will be of the order $\sqrt{d}$. Thus, if we are thinking about S rescaled by its mean, then $\frac{S}{\mathbb{E}(S)} \xrightarrow{P} 1$. Also, since $\frac{|\mathbb{E}(S) - \mathsf{Mode}(S)|}{\mathbb{E}(S)} \to 0$,

$$\frac{S}{\mathsf{Mode}(S)} \xrightarrow{P} 1.$$

Note that the mode of S is the squared of the mode of R, i.e., $\mathsf{Mode}(R) = \sqrt{d-1}$. Using the continuous mapping theorem, we conclude that

$$\frac{R}{\mathsf{Mode}(R)} \xrightarrow{P} 1.$$

Namely, all probability mass will concentrate around the mode of R when we rescale the entire distribution so that the mode occurs at radius 1! In a sense, this implies that the distribution $p(x)$ puts almost all its probability mass around the shell $\|x\|_2 = \sqrt{d}$!

4.6.3 Volume of a high-dimensional Ball

Another striking result about high-dimensional geometry is the fact that

> *most of the volumes of a high dimensional ball or cube are close to the boundary.*

To see this, note that for a d-dimensional ball with a radius R, its volume is

$$V_d(R) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} R^d,$$

where $\Gamma(\cdot)$ is the Gamma function. Thus, the ratio of a ball with unit length ($R = 1$) versus with radius $1 - \epsilon$ is

$$\frac{V_d(1-\epsilon)}{V_d(1)} = (1-\epsilon)^d.$$

When $\epsilon = r/d$, this quantity converges to e^{-r}, which decrease rapidly when r increases. Thus, most of the volume is within $\epsilon = O(1/d)$ to the boundary, which means that the majority of the volume is around the boundary. Or alternatively, if we randomly choose a point within a high dimensional ball, it is very likely that this point is within $O(1/d)$ distance to the boundary. Not only the ball, a high dimensional cube also has a similar property– most of the volume is very close to the boundary.