On Statistical Learning Theory, Oracle Inequality, and the Lasso

Colin Cui *

December 31, 2020

Abstract

This paper attempts to highlight theorems and proofs of high-dimensional learning theory and statistics achievements during the past decade. We strive to provide a window of theory and proofs to the outside world in its simplest terms. Hence, the ideas are presented to the audience from an approachable and amicable angle. In light of that, we try to use tools only requiring basic mathematical knowledge. Often times, the difficulty in these proofs is figuring out when to use what for the general audience.

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

1 Introduction

Statistics theory and high-dimensional learning have experienced a wave of transformation. From John Tukey's first EDA textbook in the early 1970s [12] to high-dimensional statistics today, the challenge has evolved due to computational power and the influx of big data [5]. First, statistics problems were concerned with estimations such as MLE, unbiasness, and asymptotic consistency. Now, we are interested in learning a model f from large-scale data that makes accurate predictions. The curse of dimensionality as mentioned in [4] is, however, a giant roadblock on the journey to high-dimension. Some prominent breakthroughs include ridge [9], Lasso [11], elastic net [17], and other regularization methods, which can be achieved through concentration inequalities. These regularization methods served as a workhorse for sparsity recovery of an underlying model. In this paper, we present the ideas with technical details in a colloquial manner that is amicable to the general audience.

2 Towards theory of high-dimensional statistics

2.1 Background

In the high-dimensional setting, the number of available variables p is often much much larger than the number of sample observations n, namely, $p \gg n$. More generally, the number of data points required to sample a high-dimensional space grows exponentially. This is the infamous curse of

*Email: colstat@gmail.com

dimensionality [1] and it is omnipresent. When 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 or very small coefficients, which leads to the notion of *regularization*.

This requires 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.

From a geometric view, the data points live inside a p-dimensional hypersphere unit-ball on high-dimensional space. Let the volume of this hyperball be denoted as vol($\mathcal{B}$). The mass volume tends to 0 as $p \to \infty$. By decreasing the radius of the hyperball by ϵ unit, and resulting in a new radius $(1-\epsilon)$, the ratio of the volumes of the hyperball shrinks by a factor of,

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

Choosing any small ϵ for the last term and hold it fixed. The ratio of the volumes goes towards zero as dimension $p \to \infty$. 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 $O(1/p)$.

2.2 Some notation

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

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

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 denote the least squares estimate as,

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

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 denote 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* depends 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 kind of restriction. 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 penalty term. We define the following popular penalized regressions.

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

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

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

The table above is far from exhaustive of the methods today. But, 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 [16].

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

2.5 Concentration inequalities and tail bounds

Most of the volume in high-dimension is empty and the mass is concentrated in a thin 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 [4]. We will use some of these inequalities in our proofs.

Definition 1 (sub-Gaussian)**.** A random variable Z is sub-Gaussian $SG(\sigma^2)$, if there exists a variance proxy σ^2 such that,

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

By definition, any Gaussian random variable implies sub-Gaussian. The general definition only requires the distribution with exponential tail.

Theorem 1 (Chernoff inequality)**.** *Let Z be a random variable with moment generating around the mean $\mathbb{E}[Z] = 0$. For any $\lambda > 0$, we have*

$$\mathbb{P}(Z \geq t) = \mathbb{P}(e^{tZ} \geq e^{t\lambda}) \leq \frac{\mathbb{E}[e^{tZ}]}{e^{t\lambda}}$$

Proof. The first inequality is trivial. By the Markov inequality, $\mathbb{P}(Z \geq \lambda) \leq \frac{\mathbb{E}[Z]}{\lambda}$ we obtain the second inequality. $\square$

The Chernoff inequality can be used to bound Bernoulli random variables. It provides 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 2 (Hoeffding's inequality)**.** *Suppose that $Z_1, ..Z_n$ are independent random variables. Let $S_n = \sum_{i=1}^n (Z_i - EZ_i)$ such that $X_i \in [a_i, b_i]$,*

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

Proof. By the Chernoff bound, we have

$$m(a) = \frac{\sigma^2 a^2}{2} - at$$
$$\mathbb{P}(S_n \geq t) \leq \min_{\lambda > 0} e^{\lambda t} \mathbb{E}\left[e^{\lambda(S_n)}\right]$$
$$e^{-st}\mathbb{E}\left[e^{ZS_n}\right] = e^{-\lambda t} \prod \mathbb{E}\left[e^{\lambda(Z_i - \mathbb{E}[Z_i])}\right[$$
$$\mathbb{P}(S_n \geq t) \leq e^{2t^2/\sum(b_i - a_i)^2}$$

We can now obtain the two tail bound by doubling the one tail bound. $\square$

Hoeffding's inequality can be applied more generally than Chernoff bound. It is not limited to only Bernoulli random variables.

Proposition 3 (Gaussian tail bound)**.** *Let Z be a Gaussian random variable with mean μ, variance σ^2, then for all $t > 0$*

$$P(Z \geq \mu + t) \leq e^{-\frac{t^2}{2\sigma^2}}$$

Proof. By the moment-generating function,

$$\mathbb{E}[e^{\lambda Z}] = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\infty} e^{\lambda z - z^2/2} dz = \frac{1}{\sqrt{2\pi}} e^{\lambda^2/2} \int_{\infty}^{\infty} e^{-(z-\lambda)^2/2} dx = e^{\lambda^2/2}$$

By the Chernoff inequality,

$$\mathbb{P}[X > t] \leq \mathbb{E}[e^{\lambda X}] e^{\lambda t}$$

let $m(\lambda) = \mathbb{E}[e^{\lambda X}]e^{\lambda t}$, we take $\inf_\lambda m(\lambda)$, we get $\lambda = t$.

$\square$

The proposition provides that majority of the probability mass is concentrated in a small region of the distribution. By the empirical rule 66%, 95%, and 99% of the mass of the Gaussian distribution is concentrated within σ, 2σ, and 3σ, respectively, of the mean.

Theorem 4 (Bernstein's inequality)**.** *Suppose that $Z_1, ..Z_n$ are independent random variables satisfying $|Z_i| \leq a$ and $EZ_i^2 = \sigma^2$. Let $S_n = \sum_{i=1}^n (Z_i - EZ_i)$, then,*

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

The proof of Bernstein is long and not trivial, we refer the reader to Lemma 2.2.11 of [14]. 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 (Sub-Gaussian tail bound)**.** *Let Z be sub-Gaussian with variance proxy σ^2, then for any $t > 0$,*

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq \exp(-\frac{t^2}{2\sigma^2})$$

Proof. Assume $Z \sim SG(\sigma^2)$, then by Markov inequality, it holds for the second inequality that

$$\mathbb{P}(Z > t) \leq \mathbb{P}(\exp(aZ) > \exp(at)) \leq \frac{\mathbb{E}[\exp(aZ)]}{\exp(at)}$$

$$\mathbb{P}(Z > t) \leq \exp(\frac{\sigma^2 a^2}{2} - at)$$

For any $a > 0$, we find the tightest bound by taking minimizing the term in side the exponential.

$$m(a) = \frac{\sigma^2 a^2}{2} - at$$
$$m'(a) = 0$$
$$\inf_a m(a) = -\frac{t^2}{2\sigma^2}$$

□

Theorem 6 (Maximal inequality)**.** *Let $Z_1, ..., Z_n \sim SG(\sigma^2)$ with $\mathbb{E}[Z] = 0$. For any $t > 0$, we have*

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

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

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" was a phrase largely credited to [2]. 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 such 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 [3]. 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 similar to the least squares estimate but restricted on $K = supp(\beta)$ with a cardinality of $k = |K|$ non-zero elements.

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

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 [8], where they showed that by letting $\lambda \asymp \sigma^2 \log p$, the oracle inequality for best subset selection satisfies

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

The proof of this is provided in [8].

2.8 Oracle inequality on slow rates Lasso

We now prove the slow rates for Lasso oracle. Let $\tilde{\beta}$ denote any other estimator than the one of interest here. We adapt a similar notation in the proof as [7].

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

This is true by the optimality condition of the Lasso formulation. There are several ways to prove this and each lead to slightly different results. We present the vanilla version in our theorem and proof.

$$\text{MSE}(X\hat{\beta}^{Lasso}) = \frac{1}{n}||X\hat{\beta}^{Lasso} - X\tilde{\beta}||_2^2 \leq 4\sigma||\tilde{\beta}||_1 \sqrt{\frac{2\log(ep/\delta)}{n}} \tag{10}$$

with probability at least $1 - \delta$.

Proof. From the middle term of 10, 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 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. *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 can be easily derived from Theorem **??** 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. *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] George EP Box. All models are wrong, but some are useful. *Robustness in Statistics*, 202:549, 1979.

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

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

[5] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

[6] 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.

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

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

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

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

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

[12] John W Tukey. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

[13] 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.

[14] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

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

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

[17] 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.