Notes to High Dimensional Statistics([Rigollet])

Hengrui Luo

Contents

Preface

> Il est facile de voir que...[1]- from *Traité de mécanique céleste* by Pierre-Simon Laplace.

Note: I did not write out every direct citation from [Rigollet] since this is already a note to it. Please check before you use any sentence from this note.

Although this is at best a faithful record of our discussions, I should confess that I do not have enough time to care for every detail so there are still many errors and obscure points in these notes. Please let me know if there is anything wrong.

This is a discussion note prepared in 2015 winter.

The primary material is [**Rigollet**] and the secondary material is [**Cressie**]. The primary material is determined by J.Zhang while the secondary one is chosen by me.

The notes and errata are composed by me, all literal errors are mine.

The *Exercise, Discussions and Examples* in notes are invented by me, I personally added some mathematical flavor as I always did. All errors are mine.

Jianhao Zhang provided a lot of help and inspiring lectures.

I prepared and leaded the discussion of Chap.1,3,4.1-4.3 and some of Chap.5 while J.Zhang completed the rest of [**Rigollet**]. We had a good time and both learnt a lot from each other, though it is a pity that the time is rather limited.

Thanks to Prof.P.Rigollet for writing such a series of high-quality notes and his explanations on some points.

Thanks to Prof.Y.Zhu for pointing out some valuable resources to us and kind help on some difficult points.

H.R.Law

[1]Translation from Wikiquote: It is therefore obvious that...

ERRATA and Complements

- p2: the last line $\|h\|_2 L$ should be $\|h\|_2^2$.
- p9: the example of smooth function is k-sparse in the trigonometric basis space seems not correct. It is approximately sparse since trigonometric basis is a Schauder basis.
- p16: Def 1.2 the random variable should not be a real number, kind of misleading notation.
- p19: $\leq e^{4\sigma^2 s^2}$ is not sharp and the proof should be done more carefully by the reader.
- p20: line 20 the "randdom" should be "random".
- p22: line 12 the "week" should be "weak".
- p23: line 4, I do not think that Taylor expansion for a measurable function can be used arbitrarily here. A more rigid way of talking is to define $exp(X) = 1 + X + \frac{X^2}{2} + \cdots + \frac{X^n}{n} + \cdots$ as a series directly. The problem is that Taylor expansion of a measurable function is not well-defined in some cases. Here we can argue by using Riemannian upper sum and dominate convergence theorem along with the Egorov Theorem to give the same inequality.
- p24: the last line should be $\mathbb{P}\left(\overline{X} > t\right) \leq \left\{\prod_{i=1}^n \mathbb{E}\left[e^{sX_i}\right]\right\} \cdot e^{-snt}$
- p27: line 8 the "compact" should be "convex".
- p36-37: $sup_{u\in\mathcal{B}_2}(\tilde{\epsilon}'u)^2$ are all the same as $max_{u\in\mathcal{B}_2}(\tilde{\epsilon}'u)^2$, no need to introduce supremum anyway.
- p39: line 13, the formula misses proxy variance, to be consistent with Theorem 2.4, it should be $\mathbb{E}\left[\text{MSE}(\mathbb{X}\theta^{\hat{L}^S}_{\mathcal{B}_1})\right] \lesssim min\left(\sigma^2 \cdot \frac{r}{n}, \sigma \cdot \sqrt{\frac{logd}{n}}\right)$.
- p40: the formula in Theorem 2.6 should be $\text{MSE}(\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)}) \lesssim \frac{\sigma^2}{n} log\begin{pmatrix} d \\ 2k \end{pmatrix} + \frac{\sigma^2}{n} \cdot k \cdot log(6) - \frac{\sigma^2}{n} log\delta$. log(6) is missing.
- p42: the proof of Corollary 2.9 is a bit messy, should be: $\mathbb{E}\left[\text{MSE}(\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)})\right]$(The risk is evaluated on the sample space of response

variable.)

$$
\begin{aligned}
&= \tfrac{1}{n}\mathbb{E}\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2\right] \\
&= \tfrac{1}{n}\int_{\mathcal{Y}}\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 d\mathbb{P}_Y \\
&= \tfrac{1}{n}\int_0^1 \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 \le u\right] du \qquad (Abelian\ summation\ technique) \\
&= \tfrac{1}{n}\int_0^\infty \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > u\right] du \qquad (change\ of\ variable : v := \tfrac{u}{n}) \\
&= \int_0^\infty \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > nv\right] dv \\
&= \int_0^H \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > nv\right] dv \qquad + \int_H^\infty \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > nv\right] dv \\
&\quad \int_0^H \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > nv\right] dv \le H\ is\ clear \\
&\le H + \int_0^\infty \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > n(v+H)\right] dv \qquad (2.6) \\
&= H + \textstyle\sum_{j=1}^{2k}\binom{d}{j} 6^{2k}\int_0^\infty e^{-\frac{n(v+H)}{32\sigma^2}} dv = \qquad H + \sum_{j=1}^{2k}\binom{d}{j} 6^{2k}\left[\tfrac{32\sigma^2}{n} e^{-\frac{nH}{32\sigma^2}}\right]
\end{aligned}
$$

- p43: line 5 from the bottom, we have not yet define sub-gaussian sequence model. So the (2.7) is not a sub-gaussian but a gaussian model.
- p55: Theorem 2.18, we must chose $2\tau = 8\sigma\sqrt{\frac{log(2d)}{n}} + 8\sigma\sqrt{\frac{log(\frac{1}{\delta})}{n}}$, the original regularization parameter is still correct (via Jensen inequality) BUT when we try to prove Theorem 3.5 we must correct it into $2\tau = 8\sigma\sqrt{\frac{log(2d)}{n} + \frac{log(\frac{1}{\delta})}{n}}$ to avoid a cross-term issue.
- p56: line 5, the correct expression should be $|\mathbb{X}_j|_2^2 \le n \cdot \left[1 + \frac{1}{14k}\right] \le 2n$. (Pointed out by J.Zhang)
- p62: line 3 from the bottom, "spam" should be "span".
- p64: line 1, "valif" should be "valid"
- p72: The definition of Sobolev ellipsoid is a bit scattered. I adopt the following definition:
  The *Sobolev ellipsoid* $\Theta(\beta, Q) := \left\{\theta \in L^2(\mathbb{N}) : \sum_{j=1}^\infty a_j^2\theta_j^2 \le Q, a_j \sim (\pi j)^{2\beta}\ as\ j \to \infty\right\}$ the Q again specify the radius of sequence norm.
- p88: line 10, $|\lambda_j|^2 \le 9min(\tau^2, |\lambda_j|^2)$ is the correct bound, $|\lambda_j|^2 \le 3min(\tau^2, |\lambda_j|^2)$ is not correct unless $\sqrt{3}\tau$ is the critical point.
- p90: The last line, $\overline{Y}$ should be $\overline{\mathbb{Y}}$.
- p92: line 14, the transpose is lost from here one, the correct one should be $x^{\mathrm{T}}(\hat{\Sigma} - I_d)y = \frac{1}{n}\left\{(X_i^{\mathrm{T}}x)^{\mathrm{T}}(X_i^{\mathrm{T}}y) - \mathbb{E}(X_i^{\mathrm{T}}x)(X_i^{\mathrm{T}}y)\right\}$ and the rest of proof should follow this notation.
- p102-103: (5.2), $\hat{\theta_n}$ is the same as $\hat{\theta}$ in (5.4), in order to be consistent. (Pointed out by J.Zhang)
- p106: line 8, It should be "...is minimized for $R, R^\star \in \Omega$(The associated sigma field) such that $\nu(R = R^\star) = 1$."
- p110: Theorem 5.10(Fano's inequality), the result is not valid for $M = 2$ since $log(M-1) = -\infty$ when $M = 2$.
- p112: Theorem 5.11, ϕ should be the decaying rate $\phi(\Theta)$.
- p115: Theorem 5.14, C_1, C_2 are irrelevant in this result.

- p117: line 4, this line is redundant.
- p118: line 17, It should be $R\min\left(\frac{R}{8}, \beta^2\sigma\frac{log(ed/\sqrt{n})}{8n}\right)$, there is a redundant bracket and the sigma square is not correct. (Pointed out by J.Zhang)
- p118: Corollary 5.16, this corollary is a bit messy, first $\phi(\mathcal{B}_0(R))$ should be $\phi(\mathcal{B}_1(R))$, be careful since k represent subscripts and the bound for dimension. And the constant C should be described separately.

Preliminaries

0.1. Regression Analysis and Prediction Risk

(1) Feature-label model.
   (a) How is it defined? response/covariates
   (b) How is it related to ANCOVA?
   (c) What is the regression function? $f(x) = \mathbb{E}[Y|X = x], x \in \mathcal{X}$. The regression function gives a simple summary of this conditional distribution.
   (d) What is BLUP? Review of **[Dean&Voss]**.

(2) Best prediction and prediction risk.
   (a) The measurablility in this note is the Borel-Lebesgue measurability.
   (b) What is the loss function of the model? $|Y - g(X)|^2$
   (c) What is a small random variable? $\mathbb{E}\left[Z^2\right] = [\mathbb{E}Z]^2 + Var[Z] \to 0$
   (d) What is the risk function of the model? $\int_{\mathcal{X}} |Y - g(X)|^2 dX = \mathbb{E}[Y - g(X)]^2$, L^2risk.

   COROLLARY 1. $\mathbb{E}[Y - g(X)]^2 \geq \mathbb{E}[Y - \mathbb{E}[Y|X = x]]^2$ *the equality holds for all real-valued* $\mathcal{X}$*-measurable function* $g(X)$ *iff* $g(X) \overset{a.s.}{=} \mathbb{E}[Y|X = x]$.

      (i) In fact it also holds for any complete space over algebraically closed field.
      (ii) It can be proven that this also holds for any random variable sequence convergent to a $\mathcal{X}$-measurable function a.s.

   DEFINITION 2. The regression function based on given data has the best prediction property. $\mathbb{E}[Y - \mathbb{E}[Y|X = x]]^2 = inf_g \mathbb{E}[Y - g(X)]^2$

(3) Prediction and estimation
   (a) What is the statistical problem? Do not have access to the conditional distribution.
      (i) This is a point which I disagree with, in classical way we can always approximate it, and so can we now.
      (ii) Basically, the author is a practitioner whose focus is that we cannot make valid use of a conditional distribution when the sample is complicated structured.
   (b) What is the overall GOAL? The goal of regression function estimation is to use the data to construct an estimator $\hat{f} : \mathcal{X} \to \mathcal{Y}$ that minimizes the risk function.

   LEMMA 3. $R(\hat{f}_n) := \mathbb{E}\left[Y - \hat{f}_n\right]^2 = \mathbb{E}[Y - f]^2 + \mathbb{E}\left[f - \hat{f}_n\right]^2$, *where* $f := \mathbb{E}[Y|X = x]$

(c) It is equivalent to consider $\mathbb{E}\left[f-\hat{f}_n\right]^2$ because we want to get consistent estimators.
(d) Two kinds of bound on accuracy
  (i) The moment bound. Discussion: Why Edgeworth bound is not good enough?
  (ii) The concentration bound(PAC bound). Discussion: Spectrum of a manifold.
  (iii) Often the concentration bound follows from the moment bound. See the lemma below. Example: Hoeffding inequality.

LEMMA 4. *(Hoeffding inequality)* $\mathbb{P}\left[\left\|\hat{f}_n-f\right\|_2^2-\mathbb{E}\left\|\hat{f}_n-f\right\|_2^2>t\right]\sim O(e^{-n}), n\to\infty$.

(e) Proof of Hoeffding inequality locally on a Riemannian manifold will give us a bound of maximum length of geodesics.
(f) Why L^2-risk? Reflexive Hilbert space. Allows the Lemma 3 to hold.
(g) What are other ways of measuring error?
  (i) Bounding a distance(pseudo-distance).
  (ii) Bounding the risk function(L^2-risk here).
  (iii) Some examples. Do the Sup-norm because it directly relates to the maximum complexity.
  (iv) I found [**Rigollet**] is talking about something unclear here, by "global" and "local" he just mean the global or local behavior of the regression function/estimator while he use the same terms for a distribution later. This should be clarified.

DEFINITION 5. The classification risk of a classifier(regression function with discrete range) $h:\mathcal{X}\to\{0,1\}$ is defined by $R(h)=\mathbb{P}\left(Y\neq h(X)\right)$

(h) The sample spaces are specified to be affine spaces in order to yield Lebesgue measure invariant in "most" cases. Note the Lebesgue measure is not invariant in some non affine spaces. Some works has been done on real algebraic sets.

0.2. Models and Methods

(1) Empirical risk minimization
  (a) Empirical risk minimization(ERM) is the leading criterion of how to construct a **linear** estimator $\hat{f}_n$. We should also known that in high dimensional cases this could be surpassed by the so-called James-Stein estimator. Another comment is that we actually saw this before as "moment replacement method".
  (b) Ridge regression and penalty review. Regularization techniques. LASSO preview(LARS realization).
  (c) ERM and LSE coincides on linear models. C.R.Rao's comment on higher moment estimation. Incomplete cases.
  (d) Identifiability. I cannot agree with the author here about mis-specified models, they can sometimes have pretty good risk if the sampling

points are not random enough. Example: samples coming from a single source.

DEFINITION 6. The oracle inequality $\mathbb{E}\left[\left\|\hat{f}_n - f\right\|_2^2\right] \leq \|\overline{f} - f\|_2^2 + \phi_n, where \left\|\hat{f}_n - \overline{f}\right\|_2^2 \leq \phi_n$. The empirical term is bounded by a sequence of positive numbers.

(e) The "oracle" comes from the semi-supervised learning, where human intervention sometimes gives extra correct pair of feature-label. As the [**Rigollet**] said "It cannot be constructed without the knowledge of the unknown f." Example: Binary tree guessing problem.

(2) High dimension and sparsity
   (a) What do you mean by "high-dimension"? "By high dimension, we informally mean that the model has more 'parameters' than there are observations. The word 'parameter' is used here loosely and a more accurate description is perhaps *degrees of freedom*." For definition of degree of freedom, see [**Kendall1**].

   DEFINITION 7. A vector $\theta \in \mathbb{R}^d$ is said to be k-sparse for some integer $k \leq d$ is it has at most k non-zero coordinates. We denote by $|\theta|_0 = \sum_{j=1}^d 1_{\theta_j \neq 0}$ the number of nonzero coordinates of θ is also known as sparsity.

   (b) An example of approximately sparse: $\left(e^{-1}, \cdots, e^{-n}, \cdots\right)$. Now it is easy to see why we need Sobolev norm when we discuss the sparsity of functions, which is somehow parallel to this definition.
   (c) What are the differences between sparsity and dense set? Actually the notion of sparsity of a vector of an affine space is different from a set being sparse/dense in a topological space. All collection of countable vectors are sparse set in the regular affine space while they might not be k-sparse for any integer k. We should be very careful because the [**Rigollet**] gave a subtle example of a smooth function being sparse in a functional space where the Schauder basis is the trigonometric series. This example is **incorrect** if the author is talking about $|\theta|_0$ since $|\theta|_0$ of any smooth function could be infinite. In fact a smooth function is approximately sparse in such a space.
   (d) Why sparsity is important? Sparsity is important because "if we knew that θ^* belongs to one of these subspaces, we could simply drop irrelevant coordinates and obtain an oracle inequality such as..."

(3) Nonparametric regression
   (a) What do you mean by nonparametric? The parameter to estimate is infinite dimensional. In nonparametric course, we have the same meaning of this term when doing nonparametric inference based on ranks of the data. This is the same case because (in some cases) the parameter can be identified to an infinite sequence of real numbers (using Schauder basis or approximation, so the case here is countably infinity where ranks still apply).
   (b) The general idea is to impose some restraints on the coefficients of the function under a certain basis, which is equivalent to restricting

the functions under consideration to a certain class. Example: The class of functions whose zeroth Fourier coefficients are zero. The class of functions whose 0-residual is zero.

(c) Following the general idea we require some smoothness on the class of functions under consideration.[2] As mentioned, the main difference is that we make assumptions on the regression function in order to control the approximation error. [3]

DEFINITION 8. The stochastic term of a Fourier series cutoff at k_0 with the constraint $|\alpha_k| \leq C |k|^{-\gamma}, \gamma \geq \frac{1}{2}$ is $\mathbb{E} \sum_{|k| \leq k_0} (\hat{\alpha_k} - \alpha_k)^2$.

(4) Matrix models
   (a) Review of random matrix: whose entries are random variables.
   (b) The key problem is that *sparse* matrix does not necessarily lead to *sparse* eigenspace. The first sparse mean many zero coefficients while the second sparse means the k-sparsity in a larger affine space. Example: Netfilx data.
   (c) What is high-dimensional estimation of covariance? Example: [**Dean&Voss**]'s random effect chapter.

(5) Optimality and minimax lower bounds

[2]Actually the majority of [**Rigollet**] follows a sub-harmonic framework implicitly.

[3][**Rigollet**] pp.12: Instead, sparsity or approximate sparsity is a much weaker notion than the decay of coefficients $\{\alpha_k\}$ presented above. In a way, sparsity only impose that after ordering the coefficients present a certain decay, whereas in nonparametric statistics, the order is set ahead of time : we assume we have found a basis that is ordered in such a way that coefficients decay at a certain rate. Example: umbrella hypothesis in nonparametric inference.
This is of course the case for otherwise we can prove dynamics of a manifold in a much easier manner...

CHAPTER 1

Gaussian and Sub-gaussian Models

1.1. Gaussian tails and MGF

(1) Gaussian tail and Markov tail
The Markov inequality gives us a bound on the tail probability of a general random variable regardless the underlying distribution, however, if we know the underlying distribution is Gaussian(OR by LLN we deduce this fact), then the bound can be improved:

PROPOSITION 9. *$X \sim \mathcal{N}(\mu, \sigma^2)$ then $\mathbb{P}(|X-\mu| > t) \leq \frac{1}{\sqrt{2\pi}} \cdot \frac{e^{-\frac{t^2}{2\sigma^2}}}{t}$, this is called the Gaussian tail OR Gaussian bound which gives the bound of tail probability of a Gaussian random variables.*

(2) Law of succession[1]This paper gave a method of doing modification of bound inbetween the Markov bound and Gaussian bound. Also, we can use a continuous distribution to yield a bound for discrete distributions, tough that will usually cause some overly liberal inference.
(3) MGF for Gaussian random variables

1.2. Sub-gaussian random variables and Chernoff bounds

(1) What is a sub-gaussian random variable?

DEFINITION 10. A random variable $X : \mathcal{X} \to \mathbb{R}$ is said to be sub-gaussian with variance proxy σ^2 if $\mathbb{E}X$=0 and its MGF satisfies $M_X(s) \leq exp\left(\frac{\sigma^2 s^2}{2}\right), \forall s \in \mathbb{R}$. We denote it by $X \sim \text{subG}(\text{œ}^2)$. This notation denotes a family of distributions rather than a single distributions.

(2) Is there hyper-gaussian random variable? Can you define it parallel to hyper-harmonic functions?
A comment here is that the framework of sub-gaussian random variables is exactly parallel to sub-harmonic analysis.
(3) What is its generalization into higher dimension? Sub-gaussian vectors. Any of simplex linear combinations of its components is a one-dimensional sub-gaussian random variable. Sub-gaussian matrices. Any of unit vectors in its row-column space is a one-dimensional sub-gaussian random variable. Actually we can define a sub-gaussian inner product space as we did for harmonic space.
Attention to the definition, the linear combination can be a simplex one

[1][Edwin B. Wilson]Probable Inference, the Law of Succession, and Statistical Inference, Journal of the American Statistical Association Vol. 22, No. 158 (Jun., 1927), pp. 209-212

or not, the difference is reflected on the scaling of the variance proxy.
Exercise: Use g-inverse to rewrite the definition of sub-gaussian matrix.

(4) Is there an equivalent definition of sub-gaussian random variables? Chernoff bound obtained by optimizing the quadratic form in exponential.

LEMMA 11. *$X \sim \text{subG}(\sigma^2)$, then for any $t > 0$, $\mathbb{P}(|X - 0| > t) \leq exp\left(-\frac{t^2}{2\sigma^2}\right)$*

(5) The **Chernoff bound** translates the bound on MGF to the bound on tails. The following gives the converse result:
Exercise: Try to prove the following result using Stein's lemma of calculating the gaussian moments. See Chap.3 of [**Casella&Berger**].

LEMMA 12. *X is a random variable such that $\mathbb{P}(|X - 0| > t) \leq exp\left(-\frac{t^2}{2\sigma^2}\right)$, then for any positive integer $\mathbb{E}|X|^k \leq (2\sigma^2)^{\frac{k}{2}} k\Gamma(\frac{k}{2})$*

COROLLARY 13. *$\left[\mathbb{E}|X|^k\right]^{\frac{1}{k}} \leq (\sigma e^{\frac{1}{e}})\sqrt{k}, k \geq 2$ and $\left[\mathbb{E}|X|^k\right]^{\frac{1}{k}} \leq \sigma\sqrt{2\pi}, k = 1$. So there is a natural embedding of such random variables described in the above lemma into the $L^p, p \geq 1$ spaces. This is exactly parallel to harmonic analysis.*

LEMMA 14. *X is a random variable such that $\mathbb{P}(|X - 0| > t) \leq exp\left(-\frac{t^2}{2\sigma^2}\right)$, then $M_X(s) \leq e^{4\sigma^2 s^2} \iff X \sim \text{subG}(8\sigma^2)$*

DEFINITION 15. (Bridge definition between the harmonic analysis and the spatial statistics) $X \in \mathcal{X}^* \sim \text{subG}(\sigma^2) \Longrightarrow$its MGF is a **rapidly-decay**[2] (Lebesgue-measurable) function on the space $L^p_{\mathbb{P}}(\mathcal{X})$.

(6) Using the property of MGFs of the independent sums:

LEMMA 16. *$\mathbb{P}\left(\left|a^T X - 0\right| > t\right) \leq exp\left(-\frac{t^2}{2\sigma^2 \cdot a^T a}\right)$*

(7) Hoeffding's Lemma: Random variables that are bounded uniformly are sub-gaussian random variables. [3]. Exercise: Try to use Definition 15 to make a direct insight into this fact.
In Hoeffding's lemma, compared with the Markov's inequality, the identically distributed assumptions is dropped and a weaker condition of boundedly supported is added.

LEMMA 17. *(Hoeffding's 1963) Let $X_i, i = 1, \cdots n$ be independent random variables that are a.s. support on $[a_i, b_i]$. Then we have $\mathbb{P}\left(\left|\overline{X} - \mathbb{E}\overline{X}\right| > t\right) \leq exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$. Using Definition 15, we know that such a sequence of random variables, if converges, converges to a sub-gaussian random variables.*

[2]The rapidly decreasing functions are those $f(x) \sim o(x^{-N}), \forall N, x \to \infty$
[3]With a variance proxy that depends on the size of their support.

1.3. Sub-exponential random variables

(1) Can we make the definition 15 a n.s. one? Yes, and that is the reason of introducing sub-exp random variables. Sub-exponential random variables are a larger family than the sub-gaussian family.

LEMMA 18. *Let X be a symmetrically distributed random variable such that $\mathbb{P}\left(\left|\overline{X}-0\right|>t\right)\leq exp\left(-\frac{2t}{\lambda}\right),\exists\lambda>0$. Then for any positive integer $k\geq 1,\mathbb{E}\left|X\right|^{k}\leq\lambda^{k}k!$. So there is a natural embedding of such random variables described in the above lemma into the $L^{p},p\geq 1$ spaces.*[4]

COROLLARY 19. $\mathbb{E}\left[e^{sX}\right]\leq e^{2s^{2}\lambda^{2}},\forall\left|s\right|\leq\frac{1}{2\lambda}$

(2) What is a sub-exponential random variable?

DEFINITION 20. A random variable $X:\mathcal{X}\to\mathbb{R}$ is said to be sub-exponential with parameter λ if $\mathbb{E}X$=0 and its MGF satisfies $M_{X}(s)\leq exp\left(\frac{\lambda^{2}s^{2}}{2}\right),\forall s\in\left[-\frac{1}{\lambda},\frac{1}{\lambda}\right]$. We denote it by $X\sim\text{subE}(\breve{})$. This notation denotes a family of distributions rather than a single distributions.

LEMMA 21. *Let $X\sim\text{subG}(\sigma^{2})$ then $Z=X^{2}-\mathbb{E}X^{2}\sim\text{subE}(16\sigma^{2})$.*

(3) Bernstein's inequality. Since the sub-exponential family is a larger family, the inequalities for sub-exponential families are larger deviations controlled by a weaker bound.

THEOREM 22. *(Bernstein's 1930s) Let $X_{i},i=1,\cdots n$ be independent random variables that are $X_{i}\sim\text{subE}(\lambda)$. Then we have $\mathbb{P}\left(\left|\overline{X}-0\right|>t\right)\leq exp\left(-\frac{n}{2}\left(\frac{t^{2}}{\lambda}\wedge\frac{t}{\lambda}\right)\right)$.*

(4) Comparison for the inequalities in this chapter:

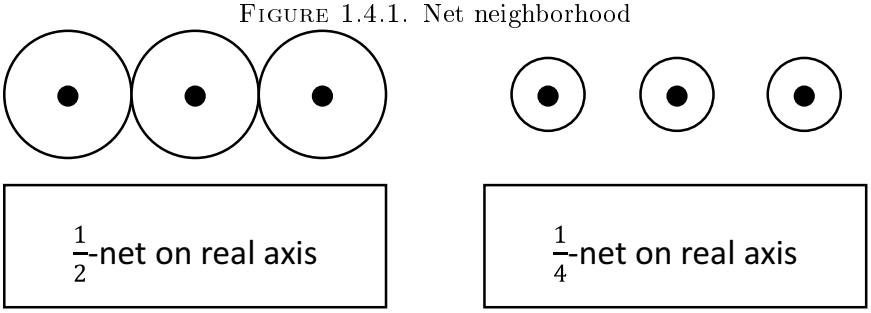
Name	Assumptions	Expressions
Gaussian-tail	$X\sim\mathcal{N}(\mu,\sigma^{2})$	$\mathbb{P}\left(\left\|X-\mu\right\|>t\right)\leq\frac{1}{\sqrt{2\pi}}\cdot\frac{e^{-\frac{t^{2}}{2\sigma^{2}}}}{t}$
Markov	X_i independent, identically distributed	$\mathbb{P}\left(\left\|X-\mu\right\|>t\right)\leq\frac{Var(X)}{t}$
Hoeffding	X_i independent	$\mathbb{P}\left(\left\|\overline{X}-\mathbb{E}\overline{X}\right\|>t\right)\leq exp\left(-\frac{2n^{2}t^{2}}{\sum_{i=1}^{n}(b_{i}-a_{i})^{2}}\right)$
Chernoff	$X_i\sim\text{subG}(\text{œ}^2)$	$\mathbb{P}\left(\left\|X-0\right\|>t\right)\leq exp\left(-\frac{t^{2}}{2\sigma^{2}}\right)$
Bernstein	$X_i\sim\text{subE}(\lambda)$	$\mathbb{P}\left(\left\|\overline{X}-0\right\|>t\right)\leq exp\left(-\frac{n}{2}\left(\frac{t^{2}}{\lambda}\wedge\frac{t}{\lambda}\right)\right)$
Maximal*	$X_i\sim\text{subG}(\text{œ}^2)$	$\mathbb{P}\left[max_{1\leq i\leq N}X_{i}\right]\leq Ne^{-\frac{t^{2}}{2\sigma^{2}}}$

1.4. Maximal inequalities

(1) The maximal inequality: Exercise: Use this result to generalized the Tukey HSD for general contrasts.

[4]Use Stirling's formula to yield a strict bound in the next corollary.

FIGURE 1.4.1. Net neighborhood



THEOREM 23. *Let $X_i, i = 1 \cdots, N$ be random variables such that $X_i \sim \text{subG}(\sigma^2)$ then*
$\mathbb{E}\left[max_{1\leq i\leq N} X_i\right] \leq \sigma\sqrt{2log(N)}$,$\mathbb{E}\left[max_{1\leq i\leq N} |X_i|\right] \leq \sigma\sqrt{2log(2N)}$
$\mathbb{P}\left[max_{1\leq i\leq N} X_i > t\right] \leq Ne^{-\frac{t^2}{2\sigma^2}}$,$\mathbb{P}\left[max_{1\leq i\leq N} |X_i| > t\right] \leq 2Ne^{-\frac{t^2}{2\sigma^2}}$

(2) Discussion: Can we generalize this result to supremum for a sequence of sub-gaussians? Why or why not?
Exercise: Generalize this result for linear transformed sub-gaussian vectors. Compared with the following results.
The comment [**Rigollet**] p26: "Extending these results to a maximum over an infinite set may be impossible. For example, if one is given an infinite sequence of i.i.d $X_i \sim \mathcal{N}(0, \sigma^2)$, then for any $N \geq 1$, we have for any $t > 0$, $\mathbb{P}\left[max_{1\leq i\leq N} X_i < t\right] = \mathbb{P}\left[X_1 < t\right]^N \to 0, N \to \infty$."
(3) Discussion: The correlation structure of a martingale.
(4) Maximum over a convex polytope.
   (a) What is a convex polytope? A convex set with a finite number of extreme points. Simplex is a convex polytope.
   (b) What is so special about convex polytope? The extreme values are reached and can only be reached at extreme points. This allows us to proceed like the convex optimization in most cases. Discussion: Kuhn's simplex approximation algorithm. [5]
   (c) The general result is no difference from standard convex analysis texts. The convex combination will not change the maximum.
(5) Maximum over a L^2 ball.
   (a) What is a sphere? What is a ball? What is a normed ball? What is a partition of unity?
   (b) What is a net neighborhood family?
   If (M, d) is a metric space, and $X \subset M$, then the *packing radius* of X is $\frac{1}{2}inf_{x\neq y} d(x, y)$.
   If the packing radius is r, then open balls of radius r centered at the points of X will all be disjoint from each other.
   The *covering radius* of X is the $inf_{\forall x\in X, \#\{d(x,y)\leq r, x\neq y\}\geq 1} r$.
   It is the smallest radius such that closed balls of that radius centered at the points of X have all of M as their union.

[5][H.W. Kuhn]Simplicial approximation of fixed points. Proc. Natl. Acad. Sci. 61, 12381242 (1968)

An ϵ-*packing* is a set X of packing radius $\geq \frac{\epsilon}{2}$
An ϵ-*covering* is a set X of covering radius $\leq \epsilon$
An ϵ-net is a set that is both an ϵ-*packing* and an ϵ-*covering*.
Exercise: Every compact set has a finite ϵ-net for any $\epsilon > 0$.

LEMMA 24. *The unit ball in $L^2(\mathbb{R}^d)$ has a ϵ-net w.r.t the Euclidean distance. This ϵ-net has cardinality $\leq (\frac{3}{\epsilon})^d$.*

(6) The maximum theorem for a normed ball.

CHAPTER 2

Linear Regression Model

2.1. Fixed design linear regression

(1) What is a random/fixed design linear regression? The usual fixed design with a sub-gaussian error/noise term. $Y_i = f(X_i)+\epsilon_i, i = 1, \cdots, n; f(X) = X'\theta*$. Depending on the nature of the design points being random variables or constants, we will favor a different measure of risk.
   (a) Fixed design
   $Y = \mathbb{X}'\theta * +\epsilon$, where X is simply a matrix with constant terms and ϵ is a random vector.
   (b) Random design
   $Y = \mathbb{X}'\theta * +\epsilon$, where X is a matrix with random variable terms and ϵ is a random vector.

(2) What is the difference between random couples and random 2-vectors? Whether there is a joint distribution.

(3) What is the GOAL? Given data $(X_1, Y_1), \cdots, (X_n, Y_n)$ we construct a "best" predictor $\hat{f}(X_{n+1})$ for Y_{n+1}. Example: log of tumor-volume.

(4) What is the risk function for random design? $R(\hat{f}_n) = \mathbb{E}\left[Y_{n+1} - \hat{f}_n(X_{n+1})\right]^2 =$ $\mathbb{E}\left[Y_{n+1} - f(X_{n+1})\right]^2 + \left\|\hat{f}_n - f\right\|^2_{L^2(P_X)}$ where the latter measurement is taken on the marginal probability measure of X_{n+1}. Discussion: Why choose this marginal measure? Because we want to predict X_{n+1}

COROLLARY 25. *For a random design,* $R(\hat{f}_n) = \sigma^2 + \left\|\hat{f}_n - f\right\|^2_{L^2(P_X)}$ *where the* σ^2 *is the variance proxy of the sub-gaussian error.*

(5) What is the risk function for fixed design? $\sigma^2 = 0$. Compare with the fixed and random models in [**Dean&Voss**]. There is no marginal measure for X_{n+1} here.

(6) What is a denoising problem? Construct a "best" predictor in a fixed design. Example: Regular design is to interpolation between points under smoothness assumptions.

(7) What is the risk function for fixed design? MSE. Exercise: Try to figure out how MSE should be in case of ANCOVA.

DEFINITION 26. The mean square error is $\text{MSE}(\hat{f}_n) := \frac{1}{n}\sum_{i=1}^n \left[\hat{f}_n(x_i) - f(x_i)\right]^2$, using the design matrix notation in linear theory we can write $\text{MSE}(\hat{f}_n = \mathbb{X}\hat{\theta}) := \frac{1}{n}\left|\mathbb{X}(\hat{\theta} - \theta*)\right|_2^2$.

2.2. LSE

(1) What is LSE? The minimum L^2-norm solution $\hat{\theta}$ to the consistent equation given by $\mathbb{X}\theta = Y$. $\theta^{\hat{LS}} \in argmin_{\theta\in\mathbb{R}^d}|Y-\mathbb{X}\theta|_2^2$. Exercise: Review 3 basic types of g-inverses.

PROPOSITION 27. *The LSE $\mu^{\hat{LS}} := \mathbb{X}\theta^{\hat{LS}} \in \mathbb{R}^n$ satisfies $\mathbb{X}'\mu^{\hat{LS}} = \mathbb{X}'Y$, and an explicit form (however this is not the unique explicit form) of this g-inverse is* [1] $\theta^{\hat{LS}} = (\mathbb{X}'\mathbb{X})^- Y$.

THEOREM 28. *(MSE for LSE) For a fixed design $Y = \mathbb{X}'\theta* + \epsilon, \epsilon \sim$ subG(σ^2), then the LSE $\theta^{\hat{LS}} = (\mathbb{X}'\mathbb{X})^- Y$ satisfies*

$\mathbb{E}\left[\text{MSE}(\mathbb{X}\theta^{\hat{LS}})\right] \lesssim \sigma^2 \cdot \frac{r}{n} \iff \mathbb{E}\left[\text{MSE}(\mathbb{X}\theta^{\hat{LS}})\right] \leq C\sigma^2 \cdot \frac{r}{n}, \exists C > 0 \iff \mathbb{E}\left[\text{MSE}(\mathbb{X}\theta^{\hat{LS}})\right] \sim O(\sigma^2 \cdot \frac{r}{n})$

where $r = rank(\mathbb{X}'\mathbb{X})$. Moreover, for any $\delta > 0$ with probability $1-\delta$it holds that

$\text{MSE}(\mathbb{X}\theta^{\hat{LS}}) \lesssim \sigma^2 \cdot \frac{r-log(\delta)}{n}$ *by using the Chernoff bound.*

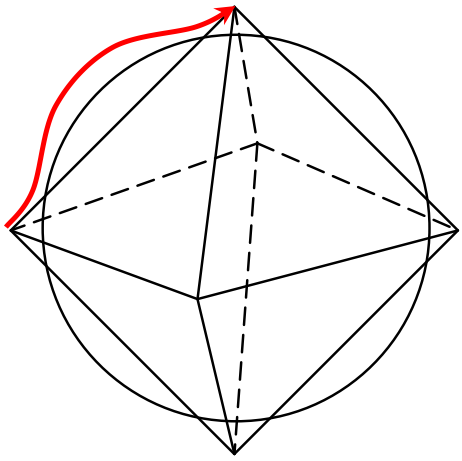
COROLLARY 29. *If $d \leq n$ and $\left|\theta^{\hat{LS}} - \theta*\right|_2^2 \leq \frac{\text{MSE}(\mathbb{X}\theta^{\hat{LS}})}{\lambda_{min}(B)}, B := \frac{\mathbb{X}'\mathbb{X}}{n}$ with rank d.*

(a) Discussion: Is the $sup_{u\in\mathcal{B}_2}(\tilde{\epsilon}'u)^2$ the same as $max_{u\in\mathcal{B}_2}(\tilde{\epsilon}'u)^2$? Why? I think they are the same here since a normed space (L^2 space here) is complete iff its unit ball is. So there is no need to implement supremum here, also for consistency with maximal inequalities used in following proof.
(b) Discussion: Is this bound sharp? No, cause the proof of Theorem 1.19 does not give a sharp bound.
(c) Exercise: Try to improve this bound using Rayleigh Theorem.
(d) Exercise: Try to calculate an example when $d = n$.
(e) It is interesting that the author regard the bounding as a technique to remove the dependency, which is not a usual point of view. pp.37. The technique here has one step redundant,which is to introduce the intermediate basis Φ.

(2) Why we should make constraint on LSE? Constraints imposed on the domain is helpful to reduce the optimization(solving of consistent system) problem. Review maximal inequalities. Example: REML.
(3) Why we need symmetry? Regular setup in functional analysis.
(4) What is a RLSE on a convex symmetric set? $\theta_K^{\hat{LS}} \in argmin_{\theta\in K\subset\mathbb{R}^d}|Y-\mathbb{X}\theta|_2^2 \overset{symmetry}{\Rightarrow} \left|\mathbb{X}\theta^{\hat{LS}} - \mathbb{X}\theta*\right|_2^2 \leq 2\epsilon'\left(\mathbb{X}\theta^{\hat{LS}} - \mathbb{X}\theta*\right) \leq 2sup_{\theta\in K-K}\epsilon'\mathbb{X}\theta$
(5) How to interpret **Gaussian width**? The Gaussian width $2sup_{\theta\in K-K}\epsilon'\mathbb{X}\theta \overset{symmetry}{=} 4sup_{v\in\mathbb{X}K}\epsilon'v$ is the measure of the width of the image$\mathbb{X}K$.
(6) The linear transformed convex set is still convex, this fact is used to obtained the constrained LSE in L^1. However, the vertices may not be exactly the images of the original vertices, depending on the rank of the linear transformation.

[1] See [**Rao1**] Chap 3

FIGURE 2.2.1. illustration of elbow effect



THEOREM 30. *(MSE for LSE restricted on $\mathcal{B}_1 \subset \mathbb{R}^d$) For a fixed design $Y = \mathbb{X}'\theta* + \epsilon, \epsilon \sim \text{subG}(\sigma^2)$, and the columns of $\mathbb{X}$ satisfy $max_j |\mathbb{X}_j| \leq \sqrt{n}$ then the LSE $\theta^{\hat{L}S}_{\mathcal{B}_1}$ satisfies*

$\mathbb{E}\left[\text{MSE}(\mathbb{X}\theta^{\hat{L}S}_{\mathcal{B}_1})\right] \lesssim \sigma \cdot \sqrt{\frac{logd}{n}}$

Moreover, for any $\delta > 0$ with probability $1 - \delta$it holds that

$\text{MSE}(\mathbb{X}\theta^{\hat{L}S}_{\mathcal{B}_1}) \lesssim \sigma \cdot \sqrt{\frac{logd - log\delta}{n}}$ *by using the Chernoff bound.*

COROLLARY 31. $\mathbb{E}\left[\text{MSE}(\mathbb{X}\theta^{\hat{L}S}_{\mathcal{B}_1})\right] \lesssim min\left(\sigma^2 \cdot \frac{r}{n}, \sigma \cdot \sqrt{\frac{logd}{n}}\right)$ *The elbow takes place around $r \simeq n$, which means the design matrix is almost of full rank.*

(a) The key of extending the previous theorem into another convex set is to use the Gaussian width to control the difference between the MSE over L^2 ball and L^1 ball.
(b) What is elbow effect? This is primary a method of selecting variables. But now we can see that visually this can be obtained by always taking the optimal bound of MSE(risk). See below the red path is the optimal bound obtained on one surface of $\mathcal{B}_1 \cup \mathcal{B}_2 \subset \mathbb{R}^3$:
(c) The sparse norm (counting nonzero components in a vector) can also be used to obtain yet another bound:

THEOREM 32. *(MSE for LSE restricted on $\mathcal{B}_0(k) \subset \mathbb{R}^d, k \leq \frac{d}{2}$, the set of k-sparse vectors) For a fixed design $Y = \mathbb{X}'\theta * + \epsilon, \epsilon \sim \text{subG}(\sigma^2)$, then the LSE $\theta^{\hat{LS}}_{\mathcal{B}_0(k)}$ satisfies for any $\delta > 0$ with probability $1 - \delta$it holds that*

$\text{MSE}(\mathbb{X}\theta^{\hat{LS}}_{\mathcal{B}_0(k)}) \lesssim \frac{\sigma^2}{n} log\begin{pmatrix} d \\ 2k \end{pmatrix} + \frac{\sigma^2}{n} \cdot k \cdot log(6) - \frac{\sigma^2}{n} log\delta$ *by using the Chernoff bound. And further by Stirling formula:*

$\text{MSE}(\mathbb{X}\theta^{\hat{LS}}_{\mathcal{B}_0(k)}) \lesssim \frac{2\sigma^2 k}{n} log\left(\frac{ed}{2k}\right) + \frac{\sigma^2}{n} \cdot k \cdot log(6) - \frac{\sigma^2}{n} log\delta$

COROLLARY 33. *For a fixed design* $Y = \mathbb{X}'\theta* + \epsilon, \epsilon \sim \text{subG}(\sigma^2)$, *then the LSE* $\theta^{L\hat{S}}_{\mathcal{B}_0(k)}$ *satisfies* $\mathbb{E}\left[\text{MSE}(\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)})\right] \lesssim \frac{\sigma^2 k}{n} log\left(\frac{ed}{k}\right)$. [2]

PROOF. Write it out in definition:
$\mathbb{E}\left[\text{MSE}(\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)})\right]$(The risk is evaluated on the sample space of response variable.)

$$
\begin{aligned}
&= \frac{1}{n}\mathbb{E}\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2\right] \\
&= \frac{1}{n}\int_{\mathcal{Y}}\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 d\mathbb{P}_Y \\
&= \frac{1}{n}\int_0^1 \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 \le u\right] du \qquad (\textit{Abelian summation technique}) \\
&= \frac{1}{n}\int_0^\infty \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > u\right] du \qquad (\textit{change of variable} : v := \tfrac{u}{n}) \\
&= \int_0^\infty \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > nv\right] dv \\
&= \int_0^H \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > nv\right] dv \qquad + \int_H^\infty \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > nv\right] dv \\
&\quad \int_0^H \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > nv\right] dv \le H \textit{ is clear} \\
&\le H + \int_0^\infty \mathbb{P}_Y\left[\left|\mathbb{X}\theta^{L\hat{S}}_{\mathcal{B}_0(k)} - \mathbb{X}\theta*\right|_2^2 > n(v+H)\right] dv \qquad (2.6) \\
&= H + \sum_{j=1}^{2k}\binom{d}{j} 6^{2k}\int_0^\infty e^{-\frac{n(v+H)}{32\sigma^2}} dv = \qquad H + \sum_{j=1}^{2k}\binom{d}{j} 6^{2k}\left[\frac{32\sigma^2}{n} e^{-\frac{nH}{32\sigma^2}}\right]
\end{aligned}
$$

Now we turn to nonparametric setup ("BIG DATA"). □

2.3. The gaussian sequence model

(1) Why study the gaussian sequence model? [**Rigollet**] pp.42"The main reason for its popularity is that it carries already most of the insight of nonparametric estimation." We can use this as a prism to explore the infinite dimensional world.
(2) What is a gaussian sequence model? $Y = \theta* + \epsilon, \epsilon \sim N(0, \sigma^2)$
(3) What is a sub-gaussian sequence model? $Y = \mathbb{X}'\theta* + \epsilon, Y \in \mathbb{R}^d, \epsilon \sim \text{subG}_\text{d}(\sigma^2)$ which is a d-dimensional sub-gaussian error.
(4) Discussion: what is a direct problem and what is an inverse problem? Fisher's fiducial probability.
(5) What are the differences between them? The error assumption and the orthogonal design assumption.

DEFINITION 34. The orthogonal design assumption or ORT is $\frac{\mathbb{X}'\mathbb{X}}{n} = I_d$[3]. The design matrix is assumption to be orthogonal. If the reader likes, we can refer [**Dean&Voss, Rao2**] for more details why orthogonal designs are more favorable, a major reason is that the contrasts are

[2] The proof of Corollary 2.9 is not clear in [**Rigollet**].

[3] When the assumption $d \le n$ is not satisfied, we call this a high-dimensional case. In usual design of experiment course, if there are more parameters than observations, some parameters are not estimable. But for high-dimensional course, somehow these parameters are estimable with some worse behaviors than BLUE.

orthogonally spanning the functional space of estimable functions. In particular it means that the d columns of $\mathbb{X}$are orthogonal in $\mathbb{R}^n$ and all have norm $\sqrt{n}$.

(6) Exercise: Write the SLR using this ORT assumption.
J.Zhang pointed out that ORT assumptions makes it easier to compute LSE and gives a direct interpretation of thresholding: the thresholding makes the components of $\hat{\theta} < \tau$ become zero and thus gives an approximation to the k-sparse version of the parameter vector θ.

COROLLARY 35. *Under ORT,* $\text{MSE}[\mathbb{X}\hat{\theta}] = \left|\hat{\theta} - \theta*\right|_2^2$ *for SLR. Moreover, under ORT, SLR is equivalent to the sub-gaussian sequence model.*

(7) What is the difference between this model and the hierarchical model? Discussion: Stein's two-step procedure.

(8) What is the philosophy of taking thresholding for approximate sparse vector? Well, loosely speaking, we have already obtained the bound on the MSE of a specific estimator and now we drop some "small" components of a vector and see how much we lose on the MSE. Discussion: Neural networks and thresholding functions.

PROPOSITION 36. *If* $\text{MSE}[\mathbb{X}\hat{\theta}] = |y - \theta*|_2^2$ *and each of its components* $(y - \theta*)_i \leq \tau$ *is* ***uniformly*** *bounded, then we lose at most* 2τ *by choosing* $\hat{\theta}_j = 0$. *Equivalently, we get* ***1 degree of sparsity*** *in trade of* 2τ ***MSE risk***.

(9) What is hard thresholding? What is soft thresholding? The goal is to obtain a uniform bound for components. By choosing good thresholdings we can achieve that.

THEOREM 37. *(The approximation of thresholding estimators for some recurrent event) With the linear regression model* $Y = \mathbb{X}\theta^* + \epsilon$ *under ORT assumption (OR equivalently, the sub-gaussian sequence model). Then the hard thresholding estimator* $\hat{\theta_{HRD}}$ *with threshold* $\tau = \sigma\sqrt{\frac{2log\left(\frac{2d}{\delta}\right)}{d}}$, *then the following properties hold for the* $\delta-$*recurrent events* A.*(i.e.* $\mathbb{P}(A) \geq 1 - \delta$*)*

(i)If $|\theta^*|_0 = k$, $\text{MSE}\left(\mathbb{X}\hat{\theta_{HRD}}\right) = \left|\hat{\theta_{HRD}} - \theta^*\right|_2^2 \lesssim \sigma^2 \frac{klog\left(\frac{2d}{\delta}\right)}{n}$

(ii)If $min_{j \in supp(\theta^*)} \left|\theta_j^*\right| > 3\tau$,*then* $supp(\hat{\theta_{HRD}}) = supp(\theta^*)$

2.4. High-dimensional linear regression

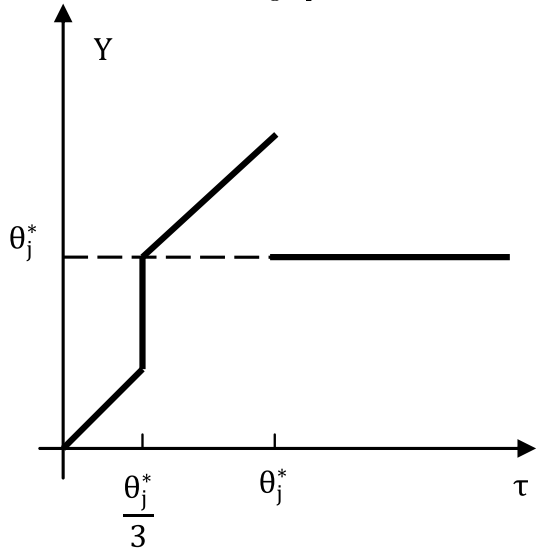
(1) The line for this section is to apply BIC and LASSO risk onto the SLR with fixed design. Discussion: g-inverse modification techniques.

LEMMA 38. *The* $\hat{\theta^{HRD}}$ *is the minimum norm solution to the SLR in* $\mathbb{R}^d$ *with norm* $|y - \theta|_2^2 + 4\tau^2 |\theta|_0$.

The $\hat{\theta^{SFT}}$ *is the minimum norm solution to the SLR in* $\mathbb{R}^d$ *with norm* $|y - \theta|_2^2 + 4\tau |\theta|_1$.

(2) How will these thresholding estimator become if ORT is satisfied? BIC and LASSO estimator.

FIGURE 2.3.1. The graph of the function



(3) "Extremely useful" duality [4] extension. The duality between the LS-risk and minimax-risk spaces gives following theorem and the extension of following theorem provides a parallel duality to the BIC and LASSO estimator.

THEOREM 39. *Let* $\mathbf{M}$ *be a positive definite matrix. Then* $(\mathbf{A}^*)^-_{m(\mathbf{M})} = \left[\mathbf{A}^-_{l(\mathbf{M}^-)}\right]^*$.

(4) A list of oracle inequality results we are to prove for parameter estimators within the setting of linear regression.
We did discuss some restricted estimators since sometimes a random model with singular dispersion matrix can be reduced to a nonsingular model with restraints. See [**Rao1**] Sec 7.4.

[4] [**Rao1**] Theorem 3.2.4

TABLE 1. List of oracles $\mathbb{P}\left\{\left[\left|\hat{\theta_n}-\theta^*\right|_2^2\right]\leq C\phi(\Theta)\right\}=1-\frac{1}{d^2}\iff$
$\mathbb{E}\left[\left|\hat{\theta_n}-\theta^*\right|_2^2\right]\leq C\phi(\Theta)$
See pp.102 of [**Rigollet**].

Name	Oracle Inequality $\mathbb{P}\{*\}=1-\delta$	τ	Thm in [Rigollet]
$Y_{n\times 1}=\mathbb{X}_{n\times d}\theta^*_{d\times 1}+\epsilon,\epsilon\sim \mathrm{subG}_n(\sigma^2)$			
LSE	$\frac{1}{n}\mathbb{E}\left\|\mathbb{X}\hat{\theta^{LS}}-\mathbb{X}\theta^*\right\|\lesssim\sigma^2\frac{r}{n}$		2.2
$\mathcal{B}_1$−LSE $\begin{cases}\left\|\mathbb{X}_j\right\|_2\leq\sqrt{n}\\ \theta^*\in\mathcal{B}_1\end{cases}$	$\frac{1}{n}\mathbb{E}\left\|\mathbb{X}\hat{\theta}^S_{\mathcal{B}_1}-\mathbb{X}\theta^*\right\|\lesssim\sigma\sqrt{\frac{log\,d}{n}}$		2.4
$\mathcal{B}_0(k)$−LSE $\theta^*\in\mathcal{B}_0(k)$	$\frac{1}{n}\mathbb{E}\left\|\mathbb{X}\theta^{\hat{L}S}_{\mathcal{B}_0(k)}-\mathbb{X}\theta^*\right\|\lesssim\frac{\sigma^2}{n}\begin{pmatrix}d\\2k\end{pmatrix}+\frac{\sigma^2k}{n}+\frac{\sigma^2}{n}log\frac{1}{\delta}$		2.6
	$\frac{1}{n}\mathbb{E}\left\|\mathbb{X}\theta^{\hat{L}S}_{\mathcal{B}_0(k)}-\mathbb{X}\theta^*\right\|\lesssim\frac{\sigma^2}{n}log\left(\frac{ed}{2k}\right)+\frac{\sigma^2k}{n}log6+\frac{\sigma^2}{n}log\frac{1}{\delta}$		2.8
HRD	$\frac{1}{n}\left\|\mathbb{X}\hat{\theta^{HRD}}-\mathbb{X}\theta^*\right\|_2^2\lesssim\frac{\sigma^2k}{n}log\frac{2d}{\delta}$	$2\tau=2\sigma\sqrt{\frac{2log(\frac{2d}{\delta})}{n}}$	2.11
BIC	$\frac{1}{n}\left\|\mathbb{X}\hat{\theta^{BIC}}-\mathbb{X}\theta^*\right\|_2^2\lesssim\left\|\theta^*\right\|_0\sigma^2log\frac{ed}{\delta}$	$\tau^2=16\,(log6)\cdot\frac{\sigma^2}{n}+32\frac{\sigma^2log(ed)}{n}$	
LASSO(slow) $\frac{\left\|\mathbb{X}_j\right\|_2}{\sqrt{n}}\leq$	$\frac{1}{n}\left\|\mathbb{X}\hat{\theta^{LASSO}}-\mathbb{X}\theta^*\right\|_2^2\leq 4\left\|\theta^*\right\|_1\sigma\sqrt{\frac{2log(2d)}{n}+\frac{2log(\frac{1}{\delta})}{n}}$	$2\tau=8\sigma\sqrt{\frac{2log(2d)}{n}+\frac{log(\frac{1}{\delta})}{n}}$	
LASSO(fast) $\begin{cases}\mathbb{X}\sim INC(k)\\ \theta^*\in\mathcal{B}_0(k)\end{cases}$	$\frac{1}{n}\left\|\mathbb{X}\hat{\theta^{LASSO}}-\mathbb{X}\theta^*\right\|_2^2\lesssim\frac{k\sigma^2}{n}log\frac{2d}{\delta}$	$2\tau=8\sigma\sqrt{\frac{2log(2d)}{n}+\frac{log(\frac{1}{\delta})}{n}}$	2.18
$Y_{n\times 1}=f\left(\mathbb{X}_{n\times 1}\right)+\epsilon,\epsilon\sim \mathrm{subG}_n(\sigma^2)$			
LSE	$\frac{1}{n}\left\|\hat{f}-f\right\|_2^2\leq inf_{\theta\in\mathbb{R}^M}\frac{1}{n}\left\|f_\theta-f\right\|_2^2+C\sigma^2\frac{M}{n}log(\frac{1}{\delta})$		3.3
BIC	$\frac{1}{n}\left\|f^{\hat{BIC}}-f\right\|_2^2\leq inf_{\theta\in\mathbb{R}^M}\frac{1+\alpha}{1-\alpha}\frac{1}{n}\left\|f_\theta-f\right\|_2^2+\frac{C\sigma^2}{\alpha(1-\alpha)n}\left\|\theta\right\|_0log(eM)+\frac{C\sigma^2}{\alpha(1-\alpha)n}log(\frac{1}{\delta})$	$\tau^2=\frac{16\sigma^2}{\alpha n}log(6eM)$	3.4
LASSO $\left[f_j(X_i)\right]_{ij}$ $INC(k)\sim$	$\frac{1}{n}\left\|f\hat{LASSO}-f\right\|_2^2\leq inf_{\theta\in\mathbb{R}^M}\frac{1+\alpha}{1-\alpha}\frac{1}{n}\left\|f_\theta-f\right\|_2^2+\frac{C\sigma^2}{\alpha(1-\alpha)n}\left\|\theta\right\|_0log(eM)+\frac{C\sigma^2}{\alpha(1-\alpha)n}log(\frac{1}{\delta})$	$2\tau=8\sigma\sqrt{\frac{2log(2M)}{n}+\frac{log(\frac{1}{\delta})}{n}}$	3.5

CHAPTER 3

Mis-specified Linear Models

3.1. Oracle inequalities

(1) The sub-normality assumption we made in previous chapter is way too strong. Sometimes we shall discuss the departure from the underlying assumptions for the normal linear models. In usual linear model course like [**Dean&Voss, Rao1**], we discussed what will happen if the variance of error terms cannot be regarded as equal; if the error terms are not normal any more; if the errors are correlated. Now we are concerned with the loss of linearity. This is different from the old discussions.
The approach adopted by [**Rao1**] is no longer valid since the expectation and variance cannot characterize the model distribution any longer.
When the chosen model $Y_i = f(X_i) + \epsilon_i, i = 1, \cdots, n$ does NOT give much twisted result when the assumptions of this model is departured, we call the model ***robust***.

(2) What is the difference between multivariate statistics and the high-dimensional statistics? Although their models are often the same one, the multivariate statistics focus on the distribution behavior of a specific statistic while the high-dimensional statistics focus on the error bound given by sparsity in high dimensions.

(3) We have no chance of finding a consistent estimator if we do not know the correct model. However, if we know something absolutely correct, we got an oracle. Oracle inequality can be used to say something to the estimators. Estimators are about the parameters; predictors are about responses.

DEFINITION 40. An oracle is a quantity that cannot be constructed without knowledge of the quantity of interest itself. The oracle we talked about here is the regression function $f(\mathbf{X})$. For all matter of purposes, an oracle can be viewed as an estimator in a given family that can be constructed with an infinite amount of data.

(4) What is the general form of an oracle within a family of functions represented w.r.t a basis consisting of functions? $f \approx \varphi_\theta := \sum_{j=1}^{M} \theta_j \varphi_j$. If the basis $\mathcal{H}$ is chosen to be the dual space of $\mathbb{R}^M$, then the GOAL is to estimate f using linear functions.

(5) What are the estimators under study?

Estimator	Expression
LSE	$\hat{\theta^{LS}} := argmin_{\theta \in \mathbb{R}^M} \frac{1}{n} \sum_{i=1}^{n} [Y_i - \varphi_\theta(X_i)]$

RLSE	$\hat{\theta_K^{LS}} \coloneqq argmin_{\theta\in K}\frac{1}{n}\sum_{i=1}^{n}[Y_i-\varphi_\theta(X_i)]$
BIC	$\hat{\theta^{BIC}} \coloneqq \left\{argmin_{\theta\in\mathbb{R}^M}\frac{1}{n}\sum_{i=1}^{n}[Y_i-\varphi_\theta(X_i)]+\tau^2\left\|\theta\right\|_0\right\}$
LASSO	$\hat{\theta^{L}} \coloneqq argmin_{\theta\in\mathbb{R}^M}\left\{argmin_{\theta\in\mathbb{R}^M}\frac{1}{n}\sum_{i=1}^{n}[Y_i-\varphi_\theta(X_i)]+2\tau^2\left\|\theta\right\|_1\right\}$

(6) What is the formal definition of oracle? What is oracle inequality?

DEFINITION 41. Let $R(\cdot)$ be a risk function on the basis space $\mathcal{H}=\{\varphi_1,\cdots,\varphi_M\}$ be a family of functions in the smooth dual space of $\mathbb{R}^d, K\subset\mathbb{R}^M$. The oracle on K w.r.t $R(\cdot)$ is denoted by $\varphi_{\bar{\theta}}, \bar{\theta}\in K$ where $R(\varphi_{\bar{\theta}})\leq R(\varphi_\theta), \forall\theta\in K$. This risk $R_K\coloneqq R(\varphi_{\bar{\theta}})$ is called oracle risk on K.
An estimator $\hat{f}$ to f is said to satisfy an oracle inequality over K with remainder ϕ in expectation(or in high probability/recurrent, via Chernoff bound) if there exists a constant $C\geq 1$ s.t. $\mathbb{E}R(\hat{f})\leq Cinf_{\theta\in K}R(\varphi_\theta)+\phi_{n,M}(K)\Longleftrightarrow\mathbb{P}\left\{R(\hat{f})\leq Cinf_{\theta\in K}R(\varphi_\theta)+\phi_{n,M,\delta}(K)\right\}\geq 1-\delta,\forall\delta>0.$
Exercise: Use Taylor expansion and the L^2-risk, calculate the remainder of any functions you are interested in. Now can we have a uniform C for an oracle inequality? Why or why not? What if we restrict the convex set onto unit ball?
Discussion: Use Laurent series and the $L^1_{\mathbb{C}}-$risk, calculate the remainder of any analytic functions you are interested in. What did Liouville Theorem tells you?

(a) Let me give some comments on this definition. The first is that we have already encountered oracle in Chap.2 of [**Rigollet**], whose risk function is MSE $\left(\mathbb{X}\hat{\theta}\right)$ then. We know that we have results on $\mathbb{E}$MSE, so we actually dropped the remainder term in exchange of a looser bound. One oracle inequality is Theorem 32 $\mathbb{P}\left\{\text{MSE}(\mathbb{X}\theta^{\hat{L}S}_{\mathcal{B}_1})\lesssim\sigma\cdot\sqrt{\frac{logd-log\delta}{n}}\right\}>$ $1-\delta$, this is the oracle for $\hat{Y}=\mathbb{X}\theta^{\hat{L}S}$, with MSE risk and the oracle is not restrained to any subset. However, we do not obtained the oracle risk and did not yield the remainder.
(b) Why we need oracle? Because oracle gives us a probability statement about how good a estimator will be. And more importantly, it also told us "how far" an estimator is from a "best" estimator under this risk. Oracle inequality is useful since we do not have to know what is exactly the "best" estimator in order to compare our current estimator to the "best" one.
This technique can be traced back to two founders of modern decision theory, see [**Chernoff&Mose**].

(7) Oracle inequality for LSE

THEOREM 42. *Let $Y=f(X)+\epsilon,\epsilon\sim\text{subG}_\text{n}(\text{œ}^2)$, then the LSE $\theta^{\hat{L}S}$ satisfies for some constant $C>0,\forall\alpha\in(0,1)$*
$\mathbb{E}\text{MSE}(\varphi_{\theta^{\hat{L}S}})\leq inf_{\theta\in\mathbb{R}^M}MSE(\varphi_\theta)+C\frac{\sigma^2M}{n}\Longleftrightarrow$
$\mathbb{P}\left\{\text{MSE}(\varphi_{\theta^{\hat{L}S}})\leq inf_{\theta\in\mathbb{R}^M}\text{MSE}(\varphi_\theta)+C\frac{\sigma^2M}{n}log(\frac{1}{\delta})\right\}>1-\delta.$

(8) Oracle inequality for BIC
The interpretation of the theorem: (pp.64) It implies that the BIC estimator will mimic the "best" trade-off between the approximation error MSE('·) and the complexity of θ as measured by its sparsity.

THEOREM 43. *Let $Y = f(X) + \epsilon, \epsilon \sim \mathrm{subG_n}(\text{œ}^2)$, then the BIC $\theta^{\hat{BIC}}$ satisfies for some constant $C > 0$*

$\mathbb{E}\mathrm{MSE}(\varphi_{\theta^{\hat{BIC}}}) \leq inf_{\theta \in \mathbb{R}^M} \left[\frac{1+\alpha}{1-\alpha} \mathrm{MSE}(\varphi_\theta) + C \frac{\sigma^2}{\alpha(1-\alpha)n} |\theta|_0 log(eM) \right] + C \frac{\sigma^2}{\alpha(1-\alpha)n} \iff$

$\mathbb{P}\left\{ \mathrm{MSE}(\varphi_{\theta^{\hat{BIC}}}) \leq inf_{\theta \in \mathbb{R}^M} \left[\frac{1+\alpha}{1-\alpha} \mathrm{MSE}(\varphi_\theta) + C \frac{\sigma^2}{\alpha(1-\alpha)n} |\theta|_0 log(eM) \right] + C \frac{\sigma^2}{\alpha(1-\alpha)n} log(\frac{1}{\delta}) \right\} > 1 - \delta.$

where the regularization parameter for BIC is $\tau^2 = \frac{16\sigma^2}{\alpha n} log(6eM), \alpha \in (0,1)$

Discussion: Young's inequality. I provided the following proof as a replacement of that one in [**Rigollet**], which is rigorous and more explicit[1]:

PROOF. The proof consists of following steps.
(i) Estimation of MSE.(Theorem 2.14)

$\mathrm{MSE}(\mathbb{X}\theta^{\hat{BIC}}) = \frac{1}{n} \left| \mathbb{X}\theta^{\hat{BIC}} - \mathbb{X}\theta^* \right|_2^2$(definition of MSE)

$\frac{1}{n} \left| Y - \mathbb{X}\theta^{\hat{BIC}} \right|_2^2 + \tau^2 \left| \theta^{\hat{BIC}} \right|_0 \leq \frac{1}{n} |Y - \mathbb{X}\theta^*|_2^2 + \tau^2 |\theta^*|_0$(definition of BIC estimator, and the regularization parameter τ is involved here.)

Since the model is $Y = \mathbb{X}\theta^* + \epsilon$, we plug in on both sides to yield:

$\frac{1}{n} \left| \mathbb{X}\theta^* + \epsilon - \mathbb{X}\theta^{\hat{BIC}} \right|_2^2 + \tau^2 \left| \theta^{\hat{BIC}} \right|_0 \leq \frac{1}{n} |\mathbb{X}\theta^* + \epsilon - \mathbb{X}\theta^*|_2^2 + \tau^2 |\theta^*|_0$

$\frac{1}{n} \left| \mathbb{X}\theta^* + \epsilon - \mathbb{X}\theta^{\hat{BIC}} \right|_2^2 + \tau^2 \left| \theta^{\hat{BIC}} \right|_0 \leq \frac{1}{n} |\epsilon|_2^2 + \tau^2 |\theta^*|_0$

$\left| \mathbb{X}\theta^* + \epsilon - \mathbb{X}\theta^{\hat{BIC}} \right|_2^2 + n\tau^2 \left| \theta^{\hat{BIC}} \right|_0 \leq |\epsilon|_2^2 + n\tau^2 |\theta^*|_0$

$\left| \mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right|_2^2 - 2\epsilon' \left(\mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right) + |\epsilon|_2^2 + n\tau^2 \left| \theta^{\hat{BIC}} \right|_0 \leq |\epsilon|_2^2 + n\tau^2 |\theta^*|_0$(Expand the whole expression as a quadratic form in $\mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}}$,$\epsilon$, cancel those redundant terms)

$\left| \mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right|_2^2 - 2\epsilon' \left(\mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right) + n\tau^2 \left| \theta^{\hat{BIC}} \right|_0 \leq n\tau^2 |\theta^*|_0$

$\left| \mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right|_2^2 \leq 2\epsilon' \left(\mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right) - n\tau^2 \left| \theta^{\hat{BIC}} \right|_0 + n\tau^2 |\theta^*|_0$

Note that this is a quadratic form in $\Delta := \left(\mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right)$, which can be written as:

$\Delta^2 \leq 2\epsilon'\Delta - n\tau^2 \left| \theta^{\hat{BIC}} \right|_0 + n\tau^2 |\theta^*|_0$

If we want to estimate the Δ, then we have a domain bounded by the parabola when both $\theta^{\hat{BIC}}, \theta^*, (\epsilon' u)^2$ are fixed. Now it is not fixed BUT dependent on the observed data(design matrix $\mathbb{X}$), so we need to figure out the envelope of this family of parabolas by some transformations.

$2\epsilon'\Delta = 2\epsilon' \left(\mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right) = 2\epsilon' \frac{\left(\mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right)}{\left| \left(\mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right) \right|_2} \cdot \left| \left(\mathbb{X}\theta^* - \mathbb{X}\theta^{\hat{BIC}} \right) \right|_2 = 2\epsilon' u \cdot |\Delta|_2 \leq 2 (\epsilon' u)^2 + \frac{1}{2}\Delta^2$(Young's inequality)

[1] *Note that $|*|$ is the dual norm of the usual norm on the functional space$\|*\|$.*

This inequality can be combined with the quadratic inequality and yield:

$\Delta^2 \leq 2\epsilon'\Delta - n\tau^2 \left|\theta^{\hat{BIC}}\right|_0 + n\tau^2 \left|\theta^*\right|_0 \leq 2\left(\epsilon' u\right)^2 + \frac{1}{2}\Delta^2 - n\tau^2 \left|\theta^{\hat{BIC}}\right|_0 + n\tau^2 \left|\theta^*\right|_0$

$\frac{1}{2}\Delta^2 \leq 2\left(\epsilon' u\right)^2 - n\tau^2 \left|\theta^{\hat{BIC}}\right|_0 + n\tau^2 \left|\theta^*\right|_0$

$\Delta^2 \leq 4\left(\epsilon' u\right)^2 - 2n\tau^2 \left|\theta^{\hat{BIC}}\right|_0 + 2n\tau^2 \left|\theta^*\right|_0$ which is exactly (2.15) in **[Rigollet]**.

We enlarge the family of the parabolas, hence the envelope is also enlarged. The envelope of this family is still not completely determined since $\left(\epsilon' u\right)^2$ is not yet determined. We must impose a bound on this term since we want a completely determined envelope.

$sup_{\theta\in\mathbb{R}^d} = max_{1\leq k\leq d} max_{|S|=k} sup_{supp(\theta)=S}$

$4\left(\epsilon' u\right)^2 - 2n\tau^2 \left|\theta^{\hat{BIC}}\right|_0 \leq sup_{\theta\in\mathbb{R}^d}\left\{4\left(\epsilon' u\right)^2 - 2n\tau^2 \left|\theta^{\hat{BIC}}\right|_0\right\}$

$= max_{1\leq k\leq d} max_{|S|=k} sup_{supp(\theta)=S}\left\{4\left(\epsilon' u\right)^2 - 2n\tau^2 k\right\}$

$\leq max_{1\leq k\leq d} max_{|S|=k} sup_{u\in\mathcal{B}_2^{rank(Col(S))}}\left\{4\left(\epsilon'\Phi_{S,*}u\right)^2 - 2n\tau^2 k\right\}$

Here we shall explain why we insert each supremum. Now $sup_{u\in\mathcal{B}_2^{rank(Col(S,*))}}$ is inserted because we want to use the maximal inequality on $\mathcal{B}_2^{rank(Col(S),*)}$ to bound on $\epsilon'\Phi_{S,*}u$, where $rank(Col(S,*))$ is the rank of the augmented matrix $\left[S\vdots supp(\theta)\right]$; $max_{|S|=k}$ is inserted because we want to get rid of $\left|\theta^{\hat{BIC}}\right|_0$; The outest max must be imposed for equivalent bound. Note that the term $2n\tau^2 \left|\theta^{\hat{BIC}}\right|_0$ is also included into this "sup-out", so the only thing left outside is the $+2n\tau^2 \left|\theta^*\right|_0$, which we deal at the end.

$\mathbb{P}\left(sup_{u\in\mathcal{B}_2^{rank(Col(S))}}\left\{4\left(\epsilon'\Phi_{S,*}u\right)^2 - 2n\tau^2 k\right\} \geq t_0\right) \leq 2\cdot 6^{rank(Col(S,*))} \cdot exp(-\frac{t_0}{8\sigma^2})$

Use the union bound technique to yield

$\mathbb{P}\left\{max_{1\leq k\leq d} max_{|S|=k} sup_{u\in\mathcal{B}_2^{rank(Col(S))}}\left\{4\left(\epsilon'\Phi_{S,*}u\right)^2 - 2n\tau^2 k\right\} \geq t\right\} \leq \sum_{k=1}^{d}\sum_{|S|=k}\mathbb{P}\left(sup_{u\in\mathcal{B}_2^{rank(Col(S))}}\left\{4\left(\epsilon'\Phi_{S,*}u\right)^2 - 2n\tau^2 k\right\} \geq \frac{t}{4} + \frac{1}{2}n\tau^2 k\right)$

The $\epsilon-$net argument is still working here by noticing that

$\mathbb{P}\left(sup_{u\in\mathcal{B}_2^{rank(Col(S))}}\left\{4\left(\epsilon'\Phi_{S,*}u\right)^2 - 2n\tau^2 k\right\} \geq \frac{t}{4} + \frac{1}{2}n\tau^2 k\right) \leq 2\cdot 6^{rank(Col(S,*))} \cdot exp(-\frac{\frac{t}{4}+\frac{1}{2}n\tau^2 k}{8\sigma^2})$ (maximal inequality, this step involves the $1-\delta$ probability, so from this step on we have all probability statement.)

$\leq 2\cdot exp(-\frac{\frac{t}{4}+\frac{1}{2}n\tau^2 k}{8\sigma^2} + [rank(Col(S,*))]\, log(6))$

$\leq 2\cdot exp(-\frac{\frac{t}{4}+\frac{1}{2}n\tau^2 k}{8\sigma^2} + [k + \left|\theta^*\right|_0]\, log(6))$

Plug this inequality back into the union bound we yield

$\Delta^2 \leq 4\left(\epsilon' u\right)^2 - 2n\tau^2 \left|\theta^{\hat{BIC}}\right|_0 + 2n\tau^2 \left|\theta^*\right|_0 \leq 4\cdot max_{1\leq k\leq d} max_{|S|=k} sup_{u\in\mathcal{B}_2^{rank(Col(S))}}\left\{4\left(\epsilon'\Phi_{S,*}u\right)^2 - 2n\tau^2 k\right\} + 2n\tau^2 \left|\theta^*\right|_0$

$\leq 4\cdot\sum_{k=1}^{d}\sum_{|S|=k}\mathbb{P}\left(sup_{u\in\mathcal{B}_2^{rank(Col(S))}}\left\{4\left(\epsilon'\Phi_{S,*}u\right)^2-2n\tau^2k\right\}\geq\frac{t}{4}+\frac{1}{2}n\tau^2k\right)+\Big|2n\tau^2\left|\theta^*\right|_0$

$\leq\left\{4\cdot\sum_{k=1}^{d}\begin{pmatrix}d\\k\end{pmatrix}\cdot 2\cdot exp(-\frac{\frac{t}{4}+\frac{1}{2}n\tau^2k}{8\sigma^2}+[k+|\theta^*|_0]\,log(6))\right\}+2n\tau^2\left|\theta^*\right|_0\Big|$

$\leq\left\{4\cdot\sum_{k=1}^{d}exp(-\frac{t}{32\sigma^2}-klog\,ed+|\theta^*|_0\,log(12))\right\}+2n\tau^2\left|\theta^*\right|_0$(using Stirling's formula to yield a same order bound)

Now we should deal with the $+2n\tau^2\left|\theta^*\right|_0$ term, which contains only constant term once the sparsity is specified.

$\lesssim|\theta^*|\cdot\sigma^2log(\frac{ed}{\delta})$.

Note that MSE $=\frac{1}{n}\Delta^2$.

(ii)Into the dual space.(Theorem 3.4)

This proof shows the usage of "middleman trick". Note that we should just interpret that $\varphi_{\theta^{\hat{BIC}}}=\left(\theta_i^{\hat{BIC}}\right)_i$ is the coordinate under the basis $\{\varphi_1,\cdots,\varphi_M\}$ where those linear functions are actually functionally independent.

$\text{MSE}(\varphi_{\theta^{\hat{BIC}}})=\frac{1}{n}\left|\varphi_{\theta^{\hat{BIC}}}-f\right|_2^2=\frac{1}{n}\left|\varphi_{\theta^{\hat{BIC}}}-(Y-\epsilon)\right|_2^2$

$\left|\varphi_{\theta^{\hat{BIC}}}-f\right|_2^2=\left|\varphi_{\theta^{\hat{BIC}}}-(Y-\epsilon)\right|_2^2=\left|\varphi_{\theta^{\hat{BIC}}}-Y\right|_2^2+2\epsilon'\left(\varphi_{\theta^{\hat{BIC}}}-Y\right)+\Big|\epsilon^2$

$|\varphi_\theta-f|_2^2=|\varphi_\theta-(Y-\epsilon)|_2^2=|\varphi_\theta-Y|_2^2+2\epsilon'\left(\varphi_\theta-Y\right)+\epsilon^2$

$\left|\varphi_{\theta^{\hat{BIC}}}-Y\right|_2^2+n\tau^2\left|\theta^{\hat{BIC}}\right|_0\leq|\varphi_\theta-Y|_2^2+n\tau^2\left|\theta\right|_0$ (definition of the BIC estimator)

The three inequalities above together imply that

$\left|\varphi_{\theta^{\hat{BIC}}}-f\right|_2^2+n\tau^2\left|\theta^{\hat{BIC}}\right|_0\leq|\varphi_\theta-f|_2^2+2\epsilon'\left(\varphi_{\theta^{\hat{BIC}}}-Y\right)+n\tau^2\left|\theta\right|_0$

It suffices to prove the cross-product term is bounded by something as we did for the term above.

$2\epsilon'\left(\varphi_{\theta^{\hat{BIC}}}-Y\right)=2\epsilon'\left(\frac{\varphi_{\theta^{\hat{BIC}}}-Y}{\left|\varphi_{\theta^{\hat{BIC}}}-Y\right|_2}\right)\cdot\left|\varphi_{\theta^{\hat{BIC}}}-Y\right|_2\leq\frac{2}{\alpha}\left[\epsilon'\left(\frac{\varphi_{\theta^{\hat{BIC}}}-Y}{\left|\varphi_{\theta^{\hat{BIC}}}-Y\right|_2}\right)\right]^2+\Big|\frac{\alpha}{2}\cdot\left|\varphi_{\theta^{\hat{BIC}}}-Y\right|_2^2$(Young's inequality)

$\frac{\alpha}{2}\cdot\left|\varphi_{\theta^{\hat{BIC}}}-Y\right|_2^2=\frac{\alpha}{2}\cdot\left|\varphi_{\theta^{\hat{BIC}}}-f+f-Y\right|_2^2\leq\alpha\left|\varphi_{\theta^{\hat{BIC}}}-f\right|_2^2+\alpha\left|f-Y\right|_2^2$(Triangle inequality)$\tau^2=\left[16\sigma^2\cdot\frac{2log(2M)}{n}+32\sigma^2\sqrt{\frac{4log(\frac{2M}{\delta})}{n^2}}+16\sigma^2\cdot\frac{2log(\frac{1}{\delta})}{n}\right]\Big|$

Now plug all these inequalities into the previous bound and get as we wanted □

COROLLARY 44. *If the linear model is the true model,* $\text{MSE}(\varphi_{\bar{\theta}})=0$

(9) Oracle inequality for LASSO

THEOREM 45. *Let* $Y=f(X)+\epsilon,\epsilon\sim\text{subG}_\text{n}(\text{œ}^2)$*, then the LASSO* $\hat{\theta^L}$ *satisfies for some constant* $C>0$

$$\mathbb{E}\text{MSE}(\varphi_{\hat{\theta^L}})\leq inf_{\theta\in\mathbb{R}^M,|\theta|_0\leq k}\left[\frac{1+\alpha}{1-\alpha}\text{MSE}(\varphi_\theta)+C\frac{\sigma^2}{\alpha(1-\alpha)n}\left|\theta\right|_0log(eM)\right]+C\frac{\sigma^2}{\alpha(1-\alpha)n}\Longleftrightarrow\Big|$$

$$\mathbb{P}\left\{\text{MSE}(\varphi_{\hat{\theta^L}}) \leq inf_{\theta\in\mathbb{R}^M,|\theta|_0\leq k}\left[\frac{1+\alpha}{1-\alpha}\text{MSE}(\varphi_\theta)+C\frac{\sigma^2}{\alpha(1-\alpha)n}|\theta|_0 log(eM)\right]+C\frac{\sigma^2}{\alpha(1-\alpha)n}log(\frac{1}{\delta})\right\}\Bigg|$$

$$> 1-\delta where\ the\ regularization\ parameter\ for\ BIC\ is\ \tau = 4\sigma\left[\sqrt{\frac{2log(2M)}{n}}+\sqrt{\frac{2log(\frac{1}{\delta})}{n}}\right], \alpha \in \Bigg|$$

$(0,1)$ *and the design matrix satisfies*$\mathsf{INC}(k)$

Discussion: The usual technique of proving oracle inequality is the "middleman trick" and the "denominator trick". The "middleman" allows us to obtain the departure between real estimator and the given estimator; the "denominator" allows us to optimize over a unit ball. (which is more tractable than a simple convex set.)

(10) What is Zipf's law? The *Zipf-Mandelbrot law* is a empirical power-law distribution on ranked data. Zipf's law is most easily observed by plotting the data on a log-log graph, with the axes being $log(rank\,order)$ and $log(frequency)$. The data conform to Zipf's law to the extent that the plot is linear. Here the Zipf's law is referred because we are using a probability statement. We tend to believe that the probability of a sparse vector occurring is following Zipf's law, with coefficients decaying polynomially.

(11) Maurey's argument: trade-off between sparsity and MSE.

THEOREM 46. *Let* $\{\varphi_1,\cdots,\varphi_M\}$ *be a basis normalized in such a way that* $max_{1\leq j\leq M}|\varphi_j|_2 \leq D\sqrt{n}$. *Then for* $\forall 1\leq k\leq M$ *and positive* R*:*

$$min_{\theta\in\mathbb{R}^M,|\theta|_0\leq 2k}\text{MSE}('_0) \leq \min_{`\in\mathbb{R}^M,|`|_1\leq R}\text{MSE}('_0)+\frac{D^2R^2}{k}$$

3.2. Nonparametric regression

(1) Review: The wavelet basis and Fourier basis for rapid-decay functions. Fundamentals of Fourier analysis.

(2) The Sobolev space
   (a) The Sobolev space $W(\beta,L) \coloneqq \left\{f:[0,1]\to\mathbb{R}:\left\|f^{(\beta)}\right\|_2\leq L\right\}, L$ does noting special but specify the radius of function norm.
   (b) The Sobolev ellipsoid $\Theta(\beta,Q) \coloneqq \left\{\theta\in L^2(\mathbb{N}):\sum_{j=1}^\infty a_j^2\theta_j^2\leq Q, a_j\sim(\pi j)^{2\beta}\, as\, j\to\infty\right\}\Bigg|$ the Q again specify the radius of sequence norm.
   (c) A function belongs to Sobolev space $W(\beta,L)\Longleftrightarrow$Its Fourier coefficient sequence belongs to Sobolev ellipsoid $\Theta(\beta,Q=\frac{L^2}{\pi^{2\beta}})$

(3) Regular design. A fractional design is said to be regular if its aliasing structure is explicitly described by the defining contrast free group. The [**Rigollet**] used this term to indicate the sampling is equally-space, which is not very appropriate. I will also call it *systematic sampling*. The Lemma 3.13 actually said that for a systematic sampling whose cardinality is more than the cardinality of a $M-$sub trigonometric basis, then the design matrix on this sampling is ORT. (We need this Lemma to yield ORT and in order to make use of rate for LS estimator in Sec 3.1 in the proof of Sobolev oracle)
   The M- sub trigonometric basis is directly related to the $M-$truncated Fourier series mentioned below. It disregards all terms falling behind the $M-$th basis. And for a Sobolev function f, we can yield the remainder

estimation of the $M-$ truncated Fourier series obtained by calculating its Fourier coefficients.

LEMMA 47. *(Truncation estimation)For any integer $M \geq 1, \beta > \frac{1}{2}, f \in \Theta(\beta, Q)$ (must include smooth functions, so we choose $\beta \geq \frac{1}{2}$), it holds that $\left\|\varphi_{\theta^*}^M - f\right\|_2^2 \leq QM^{-2\beta}$.*

And for $M = n-1$ it holds that $\left\|\varphi_{\theta^*}^{n-1} - f\right\|_2^2 \lesssim Qn^{2-2\beta}$

COROLLARY 48. *(Fourier's oracle)* $\left\|\varphi_{\theta^{\hat{L}S}}^M - \varphi_{\theta^*}^M\right\|_2^2 \lesssim \frac{1}{\alpha(1-\alpha)} Qn^{2-2\beta} + \frac{\sigma^2 M}{1-\alpha} log(\frac{1}{\delta})$

(4) Oracle for Sobolev models
This is actually a classic estimation of remainder of $M-$ truncated Fourier series when the regression function of certain Sobolev class. This result, although stated only for $L^2[0,1]$, can be extended to any connected complete spaces without difficulty. The basic idea is to apply middleman's trick on $\left(\varphi_{\theta^{\hat{L}S}}^{n-1} - f\right) = \left(\varphi_{\theta^{\hat{L}S}}^{n-1} - \varphi_{\theta^*}^{n-1}\right) + \left(\varphi_{\theta^*}^{n-1} - f\right), \theta^*$ being the TRUE parameter for the regression function. The first term is bounded by Fourier's oracle and the second term is bounded by truncation estimation. So now the estimator is estimating a truncated version of TRUE regression function. Here we have two step approximation(One is the truncation the other is estimator.).

THEOREM 49. *For $\beta \geq \frac{1+\sqrt{5}}{4}$[2], the sub-gaussian Sobolev regression model $Y = f(X) + \epsilon, f \in W(\beta, L) = \widehat{\Theta(\beta, Q)}, \epsilon \sim \mathrm{subG_n}(\mathfrak{o}\mathfrak{e}^2), \mathfrak{o}\mathfrak{e}^2 \leq 1$[3] $M = \left\lceil n^{\frac{1}{2\beta+1}} \right\rceil \leq n-1$.*

Then $\mathbb{P}\left\{\left\|\varphi_{\theta^{\hat{L}S}}^{n-1} - f\right\|_2^2 \leq C_{\beta,Q,\delta}\left[n^{-\frac{2\beta}{2\beta+1}} + \sigma^2 \frac{log(\frac{1}{\delta})}{n}\right]\right\} > 1-\delta$.

COROLLARY 50. *For $\beta \geq \frac{1+\sqrt{5}}{4}$, the sub-gaussian Sobolev regression model $Y = f(X) + \epsilon, f \in W(\beta, L) = \widehat{\Theta(\beta, Q)}, \epsilon \sim \mathrm{subG_n}(\mathfrak{o}\mathfrak{e}^2), \mathfrak{o}\mathfrak{e}^2 \leq 1M = \left\lceil n^{\frac{1}{2\beta+1}} \right\rceil \leq n-1$.*

Then $\mathbb{P}\left\{\left\|\varphi_{\theta^{\hat{L}}}^{n-1} - f\right\|_2^2 \leq C_{\beta,Q,\delta}\left[n^{-\frac{2\beta}{2\beta+1}} + \sigma^2 \frac{log(\frac{1}{\delta})}{n}\right]\right\} > 1-\delta$. and $\mathbb{P}\left\{\left\|\varphi_{\theta^{\hat{B}IC}}^{n-1} - f\right\|_2^2 \leq C_{\beta,Q,\delta}\left[n^{-\frac{2\beta}{2\beta+1}} + \sigma^2 \frac{log(\frac{1}{\delta})}{n}\right]\right\} > 1-\delta$.

(5) Adaptive estimation. The adaptivity stated here is actually different from those we encountered in Chap.3. There we talked about adaptivity about sparsity, which means that we bounded MSE error *uniformly* even we do not know the actual $k-$dimensional subspace a $k-$sparse vector lies; Here we talked about adaptivity about smoothness parameter β, which means that we bounded L^2 error *uniformly* even we do not know the actual smoothness the regression function satisfies. In both cases, in order to get

[2]This is a sharp bound due to solving $n^{1-2\beta} \leq n^{-\frac{2\beta}{2\beta+1}}$.

[3]This bound on variance proxy is not essential, just to ensure that $\epsilon \in [-1,1]$, and can be modified by scaling.

a uniform bound, we pay the price of a log factor in the oracle inequality. The oracle is still here since we do not have to know the TRUE regression function to get the bound.

CHAPTER 4

Matrix estimation

4.1. Basic facts about matrices

(1) Review of generalized inverses, bilinear simultaneous reduction and special types of matrix equations. A brief mention of Vec/Vech operators.
Exercise: Recall and prove the tensor derivation notation. $\frac{\partial}{\partial x_{ij}} f = f_{s\times t} \otimes_{\mathbb{F}^*_{m\times n} \otimes \mathbb{F}^*_{s\times t}}$ $\left(\frac{\partial}{\partial x_{ij}}\right)_{n\times m}$ for $f : \mathbb{F}_{m\times n} \to \mathbb{F}_{s\times t}$.

(2) What can be used to measure the complexity of the matrix? The rank. Actually this is intrinsically related to the sparsity.

(3) What are the interesting statistical problems involving matrices? Multivariate regression, covariance matrix estimation, PCA.

(4) Why should we study matrices? They are the canonical representation of operators from finite dimensional vector spaces to finite dimensional ones.

(5) Singular value decomposition(SVD). An existence proof. For what kinds of matrices can we have SVD? What will SVD coincide with when the matrix is nonsingular and square?

DEFINITION 51. (norms for matrix) $|X|_0 \coloneqq |VecX|_0 = |VechX|_0$;$|X|_p \coloneqq |VecX|_p = |VechX|_p$ The Hilbert-Schmidt norm defined on the matrix space is induced by the inner product $\langle A, B\rangle \coloneqq tr(A^*B)$. Spectral norm is the $\|X\|_q \coloneqq |Vec(SpecX)|_q$, $\|X\|_2$ is the Frobenius norm, some applications can be found using complex analysis; $\|X\|_1$ is called the nuclear norm; $\|X\|_\infty$ is called the operator norm.

(6) What are the interpretations of the inequalities? Weyl's (linear) manifold principal curvature bound.
Weyl's $max_{1\le k\le min(m,n)} |\lambda_k(A) - \lambda_k(B)| \le \|A - B\|_\infty$
Hoffman-Weilandt's $\sum_{1\le k\le min(m,n)} |\lambda_k(A) - \lambda_k(B)|^2 \le \|A - B\|_2$
Holder's $\langle A, B\rangle \le \|A\|_p \|B\|_{p'}$, $\frac{1}{p} + \frac{1}{p'} = 1$
Discussion: Proof for nonsingular case. An extension to singular case using SVD.

4.2. Multivariate regression

(1) What is the multi-regression model? $\mathbb{Y} = \mathbb{X}\Theta^* + E, \mathbb{Y}_{n\times T}, \mathbb{X}_{n\times d}, \Theta_{d\times T}, E \sim$ subG$_{n\times T}(\sigma^2)$ is the random matrix with sub-gaussian random variables. An equivalent model of doing this is $Vec(\mathbb{Y}) = \mathbb{X} \otimes \mathbf{1}_{1\times T} Vec\,(\Theta^*) + Vec(E)$.
Discussion: Parallel computing design and processor programming using Vec/Vech operators.

(2) The parameter matrix shares the sparsity pattern of the design matrix.

THEOREM 52. *(Problem 4.1, the Chernoff bound for MLR) Consider the multi-regression model* $\mathbb{Y} = \mathbb{X}\Theta^* + E, \mathbb{Y}_{n\times T}, \mathbb{X}_{n\times d}, \Theta_{d\times T}, E \sim \text{subG}_{n\times T}(\sigma^2)$*, then the followings are true:*

(i) There exists an estimator $\hat{\Theta} \in \mathbb{R}^{d\times T}$ [1] *such that* $\mathbb{P}\left\{\frac{1}{n}\left\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\right\|_2^2 \lesssim \sigma^2\frac{rT}{n}\right\} \geq 1-\delta, \forall\delta > 0$

(ii) There exists an estimator $\hat{\Theta} \in \mathbb{R}^{d\times T}$ *such that* $\mathbb{P}\left\{\frac{1}{n}\left\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\right\|_2^2 \lesssim \sigma^2\frac{|\Theta^*|_0 log(ed)}{n}\right\} \geq 1-\delta, \forall\delta > 0$

PROOF. The key is that under ORT, the SLR is equivalent to subgaussian sequence model. (pp.43) We can directly apply the results by using Vec operator reducing multi-regression into SLR. $\mathbb{Y} = \mathbb{X}\Theta^* + E \iff Vec(\mathbb{Y}) = \mathbb{X}\otimes\mathbf{1}_{1\times T}Vec(\Theta^*) + Vec(E)$, by Theorem 2.2, we have

$$\frac{1}{n}\mathbb{E}\left\|\mathbb{X}\Theta^{\hat{L}S} - \mathbb{X}\Theta^*\right\|_2^2 \overset{def}{=} \frac{1}{n}\mathbb{E}\left\|\mathbb{X}\otimes\mathbf{1}_{1\times T}\widehat{Vec(\Theta)}^{LS} - \mathbb{X}\otimes\mathbf{1}_{1\times T}Vec(\Theta^*)\right\|_2^2$$

$\lesssim \sigma^2\frac{rank[(\mathbb{X}\otimes\mathbf{1}_{1\times T})^*(\mathbb{X}\otimes\mathbf{1}_{1\times T})]}{n} = \sigma^2\frac{rT}{n}$ with a probability $1-\delta$. This proves (i).

By Corollary 2.9, we have

$$\frac{1}{n}\mathbb{E}\left\|\mathbb{X}\Theta^{\hat{L}S}_{\mathcal{B}_0} - \mathbb{X}\Theta^*\right\|_2^2 \overset{def}{=} \frac{1}{n}\mathbb{E}\left\|\mathbb{X}\otimes\mathbf{1}_{1\times T}\widehat{Vec(\Theta)}^{LS}_{\mathcal{B}_0} - \mathbb{X}\otimes\mathbf{1}_{1\times T}Vec(\Theta^*)\right\|_2^2 \lesssim \sigma^2\frac{|\Theta^*|_0\, log(ed)}{n}$$

. This proves (ii). □

(3) Now we can see by observing the shared sparsity pattern, we can simply yield

$$\frac{1}{n}\mathbb{E}\left\|\mathbb{X}\otimes\mathbf{1}_{1\times T}\widehat{Vec(\Theta)}^{LS}_{\mathcal{B}_0} - \mathbb{X}\otimes\mathbf{1}_{1\times T}Vec(\Theta^*)\right\|_2^2 \lesssim \sigma^2\frac{kTlog(ed)}{n}$$

on pp.85, where k is the number of nonzero coordinates in each column of Θ^*.

(4) What is a better way to control this bound? Control the sparsity of the Θ^*(OR equivalently the rank of it), if we do not know it, just assume it.
   (a) Under the ORT assumption we have $\mathbb{Y} = \mathbb{X}\Theta^* + E \iff y = \Theta^* + F, y = \frac{1}{n}\mathbb{X}^*\mathbb{Y}, F = \frac{1}{n}\mathbb{X}^*E \sim \text{subG}_{d\times T}(\frac{\sigma^2}{n})$
   (b) When $|\Theta^*|_0$ is small, then the Theorem 52 is giving a bound if we use thresholding estimators.
   (c) When $rank(\Theta^*)$ is small, equivalently sparse in its eigen basis, we could simple estimate the singular values of Θ^* by HRD.

(5) What is the singular value thresholding? The singular value thresholding estimator with threshold $2\tau \geq 0$ id defined by $\Theta^{\hat{S}VT} := \sum_j \hat{\lambda_j}\chi_{(|\lambda_j\hat{>}2\tau|)}\hat{u_j}\hat{v_j}^*$ where $y = \sum_j \hat{\lambda_j}\hat{u_j}\hat{v_j}^*$is the SVD of observation y. We are actually using hard thresholding on $Vec(Spec(\hat{\Theta}^*))$ where $\hat{\Theta}^*$ is a moment estimator under ORT.

LEMMA 53. *Let* A *be a* $d\times T$ *random matrix s.t.* $A \sim \text{subG}_{\text{d}\times\text{T}}(\text{œ}^2)$*, Then* $\mathbb{P}\left\{\|A\|_\infty \leq 4\sigma\sqrt{log(12)(d\vee T)} + 2\sigma\sqrt{2log(\frac{1}{\delta})}\right\} \geq 1-\delta, \forall\delta > 0$

[1] This is the matrix space notation, which can also be interpret as dual space notation.

THEOREM 54. *The singular value estimator* $\Theta^{\hat{SVT}}$ *with threshold* $2\tau = 8\sigma\sqrt{\frac{log(12)(d\vee T)}{n}} + 4\sigma\sqrt{\frac{2log(\frac{1}{\delta})}{n}}$ *has oracle inequality*

$$\mathbb{P}\left\{\frac{1}{n}\mathbb{E}\left\|\mathbb{X}\Theta^{\hat{SVT}} - \mathbb{X}\Theta^*\right\|_2^2 \lesssim \frac{\sigma^2 rank(\Theta^*)}{n}\left(d\vee T + log(\frac{1}{\delta})\right)\right\} \geq 1-\delta$$

.

(6) What is penalization by rank? Putting a penalization term which reflects the increasing of $rank(\Theta^*)$ in the risk function.

$$\Theta^{\hat{RK}} := argmin_\Theta\left\{\frac{1}{n}\mathbb{E}\left\|\mathbb{Y} - \mathbb{X}\Theta\right\|_2^2 + 2\tau^2 rank(\Theta)\right\}$$

. This estimator is called estimator by rank penalization with regularization parameter τ^2

(a) The strategy of dealing with a penalized estimator is a bit different from the classical scheme. The classical estimators are constructed explicitly, and their consistency, distributions and limiting behaviors are studied. However, for penalized estimator like HRD, BIC and others we will encounter, we usually bound the risk by proving some risk bound. Sometimes we will see that this bound is related with the estimators, like regularization parameters; in other occasions the bound are just related to the dimensional of the data, which is the subject of these notes.

(b) Usual techniques are: "sup-out", "middle-man trick between unbiased and penalized", "climbing ladder technique onto a norm ball"(optimize it over a convex polytope).

(c) "It follows from the following theorem that the estimator by rank penalization enjoys the same properties as the singular value thresholding estimator even when $\mathbb{X}$ does not satisfy the the ORT. This is reminiscent of the BIC estimator which enjoys the same properties as the hard thresholding estimator."

LEMMA 55. $min_{\Theta\in\mathbb{R}^{d\times T}}\left\|\mathbb{Y} - \mathbb{X}\Theta\right\|_2^2 + 2\tau^2 rank(\Theta)$
$= min_k\left\{\frac{1}{n}min_{\Theta\in\mathbb{R}^{d\times T}, rank(\Theta)\leq k}\left\|\mathbb{Y} - \mathbb{X}\Theta\right\|_2^2 + 2\tau^2 k\right\}$
Therefore an efficient way of computing the LHS is to project onto $span(X)$ *as* $\overline{\mathbb{Y}}$.

THEOREM 56. *The rank penalized estimator* $\Theta^{\hat{RK}}$ *with threshold* $2\tau = 8\sigma\sqrt{\frac{log(12)(d\vee T)}{n}} + 4\sigma\sqrt{\frac{2log(\frac{1}{\delta})}{n}}$ *has oracle inequality*

$$\mathbb{P}\left\{\frac{1}{n}\mathbb{E}\left\|\mathbb{X}\Theta^{\hat{RK}} - \mathbb{X}\Theta^*\right\|_2^2 \lesssim \frac{\sigma^2 rank(\Theta^*)}{n}\left(d\vee T + log(\frac{1}{\delta})\right)\right\} \geq 1-\delta$$

(d) Any minimizer of $\mathbb{X}\Theta \mapsto \left\|\mathbb{Y} - \mathbb{X}\Theta\right\|_2^2$ over matrices of rank at most k can be obtained by truncating the SVD of $\overline{\mathbb{Y}}$ at order k. Actually the argument is equivalent to $A^-_{l(\mathbf{M})} = \left(A'_{m(\mathbf{M})}\right)^-$ which is directly from g-inverses.

(7) Bonus Exercise*[2]:What is the explicit form of a solution to the minimization problem of rank penalization? Try to derive it using g-inverse.

4.3. Covariance matrix estimation

(1) What is the empirical covariance matrix? $\hat{\Sigma} \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$

THEOREM 57. *Let $X_1, \cdots, X_n$ be n i.i.d sub-gaussian random vectors s.t. $\mathbb{E}[XX']$ where $X \sim \mathrm{subG_d}(\|\Sigma\|_\infty)$. Then*

$$\mathbb{P}\left\{\left\|\hat{\Sigma}-\Sigma\right\|_\infty \lesssim \|\Sigma\|_\infty \left(\sqrt{\frac{d+\log(\frac{1}{\delta})}{n}} \vee \frac{d+\log(\frac{1}{\delta})}{n}\right)\right\} \geq 1-\delta, \forall \delta > 0$$

COROLLARY 58. *Under the previous Theorem,*

$$\mathbb{P}\left\{\left|\widehat{Var(X'u)} - Var(X'u)\right| \lesssim \|\Sigma\|_\infty \left(\sqrt{\frac{d+\log(\frac{1}{\delta})}{n}} \vee \frac{d+\log(\frac{1}{\delta})}{n}\right)\right\} \geq 1-\delta, \forall \delta > 0, \forall u \in S^{d-1}$$

.

4.4. PCA and sparse PCA

(1) What is PCA? Principal analysis is a means of dimension reduction based on SVD of the estimated covariance matrix. We take the first k largest SV and take corresponding variables in the model. The data can be projected on these dimensions without great loss of generality.
(2) What is a isotopic noise? It is a random variable with homoscedasity property[3].
(3) What is a spiked covariance model? Assume $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} (v'Y_1)\, v + Z_1 \\ \vdots \\ (v'Y_n)\, v + Z_n \end{pmatrix} =$
$v' \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} v + \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}, v \in S^{d-1}, Z_i \overset{i.i.d}{\sim} N_d(0, \sigma^2 I_d)$. A linear model $Y = X\beta + \epsilon$ is called a spiked model if $Cov(Y) = \Sigma = \theta vv' + I_d, \theta > 0, v \in S^{d-1}$, which means we have a perturbation/spike in the direction v.

COROLLARY 59. *In a spiked model with covariance Σ, $\|\Sigma\|_\infty = \theta + 1$.*

(a) Exercise: Prove that for a Hermitian matrix the eigenvectors of a specific eigenvalue occur in pairs.
(b) Discussion: Due to the "occurring in pairs" behavior of eigenvectors, what is a better space of choosing spike? Real projective space.
(c) Discussion: Try to define a parallel to principal angle in the space you choosen in (b).

---

[2]I have considered a simple situation where the rank of Θ^* is $min(d,T)-1$, where the solution can be expressed using reflexive g-inverse and partitioned matrix's g-inverse. This is not very easy problem, especially that if the rank is too low, we should also take into consideration the asymmetry about the $\frac{rank(\Theta^*)}{2}$, interested reader can try to solve this problem. A known related result is given in [**Rao1**].

[3][**Cressie**]

(4) Davis-Kahan sin theorem. Note that will there be any non-p.s.d estimator for covariance? Yes, a simplest example being the error variance estimator in the mixed model. An important comment is that when we talked about the largest eigenvectors then we have admitted that we have already normalized it. (Assume that all eigenvectors with norm 1, so we can talk about their principal angles.)
I think here we can use triangle inequality since $\|X\|_1 = \left|\lambda_1 + \cdots + \lambda_{R(X)}\right| \leq R(X) \cdot max_{1\leq j\leq k} |\lambda_j| \leq R(X) \cdot \left|\sum_j \lambda_j^2\right|$, the last inequality follows from C-S.

THEOREM 60. *Let Σ satisfy the spiked covariance model and $\tilde{\Sigma}$be an p.s.d. estimator of it. Let $\tilde{v}$ denote the largest eigenvector of $\tilde{\Sigma}$, then we have*

$$min\,|\pm\tilde{v} - v|_2^2 \leq 2sin^2(\angle\tilde{v}, v) \leq \tfrac{8}{\theta^2}\left\|\tilde{\Sigma} - \Sigma\right\|_\infty^2$$

COROLLARY 61. *Let $X_1, \cdots, X_n$ be i.i.d. sub-gaussian random vectors $X \in \mathbb{R}^d$ s.t. $\mathbb{E}[XX'] = \Sigma, X \sim \text{subG}_d\left(\|\Sigma\|_\infty\right), \Sigma = \theta vv' + I_d, \theta > 0$ satisfies the spiked covariance model. Then the largest eigenvector $\hat{v}$ of the empirical estimator $\hat{\Sigma}$satisfies*

$$\mathbb{P}\left\{min\,|\pm\hat{v} - v|_2 \lesssim \tfrac{1+\theta}{\theta}\left(\sqrt{\tfrac{d+log(\frac{1}{\delta})}{n}} \vee \tfrac{d+log(\frac{1}{\delta})}{n}\right)\right\} = 1 - \delta$$

(a) This corollary is a bound for $d \ll n$ case which is known as low dimensional case.
(b) When $d > n$, this bound of minimal risk is not valid and we want to resort to sparsity in high dimension and deal with an additional logarithm term specifying the subspace.

(5) Sparse PCA
Note that the sparsity in spiked covariance model is equivalent to the sparsity in principal directions: $|v|_0 = k \leq n$. Under such a circumstance, the natural estimator of quadratic form induced by covariance is $\hat{v}'\hat{\Sigma}\hat{v} = max_{u\in S^{d-1}, |u|_0=k} u'\hat{\Sigma}u$. Note that since Davis-Kahan Theorem holds for any p.s.d. estimators, this theorem also holds for other p.s.d. estimators. Discussion: What will it yield for another estimator $\tilde{v}$ where the estimator is minimal biased instead of unbiased?

THEOREM 62. *Let $X_1, \cdots, X_n$ be i.i.d. sub-gaussian random vectors $X \in \mathbb{R}^d$ s.t. $\mathbb{E}[XX'] = \Sigma, X \sim \text{subG}_d\left(\|\Sigma\|_\infty\right), \Sigma = \theta vv' + I_d, \theta > 0$ satisfies the spiked covariance model for v such that $|v|_0 \leq \frac{d}{2}$. Then the largest eigenvector $\hat{v}$ of the empirical estimator $\hat{\Sigma}$satisfies*

$$\mathbb{P}\left\{min\,|\pm\hat{v} - v|_2 \lesssim \tfrac{1+\theta}{\theta}\left(\sqrt{\tfrac{klog(\frac{ed}{k})+log(\frac{1}{\delta})}{n}} \vee \tfrac{klog(\frac{ed}{k})+log(\frac{1}{\delta})}{n}\right)\right\} = 1 - \delta$$

CHAPTER 5

Minimax Lower Bounds

5.1. Optimality in a minimax sense

(1) Two GOALs in this chapter:
   (a) Can our estimators satisfy a better bound? Sharp bound. Optimality of an estimators.
   (b) Can any other estimators give a better bound? Optimality of estimators within a specific family of estimators.
   (c) Answers to these questions are both negative. Then we can have a discussion about why should we use these estimators, they have local optimality indeed.
(2) What is the model under consideration? $Y = \theta^* + \epsilon, \epsilon \sim N_d(0, \frac{\sigma^2}{n} I_d), \mathbb{P}_{\theta^*}, \mathbb{E}_{\theta^*}$ are the associated probability measure and expectation induced by Y. GSM is a special case of the linear regression model when the design matrix satisfy ORT.
(3) Review: The minimal variance and Cramer-Rao bound.

DEFINITION 63. (Minimax optimality) An estimator is said to be minimax optimal over Θ if it satisfies

(i)$\mathbb{E}\left[\left|\hat{\theta_n} - \theta^*\right|_2^2\right] \leq C_0\phi(\Theta), C_0 > 0$

(ii)$inf_{\hat{\theta}} sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\phi^{-1}(\Theta)\left|\hat{\theta} - \theta\right|_2^2\right] \leq C_1, C_1 > 0$

OR an alternative definition is:

An estimator is said to be minimax optimal over Θ if it satisfies

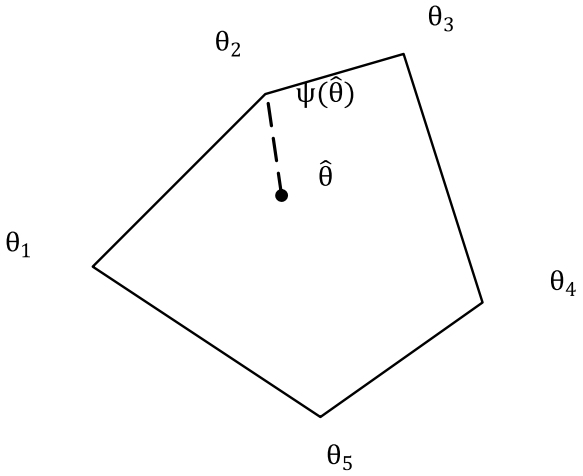
(i)$\mathbb{P}\left\{\left|\hat{\theta_n} - \theta^*\right|_2^2 \leq C_0\phi(\Theta)\right\} = 1 - \frac{1}{d^2}, C_0 > 0$

(ii)$inf_{\hat{\theta}} sup_{\theta \in \Theta} \mathbb{P}_\theta \left[\left|\hat{\theta} - \theta\right|_2^2 \geq \phi(\Theta)\right] \geq C_1, 1 \geq C_1 > 0$

5.2. Reduction to finite hypothesis testing

(1) Reduction to a finite number of hypotheses
   (a) The basic philosophy of reduction is that we can actually test a family of hypotheses and narrow down the range of possible true parameters. We use the specific hypotheses as the center of $\epsilon-$net OR extreme points of convex polytope in our argument in Chap.1. The informative theory tells that if the number of observations is too small.
   Exercise: Think why the above intuition is TRUE.

FIGURE 5.2.1. Minimum distance test



(b) To reduce the estimation problem we must find the largest possible number of hypotheses $\theta_1, \cdots, \theta_M \in \Theta$ such that$|\theta_j - \theta_k|_2^2 \geq 4\phi(\Theta)$. Then we use the $inf_{\hat{\theta}} sup_{\theta \in \Theta} \mathbb{P}_\theta \left[\left| \hat{\theta} - \theta \right|_2^2 \geq \phi(\Theta) \right] \geq inf_{\hat{\theta}} sup_{1 \leq j \leq M} \mathbb{P}_{\theta_j} \left[\left| \hat{\theta} - \theta_j \right|_2^2 \geq \phi(\Theta) \right]$.

(2) Reduction to a testing problem

(a) The minimal distance test of a specific estimator $\hat{\theta}$ is $\psi(\hat{\theta}) \coloneqq argmin_{1 \leq j \leq M} \left| \hat{\theta} - \theta_j \right|_2$

(b) $inf_{\hat{\theta}} sup_{1 \leq j \leq M} \mathbb{P}_{\theta_j} \left[\left| \hat{\theta} - \theta_j \right|_2^2 \geq \phi(\Theta) \right] \geq inf_\psi sup_{1 \leq j \leq M} \mathbb{P}_{\theta_j}[\psi \neq j]$, so it is sufficient for proving lower bounds to find $\theta_1, \cdots, \theta_M \in \Theta$ such that$|\theta_j - \theta_k|_2^2 \geq 4\phi(\Theta)$ and $inf_\psi sup_{1 \leq j \leq M} \mathbb{P}_{\theta_j}[\psi \neq j] > C_2$.

5.3. Lower bounds based on two hypotheses

(1) Neyman-Pearson Lemma

THEOREM 64. *(Neyman-Pearson's 1993, LRT is the optimal in the minimax risk sense)*

Let $\mathbb{P}_0, \mathbb{P}_1 \ll v$ be two probability measures associated with the observation random variable X, then for any test ψ, it holds that $\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) \geq \int_{\mathcal{X}} min(p_0, p_1) dv$ where equality holds for LRT $\psi^\star = \chi(p_0, p_1)$. $p_0 = \frac{d\mathbb{P}_0}{dv}, p_1 = \frac{d\mathbb{P}_1}{dv}$ are p.d.f w.r.t. the Lebesgue measure v.

(2) Total variation distance

DEFINITION 65. The total variation distance between two probability measures $\mathbb{P}_0, \mathbb{P}_1 \ll v$ on a measurable space $(\mathcal{X}, \mathcal{A}, v)$ is defined by:

$$\begin{aligned}\mathrm{TV}(\mathbb{P}_0,\mathbb{P}_1) &:= sup_{R\in\mathcal{A}}|\mathbb{P}_0(R)-\mathbb{P}_1(R)|\\ &= sup_{R\in\mathcal{A}}\int_{R\in\mathcal{A}} p_0-p_1 dv\\ &= \tfrac{1}{2}|p_0-p_1|\\ &= 1-\int_{\mathcal{A}} min(p_0,p_1)dv\\ &= 1-inf_\psi\left[\mathbb{P}_0(\psi=1)+\mathbb{P}_1(\psi=0)\right]\end{aligned}$$

(a) In view of Neyman-Pearson Lemma, we must find $\mathrm{TV}(\mathbb{P}_{\theta_i},\mathbb{P}_{\theta_j})$ as small as possible in order to maximize the lower bound for $\mathbb{P}_{\theta_i}(\psi=1)+\mathbb{P}_{\theta_j}(\psi=0)$. If we can maximize the lower bound of this "sum of two kinds of erroneous probability", then we can absolutely use LRT as an optimal test in minimax sense by the Neyman-Pearson Lemma.
(b) However, the largest possible number of hypotheses $\theta_1,\cdots,\theta_M\in\Theta$ such that$|\theta_j-\theta_k|_2^2\geq 4\phi(\Theta)$ is conflicting with our goal. We want the hypotheses as far as possible since we want to control the bound of risk; we want the hypotheses as near as possible since we want to control the bound of erroneous probabilities. The only solution is to increase the number of hypotheses to achieve both.
(c) In the most extreme cases if we test each hypotheses possible, then we find exactly the estimator should be $\hat{\theta}\equiv\theta^*$.
Discussion: what happen if we test only one hypothesis? And then compare your result with Theorem 69.

(3) Kullback-Leibler divergence

Definition 66. The Kullback-Leibler divergence between two probability measures $\mathbb{P}_0,\mathbb{P}_1\ll v$ on a measurable space $(\mathcal{X},\mathcal{A},v)$ is defined by:

$$\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_1):=\begin{cases}\int_{\mathcal{A}} log\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0}\right)d\mathbb{P}_1 & \mathbb{P}_1\ll\mathbb{P}_0\\ \infty & otherwise\end{cases}\geq 0$$

Corollary 67. $\mathrm{KL}(\otimes_i\mathbb{P}_i,\otimes_i\mathbb{Q}_i)=\sum_i KL(\mathbb{P}_i,\mathbb{Q}_i)$

Lemma 68. *(Pinsker's inequality) Let $\mathbb{P}\ll\mathbb{Q}$ be two probability measures, then* $\mathrm{TV}(\mathbb{P},\mathbb{Q})\leq\sqrt{\mathrm{KL}(\mathbb{P},\mathbb{Q})}$.

Theorem 69. *If Θ contains θ_0,θ_1 such that $|\theta_0-\theta_1|_2^2=\frac{8\alpha^2\sigma^2}{n},\alpha\in\left(0,\frac{1}{2}\right)$, then $inf_{\hat{\theta}}sup_{\theta\in\Theta}\mathbb{P}_\theta\left[\left|\hat{\theta}-\theta\right|_2^2\geq\frac{2\alpha\sigma^2}{n}\right]\geq\frac{1}{2}-\alpha$.*

5.4. Lower bounds based on many hypotheses

(1) To improve our results in Theorem 69 we need to use more than two hypotheses. In particular, in view of the above discussion, we need a set of hypotheses that spans a linear space of dimension proportional to the dimension of the parameter space. See [**Rigollet**] pp.110.
(2) Why our analysis of two hypotheses cannot directly be applied? The problem is Neyman-Pearson Lemma collapsed. But instead of the scheme $minimax\,prob\overset{Neyman-Pearson}{\to}total\,variance\overset{Pinsker}{\to}Kullback-Leibler\,div$, we use the following scheme $minimax\,prob\overset{Fano}{\to}Kullback-Leibler\,div$.

LEMMA 70. *(Fano's inequality) Let $P_1, \cdots, P_M, M \geq 2$ be a family of distributions s.t. $P_j \ll P_k$, then we have*

$$inf_\psi sup_{1\leq j\leq M} \mathbb{P}_j[\psi(X) \neq j] \geq \frac{\frac{1}{M^2}\sum_{j,k=1}^{M} \mathrm{KL}(P_j, P_k) + log2}{log(M-1)}$$

where the infimum is taken over all tests ψ with values in $\{1, \cdots, M\}$.

THEOREM 71. *If Θ contains $\theta_1, \cdots, \theta_M, M \geq 5$ such that $|\theta_i - \theta_k|_2^2 \leq \frac{2\alpha^2\sigma^2}{n} logM, \alpha \in \left(0, \frac{1}{4}\right)$, AND $|\theta_i - \theta_k|_2^2 \geq 4\phi(\Theta)$ then $inf_{\hat{\theta}} sup_{\theta\in\Theta} \mathbb{P}_\theta \left[\left|\hat{\theta} - \theta\right|_2^2 \geq \phi(\Theta)\right] \geq \frac{1}{2} - \alpha$.*

i.e. $4\phi(\Theta) \leq |\theta_i - \theta_k|_2^2 \leq \frac{2\alpha^2\sigma^2}{n}, \forall i, k \in \{1, \cdots M\}$

COROLLARY 72. *$\phi(\Theta) \leq \frac{\alpha\sigma^2}{2n} logM$ in order to let the choice of hypotheses possible.*

(3) We can use the packing model to illustrate this problem: Our goal is now pack as many balls with radius $\sqrt{\frac{\alpha\sigma^2}{2n} logM}$ into the parameter space Θ and make their centers θ as close as possible under the $|\theta|_2^2$ distances.

5.5. Applications

(1) What is Hamming distance and Hamming code?[1] The Hamming distance is the number of different coordinates between two binary vectors.

(2) Varshamov-Gilbert Stacking Lemma

LEMMA 73. *For any $\gamma \in \left(0, \frac{1}{2}\right)$, there exist binary vectors $\omega_1, \cdots, \omega_M \in \{0,1\}^d$ such that*

(i)(radius lower bound)$\rho(\omega_i, \omega_j) \geq \left(\frac{1}{2} - \gamma\right) d, \forall i \neq j$

(ii)(number lower bound)$logM = log\left\lfloor e^{\gamma^2 d}\right\rfloor \geq \frac{\gamma^2 d}{2}$

(3) Lower bounds for estimation[2]

COROLLARY 74. *The minimax rate of estimation over $\mathbb{R}^d$ in GSM is $\phi\left(\mathbb{R}^d\right) = \frac{\sigma^2 d}{n}$, which is reached by $\hat{\theta} = \theta^{\hat{LS}}$*

(4) Sparse Varshamov-Gilbert Stacking Lemma

LEMMA 75. *For any $k \in \left(1, \frac{d}{8}\right)$, there exist binary vectors $\omega_1, \cdots, \omega_M \in \{0,1\}^d, |\omega_j|_0 = k, \forall j$ such that*

(i)(radius lower bound)$\rho(\omega_i, \omega_j) \geq \frac{k}{2}, \forall i \neq j$

(ii)(number lower bound)$logM \geq \frac{k}{8} log(1 + \frac{d}{2k})$

(5) Lower bounds for sparse estimation

COROLLARY 76. *The minimax rate of estimation over $\mathcal{B}_0(k)$ in GSM is $\phi\left(\mathcal{B}_0(k)\right) = \frac{\sigma^2 k}{n} log\left(\frac{ed}{k}\right)$, which is reached by $\hat{} = \hat{\theta}^{LS}_{\mathcal{B}_\prime(\|)}$*

(6) Lower bounds for restricted estimation in $\mathcal{B}_1(R)$

---

[1] C.E.Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal 27 (July 1948), pp.379423

[2] See also [**Rao1, Rao3**], especially [**Rao1**] Chap.8 and [**Rao3**] Chap.4.

COROLLARY 77. *The minimax rate of estimation over* $\mathcal{B}_1(R) \subset \mathbb{R}^d, d \geq max(\frac{C_0 r}{\beta\sigma}\sqrt{\frac{n}{log\left(\frac{ed}{\sqrt{n}}\right)}}, n^{\frac{1}{2}+\delta}), \exists C_0, \forall \delta > 0$ *in GSM is* $\phi\left(\mathcal{B}_1(R)\right) = min\left(R^2, R\sigma\frac{logd}{n}\right)$, *which is reached by* $\hat{\theta} = \begin{cases} \theta^{L\hat{S}}_{\mathcal{B}_1(R)} & R \geq \sigma\frac{logd}{n} \\ 0 & otherwise \end{cases}$

Bibliography

[Rigollet] P.Rigollet, High Dimensional Statistics Lecture Notes, M.I.T., 2015

[Chernoff&Mose] H.Chernoff&L.E.Moses, Elementary Decision Theory, John Wiley&Sons, 1959

[Cressie] N.Cressie, Statistics for Spatial Data ,John Wiley&Sons, 1993

[McCullagh&Nelder] P.McCullagh & J.Nelder, Generalized Linear Models, 2ed, Chapman & Hall/CRC, 1991

[Dean&Voss] A.Dean & D.Voss, Design and Analysis of Experiments, Springer, 1999

[Casella&Berger] G.Casella & J.Berger, Statistical Inference, 2ed, John Wiley&Sons, 2003

[Kendall1] A.Stuart&J.K.Ord&M.Kendall, The Advanced Theory of Statistics(Volume 1: Distribution Theory), 5ed, Oxford University Press, 1987

[Kendall2] A.Stuart&J.K.Ord&M.Kendall, The Advanced Theory of Statistics(Volume 2: Classical Inference and Relationship), 5ed, Oxford University Press, 1991

[Kendall3] A.Stuart&J.K.Ord&M.Kendall, The Advanced Theory of Statistics(Volume 3: Design and Analysis, Time Series, 4ed, Hodder Arnold Publishing, 1982

[Rao1] C.R.Rao&S.K.Mitra, Generalized Inverse of Matrices and Its Applications, John Wiley&Sons, 1972

[Rao2] C.R.Rao et.al, Linear Models: Least Squares and Alternatives, 2ed, Springer, 1999

[Rao3] C.R.Rao, Linear Statistical Inference and its Applications, 2ed, John Wiley&Sons, 2001

[Ravishanker&Dey] N.Ravishanker&D.K.Dey, A First Course in Linear Model Theory, CRC Press, 2001