Statistical Learning Theory, Oracle inequality, and the Lasso

Colin Cui *

December 26, 2020

Abstract

This paper attempts to present ideas of theorems and proofs in recent advances of statistical learning theory and optimization from an approachable angle. In light of that, most of the tools we use here are from college math and statistics courses. The difficulty in these proofs is figuring out when to use what. We present them here in a gentle and friendly manner.

***Keywords*—** Sparsity, Sub-Gaussian, Risk bounds, Oracle inequalities, Optimal rates, Concentration inequalities, Gradient descent

1 Introduction

Statistics theory and machine learning have experienced a wave of transformation. From John Tukey's first EDA textbook in the early 1907s to high-dimensional statistics today, the challenge has advanced due to computation advancement and the influx of big data. In early days, the statistics problems are often concerned with estimations such as MLE, unbiasness, and their asymptotical consistency. Today, we are concerned about learning a model f from large-scale data and make accurate predictions. The curse of dimensionality is one of the biggest roadblocks on the journey of high-dimension. Some prominent breakthroughs along the way include ridge, Lasso, and other regularization methods. In this paper, we present the ideas and tools for proofs in a simple and colloquial manner that is amicable to the audience.

2 Towards theory of high-dimensional statistics

2.1 Background

In high-dimensional statistics, the number of variables p in a model is much larger than the sample size n, namely, $p \gg n$. The curse of dimensionality [1] is omnipresent. There are a large number of variables, but receive only relative small sample size, and in turn demands accurate prediction, this is a mathematical challenge for researchers. Fortunately, the underlying model f is often assumed to be sparse or weakly sparse, i.e. with many zeros coefficients or very small, which leads to the notion of *regularization*.

*456 Snowden Ln, Princeton, NJ, 08540. Email: colstat@gmail.com

We need on the order of $(1/\epsilon)^p$ evaluations on a hypersphere grid to achieve an approximation error of ϵ. In other words, the number data points needed to explore the sample space in high-dimension grows exponentially in p.

There is also a geometric view to this. One can think of a d-dimensional hypersphere unit-ball. Let the volume of this hyperball be vol($\mathcal{B}$). It tends to 0 as $d \to \infty$. When we decrease the radius of the hyperball by ϵ, resulting in a new radius $1-\epsilon$), then the volume of the hyperball shrinks by this factor:

$$\frac{vol((1-\epsilon)\mathcal{B})}{vol(\mathcal{B})} = (1-\epsilon)^d \leq e^{-\epsilon d}.$$

Choosing any small ϵ for the last term and hold it fixed. The ratio of the volumes goes towards zero as dimension $d \to \infty$. Suppose the hyperball has radius r, then most of the data points "live" on the surface of the ball between radius $r \in [(1-\epsilon), 1]$. The hyperball has a width of $O(1/d)$.

2.2 Some notation

For a vector v, our notation for ℓ_p norm is,

$$||v||_q = \begin{cases} \sum_i 1\{v_i \neq 0\}, & \text{if } q = 0 \\ (\sum_i |v_i|^q)^{1/q}, & \text{if } 0 < q < \infty \\ \max_i |v_i|, & \text{if } q = \infty \end{cases}$$

2.3 Linear model

In statistical learning theory, we are concerned with finding a model f that fits the data with some error of the form $y = f(x) + \epsilon$. In linear regression, we take f to be $f(X) = X\beta$. Let X be the matrix of input with columns $X_1, ..., X_p \in \mathbb{R}^n$, the response $y_1, .., y_n \in \mathbb{R}^n$, the coefficients $\beta_1, ..., \beta_p \in \mathbb{R}^p$ is unknown, and the errors $\epsilon_1, ..., \epsilon_n \in \mathbb{R}^n$ has $\epsilon_i \sim \mathbb{N}(0, \sigma^2)$.

$$y = X\beta + \epsilon \tag{1}$$

Give f is a linear regression, we can easily solve for the least squares estimate,

$$\hat{\beta} = X^T X^{-1} X^T y.$$

For any f, we define *true risk* to be $R(f) = \mathbb{E}[Y - f(Y)]^2$ (also called generalization error, test error, or prediction error). We estimate the true risk with an *empirical risk*, which depends on n pairs data samples $(X_1, Y_1), ..., (X_n, Y_n)$.

$$R_n = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2. \tag{2}$$

Here R_n is called the empircal risk (training error). Our goal is to find the $\hat{f}_n$ from a family of $\mathcal{F}$ in the function space by *empirical risk minimization* (ERM). Additionally, we denote $f^* = \arg\min_f \mathbb{E}[f]$ to be the best of all possible functions. Then, by the Pythagorean theorem, we decompose the estimation error

$$||\hat{f}_n - f||^2_{L_2(P_X)} = ||\hat{f}_n - f^*||^2_{L_2(P_X)} + ||f^* - f||^2_{L_2(P_X)}.$$

Now, we can define the excessive risk $\mathcal{E} \triangleq R(\hat{f}_n) - R(f^*)$ using the approximation-estimation decomposition (also called bias-variance trade-off).

$$\mathcal{E} = \underbrace{R(\hat{f}_n) - R(f^*)}_{\text{estimation error}} + \underbrace{R(\hat{f}^*) - R(f_{\mathcal{F}})}_{\text{approximation error}}. \tag{3}$$

The *esimation error* comes from using finite sample data, using the emprical risk rather than the true risk, and also from the complexity of function class $\mathcal{F}$. The *approximation error* depend on function class $\mathcal{F}$ A larger function class results in wider exploration of possible functions, and that would drive down the approximation error.

2.4 Penalized regression

When $p \gg n$ in linear regression, we can not obtain the least squares estimate. Hence, no closed-form solution in β without imposing some other restrictions. One central theme is to introduce an additional penalty term to the loss function, then we can reduce the dimension of p and recover a subset of the β coefficients and produce sparsity of the underlying regression model. This falls under the general regularization regime for constrained optimization.

$$\hat{\beta}^{pen} = \arg\min_{\beta \in \mathbb{R}^p} \left[R_n(\beta) + Pen(\beta) \right]. \tag{4}$$

Here $R_n(\beta)$ denotes empirical risk, and $Pen(\beta)$ denotes the regularization or penality term. We define the following penalized regressions.

$$\hat{\beta}^{subset} = \arg\min_{\beta \in \mathbb{R}^p} \left[\frac{1}{n} ||y - X\beta||_2^2 + \lambda ||\beta||_0 \right] \qquad \text{(best subset)} \tag{5}$$

$$\hat{\beta}^{Lasso} = \arg\min_{\beta \in \mathbb{R}^p} \left[\frac{1}{n} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \right] \qquad \text{(Lasso)} \tag{6}$$

$$\hat{\beta}^{Ridge} = \arg\min_{\beta \in \mathbb{R}^p} \left[\frac{1}{n} ||y - X\beta||_2^2 + \lambda ||\beta||_2^2 \right] \qquad \text{(Ridge)} \tag{7}$$

Table 1: ℓ_q-norm and thresholding penality

Penalty	Paper	Name	Penalty type	Advantages
ℓ_0	—	Best subset	select	unbiased, sparse
ℓ_1	Tibshirani, 1996	LASSO	shrink, select	sparse, continuous
ℓ_2	Hoerl and Kennard, 1970	Ridge	shrink	continuous
$\ell_1 + \ell_2$	Zou and Hastie, 2005	Elastic net	shrink	continuous
SCAD	Fan, Li, 2001	SCAD	shrink, select	unbiased, sparse, continuous

The goal of these methods is to produce variable selection and sparsity in β. The ℓ_0-norm penalty is one the method to produce sparsity. Methods such as AIC and BIC achieves best subset selection by penalizing the negative log-likelihood, and producing sparsity in the full model. Solving the

ℓ_0 penalty is NP-hard. So in optimization, ℓ_1 is often used as a surrogate to ℓ_0. This is called the *basis pursuit*.With this convex relaxation, there is a drawback. Under the restricted isometry properties (RIP), the ℓ_1 penalty Lasso does not exactly recover the same set of nonzero cofficients, because of the looseness of the relaxation [12].

2.5 Tail bounds and concentration inequalties

Most of mass in high-dimension is empty and the mass is concentrated in a small surface of the hypersphere. The following concentration measure in probability theory guarantees a probabilistic bound under certain mild conditions. And that they are at most ϵ away from the mean. This is also the blessing of dimensionality [3]. We will also use some in our proofs.

Proposition 1 (Gaussian tail bound)**.** *Let X be a Gaussian distribution with mean μ, variance σ^2, then for all $t > 0$*

$$P(X \geq \mu + t) \leq exp(-t^2/2\sigma^2)$$

The proposition above says that majority of the probability mass is concentrated only in a small region of the Gaussian distribution. For example, the empirical rule says that 66%, 95%, and 99% of the mass of the distribution is within σ, 2σ, and 3σ of the mean.

Definition 1 (sub-Gaussian)**.** A random vector $X \in \mathbb{R}^{\ltimes}$ is $SG(\sigma^2)$ or sub-Gaussian, if there exists a positive variance proxy σ^2 such that,

$$E[\exp(t(X - E(X)))] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right), \quad \text{for all } t \in \mathbb{R},$$

By definition, any Gaussian random variable is also a sub-Gaussian $SG(\sigma^2)$. This is a relaxation on the Gaussian random variable that we learned in statistics classes. It only requires the distribution to have an exponential tail off.

Theorem 2 (Chernoff inequality)**.** *For any $t > 0$, we have*

$$P(|S_n| \geq t) \leq \exp(-at)\, \mathbb{E}[\exp(-aS_n)].$$

The Chernoff bound applies to the Bernoulli random variables. It tells us that as long as the probability of an incorrect answer is less than 1/2 on any particular trial, then the probability that the majority of trials will give incorrect answers decreases exponentially with the number of trials.

Theorem 3 (Hoeffding's inequality)**.** *Suppose that $X_1, ..X_n$ are independent random variables. Let $S_n = \sum_{i=1}^n (X_i - EX_i)$, such that $X_i \in [a_i, b_i]$ almost surely for all $i \leq n$, then for $t > 0$*

$$P(|S_n| \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

Hoeffding's inequality is a more general use case of Chernoff bound, where is it not limited to Bernoulli random variable.

Theorem 4 (Bernstein's inequality)**.** *Suppose that $X_1, ..X_n$ are independent random variables satisfying $|X_i| \leq a$ and $EX_i^2 = \sigma^2$. Let $S_n = \sum_{i=1}^n (X_i - EX_i)$, then,*

$$P(|S_n| \geq t) \leq 2\exp\left(\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\right).$$

Bernstein inequality provides an upper bound on the sum of n random variables. This can be useful when bounding the sum of any random variables, not limited to exponential distributions as in the Chernoff inequality.

Theorem 5 (Maximal inequality)**.** *Let* $X_1, ..X_n \sim SG(\sigma^2)$ *with* $\mathbb{E}[X] = 0$*. For any* $t > 0$*, we have*

$$\mathbb{E}[\max_{1\leq i\leq p} |X_i|] \leq \sigma\sqrt{2\log(2n)}$$

$$\mathbb{P}(\max_{1\leq i\leq p} |X_j - \mathbb{E}X| \geq t) \leq 2n\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{8}$$

The sub-Gaussian random variables can be correlated. This fact will be useful when proving max norm of the Lasso estimator.

2.6 Oracle for model misspecification

"All models are wrong, but some are useful." In the real world, most underlying data generating process is probably not linear, yet linear regression is the ubiquitous choice. Though it is not correct, it is popular for its simplicity and interpretability. But there is a price to be paid using a simple model to describe a complex world.

$$R_n(\hat{f}) \leq ||f^* - f||^2_{L_2(P_X)} + \psi_{n,p}$$

This is the oracle inequality by definition from [2]. The last term $\psi_(n, p)$ depends on dimension p. This is useful later because p is large. When this last term is small, it means $\hat{f}_n$ is close to the oracle f^*, as per our risk decomposition earlier. The oracle itself is not known since we do not know the unknown f. But theoretically, if we know f^* from the function class $\mathcal{F}$, then we could drive down $\psi_{n,p}$ to as small as possible. So, oracle inequality is a convenient way to evaluate the optimality performance threshold of an estimator $\hat{f}$ against an ideal estimator f^*.

‘

2.7 Best subset selection oracle

The oracle estimator for best subset selection is essentially the same as the least squares estimator but restricted on $K = supp(\beta)$ with a cardinality of non-zero elements of $k = |K|$.

$$\hat{\beta}^{oracle} = (X_K^T X_K)^{-1} X_K^T y$$

We know from least squares that this estimator has empirical risk

$$\frac{1}{n}||X\hat{\beta}^{oracle} - X\beta||_2^2 = \sigma^2\frac{k}{n}$$

We follow the proof of [6], where they showed that by letting $\lambda \asymp \sigma^2 \log p$, the oracle inequality for best subset selection satisifies

$$\frac{1}{n}\mathbb{E}||X\hat{\beta} - X\tilde{\beta}||_2^2 \leq \frac{4\log p}{n} + \frac{2\sigma^2 k}{n} + o(1)$$

We present this result without proof, since it is already proved in the paper.

2.8 Oracle inequality on slow rates Lasso

We now prove the slow rates for Lasso oracle. Here we use $\tilde{\beta}$ denote any other estimator. We follow the setup of [5] in our notation.

$$\frac{1}{n}||y - X\hat{\beta}^{Lasso}||_2^2 \leq \frac{1}{n}||y - X\tilde{\beta}||_2^2. \tag{9}$$

This is true by the optimality condition of the Lasso formulation. We present the vanilla version of the theorem and its proofs.

Theorem 6.

$$\textit{MSE}(X\hat{\beta}^{Lasso}) = \frac{1}{n}||X\hat{\beta}^{Lasso} - X\tilde{\beta}||_2^2 \leq 4\sigma||\tilde{\beta}||_1\sqrt{\frac{2\log(ep/\delta)}{n}}$$

with probability at least $1 - \delta$.

Proof. From the left hand side of 9, we use Holder's inequality to bound the L_2 difference

$$\begin{aligned}
||X\hat{\beta}^{Lasso} - X\tilde{\beta}||_2^2 &\leq 2\epsilon^T X(\hat{\beta}^{Lasso} - \tilde{\beta}) \\
&\leq 2||X^T\epsilon||_\infty||\hat{\beta}^{Lasso}||_1 + 2||X^T\epsilon||_\infty||\tilde{\beta}||_1 \\
&\leq 4||\tilde{\beta}||_1||X^T\epsilon||_\infty
\end{aligned}$$

The max norm $||X^T\epsilon||_\infty = \max_{1\leq i\leq p}|X_j^T\epsilon|$ can also be bounded, since each $|X_j^T\epsilon|$ is Gaussian with mean zero and variance proxy $n\sigma^2$. Now using maximal inequality 8. For any $t > 0$ and $\epsilon \sim SG(\sigma^2)$

$$\max_{1\leq i\leq p}|X_j^T\epsilon| \leq \sigma\sqrt{2nlog(p/\delta)}$$

it follows that with probability at least $1 - \delta$,

$$\mathbb{P}(|X_j^T\epsilon|_\infty \geq t) \leq \sum_{j=1}^{p}\mathbb{P}(|X_j^T\epsilon| > t) \leq 2pe^{-\frac{t^2}{2n\sigma^2}}$$

□

Theorem 7 (Slow rate for the Lasso)**.** *Let $\lambda \geq \sigma\frac{1}{n}||X^T\epsilon||_\infty$, then the following holds for $\tilde{\beta}$*

$$||X\hat{\beta}^{Lasso} - X\tilde{\beta}||_2^2 \leq 4||\tilde{\beta}||_1 n\lambda$$

with probability at least $1 - \delta$.

The proof of can be easily derived from Theorem 6 by letting $\lambda = \sigma\frac{1}{n}||X^T\epsilon||_\infty$.

Lemma 1. *Choosing $\lambda = \sigma\frac{1}{n}||X^T\epsilon||_\infty$ as in Theorem 7, we have*

$$\frac{1}{n}\mathbb{E}||X\hat{\beta}^{Lasso} - X\tilde{\beta}||_2^2 \lesssim ||\hat{\beta}||_1\sigma\sqrt{\frac{\log p}{n}}$$

The lemma implies the following oracle inequality.

Theorem 8 (oracle inequality)**.** *For some estimator $\tilde{\beta}$ such that $||\tilde{\beta}||_1 \leq k$,*

$$\frac{1}{n}\mathbb{E}||X\hat{\beta}^{Lasso} - f(x)||_2^2 \leq \inf_{||\tilde{\beta}||_1 \leq k} \frac{1}{n}||\tilde{\beta}||_2^2 + 4\sigma t\sqrt{\frac{2\log p/\delta}{n}}$$

with probability at least $1 - \delta$.

The slow rate of Lasso is on the of order $\sqrt{(\log p)/n}$. We compare this with the earlier result in best selection subset, which had a rate on the order of $\log p/n$. So, the Lasso rate is much slower.

References

[1] Richard E Bellman. *Adaptive control processes: a guided tour*, volume 2045. Princeton university press, 2015.

[2] Emmanuel J Candes. Modern statistical estimation via oracle inequalities. *Acta numerica*, 15:257, 2006.

[3] David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32, 2000.

[4] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[5] Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical foundations of data science*. CRC press, 2020.

[6] Dean P Foster and Edward I George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975, 1994.

[7] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[8] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

[9] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[10] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[11] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[12] Tong Zhang et al. Multi-stage convex relaxation for feature selection. *Bernoulli*, 19(5B):2277–2293, 2013.

[13] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.